Machine Learning Multi-Stage Classification and Regression in the Search for Vector-like Quarks and the Neyman Construction in Signal Searches

by

Robert Matthew Leone

Copyright © Robert Matthew Leone 2016

CERN-THESIS-2016-229
28/11/2016

A Dissertation Submitted to the Faculty of the

Department of Physics

In Partial Fulfillment of the Requirements
For the Degree of

Doctor of Philosophy
In the Graduate College

The University of Arizona

2 0 1 6

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Robert Matthew Leone entitled Machine Learning Multi-Stage Classification and Regression in the Search for Vector-like Quarks and the Neyman Construction in Signal Searches and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

Date: November 28, 2016

Elliott C. Cheu

Date: November 28, 2016

Kenneth A. Johns

Date: November 28, 2016

Erich W. Varnes

Date: November 28, 2016

Sean P. Fleming

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Date: November 28, 2016

Dissertation Director: Elliott C. Cheu

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

Signed: Robert Matthew Leone

Acknowledgements

Experimental high energy physics is a collaborative field, and I am grateful to the many people and institutions who contributed to the completion of this research.

Professor Elliott Cheu was an exemplary advisor, providing guidance and resources that allowed for successful completion of this work. Elliott's inquisitive nature allowed for the exploration of novel ideas and methods, and I am most appreciative of his practical and critical feedback that contributed to the strength of the final results.

My other dissertation committee members, Professors Ken Johns, Erich Varnes and Sean Fleming, provided valuable critiques of the dissertation drafts that greatly improved its final quality. Ken and Erich also offered considerable guidance during the initial stages of this research.

My Z-tag search team collaborators were essential for the successful completion of the square cuts search that led to publication after two years of difficult work. Dr. Mark Cooke and Dr. Nuno Castro steered us to success, each generously providing weekly guidance. Dr. Juan Pedro Araque Espinosa provided slimmed data samples for our group that were also used in this dissertation work for the machine learning search. I am also grateful to many other members of the ATLAS experiment, too many to name individually, whose work contributed to the success of this research.

Other members of the Experimental Elementary Particle Physics group at the University of Arizona provided additional guidance for this research. Professors John Rutherford and Michael Shupe, and Dr. Peter Loch were always available to share their considerable knowledge. Dr. Caleb Lampen and Dr. Venkat Kaushik provided my initial training and later provided guidance and feedback during the early stages of this work. I am especially grateful to Caleb for introducing me to machine learning. Dr. Xiaowen Lei was helpful in providing technical guidance on occasion, and also collaborated on the estimation of fakes for this work.

The research of chapters 3 and 4 was supported in part by the Department of Energy.

Table of Contents

List of Figures **10**

List of Tables **18**

Abstract **22**

Chapter 1. Introduction **23**

Chapter 2. Theoretical Motivation for Vector-like Quarks Search . . **26**
2.1. The Standard Model 26
2.1.1. Renormalizable Relativistic Quantum Field Theory 27
2.1.2. Gauge Symmetries and Vector Bosons 30
2.1.3. Spontaneous Symmetry Breaking & Brout-Englert-Higgs Mechanism 32
2.1.4. Glashow-Weinberg-Salam Electroweak Theory 34
2.1.5. Quantum Chromodynamics 37
2.1.6. Problems with the Standard Model 40
2.2. Vector-like Quark Theories 42
2.2.1. Other Searches for Vector-like Quarks 49

Chapter 3. Data Collection **51**
3.1. Introduction to Data Collection 51
3.2. Creating p–p Collisions with the Large Hadron Collider 53
3.3. Observing p–p Collisions in the ATLAS Experiment 57
3.3.1. Tracking Charged Particles with the Inner Detector 58
3.3.2. The Electromagnetic, Hadronic and Forward Calorimeters 59
3.3.3. The Toroid Magnet System and the Muon Spectrometer 62
3.3.4. Triggering and data storage 64
3.4. Collection of the Measured Dataset 66
3.4.1. Determination of the Collision Events Primary Vertex 66
3.4.2. Determination of Electron Candidates 67
3.4.3. Determination of Muon Candidates 68
3.4.4. Determination of Hadronic Jets 69
3.4.5. Identification of Jets Containing b-quarks — b-tagging 69
3.4.6. Event Preselection 70
3.5. Construction of the Simulated Datasets 71
3.5.1. Simulated Backgrounds 72
3.5.2. Simulated Signals 74

Chapter 4. The Square Cuts Search **76**
4.1. Square Cuts Search Strategy 77

4.2. Background Corrections . . . 84
4.2.1. Estimation of the QCD Background . . . 84
4.2.2. Corrections to Simulated Data. . . . 89
4.3. Systematic Uncertainties . . . 94
4.4. Validation of the Simulated Backgrounds . . . 98
4.5. Unblinded Signal Regions . . . 101
4.6. Results of the Square Cuts Search Strategy . . . 103

Chapter 5. The Search Using Machine Learning . . . 106
5.1. Introduction to the Search Using Machine Learning . . . 106
5.2. Methodology of Machine Learning . . . 110
5.2.1. Learning Algorithms . . . 110
5.2.2. Boosted Decision Trees . . . 112
5.2.3. Multi-layer Perceptron . . . 116
5.2.4. Least Squares Linear Regression . . . 118
5.2.5. Logistic Regression (Logit Classification) . . . 120
5.3. Machine Learning in Other Particle Searches . . . 121
5.4. Data Partitioning for Machine Learning . . . 124
5.5. Stage 1 Classification with Boosted Decision Trees . . . 126
5.5.1. Selection of the Input Variables . . . 126
5.5.2. Optimization of stage 1 Boosted Decision Trees . . . 135
5.5.2.1. Automated Maximization of Signal Significance . . . 138
5.5.2.2. Automated Determination of Overtraining . . . 140
5.5.2.3. Grid Search for Optimal BDT Hyper-parameters . . . 143
5.5.2.4. Selected stage 1 BDT Classifiers . . . 148
5.6. Stage 2 Classification with Logistic Regression . . . 153
5.6.1. Selection of Logit Classification for stage 2 . . . 156
5.6.2. Logit Classification Results . . . 158
5.7. Stage 3 Machine Learning Regression . . . 170
5.7.1. Linear Regression *(selected for T_{pair}, B_{pair}, and B_{single})* . . . 172
5.7.2. Regression with Boosted Decision Trees *(selected for T_{single})* . . . 182
5.8. Modeling Validation in Control Regions . . . 187
5.8.1. Validation of stage 1 Input Variables . . . 188
5.8.2. Validation of stage 1 Boosted Decision Tree Classification . . . 194
5.8.3. Validation of stage 2 Additional Input Variables . . . 195
5.8.4. Validation of stage 2 Logit Classification . . . 197
5.8.5. Validation of stage 3 Machine Learning Regression . . . 200
5.9. Unblinded Signal Regions – Revealing the Measured Data . . . 202
5.9.1. Unblinded Results . . . 205
5.10. Statistical Inference – Setting CL_S Upper Limits . . . 215
5.10.1. Examination of Systematic Uncertainties . . . 216
5.10.2. Final Results – CL_S Upper Limits . . . 223
5.10.2.1. Square Cuts Search Results and Measuring Search Power 224
5.10.2.2. Results of the Two-stage Machine Learning Search . . . 227
5.10.2.3. Results of the Three-stage Machine Learning Search . . . 232

5.10.3. Discussion of Machine Learning Results 235
5.11. Machine Learning – Discussion 246
5.11.1. Implications for Future Research 247
5.11.1.1. Two-stage Classification 247
5.11.1.2. Machine Learning Regression 247
5.11.1.3. Unsupervised Learning 248

Chapter 6. The Neyman Construction in Signal Searches 249
6.1. A Conceptual Outline of this Chapter 252
6.2. Introduction to Inference for Signal Searches 254
6.2.1. The Multivariate Poisson Statistical Model 258
6.2.2. Neyman Confidence Region Construction 261
6.2.2.1. Example 1: Construction for the Univariate Poisson . . . 263
6.2.2.2. Example 2: Multivariate Poisson Global Ordering 266
6.2.3. Determining Acceptance Regions using Test Statistics 276
6.2.3.1. Neyman-Pearson Hypothesis Testing 279
6.2.3.2. Test Statistics Used as Local Ordering Rules 282
6.2.3.3. A Review of Pivots 283
6.2.3.4. Common Test Statistics in Signal Searches 285
6.2.4. Discussion of the CL_S Method 295
6.3. The Arizona Test Statistics 298
6.3.1. Motivation for the Arizona Test Statistics 299
6.3.2. The Test Statistic Modification Functions, Q_{inc} and Q_{rep} 309
6.3.3. Defining the Arizona Test Statistics 312
6.3.4. Performance Metrics of the Neyman Construction 316
6.3.5. Additional Examples 328
6.3.5.1. Comparison of the Acceptance Ventiles 330
6.3.5.2. Comparison of the Performance Metrics 336
6.4. Inference using Confidence-Acceptance Profiles 341
6.4.1. Motivation for Confidence-Acceptance Profiles 343
6.4.2. The Signal Confidence-Acceptance Profiles 346
6.4.3. The Background Confidence-Acceptance Profile 349
6.4.4. Relating the Three methods: FST, NPHT and NCRC 351
6.4.5. The Background Significance Ratio Function 352
6.4.6. The Signal Significance Ratio Function 357
6.4.6.1. The SSR Function and the CL_S Confidence Interval . . . 359
6.4.6.2. Combining the BSR and SSR Functions 361
6.4.7. Frequentist Interpretation of Significance Ratio Functions 363
6.4.8. The Utility of Statistical Eduction via Confidence 369
6.4.9. Combining Confidence-Acceptance Profiles via Meta-analysis . . . 372
6.4.9.1. Combining CAPs: Repeated Trials 375
6.4.9.2. Combining CAPs: Comparison to Ideal 383
6.5. Test Statistics in the Presence of Nuisance Parameters 388
6.5.1. Defining the Test Statistics Including Nuisance Parameters 392
6.5.2. Effect of Nuisance Parameters on Likelihood functions 398

6.5.3. Effect of Nuisance Parameters on Acceptance Regions 402
6.5.4. Effect of Nuisance Parameters on Confidence-Acceptance Profiles . 406
6.5.5. Effect of Nuisance Parameters on Coverage 410
6.5.5.1. Coverage Properties of Q^N_{FC} 413
6.5.5.2. Coverage Properties of Q^N_{FC} with the CL_S Method 415
6.5.5.3. Coverage Properties of Q^P_{FC} and Q^F_{FC} 417
6.5.5.4. Coverage Properties of Q^N_{D0}, Q^P_{D0} and Q^F_{D0} 419
6.5.5.5. Summary of Coverage Properties 421
6.6. Conclusion to Statistical Inference in a Signal Search 423
6.6.1. Classical Inference to Explain, After Machine Learning Predicts . . 423
6.6.2. Selection of Test Statistics 424
6.6.3. Confidence-Acceptance Profiles and Frequentist Evidential Support 426
6.6.4. Bayesian Methods for Nuisance Parameter Treatment 427

Chapter 7. Conclusion 428
7.1. Multi-stage Machine Learning 428
7.2. Other Machine Learning Results 431
7.3. The Neyman Construction in Signal Searches 432

Appendix A. The Partition Trees Learning Algorithm 433
A.1. Partition Trees 434
A.2. Partition Tree Ensembles and Semi-supervised Learning 438
A.3. Performance of Random Forests of Partition Trees 440
A.3.1. PTLA Performance with the VLQ Dataset 440
A.3.2. PTLA Performance with the Higgs Dataset 442
A.3.3. PTLA Performance with the Cancer Dataset 447
A.4. Instance-based Learning 456
A.5. Summary of the Partition Trees Learning Algorithm 460

Appendix B. Other Machine Learning Results 461
B.1. The Cross-Evaluation Partitioning Scheme 461
B.2. Unsupervised Machine Learning to Identify Signal Regions 462
B.3. Decision Tree Regression: Gradient vs. Adaptive Boosting 470

Appendix C. The Search Using Machine Learning (Addendum) 476
C.1. Stage 2 Classification with Other Methods 476
C.1.1. Support Vector Machines 476
C.1.2. Multilayer Perceptrons 476
C.1.3. Boosted Decision Trees 477
C.2. Stage 3 Linear Regression: All Tables 482
C.3. Blinded Modeling Validation: All Stage 1 Classifiers 496
C.4. Final Results – CL_S Upper Limits (Addendum) 512

Appendix D. The Neyman Construction in Signal Searches (Addendum) 527
D.1. The Arizona Test Statistics (Addendum) 528

D.1.1. Incorporating the Model Inconsistency Measure, Q_{inc} 528
D.1.2. Reducing Reporting Risk with Q_{rep} 533
D.1.3. Additional Examples: The Most Probable Test Statistic, Q_{MP} . . . 536
D.1.4. Additional Examples: The CL_S Method 538
D.2. Confidence-Acceptance Profile Properties (Addendum) 541
D.3. Additional Nuisance Parameter Results (Addendum) 544
D.3.1. Effect of Nuisance Parameters on Acceptance Regions 544
D.3.2. Effect of Nuisance Parameters on Confidence-Acceptance Profiles . 544
D.3.3. Effect of Nuisance Parameters on Coverage 544

References **558**

List of Figures

Chapter 1. Introduction

Figure 1.1. Organization of dissertation 25

Section 2.1. The Standard Model

Figure 2.1. Fundamental fermions of the Standard Model 28
Figure 2.2. Bosons of the Standard Model 29
Figure 2.3. Example Feynman diagrams 30
Figure 2.4. Feynman diagram for free neutron decay 32
Figure 2.5. Feynman diagrams illustrating generation mixing of quarks 38
Figure 2.6. Feynman diagrams illustrating gluon-gluon interactions 39
Figure 2.7. The hadronization of quarks and gluons 41

Section 2.2. Vector-like Quark Theories

Figure 2.8. Higgs production via gluon fusion 43
Figure 2.9. Decay modes for vector-like quarks 44
Figure 2.10. Vector-like quark production mechanisms 45
Figure 2.11. Vector-like top decay topologies 47
Figure 2.12. Vector-like bottom decay topologies 48

Chapter 3. Data Collection

Figure 3.2. VLQ decays producing τ leptons and neutrinos 52

Section 3.2. Creating p–p Collisions with the Large Hadron Collider

Figure 3.3. The CERN accelerator complex 55

Section 3.3. Observing p–p Collisions in the ATLAS Experiment

Figure 3.4. Observing physics objects in hadronic t decay 57

Section 3.5. Construction of the Simulated Datasets

Figure 3.5. Example Z+*jets* production 72
Figure 3.6. Example *diboson*, *triboson* and $t\bar{t}X$ production 73
Figure 3.7. Example $t\bar{t}$ production 74
Figure 3.8. Example *single top* and W+*jets* production 75

Section 4.1. Square Cuts Search Strategy

Figure 4.1. Selection of mass window and jet multiplicity requirements 77
Figure 4.2. Selection of $p_T(Z)$ and $N(b\text{-}tags)$ square cuts 79
Figure 4.3. Selection of $H_T(jets)$ and $N(forward\ jets)$ square cuts 80
Figure 4.4. T_{single} and B_{single} decay topologies 80
Figure 4.5. Selection of $m(Zb)$ as the discriminating variable 83
Figure 4.6. B_{single} and T_{pair} decay topologies 83

Section 4.2. Background Corrections

Figure 4.7. The $m(Z)$ distribution of QCD sample events after preselection . . 86
Figure 4.8. QCD background estimates in the lower side band after preselection 86
Figure 4.9. QCD background estimates after preselection 87
Figure 4.10. Distributions of QCD sample events for progressive square cuts . . 88
Figure 4.11. $N(b\text{-}tags)$ after square cuts 0–2 89
Figure 4.12. The simulated $t\bar{t}$ modeling of $m(Z)$ and $p_T(Z)$. 90
Figure 4.13. $N(b\text{-}tags)$ and $p_T(Z)$ before and after corrections 92
Figure 4.14. $N(b\text{-}tags)$ and $p_T(Z)$ before and after corrections 93

Section 4.4. Validation of the Simulated Backgrounds

Figure 4.15. Validation after square cuts 0–3 and $N(b\text{-}tags) = 1$ 99
Figure 4.16. Validation after square cuts 0–4 and inverted cuts 5a and 5b 100

Section 4.5. Unblinded Signal Regions

Figure 4.17. The unblinded $H_T(jets)$, $N(forward\ jets)$ and $m(Zb)$ variables . . 102
Figure 4.18. VLQ upper limits 104

Section 5.1. Introduction to the Search Using Machine Learning

Figure 5.1. Difficulty of binary classification in the VLQ search 108

Section 5.2. Methodology of Machine Learning

Figure 5.2. Example of a decision tree 112
Figure 5.3. The square cuts search depicted as a decision tree 113
Figure 5.4. Example of a multi-layer perceptron 116

Section 5.5. Stage 1 Classification with Boosted Decision Trees

Figure 5.5. Stage 1 input variable histograms 130
Figure 5.6. Correlations between input variables 134
Figure 5.7. BDT classifier response value distributions – $m(T_{pair}) = 650$ GeV/c^2 136
Figure 5.8. Predicted yields and ROC curve – $m(T_{pair}) = 650$ GeV/c^2 137
Figure 5.9. Significance vs. cut on classifier value – $m(T_{pair}) = 650$ GeV/c^2 . . 139
Figure 5.10. Brown's & Fisher's methods for combining tests 140
Figure 5.11. Brown-Fisher p-values – $m(T_{pair}) = 650$ GeV/c^2 145
Figure 5.12. BDT classifier response distribution – $m(B_{single}) = 500$ GeV/c^2 . . 145
Figure 5.13. Decision node variable frequencies – $m(T_{pair}) = 650$ GeV/c^2 146

Figure 5.14. Parent-child node variable pairing – $m(T_{pair}) = 650$ GeV/c^2 147
Figure 5.15. Optimized BDT hyper-parameter frequencies 151
Figure 5.16. Signal significance by signal mass 152

Section 5.6. Stage 2 Classification with Logistic Regression

Figure 5.17. Two-stage classification strategy 154
Figure 5.18. Histograms of additional stage 2 variables – T_{pair} 155
Figure 5.19. $N(forward\ jets)$ – T_{single} and B_{single} 157
Figure 5.20. Logit classification neural network 159
Figure 5.21. stage 2 logit classification – T_{pair} 160
Figure 5.22. stage 2 logit classification – B_{pair} 161
Figure 5.23. stage 2 logit classification – T_{single} 162
Figure 5.24. stage 2 logit classification – B_{single} 163
Figure 5.25. stage 2 signal separation performance – T_{pair} 164
Figure 5.26. stage 2 signal separation performance – B_{pair} 165
Figure 5.27. stage 2 signal separation performance – T_{single} 166
Figure 5.28. stage 2 signal separation performance – B_{single} 167

Section 5.7. Stage 3 Machine Learning Regression

Figure 5.29. Stage 3 $m(Q)$ OLS regression spectra – T_{pair} 178
Figure 5.30. Stage 3 $m(Q)$ OLS regression spectra – B_{pair} 179
Figure 5.31. Stage 3 $m(Q)$ OLS regression spectra – B_{single} 180
Figure 5.32. Stage 3 $m(Q)$ OLS regression performance – T_{pair}, B_{pair} & B_{single} . 181
Figure 5.33. Stage 3 BDT regression – T_{single} 183
Figure 5.34. Stage 3 BDT regression results – T_{single} 185
Figure 5.35. Stage 3 BDT regression performance – T_{single} 186

Section 5.8. Modeling Validation in Control Regions

Figure 5.36. Validation of simulated data – stage 1 input variables 188
Figure 5.37. Validation of simulated data – stage 1 650 GeV/c^2 BDT classifiers 194
Figure 5.38. Validation of simulated data – stage 2 additional features 195
Figure 5.39. Validation of simulated data – reconstructed mass spectra 196
Figure 5.40. Validation of simulated data – stage 2 logit classifiers 198
Figure 5.41. Validation of simulated data – stage 2 T_{single} logit classifier 199
Figure 5.42. Validation of simulated data – stage 3 regression 201

Section 5.9. Unblinded Signal Regions – Revealing the Measured Data

Figure 5.43. Unblinded square cuts $m(Zb)$ spectra 206
Figure 5.44. Unblinded stage 2 logit classifiers 208
Figure 5.45. Correlation of machine learning and square cuts signal regions . . . 210
Figure 5.46. Unblinded machine learning $m(Zb)$ spectra 211
Figure 5.47. Unblinded machine learning $m(Zt)$ spectra 212
Figure 5.48. Unblinded machine learning stage 3 regressand mass spectra 214

Section 5.10. Statistical Inference – Setting CL_S Upper Limits

Figure 5.49. Example of the asymmetric Gaussian-like distribution 216
Figure 5.50. Concrete example of functions of systematics 218
Figure 5.51. CL_S upper limits – square cuts search strategy 225
Figure 5.52. 95% CL_S upper limits of two hypothetical searches 226
Figure 5.53. Ratios of expected upper limits: two-stage machine learning 228
Figure 5.54. Ratios of upper limits uncertainties: two-stage machine learning . . 229
Figure 5.55. Ratios: two-stage machine learning with $m(Zt)$ 231
Figure 5.56. Ratios of expected upper limits: three-stage machine learning . . . 232
Figure 5.57. Ratios of upper limits uncertainties: three-stage machine learning . 233
Figure 5.59. 650 GeV/c^2 B_{single} stage 1 decision tree node frequencies 236
Figure 5.60. Vector-like quark decay topologies 237
Figure 5.61. Signal significances: stage 2 vs. square cuts 238
Figure 5.62. Signal significances: stage 3 vs. square cuts 239
Figure 5.63. Comparison of discriminating variables: T_{pair} 241
Figure 5.64. Comparison of discriminating variables: B_{pair} 242
Figure 5.65. Comparison of discriminating variables: T_{single} 243
Figure 5.66. Comparison of discriminating variables: B_{single} 244

Section 6.2. Introduction to Inference for Signal Searches

Figure 6.1. Example 1: The univariate Poisson model 264
Figure 6.2. Example 2: Model properties and global ordering 268
Figure 6.3. Example 2: Selected probability distributions 270
Figure 6.4. Example 2: Neyman construction using global ordering 272
Figure 6.5. Example 2: Probability distributions 273
Figure 6.6. Example 2: Selected likelihood functions 274
Figure 6.7. Example 2: Likelihood functions 275
Figure 6.8. Using a local ordering rule to find an acceptance region: example . 277
Figure 6.9. Acceptance/confidence regions delineated by $\phi^{\alpha}_{rej}(d, \mu)$: example . . 278
Figure 6.10. Using a test statistic to find a critical region: example 280
Figure 6.11. Equivalence of test inversion and Neyman construction: example . 284
Figure 6.12. Example 2: Neyman construction using Q_{D0} 288
Figure 6.13. Example 2: $Q_{D0}(d, \mu')$ spectra for selected μ' 290
Figure 6.14. Example 2: Neyman construction using Q_{FC} 292
Figure 6.15. Example 2: Neyman construction using Q_{AT} 293
Figure 6.16. Example 2: Neyman construction using Q_{MP} 294
Figure 6.17. Example 2: Neyman construction using Q_{D0}+CL_S 297

Section 6.3. The Arizona Test Statistics

Figure 6.18. Desirable test statistic criteria 300
Figure 6.19. Example 2: Motivation for the Arizona test statistics 302
Figure 6.20. Example 2: Model probabilities for $\mu = 1$ 307
Figure 6.21. Example 2: Acceptance regions of test statistics for $\mu = 1$ 308

Figure 6.22. Example 2: Undesirable features using $Q = Q_{inc} \cdot Q_{D0}$ 310
Figure 6.23. Example 2: Comparing acceptance ventiles of test statistics 313
Figure 6.24. Example 2: Comparing acceptance ventiles of test statistics+CL_S . 315
Figure 6.25. Example 3a: Comparing acceptance ventiles of test statistics . . . 333
Figure 6.26. Example 3b: Comparing acceptance ventiles of test statistics . . . 334
Figure 6.27. Example 3c: Comparing acceptance ventiles of test statistics 335

Section 6.4. Inference using Confidence-Acceptance Profiles

Figure 6.28. 90% confidence intervals of a hypothetical signal search 343
Figure 6.29. 90% confidence intervals of three hypothetical signal searches . . . 345
Figure 6.30. Hypothetical signal search acceptance deciles 345
Figure 6.31. Observed signal confidence-acceptance profiles 348
Figure 6.32. Test statistic distributions 349
Figure 6.33. Observed signal search CAPs, $S(d_{obs}, \mu)$ and $B(d_{obs}, \mu)$ 350
Figure 6.34. Test statistic distributions of three experiments 352
Figure 6.35. Test statistic cdfs of three experiments 352
Figure 6.36. SCAP probability distributions 353
Figure 6.37. The background significance ratio values for $\mu = 1$ 354
Figure 6.38. The observed background significance ratio functions 356
Figure 6.39. The signal significance ratio values for $\mu = 1$ 357
Figure 6.40. Observed signal significance ratio functions 358
Figure 6.41. 70% CL_S confidence intervals 360
Figure 6.42. Observed significance ratio functions 361
Figure 6.43. Significance ratios with low or high signal-background discrimination 362
Figure 6.44. Fisher significance test for the SSR function at μ' 365
Figure 6.45. Fisher significance test for the SSR function (all μ) 366
Figure 6.46. Fisher significance test for the BSR function at μ' 367
Figure 6.47. Fisher significance test for the BSR and SSR functions (all μ) . . . 367
Figure 6.48. Eduction in the scientific process 370
Figure 6.49. Two experiment meta-analysis probability distributions 373
Figure 6.50. Meta-analysis: CAPs for Poisson and Gaussian 376
Figure 6.51. Meta-analysis: repeated sampling Poisson model, Q_{FC} and Q_{AZ} . . 378
Figure 6.52. Meta-analysis: repeated sampling Gaussian model, Q_{FC} and Q_{AZ} . 380
Figure 6.53. Meta-analysis: repeated sampling Poisson and Gaussian, Q_{D0} . . . 382
Figure 6.54. Meta-analysis: combining 3 experiments for Poisson using Q_{FC} . . 384
Figure 6.55. Meta-analysis: combining 3 experiments for Gaussian using Q_{FC} . 386
Figure 6.56. Meta-analysis: combining 3 experiments for Poisson using Q_{D0} . . 387

Section 6.5. Test Statistics in the Presence of Nuisance Parameters

Figure 6.57. Frequentist projection due to a nuisance parameter 389
Figure 6.58. One dimensional acceptance regions via marginalization: example . 390
Figure 6.59. One dimensional confidence regions via marginalization: example . 391
Figure 6.60. Example 2: Profiled likelihood functions 400
Figure 6.61. Example 3b: Profiled likelihood functions with nuisance parameters 401

Figure 6.62. Example 3b: $\mu = 1.0$ acceptance regions for Q_{D0} 404
Figure 6.63. Example 2: Confidence-acceptance profiles (common) 408
Figure 6.64. Example 2: Confidence-acceptance profiles (Arizona) 409
Figure 6.65. Example 2: 2σ coverage study for Q^N_{FC} 414
Figure 6.66. Example 2: 2σ coverage study for Q^N_{FC} with $\mathrm{CL_S}$ method 416
Figure 6.67. Example 2: 2σ coverage study for Q^P_{FC} and Q^F_{FC} 418
Figure 6.68. Example 2: 2σ coverage study for Q^N_{D0}, Q^P_{D0} and Q^F_{D0} 420

Section 7.1. Multi-stage Machine Learning
Figure 7.1. Machine learning methodologies 430

Section A.1. Partition Trees
Figure A.1. Partitioning of training data features 434
Figure A.2. Maximum splitting of a single feature partition tree 435
Figure A.3. Selection of the splitting feature for subnode construction 436
Figure A.4. Pseudocode for the construction of a partition tree 437

Section A.2. Partition Tree Ensembles and Semi-supervised Learning
Figure A.5. Semi-supervised learning used in node splitting 439
Figure A.6. Performance of the PTLA ensemble using the vlq dataset 441
Figure A.7. A partition tree created using the vlq dataset 443
Figure A.8. The maximum P-levels of a partition tree (vlq dataset) 444
Figure A.9. Distributions of the PTLA ensemble using the vlq dataset 445
Figure A.10. Performance of the PTLA ensemble using the higgs dataset 446
Figure A.11. A partition tree created using the higgs dataset 448
Figure A.12. The maximum P-levels of a partition tree (higgs dataset) 449
Figure A.13. Distributions of the first partition tree using the higgs dataset . . . 450
Figure A.14. Distributions of the PTLA ensemble using the higgs dataset 451
Figure A.15. Performance of the PTLA ensemble using the cancer dataset . . . 452
Figure A.16. A partition tree created from the cancer dataset 453
Figure A.17. The maximum P-levels of a partition tree (cancer dataset) 454
Figure A.18. Distributions of the PTLA ensemble using the cancer dataset . . . 455

Section A.4. Instance-based Learning
Figure A.19. A partition tree viewed as a distance metric 456
Figure A.20. Various machine learning methodologies 457
Figure A.21. Instance-based scoring 458
Figure A.22. Performance of the cancer dataset 459

Section B.2. Unsupervised Machine Learning to Identify Signal Regions
Figure B.1. K-means unsupervised learning 464

Section B.3. Decision Tree Regression: Gradient vs. Adaptive Boosting
Figure B.2. Stage 3 adaptive BDT regression spectra – T_{single} 471
Figure B.3. Linear regression with adaptive boosting over time 472
Figure B.4. Linear regression with adaptive boosting analysis 473
Figure B.5. Linear regression with gradient boosting over time 474
Figure B.6. Linear regression with gradient boosting analysis 475

Section C.1. Stage 2 Classification with Other Methods
Figure C.1. Stage 2 classification using support vector machines 478
Figure C.2. Stage 2 classification using multilayer perceptrons 479
Figure C.3. Stage 2 MLP neural network 480
Figure C.4. Stage 2 classification using boosted decision trees 481

Section C.3. Blinded Modeling Validation: All Stage 1 Classifiers
Figure C.5. Validation of simulated data – stage 1 classifiers 496

Section C.4. Final Results – CL_S Upper Limits (Addendum)
Figure C.6. T_{pair} square cuts 95% CL_S upper limits 514
Figure C.7. T_{pair} stage 2 machine learning 95% CL_S upper limits 515
Figure C.8. T_{pair} stage 3 machine learning 95% CL_S upper limits 516
Figure C.9. B_{pair} square cuts 95% CL_S upper limits 517
Figure C.10. B_{pair} stage 2 machine learning 95% CL_S upper limits 518
Figure C.11. B_{pair} stage 3 machine learning 95% CL_S upper limits 519
Figure C.12. T_{single} square cuts 95% CL_S upper limits 520
Figure C.13. T_{single} stage 2 machine learning 95% CL_S upper limits ($m(Zb)$) . . 521
Figure C.14. T_{single} stage 2 machine learning 95% CL_S upper limits ($m(Zt)$) . . 522
Figure C.15. T_{single} stage 3 machine learning 95% CL_S upper limits 523
Figure C.16. B_{single} square cuts 95% CL_S upper limits 524
Figure C.17. B_{single} stage 2 machine learning 95% CL_S upper limits 525
Figure C.18. B_{single} stage 3 machine learning 95% CL_S upper limits 526

Section D.1. The Arizona Test Statistics (Addendum)
Figure D.1. Example 2: Properties of simple likelihood ratios 530
Figure D.2. Example 2: Properties of other likelihood ratios 532
Figure D.3. Example 2: Neyman construction using $Q_{rep} = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}$ 533
Figure D.4. Example 2: Neyman construction using $Q_{rep} \cdot Q_{AT}$ 535
Figure D.5. All examples: Neyman construction using the Q_{MP} 537
Figure D.6. Example 3a: Comparing acceptance ventiles of test statistics+CL_S 538
Figure D.7. Example 3b: Comparing acceptance ventiles of test statistics+CL_S 539
Figure D.8. Example 3c: Comparing acceptance ventiles of test statistics+CL_S 540

Section D.3. Additional Nuisance Parameter Results (Addendum)

Figure D.9. Example 3b: $\mu = 1.0$ acceptance regions for Q_{FC} 545
Figure D.10. Example 3b: $\mu = 1.0$ acceptance regions for Q_{AT} 546
Figure D.11. Example 3b: $\mu = 1.0$ acceptance regions for Q_{AZ1} 547
Figure D.12. Example 3b: $\mu = 1.0$ acceptance regions for Q_{AZ2} 548
Figure D.13. Example 3b: $\mu = 1.0$ acceptance regions for Q_{AZ3} 549
Figure D.14. Example 2: Confidence-acceptance profiles (common+CL_S method) 550
Figure D.15. Example 2: Confidence-acceptance profiles (Arizona+CL_S method) 551
Figure D.16. Example 2: 2σ coverage study for Q^X_{FC} 552
Figure D.17. Example 2: 2σ coverage study for Q^X_{D0} 553
Figure D.18. Example 2: 2σ coverage study for Q^X_{AT} 554
Figure D.19. Example 2: 2σ coverage study for Q^X_{AZ1} 555
Figure D.20. Example 2: 2σ coverage study for Q^X_{AZ2} 556
Figure D.21. Example 2: 2σ coverage study for Q^X_{AZ3} 557

List of Tables

Section 2.1. The Standard Model
Table 2.1. Particle fields before SSB in the GWS model 35
Table 2.2. Standard Model particle masses after SSB 36

Section 2.2. Vector-like Quark Theories
Table 2.3. Results of other searches for vector-like quarks 50

Section 3.3. Observing p–p Collisions in the ATLAS Experiment
Table 3.1. Strategy for detecting physics objects 58
Table 3.2. Parameters of the calorimeters 61
Table 3.3. The main parameters of the Muon Spectrometer 63

Section 3.4. Collection of the Measured Dataset
Table 3.4. Data measured in 2012 67

Section 4.1. Square Cuts Search Strategy
Table 4.1. The two signal regions of the square cuts search strategy 81
Table 4.2. List of hypotheses to test 82

Section 4.2. Background Corrections
Table 4.3. Estimated QCD background yields 88

Section 4.3. Systematic Uncertainties
Table 4.4. Summary of systematic uncertainties 96
Table 4.5. Estimated systematic uncertainties 97

Section 4.5. Unblinded Signal Regions
Table 4.6. Total yields of the square cuts signal regions 101

Section 5.1. Introduction to the Search Using Machine Learning
Table 5.1. List of hypotheses to test 107

Section 5.3. Machine Learning in Other Particle Searches
Table 5.2. Machine learning methods used in other research 123

Section 5.4. Data Partitioning for Machine Learning
Table 5.3. Machine learning data partitioning schemes in particle searches . . . 125

Section 5.5. Stage 1 Classification with Boosted Decision Trees
Table 5.4. Categorization of input variable information . . . 128
Table 5.5. Stage 1 classification data preselection . . . 129
Table 5.6. Correlation matrix for Brown's method . . . 142
Table 5.7. Hyper-parameter options for the stage 1 BDT grid search . . . 143
Table 5.8. Summary of stage 1 classification grid search results. . . . 150

Section 5.6. Stage 2 Classification with Logistic Regression
Table 5.9. stage 2 classification selected methods . . . 156
Table 5.10. Data sufficiency table . . . 169

Section 5.7. Stage 3 Machine Learning Regression
Table 5.11. Stage 3 regression strategy . . . 171
Table 5.12. Linear regression covariates summary statistics – T_{pair} . . . 175
Table 5.13. Linear regression parameter estimates (quadratic) – T_{pair} . . . 176
Table 5.14. Linear regression parameter estimates (non-quadratic) – T_{pair} . . . 177

Section 5.9. Unblinded Signal Regions – Revealing the Measured Data
Table 5.16. Definition of signal regions . . . 202
Table 5.17. Discriminating variables used for statistical inference . . . 203
Table 5.18. The four search strategies . . . 204
Table 5.19. Unblinded square cuts signal regions . . . 205
Table 5.20. Unblinded machine learning signal regions . . . 209

Section 5.10. Statistical Inference – Setting CL_S Upper Limits
Table 5.22. Combined backgrounds and systematics for statistical inference . . . 219
Table 5.23. Square cuts signal regions' systematic uncertainties . . . 221
Table 5.24. Machine learning signal regions' systematic uncertainties . . . 222
Table 5.25. Four methods of computing upper limits . . . 223
Table 5.26. Comparison of two-stage machine learning to square cuts . . . 230
Table 5.27. Comparison of three-stage machine learning to square cuts . . . 234

Section 6.2. Introduction to Inference for Signal Searches
Table 6.1. Three common frequentist methods of inference . . . 256
Table 6.2. Example 2: Expected background and signal . . . 266
Table 6.3. Feldman-Cousins empty set confidence regions . . . 293

Section 6.3. The Arizona Test Statistics
Table 6.4. Example 2: Four performance metrics – general properties . . . 319

Table 6.5. Example 2: Five performance metrics – limit properties 323
Table 6.6. Example 2: Two performance metrics – risk properties 327
Table 6.7. Properties of three other examples 329
Table 6.8. Summary of performance metrics' definitions 336
Table 6.9. Other examples: performance metrics 338
Table 6.10. Pros and cons of test statistics 340

Section 6.5. Test Statistics in the Presence of Nuisance Parameters

Table 6.11. Nominal nuisance parameter test statistics 395
Table 6.12. Profiled nuisance parameter test statistics 396
Table 6.13. Free nuisance parameter test statistics 397
Table 6.14. Example 3b: $\mu = 1$ signal-background discrimination with ν 405
Table 6.15. Example 2: results of 2σ coverage studies 422

Section A.2. Partition Tree Ensembles and Semi-supervised Learning

Table A.1. Datasets examined with the PTLA 438
Table A.2. Hyper-parameters used for the vlq dataset 440
Table A.3. Hyper-parameters used for the higgs dataset 442
Table A.4. Hyper-parameters used for the cancer dataset 447

Section B.1. The Cross-Evaluation Partitioning Scheme

Table B.1. Possible data partitioning: two-stage with cross-evaluation 461

Section B.2. Unsupervised Machine Learning to Identify Signal Regions

Table B.2. K-means cluster centers 468

Section C.2. Stage 3 Linear Regression: All Tables

Table C.1. Linear regression covariates summary statistics – T_{pair} 483
Table C.2. Linear regression covariates summary statistics – B_{pair} 484
Table C.3. Linear regression covariates summary statistics – B_{single} 485
Table C.4. Linear regression parameter estimates – T_{pair} 486
Table C.5. Linear regression parameter estimates – B_{pair} 487
Table C.6. Linear regression parameter estimates – B_{single} 488
Table C.7. Covariate products parameter estimates (scaled)—T_{pair} 492
Table C.8. Covariate products' parameter estimates (scaled)—B_{pair} 495

Section D.1. The Arizona Test Statistics (Addendum)

Table D.1. Example 2: Performance metrics for $Q_{rep} = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}$ 534
Table D.2. Summary of performance metrics' definitions 534
Table D.3. Summary of performance metrics' definitions 536

Table D.4. Other examples: Performance metrics of Q_{MP} 537

Abstract

A search for vector-like quarks (VLQs) decaying to a Z boson using multi-stage machine learning was compared to a search using a standard square cuts search strategy. VLQs are predicted by several new theories beyond the Standard Model. The searches used 20.3 fb^{-1} of proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 8$ TeV collected with the ATLAS detector in 2012 at the CERN Large Hadron Collider. CL_S upper limits on production cross sections of vector-like top and bottom quarks were computed for VLQs produced singly or in pairs, T_{single}, B_{single}, T_{pair}, and B_{pair}. The two stage machine learning classification search strategy did not provide any improvement over the standard square cuts strategy, but for T_{pair}, B_{pair}, and T_{single}, a third stage of machine learning regression was able to lower the upper limits of high signal masses by as much as 50%. Additionally, new test statistics were developed for use in the Neyman construction of confidence regions in order to address deficiencies in current frequentist methods, such as the generation of empty set confidence intervals. A new method for treating nuisance parameters was also developed that may provide better coverage properties than current methods used in particle searches. Finally, significance ratio functions were derived that allow a more nuanced interpretation of the evidence provided by measurements than is given by confidence intervals alone.

Chapter 1
Introduction

The first goal of this research was to perform a search for massive particles that have been theorized to exist. These vector-like quarks (VLQs) have been proposed in several theories that extend the Standard Model of particle physics in order to address some difficiencies in our current understanding of nature. A search leading to the discovery of VLQ existence would help to solve these issues, while a search ending in nondiscovery provides contraints on these beyond the Standard Model theories. The rest mass of these hypothesized VLQs is unknown, and they may exist within a large range of possible mass values. This dissertation begins by showing how a standard "square cuts" search strategy resulted in the exclusion of VLQ existence for a wide range of possible masses (chapters 2–4).

The second goal of this research was to determine if multi-stage machine learning could improve upon the results of the square cuts search (chapter 5). Use of machine learning in particle physics searches is quite common, but it is usually restricted to a single stage of machine learning. This work differentiates itself by addressing in detail several important issues when multi-stage machine learning is attempted for a particle search.

First, a search for a new particle is usually many separate searches, one for each invariant mass that the hypothesized particle could possess. The two-stage methodology proposed here was able to utilize machine learning for each hypothesized mass individually while ultimately providing a single tool that was applied to all masses simultaneously. Having a single tool was useful because it greatly reduced the human labor required to complete the search.

Second, the two-stage learning approach offered a natural partitioning (halving) of the training data. By equally splitting the training data during two-stage learning, the portion used for training at one stage was reused for validation in the other stage.

Third, training so many separate learning algorithms by hand (i.e., tuning the hyper-parameters of each learning algorithm) is prohibitively labor intensive. Automation of hyper-parameter tuning was achieved here by developing a statistical testing procedure to check for overtraining during learning. This allowed for a grid search for optimal hyper-parameters in order to identify the most powerful learning algorithms that also were not overtrained.

Finally, supervised classification is most often employed in particle searches. This work explored the novel use of supervised regression and unsupervised clustering applied to particle searches. The results of supervised regression showed great promise in extending searches to higher masses (halving upper limits from their square cuts values in some cases), while the unsupervised clustering provided a novel method for identifying control and signal regions.

The third goal of this research was to improve the statistical inference procedure used to draw conclusions from measurements (chapter 6). Frequentist parametric inference is typically used in particle searches by implementing the Neyman construction of confidence

regions in order to estimate the range (confidence interval) of likely signal strengths at which the signal may exist. If the upper limit of the confidence interval is small, then the observed measurement provides evidence against signal existence. This range of signal strengths is determined by comparing the measured data to simulations, and so accurate modeling quality of the simulated data is required for valid conclusions to be drawn.

First, the Neyman construction requires the selection of a test statistic (ordering rule) by the experimenter, and the test statistic ultimately determines the resultant confidence intervals (and their upper limits). This research examined how the test statistics commonly used in particle searches may be modified to incorporate an evaluation of the modeling quality of the simulated data. This resulted in removing problematic empty set confidence intervals, as well as providing confidence intervals such that their lengths were longer when the experimenter was less certain about the modeling quality of the simulated data.

Second, a confidence interval only provides a binary inference statement: a range of signal strengths that are accepted with all other values rejected. Unlike the credibility intervals of Bayesian inference, confidence intervals guarantee a probabilistic interpretation, that confidence intervals constructed in this manner will contain the true signal strength with a frequency equal to the desired confidence level. However, the posterior distribution of Bayesian inference allows for the comparison of the strength of evidence for individual values of the signal strength, whereas a confidence interval only has an accept vs. reject interpretation for each value of the signal strength. For example, there is no way to compare which of two signal strengths is more likely to be true if both lie within the confidence interval.

This work builds on the work of Ronald Fisher, Jerzy Neyman, Alan Birnbaum, and Deborah Mayo (fiducial distributions, the confidence principle, a theory of evidence, and severe testing, respectively) to derive significance ratio functions that provide a method of comparing the evidence for each value of the signal strength. More importantly, the significance ratio functions were shown to have a proper frequentist interpretation, i.e., a probabilistic inference statement is provided for each value of the signal strength.

Figure 1.1 shows the organization of the remainder of this dissertation. Readers interested only in the machine learning techniques that were investigated may skip directly to chapter 5, which is written in a mostly self-contained manner. Those interested in the Neyman construction may skip to chapter 6 which is completely self-contained. The new Partition Trees learning algorithm is presented in appendix A.

Chapter 2: Theoretical Motivation for Vector-Like Quarks (p.26)

- The current theory of the Standard Model.
- The new theory predicting the existence of vector-like quarks.

Chapter 3: Data Collection (p.51)

Chapter 4: The Square Cuts Search (p.76)

- The search using a standard experimental design.

Chapter 5: The Search Using Machine Learning (p.106)

- Multi-stage machine learning.
  - Stage 1: Automated grid search for optimal BDT hyper-parameters.
  - Stage 2: Logistic regression (logit classification).
  - Stage 3: Linear and BDT regressions.

Chapter 6: The Neyman Construction in Signal Searches (p.249)

- Introducing the Arizona test statistics.
- Frequentist evidence beyond confidence intervals.
  - Introducing confidence-acceptance profiles.
  - Introducing significance ratio functions.
  - Meta-analysis with confidence-acceptance profiles.

Chapter 7: Conclusion (p.428)

Appendix A: The Partition Trees Learning Algorithm (p.433)

Appendix B: Other Machine Learning Results (p.461)

Appendix C: The Machine Learning Search (Addendum) (p.476)

Appendix D: The Neyman Construction (Addendum) (p.527)

Figure 1.1: The organization of the dissertation. The chapters covering the standard square cuts search are shown with green boxes. Those covering the multi-stage machine learning search are shown in red. The chapters examining the Neyman construction applied to signal searches are shown in blue. Finally, a novel learning algorithm is described in appendix A (shown in orange).

Chapter 2
Theoretical Motivation for Vector-like Quarks Search

The Standard Model (SM) of particle physics describes the fundamental particles of the universe and the non-gravitational forces that govern their behavior. The SM includes the Brout-Englert-Higgs mechanism with its resultant Higgs boson (discussed in section 2.1.3) as an essential ingredient that is needed to describe the masses of the known particles and the chiral nature of the weak force. The observations in 2012 of a new particle produced at the CERN Large Hadron Collider (LHC) that has properties consistent with the Higgs boson and a mass of $\sim$125 GeV/c^2 provided further evidence for the validity of the SM [1, 2]. Unfortunately, the SM cannot be a complete description of the universe since it does not incorporate the force of gravity. Additionally, the SM is known to have shortcomings, particularly the hierarchy problem that describes the extreme fine tuning of the 19+ SM parameters required to describe such a low Higgs mass (discussed in section 2.1.6).

An extension of the SM is sought in order to address these shortcomings. These beyond the SM (BSM) theories predict the existence of new fundamental particles that could possibly be detected. Many of these BSM theories attempt to address the hierarchy problem by predicting the existence of *vector-like quarks* (VLQs) that are non-chiral when acted upon by the weak force. This search for VLQs was motivated by the fact that VLQs appear in so many BSM theories while having decay characteristics that make them readily distinguishable from the currently known SM fundamental particles. Additionally, most BSM theories require VLQ masses to be less than $\sim$1 TeV/c^2 in order to address fine tuning, and this is within the reach of the experiments at the LHC.

2.1. The Standard Model

The SM is a relativistic quantum field theory constructed to explain the non-gravitational interactions between the fundamental particles of the universe. A review of quantum field theories and their application to high energy particle physics may be found in [3–11]. The following review addresses the particular theoretical details pertinent to the search for VLQs.

The fundamental particles of the SM are the fermions, the force mediating bosons, and the Higgs boson. Fermions are categorized as leptons or quarks with three generations of each as depicted in figure 2.1. The up and down quarks of the first generation are the constituents of protons and neutrons, and along with the electron from the first generation of leptons, provide the majority of the visible matter of the universe. The photon is the boson that mediates the force of electromagnetism that is responsible for the bonding of atoms. The electron, muon, bottom, and top quarks were of particular interest because they provide detector signatures that are particularly useful in the search for VLQs.

As shown in figure 2.1, the SM also contains massless neutrinos. Recent experiments

have demonstrated the existence of neutrino oscillations, which imply nonzero neutrino mass [12, 13]. More observations are needed before it is known how the SM should be updated to include neutrino masses, but approximating neutrinos as massless did not lead to any observable inconsistencies with the modeling of the measured data collected for this search.

Continuous *local* gauge symmetries give rise to force mediating bosons (section 2.1.2). The SM makes extensive use of gauge symmetries in order to model the electromagnetic, weak and strong forces. Specifically, the $U(1)_Y \times SU(2)_L \times SU(3)_C$ group of continuous local transformations gives rise to the bosons that mediate these forces. Furthermore, the spontaneous symmetry breaking of the gauge group $U(1)_Y \times SU(2)_L$ via the Brout-Englert-Higgs (BEH) mechanism (section 2.1.3) predicts a massive scalar (Higgs) boson, and provides a method of introducing fermion mass into the model that allows renormalizability (section 2.1.1). The vector bosons of the Standard Model are depicted in figure 2.2.

A review of these central concepts follows: 1) the importance of renormalizability in quantum field theory that restricts the possible BSM theories, 2) the use of Gauge symmetries to construct force mediating bosons, and 3) the BEH mechanism that provides masses for the weak force bosons as well as the fermions. These concepts are then brought together to explain the unification of the electromagnetic and weak forces in the theory of Glashow, Weinberg and Salam (section 2.1.4), followed by an explanation of the strong force in the SM (section 2.1.5). Finally, the shortcomings of the SM that motivate the search for VLQs are discussed.

2.1.1. Renormalizable Relativistic Quantum Field Theory

In the SM, particles are described by quantum fields in order to model the nondeterministic, "wave-like" nature of microscopic matter. Fields are quantized so that the number of particles described by a field is a nonnegative integer. Special relativity must be included because the quantity of particles is not conserved due to the equivalence of mass and energy, and also to ensure causality so that no signal may travel faster than the speed of light. The predictions of such a theory should be consistent in any inertial frame of reference, and thus the Lagrangian density $\mathcal{L}$ of any system of particle fields must be a Lorentz scalar. In order to prevent nonlocal interactions, the Lagrangian density may only be a function of the particle fields and their first derivatives $\mathcal{L}(\boldsymbol{\phi}, \partial_\mu \boldsymbol{\phi})$.

Feynman diagrams may be used to visualize the mathematics of quantum field theory that computes the probability amplitudes of particle interactions. For example, figures 2.3 (a) and (b) depict the Feynman diagrams for two scattering processes with time increasing from left to right. In figure 2.3 (a), two electrons exchange momentum and energy via interaction with a Z boson. Momentum is conserved at each vertex of the diagram, with energy and angular momentum conservered in the total diagram. The Z boson mediating this weak force interaction between the two electrons is called a virtual particle because it does not exist before the interaction (the left hand side of the Feynman diagram) nor after (the right hand side), but only spontaneously appears allowing for the momentum-energy exchange and then disappears. Figure 2.3 (c) shows the Feynman diagram for the annihilation of an up quark with its antiparticle into an electron-positron pair via the exchange of a virtual photon.

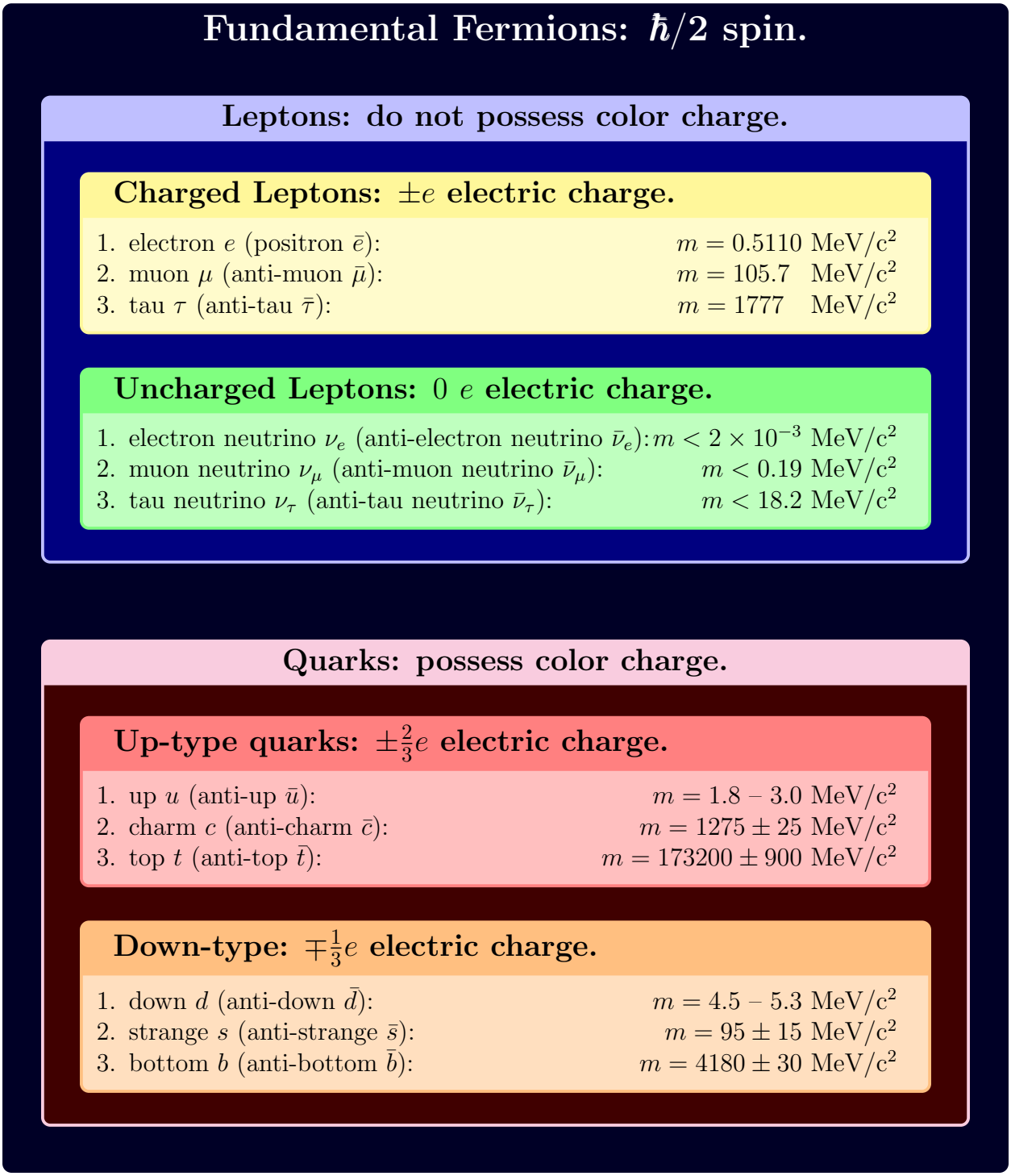


Figure 2.1: The spin $\frac{1}{2}$ fundamental fermions of the Standard Model, leptons and quarks. Three generations of each particle type are known to exist, labeled 1–3 with increasing mass (albeit masses are unknown for the neutrinos). Particle masses are listed along the right column in units of MeV/c^2 for ease of comparison.

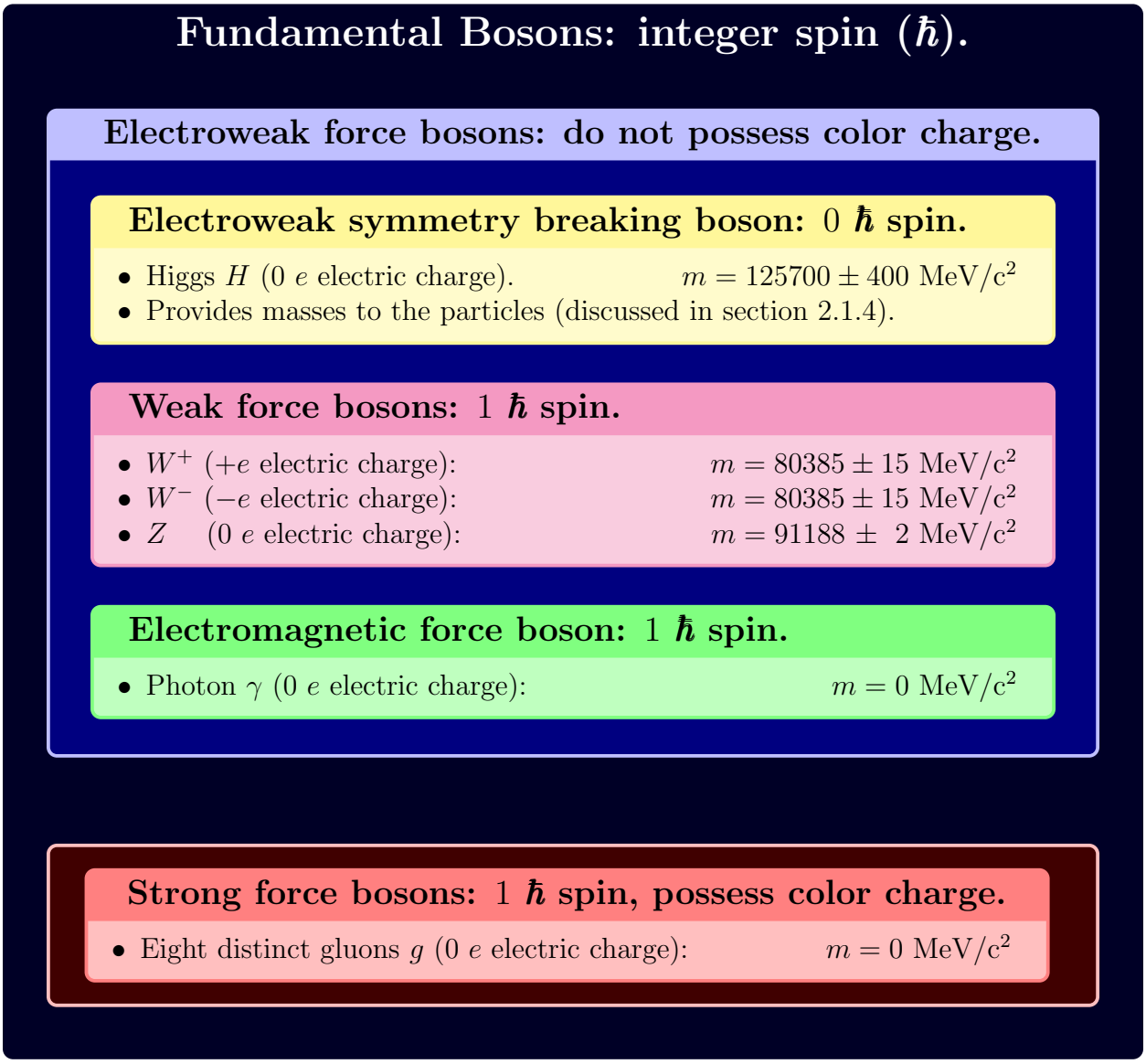


Figure 2.2: The bosons of the Standard Model. Particle masses are listed along the right column in units of MeV/c^2 for ease of comparison.

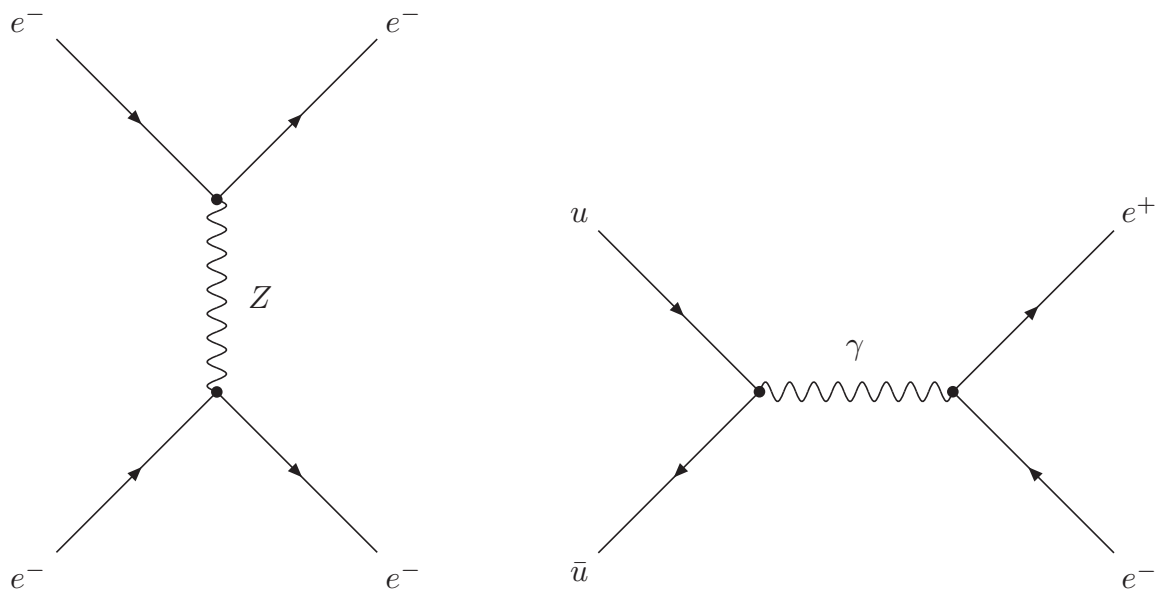


Figure 2.3: Example Feynman diagrams with time increasing from left to right: (a) two electrons scatter through the exchange of a virtual Z boson, and (b) an up quark and anti-up quark annihilate to produce an electron-positron pair via a virtual photon.

Higher order terms in the calculation of probability amplitudes may be represented by including additional virtual particles into the Feynman diagram. Many of these additions introduce divergent terms into the calculations. These can be accommodated in the theory by introducing a renormalization scale, taking a finite number of observed values from experiment (charge, mass, etc.), and absorbing the infinities into unobservable quantities. If the observable quantities of the theory remain finite in the limit as the renormalization scale approaches infinity, then the theory is deemed renormalizable and makes finite valued predictions. The constraint that the theory be renormalizable greatly restricts the form of the Lagrangian density that is permissible. The SM is a renormalizable theory, and BSM theories also seek to maintain renormalizability.

2.1.2. Gauge Symmetries and Vector Bosons

Noether's theorem provides that continuous symmetries of nature which preserve the laws of physics give rise to conserved quantities in physical theories. For example, the invariance of physical laws under translation in time yields the conservation of energy, invariance of spatial translation yields conservation of momentum, and rotational invariance gives angular momentum conservation. Similarly, global (independent of spacetime) gauge invariance in quantum field theory produces conserved currents. For example, the global $U(1)$ continuous phase symmetry of a free fermion field $\psi \rightarrow e^{-ia}\psi$, where $|a| = 1$, leads to fermion number conservation.

Imposition of local continuous gauge symmetries is how the non-gravitational forces of nature arise in the SM. Local continuous symmetries in quantum field theory give rise to new vector fields, the force mediating gauge bosons. For example, if the fermion fields are required

to satisfy a local $U(1)$ phase symmetry where $a(x^\mu)$ becomes a function of spacetime, then this local gauge transformation introduces a new massless vector field, namely the photon that mediates the electromagnetic force in the theory of quantum electrodynamics. The local $U(1)$ symmetry group is Abelian and provides a single massless gauge boson, but a non-Abelian local symmetry will provide more than one gauge boson. For example, the special unitary group $SU(2)$ will generate three massless gauge bosons.

Example: A Simple Quantum Electrodynamics. Let ψ be a four component Dirac spinor field representing a spin $\frac{1}{2}$ fermion with charge q and mass m. The Lagrangian density for this field is given by

$$\mathcal{L} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi, \tag{2.1.1}$$

where γ^μ are the Dirac matrices with dimension 4 representation and $\bar{\psi} = \psi^\dagger\gamma^0$. Impose a local $U(1)$ gauge symmetry so that $\psi \to e^{-iq\theta(x^\mu)}\psi$, where $\theta(x^\mu)$ is a function of spacetime. The Lagrangian will no longer be Lorentz invariant unless a massless vector field A_μ is introduced that transforms as $A_\mu \to A_\mu + \partial_\mu A_\mu$ and interacts with the fermion by

$$\mathcal{L} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - q\bar{\psi}\gamma^\mu\psi A_\mu, \tag{2.1.2}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$.

Fermion fields may be described with left and right-handed chiral components, $\psi = \psi_L + \psi_R \equiv \left(\frac{1-\gamma^5}{2}\right)\psi + \left(\frac{1+\gamma^5}{2}\right)\psi$ where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$, so that the Lagrangian for the free field may be rewritten

$$\mathcal{L} = i\bar{\psi}_L\gamma^\mu\partial_\mu\psi_L + i\bar{\psi}_R\gamma^\mu\partial_\mu\psi_R - m\left(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L\right). \tag{2.1.3}$$

Left and right-handed components of the fermion field mix under Lorentz transformations. This rewritten Lagrangian retains Lorentz invariance because $\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L$ balances this mixing.

However, the weak force is known to interact differently with the left and right-handed components of fermions. Figure 2.4 provides an example of a typical weak force interaction, the decay of a free neutron to a proton, electron and anti-electron neutrino. An initial down quark of the neutron is exchanged with an up quark via a virtual W^- boson that also leads to the creation of an electron and anti-electron neutrino. Only the left-handed components of the d and u contribute to the probability amplitude of the interaction vertex of the W^- with the d and u. Similarly, only the left-handed component of the e^- and right-handed component of the $\bar{\nu}_e$ contribute to the probability amplitude of the interaction vertex of the W^- with the e^- and $\bar{\nu}_e$.

Inclusion of the weak interaction into the Lagrangian density will violate Lorentz invariance unless the fermion mass is set to zero. It is the chiral nature of the weak force that does not allow for the inclusion of fermion masses directly into the Lagrangian density in the above manner.

The SM imposes a local gauge symmetry of $U(1) \times SU(2) \times SU(3)$, where one massless vector boson is due to the $U(1)$ symmetry, three from $SU(2)$, and eight from $SU(3)$. At this point, the number of vector bosons agrees with current experimental understanding, one photon mediating electromagnetism, three bosons mediating the weak force, and 8 gluons mediating the strong force.

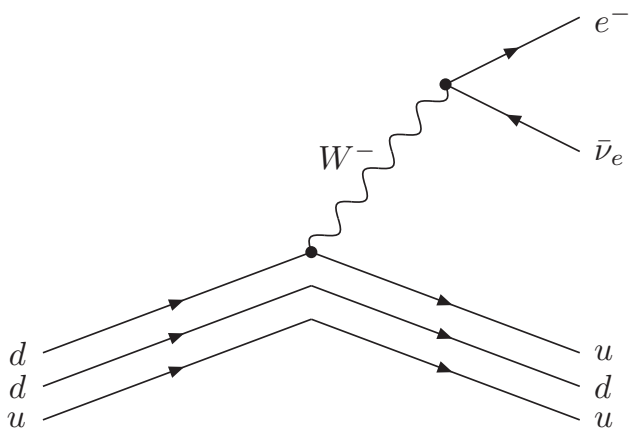


Figure 2.4: The Feynman diagram for the decay of a neutron to a proton, electron and anti-electron neutrino via the exchange of a virtual W^- boson.

The use of continuous local gauge symmetries provides a graceful method for describing the non-gravitational forces of nature. Unfortunately, the force mediating bosons that this produces are massless while the W^+, W^- and Z bosons that mediate the weak force are known to have mass. Furthermore, fermion mass terms cannot be directly inserted into the Lagrangian density because fermions experience a chiral weak force interaction. This conundrum of how to give masses to bosons can be resolved with an elegant procedure called spontaneous symmetry breaking. The spontaneous symmetry breaking implemented in the Brout-Englert-Higgs (BEH) mechanism provides masses for the weak force bosons as well as the fermions by introducing a new particle, the Higgs boson.

2.1.3. Spontaneous Symmetry Breaking & Brout-Englert-Higgs Mechanism

Spontaneous symmetry breaking (SSB) begins with the assumption that the ground state of some scalar fields is degenerate so that many possible configurations of the fields will yield states with the same minimum energy. It is further assumed that these degeneracies satisfy some global symmetry. When a ground state is selected, this global symmetry is broken, and in this process, new massive and massless scalar bosons are introduced to the theory. The Higgs boson is introduced into the SM in this manner.

SSB can be implemented with several types of symmetries as shown in the next three examples using classical field theory. The first example shows how a simple discrete symmetry for a massless particle may be broken to produce a mass term in the Lagrangian density for the particle. The second example shows how the SSB of a continuous global symmetry for a massless particle can produce a massive scalar particle field, but also produces a massless Goldstone boson that is undesirable since no such Goldstone bosons are observed in nature. The third example is the SSB of a continuous local gauge symmetry. This shows how a gauge boson may acquire mass while preventing Goldstone bosons from appearing in the theory. It is in a similar manner to this third example that the BEH mechanism provides masses for the weak force bosons of the SM.

Example 1: Discrete Global Symmetry Breaking. Begin with the Lagrangian density for a real scalar field

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \phi\right)^2 + \frac{m^2}{4}\phi^2 - \frac{\lambda^2}{8}\phi^4, \tag{2.1.4}$$

with positive constants m and λ, which has the field potential $V(\phi) = -\frac{m^2}{4}\phi^2 + \frac{\lambda^2}{8}\phi^4$. This Lagrangian density has the discrete symmetry $\phi \to -\phi$. The minimum energy ground state determined by minimizing the potential is not given by $\phi = 0$, but rather $\phi = \pm\frac{m}{\lambda}$. The field may be expanded around either choice of ground state $\phi = \frac{m}{\lambda} + \sigma$, which gives

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \sigma\right)^2 - \frac{m^2}{2}\sigma^2 - \frac{m\lambda}{2}\sigma^3 - \frac{\lambda^2}{8}\sigma^4, \tag{2.1.5}$$

which now describes a real scalar field σ with mass m and self interaction terms. The Lagrangian density no longer posses the original symmetry due to the cubic term and thus the original symmetry has been broken and a new particle introduced into the theory.

Example 2: Continuous Global Symmetry Breaking – The Simplified Linear Sigma Model. If the Lagrangian density describes a particle with four real scalar fields $\bar{\phi} = (\phi_1, \phi_2, \phi_3, \phi_4)$,

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \bar{\phi}\right)^2 + \frac{m^2}{4}\bar{\phi}^2 - \frac{\lambda^2}{8}\bar{\phi}^4, \tag{2.1.6}$$

then the Lagrangian density is symmetric under the $O(4)$ continuous rotations of the field. The potential now has a minimum $|\phi| = \frac{m}{\lambda}$ that is infinitely degenerate under the rotations, and one may arbitrarily choose the ground state $\bar{\phi}_0 = \left(0, 0, 0, \frac{m}{\lambda}\right)$ and expand around it, $\bar{\phi} = \left(\pi_1, \pi_2, \pi_3, \frac{m}{\lambda} + \sigma\right)$. This gives the Lagrangian density (writing $\bar{\pi} = (\pi_1, \pi_2, \pi_3)$)

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \bar{\pi}\right)^2 + \frac{1}{2}\left(\partial_\mu \sigma\right)^2 - \frac{m^2}{2}\sigma^2 - \frac{m\lambda}{2}\sigma^3 - \frac{\lambda^2}{8}\sigma^4 - \frac{m\lambda}{2}\bar{\pi}^2\sigma - \frac{\lambda^2}{4}\bar{\pi}^2\sigma^2 - \frac{\lambda^2}{8}\bar{\pi}^4, \tag{2.1.7}$$

which now describes three massless real scalar fields $\bar{\pi}$ and a massive real scalar field σ with mass m (and interaction and self-interaction terms). The $O(4)$ symmetry has been broken and replaced with the $O(3)$ rotational symmetry of the $\bar{\pi}$ fields. Goldstone's theorem describes that whenever a continuous symmetry is broken, a massless scalar (Goldstone) boson appears in the theory [14]. No scalar massless bosons have ever been detected, but it is possible to absorb Goldstone boson terms of the Lagrangian into those of known particles through the SSB of gauge symmetries. Also, it can be shown that once the classical field is quantized, the resultant quantum field theory is renormalizable.

Example 3: Continuous Local Gauge Symmetry Breaking. The Lagrangian density in this example describes a complex scalar field $\phi = \frac{1}{\sqrt{2}}\left(\phi_r + i\phi_i\right)$ with charge q that is symmetric under a local $U(1)$ gauge symmetry, $\phi \to e^{-iq\theta(x^\mu)}\phi$. The local gauge symmetry requires a massless vector boson A^μ as described in section 2.1.2 so that the Lagrangian density before symmetry breaking is

$$\mathcal{L} = \left(\partial_\mu - iqA_\mu\right)\phi^\dagger\left(\partial_\mu + iqA^\mu\right)\phi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{m^2}{4}\phi^2 - \frac{\lambda^2}{8}\phi^4. \tag{2.1.8}$$

The ground state is achieved when $A^\mu = 0$ and ϕ is constant with $|\phi| = \frac{m}{\lambda}$. Note however that the symmetry transformation $\phi \to e^{-iq\theta(x^\mu)}\phi$ mixes the real and imaginary components of ϕ

so that the minimum is infinitely degenerate. Symmetry is broken by selecting a particular gauge for $\theta(x^\mu)$, the most convenient choice being whichever makes $\phi' = e^{-iq\theta(x^\mu)}\phi$ positive real valued so that the ground state is $\phi_0 = \frac{m}{\lambda}$. When the theory is quantized, the classical ground state becomes a vacuum expectation value (VEV). Expanding about the ground state $\phi = \frac{m}{\lambda} + \frac{h}{\sqrt{2}}$ with h a real valued field due to the choice of gauge gives the Lagrangian density

$$\begin{aligned}\mathcal{L} &= (\partial_\mu - iqA_\mu)\left(\frac{m}{\lambda} + \frac{h}{\sqrt{2}}\right)(\partial_\mu + iqA^\mu)\left(\frac{m}{\lambda} + \frac{h}{\sqrt{2}}\right) - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \\ &\quad + \frac{m^2}{4}\left(\frac{m}{\lambda} + \frac{h}{\sqrt{2}}\right)^2 - \frac{\lambda^2}{8}\left(\frac{m}{\lambda} + \frac{h}{\sqrt{2}}\right)^4 \\ &= \frac{1}{2}(\partial_\mu h)^2 - \frac{m^2}{4}h^2 - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{q^2m^2}{\lambda^2}A_\mu A^\mu \\ &\quad + \sqrt{2}\frac{q^2m}{\lambda}A_\mu A^\mu h + \frac{q^2}{2}A_\mu A^\mu h^2 \\ &\quad - \frac{\sqrt{2}m\lambda}{8}h^3 - \frac{\lambda^2}{32}h^4. \end{aligned} \tag{2.1.9}$$

This example shows that when a local gauge symmetry is spontaneously broken, a massive scalar field is created, a mass has been provided for the gauge boson, and there are no longer any massless Goldstone bosons.

In a similar manner, the BEH mechanism begins with a complex scalar doublet field called the Higgs field, and spontaneously breaks a continuous local gauge symmetry, $U(1) \times SU(2)$. This produces a new massive scalar particle called the Higgs boson, three massive gauge bosons for the weak force, and no massless Goldstone bosons. Fermion mass terms may be constructed by including Yukawa interactions that couple the fermions to the Higgs field. When the Higgs field undergoes SSB, fermion mass terms are generated in the Lagrangian density in a manner that does not violate Lorentz invariance.

2.1.4. Glashow-Weinberg-Salam Electroweak Theory

It is the chiral nature of the weak force, acting differently on the left and right-handed components of the fermion fields, that causes the difficulty in providing masses for the fermions. The Glashow-Weinberg-Salam electroweak theory (GWS) unifies the chiral weak force with the electromagnetic force, and in doing so provides masses for the fermions and three weak force bosons by implementing the BEH mechanism.

In GWS, four vector bosons are constructed by imposing a $U(1)_Y \times SU(2)_L$ local gauge symmetry. The $U(1)_Y$ symmetry provides a massless gauge boson B^0 with gauge coupling constant g_B. The $SU(2)_L$ symmetry provides three massless gauge bosons A^1, A^2 and A^3 with coupling g_A. The $SU(2)_L$ symmetry addresses the chiral nature of the weak force by acting on doublets of left-handed fermion chirality states (right-handed for antiparticles). This is constructed as

$$\begin{pmatrix} \nu_l \\ l^- \end{pmatrix}_L \text{ and } \begin{pmatrix} q_u \\ q_d \end{pmatrix}_L, \tag{2.1.10}$$

for leptons and quarks, respectively, where $l = e$, μ, or τ with ν_l the corresponding flavor

neutrino, and where q_u is an up type quark and q_d is down type.

The $SU(2)_L$ symmetry acts on these left-handed doublets where two of its three generators, T_1 and T_2, mix the upper and lower components (flavor changing), which gives rise to the two massless $W^{\pm}$ gauge bosons. The left-handed fermion doublet is an eigenstate of the third generator T_3 with eigenvalues $\pm\frac{1}{2}$ for the upper and lower components respectively. This third generator provides a weak neutral current mediated by A^3. The electromagnetic charge Q is related to the quantum numbers for Y and T_3 by $Q = T_3 + \frac{Y}{2}$. The right-handed fermion chirality states are singlets of $SU(2)_L$ (left-handed for anti-particles), and this produces the chirality of the weak interaction in the SM.

The $U(1)_Y$ gauge symmetry describes fermion hypercharge with accompanying massless gauge boson B^0, and is used to couple the chiral electromagnetic interaction to the chiral weak interaction after SSB. Finally, in addition to the 12 fermions and four massless gauge bosons, a complex scalar Higgs field ϕ_H is introduced that is a doublet of $SU(2)_L$ and has hypercharge $+1$ under $U(1)_Y$. Yukawa couplings of the Higgs to the fermions are also included. The particle fields of the theory before spontaneous symmetry breaking are described by table 2.1.

The BEH mechanism is now employed and the $U(1)_Y \times SU(2)_L$ gauge symmetry spontaneously broken with a VEV of v and real valued scalar field H so that the Higgs

	particle field	weak hypercharge Y/Q	T_3 quantum number #/e	electric charge Q/e
leptons	l_L^-	-1	$-\frac{1}{2}$	-1
	$\nu_{l,L}$	-1	$+\frac{1}{2}$	0
	l_R	-2	0	-1
quarks	$q_{u,L}$	$+\frac{1}{3}$	$+\frac{1}{2}$	$+\frac{2}{3}$
	$q_{d,L}$	$+\frac{1}{3}$	$-\frac{1}{2}$	$-\frac{1}{3}$
	$q_{u,R}$	$+\frac{4}{3}$	0	$+\frac{2}{3}$
	$q_{d,R}$	$-\frac{2}{3}$	0	$-\frac{1}{3}$
$U(1)_Y$ gauge	B^0	0	0	0
$SU(2)_L$ gauge	A^+	0	$+1$	$+1$
	A^-	0	-1	-1
	A^3	0	0	0
scalar Higgs field	ϕ_H	$+1$	$-\frac{1}{2}$	0

Table 2.1: Particle fields before SSB in the GWS model with fermions separated by chirality state. For simplicity, $A^{\pm}$ are defined as $(A^1 \mp iA^2)/\sqrt{(2)}$.

field expanded about the ground state is $\phi_H = \left(0, (v+H)/\sqrt{2}\right)$. This is the process of *electroweak symmetry breaking*. One gauge symmetry remains, $U(1)_{EM}$, which provides the symmetry needed for the electromagnetic interaction mediated by a massless photon. The Lagrangian density is rearranged by defining new fields (and writing $\tan(\theta_W) \equiv g_B/g_A$)

$$W^{\pm} = \left(A^1 \mp iA^2\right), \tag{2.1.11}$$

$$Z = \cos(\theta_W) A^3 - \sin(\theta_W) B^0, \tag{2.1.12}$$

$$\gamma = \sin(\theta_W) A^3 + \cos(\theta_W) B^0, \tag{2.1.13}$$

with the mass term for the γ field being eliminated. The γ field now corresponds to the massless photon stemming from the remaining $U(1)_{EM}$ gauge symmetry. The remaining mass terms give $m_{W^+} = m_{W^-} = vg_A/2$ and $m_Z = m_W/cos(\theta_W)$ with the mass of the Higgs (and VEV) determined by the two parameters of the quartic potential for ϕ_H used in the BEH mechanism. The measured masses of the known particles are shown in table 2.2 [3].

	particle field	electric charge Q/e	mass MeV/c^2
leptons	e^-	-1	0.5110
	μ^-	-1	105.7
	τ^-1	-1	1777
	ν_e	0	$< 2 \times 10^{-3}$
	ν_μ	0	< 0.19
	ν_τ	0	< 18.2
quarks	u	$+\frac{2}{3}$	$1.8 - 3.0$
	c	$+\frac{2}{3}$	1275 ± 25
	t	$+\frac{2}{3}$	173200 ± 900
	d	$-\frac{1}{3}$	$4.5 - 5.3$
	s	$-\frac{1}{3}$	95 ± 5
	b	$-\frac{1}{3}$	4180 ± 30
gauge bosons	γ	0	0
	$W^{\pm 1}$	± 1	80385 ± 15
	Z	0	91188 ± 2
	g	0	0
Higgs	H	0	125700 ± 400

Table 2.2: Standard model particle masses after SSB in the GWS model. The eight strong force mediating gluons g are included here for completeness.

Additional Yukawa couplings of the fermions to the Higgs field are included in the Lagrangian density to provide masses for the fermions. After SSB, interaction terms appear that mix the lepton generations and that mix the quark generations. If massless neutrinos are assumed, the mass eigenstates of the leptons may be made to coincide with the weak eigenstates. This results in the accidental symmetry (a symmetry not imposed in the original Lagrangian density) of $U_e(1) \times U_\mu(1) \times U_\tau(1)$, which separately conserves the lepton numbers within each of the three lepton generations. Thus the weak interaction will only mix lepton flavors within each separate generation, such as electrons only with electron neutrinos. This symmetry is only approximate in the case of massive neutrinos, but this approximation does not affect the results of the particle physics search conducted here.

In the case of the quarks, weak eigenstates do not correspond to mass eigenstates. Since it is the mass eigenstate of a particle that is observed in nature, the weak interaction mixes *between* quark generations, and the only accidental symmetry is a $U_q(1)$ conservation of total quark number. The cross-generational mixing of the quarks by the weak interaction is described by the Cabibbo-Kobayashi-Maskawa (CKM) matrix which delineates the amount of mixing between the quark mass eigenstates. The CKM matrix has been measured to be nearly diagonal so that cross-generational mixing is rare, but small off-diagonal elements do allow for some generation mixing.

Generational mixing of the quarks allows the decay of higher generation quarks to first generation quarks, which are the only ones observed to be stable in nature. This is demonstrated in figures 2.5 (a) and (b) with the decay of a bottom quark. Figure 2.5 (a) shows that without generation mixing, the weak force could only exchange the bottom quark for a top quark, which is much more massive than the bottom quark. Figure 2.5 (b) shows that generation mixing allows the bottom quark to decay to first and second generation quarks (and possibly leptons). The off diagonal elements of the CKM matrix that quantify this mixing are particularly small for the b quark, which gives the bottom quark a rather long lifetime.

On the other hand, charged leptons of higher generations may freely decay to lower generation leptons (or possibly quarks) without the need for generation mixing because the same generation neutrino is nearly massless. This is demonstrated in figure 2.5 (c) where a muon (second generation) decays to first generation leptons or quarks.

2.1.5. Quantum Chromodynamics

The strong force is incorporated into the SM by employing a non-Abelian $SU(3)_C$ gauge symmetry with the eight generators of $SU(3)_C$ represented as 3-by-3 matrices providing eight massless spin 1 gauge bosons, the gluons. The quarks on which the gluons interact transform as triplets of $SU(3)_C$, while the leptons which do not interact with the strong force transform as singlets. The three quark fields of this triplet are labeled by a color charge, red, green and blue, while the antiquark is labeled by anticolor charge

$$\begin{pmatrix} q_r \\ q_g \\ q_b \end{pmatrix}_C \quad \text{and} \quad \begin{pmatrix} \bar{q}_{\bar{r}} \\ \bar{q}_{\bar{g}} \\ \bar{q}_{\bar{b}} \end{pmatrix}_C . \tag{2.1.14}$$

The interaction of gluons upon the quark triplets may be described in terms of color-anticolor pairs. An example of such a pair is $r\bar{g}$ which rotates the fields q_g to q_r, or for the antiquark,

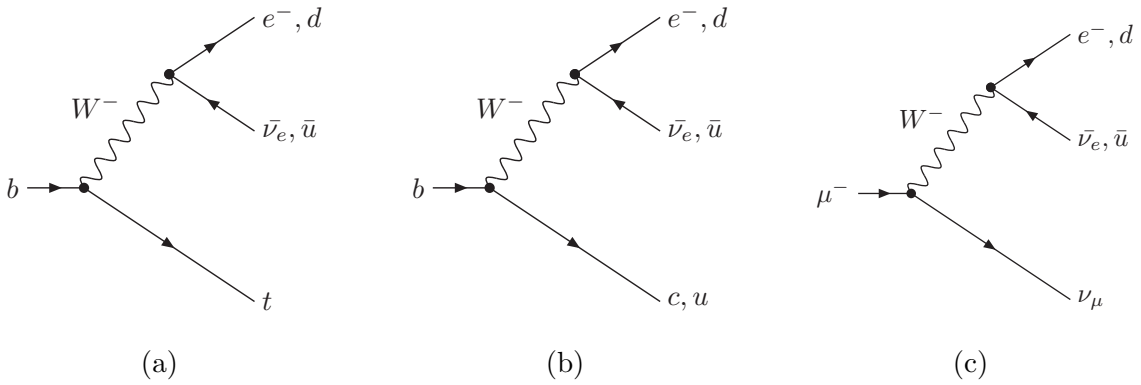


Figure 2.5: Feynman diagrams demonstrating the decay of a higher generation fermions to lower generation. (a) Without generation mixing by the weak force, a bottom quark could only be exchanged with a top quark. (b) Generation mixing as quantified in the CKM matrix allows the bottom quark to decay to first and second generation fermions. (c) A muon (second generation) decays to first generation fermions without the need for generation mixing by the weak force because the muon neutrino is nearly massless.

$\bar{q}_{\bar{r}} \rightarrow \bar{q}_{\bar{g}}$. For example, the gluon described by $\left(r\bar{b} + b\bar{r}\right)/\sqrt{2}$ exchanges q_r and q_b ($\bar{q}_{\bar{r}}$ and $\bar{q}_{\bar{b}}$ for the antiquark), and corresponds to the $SU(3)_C$ generator represented as $\left(\begin{smallmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$.

Despite being mediated by massless bosons, the strong force only has an effective range of $\sim 10^{-15}$ m. The effective range of the weak interaction is $\sim 10^{-18}$ m because of the large masses of the $W^{\pm}$ and Z weak force mediating bosons. The electromagnetic interaction has an infinite range because it is mediated by a massless photon. Furthermore, individual quarks are not observed in nature. Only hadronic matter is observed as bound states of two or more quarks. The explanation for these observations is the theory of color confinement.

The primary difference between the long-ranged electromagnetic force and the short-ranged strong force is due to the differing nature of the charge carriers, the photon versus the gluons. Because the photon arises from the Abelian $U(1)_{EM}$ gauge symmetry, there is no photon-photon interaction in the SM Lagrangian density. However, the $SU(3)_C$ gauge symmetry is non-Abelian so that gluon-gluon interactions do appear in the SM as shown in figures 2.6 (a) and (b). These figures illustrate that a gluon carries color charge and may thus interact with other gluons. This is the primary difference that leads to the theory of color confinement for the strong force giving the short interaction range.

The theory of color confinement states that no particle or bound state of particles may be observed with any net color charge, but must remain color neutral. A particle of nonzero electromagnetic charge may be observed because the electromagnetic field surrounding it and extending to infinity is mediated by uncharged photons. If a nonzero color charge particle were to be isolated in space, it would produce a color field mediated by gluons. The color charge of the gluons within this field would contain self-interaction among the gluons that would require the field to store an infinite amount of energy. Therefore, no nonzero color charge particle may be observed in isolation, only combinations that yield net neutral color charge. The color neutral combinations of quarks observed in nature are the two and three

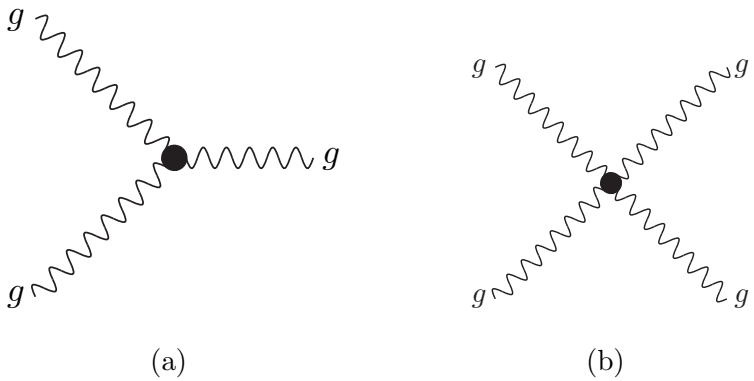


Figure 2.6: Feynman diagrams demonstrating interactions between the strong force mediating gluons due to gluons possessing color charge: (a) two gluons may combine to a third, and (b) two gluons may exchange momentum and energy.

quark bound state hadrons, further categorized as mesons comprised of two bound quarks and baryons comprised of three bound quarks.

In addition to color confinement, the strong interaction has the property of asymptotic freedom. Asymptotic freedom describes how the strong force is large when quarks interact with one another at low energies, but gets progressively smaller as the interaction energy increases. Asymptotic freedom explains why quarks are observed to be tightly bound together in nature when little energy is being exchanged between the quarks. When quarks are made to interact at higher energies, the strong force is much smaller, and this allows interaction probabilities to be calculated using perturbative QCD techniques.

The coupling strength between two quarks *increases* with separation distance. This is the opposite effect as seen in electromagnetism. In electromagnetism, the region surrounding a charged particle is filled with virtual particles shielding the bare charge of the particle. The closer another charged particle gets, the more it pierces the cloud of virtual particles and experiences a stronger interaction. However, the cloud of virtual particles surrounding a color charged particle also includes the color charged gluons, which act to increase the strong interaction strength with increasing separation distance. This asymptotic freedom supports the observation of quarks only in bounds states at roughly the femtometer scale, as well as making it difficult to accurately predict low energy strong force interactions.

The process of hadronization is a consequence of color confinement and asymptotic freedom. During a proton-proton collision at the LHC, individual quarks and gluons are often created as decay products. However, a quark or gluon has color charge, and the strong force between that quark or gluon and other color carrying particles grows as the separation distance increases. The energy stored in the field of virtual color charged particles must also increase with separation distance, where this energy is provided by the kinetic energy of the collision particles. As the ejected color carrying particle moves away from the collision point, eventually enough energy is stored in the virtual particle field to lead to the creation of one or more quark-antiquark pairs. These additional quarks then combine with the original quarks resulting in color neutral mesons and baryons as the final outgoing collision decay products. With the high energy collisions of the LHC, many hadronic particles may be created during

a single collision, forming jets of hadronic particles that may be detected. The process by which hadronization is modeled for simulations used in this study is shown in figure 2.7.

2.1.6. Problems with the Standard Model

The SM is a phenomenological model because it contains 19+ parameters, including masses and interaction strengths, which cannot be derived from first principles. The SM also suffers from the hierarchy problem, it does not describe the force of gravity, and it is unable to explain several astronomical puzzles.

The hierarchy problem alone warrants the search for new particles predicted by BSMs. The Planck energy scale where the gravitational interaction contributes nonnegligibly to elementary particle interactions is $\mathcal{O}(10^{19})$ GeV/c^2, while the electroweak energy scale at which electroweak symmetry breaking occurs is only $\mathcal{O}(10^2)$ GeV/c^2. The large difference between these scales requires fine tuning of the SM parameters to more than 30 decimal places in order to obtain the observed masses of the known particles [18]. More specifically, the SM parameters must be extremely fine-tuned for the renormalized Higgs mass to remain at the electroweak scale [9].[1]

Although the SM is an incomplete model without any description of gravitation, it does provide particle fields containing mass on which gravity acts. Several astronomical puzzles exist which may be examined from the perspective of the SM and BSMs: baryon asymmetry, dark matter, and dark energy. The baryonic material within the universe is comprised of matter and not antimatter, which gives the baryon asymmetry puzzle. Symmetry violations of the SM do not provide a solution, but do supply hints on how the SM might be extended to accommodate this observation. The dark matter puzzle refers to the observation that ∼80% of the matter density of the observable universe is likely comprised of cold dark matter of unknown origin [20]. Particles that interact weakly with other forms of matter are a prime suspect in the constituency of dark matter. Although the SM does not provide any viable candidates, the weakly interacting SM neutrinos provide an example of how such particles can be constructed in a BSM theory. Finally, ∼70% of all the matter-energy density of the universe is believed to be comprised of dark energy, leading to an acceleration of the expansion of the observable universe. The SM provides a vacuum energy density that is of the wrong magnitude to explain the observed cosmological constant, but does provide a mechanism for achieving an explanation by extension of the SM.

[1] Chapter 1 of [19] provides a more detailed description of the hierarchy problem.

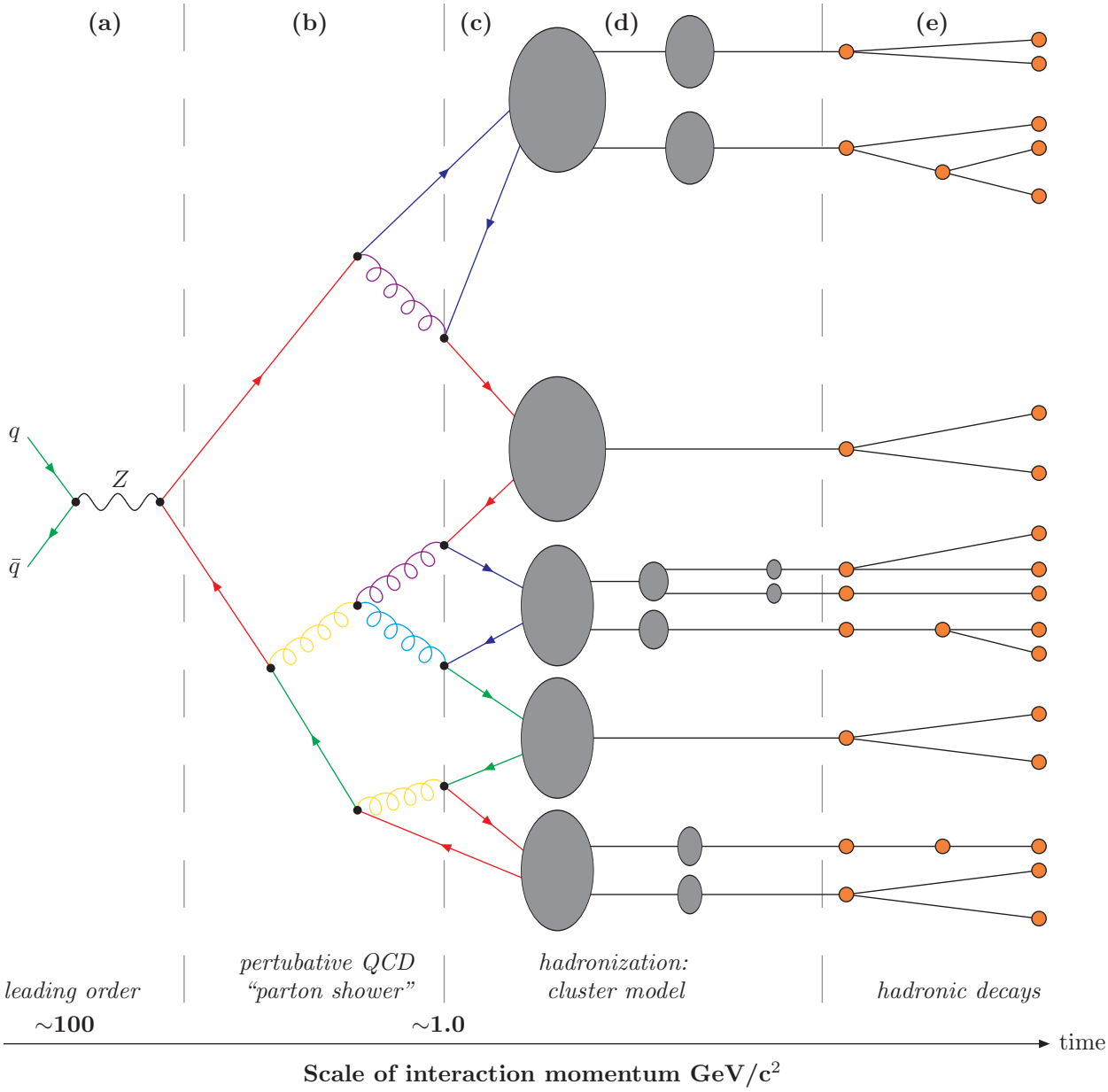


Figure 2.7: Several processes are needed to model the hadronization of quarks and gluons that are measured as hadronic jets. (a) Leading order calculations are performed at tree level. (b) Next-to-leading order calculations used to model the parton shower are possible with perturbative QCD until about 1 GeV/c^2 [15]. (c) Phase space splitting of gluons is performed, and ensures that color singlet combinations enter the clusters [16]. (d) Clusters undergo simple isotropic decay into pairs of hadrons chosen according to density of states calculations [17]. (e) The hadrons formed by the clustering algorithm have well-measured decay properties. The Lund string model may also be used in a similar manner instead of clustering.

2.2. Vector-like Quark Theories

Many beyond the SM (BSM) theories attempt to address the hierarchy problem by extending the SM to include new particles that remove the fine tuning needed to achieve the large difference between the Planck and electroweak scales while still reproducing the SM at low energies. These hypothetical particles must have large masses to have avoided detection in past experiments, but only new particles with masses < 1 TeV/c^2 could be examined in this research using measurements collected in 2012 at a collision energy of $\sqrt{s} = 8$ TeV.

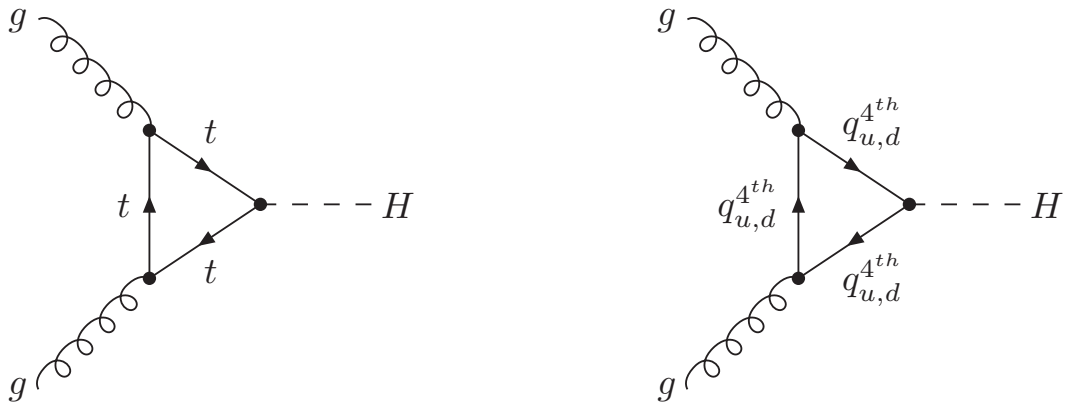
The hierarchy problem is due to the scalar nature of the Higgs boson, which gives quadratic divergences in the computation of the Higgs boson mass using perturbation theory [21]. The SM parameters must by fine-tuned to over 30 decimal places, or else when the SM is applied to energies above the electroweak scale, $\mathcal{O}(10^2$ GeV/c$^2)$, the squared mass of the Higgs boson diverges quadratically toward the Plank scale, $\mathcal{O}(10^{19}$ GeV/c$^2)$ [22]. The hierarchy problem places large constraints on Grand Unified Theories seeking to unify the known fundamental forces of nature [23]. If one considers the SM an effective field theory of some higher energy theory that is only valid up to $\sim$2 TeV, then new particles with masses $\lesssim 10$ TeV must exist [24].

Vector-like Quarks

The simplest extension to the SM is to propose the existence of a fourth generation of chiral fermions, which if found would leave the SM essentially unaltered. However, the existence of a fourth generation of chiral leptons or chiral quarks is highly constrained. Precision measurements of the Breit-Wigner width of the Z boson are in good agreement with the current three generations of chiral fermions [3]. A new chiral lepton and neutrino would require that the masses of each be above 45 GeV/c^2 in order to not affect the Z decay width. The existence of a fourth generation chiral neutrino with such a large mass would require additional explanation. Therefore, it is more natural to consider a fourth generation of chiral quarks rather than chiral leptons, especially since the SM top quark mass has been measured to be quite large $\sim$173 GeV/c^2.

The existence of a fourth generation of chiral quarks is also highly constrained in the SM. The most obvious constraint is due to the cross section for Higgs boson production, which has gluon fusion with a top quark loop as the primary production mechanism at the LHC as shown in figure 2.8 (a). If a fourth generation of chiral quarks existed as shown in figure 2.8 (b), then the Higgs cross section would be at least nine times larger than predicted by the SM, in large disagreement with current measurements.

These constraints do not apply to fermions that are not chiral in weak interactions. Vector-like quarks (VLQs) are nonchiral and appear in extensions of the SM that attempt to address the hierarchy problem. The SM Lagrangian contains both vector and axial weak interaction coupling terms for the SM fermions. This is how the SM captures the chiral nature of the weak force. Vector-like quarks (VLQs) are new fermions proposed by many BSMs and do not include an axial weak interaction term, only the vector term, and are thus "vector-like". Therefore, both the left and right-handed chirality states of VLQs transform identically under the $U(1)_Y \times SU(2)_L$ gauge symmetry of the SM. BSM theories predicting VLQs include some supersymmetry models [25, 26], composite Higgs [27, 28], Little Higgs [29,



(a) (b)

Figure 2.8: Gluon fusion with (a) a top quark loop as the primary production mechanism for the Higgs boson. (b) If chiral fourth generation of up $q_u^{4^{th}}$ and down $q_d^{4^{th}}$ type quarks existed, the Higgs production cross section would be increased by a factor of ∼9.

30], extended Higgs [31], enlarged gauge symmetry groups [32] including those for Grand Unified Theories [33], and theories with extra dimensions [34, 35].

VLQs may have a large mass that does not alter the Z boson decay width, and they do not require a Yukawa coupling to the Higgs field that leads to an increase of Higgs cross section. Furthermore, a bare mass term for VLQs may be included directly into the SM Lagrangian because it does not violate the $SU(2)_L$ local gauge symmetry. Thus the mass of VLQs does not require the use of the BEH mechanism.

Vector-like Quark Models

If the Higgs field is restricted to the complex scalar doublet as in the SM, and if VLQs are required to have renormalizable couplings to the SM quarks, then the VLQs may only appear as singlets, doublets, and triplets of $SU(2)_L$ [36, 37].

$$(T),\ (B),\ \begin{pmatrix} X \\ T \end{pmatrix},\ \begin{pmatrix} T \\ B \end{pmatrix},\ \begin{pmatrix} B \\ Y \end{pmatrix},\ \begin{pmatrix} X \\ T \\ B \end{pmatrix},\ \text{and}\ \begin{pmatrix} T \\ B \\ Y \end{pmatrix}, \tag{2.2.1}$$

where the vector-like top T is an up type VLQ with electromagnetic charge $+2/3$, and the vector-like bottom B is a down type VLQ with $-1/3$ electric charge. X and Y have electromagnetic charge of $+5/3$ and $-4/3$, respectively. Only the T and B VLQs were investigated in the search presented in this research.

The addition of VLQs to the SM necessitates enlargement of the CKM matrix to include the mixing of mass eigenstates of the additional quark content [38]. This mixing alters the coupling strength between SM quarks and weak bosons, and is constrained by current measurements. A well motivated scenario is that the usual mass hierarchy of the CKM matrix is followed so that VLQs mix primarily with third generation SM quarks [39], and

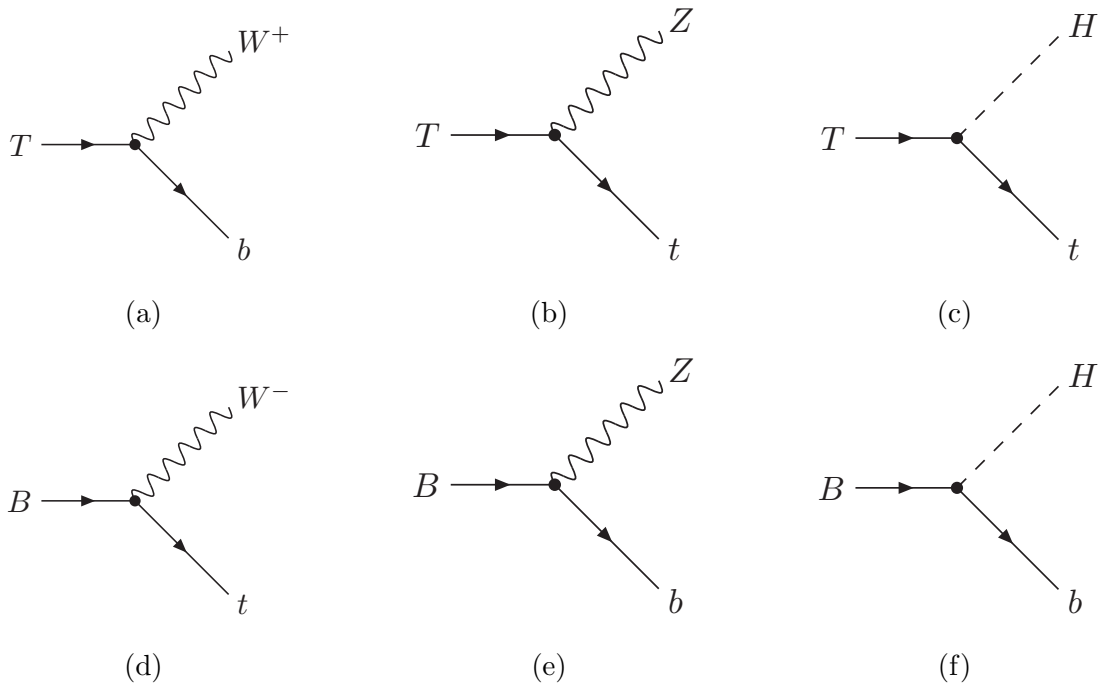


Figure 2.9: The decay modes of the vector-like quarks T and B each have a charged current decay mode as would be found with chiral fourth generation SM quarks, but also two neutral current decay modes. The vector-like top T decay modes: (a) $T \to W^+ + b$, (b) $T \to Z + t$, and (c) $T \to H + t$. The vector-like bottom B decay modes: (d) $B \to W^- + t$, (e) $B \to Z + b$, and (f) $B \to H + b$.

this assumption was made for this search.

The assumption of primarily third generation mixing provides three decay modes for each of T and B via weak interactions as shown in figures 2.9 (a)–(f). BSM models with VLQs may eliminate the Glashow–Iliopoulos–Maiani (GIM) mechanism allowing neutral current VLQ decay [40]. The t, W, Z and H daughter particles may themselves decay hadronically or leptonically to a large variety of additional decay products. In order to constrain this search, only decay topologies that contained a leptonically decaying Z boson were considered, $Z \to l^+ + l^-$, where $l = e$ or μ. The search for VLQs producing a Z boson was conducted by the Z-tag search team at ATLAS [41], and is the basis for the research presented here.

The branching ratios for the three decay modes are determined by the parameters of the specific BSM theory. The sum of the branching ratios (BRs) must equal 1,

$$BR(Q \to W + q') + BR(Q \to Z + q) + BR(Q \to H + q) = 1, \tag{2.2.2}$$

where Q is either T or B, and so a model independent search for such VLQs must span the two dimensional plane of all possible branching ratios. Such a general search was conducted by the Z-tag search team at ATLAS and was reported in [41]. However, certain models with specific branching ratios are well-motivated from theory [37, 42–45]. Figures 2 (a) and (b) of [41] show the mass-dependent branching ratios for SU(2) singlet and doublet models. The

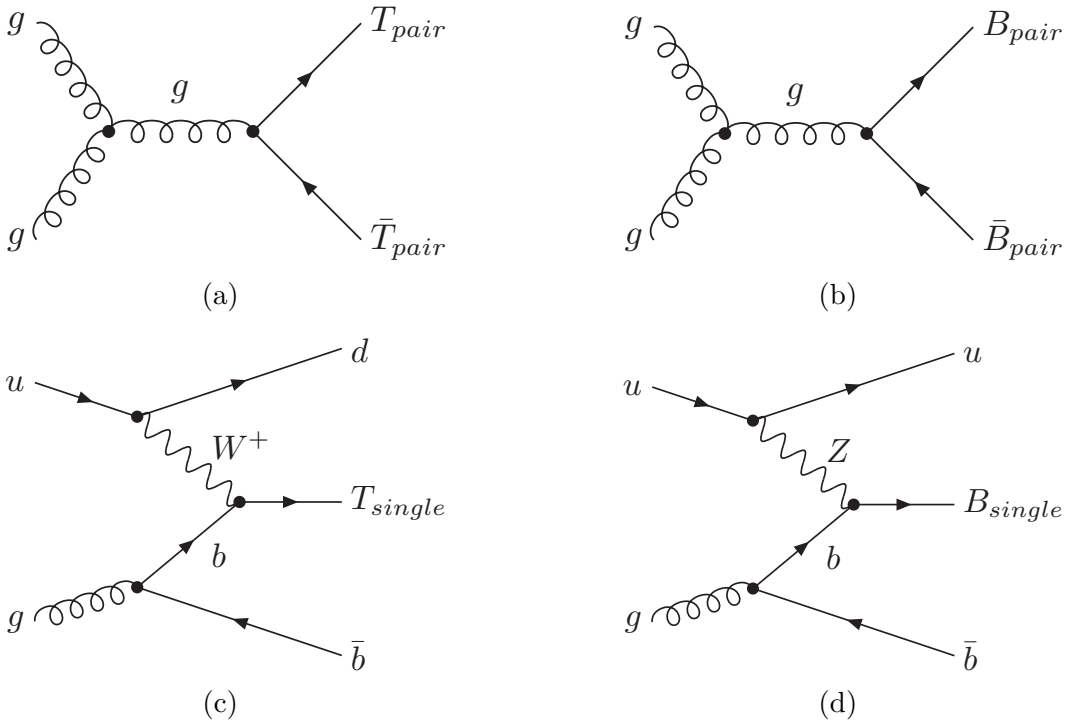


Figure 2.10: The principle production mechanisms of VLQs occurs via gluon fusion for pair produced (a) T_{pair} and (b) B_{pair}, and with an associated SM b quark for singly produced (c) T_{single} and (d) B_{single}.

decay modes producing a Z boson were of primary interest for the Z-tag search, and these figures show that the SU(2) singlet model for T or B had small branching ratios for the Z boson decay modes (solid blue lines), about ∼0.2 and ∼0.3, respectively, and depending on the VLQ mass. This dissertation research sought to improve upon a standard square cuts search strategy using machine learning, and the search for the singlet model was selected because the small branching ratio of the Z boson decay mode made this a particularly challenging search.

Vector-like Quark Production: T_{pair}, B_{pair}, T_{single}, and B_{single}

Figures 2.10 (a)–(d) show that the T and B VLQs may be produced in pairs or singly. The pair production of T_{pair} and B_{pair} from gluon fusion shown in figures 2.10 (a) and (b) gives cross sections that are model independent, and only depend only on the VLQ mass. This makes the search for pair produced VLQs attractive since it applies to a wide variety of BSM theories. The production mechanisms for the singly produced VLQs of figures 2.10 (c) and (d) rely on weak interactions so that the cross sections are model dependent.

Figure 1 (b) of [41] shows the model independent theoretical cross sections of the pair produced VLQs (solid black line) as well as those of three singly produced models. For the singly produced VLQ models parameterized by V_{Tb} and X_{Bb} (the extended CKM matrix elements), the cross sections using $V_{Tb} = 0.1$ for T_{single} and $X_{Bb} = 0.1$ for B_{single} are shown. These values are those maximally allowed by electroweak precision measurements for the

assumption of only one vector-like multiplet at the electroweak scale [37, 46]. However, if several multiplets exist, mixing strength is less constrained. A composite Higgs model was constructed for T_{single} that allowed this possibility with the TWb vertex strength parameterized by $1 \leq \lambda_T \leq 5$, and $\lambda_T = 2$ was selected for the search [47, 48]. Although a search for singly produced VLQs does not offer the generality that a search for pair produced VLQs does, the rate of pair production decreases more quickly than that of single production. In principle, a large mass VLQ may be more easily detected with a search targeting singly produced VLQs.

Figures 2.11 (a)–(d) show example T decay topologies, and figures 2.12 (a)–(d) example B decay topologies. Only topologies containing a leptonically decaying Z boson are shown. Comparing one of the most complex decay topologies, figure 2.11 (a) to one of the least, figure 2.12 (d), demonstrates the large differences in VLQ decay signatures. These differences increased the complexity of designing a VLQ search strategy that could be applied to all VLQ decay topologies. This is the topic of chapters 4 and 5.

The Z-tag search thus examined the evidence for the existence of VLQs with the following properties.

- The VLQs had +2/3 and -1/3 electric charge (the T and B, respectively).
- The VLQs mixed primarily with third generation SM quarks.
- The VLQs had a nonnegligible branching ratio to a Z boson decay product, $BR(Q \rightarrow Z + q) > 0$.
- The VLQs were produced in pairs, T_{pair} and B_{pair}, or else singly, T_{single} and B_{single}.

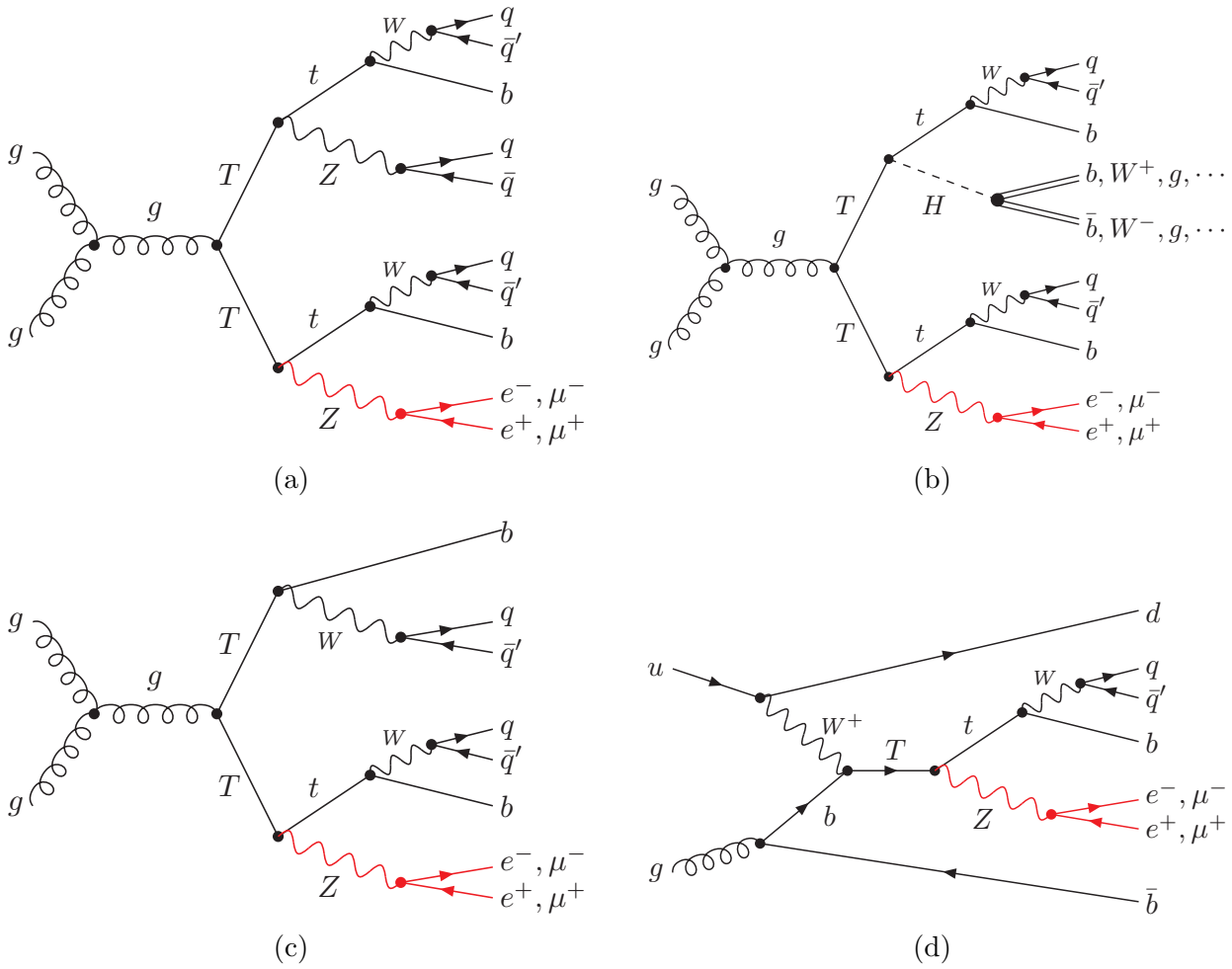


Figure 2.11: VLT decay topologies for the dilepton constraint shown in red: (a)–(c) T_{pair}, and (d) T_{single}.

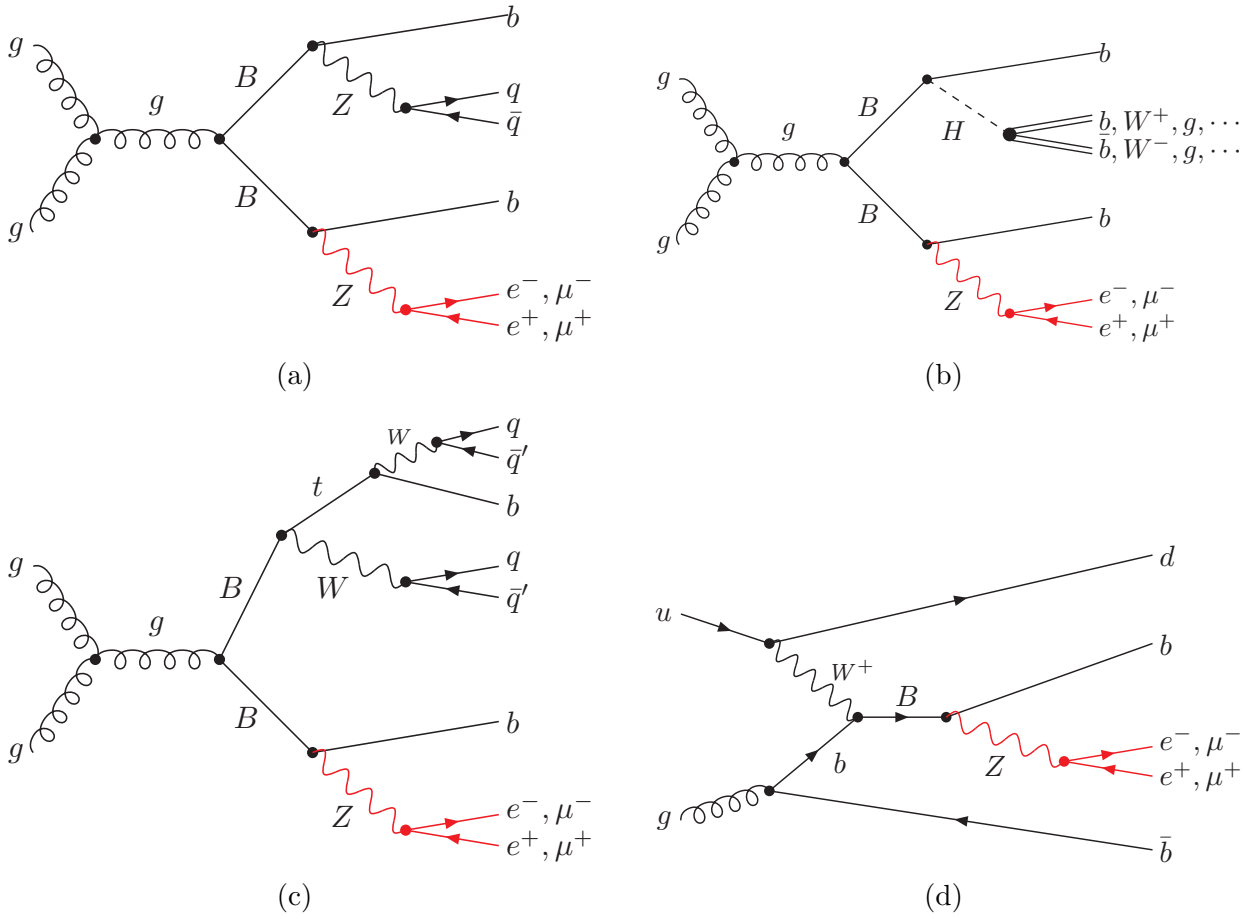


Figure 2.12: VLB decay topologies for the dilepton constraint shown in red: (a)-(c) B_{pair}, and (d) B_{single}.

2.2.1. Other Searches for Vector-like Quarks

The first published results of VLQ searches from the ATLAS and CMS experiments were in 2011. Table 2.3 summarizes the principle results of these and later searches for pair produced VLQs, where upper limits were computed to determine the range of VLQ masses excluded with 95% confidence.[2] No evidence for VLQ existence was detected in these searches. However, these earlier searches were performed with a relatively small amount of data, 1–5 fb^{-1} (collected at a collision energy of $\sqrt{s} = 7$ TeV/c^2 from LHC Run 1), which necessitated making strong assumptions about the VLQ decay processes. The ∼20 fb^{-1} at 8 TeV/c^2 from Run 2 allowed for searches of more realistic models.

The Run 1 searches for pair produced VLTs included a dilepton search for $T \to Z + t$ that set an upper mass exclusion limit of 475 GeV/c^2 by assuming $BR(T \to Z + t) = 1$ [49], a single lepton search for $T\bar{T} \to W^+b + W^-\bar{b}$ that set a mass limit of 404 GeV/c^2 by assuming $BR(T \to W + b) = 1$ [50], a dilepton search for $T\bar{T} \to W^+b + W^-\bar{b} \to l^+\nu b + l^-\bar{\nu}\bar{b}$ that set a limit of 557 GeV/c^2 by assuming $BR(T \to W + b) = 1$ [51], and a single lepton search for $T \to W + b$ that set a limit of 500 GeV/c^2 for the singlet model, but also provided the first model independent upper limits for all possible branching ratios [52].

Similarly, the Run 1 searches for pair produced VLBs included a dilepton search for $B \to Z + b$ that set a mass limit of 400 GeV/c^2 assuming $BR(B \to Z + b) = 1$ [53], and a dilepton search for $B \to Z + b$ that set a mass limit of 466 GeV/c^2 assuming $BR(B \to Z + b) = 1$, and which also found that use of boosted decision trees improved limits by ∼40% [54]. Another Run 1 search targeted both VLTs and VLBs with a single lepton search for $T \to Z + t$ and $B \to W + t$ that gave limits of 625 and 675 GeV/c^2, respectively, with the assumptions $BR(T \to Z + t) = 1$ and $BR(B \to W + t) = 1$ [55].

Run 2 searches for VLTs and VLBs were more nuanced in that they examined both specific models, such as the singlet model, and also spanned all possible branching ratios in order to provide model independent upper limits. For the VLT, this included a search for all three decay channels using boosted decision trees that set a mass limit of 687 GeV/c^2 [56], and an all-hadronic search for $T \to H + t$ using jet substructure that set a limit of 698 GeV/c^2 by assuming $BR(T \to H + t) \geq 0.8$ [57]. For the VLB, this included a multilepton search for B decaying to all three possible modes that set a limit of 740 GeV/c^2 [58], and a single lepton search for the singlet model with B decaying by all three possible modes and using boosted decision trees that set a limit of 640 GeV/c^2, and which also provided limits for all possible branching ratios [59].

Simultaneous Run 2 searches for both VLTs and VLBs included a same-sign dilepton search for the singlet model yielding a limit of 590 GeV/c^2 for the VLT, 620 GeV/c^2 for the VLB, and which also provided limits for all possible branching ratios [60]. Also, a single lepton search for $T \to Wb + X$ or $T \to Ht + X$ gave a limit of 700 GeV/c^2 for all possible branching ratios of the VLT, and 575 GeV/c^2 for the VLB [45].

The search on which this dissertation is based was conducted for VLTs and VLBs that were produced in pairs as well as singly, and where a VLQ decayed to a Z boson: $T \to Z + t$ and $B \to Z + b$ [41, 61–64]. For pair produces VLQs, this provided a VLT mass limit of 655 GeV/c^2 for the singlet and 735 GeV/c^2 for the doublet, and a VLB mass limit of 685 GeV/c^2 for singlet and 755 GeV/c^2 for the doublet [41]. Additionally, upper limits for

[2] This statistical inference procedure is the central topic of chapter 6.

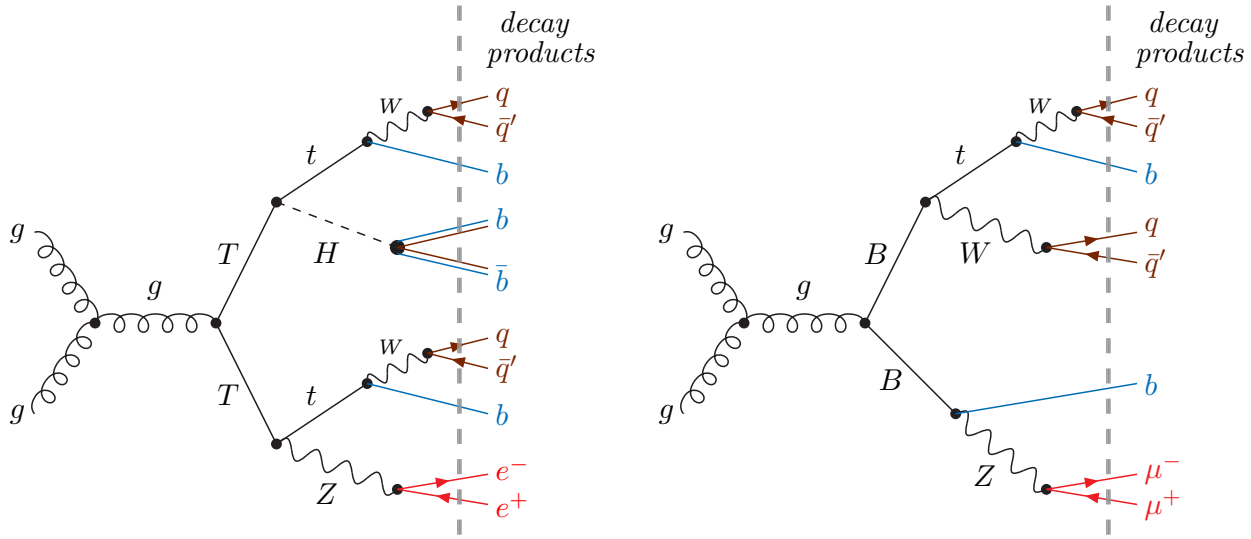
all possible branching ratios were provided for pair produced VLQs. This search also set 95% confidence level limits on VLT and VLB single production for the specific parameters discussed in section 2.2. Although these did not constrain the permitted masses of the BSM theories, they represented an early perspective on future single production searches at the LHC, and this dissertation work examined how these results may be improved with the use of machine learning.

mass limit GeV/c^2	model assumptions	publication year & [ref.]	experiment, collision energy TeV/c^2, collected data fb^{-1}
T_{pair}			
475	$BR(T \to Z + t) = 1$	2011 [49]	CMS, 7, 1.14
404	$BR(T \to W + b) = 1$	2012 [50]	ATLAS, 7, 1.04
557	$BR(T \to W + b) = 1$	2012 [51]	CMS, 7, 5.0
625	$BR(T \to Z + t) = 1$	2012 [55]	CMS, 7, 5.0
500	$SU(2)_L$ singlet, $BR(T \to W + b) \sim 0.5$	2013 [52]	ATLAS, 7, 4.7
687	all possible branching ratios	2014 [56]	CMS, 8, 19.5
655	$SU(2)_L$ singlet	2014 [41]	ATLAS, 8, 20.3
735	$SU(2)_L$ doublet	2014 [41]	ATLAS, 8, 20.3
698	$BR(T \to H + t) \geq 0.8$	2015 [57]	CMS, 8, 19.7
590	$SU(2)_L$ singlet	2015 [60]	ATLAS, 8, 20.3
700	all possible branching ratios	2015 [45]	ATLAS, 8, 20.3
B_{pair}			
400	$BR(B \to Z + b) = 1$	2012 [53]	ATLAS, 7, 2.0
358	$SU(2)_L$ singlet, $BR(B \to Z + b) \sim 0.12$	2012 [53]	ATLAS, 7, 2.0
466	$SU(2)_L$ singlet, $BR(B \to Z + b) = 1$	2012 [54]	ATLAS, 7, 4.7
675	$BR(B \to W + t) = 1$	2012 [55]	CMS, 7, 5.0
685	$SU(2)_L$ singlet	2014 [41]	ATLAS, 8, 20.3
755	$SU(2)_L$ doublet	2014 [41]	ATLAS, 8, 20.3
740	all possible branching ratios	2015 [58]	CMS, 8, 19.7
620	$SU(2)_L$ singlet	2015 [60]	ATLAS, 8, 20.3
640	$SU(2)_L$ singlet, all branching ratios	2015 [59]	ATLAS, 8, 20.3
575	all possible branching ratios	2015 [45]	ATLAS, 8, 20.3

Table 2.3: The principle results of other searches for VLQs that mix primarily with third generation SM quarks. The upper mass exclusion limits are at the 95% confidence level.

Chapter 3
Data Collection

3.1. Introduction to Data Collection



The above Feynman diagrams provide two examples of the spectra of decay products that appear from the decay of vector-like quarks (VLQs) when restricted to decay topologies that contain a leptonically decaying Z boson. The decay products shown in red are the same-flavor, opposite-sign electrons or muons that are used to identify the Z boson candidate from VLQ decay. The identification of electrons by the detector is much different from the identification of muons so that the Z-tag search is essentially two separate searches, one for electronic Z boson decays and one for muonic.

Shown in blue and brown are the Standard Model (SM) quarks that appear in VLQ decay. Quarks colored brown in the figures are the first and second generation SM quarks (up, down, charm, and strange), and those colored blue are the third generation bottom quarks. Isolated quarks and gluons do not exist in nature due to hadronization so that they appear in the detector as jets of hadronic material (discussed in section 3.4.4). The SM top quarks have lifetimes $\sim 10^{-25}$ seconds so that they decay before any hadronization is possible, and are only observed by their decay products. The lifetime of B hadrons is $\sim 10^{-12}$ seconds so that B hadrons travel several millimeters on average before decaying. This displacement often allows for the identification of the bottom quarks (discussed in section 3.4.5).

Therefore, the Z-tag search for VLQs required identifying same-flavor, opposite-sign electrons or muons that provided Z boson candidates, and hadronic jets both from first and second generation SM quark decay products, as well as hadronic jets originating from bottom quarks.

In addition to detecting these objects, the ATLAS experiment was able to a) detect photons, b) infer the presence of neutrinos, and c) tag hadronically decaying τ leptons. Photons were not often present in VLQ decay topologies, but photon signatures in the detector often mimicked those of electrons or jets necessitating steps to reduce this misidentification error. Neutrinos interact so weakly with matter that they nearly always escaped the detector. Neutrinos appeared in Z-tag decay topologies, especially when τ leptons were produced as shown in figures 3.2 (a) and (b). In the Z-tag search, no attempt was made to identify τ leptons or the presence of neutrinos, and the search strategy remained agnostic to their presence.

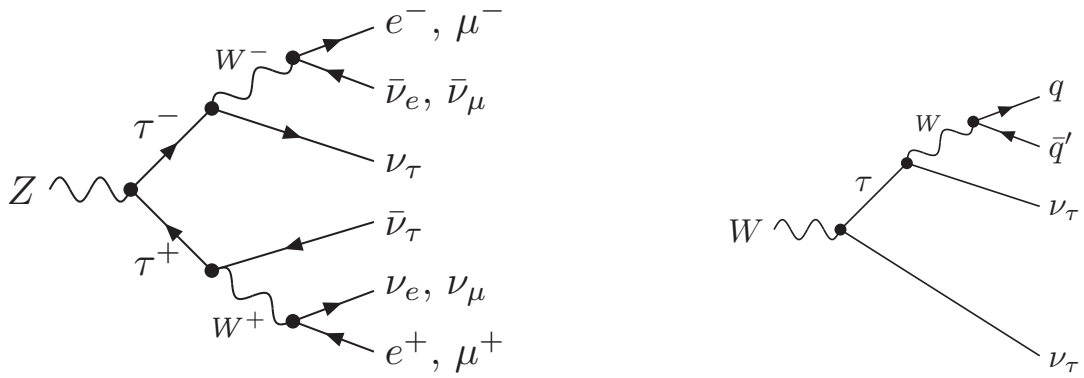


Figure 3.2: VLQ decays often produced τ leptons and neutrinos in the subsequent decays of (a) Z bosons, and (b) W bosons.

The objective of the Z-tag search for VLQs was to examine the measurements made by the ATLAS detector in order to identify the presence of the final state particles that would have been produced in VLQ decays if VLQs existed. These VLQ signatures were very similar to those produced by other SM processes, the background of the VLQ signal. Therefore, the search required comparing measurements made by the detector to simulations of the VLQ signal plus simulations of the SM background processes. In the context of machine learning, the simulated data is labeled data (signal vs./ background) because one has the simulated collision events that produce VLQs or else the simulations of the Standard Model background processes. These simulations were then used in the machine learning process of binary classification (discussed in chapter 5).

The remainder of this chapter explains the creation of the datasets used for the Z-tag search. This begins with section 3.2 that explains how the Large Hadron Collider provided the proton-proton collisions for measurement, and section 3.3 that explains how the ATLAS detector made these measurements. This is followed by section 3.4 that delineates how the measured dataset was constructed by reconstructing physics objects from the detector responses. Finally, section 3.5 explains the construction of the simulated datasets for both the SM backgrounds and the VLQ signals.

3.2. Creating p–p Collisions with the Large Hadron Collider

In order to test the existence of VLQs, protons were collided with sufficient energy to create VLQs if they existed. The principle tasks of the CERN accelerator complex were

1. to create bunches of protons focused into collimated beams,
2. to direct two proton beams to recirculate in opposing directions,
3. to accelerate the proton beams to high energies of nearly uniform velocity, and
4. to cross the beams at specific points where particle detectors were located.

This section describes the most pertinent aspects of how these tasks were achieved. Further details are described in [65–67].

The Large Hadron Collider (LHC) at CERN has four experimental detectors within a 26.7 km circumference tunnel, lying ∼100 meters underground, and situated on the border between France and Switzerland. The circular tunnel was originally constructed for the Large Electron-Positron (LEP) collider. The tunnel for the LEP was constructed with such a large circumference in order to reduce the power loss from synchrotron radiation. An ultrarelativistic charged particle accelerated perpendicular to its direction of motion emits electromagnetic radiation (synchrotron radiation) with power loss inversely proportional to the fourth power of its mass, $P_{loss} \propto m^{-4}$ [68]. The small mass of the electrons and positrons collided at LEP necessitated the large tunnel circumference. A little more than 4 km of the tunnel was comprised of eight straight sections, four experimental insertions and four utility insertions, giving an effective radius of curvature ~ 3.5 km. The more massive protons collided at the LHC had only $\sim 10^{-13}$ the synchrotron radiation power loss as the electrons and positrons collided at the LEP, and so were able to be accelerated to much higher energies.

The primary method of charged particle acceleration was the use of radio frequency (RF) oscillating electromagnetic fields made to occupy resonant cavities [69]. These resonant cavities were designed to provide large electric potential gradients in the direction of particle travel yielding acceleration of ~ 100 MeV per meter of cavity, and to also to ensure phase stability so that faster moving particles entering the cavity did not receive as much acceleration (and initially slower particles received slightly more than average acceleration), which caused the particles to cluster into bunches of nearly equal velocity. In order to achieve very high final energies, the charged particles either passed through a long series of RF accelerating cavities, such as in the CERN LINAC-2, or else were made to travel in recirculating orbits in order to encounter the same accelerating RF cavities multiple times as was done in the several synchrotrons of the CERN accelerator complex.

The energy contained in two colliding protons that is available for particle-creating collisions is described by the Mandlestam variable s. This is a relativistic invariant that describes the maximum amount of energy available for particle creation, $s = (p_1^\mu + p_2^\mu)^2$, where p_1^μ is the four-momentum of one proton and p_2^μ of the other proton. The rest mass of the proton is only 9.4×10^{-4} TeV so that each proton was accelerated to ultrarelativistic velocity in order to reach TeV-scale energies. In the collision of ultrarelativistic protons, $s \simeq 2p_1^\mu p_2^\mu$, and when the protons have equal and opposite momentum and energy E, then $s \simeq 4E^2$, or $\sqrt{s} \simeq 2E$.[1] Hence a proton-proton collision described by $\sqrt{s} = 8$ TeV indicated

[1] Compare this to a proton with energy E colliding with a stationary proton: $\sqrt{s} = \sqrt{2Em_p}$, where $m_p = 0.938$ TeV/c^2 is the proton invariant mass. If $E = 4$ TeV, then two colliding protons has nearly three

that each colliding proton had an energy of 4 TeV so that each proton was traveling with a velocity 0.99999997 times the speed of light.

In addition to the energy available for particle creation, the rate at which creation occurs also had to be known. The instantaneous luminosity $\mathcal{L}$ $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ that may be achieved by two colliding proton beams (modeled by Gaussian cross sectional profiles) is given by [70]

$$\mathcal{L} = \frac{n_{ppb}^2 N_{bunch} f_{rev}}{4\pi w_x w_y}, \tag{3.2.1}$$

where n_{ppb} is the number of protons per bunch in each of the two beams, N_{bunch} the number of bunches per beam, f_{rev} the frequency of recirculation, and w_x and w_y the Gaussian cross sectional widths. The luminosity increases with the number of protons per bunch, number of bunches, and decreases with cross sectional beam size. This luminosity may then be used to predict the rate at which some particular particle creating process X occurs by

$$\frac{dN_X}{dt} = \mathcal{L}\sigma_X, \tag{3.2.2}$$

where $\frac{dN_x}{dt}$ is the creation rate of process X, and σ_X is the cross section in cm^2 for this process that describes the probability of occurrence and depends on the collision energy. The total integrated luminosity of the entire time of data collection, $L = \int \mathcal{L}dt$, may then be used to predict how many of these events will have occurred, $N_X = L\sigma_X$.

The goal was thus to create two highly energetic beams of protons within the LHC, grouped into bunches of large numbers of protons, and to focus these bunches to a small cross sectional area at the interaction points where the detectors were located: i.e., to create high energy, high luminosity proton-proton collisions that could produce VLQs if they existed. This process began within several facilities in the CERN accelerator complex that created the proton beams and carefully controlled their properties before injecting them into the LHC.

Figure 3.3 shows the chain of accelerators that contributed to the final 8 TeV collision energy. This began at the Duoplasmatron Proton Ion Source. Hydrogen gas, H_2, was injected into a cathode chamber where 50 amp pulsed arcs stripped the H_2 of electrons, and, after a series of chemical reactions facilitated by multipole magnetic confinement, produced H^+ protons [71]. These protons were then accelerated toward a molybdenum anode across a 92 kV potential and left the duoplasmatron with a 250–320 mA beam current where they entered the LINAC-2 linear accelerator.

The LINAC-2 produced a high intensity 180 mA beam of 50 MeV protons using the 750 keV Radio Frequency Quadrupole 2 (RFQ2) accelerator, and three Alvarez drift tube tanks [72]. The protons first entered RFQ2 which had four longitudinal electrode vanes resonating in TE_{210} transverse electric field mode and pulsed at radio frequency to achieve focusing, bunching and acceleration of the protons with high efficiency [73]. Note that this was the location in the accelerator chain where protons first began to be collected into bunches. The protons left RFQ2 and entered the 33.3 m long Alvarez tubes that accelerated the protons to 10.3, 30.5, and finally 50 MeV, respectively, and were passed to the Proton Synchrotron Booster (PSB) [74].

---

times the energy available for particle creation (8 TeV) compared to the stationary proton target (2.7 TeV).

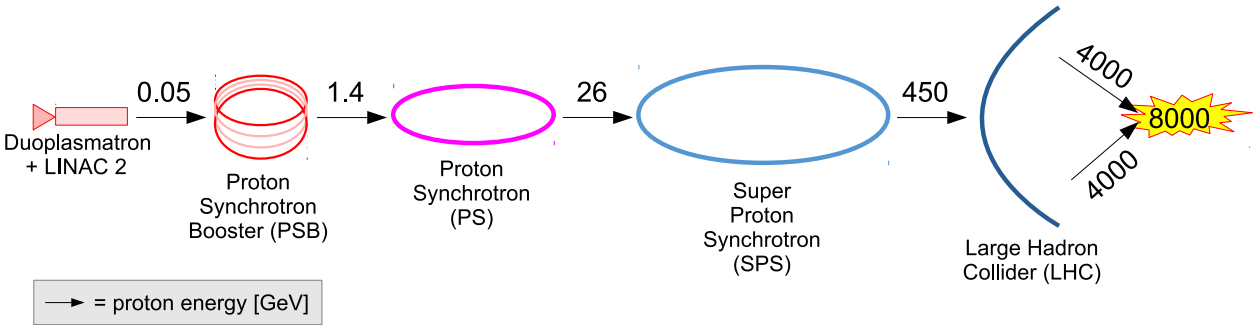


Figure 3.3: The CERN accelerator complex injection chain showing the energies of the protons as they left each component.

The PSB was the first of three synchrotrons used to accelerate the protons before injection into the LHC, and acted as a mid-energy bridge between LINAC-2 and the older Proton Synchrotron (PS) in order to increase the amount of protons entering the PS. The PSB had four 25 m radius rings with 32 dipole and 48 quadrupole magnets. The proton beam arriving from LINAC-2 was split into four separate rings by a staircase-like kicker magnet, and once filled could be accelerated to 1.4 GeV in a large variety of modes. The typical mode used for LHC high luminosity proton-proton collisions was the $h = 1$ harmonic, where h is the harmonic number equal to the number of RF cycles occurring during one particle orbit, and accumulated over 10^{12} protons per bunch with one bunch per ring [75]. The bunches in the four rings were then recombined into a single beam entering the Proton Synchrotron through a six bunch extraction filling scheme (1, 2, 3, $4 + 3$, 4). This extraction scheme was facilitated by filling all four PSB rings in one booster cycle followed by the acceleration of only two rings (3 and 4) in the next, and took 1.2 seconds for each delivery of the six bunches.

The 628 m circumference Proton Synchrotron (PS) received the six proton bunches from the PSB, filling six of seven stationary buckets with the PS operating in the $h = 7$ mode. These were split into 72 bunches [76, 77] through a process of "RF gymnastics" utilizing several h values, and accelerated to 26 GeV [76]. The empty seventh bucket provided a 320 ns beam gap that allowed for the rise time of the ejection kicker for export from the PS to the Super Proton Synchrotron.

The Super Proton Synchrotron (SPS) is a 6.9 km long synchrotron that was loaded by 3–4 cycles of the PS for a maximum of 288 bunches [67, 78]. The SPS then accelerated the beam to 450 GeV with over 10^{11} protons per bunch with 0.5 ns bunch width, and loaded the 26.7 km long LHC. Loading the LHC took several minutes, depositing 1374 bunches into each of the two LHC beams [79].

Once each of the two LHC beams were filled with 0.45 TeV protons, each beam was then accelerated to contain 4 TeV protons. This was accomplished with eight $f_0 = 400.8$ MHz RF stations per beam that accelerated during ramping, compensated for small energy losses during coasting, and provided longitudinal focusing [80].

These proton beams were made to recirculate using a series of bending magnets. The LHC contains 1232 main dipole magnet units occupying ~75% of the tunnel length, each with a length of 15 m and a mass of 30,000 kg, which bent the two proton beams in the circular sections of the tunnel [81]. The units provided an initial magnetic field strength of

0.54 T (tesla) as the proton beams were populated by the injector system, and then ramped up to 8.3 T during the acceleration of the beams.

The magnetic fields of the dipole magnets were directed perpendicular to the plane of the tunnel in order to direct the proton beams along the tunnel curvature. The magnetic fields were created by high density currents surrounding the 6.3 cm diameter beam pipes containing each beam. Such large magnetic fields were made possible by the use of superconducting currents cooled by a total of 130,000 kg of 1.9 K superfluid helium, which provided much higher thermal conductivity than would have been possible with slightly higher temperature liquid helium.

Because the proton beams traveled in opposite directions, they each had to be provided magnetic fields in opposing directions. The narrow width of the tunnel and cost considerations required that the two magnetic fields be created within the same device. In addition to the numerous bending dipole magnets, 10% of the tunnel length was comprised of 7,724 focusing and corrector magnets of various kinds: dipole, quadrupole, sextupole, octopole, and dodecapole magnets [81]. Without focusing, the beam would widen over time so that w_x and w_y would increase in equation 3.2.1 and thus decrease the collision luminosity.

Quadrupole and higher multipole magnets prevented this by providing a series of focusing and defocusing magnetic lenses. The design of the focusing magnets made extensive use of Liouville's theorem which states that the phase space area of the position and momentum transverse to the direction of the beam, described by the emittance ϵ, remained constant as each proton bunch traversed a magnetic field [82]. A quadrupole magnet focused a proton beam in one transverse direction while defocusing it in the other, thus conserving the emittance.

The width of each proton bunch was $\sim$ 200 μm on average, with longitudinal bunch spacing length of 5.6 cm, and these were focused to a small cross sectional area of $\sim$ 16 μm at the collision region near the center of the detector for the ATLAS experiment. This provided an instantaneous luminosity from 3.65 to 7.7 $\times$ 10^{33} $\mathrm{cm}^{-2}\mathrm{s}^{-1}$ [83, 84]. It was the task of the ATLAS experiment to measure the collision products created by these high luminosity proton-proton collisions.

3.3. Observing p–p Collisions in the ATLAS Experiment

The Standard Model (SM) predicts that when two protons collided, the energy they contain often leads to the creation of new particles. The created particles are usually unstable and decay into other, more stable particles over very short distances. This often occurs many times via decay chains and hadronization until final state particles are produced that are stable enough to reach the various components of the detector. It was these final state particles that were measured in the search for VLQs, and from these measurements, the presence of the initially created particles had to be inferred. There were four types of final state particles with lifetimes long enough to reach and interact with the ATLAS detector: electrons, photons, muons, and hadronic particles of many types. Neutrinos were stable final state particles, but interacted so weakly with the detector material that they passed through without detection.

The various subdetectors contained within the ATLAS experiment returned electronic responses. These responses were reconstructed into "physics objects" that could be interpreted within the context of particle physics theory, i.e., the fundamental particles of the SM. The four physics objects that were reconstructed from measurements were electron candidates, photon candidates, muon candidates, and hadronic jet candidates.

For each collision, the estimated locations and energies of the reconstructed objects provided the information needed to infer the presence of the original initial fundamental particles created in the proton-proton collisions. An illustrative example is provided by the decay of a SM top quark shown in figure 3.4. The SM predicts that the top quark will rapidly decay to a W boson and a bottom quark, b. The bottom quark will hadronize to form a jet of hadrons containing the b quark that may be identified with "b-tagging" (discussed in section 3.4.5). One possible decay channel for the W boson is to a SM u quark and a SM $\bar{d}$ quark, where each will hadronize and may be detected as two hadronic jets. The observation of three jets, one of which is b-tagged, may indicate the presence of the t quark. For any single collision, it was not possible to state with certainty the initial fundamental particles that resulted in the measured reconstructed physics objects. Therefore, the statistical analysis of many collisions was performed.

The ATLAS experiment was comprised of several detection systems and magnet systems in order to detect different kinds of particles. A right-handed cylindrical coordinate system was used with the origin at the center of the detector, and the z-axis aligned with the LHC

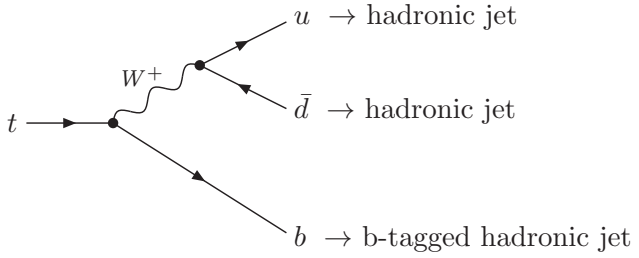


Figure 3.4: In the hadronic decay of the SM top quark, one possible way that physics objects may be detected is three hadronic jets, one of which having properties indicating the presence of a SM bottom quark (b-tagged).

beam line and increasing eastward (from the location of ATLAS), with r the radial distance from the z-axis, and ϕ the azimuthal angle beginning from a horizontal vector directed perpendicular to the beam line and toward the center of the LHC. From these, a polar angle $\theta = \tan^{-1}(r/z)$ was defined, and often the pseudorapidity angle $\eta = -\ln(\tan(\theta))$ was used. When an interaction point was determined from a collision, similar coordinates were defined with the origin located at the interaction point. Also, the angle between two vectors, v_1 and v_2, was often defined as $\Delta R(v_1, v_2) = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2}$.

A tracking system located closest to the beam line measured the trajectory of charged particles, such as electrons, muons and hadronic matter with net charge, moving in a strong magnetic field created by the solenoid magnet system that bent the trajectories of charged particles. The tracking system did not detect uncharged particles, such as photons, and this allowed for differentiating between charged and uncharged particles. Electromagnetic (EM) calorimeters surrounded the tracking system and were designed to capture electrons and photons and measure their energies. Hadronic matter and muons also interacted with the EM calorimeters, aiding in their identification and the measurement of their energies and momenta. Outside the EM calorimeters were hadronic calorimeters that captured and measured the energy of hadronic matter, as well as provided partial measurements of muon kinematics. A large toroidal magnet system surrounded the hadronic calorimeter that bent the trajectories of muons in order to estimate muon momenta, that in turn was surrounded by muon spectrometers that identified muons. This strategy is summarized in table 3.1.

increasing distance from interaction point $\longrightarrow$				
Inner Detector tracking?	Electromagnetic Calorimeter energy deposition	Hadronic Calorimeter energy deposition	Muon Spectrometer detection?	physics object detected
yes	large	very small	no	$\rightarrow$ electron
no	large	very small	no	$\rightarrow$ photon
yes	medium	large	no	$\rightarrow$ hadrons
yes	small	small	yes	$\rightarrow$ muon

Table 3.1: The strategy for detecting physics objects with the ATLAS experiment based on varying interactions with several subdetector systems.

3.3.1. Tracking Charged Particles with the Inner Detector

The Inner Detector (ID) that tracked charged particles surrounded the beam line within $|\eta| \leq 2.5$ was comprised of the Pixel Detector (PD), the Semiconductor Tracker (ST), and the Transition Radiation Tracker (TRT) [85]. The close proximity of the inner detector to the beam line resulted in a high density of tracks requiring fine granularity in order to precisely measure momentum as well as to identify the locations of primary vertices, as well as secondary vertices that often indicated the presence of b quark decays via b-tagging

(discussed in section 3.4.5).

The PD had three barrel layers and end cap disks of semiconductor wafers [86]. The barrel portions immediately surrounded the beam line [85]. Each pixel had 50 μm $\hat{\boldsymbol{\phi}}$ width and either 400 or 600 $\hat{\mathbf{z}}$ μm length that provided a resolution of $\sim$10 $r\hat{\boldsymbol{\phi}}$ μm, and 115 $r\hat{\mathbf{z}}$ μm for the barrel layers and 115 $\hat{\mathbf{r}}$ μm for the end caps [86]. Ionizing charged particles passing through the active semiconductor region created charge carries that were then detected [87]. The wafers were fabricated containing large concentrations of oxygen to increase radiation hardness, with the innermost barrel designed to be replaceable.

The PD was surrounded by the SemiConductor Tracker (SCT) that had four barrel layers of silicon microstrip detectors and nine end cap disks. The four barrel layers extended radially from 30 cm to 51 cm [88]. The SCT utilized similar semiconductor material as the PD, but the much larger volume and desire to minimize cost necessitated the use of long semiconductor strips that decreased resolution on the long dimension. The SCT provided a resolution of $\sim$17 $r\hat{\boldsymbol{\phi}}$ μm and 580 $r\hat{\mathbf{z}}$ μm for the barrel layers and 580 $\hat{\mathbf{r}}$ μm for the end caps [86].

The SCT was surrounded by the Transition Radiation Tracker (TRT). The TRT was further from the beam line than the PD and SCT so that the track density was much lower while the volume occupied by the TRT much greater, and this motivated the selection of lower resolution detection technology at lower cost. Transition (electromagnetic) radiation is created when an ultrarelativistic charged particle crosses the boundary between two materials. Less massive particles generate more transition radiation so that the TRT may distinguish electrons from hadrons. The TRT used polyimide drift tube straws that allowed for high modularity without compromising the continuous tracking concept [89]. Central particles ($|\eta| \leq 2$) crossed 35-40 straws thereby providing the desired continuous tracking. Straws in the barrel region were 144 cm long, and 37 cm in the end caps. The small straw inner radius of 4 mm ensured that the TRT response time was quick enough to meet the 50 ns bunch crossing time. Straw walls providing the cathode were made of layered polyurethane, Kapton, aluminum, graphite-polyimide, and carbon fiber bundles. A gold plated tungsten wire through the center of the straw provided the anode, and the straws were filled with gaseous Xe (70%) for X-ray absorption, CO_2 (27%) to absorb UV photons and stabilize drift velocity, and O_2 (3%) to stabilize amplification. When high energy charged particles passed through the straw gas, neutral gas molecules were ionized leading to a cascade of ionized molecules, and the high mobility free electrons that were liberated drifted in the presence of the applied electric field to the anode where they were collected [90].

The ID was surrounded by the Central Solenoid (CS) magnet system that provided the magnetic field necessary to curve the charged particle tracks for momentum measurements. The superconducting CS was only 19 cm thick to minimize interactions with particles, and supplied a strong central magnetic field of 2.0-2.6 T parallel to the beam line [91, 92]. The combined PD, SCT and TRT measurements allowed for the reconstruction of charged particle tracks, which were then used to determine the location of interaction vertices.

3.3.2. The Electromagnetic, Hadronic and Forward Calorimeters

The ATLAS detector measured the energies of the high energy electrons, photons and hadronic jets created in proton-proton collisions using several sampling calorimeters.

Sampling calorimeters were comprised of alternating layers of absorbing material and active material. The absorbing layers induced electromagnetic and strong interactions with high energy particles, and the active material layers measured the location of the particles and estimated the energy that the particles deposited in the absorbing material. These calorimeters surrounded the inner detector and central solenoid with the Electromagnetic Calorimeter (EMC) surrounded by the Hadronic Calorimeter (HC). The main parameters of the calorimeter systems are provided in table 3.2.

The EMC was designed to capture and measure the energies and locations of electrons and photons. Hadronic jets and muons also interacted with the EMC, but only deposited a small fraction of their total energies in the EMC, and then passed through to the HC. The EMC used ∼1–2 mm thick lead plates sandwiched by 0.2 mm thick stainless steel for its absorbing layers, designed in an accordion geometry to provide full coverage [86]. The active material was liquid argon that was placed between the absorbing sheets with varying thickness ∼2–4 mm depending on location in the system. Sheets of high voltage copper surrounded the liquid argon with a capacitively coupled sheet in the middle to measure drift electrons from the liquid argon ionization.

High energy electrons (or positrons) passing through the lead layers created electromagnetic showers via bremsstrahlung radiation while high energy photons underwent electron-positron creation induced by the lead nuclei, with those leptons then showering via bremsstrahlung. The shower continued as long as the photon energy remained above ∼1 MeV while electron/positron energy loss by ionization eventually dominated bremsstrahlung. As the shower passed through the liquid argon layer, ionization of the argon atoms in the presence of an applied electric field allowed for measurement of the electrons and positrons in the shower.

Bremsstrahlung energy loss due to linear acceleration, a, of particles is proportional to $a^2\gamma^6$, where $\gamma = 1/\sqrt{1 - v^2/c^2}$. For particles having the same energy and experiencing similar electromagnetic forces, but having differing masses, the special relativity formula $E = \gamma mc^2$ implies that the energy loss due to bremsstrahlung is proportion to m^{-6}. The mass of a muon is ∼200 times larger than that of an electron, and the smallest mass of a hadron (π_0) is 260 times the electron mass, so that for particles of similar initial energy, the bremsstrahlung energy loss of an electron is greater than 10^{13} times that of muons and hadrons. This high rate of bremsstrahlung of electrons and photons causes them to shower closer to the beam line (within the EMC) than hadronic particles that traverse the EMC and are collected further from the beam line in the Hadronic Calorimeter.

The Hadronic Calorimeter (HC) was designed to capture and measure the energies and location of hadronic jets. Muons also interacted with the HC, but only deposited a small fraction of their total energies in the HC, and then passed through to the Muon Spectrometer. Jets contained numerous charged hadronic particles, and although these interacted via the electromagnetic force, the energy loss due to bremsstrahlung was negligible (as explained above). Instead, hadronic jets deposited energy in the HC primarily due to strong force interactions between the hadrons and the nuclei of the absorbing material. The end caps of the HC had 25 and 50 mm thick layers of copper plates as the absorbing material. These had 8.5 mm gaps filled with liquid argon as the active media, which were split into four drift areas by parallel electrodes [93]. The barrel region had steel plates for absorbing material and used scintillating tiles for the active material, rather than liquid argon, which lowered

	Barrel		End-cap	
EM calorimeter				
Number of layers and $\|\eta\|$ coverage				
Presampler	1	$\|\eta\| < 1.52$	1	$1.5 < \|\eta\| < 1.8$
Calorimeter	3	$\|\eta\| < 1.35$	2	$1.375 < \|\eta\| < 1.5$
	2	$1.35 < \|\eta\| < 1.475$	3	$1.5 < \|\eta\| < 2.5$
			2	$2.5 < \|\eta\| < 3.2$
Granularity $\Delta\eta \times \Delta\phi$ versus $\|\eta\|$				
Presampler	0.025×0.1	$\|\eta\| < 1.52$	0.025×0.1	$1.5 < \|\eta\| < 1.8$
Calorimeter 1st layer	$0.025/8 \times 0.1$	$\|\eta\| < 1.40$	0.050×0.1	$1.375 < \|\eta\| < 1.425$
	0.025×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.1	$1.425 < \|\eta\| < 1.5$
			$0.025/8 \times 0.1$	$1.5 < \|\eta\| < 1.8$
			$0.025/6 \times 0.1$	$1.8 < \|\eta\| < 2.0$
			$0.025/4 \times 0.1$	$2.0 < \|\eta\| < 2.4$
			0.025×0.1	$2.4 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter 2nd layer	0.025×0.025	$\|\eta\| < 1.40$	0.050×0.025	$1.375 < \|\eta\| < 1.425$
	0.075×0.025	$1.40 < \|\eta\| < 1.475$	0.025×0.025	$1.425 < \|\eta\| < 2.5$
			0.1×0.1	$2.5 < \|\eta\| < 3.2$
Calorimeter 3rd layer	0.050×0.025	$\|\eta\| < 1.35$	0.050×0.025	$1.5 < \|\eta\| < 2.5$
Number of readout channels				
Presampler	7808		1536 (both sides)	
Calorimeter	101760		62208 (both sides)	
LAr hadronic end-cap				
$\|\eta\|$ coverage			$1.5 < \|\eta\| < 3.2$	
Number of layers			4	
Granularity $\Delta\eta \times \Delta\phi$			0.1×0.1	$1.5 < \|\eta\| < 2.5$
			0.2×0.2	$2.5 < \|\eta\| < 3.2$
Readout channels			5632 (both sides)	
LAr forward calorimeter				
$\|\eta\|$ coverage			$3.1 < \|\eta\| < 4.9$	
Number of layers			3	
Granularity $\Delta x \times \Delta y$ (cm)			FCal1: 3.0×2.6	$3.15 < \|\eta\| < 4.30$
			FCal1: $\sim$ four times finer	$3.10 < \|\eta\| < 3.15$,
				$4.30 < \|\eta\| < 4.83$
			FCal2: 3.3×4.2	$3.24 < \|\eta\| < 4.50$
			FCal2: $\sim$ four times finer	$3.20 < \|\eta\| < 3.24$,
				$4.50 < \|\eta\| < 4.81$
			FCal3: 5.4×4.7	$3.32 < \|\eta\| < 4.60$
			FCal3: $\sim$ four times finer	$3.29 < \|\eta\| < 3.32$,
				$4.60 < \|\eta\| < 4.75$
Readout channels			3524 (both sides)	
Scintillator tile calorimeter				
	Barrel		Extended barrel	
$\|\eta\|$ coverage	$\|\eta\| < 1.0$		$0.8 < \|\eta\| < 1.7$	
Number of layers	3		3	
Granularity $\Delta\eta \times \Delta\phi$	0.1×0.1		0.1×0.1	
Last layer	0.2×0.1		0.2×0.1	
Readout channels	5760		4092 (both sides)	

Table 3.2: Main parameters of the ATLAS calorimeters [86].

the total cost since the barrel portion occupied such a large volume [94].

The hadronic matter moving through the copper or steel absorbing material of the HC deposited energy primarily through strong force interactions as the hadrons scattered inelastically from absorber nuclei, leading to a cascade of additional hadronic particles that were measured in the active material. These interactions also led to nuclear excitations and fission in the absorber that were not measured so that energy measurements had to be calibrated to account for this unmeasured energy loss.

The Forward Calorimeter (FC) was located near the beam line where it received high particle flux. It used copper and tungsten for the absorbing material, and liquid argon for the active material, with the liquid argon contained in longitudinal tubes parallel to the beam with smaller drift gaps than the other calorimeters in order to prevent ion buildup due to its high radiation environment [86]. The FC was important for detecting forward hadronic jets used in the search for VLQs, as well as contributing to the measurement of the luminosity of the *p*-*p* collisions.

At high luminosity where multiple inelastic collisions were expected per bunch crossing, luminosity and the number of particles per bunch crossing were estimated from the number of hits per crossing in a each of several detector systems [83]. Lumiblocks denoted the period of time that collision conditions were uniform so that luminosity was constant during each lumiblock. The BCM diamond detector and the LUCID Cherenkov detector located in the extreme forward regions of the ATLAS detector provided per bunch luminosity measurements, while the HC barrel tile calorimeter and the FC gave average rates over longer time periods. These measurements provided the total integrated luminosity used in this study.

Typically, several bunch crossings occurred before a collision of interest was identified, and there were a large number of proton-proton collisions in each bunch crossing, 20.7 on average. This caused difficulty in determining which measurements by the various detectors corresponded only to the collision of interest. The products of the other collisions appeared in the detector elements as *in-time pile-up* (also called the "underlying event"). Additionally, the response time of the majority of detectors was longer than the 50 ns bunch spacing so that the products from collisions in adjacent bunch crossings were also combined with the desired measurements as *out-of-time pile-up* [95]. Pile-up was reduced, but not eliminated, by requiring that the tracks corresponding to each of the measurements traced back to the primary vertex belonging to the collision of interest that was identified by the Inner Detector.

3.3.3. The Toroid Magnet System and the Muon Spectrometer

The Hadronic Calorimeter (HC) was surrounded by barrel and end cap toroidal magnet system that provided toroidal magnetic fields of $\sim$0.5 and $\sim$1 T, respectively. These magnetic fields bent the paths of high energy muons leaving the HC and heading toward the Muon Spectrometer. The curvature of the muon tracks in the inner detector (ID) coupled with the curved trajectory of the muons through the magnetic fields of the toroidal magnet system (with final position determined by the Muon Spectrometer) allowed accurate measurement

Monitored drift tubes	**MDT**
- Coverage	$\|\eta\| < 2.7$ (innermost layer: $\|\eta\| < 2.0$)
- Number of chambers	1088 (1150)
- Number of channels	339 000 (354 000)
- Function	Precision tracking
Cathode strip chambers	**CSC**
- Coverage	$2.0 < \|\eta\| < 2.7$
- Number of chambers	32
- Number of channels	31 000
- Function	Precision tracking
Resistive plate chambers	**RPC**
- Coverage	$\|\eta\| < 1.05$
- Number of chambers	544 (606)
- Number of channels	359 000 (373 000)
- Function	Triggering, second coordinate
Thin gap chambers	**TGC**
- Coverage	$1.05 < \|\eta\| < 2.7$ (2.4 for triggering)
- Number of chambers	3588
- Number of channels	318 000
- Function	Triggering, second coordinate

Table 3.3: The main parameters of the Muon Spectrometer [86].

of the momenta of the muons [92].[2]

The Muon Spectrometer (MS) was composed of several subdetector systems: the Monitored Drift Tubes (MDT), the Cathode Strip Chambers (CSC), the Resistive-Plate Chambers (RPC), and the Thin-Gap Chambers (TGC), with the main parameters given in table 3.3. The MS provided momentum resolution of 3-10% for momenta of 10–1000 GeV/c [96].

The MDT chambers consisted of layers of 30 mm diameter aluminum drift tubes of length 70–630 cm with central tungsten-rhenium anode wire [96], and filled with an Argon (93%) CO_2 (7%) gas mixture [97]. When a charged particle passed through the drift tube, it ionized the gas mixture leading to an avalanche of drift electrons that were collected by the anode. The time difference between the bunch crossing and anode collection was used to estimate the location of the ionizing particle in the tube with $\sim$80 μm resolution.

The CSC was located in the extreme forward regions where large particle flux was present requiring higher rate capability and time resolution [86]. The CSC contained cathode strip chambers in which planes of orthogonal cathode strips made of copper clad RF4 sandwiched a series of tungsten-rhenium anode wires filled with an Ar/CO_2 gas mixture. The orthogonal strips provided location detection in both directions transverse to the muon path, with greater precision in the $\hat{\phi}$ direction. Drift electrons from the ionized gas traveled toward the anode inducing a charge distribution on the cathode strips that was measured. The position of the muon inside the chamber was extrapolated from the relative measurements of neighboring strips.

The RPC and TGC subdetectors were designed for the fast response needed for immediate detection of muon creating events in bunch crossings, as well as providing secondary location

[2] ATLAS is an acronym for "A Toroidal LHC ApparatuS", which emphasizes the importance of the muon measurement capabilities of the ATLAS detector.

measurements for the MS [98]. The resistive plate counter (RPC) had a gap filled with an unpressurized, high UV absorption gas sandwiched between high resistance plastic layers and in the presence of a uniform electric field. When a charged particle ionized the gas, the electric field was switched off in that region preventing discharge to other regions, and photons created in the process were not able to propagate outside the local region due to the gas, and so fast location of muons was possible [99]. The RPC was composed of layers of 2 mm plastic laminate electrodes sandwiching 2 mm of a gas mixture of $C_2H_2F_4$ (94.7%), C_4H_10 (5%) and SF_6 (0.3%) with a time resolution of $\sim$ 1 ns and spatial resolution of only 5–10 mm in the barrel region of the MS [100].

The TGC was comprised of 1.6 mm graphite coated cathode layers sandwiching a 1.4 mm thick layer of a highly quenching gas mixture of CO_2 (55%) and n-C_5H_{12}, and contained anode wires spaced every 1.8 mm [98]. This provided a small drift time with 99% of output signals delivered within 50 ns [101].

3.3.4. Triggering and data storage

Proton bunches were made to cross every 50 ns which corresponded to an event rate of 20–40 MHz. With an average of 20.7 proton-proton collisions per bunch crossing, this gave an approximate collision rate of $\sim$1 GHz. However, collision data could not be saved at a rate faster than 1 kHz. The raw data of a saved event contained a minimum of 1.7 megabytes (not including effects from pile-up), so less than two gigabytes of data was could be saved each second [84, 102]. Therefore, it was necessary to select the most interesting collision event to save from among as many as 40,000 other events (40 MHz/1kHz). This was accomplished with the Trigger and Data Acquisition system (TDAQ) that implemented three stages of filtering.

Events were most interesting when they provided signatures in the detector that indicated the presence of massive particles, such as the top quark, that are predicted to be associated with exotic phenomenon such as vector-like quarks (VLQs), or signatures that indicated extreme energy transfer between the colliding protons, a necessary condition to produce massive particles. These signatures indicated the presence of particular physics objects, such as electrons or muons that indicated weak force interactions, or high p_T hadronic jets, where p_T is the component of the physics object's momentum perpendicular to the beam line [103]. Various trigger requirements were implemented to specify the minimum p_T thresholds and other properties of these physics objects for the collision event to pass each of the three stages of filtering. The Level 1 trigger (L1) reduced the event rate from $\sim$40 MHz to $\sim$70 kHz, the Level 2 trigger (L2) reduced this further to $\sim$ 5 kHz, with the Event Filter (EF) creating several streams of output data for recording at 700–1,000 Hz [84].

The L1 trigger was hardware-based and pipelined to reduce measurement data obtained primarily from dedicated subsystems within the MS, EMC, HC, and FCAL to the Central Trigger Processor (CTP) [104]. The L1 trigger implemented a trigger menu of 256 criteria on the reduced data physics objects, and the data and locations (regions of interest, ROI) of those objects meeting the criteria was sent to the CTP. The CTP made an accept/reject decision for each event within 2.5 μs, and when accepting, signaled that unreduced data was to be read out from the various subdetectors and sent to the L2 trigger.

The L2 trigger was software-based and performed fast computations using more detailed

data contained within the ROIs that had been determined by the L1 trigger [105]. The L2 trigger used a trigger menu similar to the L1 trigger with more refined criteria for event acceptance. When the L2 trigger accepted an event, all data pertaining to that event from the various subdetector systems was sent to the Event Builder (EB) where it was assembled into a single data structure and sent to the EF. All data for events rejected by the L2 trigger were subsequently cleared from the various pipelines and data stores. The EF was software-based and implemented event reconstruction and selection algorithms on the fully built events sent from the DB. Events selected for acceptance were then sent to one or more of six data streams for recording. The Muons and Egamma data streams were used in the Z-tag VLQ search since these streams provided the two electrons or two muons needed to construct the leptonically decaying Z boson candidate.

3.4. Collection of the Measured Dataset

Proton bunches passed one another every 50 nanoseconds, each resulting in several proton-proton collisions. A small fraction of these collisions provided enough momentum exchange between the colliding protons to result in the creation of exotic particles such as Z bosons and Standard Model (SM) t or b quarks. Collision events containing exotic particles were important because the hypothesized vector-like quarks (VLQs) produced these when decaying. Since a VLQ could only be detected through the examination of its decay products, only collision events containing exotic decay products were of interest.

The principal task of the ATLAS detector was to 1) select collision events containing exotic decay products and 2) determine the identity and properties of the decay products of each collision event. However, the ATLAS detector did not directly detect particles, but rather their interactions with the various detector components. The resultant detector responses after a collision event were then analyzed to determine the nature of the particles that caused these responses. In other words, the properties of the physics objects (electrons, muons, quarks, etc.) had to be reconstructed from the measured data. Fast reconstruction algorithms were necessary for determining the presence of a collision event of interest every 50 ns, while more elaborate reconstruction was performed offline to provide the best estimates of the positions, energies and momenta of the candidate physics objects.

This section describes the reconstruction of physics objects that was performed so that the measured dataset used for the VLQ search contained a collection of collision events of interest, where each event provided estimates of the locations, energies and momenta of the candidate electrons, muons, etc. Table 3.4 shows the data recorded during 2012 in the Egamma (e/γ) and Muons data streams. All 1.4 billion collision events were examined to determine which were most likely to indicate the presence of the VLQ signals.

3.4.1. Determination of the Collision Events Primary Vertex

The locations of the proton-proton collisions in a single bunch crossing were estimated using the measured tracks and extrapolating them to a common point. This yielded a collection of interaction "vertices," with different sets of tracks pointing to each of the vertices. This used the vertex fitting algorithm that implemented requirements on the quality of the tracks including the number of hits in the inner detector, minimum track energies, the distance from the candidate vertex, and the uncertainty on that distance [106–108]. A large number of proton-proton interactions were observed for each bunch crossing, which resulted in many vertices obtained per event.

The location of the collision event of interest, the primary vertex, was determined from the candidate vertices using the primary vertex finding algorithm that selected the vertex with the set of tracks providing the largest sum of estimated transverse momenta, $\sqrt{\Sigma_{track} p_T^2}$ [106–108]. The primary vertex was then used to select the physics objects belonging to the collision of interest while removing the remainder of detector responses from consideration. In this manner, the collision event of interest was isolated from the other collisions of the bunch crossing.[3]

[3]The majority of "minimum bias" events obtained from loose trigger requirements were from soft collisions [109].

data period	collection dates	total integrated luminosity (fb^{-1})	N_{events} e/γ	N_{events} muons
A	Apr 04–Apr 20	0.91	43,244,708	43,661,420
B	May 01–Jun 18	5.59	177,569,377	163,597,791
C	Jul 01–Jul 24	1.64	51,382,952	80,935,368
D	Jul 24–Aug 23	3.60	113,174,461	112,743,915
E	Aug 23–Sep 17	2.86	86,690,531	81,667,914
G	Sep 26–Oct 08	1.40	43,857,431	41,182,058
H	Oct 13–Oct 26	1.65	53,569,883	49,412,952
I	Oct 26–Nov 02	1.15	36,739,595	34,765,048
J	Nov 02–Nov 26	2.94	95,315,250	88,410,196
L	Nov 30–Dec 06	0.98	30,689,178	29,040,147
A–L	Apr 04–Dec 06	22.72	732,233,366	725,416,809
			1,457,650,175	

Table 3.4: The measured data collected in 2012: the number of collision events recorded by data period in the e/γ and muons data streams.

3.4.2. Determination of Electron Candidates

Candidate electrons were reconstructed using the energy deposited in the Electromagnetic (EM) calorimeter. The energy deposited by an electron appeared as a localized cluster within the EM calorimeter cells, with the cluster location corresponding to the measured track position that the electron left as it passed through the inner detector. These electron candidates were required to have large transverse momentum, $p_T \geq 25$ GeV/c, and be located with a pseudorapidity of $|\eta_{cluster}| < 2.47$ as well as $1.37 < |\eta_{cluster}| < 1.52$ to avoid the transition region between the barrel and end-cap portions of the EM calorimeter. The candidate electron track measured by the inner detector was required to have less than a 2 mm longitudinal impact parameter to the selected primary vertex.

The Z-tag search for VLQs utilized both a dilepton channel (two leptons) and a trilepton channel. The work presented here only investigated the dilepton channel: the collision events containing exactly two electrons or else exactly two muons from the decay of a Z boson.[4] The two electron candidates were required to only meet the medium quality criteria because their combination was further required to provide an acceptable Z boson candidate. The medium quality requirements were defined as follows [110].

- The cluster had to meet shower shape criteria in both the first and second layer of the EM calorimeter, as well as hadronic leakage criteria in the hadronic calorimeter.
- The track and cluster had to meet matching criteria including a pseudorapidity difference no greater than 0.0005.
- The track had to include a hit in the innermost layer of the pixel detector to reject electrons created from photon conversion.

[4] In experimental particle physics, the term "electron" is often used to refer to both the electron and the positron, and the term "muon" for both the muon and antimuon.

- The transverse impact parameter had to be less than 5 mm.

Collision events containing electrons were selected by using a combination of electron triggers. The first trigger accepted events that contained an electron candidate with $p_T \geq 24$ GeV/c with its energy deposit in the EM calorimeter well isolated from responses in the hadronic calorimeter positioned behind the EM calorimeter. This isolation requirement reduced the amount of "faking" by non-electron causes, but reduced the ability to detect high energy electrons that would often "punch through" the EM calorimeter and enter the hadronic calorimeter. VLQs were predicted to often produce these high p_T electrons, and so the second trigger applied $p_T \geq 60$ GeV/c without any isolation requirements. Finally, the triggering electron candidate had to meet the above requirements (i.e., be matched to one of the two electrons in the dielectron event).[5]

3.4.3. Determination of Muon Candidates

Candidate muons were reconstructed by matching the track segments detected in the muon spectrometer to the tracks measured by the inner detector [111]. The muon candidates were required to have large transverse momentum, $p_T \geq 25$ GeV/c, be located with a pseudorapidity of $|\eta_{muon}| < 2.5$, and also meet several hit requirements regarding the inner detector (the tight quality criteria). The candidate muon track measured by the inner detector was required to have less than a 2 mm longitudinal impact parameter to the selected primary vertex. This was done to remove from consideration the muons created in the upper atmosphere that traveled through the earth to the ATLAS detector (i.e., rejecting cosmic muons).

Because the muon spectrometer lied furthest from the beam line, any muons created by the decay of hadronic particles had to be rejected. This was achieved by requiring that muons were well isolated from selected jets with $\Delta R \geq 0.4$. Additionally, a track-based isolation requirement with variable cone radius, r_{cone}, as a function of muon p_T was required (the "mini-isolation" criteria). This was defined by setting $r_{cone} = (10\,\text{GeV/c})/p_T$, and requiring that the scalar sum of the transverse momenta of all tracks, t_i, not associated with the muon candidate and lying within $\Delta R\,(\mu, t_i) \leq r_{cone}$ was no more than 5% of the muon p_T.

Collision events containing muons were selected by using a combination of muon triggers. The first trigger accepted events that contained a muon candidate with $p_T \geq 24$ GeV/c and also met isolation requirements, while the second trigger required $p_T \geq 36$ GeV/c without isolation. For these dimuon events to be accepted, the triggering muon candidate had to be matched to one of the two muon candidates meeting the above requirements.

[5] Jet candidates were selected after selecting *initial* electron candidates, and jets were rejected if found within $\Delta R < 0.2$ of an initial electron because jet reconstruction also created jet candidates from real electrons (i.e., false jets were removed as a first overlap removal step). *After* the jet candidates were determined, any initial electron within $\Delta R < 0.4$ of a jet was rejected to ensure that real electrons were not secondary products in the decay of hadronic material (i.e., real secondary electrons were removed as a second overlap removal step).

3.4.4. Determination of Hadronic Jets

A proton-proton collision of interest was a hard scatter process between the constituent partons (quarks and gluons) of each proton. The hard scattering between partons was mediated by the strong force, and so most often produced outgoing partons as a result.

Partons cannot exist independently in nature, and so they hadronized into color-neutral multi-parton particles, hadrons, after their creation (discussed in section 2.1.5 on page 37). Additionally, partons radiated other partons through additional strong force interactions, often called "parton showering" as initial and final state radiation). A single high energy proton-proton collision at the LHC showered to produce hundreds to thousands of outgoing particles [112], where an energetic outgoing quark might have eventually produced hundreds of outgoing hadrons through a cascade of hadronization [113].

At least two Standard Model quarks were expected to be produced in collision events containing VLQs, and so the identification and measurement of these quarks was necessary. The spray of hadrons emitted by an outgoing parton was observed as a collection of localized energy deposits in the calorimeters. The anti-k$_t$ jet reconstruction algorithm was used to combine these readings into a single estimate of the position, energy and momentum of the original outgoing parton [114]. The anti-k$_t$ algorithm was found to be infrared safe so that contributions from less energetic radiated hadrons did not much alter the final jet estimates. This algorithm was also shown to be collinear safe so that clusters originating from the same parton were unlikely to be separated into more than one jet, and to be robust against pile-up, detector noise and underlying event backgrounds [114].[6] The anti-k$_t$ algorithm was applied in the VLQ search using a clustering parameter of 0.4, and this yielded a collection of jet physics objects for each collision event.

The collection of jets provided by the anti-k$_t$ algorithm for each event was filtered for only those most likely originating from the collision of interest, the candidate jets. A track-based requirement was used to remove jets with transverse momentum less than 50 GeV/c that originated from secondary proton-proton collisions (jets with $p_T > 50$ GeV/c were not made to meet this requirement). This was achieved by selecting only such a jet where the sum of the transverse momenta of its tracks associated with the primary vertex was larger than 50% of the total transverse momenta of its tracks [118]. This was only applicable to jets that passed through the inner detector, i.e., $|\eta_{jet}| < 2.5$. "Central jets" were defined as those with $p_T \geq 25$ GeV/c and located with $|\eta_{jet}| < 2.5$, while "forward jets" were those with $p_T \geq 35$ GeV/c and $2.5 < |\eta_{jet}| < 4.5$. Only central jets were used in the VLQ search for VLQs produced in pairs, but singly produced VLQs were predicted to frequently decay to quarks with high $|\eta_{jet}|$ so that forward jets were used in addition to central jets in the search for singly produced VLQs. Finally, jets found too near any electron candidate were removed from consideration with the requirement $\Delta R(jet, electron) > 0.2$.

3.4.5. Identification of Jets Containing b-quarks — b-tagging

At least two Standard Model bottom quarks were expected to be produced in collision events containing VLQs. The identification of jets originating from hadronized b quarks was

[6] The term "underlying event" is used to refer to the final state particles created by proton-proton collisions other than the collision of interest [115–117].

useful for identifying collision events most likely due to the presence of VLQs, as well as allowing the reconstruction of the VLQ invariant mass as a powerful discriminating variable. A b-tagging algorithm was used to identify these jets from the central jet candidates. This algorithm was estimated to be only 70% efficient at correctly identifying true b quarks, and so when at least two b-tagged jets were required for collision events, only ∼50% of events containing VLQs were expected to be accepted. Nevertheless, requiring events to contain two b-tagged jets was found to remove a much larger fraction of the events due to SM backgrounds, ∼98%, making this requirement very useful for the VLQ search.[7]

As explained in section 2.1.4, the b quark had a long lifetime, and the B hadrons it hadronized into had a lifetime of ~ 1.5 ps [5]. This allowed the B hadron to travel ~ 5 mm on average from the primary vertex before decaying at a secondary vertex. The tracks of the candidate jets were extrapolated back to their interaction points, and those jets exhibiting such a secondary vertex were more likely to have originated from B hadron decay.

The MV1 b-tagging algorithm was used to identify jets originating from b quarks. This algorithm implemented an artificial neural network learning algorithm that used the outputs of three other b-tagging algorithms as its inputs [119, 120]. These were the IP3D b-tagging algorithm based on impact parameters, the SV1 b-tagging algorithm that examined the properties of reconstructed secondary vertices, and the JetFitter b-tagging algorithm that reconstructed decay chains with a Kalman filter, and where all three of these implemented a likelihood ratio technique [121]. By combining the results of three separate algorithms using machine learning, the MV1 b-tagging algorithm provided superior performance compared to any one particular algorithm.

3.4.6. Event Preselection

A collision event of interest was one containing physics objects most consistent with the decay of VLQs. This included selecting the primary vertex, reconstructing the candidate electrons, muons and jets, and requiring the electron or muon event trigger to be matched to the corresponding physics object (as described in the previous subsections). This was done for collision events with $\sqrt{s} = 8$ TeV center of mass energy that was collected in 2012 providing a measured dataset representing 20.3 ± 0.6 fb^{-1} of total integrated luminosity (discussed in section 3.3.2 on page 59).

VLQs possessed decay products that included at least three SM quarks, and so were expected to contain at least three jets in any collision event. Therefore, collision events were rejected if they did not contain at least two central jet candidates. All of these requirements determined the preselection region for the VLQ search. Additional requirements were later imposed as part of the square cuts search strategy that is described next in chapter 4.

[7] Additionally, only 20% of jets containing a c quark were falsely accepted as b-tagged while $< 0.8\%$ of those deriving from the lighter u, d and s quarks were falsely b-tagged.

3.5. Construction of the Simulated Datasets

A new theory of particle physics does not usually replace the established Standard Model (SM), but rather builds on it by predicting new particles in addition to those predicted by the SM. These beyond the Standard Model (BSM) theories that predicted vector-like quarks (VLQs) were described in chapter 2. The SM background of the VLQ search was simulated via Monte Carlo using only the predictions of the SM theory. The VLQ signal was obtained from simulations of the BSM theory.

The search for VLQs was performed by 1) comparing the simulated SM background to the actual measurements, and 2) comparing the simulated SM background *plus* the simulated VLQ signal to the measurements. If comparison 1) showed that the measured data was more consistent with the SM backgrounds, then this provided evidence for VLQ signal nonexistence. If 2) was more consistent, then this demonstrated evidence for the existence of VLQs.[8]

There were four stages of Monte Carlo simulation performed to obtain a simulated dataset. First, parton distribution functions were used to estimate the longitudinal momenta of the incoming partons within the colliding protons. Second, tree-level event generators simulated the outgoing partons according to the theory being used. Third, the showering and hadronization of these outgoing partons were simulated to provide the final state particles, the decay products of the collision. Finally, the response of the ATLAS detector to these outgoing particles was simulated.

Because the proton-proton collisions were actually strong force interactions between the partons contained in the two colliding protons, it was necessary to estimate the longitudinal momentum of these partons. I.e., the total momentum of the two colliding protons was not available for momentum transfer, only the fraction of the total momentum possessed by each of the two colliding partons. Parton distribution functions provided estimates of the probability of each parton to posses a given fraction of the total proton longitudinal momentum.

The parton density function (PDF), $f_i(x, Q^2)$, gave the probability of a parton with flavor i (u quark, $\bar{u}$ quark, d quark, g gluon, etc.) to possess fraction x of the total proton longitudinal momentum at an energy scale Q of the strong force interaction [122]. Accurate modeling by the PDF was most important for estimating the soft scattering events that caused the large background of the underlying event [123].

A tree-level event generator simulated the strong interactions between incoming partons and computed the cross sections for a fixed number of outgoing partons. I.e., it computed the matrix elements of a Feynman diagram representing the possible result of a parton-parton collision. Tree-level event generators did not model parton showering nor hadronization, but often included higher order terms that would be incorrectly modeled by the parton showering models alone. For example, the `MADGRAPH 5` [124] software that was used in the VLQ search was able to compute next-to-leading order terms, and correctly interface these with parton showering software to avoid double-counting [124].[9]

Parton showering and hadronization software was used to take the outgoing partons from

[8] The final inference result was not a binary decision regarding VLQ signal existence, but rather a summary of the evidence provided for both existence and nonexistence (described in detail in chapter 6).

[9] Also see a list of generators on page 20 of [125].

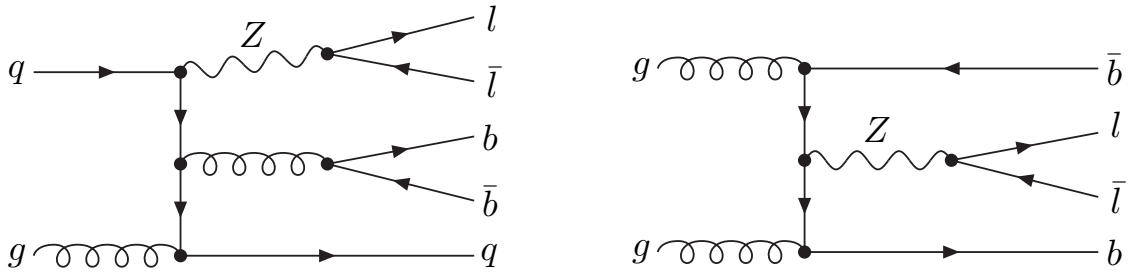
the tree-level event generator and compute the higher-order corrections of partonic cascades and the subsequent hadronization of these showered partons (and the possible decays of any unstable outgoing hadrons) [126].

Finally, accurate modeling of the response of the ATLAS detector to the outgoing final state particles was necessary. This included modeling how the outgoing particles interacted with the material of each detector component, including cables, support racks, etc., whose exact position had to be known so that accurate estimates of the measured responses could be provided. For example, the GEANT4 software used in the VLQ search modeled the geometry of ATLAS materials, various particle interactions with each type of material, detector responses to these interactions, and the transmission of the responses through the data pipelines [127].

3.5.1. Simulated Backgrounds

The Z-tag search examined the existence of VLQs having a Z boson as a decay product. Although not a requirement for preselection (and discussed fully in the next chapter), the Z-tag search required each collision event to possess two same-flavor opposite-sign leptons, l and $\bar{l}$, (electrons or muons) such that their summed four-momenta provided a Z boson candidate, $Z^\mu = l^\mu + \bar{l}^\mu$. This requirement was important in the selection of pertinent SM backgrounds.

SM backgrounds that produced Z boson as a decay product ("Z+*jets*") were examined since these often had decay products similar to those of the VLQs. Two examples of the production of the Z+*jets* background are shown in figures 3.5 (a) and (b). The Z+*jets* backgrounds were generated with SHERPA v.1.4.1 [128] with the CT10 PDFs using massive b and c quarks [129], and were normalized to next-to-next-to-leading order using the DYNNLO software [130]. The Z+*jets* backgrounds were separated into three categories, Z+*bottom*, Z+*charm*, and Z+*light*, by filtering the simulated background events into those having a b quark decay product, those without a b but having a c, and those without a b or c, respectively.[10]



(a) $q + g \to Z$+*jets* (b) $g + g \to Z$+*jets*

Figure 3.5: Two examples of Z+*jets* production containing two b quark final states. These were of particular importance in the Z-tag search for VLQs, which required the presence of two or more b-tagged jets.

[10] Additionally, in-time and out-of-time pile-up was simulated from minimum bias events using PYTHIA6 with the AMBT2B CTEQ6L1 PDFs [131].

Diboson events were SM backgrounds that produced two Z bosons, one Z and one W boson, or else two W bosons, and so were also examined since they had decay topologies similar to the VLQ signals. The diboson background was generated with SHERPA [128] using the CT10 PDFs [129] normalized to next-to-leading order with MCFM [132]. Similarly, *triboson* SM backgrounds were also examined. These were generated with MADGRAPH [124] interfaced with PYTHIA 6 [112, 133] using the CTEQ6 PDFs [134]. SM t quarks produced in pairs with an additional Z or W boson, the "$t\bar{t}X$" background, were also examined as a background that competed with the VLQ signals. The $t\bar{t}X$ background was generated with MADGRAPH 5 [124] using PYTHIA [112, 133] for parton showering and hadronization and normalized to next-to-leading order [135]. Figures 3.6 (a)–(d) show examples of the production of these backgrounds.

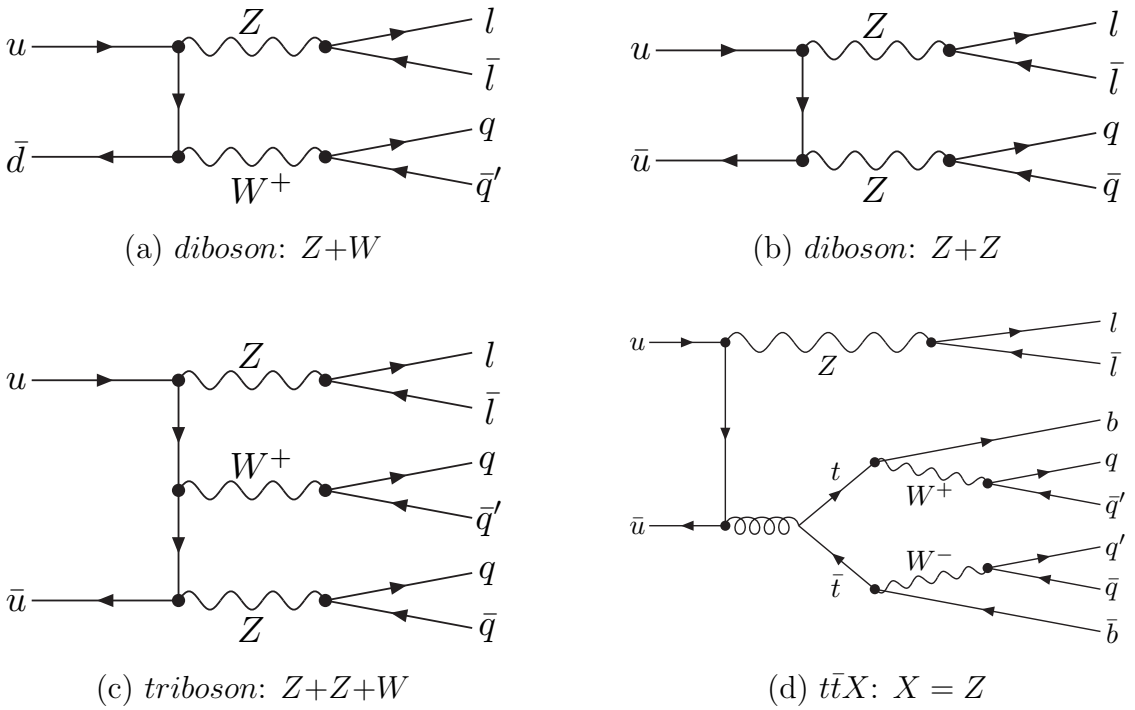
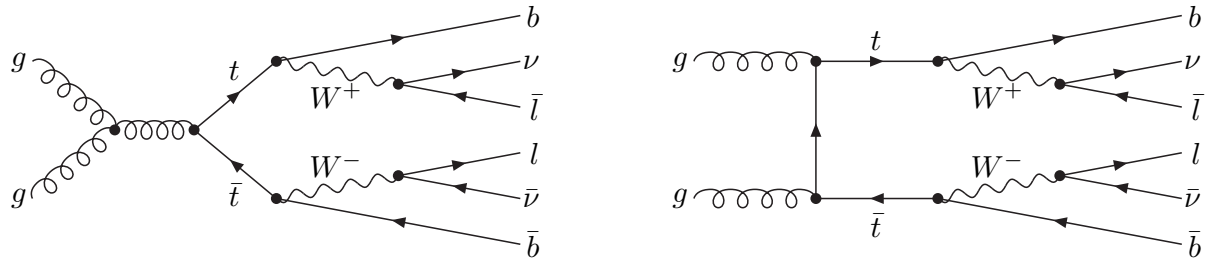


Figure 3.6: Examples of *diboson*, *triboson* and $t\bar{t}X$ production that resulted in the production of at least one Z boson.

SM backgrounds that did not have a Z boson as a decay product, but that had two leptons were also examined because the sum of the four-momenta of the two leptons were frequently confused as a Z boson candidate. The pair produced SM top quarks, "$t\bar{t}$", was one such background. (Other such backgrounds included the *diboson* and *triboson* backgrounds described above that had two or three W bosons, as well as *singletop* backgrounds with an associated W boson described below.) Figures 3.7 (a) and (b) show two examples of $t\bar{t}$ background production where both t quarks decay leptonically. The $t\bar{t}$ background was generated with POWHEG [136] using PYTHIA v.6.421 [112, 133] and the CT10 PDFs [129].

SM backgrounds that did not have a Z boson decay product, but had a single lepton as a decay product were also examined because infrequently a second "fake" lepton was measured



(a) $g + g \to t\bar{t} \to l + \bar{l} +$ other (b) $g + g \to t\bar{t} \to l + \bar{l} +$ other

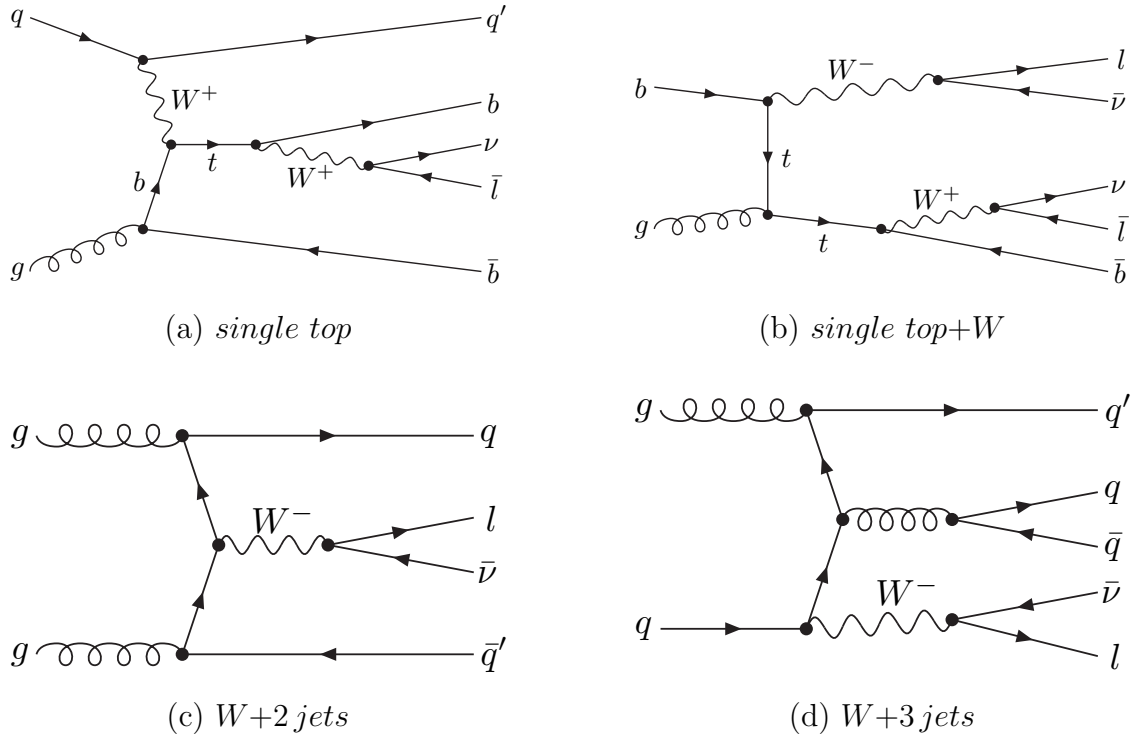
Figure 3.7: Two examples of $t\bar{t}$ production that result in two outgoing same-flavor opposite-sign leptons that may provide an "accidental" Z boson candidate.

by the detector, and the combination of the real and fake lepton produced a faked Z boson candidate. SM t quarks produced singly were examined, the "*single top*" background. Also, singly produced t quarks with an additional W boson frequently produced two real leptons that resembled a Z boson candidate. The *single top* backgrounds were generated with 1) MC@NLO [137, 138] using HERWIG 6.520.2 [139] for s-channel and t+W production, and 2) AcerMC [140] with PYTHIA [112, 133] for the t-channel, where both were normalized to next-to-leading order [141]. SM backgrounds that produced a single W boson infrequently contained an additional fake lepton leading to a faked Z boson candidate. These "W+*jets*" backgrounds were generated with ALPGEN [126] using PYTHIA [112, 133]. Figures 3.8 (a)–(d) show examples of the production of these backgrounds.

SM backgrounds interacting solely via the strong force (i.e., without the weak force to generate any true leptons), the "*QCD*" background, were examined because these infrequently contained two fake leptons that led to an faked Z boson candidate. The *QCD* background was not simulated, but instead estimated by a data-driven method that is described in the next chapter.

3.5.2. Simulated Signals

The pair produced VLQ signal datasets, T_{pair} and B_{pair}, were simulated at leading order using PROTOS v.2.2 [42, 142, 143], interfaced with PYTHIA v.6.421 [112, 133] to model parton showering, and using the MSTW 2008 LO PDFs [144]. The signal cross sections were normalized using the TOP++ v.2.0 to next-to- next-to-leading order (NNLO) including next-to-next-to-leading logarithm soft gluon resummation [145]. The T_{single} signal was simulated with MADGRAPH 5 [124] and validated with PROTOS v.2.2 [42, 142, 143]. This was interfaced with PYTHIA 8 [146] for parton showering, and used CTEQ6L1 PDFs [134]. The B_{single} signal was simulated with PROTOS v.2.2 [42, 142, 143] interfaced with PYTHIA 6 [112, 133] using the MSTW 2008 LO PDFs [144]. The response of the ATLAS detector was simulated using ATLFAST2 [147, 148] and validated using GEANT4 [127].



(a) *single top* (b) *single top+W*

(c) W+2 *jets* (d) W+3 *jets*

Figure 3.8: Examples of *single top* and W+*jets* production that resulted in at least one lepton decay product that was infrequently combined with a fake lepton to give a faked Z boson candidate.

Chapter 4
The Square Cuts Search

A *square cut* on a variable was a requirement for a collision event to be accepted if its value of the variable was above (or below) some threshold value. A square cut requirement determined a *region of interest*, a subspace of the space spanned by the possible values of the collision event's variables. A *signal region* was a region of interest that was expected to contain a large amount of collision events due to the production of vector-like quarks (VLQs) and a small amount of collision events due to the Standard Model (SM) backgrounds. A *control region* was a region of interest containing a majority of collision events due to one or more particular SM backgrounds, and containing a negligible contribution from the VLQ signal. Signal regions were used to determine the possible existence of VLQs, while control regions were used to correct any mismodeling of the simulated SM background datasets by comparing them to the measured datasets. The square cuts search strategy involved identifying the optimal set of square cut requirements to use in determining the signal and control regions.

An advantage of a square cuts search strategy was that it could be performed more quickly than a search using machine learning. An important task in the VLQ search was to validate the modeling of the simulated SM backgrounds so that valid conclusions could be drawn from statistical inference in the signal regions.[1] The use of machine learning increased the complexity of this task by requiring more validation of the simulated data, both of all the variables used as inputs to the learning algorithms, as well as the output predictor functions produced by the learning algorithms. A disadvantage of the square cuts search was the loss of signal-background discrimination power, which led to weaker conclusions regarding the possible existence of VLQs.

In addition to identifying signal regions, the Z-tag search strategy also identified a *discriminating variable* that was used for spectral discrimination in order to increase the signal-background discrimination power of statistical inference. A simple counting search would have involved simply comparing the total predicted yields of the simulated datasets to the measured yields in the signal region.[2] Spectral discrimination examined how those yields were distributed across the spectrum of a discriminating variable. Spectral discrimination required greater modeling accuracy of simulated data, but provided much stronger inference results. With regard to the work presented here, a *search strategy* was both 1) the determination of a signal region defined by square cut requirements and 2) the identification of a powerful discriminating variable for use in the spectral discrimination approach to statistical inference.

[1] Parametric statistic inference in the form of the Neyman construction of confidence regions was used to compute the upper limits of the cross sections of the VLQ signals at a 95% confidence level. This procedure is described in depth in chapter 6.

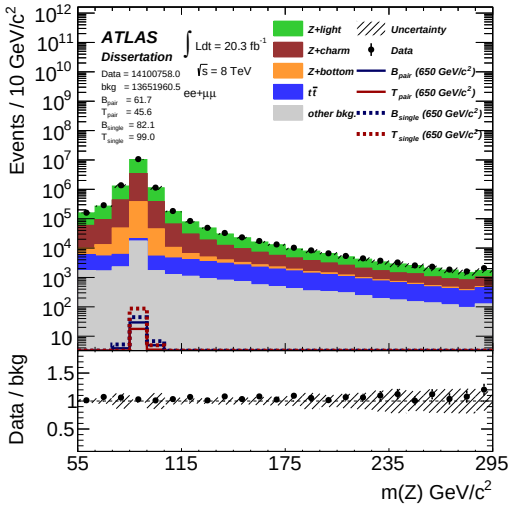
[2] An expected yield is the expected number of collision events predicted by the simulated datasets in a region of interest, and a measured yield is the number of collision events actually measured by the detector.

Several square cuts searches were performed by the ATLAS exotics Z-tag search team in the search for VLQs [41, 61–64, 149, 150]. In those searches, events containing either two leptons or else three leptons provided two orthogonal search channels (the dilepton and trilepton channels, respectively). The focus of this dissertation was the investigation of the use of machine learning methods to improve the searches in the dilepton channel, and so this chapter is restricted to the square cuts searches in only the dilepton channel. Details of the searches using the trilepton channel may be found in [41, 61–64].

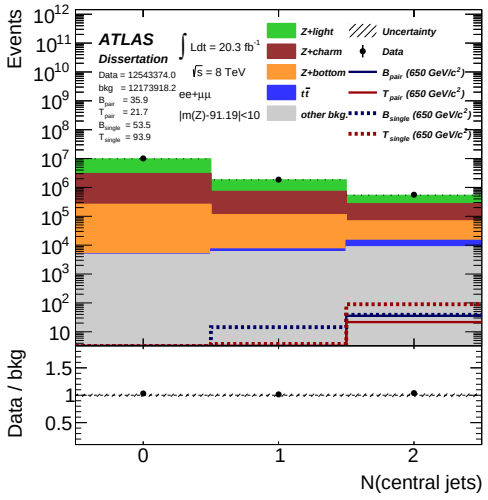
4.1. Square Cuts Search Strategy

After the preselection described in section 3.4, the dilepton Z-tag search began by adding the four-momenta of the two same-flavor, opposite-sign leptons to obtain the four-momentum of the Z-boson candidate, $Z^\mu = l^\mu + \bar{l}^\mu$. The invariant mass of the Z boson was then estimated as $m(Z) \equiv \sqrt{\eta_{\mu\nu} Z^\mu Z^\nu}$, where $\eta_{\mu\nu}$ was the Minkowski metric, and the transverse momentum was given by $p_T(Z) \equiv [\,(Z^1)^2 + (Z^2)\,]^{1/2}$.

Figure 4.1 (a) shows the expected and measured yields of $m(Z)$ after preselection. These demonstrated that collisions producing Z bosons had values of $m(Z)$ tightly centered near 91.19 GeV/c^2. This motivated imposing a mass window requirement, $81.19 < m(Z) < 101.19$ GeV/c^2, as square cut 1 to remove collision events not likely to contain a Z boson. Figure 4.1 (b) shows the multiplicity of central jets after imposing the mass window square cut. All VLQ signals produced at least two central jets so that adding the requirement $N(central\ jets) \geq 2$ as square cut 2 removed a large fraction of the SM background.



(a) After preselection.



(b) After $81.19 < m(Z) < 101.19$ GeV/c^2.

Figure 4.1: (a) The $m(Z)$ yields that motivated the 20 GeV/c^2 mass window requirement. (b) The central jet multiplicity that motivated the $N(central\ jets) \geq 2$.

The square cuts search strategy at this point was described by the following requirements that reduced the number of background collision events by 96% (with little impact on the expected signal distribution).

- **Square cut 0** – Preselection requirements of section 3.4 (14+ million events).
- **Square cut 1** – $81.19 < m(Z) < 101.19$ GeV/c^2 (12.5+ million events).
- **Square cut 2** – $N(central\ jets) \geq 2$ (500,000+ events).

The Z boson emitted from a decaying VLQ was expected to be highly energetic and thus have a large transverse momentum, $p_T(Z)$. Furthermore, all VLQ decay products included at least two b quarks, and so VLQs could be detected by the presence of b-tagged jets. Figure 4.2 (a) shows the yields of $p_T(Z)$, and figures 4.2 (b)–(d) show these yields with different b-tagging requirements. The predicted yields of the 650 GeV/c^2 signal masses of the four VLQ signal types are also shown for comparison with the SM backgrounds. These indicated that the greatest signal-background discrimination occurred with large $p_T(Z)$ and $N(b\text{-}tags) \geq 2$. Thus, two additional square cut requirements were included in the search strategy that further reduced the background by 99.9%.

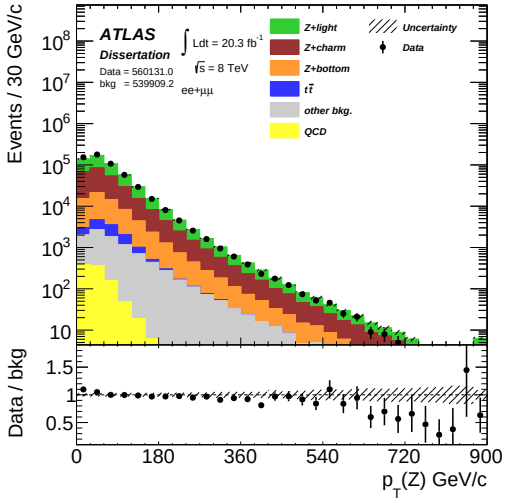
- **Square cut 3** – $p_T(Z) > 150$ GeV/c (35,000+ events).
- **Square cut 4** – $N(b\text{-}tags) \geq 2$ (∼500 events).

These figures also show that there was mismodeling of the shape of the $p_T(Z)$ yields, where too much simulated background was located at larger $p_T(Z)$. This indicated that corrections to the simulated backgrounds would be necessary. These also showed somewhat more measured events than simulated background events, and although the excess of measured data was within the range of uncertainty of the simulated background, this suggested that the QCD background should to be taken into account.[3] The corrections to the simulated backgrounds and the estimation of the QCD background is discussed in section 4.2 (page 84).

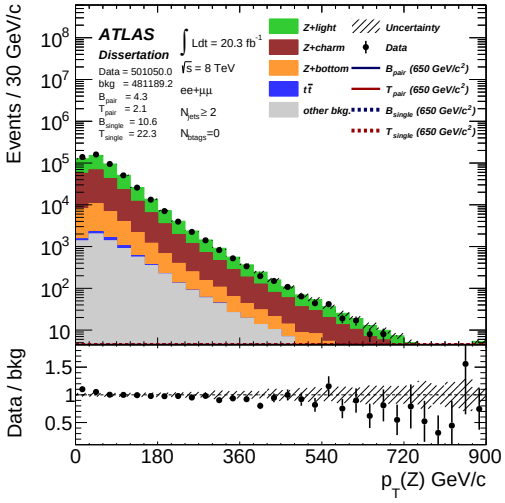
The decay of VLQs usually resulted in the detection of many high energy central jets. The sum of the transverse momenta of these jets, $H_T(jets) = \sum_i p_T(jet_i)$, was expected to be large as a result. Figure 4.3 (a) shows the expected yields of the $H_T(jets)$ variable. These show that the background yields decreased quickly with increasing $H_T(jets)$. The pair produced signal types, T_{pair} and B_{pair} (solid lines), had large values of $H_T(jets)$, and square cut 5a was defined as $H_T(jets) > 600$ GeV for T_{pair} and B_{pair}. However, T_{single} and B_{single} (dotted lines) had smaller values of $H_T(jets)$ that were similar to those of the background, and so square cut 5a could not be used for these.

Figures 4.4 (a) and (b) show the decay topologies of T_{single} and B_{single}, respectively. For these signal types, the outgoing light quark was expected to be located at larger pseudorapidity, η_q, often outside the range that defined the central jets, $\eta_q > 2.5$. Such quarks were expected to be detected as forward jets in the detector.

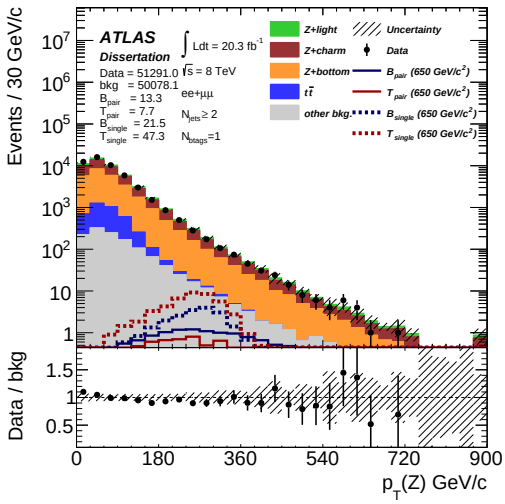
[3]The QCD background was generated by strong force interactions that did not create real leptons. However, these events were produced in such large numbers that their decay products were frequently mistaken as leptons and provided a "fake" Z boson candidate.



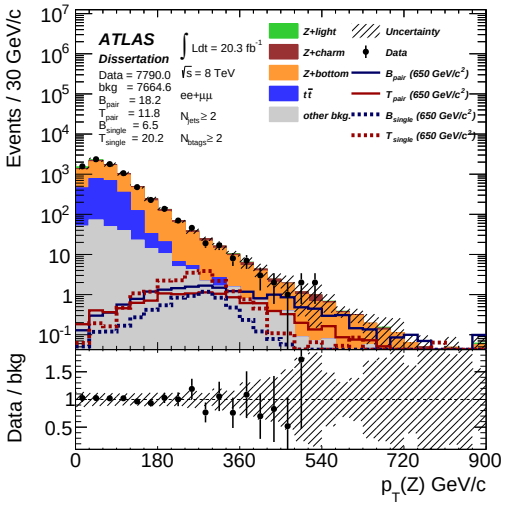
(a) Any $N(b\text{-}tags)$



(b) $N(b\text{-}tags) = 0$.

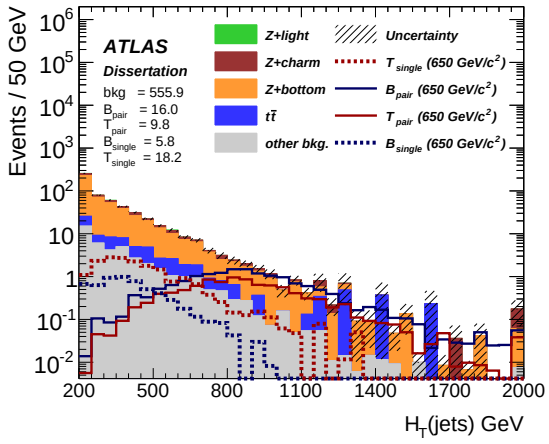


(c) $N(b\text{-}tags) = 1$.

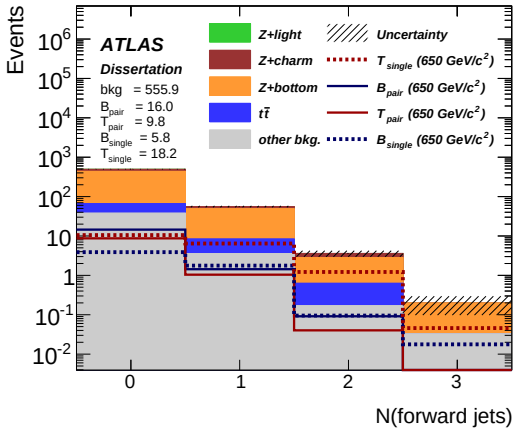


(d) $N(b\text{-}tags) \geq 2$.

Figure 4.2: (a) The $p_T(Z)$ yields after square cuts 1 & 2, and (b)–(d) $p_T(Z)$ for different $N(b\text{-}tags)$ requirements.

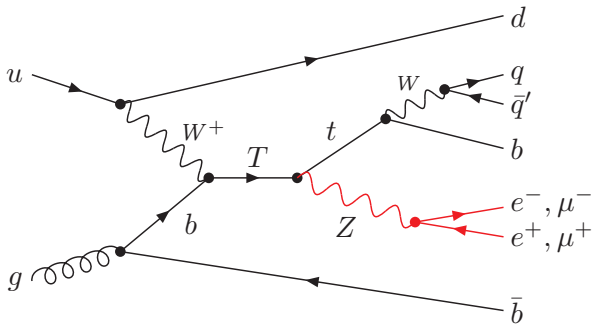


(a) $H_T(jets)$.

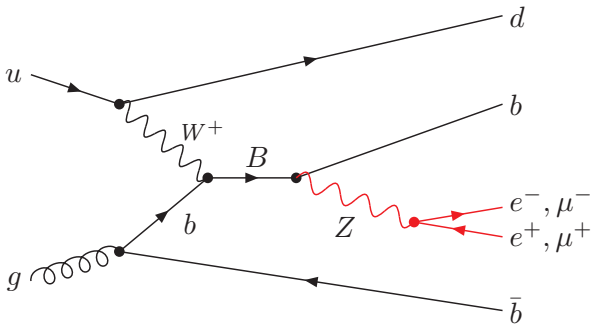


(b) $N(forward\ jets)$.

Figure 4.3: (a) The $H_T(jets)$ yields, and (b) the $N(forward\ jets)$ yields, both after square cut requirements 1–4 (with measured yields blinded to prevent experimenter bias).



(a) T_{single}



(b) B_{single}

Figure 4.4: The decay topologies of the T_{single} and B_{single} signal types, where the outgoing light quark (the d quark at the top right) was expected to be detected as a forward jet.

Figure 4.3 (b) shows the expected yields for the forward jet multiplicities. These indicated that the T_{single} and B_{single} collision events often contained a forward jet, and so square cut 5b was defined as $N(forward\ jets) \geq 1$ for T_{single} and B_{single}.

These final square cuts completed the design of the square cuts signal regions.

- **Square cut 5a (T_{pair} and B_{pair})** – $H_T(jets) > 600$ GeV/c (~35 events).
- **Square cut 5b (T_{single} and B_{single})** – $N(forward\ jets) \geq 1$ (~60 events).

Table 4.1 summarizes the two signal regions of the square cuts search strategy (one for pair produced and one for singly produced).

square cut	requirement	background yield
0	preselection (see section 3.4)	14,000,000
1	$81.19 < m(Z) < 101.19$ GeV/c^2	12,500,000
2	$N(central\ jets) \geq 2$	500,000
3	$p_T(Z) > 150$ GeV/c	35,000
4	$N(b\text{-}tags) \geq 2$	500
5a (T_{pair} & B_{pair})	$H_T(jets) > 600$ GeV	35
5b (T_{single} & B_{single})	$N(forward\ jets) \geq 1$	60

Table 4.1: The signal regions of the square cuts search strategy defined as a series of square cut requirements.

The two square cuts signal regions still predicted too many background events to draw conclusions regarding the existence of VLQs, and so a discriminating variable was sought that would enhance signal-background discrimination via spectral discrimination statistical inference. A reconstructed invariant mass was chosen as the discriminating variable. The invariant mass of the each of the four signal types, T_{pair}, B_{pair}, T_{single}, and B_{single}, was not a known parameter, i.e., VLQs could exist with an unspecified mass. Each possible VLQ invariant mass thus represented a unique signal particle, with its own cross section and unique decay kinematics. For example, the 350 GeV/c^2 T_{pair} signal mass had a much higher theoretical cross section than the 1000 GeV/c^2 T_{pair} signal mass, and so the 350 GeV/c^2 signal was expected to be created in proton-proton collisions much more frequently than the 1000 GeV/c^2 signal. Nevertheless, the 1000 GeV/c^2 signal was expected to have much more energetic decay products, and thus have decay products that were much easier to differentiate from those of the backgrounds.

Essentially, each invariant mass for each signal type represented a separate particle search. Two signal regions could not be expected to provide large signal-background discrimination for all 59 unique signal masses,[4] and the use of a discriminating variable provided a method to simultaneously discriminate the background from each of the separate signal masses. Table 4.2 lists the four signal types of the VLQs and their invariant masses that were investigated.

A collision event containing a VLQ was different from a background event because it contained the decay products from a massive particle, the VLQ itself. The strategy was to identify the specific VLQ decay products from among all the physics objects of a collision event, and to combine them to obtain a reconstruction of the VLQ. For example, the vector-like bottom quark decayed to a Z boson and b quark, $B \to Z + b$. The four-momentum of the Z boson candidate, Z^μ, was already reconstructed from the two leptons of the event, $Z^\mu = l^\mu + \bar{l}^\mu$. Therefore, identification of b^μ provided $B^\mu = Z^\mu + b^\mu$,

[4]The use of only two signal regions was in order to reduce the amount of labor required to complete the VLQ search. A large amount of validation was required to ensure that each signal region would provide accurate results. Therefore, defining a separate signal region for each of the 59 signal masses was impractical.

signal type	signal masses GeV/c^2	# signal masses
T_{pair}	350–1100	16
B_{pair}	350–1100	16
T_{single}	400–1050	14
B_{single}	400–1200	13
total		59

Table 4.2: List of four vector-like quark signal types and their signal invariant masses, where a statistical inference procedure was performed separately for each signal mass.

from which the invariant mass of the B was estimated, $m(Zb) \equiv \sqrt{\eta_{\mu\nu} B^\mu B^\nu}$. The b quark originating from the massive B was expected to be highly energetic, and so the b-tagged jet with the largest transverse momentum was chosen as the candidate b quark from B decay.

Figure 4.5 shows the expected distribution of the $m(Zb)$ variable. For the 650 GeV/c^2 B_{pair} and B_{single} signals (blue solid and dotted lines), $m(Zb)$ was observed to have a narrow distribution that peaked near 650 GeV/c^2 (and similarly for the other signal masses not shown). For the background, $m(Zb)$ was constructed from physics objects that did not originate from B decay, and so the background distribution of $m(Zb)$ was wide and decreasing.

The selection of $m(Zb)$ as the discriminating variable of the search strategy was ideally suited for use with the B_{pair} and B_{single} signal types since its distribution provided a strong peak at the signal mass, and so was applicable over the wide range of signal masses. However, figure 4.5 also shows that the $m(Zb)$ distribution for the 650 GeV/c^2 T_{pair} and T_{single} signals (red solid and dotted lines) was wider and peaked below 650 GeV/c^2 (and similarly for the other signal masses not shown). The vector-like top decayed to a Z boson, a W boson and a b quark, $T \to Z + t \to Z + W + b$, and so $m(Zb)$ did not fully reconstruct the T mass because it did not contain the contribution from the W boson. The result was that the $m(Zb)$ did not perform as well for T_{pair} and T_{single} because those signal distributions were shifted to lower values where more of the background distribution was located.

The W boson decayed to two SM quarks, which would have been detected as one or two jets.[5] Several attempts were made to identify the W boson decay products so that the vector-like top mass could be fully reconstructed for use as a discriminating variable. These attempts were largely unsuccessful.[6]

Figures 4.6 (a) and (b) compare example decay topologies of the B_{single} signal type (the least complex) and the T_{pair} signal type (the most complex). For B_{single}, there were far fewer decay products from which to choose in reconstructing the vector-like bottom mass, while for T_{pair}, identifying the decay products of the vector-like top was prone to error. A novel use of machine learning regression was found to address these difficulties and is discussed in the next chapter (section 5.7 on page 170).

[5] The restriction of the search to two leptons from a Z boson decay ruled out leptonic W boson decay.

[6] With the exception of one method of $m(Zt)$ full reconstruction for T_{single} discussed in footnote 12 on page 170. Unfortunately, this $m(Zt)$ did not perform any better than $m(Zb)$.

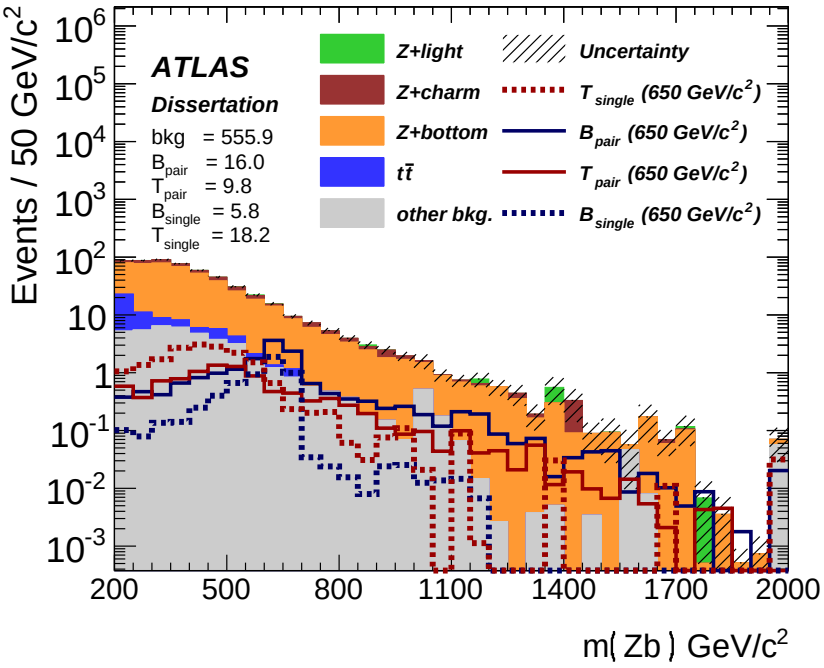
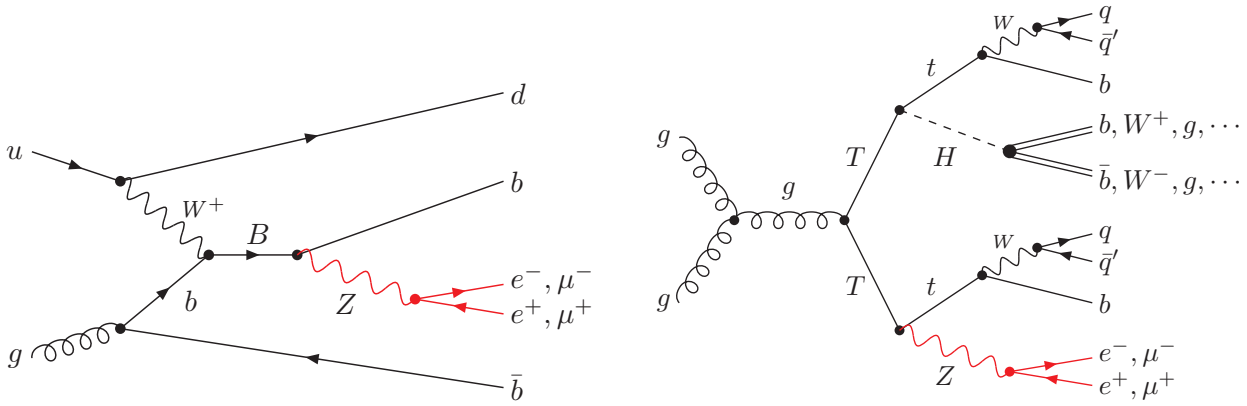


Figure 4.5: The expected $m(Zb)$ yields after square cut requirements 1–4 (with measured yields blinded to prevent experimenter bias).



(a) B_{single} (b) T_{pair}

Figure 4.6: Contrasting the decay topologies of the B_{single} vs. the T_{pair} signal types.

Therefore, $m(Zb)$ was used as the discriminating variable for all signal types. This provided a fully reconstructed B mass for B_{pair} and B_{single}, and a partially reconstructed T mass for T_{pair} and T_{single}. This selection of the two square cuts signal regions and the $m(Zb)$ discriminating variable determined the square cuts search strategy.[7]

[7]The figures shown in this section did not contain QCD background estimates or corrections to the simulated backgrounds. Once these tasks were completed, these figures were recomputed and found to support the same selection of the signal region requirements and the use of $m(Zb)$ as the discriminating variable.

4.2. Background Corrections

The final result of a signal search is an inference statement summarizing the evidence for signal existence based on experimental measurements. This is accomplished by comparing the measured data to 1) the simulated background data, and 2) the simulated background plus simulated signal data. If the signal does not exist, then the measured data should be more consistent with 1), while greater consistency with 2) provides evidence for the signal existing above the background. In either case, the simulated background is used for the comparison, and so accurate modeling of the background is imperative for valid conclusions to be drawn.

Figures 4.2 (a)–(d) indicated that the QCD background needed to be taken into account, and that the transverse momentum of the Z boson candidate, $p_T(Z)$, was not accurately modeled by the simulated data. Additional studies showed that the b-tagging was not accurately modeled. This was particularly problematic since square cut 4, $N(b\text{-}tags) \geq 2$, reduced the expected background by ∼99%. This section describes the data-driven estimation of the QCD background and the corrections to the simulated backgrounds that were necessary in determining the possible existence of VLQs.

4.2.1. Estimation of the QCD Background

QCD background events passed the preselection requirements whenever two jets were misidentified as electrons or muons ("faked" leptons).[8] The QCD background was difficult to accurately simulate, and so a data-driven approach was implemented. The data-driven approach involved obtaining a sample of *measured data* that would accurately model the QCD background, the "QCD sample." Four objectives had to be achieved for the data-driven approach to be valid.

- The QCD sample had to be comprised of measured events most likely to originate from the QCD background, and not from other backgrounds such as the Z+$jets$ background (i.e., double-counting other backgrounds had to be avoided).
- None of the measured data events of the QCD sample could be the same as any of the events of the measured data used for statistical inference (i.e., orthogonal datasets for statistical independence).
- The QCD sample had to accurately model the yields of the QCD background in any signal or control region used in the VLQ search (i.e., events weighted to give the correct cross section normalization).
- The QCD sample had to accurately model the decay kinematics of the QCD background so that the yield distributions of variables such as $H_T(jets)$ were accurate (i.e., accurate variable "shapes").

Once these objectives were met, the QCD sample of measured events was treated in the same manner as the datasets of the simulated backgrounds.

[8] Standard Model processes that "faked" a Z boson by producing one real lepton and one fake lepton were modeled using Monte Carlo simulations (such as with the W+$jets$ background).

The use of inverted lepton reconstruction quality requirements provided a method of obtaining a measured dataset that was orthogonal to the measured data used for statistical inference. The faking of leptons by the jets of the QCD background occurred at varying levels of lepton reconstruction quality. Many QCD background jets faked leptons with looser reconstruction quality requirements, while many fewer faked leptons with tighter reconstruction criteria.

The lepton requirements used for physics object reconstruction was described in section 3.4.2 and section 3.4.3, which included the `medium` quality criteria for electrons and the `mini-isolation` criteria for muons. The QCD sample was obtained by reconstructing electrons using the `loose` quality criteria, and then rejecting events when the electrons also passed the `medium` criteria. Thus, the electrons satisfied a `loose-but-not-medium` criteria, and thus dataset orthogonality was achieved. In order to restrict background double-counting from Z+$jets$ and $t\bar{t}$, the two electrons were required to have the same electric charge. Similarly, muons for the QCD sample were reconstructed as `muon-but-not-mini`, where these muons met all other requirements, but failed the `mini-isolation` criteria.[9] These inverted requirements were used to produce a sample in which the "electrons" or "muons" were actually misidentified jets.

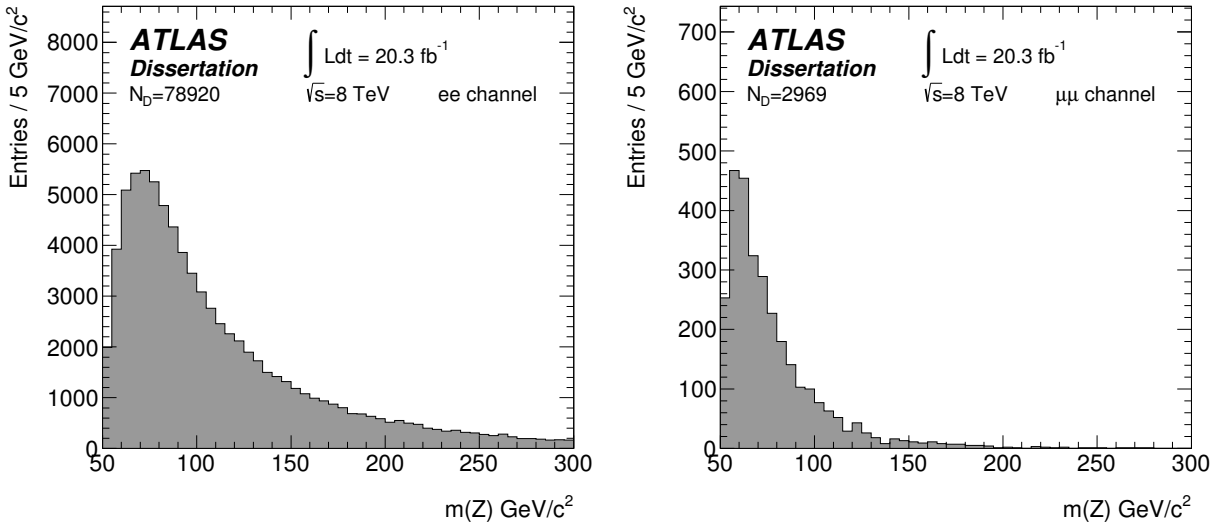
The Z-tag VLQ search required that the Z boson candidate have an invariant mass within 10 GeV/c^2 of 91.19 GeV/c^2 (square cut 1, $81.19 < m(Z) < 101.19$ GeV/c^2). The QCD sample was normalized to the cross section of the QCD background by examining the side bands surrounding this $m(Z)$ mass window. More specifically, the range $50 < m(Z) < 75$ GeV/c^2 was used to estimate the normalization and the range $110 < m(Z) < 135$ GeV/c^2 was used to determine the systematic uncertainty due to this procedure [63].

Let N_{sample} be the number of QCD sample events contained in a side band region, let N_{data} be the number of test data events, and let $N_{sim\text{-}bkg}$ be the total yield of the simulated, non-QCD backgrounds. The yield of the QCD background in the side band was estimated as the `data-minus-other-backgrounds`, $N_{QCD} \equiv N_{data} - N_{sim\text{-}bkg}$. The normalization of the QCD sample was then given by $f_{sample} \equiv N_{QCD}/N_{sample}$. This scale factor was estimated as 0.61 ± 0.07 for the ee channel and 3.32 ± 1.08 for the $\mu\mu$ channel.[10]

Figures 4.7 (a) and (b) show the number of events of the QCD sample as a function of $m(Z)$ after preselection. These indicated that there was little contamination from the Z+$jets$ background because no peak was observed near 91.19 GeV/c^2. Figures 4.8 (a) and (b) show the resultant estimated QCD background in the lower side band used for normalization.

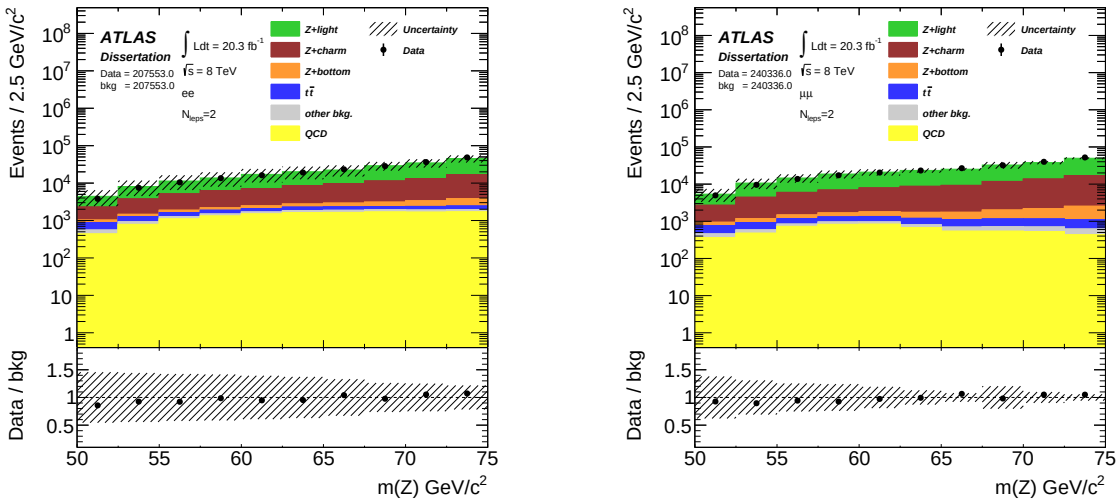
[9] Additionally, at least one electron had to fail an isolation requirement by having E_T-cone20 > 5 GeV, and at least one muon had to have a d_0 significance of at least 5.

[10] The large factor for the $\mu\mu$ channel was due to the difficulty in obtaining a large QCD sample that was not contaminated by other backgrounds that were already modeled via Monte Carlo simulation. This led to a large fractional uncertainty for the QCD background in the $\mu\mu$ channel. Nevertheless, table 4.3 shows that the predicted number of $\mu\mu$ QCD events was insignificant (<0.02% of the total predicted background).



(a) ee channel. (b) $\mu\mu$ channel.

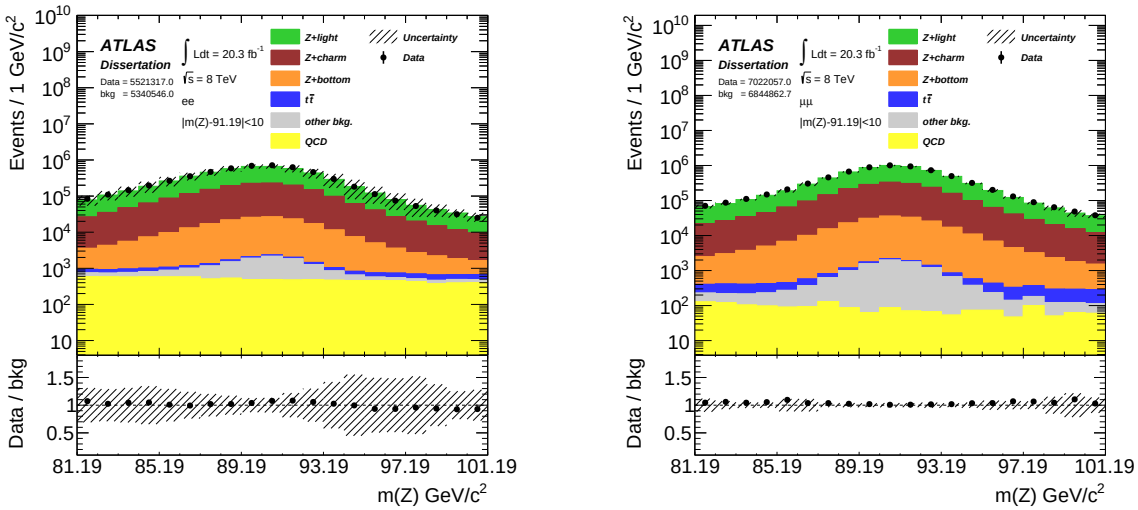
Figure 4.7: The $m(Z)$ distribution of QCD sample events after preselection (with inverted electron and muon reconstruction criteria).



(a) ee channel. (b) $\mu\mu$ channel.

Figure 4.8: The $m(Z)$ distribution using the estimated QCD background in the lower side band used for normalization of the QCD sample.

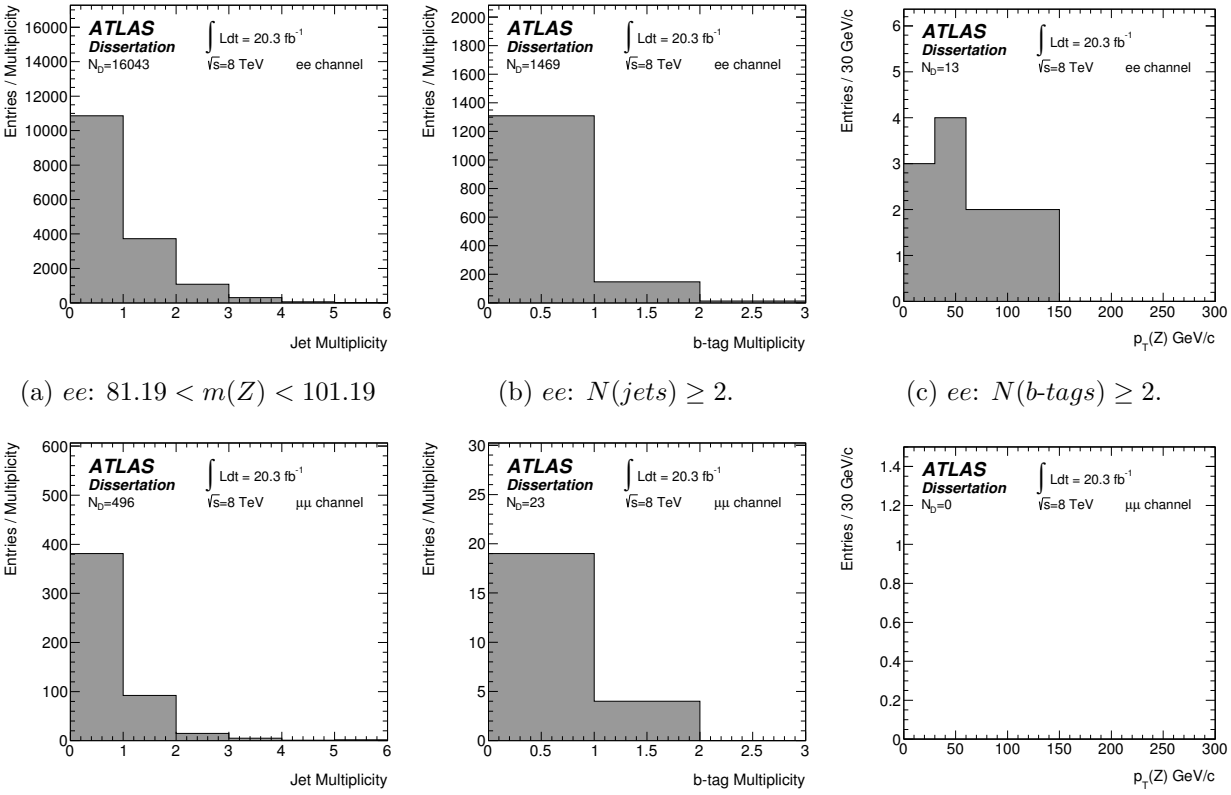
Figures 4.9 (a) and (b) show the resultant estimated QCD background in the mass window, $81.19 < m(Z) < 101.19$ GeV/c^2, that was used for the VLQ search (square cut 1). These showed that the expected background yields in the mass window were still lower than the measured yields by ∼3% after including the estimated QCD background, and although within the bounds of measurement uncertainties, possibly indicated the need for corrections to the non-QCD simulated backgrounds (discussed in the next section).



(a) ee channel. (b) $\mu\mu$ channel.

Figure 4.9: The $m(Z)$ distribution using the estimated QCD background in the $81.19 < m(Z) < 101.19$ GeV/c^2 mass window used for the VLQ search.

The QCD sample was next used to examine the expected QCD background distributions for progressive square cuts leading to the final signal regions. Figures 4.10 (a)–(f) show the how the number of events in the QCD sample rapidly decreased as the square cuts were applied. Table 4.3 summarizes these results, along with the statistical uncertainties. The statistical uncertainty on each yield was the combination of the uncertainty from the limited number of events in the QCD sample and the uncertainty in the estimation of the normalization scale factor. The final result of the QCD background estimation was that no QCD background was present in the final signal regions.



(a) ee: $81.19 < m(Z) < 101.19$ (b) ee: $N(jets) \geq 2$. (c) ee: $N(b\text{-}tags) \geq 2$.

(d) $\mu\mu$: $81.19 < m(Z) < 101.19$. (e) $\mu\mu$: $N(jets) \geq 2$. (f) $\mu\mu$: $N(b\text{-}tags) \geq 2$.

Figure 4.10: The distributions of the QCD sample events for progressive square cuts. Note that none of the events in the ee channel of (c) pass the additional square cut $p_T(Z) > 150$ GeV/c.

	ee channel		$\mu\mu$ channel	
square cut	unscaled	scaled	unscaled	scaled
$81.19 < m(Z) < 101.19$ GeV/c^2	16050	9838.7±1162.4	496	1647.2±539.0
$N_{jets} \geq 2$	1469	900.5±108.7	23	76.4±29.4
$N_{b-jets} \geq 2$	13	8.0±2.4	0	0.00±0
$p_T(Z) > 150$ GeV	0	0.00±0	0	0.00±0
$H_T(jets) > 600$ GeV	0	0.00±0	0	0.00±0
$N(forward\ jets) \geq 1$	0	0.00±0	0	0.00±0

Table 4.3: The counts of the QCD sample and the estimated QCD background yields from the scaled counts for progressive square cut requirements.

4.2.2. Corrections to Simulated Data.

Figure 4.11 shows the yields for $N(b\text{-}tags)$ after square cuts 0–2 (the square cuts are listed in table 4.1 on page 81). This indicated that the modeling of $N(b\text{-}tags)$ was mostly accurate. Nevertheless, the square cut 4 requirement, $N(b\text{-}tags) \geq 2$, reduced the background yields by ∼99%, and so greater accuracy of the b-tag modeling was required.[11]

The Z+*jets* background was the dominant background after square cut 2, comprising ∼97% of the total background, and so corrections to the b-tagging were most important for this background. Before correcting the Z+*jets* background, it was important to know if the fraction of Z+*jets* background containing b quarks or c quarks ("heavy flavor composition") was accurate. For example, after square cut 2, the simulations predicted that 10.3% of the Z+*jets* background events contained a b quark decay product, and this prediction needed to be checked against the measured data. The Z+*jets* background events had been filtered into the Z+*bottom*, Z+*charm* and Z+*light* subsets based on whether each event had a b or c quark (discussed in section 3.5.1, page 72), and this allowed an examination of the modeling accuracy of the heavy flavor composition.

This was achieved using two applications of the JetFitterCombNN artificial neural network (ANN) learning algorithm to examine the modeling accuracy of the Z+*jets* flavor composition (as was similarly done in [151]). For input variables, JetFitterCombNN used six variables from the JetFitter b-tagging algorithm and the output of the IP3D b-tagging algorithm.[12] JetFitterCombNN was then trained separately to generate two classification response functions, $p_b(jet)$ and $p_u(jet)$. The function $p_b(jet)$ returned a large value for jets most likely to indicate the presence of b quarks, and $p_u(jet)$ the presence of only light

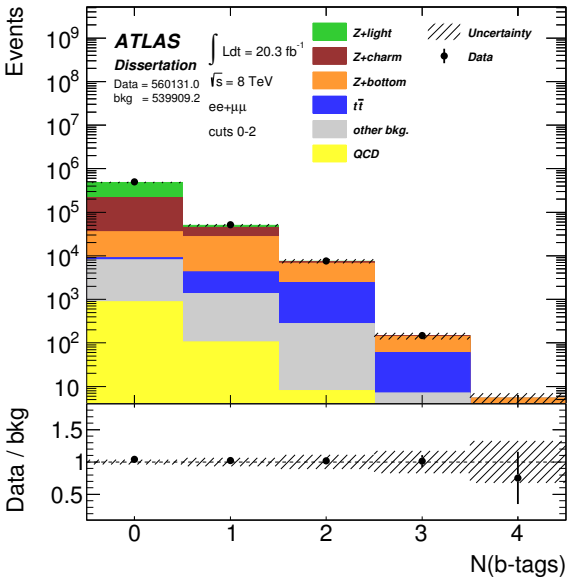


Figure 4.11: The distribution of $N(b\text{-}tags)$ after square cuts 0–2.

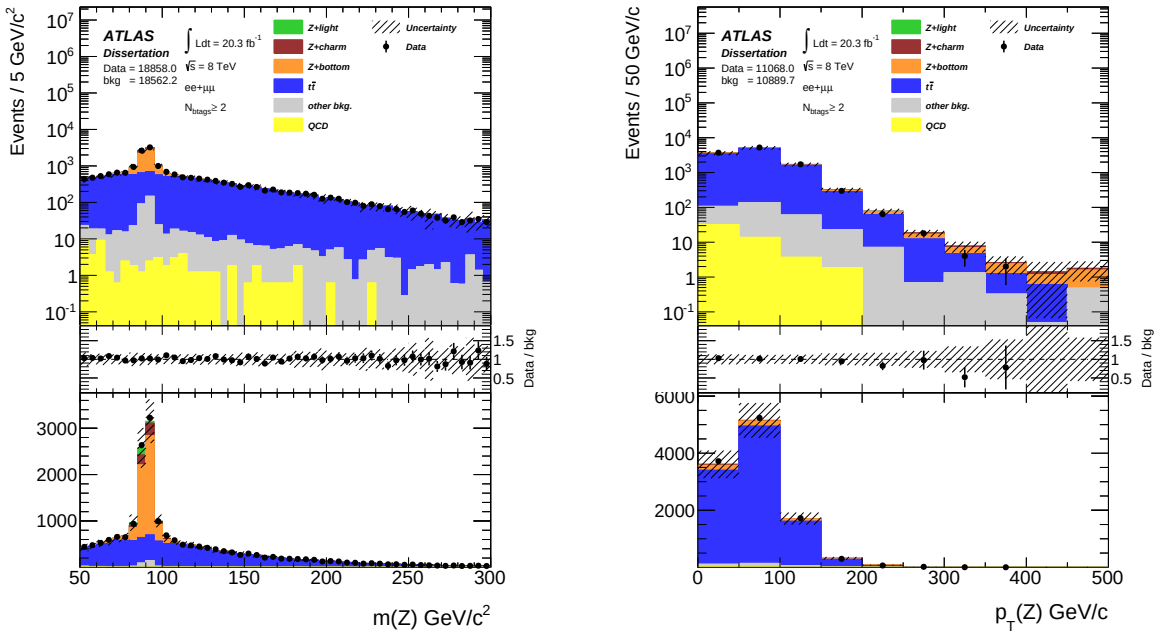
[11] The corrections to the Z+*jets* background described in this section were implemented separately for the dielectron and dimuon channels due to differing levels of modeling inconsistency that were caused by the different performance of object reconstruction for electrons vs. muons.

[12] The JetFitter, IP3D and MV1 flavor-tagging algorithms were described in section 3.4.5 on page 69.

quarks (non-b and non-c). The Z-tag search for VLQs required the presence of at least two b-tagged jets (square cut 4) using the MV1 ANN b-tagging algorithm, which itself generated a classification response function, $MV1(jet)$. For each event in the Z-tag search, the two jets, j_1 and j_2, with the largest values of $MV1(j_i)$ were selected. The event was then assigned the value JetFitter($MV1\ b_1 + b_2$) $\equiv \log\left(p_b(j_1)/p_u(j_1)\right) + \log\left(p_b(j_2)/p_u(j_2)\right)$, and this event-level variable was used to study the flavor composition of the Z+$jets$ background. Figures 5.11 (a)–(c) of [62] show the yield distributions of the JetFitter($MV1\ b_1 + b_2$) for each b-tag multiplicity, and demonstrated that the heavy flavor composition was well-modeled within the experimental uncertainties. This justified the use of flavor-independent corrections to the Z+$jets$ background.

The modeling accuracy of the $t\bar{t}$ background was also examined before implementing any corrections to the Z+$jets$ background. Figure 4.12 (a) shows the $m(Z)$ yields after square cuts 0 and 4, the preselection and $N(b\text{-}tags) \geq 2$.[13] By removing events within the Z boson mass window, the remaining events were primarily the $t\bar{t}$ background (91.6%). Figure 4.12 (b) shows the $p_T(Z)$ yields after excluding the mass window events, and indicated that $t\bar{t}$ was not the cause of $p_T(Z)$ mismodeling. These demonstrated $m(Z)$ and $p_T(Z)$ were well-modeled by the $t\bar{t}$ simulations when two or more b-tags were required.

The correction of $N(b\text{-}tags)$ for the Z+$jets$ background was performed similarly to



(a) $m(Z)$ (b) $p_T(Z)$

Figure 4.12: (a) The $m(Z)$ yields after preselection and requiring $N(b\text{-}tags) \geq 2$, which had Z+$bottom$ and $t\bar{t}$ as the two dominant backgrounds. The $t\bar{t}$ background was dominant outside the Z boson mass window, $81.19 < m(Z) < 101.19$ GeV/c^2. (b) The $p_T(Z)$ yields after inverting the mass window, $m(Z) < 81.19$ GeV/c^2 or $101.19 < m(Z)$ so that $t\bar{t}$ was the dominant background.

[13] The $t\bar{t}$ background events always generated two b quarks as decay products.

the QCD background estimation: a control region orthogonal to the signal regions was determined, and a data driven estimate was made. Specifically, square cut 3 was inverted so that $p_T(Z) < 150$ GeV/c and divided into two control regions, CR1 and CR2. Control region CR1 was defined by square cuts 0–2 and $p_T(Z) < 100$ GeV/c and CR2 was defined by square cuts 0–2 and $100 < p_T(Z) < 150$ GeV/c. CR1 was used to provide the correction reweighting factor, g_i, where i was the number of b-tagged jets in the event, $i = 0$, 1, 2 or 3. CR2 was used to estimate the uncertainty of g_i. The reweight factor was computed as

$$g_i = \frac{N_{data,i} - N_{nonZj,i}}{N_{Zj,i}}, \tag{4.2.1}$$

where $N_{data,i}$ was the number of observed events in the control region with i b-tags, $N_{nonZj,i}$ was the total yield of all other backgrounds, and $N_{Zj,i}$ was the Z+$jets$ yield.

Figures 44 (a) and (b) of [63] show the resultant values of the $N(b\text{-}tags)$ reweight factors and its uncertainties with solid green lines.[14] Note that the reweight factors for Sherpa were approximately 1.0 for each b-tag multiplicity and had small uncertainty.

The square cuts search strategy implemented the requirement $p_T(Z) > 150$ GeV/c as cut 3 that provided a 70% reduction in background. However, figures 4.2 (b)–(d) showed that the $p_T(Z)$ distribution was mismodeled for any b-tag multiplicity. Z+$jets$ was the dominant background after square cuts 0–2, while figure 4.12 (b) showed that the mismodeling was not caused by the $t\bar{t}$ simulations, and so the $p_T(Z)$ reweighting was only applied to the Z+$jets$ background.[15] Because square cut 4 restricted the signal region to $N(b\text{-}tags) \geq 2$, the other b-tag multiplicities were used as control regions with CR3 given as $N(b\text{-}tags) = 1$ used for the reweight factor estimates and CR4 given as $N(b\text{-}tags) = 0$ used to measure additional uncertainty in the reweight factor.[16]

First, reweight factors were derived for several bins of the $p_T(Z)$ distribution,

$$h_j = \frac{N_{data,j} - N_{nonZj,j}}{N_{Zj,j}}, \tag{4.2.2}$$

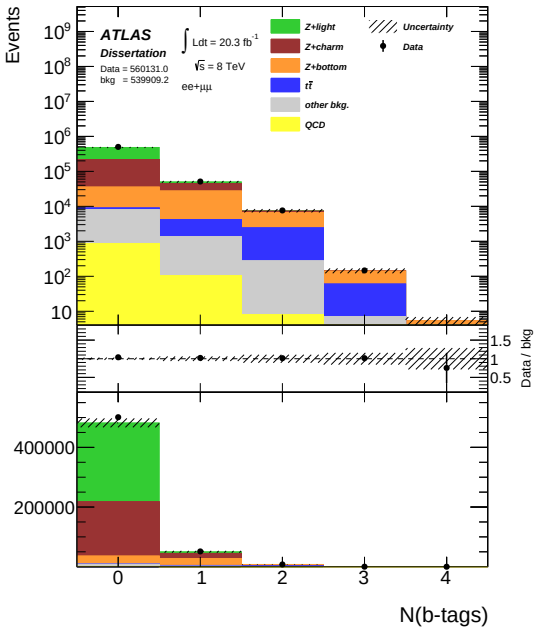
for each j^{th} bin. These values are shown with black dots in figures 61 (a) and (b) of [63]. These values were then fit with a third order polynomial (dashed blue line) that was used as the nominal reweight factor applied to each event. The 95% confidence interval for the fitted polynomial is shown with a green band. Systematic uncertainties were due to the number of events available to estimate each h_j (vertical error bars), the fit uncertainty (green band), as well as the fit function of the alternate control region, CR4 (not shown).

Figures 4.13 (a)–(d) and figures 4.14 (a)–(d) show the $N(b\text{-}tags)$ and $p_T(Z)$ distributions before and after these corrections. These show that the corrections to the b-tagging normalization were minor, while those to the $p_T(Z)$ normalization improved the modeling accuracy of the Z+$jets$ background.

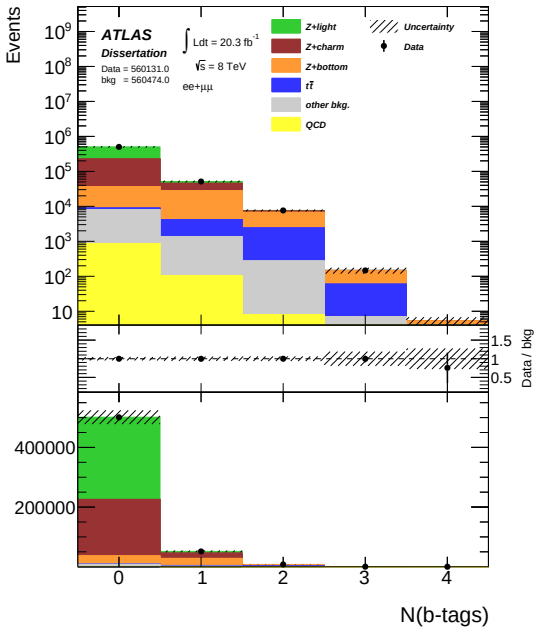
[14] This correction was performed for the primary simulated Z+$jets$ background, "Sherpa", as well as for an alternative simulation of Z+$jets$, "Alpgen", that was used as a robustness check during the VLQ search.

[15] After square cuts 0–2, Z+$jets$ comprised 98% of the total background for $N(b\text{-}tags) = 0$, 92% for $N(b\text{-}tags) = 1$, and 68% for $N(b\text{-}tags) \geq 2$. Also, the $p_T(Z)$ mismodeling was independent of the number of b-tagged jets, so that the $p_T(Z)$ correction was not a function of $N(b\text{-}tags)$.

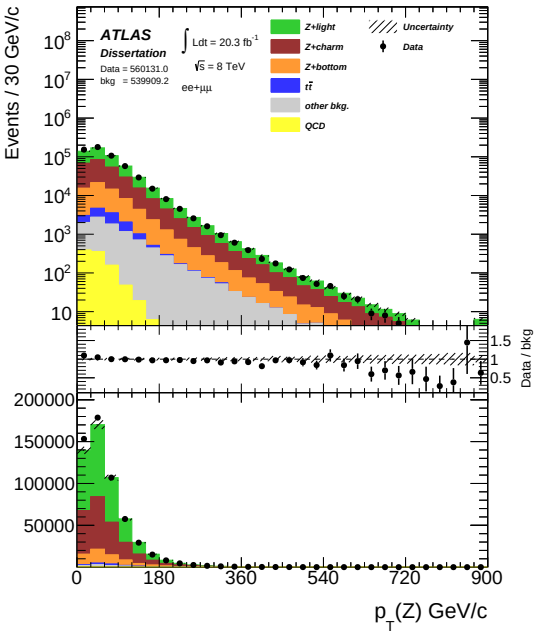
[16] CR3 and CR4 were also defined as meeting square cuts 0–2.



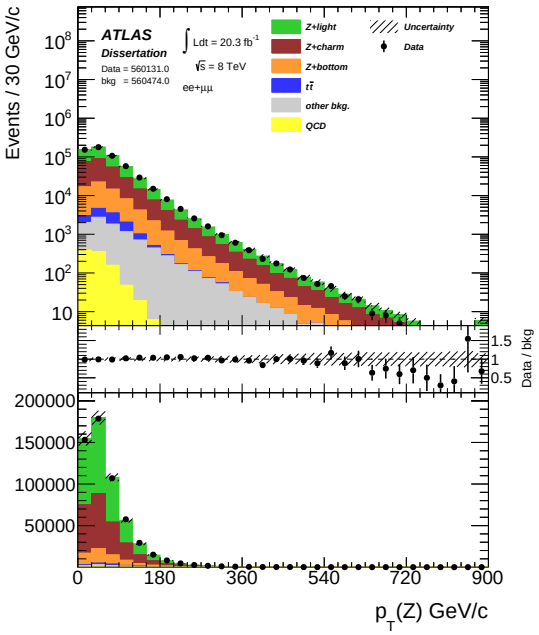
(a) $N(b\text{-}tags)$: uncorrected.



(b) $N(b\text{-}tags)$: corrected.

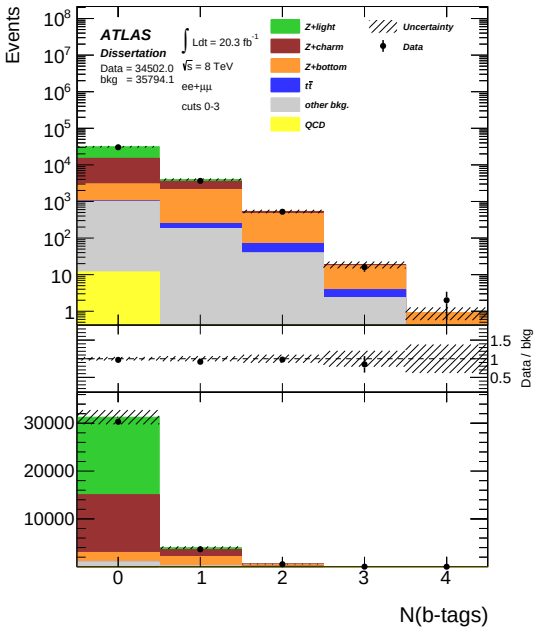


(c) $p_T(Z)$: uncorrected.

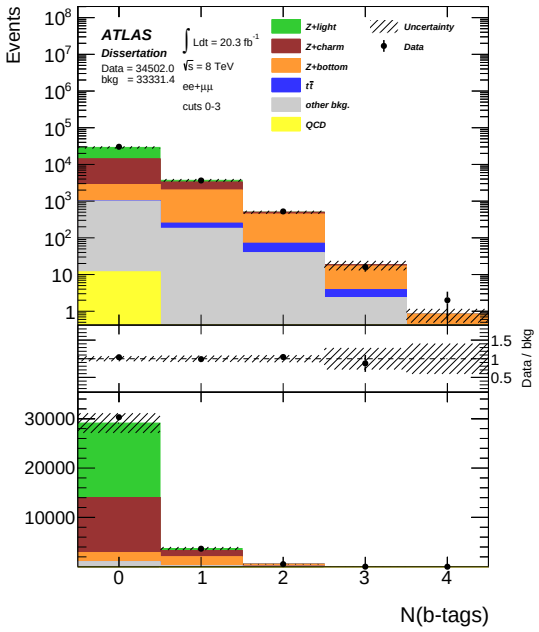


(d) $p_T(Z)$: corrected.

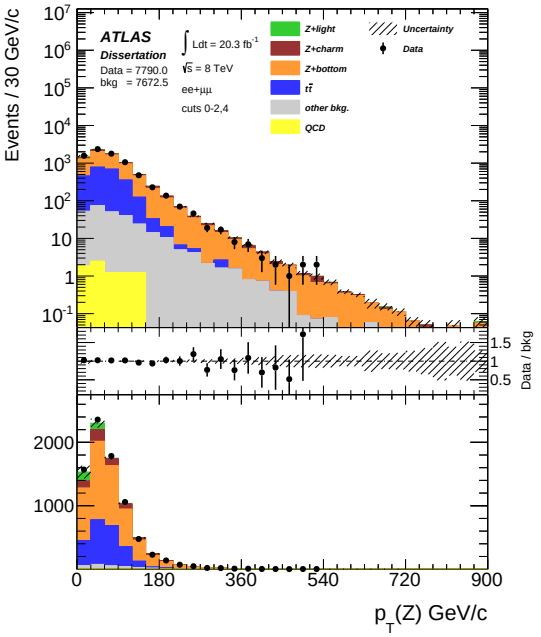
Figure 4.13: The distributions of $N(b\text{-}tags)$ and $p_T(Z)$ after square cuts 0–2: (a) $N(b\text{-}tags)$ before and (b) after corrections, and (c) $p_T(Z)$ before and (d) after corrections.



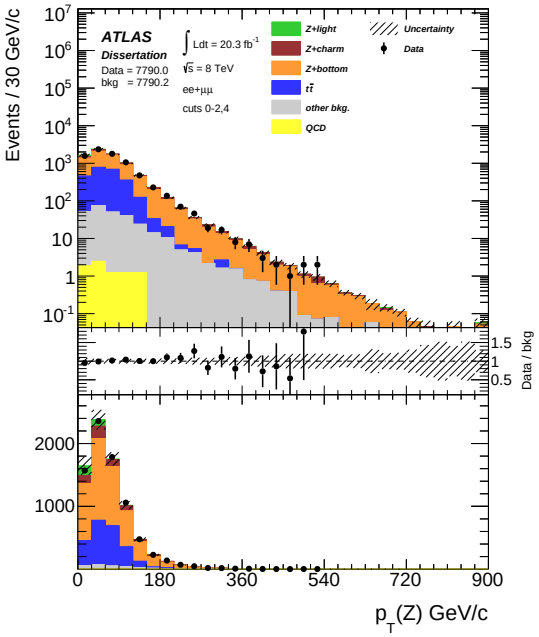
(a) $N(b\text{-}tags)$: uncorrected.



(b) $N(b\text{-}tags)$: corrected.



(c) $p_T(Z)$: uncorrected.



(d) $p_T(Z)$: corrected.

Figure 4.14: The distribution of $N(b\text{-}tags)$ after square cuts 0–3: (a) before and (b) after corrections. The $p_T(Z)$ distribution after square cuts 0–2 & 4: (c) before and (d) after corrections.

4.3. Systematic Uncertainties

The predicted yields obtained via simulation were not known with infinite precision. Uncertainties from a variety of sources had to be estimated, often using auxiliary measurements made in control regions (CRs). Uncertainty in the simulated signal and background yields degraded the quality of the search results because it made it more difficult to differentiate between the background-only expectations and the signal plus background expectations. Specifically, larger uncertainty in the simulated predictions led to larger upper limits for the cross sections of the VLQ signals.

The simulated backgrounds and VLQ signals were produced so that there were many more simulated events used to model a smaller number of measured events. For example, 230,000 unweighted simulated $t\bar{t}$ events passing preselection were used to model a predicted yield of 42,000 events. Using a large number of simulated events was done to provide accurate modeling of the distributions of the many variables, and so simulated events were weighted with values typically smaller than 1.0. A given background (or signal), Y, was expected to produce $N^Y_{prod.}$ collision events given by

$$N^Y_{prod.} = \mathcal{L}\,\sigma_Y, \tag{4.3.1}$$

where $\mathcal{L}$ was the total integrated luminosity,[17] and σ_Y was the theoretical cross section of that background.

If $N^Y_{sim.}$ events were simulated to model $N^Y_{prod.}$, then the event weight for all those background events would have been $N^Y_{prod.}/N^Y_{sim.}$. However, the detector did not measure all $N^Y_{prod.}$ events with 100% efficiency, nor was detection efficiency constant for each particular event of that background, therefore, event-dependent weights were used to model the detector efficiency.

For background Y, each simulated collision event, i, was weighted with a total event weight, w^Y_i, in order to correctly predict the fraction of type Y events that event i represented by accounting for detector efficiencies. These total event weights gave the total predicted yield in a region of interest (i.e., using only events passing the square cut requirements for that region) of

$$N^Y_{ROI} = \mathcal{L}\,\sigma_Y\,\frac{\sum_i w^Y_i}{N^Y_{sim.}}. \tag{4.3.2}$$

Total event weights were thus corrections that were applied to the simulated datasets to account for how each simulated collision event was expected to interact with the various subdetector systems of the ATLAS experiment. Each total event weight was a combination of several other weights corresponding to specific properties of a given event. For example, weights were applied to an event based on the number of jets it had, and the energies and locations of those jets in the detector. These various weights were determined by auxiliary measurements, and thus contained measurement uncertainty. These measurement uncertainties in w_i propagated to the estimation of N^Y_{ROI}, and therefore had to be taken into account during statistical inference. Also note that $\mathcal{L}$ was estimated from measurements

[17] The total integrated luminosity was discussed in chapter 3 and was proportional to the number of collisions that occurred during a data taking period. Therefore, $\mathcal{L}\cdot\sigma_Y$ gave the number of those collision events attributable to a type Y occurrence.

(see section 3.3.2), and corrections were also made to theoretical cross sections based on measurements. Each of the uncertainties affecting the estimation of the total predicted yields was referred to as a "systematic uncertainty."

Table 4.4 shows a summary of the types of systematic uncertainties that affected the VLQ search. The total integrated luminosity for the VLQ search was estimated as 20.3 fb^{-1}, but this measurement was estimated to have a 2.8% systematic uncertainty. Therefore, all simulated datasets predicted total yields of $N^Y_{ROI} \pm 2.8\%$ that were used for comparison with the measured datasets in a given region of interest (ROI). As more systematic uncertainties were included, the power to differentiate between the background-only hypothesis and the background plus signal hypothesis was decreased.

The reconstruction of jets led to large systematic uncertainties. Jet reconstruction (discussed in section 3.4.4) involved examining clusters of energy deposits in the calorimeters. The calorimeters did not collect all the energy of the hadronic jets, and so the total jet energy needed to be scaled to a total estimated value. The jet energy scaling was also corrected for pile-up, and these factors contributed 1–5% to the systematic uncertainties of jet reconstruction [152]. The transverse momentum balance of dijet events allowed for measurement of the resolution of jet energies, the variability of energy deposition in material during hadronic showering. This variability gave additional uncertainty of several percent (the jet energy resolution uncertainty) [153].

The signal regions of the VLQ search required the presence of at least two b-tagged jets (square cut 4). The b-tagging of hadronic jets was discussed in section 3.4.5 (page 69). Systematic uncertainties due to b-tagging were estimated at 2–6% [120].

The VLQ search also relied on the correct identification and measurement of exactly two electrons or else exactly two muons. Uncertainties in the identification efficiencies, electron energy scale, and the energy and momentum resolution were on the order of 5% [110, 111].

Additional systematic uncertainty for the $Z{+}jets$ backgrounds was due to the corrections to the $N(b\text{-}tags)$ and $p_T(Z)$ variables described in section 4.2.2 (page 89). These were due to the amount of simulated data used to derive the reweighting factors, control region uncertainty estimated using alternate control regions, and the uncertainty of the cubic fit for the $p_T(Z)$ correction.

Finally, the "Monte Carlo Statistical" uncertainty due to a limited volume of simulation data used to predict yields in ROIs was also incorporated. This uncertainty began quite small, but increased as square cut requirements were progressively made, leaving fewer and fewer simulated collision events for use in estimating yields.

systematic uncertainty	description
luminosity	Estimation of total amount of collisions measured.
cross section	Estimation of rate of particle creation in collisions.
jet reconstruction	Estimation of the energy scale, the energy resolution, the reconstruction efficiency, and the jet vertex fraction cut efficiency.
b-tagging corrections	Estimation of the efficiency of correctly tagging a heavy flavor quark and the rate of mistagging.
electron reconstruction	Estimation of the triggering efficiency, the efficiency of correctly identifying an electron, the rate of incorrectly identifying a hadronic jet as an electron, the energy scale and resolution.
muon reconstruction	Estimation of the triggering efficiency, the efficiency of correctly identifying a muon, the momentum scale and resolution.
Z+$jets$ reweight factor	The uncertainties of the reweighting factors used for the corrections to the $N(b\text{-}tags)$ and $p_T(Z)$ variables due to the amount of data used in estimation, and also for $p_T(Z)$, the 95% confidence interval of the fitted 3^{rd} order polynomial.
Z+$jets$ alternate CR	The control region uncertainty determined by estimating the reweighting factors in alternate control regions.
MC statistical	The uncertainty due to the limited number of simulated events used in making estimates (a measure of simulated data insufficiency).

Table 4.4: Summary of the systematic uncertainties of the VLQ search.

The impact of these systematic uncertainties on the background predictions[18] in the two signal regions (SRs) are shown in table 4.5, as well as for the 650 GeV/c^2 signal masses of the four VLQ signal types. Jet reconstruction gave the largest contribution to the total systematic uncertainty of the background, particularly for the SR applied to singly produced VLQs. This was due to the use of forward jets for that SR, which typically had larger measurement uncertainties than central jets. For both signal regions, *Z+jets* was the dominant background comprising 75% of the SR applied to the pair produced VLQs, and 84% for the SR applied to singly produced VLQs. This led to the *Z+jets* corrections giving the second largest contribution to the total systematic uncertainty. Also note that data insufficiency was problematic for the T_{single} and B_{single} simulated datasets, where these had nontrivial MC statistical uncertainty.

	predicted yield (#)	MC stat. (%)	Z+j corr. (%)	jet reco. (%)	b-tag (%)	lepton reco. (%)	cross section (%)	lumi. (%)
H_T(jets) > 650 *GeV* *(pair produced VLQ signal region, square cut 5a)*								
Z+jets	24.0	5.0	8.6	7.8	5.0	2.5	1.0	N/A
$t\bar{t}$ + *single t*	5.0	24.4	N/A	16.1	7.9	7.9	5.7	2.8
diboson + $t\bar{t}X$ + *other*	3.7	8.2	N/A	7.0	7.7	9.3	29.4	2.8
total backgrounds	32.6	5.6	6.3	9.0	5.7	2.1	4.2	0.7
650 GeV/c^2 T_{pair}	8.4	2.8	N/A	1.2	4.1	0.9	N/A	2.8
650 GeV/c^2 B_{pair}	13.9	2.3	N/A	1.6	5.9	2.4	N/A	2.8
N(forward jets) ≥ 1 *(singly produced VLQ signal region, square cut 5b)*								
Z+jets	47.0	4.1	7.2	11.6	3.5	2.3	0.6	N/A
$t\bar{t}$ + *single t*	6.1	20.8	N/A	17.1	3.9	5.6	5.7	2.8
diboson + $t\bar{t}X$ + *other*	2.9	10.0	N/A	11.2	8.7	9.5	29.7	2.8
total backgrounds	55.9	4.3	6.0	12.2	3.8	2.4	2.1	0.4
650 GeV/c^2 T_{single}	7.7	7.6	N/A	3.0	5.1	4.5	N/A	2.8
650 GeV/c^2 B_{single}	1.9	5.1	N/A	8.5	8.5	1.5	N/A	2.8

Table 4.5: The estimated systematic uncertainties in the two signal regions of the VLQ search.

[18] Some backgrounds were grouped together when they had similar predicted yield frequency distributions across the discriminating variable spectrum, similar cross section uncertainties, and generally similar fractional systematic uncertainties in each of the bins of the discriminating variable. The $t\bar{t}$ and *single top* backgrounds shared these characteristics, as well as *diboson* and $t\bar{t}X$. All other backgrounds contributed insignificantly to the signal regions and were thus grouped with *diboson* and $t\bar{t}X$ for convenience.

4.4. Validation of the Simulated Backgrounds

The VLQ search to this point was blinded to measured data in regions of interest where the predicted VLQ signal yield represented a sizable fraction of the total predicted yields of the background. Performing a blinded particle search was important because it prevented experimenter bias from affecting decisions regarding the design of the search strategy and the evaluation of the modeling accuracy of the background simulations.

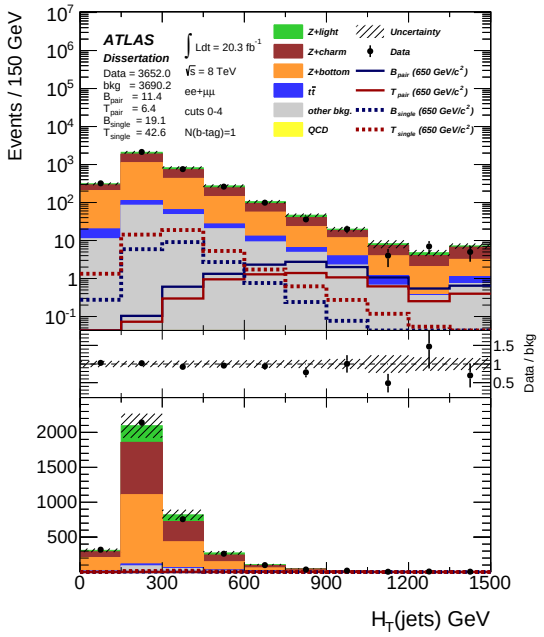
Validation of the simulated backgrounds in control regions was necessary to demonstrate accurate modeling of the Standard Model (SM) processes. Without accurate modeling, no claim of observing measured data inconsistent with the SM could be made (and thus no claim that such measurements supported VLQ existence).

The most pertinent variables that were validated were the $H_T(jets)$ and $N(forward\ jets)$ variables used to define square cuts 5a and 5b, respectively, which determined the signal regions of the search, as well as $m(Zb)$ that was used as the discriminating variable for statistical inference. The many hundreds of additional validation studies analyzed for the Z-tag VLQ search may be found in [63].

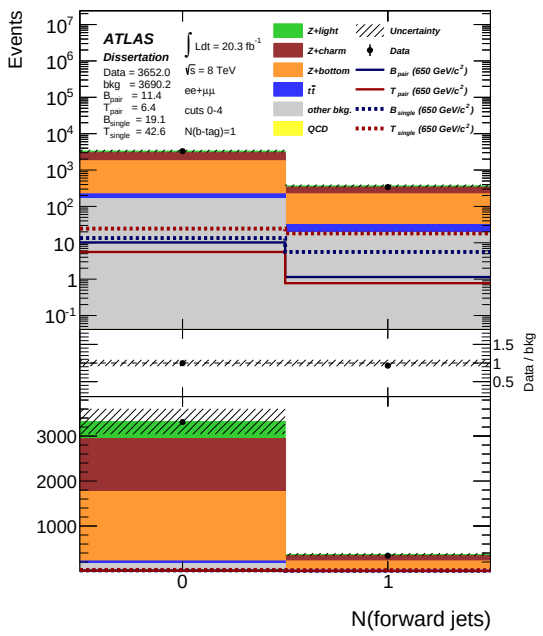
Figures 4.15 (a)–(c) show the validation of the $H_T(jets)$, $N(forward\ jets)$ and $m(Zb)$ variables in the control region defined by square cuts 0–3 and requiring exactly one b-tagged jet (where cut 4 requires two or more b-tagged jets). These showed good agreement within the range of systematic uncertainties.[19]

Figures 4.16 (a) and (b) show the validation of $H_T(jets)$ and $m(Zb)$ after cuts 0–4 and inverting square cut 5a, $H_T(jets) < 600$ GeV. Figures 4.16 (c) and (d) show the validation of $N(forward\ jets)$ and $m(Zb)$ after cuts 0–4 and inverting square cut 5b, $N(forward\ jets) = 0$. The modeling accuracy of the simulated background was deemed valid, and so the VLQ search proceeded to the unblinding of the signal regions.

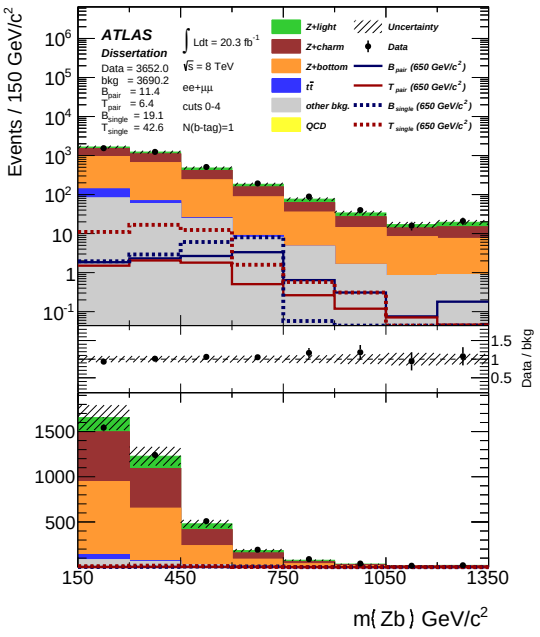
[19] The uncertainties shown with diagonal hashing include all systematic uncertainties.



(a) $H_T(jets)$.

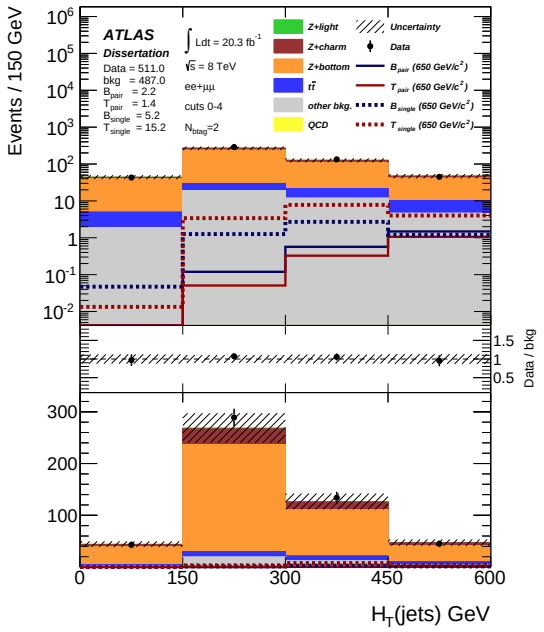


(b) $N(forward\ jets)$.

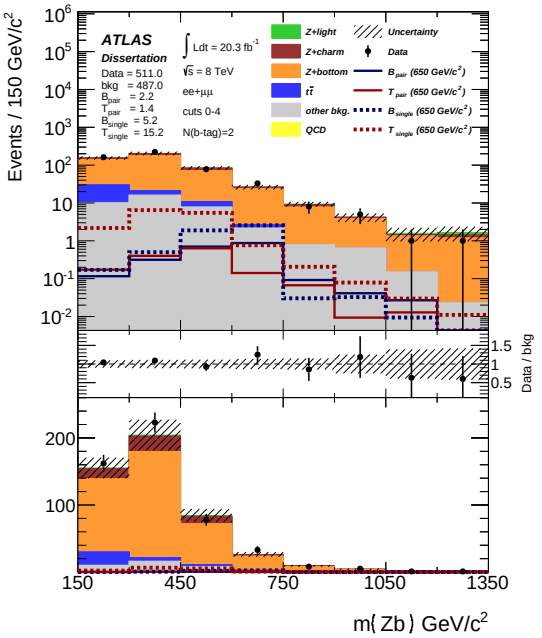


(c) $m(Zb)$.

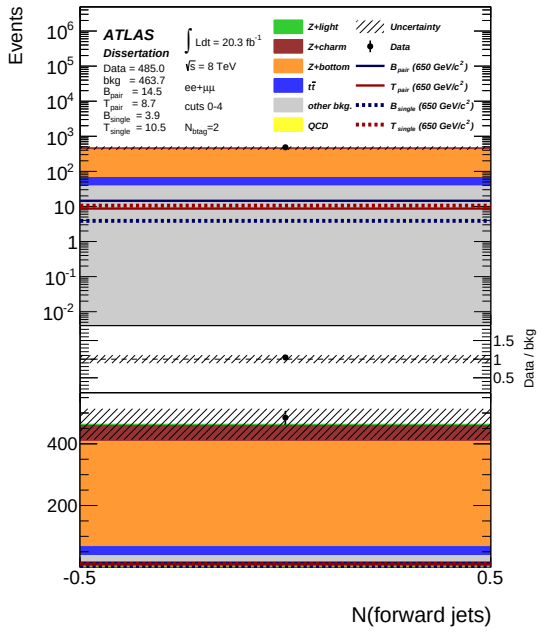
Figure 4.15: Validation in the control region defined by square cuts 0–3 and partially inverting square cut 4 with $N(b\text{-}tags) = 1$ to remain orthogonal to the signal regions. (The last bin of (b) shows $N(forward\ jets) \geq 1$.)



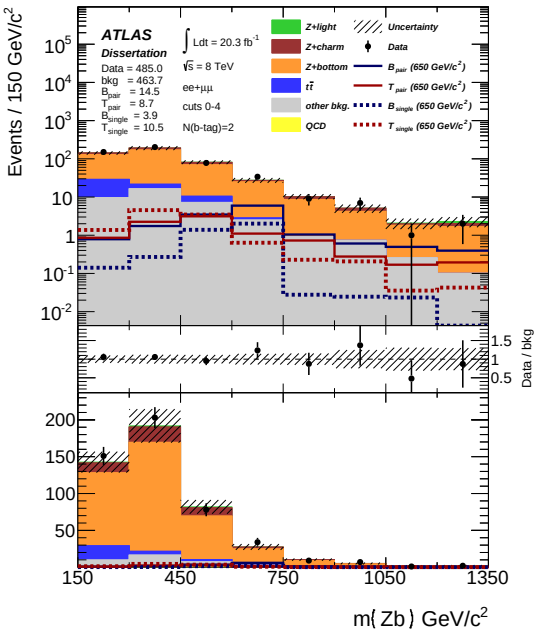
(a) $H_T(jets)$.



(b) $m(Zb)$.



(c) $N(forward\ jets)$.



(d) $m(Zb)$.

Figure 4.16: Validation in the control region defined by square cuts 0–4 and inverting cuts 5a and 5b to remain orthogonal to the signal regions: (a)–(b) inverted 5a, $H_T(jets) < 600$ GeV, and (c)–(d) inverted 5b, $N(forward\ jets) = 0$.

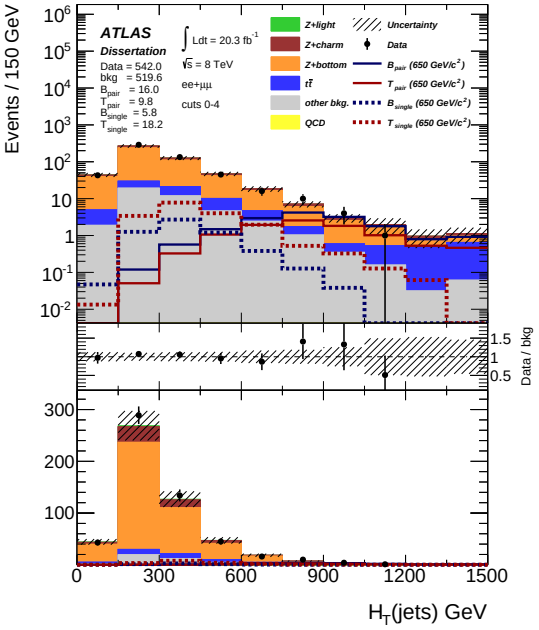
4.5. Unblinded Signal Regions

After successfully demonstrating the accuracy of the background modeling, the signal regions were unblinded. Table 4.6 provides the total predicted and measured yields in the two signal regions, including those of the 650 GeV/c^2 signal masses. These implied that the measured data was more consistent with the background-only hypothesis for the T_{pair}, B_{pair} and T_{single} VLQ signal types. The expected yield of the 650 GeV/c^2 B_{single} signal mass was not large enough to imply either existence or nonexistence. This provided additional motivation for the use of spectral discrimination rather than simple counting during statistical inference.

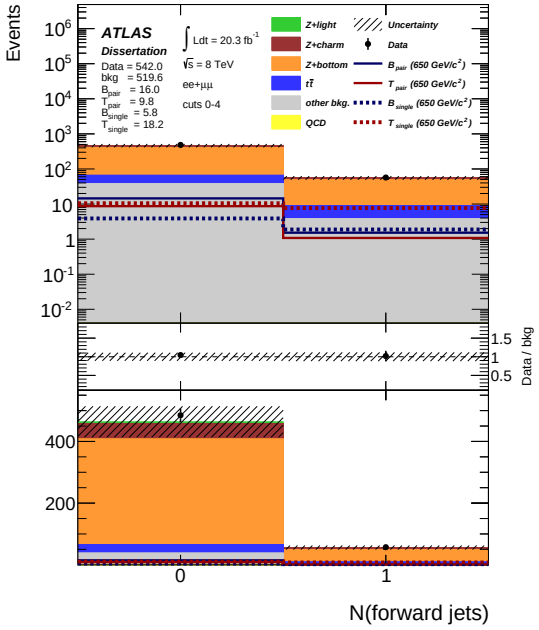
	cuts 0–5a	cuts 0–5b
Z+jets	$24.0 \pm 1.2^{+4.8}_{-4.6}$	$47.0 \pm 1.9^{+10.6}_{-7.4}$
other backgrounds	$8.6 \pm 1.3^{+1.8}_{-2.3}$	$9.0 \pm 1.3^{+2.9}_{-2.4}$
total background	$32.6 \pm 1.7^{+4.8}_{-5.7}$	$55.9 \pm 2.3^{+11.1}_{-9.7}$
measured data	31	57
650 GeV/c^2 T_{pair}	$8.4 \pm 0.2^{+0.6}_{-0.5}$	N/A
650 GeV/c^2 B_{pair}	$13.9 \pm 0.3^{+0.9}_{-1.2}$	N/A
650 GeV/c^2 T_{single}	N/A	$7.7 \pm 0.6^{+0.8}_{-0.6}$
650 GeV/c^2 B_{single}	N/A	$1.9 \pm 0.1^{+0.3}_{-0.2}$

Table 4.6: The total yields of the square cuts signal regions, where square cut 5a was $H_T(jets) > 600$ GeV and 5b was $N(forward\ jets) \geq 1$. Simulated predicted yields are given as *predicted yield* $\pm$ *MC stat.* $^{+up}_{-down}$ *other systematics.*

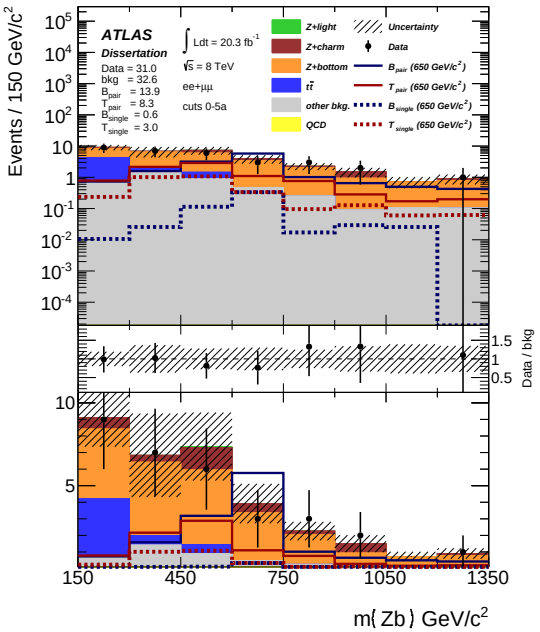
Figures 4.17 (a) and (b) show the unblinded $H_T(jets)$ and $N(forward\ jets)$ variables that were used to determine the final signal regions. Figures 4.17 (c) and (d) show the unblinded $m(Zb)$ discriminating variable in the two signal regions. These were used to estimate the upper limits at 95% confidence level of the measured cross sections for the 59 signal masses of the four VLQ signal types.



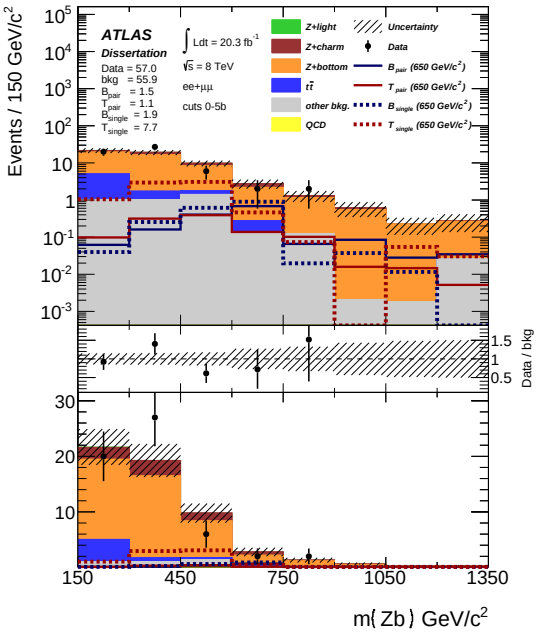
(a) $H_T(jets)$ unblinded.



(b) $N(forward\ jets)$ unblinded.



(c) $m(Zb)$ for T_{pair} and B_{pair}.



(d) $m(Zb)$ for T_{single} and B_{single}.

Figure 4.17: (a) The unblinded $H_T(jets)$, where $H_T(jets) > 600$ GeV determined the signal region for T_{pair} and B_{pair}. (b) $N(forward\ jets)$ where the second bin gives $N(forward\ jets) \geq 1$, which determines the signal region for T_{single} and B_{single}. (c)–(d) The unblinded $m(Zb)$ discriminating variable in the signal regions.

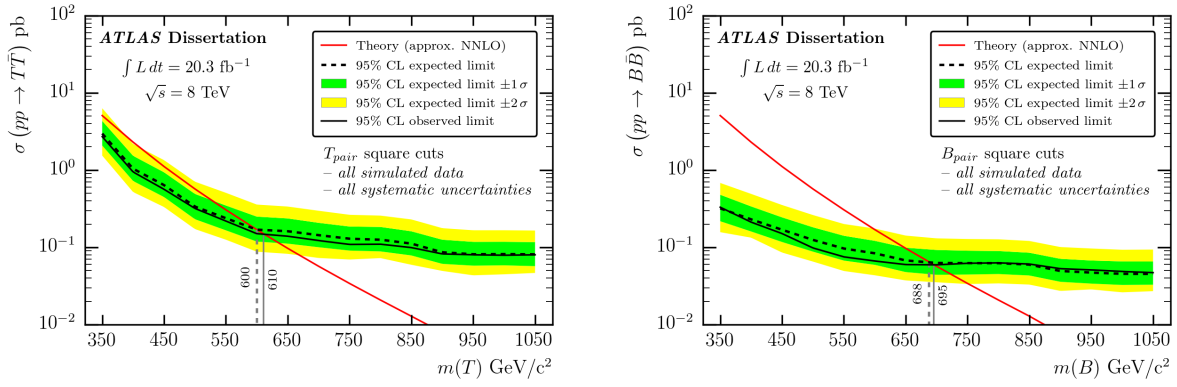
4.6. Results of the Square Cuts Search Strategy

Confidence intervals (upper limits) were computed for the expected and measured cross sections of each of the 59 signal masses of the four VLQ signal types at a 95% confidence level. This was done using the Neyman construction of confidence regions method of statistical inference that is the central topic of chapter 6. More specifically, the discriminating variable spectra were modeled with the multivariate Poisson model (see section 6.2.1 on page 6.2.1), and the upper limits were computed using the DZero recommended test statistic (section 6.2.3.4, page 285) and implementing the CL_S method (page 289 and also section 6.2.4, page 295). All systematic uncertainties were modeled as nuisance parameters with asymmetric Gaussian-like prior distributions sing the "MCLimit" interpolation and extrapolation method (section 6.2.1, page 258), and implemented using the Profiling treatment for all nuisance parameters in a hybrid Bayesian-frequentist manner (section 6.5.1, page 392). For robustness, the upper limits were recomputed using the Null treatment method of nuisance parameters (section 6.5.1), which nearly exactly reproduced the profiled results.[20] These were also recomputed without systematic uncertainties, the "stat only" upper limits, and were found to be consistent with the profiled results (slightly smaller upper limits as expected, but no significant discrepancies were observed). These robustness checks are provided in appendix section C.4, page 512.

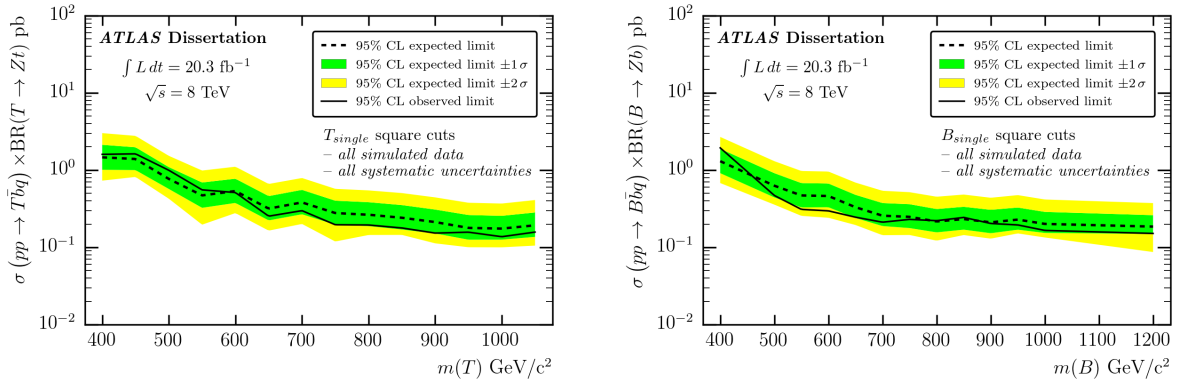
For a given signal mass, the computation of an 95% confidence level upper limit on its cross section, σ_u, represented a confidence interval with a fixed lower limit of 0 so that the true cross section, σ_{true}, was said to lie within the interval $[0, \sigma_u]$ with 95% confidence. This confidence interval was then used to infer whether or not the VLQ signal existed at that signal mass with the predicted theoretical cross section. For example, if σ_{th} was the theoretical cross section of a VLQ for a particular invariant mass, and if $[0, \sigma_u]$ was the 95% confidence interval with its upper limit below the theoretical cross section, $\sigma_u < \sigma_{th}$, then the evidence provided by the measurement was said to exclude the existence of the VLQs with that invariant mass with 95% confidence. This is because 95% confidence intervals contain the true value of the cross section in 95% of experiments, and if the true VLQ cross section was below the theoretical cross section, then the theory hypothesizing VLQ existence was incompatible with the result of the measurement.

Figures 4.18 (a)–(d) provide the resultant upper limits for the VLQ square cuts search. These upper limits were for the VLQ models described in section 2.2 (page 42). VLQs produced in pairs were model independent with cross sections that only depended on the signal mass. Only the decay branching ratios for these were determined by the SU(2) singlet model parameters. For T_{pair} and B_{pair}, limits for all other possible models (all possible decay branching ratios) were computed by the Z-tag search team. These results may be found in [41, 61–64, 149, 150].

[20] A previous study also found that upper limits computed with either the Profiling or else the Null methods were observed to be nearly identical [154].



(a) T_{pair} (b) B_{pair}



(c) T_{single} (d) B_{single}

Figure 4.18: The 95% CL_S upper limits of the four VLQ signal types computed using the DZERO test statistic, the PROFILING treatment of nuisance parameters, and implementing the CL_S method.

Figures 4.18 (a) and (b) show that the T_{pair} and B_{pair} SU(2) singlet models were excluded for signal masses below 610 GeV/c^2 and 695 GeV/c^2, respectively, with the observed cross section upper limit (solid black line) intersecting the theoretical cross section (red line) at those values. These also show that the observed upper limits were consistent with the background-only expectations (dashed black lines), and well within the one and two standard deviation uncertainties of those expectations (green and yellow bands, respectively).[21]

[21] A one sigma upper limit represents the value that covers (with 95% confidence) 84.1% of outcomes (by probability mass) generated by the background-only hypothesis. See the discussion of figure 6.17 (c) on page 297.

The theoretical cross sections are not shown in figures 4.18 (c) and (d) for the T_{single} and B_{single} models because the search was not sensitive to these signal types (i.e., the theoretical cross sections were well below the upper limits obtained from the search).[22] The T_{single} and B_{single} models were not model independent, with both the theoretical cross sections and branching ratios depending on the model parameters (discussed in section 2.2). Nevertheless, these results were the first such measurements reported in the literature (as first reported by the Z-tag search team in [41]). Furthermore, the theoretical cross sections (at $\sqrt{s} = 8$ TeV collision energies) of the singly produced VLQs were larger than those of the pair produced VLQs at large VLQ invariant masses, and this motivated obtaining these early results for singly produced VLQs. The results of this search provided evidence against VLQ existence for invariant masses below ∼600–700 GeV/c^2. As more data is collected at higher collision energies by ATLAS in the coming years, the search for the T_{single} and B_{single} signal types may provide the best opportunity for detection of VLQs having larger invariant masses at the TeV scale.

[22] Figures showing upper limits, such as figures 4.18 (a)–(d), were computed for 13–16 specific signal masses, and values for masses between these were obtained by exponential interpolation. For figures 4.18 (c) and (d), this interpolation resulted in a "choppy" appearance. This was due to data insufficiency (a small number of simulated signal events), particularly for T_{single} in figure (c). These could have been made less choppy by widening the bins of the $m(Zb)$ discriminating variable, but at the cost of larger upper limits.

Chapter 5
The Search Using Machine Learning

"Human performance suffers because we are, deep down, deterministic thinkers with an aversion to probabilistic strategies that accept the inevitability of error. We insist on looking for order in random sequences."
– Philip E. Tetlock [155]

"Beware of geeks bearing formulas."
– Warren E. Buffett [156]

5.1. Introduction to the Search Using Machine Learning

The most common use of machine learning in experimental particle physics searches is to classify signal events from background events [157]. However, many particle physics searches still implement a standard square cuts search methodology (discussed in chapter 5), which implies a cost-benefit tradeoff. The benefit of machine learning is the possibility of improved search significance (defined in section 5.5.2.1). The main drawbacks are the increased complexity of the search that requires additional validation, and the potential loss of valuable data spent in training learning algorithms that may lead to increased statistical uncertainty and thus degrade the power of the final statistical inference. The above quote by Tetlock describes a weakness of human learning. We are wired to look for patterns everywhere and are prone to discover them when they don't exist by misinterpreting noise. Machine learning algorithms may also be misled by noise, especially as the volume of data available for training becomes limited ("data insufficiency"). Therefore, validation of machine learning must be conducted to avoid drawing erroneous conclusions.

The work presented here built on that of the ATLAS exotics Z-tag search team that used the standard square cuts approach to search for the existence of vector-like quarks [41, 61–64, 149, 150]. This chapter presents research that examined how various machine learning techniques were used to extend the reach of that search.[1]

[1] Most machine learning was performed using TMVA: Toolkit for Multivariate Data Analysis with ROOT v.4.2.0 included within the C++ software package ROOT v.5.34.18 [158–160]. Figures that have "TMVA: Toolkit for Multivariate Data Analysis" printed at the top were often created by hand specifically for this work or were modified versions of default TMVA figures in order to obtain correct handling of all event weights so that predicted yields, histograms and statistical errors would be exact. Some new figures were also useful in hyper-parameter optimization (described in section 5.5.2 and section 5.5.2.3). Unsupervised machine learning was conducted with Scikit-learn v.0.14.1 using Python v.2.7.3 which utilized the scientific programming modules NumPy v.1.9.2 and SciPy v.0.14.0, with figures created using MatPlotLib v.1.4.3 [161–163].

Early attempts made at improving upon the square cuts Z-tag search using single-stage machine learning (supervised classification techniques) failed to improve the Z-tag search. Although, data insufficiency due to sparsely populated signal regions[2] was an ever present challenge, the primary obstacle was the similarity of signal and background decay topologies, i.e., their similarity in the values of the input variables. These early failures of single-stage learning motivated the use of multi-stage learning and machine learning regression in addition to classification.

Figure 5.1(a)–(c) illustrate the difficulty of this task. Figures 5.1(a) and (b) show the distributions of two pairs of input variables used for machine learning (discussed in section 5.5.1) using the simulated background data set and one of the simulated signal data sets. These figures demonstrate the large degree to which this particular VLQ signal's input variables' topologies were similar to those of the background. Figure 5.1(c) shows the square cuts of chapter 4 that were applied to the distribution shown in figure 5.1(a).

Table 5.1 lists the four signal types of vector-like quarks and their invariant masses ("signal masses") that were investigated. There were 13–16 signal masses for each signal presenting a total of 59 individual searches that were conducted. Often when machine learning is used in a particle search, not all of the signal masses are used in the training of the learning algorithm (as further described in section 5.3). Ideally, training would be implemented for each unique signal mass. However, 59 unique searches is impractical as this would require an enormous amount of modeling validation. Nevertheless, signal topologies vary greatly over such a wide range of signal masses. If only central signal masses are used in training, then search sensitivity may degrade for other masses. For each signal type, use of multi-stage machine learning provided a way to train each signal mass separately in the first stage, and then combine the results with a second stage of learning into a single classification tool that applied to all masses simultaneously.

signal type	signal masses GeV/c^2	# signal masses
T_{pair}	350–1100	16
B_{pair}	350–1100	16
T_{single}	400–1050	14
B_{single}	400–1200	13
total		59

Table 5.1: List of four vector-like quark signal types and their signal masses, where a statistical inference procedure was performed separately for each signal mass.

[2]A signal region is the region of interest used for statistical inference. It is defined a set of requirements on one or more of the dataset variables. For example, a collision event was required to have a value of 150 GeV/c or higher in the $p_T(Z)$ variable for the event to be included in the each of the signal regions analyzed in this chapter.

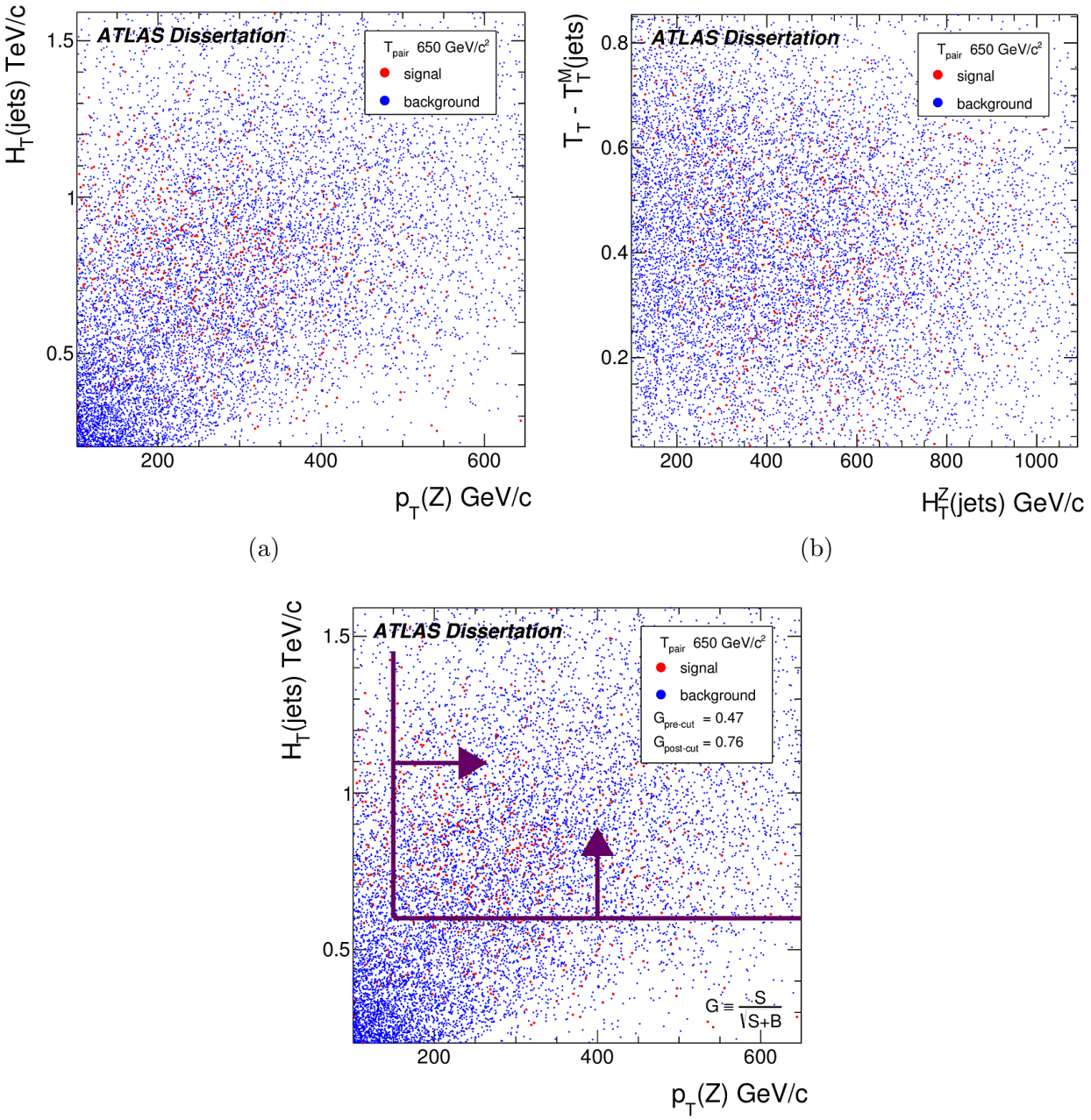


Figure 5.1: Demonstration of difficulty of signal and background separation in the search for VLQs using the $m(T_{pair}) = 650$ GeV/c^2 signal data as an example. The large amount of overlap between signal and background is shown (unweighted) for the input variable pairs (a) $H_T(jets)$ vs. $p_T(Z)$, and (b) $T_T - T_T^M(jets)$ vs. $H_T^Z(jets)$ (discussed in section 5.5.1). (c) The significance (discussed in section 5.5.2.1) achieved by the application of the standard square cuts methodology of chapter 4 with $p_T(Z) >$ 150 GeV/c and $H_T(jets) > 600$ GeV/c.

The remainder of this chapter describes how multi-stage machine learning classification and regression were used in the search for vector-like quarks, and proceeds as follows.

- Section 5.2 (p. 110) describes the machine learning algorithms that were used.
- Section 5.3 (p. 121) provides a literature review of how machine learning has been used in other particle physics searches.
- Section 5.4 (p. 124): the data partitioning schemes that were used for multi-stage learning.
- Section 5.5 (p. 126) explains the first stage of machine learning for the search. This section examined how hyper-parameter optimization was automated so that a grid search could be conducted to find the best 59 separate classifiers for each specific signal mass.
- Section 5.6 (p. 153) describes how the 59 resultant classifier response values from stage 1 learning were then used as input variables for a second stage of machine learning *classification*.
- Section 5.7 (p. 170): describes stage 3 learning where the stage 1 classifiers were used as input variables for machine learning *regression*, and combined with the results of the stage 2 classifiers in order to predict the signal masses as regressands.
- Section 5.8 (p. 187) shows how the simulated data accurately modeled the measured data after machine learning.
- Section 5.9 (p. 202) shows the unblinding of the searches by comparing the measured data to the simulated data in the signal regions.
- Section 5.10 (p. 215) shows the final results, the upper limits on the signal strengths (cross sections).
- Section 5.11 (p. 246) summarizes the methodologies that were developed here and their results.

5.2. Methodology of Machine Learning

This section provides a brief summary of machine learning methodology as it pertains to this work. It is not meant to be an exhaustive description of machine learning techniques as such literature is readily available (such as [157, 164–168]).

Machine learning is the use of learning algorithms to discover patterns in training data. These learned patterns can then be used to predict outcomes in non-training (test) data. *Supervised learning* describes the process of determining patterns in training data input variables that correlate with a known outcome, and then predicting the unknown outcome of test data, such as whether or not each test datum is from a particular category (classification), or predicting a continuous regressand (regression) for the test data. *Unsupervised learning* is used to identify clusters of training data which share similar input variables in the absence of any known outcome, and then predicting to which cluster each test datum belongs. Supervised learning is most often used in particle physics searches since the simulated data is already known to be generated either by the Standard Model (background events) or else by the new theory (signal events) [157]. (A short investigation of *unsupervised learning* applied to the search for VLQs is reported in section B.2.)

During supervised classification, signal and background training data are provided to the learning algorithm, which then attempts to find the best function of input variables for separating the signal from the background. The output is a classifier function that returns a response value for each datum event relating to the probability that the event is from the signal training data. Alternatively, if there exists a continuous target variable which differs between the signal and background data, supervised regression may be used to forecast this target value. Since a new theory of particle physics hypothesizes a new particle with an invariant mass, supervised regression may be used to predict this signal mass.

5.2.1. Learning Algorithms

In supervised learning, a machine learning algorithm is given a subset of the simulated data on which to train, $D_{train} = \{(y_i, \mathbf{X}_i)\}$, with input variables $\mathbf{X}_i$ and known outcome y_i. The outcome y_i may be discrete classes (such as the binary classification of signal vs. background), or may be a continuous regression target (such as the signal mass of a particle). The learning algorithm then seeks to find the best predictor function $H_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\mathbf{X}_i)$, where $\boldsymbol{\alpha}$ is a set of learning parameters that the learning algorithm seeks to optimize, and $\boldsymbol{\beta}$ is a set of fixed hyper-parameters that are tuned by hand. The final predictor function is found by sequentially generating j predictor functions $h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}}(\mathbf{X}_i)$ by varying the learning parameters $\boldsymbol{\alpha}_j$.

Each predictor function $h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}}$ provides predicted outcomes for each of the training data $\hat{y}_{\boldsymbol{\alpha}_j,\boldsymbol{\beta},i} = h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}}(\mathbf{X}_i)$ with some measurable error $y_i - \hat{y}_{\boldsymbol{\alpha}_j,\boldsymbol{\beta},i}$ since the true outcome of the training data is known in supervised learning. A loss function of this error $L(y_i - \hat{y}_{\boldsymbol{\alpha}_j,\boldsymbol{\beta},i})$ determines the accuracy of a given predictor function for a given datum. Learning the $\boldsymbol{\alpha}$ parameters occurs sequentially as the algorithm explores the selection of $\boldsymbol{\alpha}_j$ which decreases the total loss of the training data,

$$\mathcal{L}\left(\boldsymbol{\alpha}_j, \boldsymbol{\beta}, D_{train}\right) = \sum_{i \in D} L(y_i - \hat{y}_{\boldsymbol{\alpha}_j,\boldsymbol{\beta},i}), \tag{5.2.1}$$

One may regard each predictor functions $h_{\boldsymbol{\alpha}_j,\boldsymbol{\beta}}$ as a separate statistical model, and the machine learning that occurs as the adjustment of the model parameters $\boldsymbol{\alpha}_j$. As with any statistical model, the more complex the model, the more likely it is to overtrain (overfit). Prevention of overtraining may be achieved by restricting the learning algorithm to less complex statistical models (see chapter 4 of [165] or chapter 3 of [164]), or by limiting the number of models examined during learning.

As more statistical models $\boldsymbol{\alpha}_j$ are evaluated during learning, the more likely overtraining is to occur. This is because the training data set D_{train} represents a random sampling from a true and unknowable probability density. The sampled training data will randomly deviate from the true density to some degree, and the learning algorithm may "learn" these statistical fluctuations and go on to yield poor out-of-sample predictions. For example, consider a training data set with 20 input variables used for classification of signal and background. If each variable is binned into 10 values (a coarse level of granularity), then the "volume" of the entire space of input variables is 10^{20}. Such a large volume can never be spanned by a reasonable amount of training data, and so the sampling will always be sparse. As the number of models attempted by the learning algorithm increases, it will be more likely that statistical fluctuations inherent to the training data will be be used for signal and background separation. This will result in poor out-of-sample performance, i.e., increased generalization error.

Bootstrapping methods are often implemented to mitigate the effects of such sampling bias that leads to generalization error, namely bagging or "bootstrap aggregating". Bagging is the procedure of providing only a portion of the training data to the learning algorithm during each iteration of learning (sampling from D_{train} with replacement) [169]. Bagging guards against any particular statistical fluctuation of the training sample from being used at every iteration by the learning algorithm.

Validation is the most important procedure used in machine learning to prevent overtraining. A validation data set is a portion of labeled data that is not included in training, which may then provide an out-of-sample estimate of the expected performance on unlabeled test data [165]. Of course, reducing the amount of data available for training will reduce the performance of the final predictor function.

The machine learning strategy in the search for VLQs was to first find the optimal learning algorithm to use to separately train each of the 59 signal masses. The resultant 59 stage 1 classifiers were then used as inputs to four stage 2 learning algorithms, one for each signal type, T_{pair}, B_{pair}, T_{single}, and B_{single}. In this manner, training would occur for each signal mass separately in stage 1, while after stage 2, there would only be one classifier for each signal type so that only four separate analyses would be performed as was done with the standard square cuts analysis of chapter 4.

5.2.2. Boosted Decision Trees

The decision tree learning algorithm constructs a collection of decision nodes that are selected to best separate signal and background classes as shown in the example of figure 5.2 (compare this to the square cuts search of chapter 4 depicted as a decision tree in figure 5.3). The decision tree is a collection of decision nodes where all the data enters at the top node and is split into groups at each subsequent decision node until being collected into separate groups at each of the final "leaf" nodes[3]. The splitting of the data at a decision node is the result of applying a square cut on one of the input variables.

The selection of which variable to use for splitting for a given decision node, and what value of that variable to use for the square cut is determined during the training of the decision tree. First, all signal and background training data, $D_{T,S}$ and $D_{T,B}$, enter the top node of the tree (possibly bagged subsets), and each input variable is examined for the cut value that provides the greatest separation score (described below). The training data events are then split into two groups according to the selected variable (greater or less than the cut value). In turn, each of these two groups become intermediate decision nodes for further splitting, which continues until no positive splitting score is obtained for any variable, or else a minimum number of events in each node is reached, with these final nodes becoming leaf nodes. The leaf nodes are then used for evaluation based on the proportion of signal training data to background in each leaf node.

The separation score for a node is determined by a splitting function, which for classification done in this research was the Gini index that measures the impurity of a node. For a given node containing N_S signal events and N_B background events, the node has signal fraction $f_S = \frac{N_S}{N_S+N_B}$, and the Gini index for this node is $G = f_S(1 - f_S)$. Note that the Gini index is equivalent to the variance of a Bernoulli distribution modeling the probability

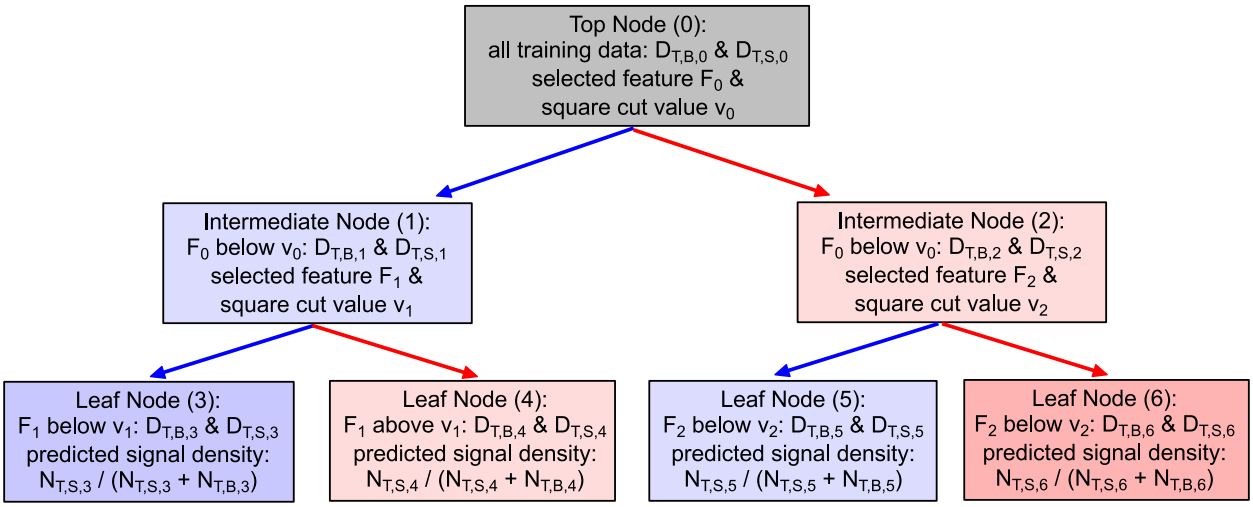


Figure 5.2: Example of a simple decision tree. Increasingly blue represents a higher concentration of background data while increasingly red is signal.

[3]The specific implementation used in this research was the CART algorithm: Classification and Regression Trees [164], as implemented by the TMVA program [158], where the term "decision tree" in this work is used to refer to both classification and regression trees.

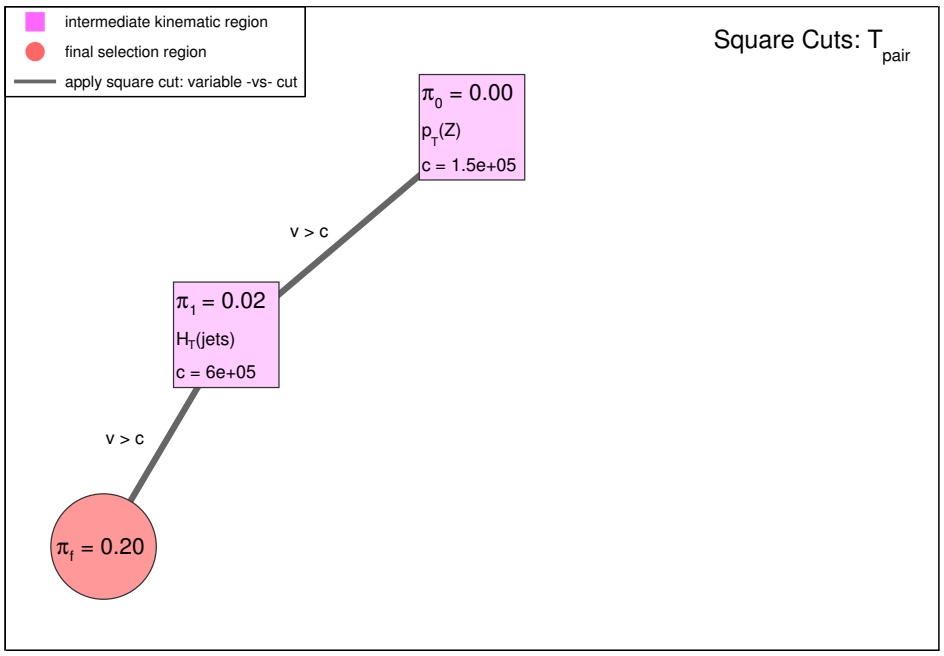


Figure 5.3: The square cuts search for the T_{pair} signal of chapter 4 may be depicted as a decision tree containing a single path with three nodes. Purity begins very small, $N_S/(N_S + N_B) < 0.005$, but increases to $\sim$20% after two square cuts.

of selecting a signal event at this node as equal to f_S. A proposed splitting would divide the signal and background into two groups $(N_{S,1}, N_{B,1})$ and $(N_{S,2}, N_{B,2})$ with G_1 and G_2, respectively. The separation score is then given by $G - \frac{N_{S,1}+N_{B,1}}{N_S+N_B}G_1 - \frac{N_{S,2}+N_{B,2}}{N_S+N_B}G_2$, which is a measure of the decrease in the variance of signal selection by splitting the node into two subnodes. Other splitting functions (cross entropy, misclassification error, and significance) were examined, but differed little in performance. For regression, the node splitting function used was the average squared error $\frac{1}{N}\sum(y - \hat{y})^2$.

If a single decision tree was to be used, it would need to be grown deeply with many levels of subnodes in order to provide substantial separation, and this usually leads to overtraining. Instead of growing a single, deep decision tree, an alternative is to grow many shallow decision trees, and combine their separate predictions into a single "forest of weak learners", which in aggregate can provide a powerful learning algorithm [170]. This ensemble method greatly reduces overtraining since statistical fluctuations of training data are less likely to be learned by shallow trees, and even if some trees are overtrained, they individually contribute only a small amount to the final result.

It is natural to wonder why all the trees in a forest are not identical since they follow from the same learning algorithm. The use of bagging can lead to slightly differentiated trees since each tree learns from a different subsample of the training data. However, randomizing the forests may be used to create highly variable trees. During tree construction in a randomized forest, only a random subset of the input variables are available for use during node splitting. Considering the many nodes of a tree, the large number of possible permutations of selected input variables greatly differentiates each tree of the forest. Furthermore, restricting nodes to only random subsets of the input variables prevents any particular input variable from

being utilized much more frequently than the others, and this also aids in the reduction of overfitting by restricting the reliance on only the most discriminating variables during tree construction. Boosting techniques also result in differentiated trees of a forest, and can greatly enhance the performance of the final classifier.

The performance of the decision tree forest may be enhanced when boosting is introduced, wherein subsequent trees of the forest attempt to improve upon the results of the previous trees. Two boosting techniques were implemented in this work: adaptive boosting and gradient boosting.

The AdaBoost algorithm was used for adaptive boosting of the classification decision trees in this work. AdaBoost works by increasing the weights of events that were misclassified by the previous tree iteration, and decreasing the weights of events that were correctly classified [158, 171–173]. In this manner, learning occurs by iteratively adjusting the weight of each datum that is to be used in node splitting decisions. Specifically, a boost scale factor k_t is derived for the decision tree at iteration t,

$$k_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \tag{5.2.2}$$

where ϵ_t is the weighted error rate of misclassified events of the tree and must be less than 0.5,

$$\epsilon_t = \sum_{i \in mis.} w_{t,i} \Big/ \sum_i w_{t,i}, \tag{5.2.3}$$

with $w_{t,i}$ the boosting weight of event i for iteration tree t. The scale factor, along with a user selected learning rate $0 < \lambda_b < 1$, then determine each data weight for the next iteration,

$$w_{i,t+1} \to e^{\pm k_t \lambda_b} w_{i,t}, \tag{5.2.4}$$

with "+" for misclassified events at iteration t (a weight increase), and "−" for correctly classified events (a weight decrease). These new weights are then renormalized so that the sum of event weights remains constant during learning. Learning terminates at iteration T with the ensemble prediction for a given event i determined by

$$H(\mathbf{X}_i) = \frac{1}{T}\sum_t^T \lambda_b k_t h_t(\mathbf{X}_i), \tag{5.2.5}$$

where $h_t(\mathbf{X}_i)$ is the prediction for event i by tree t.

An alternative to adaptive boosting is gradient boosting (technically, the AdaBoost algorithm is a special case of gradient boosting [172]). Gradient boosting constructs a prediction function $H_T(\mathbf{X})$ by iteratively applying a gradient descent procedure during T decision tree iterations [158, 174]. Let $H_\tau(\mathbf{X})$ be the current predictor function constructed from the first τ trees with

$$H_\tau(\mathbf{X}) = \sum_t^\tau \lambda_t h_t(\mathbf{X}), \tag{5.2.6}$$

and with some tree weighting parameter λ_t. At this stage of iteration, $H_\tau(\mathbf{X}_i)$ generates

residuals

$$r_{\tau,i} = y_i - H_\tau(\mathbf{X}_i). \tag{5.2.7}$$

Noting that if a "perfect" predictor function (of the training data only) could be obtained by using $H_\tau(\mathbf{X}_i) - r_{\tau,i}$, then the next iteration tree could then trained on the (negative) residuals, $r_{\tau,i}$, so that $H_\tau(\mathbf{X}_i) + h_{\tau+1}(r_{\tau,i})$ would provide an improved predictor. A more general approach is to replace the use of residuals with the gradient of some loss function, $L(y_i, H_\tau(\mathbf{X}_i))$, so that the next iteration tree is fit to the negative gradient of the loss function in order to minimize the expected loss after the next iteration,

$$g_{\tau,i} = -\left.\frac{\partial L(y_i, H(\mathbf{X}_i))}{\partial H(\mathbf{X}_i)}\right|_{H=H_\tau}. \tag{5.2.8}$$

Note that $g_{\tau,i}$ equals the negative residual when the squared error loss function, $L(y_i, H_\tau(\mathbf{X}_i)) = \frac{1}{2}(y_i - H_\tau(\mathbf{X}_i))^2$ is used. The updated predictor function is then taken as

$$H_{\tau+1} = H_\tau + \lambda_\tau g_{\tau,i}, \tag{5.2.9}$$

where λ_τ is given by loss minimization

$$\lambda_\tau = \arg\min_\lambda L(H_\tau + \lambda g_{\tau,i}). \tag{5.2.10}$$

Finally, a degree of regularization may be implemented by instead updating the predictor function to

$$H_{\tau+1} = H_\tau + \nu\lambda_\tau g_{\tau,i}, \tag{5.2.11}$$

where $0 < \nu < 1$ is the "shrinkage" hyper-parameter that prevents the update from realizing the full gradient. The shrinkage regularization increases the stability of the learning algorithm by preventing rapid gradient descent.[4]

[4] A technical discussion of the Vapnik-Chervonenkis generalization bound that predicts out-of-sample performance may be found in [165]. For binary classification, let $E_{IS}(h)$ be the fraction of misclassified training data and $E_{OOS}(h)$ be the expected misclassification rate for out-of-sample test data, where h is the predictor function created by the learning algorithm. Then for a given tolerance level δ, E_{OOS} is bounded above with $E_{OOS} \leq E_{IS} + \sqrt{\frac{8}{N_{train}} \ln\left(\frac{4m_{\mathcal{H}}(N_{train})}{\delta}\right)}$ with probability $1 - \delta$, and where $m_{\mathcal{H}}(N_{train})$ is a function describing the growth of complexity of the learning algorithm with the amount of training data. A forest of N_{tree} decision trees of depth N_{depth}, and with $N_{variables} \gg N_{depth}$, has an upper bound on its VC dimension of $D_{VC} \leq 2^{N_{depth} N_{tree}}$. By Sauer's lemma, this gives the polynomial bound $m_{\mathcal{H}}(N) < 2N^{D_{VC}}$. Therefore, the bound on out-of-sample error (as a proxy for the likely amount of overtraining) is expected to increase with $\sqrt{8\ln(8/\delta)/N_{train} + 2^{N_{tree}N_{depth}+3}\ln(N_{train})/N_{train}}$. With 5% tolerance and limited training data, N_{train}~100, this approximate generalization bound grows quickly with BDT complexity, $\sim 0.6 \cdot 2^{N_{tree}N_{depth}/2}$. This predicts that shallow decision trees or forests with a small number of trees are likely needed to prevent overfitting in cases of data insufficiency.

5.2.3. Multi-layer Perceptron

The multi-layer perceptron (MLP) is a feedforward neural network organized into layers as shown in the example of figure 5.4. An MLP may be considered as a series of layers of logistic regression models with the output of one layer becoming the input to the next [164]. Unlike the decision nodes of the decision tree learning algorithm that perform discrete binary decisions on a single input variable (square cuts), each of the neural nodes of the first layer take all input variables simultaneously as inputs, and each produces a continuous output value. These outputs are in turn used as inputs to the next layer until reaching a final output layer that produces the output prediction.

The MLP is able to approximate a large class of functions [175], and therefore has the potential to closely approximate a true underlying probability density. However, experimentation here revealed that as the volume of training data became more limited, the MLP's power became a liability, often leading to overtraining unless extensive hyper-parameter tuning was employed. Very often in this work, MLPs were rejected in favor of decision trees due to limitations in the computational resources needed to optimally tune the MLP hyper-parameters and to select the optimal subset of input variables. However, when no hidden layers are used (and using cross-entropy loss), the MLP is equivalent to simple logistic regression [176], and this single node MLP was used often in this work to provide logistic regression results (logistic regression is discussed in section 5.2.5).

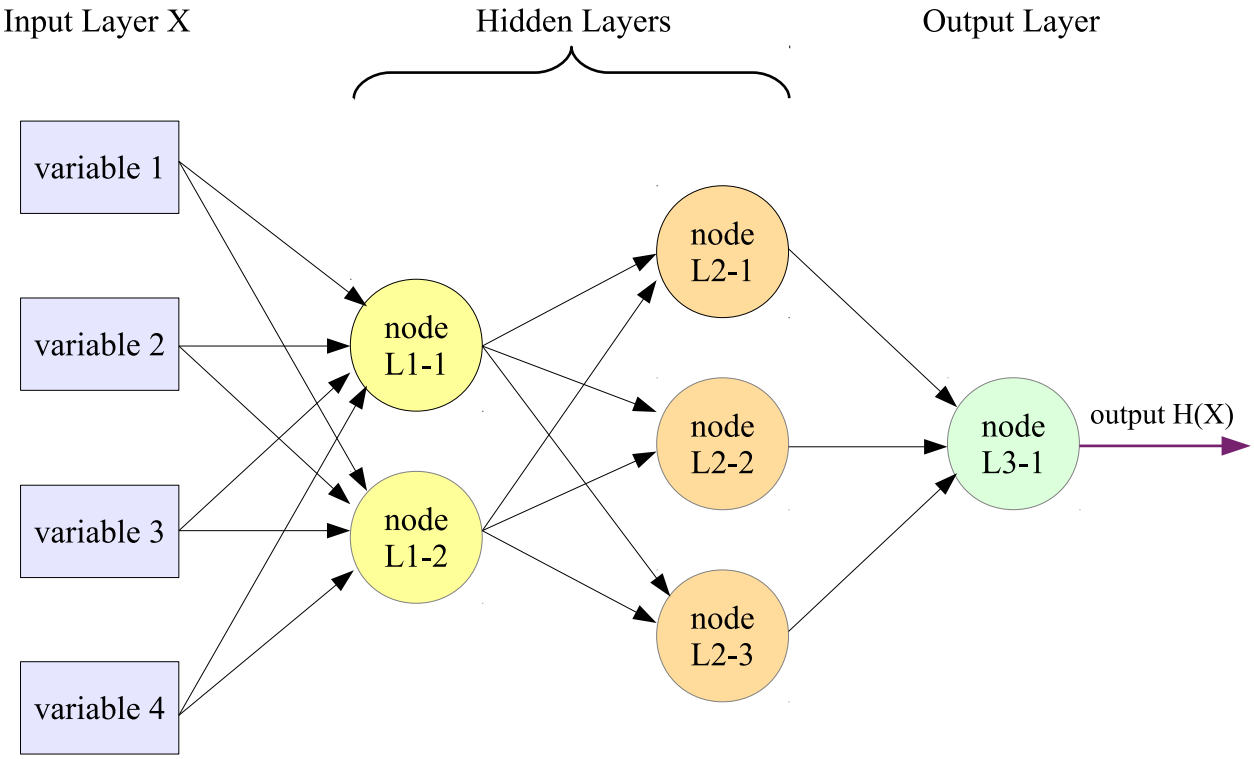


Figure 5.4: Example of a simple multi-layer perceptron with an input layer containing four input variables, two hidden layers containing two and three neural nodes, respectively, and an output layer with a single neural node for a single output. (Note that bias terms for each layer are not shown in the figure.)

The mechanics of the MLP are best described by examining how the MLP produces a final output prediction for a particular datum, $H(\mathbf{X}_i)$. Let $Z_{l,m}$ be the output of the m^{th} node in layer l for this datum, including $l = 0$, the values of the input variables. Also, for every layer a constant bias term $Z_{l,0} = 1$ is included as input for the next layer. The n^{th} node in the subsequent layer $l + 1$ applies weights $W_{l+1,m,n}$ to each of the previous layer's outputs to determine the "activation" value for this datum at this node[5],

$$a_{l+1,n} = \sum_m W_{l+1,m,n} Z_{l,m}. \tag{5.2.12}$$

The output from this node $Z_{l+1,n}$ then follows from an activation function[6],

$$Z_{l+1,n} = \mathcal{A}(a_{l+1,n}), \tag{5.2.13}$$

with the final prediction being the output from the single node of the last output layer L,

$$H_{\mathbf{W}}(\mathbf{X}_i) = Z_{L,1}. \tag{5.2.14}$$

It is in this manner that the final output value for a datum is determined, where the activation function outputs from one layer become the inputs to the next layer "feeding forward" with the neural node weights uniquely determining the output of the MLP.

During a training cycle t, node weights are updated sequentially for every event i (thus online) based on the gradient of the loss function for that datum,

$$\mathbf{W}_{(t+1)\cdot i} = \mathbf{W}_{t\cdot i} - \eta \frac{\partial L(y_i, H_{W_{t\cdot i}}(\mathcal{X}_i))}{\partial W_{t\cdot i}}, \tag{5.2.15}$$

where η is a user selected learning rate. The term "backpropagation" follows from the computation of the loss function gradient, which begins from the output layer and continues to the input layer sequentially through the hidden layers (see section 16.5.4 of [164] for an example).

[5] A synaptic function may also be used to transform the weighted input, $a_{l+1,n} = \sum_m \mathcal{S}(W_{l+1,m,n} Z_{l,m})$, but none was used in this work.

[6] The activation function used in this work was the sigmoid (logistic) function, $\mathcal{A}(a) = \frac{1}{1+e^{-a}}$, the only option available with the TMVA software at the time this analysis was conducted.

5.2.4. Least Squares Linear Regression

Least squares linear regression was used extensively in this work for supervised regression when more sophisticated machine learning techniques failed to provide useful or valid results. A linear regression model has the form,

$$\begin{aligned} y_i &= \mathbf{X}_i\boldsymbol{\beta} + u_i, \text{ (for datum } i\text{), or} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \text{ (in matrix form),} \end{aligned} \tag{5.2.16}$$

where $\mathbf{X}$ is the input variables of the data with the addition of a constant bias term $X_0 = 1$, and $\boldsymbol{\beta}$ is an array of fit parameters of length $N_{variables}+1$. Each datum is modeled to contain a random disturbance u_i that alters the value of the dependent variable y_i, with u_i assumed to be normally distributed with zero mean and modeled with some covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_u)$. The disturbance covariance $\boldsymbol{\Sigma}_u$ is an $N_{train} \times N_{train}$ matrix, and for the simplest model may be assumed to be a constant $\boldsymbol{\Sigma}_u = \sigma^2\mathbf{I}$ for all data points, i.e., the homoscedastic assumption of ordinary least squares (OLS). Note that in particle physics searches, each simulated datum is independent from all others so that $\boldsymbol{\Sigma}_u$ must always be diagonal. However, the disturbances may not be identically distributed because the simulated backgrounds and signals arise from differing physical processes. Usually $\boldsymbol{\Sigma}_u$ is unknown and must be estimated from the training data, generating an additional 1 or more additional fit parameters.

Interactions between the input variables may be included in the model with the addition of a quadratic term

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{X}\mathbf{B}\mathbf{X} + \mathbf{u}, \tag{5.2.17}$$

where $\mathbf{B}$ is a symmetric $N_{variables} \times N_{variables}$ matrix. When implemented in this work, the diagonal elements of $\mathbf{B}$ were set to zero to disallow quadratic terms for individual variables, so that $\mathbf{B}$ contained an additional $0.5\,(N_{variables} - 1)^2$ fit parameters.

The least squares linear regression model is uniquely determined by its fit parameters. Unlike the learning parameters of the previously described machine learning methods, the fit parameters are computed at once as the exact solutions to the minimization of the quadratic loss function, $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2$, (the "least squares"), which gives unbiased fit parameter estimates [177]

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{5.2.18}$$

$$\hat{\sigma}^2 = \frac{\mathbf{e}^2}{N_{data} - 2}, \tag{5.2.19}$$

where $\mathbf{e}$ are the training data residuals, $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. This provides the predictor function

$$H_{\hat{\beta},\hat{\sigma}}(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{5.2.20}$$

The simulated data of particle physics contains event weights spanning many orders of magnitude. This suggests the use of weighted least squares (WLS), in which the disturbance covariance matrix is modeled as $\boldsymbol{\Sigma} = \sigma^2\text{diag}(1/w_i)$ [178]. For the case of WLS or heteroscedastic disturbances in general (generalized least squares, GLS) with $\boldsymbol{\Omega} \equiv \frac{1}{\sigma^2}\boldsymbol{\Sigma}$,

the fit parameters are estimated as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}, \tag{5.2.21}$$

$$\hat{\sigma}^2 = \frac{\mathbf{e}_w^2}{N_{data} - 2}, \tag{5.2.22}$$

where $\mathbf{e}_w$ are the weighted training data residuals, $\mathbf{e}_w = \mathbf{w}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. This provides the predictor function

$$H_{\hat{\beta},\hat{\sigma}}(\mathbf{X}) = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}. \tag{5.2.23}$$

It is often useful to examine the uncertainty of the fit parameters of the trained regression model. The covariance matrix of $\boldsymbol{\beta}$ is given by

$$\text{cov}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}'\mathbf{X})^{-1}\frac{\mathbf{e}^2}{N_{train} - N_{variables}} \text{ (OLS)} \tag{5.2.24}$$

$$= (\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\frac{\mathbf{e}_w^2}{N_{train} - N_{variables}} \text{ (WLS or GLS)}. \tag{5.2.25}$$

with the diagonal entries providing the variance estimates of the $\hat{\boldsymbol{\beta}}$ fit parameters.

An often used measure of the model fit quality is the coefficient of determination,

$$R^2 = 1 - \frac{\mathbf{e}^2}{\bar{\mathbf{e}}^2} \geq 0, \tag{5.2.26}$$

where $\bar{\mathbf{e}} = \mathbf{y} - \bar{y}$ is the difference between the observed y values and their average. R^2 is often adjusted to account for increasing model complexity, so that the adjusted R^2 value is

$$\bar{R}^2 = 1 - (1 - R^2)\frac{N_{train} - 1}{N_{train} - N_{variables}}. \tag{5.2.27}$$

A low R^2 value indicates a poor fit to the training data, and implies poor predictability when applied to the test data.

One may investigate the validity of the modeling of the disturbances with constant covariance, $\boldsymbol{\Sigma}_u = \sigma^2\mathbf{I}$, by performing a Breusch-Pagan test. This is done by performing a second stage regression on the squared residuals from the first regression, and forming the test statistic $N_{train}R^2_{2^{nd}}$, which has the distribution $\chi^2_{(N_{variables}-1)}$. A low p-value of this statistic indicates rejection of the null hypothesis of homoscedasticity [179].

When the homoscedastic assumption is deemed invalid, one may apply a heteroscedasticity correction to the estimated uncertainties of the fit parameters. This is achieved by analysis of the regression residuals $\mathbf{e}$, and for small sample sizes, this may be done with the approximation

$$\text{cov}[\hat{\boldsymbol{\beta}}] \approx (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\text{diag}\left[\frac{e_i^2}{(1 - \Gamma_{i,i})^2}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \tag{5.2.28}$$

where $\Gamma = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ [180, 181].

After correcting for heteroscedastic disturbances, it is natural to consider estimating the heteroscedasticity of the disturbance covariance matrix $\boldsymbol{\Sigma}_u$. This may be done via two-stage generalized least squares wherein an auxiliary regression is performed on the logarithm of

the squared residuals [182]

$$\ln\left[\mathbf{e}^2\right] = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\nu}, \tag{5.2.29}$$

where $\boldsymbol{\nu} \sim \mathcal{N}(0, \sigma^2_{2^{nd}}\mathbf{I})$ for OLS, or $\boldsymbol{\nu} \sim \mathcal{N}(0, \sigma^2_{2^{nd}}\mathbf{W}^{-1})$ for WLS. The $N_{variables} + 1$ fit parameters $\boldsymbol{\alpha}$ may then be used to estimate $\boldsymbol{\Sigma}$ and a final regression is then performed.

5.2.5. Logistic Regression (Logit Classification)

In the context of machine learning, the above linear least squares regression is a method of supervised *regression*. Supervised binary classification may also be performed in a similar manner using logistic regression. The logistic regression linear model has the form [157]

$$\ln\frac{\mathbf{p}}{1-\mathbf{p}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \tag{5.2.30}$$

where $\mathbf{p}$ is the probability of a datum belonging to the first class (and $1 - \mathbf{p}$ the probability of being in the second class). In the training data where the correct classification is known, $\mathbf{p}$ is set to 0 or 1 and the model solved for the fit parameters $\hat{\boldsymbol{\beta}}$. Rearranging for $\mathbf{p}$ yields

$$\mathbf{p} = \frac{1}{1 + \exp(-\mathbf{X}\boldsymbol{\beta})}, \tag{5.2.31}$$

which is the logistic (sigmoid) function. Unlike least squares linear regression for which there is an analytic solution to the quadratic loss minimization, the fit parameters of logistic regression $\boldsymbol{\beta}$ must be obtained numerically (see chapter 4 of [176] for a maximum likelihood implementation). This yields the prediction function

$$H_{\hat{\boldsymbol{\beta}}}(\mathbf{X}) = \frac{1}{1 + \exp(-\mathbf{X}\hat{\boldsymbol{\beta}})}, \tag{5.2.32}$$

which predicts the classification probability for the test data. The use of the logistic function in the logistic regression model provides class probability estimates that are bounded appropriately $0 \leq \mathbf{p} \leq 1$.

5.3. Machine Learning in Other Particle Searches

Several different approaches have been made in implementing machine learning in particle physics experiments. Generally, experiments can be categorized by the two types of statistical inference performed in the final signal region: simple counting and spectral discrimination. Simple counting compares the total predicted yields in a signal region, while spectral discrimination examines how those yields are distributed across the spectra of one or more discriminating variables. Spectral discrimination requires greater accuracy of simulated data in modeling the measured data, and a larger amount of data to fit the spectrum accurately.

A short review of recent experimental particle physics searches shows the general methods in which machine learning is utilized with a summary provided in table 5.2. The simplest approach is to train a single classification learning algorithm to produce a classifier value for each event, and then remove data below the selected classification value that maximizes the significance (discussed in section 5.5.2.1). With the remaining data, statistical inference may be performed with simple counting [183–185], or else using a spectrum, such as a reconstructed mass [54, 185–188]. Instead of cutting on a particular classifier response value, often the classifier response distribution itself provides the spectrum for statistical inference [56, 59, 185, 189–201], or is used together with a reconstructed mass (or similar) spectrum, and statistical inference is performed on the two dimensional spectrum [188, 202, 203].

The type of learning algorithm selected for use in particle searches was mostly limited to BDTs and artificial neural networks (ANNs). Very little use of SVMs was observed in the literature (a SVM was found to provide weaker results than an ANN in [204], and the lack of use of SVMs in experimental particle physics is discussed in [205]). Other methods such as generalized linear models and Gaussian mixture models were found to perform as well as BDTs and MLPs in [206], but are rare in the literature. Use of BDTs were more often observed ([59, 183–186, 188, 189, 191, 193–203, 207–209]) than ANNs ([187, 190, 198]).

One reason for the popularity of BDTs may be due to a lower sensitivity to input variable selection, particularly when random forests are used. In direct comparisons, [198] found that an ANN performs better than a BDT when constrained to only 4 input variables, while in [210] it was found that BDTs outperformed ANNs as the number of variables used was increased. Also, it is often much easier to successfully implement a BDT, which tends to return reasonable signal separation with little overtraining over a wide range of hyper-parameters. Furthermore, in [211] it found that BDTs were less sensitive to shifting and smearing of the input variables. Unlike perceptrons, decision trees offer the advantage of analysts being able to more easily examine the interactions among the variables discovered during learning [212].

Simulated data that is used in training learning algorithms may not be reused for statistical inference. This can exacerbate data insufficiency issues and degrade testing power. Cross-evaluation techniques can allow the reuse of all simulated training data during final statistical inference [194, 201, 207]. (See section B.1 for further discussion of cross-evaluation.)

The output of one or more learning algorithms can become the input to another learning algorithm during multi-stage learning. This technique was infrequently found in the literature. In one example, two BDTs identifying two different types of events were used

as inputs to a second stage BDT that was applied to all events [199]. In another, two BDTs identifying the presence of two different reconstructed objects were inputs to a second stage BDT (along with an additional input variable) [209]. And finally, 10 input variables were combined with the output of a first stage BDT as input to a second stage BDT [200].

The analysis described in this dissertation trained separate BDTs for each signal mass, which were then used as inputs (along with three additional input variables) to a variety of second stage learning algorithms. It should be noted that multi-stage machine learning is used in nearly all recent ATLAS experiments that incorporate the b-tagging of jets in events. The JetFitterCOMBNN and MV1 tagging algorithms use ANN classification to combine results from simpler algorithms which themselves use ANN classification [213]. Nevertheless, it seems that few particle searches use multi-stage techniques in event level discrimination.

Reports of machine learning regression are rare with the nearly all particle searches using classification. In [214], BDT regression was used to more accurately estimate b-jet energies, while in [215, 216], BDT regression was used to improve the reconstructed invariant mass of electron pairs and the Higgs boson, respectively. In [204], regression with several learning algorithms was performed in order to predict expected signal yields for arbitrary SUSY parameters from a set of known parameters. The analysis described in this dissertation performed machine learning regression with the VLQ invariant mass as the target variable (where zero mass was used for background events), and found that BDTs (with gradient boosting, but not adaptive boosting) provided good results. In [217, 218], similar techniques were experimented with for Higgs mass estimation using BDTs with adaptive boosting, but encountered problems with the resultant regressand spectrum similar to unreported results in this work using adaptively boosted regression trees (while the use of gradient boosted regression trees discussed in section 5.7.2 did not show these problems).

learning algorithm	number of variables	statistical inference method or measurement method	publication year & [ref.]
BDT	7,13	classifier cut→reconstructed mass spectrum	2012 [54]
BDT	4	classifier cut→simple count	2013 [184]
BDT	39	classifier spectrum	2013 [197]
BDT	6	classifier cut→reconstructed mass spectrum	2014 [186]
BDT	3	classifier spectrum	2014 [202]
BDT	7	classifier cut→simple counting	2014 [183]
BDT	7–12	classifier spectrum	2014 [56]
BDT	6,9	classifier spectrum	2014 [194]
BDT	8	classifier spectrum	2014 [195]
BDT	7	classifier spectrum	2014 [196]
BDT	3	classifier spectrum	2014 [203]
BDT	5	classifier spectrum	2014 [201]
BDT	8–9	classifier spectrum	2014 [199]
BDT	12–14	classifier spectrum	2015 [189]
BDT	12	classifier spectrum	2015 [59]
BDT	19	fit to classifier spectrum	2015 [191]
BDT	5-9	classifier spectrum	2015 [193]
BDT	12	classifier spectrum	2015 [59]
BDT	16	classifier spectrum	2015 [188]
ANN	11–13	fit to classifier spectrum	2014 [190]
ANN	2	fit to classifier spectrum	2014 [192]
ANN	13	classifier cut→reconstructed mass spectrum	2014 [185]
ANN	4	classifier spectrum	2015 [198]
ANN	12	classifier cut→mass measurement	2015 [187]

Table 5.2: Summary of machine learning methods implemented in other particle physics studies. Boosted decision trees and artificial neural networks are the most common algorithms used.

5.4. Data Partitioning for Machine Learning

In many applications of machine learning, a 3-way partitioning is required, one partition for training, one for validating the training, and one for statistical inference. This is done in order to prevent accidental validation of the learning algorithm where the predictions for the training data are very similar to the predictions for the validation data, but the result of learning still produces large generalization error. Accidental validation may occur when analysts tune hyper-parameters by repeatedly training the learning algorithm using the same training and validation data. Due to the nature of particle searches, this 3-way partitioning is unnecessary (see footnote 9 of [159]). Machine learning is only performed on the simulated data, not the measured data, and there are two forms of validation: validating the learning algorithm, and validating the modeling accuracy of the simulated data. Simulated data used in training must still be excluded from statistical inference, but simulated data used for training validation can be reused in statistical inference after successful modeling validation is demonstrated.

Three typical data partitioning schemes for machine learning in particle searches are shown in tables 5.3(a)–(c). Table 5.3(a) shows the simplest scheme which divides each of the simulated data into one partition used for training, and a second partition used for 1) validation of the training, 2) validation of the simulations, and 3) statistical inference. This is acceptable when the second partition contains a sufficient amount of simulated data to minimize statistical fluctuations. However, for some searches, the loss of some of the simulated data to training degrades the testing power. Table 5.3(b) shows how cross-evaluation overcomes data insufficiency by evenly splitting all data, simulated *and* measured, performs statistical inference on each partition separately, and then combines the results (see section 7.1.1 of [201]). Table 5.3(c) shows the partitioning scheme for two-stage learning. Data used in one stage of training is available for reuse in the validation of another training stage (although all training data must be excluded from final statistical inference). This scheme is suited to difficult signal separation tasks where combining multiple learning algorithms is needed to improve significance. Appendix B.1 describes the possibility of combining the multi-stage partitioning scheme with the cross-evaluation scheme.

	single stage
partition 1	train
partition 2	validate[1,2] & test

(a)

	test 1	test 2
partition 1	train	validate[1,2] & test
partition 2	validate[1,2] & test	train

(b)

	stage 1	stage 2
partition 1	train	validate[1]
partition 2	validate[1]	train
partition 3	validate[2] & test	validate[2] & test

(c)

Table 5.3: The machine learning data partitioning schemes used in particle searches. Validate[1] refers to validation of the learning algorithm (checking for overtraining) while validate[2] refers to validation of the modeling accuracy of simulated data compared to measured data.
(a) A simple split of simulated data (often $50\% : 50\%$).
(b) Cross-evaluation split that performs two independent statistical inference procedures. Both simulated data and measured data are evenly split.
(c) 3-way split of simulated data for two-stage learning, $X\% : X\% : 100 - 2X\%$, where X is the minimum necessary for accurate machine learning (selected to be 15% for the T_{pair} and B_{pair} signals, as well as for the backgrounds, and 25% for the T_{single} and B_{single} signals in this work).

5.5. Stage 1 Classification with Boosted Decision Trees

BDTs using random forests and bagging were selected as the learning algorithm for stage 1 classification because they consistently provided the greatest significance with the least amount of overtraining[7]. BDTs with random forests have been observed to provide superior performance for certain learning tasks [219–223]. The use of bagging in addition to boosting can reduce sensitivity to statistical noise [224]. The choice of gradient boosting was motivated by the observation that it outperformed adaptive boosting in most test cases.

Results from the optimized BDT learning algorithm for the $m(T_{pair}) = 650$ GeV/c^2 signal is often presented in this section to illustrate the typical visual and numerical analysis performed when examining the performance of the many BDTs. The hyper-parameters used for this BDT were 1) a shrinkage rate of 0.2, 2) a 50% bagging fraction at each iteration, 3) a forest of 150 trees, 4) a maximum tree depth of 2, and 5) three random input variables provided to each tree node (a randomized forest). The selections of the stage 1 BDT hyper-parameters for each signal mass were determined by conducting a grid search (discussed in section 5.5.2).

5.5.1. Selection of the Input Variables

In many statistical analyses, much effort has often been devoted to the selection of subsets of the available input variables ("feature selection") that were observed to provide either the greatest predictive ability (for machine learning) or greatest explanatory ability (for parametric statistical inference, discussed in chapter 6). In [225], Leo Breiman discussed his finding that small perturbations to the training data set may lead to different selection of input variable subsets. More specifically, he found that many different data models compete for most accurately describing the training data, each with its own unique interpretation of how the input variables predict or explain the training data, and that small changes to the training data effect which model is selected as optimal. This finding provided motivation for removing only a small number of available input variables and for the use randomized forests of decision trees (with bagging) in stage 1 learning. Essentially, each tree of the forest optimized its own selection of an input variable subset using only a portion of the training data, and aggregating the results of many trees provided stability against reliance on any one particular input variable subset [226].

Many input variables were investigated, and the best results were obtained with a spectrum of variables that contained a variety of useful information regarding the underlying physical processes. These included transverse shape variables that provide useful information for collision events in hadron colliders [227, 228]. The input variables selected for stage 1 classification are described below.

1. $\mathbf{p_T(Z)}$ — The transverse momentum of the reconstructed leptonic Z boson.
2. $\mathbf{|y(Z)|}$ — The rapidity of the reconstructed leptonic Z boson.
3. $\mathbf{\Delta R(l_1, l_2)}$ — The pseudorapidity difference between the two leptons used in the reconstruction of the leptonic Z boson.
4. $\mathbf{|\Delta\phi(l_1, l_2)|}$ — The azimuthal angle difference between the two leptons used in the

[7] In 1996, Leo Breiman referred to boosted decision trees as the best off-the-shelf classifier in the world [176].

reconstruction of the leptonic Z boson.

5. $\mathbf{\Delta R(Z, b_1)}$ — The pseudorapidity difference between the reconstructed leptonic Z boson and the leading b-tagged jet.
6. $\mathbf{|\Delta\phi(Z, T_T)|}$ — The azimuthal angle difference between the reconstructed leptonic Z boson and the transverse thrust direction ($\hat{n}_T$). Transverse thrust is defined as

$$T_T = \max_{\hat{n}_T} \frac{\sum_i |\vec{p}_{\perp i} \cdot \hat{n}_T|}{\sum_i |\vec{p}_{\perp i}|}, \tag{5.5.1}$$

where $\vec{p}_{\perp i}$ is the two dimensional momentum vector in the transverse plane of the i^{th} jet. Possible values of T_T range from $2/\pi$ for a circularly symmetric jet configuration to 1.0 for a back-to-back dijet event.

7. $\mathbf{H_T^Z(jets)}$ — $H_T(jets) = \sum_i |\vec{p}_{\perp i}|$ is the sum of the transverse momenta of the jets, which was a dominant discriminating input variable in the square cuts analysis (for the T_{pair} and B_{pair} signal types only). However, it was determined that the sum of the projection of the jets' transverse momenta onto the transverse direction of the reconstructed leptonic Z boson also provided discrimination power, and was applicable for all four signal types. $H_T^Z(jets)$ is defined by[8]

$$H_T^Z = \sum_i \frac{|\vec{p}_{\perp i} \cdot \vec{p}_{\perp i}^Z|}{p_T(Z)}. \tag{5.5.2}$$

8. $\mathbf{T_T - T_M(jets)}$ — The transverse thrust minor measures energy flow out of the plane formed by $\hat{n}_T$ and $\hat{z}$, defined as

$$T_M = \frac{\sum_i |\vec{p}_{\perp i} \times \hat{n}_T|}{\sum_i |\vec{p}_{\perp i}|}. \tag{5.5.3}$$

T_M ranges in value from 0 for a back-to-back dijet event to $2/\pi$ for a circularly symmetric jet configurations. T_T and T_M are highly correlated, and it was observed that $T_T - T_M(jets)$ provided the most useful shape description ranging from 0 for circularly symmetric jet configurations to 1 for back-to-back dijet events.

9. $\mathbf{S_T(jets)}$ — Like transverse thrust, the transverse sphericity also measures the symmetry of jet momenta in the transverse plane, ranging from 0 for back-to-back dijet events to 1.0 for circularly symmetric jet configurations. It is defined as

$$S_T = \frac{2\lambda_2^{xy}}{\lambda_1^{xy} + \lambda_2^{xy}}, \tag{5.5.4}$$

where $\lambda_1^{xy} > \lambda_2^{xy}$ are the eigenvalues of the transverse momentum tensor,

$$S^{xy} = \sum_i \frac{1}{|\vec{p}_{\perp i}|^2} \begin{pmatrix} p_{xi}^2 & p_{xi}p_{yi} \\ p_{xi}p_{yi} & p_{yi}^2 \end{pmatrix}. \tag{5.5.5}$$

10. $\mathbf{p_T(b_1)}$ — The transverse momentum of the b-tagged jet with highest transverse momentum, "leading".
11. $\mathbf{\Delta R(b_1, b_2)}$ — The pseudorapidity difference between the two leading b-tagged jets.

[8] I am grateful to Professor Erich Varnes for helping to design this variable in July 2014.

12. $\mathbf{p_T(a_1)}$ — The transverse momentum of the leading jet that failed the b-tagging criteria ("anti-b-tagged").
13. $\mathbf{\Delta R(b_1, a_1)}$ — The pseudorapidity difference between the leading b-tagged jet and leading anti-b-tagged jet.
14. $\mathbf{\Delta R(b_2, a_1)}$ — The pseudorapidity difference between the second leading b-tagged jet and leading anti-b-tagged jet.

A few additional input variables were held back for later stage machine learning. These were

- $\mathbf{H_T(jets)} = \sum_i p_T\left(i^{th}\ jet\right)$.
- $\mathbf{H_T(jets + Z)} = p_T(Z) + H_T(jets)$.
- $\mathbf{N(forward\ jets) \geq 1}$ — The presence of a forward jet, true or false.
- $\mathbf{m(Zb)}$ — The mass of the reconstructed leptonic Z boson plus leading b-tagged jet.

Table 5.4 describes the types of information contained within each input variable. Variables may be grouped according to their complexity, being derived from a single reconstructed object to many objects. They may also be grouped according to the type of physics they represent: leptonic, hadronic or both, which corresponds to the type of reconstruction performed. And finally, they may be grouped by the type of measurement made to obtain them: energy (and momentum) based, position based or both, where energy measurements will usually contain more measurement error than location measurements. Optimization of variable selection for machine learning is difficult. Categorizing variables along these three information dimensions ensured that all information types were represented in the training. (A similar categorization of a large number of BDT input variables was performed in table 3 of [197], and for a Bayesian neural network in section 7.3 of [185].)

	increasing complexity →					
	single object		**paired objects**		**multi-object**	
leptonic physics	{E}	$p_T(Z)$	{L}	$\Delta R\,(l_1, l_2)$		
	{L}	$\lvert y(Z)\rvert$	{L}	$\lvert\Delta\phi\,(l_1, l_2)\,\rvert$		
hadronic physics	{E}	$p_T\,(b_1)$	{L}	$\Delta R\,(b_1, b_2)$	{E}	$H_T(jets)^{\dagger}$
	{E}	$p_T\,(a_1)$	{L}	$\Delta R\,(b_1, a_1)$	{E,L}	$T_T - T_T^M(jets)$
	{L}	$has\ fwd\ jet^{\dagger}$	{L}	$\Delta R\,(b_2, a_1)$	{E,L}	$S_T(jets)$
combined leptonic & hadronic physics			{L}	$\Delta R\,(Z, b_1)$	{E}	$H_T(jets + Z)^{\dagger}$
			{E,L}	$m\,(Zb)^{\dagger\dagger}$	{L}	$\lvert\Delta\phi\,(Z, T_T)\,\rvert$
					{E,L}	$H_T^Z(jets)$

{E} – Energy based measurement. {L} – Location based measurement. {E,L} – Both.
† *Variables used in second stage classification.* †† *Variables used in third stage regression.*

Table 5.4: Input variables categorized along three information dimensions: physics category (vertical), variable complexity (horizontal), and measurement category (colorized and labeled).

The choice of input variables was limited by the data preselection performed before training. For example, in order to use $\Delta R(b_2, a_1)$, only events with at least two b-tagged jets and one anti-b-tagged jet were used in the training. This variable provided strong discrimination power because it contained useful information regarding the underlying physics processes. However, its inclusion increased data insufficiency since the necessary preselection removed a significant amount of data before training. It was possible to use include a larger data set for training in which this variable was set to some default value for events where it was not defined (e.g., -1), or to an alternative definition (e.g., allowing the anti-b-tagged jet with properties closest to a b-tagged jet to proxy as a second b-tagged jet). However, such attempts to enlarge the amount of data available for training weakened the discriminating power of the BDTs.

Many combinations of input variables and preselections were examined such as relaxing b-tagging preselection requirements, and examining variables that did not contain b-tagging information, such as $\Delta R(j_1, j_2)$, the pseudorapidity difference between the first and second leading jets. It was observed that the "topology" of b-tagged and anti-b-tagged jets provided such strong discrimination power as to warrant the more restrictive preselection requirements. Based on this observation, the preselection requirement $p_T(Z) > 150$ GeV/c was relaxed to $p_T(Z) > 100$ GeV/c for stage 1 classification. Even though these additional events are excluded from the final signal region due to their use in determining some systematic uncertainties, they still contained useful information regarding b-tagging topology that was extracted during training. The preselection implemented for stage 1 classification is shown in table 5.5.

preselection requirements	description
1 preliminary requirements	Passes triggers, object selection, etc. See section 3.4.6.
2 $Z = l_1 + l_2$	Same flavor opposite sign electrons or muons with 81.19 GeV/c^2 $< m(l_1 l_2) <$ 91.19 GeV/c^2.
3 $N(leptons) = 2$	Exactly two leptons (the leptonic Z boson candidate).
4 $p_T(Z) > 100$ GeV/c	Excluding low transverse momentum Z bosons that are used in correcting the b-tagged jet multiplicity.
5 $N(b\text{-}tagged\ jets) \geq 2$	Require at least two b-tagged jets.
6 $N(anti\text{-}b\text{-}tagged\ jets) \geq 1$	Require at least one anti-b-tagged jet.

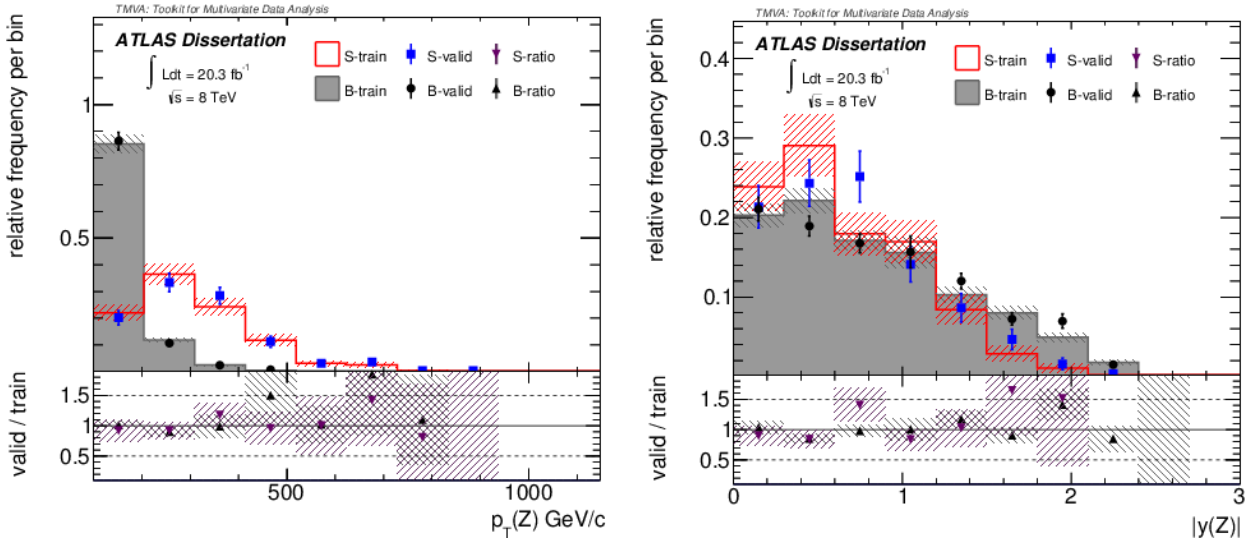
Table 5.5: Stage 1 classification data preselection.

Figures 5.5(a)–(n) shows the histograms of the input variables for the background and the 650 GeV/c^2 T_{pair} signal mass for both the training and validation partitions. The difference between the signal and background is pronounced for some variables while less so for others. It would at first appear that more than half of the variables seemingly offer little discriminating power. However, these marginal distributions are misleading as all these

variables were found to have some discriminating power when the BDTs combined them while exploring the larger multidimensional space. For example, the optimal BDT obtained for the $m(T_{pair}) = 650$ GeV/c^2 signal most often coupled the jet shape variable $T_T - T_M(jets)$ with the b-tagging variables $p_T(b_1)$ and $\Delta R(b_1, a_1)$. Thus, the signal and background histograms for $T_T - T_M(jets)$ ultimately did not indicate the usefulness of this variable.

For some of these input variables, there is some noticeable disagreement between training and validation distributions. This was expected due to the amount of data insufficiency, especially for variables inherently containing larger measurement error such as the jet shape variables. These differences have the potential to cause overtraining if the BDTs learn too much from this "noise". A robust method for checking for overtraining was developed and is described in section 5.5.2.2.

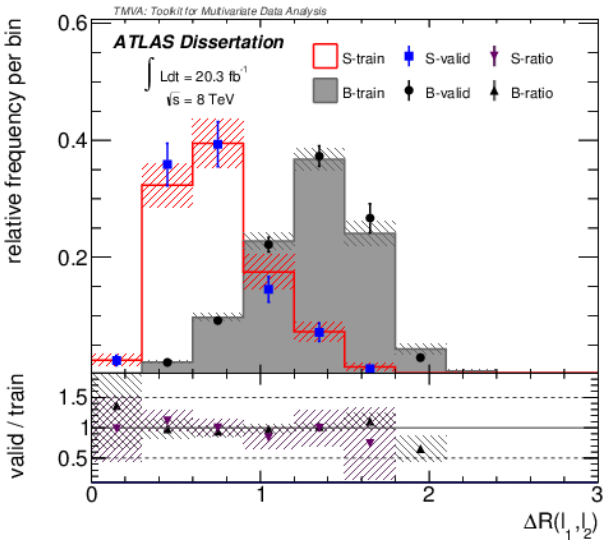
The 14 previously listed variables were used to train BDTs for each of the 59 signals. Variable selection was based on separation performance in aggregate. In the case that a variable was not useful for a particular signal, the BDT automatically ignored it because the BDT learning algorithm chose the variable to use at each node based on its ability to separate signal and background. The use of randomization decreased the BDT's ability to ignore less discriminating variables to some extent, so it was still important that most of the variables were useful for signal separation for most of the signals. The frequency which each variable was ultimately used is discussed in section 5.5.2.4.



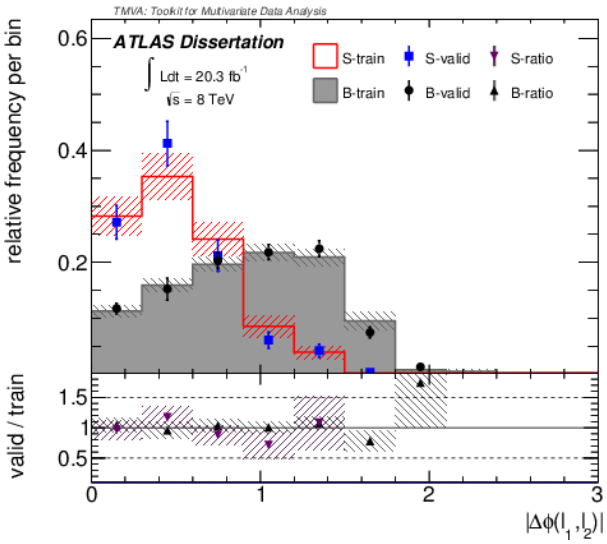
(a) $p_T(Z)$ (b) $|y(Z)|$

Figure 5.5: Input variable histograms for the $m(T_{pair}) = 650$ GeV/c^2 signal. Signal training data is shown with red lines, training background with gray bars, validation signal with blue squares, background validation with black circles. The lower "valid/train" plot shows the ratio of the validation to training distributions, violet downward triangles for signal and black upward triangles for background.

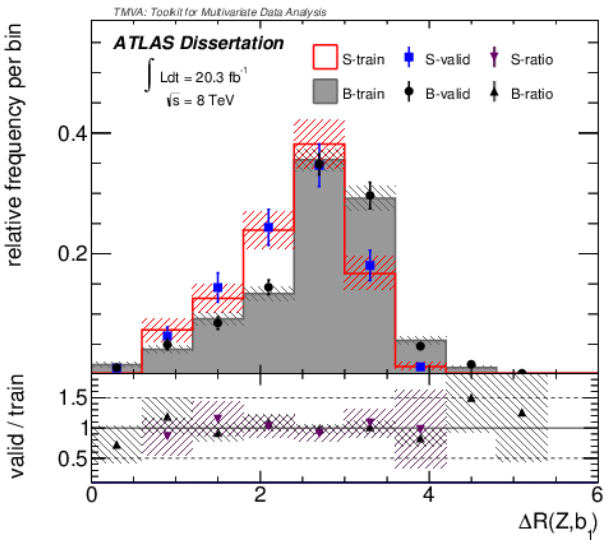
Figure 5.5 Continues on next 3 pages.



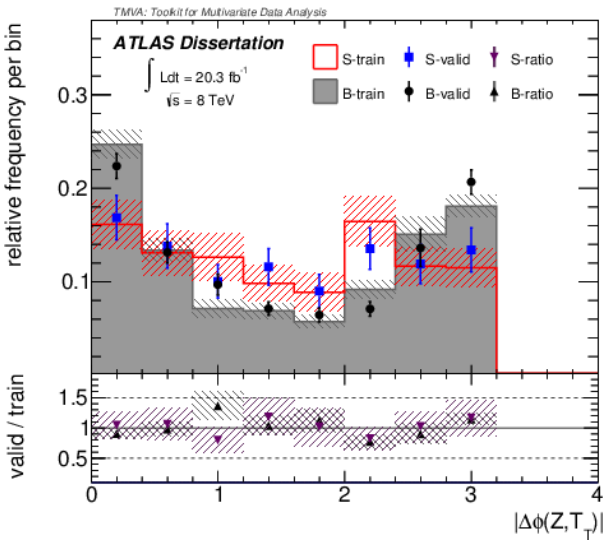
(c) $\Delta R(l_1, l_2)$



(d) $|\Delta\phi(l_1, l_2)|$

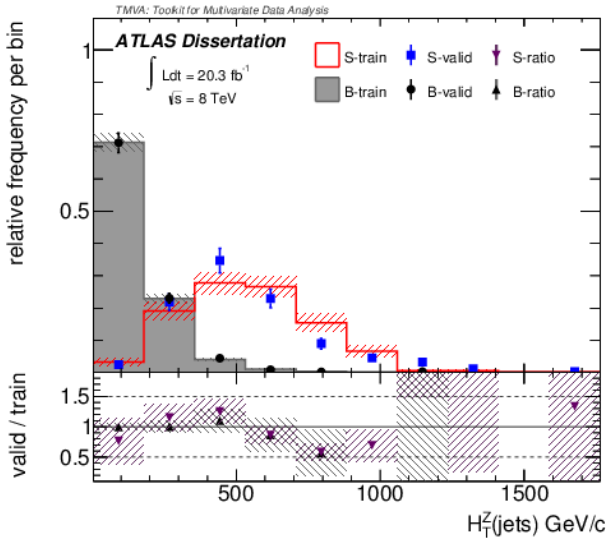


(e) $\Delta R(Z, b_1)$

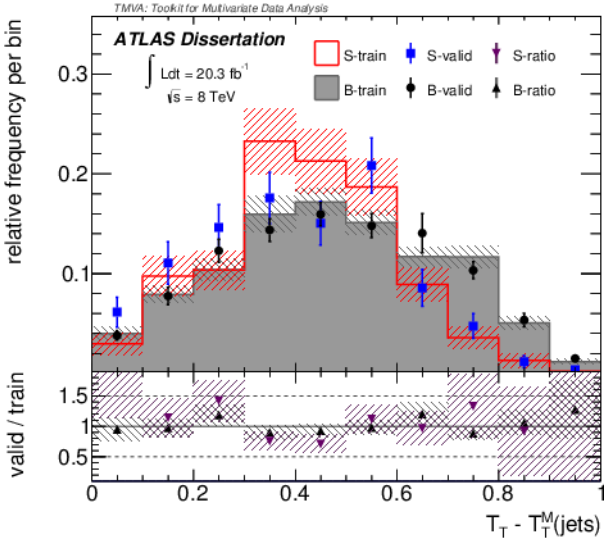


(f) $|\phi(Z, T_T)|$

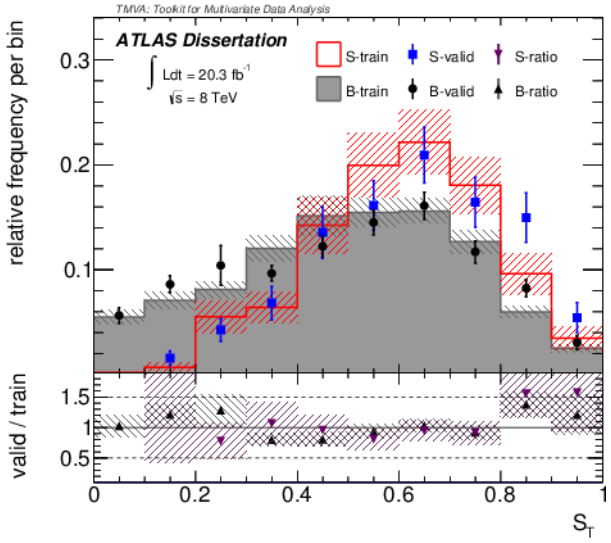
Figure 5.5: – *Continued, continues on next 2 pages.*



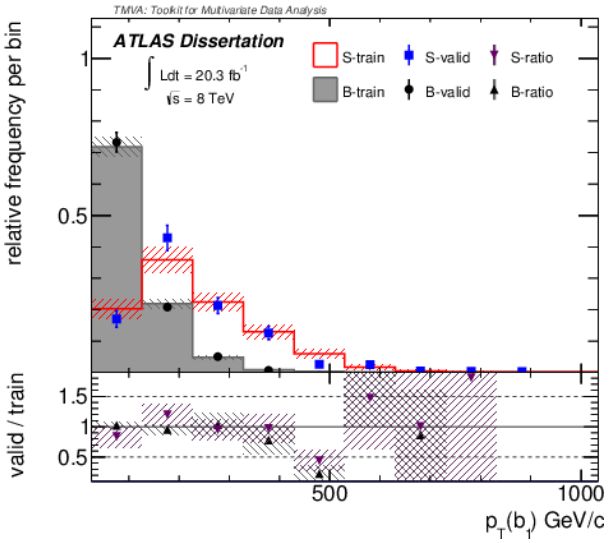
(g) $H_T^Z(jets)$



(h) $T_T - T_M(jets)$

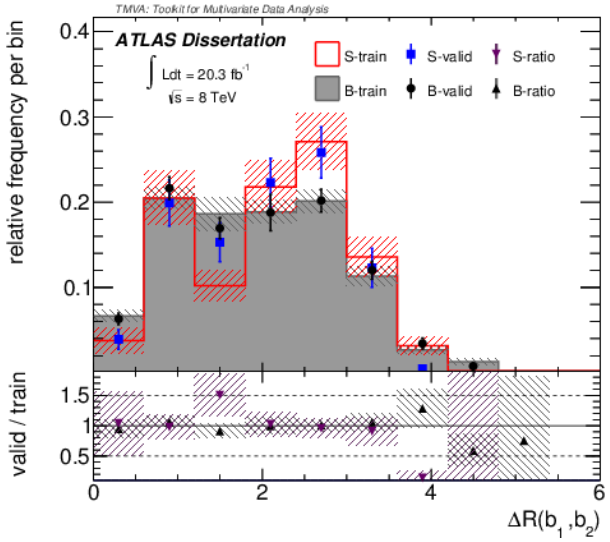


(i) $S_T(jets)$

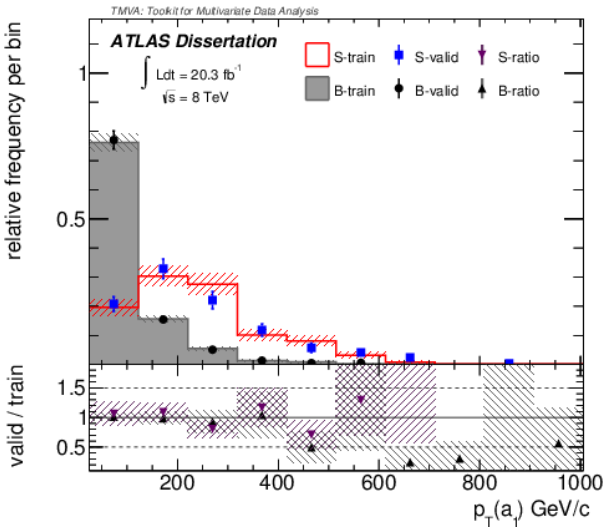


(j) $p_T(b_1)$

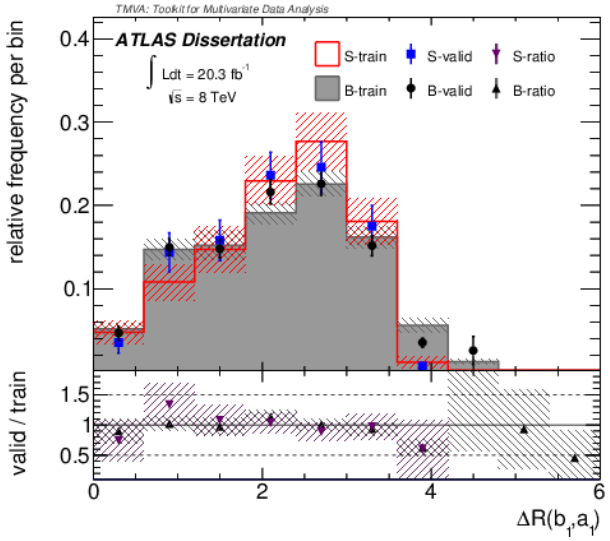
Figure 5.5: – *Continued, continues on next page.*



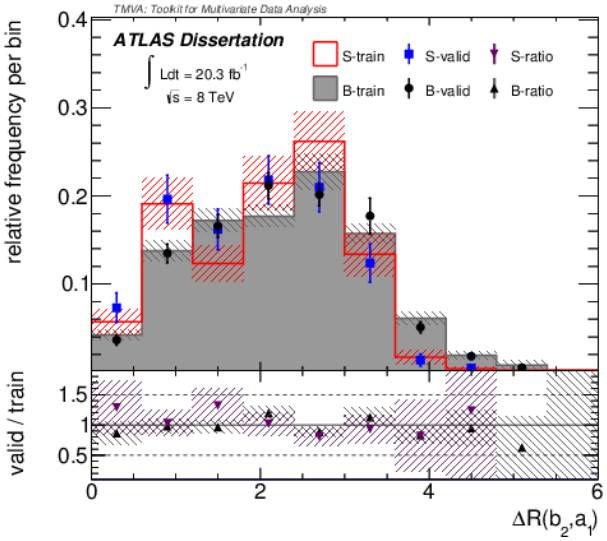
(k) $\Delta R(b_1, b_2)$



(l) $p_T(a_1)$



(m) $\Delta R(b_1, a_1)$

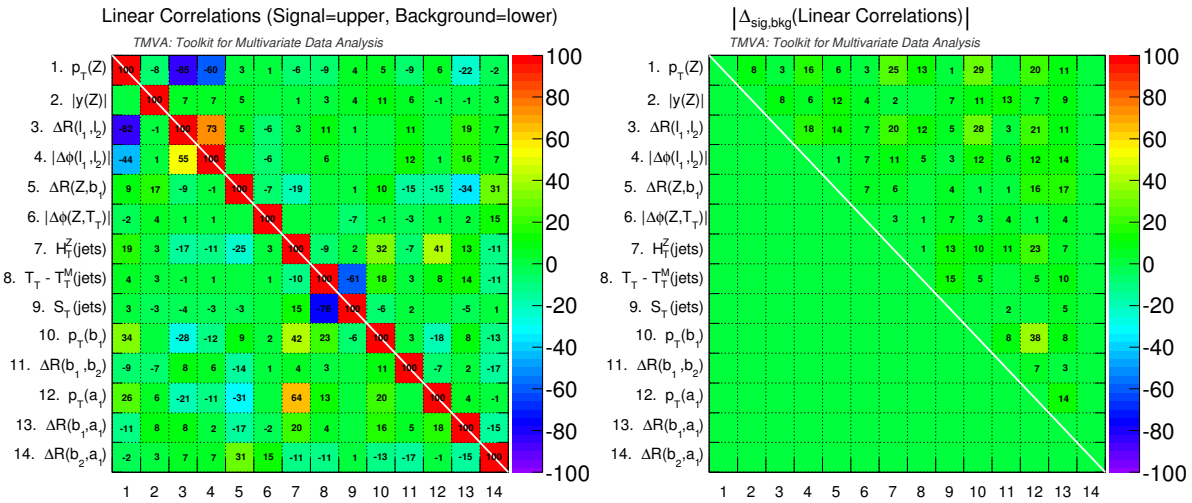


(n) $\Delta R(b_2, a_1)$

Figure 5.5: – *Continued, last page.*

Figure 5.6(a) shows the correlations of the input variables for the $m(T_{pair}) = 650$ GeV/c^2 signal and background training data. Only a few of the variable pairs have high correlation, such as the lepton position variables $\Delta R(l_1, l_2)$ and $|\Delta\phi(l_1, l_2)|$. High correlations may seem to imply a redundancy of useful information. However, since the BDTs separated signal from background using multidimensional, nonlinear relationships, this was not necessarily true. For example, $\Delta R(l_1, l_2)$ and $|\Delta\phi(l_1, l_2)|$ contained similar information, especially since $|\Delta\phi(l_1, l_2)|$ is a component in the computation of $\Delta R(l_1, l_2)$, $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$. Nevertheless, attempts to remove one or the other from training resulted in weaker separation power. It was also observed that replacing $\Delta R(l_1, l_2)$ with $|\Delta\eta(l_1, l_2)|$ gave slightly worse performance.

One indication that $\Delta R(l_1, l_2)$ and $|\Delta\phi(l_1, l_2)|$ are both useful is that their correlation was different between the signal data and background data, 0.73 for signal and 0.55 for background. Variables with noticeably different correlations between signal and background could not be removed because this indicated differing correlation structure which the BDTs would be able to utilize. Figure 5.6(b) shows the differences of the correlations between signal and background variables for the $m(T_{pair}) = 650$ GeV/c^2 signal. Matrices of correlation differences were often examined during input variable selection.



(a) (b)

Figure 5.6: (a) Correlations between input variables for the $m(T_{pair}) = 650$ GeV/c^2 signal. Signal training data is shown in the upper triangular matrix and background training in the lower. (b) The magnitude of the difference of the signal and background correlations.

5.5.2. Optimization of stage 1 Boosted Decision Trees

> *"Actually a person does not* really *understand something until he can teach it to a* computer, *i.e., express it as an algorithm."*
>
> – Donald Knuth [229]

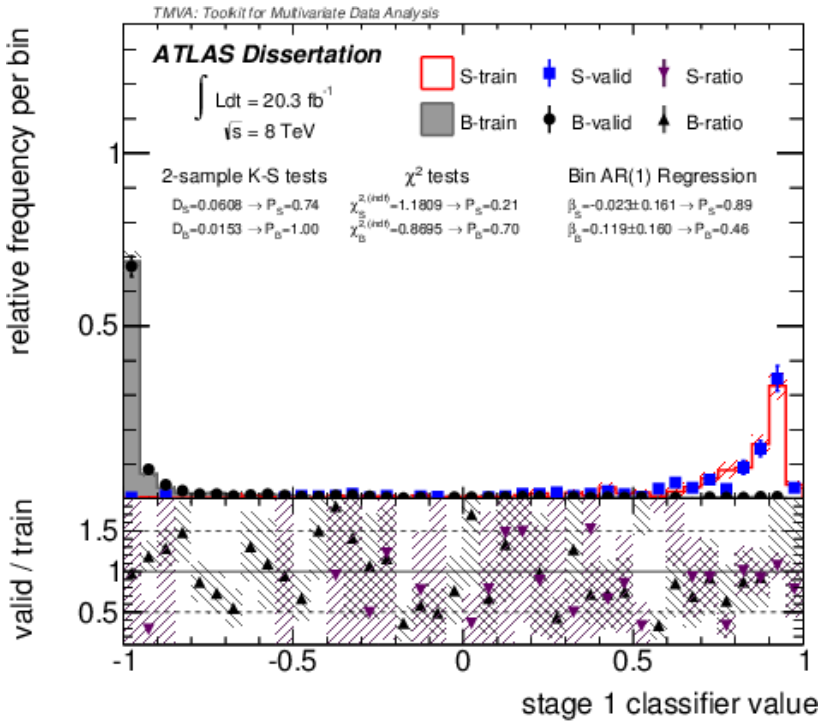
When a single signal mass is being investigated, it is reasonable for the analyst to tune learning algorithm hyper-parameters by hand. This was not feasible with 59 unique signal masses requiring 59 separate BDTs, thus some level of automation was required. This meant that a machine had to be taught how to view the results of a learning algorithm as the analyst does, translating the various visualizations into quantifiable measures: what to maximize and how to check for overtraining.

The hyper-parameter optimization to find the best 59 BDTs was done via grid search described in section 5.5.2.3 in a manner similar to that performed in [230]. A grid search for hyper-parameter optimization offered the advantage of simplicity, reliability and likely superior results [231]. The specific objective of the search for VLQs required high signal significance and low overtraining. These values were quantified as described next in sections 5.5.2.1–5.5.2.2.

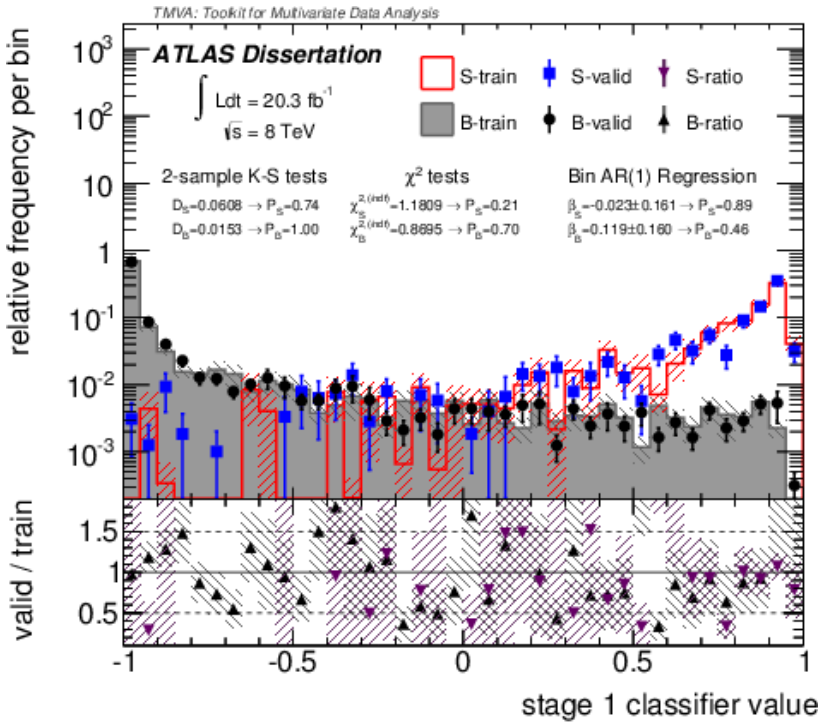
Figures 5.7(a) and (b) show histograms of the signal and background classifier response values of the optimized BDT for the $m(T_{pair}) = 650$ GeV/c^2 signal that was obtained from the grid search. Visual inspection shows that there is good signal-background separation, and no clear indication of overtraining. Some statistical tests also do not indicate the presence of overtraining (p-values > 0.05, described in detail in section 5.5.2.2).

Instead of examining these histograms, it was often useful to compare the cumulative distributions of the predicted yields (reversed) as shown in figure 5.8(a). This allowed quick estimation of a classifier value which would achieve a desired level of signal significance. For instance, if $N_S \geq N_B$ was the objective (a significance of $10/\sqrt{20} = 2.2$), data with classifier values below $\sim$0.85 could be discarded.

Figure 5.8(b) shows the receiver operating characteristic (ROC curve) for this BDT. The ROC curve provides a trajectory of coordinates $(\epsilon_S(v), r_B(v))$ where ϵ_S is the signal efficiency (acceptance rate) and r_B is the background rejection rate. These are functions of v, a cut on the classifier response value, where only data events above v are retained. As the cut value v is decreased, the background rejection rate decreases while the signal efficiency increases, i.e., traces the ROC curve from left to right. This ROC curve is constructed so that position along the x-axis of the training and validation data ROC curves corresponds to the same cut value v.

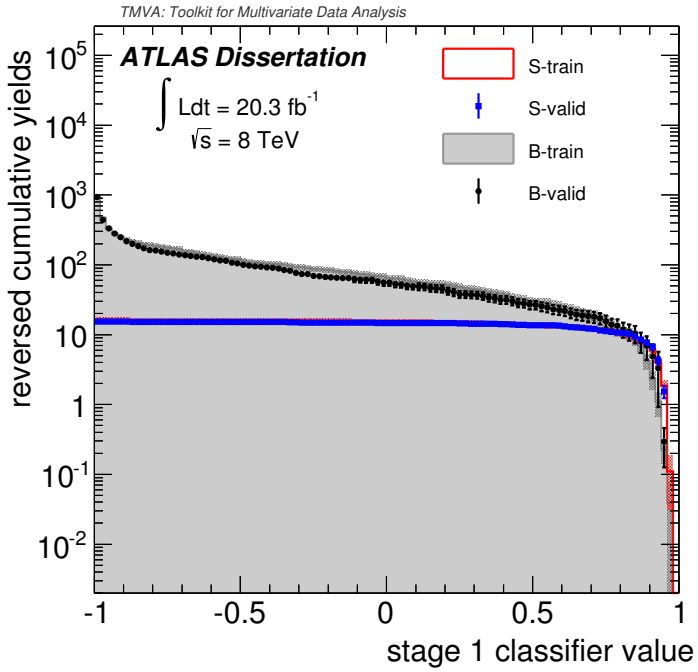


(a)

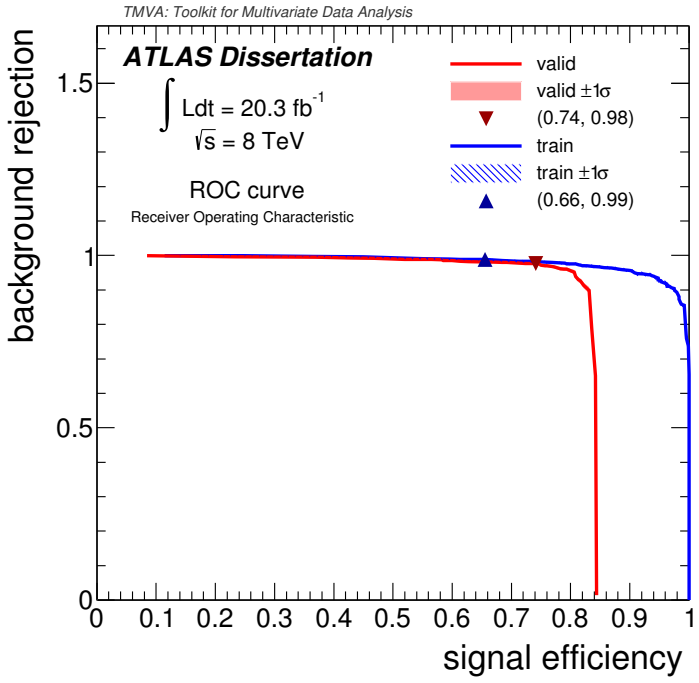


(b)

Figure 5.7: BDT classifier response value distributions for the $m(T_{pair}) = 650$ GeV/c^2 signal: (a) linear and (b) logarithmic.



(a)



(b)

Figure 5.8: The $m(T_{pair}) = 650$ GeV/c^2 signal and background: (a) cumulative yield (reversed) vs. BDT classifier response value, and (b) the receiver operating characteristic (ROC curve) where triangles indicate the positions of maximum significance.

5.5.2.1. Automated Maximization of Signal Significance

In a particle search, the measured data is generated by one of two processes that provide two competing hypotheses: the null hypothesis that the signal does not exist and that the observations are due only to Standard Model processes, or the signal hypothesis that the signal does exist and that the observations are caused by both the Standard Model and the new physics that generates the signal. Because the final statistical inference is numerically intensive (such as with the CL_S method described in chapter 6), it is often desirable to determine a simple figure of merit to estimate the expected significance that may be achieved.

The stage 1 classification results obtained in this study were not to be used directly in statistical inference, but rather as inputs to further stages of machine learning. Nevertheless, a simple variable that indicated good signal identification was desired, and this "significance" figure of merit, G, was defined as

$$G = \frac{N_S}{\sqrt{N_S + N_B}}, \tag{5.5.6}$$

where N_S and N_B were the predicted yields of the signal and backgrounds, respectively.[9] More specifically, the performance of the classifier was assigned the maximum value of G achieved by *some* square cut on its classifier value. Note that G may be viewed as the maximum signal-to-noise ratio achieved by the classifier, and returns values similar to the approximate median discovery significance (AMS) [232, 233] (a theoretical justification is discussed in [234]).

Figure 5.9 shows the significance obtained for different cuts on the classifier value of the selected BDT classifier for the $m(T_{pair}) = 650$ GeV/c^2 signal. The variance of G was estimated from Taylor expansion giving

$$\sigma_G^2 = N_S \left[\frac{(N_S/2 + N_B)^2}{(N_S + N_B)^3} \right] + N_B \left[\frac{(N_S/4)^2}{(N_S + N_B)^3} \right], \tag{5.5.8}$$

and in figure 5.9 shows good agreement between training and validation data. It was often observed that maximum G was achieved with very high σ_G. Therefore, the performance of each BDT was conservatively evaluated using the maximum of $G - \sigma_G$.

[9] Consider binary classification of "signal" and "background" where one seeks to determine the probability of *the existence* of the signal class. The measured data may be produced by the background-only process modeled by a Poisson distribution, $d_B \sim \mathcal{P}(N_B)$ with N_B the expected yield of background data, or by the signal-plus-background process, $d_{BS} \sim \mathcal{P}(N_B + N_S)$ with N_S the expected yield of signal data. A classifier was deemed useful if it was able to differentiate between these two distributions: random samples taken from each distribution were likely to be significantly different. Making Gaussian approximations for the Poisson distributions, one obtains the approximate Gaussian distribution for $\Delta_d = d_{BS} - d_B$ as

$$\Delta_d \sim \mathcal{N}\left([N_B + N_S] - [N_B], [N_S + N_B] + [N_B] - \mathrm{cov}[d_B, d_{BS}]\right) = \mathcal{N}(N_S, N_S + N_B). \tag{5.5.7}$$

Letting $Z = (N_S - \Delta_d)/\sqrt{N_S + N_B}$ so that $Z \sim \mathcal{N}(0,1)$, then $P(\Delta_D \leq 0) = P(Z \geq G)$ so that G provided an approximate Z-score (significance) with larger values indicating better differentiation between the background-only and signal-plus-background processes. These arguments assume that N_S and N_B are known with certainty, necessary in order to avoid numerically expensive computation. A more formal likelihood argument provides the similar approximate median significance (AMS) of [232].

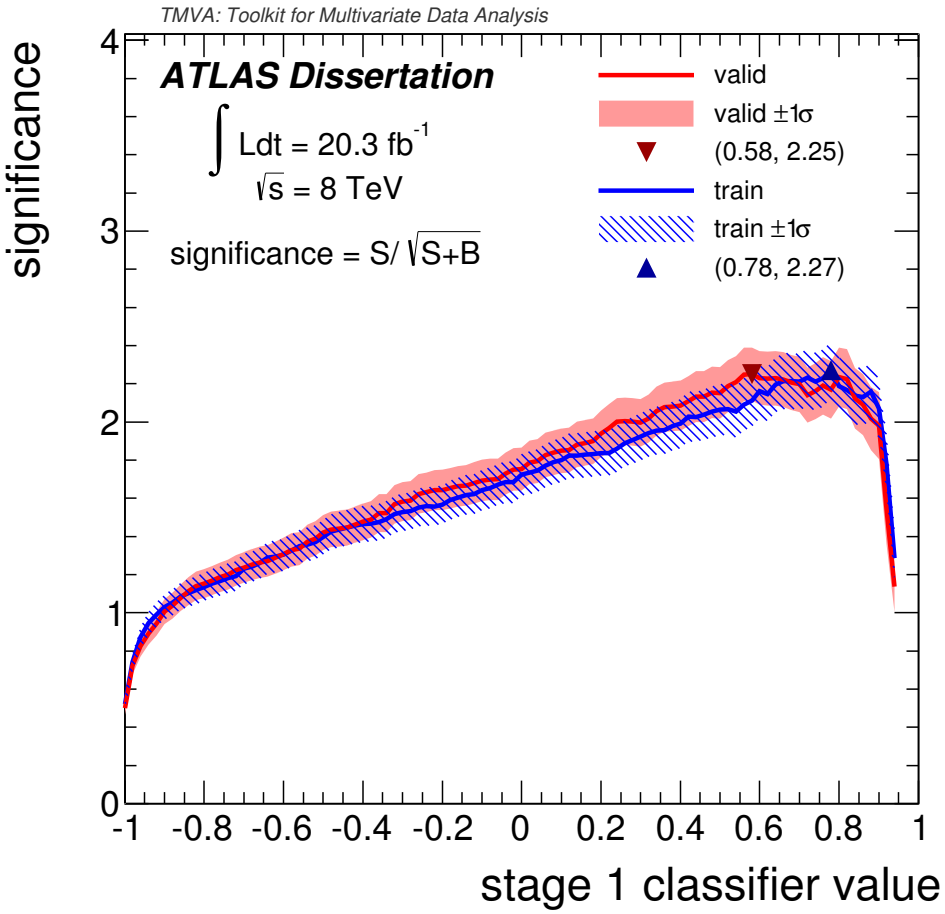


Figure 5.9: Significance vs. cut on classifier value for the $m(T_{pair}) = 650$ GeV/c^2 signal.

5.5.2.2. Automated Determination of Overtraining

An overtrained learning algorithm does not *per se* lead to invalid statistical inference results in a particle search. Results are only degraded when a bias is introduced into the simulated data which causes it to poorly model the measured data. However, it was observed that overtraining usually lead to this bias, while little or no overtraining did not. (Inspection of simulated data for bias was always conducted after each stage of ML. This should not be confused with the inspection for overtraining during machine learning.) The check for overtraining was more important than maximization of significance because a weaker classifier that lead to a valid statistical inference result was always preferable to a stronger classifier that lead to invalid results due to bias.

Four statistical tests were used to determine the presence of overtraining by comparing training and validation distributions: 1) the two-sample Kolmogorov-Smirnov (K-S) test, 2) the two-sample Anderson-Darling (A-D) test, 3) a modified two-sample chi-squared (χ^2) test [235], and 4) an additional *ad hoc* test to detect any "shape shifting" errors (explained in detail below).

These four tests were performed independently on the two signal distributions (training data versus validation data) and the background distributions yielding eight separate p-values. A low p-value for a given test indicated a low probability that the observed disagreement between two distributions occurred by chance (i.e., a low p-value for any of the four tests was indicative of possible overtraining). In order to facilitate the automated check for overtraining, a single p-value combining all eight p-values was desired. Because the four p-values for the tests of the signal (or else background) distributions were highly correlated, they were combined into a single p-value using Brown's method that accounts for correlated p-values (described fully below). This yielded two p-values, one for the combined four tests of the signal distributions and one for the background. The signal and background distributions were uncorrelated so that these two p-values were then combined into a final p-value using Fisher's method, the Brown-Fisher p-value. This scheme is depicted in figure 5.10.

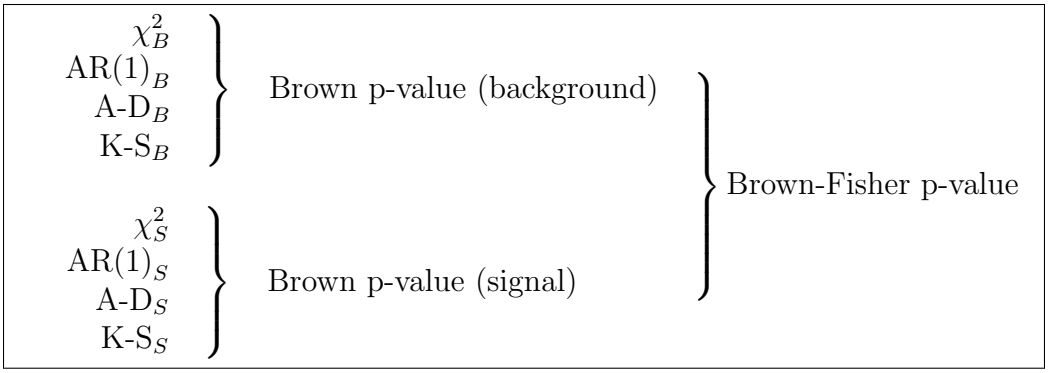


Figure 5.10: The strategy for implementing Brown's method and Fisher's method for combining the resultant p-values of the eight correlated statistical tests into one final p-value that was used for automated detection of overtraining.

This final p-value allowed the grid search for optimal BDT hyper-parameters to be automated by providing a single test statistic that correlated with the amount of overtraining. This method was conservative since low final p-value did not necessarily imply overtraining, but experimentation showed that this usually was the case. Additionally, the selected 59 BDTs from the automated grid search were visually inspected for overtraining.

The K-S test assumes that the two sample distributions are from the same underlying distribution and examines the maximum disagreement between their cumulative distribution functions (CDFs). However, the K-S test is most sensitive to disagreement in the central region of the CDFs. Efficient BDTs typically generate highly asymmetric classifier value distributions. The A-D test also measures the disagreement between observed CDFs, but more heavily weights the disagreements near the tails of the CDFs [157, 236].

It was determined that the K-S, A-D and χ^2 tests did not always accurately detect the presence of overtraining when there was a small shift in the shape of the classifier value distribution. Such shifts occurred when the classifier values predicted for the training data were systematically higher than that of the validation data. This systematic shift infrequently occurred in a manner small enough to cause little disagreement in the CDFs and thus went undetected by the K-S and A-D tests, and also to have small residual errors in each of the bins so that this overtraining was not detected by the χ^2 test. In order to check for this "shape error", an *ad hoc* test was implemented by examining bin-by-bin correlation of the normalized residuals. Let

$$\delta_i = \frac{f_{train,i} - f_{valid,i}}{\sqrt{\sigma^2_{train,i} + \sigma^2_{valid,i}}}, \tag{5.5.9}$$

be the normalized residual of the difference in training and validation frequencies in the i^{th} bin of the classifier value histogram, with σ_i the Gaussian approximate error of $\sqrt{N_i}/N$. Then examine the autoregressive model AR(1)

$$\delta_{i+1} = \beta\delta_i + \epsilon, \tag{5.5.10}$$

where ϵ is white noise, and compute $\hat{\beta}$ by ordinary least squares. If there was a systematic shape error, then $\hat{\beta}$ would differ significantly from zero and a p-value was obtained. This AR(1) test was useful in automating the detection of systematic shape error due to overtraining. For example, the AR(1) test shown in figure 5.7 (a) provided p-values of 0.46 for background distributions and 0.89 for signal, so that no systematic shape error was detected. Examining the ratio plot in the bottom panel of figure 5.7 (a), one sees that each of the background and signal ratio distributions oscillate in a likely random manner above and below 1.0.

Each of these tests, K-S, A-D, χ^2, and AR(1), provided eight useful indications of possible overtraining (four for the background distributions, and four for signal). The next task was to combine these eight p-values into a single measure that would aid in automated overtraining detection. Fisher's combined probability test (Fisher's method) is a composite test of several independent tests that combines the k separate tests' p-values into a single test statistic, $X_F = -2\sum_{i=1}^{k} \ln(p_i)$ where $X_F \sim \chi^2_{2k}$ so that $E[X_F] = 2k$ and $var[X_F] = 4k$, which is then used to compute a single p-value. Fisher's method is sensitive to small p-values so that if one of the individual tests for overtraining was positive, the composite test was also likely

to test positive for overtraining [237]. This was a desirable feature since the goal was simply to identify BDTs that were likely overtrained, and each of the eight tests was sensitive to overtraining in different ways.

However, Fisher's method could not be used directly because these eight tests were correlated having been applied on the same data. Brown's method extends the Fisher test to approximately accommodate these correlations [238]. Brown's method is to assume that $X_B = -2\sum_{i=1}^{k}\ln(p_i)$ is approximately $X_B/c \sim \chi^2_{2k^*}$ with scale factor c and degrees of freedom k^* determined by [239]

$$c = \frac{var[X_B]}{4k}, \text{ and } \; k^* = \frac{4k^2}{var[X_B]}, \tag{5.5.11}$$

where

$$var[X_B] = 4k + 2\sum_{i=1}^{k}\sum_{j=i+1}^{k} cov[-2\ln(p_i), -2\ln(p_j)]. \tag{5.5.12}$$

Therefore, if the covariance between the individual tests is known, a single p-value may be computed from the $\chi^2_{2k^*}$ distribution.

In order to apply Brown's method, it was necessary to estimate the covariance matrix of the eight tests. This was estimated from a large number of learning algorithm results (>20,000). Each of the four tests (χ^2, AR(1), A-D, and K-S) was applied separately to background and signal distributions, and inspection of the resultant 8×8 correlation matrix indicated similar correlations by test type, therefore the covariance matrix was re-estimated by combining the background and signal results. The 4×4 correlation matrix corresponding to the covariance matrix used to compute p-values using Brown's method is shown in table 5.6. This shows that the *ad hoc* AR(1) test was least correlated with the other three tests, providing further motivation for its use.

	χ^2	AR(1)	A-D	K-S
χ^2	**1.000**	0.498	0.649	0.647
AR(1)	0.498	**1.000**	0.454	0.434
A-D	0.649	0.454	**1.000**	0.926
K-S	0.647	0.434	0.926	**1.000**

Table 5.6: Correlation matrix of p-values for the non-independent hypothesis tests over all stage 1 classifiers.

Therefore, the four correlated tests for the background were combined into a single p-value using Brown's method, and similarly the four correlated tests for signal were combined into a second p-value. Since the background and signal distributions were independent, these two p-values were combined using Fisher's method to obtain a single p-value, labeled the "Brown-Fisher p-value" (B-F p-value). This single p-value was then used for automated inspection for overtraining.

5.5.2.3. Grid Search for Optimal BDT Hyper-parameters

For each of the 59 signals, an automated grid search for the optimal BDT was performed by examining many BDTs with different hyper-parameters, rejecting those exhibiting overtraining, and selecting the one with the highest significance. First, several rounds of searches were performed in order to identify reasonable ranges of the hyper-parameters to explore. The final search was conducted over the hyper-parameter grid defined in table 5.7.

Figures 5.11(a) and (b) show the results of the hyper-parameter grid search for optimal BDT for the $m(T_{pair}) = 650$ GeV/c^2 signal. Figure (a) shows the 384 points corresponding to each hyper-parameter combination with each point giving the maximum significance achieved vs. its computed B-F p-value. The general trend was a negative slope indicating that overtrained BDTs tended to have higher signal separation power. That is why the automated check for overtraining was necessary, as maximization of the significance alone would nearly always lead to the selection of overtrained BDTs.

A high B-F p-value indicated a low probability of overtraining. Visual inspection of many results led to the following rule of thumb for selection of the optimal BDT. The BDT with maximum $G - \sigma_G$ was selected from those with B-F p-value above 0.5. If no BDT had a B-F p-value above 0.5, then the maximum of those with B-F p-values above 0.4 were selected, and so on with additional break points of 0.35, 0.30, 0.25, and 0.20. Figure 5.11(b) shows the frequency of the B-F p-values obtained for the $m(T_{pair}) = 650$ GeV/c^2 signal. This demonstrated the availability of non-overtrained (high B-F p-value) BDTs for selection. For this signal, over 40% of the BDTs met the B-F p-value> 0.5 threshold, with the selected BDT being the one of these with highest significance (or rather, the one with $\max(G - \sigma_G)$).

For the pair produced signal types, all selected BDTs for each signal mass had B-F p-values $\geq$0.5. For T_{single}, the selected BDT with lowest B-F p-value was 0.41. For B_{single} the lowest B-F p-value was 0.41, except for the 500 GeV/c^2 signal where the B-F p-value was 0.22. All selected BDTs were visually inspected for overtraining, and no overtraining was detected

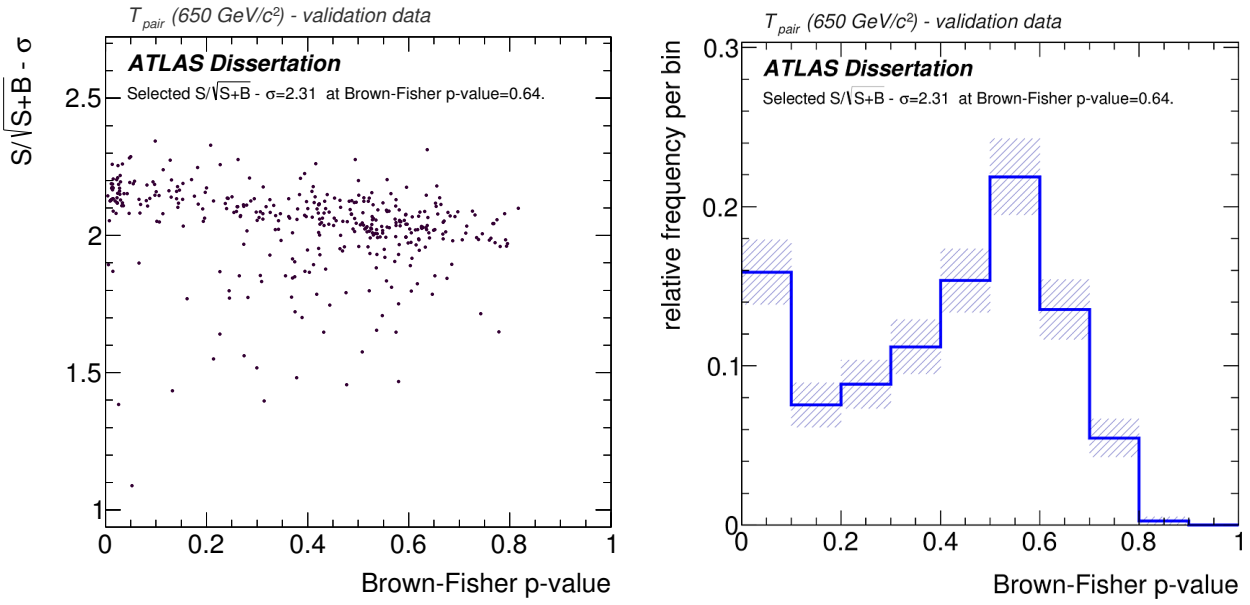
hyper-parameter	description	range of options	# of options
max depth	Maximum # of levels of node splitting allowed per weak classifier tree.	2 or 3	2
$\frac{\text{var}}{\text{node}}$	# of randomly chosen variables available per node.	1, 2 or 3	3
shrinkage	Gradient boosting learning rate.	0.03, 0.05, 0.07, 0.10, 0.15, 0.20, 0.25, 0.30	8
N_{trees}	# of weak classifier trees used in BDT forest.	10, 50, 100, 150, 250, 500, 750, 1000	8
# combinations			384

Table 5.7: The hyper-parameter options examined in the grid search for the optimal BDT. The same options are used for the 59 separate grid searches for the 59 signal masses.

in any except the $m(B_{single}) = 500$ GeV/c^2 signal. Figure 5.12 shows the histogram of BDT classifier response values for the B_{single} 500 GeV/c^2 signal, which had the lowest B-F p-value of 0.22. This B_{single} mass proved to be the most difficult for signal-background separation regardless of attempts at expanding the hyper-parameter options for the grid search. In this case, there was some discrepancy between the signal's training and validation distributions that implied decreased signal-separation ability when applied to test data. However, no overtraining was observed in the background that could have led to the introduction of bias in the simulated background modeling.

Once the optimal BDT has been selected for a signal mass, summary statistics of the BDT characteristics were examined. Figure 5.13 shows the frequencies of the input variables used in the decision nodes of the selected BDT random forest for the $m(T_{pair}) = 650$ GeV/c^2 signal. The first column indicates that this BDT relied most on the $H_T^Z(jets)$ variable for discrimination, and least on $\Delta R(b_2, a_1)$. Nevertheless, the spread in frequency is quite small, 12.1%, so that all of the variables provided information to the learning algorithm. Without the use of random variable selection at each decision node, BDTs were often seen to rely predominantly on two or three input variables and ignore the remainder. This reliance on just a few powerful discriminating variables was observed to result in overtraining more often than with random forests. Also, variable node frequency charts such as this were often helpful in determining the final selection of input variables.

Figure 5.14 shows the frequencies of the input variables used in parent-child decision node pairings for the selected BDT random forest of the $m(T_{pair}) = 650$ GeV/c^2 signal. A "pairing" of variables indicates that the second variable was used for decision node splitting in the child node of the parent node, while the parent node used the first variable for splitting. This plot provided a detailed description of the most useful pairings of variables determined by the learning algorithm, a kind of "information map" depicting the relationships discovered during the learning process. For example, half of the time that $\Delta R(Z, b_1)$ was used in a parent node, $p_T(Z)$ was chosen for use in a child node. The correlation between these two variables was nearly zero (see figure 5.6(a)) so that the relationship between these two variables that the BDT discovered was highly nonlinear. Another example is that $\Delta R(b_2, a_1)$ was never used as a parent node. Nevertheless, it was used 4.3% of the time as a final child node by more than half of the other variables. Generally, this map shows that information of different kinds was often combined during the learning process, such as b-tagging variables being combined with leptonic variables. This was in line with the goal of choosing input variables that spanned several information dimensions (described in table 5.4), and further supports the use of randomized forests in extracting these useful relationships.



(a) (b)

Figure 5.11: Examination of the Brown-Fisher p-values of the stage 1 BDT grid search for the $m(T_{pair}) = 650$ GeV/c^2 signal. (a) Significance (validation data only) vs. Brown-Fisher p-value for the 384 hyper-parameter combinations, and (b) the histogram of the Brown-Fisher p-values showing that over 40% met the 0.5 threshold.

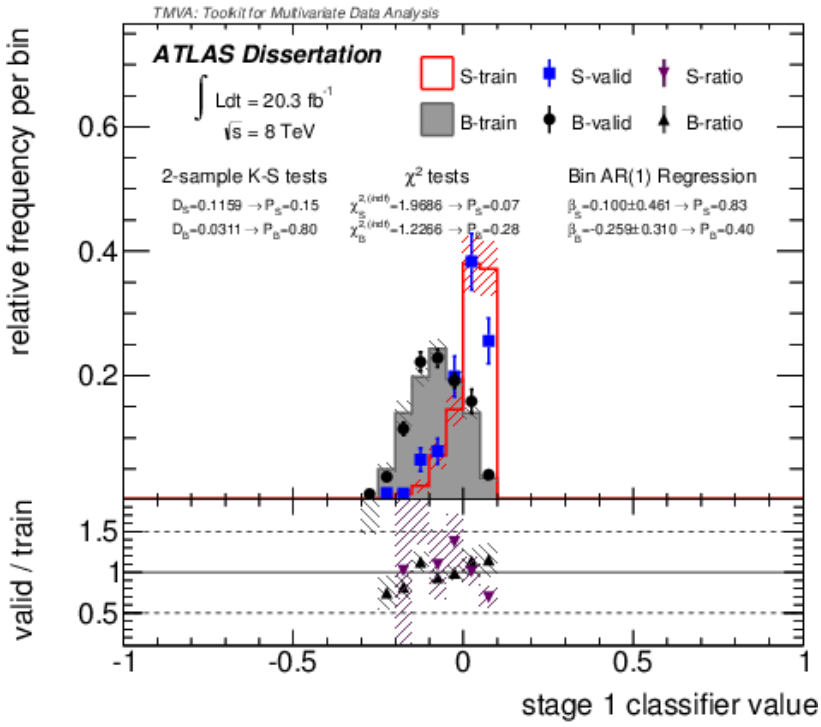


Figure 5.12: The histogram of BDT classifier response values for the most difficult signal to separate from background, $m(B_{single}) = 500$ GeV/c^2.

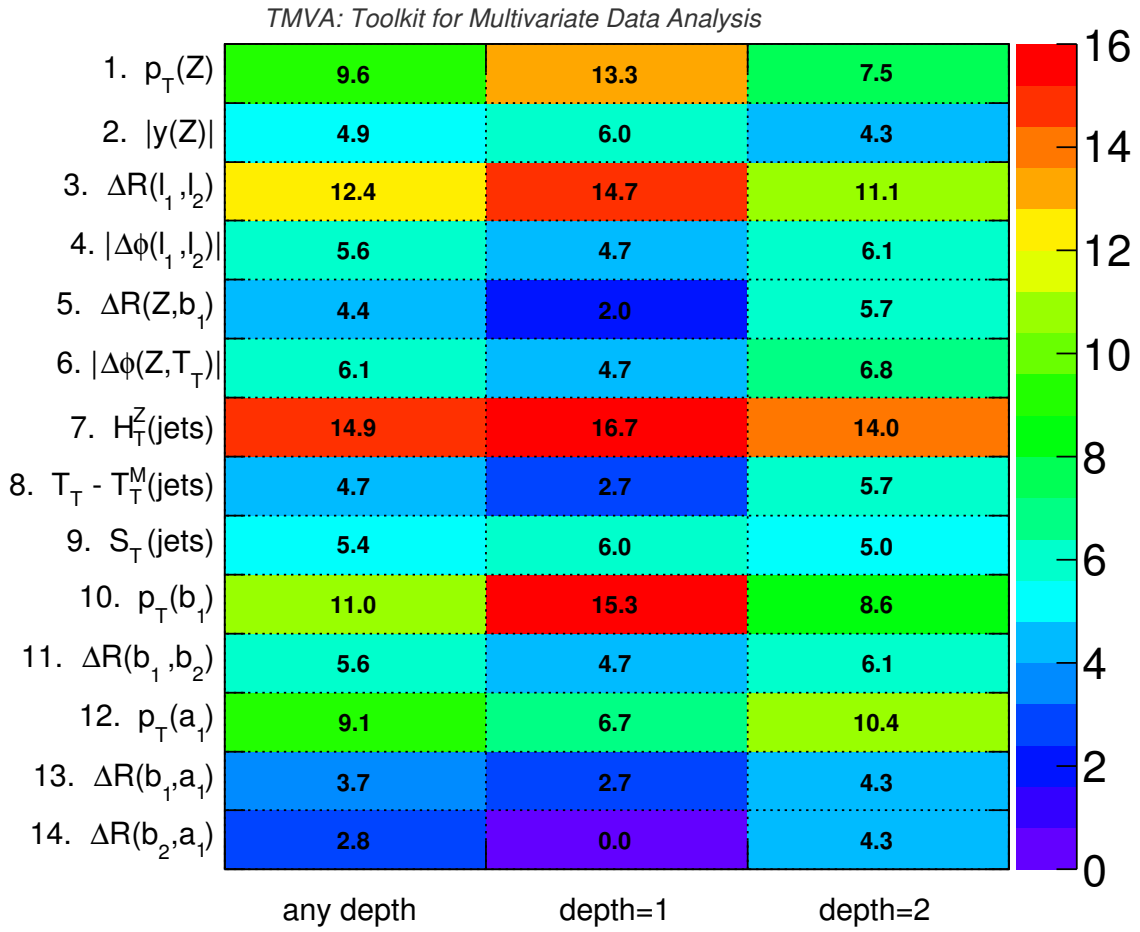


Figure 5.13: Relative frequencies (in percentages) of the input variables used for the decision nodes of the selected BDT random forest for the $m(T_{pair}) = 650$ GeV/c^2 signal.

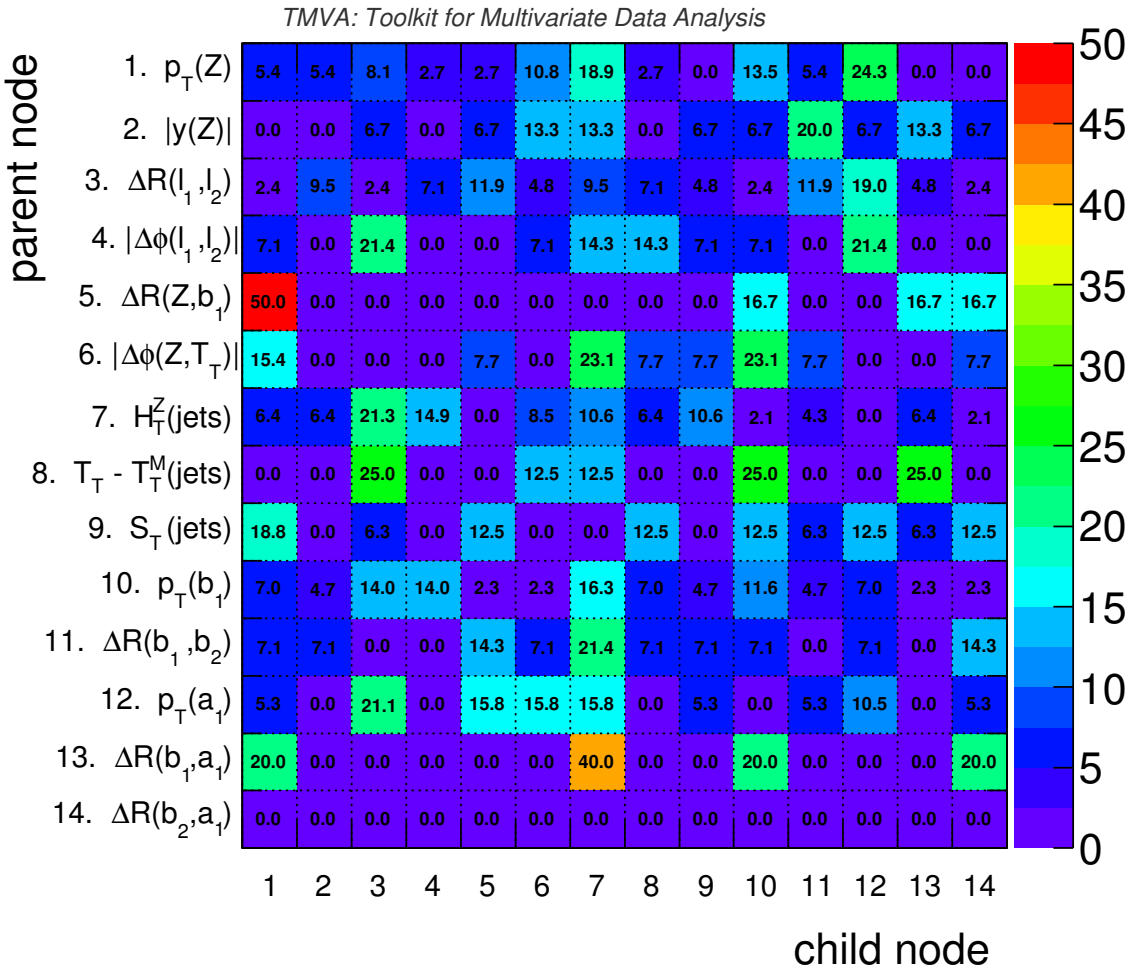


Figure 5.14: Relative frequencies (in percentages) of input variables used for parent-child decision node pairs of the selected BDT random forest for the $m(T_{pair}) = 650$ GeV/c^2 signal (rows add to 100% except where all are 0).

5.5.2.4. Selected stage 1 BDT Classifiers

Table 5.8 summarizes the results of the grid search for optimal BDT hyper-parameters. The most complex BDTs would have been obtained with a maximum depth of 3, 3 random variables per node, and 1000 trees for the random forest. Data insufficiency prevented the selection of such BDTs as these were more prone to overtraining. Only one of the 59 optimized BDTs used all 1000 trees possible. In fact, quite a few used the minimum of 10 trees. With so few trees, the separation power for these BDTs was lower than average. Because the stage 1 BDTs were inputs for later stage 2 learning (along with additional variables), it was anticipated that even weaker stage 1 BDTs would be informative when combined with the others.

Figure 5.15 shows the frequency distributions of the hyper-parameters of the selected BDTs for all signals. This figure was used often before the final search was implemented in order to determine the range of hyper-parameter options to search. For instance, using a maximum depth of 4, or using more than 3 random variables per node nearly always resulted in some amount of overtraining. It is interesting that the high level of randomization used in tree formation (1 to 3 input variables randomly provided to each node) gave superior results [240]. This provided support that the ultimate goal of information extraction from each of the variables was accomplished by forcing the BDT learning algorithm to sample from each of them via randomization. The gradient boosting learning rate (shrinkage) did not exhibit any strong trend and so required a larger range for the search.

Figure 5.16 (a)–(d) shows the maximized $G - \sigma_G$ by signal mass for the four VLQ signal types. The general trend was an exponential decay corresponding to the decreasing cross sections with increasing signal mass. Figure 5.16 (d) shows that certain B_{single} signal masses are much more difficult to separate from background than other B_{single} signal masses (note that the gaps at 450 and 1050-1150 are correct because no simulations were generated for these signal masses). This is sensible because the B_{single} signal topology has input variables most like the background.

Signal	Mass GeV/c^2	Selected Brown-Fisher p-value	$\frac{S}{\sqrt{S+B}} - \sigma$ (at p-value)	Optimal parameter settings max depth	$\frac{\text{vars}}{\text{node}}$	shrinkage	N_{trees}
T_{pair}	350	0.52	13.76	2	3	.1	10
	400	0.69	10.41	2	2	.025	50
	450	0.54	7.38	2	3	.025	50
	500	0.61	5.77	3	2	.05	50
	550	0.52	3.93	2	2	.3	150
	600	0.50	2.88	2	2	.25	10
	650	0.64	2.31	2	3	.2	150
	700	0.55	1.55	2	1	.15	100
	750	0.51	1.16	2	3	.3	50
	800	0.52	0.90	3	3	.075	150
	850	0.60	0.79	2	3	.25	50
	900	0.59	0.69	3	3	.25	10
	950	0.56	0.50	2	3	.05	500
	1000	0.52	0.41	3	1	.3	50
	1050	0.70	0.33	2	2	.025	100
	1100	0.56	0.07	2	1	.1	100
T_{singlet}	400	0.41	2.28	2	3	.3	10
	450	0.55	2.18	2	3	.05	10
	500	0.41	2.07	3	3	.05	250
	550	0.62	1.82	2	1	.15	10
	600	0.57	1.22	3	2	.025	50
	650	0.48	1.23	3	3	.025	500
	700	0.54	0.97	2	1	.1	50
	750	0.53	0.87	2	3	.3	150
	800	0.49	0.53	3	1	.3	50
	850	0.42	0.53	3	2	.025	1000
	900	0.45	0.38	3	3	.075	10
	950	0.49	0.41	2	1	.15	500
	1000	0.41	0.34	2	2	.2	500
	1050	0.45	0.28	3	2	.1	250

Table 5.8—*continues on next page*

Signal	Mass GeV/c^2	Selected Brown-Fisher p-value	$\frac{S}{\sqrt{S+B}} - \sigma$ (at p-value)	Optimal parameter settings max depth	$\frac{\text{vars}}{\text{node}}$	shrinkage	N_{trees}
B_{pair}	350	0.63	21.36	2	3	.2	10
	400	0.50	14.87	2	1	.1	750
	450	0.62	10.17	3	3	.025	500
	500	0.53	6.98	2	3	.15	500
	550	0.72	4.95	3	1	.2	100
	600	0.51	3.26	3	3	.3	10
	650	0.63	2.84	2	3	.1	500
	700	0.63	2.08	2	3	.3	250
	750	0.52	1.55	3	3	.075	500
	800	0.51	1.15	2	1	.2	100
	850	0.56	1.05	2	1	.075	100
	900	0.51	0.89	2	1	.15	50
	950	0.50	0.59	3	2	.15	100
	1000	0.61	0.48	3	2	.05	250
	1050	0.66	0.37	3	3	.15	50
	1100	0.56	0.08	3	1	.05	250
B_{singlet}	400	0.46	0.45	3	2	.15	100
	500	0.22	0.10	2	1	.15	10
	550	0.72	0.20	2	1	.075	50
	600	0.51	0.24	2	2	.05	50
	650	0.44	0.36	3	2	.075	100
	700	0.42	0.39	2	1	.15	50
	750	0.68	0.11	2	2	.025	50
	800	0.49	0.05	2	2	.3	10
	850	0.51	0.21	2	2	.075	100
	900	0.62	0.23	2	1	.3	50
	950	0.48	0.10	3	1	.25	100
	1000	0.41	0.03	3	3	.3	10
	1200	0.51	0.12	3	1	.3	150

Table 5.8: Summary of 59 stage 1 classifier grid searches for each signal mass. Brown-Fisher p-value for overtraining hypothesis tests indicates no presence of overtraining (confirmed by visual inspection). Variability among selected optimal parameter settings, especially N_{trees}, indicate that background is more difficult to separate from certain signal masses than others.

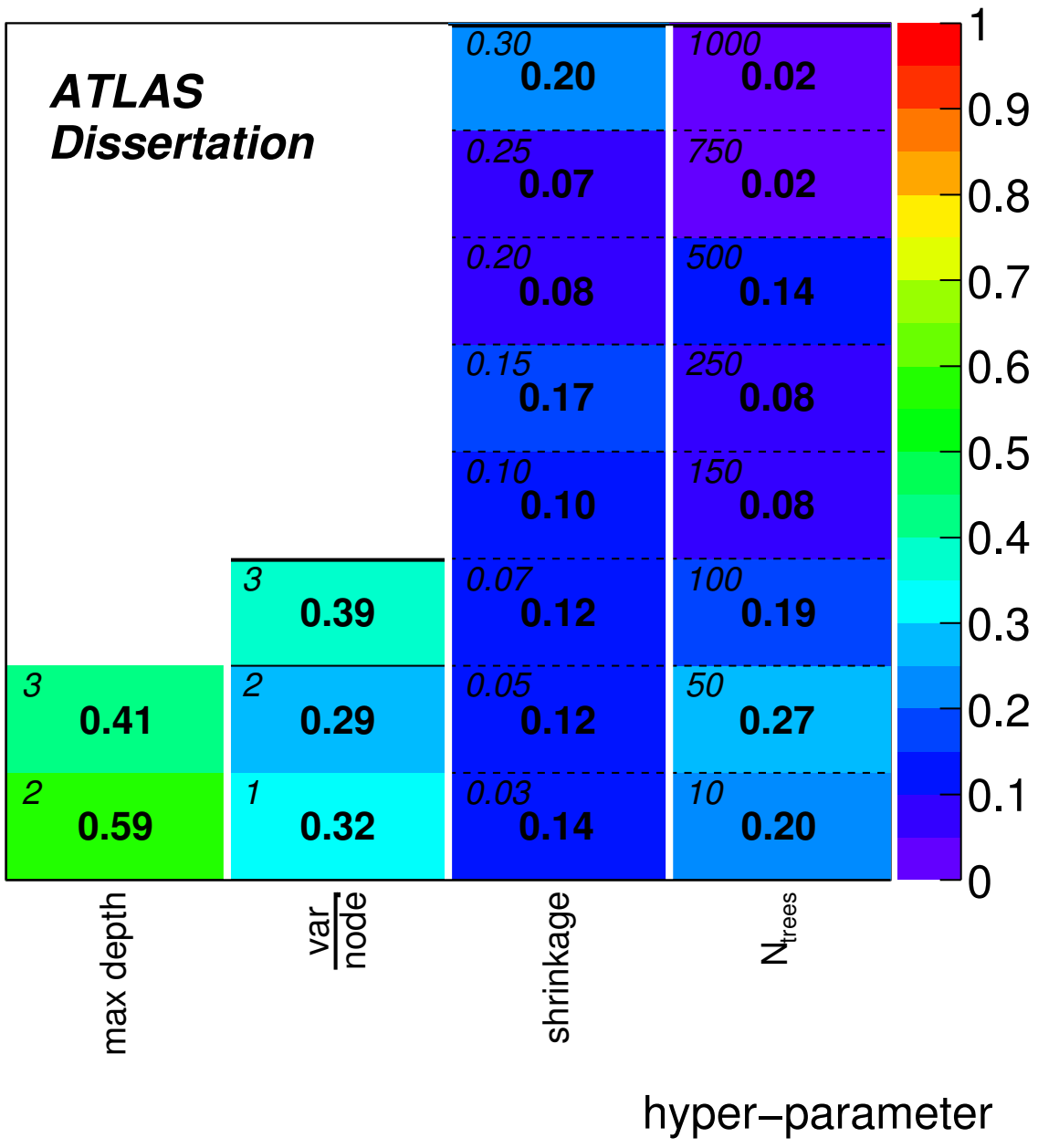


Figure 5.15: Relative frequencies of hyper-parameter options obtained in the grid search for optimal BDTs for the 59 signals. The number in the upper left of each cell is the hyper-parameter option with the frequency with which it was selected in the center.

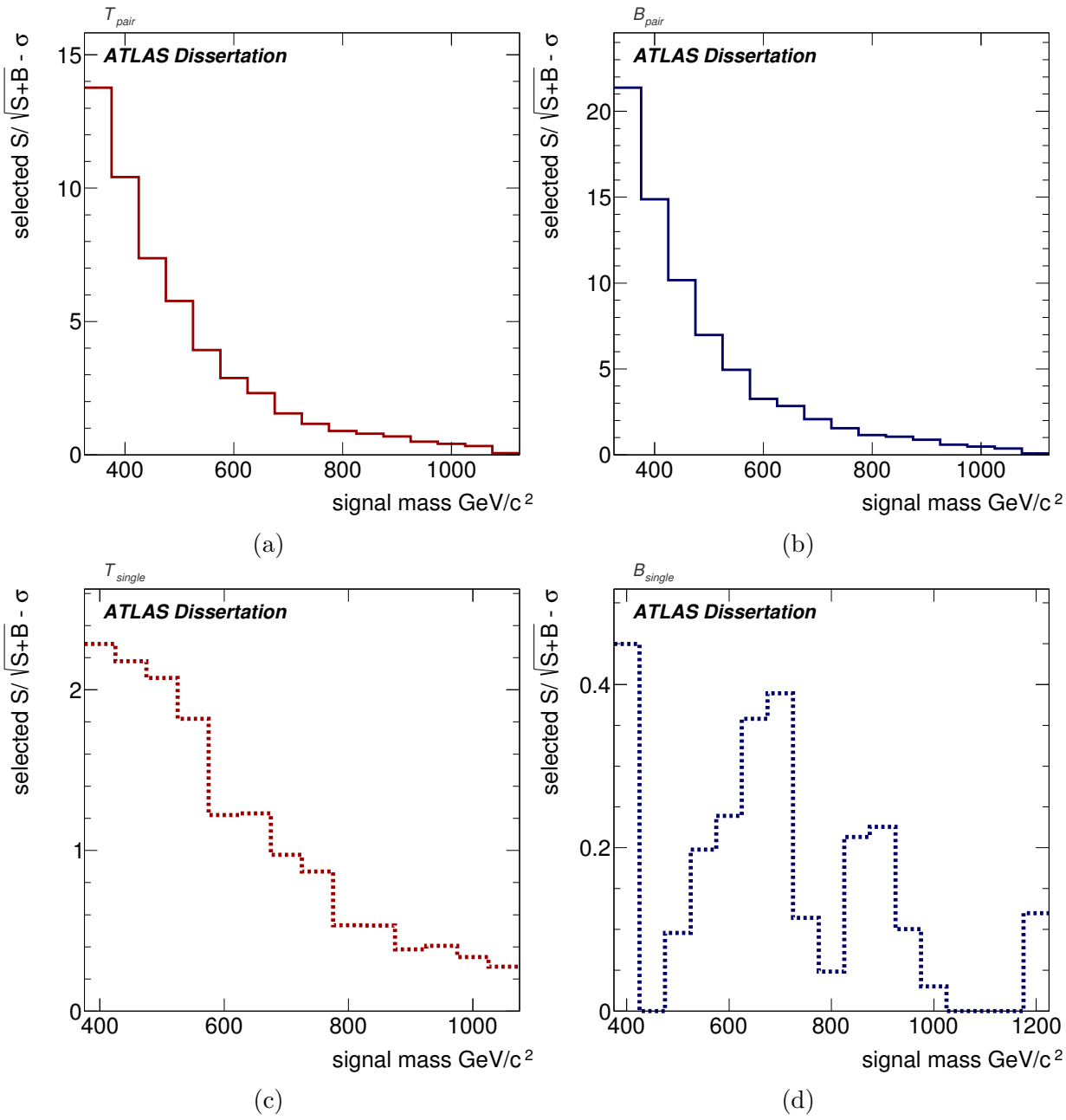


Figure 5.16: Signal significance by signal mass: (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}. (Note that $m(B_{single}) = 450$, and 1050-1150 GeV/c^2 were not included in the search).

5.6. Stage 2 Classification with Logistic Regression

The strategy of stage 2 was to use the many stage 1 BDT classifier functions constructed per signal mass as input variables for a second stage classification learning algorithm (not necessarily a BDT). This was done for each of the four signal types as shown in figure 5.17. BDTs had been found to be the superior learning algorithm for stage 1 due to the desire to incorporate as many types of information as possible into the training (as many useful input variables as possible) while avoiding overtraining. These stage 1 BDTs extracted this information, encapsulating it within their class prediction functions.

The additional variables $H_T(jets)$ and $H_T(jets + Z)$ contained useful information regarding the total energy of each event. Histograms of these are shown in figures 5.18 (a) and (b), respectively. These were included in stage 2 classification because for each event, it was hypothesized that these allowed the stage 2 learning algorithm to determine which of the stage 1 classifiers was most appropriate for a given event. For example, a signal event with a large $H_T(jets)$ was more likely to correspond to a higher signal mass, in which case the stage 1 classifiers trained on higher signal masses would be more useful for signal separation for this event.

The $N(forward\ jets) \geq 1$ variable was used during stage 2 training instead of stage 1 training. Its histogram is shown in figure 5.18 (c). It provided useful information for signal separation for the singly produced signals, B_{single} and T_{single}, but its inclusion during stage 1 caused degradation of performance due to its binary range and low correlation to other variables. Attempts to include $N(forward\ jets) \geq 1$ during stage 1 training essentially had the effect of splitting the training data into two subsets during learning, those with a forward jet and those without, and thus decreased the usefulness of the other input variables. This demonstrates a further advantage of multi-stage learning: variables containing useful information that prove problematic to use in conjunction with other variables may be moved to later stages of learning. The use of the $N(forward\ jets) \geq 1$ variable during stage 2 learning was not necessary for the T_{pair} and B_{pair} signal types, but its inclusion was not observed to effect the stage 2 classification results, and having common input variables for all four signal types simplified the analysis.

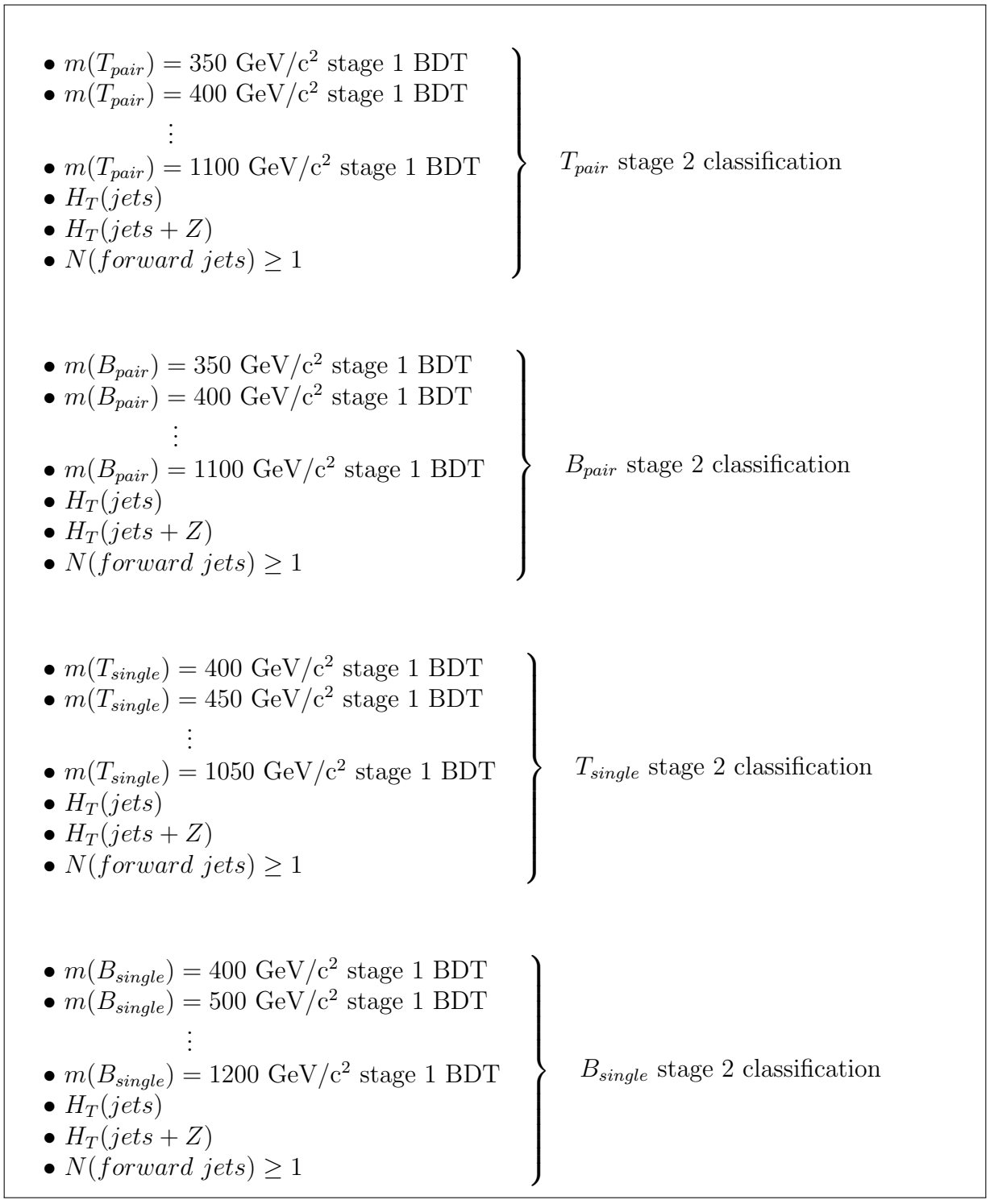
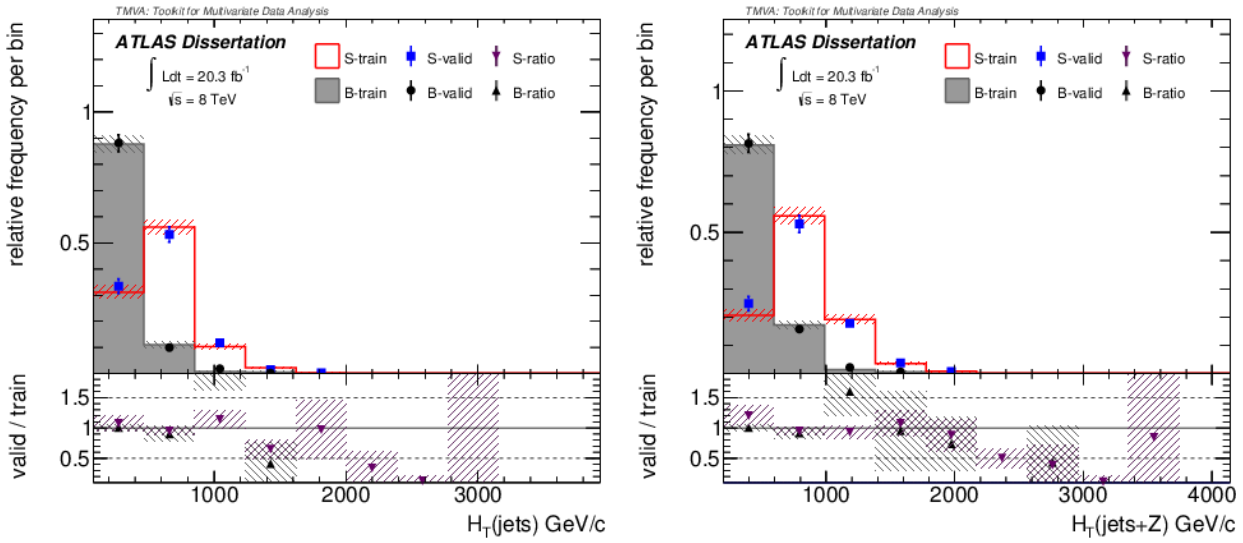
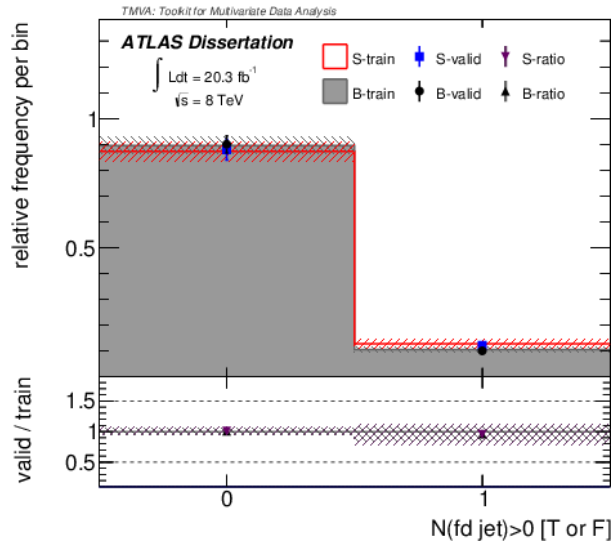


Figure 5.17: The two-stage classification strategy for each of the four signal types. The BDT classifier response value outputs for each signal mass from stage 1 machine learning (and three additional variables) become inputs for the four stage 2 classifications.



(a) (b)



(c)

Figure 5.18: Histograms of the additional stage 2 variables for the T_{pair} signal type: (a) $H_T jets$, (b) $H_T(jets + Z)$, and (c) $N(forward\ jets) \geq 1$.

5.6.1. Selection of Logit Classification for stage 2

The choice of the type of learning algorithm to use for stage 1 was motivated by the desire to extract information from many input variables without overtraining, and BDTs proved the most capable. At stage 2, the majority of the information contained in the input variables had already been extracted, contained in the stage 1 class prediction functions, and a simple learning algorithm was needed that would "stitch" these classifiers together (along with the three additional variables). BDTs, MLPs, and SVMs were examined for use as the stage 2 classifier, but usually resulted in overtraining (discussed in appendix section C.1). However, simple logistic regression ("logit classification") was found to perform well with little or no overtraining.

Table 5.9 shows strategy for stage 2 learning. After the stage 2 classifier was obtained by logit classification, a decision rule was implemented (square cut on the stage 2 classifier value), and statistical inference proceeded using the reconstructed mass spectra (discussed further in section 5.7). Also, an additional $p_T(Z) > 150$ GeV/c square cut had to be performed. Recall that $p_T(Z) < 150$ GeV/c was used as a control region, and could not be used for any statistical inference. Nevertheless, it was often useful to include events with $100 < p_T(Z) < 150$ GeV/c during training to address data insufficiency, and also because these events still contained useful information, particularly regarding the b-tagging input variables.

The decision rule implemented using the results of stage 2 logit classification was a cut on the classifier response value to form the signal region (i.e., keeping data only above a certain threshold of the classifier response value). This method was preferable to statistical inference on the classifier response value itself since the sharply peaked $m(Zb)$ spectrum provided further signal separation, and preferable to a two dimensional spectrum of $m(Zb)$ versus classifier response value due to data insufficiency. The cut value was chosen with high

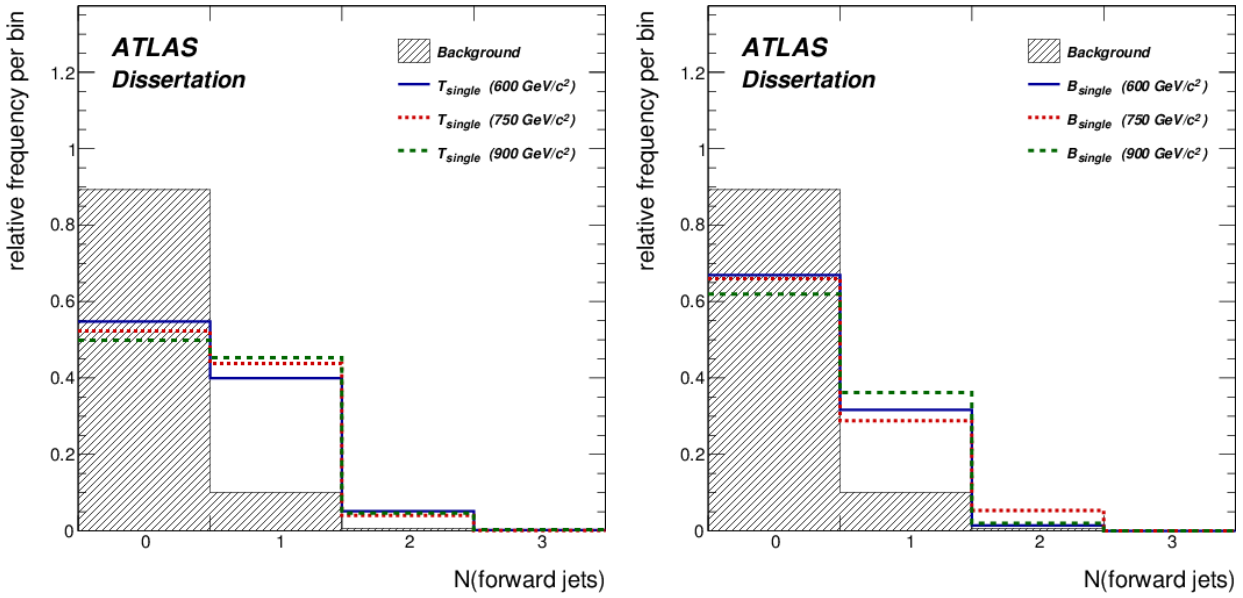
$Q_{production}$	selected stage 2 classification learning algorithm	▷ stage 2 decision rule • additional requirement	spectrum for statistical inference
T_{pair}	logit classification	▷ logit classifier value > 0.8 • $p_T(Z) > 150$ GeV/c	$m(Zb)$
B_{pair}	logit classification	▷ logit classifier value > 0.8 • $p_T(Z) > 150$ GeV/c	$m(Zb)$
T_{single}	logit classification	▷ logit classifier value > 0.8 • $p_T(Z) > 150$ GeV/c • $N(forward\ jets) \geq 1$	$m(Zb)$ or $m(Zt)$
B_{single}	logit classification	▷ logit classifier value > 0.8 • $p_T(Z) > 150$ GeV/c	$m(Zb)$

Table 5.9: The selected stage 2 classification learning algorithms for each of the four signal types.

significance in mind, but was mainly determined by the need to retain sufficient data to form an $m(Zb)$ spectra for statistical inference with cut values higher than 0.8 resulting in too sparsely populated $m(Zb)$ spectra for accurate statistical inference.

As square cut values for the decision rule were explored (determining the classifier response value threshold), it was observed that the $m(Zb)$ spectra to be used for statistical inference retained a consistent shape. This indicated that no "false peaks" were being generated as the square cut value was increased, which could have led to false detection or else poor exclusion limits (i.e., little or no correlation between the $m(Zb)$ spectra and the stage 2 classifier value was observed for any of the four signal types).

Figures 5.19(a) and (b) show the histograms of the forward jet multiplicity of the singly produced VLQ signal types. For the T_{single} signal, requiring a forward jet was found to provide increased significance and good modeling accuracy, even though the presence of a forward jet had been included during the training of the stage 2 classifier. This implied that for T_{single}, signal separation for events without a forward jet was poor. However, this was not the case for B_{single}, which had a much larger fraction of events without a forward jet so that retaining events without a forward jet improved significance.



(a) (b)

Figure 5.19: Histograms of $N(forward\ jets)$ for selected signal masses of the singly produced VLQ signal types: (a) T_{single}, and (b) B_{single}.

5.6.2. Logit Classification Results

Figure 5.20 shows the single node MLP neural networks of the TMVA software that provided logistic regression. These indicated that the T_{pair}, B_{pair} and T_{single} signals utilized the $H_T(jets)$ variable more than the B_{single} signal (compare the size of the connecting green vertices between the $H_T(jets)$ input node and the center node, which indicates the size of the node weighting). This was because the decay topology of the B_{single} signal type typically produced fewer jets than the other three signal types. Figure 5.20 also shows that the T_{pair} and B_{pair} signal types had more similar node weightings since their decay topologies were expected to be more similar to one another than those of T_{single} or B_{single}, which were each expected to have more unique topologies.

Figures 5.21–5.24 show the results of stage 2 logit classification. Each demonstrated good signal separation and negligible overtraining except for the T_{single} signal. The T_{single} signal shows some overtraining for the signal distribution. Unlike overtraining of backgrounds which can result in the degradation of modeling accuracy, some overtraining of the signal distribution is not critically important since it has no effect on the modeling accuracy of the simulated backgrounds. Figure 5.23(c) indicates that this overtraining resulted in a tolerable decrease in significance.

Figures 5.25–5.28 detail the performance of signal separation for the four signal types as well as the significance that is obtained for cuts on the stage 2 classifier value. These generally indicated that the significance for the signal masses was largest for the B_{pair} signal type, closely followed by T_{pair}, then T_{single}, and finally quite poor for B_{single}.

Table 5.10 shows the data sufficiency of the simulated data: the unweighted number of simulated events used for the analyses. Before implementing the decision rule of retaining only events reaching a stage 2 classifier response value of 0.8, the T_{pair} and B_{pair} signals used random partitioning of the simulated data that assigned ∼15% of the simulated events for machine learning training, ∼15% for machine learning validation, and ∼70% for the final statistical inference[10]. The T_{single} and B_{single} signals were more difficult to separate, and required more training data with a 25%:25%:50% partitioning.

[10]Note that the amount of simulated data used for training and validation were equal because the training and validation partitions were switched between stage 1 and stage 2 according to partitioning scheme shown in table 5.3 (c).

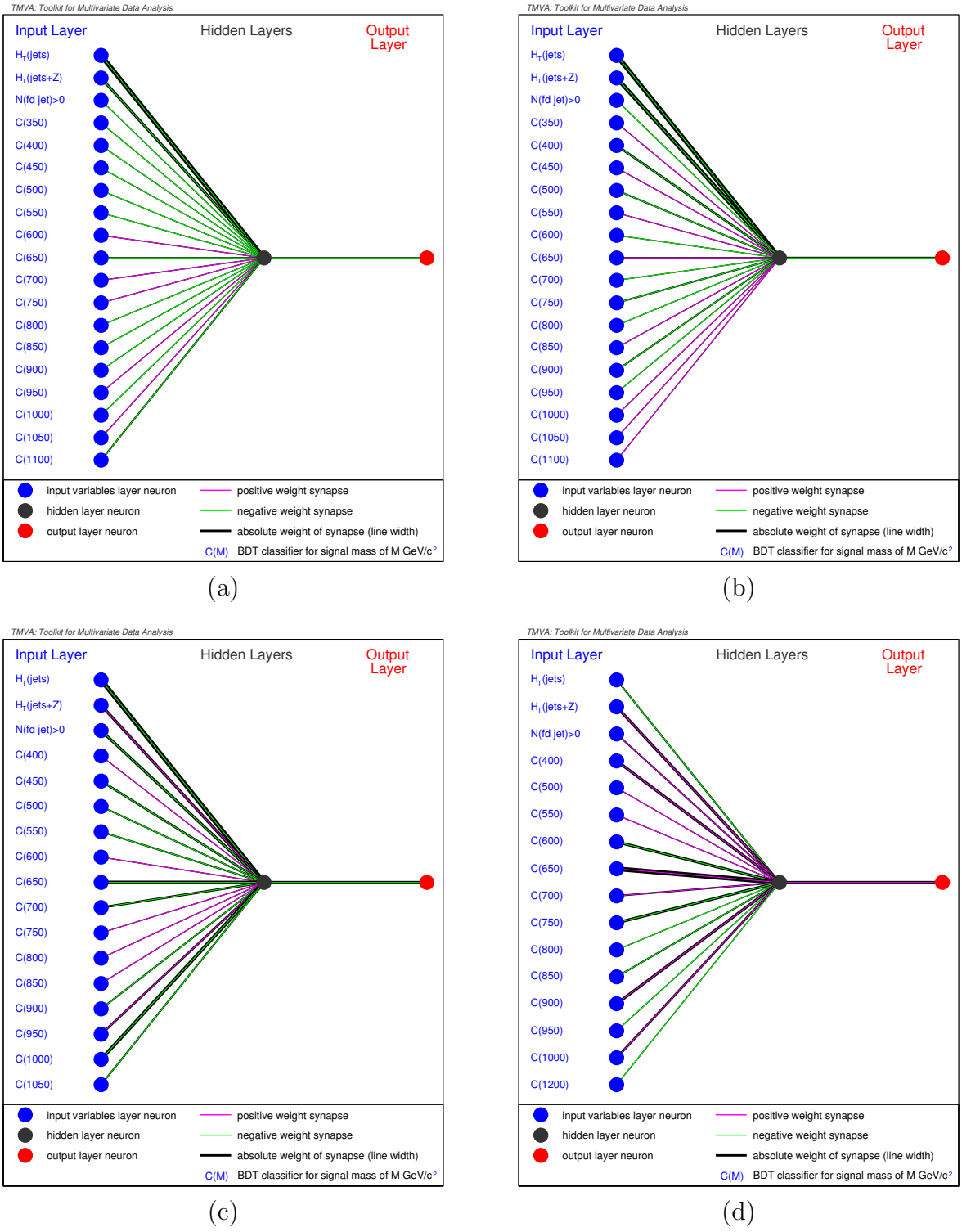


Figure 5.20: The single node neural networks for the multilayer perceptrons with sigmoid neural response that provide logit classification for the four signal types: (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

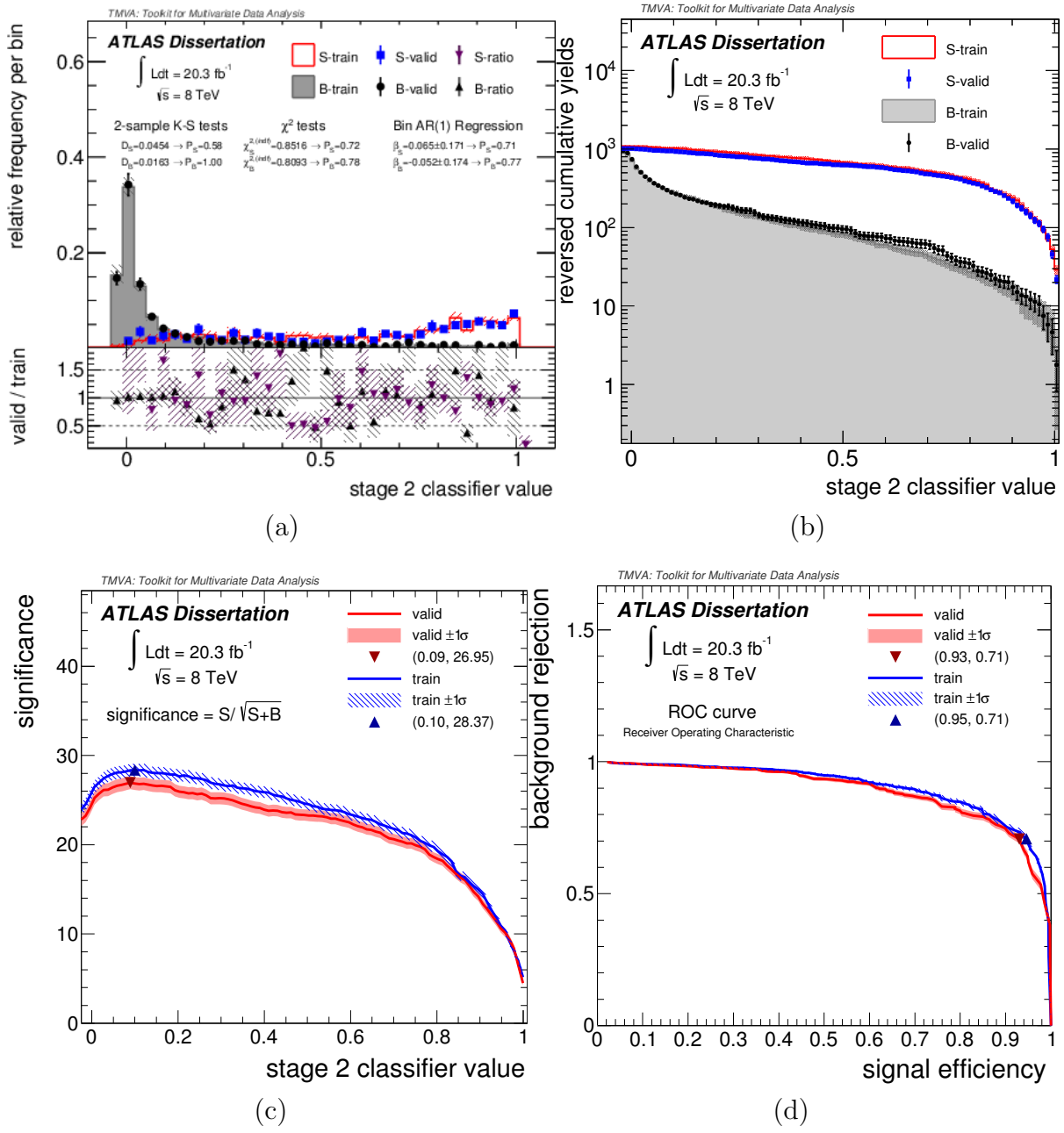


Figure 5.21: The results of logit classification for the T_{pair} signal type: (a) histogram, (b) (reversed) cumulative distribution of the predicted yields, (c) significance, and (d) receiver operator characteristic (ROC curve).

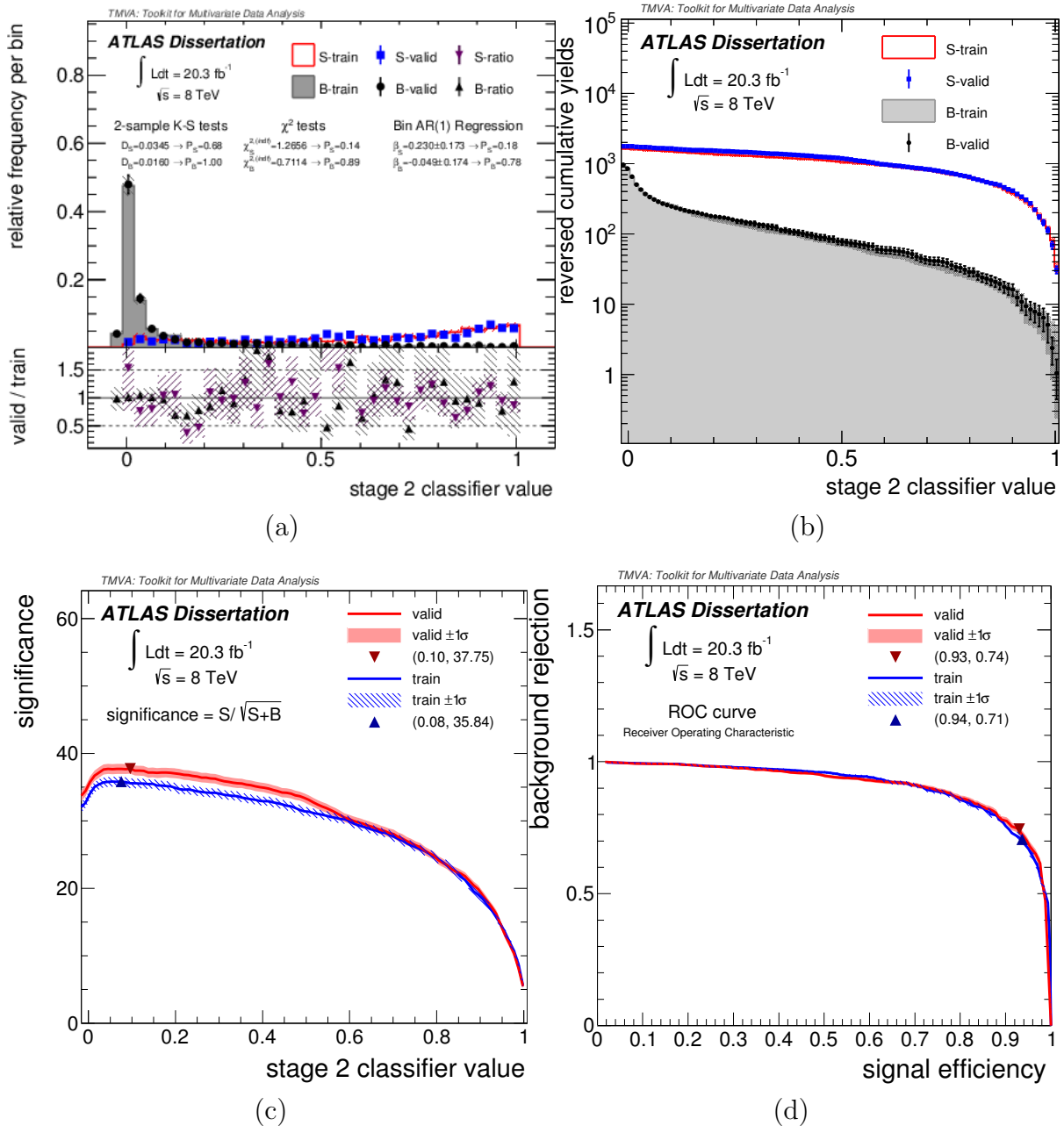


Figure 5.22: The results of logit classification for the B_{pair} signal type: (a) histogram, (b) (reversed) cumulative distribution of the predicted yields, (c) significance, and (d) receiver operator characteristic (ROC curve).

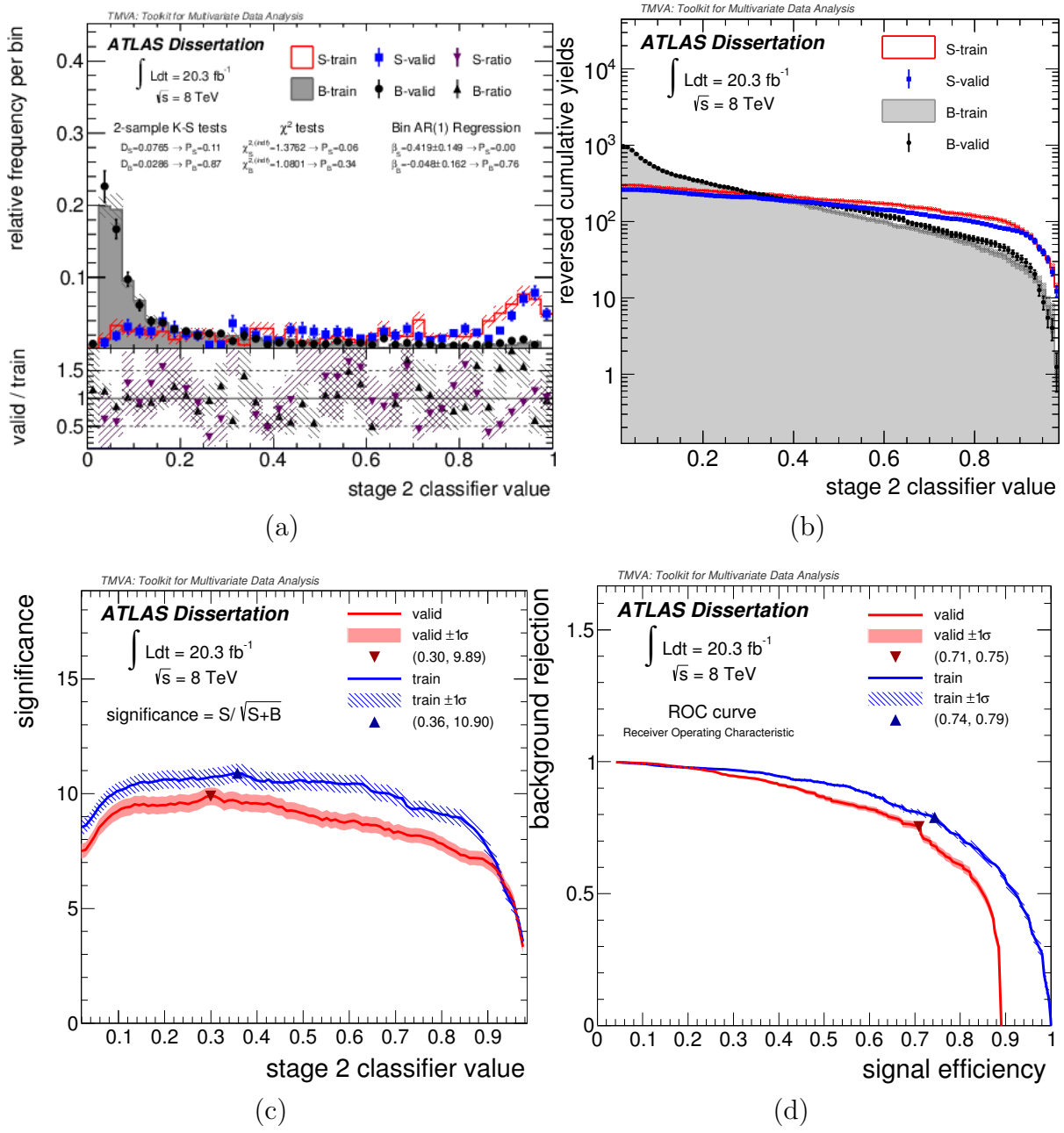


Figure 5.23: The results of logit classification for the T_{single} signal type: (a) histogram, (b) (reversed) cumulative distribution of the predicted yields, (c) significance, and (d) receiver operator characteristic (ROC curve).

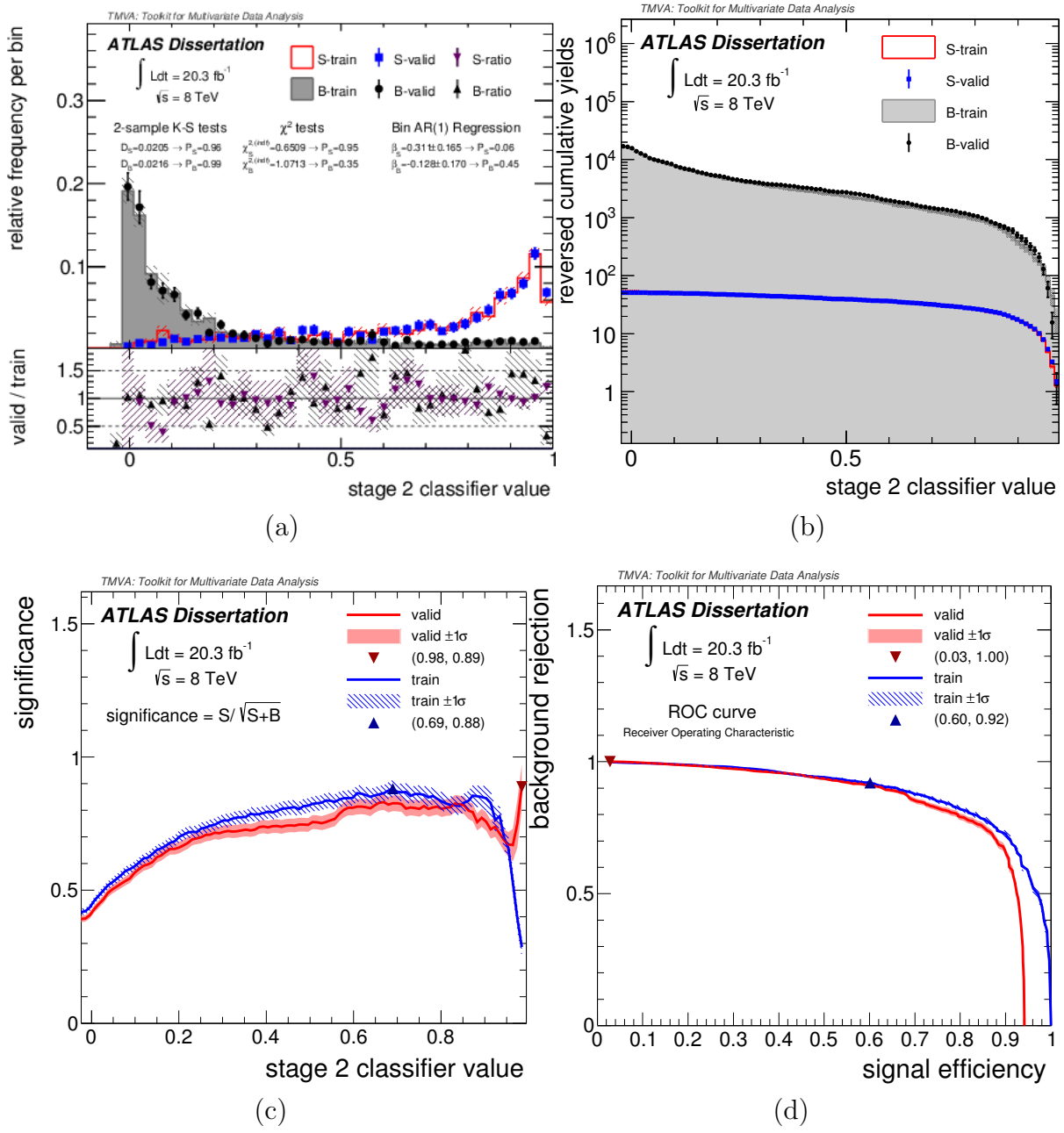
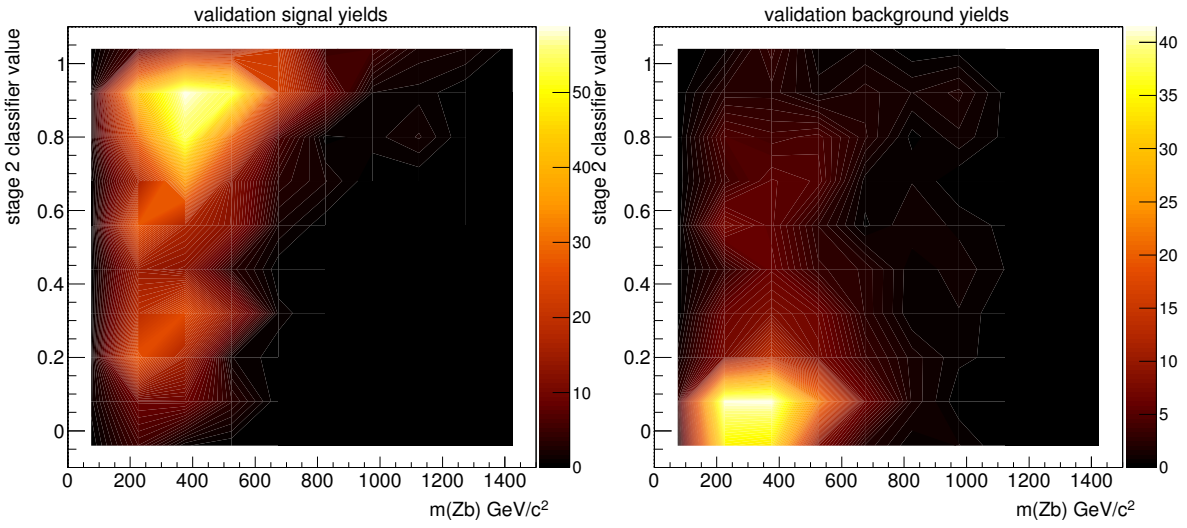
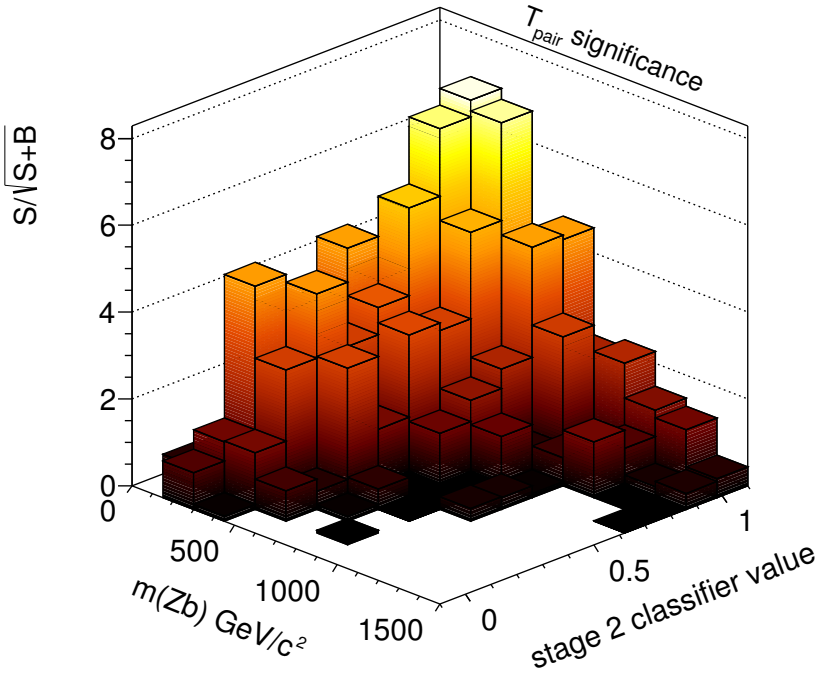


Figure 5.24: The results of logit classification for the B_{single} signal type: (a) histogram, (b) (reversed) cumulative distribution of the predicted yields, (c) significance, and (d) receiver operator characteristic (ROC curve).

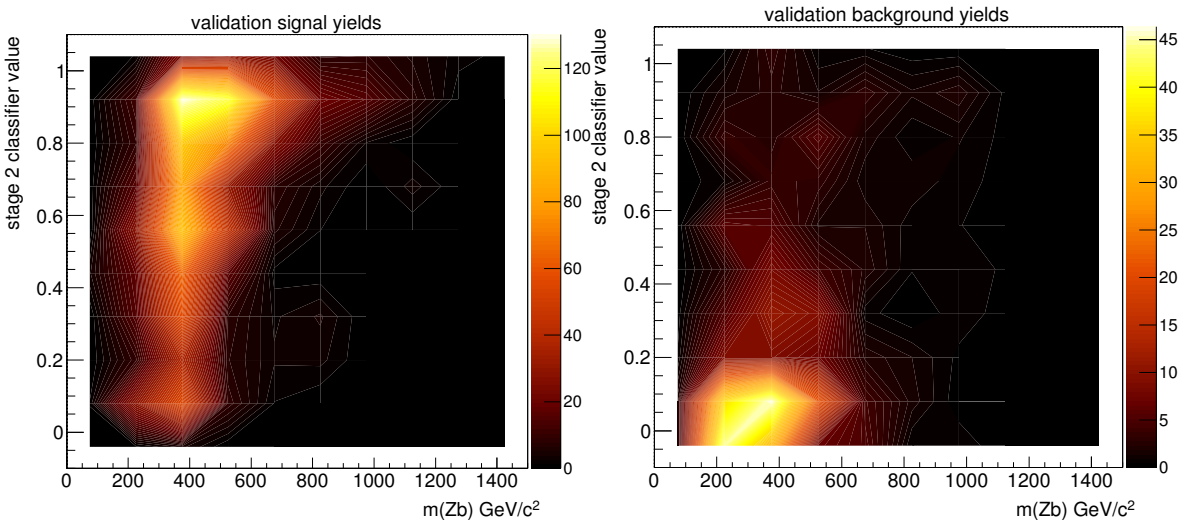


(a) (b)

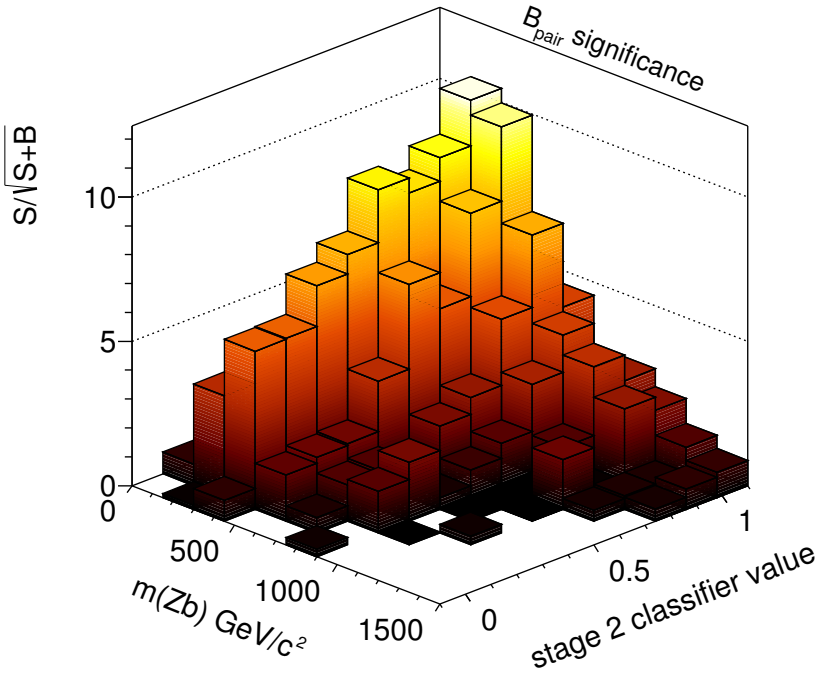


(c)

Figure 5.25: Heat maps showing the predicted yields of the stage 2 logit classifier response value over the $m(Zb)$ spectrum for the T_{pair} signal type: (a) validation signal, and (b) validation background. (c) A lego plot showing the significance achieved by validation data remaining above any cut on given classifier response over the $m(Zb)$ spectrum.

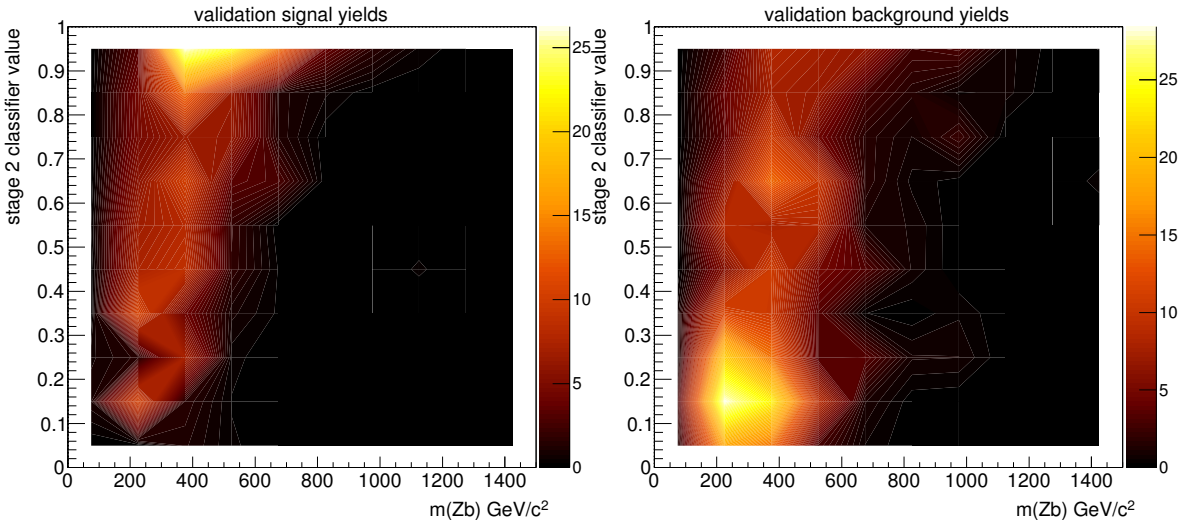


(a) (b)

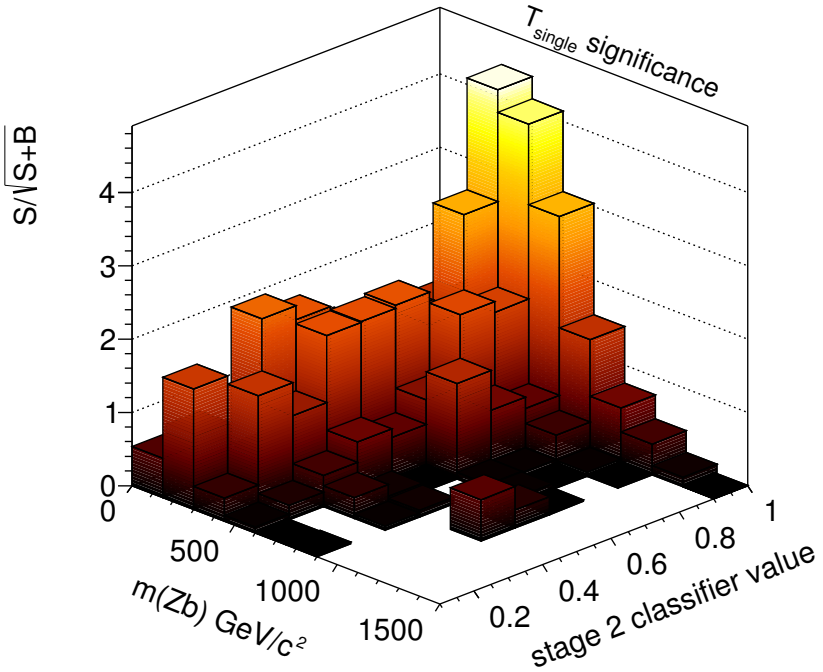


(c)

Figure 5.26: Heat maps showing the predicted yields of the stage 2 logit classifier response value over the $m(Zb)$ spectrum for B_{pair} the signal type: (a) validation signal, and (b) validation background. (c) A lego plot showing the significance achieved by validation data remaining above any cut on given classifier response over the $m(Zb)$ spectrum.

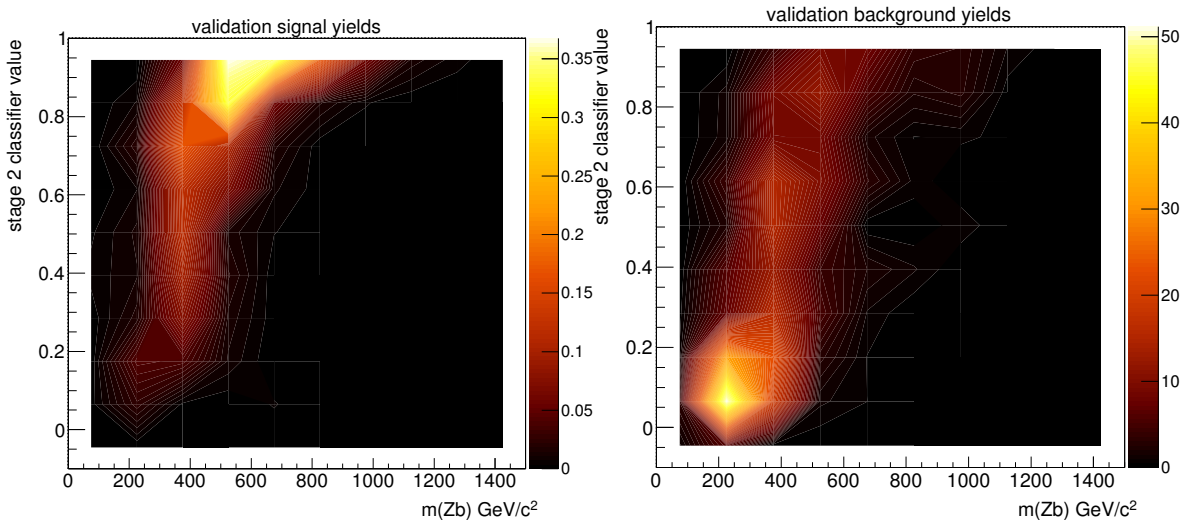


(a) (b)

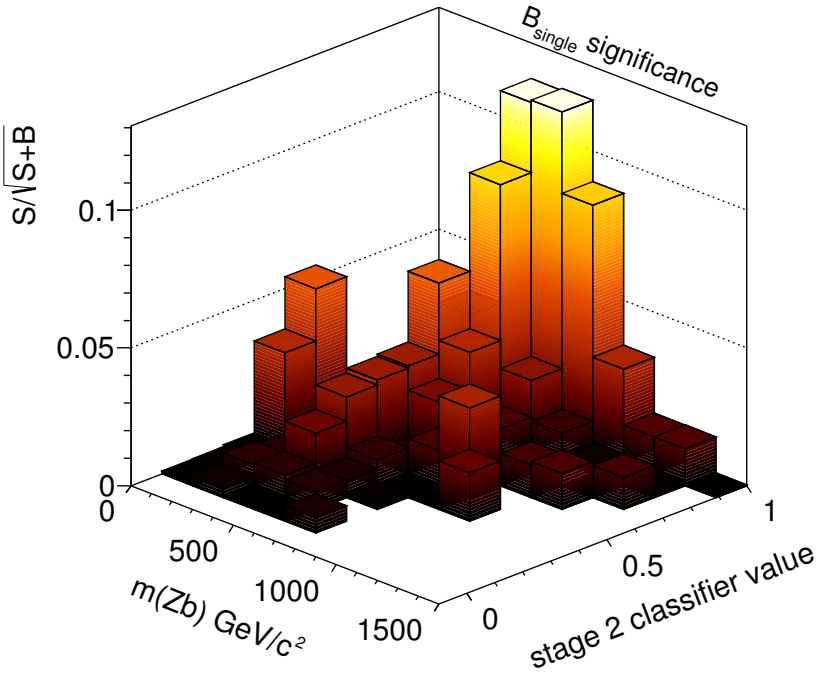


(c)

Figure 5.27: Heat maps showing the predicted yields of the stage 2 logit classifier response value over the $m(Zb)$ spectrum for the T_{single} signal type: (a) validation signal, and (b) validation background. (c) A lego plot showing the significance achieved by validation data remaining above any cut on given classifier response over the $m(Zb)$ spectrum.



(a) (b)



(c)

Figure 5.28: Heat maps showing the predicted yields of the stage 2 logit classifier response value over the $m(Zb)$ spectrum for the B_{single} signal type: (a) validation signal, and (b) validation background. (c) A lego plot showing the significance achieved by validation data remaining above any cut on given classifier response over the $m(Zb)$ spectrum.

dataset	before decision rule			after decision rule			$p_T(Z) > 150$ GeV/c*		
	n_{train}	n_{valid}	n_{test}	n_{train}	n_{valid}	n_{test}	n_{train}	n_{valid}	n_{test}
T_{pair} 350 GeV/c^2	217	189	953	46	40	158	39	37	130
T_{pair} 400 GeV/c^2	282	275	1346	87	94	367	78	80	321
T_{pair} 450 GeV/c^2	315	286	1571	125	122	625	105	109	555
T_{pair} 500 GeV/c^2	397	310	1714	229	177	904	202	158	801
T_{pair} 550 GeV/c^2	391	364	1816	266	259	1232	238	225	1099
T_{pair} 600 GeV/c^2	470	477	2174	357	377	1687	327	364	1548
T_{pair} 650 GeV/c^2	417	374	1866	355	329	1570	337	304	1473
T_{pair} 700 GeV/c^2	795	843	3862	695	740	3396	662	707	3215
T_{pair} 750 GeV/c^2	433	440	1922	387	399	1759	365	385	1668
T_{pair} 800 GeV/c^2	382	387	1863	363	359	1721	344	346	1653
T_{pair} 850 GeV/c^2	412	366	1871	386	349	1763	368	339	1699
T_{pair} 900 GeV/c^2	498	471	2115	476	443	2024	468	428	1984
T_{pair} 950 GeV/c^2	464	435	2237	444	421	2163	432	407	2109
T_{pair} 1000 GeV/c^2	411	479	2167	401	466	2102	392	455	2039
T_{pair} 1050 GeV/c^2	428	414	2146	422	398	2093	411	390	2041
T_{pair} 1100 GeV/c^2	434	392	1932	426	387	1895	417	382	1846
background	3392	3362	16360	277	261	1355	253	228	1221
B_{pair} 350 GeV/c^2	400	427	1936	83	94	442	76	85	398
B_{pair} 400 GeV/c^2	434	436	2196	167	156	779	149	132	703
B_{pair} 450 GeV/c^2	476	481	2119	235	240	1065	210	218	957
B_{pair} 500 GeV/c^2	494	461	2289	331	303	1505	305	275	1383
B_{pair} 550 GeV/c^2	465	481	2336	352	353	1787	334	325	1649
B_{pair} 600 GeV/c^2	423	444	2006	348	380	1668	329	355	1572
B_{pair} 650 GeV/c^2	504	473	2305	449	418	2038	427	400	1941
B_{pair} 700 GeV/c^2	957	994	4509	887	901	4129	854	859	3952
B_{pair} 750 GeV/c^2	490	480	2231	461	459	2090	444	448	2013
B_{pair} 800 GeV/c^2	468	461	2190	446	444	2117	435	431	2040
B_{pair} 850 GeV/c^2	482	435	2138	470	418	2078	456	410	2031
B_{pair} 900 GeV/c^2	560	567	2455	549	555	2397	538	543	2340
B_{pair} 950 GeV/c^2	539	536	2454	530	531	2398	514	521	2352
B_{pair} 1000 GeV/c^2	562	477	2406	553	466	2371	546	461	2330
B_{pair} 1050 GeV/c^2	512	512	2273	502	507	2239	489	499	2209
B_{pair} 1100 GeV/c^2	510	468	2191	504	465	2167	496	460	2138
background	3392	3362	16360	273	254	1331	252	229	1225

Table 5.10—*continues on next page*

dataset	before decision rule			after decision rule			$p_T(Z) > 150$ GeV/c*		
	n_{train}	n_{valid}	n_{test}	n_{train}	n_{valid}	n_{test}	n_{train}	n_{valid}	n_{test}
T_{single} 400 GeV/c^2	102	110	236	15	15	25	12	13	18
T_{single} 450 GeV/c^2	116	105	218	28	18	43	18	14	26
T_{single} 500 GeV/c^2	139	143	295	43	49	88	30	38	70
T_{single} 550 GeV/c^2	163	141	292	72	66	120	51	50	88
T_{single} 600 GeV/c^2	158	155	333	82	75	190	53	49	122
T_{single} 650 GeV/c^2	162	165	354	110	111	220	65	54	113
T_{single} 700 GeV/c^2	179	155	330	136	101	241	61	51	120
T_{single} 750 GeV/c^2	150	176	349	115	137	256	50	60	133
T_{single} 800 GeV/c^2	133	162	348	106	138	281	54	61	125
T_{single} 850 GeV/c^2	152	190	344	133	159	295	58	69	121
T_{single} 900 GeV/c^2	164	161	351	142	139	303	64	71	128
T_{single} 950 GeV/c^2	177	169	339	155	146	294	68	65	119
T_{single} 1000 GeV/c^2	165	170	336	146	156	289	64	63	143
T_{single} 1050 GeV/c^2	135	179	297	119	158	270	54	68	126
background	3392	3362	16360	429	423	2133	85	84	489
B_{single} 400 [GeV]	217	232	442	33	36	70	33	33	67
B_{single} 500 [GeV]	226	231	484	91	108	206	90	108	202
B_{single} 550 [GeV]	263	234	470	144	127	283	144	127	279
B_{single} 600 [GeV]	242	204	481	151	145	303	151	143	299
B_{single} 650 [GeV]	228	230	468	163	180	354	163	176	353
B_{single} 700 [GeV]	215	230	458	175	188	358	173	188	357
B_{single} 750 [GeV]	208	203	467	162	172	384	160	171	383
B_{single} 800 [GeV]	209	234	458	176	209	388	175	209	387
B_{single} 850 [GeV]	212	219	418	181	189	352	181	188	352
B_{single} 900 [GeV]	218	226	389	200	198	343	200	198	343
B_{single} 950 [GeV]	203	196	413	180	176	377	180	176	377
B_{single} 1000 [GeV]	210	195	401	190	178	362	190	178	362
B_{single} 1200 [GeV]	175	199	338	161	185	314	161	185	314
background	3392	3362	16360	399	393	1895	395	391	1886

Table 5.10: A data sufficiency table showing the number of simulated events (unweighted) used for machine learning training (n_{train}), learning validation (n_{valid}), and for final statistical inference (n_{test}). The decision rule was a square cut on the stage 2 logit classifier response value (retaining only events reaching a threshold of 0.8). The $p_T(Z) > 150$ GeV/c requirement removed all events that were also used previously for a control region (described in chapter 4).

* *The additional requirement of N(forward jets)≥1 was used for the T_{single} signal.*

5.7. Stage 3 Machine Learning Regression

The reconstructed mass of a candidate particle was obtained by summing the energy-momentum four-vectors ("four-momenta") of the physics objects thought most likely to represent the decay products of the candidate particle. The resultant four-momentum then provided an measurement of the mass of the candidate particle. For instance, the vector-like bottom quark, B, decayed to a Z boson and a Standard Model bottom quark, b. In the Z-tag search, the four-momentum of the candidate Z boson was estimated as the sum of the four-mementa of the two same-flavor, opposite-sign leptons in the event, $Z^\mu = l^\mu + \bar{l}^\mu$.[11] The b quark candidate was estimated as the four-momentum of the hadronic jet with highest p_T that was b-tagged, j_b^μ. Therefore, an estimate of the four-momentum of the candidate B was obtained from the sum $B^\mu \approx Z^\mu + j_b^\mu$, and from this, an estimate of the invariant mass of the B was obtained, $m(Zb) \equiv -\sqrt{\eta_{\mu\nu} B^\mu B^\nu}/c^2$, where $\eta_{\mu\nu}$ was the Minkowski metric tensor.

For the B, $m(Zb)$ provided a good estimate of the mass of the B quark for both the B_{pair} and B_{single} signal types. However, the vector-like top quark, T, decayed to a Z boson and a Standard Model top quark, t, with the t further decaying to a b quark and W boson, $T \to Z + \bar{b} + W^+$. In this case, $m(Zb)$ only provided a partial reconstruction of the estimated invariant mass of T since it was missing the contribution from the W boson four-momentum. This partially reconstructed vector-like top mass weakened the power of final statistical inference. During the analysis that led to publication [41], many attempts to fully reconstruct the T mass were attempted, but none was successful. Although in this dissertation a full $m(Zt)$ reconstruction was obtained for the T_{single} signal type,[12] none was found for the T_{pair} signal type.

Regression using machine learning offered an alternate approach to using a reconstructed mass spectrum for statistical inference. Rather than estimate the signal masses using only the information contained in the four-momenta of the candidate decay products, regression utilized all the information of the event variables in order to estimate the signal masses. In other words, the signal mass became the target variable that the regressand attempted to predict.

The signal mass was known for each of the signals, of course, so each signal mass was input as the target variable for training the signal data sets. The background contains no signal, and so it was ambiguous what the target should be for these data sets. For simplicity, the target variable was set to 0 GeV/c^2 for the training of the background data sets. In this manner, machine learning regression provided a spectrum for the predicted signal mass that was used for statistical inference.

---

[11] The four-momentum of a physics object, l^μ, is given by its total energy and three dimensional momentum components, $l^\mu = [E, p_x, p_y, p_z]$.

[12] The reconstruction of the vector-like top mass was accomplished by attempting to identify the decay products of the Standard Model top quark, t, to form a candidate t quark, and then combine these with the Z boson candidate. Attempts to separately identify the t quark's W boson decay daughter proved too problematic in the busy topologies of the collision events. Instead, the candidate b quark four-momentum, b^μ, with the highest transverse momentum was used as a "seed" for the initial four-momentum of the t, t_0^μ. Next, the four-momentum of the highest p_T anti-b-tagged jet, $j_{a,1}^\mu$, within $\Delta R(b, j_{a,1}) \leq 1.5$ was added to give $t_1^\mu = b^\mu + j_{a,1}^\mu$. Next, the highest p_T anti-b-tagged jet, $j_{a,2}^\mu$, within $\Delta R(t_1, j_{a,2}) \leq 1.5$ was then added to the t candidate to give $t_2^\mu = b^\mu + j_{a,1}^\mu + j_{a,2}^\mu$. This process continued until no further anti-b-tagged jets were within $\Delta R(t_n, j_{a,n}) \leq 1.5$. The invariant mass of this t_n candidate top quark is labeled $m(bW)$.

The input variables for stage 3 regression were the same as for stage 2 classification: the stage 1 classifiers for each signal mass, $H_T(jets)$, $H_T(jets+Z)$, and $N(forward\ jets) \geq 1$, as well as including the reconstructed $m(Zb)$ mass. The preselection used for stage 3 regression was the same as for stage 2 classification, with the addition of the $p_T(Z) > 150$ GeV/c square cut to remove the portion of the data containing a control region, as shown in table 5.11.

Note that machine learning was *not* used to identify the specific decay products of the vector-like quarks, which would have resulted in a candidate vector-like quark giving a fully reconstructed vector-like quark mass. Identification of decay products was viewed as an unnecessarily complex task. The complicated decay topologies of the collision events made specific identification of decay products prone to error (especially for vector-like quarks produced in pairs). For each collision event, an estimate of the vector-like quark invariant mass was ultimately desired, and so rather than attempt to identify the decay products of the vector-like quark as an intermediate step, this method of machine learning simply proceeded to the estimate in a single step.[13]

Simulated data not passing the stage 2 decision rule was also included in stage 3 training to address data insufficiency. Several learning algorithms were attempted, with final selection of best methods shown in table 5.11. These methods are described in the following subsections.

$Q_{production}$	∘ additional preselection ▷ stage 2 decision rule • additional requirement	regression spectrum for statistical inference
T_{pair}	∘ $p_T(Z) > 150$ GeV/c ▷ logit classifier value > 0.8	OLS with quadratic terms
B_{pair}	∘ $p_T(Z) > 150$ GeV/c ▷ logit classifier value > 0.8	OLS with quadratic terms
T_{single}	∘ $p_T(Z) > 150$ GeV/c ▷ logit classifier value > 0.8 • $N(forward\ jets) \geq 1$	BDT regression
B_{single}	∘ $p_T(Z) > 150$ GeV/c ▷ logit classifier value > 0.8	OLS without quadratic terms

Table 5.11: The strategy for stage 3 machine learning regression for the four signal types.

[13] Because this method used the results of stage 1 machine learning as input variables, and because it used the results of stage 2 to define a final signal region, it was labeled "stage 3".

5.7.1. Linear Regression *(selected for T_{pair}, B_{pair}, and B_{single})*

> *"As a rule of thumb, when faced with learning problems, it is generally a winning strategy to try a linear model first."*
> – Yaser S. Abu-Mostafa, et al. [165]

MLP and both adaptive and gradient BDT learning algorithms were attempted for stage 3 regression, but none performed as well as standard least squares linear regression (with the exception of gradient boosted decision trees in some cases). The four least squares linear regression models that were investigated are briefly described here (a more complete discussion was provided in section 5.2.4).

1. **Ordinary Least Squares** (OLS) — The OLS model was $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where the per event disturbances u_i were independent and identically distributed (i.i.d.) with $u_i \sim \mathcal{N}(0, \sigma^2)$, and where the disturbance variance σ has to be be estimated.
2. **Weighted Least Squares** (WLS) — The WLS model was $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, where the per event disturbances u_i were independent, but each had variance proportional to the inverse of each event weight, $u_i \sim \mathcal{N}(0, \sigma^2/w_i)$, and where σ still had to be estimated. Because the signal weights were much smaller than background, the signal training data was scaled so that the sum of signal training data weights equaled the sum of background training data weights, and this equally weighted signal and background was used to fit the weighted least squares models. This scaling was removed before computing the Breusch-Pagan test statistic and goodness of fit statistics.
3. **Two-stage Generalized Least Squares** (GLS) — When heteroscedasticity was detected by applying a Breusch-Pagan test, σ could no longer be assumed to be uniform for each of the data. This implied a model where the disturbances were no longer identically distributed, $u_i \sim \mathcal{N}(0, \sigma_i^2)$, or in vector notation, $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma_u})$, where $\boldsymbol{\Sigma_u}$ was the $N_{data} \times N_{data}$ disturbance covariance matrix. Because each event of the simulated particle physics data was independent of all other events, $\boldsymbol{\Sigma_u}$ was diagonal so that only the diagonal entries had to be estimated. This was be done by first regressing with OLS to obtain initial residuals, and then performing a second regression that estimated $\boldsymbol{\Sigma_u}$. With the estimated $\boldsymbol{\Sigma_u}$, a final regression was performed.
4. **Two-stage Generalized Weighted Least Squares** (WGLS) — This model was the same as the previous one, but incorporated the event weights.

It is important to restate here that the goal of these regressions was to obtain a useful *predictor* function of the signal masses that demonstrated little overtraining, and where its use did not lead to mismodeling of the measured data by the simulated data. The goal was not to provide *explanatory* support for some postulated model. In that case, linear regression is a useful tool because its results may be easily interpreted. Here, linear regression was a useful tool because its simplicity reduced the chance of overtraining and mismodeling while still providing useful signal mass predictions.

For each of these four regression models, a separate model was examined that included quadratic terms between each of the input variables, X_i and X_j (the regression *covariates*), in

order to model correlations between the input variables. This was accomplished by including a term $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{XBX} + \mathbf{u}$, where $\mathbf{B}$ was a symmetric $N_{variables} \times N_{variables}$ matrix. The diagonal entries of $\mathbf{B}$ were set to zero to disallow quadratic terms for individual variables, i.e., the X_i^2 terms were removed. The quadratic terms were added because there were significant correlations between several of the input variables (an example is shown in figure 5.33(b)).

Table 5.12 provides the summary statistics of the input variables used in stage 3 regression for the T_{pair} signal type. The training and validation partitions were the same as those used in stage 2 classification with the additional $p_T(Z) > 150$ GeV/c cut. Each variable was scaled by its standard deviation measured from the training partition in order to more easily compare the resultant regression coefficients. Similar summary statistics were obtained for the B_{pair} and B_{single} signals, and are provided in appendix C.2 (linear regression was not the selected stage 3 regression method for the T_{single} signal type).

Event weights for the simulated data included the theoretical cross section, total integrated luminosity, etc., representing the actual estimated yield of each individual event. There were ten orders of magnitude between the smallest weighted events and largest. This presented a challenge for the selection of the appropriate regression model (whether or not to use event weights) because the signal events had much smaller weights than most background, and because the background events most difficult to separate from signal also had small weights. For this reason, regression models both ignoring event weights (OLS and GLS) or including event weights (WLS and WGLS) were examined.

Tables 5.13 and 5.14 show the results of the eight regression models applied to the T_{pair} signal type: the last four columns of 5.13 show the results of the four models without the additional quadratic terms, and the last four columns of 5.14 show the results when the quadratic terms are included. The upper half of each table provides the parameter estimates ("betas") for the variables scaled by variance, which were useful for examining the predictive importance of each variable relative to the others, while the lower half removes the scaling. The standard errors of the parameter estimates are provided, along with the heteroscedasticity-corrected errors which are shown in parentheses [241].

The central rows of each table provide the estimated variance parameter, $\hat{\sigma}$ (not applicable for GLS and WGLS), the result of the Breusch-Pagan test for heteroscedasticity (not applicable for GLS and WGLS), and the several adjusted R^2 goodness of fit statistics:

- $\bar{R}^2_{train,unwtd.}$ — The training data sets without using event weights.
- $\bar{R}^2_{valid,unwtd.}$ — The validation data sets without using event weights.
- $\bar{R}^2_{train,wtd.}$ — The training data sets using event weights.
- $\bar{R}^2_{train,wtd.}$ — The validation data sets using event weights.

Similar results for the B_{pair} and B_{single} signals are provided in appendix C.2.

The Breusch-Pagan test indicated substantial heteroscedasticity for the OLS and WLS models of both tables 5.13 and 5.14 (without and with $\frac{1}{2}\mathbf{XBX}$, respectively). When generalized least squares was attempted to address heteroscedasticity, $\bar{R}^2_{train,unwtd.}$ was observed to increase substantially, from ∼0.6 to ∼0.8–0.9. However, this trend was not observed with the validation data sets as seen with $\bar{R}^2_{valid,unwtd.}$, which implied that the simplest method of dealing with heteroscedasticity, two-stage generalized least squares, was unsuccessful on out-of-sample data. It was also observed that OLS provided as good or better goodness of fit than WLS. Further inspection (not shown) revealed that WLS regression introduced bias that degraded the modeling accuracy of the simulated data. This was

not unexpected since the ten orders of magnitude spanned by the event weights essentially reduced the data set to only those events with reasonably large weights.

When event weights were included, $\bar{R}^2_{train,wtd.}$ and $\bar{R}^2_{valid,wtd.}$ were found to be much lower than their unweighted counterparts. This was likely due to highly weighted background events receiving predictions larger than 0 GeV/c^2. Nevertheless, $\bar{R}^2_{train,wtd.}$ and $\bar{R}^2_{valid,wtd.}$ did indicate that 1) out-of-sample performance ($\bar{R}^2_{valid,wtd.}$) was better than in-sample performance ($\bar{R}^2_{train,wtd.}$) for the OLS and WLS regressions, and that 2) GLS and WGLS models did not successfully address heteroscedasticity.

Based on these observations, OLS was the model selected for use in stage 3 regression for the T_{pair}, B_{pair}, and B_{single} signal types (the selection of BDT regression for T_{single} is discussed next in section 5.7.2). OLS provided the most consistently similar values of $\bar{R}^2_{train}$ compared to $\bar{R}^2_{valid}$, indicating little overtraining. The OLS model including quadratic terms was found to provide better performance for the T_{pair} and B_{pair} signals while the OLS model without quadratic terms provided better performance for the B_{single} signal (see appendix C.2 for all selected linear regression results).

Figures 5.29(a)–(d) show the results of OLS regression for the T_{pair} signal type. The histogram (a) and expected yields (b) demonstrated the absence of overtraining. Figures 5.29(c) and (d) show the expected yields of the 650 and 750 GeV/c^2 signal masses, respectively. The regressand masses for these signals peaked at values somewhat above their target signal mass values (650 and 750 GeV/c^2, respectively). Nevertheless, even before implementation of the decision rule (a cut on the stage 2 classifier response value), the regressand mass distributions for these signals had higher expected yields for the signal distributions than for the backgrounds near their maxima. Figures 5.30–5.31 show the OLS regression results for the B_{pair} and B_{single} signal types.

Figures 5.32 (a)–(c) provide further evaluation of the performance of OLS regressions for the T_{pair}, B_{pair}, and B_{single} signal types by comparing the signal masses, $m(Q)$, to the predicted regressand values using the validation signal data sets. These indicated that the regressand mass predictions were more accurate for higher signal masses.

The utility of replacing reconstructed mass spectra with regressand mass spectra is examined in section 5.9, where the results of statistical inference are examined. The results of this section demonstrate that regression for the signal mass using is possible when combined with multi-stage machine learning techniques. This may be particularly useful for particle physics searches where a full mass reconstruction is not possible, such as here with the T_{pair} signal type, or in searches with decay topologies containing undetectable particles, such as neutrinos.

covariates **y** and **X**	data sample	mean	standard deviation	minimum value	maximum value
$m\,(Q)$ GeV/c^2	train	617.79	356.44	0.00	1100.00
	valid	620.88	357.11	0.00	1100.00
1. $H_T(jets)$ GeV/c	train	878.75	420.32	97.69	2807.12
	valid	880.75	409.25	110.33	3927.52
2. $H_T(jets + Z)$ GeV/c	train	1237.15	512.47	278.94	3961.74
	valid	1237.75	494.60	280.80	4140.85
3. $m\,(Zb)$ GeV/c^2	train	578.11	280.45	113.68	3103.37
	valid	586.63	279.41	101.95	2370.12
4. $N(fd\ jet) > 0$ [T or F]	train	0.08	0.28	0.00	1.00
	valid	0.09	0.28	0.00	1.00
5. Classifier (350 GeV/c^2)	train	0.00	0.05	-0.29	0.04
	valid	0.00	0.05	-0.29	0.04
6. Classifier (400 GeV/c^2)	train	0.03	0.07	-0.26	0.07
	valid	0.02	0.07	-0.26	0.07
7. Classifier (450 GeV/c^2)	train	0.03	0.08	-0.31	0.09
	valid	0.03	0.08	-0.31	0.09
8. Classifier (500 GeV/c^2)	train	0.08	0.15	-0.55	0.19
	valid	0.08	0.15	-0.55	0.19
9. Classifier (550 GeV/c^2)	train	0.57	0.59	-1.00	0.98
	valid	0.56	0.60	-1.00	0.98
10. Classifier (600 GeV/c^2)	train	0.12	0.20	-0.58	0.28
	valid	0.12	0.20	-0.59	0.28
11. Classifier (650 GeV/c^2)	train	0.49	0.65	-1.00	0.97
	valid	0.49	0.65	-1.00	0.97
12. Classifier (700 GeV/c^2)	train	0.46	0.53	-0.96	0.92
	valid	0.46	0.53	-0.94	0.92
13. Classifier (750 GeV/c^2)	train	0.39	0.60	-0.99	0.87
	valid	0.40	0.60	-0.99	0.87
14. Classifier (800 GeV/c^2)	train	0.27	0.59	-0.99	0.75
	valid	0.28	0.59	-0.99	0.75
15. Classifier (850 GeV/c^2)	train	0.27	0.59	-0.99	0.78
	valid	0.28	0.59	-0.99	0.78
16. Classifier (900 GeV/c^2)	train	0.07	0.29	-0.71	0.33
	valid	0.07	0.29	-0.71	0.33
17. Classifier (950 GeV/c^2)	train	0.32	0.74	-1.00	0.96
	valid	0.33	0.73	-1.00	0.96
18. Classifier (1000 GeV/c^2)	train	0.25	0.62	-0.99	0.89
	valid	0.26	0.62	-0.99	0.90
19. Classifier (1050 GeV/c^2)	train	0.00	0.24	-0.59	0.27
	valid	0.01	0.24	-0.59	0.27
20. Classifier (1100 GeV/c^2)	train	0.08	0.49	-0.92	0.68
	valid	0.09	0.49	-0.92	0.68
event weights (**w**)	train	0.14	0.41	7.997e-10	8.072e+00
	valid	0.13	0.39	3.104e-09	5.516e+00
		N_{bkg}	$\sum_{bkg}(w_i)$	N_{sig}	$\sum_{sig}(w_i)$
	train	1235	246.27	5078	642.40
	valid	1203	240.54	4878	549.18

Table 5.12: Covariates summary statistics for the T_{pair} signal: in-sample simulated data for training ("train") and out-of-sample simulated data for validation ("valid").

independent covariate $\mathbf{X}$	$\boldsymbol{\beta}_{OLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\boldsymbol{\beta}_{WLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\boldsymbol{\beta}_{GLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2)$	$\boldsymbol{\beta}_{WGLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
constant	0.58 ±0.048 (±0.053)	0.28 ±0.048 (±0.061)	0.56 ±0.021	0.32 ±0.021
1. $H_T(jets)$	0.13 ±0.034 (±0.036)	0.01 ±0.038 (±0.045)	0.07 ±0.015	-0.10 ±0.018
2. $H_T(jets + Z)$	0.24 ±0.038 (±0.041)	0.44 ±0.043 (±0.052)	0.27 ±0.017	0.50 ±0.020
3. $m(Zb)$	0.03 ±0.011 (±0.012)	0.01 ±0.011 (±0.014)	0.05 ±0.005	0.02 ±0.005
4. $N(fd\ jet) > 0$	0.01 ±0.008 (±0.008)	0.01 ±0.007 (±0.007)	0.01 ±0.003	0.01 ±0.003
5. Classifier (350)	-0.02 ±0.017 (±0.017)	-0.03 ±0.012 (±0.009)	-0.01 ±0.005	-0.01 ±0.003
6. Classifier (400)	0.15 ±0.031 (±0.035)	0.07 ±0.023 (±0.021)	0.05 ±0.011	-0.01 ±0.008
7. Classifier (450)	-0.09 ±0.041 (±0.048)	-0.09 ±0.033 (±0.032)	-0.05 ±0.016	-0.03 ±0.011
8. Classifier (500)	0.04 ±0.031 (±0.031)	0.04 ±0.022 (±0.017)	0.02 ±0.011	0.00 ±0.007
9. Classifier (550)	0.14 ±0.033 (±0.040)	0.14 ±0.026 (±0.029)	0.27 ±0.017	0.21 ±0.013
10. Classifier (600)	-0.04 ±0.020 (±0.022)	-0.05 ±0.017 (±0.018)	-0.01 ±0.008	-0.02 ±0.006
11. Classifier (650)	0.03 ±0.043 (±0.054)	0.25 ±0.038 (±0.049)	-0.06 ±0.022	0.17 ±0.021
12. Classifier (700)	-0.13 ±0.051 (±0.061)	-0.22 ±0.045 (±0.052)	-0.14 ±0.024	-0.17 ±0.023
13. Classifier (750)	0.04 ±0.054 (±0.066)	-0.12 ±0.052 (±0.064)	0.11 ±0.026	-0.06 ±0.027
14. Classifier (800)	0.18 ±0.050 (±0.064)	0.23 ±0.048 (±0.064)	0.13 ±0.025	0.15 ±0.027
15. Classifier (850)	0.02 ±0.057 (±0.070)	0.04 ±0.056 (±0.074)	0.02 ±0.027	0.00 ±0.028
16. Classifier (900)	-0.03 ±0.029 (±0.033)	-0.04 ±0.026 (±0.029)	-0.03 ±0.013	-0.08 ±0.011
17. Classifier (950)	0.13 ±0.042 (±0.049)	0.26 ±0.042 (±0.054)	0.19 ±0.019	0.39 ±0.021
18. Classifier (1000)	-0.01 ±0.046 (±0.055)	-0.09 ±0.045 (±0.056)	-0.11 ±0.020	-0.19 ±0.021
19. Classifier (1050)	0.01 ±0.050 (±0.053)	-0.03 ±0.048 (±0.052)	0.02 ±0.020	0.06 ±0.018
20. Classifier (1100)	0.07 ±0.041 (±0.044)	0.12 ±0.043 (±0.050)	0.13 ±0.017	0.12 ±0.018
$\hat{\sigma}$	0.62	0.58	n/a	n/a
Breusch-Pagan (p-value)	388 (< 0.01)	523 (< 0.01)	n/a	n/a
$\bar{R}^2_{train,unwtd.}$	0.62	0.59	0.82	0.81
$\bar{R}^2_{valid,unwtd.}$	0.61	0.60	0.61	0.59
$\bar{R}^2_{train,wtd.}$	0.06	-0.02	0.58	0.59
$\bar{R}^2_{valid,wtd.}$	0.12	0.09	0.10	0.05
Unscaled $\boldsymbol{\beta}$ estimates:				
constant	208 ±17.1 (±18.8)	99 ±17.0 (±21.8)	199 ±7.6	114 ±7.6
1. $H_T(jets)$ GeV/c	0.11 ±0.029 (±0.030)	0.01 ±0.033 (±0.038)	0.06 ±0.013	-0.09 ±0.01
2. $H_T(jets + Z)$ GeV/c	0.17 ±0.027 (±0.028)	0.30 ±0.030 (±0.036)	0.19 ±0.012	0.35 ±0.01
3. $m(Zb)$ GeV/c^2	0.04 ±0.014 (±0.016)	0.01 ±0.014 (±0.017)	0.06 ±0.006	0.03 ±0.01
4. $N(fd\ jet) > 0$ [T or F]	13 ±10.1 (±10.3)	17 ±9.4 (±9.4)	13 ±4.3	8 ±3.7
5. Classifier (350)	-138 ±118.1 (±115.6)	-235 ±81.5 (±60.6)	-71 ±34.7	-105 ±20.5
6. Classifier (400)	793 ±166.7 (±183.6)	379 ±124.0 (±111.7)	260 ±60.3	-64 ±40.1
7. Classifier (450)	-407 ±184.3 (±213.2)	-407 ±147.1 (±144.1)	-238 ±71.9	-137 ±48.6
8. Classifier (500)	99 ±76.1 (±75.5)	94 ±52.9 (±41.7)	39 ±26.6	0 ±17.6
9. Classifier (550)	84 ±19.7 (±24.3)	87 ±15.6 (±17.7)	163 ±10.0	127 ±8.1
10. Classifier (600)	-76 ±36.9 (±39.7)	-92 ±30.8 (±32.1)	-17 ±14.6	-40 ±10.1
11. Classifier (650)	14 ±23.5 (±29.7)	139 ±20.6 (±26.7)	-32 ±12.0	95 ±11.6
12. Classifier (700)	-87 ±34.4 (±41.0)	-147 ±30.1 (±35.1)	-95 ±16.1	-113 ±15.1
13. Classifier (750)	25 ±32.2 (±38.8)	-70 ±30.5 (±38.0)	64 ±15.3	-38 ±15.7
14. Classifier (800)	111 ±29.9 (±38.7)	136 ±28.7 (±38.4)	81 ±15.3	89 ±16.0
15. Classifier (850)	10 ±34.3 (±42.0)	23 ±33.7 (±44.5)	11 ±16.2	2 ±17.0
16. Classifier (900)	-39 ±36.1 (±40.6)	-52 ±32.0 (±35.1)	-43 ±15.4	-95 ±13.5
17. Classifier (950)	60 ±20.3 (±23.6)	125 ±20.3 (±26.1)	92 ±9.3	187 ±9.9
18. Classifier (1000)	-6 ±26.4 (±31.2)	-51 ±25.7 (±31.9)	-63 ±11.6	-107 ±11.9
19. Classifier (1050)	12 ±74.4 (±79.3)	-40 ±71.6 (±78.6)	36 ±29.8	90 ±26.4
20. Classifier (1100)	52 ±30.0 (±32.0)	85 ±31.0 (±36.2)	91 ±12.3	86 ±13.0

Table 5.13: Linear regression parameter estimates for the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ (without quadratic terms) with four alternate models of the disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{model})$ for the T_{pair} signal type. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]), and the Breusch-Pagan test for heteroscedasticity is provided [179] along with adjusted R^2 goodness of fit statistics.

independent covariate $\mathbf{X}$	$\boldsymbol{\beta}_{OLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\boldsymbol{\beta}_{WLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\boldsymbol{\beta}_{GLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2)$	$\boldsymbol{\beta}_{WGLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
constant	-0.08 ±0.270 (±0.329)	-0.54 ±0.255 (±0.333)	0.33 ±0.118	0.26 ±0.114
1. $H_T(jets)$	0.41 ±0.252 (±0.333)	-0.11 ±0.249 (±0.335)	0.59 ±0.111	0.16 ±0.114
2. $H_T(jets + Z)$	0.70 ±0.301 (±0.396)	1.32 ±0.295 (±0.390)	0.28 ±0.133	0.66 ±0.134
3. $m(Zb)$	0.01 ±0.059 (±0.071)	0.01 ±0.057 (±0.075)	0.01 ±0.025	-0.01 ±0.024
4. $N(fd\ jet) > 0$	-0.04 ±0.048 (±0.052)	0.02 ±0.052 (±0.056)	0.01 ±0.021	0.02 ±0.022
5. Classifier (350)	-0.05 ±0.117 (±0.126)	-0.25 ±0.111 (±0.135)	-0.14 ±0.041	-0.25 ±0.038
6. Classifier (400)	-0.44 ±0.233 (±0.265)	-0.06 ±0.220 (±0.304)	-0.35 ±0.097	0.05 ±0.095
7. Classifier (450)	-0.25 ±0.306 (±0.381)	-0.71 ±0.307 (±0.406)	0.17 ±0.127	-0.43 ±0.125
8. Classifier (500)	-0.57 ±0.317 (±0.368)	-1.07 ±0.293 (±0.363)	-0.39 ±0.110	-0.43 ±0.097
9. Classifier (550)	0.77 ±0.263 (±0.363)	0.61 ±0.253 (±0.440)	0.25 ±0.146	0.21 ±0.147
10. Classifier (600)	0.03 ±0.163 (±0.194)	-0.30 ±0.160 (±0.213)	-0.02 ±0.064	-0.38 ±0.062
11. Classifier (650)	0.03 ±0.343 (±0.492)	0.31 ±0.333 (±0.543)	0.07 ±0.188	0.36 ±0.196
12. Classifier (700)	-1.04 ±0.411 (±0.516)	-0.85 ±0.408 (±0.569)	-0.91 ±0.182	-0.65 ±0.188
13. Classifier (750)	0.30 ±0.429 (±0.562)	0.09 ±0.424 (±0.619)	0.08 ±0.206	-0.43 ±0.215
14. Classifier (800)	0.61 ±0.371 (±0.545)	1.61 ±0.364 (±0.590)	0.60 ±0.193	1.02 ±0.211
15. Classifier (850)	0.07 ±0.437 (±0.575)	0.50 ±0.440 (±0.699)	0.06 ±0.210	0.03 ±0.228
16. Classifier (900)	-0.21 ±0.210 (±0.239)	-0.42 ±0.207 (±0.258)	-0.33 ±0.087	-0.22 ±0.087
17. Classifier (950)	0.08 ±0.356 (±0.454)	-0.45 ±0.355 (±0.545)	0.45 ±0.166	0.46 ±0.178
18. Classifier (1000)	-0.29 ±0.355 (±0.464)	-0.88 ±0.359 (±0.523)	-0.07 ±0.157	-1.12 ±0.167
19. Classifier (1050)	0.22 ±0.297 (±0.328)	0.79 ±0.295 (±0.368)	0.02 ±0.126	0.80 ±0.133
20. Classifier (1100)	1.15 ±0.286 (±0.331)	1.33 ±0.288 (±0.377)	0.86 ±0.119	1.28 ±0.128
$\hat{\sigma}$	0.59	0.53	n/a	n/a
Breusch-Pagan (p-value)	802 (< 0.01)	1165 (< 0.01)	n/a	n/a
$\bar{R}^2_{train,unwtd.}$	0.65	0.61	0.92	0.91
$\bar{R}^2_{valid,unwtd.}$	0.62	0.59	0.60	0.57
$\bar{R}^2_{train,wtd.}$	0.11	0.03	0.81	0.80
$\bar{R}^2_{valid,wtd.}$	0.04	-0.03	-0.00	-0.07
Unscaled $\boldsymbol{\beta}$ estimates:				
constant	-27 ±96.4 (±117.2)	-193 ±90.9 (±118.8)	116 ±41.9	93 ±40.7
1. $H_T(jets)$ GeV/c	0.35 ±0.214 (±0.282)	-0.09 ±0.212 (±0.284)	0.50 ±0.094	0.13 ±0.10
2. $H_T(jets + Z)$ GeV/c	0.49 ±0.210 (±0.275)	0.92 ±0.205 (±0.271)	0.19 ±0.092	0.46 ±0.09
3. $m(Zb)$ GeV/c^2	0.01 ±0.075 (±0.090)	0.01 ±0.073 (±0.096)	0.01 ±0.032	-0.01 ±0.03
4. $N(fd\ jet) > 0$ [T or F]	-48 ±61.5 (±66.2)	30 ±66.0 (±71.5)	7 ±26.7	22 ±28.6
5. Classifier (350)	-380 ±817.0 (±881.6)	-1751 ±778.8 (±943.6)	-992 ±289.5	-1751 ±268.4
6. Classifier (400)	-2333 ±1240.0 (±1408.5)	-315 ±1170.4 (±1617.5)	-1882 ±517.3	254 ±503.3
7. Classifier (450)	-1101 ±1367.9 (±1701.3)	-3182 ±1370.6 (±1815.2)	765 ±566.4	-1919 ±559.0
8. Classifier (500)	-1399 ±774.0 (±898.0)	-2617 ±714.8 (±885.8)	-961 ±269.0	-1054 ±237.2
9. Classifier (550)	465 ±158.0 (±218.3)	365 ±152.3 (±264.2)	151 ±87.5	129 ±88.6
10. Classifier (600)	62 ±294.0 (±349.0)	-545 ±287.3 (±383.2)	-35 ±114.5	-678 ±111.0
11. Classifier (650)	14 ±188.1 (±269.7)	170 ±182.7 (±297.6)	40 ±102.8	199 ±107.4
12. Classifier (700)	-697 ±276.0 (±346.5)	-569 ±274.1 (±381.5)	-607 ±122.4	-437 ±126.3
13. Classifier (750)	179 ±253.5 (±332.3)	52 ±250.6 (±365.8)	46 ±121.5	-253 ±126.8
14. Classifier (800)	371 ±223.6 (±328.5)	970 ±219.1 (±355.6)	360 ±116.5	614 ±127.0
15. Classifier (850)	41 ±263.4 (±346.5)	303 ±265.3 (±421.3)	39 ±126.8	20 ±137.4
16. Classifier (900)	-255 ±257.6 (±293.1)	-515 ±254.5 (±316.3)	-408 ±106.2	-276 ±106.5
17. Classifier (950)	39 ±171.1 (±218.4)	-216 ±171.0 (±262.3)	217 ±80.0	222 ±85.6
18. Classifier (1000)	-164 ±203.1 (±265.3)	-503 ±205.5 (±299.0)	-41 ±89.7	-641 ±95.7
19. Classifier (1050)	324 ±445.1 (±492.0)	1185 ±442.5 (±550.8)	33 ±189.0	1203 ±199.3
20. Classifier (1100)	838 ±207.8 (±240.0)	967 ±209.2 (±274.0)	624 ±86.1	928 ±92.8

Table 5.14: Linear regression parameter estimates for the regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{XBX} + \mathbf{u}$ (with quadratic terms) with four alternate models of the disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{model})$ for the T_{pair} signal type. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]), and the Breusch-Pagan test for heteroscedasticity is provided [179] along with adjusted R^2 goodness of fit statistics.

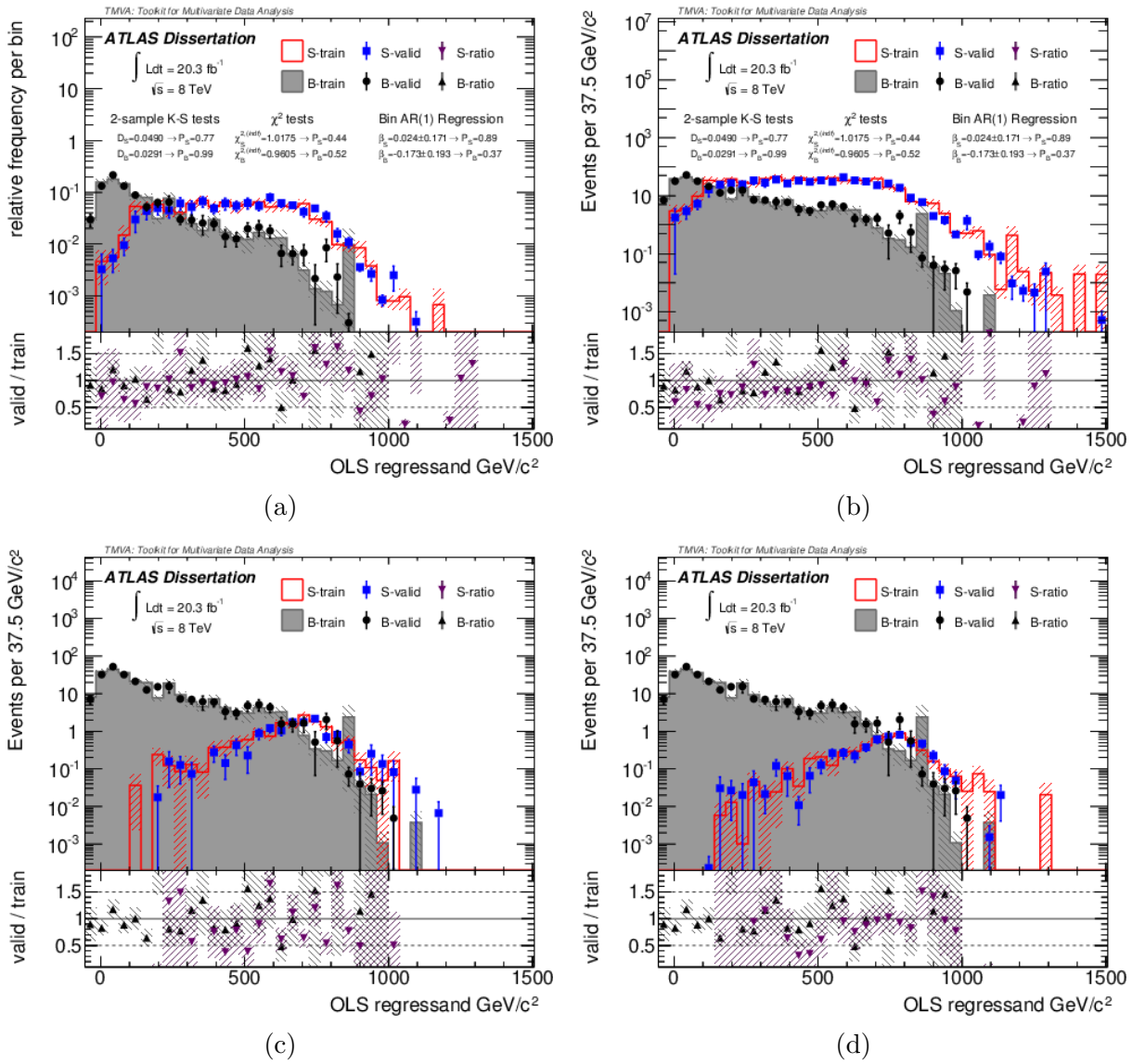


Figure 5.29: Stage 3 OLS regression spectra for the T_{pair} signals (before cut on stage 2 classifier response value). (a) The distributions for all signal masses combined. The predicted yields for (b) all signal masses combined, (c) the $m(T_{pair}) = 650$ GeV/c^2 signal mass, and (d) the $m(T_{pair}) = 750$ GeV/c^2 signal mass.

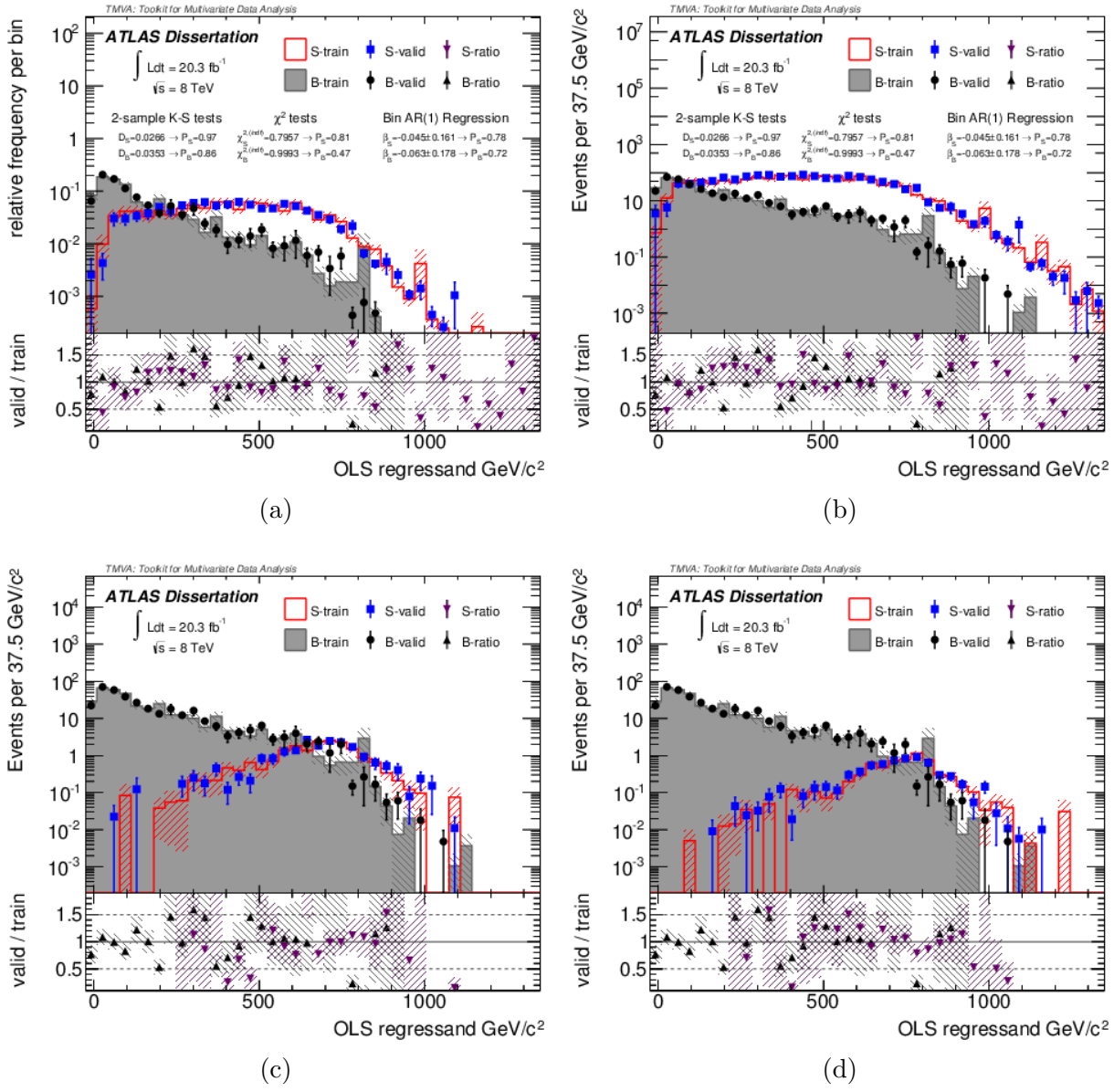


Figure 5.30: Stage 3 OLS regression spectra for the B_{pair} signals (before cut on stage 2 classifier response value). (a) The distributions for all signal masses combined. The predicted yields for (b) all signal masses combined, (c) the $m(B_{pair}) = 650$ GeV/c^2 signal mass, and (d) the $m(B_{pair}) = 750$ GeV/c^2 signal mass.

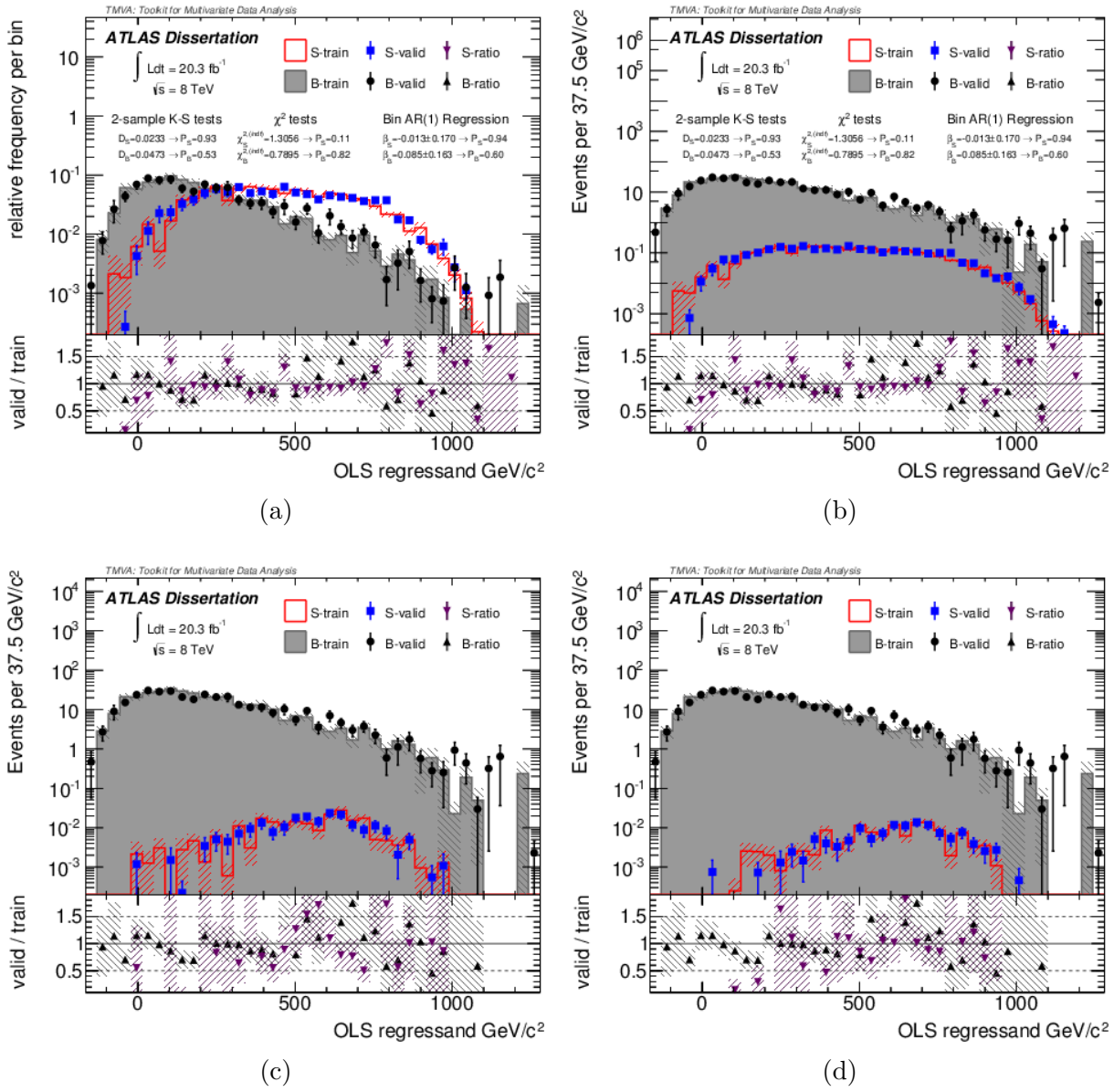


Figure 5.31: Stage 3 OLS regression spectra for the B_{single} signals (before cut on stage 2 classifier response value). (a) The distributions for all signal masses combined. The predicted yields for (b) all signal masses combined, (c) the $m(B_{single}) = 650$ GeV/c^2 signal mass, and (d) the $m(B_{single}) = 750$ GeV/c^2 signal mass.

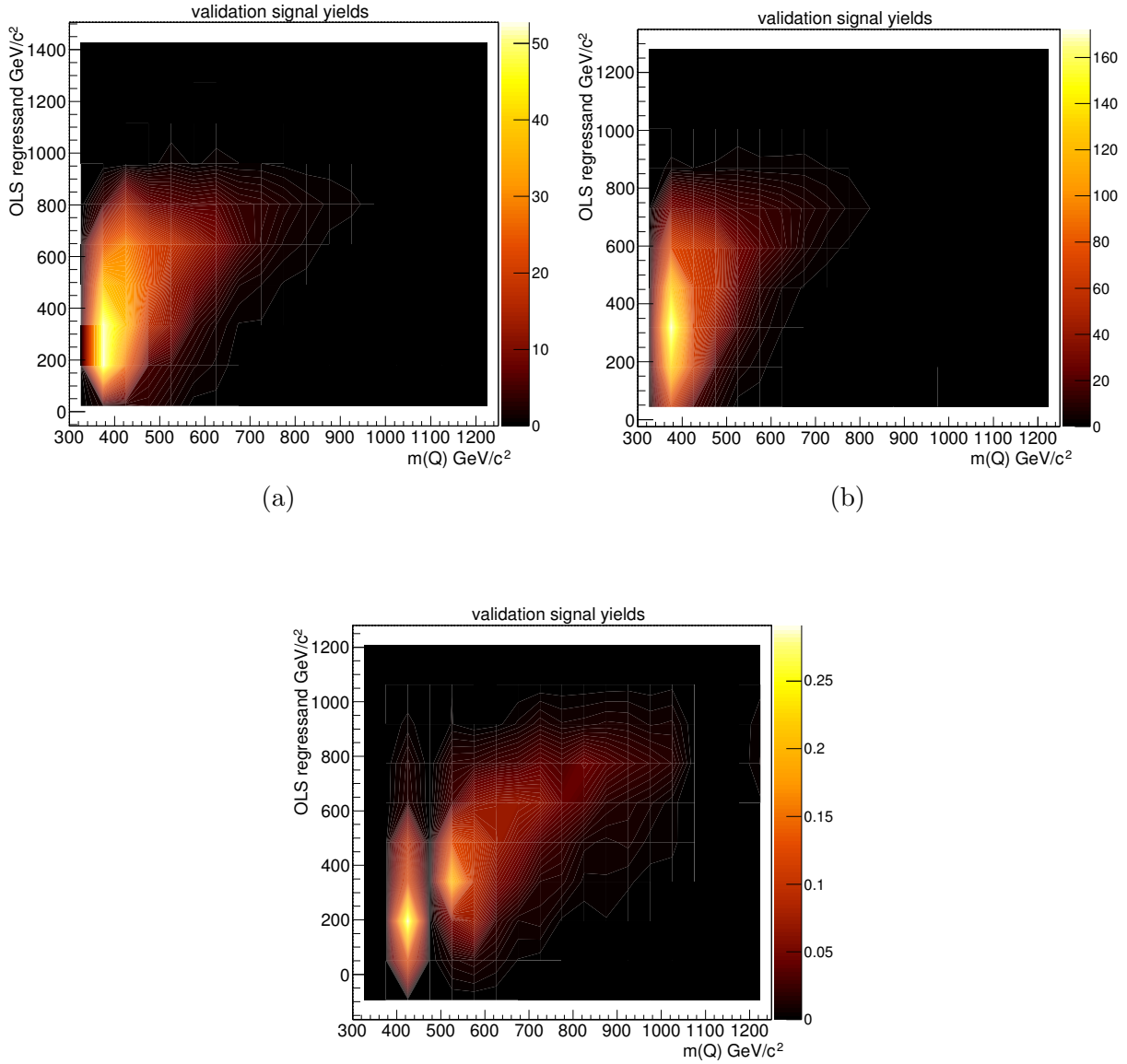


Figure 5.32: Heat maps showing the performance of the stage 3 OLS regression before cutting on the stage 2 classifier response value. The x-axis is the actual (theoretical) signal mass and the y-axis is the OLS regressand prediction of this mass, where perfect prediction would be represented by the diagonal line $y = x$. (a) T_{pair} signals, (b) B_{pair} signals, and (c) B_{single} signals.

5.7.2. Regression with Boosted Decision Trees *(selected for T_{single})*

BDT regression with gradient boosting also provided good regressand predictions of $m(Q)$, though usually not as accurately as linear regression. For the T_{single} signal type, BDT regression provided slightly better simulation performance and so was chosen as the method for stage 3 regression for T_{single}. This BDT used a learning rate (shrinkage) of 0.1, 500 trees for the forest, 50% bagging of training data for each tree, a maximum depth of 3 for each tree, and two randomly chosen variables for each decision node.

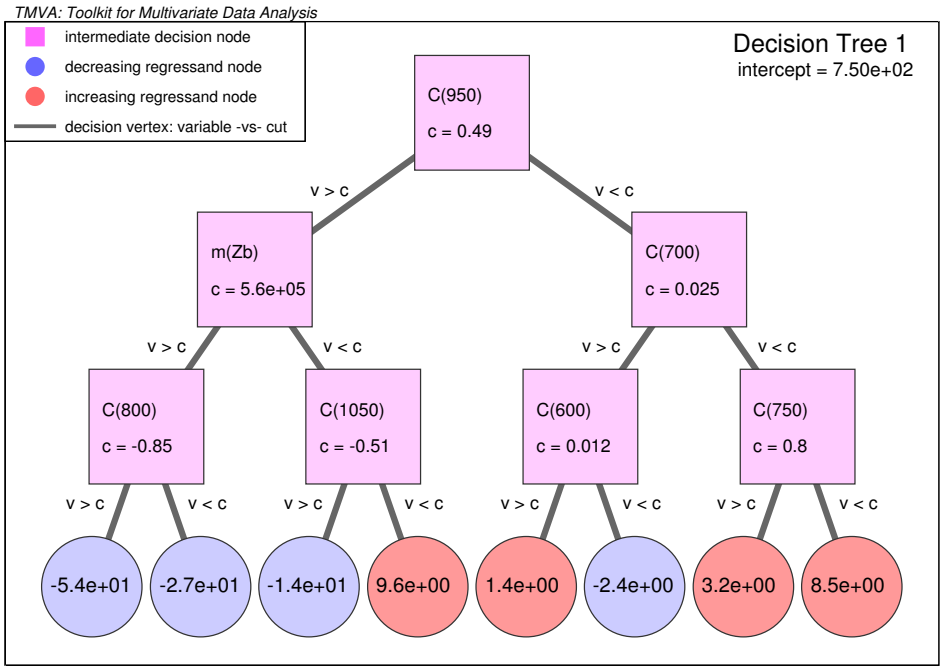
Figure 5.33(a) shows the first regression tree of the random forest as an example. The majority of the variables used in the decision nodes for this tree were the middle to high signal mass stage 1 classifiers. The leaf nodes gave regression values between -54 to 14 GeV/c^2 and had an intercept value of 750 GeV/c^2 so that this particular tree returned regressand values between 696 and 764 GeV/c^2. This range did not span the full ∼0 to ∼1050 GeV/c^2 range of the background and signal masses, and was thus a weak learner, needing to be combined with many such trees if the full range was to be estimated.

Figure 5.33(b) shows the correlations between the input variables. The $H_T(jets)$ variable was highly correlated with $H_T(jets + Z)$, as were the many of the higher signal mass stage 1 classifiers. However, instead of removing variables to reduce complexity, the use of randomized forests of shallow regression trees allowed for more even sampling of these variables. This is shown in figure 5.33(c) where, for instance, $H_T(jets)$ was used in 5.8% of tree nodes and $H_T(jets + Z)$ in 5.7%. The frequency of use in decision nodes of the stage 1 classifiers ranged from 4.9% to 6.6%. Figure 5.33(d) shows that a large degree of heterogeneous pairing was obtained between the variables.

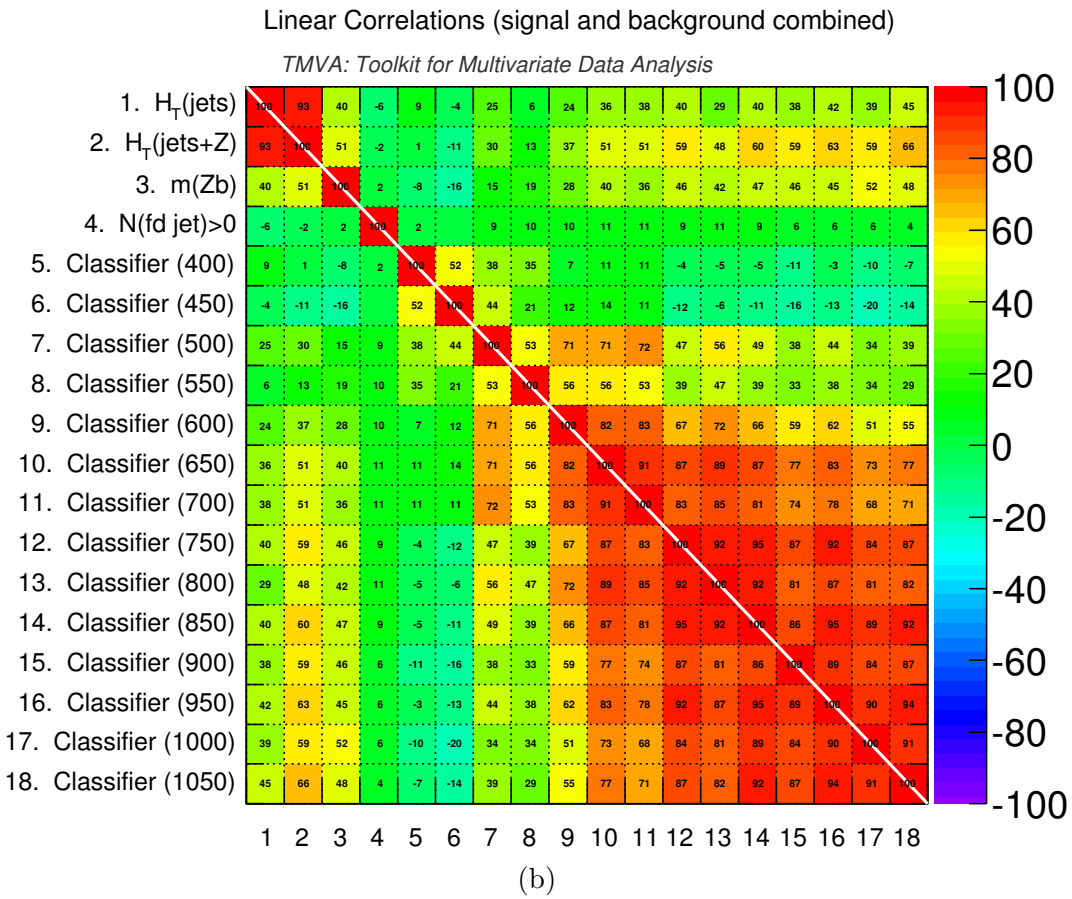
Figures 5.34(a)–(d) show the result of BDT regression for the T_{single} signal type. Figures 5.34(a) and (b) show histogram and predicted yields, respectively, for all the signal masses combined. The statistical tests provided in the figures imply the presence of overtraining for both signal and background distributions. However, the K-S test shown in the figure was binning dependent while the binning independent K-S and A-D tests (not shown) gave p-values of 0.55 and 1.00, respectively, for the background so that in no overtraining of the background was detected. Figures 5.34(c) and (d) show the expected yields of the 650 and 750 GeV/c^2 signal masses, respectively. The regressand masses for these signals peaked at values near their signal mass values.

Figure 5.35 compared the signal mass, $m(Q)$, to the predicted regressand signal mass. As with OLS, the BDT mass regressand performed better at higher signal masses.

The results of using the gradient BDTs to provide the regression mass spectrum for statistical inference of the T_{single} signal type are described in section 5.9. This section demonstrated that gradient BDTs used for regression may provide an alternative to the least squares linear regression models that were examined in the previous section.

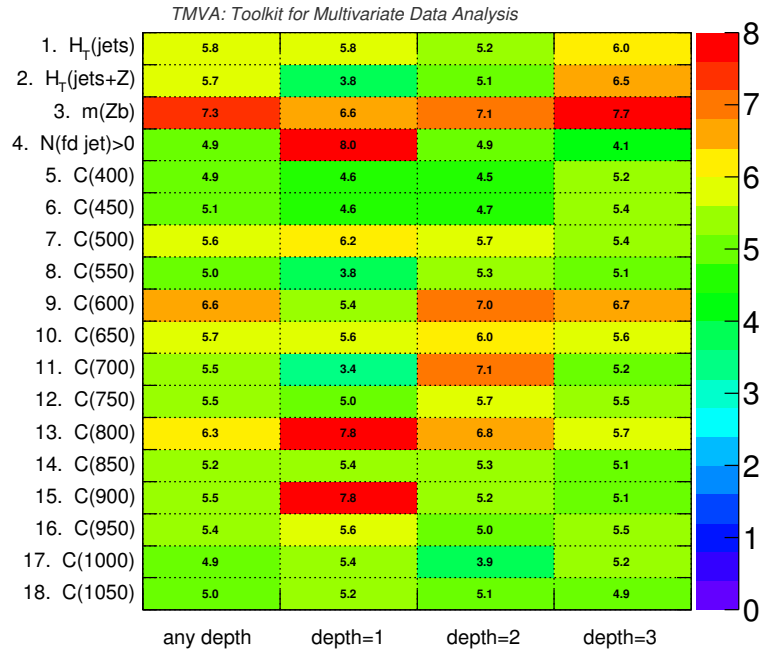


(a)

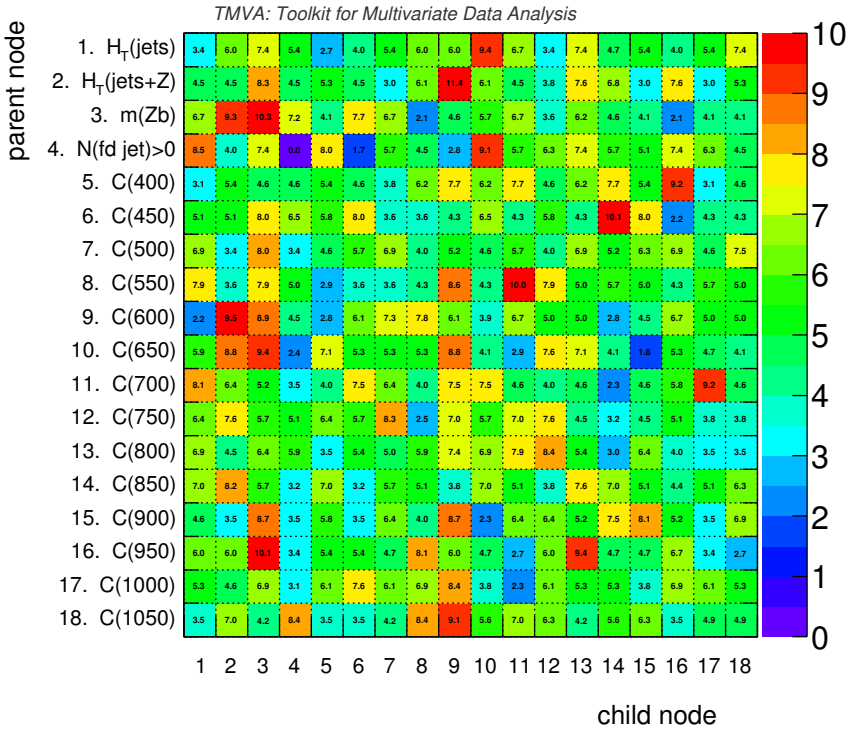


(b)

Figure 5.33: – *Continues on next page.*



(c)



(d)

Figure 5.33: Implementation of stage 3 BDT regression for the T_{single} signal type before cutting on the stage 2 classifier response value: (a) the resultant first tree in the BDT random forest, (b) correlations between the input variables, (c) frequency of variables used in decision nodes, and (d) variable pairings frequencies of the parent–child decision nodes (rows add to 100%).

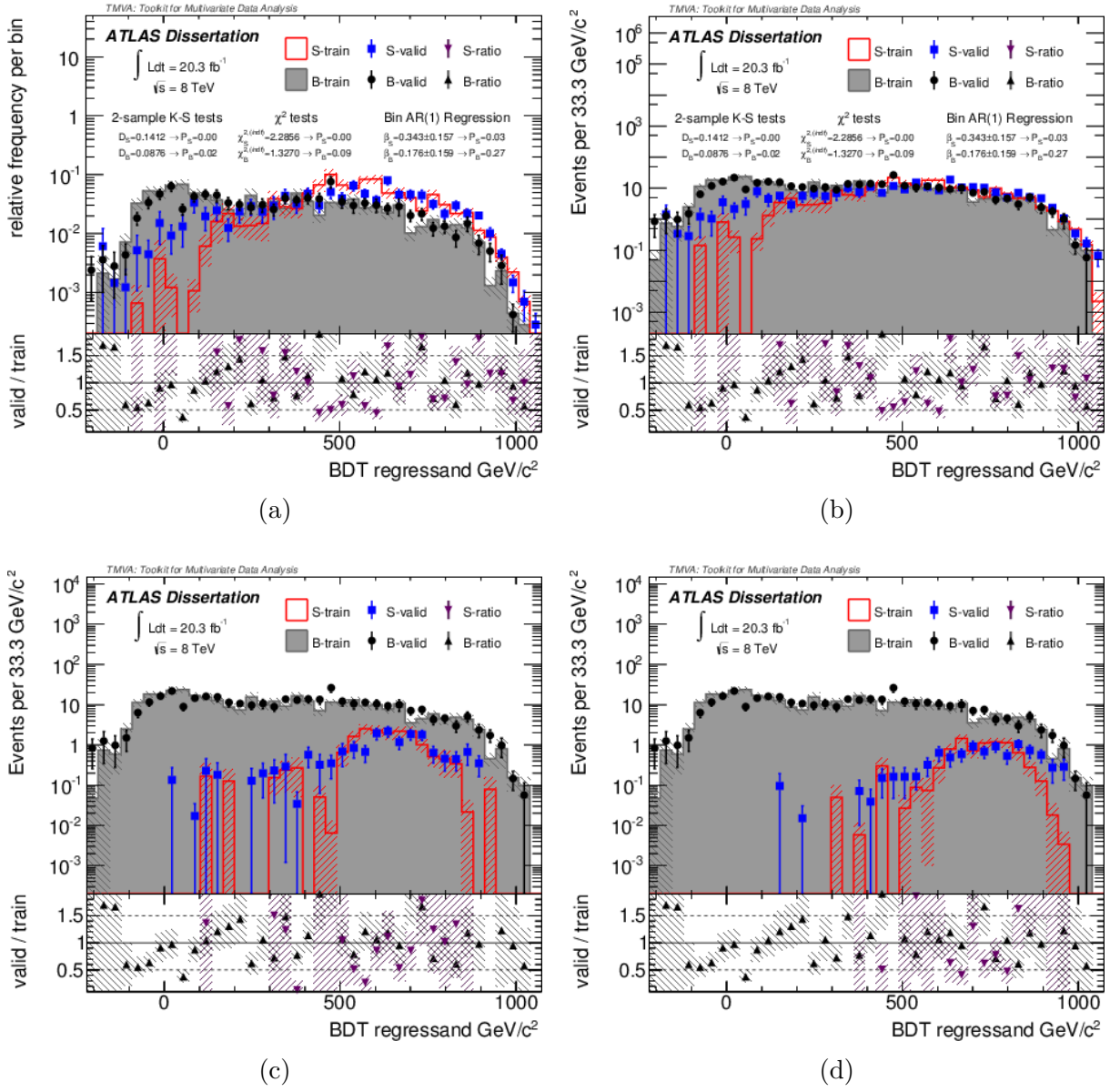


Figure 5.34: Stage 3 BDT regression spectra for the T_{single} signal type before cutting on the stage 2 classifier response value. (a) The histogram for all signal masses combined. The predicted yields of (b) all signal masses combined, (c) the $m(T_{single}) = 650$ GeV/c^2 signal, and (d) the $m(T_{single}) = 750$ GeV/c^2 signal.

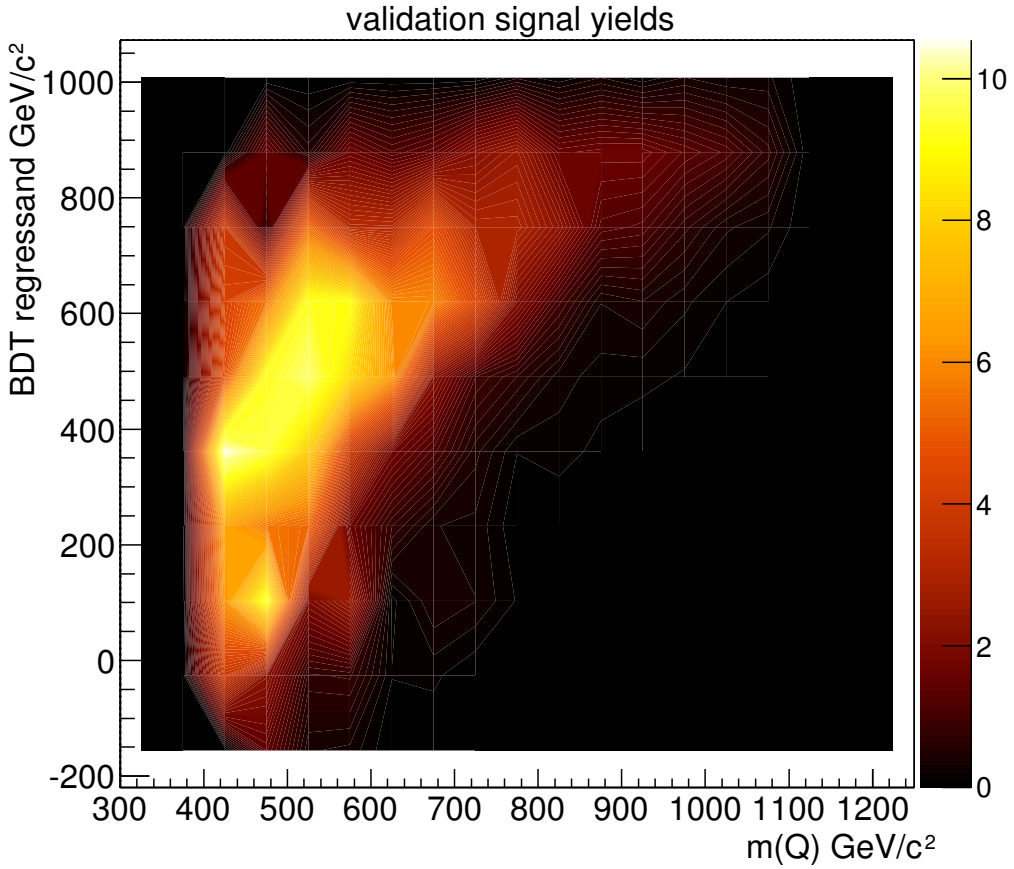


Figure 5.35: A heat map showing the predicted mass of stage 3 BDT regression vs. signal mass $m(Q)$ for the T_{single} signal type before cutting on the stage 2 classifier response value.

5.8. Modeling Validation in Control Regions

This section documents the validation of the modeling quality of the simulated background data that was accomplished by comparing the expected yields of the simulated Standard Model backgrounds to those actually measured in control regions [14]. This was done to demonstrate the accuracy of the simulated backgrounds before examining the predicted and measured yields in the final signal regions that were used for the final statistical inference (estimating upper limits on the strength of the signals). All of the estimated systematic uncertainties (described in chapter 4) are shown in these validation histograms. Significant deviation of the measured yields outside the range of systematic uncertainties of the expected yields may imply possible mismodeling by the simulated data.

First, modeling validation is shown for the stage 1 input variables in the preselection region that was used for training the learning algorithms, and is described in table 5.5 that is repeated below for convenience. Then modeling validation is shown for the resultant stage 1 BDT classifiers. This is followed by modeling validation of the stage 2 logit classifiers, and lastly the validation of the stage 3 regressions [15].

A large number of such validation histograms in various control regions were examined, and these indicated good modeling of the measured data by the simulations. Only the most pertinent of these histograms are provided here [16].

preselection requirements		description
1	preliminary requirements	Passes triggers, object selection, etc. See section 3.4.6.
2	$Z = l_1 + l_2$	Same flavor opposite sign electrons or muons with 81.19 GeV/c^2 $< m(l_1 l_2) <$ 91.19 GeV/c^2.
3	$N(leptons) = 2$	Exactly two leptons (the leptonic Z boson candidate).
4	$p_T(Z) > 100$ GeV/c	Excluding low transverse momentum Z bosons that are used in correcting the b-tagged jet multiplicity.
5	$N(b\text{-}tagged\ jets) \geq 2$	Require at least two b-tagged jets.
6	$N(anti\text{-}b\text{-}tagged\ jets) \geq 1$	Require at least one anti-b-tagged jet.

Table 5.5: Stage 1 classification data preselection. *(Repeated from page 129.)*

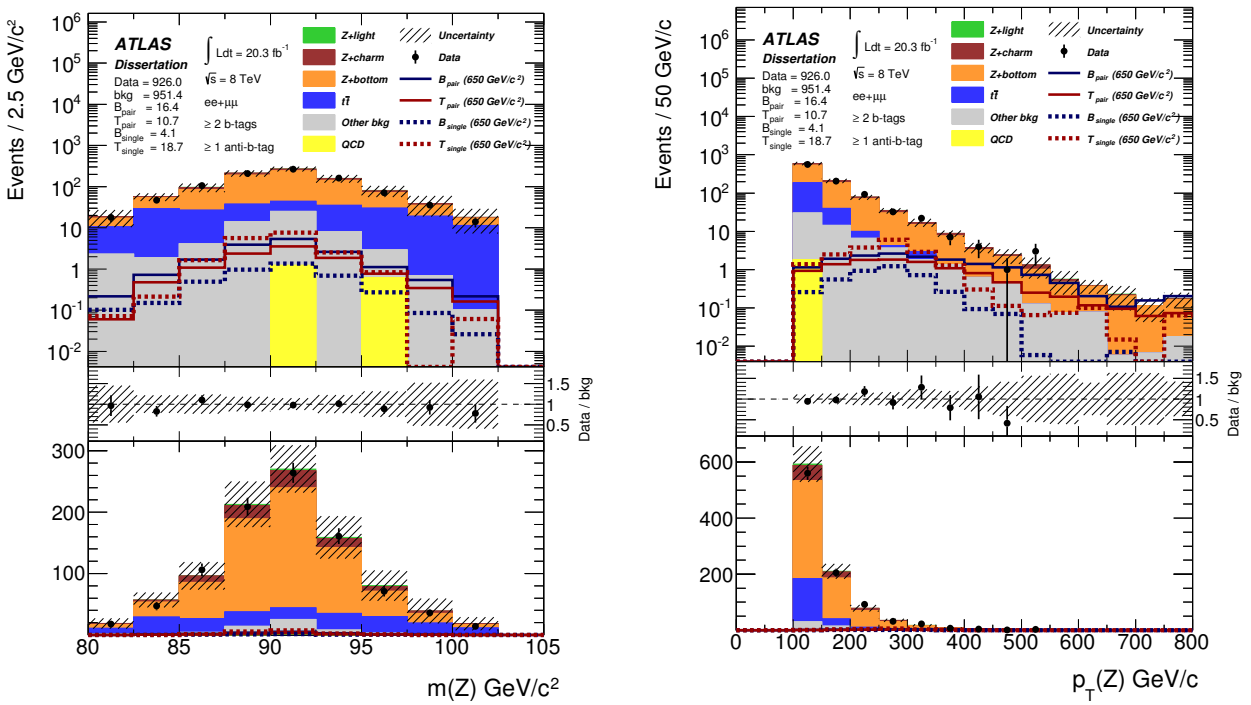
[14] A *control region* describes a set of input variable requirements such that the simulated data passing the requirements is expected to have insignificant expected signal yield in comparison to the expected background yield. In such a region, the measured data may be compared to the expected background yield in order to establish the modeling accuracy of the simulated background data. A *signal region* is a set of requirements such that the simulated and measured data passing the requirements are used for the final statistical inference.

[15] In addition to showing the expected background yields and measured yields, the expected yields of the 650 GeV/c^2 signal masses of the four signal types (T_{pair}, B_{pair}, T_{single}, and B_{single}) are also shown in the validation histograms for reference.

[16] Other control regions (not shown) involved removing preselection cut 6 shown in table 5.5, and also altering preselection cut 5 to $N(b\text{-}tagged\ jets) = 1$ and to $N(b\text{-}tagged\ jets) = 0$.

5.8.1. Validation of stage 1 Input Variables

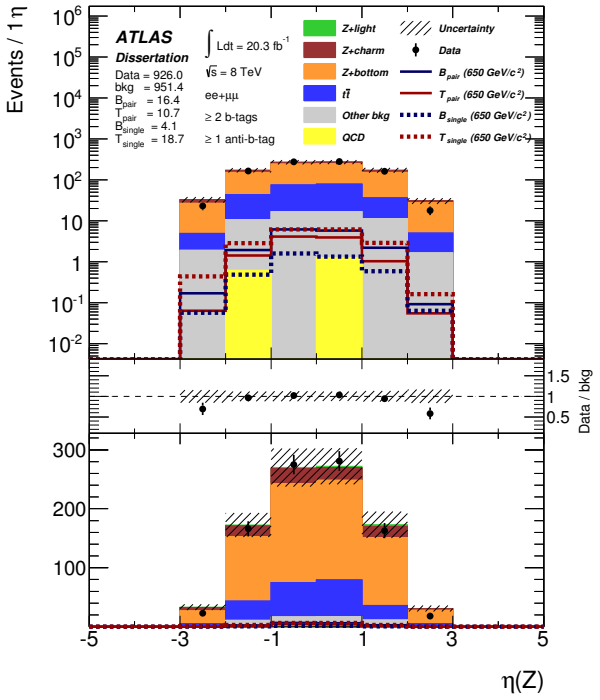
Figures 5.36 (a)–(u) show the validation histograms for the stage 1 input variables and some additional supporting variables in the stage 1 preselection region (defined in table 5.5 on page 129). Figures (a)–(f) show that the spectra of various properties of the reconstructed leptonic Z boson was well modeled with minor discrepancies well covered by the total uncertainties in nearly all cases. Figures (g)–(j) demonstrate good modeling accuracy of the two leading b-tagged jets, and figures (k)–(l) the accurate modeling of the leading anti-b-tagged jet, while figures (m)–(n) show that the combined b-tagged jets and leading anti-*b*-tagged jet variables were well modeled. Figures (o)–(r) show that the transverse jet shape variables were well modeled. Lastly, figures (s)–(u) show good modeling of variables constructed using the Z boson combined with the leading b-tagged jet (figure (s)), or else combined with all jets (figures (20)–(21)).



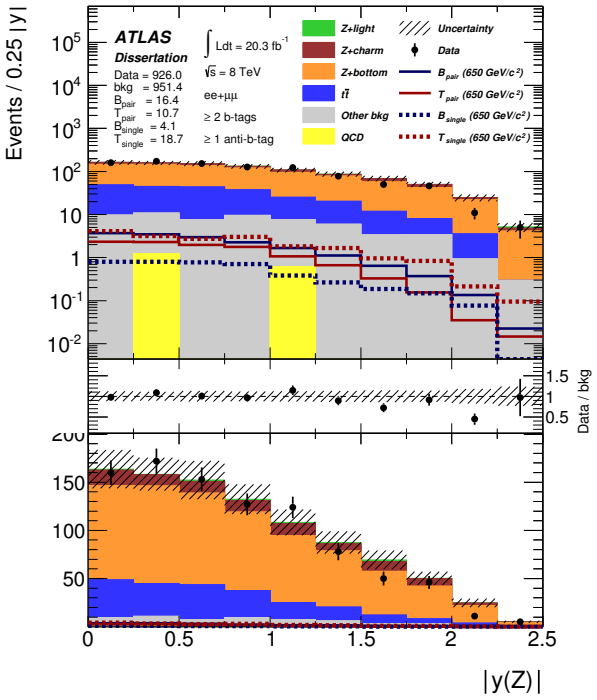
(a) $m(Z)$ (b) $p_T(Z)$

Figure 5.36: Histograms showing the simulated data validation in the preselection region of the input variables (and other supporting variables) used for stage 1 classification. Each subfigure has three components: upper showing predicted and measured yields on a logarithmic scale, middle showing the ratio of measured data to expected background, and lower showing the same as the upper on a linear scale. All simulated data shown is only from the testing data partition. The 650 GeV/c^2 signal mass histograms for each of the four signal types are also shown. Uncertainty shown includes both statistical and all systematic uncertainties. (a) $m(Z)$, (b) $p_T(Z)$.

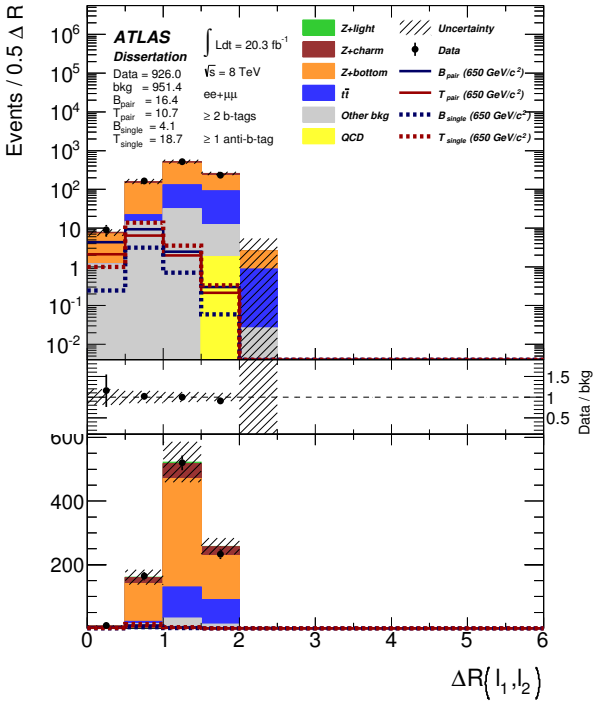
Figure 5.36 Continues on next 5 pages.



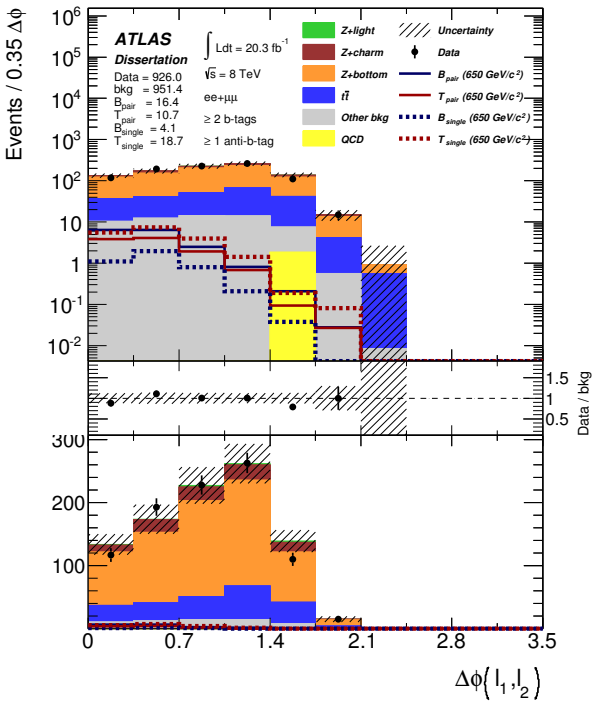
(c) $\eta(Z)$



(d) $|y(Z)|$



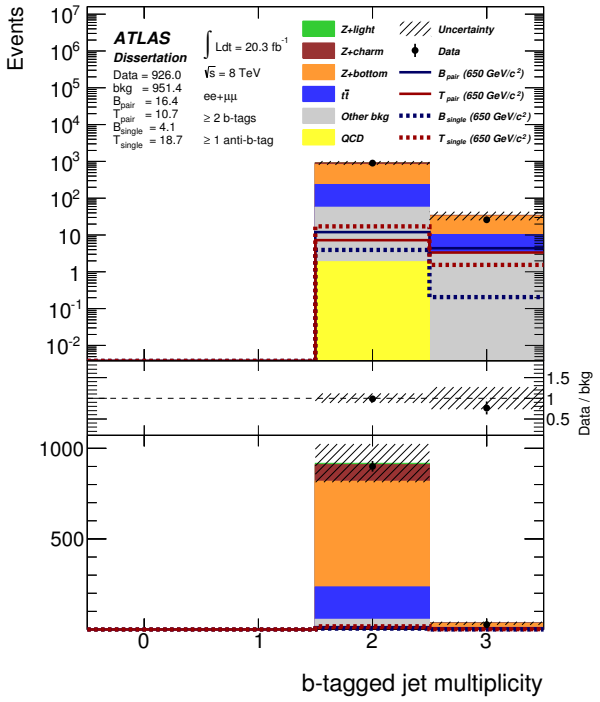
(e) $\Delta R(l_1, l_2)$



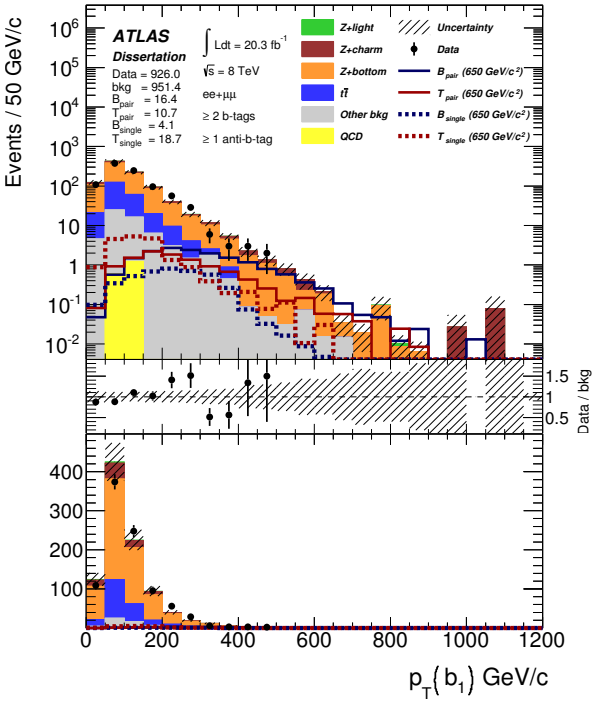
(f) $\Delta\phi(l_1, l_2)$

Figure 5.36: (c) $\eta(Z)$, (d) $|y(Z)|$, (e) $\Delta R(l_1, l_2)$, (f) $\Delta\phi(l_1, l_2)$.

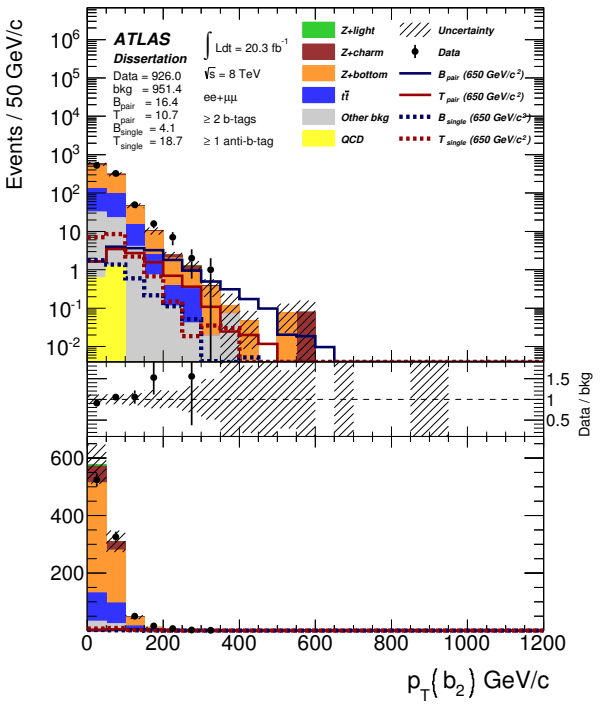
– *Continued, continues on next 4 pages.*



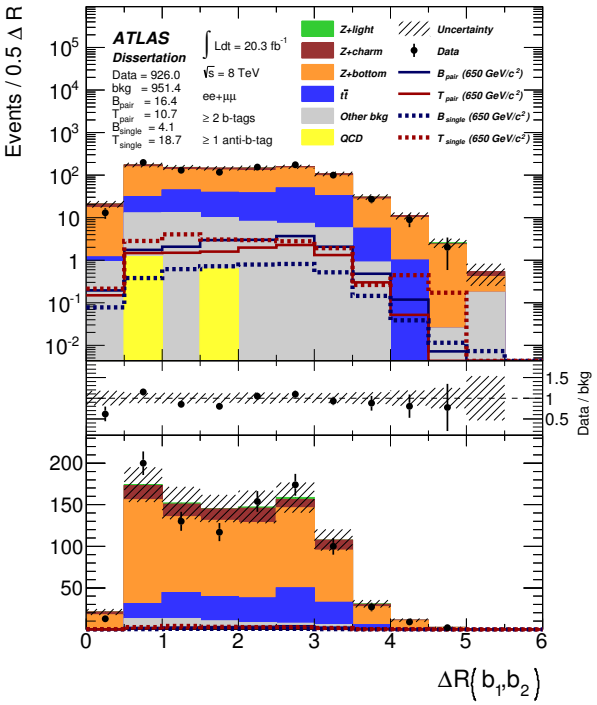
(g) $N(b\text{-}tagged\ jets)$



(h) $p_T(b_1)$



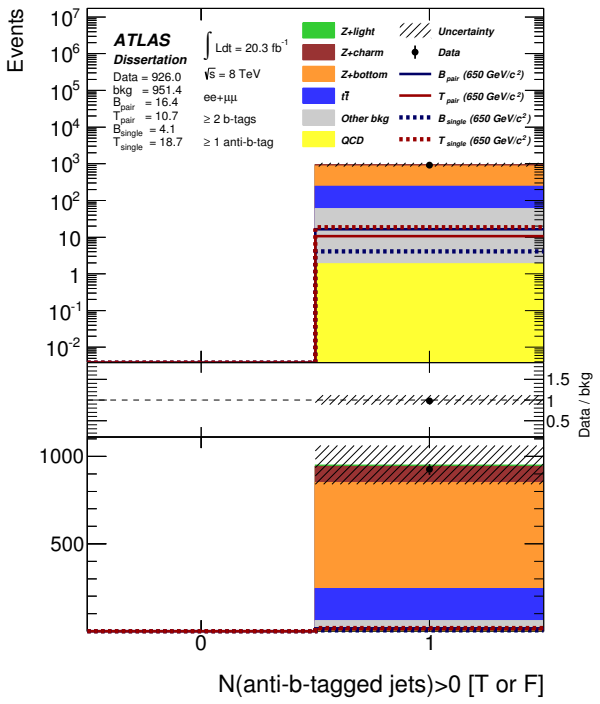
(i) $p_T(b_2)$



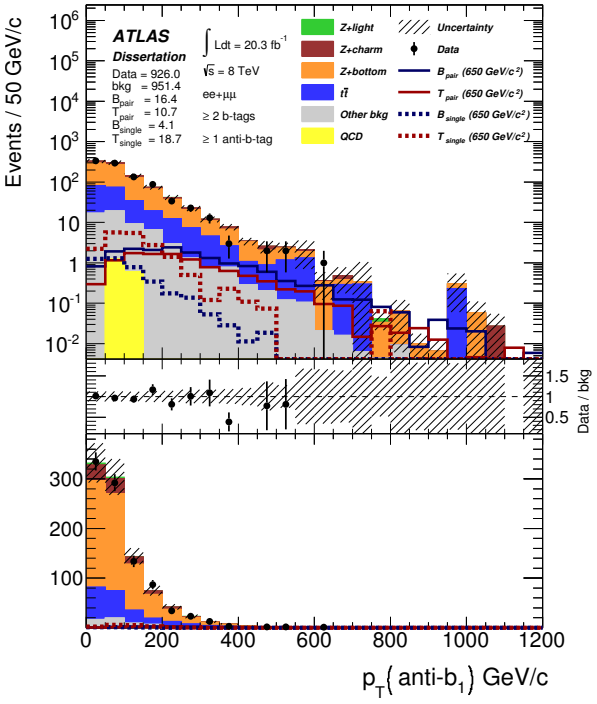
(j) $\Delta R(b_1, b_2)$

Figure 5.36: (g) $N(b\text{-}tagged\ jets)$, (h) $p_T(b_1)$, (i) $p_T(b_2)$, (j) $\Delta R(b_1, b_2)$.

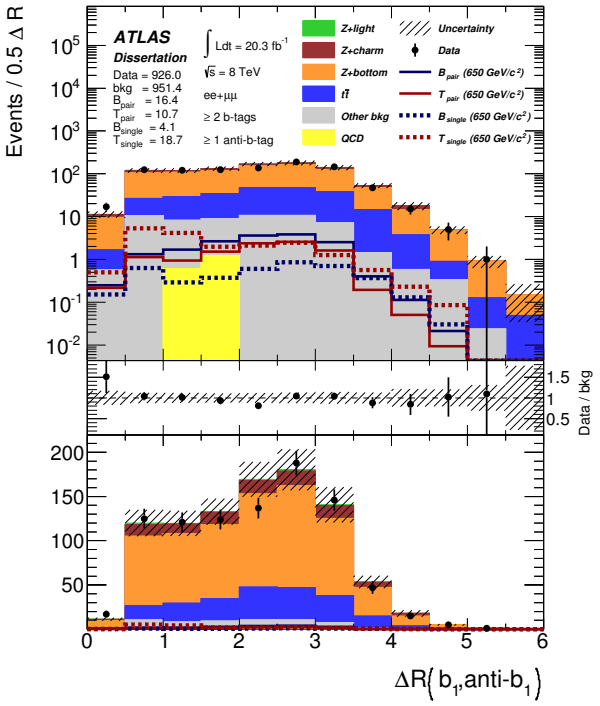
– *Continued, continues on next 3 pages.*



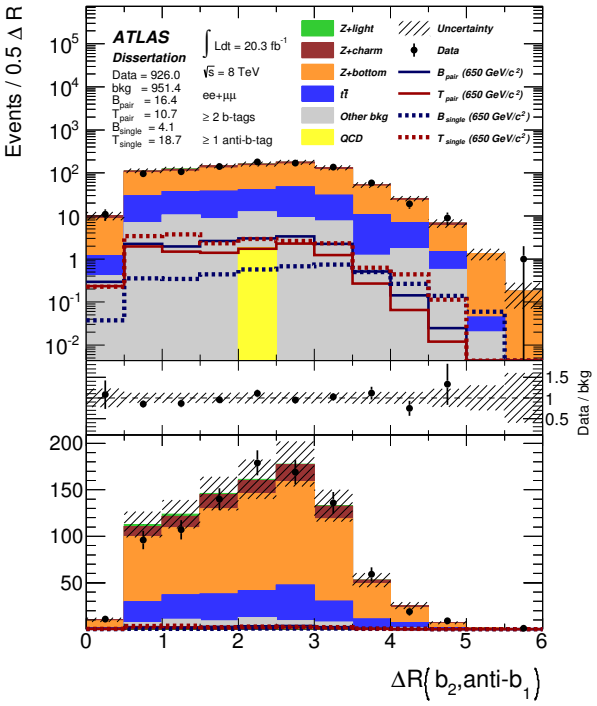
(k) $N(anti\text{-}b\text{-}tagged)$



(l) $p_T(anti\text{-}b_1)$



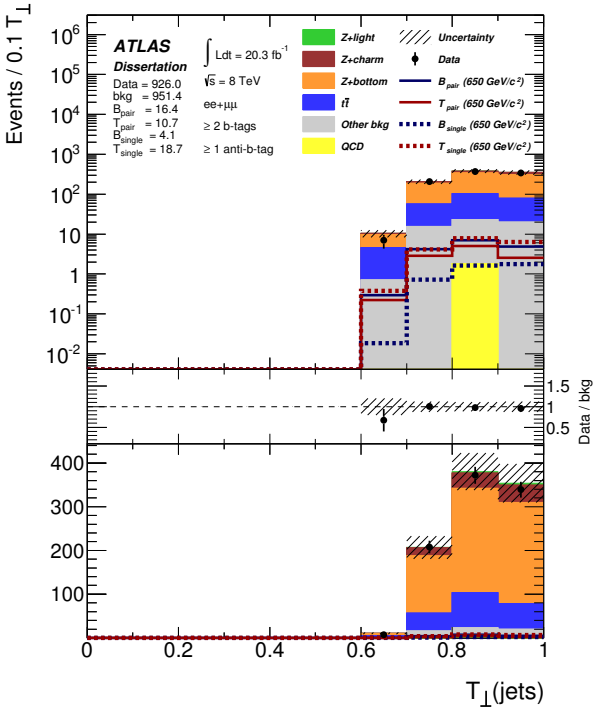
(m) $\Delta R(b_1, anti\text{-}b_1)$



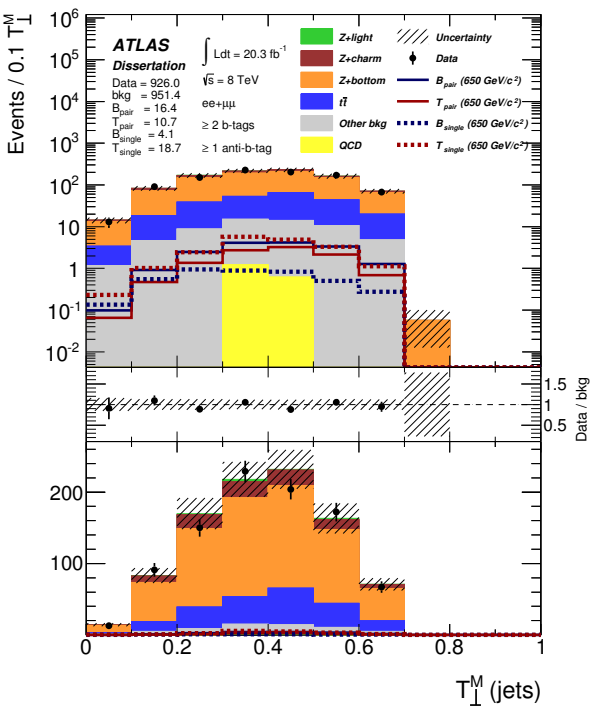
(n) $\Delta R(b_2, anti\text{-}b_1)$

Figure 5.36: (k) $N(anti\text{-}b\text{-}tagged\ jets$, (l) $p_T(anti\text{-}b_1)$, (m) $\Delta R(b_1, anti\text{-}b_1)$, (n) $\Delta R(b_2, anti\text{-}b_1)$.

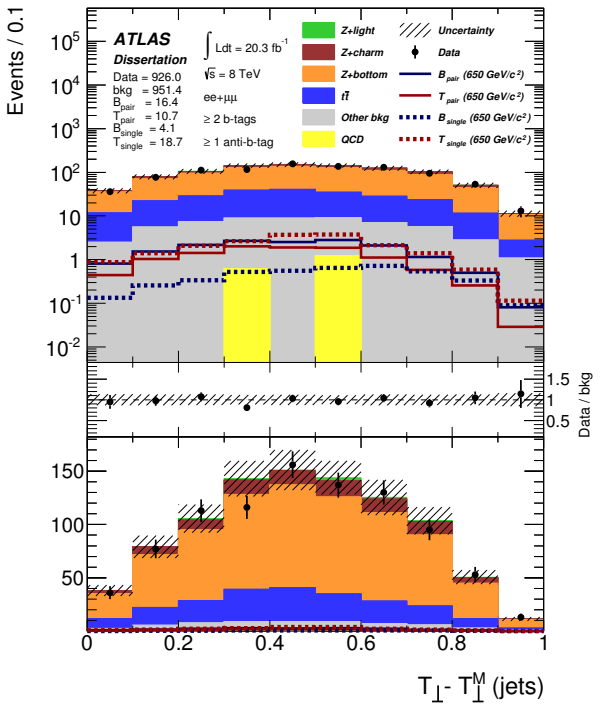
– *Continued, continues on next 2 pages.*



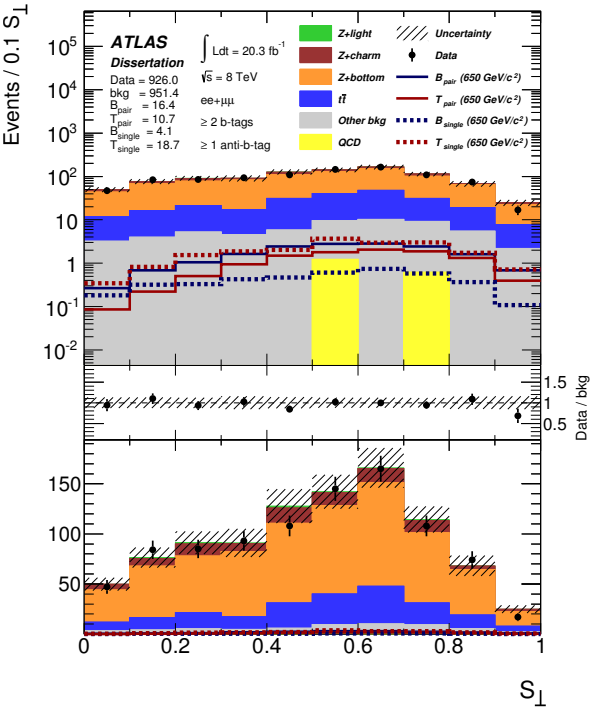
(o) $T_\perp(jets)$



(p) $T_\perp^M(jets)$



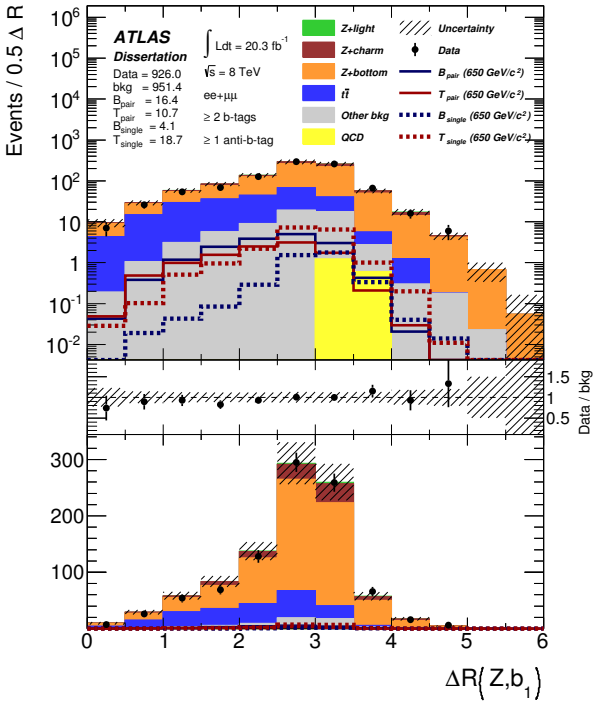
(q) $T_\perp - T_\perp^M(jets)$



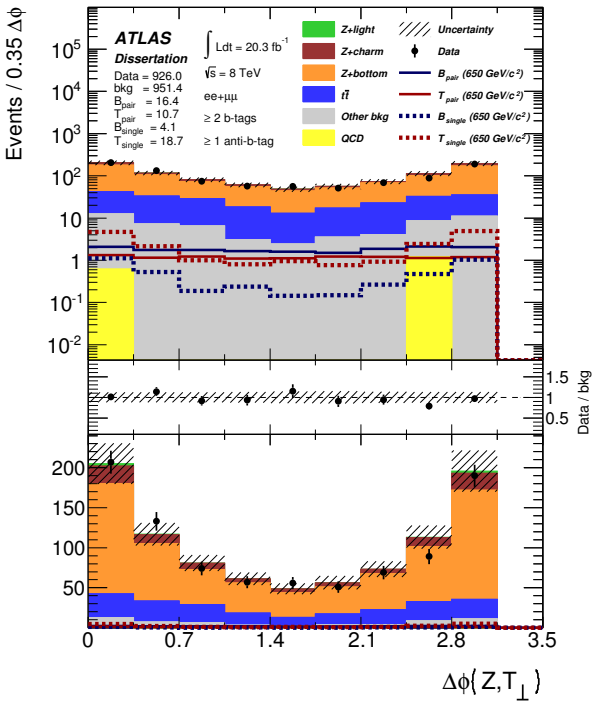
(r) $S_\perp$

Figure 5.36: (o) $T_\perp(jets)$, (p) $T_\perp^M(jets)$, (q) $T_\perp - T_\perp^M(jets)$, (r) $S_\perp$.

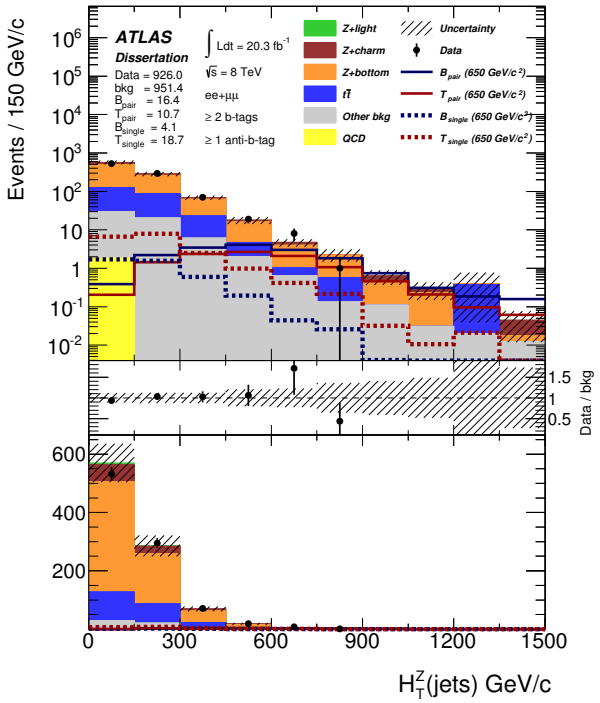
– *Continued, continues on next page.*



(s) $\Delta R(Z, b_1)$



(t) $\Delta\phi(Z, T_\perp)$



(u) $H_T^Z(jets)$

Figure 5.36: (s) $\Delta R(Z, b_1)$, (t) $\Delta\phi(Z, T_\perp)$, (u) $H_T^Z(jets)$.

– *Continued, last page.*

5.8.2. Validation of stage 1 Boosted Decision Tree Classification

Figures 5.37 (a)–(d) show the stage 1 BDT classifier response values for the 650 GeV/c^2 signal masses of each of the four signal types. The validation of the remaining 55 signal masses are provided in figure C.5 of appendix C.3. Modeling accuracy of the stage 1 classifiers was consistently good other than a few outliers that were deemed reasonable due to expected statistical fluctuations of analyzing so many comparisons.

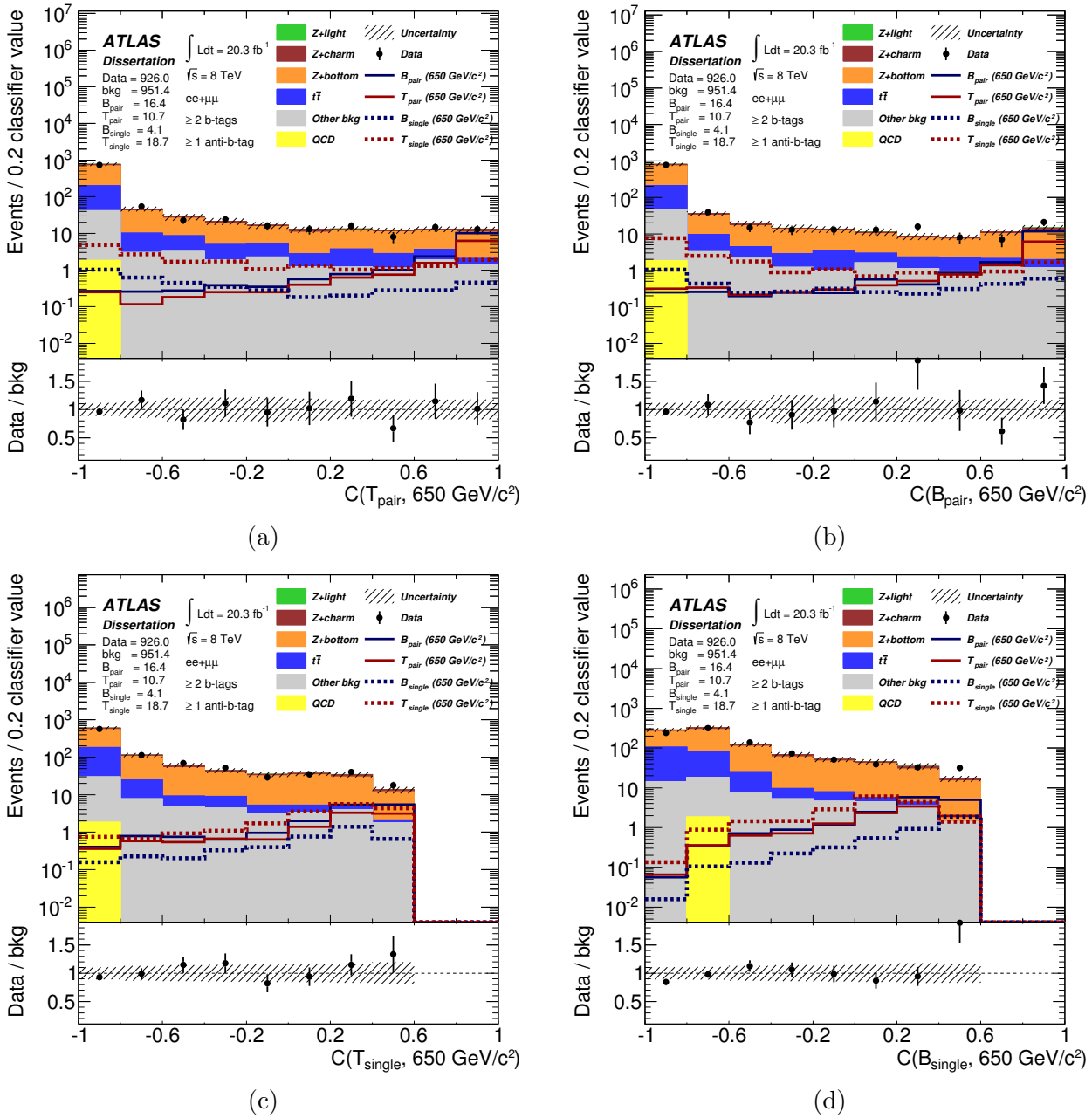
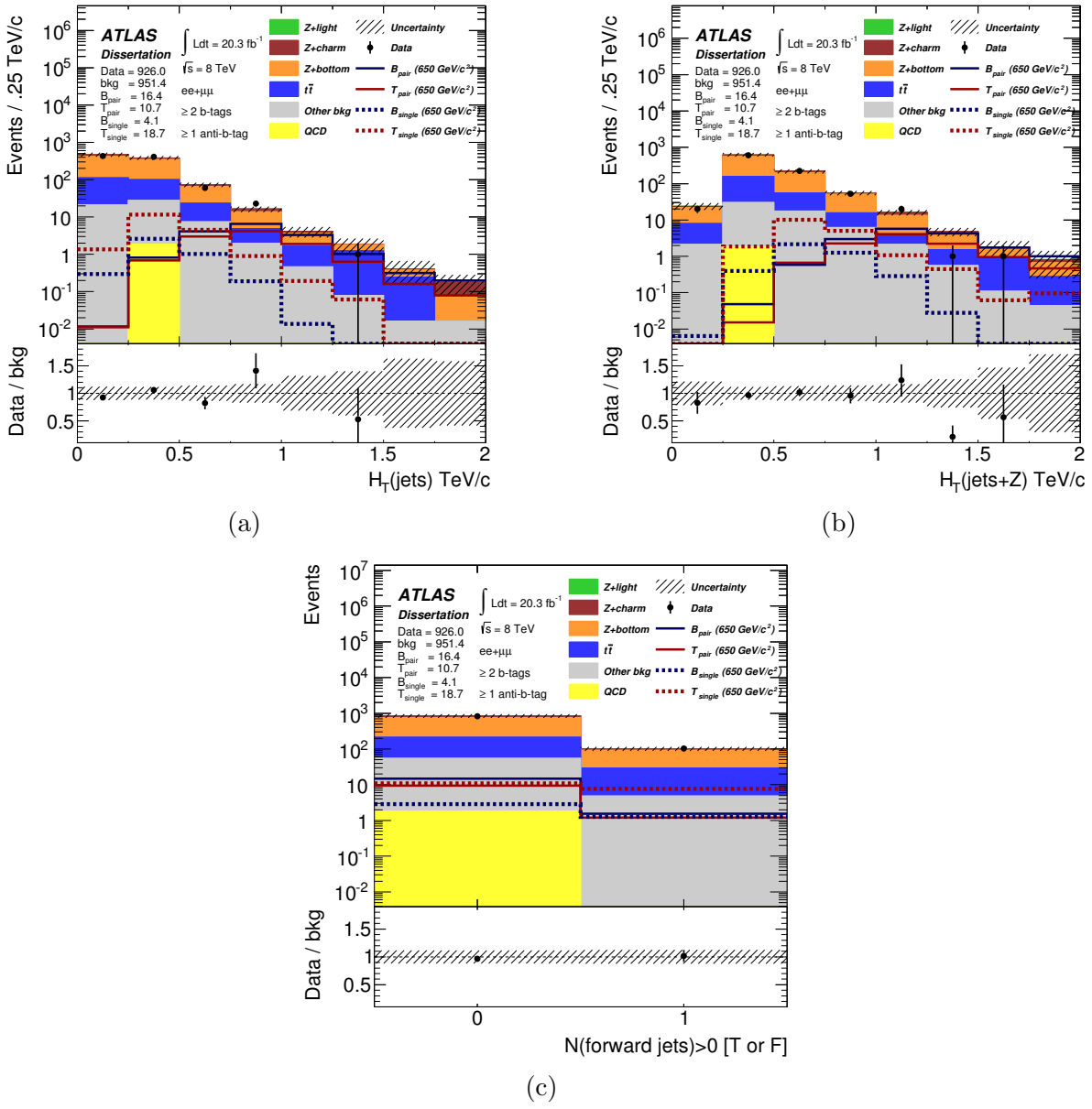


Figure 5.37: Simulated data validation in the preselection region of the resultant stage 1 BDT classifiers for the 650 GeV/c^2 signal masses of the four signal types (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

5.8.3. Validation of stage 2 Additional Input Variables

In addition to the stage 1 BDT classifiers, three other additional input variables were used for stage 2 logit classification: $H_T(jets)$, $H_T(jets + Z)$ and $N(forward\ jets) \geq 1$. The histograms for these are shown in figures 5.38 (a)–(c), and demonstrated good modeling accuracy.



(a)

(b)

(c)

Figure 5.38: Simulated data modeling validation in the preselection region of the stage 2 additional input variables (a) $H_T(jets)$, (b) $H_T(jets + Z)$, and (c) $N(forward\ jets) \geq 1$.

Figures 5.39 (a)–(c) show an intermediate check of the reconstructed mass spectra used in later statistical inference: (a) $m(Zb)$, the fully reconstructed vector-like bottom mass (or partially reconstructed vector-like top mass), (b) $m(bW)$, the reconstructed standard model top quark mass (see footnote 12 on page 170) that was used to fully reconstruct the vector-like top mass for the T_{single} signal type, and (c) $m(Zt)$, the fully reconstructed vector-like top mass used for the T_{single} signal type. These indicated good modeling quality at stage 1 preselection.

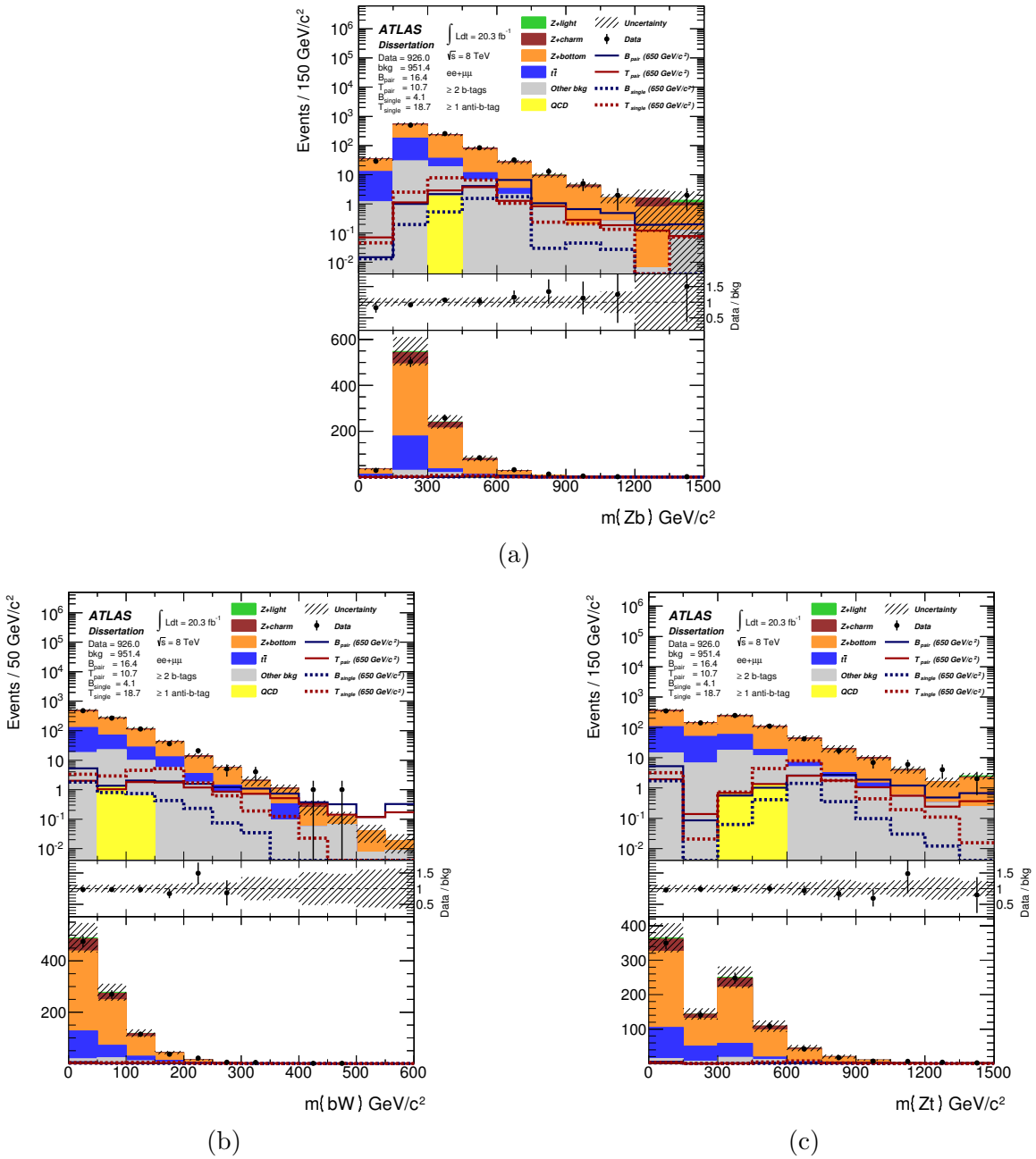


Figure 5.39: Simulated data modeling validation in the preselection region of the reconstructed mass spectra: (a) $m(Zb)$, (b) $m(bW)$, and (c) $m(Zt)$.

5.8.4. Validation of stage 2 Logit Classification

The selected method for stage 2 classification was logistic regression (logit classification) as described in section 5.6.2. The modeling validation for these classifiers are shown in figures 5.40 (a)–(d). These only show the classifier response values below the cut values used for the decision rules (and so remain blinded to the final signal regions). Figure 5.41 shows the T_{single} logit classifier modeling validation after an additional requirement of $N(forward\ jets) \geq 1$ was implemented (this additional requirement is explained in section 5.6 on page 153).

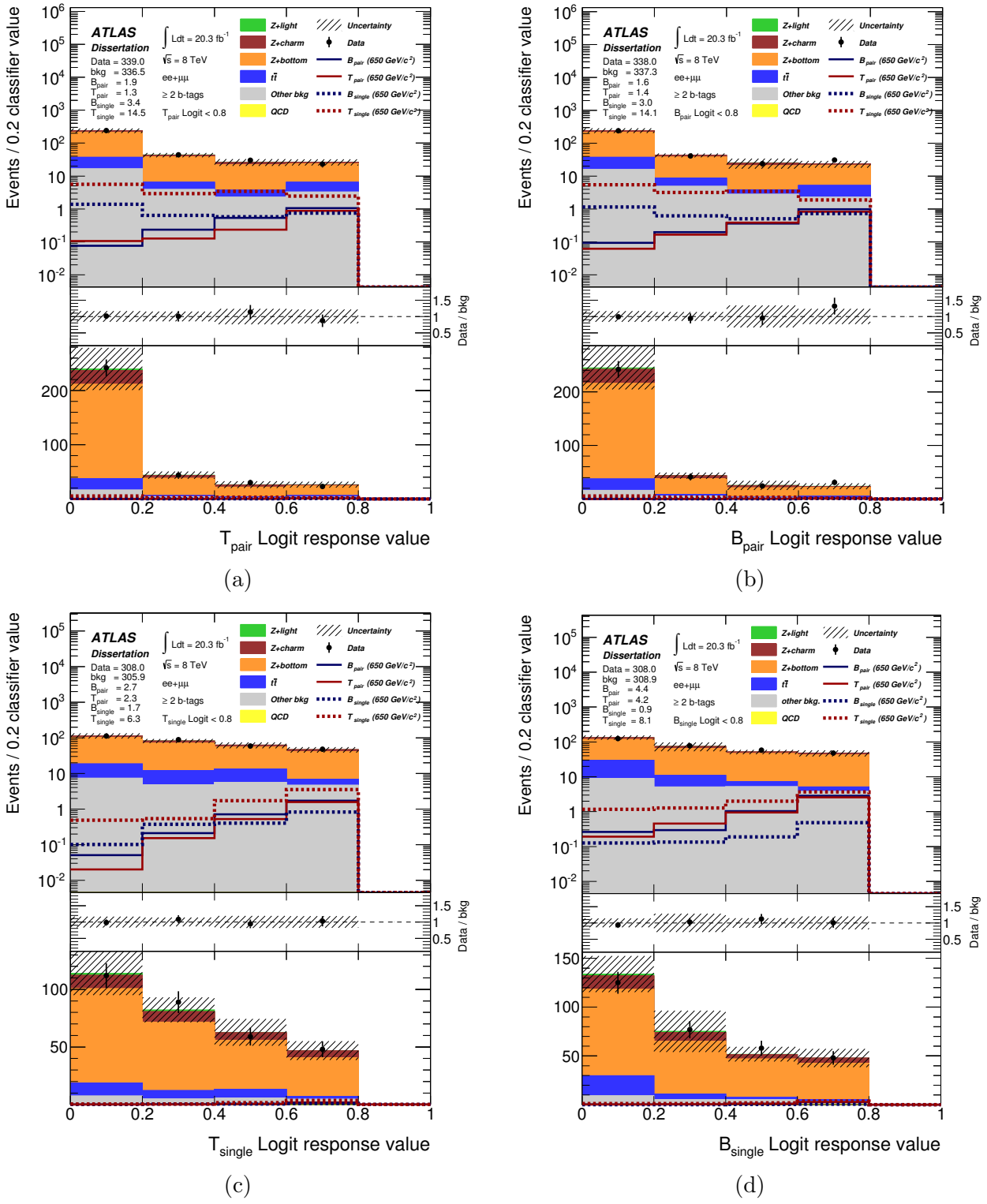


Figure 5.40: Simulated data modeling validation in the preselection region of the stage 2 logit classifiers (with the decision rules blinded) for the signal types (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

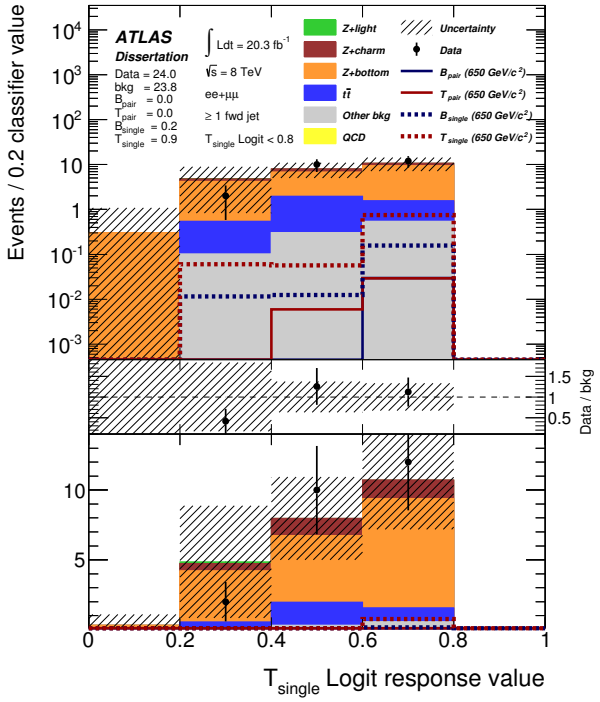


Figure 5.41: Simulated data modeling validation in the preselection region of the stage 2 T_{single} logit classifier after the additional requirement of $N(forward\ jets) \geq 1$ (with the decision rule blinded).

5.8.5. Validation of stage 3 Machine Learning Regression

The selected methods for stage 3 regression was OLS regression including quadratic terms for the T_{pair} and B_{pair} signal types, BDT regression for the T_{single} signal type, and OLS regression without quadratic terms for the B_{single} signal type (these selections are described in table 5.11 on page 171). The successful modeling validation for these are shown in figures 5.42 (a)–(d) before implementation of the stage 2 classification decision rules [17].

These also show that for the 650 GeV/c^2 signal masses, each regressand peaked near 650 GeV/c^2 for its targeted signal type. For example, in figure (a), the solid red line showing the expected yields of the T_{pair} signal type has the greatest expected yield in the histogram bin containing the 650 GeV/c^2 regressand value [18].

[17] Recall that the final signal region is defined by requiring a stage 2 logit classifier response value of 0.8 or greater.

[18] Although histograms showing the other 55 signal masses are not reported here, these also showed that machine learning regression was successful in correctly estimating the signal invariant masses with the expected signal yields peaking at the correct masses.

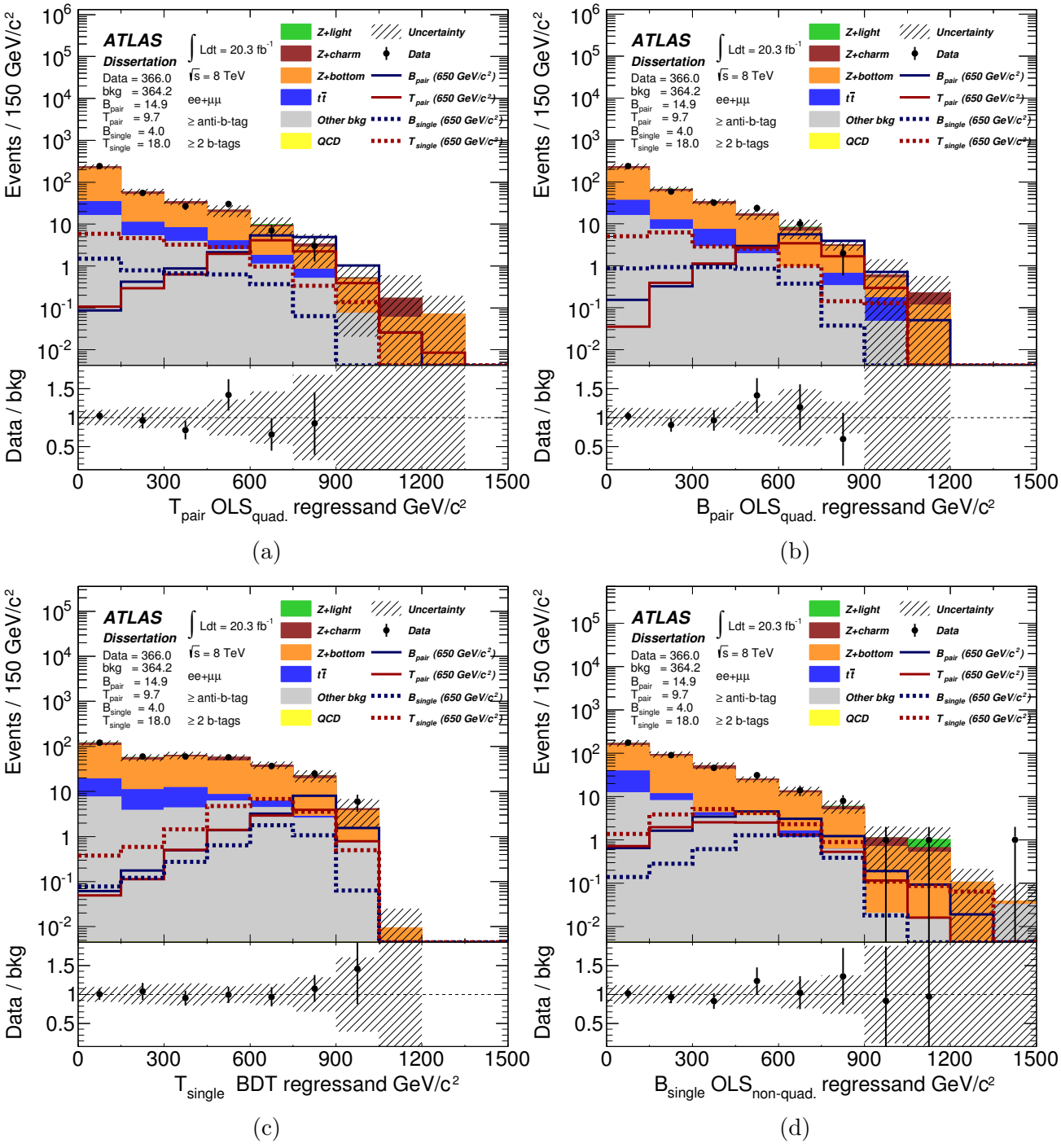


Figure 5.42: Simulated data modeling validation of the stage 3 regressions after stage 1 preselection and the additional $p_T(Z) > 150$ GeV/c requirement for the signals (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

5.9. Unblinded Signal Regions – Revealing the Measured Data

The strategy of the signal search was to determine the requirements that best reduced the data to only those collision events that were most signal-like and least background-like. These requirements defined the *signal region* that was then used to perform the final statistical inference to determine the upper limits of the strength of the signals[19]. The statistical inference procedure was made more powerful by examining the expected and measured yields over the spectrum of a discriminating variable (rather than simply examining the yields that pass the signal region requirements). Therefore, the statistical analysis used to evaluate the evidence for signal existence required 1) the definition of the signal region requirements and 2) the selection of a discriminating variable.

Table 5.16 shows the signal region requirements for the square cuts search of chapter 4 and the multi-stage machine learning-based search. For the square cuts search, requiring a large value of $H_T(jets)$ was found to greatly increase the ratio of expected signal yields to background yields for the T_{pair} and B_{pair} signal types. The T_{single} and B_{single} signal types had decay topologies that resulted in smaller values of $H_T(jets)$, and so the presence of a forward jet was required instead (discussed in chapter 4). For the machine learning search, the anti-b-tagged jet with the largest transverse momentum was used as an input variable to the stage 1 BDT learning algorithms, necessitating requirement 6. The stage 2 logit classification response value was then required to be at least 0.8, so that only the collision events that were predicted to be most signal-like were retained. Finally, at least one forward jet was required for the T_{single} signal type as explained in section 5.6 on page 153).

	$\mathbf{T_{pair}}$	$\mathbf{B_{pair}}$	$\mathbf{T_{single}}$	$\mathbf{B_{single}}$
all search methods and signal types:				
1	Preliminary requirements (see section 3.4.6).			
2	Exactly two same flavor, opposite sign leptons providing a Z boson candidate.			
3	$\lvert m(Z) - 91.19\ \text{GeV/c}^2\rvert \leq 10\ \text{GeV/c}^2$.			
4	$p_T(Z) \geq 150$ GeV/c.			
5	$N(\textit{b-tagged jets}) \geq 2$.			
square cuts:				
6	$H_T(jets) \geq 600$ GeV/c	$H_T(jets) \geq 600$ GeV/c	$N(\textit{forward jets}) \geq 1$	$N(\textit{forward jets}) \geq 1$
machine learning:				
6	$N(\textit{anti-b-tagged}) \geq 1$	$N(\textit{anti-b-tagged}) \geq 1$	$N(\textit{anti-b-tagged}) \geq 1$	$N(\textit{anti-b-tagged}) \geq 1$
7	stage 2 logit ≥ 0.8	stage 2 logit ≥ 0.8	stage 2 logit ≥ 0.8	stage 2 logit ≥ 0.8
8	n/a	n/a	$N(\textit{forward jets}) \geq 1$	n/a

Table 5.16: A summary of the signal region requirements for the square cuts searches and the machine learning searches for the four signal types. Both search methods have the first five requirements in common.

[19] The upper limit of the signal strength was expressed as the upper limit of the theoretical cross section of the hypothesized vector-like quark, and that was supported by the evidence provided by the measured data using the Neyman construction discussed in the next chapter.

The square cuts search of chapter 4 used the reconstructed mass $m(Zb)$ as the discriminating variable for statistical inference in the signal regions. This worked well for the B_{pair} and B_{single} signal types because it fully reconstructed the vector-like bottom quark mass. For example, the spectrum of the 650 GeV/c^2 signal mass for B_{pair} or B_{single} was sharply peaked near $m(Zb) = 650$ GeV/c^2 as is shown in figure 5.39 (a) on page 196 (blue lines).

However, $m(Zb)$ did not fully reconstruct the vector-like top quark mass since it did not contain the contribution from the W boson decay product. For T_{pair} and T_{single}, this resulted in a less well-defined $m(Zb)$ spectrum. For example, the spectrum of the 650 GeV/c^2 signal mass for T_{pair} or T_{single} had a wider spread than for B_{pair} or B_{single} and was peaked below $m(Zb) = 650$ GeV/c^2 (also shown in figure 5.39 (a) with red lines).

Many strategies to fully reconstruct the vector-like top quark mass, $m(Zt)$, were attempted. This required identifying the t quark by its decay products, b and W. Unfortunately, the W boson further decayed to two SM quarks,[20] $W \to q + \bar{q}'$, which were detected as one or two hadronic jets. Because the typical collision event contained many hadronic jets, it was difficult to correctly identify which jets corresponded to the decay of the W boson. This was especially problematic for the T_{pair} signal type because its decay topologies contained many more jets than for T_{single}.

One method of fully reconstructing the vector-like top mass, $m(Zt)$, was found for T_{single} (described in footnote 12 on page 170). This was done by first identifying a candidate t quark with reconstructed mass $m(bW)$. The spectrum of $m(bW)$ for the 650 GeV/c^2 T_{single} signal mass is is shown in figure 5.39 (b) with the dashed red line, which peaks near 172 GeV/c^2, the invariant mass of the SM top quark. Figure 5.39 (c) shows the spectrum of $m(Zt)$ after combining the SM top quark candidate with the Z boson candidate, which successfully peaked near 650 GeV/c^2 (dashed red line).

As discussed in section 5.7, stage 3 machine learning *regression* offered a solution to obtaining estimates of the vector-like quark invariant masses that did not rely on the problematic procedure of identifying specific decay products within complex collision events. For each collision event, machine learning simply proceeded to the estimate in a single step using the results of stage 1 and stage 2 machine learning.

Therefore, three discriminating variables were used for statistical inference of the four signal types as summarized in table 5.17.

discriminating variable	signal type			
	T_{pair}	B_{pair}	T_{single}	B_{single}
reconstructed $m(Zb)$	✓	✓	✓	✓
reconstructed $m(Zt)$			✓	
stage 3 regressand $m(Q)$	✓	✓	✓	✓

Table 5.17: The three discriminating variables used for statistical inference.

[20] The W boson may also decay to a lepton and neutrino, however the requirement of exactly two leptons per collision event (requirement 2 in table 5.16) removed these collision events from the signal regions.

These signal regions and discriminating variables defined four unique search strategies that are summarized in table 5.18. These were used for the statistical inference procedure of implementing the Neyman construction to determine upper limits of the signal strength (fully explained in the next chapter). The upper limit results of these strategies allowed comparison of the performance of the traditional square cuts search to machine learning classification (two-stage), as well as to machine learning regression (three-stage).

	search strategy	=	signal region	+	discriminating variable
1)	traditional square cuts	=	square cuts	+	reconstructed $m(Zb)$
2)	two-stage machine learning	=	machine learning	+	reconstructed $m(Zb)$
3)*	two-stage machine learning	=	machine learning	+	reconstructed $m(Zt)$
4)	three-stage machine learning	=	machine learning	+	stage 3 regressand mass

Table 5.18: The four search strategies used for the final statistical inference.

**Only used for the T_{single} signal type.*

5.9.1. Unblinded Results

The analysis was unblinded by examining the spectra of the discriminating variables in the final signal regions using both the measured data and the simulated data. These signal region spectra were used for the final statistical inference of examining the evidence for signal existence provided by the measured data.

Tables 5.19 (a) and (b) show the expected and measured yields of the square cuts search strategy signal regions. Each expected yield was expressed as a central value, followed by the uncertainty due to a limited number of simulated events (referred to as "Monte Carlo statistical uncertainty" in particle physics), and finally the asymmetric estimated systematic uncertainty: *central val.* $\pm$ *stat. unc.* $^{+syst.\ unc.}_{-syst.\ unc.}$. The expected yields of the 650 GeV/c^2 signal masses are also shown for reference.[21]

These tables show that the measured yields were near the central values of the expected total backgrounds and within the uncertainty bounds. For the B_{pair} signal type, table 5.19 (a) indicated that signal existence with a mass of 650 GeV/c^2 was strongly disfavored. This was because the measured yield was well below the sum of the expected yields of the signal and background. For the T_{pair} and T_{single} signal types, the measured yields were near the lower uncertainty bounds of the 650 GeV/c^2 signal plus background estimated yields. The central value of the 650 GeV/c^2 B_{single} signal plus background estimated yields were near the measured yields. Thus the result at this point was ambiguous for T_{pair}, T_{single} and B_{single}.

data sample	ee+$\mu\mu$ channels
Z+light	$0.1\pm0.06^{+0.35}_{-0.04}$
Z+charm	$4.0\pm0.62^{+1.13}_{-1.08}$
Z+bottom	$19.9\pm1.00^{+3.30}_{-3.51}$
$t\bar{t}$	$4.6\pm1.16^{+0.59}_{-0.94}$
other background	$4.0\pm0.47^{+1.23}_{-1.40}$
total background	$32.6\pm1.72^{+4.77}_{-5.66}$
measured data	31
B_{pair} (650 GeV/c^2)	$13.86\pm0.33^{+0.90}_{-1.19}$
T_{pair} (650 GeV/c^2)	$8.35\pm0.24^{+0.59}_{-0.52}$

(a)

data sample	ee+$\mu\mu$ channels
Z+light	$0.1\pm0.08^{+0.36}_{-0.02}$
Z+charm	$7.2\pm1.00^{+1.59}_{-1.65}$
Z+bottom	$39.7\pm1.65^{+8.69}_{-5.68}$
$t\bar{t}$	$5.2\pm1.15^{+1.75}_{-1.29}$
other background	$3.8\pm0.62^{+1.11}_{-1.14}$
total background	$55.9\pm2.32^{+11.10}_{-9.73}$
measured data	57
B_{single} (650 GeV/c^2)	$1.88\pm0.10^{+0.25}_{-0.21}$
T_{single} (650 GeV/c^2)	$7.70\pm0.58^{+0.83}_{-0.64}$

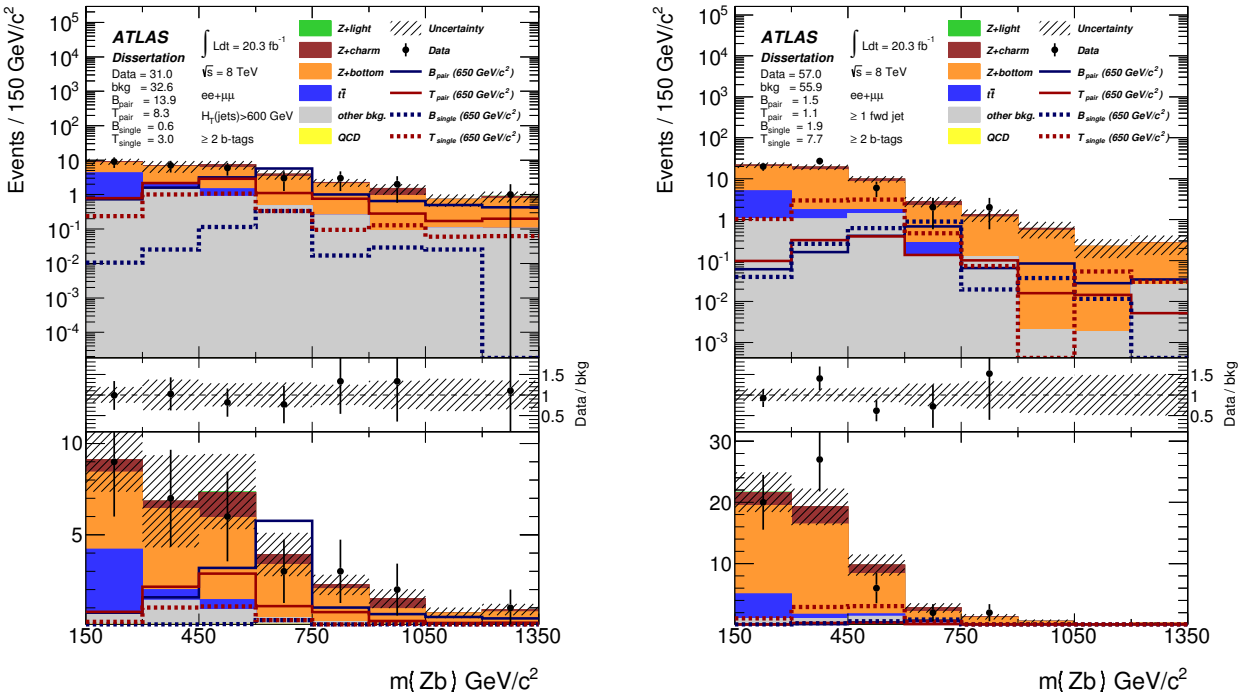
(b)

Table 5.19: ***(Square Cuts)*** — The expected and measured yields of the unblinded square cuts search strategy: (a) using $H_T(jets) \geq 600$ GeV/c for square cuts signal region requirement 6 for T_{pair} and B_{pair} (see table 5.16), and (b) using $N(forward\ jets) \geq 1$ for T_{single} and B_{single}.

[21] The representative signal mass of 650 GeV/c^2 was selected for display in such tables and figures as this was near the current threshold value of previous experiments, below which vector-like quarks already were observed to have little evidence of existing.

The ambiguity of the 650 GeV/c^2 signals existence when examining the total yields in the signal regions motivated the use of a discriminating variable spectrum within the signal regions to further distinguish between the signal and background. Figures 5.43 (a) and (b) show the $m(Zb)$ discriminating variable spectra for the square cuts search strategy. These show that the expected yields of the background were larger at lower values of $m(Zb)$ than those of the 650 GeV/c^2 signals.

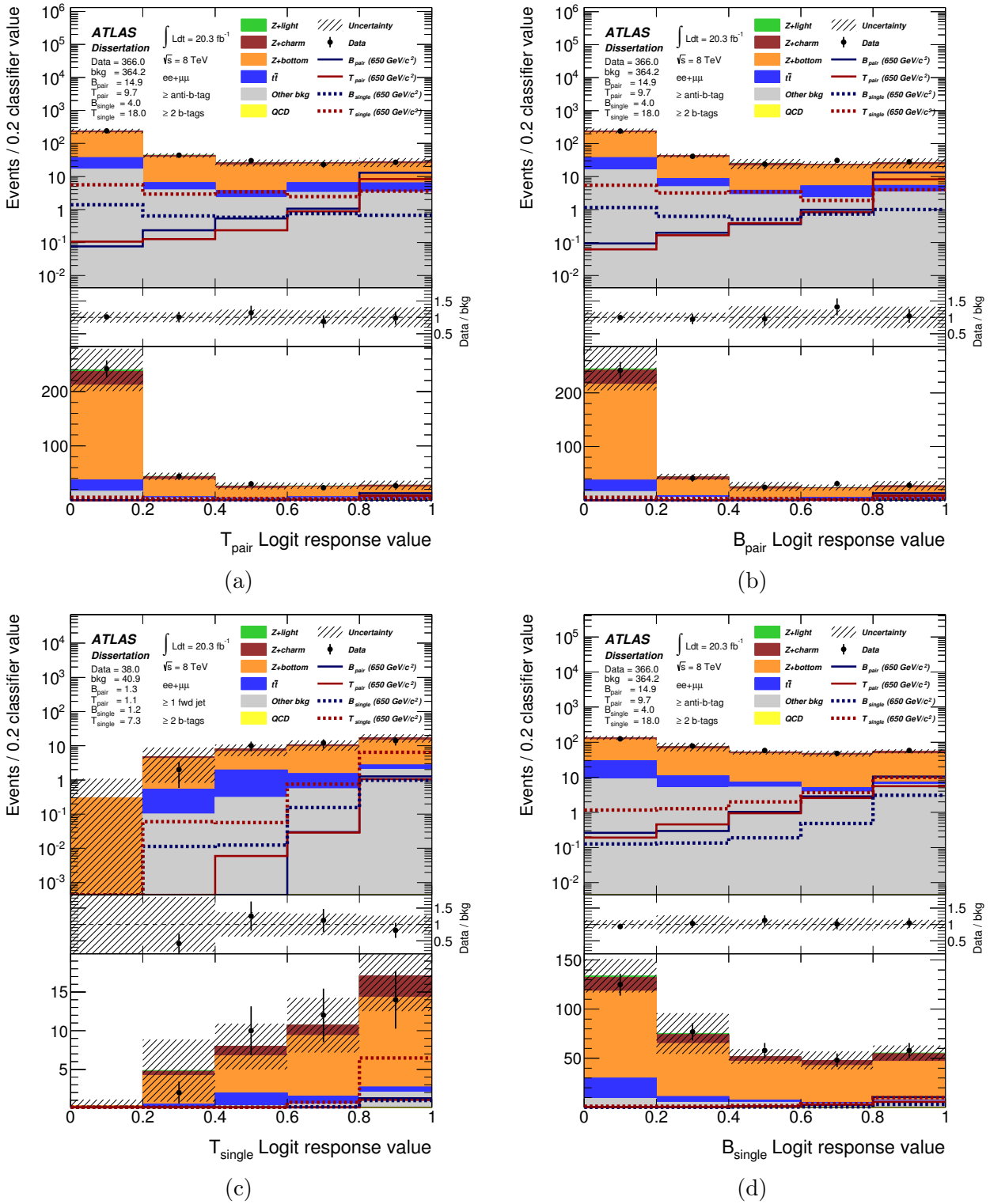
In figure 5.43 (a), the measured yields appeared more consistent with the expected background yields than with the background plus 650 GeV/c^2 T_{pair} and B_{pair} signals (solid lines). In figure 5.43 (b), the result was more ambiguous because the 650 GeV/c^2 T_{single} and B_{single} signals (dashed lines) had much smaller expected yields than the background. This was problematic because even though these showed that there was little evidence for signal existence (at 650 GeV/c^2), it led to poor inference results of large upper limits that essentially provided little useful information. This was the main motivation for seeking a machine learning solution.



(a) (b)

Figure 5.43: ***(Square Cuts – m(Zb))*** — The $m(Zb)$ discriminating variable spectra in the unblinded square cuts signal regions: (a) using $H_T(jets) \geq 600$ GeV/c for square cuts signal region requirement 6 for T_{pair} and B_{pair} (see table 5.16), and (b) using $N(forward\ jets) \geq 1$ for T_{single} and B_{single}.

Figures 5.44 (a)–(d) show the stage 2 logit classifier response values. The last bin of each figure is unblinded, showing the measured yields for values above 0.8 that define the machine learning signal regions (and also implementing the other machine learning signal region requirements, see table 5.16). These show that the measured yields appeared to be consistent with the expected background. However, the small expected yields of the 650 GeV/c^2 signals made conclusions for signal existence ambiguous, indicating the need for the use of a discriminating variable. This is further demonstrated by tables 5.20(a)–(d) that compare the expected yields to the measured yields in these machine learning signal regions.



(a) (b) (c) (d)

Figure 5.44: ***(Machine Learning)*** — The unblinded stage 2 logit classifiers for the signal types (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}. The last bins of each show the yields of the machine learning signal regions.

data sample	ee+$\mu\mu$ channels
Z+light	$0.0\pm0.01^{+0.39}_{-0.00}$
Z+charm	$3.7\pm0.68^{+0.89}_{-1.28}$
Z+bottom	$17.7\pm1.11^{+2.71}_{-7.40}$
$t\bar{t}$	$3.0\pm1.17^{+1.16}_{-0.50}$
other background	$3.3\pm0.35^{+0.85}_{-0.58}$
total background	$27.7\pm1.78^{+3.95}_{-7.99}$
measured data	27
T_{pair} (650 GeV/c^2)	$8.4\pm0.29^{+0.70}_{-0.46}$

(a)

data sample	ee+$\mu\mu$ channels
Z+light	$0.4\pm0.35^{+0.73}_{-0.00}$
Z+charm	$3.7\pm0.69^{+0.87}_{-1.67}$
Z+bottom	$17.5\pm1.09^{+2.87}_{-7.23}$
$t\bar{t}$	$1.9\pm0.94^{+0.82}_{-0.26}$
other background	$3.4\pm0.37^{+0.48}_{-1.04}$
total background	$26.9\pm1.68^{+4.12}_{-8.28}$
measured data	28
B_{pair} (650 GeV/c^2)	$13.3\pm0.39^{+1.69}_{-0.67}$

(b)

data sample	ee+$\mu\mu$ channels
Z+light	$0.0\pm0.00^{+0.00}_{-0.00}$
Z+charm	$2.8\pm0.75^{+0.43}_{-1.87}$
Z+bottom	$11.6\pm1.01^{+2.91}_{-2.39}$
$t\bar{t}$	$0.7\pm0.37^{+0.80}_{-0.07}$
other background	$2.1\pm0.61^{+0.34}_{-0.98}$
total background	$17.1\pm1.45^{+3.42}_{-4.63}$
measured data	14
T_{single} (650 GeV/c^2)	$6.5\pm0.75^{+0.89}_{-0.62}$

(c)

data sample	ee+$\mu\mu$ channels
Z+light	$0.8\pm0.45^{+0.15}_{-0.69}$
Z+charm	$8.0\pm1.09^{+0.54}_{-5.21}$
Z+bottom	$39.3\pm1.68^{+10.46}_{-4.02}$
$t\bar{t}$	$1.0\pm0.54^{+0.77}_{-0.17}$
other background	$6.3\pm0.78^{+1.95}_{-2.20}$
total background	$55.4\pm2.26^{+9.24}_{-8.88}$
measured data	58
B_{single} (650 GeV/c^2)	$3.08\pm0.18^{+0.22}_{-0.66}$

(d)

Table 5.20: ***(Machine Learning)*** — The unblinded machine learning signal regions: (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

Figures 5.45 (a)–(d) show Venn diagrams of the expected yields in the square cuts signal regions and machine learning signal regions for each of the four signal types (the middle number of each shows the amount of expected yields common to both signal regions). These indicated 1) that the signal regions for T_{pair} and B_{pair} shared a large portion of the expected background and signal in common, 2) that the T_{single} machine learning signal region was a subset of the square cuts signal region, and 3) that the B_{single} square cuts and machine learning signal regions were quite dissimilar. These correlations between the square cuts and machine learning signal regions were helpful in understanding the final results of statistical inference presented in the next section.

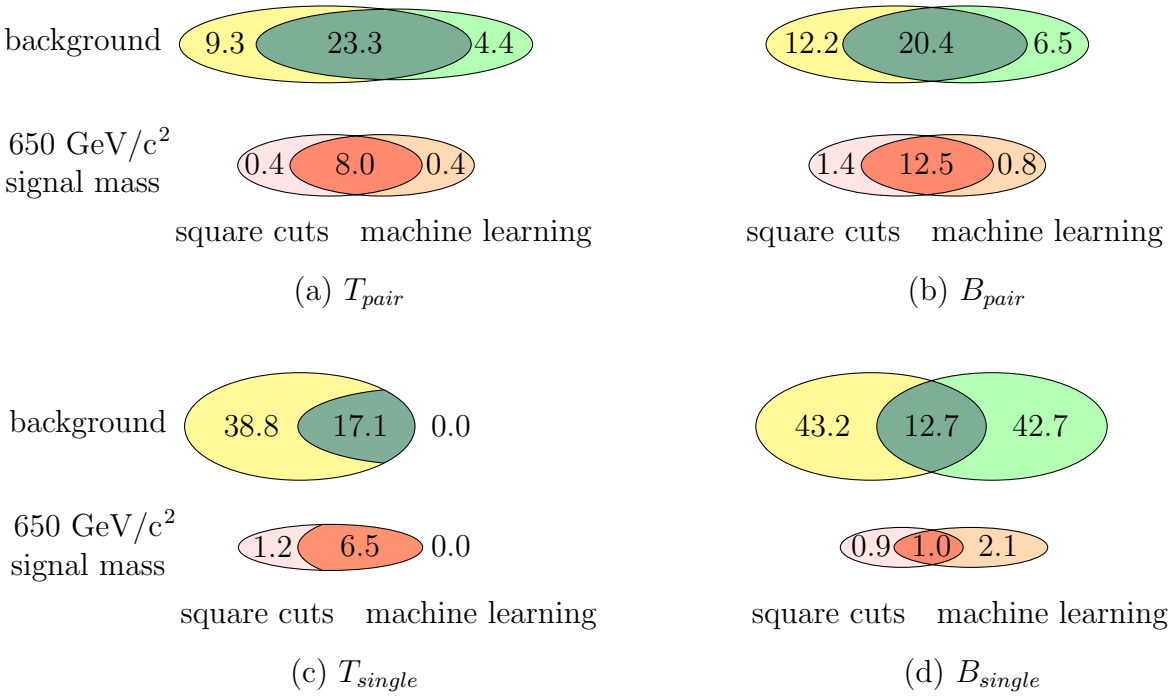


Figure 5.45: Venn diagrams of expected yields showing the similarities between the square cuts signal regions and machine learning signal regions for T_{pair}, B_{pair} and T_{single}, and the differences for those of B_{single}

Figures 5.46 (a)–(d) show the $m(Zb)$ discriminating variable spectra in the machine learning signal regions. The measured yields appeared consistent with the expected background yields, mostly lying within the estimated background uncertainty bounds. Comparing these to the square cuts spectra of figures 5.43 (a) and (b), it was not apparent which method provided stronger evidence of signal nonexistence (at 650 GeV/c^2).

The solid red line of figure 5.46 (a) shows the expected yields of the 650 GeV/c^2 T_{pair} signal. Unlike the solid blue line of figure 5.46 (b), the expected yields of the 650 GeV/c^2 B_{pair} signal that peaked sharply at 650 GeV/c^2, the T_{pair} signal peaked below 650 GeV/c^2 and exhibited the problem of using $m(Zb)$ that only partially reconstructs the invariant mass of the vector-like top quark. The red dashed line of figure 5.46 (c) and blue dashed line of figure 5.46 (d) indicated that $m(Zb)$ was only somewhat successful in estimating the invariant masses of the vector-like top and bottom quarks, respectively.

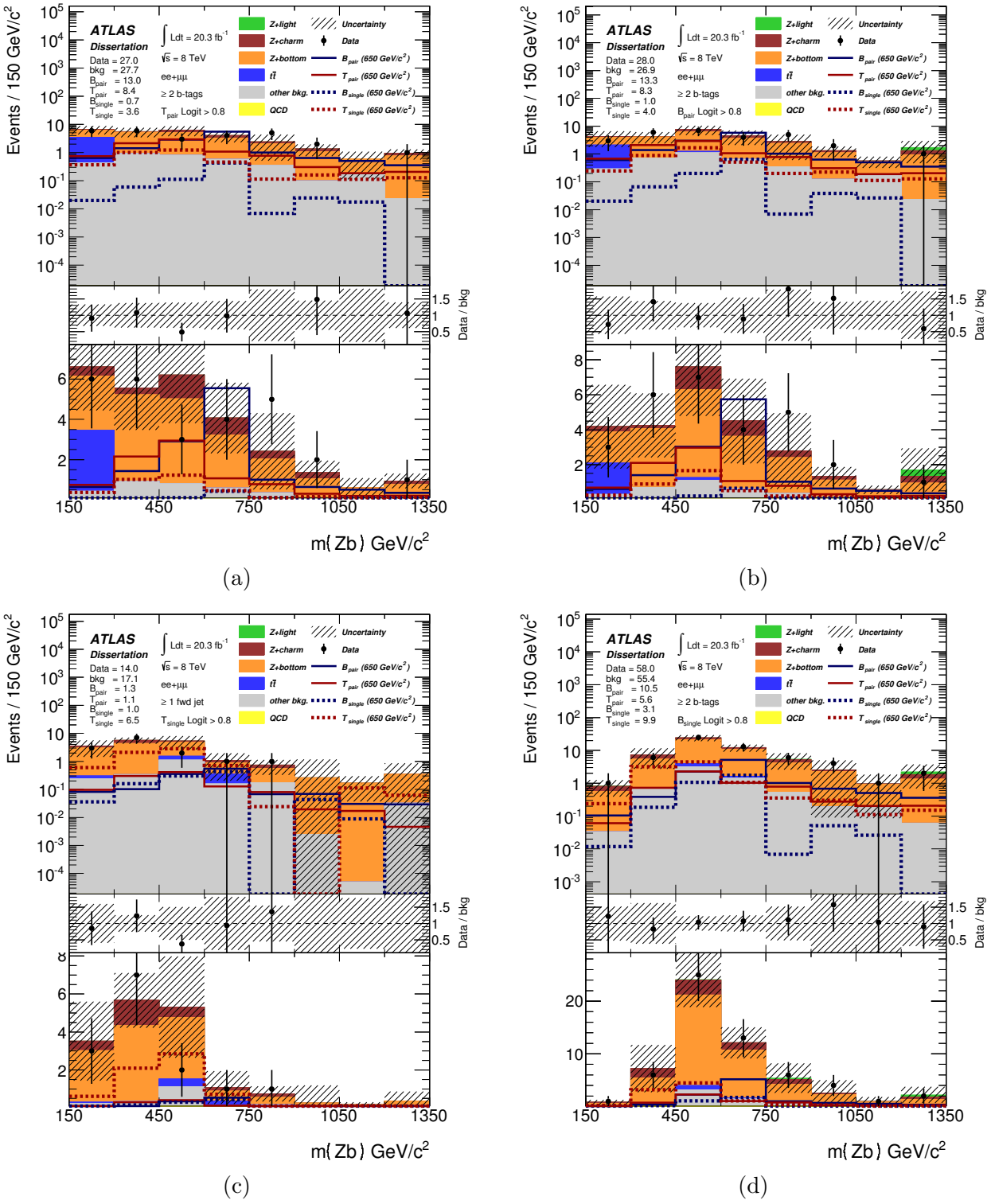
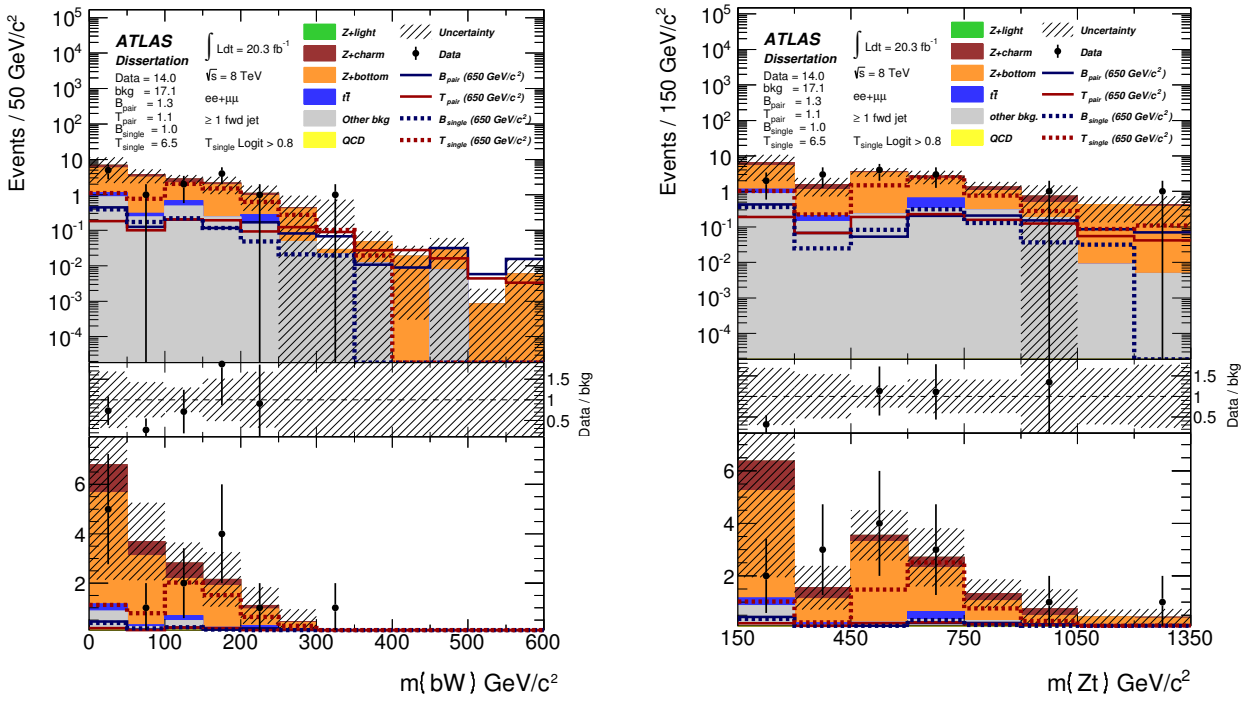


Figure 5.46: ***(Machine Learning – m(Zb))*** — The $m(Zb)$ discriminating variable spectra in the unblinded machine learning signal regions: (a) T_{pair}, (b) B_{pair}, (c) T_{single} (with $N(forward\ jets) \geq 1$), and (d) B_{single}.

The $m(Zb)$ discriminating variable provided only a partial reconstruction of the vector-like top mass because it did not contain the contribution from the W boson decay product (since $T \to Z + t \to Z + \bar{b} + W^+$). For the T_{single} signal type, a full reconstruction of the T mass was found using a simple clustering algorithm to identify the decay products of the T (see footnote 12 on page 170). However, no method was found that successfully identified the decay products of T when pair produced (i.e., the T_{pair} signal type). This was because the decay topology of T_{pair} was much more complex than T_{single}.[22]

Figure 5.47 (a) shows the spectrum of the reconstructed Standard Model (SM) top quark mass, $m(bW)$. Although figure 5.39 (b) (page 196) shows good modeling accuracy of $m(bW)$ in the preselection region, figure 5.47 (a) indicated some possible modeling inaccuracy of $m(bW)$ in the machine learning signal region.

The candidate SM top quark, t, was estimated as an intermediate step in fully reconstructing the vector-like top mass, $m(Zt)$. Figure 5.47 (b) shows the $m(Zt)$ mass spectrum that was used as an alternate discriminating variable for the search for the T_{single} signal type. The dashed red line shows the spectrum of the 650 GeV/c^2 T_{single} signal mass. Unlike the $m(Zb)$ spectrum of figure 5.46 (d), the $m(Zt)$ spectrum correctly peaked at 650 GeV/c^2 and exhibited smaller spread, and thus provided a stronger discriminating variable for the T_{single} search.



(a) (b)

Figure 5.47: ***(Machine Learning – m(Zt))*** — The full reconstruction of the T_{single} invariant mass, $m(Zt)$, in the machine learning signal region. (a) The spectrum of the reconstructed SM top quark mass, $m(bW)$, used to obtain (b) $m(Zt)$.

[22] The T_{pair} decay topology usually contained two hadronically decaying SM top quarks so that identifying the correct t candidate to combine with the Z boson candidate was problematic.

Two-stage machine learning *classification* was used as an alternative approach to the square cuts strategy of identifying signal regions. Machine learning *regression* was used as an alternative method of obtaining a discriminating variable. In order for the reconstructed masses, $m(Zb)$ and $m(Zt)$, to have been useful discriminating variables in these searches, the specific decay products of the vector-like quarks needed to by correctly identified. The complex decay topologies of the vector-like quark signal, as well as those for the competing backgrounds, made this identification difficult.[23]

Machine learning regression was used here to estimate the signal invariant masses *without* identifying the specific vector-like quark decay products. This was accomplished simply by providing the information contained in the collision events to the learning algorithms, and allowing machine learning regression to determine the signal invariant masses (as described fully in section 5.7).

Figures 5.48(a)–(d) show the stage 3 regression mass spectra that were used as an alternative method of providing a discriminating variable. These show that the regressand masses for the 650 GeV/c^2 signals correctly peaked at 650 GeV/c^2. For example, the solid red line in (a) reaches a maximum in the histogram bin containing 650 GeV/c^2. In general, the stage 3 regressand masses tended to provide more powerful signal-background discrimination than the reconstructed masses, $m(Zb)$ and $m(Zt)$, for higher signal masses. This was because the regressand mass spectra for higher signal masses peaked at the correct mass value while tending to have a smaller spread than the reconstructed mass spectra.

[23] One complaint often made against the use of machine learning in particle searches is that it complicates the analysis, and increases the workload of the search team and possibly the length of time to obtain a final result. Although this is generally true, machine learning regression for the reconstructed mass that is to be used as the discriminating variable may save time and effort in cases of busy decay topologies or where the signal has several different decay paths (both true for vector-like quarks).

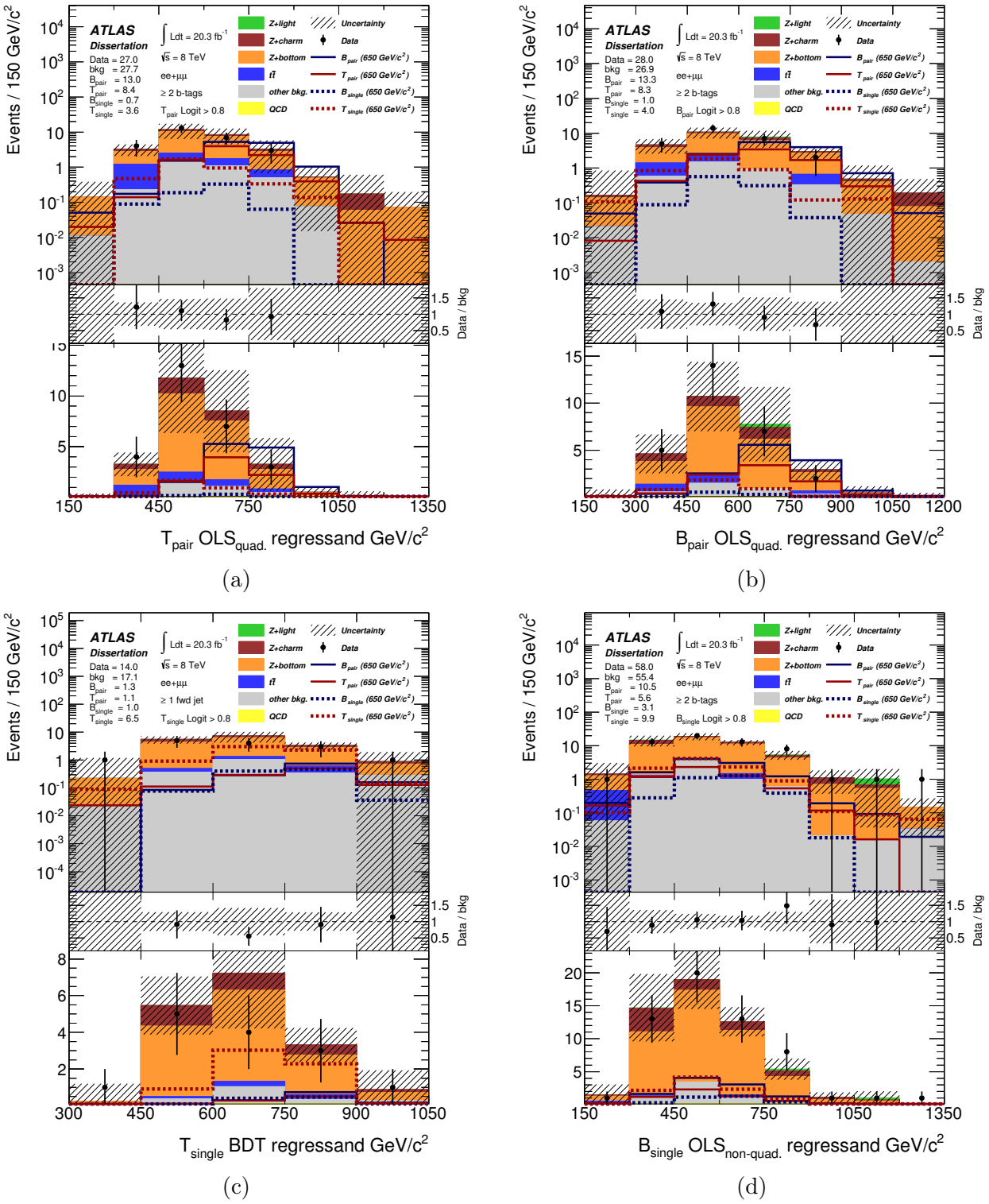


Figure 5.48: ***(Machine Learning – stage 3 regressand mass)*** — The unblinded stage 3 regressand mass spectra in the machine learning signal regions: (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

5.10. Statistical Inference – Setting CL_S Upper Limits

The final result of a signal search is obtained by performing statistical inference on the unblinded data in order to quantify the evidence for signal existence. For the vector-like quark searches, the Neyman construction of confidence regions was used to compute 95% confidence level confidence intervals for their measured cross sections. These confidence intervals were then compared to the theoretical cross sections predicted by the vector-like quark theories in order to determine the existence of the vector-like quarks (at 95% confidence).

For example, if σ_{th} was the theoretical cross section of a vector-like quark for a particular invariant mass, and if $[0, \sigma_u]$ was the 95% confidence interval with its upper limit below the theoretical cross section, $\sigma_u < \sigma_{th}$, then the evidence provided by the measurement was said to exclude the existence of the vector-like quark (for that mass) with 95% confidence. This is because 95% confidence intervals contain the true value of the cross section in 95% of experiments. If the true vector-like quark cross section was below the theoretical cross section, then the theory hypothesizing vector-like quark existence was incorrect.

Using only the simulated data, Monte Carlo simulations were used to estimate the *expected* upper limits that would have been obtained in the case of signal nonexistence.[24] Examining the expected upper limits allowed a comparison of the performance of each search strategy that was independent of the actual measured data. The goal of machine learning was to improve upon the results obtained using the traditional square cuts strategy. Table 5.18 that shows the four search strategies is repeated here for convenience.

	search strategy	=	signal region	+	discriminating variable
1)	traditional square cuts	=	square cuts	+	reconstructed $m(Zb)$
2)	two-stage machine learning	=	machine learning	+	reconstructed $m(Zb)$
3)*	two-stage machine learning	=	machine learning	+	reconstructed $m(Zt)$
4)	three-stage machine learning	=	machine learning	+	stage 3 regressand mass

Table 5.18: The four search strategies. *(Repeated from page 204.)*

** Only used for the T_{single} signal type.*

[24] See section 6.2.4.

5.10.1. Examination of Systematic Uncertainties

The machine learning search included the same systematic uncertainties that were used in the square cuts search, (discussed in section 4.3 on page 94). Systematic uncertainties were not used in the training of learning algorithms, instead using the traditional approach to first train without them, and then evaluate their impact on the resultant predictor functions [242]. The use of machine learning did not generate any additional systematic uncertainty [243].[25] This was because the systematic uncertainties were already estimated for each of the input variables used for machine learning.

For example, let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ be the set of n input variables used for machine learning, and let $y = f(\mathbf{X})$ be the predictor function constructed via machine learning that returns the response value y. In a particle search, Bayesian prior distributions are estimated for each systematic uncertainty as asymmetric Gaussian-like distributions with two parameters: a one standard deviation upward shift, and a one standard deviation downward shift (discussed in section 6.2.1 on page 258, see also section 3.3 of [244] that explicitly describes the "MCLimit" interpolation and extrapolation method). Figure 5.49 provides an example of an asymmetric Gaussian-like probability distribution.

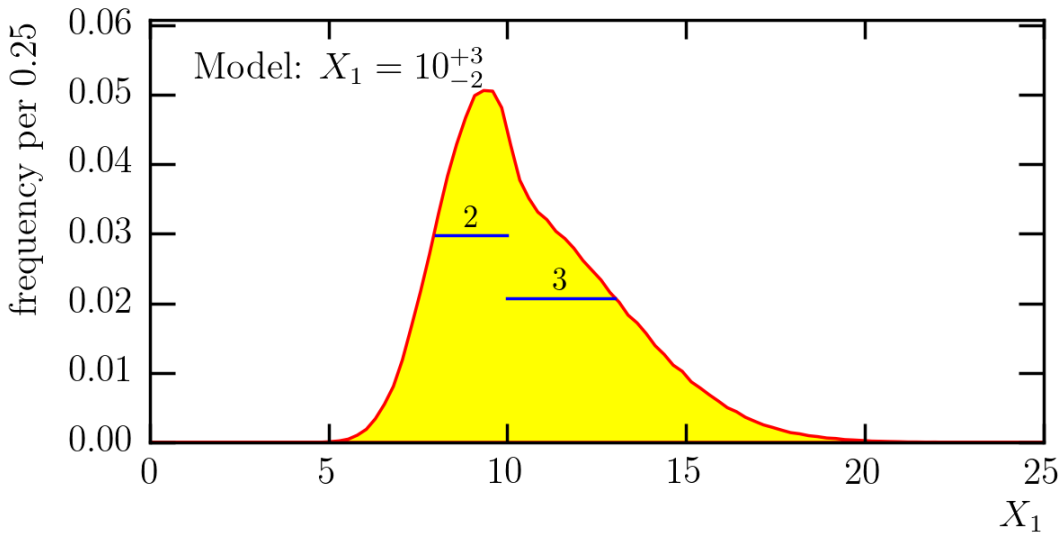


Figure 5.49: An example of the asymmetric Gaussian-like distribution for the hypothetical input variable, X_1, that has a central value of 10, an upward one standard deviation of 3, and a downward one standard deviation of 2. This was computed with 10^6 Monte Carlo samples implementing the "MCLimit" method of interpolation and extrapolation.

[25] One caveat to this statement is that simulated data used to train learning algorithms may not be used for statistical inference. This increased the uncertainty due to limited simulated data, often called "statistical uncertainty" in particle searches.

A systematic uncertainty may impact more than one of the input variables so that one may express the prior distribution of $\mathbf{X}$ with respect to the systematic uncertainty as $\{X_{1\,-\sigma_{down,1}}^{+\sigma_{up,1}}, \ldots, X_{n\,-\sigma_{down,n}}^{+\sigma_{up,n}}\}$. In order to determine the effect of the systematic uncertainty on the response value, one merely examines the effect of both the upward shift, and then the downward shift: $y + \sigma_{up,y} = f(X_1 + \sigma_{up,1}, \ldots, X_n + \sigma_{up,n})$, and $y - \sigma_{down,y} = f(X_1 - \sigma_{down,1}, \ldots, X_n - \sigma_{down,n})$. In this manner, systematic uncertainties were propagated through the predictor functions of the input variables.

The validity of this method of systematic uncertainty propagation was illustrated with a concrete example shown in figure 5.50. This example examined the propagation of a single systematic uncertainty through a complicated function of three input variables. The systematic uncertainty affected the input variables differently, so that each was modeled with a different asymmetric Gaussian-like distribution. Note that the input variables were perfectly correlated. For example, if a single random sample of the systematic uncertainty represented a +1.5 standard deviation upward, then X_1 was shifted upward by 4.5, X_2 by 3, and X_3 by 7.5.

Ten million Monte Carlo samples were generated to provide the distribution of $f(X_1, X_2, X_3)$ (implementing "MCLimit" interpolation and extrapolation). The mean, upward standard deviation and downward standard deviation of $f(X_1, X_2, X_3)$ were then computed and compared to the systematic uncertainty propagation method. The results of the two methods were shown to agree.

The example of figure 5.50 showed the result of propagating a single systematic uncertainty. However, all the systematic uncertainties of the particle search were independent from one another, and so the effect of each systematic uncertainty on the predictor function response value was computed separately for each uncertainty. Note that this method of propagating systematic uncertainties was also valid for the reconstructed mass discriminating variables, which were also complicated functions of several input variables.[26,27]

After the systematic uncertainties were propagated into the discriminating variables, the systematic uncertainties were examined for each of the background types within each of the bins of the discriminating variable spectra.[28] It was observed that some background types shared common relative systematic uncertainty in their expected yields for each of the discriminating variable bins. This allowed these backgrounds to be combined in order to reduce the computational complexity of computing upper limits.[29] Table 5.22 shows

[26]The input variables used for mass reconstruction were expressed as energy-momentum four-vectors ("four-momenta"). If p_1^μ, p_2^μ and p_3^μ were three four-momenta used to reconstruct an invariant mass, then the complicated, nonlinear function of the twelve input variables was given as:

$$f(p_1^1, p_1^2, p_1^3, p_1^4, p_2^1, p_2^2, p_2^3, p_2^4, p_3^1, p_3^2, p_3^3, p_3^4) = \sqrt{\eta_{\mu\nu}(p_1^\mu + p_2^\mu + p_3^\mu)(p_1^\nu + p_2^\nu + p_3^\nu)},$$

where $\eta_{\mu\nu}$ was the Minkowski metric.

[27]One technical detail is that simulated data used for machine learning training must also be removed from the estimation of systematic uncertainties.

[28]See footnote 9.

[29]Rather than examining histograms such as shown in section 5.8, a large number of tables were created (unreported here) that showed the relative systematic uncertainty in each of the discriminating variable bins for each background type and each signal mass. I.e., these tables showed every value that was to be input into the MCLIMIT software that was used to compute the upper limits [245] using the DZero recommended test statistic and the Null treatment of nuisance parameters (discussed in chapter 6). These were visually inspected in order to determine trends that would allow the reduction of the number of

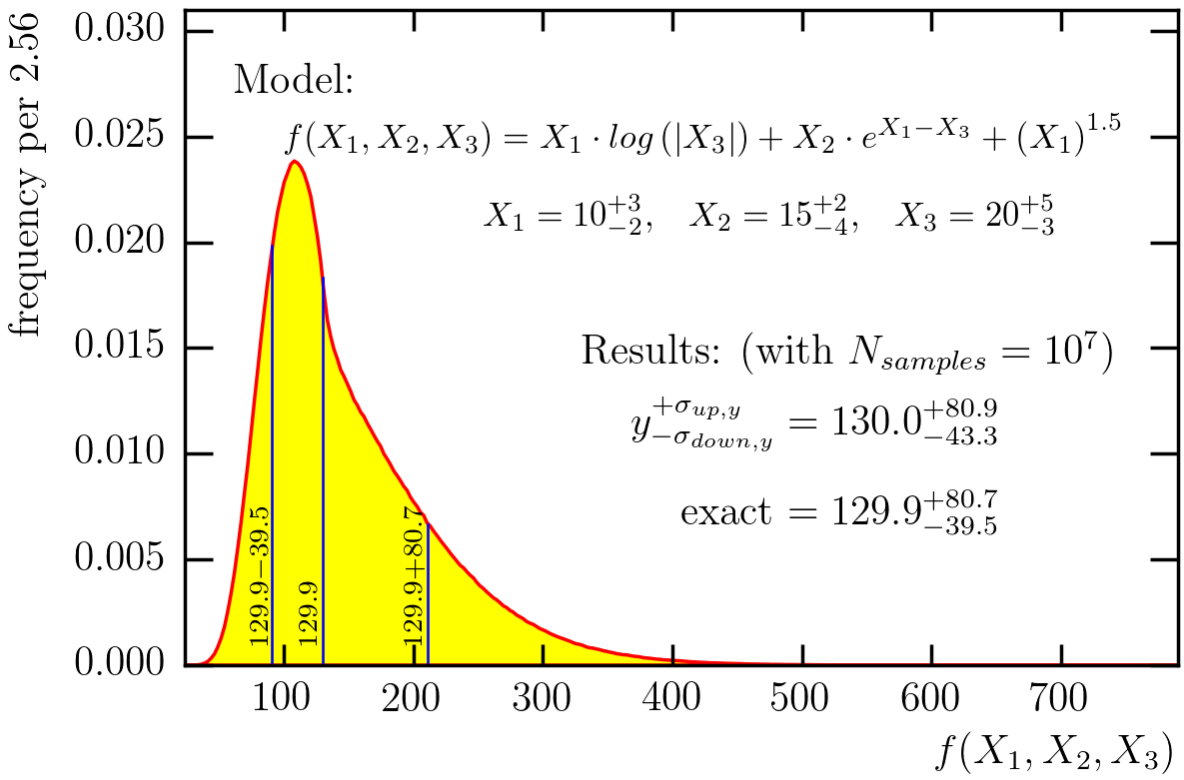


Figure 5.50: A concrete example of how systematic uncertainties propagate through a function of the input variables. The model was a complicated function of three input variables, X_1, X_2 and X_3, where each input variable was modeled by an asymmetric Gaussian-like distribution of a single systematic uncertainty (i.e., the three variables were perfectly correlated). The formula for $y^{+\sigma_{up,y}}_{-\sigma_{down,y}}$ was obtained by simply plugging in the one standard deviations of the input variables into the function. This was compared to 10^7 Monte Carlo simulations that estimated the exact distribution (yellow area).

the seven backgrounds types that were combined into three combined background types for computing upper limits.

Next the 23 systematic uncertainties for the three combined backgrounds were combined into five groups to further reduce the computational complexity. These groups were defined as:

1. ***jet reconstruction*** — The uncertainties associated with estimating the jet physics objects (four combined).
2. ***flavor tagging*** — The uncertainties associated with identifying the jet flavor (three combined).
3. ***lepton reconstruction + luminosity*** — The uncertainties of estimating the lepton physics objects plus the uncertainty in the measurement of the total integrated luminosity (11 combined).
4. **Z + jets *corrections*** — The uncertainties that were generated in estimating corrections to the $Z + jets$ background (four combined).
5. ***cross section*** — The uncertainty of the background cross sections was not combined with other systematic uncertainties.

background type	combined background
$Z + light$ $Z + charm$ $Z + bottom$	$Z + jets$
$t\bar{t}$ $single\ top$	$t\bar{t} + single\ top$
$diboson$ $t\bar{t}X$	$diboson + t\bar{t}X$

Table 5.22: The reduction of nuisance parameters achieved by combining background types with similar relative systematic uncertainties and by combining uncorrelated systematic uncertainties according to their similarities. 7 background types were reduced to 3, and 23 systematic uncertainties were reduced to 5. (Other backgrounds with insignificant contributions to the total background, <0.01%, were merged with $diboson + t\bar{t}X$).

systematic uncertainties. As discussed in section 6.2.1, each systematic uncertainty is modeled by a nuisance parameter. Many nuisance parameters increase the complexity of computing upper limits. For example, a discriminating variable spectrum with 5 bins, 3 background types, and 6 global systematic uncertainty types results in $5 \cdot (3 + 1) + 6 = 26$ nuisance parameters (5 statistical plus 6 global systematics for 3 background plus 1 signal mass), and noting that the upper limits must be computed separately for each individual signal mass.

Whereas the backgrounds were combined by simple addition,[30] combining systematics required performing Monte Carlo simulations. For each group of systematics to be combined, the Monte Carlo simulations of the systematics were performed with the same method as applied by the MCLIMIT software in modeling systematic uncertainties.[31] These simulations provided the distribution of the expected yields for each bin of the discriminating variable (performed separately for each combined background group). From these bin distributions, the combined one standard deviation upward and downward shifts were obtained.

In terms of computation complexity, the grouping of backgrounds and systematic uncertainties reduced the number of specific numerical values that were entered into the MCLIMIT software from more than 2300 to ∼240 (for each signal mass tested) giving ∼90% reduction.[32]

Tables 5.23 (a) and (b) show the systematic uncertainties in the square cuts signal regions after combining backgrounds and systematic uncertainties. The first row of each table shows the expected yields for reference. The row labeled "*MC stat.*" shows the uncertainty due to limited simulated data (often called Monte Carlo statistical uncertainty in particle searches). Tables 5.24 (a)–(d) shows the systematic uncertainties of the machine learning signal regions.

These results indicated that because $Z + jets$ was the dominant background, its systematic uncertainties had the greatest impact on the magnitudes of the upper limits discussed in the next section. They also show that jet and lepton reconstruction were the largest two systematic uncertainties for all signal regions.

[30] Given two backgrounds, A and B, to be combined by simple addition, let b_i^A and b_i^B be the expected background yields in bin i of the discriminating variable, let $\sigma_i^{A,u}$ and $\sigma_i^{B,u}$ be the one standard deviation upward shifts of a systematic uncertainty, and $\sigma_i^{A,d}$ and $\sigma_i^{B,d}$ the downward shifts. Then the combined background expected yield is $b_i = b_i^A + b_i^B$ with uncertainties $\sigma_i^u = \sigma_i^{A,u} + \sigma_i^{B,u}$ and $\sigma_i^d = \sigma_i^{A,d} + \sigma_i^{B,d}$. This is valid because σ^A and σ^B are perfectly correlated (for the same systematic) so that

$$\sigma = \sqrt{(\sigma^A)^2 + (\sigma^B)^2 + 2 \cdot cov(A,B)} = \sqrt{(\sigma^A)^2 + (\sigma^B)^2 + 2\sigma^A\sigma^B} = \sigma^A + \sigma^B.$$

[31] All systematic uncertainties were assumed to be independent of one another. From section 3.3 of [244], the additive model was used (with resampling for negative yield outcomes), and using mclimit interpolation and extrapolation (the most conservative option). 100,000 simulations were performed for each group of systematics to be combined (per each combined background group).

[32] This accounted for all symmetric and asymmetric systematic uncertainties, their values for all background types (plus one signal) in each discriminating variable bin (assuming 8 bins), and the nominal and statistical uncertainties of the background types and signal in each bin. This count was for a single signal mass within a single search strategy (there were ∼200 combinations). A cluster of 51 nodes required over six weeks to compute the upper limits even after this reduction (when profiling the nuisance parameters).

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	24.0	5.0	3.7	32.6
MC stat. (%)	5.0	24.4	8.2	5.6
jet reco. (%)	7.8	16.1	7.0	9.0
flavor tagging (%)	5.0	7.9	7.7	5.7
lepton reco. + *lumi.* (%)	2.9	8.4	9.8	2.2
cross section (%)	1.0	5.7	29.4	4.2
$Z + jets$ *corrections* (%)	8.6	N/A	N/A	6.3

(a) T_{pair} & B_{pair}

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	47.0	6.1	2.9	55.9
MC stat. (%)	4.1	20.8	10.0	4.3
jet reco. (%)	11.6	17.1	11.2	12.2
flavor tagging (%)	3.5	3.9	8.7	3.8
lepton reco. + *lumi.* (%)	2.9	8.4	9.8	2.2
cross section (%)	0.6	5.7	29.7	2.1
$Z + jets$ *corrections* (%)	7.2	N/A	N/A	6.0

(b) T_{single} & B_{single}

Table 5.23: Summary of the systematic uncertainties of the square cuts signal regions.

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	21.4	3.0	3.3	27.7
MC stat. (%)	6.1	38.6	10.3	6.9
jet reco. (%)	8.5	1.3	8.1	7.6
flavor tagging (%)	5.6	14.4	6.9	6.7
lepton reco. + *lumi.* (%)	7.4	6.8	4.9	7.1
cross section (%)	1.1	5.9	29.3	4.1
$Z + jets$ *corrections* (%)	9.1	N/A	N/A	7.0

(a) T_{pair}

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	21.5	2.0	3.4	26.9
MC stat. (%)	6.2	47.8	10.7	6.8
jet reco. (%)	9.4	12.1	7.2	9.3
flavor tagging (%)	5.1	8.7	4.7	4.9
lepton reco. + *lumi.* (%)	3.6	43.3	23.7	9.1
cross section (%)	1.2	6.2	29.1	4.1
$Z + jets$ *corrections* (%)	9.9	N/A	N/A	7.9

(b) B_{pair}

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	14.4	1.2	1.5	17.1
MC stat. (%)	9.1	54.5	16.0	10.0
jet reco. (%)	13.0	6.1	8.4	12.1
flavor tagging (%)	3.4	7.3	8.0	3.8
lepton reco. + *lumi.* (%)	6.3	3.3	2.4	5.7
cross section (%)	0.9	5.3	29.3	2.9
$Z + jets$ *corrections* (%)	8.4	N/A	N/A	7.1

(c) T_{single}

	$Z + jets$	$t\bar{t} +$ *single top*	$diboson + t\bar{t}X$	total
expected yields (#)	48.1	1.6	5.7	55.4
MC stat. (%)	4.3	50.2	9.7	4.3
jet reco. (%)	7.8	12.8	7.3	7.6
flavor tagging (%)	7.4	4.8	7.2	7.3
lepton reco. + *lumi.* (%)	3.1	12.5	1.0	3.1
cross section (%)	1.2	5.4	29.7	3.2
$Z + jets$ *corrections* (%)	8.9	N/A	N/A	7.7

(d) B_{single}

Table 5.24: Summary of the systematic uncertainties of the machine learning signal regions.

5.10.2. Final Results – CL_S Upper Limits

The goal of machine learning was to provide smaller expected upper limits than were obtained using the traditional square cuts search strategy. The upper limits of a search strategy were most dependent on the power of the signal-background separation provided by the strategy's signal region and discriminating variable. However, they were also affected by the amount of simulated data available for statistical inference, and the size of the systematic uncertainties.

The machine learning searches had less simulated data available for statistical inference than the square cuts searches because a portion of the simulated data was used for machine learning training (explained in section 5.4, page 124).[33] In order to examine if the loss of simulated data to training negatively impacted the upper limits of the machine learning searches (in comparison to the square cuts searches), the upper limits for the square cuts searches were recomputed after restricting the square cuts searches to the same testing partition data that was available to the machine learning searches.[34]

The length of a confidence interval increases when there is more uncertainty present in an experiment. The many systematic uncertainties in the searches had a substantial impact on the size of the upper limits. Because estimates of the systematic uncertainties were different for each of the searches, it was possible that a particular search strategy resulted in larger systematic uncertainties, and thus larger upper limits. In order to detect this effect, the upper limits were recomputed for all searches without including the systematic uncertainties (often called "stat only limits" in particle searches).

Thus the upper limits were computed in four different ways to examine the effects of loss of training data and the impact of systematic uncertainty as summarized in table 5.25.

Use systematic uncertainties?	Restrict square cuts searches to machine learning testing partition? **no**	**yes**
yes	The complete upper limits using all available information.	Examine the effect of loss of training data for machine learning.
no	Examine the effect of differing systematic uncertainties.	Examine both effects simultaneously.

Table 5.25: The four methods used to compute upper limits in order to examine the effects of the loss of simulated data to training learning algorithms and the differing systematic uncertainties.

[33] Recall from table 5.3 (page 125) that 30% of the simulated data was used for machine learning training for the background and T_{pair} and B_{pair} signal types, while 50% was needed for T_{single} and B_{single}.

[34] These "restricted" square cuts searches required separate modeling validation and estimation of systematic uncertainties (not reported here).

5.10.2.1. Square Cuts Search Results and Measuring Search Power

Figures 5.51 (a)–(d) show the resultant upper limits of the measured cross sections for the square cuts search strategy for the four signal types.[35] The dashed black line shows the expected upper limit for each of the signal masses under the assumption of signal nonexistence.[36] The expected upper limits for the signal masses best described the power of a search strategy, with smaller values indicating that the strategy was expected to produce smaller confidence intervals and thus better able to either detect signal presence or else rule out signal existence (at 95% confidence). The green and yellow bands describe the uncertainty of the expected upper limits.[37] The solid black line shows the observed upper limits that were determined from the measured data. The observed upper limits were found to lie near the expected upper limits (within its uncertainty bounds) for all search strategies, signal types and signal masses. This meant that the results of all searches were consistent with vector-like quark nonexistence.

The smoothness of the upper limits across the signal masses decreased as simulated data insufficiency became more problematic. This can be seen in the T_{single} upper limits in figure 5.51 (c) and to a lesser extent B_{single} in figure 5.51 (d). The number of simulated events entering the square cuts signal region explained the relative amount of smoothness in the upper limits of the four signal types. For example, the number of unweighted simulated events for the 650 GeV/c^2 signals were: $N(T_{pair}) = 2076$, $N(B_{pair}) = 2834$, $N(T_{single}) = 273$, and $N(B_{single}) = 433$.

The solid red lines of figures 5.51 (a) and (b) show the theoretical cross sections of the pair produced vector-like quarks. For the signal masses where the observed upper limits of the measured cross sections were below the theoretical cross sections, vector-like quark existence at those signal masses was said to be "excluded" (from existence with 95% confidence). For example, figure 5.51 (a) shows that the pair-produced vector-like quark was excluded for signal masses below 610 GeV/c^2.

The values of the expected upper limits in relation to the theoretical cross sections described a different measure of the power of a search. These showed the range of signal masses that were expected to be excluded by the search under the assumption of signal nonexistence. For example, figure 5.51 (a) shows that the T_{pair} square cuts search was expected to exclude signal masses below 589 GeV/c^2.

The theoretical cross sections of the singly produced vector-like quarks were not shown because they fell well below all expected and measured signal cross sections. This meant that the searches for T_{single} and B_{single} were not powerful enough to exclude any signal masses. This result was consistent with that obtained in the work presented by ATLAS exotics Z-tag search team that used the standard square cuts strategy [41].

The magnitude of the uncertainty of the expected upper limits (vertical distance of the

[35] Specifically, the discriminating variable spectra were modeled with the multivariate Poisson model (described in section 6.2.1 on page 258). The DZERO recommended test statistic (section 6.2.3.4, page 285) and CL_S method (page 289 and section 6.2.4, page 295) were used in the Neyman construction. Nuisance parameters were treated using the NULL method (section 6.5, page 388).

[36] The upper limits were computed for 13-16 specific signal masses, and exponential interpolation provided upper limit values between these masses.

[37] Loosely referred to as $\pm 1\,\sigma$ and $\pm 2\,\sigma$ uncertainties, respectively. See figure 6.17 (c) on page 297 for the full explanation of the computation of these uncertainty bands.

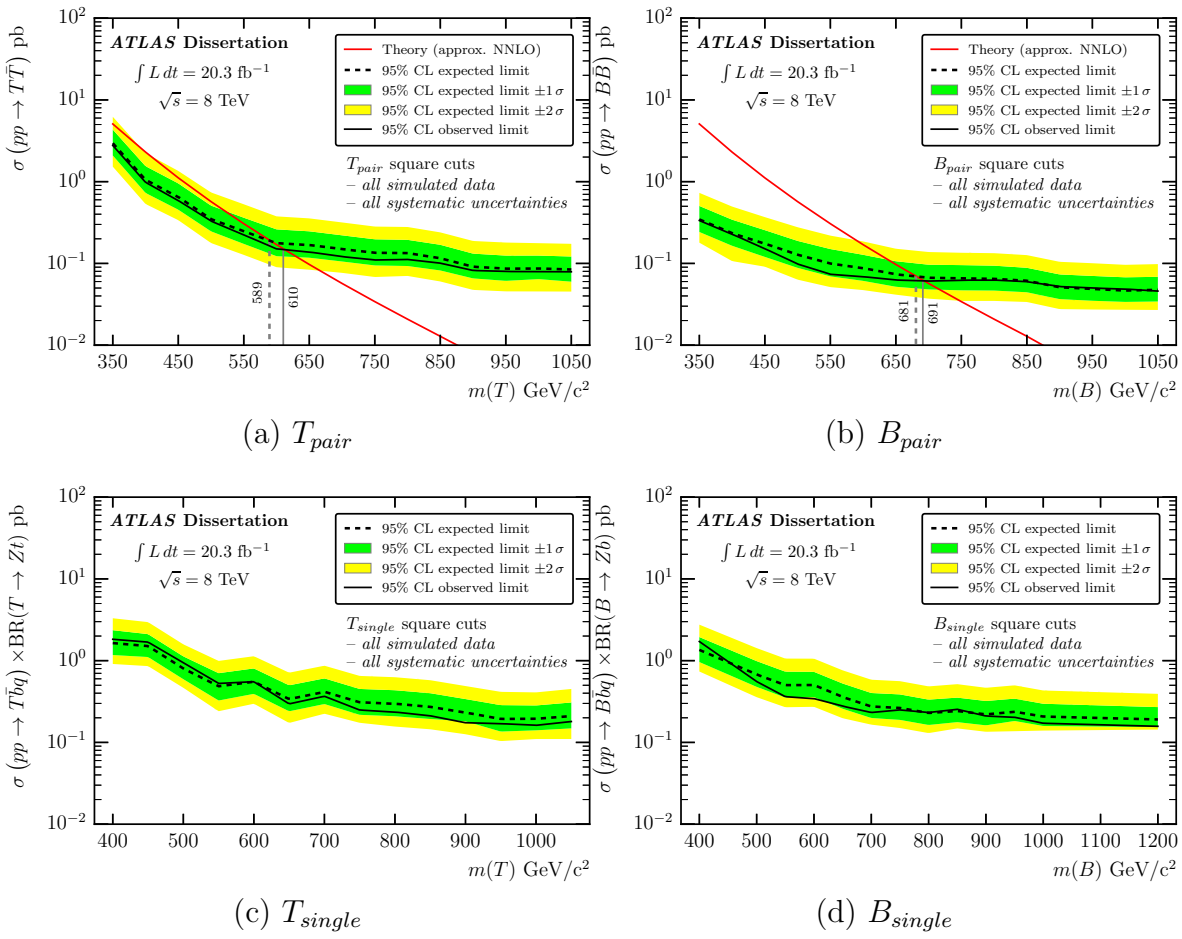


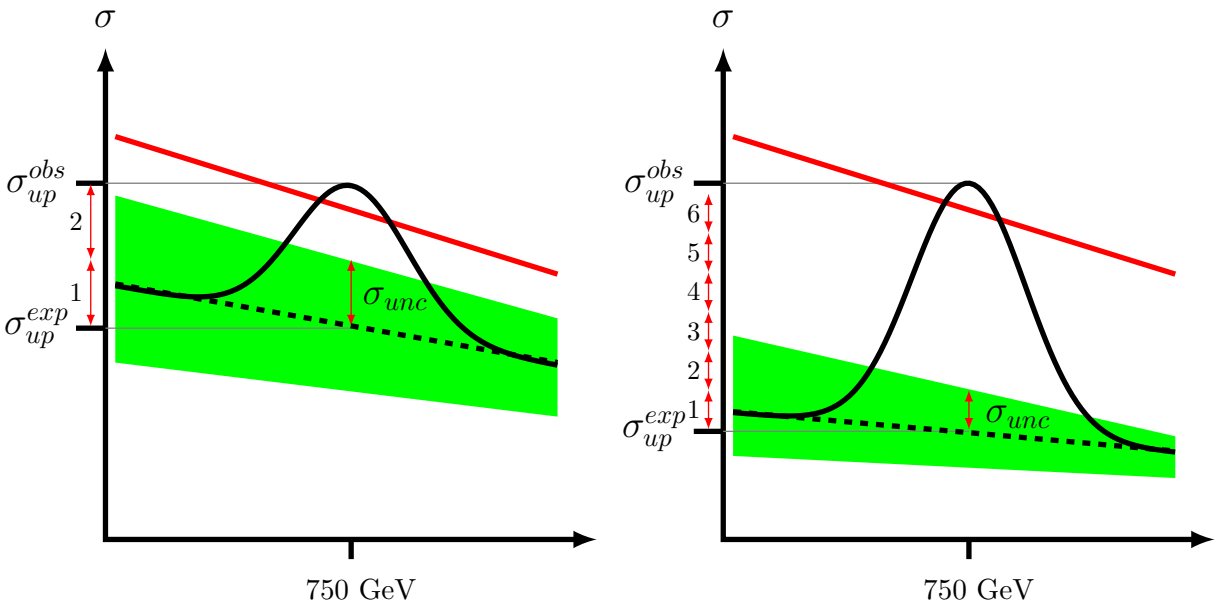
Figure 5.51: The 95% confidence level CL_S upper limits for the square cuts search strategy: the square cuts signal region and the $m(Zb)$ discriminating variable.

green bands, for example) was quite important in describing the power of a search strategy to *detect* the presence of a particle if it did exist. An example illustrates why this was so. Figures 5.52 (a) and (b) show the hypothetical results of two different search strategies for the same particle, which is assumed to exist with an invariant mass of 750 GeV/c^2. Strategy A is not as powerful as strategy B because its expected upper limits are twice as large as strategy B (dashed black lines). Also, the uncertainty of the expected upper limits in strategy A are twice as large as in strategy B (green areas).[38]

Because the particle exists at 750 GeV/c^2, the observed (measured) upper limit at 750 GeV/c^2 reaches the theoretical cross section (also assuming the particle exists with the cross section predicted by theory), and thus a "bump" has been detected. However, the bump found with strategy A is less meaningful because it only represents a two standard deviation from expectations, while the bump found with strategy B represents more than six standard deviations.[39]

[38]It was observed that when the expected upper limits of one strategy were twice as large as another, its upper limit uncertainty tended to be twice as large.

[39]Note that one does not simply count standard deviations in this manner. Nevertheless, this illustrates



(a) search strategy A. (b) search strategy B.

Figure 5.52: Two hypothetical search strategies for the same particle. Assuming the particle exists at 750 GeV/c^2, the observed upper limit, σ_{up}^{obs}, of each search (solid black line) is found above the theoretical cross section (red line) at 750 GeV/c^2. Hence at 750 GeV/c^2, σ_{up}^{obs} is much larger than the upper limit that would be expected if the particle did not exist, σ_{up}^{exp} (dashed black line). Also, σ_{up}^{obs} is outside the one standard deviation uncertainty bound, σ_{unc}, of the expected upper limit at 750 GeV/c^2 (red arrow). I.e., an unexpected "bump" has been discovered. Search strategy A has twice the expected upper limit as strategy B, as well as twice the uncertainty. Loosely speaking, the bump is two standard deviations above the expectations of strategy A, but is more than six standard deviations above the expectations of strategy B, a consequence of both the difference in the expected upper limits *and* the difference in their uncertainties.

Therefore, three measures of search power were used to compare the results of the four search strategies, all based on the expected upper limits:[40]

- the magnitude of the expected upper limits,
- the expected signal mass exclusion range,
- and the one standard deviation uncertainties in the expected upper limits.

The remainder of this section compares the upper limits of the machine learning search strategies to those of the square cuts search strategy. Only comparison figures are shown. All figures showing upper limits are provided in appendix section C.4 (page 512). A discussion of the explanation of these results is given in section 5.10.3 (page 235).

the importance of reducing the upper limit uncertainty.

[40]The expected values were used to compare the power of the searches as a means to evaluate the *design quality* of the search strategies, rather than comparing the observed results.

5.10.2.2. Results of the Two-stage Machine Learning Search

Figures 5.53 (a)–(d) show the ratios of the two-stage machine learning search strategy expected upper limits to those of the square cuts search strategy. Both these strategies used the $m(Zb)$ discriminating variable, and so only differed in the selection of signal regions. For a given signal mass, a ratio of 1.0 indicated no difference in its expected upper limit computed using two-stage classification versus using square cuts.

Limits were computed in four ways as described in table 5.25 on page 223 thus providing four different ratios. The upper limit ratios differed for each signal mass, and so a simple average ratio of all the signal masses was computed. These averages are shown in the figure legends. Blue lines show the upper limit ratios that were computed using all systematic uncertainties, and red lines when computed without the systematic uncertainties. Solid lines show the results when all simulated data was used for the square cuts searches, and dashed lines when the square cuts searches were restricted to the same testing partition data that was used by the machine learning searches.

The blue line of figure 5.53 (a) shows that for the T_{pair} signal type, two-stage machine learning on average performed as well as the square cuts strategy (a value of 1.00 in the legend). The dashed blue line shows that when applied to the same simulated data, two-stage machine learning on average provided a 5% reduction in the expected upper limits (0.95 in legend). Figure 5.53 (b) shows that for B_{pair}, machine learning performed 3% worse than square cuts, although it performed somewhat better at low signal masses and worse at higher masses. Thus for the pair produced vector-like quarks, two-stage classification offered roughly the same search power as the square cuts strategy.

Figure 5.53 (c) shows that two-stage machine learning improved the T_{single} signal type with an average 17% reduction of the upper limits, while figure 5.53 (d) shows that for B_{single}, two-stage machine learning had less search power than the square cuts strategy. A discussion of the explanation of these results is given in section 5.10.3 (page 235).

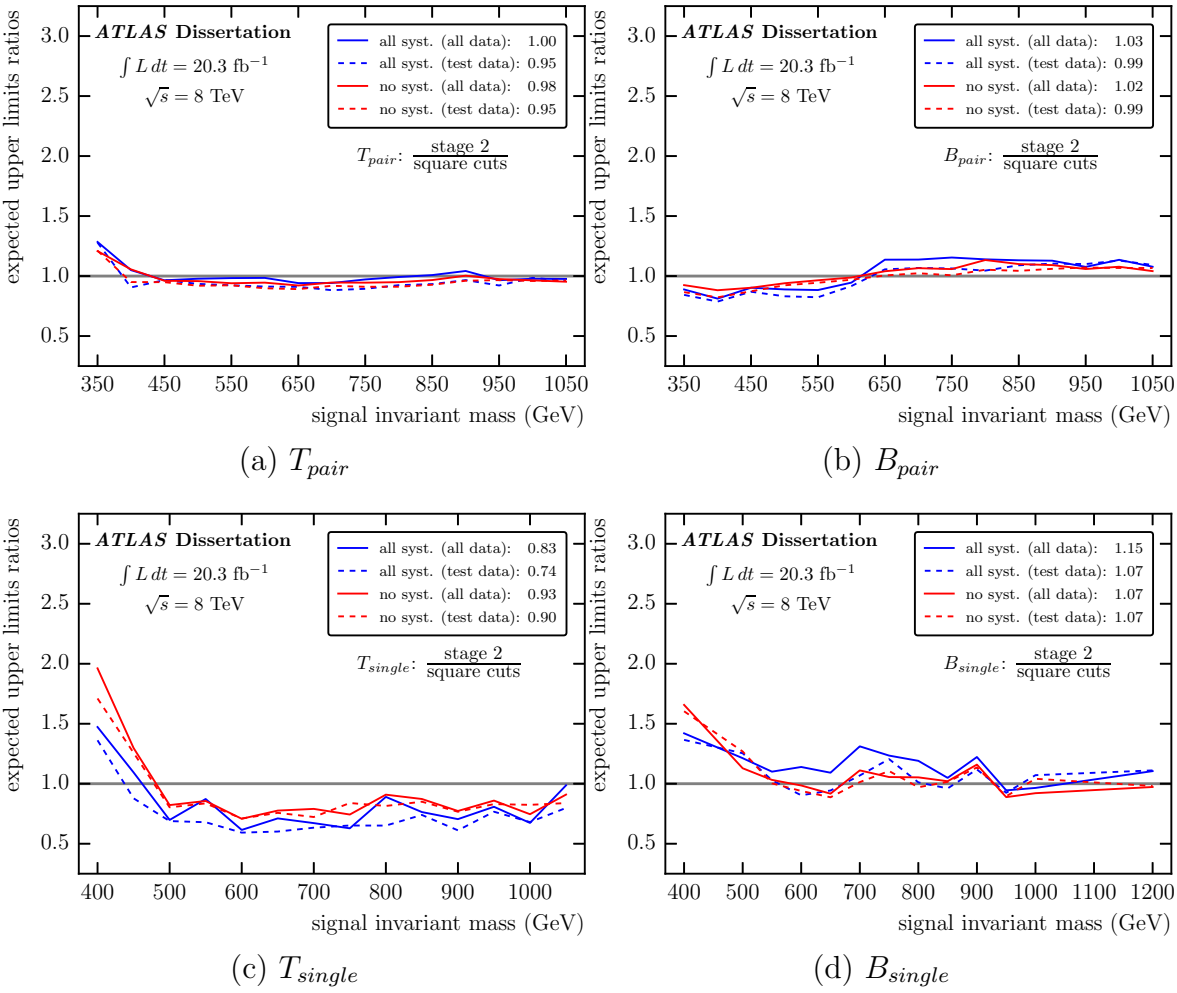


Figure 5.53: The ratios of the 95% confidence level CL_S upper limits: the two-stage machine learning search strategy to the square cuts search strategy. The average values of the ratios are shown in the legend.

Figures 5.54 (a)–(d) show the ratios of the one standard deviation uncertainties of the expected limits. The upper limit uncertainties were asymmetric with the plus one standard deviation larger than the minus one standard deviation, and so these two values were averaged (for example, half the vertical distance of the green band in figure 5.51 (a)). These uncertainty ratios were nearly equal to the ratios of the upper limits of figures 5.53 (a)–(d), indicating that the uncertainty of an upper limit was typically directly proportional to the value of the upper limit itself. For example, if machine learning reduced the expected upper limit by 50%, then it also reduced the upper limit uncertainty by roughly 50%.

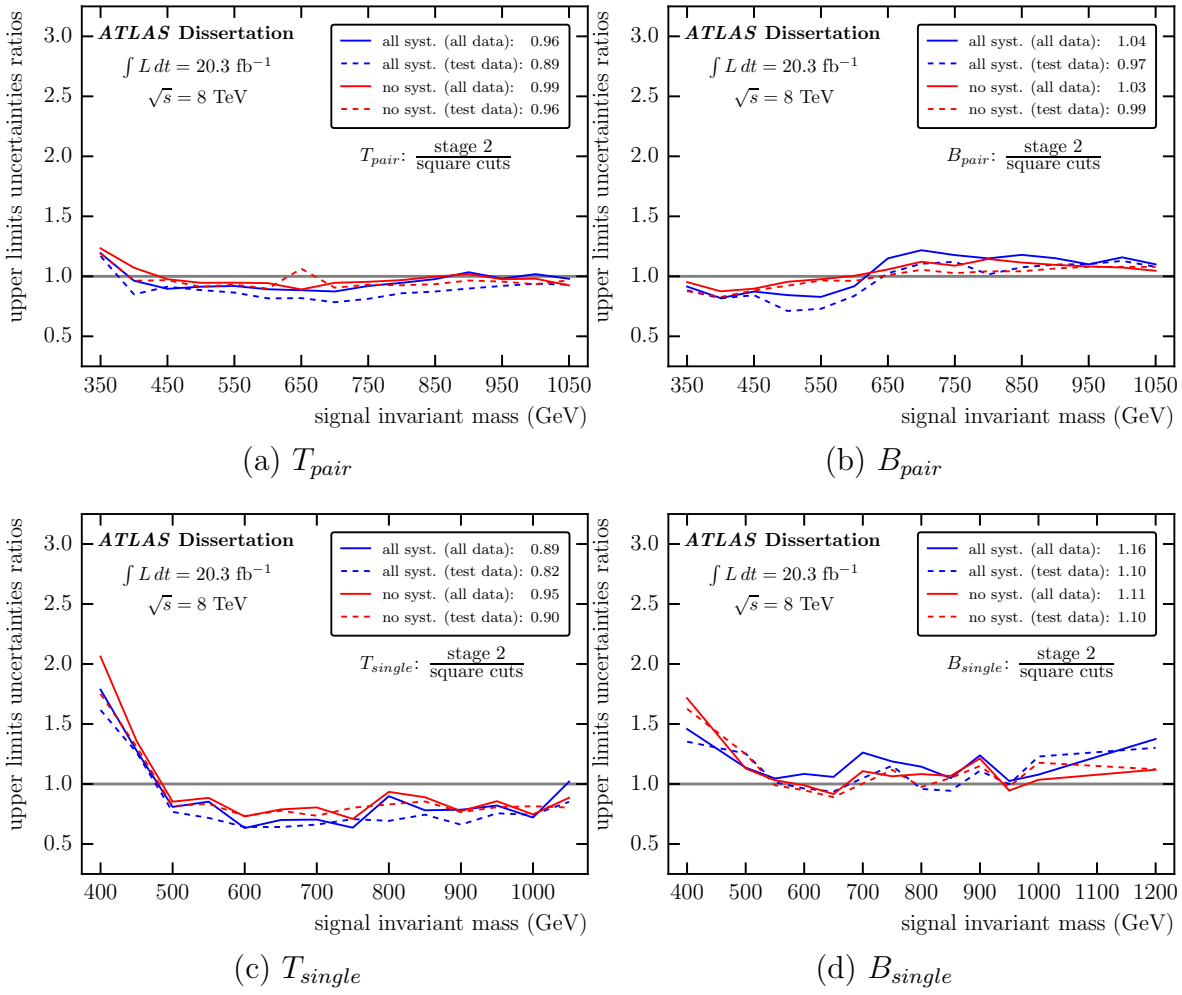


Figure 5.54: The ratios of the upper limit uncertainties: the two-stage machine learning search strategy to the square cuts search strategy.

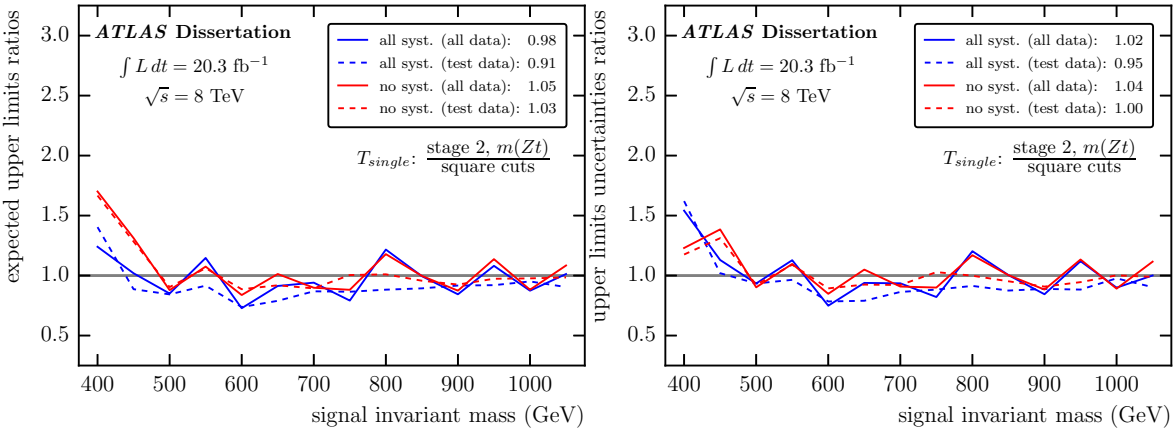
Table 5.26 summarizes the comparison of results between the two-stage machine learning strategy and the square cuts strategy. This includes the ratio of expected upper limits averaged over all the signal masses, as well as for the particular signal mass of 1000 GeV/c^2 that is provided in order to examine the search power at higher signal masses. The maximum of the expected range of excluded signal masses for the two strategies is also provided for the T_{pair} and B_{pair} signal types. These results show that two-stage machine learning search strategy offered little improvement over the square cuts strategy for the pair produced signal types, T_{pair} and B_{pair}, offered a significant reduction in the upper limits of T_{single}, and performed worse than square cuts for B_{single}.

	T_{pair}	B_{pair}	T_{single}	B_{single}
average ratio of expected upper limits				
all syst. (all data)	1.00	1.03	0.83	1.15
all syst. (test data)	0.95	0.99	0.74	1.07
no syst. (all data)	0.98	1.02	0.93	1.07
no syst. (test data)	0.95	0.99	0.90	1.07
1000 GeV/c^2 ratio of expected upper limits				
all syst. (all data)	0.98	1.14	0.67	0.96
all syst. (test data)	0.99	1.13	0.68	1.07
no syst. (all data)	0.96	1.08	0.75	0.92
no syst. (test data)	0.96	1.07	0.82	1.04
max of expected mass exclusion range				
square cuts	589	681	N/A	N/A
two-stage	593	667	N/A	N/A
average ratio of upper limit uncertainty				
all syst. (all data)	0.96	1.04	0.89	1.16
all syst. (test data)	0.89	0.97	0.82	1.10
no syst. (all data)	0.99	1.03	0.95	1.11
no syst. (test data)	0.96	0.99	0.90	1.10

Table 5.26: Summary of the comparison of the two-stage machine learning strategy to the square cuts strategy.

Recall that the $m(Zb)$ discriminating variable only partially reconstructed the vector-like top quark invariant mass. For T_{single}, a full reconstruction was obtained with the $m(Zt)$ discriminating variable. Figures 5.55 (a) and (b) show the comparison of results using two-stage machine learning with $m(Zt)$ as the discriminating variable to the square cuts strategy using $m(Zb)$. The small amount of T_{single} signal simulated data entering the final signal regions caused the lack of smoothness of the upper limits across the signal masses (see the third panel of table 5.10 on page 169), and this was reflected in the ratio comparison figures.

These results indicated that the algorithm used to fully reconstruct the T_{single} mass, $m(Zt)$, did not provide any benefit over the partial reconstruction, $m(Zb)$ (this algorithm is given in footnote 12 on page 170). Comparing figure 5.55 (a) to figure 5.53 (c) shows that $m(Zt)$ performed worse as a discriminating variable than $m(Zb)$. This demonstrated the need for a machine learning solution in cases of difficult invariant mass reconstruction.



(a) T_{single} expected upper limits (b) T_{single} upper limit uncertainties

Figure 5.55: Ratio results for T_{single} two-stage machine learning search strategy (using the $m(Zt)$ reconstructed mass) to the square cuts search strategy. (a) The ratios of the expected upper limits and (b) the ratios of the upper limit uncertainties.

5.10.2.3. Results of the Three-stage Machine Learning Search

Machine learning regression provided a solution to estimating the invariant mass of the vector-like quarks (described in section 5.7 on page 170). Figures 5.56 (a)–(d) show the expected upper limit ratio comparisons of the machine learning strategy using the stage 3 regressand mass as the discriminating variable to the square cuts strategy (that used $m(Zb)$).

For T_{pair}, B_{pair} and T_{single}, the expected upper limits provided by three-stage machine learning were larger at the smallest signal masses, but decreased more quickly with increasing signal mass than with the square cuts strategy, and reached roughly half of those of square cuts for signal masses ≥ 1000 GeV/c^2.[41] For B_{single}, three-stage machine learning performed poorly compared to the square cuts strategy.

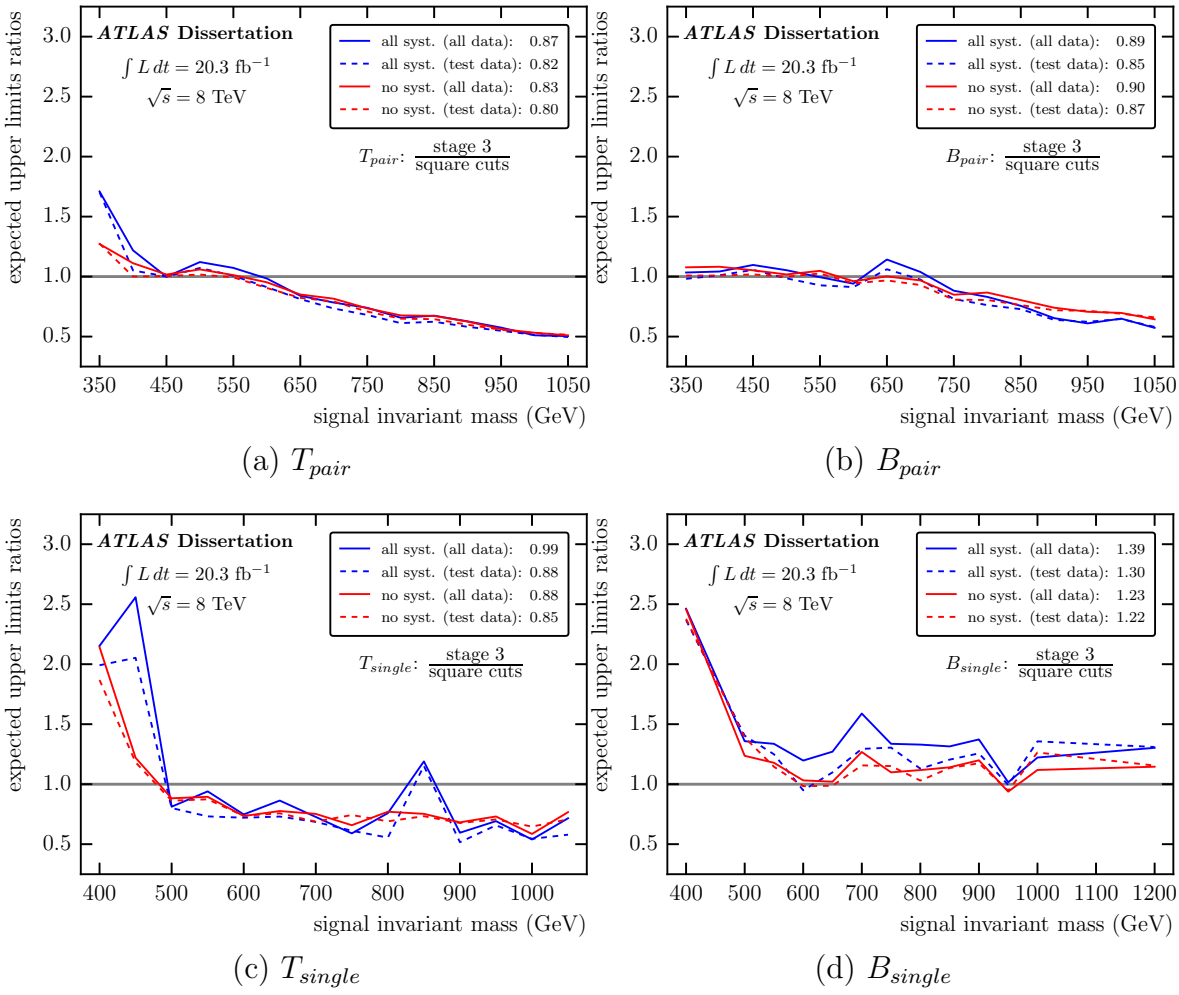


Figure 5.56: The ratios of the 95% confidence level CL_S upper limits: the three-stage machine learning search strategy to the square cuts search strategy. The average values of the ratios are shown in the legend

[41] The "bump" at 850 GeV/c^2 in figure 5.56 (c) was caused by a larger estimated systematic uncertainty in the third bin of the stage 3 regressand mass for the 850 GeV/c^2 signal mass than the neighboring 800 and 900 GeV/c^2 signal masses.

Figures 5.57 (a)–(d) show the comparisons of the upper limit uncertainties. These closely followed the expected upper limit ratios of figures 5.56 (a)–(d). Recall from figure 5.52 (page 226) that the simultaneous decrease in the expected upper limit *and* the decrease in upper limit uncertainty greatly increases the ability of a search strategy to detect the presence of a signal.

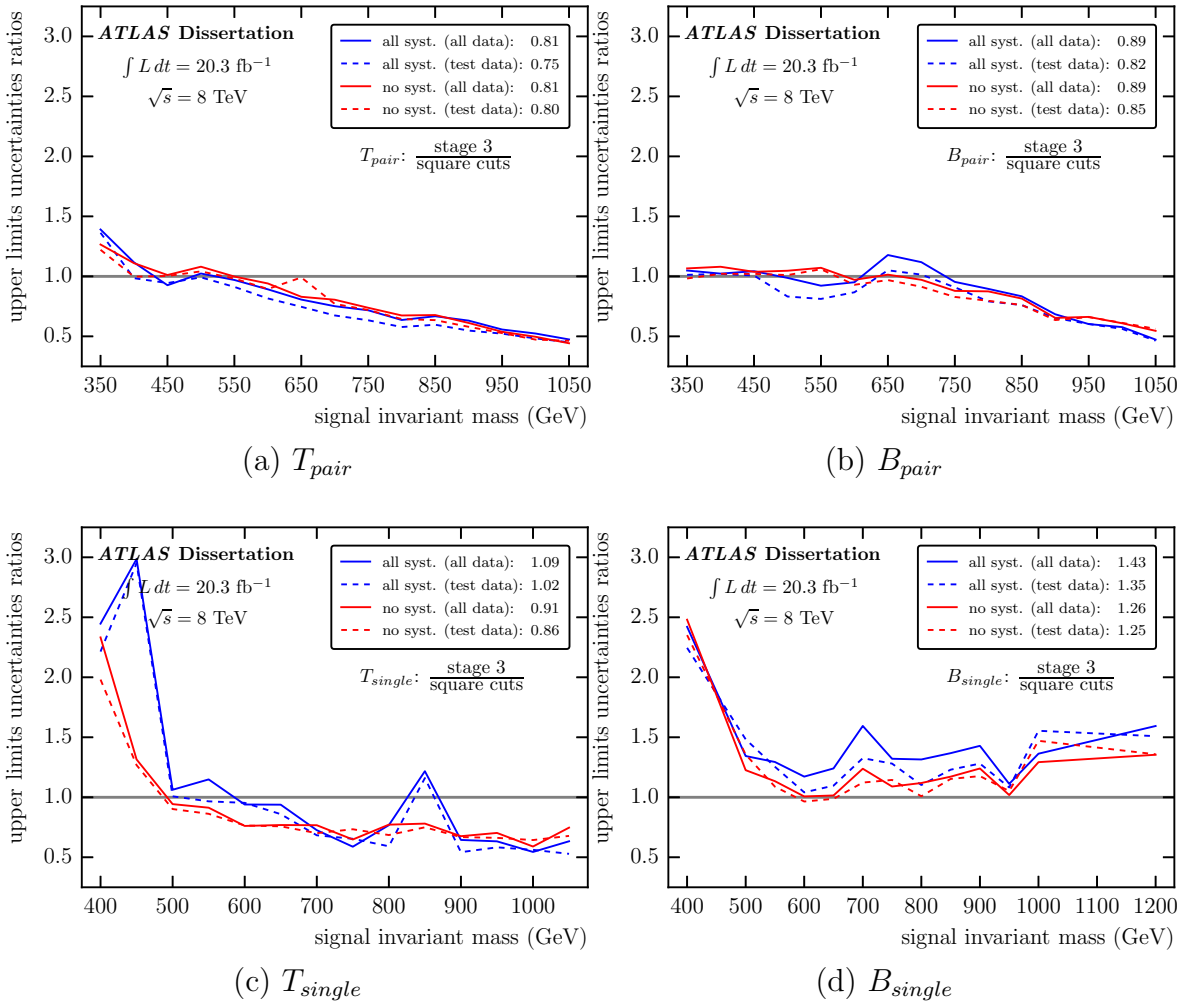


Figure 5.57: The ratios of the upper limit uncertainties: the three-stage machine learning search strategy to the square cuts search strategy. The average values of the ratios are shown in the legend

Table 5.27 summarizes the comparisons of the three-stage machine learning search strategy to the square cuts strategy. For the T_{pair}, B_{pair} and T_{single} signal types, the three-stage machine learning strategy outperformed the square cuts strategy on average. More importantly, the increase in performance grew with increasing signal mass. For example, the expected upper limits of the 1000 GeV/c^2 signal massed for T_{pair} and T_{single} were nearly half as large when using machine learning. The example discussed in figure 5.52 (page 226) illustrates how such an improvement may have a drastic increase in a search's sensitivity.

	T_{pair}	B_{pair}	T_{single}	B_{single}
average ratio of expected upper limits				
all syst. (all data)	0.87	0.89	0.99	1.39
all syst. (test data)	0.82	0.85	0.88	1.30
no syst. (all data)	0.83	0.90	0.88	1.23
no syst. (test data)	0.80	0.87	0.85	1.22
1000 GeV/c^2 ratio of expected upper limits				
all syst. (all data)	0.51	0.65	0.54	1.22
all syst. (test data)	0.51	0.65	0.54	1.36
no syst. (all data)	0.53	0.70	0.59	1.12
no syst. (test data)	0.53	0.69	0.65	1.26
max of expected mass exclusion range				
square cuts	589	681	N/A	N/A
three-stage	589	681	N/A	N/A
average ratio of upper limit uncertainty				
all syst. (all data)	0.81	0.89	1.09	1.43
all syst. (test data)	0.75	0.82	1.02	1.35
no syst. (all data)	0.81	0.89	0.91	1.26
no syst. (test data)	0.80	0.85	0.86	1.25

Table 5.27: Summary of the comparison of the three-stage machine learning strategy to the square cuts strategy.

5.10.3. Discussion of Machine Learning Results

Figures 5.45 (a)–(d) (repeated below for convenience) show that the square cuts signal regions and the machine learning signal regions of T_{pair}, B_{pair} and T_{single} overlapped considerably. Recall that two-stage machine learning used classification to determine signal regions, but used the same $m(Zb)$ discriminating variable as the square cuts search strategy.

For T_{pair} and B_{pair}, the square cuts and machine learning signal regions were very similar (sharing many of the same simulated data events). This explains why there was little difference in the results of the two-stage machine learning and the square cuts search strategies. Since both of these used the same $m(Zb)$ discriminating variable, they were actually very similar searches. This further motivated the use of the machine learning regression to improve the search power of the discriminating variable after machine learning classification failed to significantly improve the selection of the signal regions.

For T_{single}, the machine learning signal region was a subset of the larger square cuts signal region.[42] In this case, machine learning classification was able to identify a subregion of the square cuts signal region that provided better signal-background discrimination (and thus the lower expected limits with the two-stage machine learning strategy).

For B_{single}, the machine learning and square cuts signal regions were quite dissimilar. Two-stage learning performed worse than square cuts so apparently the signal region determined by machine learning classification had poorer signal-background discrimination. The reason for this is explained next.

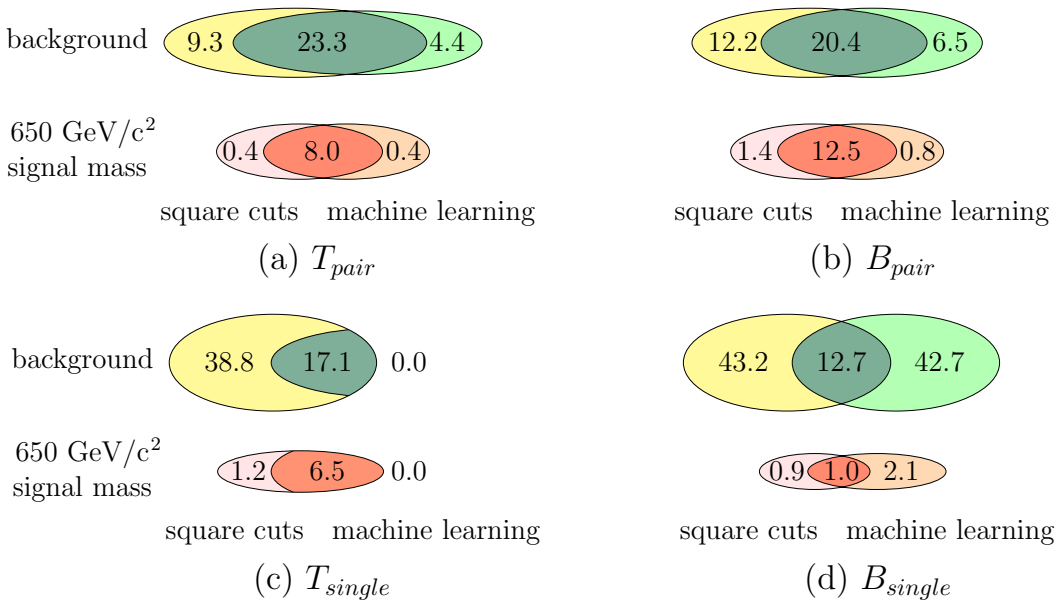


Figure 5.45: Venn diagrams of expected yields showing the similarities between the square cuts signal regions and machine learning signal regions for T_{pair}, B_{pair} and T_{single}, and the differences for those of B_{single}. *(Repeated from page 210.)*

[42] Recall that the requirement $N(forward\ jets) \geq 1$ was added to the T_{single} machine learning signal region, and this caused it to become a subset of the square cuts signal region.

Figure 5.59 shows the frequencies with which the input variables were used to construct the forest of decision trees in stage 1 learning for the 650 GeV/c^2 B_{single} signal (with this mass as a particular example). This shows that the input variables, $p_T(b_1)$, $\Delta R(l_1, l_2)$, $p_T(Z)$, and $\Delta R(Z, b_1)$ (numbered 10, 3, 1, and 5, respectively), were most often used and so likely provided the greatest signal-background discrimination. These input variables related to the Z boson candidate: $\Delta R(l_1, l_2)$ and $p_T(Z)$, to the b quark candidate: $p_T(b_1)$, and to their relation to one another: $\Delta R(Z, b_1)$.[43] For reasons explained in the next paragraphs, many of the other input variables were not likely to contain useful information regarding the decay products of the B_{single} signal type. However, recall that randomization was implemented so that all of the input variables were forced to be used to some extent in the construction of the trees. By forcing the use of less relevant input variables at the expense of useful ones, the power of the two-stage machine learning search was degraded from that of the square cuts strategy, which relied on much fewer but more relevant input variables.

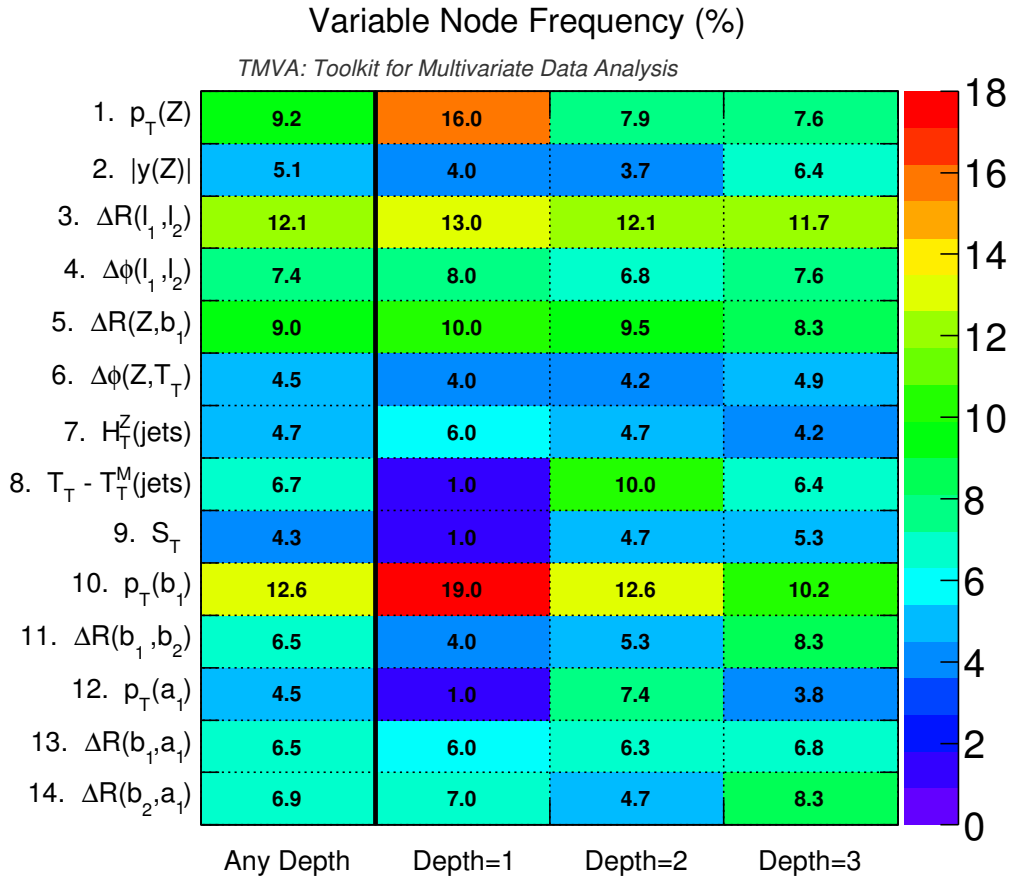


Figure 5.59: The frequencies that the input variables were used in the splitting nodes of the stage 1 partially randomized gradient-boosted decision tree forest for the 650 GeV/c^2 B_{single} signal.

[43] For all B_{single} signal masses, the stage 1 decision trees most often used the input variables: $p_T(b_1)$, $\Delta R(l_1, l_2)$ and $p_T(Z)$. This supports the conclusion drawn from this example of the 650 GeV/c^2 signal mass.

Figures 5.60 (a)–(h) show the decay products of the four signal types. For T_{pair} and B_{pair}, the second vector-like quark (the one not decaying to a leptonically decaying Z boson) decayed to either another Z boson, a W boson, or else a Higgs boson (figures (a)–(c) and (d)–(f), respectively).

Figure 5.60 (h) shows that the decay of B_{single} was much less complex than the other signal types with only one Z boson, two b quarks, and one d quark. This explained why the input variables relating to the Z boson and the b quark with highest tranverse momentum were the ones most frequently used in construction of the decision trees, and why many of the other input variables were less relevant for identifying the B_{single} signal. Because the results of stage 3 regression relied on those of the previous stages, the regression for the B_{single} signal mass also was unsuccessful in improving the search power over the square cuts strategy. The utility of machine learning is its ability to take complicated datasets of input variables and extract information about the complex relationships between those variables. It also seems likely that the simulated data of the B_{single} signal type did not contain enough complexity for machine learning to be useful.

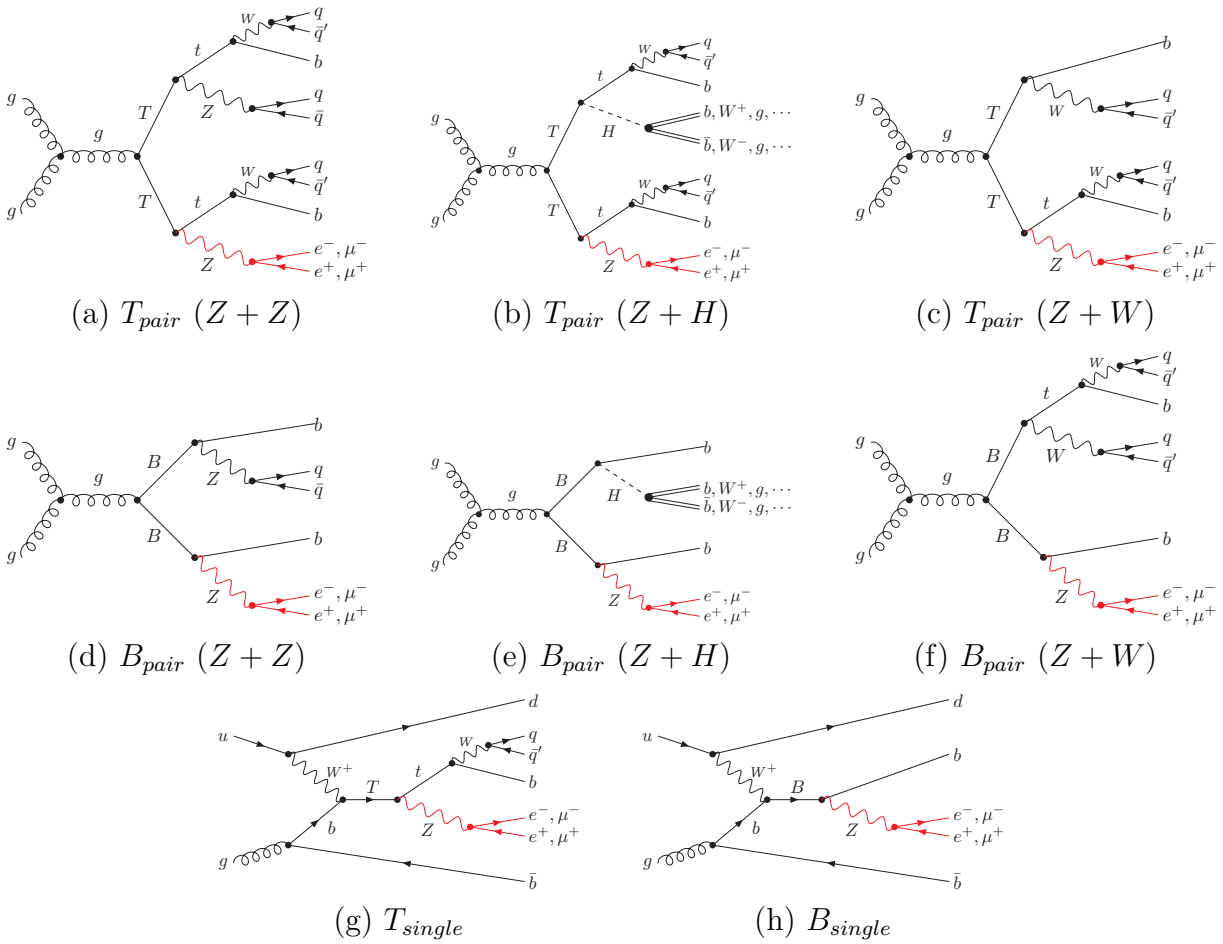
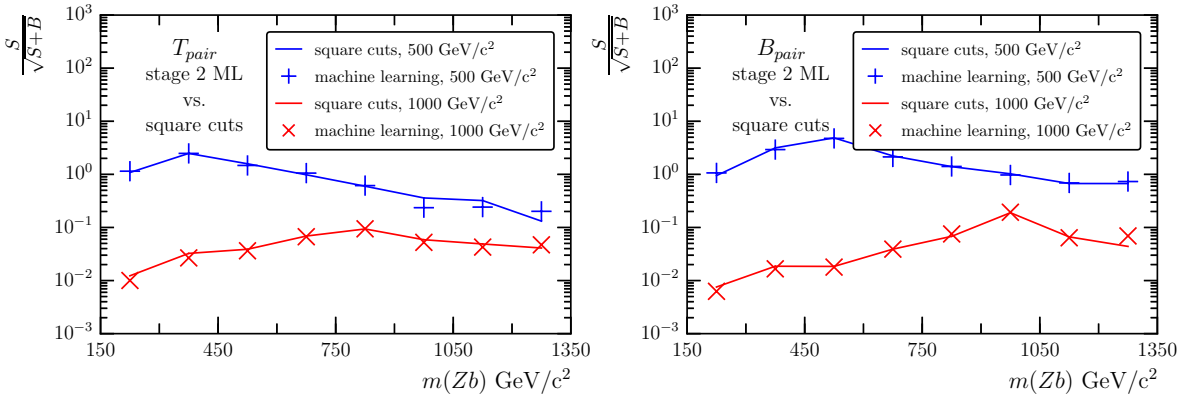


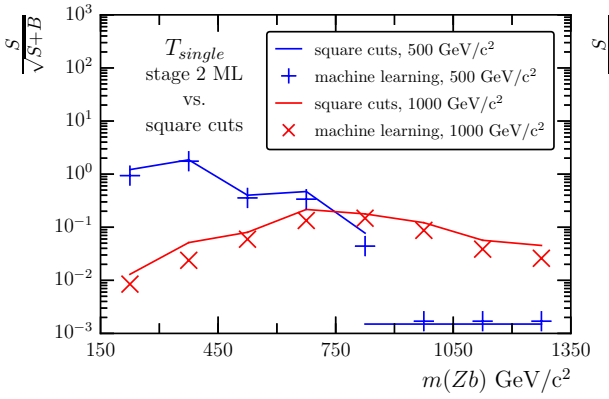
Figure 5.60: Decay topologies for the vector-like quarks containing a leptonically decaying Z boson.

The signal significance, $S/\sqrt{S+B}$ (where S is the number of signal events in the signal region, and B the number of background), may be used as the figure of merit that quantifies the search power of a signal region.[44] Signal significance was previously used to optimize stage 1 learning algorithms. Figures 5.61 (a)–(d) compare the search power of the two signal regions provided by 1) two-stage machine learning ("+" and "×" markers) and 2) square cuts (solid lines). More specifically, these show $S_i/\sqrt{S_i+B_i}$ for the i^{th} bin of the discriminating variable, where both machine learning and square cuts used the same $m(Zb)$ discriminating variable. The number of signal events entering $S_i/\sqrt{S_i+B_i}$ depends on which signal mass is chosen, and so both the 500 GeV/c^2 signal mass (blue) and the 1000 GeV/c^2 signal mass (red) are shown. These indicated that both strategies performed similarly for both these signal masses as suggested by the upper limit results of figures 5.53 (a)–(d) on page 228.

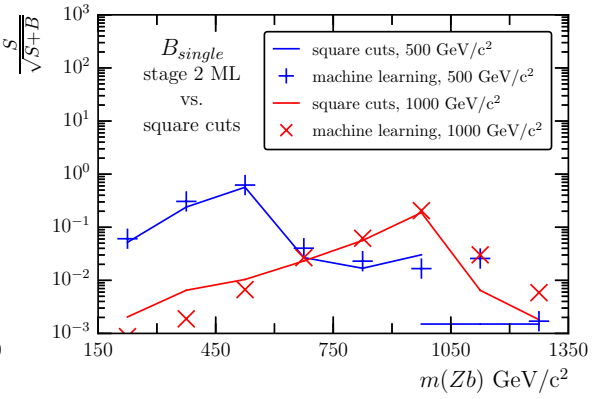


(a) T_{pair}

(b) B_{pair}



(c) T_{single}

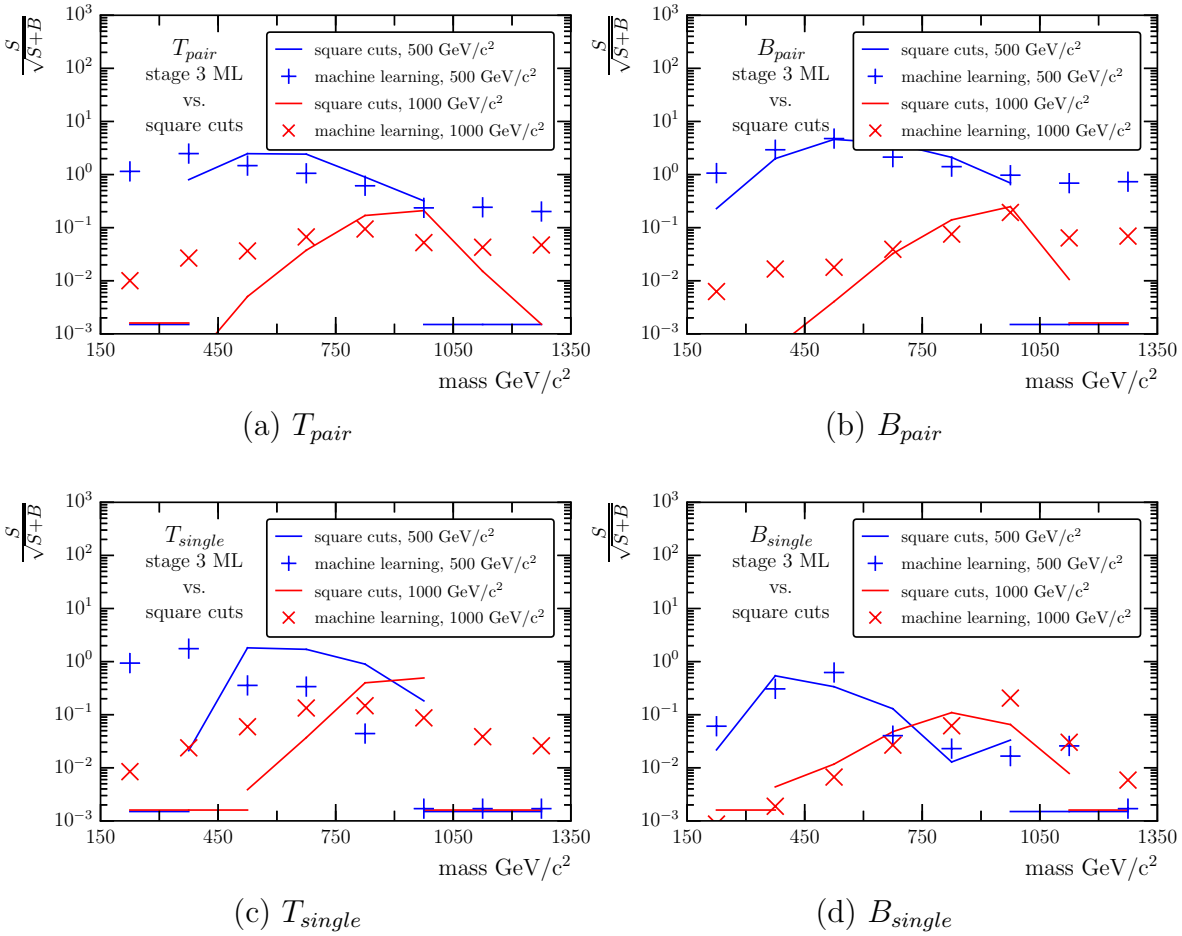


(d) B_{single}

Figure 5.61: Stage 2 machine learning vs. square cuts search strategies for the 500 and 1000 GeV/c^2 signal masses. A comparison of the signal significance, $\frac{S}{\sqrt{S+B}}$, in each bin of the $m(Zb)$ discriminating variable.

[44] The approximate median significance (AMS) of [232] is another choice, but gave very similar results. See footnote 9 on page 138.

Figures 5.62 (a)–(d) show a comparison of the signal significances of the three-stage regression strategy compared to the square cuts strategy, where $m(Zb)$ was used as the discriminating variable for the square cuts searches and the stage 3 regressand was used for machine learning. For T_{pair}, B_{pair} and T_{single}, these demonstrated that the stage 3 regressand provided superior signal significance for the 1000 GeV/c^2 signal masses, leading to lower expected upper limits at 1000 GeV/c^2 as shown in figures 5.56 (a)–(d) (page 232).



(a) T_{pair} (b) B_{pair}

(c) T_{single} (d) B_{single}

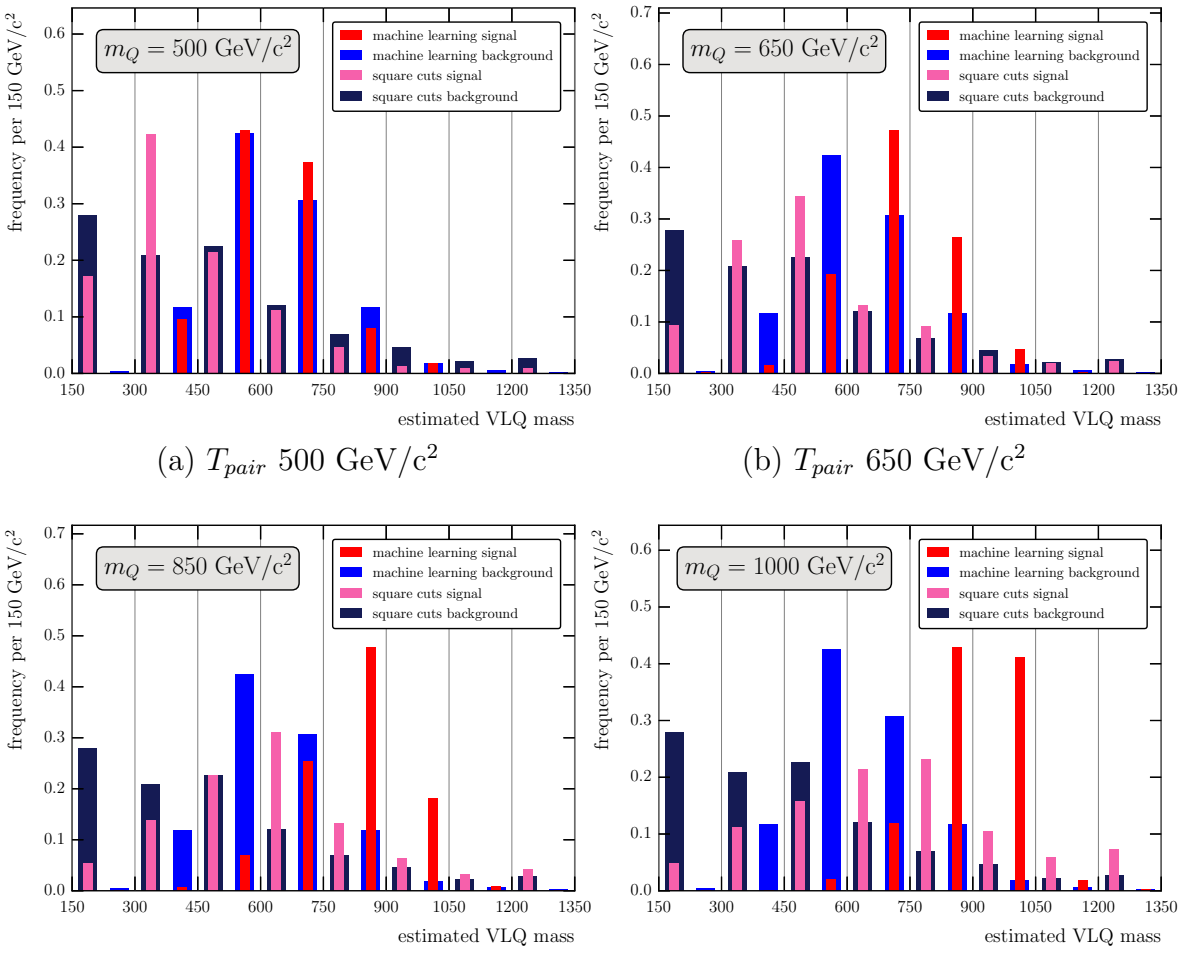
Figure 5.62: Stage 3 machine learning vs. square cuts search strategies for the 500 and 1000 GeV/c^2 signal masses. A comparison of the signal significance, $\frac{S}{\sqrt{S+B}}$ (see footnote 9 on page 138), in each bin of the discriminating variable ($m(Zb)$ for square cuts and the stage 3 regressand for machine learning).

Figures 5.63 (a)–(d) compare the frequency distributions (as histograms) of the discriminating variables used to estimate the VLQ mass spectra: 1) the stage 3 machine learning regressand mass (red and blue), and 2) the $m(Zb)$ reconstructed mass (pink and navy). Figure 5.63 (a) shows these for the 500 GeV/c^2 T_{pair} signal mass, (b) for 650 GeV/c^2, (c) for 850, and (d) for 1000. Vertical gray lines delineate the boundaries between the bins of the discriminating variable so that four histogram bars are shown in each bin.

It is important to know how a discriminating variable performs on the simulated signal vs. the simulated background. Both the stage 3 regressand and $m(Zb)$ reconstructed mass are predictor functions that take a collision event's variables as inputs and return estimates of the VLQ mass. When the predictor function is applied to simulated signal events for a particular signal mass, m_Q, it should return values close to m_Q. When the predictor function is applied to simulated background events, it should return values far from m_Q so that signal-background discrimination is achieved. Thus in figures 5.63 (a)–(d), the stage 3 regressand frequency for the signal mass (red) is overlaid on top of its corresponding background prediction (blue), and the $m(Zb)$ signal (pink) is overlaid on top of its background prediction (navy). Note that the background distributions are independent of the m_Q signal mass (i.e., they are the same for all four figures).

A few observations illustrate why the stage 3 regressand performed better than $m(Zb)$ for high signal masses.

- The stage 3 regressand applied to the simulated signals (red bars) correctly peaked in the bin containing the true signal mass, m_Q.
- The $m(Zb)$ variable applied to simulated signals (pink) peaked below the correct bin because $m(Zb)$ only partially reconstructed the T mass (it missed the contribution from the W boson).
- The stage 3 regressand distributions for signal (red) were more tightly centered around m_Q than the $m(Zb)$ signal distributions (pink).
- The stage 3 regressand applied to background (blue) peaked in the 450–600 bin while the $m(Zb)$ background peaked in the 150–300 bin. The difference in these background distributions ultimately caused the three-stage machine learning upper limits to be *worse* for the low signal masses (the VLQ search began with 350 GeV/c^2 as the lowest signal mass investigated).
- The stage 3 regressand signal for $m_Q = 1000$ GeV/c^2 shown in figure 5.63 (d) provided very strong signal-background separation, which explains why the upper limits for this signal mass were improved by $\sim$50%.

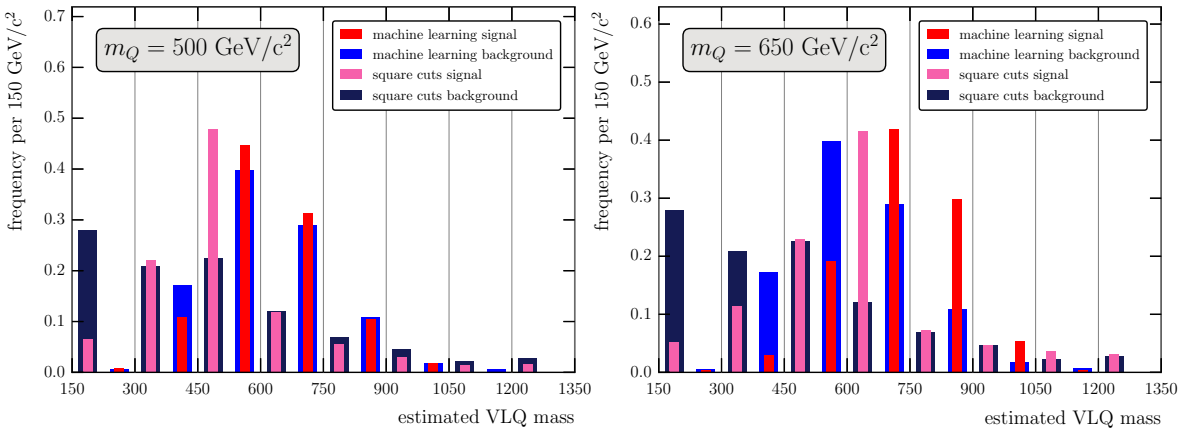


(a) T_{pair} 500 GeV/c^2 (b) T_{pair} 650 GeV/c^2

(c) T_{pair} 850 GeV/c^2 (d) T_{pair} 1000 GeV/c^2

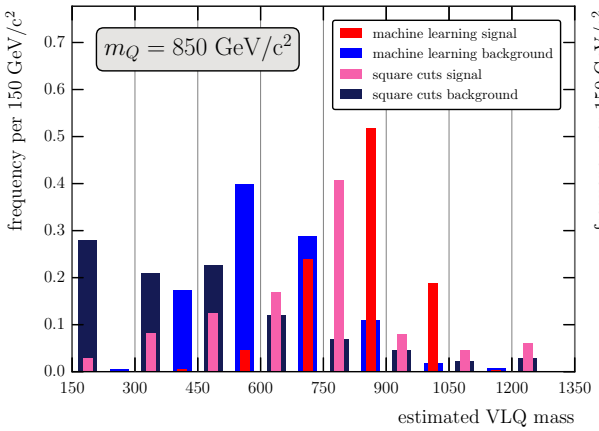
Figure 5.63: Comparison of the frequency distributions of the stage 3 regressand and the $m(Zb)$ reconstructed mass for the T_{pair} signal type. Vertical gray lines show the bin edges with four histogram bars per bin: the stage 3 regressand applied to simulated signals (red) and simulated backgrounds (blue), and the reconstructed mass, $m(Zb)$, applied to signal (pink) and background (navy). Four representative signal masses are shown: (a) 500 GeV/c^2, (b) 650 GeV/c^2, (c) 850 GeV/c^2, and (d) 1000 GeV/c^2.

Figures 5.64 (a)–(d) compare the frequency distributions of the stage 3 regressand and $m(Zb)$ for the B_{pair} signal type. The results for B_{pair} were similar to T_{pair} with one difference. The $m(Zb)$ reconstructed mass applied to signal (pink) correctly peaked in the bin containing m_Q. Nevertheless, figure 5.64 (d) shows that the distribution for the stage 3 regressand applied to signal (red) provided superior signal-background discrimination, explaining why three-stage machine learning gave a ∼40% improvement over square cuts at high signal masses.

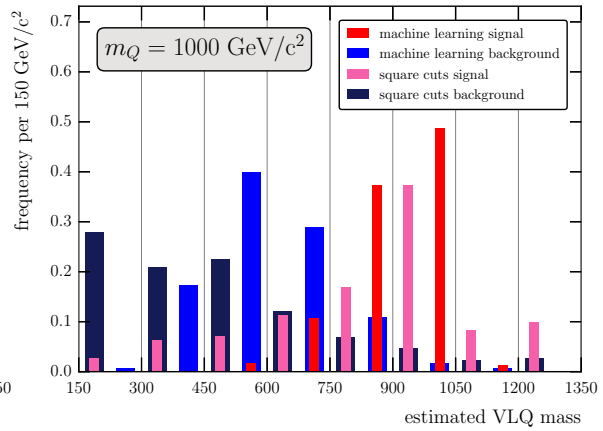


(a) B_{pair} 500 GeV/c^2

(b) B_{pair} 650 GeV/c^2



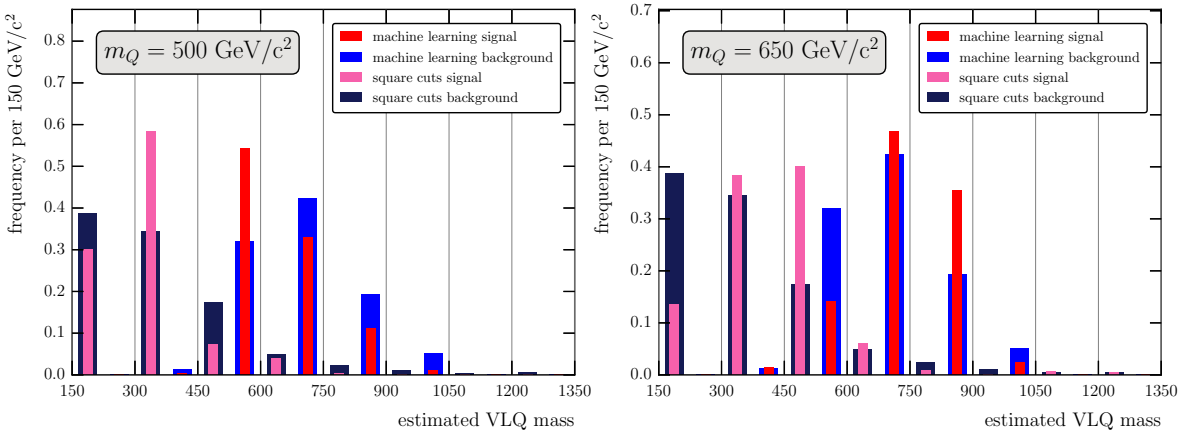
(c) B_{pair} 850 GeV/c^2



(d) B_{pair} 1000 GeV/c^2

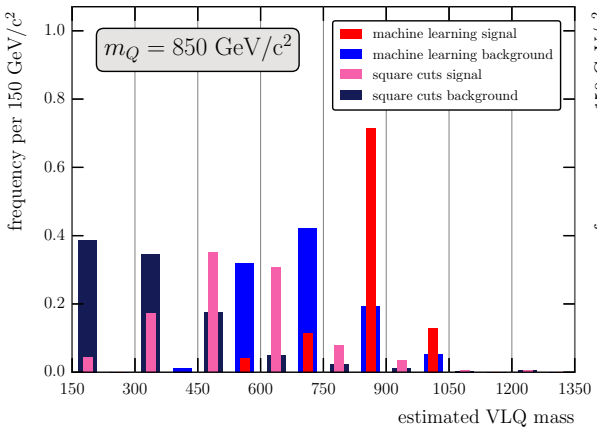
Figure 5.64: Comparison of the frequency distributions of the stage 3 regressand and the $m(Zb)$ reconstructed mass for the B_{pair} signal type.

Figures 5.65 (a)–(d) compare the frequency distributions of the stage 3 regressand and $m(Zb)$ for the T_{single} signal type. The results for T_{single} were similar to T_{pair}.

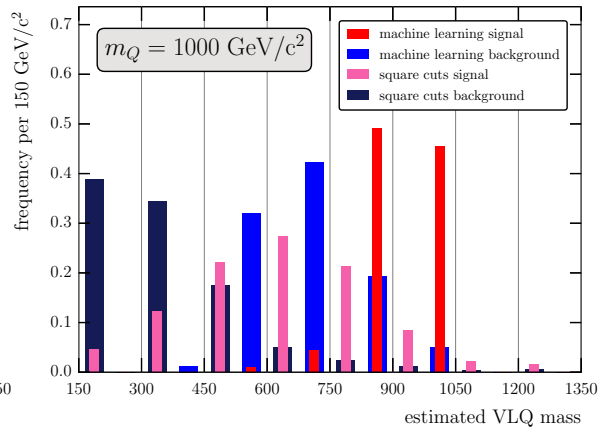


(a) T_{single} 500 GeV/c^2

(b) T_{single} 650 GeV/c^2



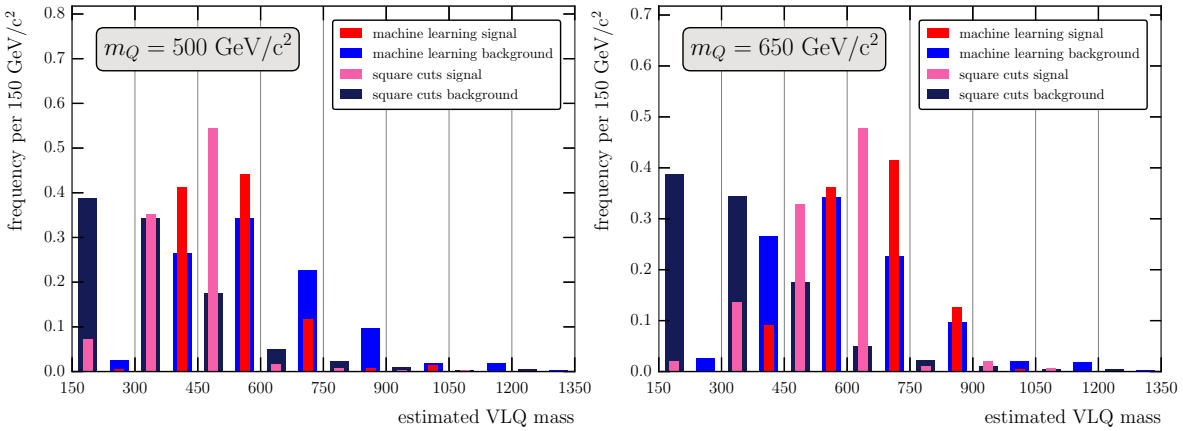
(c) T_{single} 850 GeV/c^2



(d) T_{single} 1000 GeV/c^2

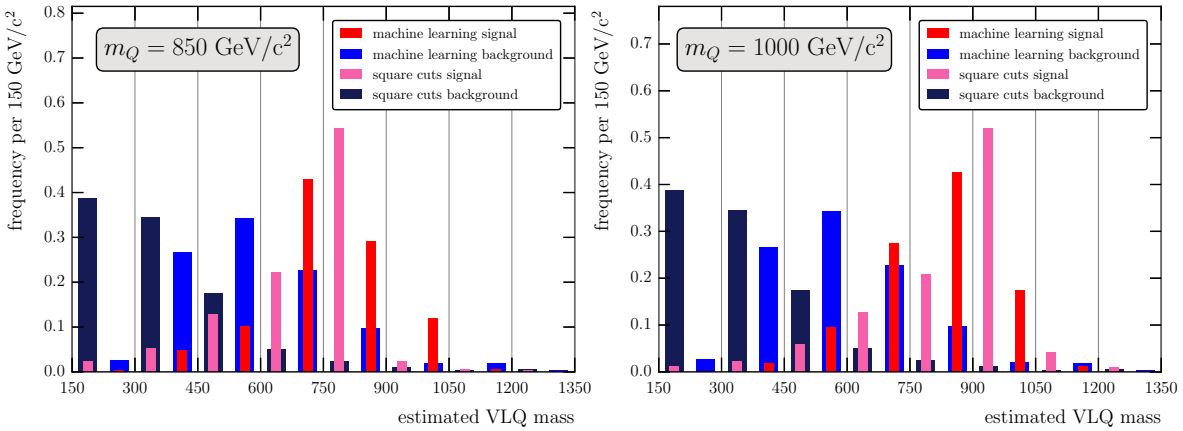
Figure 5.65: Comparison of the frequency distributions of the stage 3 regressand and the $m(Zb)$ reconstructed mass for the T_{single} signal type.

Figures 5.66 (a)–(d) compare the frequency distributions of the stage 3 regressand and $m(Zb)$ for the B_{single} signal type. For B_{single}, the stage 3 regressand mass distributions applied to signal (red) were less tightly centered around m_Q than those for $m(Zb)$ (pink). Additionally, the stage 3 regressand background distribution (blue) was shifted to higher values than those of $m(Zb)$ (navy), leading to worse signal-background discrimination. This caused the three-stage machine learning results to be worse for B_{single} than when using the square cuts search strategy.



(a) B_{single} 500 GeV/c^2

(b) B_{single} 650 GeV/c^2



(c) B_{single} 850 GeV/c^2

(d) B_{single} 1000 GeV/c^2

Figure 5.66: Comparison of the frequency distributions of the stage 3 regressand and the $m(Zb)$ reconstructed mass for the B_{single} signal type.

Therefore, the results of the machine learning strategies were explained by the following observations.

- $\mathbf{T_{pair}}$ and $\mathbf{B_{pair}}$ — Two stages of classification were unable to isolate a signal region with better signal-background discrimination than with square cuts. This was because the backgrounds entering the square cuts and machine learning signal regions were mostly the same, and while the machine learning signal region did have slightly less background, the background that was removed occurred at low values of the $m(Zb)$ discriminating variable, which did not contribute to the computation of the upper limits (compare figure 5.46 (a) to figures 5.48 (a) and (b)). However, machine learning regression was able to provide a better discriminating variable for higher signal masses (and worse for smaller) by shifting more background to lower signal masses of the regressand spectrum.

- $\mathbf{T_{single}}$ — Both the two-stage classification and the third stage of regression yielded better results than the square cuts searches for all but the smallest signal masses. For the two-stage classification, the machine learning signal region was a subset of the square cuts signal region, and unlike T_{pair} or B_{pair}, machine learning classification removed background from the square cuts signal region for higher values of $m(Zb)$. This led to the smaller upper limits of the two-stage strategy. The third stage regression lowered the upper limits even further by providing more discrimination at higher signal masses by 1) shifting the background to lower masses, and 2) providing a more accurate discriminating variable.

- $\mathbf{B_{single}}$ — Both the two and three-stage machine learning strategies performed worse than the square cuts search strategy. The two-stage classification failed because the sparse decay topology of B_{single} meant that many of the input variables were less relevant, but the use of randomization in constructing the decision trees forced these input variables to be used, leading to worse signal-background discrimination than with square cuts. The third stage regression used the flawed results of the first two stages, and thus delivered a discriminating variable that was less accurate than $m(Zb)$, as well as shifting the background to larger mass values.

5.11. Machine Learning – Discussion

The search for vector-like quarks (VLQs) was motivated by the desire to address the hierarchy problem of the Standard Model (SM), and by the fact that VLQs appear in many beyond the Standard Model (BSM) theories. VLQs produced in pairs are model independent so that search results are directly applicable to all BSM theories predicting the existence of VLQs. As with the square cuts search presented in chapter 4, the machine learning search for VLQs yielded no evidence of their existence.

More importantly, novel techniques for applying machine learning to particle physics searches were investigated. This was motivated by an initial failure to improve upon the square cuts search by using single stage machine learning approaches that are common in experimental particle physics. These techniques included

- two-stage machine learning *classification* to determine signal regions,
- a novel third stage of machine learning *regression* that provided more powerful discriminating variables, (including the use of some regression methods more commonly found in econometrics research),
- new statistical methods for the detection of overtraining that allowed the automation of the optimization of learning algorithm hyper-parameters, and
- the exploration of applying *unsupervised* machine learning to determine signal and control regions (reported in appendix section B.2 on page 462).

Several specific results were obtained. First, each new theoretically hypothesized particle comes with a spectrum of possible signal masses to be tested. The two-stage machine learning search strategy provided a method to maximize the power of the search for each mass separately while combining the separate stage 1 classifiers into a single tool. Thus it gave the flexibility of mass specific signal detection with the practicality of a single search for all masses simultaneously. Additionally, the two-stage approach provided a natural partitioning of the simulated data by allowing the training and validation partitions to be switched in the second stage.

A statistical procedure for automating a grid search for optimal learning hyper-parameters was developed. This involved maximizing statistical significance while avoiding overtraining. Brown's method and Fisher's method were used to combine the results of four correlated statistical tests for overtraining (including a new autoregressive model to test for slight "shape shift" overtraining errors) into a test for overtraining.

It was observed that the second stage of machine learning required the use of less powerful learning algorithms. This was necessary to disallow the overtraining that accompanied more powerful stage 2 learning algorithms. Simple logistic regression (logit classification) was found to provide the best results for stage 2 classification.

The results of the two-stage machine learning search strategy did not provide any significant improvement over the square cuts search strategy. This motivated the use of a third stage of machine learning regression to provide a more powerful discriminating variables. Machine learning regression (as opposed to classification) used in this manner has not been observed in the particle search literature. By regressing for the signal mass, the power of the machine learning search grew with the invariant mass of the signal. At the highest signal masses, three-stage machine learning provided a significant improvement

over the square cuts search by reducing the upper limits by *half*. This result implies that the reach of particle searches may be extended to much higher signal masses with this technique.

Several regression approaches were attempted, but the best results were usually obtained with techniques familiar to econometrics. It was also found that only decision trees with gradient boosting provided reliable regression results when compared to other machine learning algorithms (with spurious regression results obtained using adaptive boosted decision trees, multi-layer perceptrons, and support vector machines).

Additionally, a brief exploration of unsupervised clustering offered a novel approach for the identification of signal and control regions. Control regions are usually needed for data-driven corrections to the simulated data, but are frequently difficult to identify. Locating signal regions with high statistical significance is always desirable.

5.11.1. Implications for Future Research

The research presented here represents a first attempt at several novel applications of machine learning to particle physics searches. The results of these new methods justifies further investigation.

5.11.1.1. Two-stage Classification

First, an attempt should be made to combine the two-stage partitioning scheme with the cross evaluation scheme as this may achieve *lossless simulated data* machine learning (discussed in more depth in appendix section B.1). Second, the use of an *autoregressive model* for the identification of shape shift error of two distributions (equations 5.5.9-5.5.10 on page 141) could be useful in particle searches in addition to the more usual tests (χ^2, etc.), but requires further examination.

5.11.1.2. Machine Learning Regression

The regression methods used here relied heavily on techniques commonly found in econometrics: ordinary least squares (OLS), weighted least squares (WLS), tests for heteroscedasticity, heteroscedastic corrected errors, and two-stage generalized least squares regression (as another method of dealing with heteroscedasticity), but simple ordinary least squares provided the most reliable results.

Unlike time series forecasting where off-diagonal terms of the disturbance covariance matrix are usually a primary concern, simulated data in particle searches is uncorrelated by construction, i.e., the disturbance covariance matrix is guaranteed diagonal. Because the particle physics background data arrive from several different processes of varying uncertainty, it is natural to assume that the disturbance covariance matrix is not uniform, even if diagonal. This would suggest the disturbance covariance matrix could be modeled as blocks of diagonal matrices with a constant variance block for each background process. However, even within a given background process, measurements of the various input variables have differing systematic errors, and this motivated the attempt to include heteroscedasticity in a more general manner (by estimating each element of the diagonal covariance matrix separately during two-stage regression). There are many techniques

available to achieve this that should be explored, especially given the strength of the results of the three-stage machine learning searches.

Further complicating matters, simulated data in particle searches have weights that span many orders of magnitude. In WLS, these weights are included in the disturbance covariance matrix. It is therefore unclear why the best results were not obtained with a WLS-based regression model, except perhaps that data insufficiency disallowed the modeling of so much information into the disturbance covariance matrix: event weights, background process specific uncertainty, and input variable systematic uncertainty. If this style of machine learning is to be implemented in particle searches, additional investigation is needed in order to determine the best approach to regression modeling.

5.11.1.3. Unsupervised Learning

As discussed in section 5.3, it has become acceptable in particle searches to define signal regions using the spectrum of a decision tree classifier. Because decision trees are inherently "local learners" that group data with common input variable values into leaf nodes, it seems natural that data with common variable values might be grouped together by other means in order to define signal (or control) regions. Unsupervised clustering techniques may provide such a method as discussed in section B.2 where k-means clustering and Gaussian mixture models were applied.

Unsupervised clustering of simulated background data identifies regions of background data with common variable values that would be impossible to locate manually, and are independent of any signal distributions. Once background clusters are obtained, signal and control regions may be identified as particular clusters according to how much signal data populates each of the clusters. This process might be most useful near the beginning of a search using a small number of elementary kinematic variables to obtain the signal and control regions, with more powerful machine learning tools applied later to the clusters with high signal significance.

Chapter 6
The Neyman Construction in Signal Searches

"If long run properties are taken seriously, anything which might have happened in the study must be taken into account."

– James K. Lindsey [246]

"The loss resulting from an incorrect acceptance of the hypothesis, for example, will not be the same for all alternatives. The more the alternative differs from the hypothesis, the more serious are the consequences of such an error."

– Erich L. Lehmann [247]

"The upper limit [...] *is a particular instance of a* hypothetical *long run of statements, a proportion* $1-\alpha$ *of which will be true, always, of course, assuming our model is sound. We can, at least in principle, make such a statement for each* α *and thereby generate a collection of statements, sometimes called a* confidence distribution*."*

– David R. Cox [248]

In chapter 5, machine learning algorithms are used to construct functions of many input variables in order to predict the presence of a signal above a background. These predictor functions are designed to maximize discrimination between signal and background while preventing overtraining/overfitting that degrades predictive capacity. The algorithms are trained and validated with simulated data, and further evaluated with measured data in control regions where only background is expected to be present. Once a predictor function is deemed to be valid, the predictor function is applied to the simulated background and signal data in the final signal region. The predictor function is also applied to the measured data (although the result of this remains blinded until the completion of the design of the final statistical test). The result of this is three datasets containing the values of a discriminating variable for a series of outcomes: a dataset of measured (observed) outcomes, a dataset of simulated (expected) background outcomes, and a dataset of simulated (expected) signal outcomes.

Once these final datasets are provided by machine learning, these datasets must be compared to see if the measured outcomes are more consistent with the simulated background outcomes, or else more consistent with the combined simulated signal and background outcomes. This chapter discusses the methods used to quantify this comparison: the design of the statistical procedure that provides the final inference result from which conclusions are drawn and decisions are made.

The final result in the search for a signal must be a statistical statement describing how likely the signal is to exist. The result is based on comparing *real-world* observations to

theoretical statistical models which describe the probabilities of obtaining those observations. The result should be *sensible* so that it it is consistent with our understanding of the specific complexities of the experiment at hand (i.e., not solely relying on a general purpose mathematical procedure). The result should also be *useful* in decision making and for the advancement of scientific understanding in general [249].

The beginning quotations describe the three central aims of this chapter in relation to delivering sensible and useful inference results. First, statistical models of the real-world often include possible outcomes that are highly improbable. If any of these improbable outcomes were actually witnessed after the experiment was conducted, the experimenter is rightly justified to have doubts regarding the quality of the statistical modeling (i.e., become suspicious that the experimental design was misspecified). Rather than reject the experiment outright and not report any result, these doubts can be incorporated by modifying the test statistic used for inference so that useful results may still be reported, or if not particularly useful, at least provide results with an unambiguous interpretation. More importantly, if the experimenter is confident that the correct inference statement is provided by the statistical test for *any* possible outcome, no matter how improbable, then the experimenter will not be tempted to alter the test *post hoc* if an unusual outcome is observed.

Second, many statistical procedures utilize likelihood ratios in order to maximize the power of discriminating between the hypothesis that the signal exists with some signal strength μ (defined in section 6.2.1) and the hypothesis that only the background exists (or some other suitable alternative) [250]. In the case of frequentist confidence interval construction for the parameter of interest μ, the use of likelihood ratios as test statistics may provide uniformly most accurate (UMA) confidence intervals, defined as those which maximize the power to exclude incorrect values of μ [251]. Confidence intervals that are UMA have the shortest expected length (defined in section 6.3.4), a useful property when attempting to infer the best values to describe a quantity that is known to exist. However, confidence intervals with shortest length may be far from μ_{true} when they do not contain μ_{true}. It may be more desirable to construct somewhat longer confidence intervals in a manner such that even if μ_{true} is not within the confidence interval, then μ_{true} may be expected to be not too far outside the confidence interval [252]. One may consider this the competing effects of two kinds of risk: the risk to include false values of $\mu \neq \mu_{true}$ in the confidence interval versus the risk of μ_{true} being far from the confidence interval when not contained within the confidence interval. If a loss function is defined that is proportional to the distance of μ_{true} to the confidence interval, and *reporting risk* is defined as the frequentist expectation of this loss, the Arizona test statistics developed here are shown to reduce the risk of reporting a confidence interval that may be distant from μ_{true}.

Finally, if the goal of statistical inference is to aid in making nuanced decisions, such as whether or not to continue expending resources to continue a signal search as new data is obtained, then a frequentist confidence interval at some specified confidence level is not as informative as, for instance, a Bayesian posterior distribution (described in section 6.2). However, frequentist methods can also provide distribution-like information when confidence intervals are constructed at all levels. By providing more information than a single confidence interval, such frequentist methods may be more useful in making decisions based on experimental outcomes, especially in situations where Bayesian analysis is inappropriate. The construction and proper interpretation of frequentist *confidence-acceptance profiles*

(CAP) are developed here, as well as methods for combining CAPs from multiple experiments via meta-analysis.

The frequentist inference method of interval estimation is often more useful in addressing the question of signal existence than significance testing or hypothesis testing (described in section 6.2). Significance testing seeks to answer the question, "How much does my observation deviate from the expectations of my model?" which is only useful for ruling out one particular model. Hypothesis testing attempts to answer, "Based on my observation, should I accept my primary model or reject it in favor of any alternative models?" which only allows for a binary decision to be made (accept the primary model or reject it in favor of alternatives). The appeal of constructing confidence intervals is that it seeks to answer the question, "Based on my observation, which models are most likely to be correct?" which is treats all models equally in the search to find the best models that explain the observed data.

There are several good texts on statistical inference for particle physics [157, 253–255], as well as those for general use, [246–248, 251, 256]. The goal of this chapter is not to survey this extensive literature, but rather to delineate the aspects most pertinent to a signal search, and to explain new methods developed here. After presenting some introductory material, section 6.2 explains the Neyman confidence region construction (NCRC) method of frequentist interval estimation, and how test statistics from Neyman-Pearson hypothesis testing (NPHT) may be incorporated into NCRC. Then in section 6.3, the problems with current methods are delineated, and the Arizona test statistics are defined to address these issues. After this in section 6.4, confidence-acceptance profiles (CAPs) are defined and shown to provide more detailed frequentist inference information beyond the usual confidence intervals, as well as examining how to combine CAPs from multiple experiments via meta-analysis. This is followed by a brief study of how systematic uncertainties are incorporated as nuisance parameters, and the frequentist coverage properties that are found when using hybrid Bayesian-frequentist treatments of these nuisance parameters (section 6.5). The chapter ends with a brief summary discussion of the importance of the new methods.

6.1. A Conceptual Outline of this Chapter

The following outline is provided to give the reader a clearer understanding of the conceptual structure of this chapter.

1. There are 3 common methods of frequentist inference. The concepts of these overlap a great deal, which often leads to confusion (especially regarding the parameter α).
   - *Fisher Significance Testing* (FST)[1]: A test statistic is used to test a single hypothesis. The result is a p-value, and α is the significance level (when used).
   - *Neyman-Pearson Hypothesis Testing* (NPHT): A test statistic is used to test a null hypothesis against an alternative hypothesis. The result is critical region that determines acceptance of the null hypothesis or else acceptance of the alternative hypothesis, and α is the test size and $(1-\beta)$ the test power.[2]
   - *Neyman Confidence Region Construction* (NCRC): An ordering rule is used for all possible hypotheses to determine an acceptance region for each hypothesis. The result is a confidence region (or interval) that determines which hypotheses are accepted or else rejected. The confidence level is $(1-\alpha)$. ***This method is the primary focus of this chapter.***
2. In NCRC, a test statistic from NPHT may be used as the ordering rule. The test statistic used as an ordering rule for NCRC has an *separate interpretation* when viewed in NPHT, which can lead to confusion.
3. The multivariate Poisson distribution may be used to model a signal search experiment. The parameter of interest is the signal strength, μ, for which one seeks to obtain a confidence interval that provides evidence for which values of μ might equal μ_{true}.
4. Common test statistics used in particle physics are Q_{FC}, Q_{D0}, and Q_{AT}. Each of these test statistics produces different confidence intervals for the same dataset.
   - ***Problem:*** Sometimes empty set confidence regions are produced so that no hypothesis is accepted, which is difficult to interpret. Sometimes very small confidence intervals are produced that are very often misinterpreted as a statement of high precision for the accepted values of μ.
   - ***Solution:*** One can modify the test statistics so that the lengths of the confidence intervals more accurately reflect the precision of the experiment. Modification is based on the common sense assumption that the experimenter is able to state which measurements are most signal-like and which are least signal-like.
5. The test statistics, Q_{FC}, Q_{D0}, and Q_{AT}, are based on likelihood ratios, which are known to produce the shortest confidence intervals.
   - ***Problem:*** If μ_{true} is not in the confidence interval, NCRC does not provide any statement regarding how far μ_{true} is from the confidence interval.
   - ***Solution:*** Compute the *expected* distance of μ_{true} from the confidence interval, named "reporting risk". Again, modify the test statistics so that this expected distance is reduced. These are the *Arizona test statistics*.

[1] Often called "null hypothesis testing" or "null hypothesis significance testing".
[2] NPHT is explained in section 6.2.3.1.

6. The result of NCRC is a confidence interval that allows a simple accept or reject statement for each value of μ.
   - ***Problem:*** A confidence interval for a given confidence level does not allow one to compare the strength of evidence for different values of μ inside the confidence interval.
   - ***Solution:*** Produce as the result a function that shows the confidence intervals at *all* confidence levels simultaneously. This function is named the *confidence-acceptance profile* (CAP).[3]
     - ***Problem:*** It is ambiguous if the CAP has any frequentist meaning beyond providing the confidence intervals at all confidence levels.
     - ***Solution:*** One may use CAPs to construct functions that show the relative evidence for each value of μ while providing frequentist error estimates for these functions. The are the *signal significance ratio (SSR) function* and the *background significance ratio (BSR) function*, which gives the strength of evidence for signal nonexistence for each value of μ, and the strength of evidence for signal existence at each value of μ, respectively.
     - ***Problem:*** Researchers may not feel the need to report CAPs.
     - ***Solution:*** If researchers for different experiments report CAPs for the same parameter of interest, one may easily combine the different CAPs via *meta-analysis* to obtain a combined experimental result.
7. Systematic uncertainties are incorporated into the multivariate Poisson model as nuisance parameters.
   - ***Problem:*** Find reasonable alternative methods for the treatment of nuisance parameters since the only logically valid frequentist method, projection, will lead to severe overcoverage.
   - ***Solution:*** Perform a hybrid Bayesian-frequentist method of marginalization over the nuisance parameters, with three different methods of incorporating the nuisance parameters into test statistics. Two methods are known, the nominal and profiling methods, while a new method is also examined.
     - ***Problem:*** Measure how effective each of these three methods are.
     - ***Solution:*** Conduct 2σ-coverage studies that show the nominal and profiling methods perform about the same, while the new method appears to provide superior frequentist coverage properties.

[3] A confidence-acceptance profile is different from Fisher's fiducial distribution and other confidence distributions because it is *not* a distribution.

6.2. Introduction to Inference for Signal Searches

The task of a signal search is to use inductive reasoning to infer the most reasonable values of the signal strength parameter μ based on the observational data. Inductive reasoning begins with *real-world observations* and attempts to determine the most plausible *theoretical models* that best describe the causation of the observations, i.e., making sense of data. Parametric statistical inference restricts the theoretical models examined during induction to a family of probability distributions of a specific form with one or more unknown parameters to be determined.

There are two general methodologies of parametric statistical inference often employed: *Bayesian* and *frequentist*. The principal difference between Bayesian and frequentist inference is the interpretation of probability. A Bayesian probability represents a degree of belief in the possibility of an outcome while a frequentist probability is simply the long run frequency of occurrence of an outcome.

> *"'Make up a prior and solve the problem!' a staggering affront to the scientific process."*
>
> – Donald Fraser [257]

The advantage of Bayesian inference is that it results in a posterior probability distribution for the model parameters, which may be used to make probabilistic statements about the possible values of the parameters. The disadvantage of Bayesian inference is that it requires the subjective selection of a prior probability distribution for the model parameters.[4] The use of priors may be sensible in many circumstances, but is less appropriate in a search for a signal that is not known to exist because the subjective degree of prior belief in existence or nonexistence affects the resultant posterior distribution. Essentially, Bayesian inference allows one to make probability statements using a posterior distribution [259]:

$$p\left(\mu \,|\, d_{obs};\, \pi(\mu)\right), \tag{6.2.1}$$

which describes the probability of μ being equal to μ_{true} conditional on the observed data, d_{obs}, and using a prior probability distribution, $\pi(\mu)$, that summarizes the experimenter's beliefs before any measurement is taken.

More specifically for a signal search, let the probability of observing real world data d be $p(d;\, \mu)$, where $\mu \geq 0$ is a parameter that models the strength of the signal. A value of $\mu = 1$ represents the nominal expected strength of the signal if it exists at the level predicted by theory, and small values of μ imply the nonexistence of the signal, far below those expected if the signal were to exist.

In Bayesian inference, μ is promoted to the status of a random variable so that the parametrized model becomes a conditional model, $p(d;\, \mu) \rightarrow p(d \,|\, \mu)$, and then Bayes' theorem is used to compute the posterior distribution,

$$p(\mu \,|\, d_{obs}) \;\propto\; \pi(\mu) \cdot p(d_{obs} \,|\, \mu), \tag{6.2.2}$$

[4] Bayes himself was aware of the logical inconsistency of assuming a prior for situations where one has no previous knowledge. This is likely the reason Bayes did not publish his result (it was published posthumously). See *Scholium* in section II of [258].

where d_{obs} is the actual data observation, $\pi(\mu)$ is the chosen prior probability distribution of μ, and $p(d_{obs} \mid \mu)$ is the likelihood function corresponding to d_{obs} [253]. In Bayesian inference, the likelihood function is the function of the *random variable* μ that is obtained from the conditional probability distribution, $p(d' \mid \mu)$, for a fixed outcome d'. (In frequentist inference, the likelihood function is the function of the *model parameter* μ that is obtained from the parametrized probability distribution, $p(d'; \mu)$, for a fixed outcome d'.[5]) The Bayesian posterior distribution, $p(\mu \mid d_{obs})$, summarizes the evidence provided by the actual observed outcome, d_{obs}, and may then be used to provide statistics and credibility interval estimates for values of μ.

> *"Many Bayesians became Bayesians because the likelihood principle left them little choice."*
>
> – James Berger, et al. [261]

Bayesian inference satisfies the likelihood principle that implies that only the observed measurement is relevant for inference, and not any other outcome that could have possibly occurred [261]. The likelihood principle states that if two outcomes d_1 and d_2 yield proportional likelihood functions, $p(d_1; \mu) \propto p(d_2; \mu)$, then these provide equivalent evidence for inference[6] regarding μ [264]. The likelihood principle and the proof that it follows from more primitive concepts (with this proof refuted in [266]) provides a theoretical justification for Bayesian inference [261].

Experimental particle physics approaches statistical inference in a pragmatic fashion that often incorporates elements of frequentist and Bayesian approaches. For example, when measuring the invariant mass of a particle that is known to exist, a Bayesian prior may be constructed from knowledge of previous measurements [267]. However, when measuring the possible cross section of a particle not known to exist, a frequentist approach is more appropriate.

For frequentist inference, μ is never viewed as a random variable but rather assumed to exist with a single true value, μ_{true}. The results of frequentist inference describe what one should expect if the experiment was repeated many times. Three common frequentist methods are often employed in particle physics: the Fisher significance testing (FST), the Neyman-Pearson hypothesis testing (NPHT), and the Neyman confidence region construction (NCRC). Table 6.1 provides a brief summary of these methods.

FST is the simplest form of inference where the observed outcome d_{obs} is evaluated for

[5] Notation varies among authors, and to avoid clutter, the statement "$p(d; \mu)$" will refer to either the frequentist parametrized probability distribution of the random variable d (possible outcomes) for fixed parameter value μ, or else to the likelihood function of variable parameter μ for fixed outcome d, depending on the context of usage. This is also done in [260].

[6] A pragmatist's argument against the likelihood principle may be given as follows. If $p(d_1; \mu) = 10^{-3} p(d_2; \mu)$ for all μ (i.e., proportional likelihood functions), then clearly $p(d_1) = 10^{-3} p(d_2)$ so that d_1 is a much less probable experimental outcome than d_2. For the purpose of drawing conclusions from measured data, the pragmatist should be less convinced of any conclusion inferred when observing the much less probable outcome d_1 than the much more probable outcome d_2. Such doubt of conclusions based on highly improbable outcomes cannot be incorporated into methods that satisfy the likelihood principle. Apparently Birnbaum held a similar doubt of the likelihood principle as it violated his definition of *unbiased evidence* [262] (see also section 26.25 of [263]); this doubt despite himself having originally defined the likelihood principle [264]. See also [261, 265].

Fisher Significance Testing (FST)

Inputs:

H_N: A single test hypothesis (a particular value of μ).

$Q(d)$: A test statistic (function of possible outcomes d).

α: A predetermined significance level (optional).

Result:

The p-value determined using $Q(d_{obs})$ that gives a measure of significance.

Neyman-Pearson Hypothesis Testing (NPHT)

Inputs:

$H_N : \mu = \mu'$. The null hypothesis that the model parameter value is μ'.

H_A: An alternative hypothesis used for comparison (alternative values of μ).

$Q(d, \mu)$: A test statistic (function of outcomes d and parameter values μ).

α: The test size (desired frequency of type I error).

Intermediate Results:

Q^*_α: The critical value (determines the critical region).

$(1 - \beta)$: The test power (probability of type II error).

Final Result:

Decide to accept or reject H_N by comparing $Q(d_{obs}, \mu')$ to Q^*_α.

Neyman Confidence Region Construction (NCRC)

Inputs:

H_μ: A separate hypothesis for each value of μ.

$Q(d, \mu)$: An ordering rule (function of outcomes d and parameter values μ).

$(1 - \alpha)$: The confidence level (frequency that confidence region is incorrect).

Intermediate Results:

$Q^*_\alpha(\mu)$: The critical values (determines the acceptance region for each μ).

$\phi(d, \mu)$: The indicator function that shows the acceptance region for each μ.

$C(d, \mu)$: The confidence regions for each possible outcome, d.

Final Result:

$C(d_{obs}, \mu)$: The observed confidence region (values of μ that are accepted).

Table 6.1: A summary of the three common frequentist methods for parametric inference where a stochastic model relates the parameter of interest μ to the possible outcomes d. FST seeks to determine how significantly improbable the observed outcome, d_{obs}, is with respect to the test model, H_N, and is useful for providing evidence against H_N. NPHT compares a specific null hypothesis model, H_N, to alternative models, H_A, and seeks to determine which of H_N or H_A should be accepted and the other rejected, given a predetermined test size, α. NCRC examines all possible model parameter values as unique hypotheses, H_α, treated equally so that there is no preferred null hypothesis. NCRC seeks to determine the values of μ that are accepted (which of H_α are accepted), given a predetermined confidence level, $(1 - \alpha)$.

consistency with a single hypothesis, H_N, described by a probability model, $p(d)$. This involves examining the distribution of a test statistic $Q(d)$ with distribution $p(q = Q(d))$ determined by the model, and integrating over a portion of the distribution with a boundary determined by the observed test statistic value, $q_{obs} = Q(d_{obs})$. The smaller this integral is (called a p-value), the less consistent the observed outcome is with H_N.

The exact region of integration is not specified in FST [263], and the experimenter must specify the order in which outcomes d enter the integral for the p-value. If $Q(d)$ is designed so that the outcomes that are least consistent with the model receive the smallest values of $Q(d)$, then the p-value is obtained by $\int_{-\infty}^{q_{obs}} p(q)\, dq$.

Before the measurement of d_{obs}, one may determine a significance level, α, as a threshold below which the p-value may be claimed to demonstrate significant inconsistency with H_N, but this is not part of the original procedure,[7] and the smaller of two p-values always represents greater inconsistency with H_N even if both p-values are below α. A large p-value does not demonstrate that H_N is correct, but a small p-value indicates that if H_N is correct, then it is improbable to measure outcomes at least as inconsistent with H_N as d_{obs}, and in this way provides evidence against H_N being correct.

NPHT is described in detail in section 6.2.3.1 and NCRC in section 6.2.2. These are more complex than FST because they examine the evidence for multiple hypotheses simultaneously. In NPHT, a preferred model hypothesis, H_N, is compared to one or more alternative hypothesis models, H_A, in order to determine whether or not d_{obs} supports accepting H_N over H_A. In NCRC, all possible hypotheses are given equal status, and evidence is examined to determine which hypotheses are most supported by d_{obs}. In all three methods, the values of Q provide an ordering of the possible outcomes so that tail probabilities may be computed. The experimenter's choice of Q provides a flexibility to select inference results with desirable properties, such as UMA confidence intervals, or results that are the most sensible according to the specifics of the experiment (discussed in section 6.3.1). Concepts from each of the approaches will be synthesized in section 6.4.

Frequentist inference in a signal search is often used to construct a confidence interval for the signal strength parameter μ, usually at a 95% confidence level in a search. A resultant confidence interval that contains small values of μ close to zero while not containing $\mu \geq 1$ provides evidence for signal nonexistence. A confidence interval that contains $\mu = 1$ but not small values close to zero provides evidence for signal existence.

Frequentist interval estimation is useful because it follows from the *confidence principle* that guarantees the rate at which inference statements are correct in the long run. A confidence level of 95%, for instance, implies that as experiments are performed many times, the long run frequency with which μ_{true} lies inside the reported confidence intervals is 0.95. No method of inference can provide absolute truth, but the confidence principle does give the rate of correctly obtaining the truth in the long run of experimentation.

[7] See section 2.2 of [268] for a discussion of the origin of $\alpha = 0.05$.

6.2.1. The Multivariate Poisson Statistical Model

A signal search relying on a counting experiment may be modeled with a Poisson distribution,

$$p(d;\mu) = \frac{(\mu s + b)^d}{d!} e^{-(\mu s + b)}, \tag{6.2.3}$$

where d is the random variable representing the number of events detected (the possible outcomes) and b is a constant that represents the expected number of background events (obtained from simulated datasets). The signal strength μ is the parameter of interest that is to be inferred with $\mu = 0$ providing the background-only model when needed, and s is a constant representing the number of expected signal events according to the new theory (also obtained from simulated datasets). Note that $\mu = 1.0$ corresponds to the nominal signal model with s expected signal events.[8] The signal strength parameter, μ, is thus a multiplier to the nominal expected signal yield, s, so that inferring the confidence interval for μ not only provides a concise summary of the strength of evidence for signal existence provided by the observed outcome, but also a description of the amount of signal that may exist above the background.

It is often the case that rather than simply counting events, a measurement is made on a discriminating variable for each event and its value recorded. The discriminating variable may then be binned into ranges of values so that a count in each bin of the discriminating variable is obtained. This results in a multivariate Poisson model,

$$p(d;\mu) = \prod_{i}^{N_{bin}} \frac{(\mu s_i + b_i)^{d_i}}{d_i!} e^{-(\mu s_i + b_i)}, \tag{6.2.4}$$

where $d = [d_1, d_2, \ldots, d_{N_{bin}}]$ are the possible event counts in each of the discriminating variable bins, and b_i and s_i are the expected background and signal for the i^{th} bin.[9]

The estimates of s_i and b_i are taken from weighted simulated datasets,

$$s_i = \sum_{j}^{N_i^{s\,sim}} w_{i,j}^s \text{ and } b_i = \sum_{j}^{N_i^{b\,sim}} w_{i,j}^b, \tag{6.2.5}$$

where $w_{i,j}$ is the weight given to event j arriving in bin i, and N_i^{sim} is the total number of events arriving in bin i. The systematic uncertainty due solely to the amount of simulated data available for these measurements (often referred to as "statistical" uncertainty), is

[8] Unlike the possible outcomes d that must be integers, the constants b and s need not be integers due to weighting used in the simulated datasets.

[9] The collection of random variables d_i, the counts in each bin of the discriminating variable, provide the *approximately* jointly sufficient statistics necessary to make inferences about μ. These are approximate because some information contained in the spectrum of the discriminating variable is lost when binning is imposed. The binning of the discriminating variable is useful in modeling the systematic uncertainties of s_i and b_i with nuisance parameters.

estimated from the weights as

$$\sigma_i^s = \left(\sum_j^{N_i^{s\,sim}} \left(w_{i,j}^s\right)^2 \right)^{1/2} \text{ and } \sigma_i^b = \left(\sum_j^{N_i^{b\,sim}} \left(w_{i,j}^b\right)^2 \right)^{1/2}. \tag{6.2.6}$$

Although in theory one may always reduce this uncertainty to any level by generating more simulated data, it is often the case in practice that one must make the best inference possible with the simulated data at hand. Inclusion of this uncertainty in the model, which is uncorrelated between bins, is accomplished with

$$p(d;\mu,\boldsymbol{\nu}^s,\boldsymbol{\nu}^b) = \prod_i^{N_{bin}} \frac{\left(\mu\,\nu_i^s s_i + \nu_i^b b_i\right)^{d_i}}{d_i!}\, e^{-\left(\mu\nu_i^s s_i + \nu_i^b b_i\right)}\, f_i^s(\nu_i^s) f_i^b(\nu_i^b), \tag{6.2.7}$$

where $f_i^s(\nu_i^s)$ is the Bayesian prior probability distribution assigned to the signal nuisance parameter ν_i^s, given by a truncated Gaussian,[10]

$$f_i^s(\nu_i^s) = \frac{1}{\mathcal{C}\sigma_i^s\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{s_i}{\sigma_i^s}\right)^2\left(\nu_i^s - 1\right)^2},\ k_i^s \geq 0, \tag{6.2.8}$$

with a normalization constant $\mathcal{C}$ from truncation (and is similarly defined for the background nuisance parameters $f_i^b(\nu_i^b)$). These functions have the effect of adjusting the nominal value of the expected signal in each bin, s_i, by the uncertainty due to limited simulation data (and similarly for background).

In the case of multiple backgrounds indexed by j, the model becomes

$$p(d;\mu,\boldsymbol{\nu}^s,\boldsymbol{\nu}_j^b) = \prod_i^{N_{bin}} \left(\frac{\left(\mu\,\nu_i^s s_i + \sum_j \nu_{j,i}^b b_{j,i}\right)^{d_i}}{d_i!}\, e^{-\left(\mu\nu_i^s s_i + \sum_j \nu_{j,i}^b b_{j,i}\right)}\, f_i^s(\nu_i^s) \prod_j f_{j,i}^b(\nu_{j,i}^b) \right), \tag{6.2.9}$$

but this additional notational complexity will be omitted in much of what follows to retain clarity. Note that if there are N_{bkg} backgrounds, then this type of uncertainty generates $(N_{bkg}+1)\cdot N_{bin}$ unique nuisance parameters of this type.

Additional systematic uncertainties arise from measurement errors as described in section 4.3 (page 94). These uncertainties are assumed to be uncorrelated from one another. However, the effect of each systematic uncertainty is correlated between the bins of the discriminating variable. For example, if the total integrated luminosity is actually 2% above the nominal estimated value, then this will increase the expected yields in each bin uniformly by 2%. Indexing each of these systematic uncertainties by k, they are incorporated into the

[10] For alternative Bayesian priors, see section 3.2 of [244].

model as

$$p(d;\mu,\boldsymbol{\nu}^s,\boldsymbol{\nu}^b,\boldsymbol{\nu}) = \prod_k^{N_{syst}} f_k(\nu_k) \left(\prod_i^{N_{bin}} \frac{\left(\mu\,\nu_i^s g_i^s(\boldsymbol{\nu}) s_i + \nu_i^b g_i^b(\boldsymbol{\nu}) b_i\right)^{d_i}}{d_i!} e^{-\left(\mu\nu_i^s g_i^s(\boldsymbol{\nu}) s_i + \nu_i^b g_i^b(\boldsymbol{\nu}) b_i\right)} f_i^s(\nu_i^s) f_i^b(\nu_i^b) \right), \tag{6.2.10}$$

where $\boldsymbol{\nu} = \{\nu_k\}$, with each of these nuisance parameters assigned a Bayesian prior of a standard normal distribution $f_k(\nu_k) \sim \mathcal{N}[0,1]$, and where $g_i^s(\boldsymbol{\nu})$ is a function that returns the adjustment to the nominal value s_i due to the $\boldsymbol{\nu}$ (and similarly for background). For example, if the nominal expected signal count for the first bin is $s_1 = 10$, the bin systematic is $\nu_1^s = 1.1$, and the measurement systematic uncertainty function is $g_1^s(\boldsymbol{\nu}) = 0.8$, then the expected signal count including systematic shifts is $10 \times 1.1 \times 0.8 = 8.8$.

Note that the background-only model is given by $p(d;\mu = 0,\boldsymbol{\nu}^b,\boldsymbol{\nu})$, i.e., does not depend on any nuisance parameters solely associated with signal. The Bayesian priors of nuisance parameters used to model empirically estimated systematic uncertainties (based on previous measurements) represent the experimenters' beliefs of the probabilities of systematic shifts in the expected background and signal counts [259].

The adjustment function $g_i^s(\boldsymbol{\nu})$ may be decomposed into component functions for each systematic uncertainty,[11]

$$g_i^s(\boldsymbol{\nu}) = 1 + \sum_k \left(g_{k,i}^s(\nu_k) - 1 \right), \tag{6.2.11}$$

where $g_{k,i}^s(\nu_k)$ is determined from the simulated data (and similarly for each of the backgrounds). This is done for systematic k by separately determining the $+1\,\sigma_k^s$ and $-1\,\sigma_k^s$ variation of the signal's discriminating variable spectrum across the bins. The bin-by-bin ratios of each of these spectra to that of the nominal spectra provide the adjustment to the nominal bin values in the case that $\nu_k = +1$ and $\nu_k = -1$, respectively. For example, if the results for the first bin are $s_0 = 10$, $s_0^{+\sigma_k} = 12$, and $s_0^{-\sigma_k} = 7$, then one obtains the adjustments $g_{k,0}^s(+1) = 1.2$ and $g_{k,0}^s(-1) = 0.7$ for these two values of the nuisance parameter, ± 1. The values of $g_{k,0}^s(\nu_k)$ when $|\nu_k| < 1$ must be obtained by interpolation and by extrapolation when $|\nu_k| > 1$. Several interpolation-extrapolation methods are possible, and here the "MCLimit" method was selected [244].[12]

[11] The additive model was selected over the multiplicative model (both described in section 3.3 of [244]) because a) it does not induce correlations between the component adjustment functions, and b) it is more reasonable to expect that an adjustment of +10% and another adjustment of -10% to the nominal value should exactly cancel. The drawback of the additive model is that it can produce unphysical outcomes with $g_i(\boldsymbol{\nu}) < 0$, especially when many large systematic uncertainties are present. When this occurs, the multiplicative model may be used instead with $g_i(\boldsymbol{\nu}) = \prod_k g_{k,i}(\nu_k)$.

[12] Several interpolation-extrapolation methods were examined in section 3.3 of [244], but MCLimit provided the most conservative extrapolation (see figure 2 of [244]).

6.2.2. Neyman Confidence Region Construction

The Neyman confidence region construction method (NCRC) is used to perform frequentist interval estimation [267, 269, 270]. A confidence region for a parameter of interest μ with confidence level $(1-\alpha)$ where $0 \leq \alpha \leq 1$, is the set of values of μ that an experimenter is confident at level $(1-\alpha)$ to contain the true value, μ_{true}. Confidence at level $(1-\alpha)$ implies that if experiments were performed many times and a confidence region reported each time, the long run frequency of confidence regions not containing μ_{true} would equal α. Confidence is thus the frequentist probability that the *reporting procedure* provides correct results with a frequency of at least $(1-\alpha)$. However, confidence does not describe any probabilistic claim regarding the confidence region of any single experimental result.[13]

If the reporting procedure provides confidence at exactly level $(1-\alpha)$ for all possible values of μ, then it is said to provide exact frequentist coverage. If for *some* values of μ, this frequency is greater than $(1-\alpha)$, the procedure is said to overcover. If for *any* values of μ, this frequency is less than less than $(1-\alpha)$, the procedure is said to undercover, regardless of the possible overcoverage for other values of μ. For discrete random variables d, it is not always possible to provide exact coverage. In this case, overcoverage is preferable to undercoverage so that in the long run, confidence regions not containing μ_{true} are reported with a frequency of less than α.

Blinded statistical inference begins by constructing a $(1-\alpha)$ level confidence region $CR_\alpha(d)$ for every possible experimental outcome d. Once all $(1-\alpha)$ level confidence regions have been constructed for any outcome d, the actual observed outcome d_{obs} is then unblinded, and the unique confidence region $CR_\alpha(d_{obs})$ is reported. Blinding forces the experimenter to determine *a priori* how any actual outcome is to be interpreted in order to remove any human bias from reporting, such as the "flip-flopping" bias in experimental particle physics discussed in [272].

If a confidence region at level $(1-\alpha)$ is the set of all μ values between a lower bound l and upper bound u, then it is called a (two-sided) confidence interval, $CI_\alpha = [l, u]$. For a signal search, μ is restricted to be nonnegative so that $l \geq 0$. If a confidence region is given as $[0, u]$ with zero a fixed constant, then it is called the upper confidence limit, $UCL_\alpha = [0, u]$ (or alternatively called the one-sided confidence interval). Since blinded inference must account for any possible outcome d, then these bounds are functions of d and thus are also random variables. Thus before unblinding, one has confidence intervals or upper limits that are random variables, $CI_\alpha(d) = [l(d), u(d)]$ or $UCL_\alpha(d) = [0, u(d)]$, respectively.

A confidence region at level $(1-\alpha)$ may be determined by a non-randomized decision rule[14] defined by a rejection indicator function, $\phi^\alpha_{rej}(d, \mu) = 0$ or 1, that is a function of possible real world outcomes d and parameter values μ. The indicator function $\phi^\alpha_{rej}(d, \mu)$

[13] Based on discussions with Feldman and Cousins, Neal in [271] complained that the scientific use of frequentist confidence intervals is unclear if additional decisions beyond simple acceptance and rejection must be made. Neal's real issue was that conclusions drawn from Feldman-Cousins confidence intervals for certain possible outcomes were irrational. He was correct in this, and this is what the Arizona test statistics seek to address (discussed in section 6.3).

[14] The use of randomized decision rule where ϕ^α_{rej} may take values between 0 and 1 is seldom used in particle physics. Note that if d' was observed and $\phi^\alpha_{rej}(d'; \mu') = 0.5$, then one would need to flip a coin to determine whether or not to include μ' in the confidence interval, a procedure many scientists would not be comfortable with.

returns a binary outcome for the decision to reject (= 1) or accept (= 0) the value of μ (for a given d). For a fixed μ', the set of d such that $\phi^{\alpha}_{rej}(d, \mu') = 0$ defines an *acceptance region* for μ' and for the procedure to deliver $(1 - \alpha)$ confidence, the probability of acceptance for μ' must be at least $(1 - \alpha)$ so that

$$\sum_{d'} p(d'; \mu') \left(1 - \phi^{\alpha}_{rej}(d', \mu')\right) \geq 1 - \alpha. \tag{6.2.12}$$

If this condition holds for all μ, then for any fixed d', the set of values for μ such that $\phi^{\alpha}_{rej}(d', \mu) = 0$ defines the confidence region for d'.

In order to guarantee the absence of undercoverage at level $(1 - \alpha)$, $\phi^{\alpha}_{rej}(d, \mu)$ must be defined so that the risk of undercoverage is less than α. Risk is the expectation of loss, and any outcome d' that results in rejecting μ_{true} with $\phi^{\alpha}_{rej}(d', \mu_{true}) = 1$ is thus an error. Since μ_{true} is unknown, $\phi^{\alpha}_{rej}(d, \mu)$ must be defined so that for any possible value μ', the risk of undercoverage for μ' is less than α,

$$R^{\alpha}_{size}(\mu') \equiv E_{\mu'}[\phi^{\alpha}_{rej}(d', \mu')] = \sum_{d'} p(d'; \mu')\, \phi^{\alpha}_{rej}(d', \mu') \leq \alpha, \tag{6.2.13}$$

where D is the set of all possible outcomes for d (and may be replaced with integration when D is continuous). The inequality of equation 6.2.13 expresses that $R^{\alpha}_{size}(\mu)$ must never be greater than α to prevent undercoverage, but in order to minimize the amount of overcoverage, $R^{\alpha}_{size}(\mu)$ should be close to α without being greater. The subscript "size" is used because of the similarity of this risk to the Neyman-Pearson hypothesis testing (NPHT) (discussed in section 6.2.3.1), where the size, α, of a hypothesis test is the integral over the tail of the probability distribution of the null hypothesis model, and where α is predetermined by the experimenter. For each value of μ during the Neyman construction, the acceptance region is determined with regard to the predetermined confidence level, $(1 - \alpha)$, i.e., one takes H_{μ} as the null hypothesis and selects the acceptance region that ensures the correct size. This connection is emphasized in section 6.2.3 that shows how a test statistic from NPHT may be used to determine the acceptance regions during the Neyman construction.

6.2.2.1. Example 1: Construction for the Univariate Poisson

The following simple example demonstrates how Neyman confidence regions at the 95% confidence level are constructed for the univariate Poisson model of equation 6.2.3, a simple counting experiment. Let the expected background and signal counts be given as $b = 4$ and $s = 4$, respectively, without any uncertainty. This gives the model probability distribution [15] as,

$$p(d;\, \mu) = \frac{(\mu\, 4 + 4)^d}{d!} e^{-(\mu\, 4+4)}, \tag{6.2.14}$$

where d is the possible outcome count that could be measured. The values of the probability distribution for selected values of μ are shown in figure 6.1 (a) in each horizontal row, and only showing these probabilities for possible counts ranging from 0 to 11. Figure 6.1 (a) shows that the mode of the probability distribution increases with increasing μ, from a possible count of 3 and 4 for $\mu = 0$ to a count of 9 for $\mu = 1.4$.

Confidence regions are obtained by first determining acceptance regions for each value of μ, i.e. each horizontal row. An acceptance region is a subset of the possible outcomes with a total probability of at least 95%. Garwood [273] was the first to compute upper confidence limits for the univariate Poisson model. This is done by summing from left to right (in each row) until reaching no more than 5% total probability. Figure 6.1 (b) shows with black crosses the possible outcome counts d for each row that are rejected from the acceptance region. For each horizontal row, the sum of probabilities that are rejected is less than or equal to 0.05. The rejection or acceptance of possible counts defined the indicator function, $\phi_{rej}(d, \mu)$, as shown in figure 6.1 (c), where outcomes d that are assigned to have $\phi_{rej}(d, \mu) = 1$ correspond to those shown with black crosses in figure 6.1 (b). This result is shown graphically in figure (d) with red horizontal bars delineating the acceptance region, $\phi_{rej}(d, \mu) = 0$, for each value of μ (and black horizontal bars for rejection, $\phi_{rej}(d, \mu) = 1$).

Note that due to the discrete values of d, this method often overcovers with less than 5% probability being rejected in each row. Crow and Gardner [275] and Casella [276] developed algorithms to reduce the amount of this purely mathematical overcoverage for the univariate Poisson model. However, mathematical overcoverage is not nearly as severe for the multivariate Poisson model, and it is difficult to extend these algorithms from univariate to multivariate models. In any case, one can argue that the selection of which outcomes d to include in an acceptance region should be based on the specific details of the experiment at hand, and not a purely mathematical optimization procedure (this is additional motivation for the Arizona test statistics discussed in section 6.3).

Once the acceptance regions (horizontal) are determined, the confidence region for each d is obtained simply by scanning vertically at that value of d. Any value of μ that lies in the acceptance region for d is accepted into the confidence region for d (and rejected otherwise) as is shown with red vertical bars in figure 6.1 (e) (with black vertical bars indicating values of μ outside the confidence regions). After the actual experimental outcome d_{obs} is obtained,

[15] The term "probability distribution" or "distribution" will refer to either discrete mass functions or else continuous density functions, with "cumulative distribution" used for the cumulative distribution functions of either.

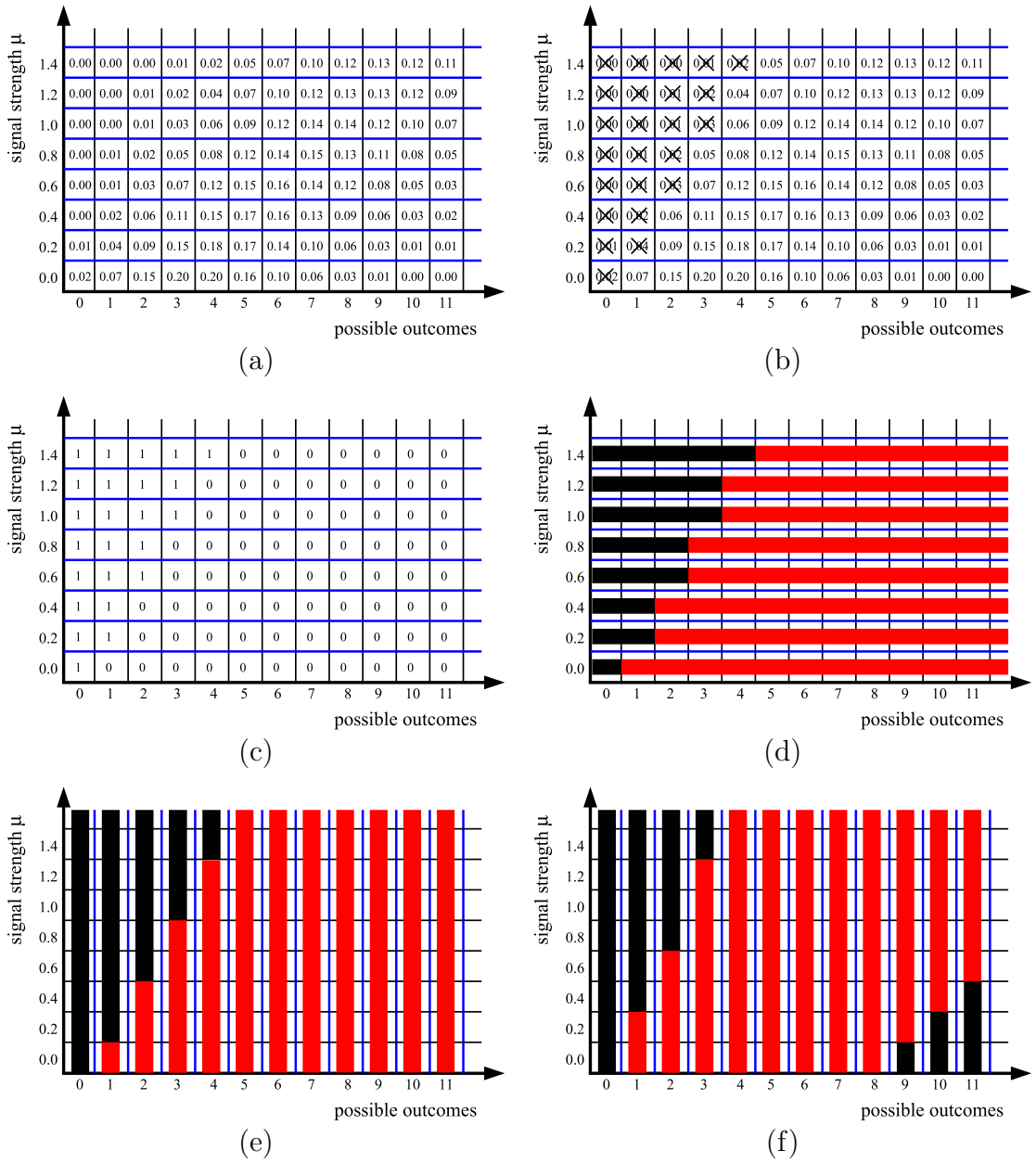


Figure 6.1: Example 1 – The univariate Poisson model with $b = 4$ and $s = 4$. (a) Horizontal rows show the probability distributions, $p(d; \mu)$, for selected values of μ and d. (b) Constructing acceptance regions by rejecting d from left to right until a probability of no more the 0.05 is rejected. (c) Acceptance regions define the indicator function, $\phi_{rej}(d, \mu)$, with $\phi_{rej}(d, \mu) = 1$ for rejection and $\phi_{rej}(d, \mu) = 0$ for acceptance. (d) Resultant acceptance regions shown with horizontal red bars. (e) The final Garwood [273] upper confidence limits for each d shown with vertical red bars. (f) The two-sided confidence intervals of Ricker [274].

the confidence region for that specific value of d is the one reported.

It is assumed that this construction is done without knowledge of any actual measurement (blinded) or occurs before measurement in order to prevent experimenter bias in selecting acceptance regions. Figure 6.1 (f) shows the two-sided confidence intervals obtained by the method of Ricker [274] where the removal of d from each acceptance region is done from both left and right sides, removing no more than 2.5% probability from each side. These provide different confidence intervals than those of figure 6.1 (e), demonstrating the need for blinding. For example, if the experimenter knew that a count of $d_{obs} = 3$ had been measured before selecting between the Garwood and Ricker methods, then the experimenter could cherry pick which confidence interval to report. The process of confidence region construction is subjective in the sense that the experimenter must decide which criteria to use to determine each acceptance region. The criteria selected for constructing acceptance regions should be based on the experimenter's best judgment relying on their knowledge of the experiment. This is discussed in section 6.3.1.

These resultant confidence intervals for this simple example illustrate an important issue. The outcome $d = 0$ results in reporting the empty set as the confidence region for both construction methods. Even if the signal did not exist ($\mu = 0$), there is a $\sim$2% probability of obtaining this outcome (given by the first cell of the bottom row of figure 6.1 (a)). Depending on the consequences of obtaining an empty confidence interval, the experimenter obtaining such a result might seek instead to perform a different inference procedure. This type of flip-flopping has been shown to alter the coverage probabilities of the Neyman construction [272], and in general raises doubts about the veracity of any reported result where the experimenter *could* have changed their inference procedure if an unfavorable outcome was obtained. Feldman and Cousins were primarily concerned with an experimenter observing a large count of d_{obs} and deciding *post hoc* to report two-sided intervals instead of one-sided [272]. Their "unified method" addressed this type of flip-flopping by introducing a test statistic based on likelihood ratios (discussed in section 6.2.3.4), but this does not resolve possible flip-flopping for small counts of d_{obs}. This problem is addressed with the Arizona test statistics in section 6.3.

6.2.2.2. Example 2: Multivariate Poisson Global Ordering

The following example is used to illustrate the Neyman confidence region construction for the multivariate Poisson model, but more importantly, to explain the methodology used to visually examine inference results for multivariate statistical models. Specifically, each possible experimental outcome d is a vector of N_{bin} counts for each of the bins of the discriminating variable spectrum. For example, the discriminating variable has two bins in the bivariate Poisson model, "low" and "high", so that one possible outcome is to measure four counts in the first bin and three in the second, $d = [4, 3]$.

For a multivariate model, there is no natural ordering of the possible vector outcomes d with which to make a one dimensional axis for d. In the bivariate Poisson model for example, should $[0, 1]$ be ordered before or after $[1, 0]$? Nevertheless, it is useful to define a *global ordering* of d so that plots may be made for analysis with d plotted along a single axis. By determining such a global ordering rule, comparing inference results from different methodologies is more feasible. Another issue in making plots using the possible vector outcomes d along a single axis is that there are infinitely many possible outcomes in any Poisson model. It it reasonable to only plot the "in-bounds" set of the most probable outcomes, D_{in}, while the remaining set of "out-of-bounds" outcomes, D_{oob}, are outside the visible range of a plot.

In this example, the final discriminating variable is simply binned into one of two binary categories, bin 1: "low-valued" and bin 2: "high-valued", with expectations of the background and signal counts in each bin as shown in table 6.2. These values are assumed to be precisely known so that no nuisance parameters are included to model their uncertainty. The bivariate Poisson distribution modeling this process is

$$p(d;\mu) = \left(\frac{(\mu\, s_1 + b_1)^{d_1}}{d_1!}\, e^{-\mu\, s_1 - b_1}\right)\left(\frac{(\mu\, s_2 + b_2)^{d_2}}{d_2!}\, e^{-\mu\, s_2 - b_2}\right), \tag{6.2.15}$$

where $d = [d_1, d_2]$ represents the possible observed counts for both bins, and μ is the signal strength parameter of interest to be inferred.

	bin 1	bin 2
expected background, b_i	4.0	3.0
expected signal, s_i	1.0	3.0

Table 6.2: Example 2 – Expected background and signal counts for a binary output discriminating variable in a hypothetical signal search.

Figures 6.2 (a) and (b) show the probability distribution of the background-only model (gray area), and the maximum likelihood estimate (red crosses), for possible outcomes d (described below). These demonstrate two different global orderings of the possible outcomes, d, shown along the x-axis. The two global orderings are the "background order", shown in figure 6.2 (a) and the "MLE order" (explained below), shown in figure 6.2 (b). These figures show in gray the probability distribution of the background-only model ($\mu = 0$) for 431 enumerated "in-bounds" outcomes $d \in D_{in}$. The probability scales are provided on the right

hand sides of the plots. A gray bar on the far right of each denotes the total probability of the remaining "out-of-bounds" infinite set of $d \in D_{oob}$.

These figures also show the mode of the background-only model, $d^b_{mode} = [4, 3]$, in dashed cyan. It is useful to know the location of d^b_{mode} since it is the most probable outcome if the signal does not exist. Also shown in dashed green is $d^{s+b}_{mode} = [5, 6]$, the mode of the signal model assuming nominal signal strength, $\mu = 1$. Finally, red crosses denote $\tilde{\mu}_u(d)$, the unrestricted maximum likelihood estimate (MLE) for each $d \in D_{in}$. The MLE is the value of μ that maximizes the likelihood function $p(d; \mu)$ for a given d. These are "unrestricted" so that negative values are permissible. The scales for the values of μ are provided on the left hand sides of the plots. The value $\tilde{\mu}_u(d')$ that maximizes $p(d'; \mu)$ for a specific outcome d' is numerically computed by simple bisection to a precision of 1×10^{-14}. This high level of precision is not particularly useful, and is lowered to 1×10^{-3} for nuisance parameter studies which are more computationally demanding.

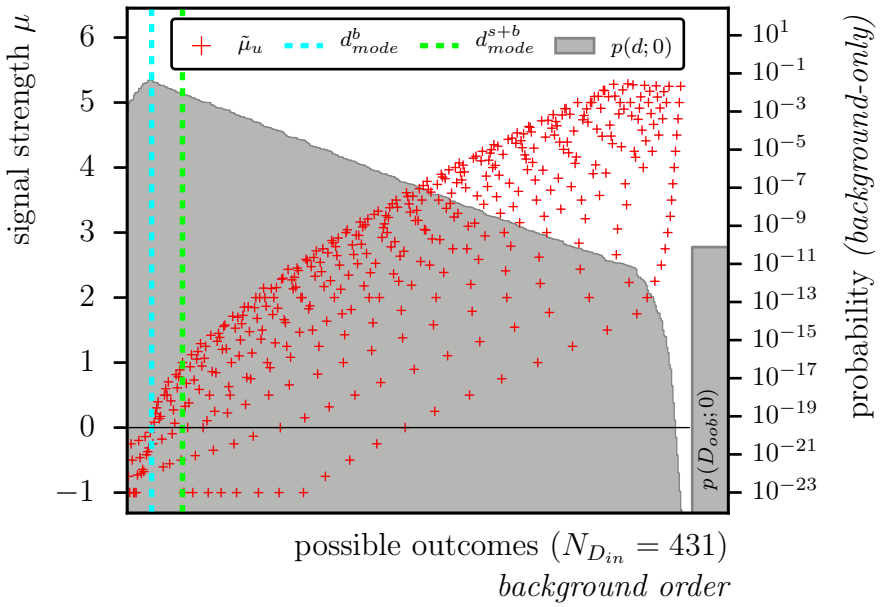
The selection of the 431 enumerated possible outcomes, $d \in D_{in}$, is determined by the following algorithm. First $d^b_{mode} = [4, 3]$ is selected. Then all $d = [d_1, d_2]$ below d^b_{mode} are selected, defined as having both $d_1 < 4$ and $d_2 < 3$. Then all d above d^b_{mode} are added in decreasing probability until the probability sum of D_{oob} is less than some predetermined value, 10^{-10} in this case. This allows the claim to be made that confidence regions for the most probable configurations according to the background-only model are to be analyzed since less than 10^{-10} of total probability lies outside D_{in}. Next, the maximum of MLE $\tilde{\mu}^{max}_u$ of those d currently enumerated is computed, and any configurations with $\tilde{\mu}_u(d) \leq \tilde{\mu}^{max}_u$ are also enumerated. Thus $d \in D_{in}$ now comprises of both the most probable configurations with no more than 10^{-10} of total probability remaining in D_{oob} and all configurations with MLE below $\tilde{\mu}^{max}_u$. It can be seen in figure 6.2 (a) that many of these configurations are highly improbable with $p(d; \mu)$ less than 10^{-20} for some $d \in D_{in}$.

In figure 6.2 (a), the possible outcomes d are globally ordered so that each $d = [d_1, d_2]$ to the left of d^b_{mode} has both $d_1 < 4$ and $d_2 < 3$, and these are ordered in increasing probability. The remaining configurations d to the right of d^b_{mode} are ordered in decreasing probability. The background global ordering is meant to emulate the appearance of the probability distribution of a univariate Poisson distribution, $p(x; \lambda) = \lambda^x / x!\, e^{-\lambda}$, which is typically plotted with x increasing.

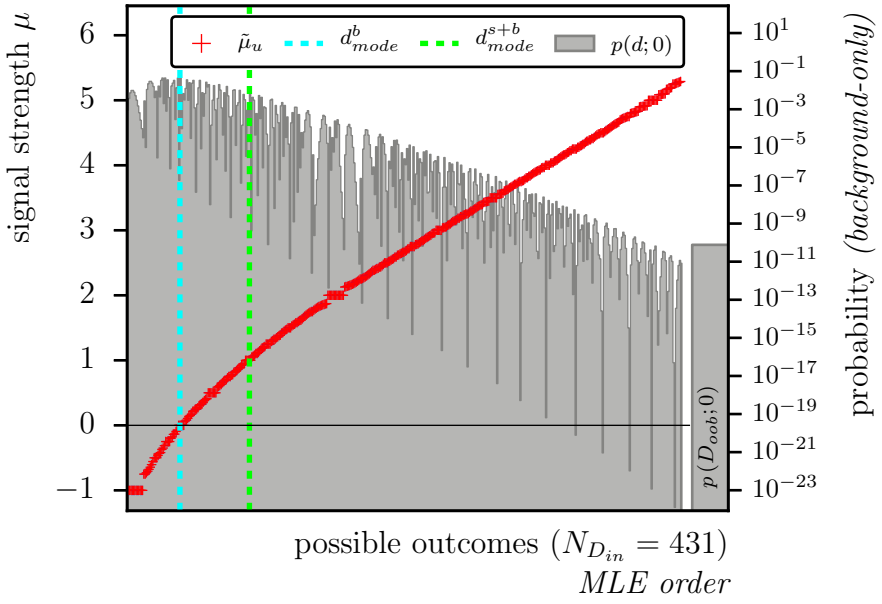
Figure 6.2 (b) shows the results of reordering the x-axis of figure (a) with the MLE global ordering. The MLE global order is defined as ordering $d \in D_{in}$ with increasing $\tilde{\mu}_u(d)$. In the case of equal values of $\tilde{\mu}_u(d)$, a secondary ordering rule is implemented with increasing $p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)$, where $\tilde{\mu}_p = max[0, \tilde{\mu}_u]$ is the MLE restricted to be "physical" since the true signal strength parameter may never be negative. The MLE global ordering is useful for visualizing the confidence regions constructed using test statistics based on likelihood ratios (discussed in section 6.2.3).[16]

Under the MLE ordering shown in figure 6.2 (b), the probability distribution for the background-only hypothesis no longer varies smoothly as with the background ordering. However, figures 6.3 (a)–(c) demonstrate that for the multivariate Poisson distribution, no

[16] The secondary global ordering rule is usually only implemented for configurations with $\tilde{\mu}_u(d) = \tilde{\mu}^{min}_u$, where $\tilde{\mu}^{min}_u$ is the minimum value of μ permissible in the multivariate Poisson model, below which the model is undefined, $\tilde{\mu}^{min}_u = \max_{i\, bins}[-b_i/s_i]$. Note that in both background and MLE global ordering, $d = [0, 0]$ remains the leftmost configuration on the x-axis.



(a)



(b)

Figure 6.2: Example 2 – The bivariate Poisson model properties and global ordering. The gray area shows the probability distribution of the background-only model, red crosses show the values of the MLEs for each outcome d, a dashed cyan line shows d^b_{mode}, and a dashed green line shows d^{s+b}_{mode}. (a) Global ordering by the background-only model and (b) the same as (a) but the x-axis is ordered by the maximum likelihood values.

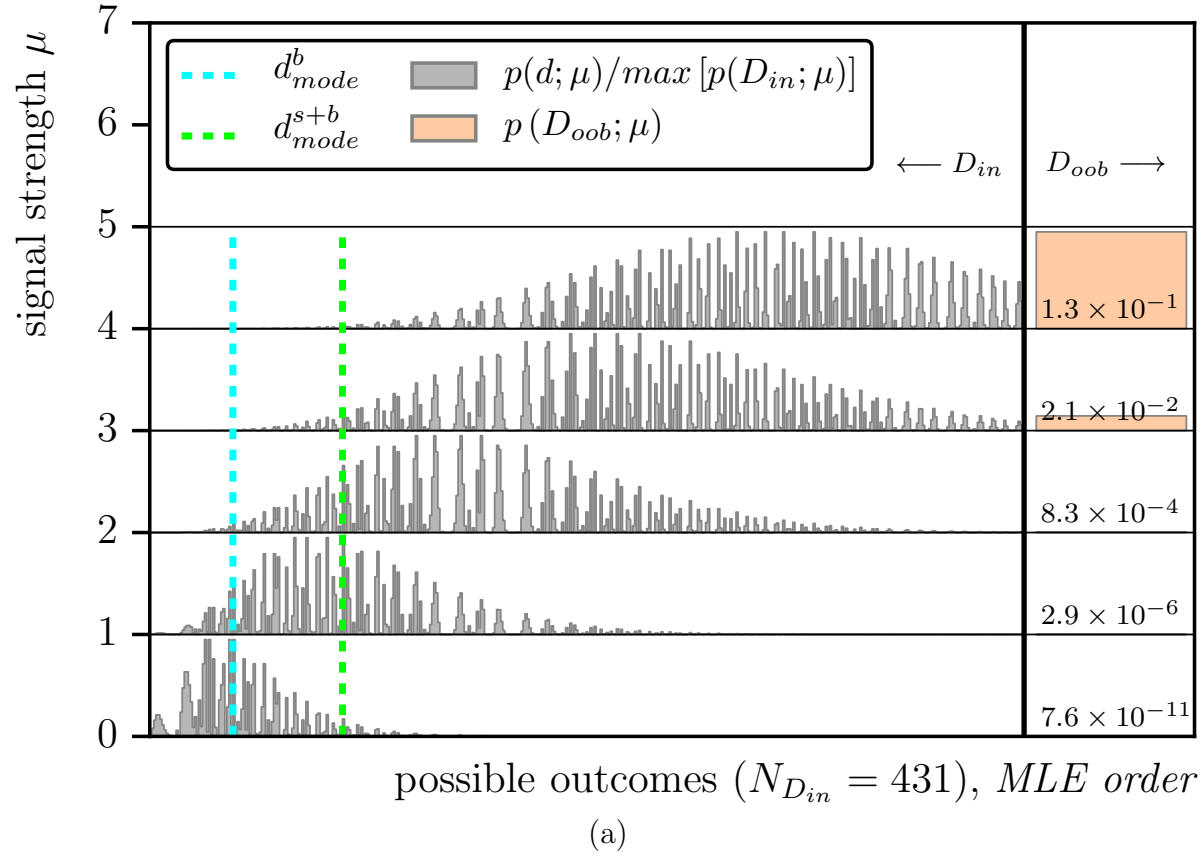
global ordering provides the appearance of smoothly varying distribution, $p(d \in D_{in}; \mu)$, for all values of μ simultaneously. Figure 6.3 (a) shows the probability distribution using the MLE global ordering for selected values of μ: 0, 1, 2, 3, and 4. None of these distributions varies smoothly. Figure 6.3 (b) shows that when background ordering is used, the probability distribution for $\mu = 0$ varies smoothly, but the other probability distributions do not. Figure 6.3 (c) shows that with D_{in} ordered by decreasing probability for the $\mu = 1$ probability distribution, then only this distribution varies smoothly. Some graphical results may seem strange when depicted in a one dimensional figure because the multivariate Poisson probability distributions do not simultaneously vary smoothly under *any* global ordering. Nevertheless, the MLE global ordering is useful for visually examining inference results based on likelihood ratio test statistics.

Once a global ordering is determined, globally ordered Neyman construction is possible. The MLE global ordering is used to construct one and two-sided confidence intervals in the style of Garwood and Ricker. This is performed numerically on a linear grid of 320 values of μ ranging from 0 to 8. Figure 6.4 (a) shows the result of determining acceptance regions in the style of Garwood [273] by removing $d \in D_{in}$ from left to right from the acceptance region of each value of μ until a total probability of no more than 5% is removed. The area in red shows the acceptance regions for each μ when scanned horizontally, and the resultant confidence intervals for each $d \in D_{in}$ when scanned vertically. The modes d^{b}_{mode} and d^{s+b}_{mode} are provided for reference, as well as the physical MLEs, $\tilde{\mu}_p(d)$.

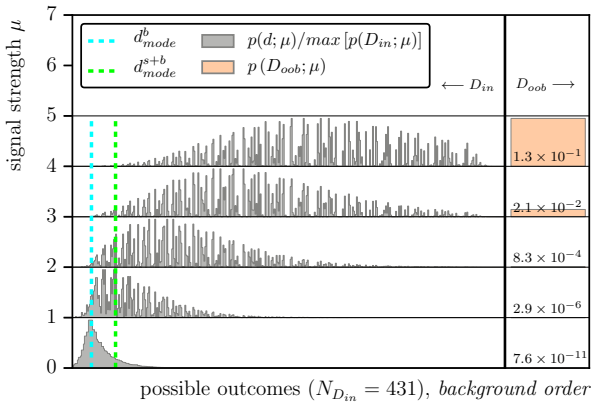
The right hand side of figure 6.4 (a) shows orange and white areas contained in a tall box with a gray border. This part of the figure accounts for the total probability of the unenumerated, infinitely many outcomes of D_{oob}. The width of the box represents 100% probability. For any value of μ, the width of the orange area for that μ shows the amount of total probability within the acceptance region for that μ. For example, for $\mu \sim 5.5$, the width of the orange area is half the total width of the gray box indicating that 50% of the accepted outcomes (by probability) lie in D_{oob}. Since the confidence level is 95%, the sum of probability over the accepted outcomes in D_{in} must be $\sim$45%. Though difficult to see, any blue area next to the orange accounts for the total probability that is rejected from the acceptance region for outcomes in D_{oob}.[17] For $\mu = 8.0$, nearly all of the total probability lies in D_{oob} so that the width of its orange and blue areas nearly spans the gray box. For $\mu = 8.0$, apparently the total probability of the rejected outcomes in D_{in} account for very little of the probability distribution of $p(d; 8.0)$. An actual observed outcome $d_{obs} \in D_{oob}$ would provide evidence for signal existence with a strength much larger than $\mu = 1$, but the enumeration would need to be expanded in order to construct the confidence interval for such an observation.

Figure 6.4 (b) shows the result of determining acceptance regions in the style of Ricker [274] by removing $d \in D_{in}$ from both the left and right until no more than 2.5% total probability is removed from each end. Both methods of construction yielded empty confidence intervals for the many of the leftmost configurations, although this is less pronounced for the two-sided confidence intervals. These two methods produce confidence

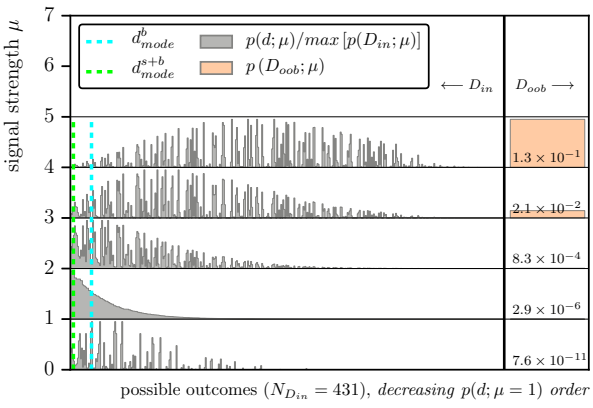
[17] Because rejection occurs from left to right in figure 6.4 (a), very little rejection occurs in D_{oob} so that there is virtually no blue area to view (a tiny sliver is visible near $\mu = 8.0$ with high magnification). A small amount of blue area is visible in figure 6.4 (b), where it appears as a dark edge on the right side of the orange area. Note that this technical detail is not particularly important for this discussion.



(a)



(b)



(c)

Figure 6.3: Example 2 – The bivariate Poisson model selected probability distributions with different global ordering demonstrate that no global ordering provides smoothly varying distributions for all μ simultaneously. (a) MLE ordering, (b) background ordering, and (c) global ordering by decreasing probability of $p(d;\mu = 1)$.

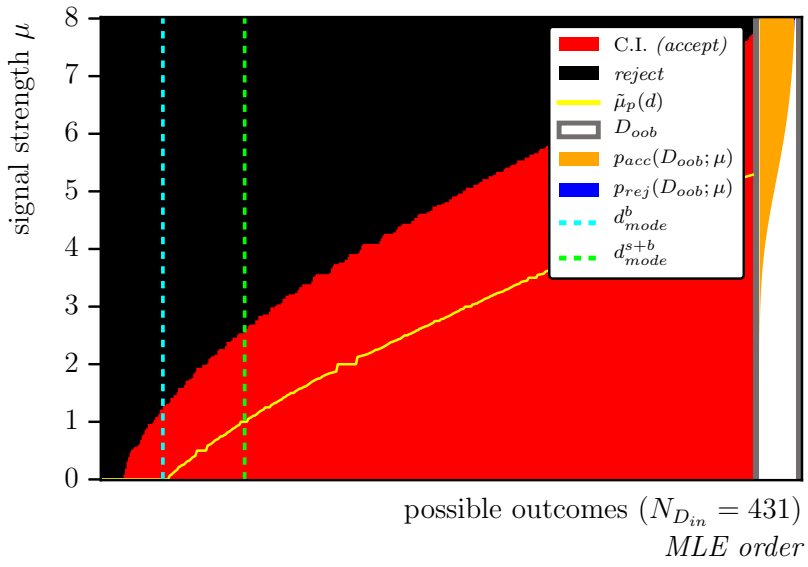
intervals that are surprisingly similar to those produced by common methods that use likelihood ratio test statistics (section 6.2.3.4).

A few additional properties of the multivariate Poisson signal model are examined next. Figure 6.5 (a) shows the probability distributions and likelihood functions for $\mu \in [0, 8]$ and $d \in D_{in}$ using the background ordering. If one selects a value μ' and views the horizontal row corresponding to that μ', then one obtains the model distribution, $p(d; \mu')$, for the 431 enumerated outcomes ($d \in D_{in}$). If instead one selects an outcome d' and views its vertical column, the one obtains the likelihood function, $p(d'; \mu)$ for $\mu \in [0, 8]$. In figure 6.5 (a), the horizontal row for $\mu = 0$ corresponds to the background-only model distribution shown in figure 6.2(a). The probability distributions (horizontal rows) for increasing μ become less background-like and are increasingly rippled. Meanwhile, the several dark vertical bands show that some configurations of $d \in D_{in}$ are simply much less probable than others regardless of the value of μ.

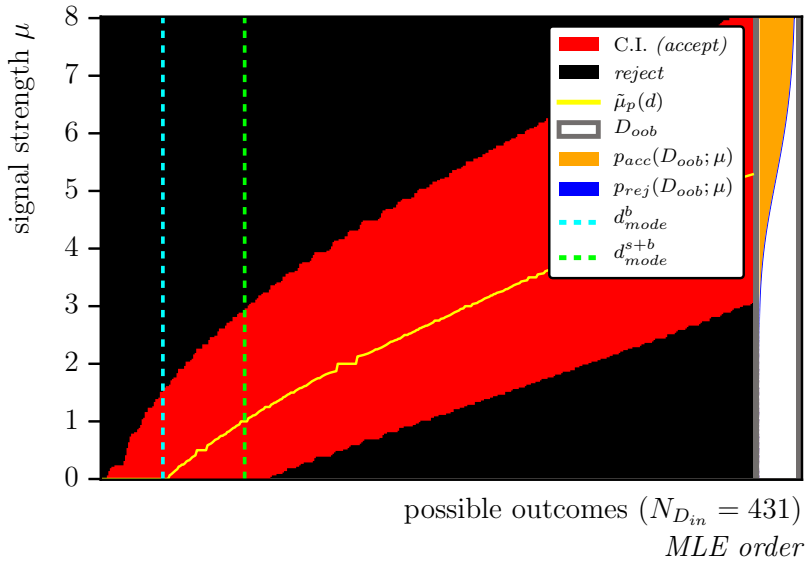
Figure 6.5 (b) shows these probability distributions under the MLE global ordering. The yellow point delineating $\tilde{\mu}_p(d)$ for a fixed d' is the maximum of the likelihood function (vertical column), $p(d'; \mu)$. The likelihood function for every d decreases smoothly from its maximum at $p(d; \tilde{\mu}_p(d))$ as values of μ become more distant from $\tilde{\mu}_p(d)$.

This is illustrated more clearly in figures 6.6 (a)–(f) that show the likelihood functions for selected configurations d. These demonstrate that the likelihood functions for the multivariate Poisson model always have a single maximum value, $\tilde{\mu}_u(d)$, although this may sometimes be unphysical (negative).[18] Figure 6.7 shows the likelihood functions with greater detail than figure 6.5 (b).

[18] Figures 6.6 (a)–(f) also show a positive width value that is merely provided to aid the reader in comparing the shapes of the likelihood functions.

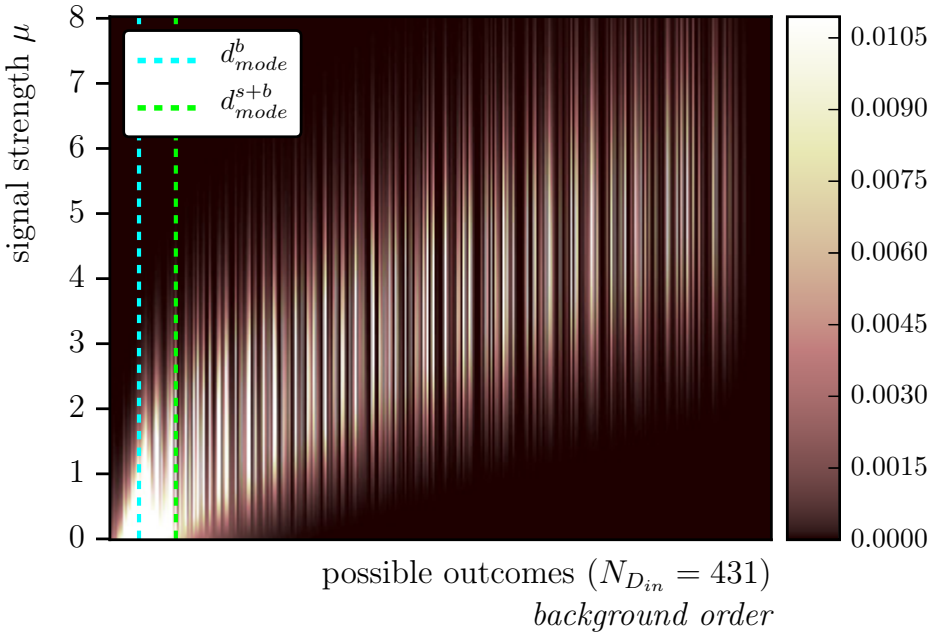


(a)

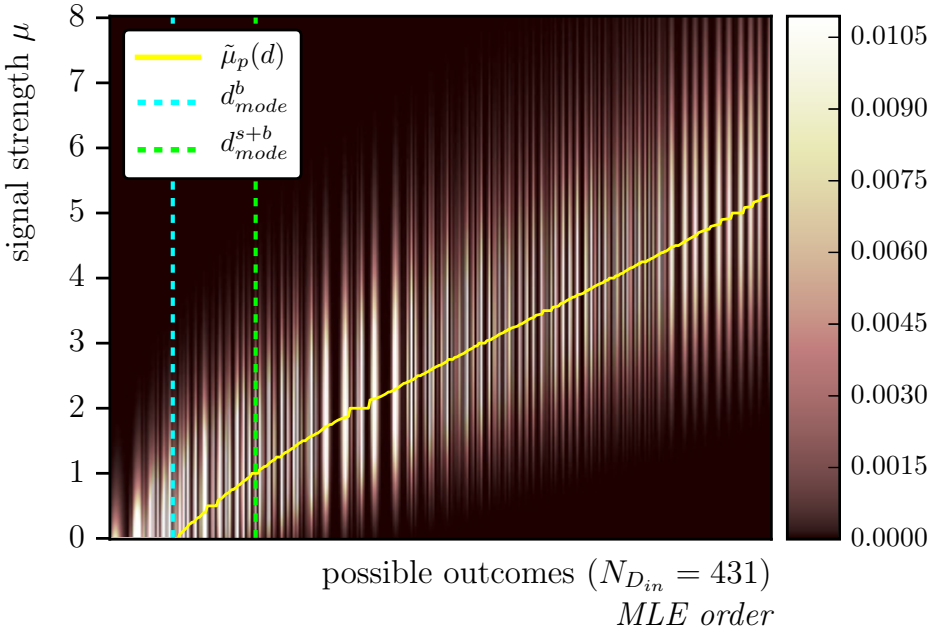


(b)

Figure 6.4: Example 2 – The Neyman construction of confidence regions using global ordering for the bivariate Poisson model: (a) one-sided confidence intervals (upper limits) determined in the style of Garwood [273], and (b) two-sided confidence regions in the style of Ricker [274]. Red horizontal rows show the acceptance region for each μ while red vertical columns show the confidence region for each d. A yellow line shows the maximum likelihood of the likelihood function for each d. The right side bar tabulates the amount of acceptance and rejection that occurs in D_{oob} for a given μ (detailed description in the text).



(a)



(b)

Figure 6.5: Example 2 – The probability distributions (horizontal rows) and likelihood functions (vertical columns) of the bivariate Poisson model: (a) background ordering, (b) MLE ordering that also shows the maximum likelihood value, $\tilde{\mu}_p(d)$, for each $d \in D_{in}$.

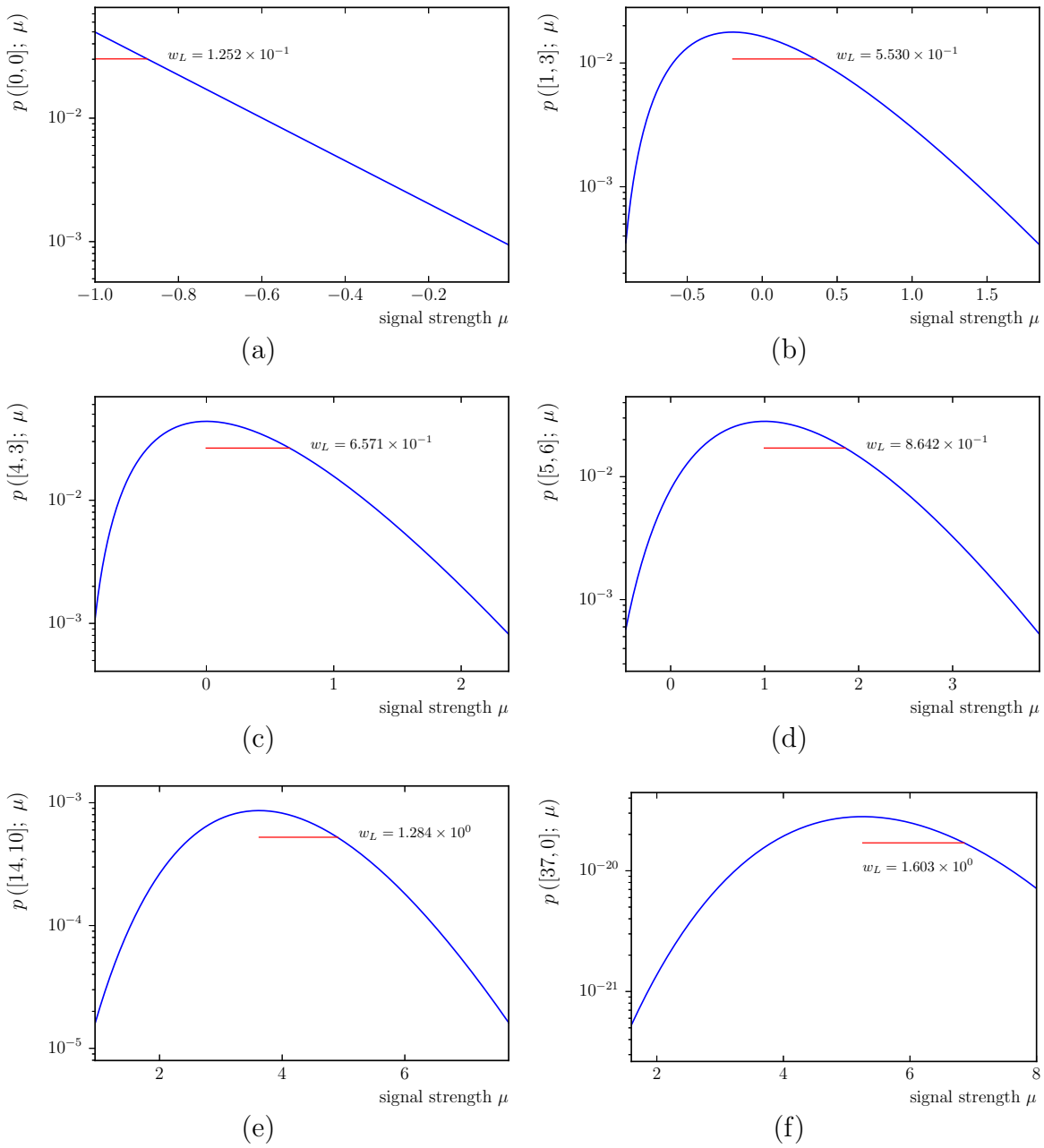


Figure 6.6: Example 2 – The bivariate Poisson model: selected likelihood functions, $p(d; \mu)$, for several possible outcomes d'. Horizontal red line segments denote the width of the likelihood function, w_L, determined as the location, μ', from the maximum, $\tilde{\mu}_p$, where $p(d';\ \mu') = e^{-1/2}\, p(d';\ \tilde{\mu}_p)$, which corresponds to a "$+1\sigma$" change when viewing the likelihood function as an unnormalized Gaussian distribution. (a) $d = [0, 0]$, the first outcome of D_{in}, (b) $d = [1, 3]$, (c) $d^b_{mode} = [4, 3]$, (d) $d^{s+b}_{mode} = [5, 6]$, (e) $d = [4, 10]$, and (f) $d = [37, 0]$, the last outcome of D_{in} when using the MLE ordering.

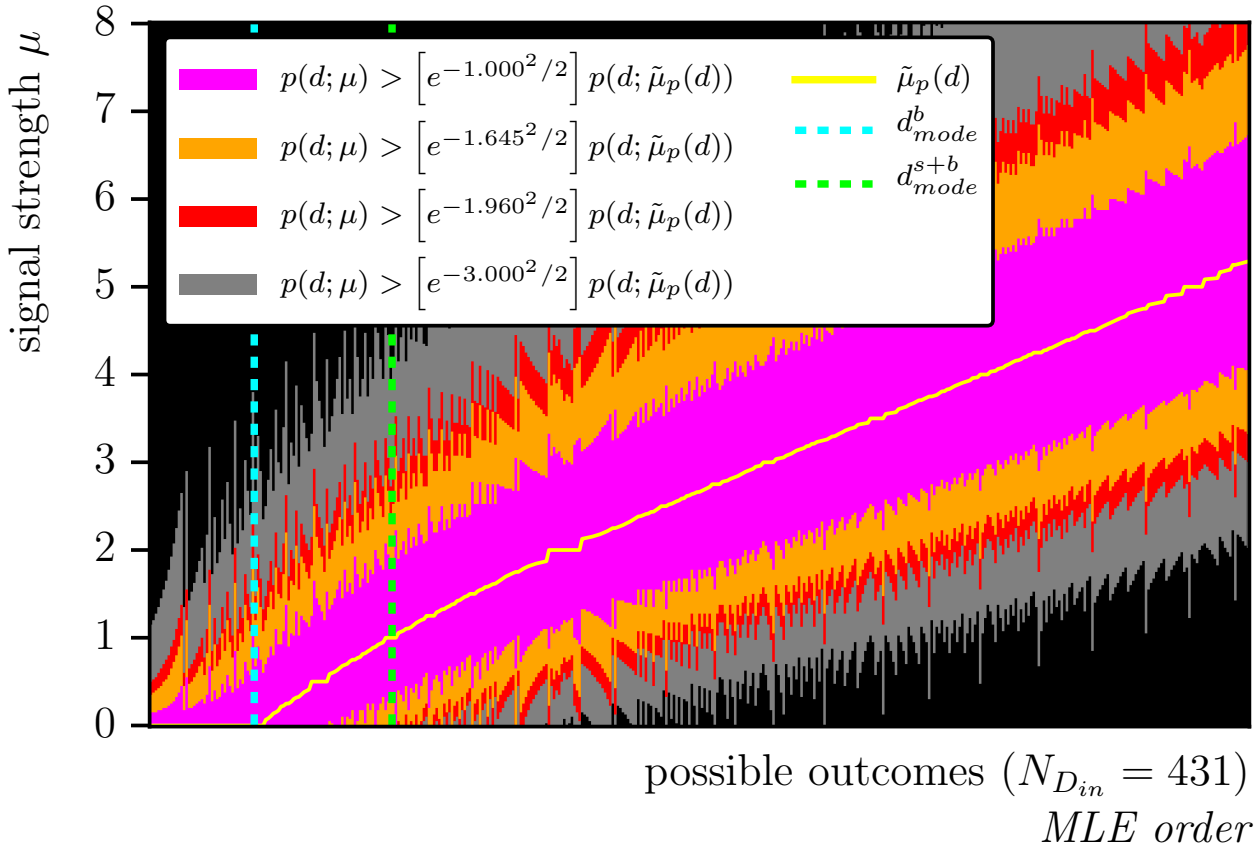
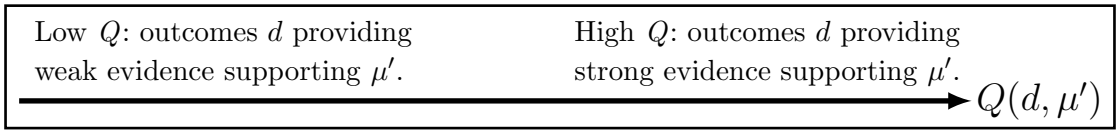


Figure 6.7: Example 2 – The likelihood functions of the bivariate Poisson model for the enumerated outcomes plotted with the MLE global order. The likelihoodist method [69, 255] of solving for μ where the likelihood function has decreased to a given fraction of its maximum value (treating the likelihood function as an unnormalized Gaussian probability distribution) is implemented for: "$\pm 1\,\sigma$" (68.3%) in violet, "$\pm 1.645\,\sigma$" (90%) in orange, "$\pm 1.96\,\sigma$" (95%) in red, and "$\pm 3\,\sigma$" (99.7%) in gray.

6.2.3. Determining Acceptance Regions using Test Statistics

Neyman confidence regions are determined by the how the possible outcomes d are selected for inclusion in the acceptance regions of each value of μ. In the previous examples, this was accomplished by summing tail probabilities where the ordering of d in the tails was defined globally, i.e., the same ordering regardless of the value of μ. However, the order in which outcomes d are selected may depend on μ, and hereby called a "local ordering rule".

For a specific value μ', a local ordering rule, $Q(d, \mu')$, assigns a (possibly non-unique) real value to each outcome d. The values of $Q(d, \mu')$ determine the order in which outcomes d are excluded from each acceptance region. The local ordering rule for a particular value μ' may be arbitrary,[19] but a sensible design is for $Q(d, \mu')$ to order the possible outcomes d from those providing weak evidence supporting the existence of the signal with signal strength μ', according to some criteria, to those d yielding the strongest evidence for μ'.[20] This is illustrated below by viewing $Q(d, \mu')$ as an axis along which the possible outcomes d are positioned.



With such a local ordering rule defined for μ', an acceptance region may be constructed by excluding outcomes d providing the weakest evidence for μ'. The actual values of $Q(d, \mu')$ are irrelevant; only their values relative to one another matter for determining which outcomes d are to be included in an acceptance region.[21]

Given a $(1 - \alpha)$ confidence level, there is a critical value $Q^*_\alpha(\mu')$ such that all configurations d are excluded from the acceptance region if $Q(d, \mu') \leq Q^*_\alpha(\mu')$. The critical value for each value of μ, $Q^*_\alpha(\mu)$, are chosen to guarantee frequentist coverage,

$$\sum_{q_i \leq Q^*_\alpha(\mu)} \left(\sum_{Q(d',\mu)=q_i} p(d'; \mu) \right) \leq \alpha, \, \forall \mu, \tag{6.2.16}$$

where the second summation accounts for configurations with equal values of Q. Equation 6.2.16 could be satisfied for any small enough value of $Q^*_\alpha(\mu)$, but the critical value is also chosen to minimize overcoverage so that any increase in $Q^*_\alpha(\mu)$ would reject

[19] A local ordering rule may be defined in any manner, even randomly, but if care is not taken, the resultant confidence regions may be highly discontinuous and of little use for drawing conclusions from the observed outcome.

[20] The local ordering rule for μ' might also be determined by examining the strength of evidence provided for values of μ other than μ'. This motivates the use of likelihood ratios as test statistics.

[21] A local ordering rule that gives the same value to many outcomes d may be problematic since these outcomes would be unordered relative to one another, and this may lead to overcovering and other difficulties. This occurs with the ATLAS recommended test statistic discussed in section 6.2.3.4.

additional outcomes and violate equation 6.2.16,

$$\sum_{\substack{q_i \\ q_{i-1} \le Q^*_\alpha(\mu)}} \left(\sum_{Q(d',\mu)=q_i} p(d';\mu) \right) > \alpha, \; \forall \mu \tag{6.2.17}$$

where it is assumed the values q_i are indexed in increasing magnitude so that the inclusion the next value of Q in the summation would result in less than $(1-\alpha)$ confidence.

These may also be written in terms of the probability distribution of $Q(d,\mu')$ by noting that $p(q;\,\mu) = \sum_{d'} p(d';\,\mu \,|\, q = Q(d',\mu))$, so that for all μ,

$$\sum_{q_i \le Q^*_\alpha(\mu)} p(q_i;\,\mu) \le \alpha \quad \text{and} \quad \sum_{\substack{q_i \\ q_{i-1} \le Q^*_\alpha(d;\,\mu)}} p(q_i;\,\mu) > \alpha. \tag{6.2.18}$$

This is illustrated in figure 6.8 by viewing a hypothetical probability distribution, $p(Q;\,\mu')$, of some local ordering rule $Q(d,\mu')$ for the value μ'. The outcomes are discrete so that the output values of $Q(d,\mu')$ are also discrete: q_0, q_1, ... For 90% confidence, $Q^*_{0.1}(\mu')$ must be selected to reject only the first two values with $p(q_0;\mu') + p(q_1;\mu') = 9\%$ which results in 1% overcoverage.

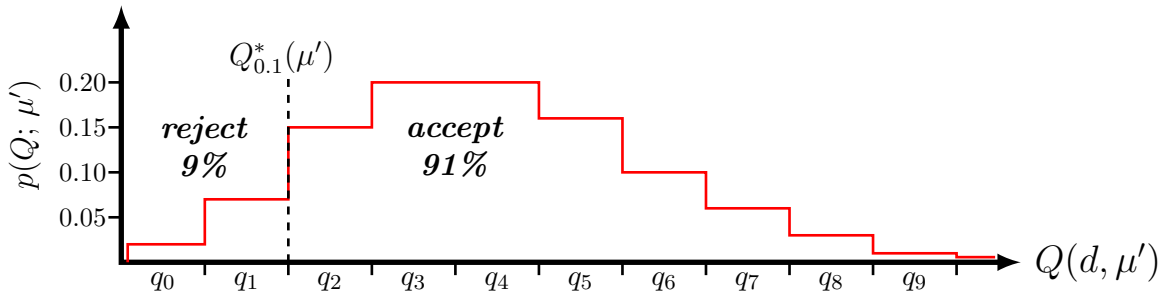


Figure 6.8: An example demonstrating the determination of an acceptance region using a local ordering rule, $Q(d,\mu)$.

The critical value $Q^*_\alpha(\mu)$ defines the non-random rejection indicator function at $(1-\alpha)$ confidence level as

$$\phi^\alpha_{rej}(d,\mu) = \begin{cases} 1 & \text{if } Q(d,\mu) \le Q^*_\alpha(\mu) \\ 0 & \text{if } Q(d,\mu) > Q^*_\alpha(\mu) \end{cases}, \; \forall \mu. \tag{6.2.19}$$

The rejection indicator function, $\phi^\alpha_{rej}(d,\mu)$, is useful for analyzing the results of the Neyman construction because it simultaneously summarizes the location of the acceptance regions and the location of the confidence regions. Figure 6.9 shows a hypothetical example demonstrating this. The rejection indicator function is especially useful for software applications where a grid of coordinates (d,μ) is defined (possible outcome vs. parameter value) because it allows rapid computation of various measures of confidence interval properties, such as $R^\alpha_{size}(\mu)$ defined in equation 6.2.13.

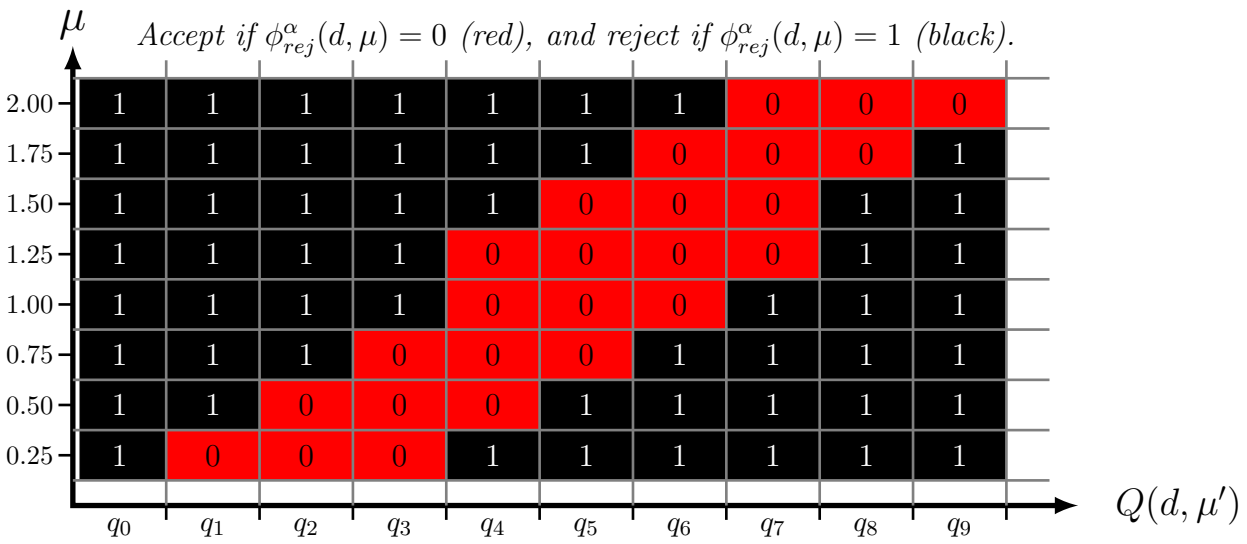


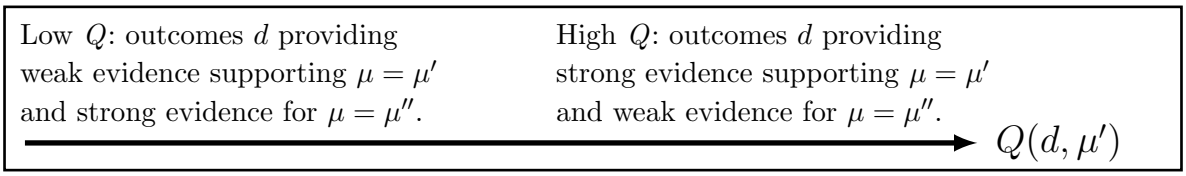
Figure 6.9: A hypothetical example demonstrating how the rejection indicator function, $\phi^{\alpha}_{rej}(d,\mu)$, specifies both the acceptance regions (horizontal red) and the confidence regions (vertical red) after performing the Neyman construction using test statistic $Q(d,\mu')$ for some confidence level $(1-\alpha)$.

6.2.3.1. Neyman-Pearson Hypothesis Testing

The statistical procedure of Neyman-Pearson hypothesis testing (NPHT) is closely related to the process of Neyman confidence region construction (NCRC). Assuming that the signal strength parameter μ is equal to one of only two values, μ' and μ'', a NPHT may be performed to compare the *null hypothesis* that $H_N : \mu = \mu'$ against the *alternative hypothesis* that $H_A : \mu = \mu''$. The result of the NPHT is an acceptance region, a set of possible outcomes for which H_N is accepted and H_A rejected, while the remaining outcomes form a rejection region, a set of possible outcomes where H_N is rejected and H_A accepted.[22] A test statistic is used in NPHT to determine the acceptance and rejection regions. This is similar to the NCRC described in the previous section where for each μ, a local ordering rule is used to define acceptance and rejection regions. These two methods are made even more similar when a test statistic from NPHT is used as the local ordering rule in NCRC.

Two types of error may occur in NPHT. A type I error occurs when H_N is actually true but rejected while a type II error occurs when H_A is true but rejected. The test size, α, is the probability of type I error and is set to a desired level by the experimenter. The selection of a test statistic and the test size allow for the determination of the type II error probability, β. The test power is $(1 - \beta)$ so that greater power is desirable as it corresponds to smaller type II error probability. The test size and power are determined before any knowledge of d_{obs} is obtained in order to prevent any bias by the experimenter. Once d_{obs} is obtained, its test statistic value determines whether H_N or H_A is accepted.

A test statistic,[23] $Q(d, \mu')$, is a function that returns a value for each possible outcome d. Similar to the local ordering rule used in the NCRC, a sensible design for a test statistic is to return high values for outcomes that simultaneously provide strong evidence for H_N and weak evidence for H_A according to some criteria.



This is illustrated in figure 6.10 by viewing two hypothetical probability distributions, $p(q; \mu')$ shown in red and $p(q; \mu'')$ shown in blue, where q_i are discrete output values of $Q(d, \mu')$. Here the desired test size is 10%, but due to the discrete nature of the outcomes, a critical value $Q^*_{0.1}$ can only be found that achieves a type I error probability of 9%. The test power, $(1 - \beta)$, is the probability of accepting H_A when it is true, 25% in this case.

[22] Technically, the set of outcomes for which the null hypothesis is rejected is called the critical region. For the purpose of continuity with concepts from confidence region construction, the set of outcomes for which the null hypothesis is accepted is referred to as the acceptance region, and those rejected giving the rejection region.

[23] A note on the notation for test statistics and local ordering rules. Both are written as $Q(d, \mu')$, where Q is a function of the possible outcomes d. The μ' in the test statistic function indicates that the null hypothesis is $\mu = \mu'$. The μ' in the local ordering rule indicates that the ordering is for the acceptance region of the value μ'. In reality, $Q(d, \mu')$ is often parametrized by more than one value of μ, such as with likelihood ratios, e.g., $Q(d, \mu') = p(d; \mu')/p(d; \mu'')$.

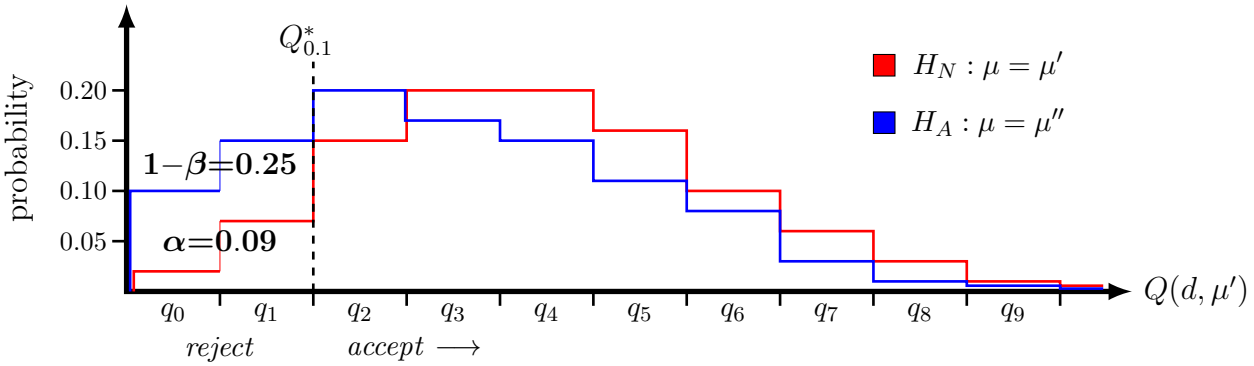


Figure 6.10: An example demonstrating the determination of the acceptance and rejection (critical) regions using a test statistic, $Q(d, \mu')$.

A test statistic, $Q(d, \mu)$, in NPHT orders the outcomes which allows acceptance regions to be determined for any desired type I error probability, α. Therefore, a test statistic may also be used as a local ordering rule, $Q(d, \mu)$, in NCRC to construct acceptance regions for any desired level of confidence, $(1 - \alpha)$. Note that the parameter α has a slightly different interpretation between these two inference methods. For NPHT, if H_N is true, then the test size α is the probability that one incorrectly rejects H_N based on the observed outcome. For the NCRC, $(1 - \alpha)$ is the confidence level, and α is the probability that the observed outcome is not contained in the acceptance regions of each value of μ, including μ_{true}. It is in this sense that the NCRC treats all models equally since the confidence level $(1 - \alpha)$ applies to the acceptance regions of all values of μ simultaneously, while in NPHT the test size α only applies to the null hypothesis and not any alternative.

The test size and test power of NPHT may be reframed in terms of risk functions. In section 6.3.4, reporting risk is defined that motivates the use of the Arizona test statistics, and so it is instructive to discuss here two particular risk functions: $R^{\alpha}_{size}(\mu')$ and $R^{\alpha}_{pow}(\mu', \mu'')$. The risk of power loss, $R^{\alpha}_{pow}(\mu', \mu'')$ (defined below), provides motivation for the use of likelihood ratios as test statistics (NPHT) and local ordering rules (NCRC).

$R^{\alpha}_{size}(\mu')$ is defined in equation 6.2.13 (and repeated below for convenience) as the total probability of the rejection region in the Neyman construction using the signal model $p(d; \mu')$. The experimenter predetermines the confidence level, $(1-\alpha)$, so that $R^{\alpha}_{size}(\mu') = \alpha$ is always true by construction (or possibly slightly less than α for discrete models). Examining equation 6.2.20, one sees that the loss function corresponding to this risk is simply the rejection indicator function, $L^{\alpha}_{size}(d, \mu') = \phi^{\alpha}_{rej}(d, \mu')$. Strictly speaking, there is no rejection indicator function in NPHT, but one may define it similarly as in the Neyman construction so that $R^{\alpha}_{size}(\mu')$ is defined for NPHT using the predetermined test size, α, and the null hypothesis model distribution, $p(d; \mu')$,

$$R^{\alpha}_{size}(\mu') \equiv E_{\mu'}[\phi^{\alpha}_{rej}(d', \mu')] = \sum_{d'} p(d'; \mu')\, \phi^{\alpha}_{rej}(d', \mu') \leq \alpha. \tag{6.2.20}$$

$R^{\alpha}_{size}(\mu')$ does not depend on the alternative hypothesis model distribution. However, the power loss risk (defined below) is determined using the alternative hypothesis model distribution instead of the null hypothesis model. The test power, $(1-\beta)$, is the alternative model probability of an outcome being in the rejection region. Therefore, β is the type II error probability, the risk of accepting H_N when really H_A is true. This is simply the sum of probability masses[24] for outcomes in the acceptance region using the alternative hypothesis, $p(d;\,\mu'')$,

$$R^{\alpha}_{pow}(\mu',\mu'') \equiv E_{\mu''}\left[1-\phi^{\alpha}_{rej}(d',\mu')\right] = \sum_{d'} p(d';\mu'')\,\left[1-\phi^{\alpha}_{rej}(d',\mu')\right] = \beta, \tag{6.2.21}$$

where the subscript "pow" is used to denote that this is the risk of loss of power because increasing β is equivalent to decreasing power, $1-\beta$. Note that the power loss risk is a function of both μ' and μ'' since it is the rejection indicator function for μ' that determines the risk of power loss, while the probability distribution for μ'' is used to obtain the expected value. The loss function corresponding to $R^{\alpha}_{pow}(\mu',\mu'')$ is simply one minus the rejection indicator function, $L^{\alpha}_{pow}(d;\,\mu',\mu'') = 1-\phi^{\alpha}_{rej}(d;\mu')$.

In NPHT, $R^{\alpha}_{pow}(\mu',\mu'')$ may be minimized (so that power is maximized) by selecting the likelihood ratio as the test statistic, $Q(d,\mu') = \frac{p(d;\,\mu')}{p(d;\,\mu'')}$, as proven by the Neyman-Pearson lemma [250]. This is sensible since outcomes d with large $p(d;\,\mu')$ and small $p(d;\,\mu'')$ provide strong evidence supporting μ' and weak evidence for μ''.

The Neyman-Pearson lemma motivates the use of likelihood ratios as test statistics in order to minimize power loss risk, but the denominator of the likelihood ratio depends on which alternative hypothesis the experimenter selects. For a signal search with null hypothesis $H_N : \mu = \mu'$, the selected alternative hypothesis may be the background-only model, $H_A : \mu = 0$.[25] The use of the background-only model as the alternative hypothesis yields the DZero recommended test statistic often used in particle physics signal searches (and discussed in the following sections).

However, one may also compare the null hypothesis to many values of μ simultaneously as the alternative hypothesis.[26] Two common choices for compound alternative hypotheses in a signal search are $H_A : \mu \neq \mu'$ which tests μ' against all possible alternatives, and $H_A : \mu < \mu'$ which tests μ' against all values of μ less than μ'. These two selections of alternative hypotheses yield the Feldman-Cousins and ATLAS recommended test statistics often used in particle physics signal searches (also discussed in the following sections).

When likelihood ratio test statistics are used as the local ordering rules in the Neyman construction, the resultant confidence intervals may be uniformly most accurate (UMA). UMA confidence intervals typically have shorter length than other confidence intervals

[24] Recall that the probability distribution for a continuous random variable describes a probability *density*, while that for a discrete random variable describes a probability *mass*.

[25] The background-only model will often be written using $\mu = \tilde{\mu}_0$ instead of $\mu = 0$ to emphasize that when nuisance parameters are included, the background-only model will be contain fewer nuisance parameters than the signal model as described in section 6.2.1.

[26] Additionally, the null hypothesis may contain many values of μ. Such compound-vs-compound tests are seldom employed in a particle search. Nevertheless, some were examined, but the results are not reported here because they did not provide any improvements over the other methods: simple-vs-simple and simple-vs-compound.

at equal confidence level, which may be a desirable feature.[27] Nevertheless, the use of likelihood ratios as test statistics implies the goal of minimization of power loss risk, and an experimenter may desire to minimize a different risk function. Neyman and Pearson did note that if decisions that rely on observed outcomes are more elaborate than simple acceptance and rejection of competing hypotheses, then "[...] it is possible that other conceptions of relative value may be introduced," [250]. Additionally, the experimenter must select which alternative hypothesis to use, which determines the form of the denominator in the likelihood ratio.

6.2.3.2. Test Statistics Used as Local Ordering Rules

A test statistic, $Q(d, \mu')$, is a function of possible outcomes d that is used in Neyman-Pearson hypothesis testing (NPHT) to order outcomes from those providing strong evidence for H_A and weak evidence for $H_N : \mu = \mu'$ to the outcomes providing strong evidence for H_N and weak evidence for H_A. A local ordering rule, $Q(d, \mu)$, is a function of outcomes used in Neyman confidence region construction (NCRC) to order outcomes from those providing the least evidence for μ' to those providing the most.

Both these functions order outcomes according to some desired criteria. The difference lies in their design. For NPHT, the test statistic must perform the dual task of evaluating the strength of evidence provided by outcomes supporting H_N and also the evidence for H_A.[28] In determining the acceptance region for a particular value μ' during the NCRC, the local ordering rule need only evaluate the strength of evidence provided by outcomes supporting μ'. Nevertheless, there may be good reason in NCRC to consider alternative hypotheses when evaluating the evidence for μ'. One might consider defining weak evidence for μ' as strong evidence for some other value of μ. If this is the case, then one may consider using a test statistic from NPHT as the local ordering rule for NCRC.

The next section discusses some test statistics used as local ordering rules in the NCRC that are common in particle physics signal searches. Each test statistic used as a local ordering rule in NCRC has $H_{\mu'} : \mu = \mu'$ as its null hypotheses when constructing the acceptance region for the particular value μ'. Where these test statistics differ is in the choice of the alternative hypothesis with which to compare to $H_{\mu'}$.

[27] UMA confidence intervals are not explicitly defined here, and are only mentioned to demonstrate that a local ordering rule may be selected by an experimenter due to the desired properties of the confidence intervals that it produces (as opposed to the properties of the acceptance regions). Whether or not a likelihood ratio delivers UMA confidence intervals depends on several factors that are not discussed here. Regardless, the statement is correct in spirit, likelihood ratios usually produce shorter confidence intervals than other methods (see [277] for a comparison of likelihood ratio confidence intervals to those of Crow-Gardner).

[28] See [278] for a related historical discussion regarding *one theory vs. two.*

6.2.3.3. A Review of Pivots

[29]It is instructive to demonstrate the connection between the Neyman construction using test statistics and the procedure of test statistic inversion in order to determine confidence intervals. A test statistic with a probability distribution that is independent of the model parameters is called a *pivot*. A pivot is useful because it allows for direct computation of confidence intervals without requiring performing the Neyman construction. However, test inversion is nevertheless equivalent to the Neyman construction, as is shown next.

For example, let $\vec{x}$ be N random samples from a normal distribution with unknown mean and unit variance, $x_i \sim \mathcal{N}(\mu, 1)$. Then the test statistic $z = \frac{\bar{x}-\mu}{1/\sqrt{N}}$, where $\bar{x}$ is the sample mean, has a standard normal distribution, $z \sim \mathcal{N}(0, 1)$. Because the distribution for z is independent of μ, it is a pivot. Noting that $\int_{-2}^{2} z \approx 0.95$, one can make the probability statement

$$p\left(-2 \leq z \leq 2\right) \approx 0.95, \tag{6.2.22}$$

$$p\left(-2 \leq \frac{\bar{x}-\mu}{1/\sqrt{N}} \leq 2\right) \approx 0.95, \tag{6.2.23}$$

so that rearranging gives

$$p\left(\bar{x} - 2/\sqrt{N} \leq \mu \leq \bar{x} + 2/\sqrt{N}\right) \approx 0.95, \tag{6.2.24}$$

which defines the two-sided 95% confidence intervals of μ for any possible outcome of $\bar{x}$,

$$\bar{x} - 2/\sqrt{N} \leq \mu \leq \bar{x} + 2/\sqrt{N}. \tag{6.2.25}$$

Once the experiment is unblinded and $\bar{x}_{obs}$ is observed, or $z_{obs} = \frac{\bar{x}_{obs}-\mu}{1/\sqrt{N}}$, the confidence interval $\bar{x}_{obs} - 2/\sqrt{N} \leq \mu \leq \bar{x}_{obs} + 2/\sqrt{N}$ is reported.[30]

This is equivalent to the result that would be obtained using the full Neyman construction. Figure 6.11 shows a graphical description of the Neyman construction for this example. Blue lines represent the probability distributions for the values of μ at the endpoints of the confidence interval. (One would not know *a priori* to select these particular values of μ and would perform this operation for every value of μ.) These are figurative as the vertical height above each dashed line representing the probability distribution value does not correspond to the labeled y-axis. The red area under each probability distribution represents the central acceptance region chosen to yield 95% of each distribution, while the areas in black

[29]The reader may choose to skip this section. The following review is provided to show the connection of the Neyman construction to the common method of determining confidence intervals found in introductory statistics texts, which is based on using local ordering rules that are *pivots*. The conclusion is that using pivots is equivalent to the Neyman construction, but that restricting the selection of the local ordering rule to only pivots limits the experimenter, and that a pivot may not exist for a given parametric model.

[30]Note that the observed confidence interval should not be written as $\mu = \bar{x}_{obs} \pm 2/\sqrt{N}$ since this would imply the central value has some special meaning (perhaps as a point estimate of μ). The Neyman construction only provides confidence regions with no relative importance ascribed to any value of μ contained in the confidence region. Therefore, the observed confidence interval should be written as $[-2/\sqrt{N}, 2/\sqrt{N}]$.

represent the values of z that are rejected. With acceptance regions constructed horizontally, this graph is then viewed vertically to determine confidence regions. After unblinding and z_{obs} is obtained, the specific vertical confidence interval at z_{obs} is reported, shown with a vertical red line (and vertical black lines to delineate values of μ outside the confidence interval).[31]

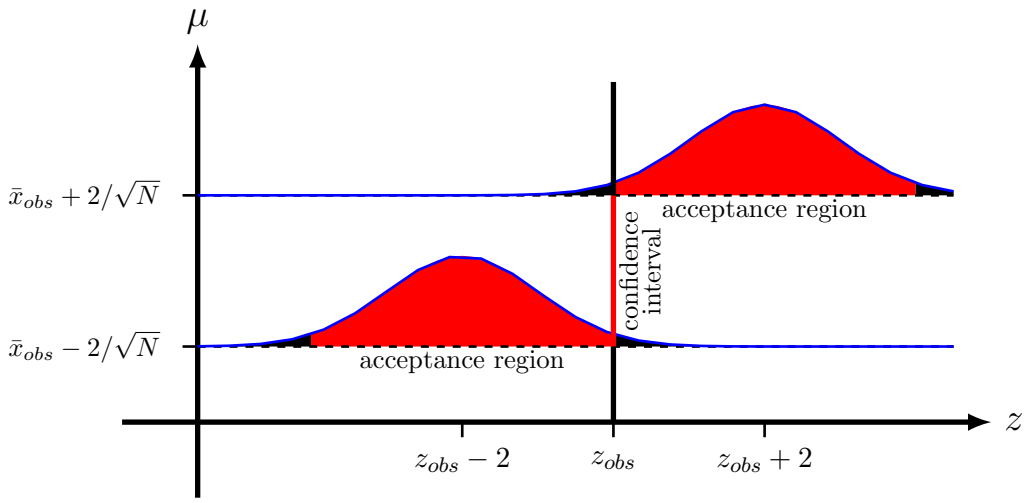


Figure 6.11: An example demonstrating the equivalence of test inversion using a pivot and the full Neyman construction in the determination of confidence intervals.

Test inversion is useful since it simplifies the computational effort needed to determine confidence regions. However, when a suitable pivot cannot be identified or is only approximately valid, the full Neyman construction may always be used to produce exact confidence regions. Also, test statistics that do not happen to be pivots may be more appropriate for a particular experiment in determining which outcomes d provide strong versus weak supporting evidence for a given value of μ. The full Neyman construction allows for the use of the most sensible test statistic, as opposed to test inversion which restricts the analysis to test statistics that are also pivots.

[31] Note that if the variance σ was unknown in this example, then Gosset's ("Student's") t-test statistic has a known distribution that is independent of μ and σ, and so can be used to construct confidence intervals without using the full Neyman construction.

6.2.3.4. Common Test Statistics in Signal Searches

A test statistic, $Q(d, \mu)$, is a function of the possible outcomes used in NPHT to order the outcomes from those most like H_A (outcomes with small values of $Q(d, \mu)$) to those most like H_N (large values). The Neyman-Pearson lemma states that using the likelihood ratio of the two hypotheses as the test statistic provides the most powerful function to discriminate between H_N and H_A [250].

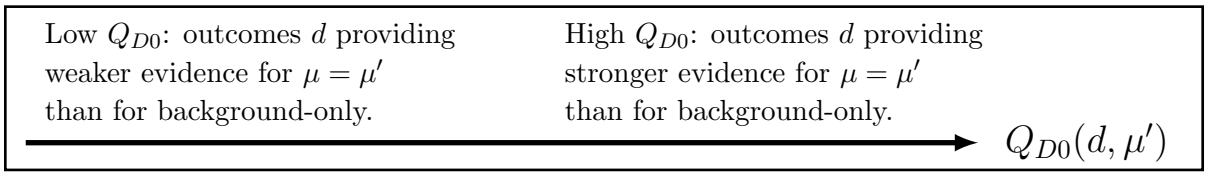
Three common likelihood ratio test statistics are used in particle searches to construct confidence regions, the DZero (D0) recommended test statistic [279], the Feldman-Cousins test statistic [272], and the ATLAS recommended test statistic [280].[32] Each of these three use a different likelihood ratio in order to determine the local ordering of the outcomes d when determining acceptance regions for any given confidence level $(1 - \alpha)$. In some sense, each test statistic uses its own criteria to score the outcomes d, with the highest scores awarded to outcomes that give the best evidence of signal presence. Each test statistic results in different frequentist confidence intervals, and so the experimenter must decide which judging criteria is most appropriate in providing inference results that can be used to make decisions based on real-world observations. Therefore, it is ultimately the kind of decision that needs to be made that determines which test statistic is the most useful. If multifaceted decisions need to be made, there may not be any optimal test statistic.

Ignoring possible nuisance parameters (discussed in section 6.5), the D0 test statistic is

$$Q_{D0}(d, \mu') = \frac{p(d;\, \mu')}{p(d;\, \tilde{\mu}_0)}, \tag{6.2.26}$$

where $p(d;\, \tilde{\mu}_0)$ is the likelihood function of the background-only model. The function notation μ' is written to remind the reader that a specific value of μ is being evaluated by the test statistic. This implies that $0 < Q_{D0}(d, \mu')$ with no upper bound since $p(d;\, \mu')$ may be larger than $p(d;\, \tilde{\mu}_0)$.

For NPHT, this corresponds to the simple-vs-simple test:[33] $H_N : \mu = \mu'$ and $H_A : \mu = \tilde{\mu}_0$, and so for NCRC, each signal model given by the value μ' is evaluated in comparison to the background-only model. This is depicted with the following figure.



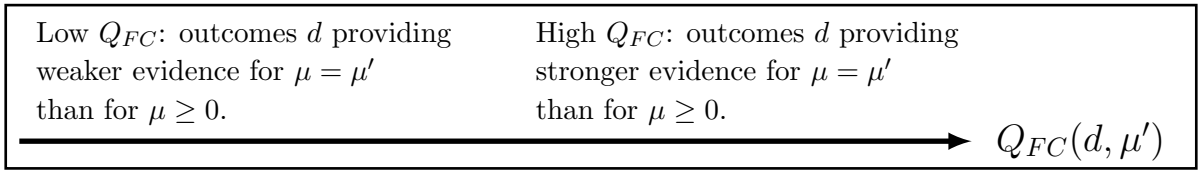
[32] A recent presentation [281] examined four frequentist interval estimation methods in particle physics: "Delta LLR", "CL_S", "Neyman construction", and "p-value". These are equivalent to the presentation here. Delta LLR uses the asymptotic properties of the ATLAS test statistic, approximating the full Neyman construction. CL_S uses the D0 test statistic along with the CL_S method discussed here in section 6.2.4. By Neyman construction, they mean the Feldman-Cousins test statistic. Their p-value method is equivalent to $Q_{MP}(d, \mu') = p(d;\, \mu')$, which accepts the most probable (MP) outcomes first, i.e., locally orders outcomes by probability mass (also discussed below).

[33] Some texts reverse the definition of H_N and H_A given here, but then also reverse the definition of test size and power (see pages 123–125 of [253]). Terminology consistent with standard statistics practice is followed here.

The Feldman-Cousins test statistic is given by[34]

$$Q_{FC}(d, \mu') = \frac{p(d;\, \mu')}{p(d;\, \tilde{\mu}_p)}, \tag{6.2.27}$$

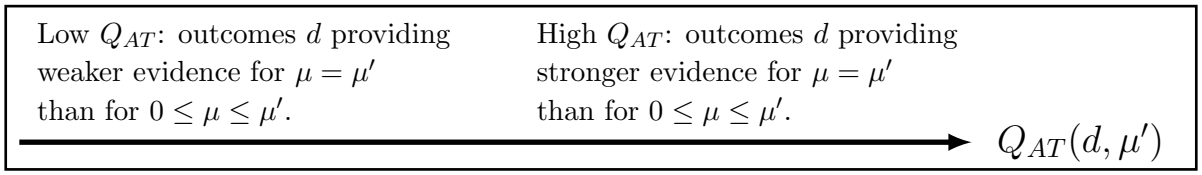
where $\tilde{\mu}_p$ is the physical (non-negative) MLE of the likelihood function that is given for each outcome d', given by $\tilde{\mu}_p = \arg\max_{\mu \geq 0}[p(d';\, \mu)]$.[35] Because $p(d;\, \mu') \leq p(d;\, \tilde{\mu}_p)$, the test statistic values are bounded, $0 \leq Q_{FC}(d, \mu') \leq 1$. For NPHT, this corresponds to the simple-vs-compound test: $H_N : \mu = \mu'$ and $H_A : \mu \neq \mu'$, $\mu \geq 0$. For NCRC, each μ' is evaluated in comparison to the most likely physical value of μ (depicted in the below figure).



The ATLAS test statistic is given by

$$Q_{AT}(d, \mu') = \frac{p(d;\, \mu')}{p(d;\, \tilde{\mu}_b)}, \tag{6.2.28}$$

where $\tilde{\mu}_b$ is the bounded MLE of the likelihood function that is required to be non-negative and less than or equal to the value μ', given by $\tilde{\mu}_b = \arg\max_{0 \leq \mu \leq \mu'}[p(d';\, \mu)]$. Because $p(d;\, \mu') \leq p(d;\, \tilde{\mu}_b)$, the test statistic values are bounded, $0 \leq Q_{AT}(d, \mu') \leq 1$. For NPHT, this corresponds to the simple-vs-compound test: $H_N : \mu = \mu'$ and $H_A : 0 \leq \mu \leq \mu'$. For NCRC, μ' is compared to the most likely physical value of μ bounded above by μ' (depicted in the below figure).



The Feldman-Cousins and ATLAS test statistics both satisfy $0 < Q_{FC}(d, \mu') \leq 1$ and $0 < Q_{AT}(d, \mu') \leq 1$. However, $Q_{FC}(d, \mu') = 1$ only when μ' equals the physical MLE, $\tilde{\mu}_p$, while $Q_{AT}(d, \mu') = 1$ for all μ' that are smaller than $\tilde{\mu}_p$. This often results in many outcomes d yielding the maximum ATLAS test statistic value of 1.

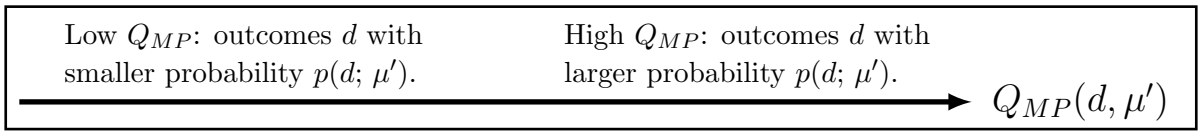
The most probable (MP) ordering is less frequently used in particle physics to construct confidence intervals,

$$Q_{MP}(d, \mu') = p(d;\, \mu'). \tag{6.2.29}$$

[34] In their original paper [272], Feldman and Cousins provided an algorithmic version of their Neyman construction, which is equivalent to the presentation here using a test statistic.

[35] Since an MLE, $\tilde{\mu}$, of the likelihood function, $p(d;\, \mu)$, is obtained for each outcome d, the MLE is a function of the outcomes, $\tilde{\mu}(d)$, but this notation is cumbersome and so is suppressed here.

Rather than base the local ordering rule on likelihood ratios, Q_{MP} forms acceptance regions from the most probable outcomes. This test statistic does not have an associated Neyman-Pearson hypothesis test, but rather may be interpreted as a Fisher significance test (FST) with $H_N : \mu = \mu'$ the only model evaluated. The most improbable outcomes in FST yield the smallest p-values so that Q_{MP} orders outcomes by increasing p-value. For NCRC, the highest probability outcomes d are accepted first (depicted in the below figure).



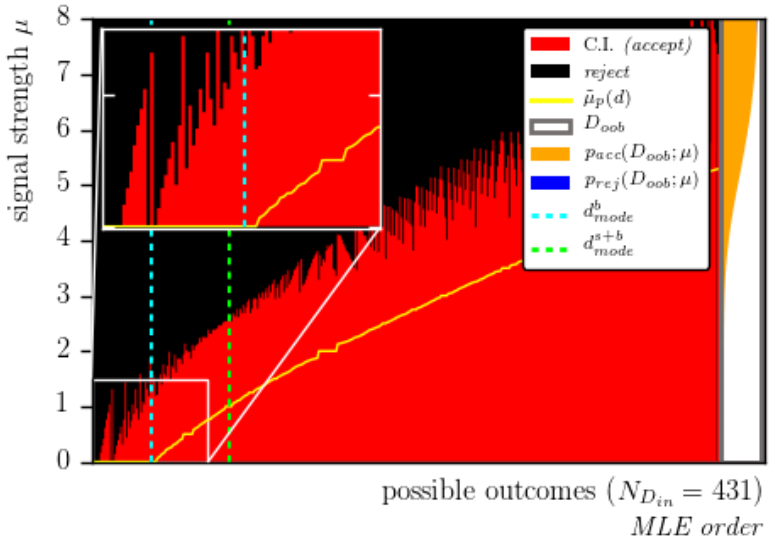
Neyman construction with Q_{D0} applied to example 2.

Figure 6.12 (a) shows the 95%-confidence regions obtained when the D0 test statistic is applied to the bivariate Poisson model of example 2 (section 6.2.2.2). An inset is provided to better examine the results for outcomes d at the far left. At 95% confidence, this results in one-sided confidence intervals (upper limits) for each d, and returns the empty set as the confidence region for some values of d on the left side of the plot.

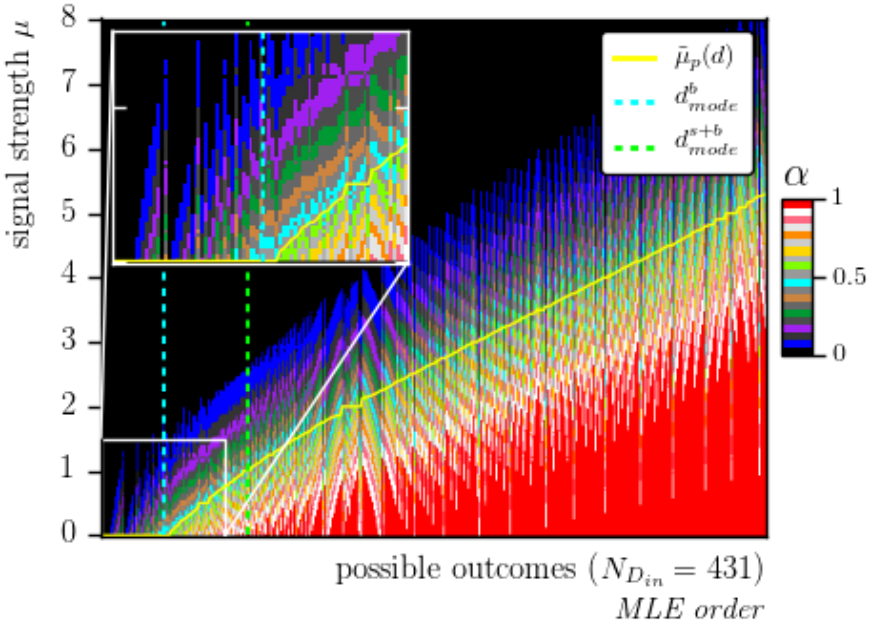
Note that for all Neyman constructions performed in this chapter, 100,000 Monte Carlo simulated outcomes are obtained *for each* μ' from the signal model, $p(d; \mu')$, as well as 100,000 from the background-only model. The values of μ for which acceptance regions are computed are from an equally-spaced grid in steps of 0.025 that extends to the maximum value of μ shown on the μ axis of each generated figure showing results.

One-sided confidence intervals are produced for any level of confidence using the D0 test statistic. The reason for this will be explained in section 6.3, but now it is more instructive to view the performance of the test statistic at several confidence levels simultaneously. This is accomplished by showing the *acceptance ventiles* in figure 6.12 (b). When constructing the acceptance region for a particular value μ' (horizontal row), the distribution of $Q(d, \mu')$ according to the signal model $p(d; \mu')$ may be dividing into twenty quantiles (ventiles). Outcomes in the lowest quantile provide the least evidence for μ' and are colored with a black area. Those in the highest quantile are the most signal-like and are colored with a red area. Thus the acceptance ventiles are constructed horizontally for each value of μ.

When viewed vertically, these acceptance ventiles provide the confidence intervals at confidence levels in increments of 5% confidence level. One should view figure 6.12 (b) as the red area being stacked on top of the white area (so that the white area extends underneath the red area), and so on with the dark blue area lying on top of a black field. For a given outcome d' (vertical column corresponding to d'), the red area shows the 5% confidence level intervals, the white area shows the 10% confidence intervals, and so on with the dark blue area the 95% confidence intervals. The confidence level is given by $(1 - \alpha)$, the sum of probability masses of outcomes within the acceptance region, where α gives the probability of the rejection region. Figure 6.12 (a) can be recovered from figure 6.12 (b) by selecting coordinates (d, μ) which have an acceptance ventile of no more than 95% confidence level ($\alpha > 0.05$), the coordinates that are shown with non-black color.



(a)



(b)

Figure 6.12: Example 2 – The Neyman construction using the D0 test statistic, $Q_{D0}(d, \mu)$: (a) 95% confidence intervals and (b) the test statistic acceptance ventiles.

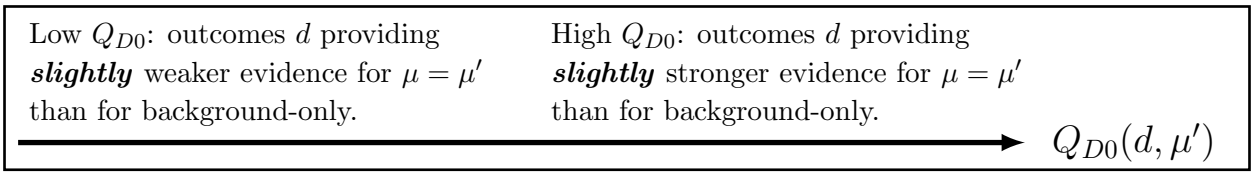
Although the final result of the Neyman construction is obtained by vertically scanning figure 6.12 (b) to obtain confidence intervals at the desired confidence level, figure 6.12 (b) is constructed *horizontally* across all d for each separate value of μ. Figures 6.13 (a)–(f) help to explain how this is achieved. For a given value of μ, Monte Carlo simulations of possible outcomes are generated for both the signal model using $p(d \in D_\mu; \mu)$ and the background-only model using $p(d \in D_0; \tilde{\mu}_0)$, where it is noted that the sets of simulated outcomes will often differ for the two models, D_μ for signal and D_0 for background-only. The values of the test statistic for the simulated outcomes are then plotted in histograms,[36] which provide the probability distributions of both the signal model, $p(q; \mu)$, shown in red, and the background-only model, $p(q; \tilde{\mu}_0)$, shown in gray.

The signal model probability distribution is then sectioned into twenty 5% quantiles, ventiles, of increasing q. All ventiles are not shown to avoid clutter, but a blue line shows the location of the lowest 5% ventile where the sum of $p(q; \mu)$ equals 0.05 for q below the blue line. This lowest ventile is equivalent to the critical value, $Q^*(\mu)$, that is used to determine the acceptance region at 95% confidence.

Introducing the CL_S method.

Figure 6.13 (a) demonstrates that when μ is small, 0.025, the signal and background-only probability distributions are very similar. This is sensible because a weak signal should be difficult to differentiate from background. As the value of μ increases as shown in the remaining figures, these probability distributions become increasingly distinct with the signal model shifting to larger values of the test statistic than the background-only model.

If it were the case that $\mu_{true} = 0.025$, then the outcomes d with test statistic values that fall below the critical value in figure 6.13 (a) would be rejected from the acceptance region, even though the test statistic is unable to differentiate between the signal and background-only models. In this case, the D0 test statistic is only able to provide a feeble spectrum of statistical statements for the possible outcomes (depicted in the below figure).



This motivates the use of the CL_S method [282]. The CL_S method is discussed further in section 6.2.4, but it is most natural to introduce the CL_S method here using figures 6.13 (a)–(f). The aim of the CL_S method is to purposefully overcover during Neyman construction by including more outcomes d in the acceptance regions when the test statistic cannot differentiate between the signal and background-only models.

[36] The binning used to draw the histograms of figures 6.13 (a)–(f) is automated in the software for clarity of presentation, while the exact spectrum is used for computation.

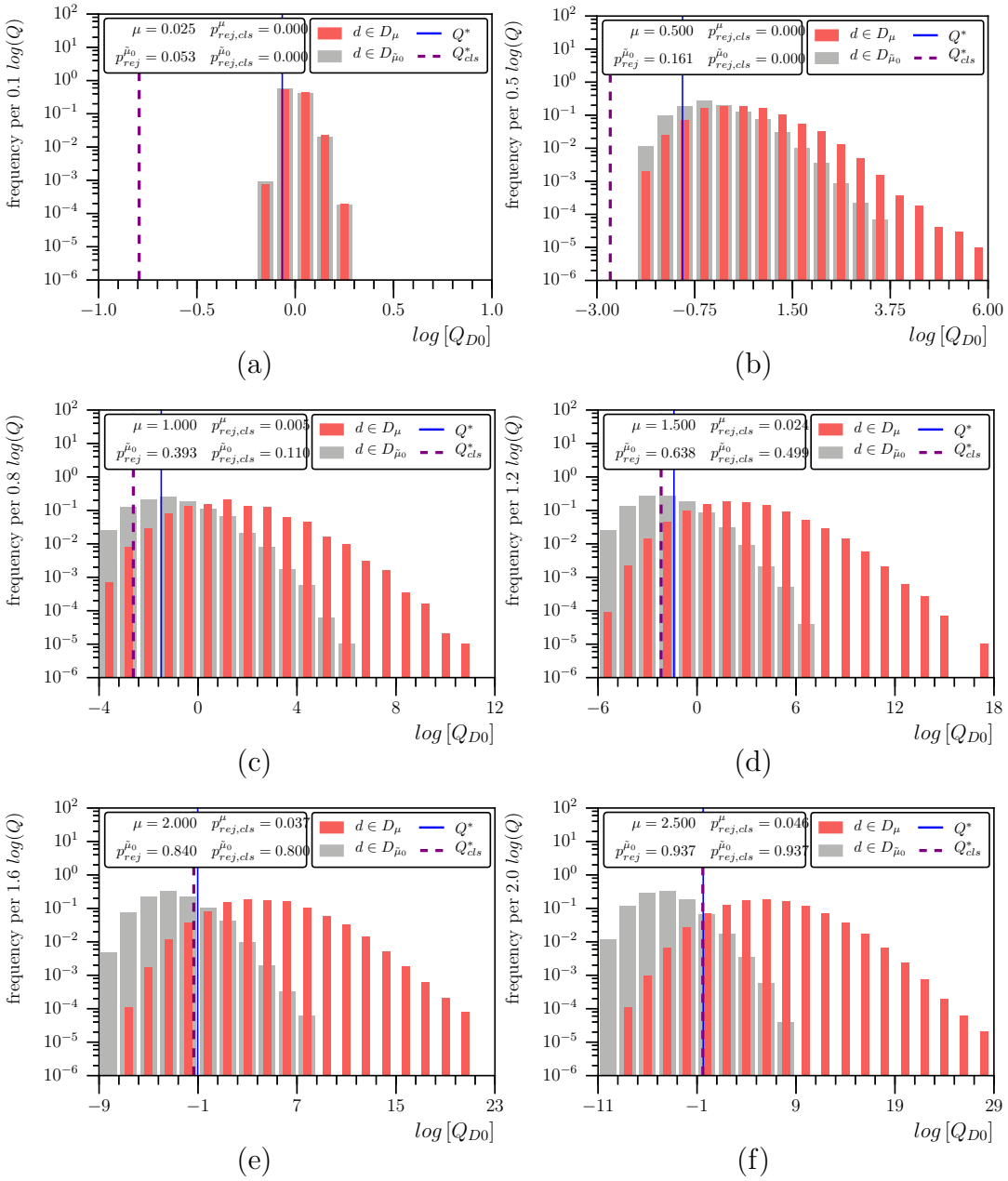


Figure 6.13: $Q_{D0}(d, \mu')$ spectra of Monte Carlo simulations of signal model (red) and background-only model (gray) for selected μ': (a) $\mu' = 0.025$, (b) $\mu' = 0.5$, (c) $\mu' = 1.0$, (d) $\mu' = 1.5$, (e) $\mu' = 2.0$, and (f) $\mu' = 2.5$. Blue lines delineate the 95% confidence critical value, $Q^*(\mu')$, while dashed purple lines show the critical value using the CL_S method, $Q^*_{cls}(\mu')$. It is often not possible to achieve $\alpha_{cls} = 0.05$ in cases of poor signal-background discrimination so that $Q^*_{cls}(\mu')$ is set to below the distribution.

This is achieved by defining the CL_S confidence level, $(1-\alpha_{cls})$, and the corresponding CL_S critical value, $Q^*_{cls}(\mu)$, as those values that solve

$$\frac{p(q \leq Q^*_{cls}(\mu);\, \mu)}{p(q \leq Q^*_{cls}(\mu);\, \tilde{\mu}_0)} = \alpha_{cls}. \tag{6.2.30}$$

If one considers the hypothesis test of $H_N : \mu = 0.025$ and $H_A : \mu = \tilde{\mu}_0$, with test size α^* and test power $(1-\beta^*)$, the CL_S method effectively divides the probability of falsely rejecting the signal model (size) by the probability to correctly accept the background-only model (power),[37] $\alpha_{cls} = \alpha^*/(1-\beta^*)$. The CL_S confidence level is sometimes viewed as the ratio of two p-values [283], although the p-values of FST are derived from observed measurements while α^* and β^* are predetermined from the statistical models.

Figures 6.13 (a)–(f) show the CL_S critical value, $Q^*_{cls}(\mu)$, for a CL_S confidence level of $(1-\alpha_{cls}) = 0.95$ with a dashed purple line. In figure (c), for example, the sum of the red bars to the left of the dashed purple line, divided by the sum of the gray bars to the left of that line, equals 0.05. In figures (a) and (b), the signal and background-only model probability distributions are too similar to achieve this 0.95 CL_S confidence; there is no value of $Q^*_{cls}(\mu)$ that solves equation 6.2.30 and so the critical value is set to include all outcomes as a default.[38] As μ increases and the probability distributions of the signal and background-only models become more easily differentiated, the CL_S critical value, $Q^*_{cls}(\mu)$, converges to $Q^*(\mu)$, the critical value of the Neyman construction. The CL_S method is discussed further in section 6.2.4.

Neyman construction with Q_{FC} applied to example 2.

Figure 6.14 shows the confidence regions obtained when the Feldman-Cousins test statistic is applied to the bivariate Poisson model of example 2 (section 6.2.2.2). The test statistic ventiles are shown which allow one to examine confidence regions for confidence levels $(1-\alpha)$ in steps of 0.05. The Feldman-Cousins test statistic delivers one-sided 95% confidence intervals for outcomes d on the left, and two-sided 95% confidence intervals on the right. For d where $\tilde{\mu}_p = 0$ on the left side of figure 6.14 (a) and enlarged in the inset, most confidence intervals do not reach the 5% confidence level (no red in the vertical column), with the leftmost d only reaching 75% confidence and not lower (the second gray is $\alpha \in [0.20, 0.25)$). Therefore, even though for example 2 the 95% confidence regions are never the empty set for any d, there are empty sets for the confidence regions with lower than 75% confidence.

It is unclear at what confidence level the Feldman-Cousins test statistic guarantees that no empty set confidence regions are produced. The null outcome, $d_i = 0$ for each i^{th} bin (always plotted furthest to the left), typically has the shortest confidence intervals when using Q_{FC}, and so the appearance of an empty set may be expected for the null outcome

[37] In the CL_S method, α_{cls} is predetermined and for each μ, one solves for $Q^*_{cls}(\mu)$ that gives $\alpha_{cls} = \alpha^*/(1-\beta^*)$. Thus the true test size, α^*, will be different for different values of μ. However, $\alpha^* \leq \alpha_{cls}$ for all μ since $0 \leq (1-\beta^*) \leq 1$ so that overcoverage is guaranteed.

[38] The choice to set $Q^*_{cls}(\mu)$ to include all outcomes when equation 6.2.30 cannot be solved is not part of the original definition of the CL_S method. The usual implementation of the CL_s method is to search for the value of μ that provides the upper limit for the desired α_{cls}. However, some decision must be made if one seeks to examine acceptance regions for *all* values of μ.

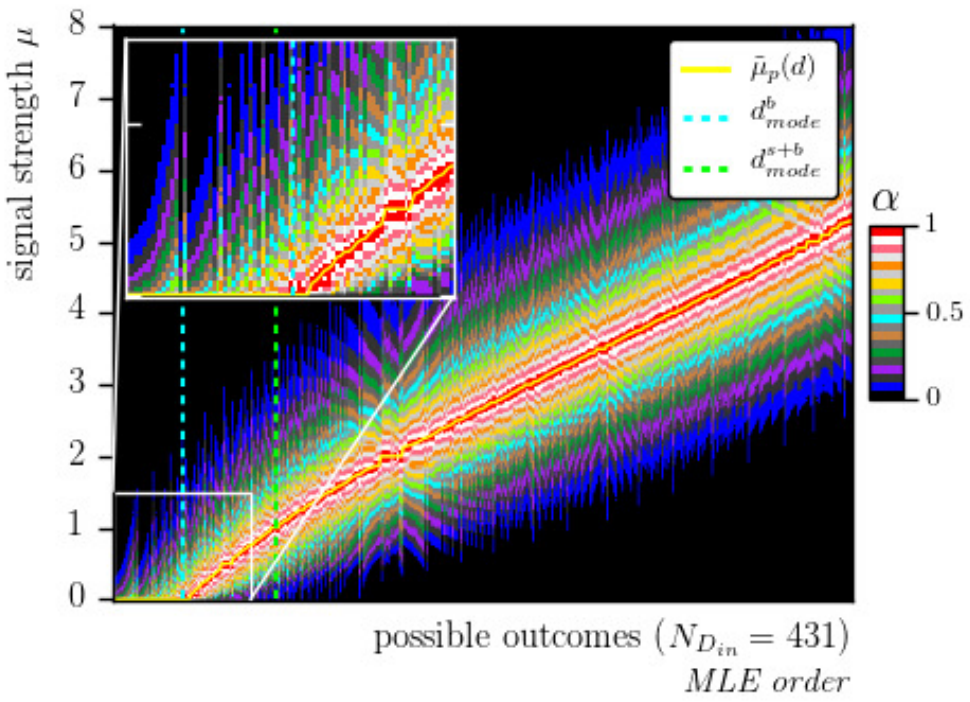


Figure 6.14: The test statistic acceptance ventiles of the Neyman construction for example 2 using the Feldman-Cousins test statistic, $Q_{FC}(d, \mu)$.

before other outcomes. In their example of table 1 of [272], Feldman and Cousins found that for a single bin counting experiment with $b = 3$, the outcome $d = 0$ is contained in the $\mu' = 0.5$ acceptance region at 85.8% confidence level, but not lower (recall that lower confidence level intervals are shorter than higher confidence level intervals). When Q_{FC} is applied to the two binned example 2, $d_i = 0$ is excluded from the $\mu' = 0.5$ acceptance region at confidence levels less than 82%. Setting $\mu' = 0.0001$ as a proxy to determine if the confidence interval for $d = 0$ is empty, $d = 0$ is rejected at 59% confidence level (and lower). One could conclude that for this example that no empty set regions are produced at confidence levels of 60% or higher.

Table 6.3 shows the result of using the Feldman-Cousins test statistic for a multivariate Poisson model with up to 5 bins, where $b = 3$ for each bin. This does seem to imply that Q_{FC} returns non-empty confidence intervals for confidence levels of ∼60% and higher (consistent with some additional scenarios examined but not reported here).

The lack of empty set confidence regions (at least for a reasonably high confidence level) is an attractive feature of the Feldman-Cousins test statistic since this prevents the possibility of the experimenter having essentially nothing to report. However, as was pointed out by Neal in [271], reporting a small frequentist confidence interval may be misinterpreted as a statement of high precision. This would be a misunderstanding of the difference between frequentist and Bayesian statistics, but nevertheless reflects the need for useful and sensible inference results with which to make decisions. The Arizona test statistics and the confidence-acceptance profiles seek to address this issue (discussed in sections 6.3 and 6.4,

	N_{bins} (with $b_i = 3$ in each bin)				
	2	3	4	5	
$\mu = 0.5$	82%	94%	96%	97%	99%
$\mu = 0.0001$	59%	60%	58%	57%	56%

Table 6.3: The confidence level at which the outcome $d = 0$ is rejected from the acceptance region when using the Feldman-Cousins test statistic.

respectively).

Neyman construction with Q_{AT} applied to example 2.

Figure 6.15 shows the confidence regions obtained when the ATLAS test statistic is applied to example 2. These are one-sided intervals with the existence of some empty set confidence regions. The 95% confidence intervals (dark blue) are similar to those produced by the D0 test statistic (figure 6.12 (b)). However, the ventile structure produced by Q_{AT} is uniform (medium gray area of $\alpha \in [0.50, 0.55)$) for $\mu' \leq \tilde{\mu}_p$, unlike that of Q_{D0} for which ventiles are graduated in this region. For an outcome d', any value μ' less than or equal to the physical MLE, $\tilde{\mu}_p$, returns the maximum value of the ATLAS test statistic, $Q_{AT}(d', \mu' \leq \tilde{\mu}_p, \mu) = 1$. This leads to many outcomes receiving the same test statistic value of 1, especially for smaller values of μ. This necessitates an exact definition for how quantiles are computed. When determining the acceptance region for μ', an outcome d' with test statistic value q', is defined as being in the α-quantile if $p(q < q') = \alpha$, i.e., quantiles are defined from below. If more than 5% (in probability) of the test statistic values equals the maximum, then no outcome can have $\alpha \geq 0.95$ since less than 95% of the probability is strictly below the maximum.

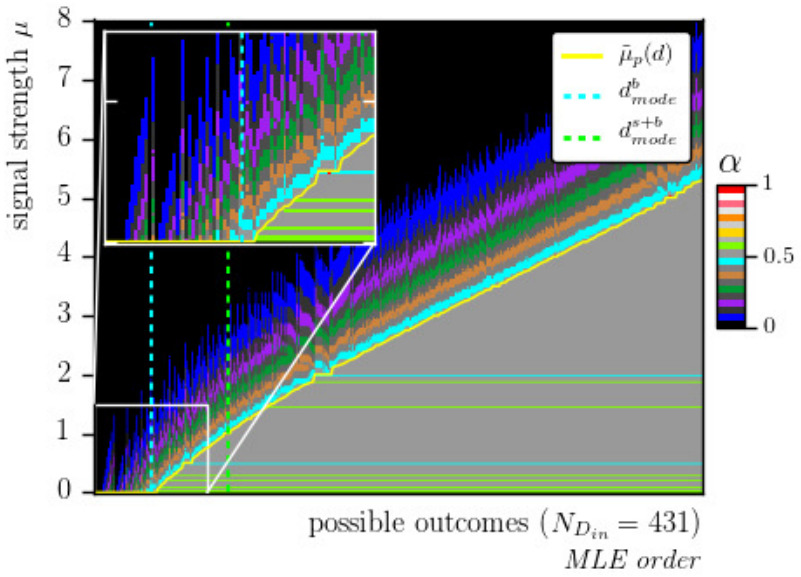


Figure 6.15: The test statistic acceptance ventiles of the Neyman construction for example 2 using the ATLAS test statistic, $Q_{AT}(d, \mu)$.

Neyman construction with Q_{MP} applied to example 2.

Figures 6.16 shows the acceptance ventiles obtained using the most probable test statistic, $Q_{MP} = p(d; \mu)$. These simply reflect the magnitude of the probability distribution itself, which can be seen by comparing to figure 6.5 (b) on page 273. Like the previous test statistics' results, the general trend is for the confidence regions to follow the physical MLE, $\tilde{\mu}_p$. However, use of Q_{MP} results in many empty confidence regions, and these do not necessarily occur for the same outcomes as for Q_{D0} and Q_{AT} (note the difference at the extreme left).

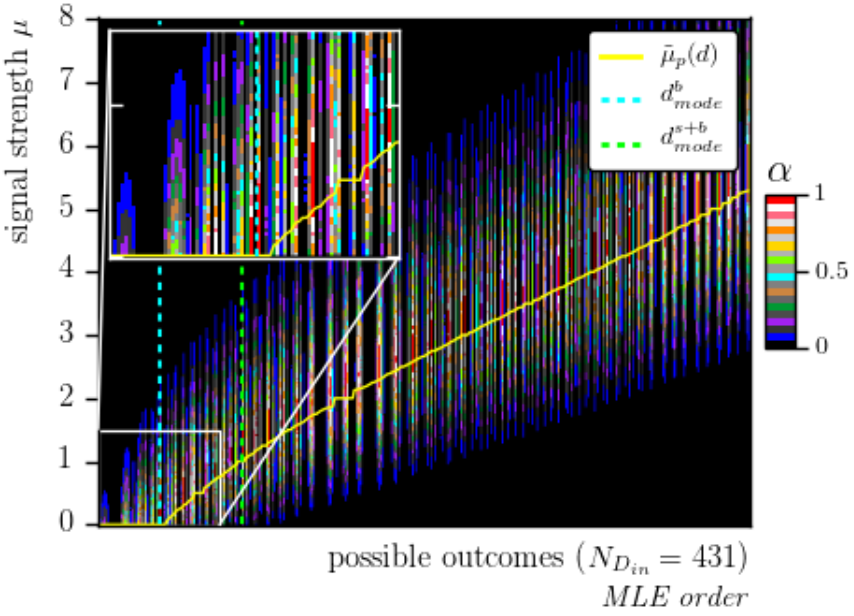


Figure 6.16: The test statistic acceptance ventiles of the Neyman construction for example 2 using the most probable test statistic, $Q_{MP}(d, \mu)$.

6.2.4. Discussion of the CL_S Method

As mentioned in the previous section, the CL_S method attempts to address the situation where a test statistic provides poor discrimination between the signal model and the background-only model [284, 285]. This is done by lowering the critical values, $Q^*(\mu)$, to $Q^*_{cls}(\mu)$ in accordance with equation 6.2.30 so that more outcomes enter the acceptance regions. This increases the size of the acceptance region (overcovers), which causes the confidence intervals to become longer than when using $Q^*(\mu)$ as the critical value. The CL_S method provides an upper limit that approximates an upper limit that would be obtained in a search with no expected background [286, 287]. It should be noted that the CL_S method has been criticized for its lack of foundation in theoretical statistics [288]. This criticism is partially addressed with the signal significance ratio (SSR) function defined in section 6.4.6.

The D0 test statistic, Q_{D0}, uses the background-only model as the alternative hypothesis, and by the Neyman-Pearson lemma, Q_{D0} will provide the strongest discrimination between the signal and background-only model. However, a test statistic may be designed to compare the signal model to any alternative model hypotheses, such as with Q_{FC} and Q_{AT}. The CL_S method causes a larger increase in confidence interval length when using test statistics other than Q_{D0} because those do not use the background-only model as the alternative hypothesis. The CL_S method was first applied to the D0 test statistic [282, 286, 287, 289], but has also been applied to the ATLAS test statistic [280], which produces one-sided intervals similar to the D0 test statistic.[39]

The main focus of this chapter is on the properties of various test statistics used in the Neyman construction without implementing the CL_S method. Nevertheless, results using the CL_S method will often be presented due to its frequent use in particle physics.

Figures 6.17 (a)–(f) provide a comparison of the Neyman construction using Q_{D0} for example 2 both with and without the CL_S method. Figure (a) shows the 95% confidence intervals without using the CL_S method (shown previously in figure 6.12 (a)), and figure (b) shows result of implementing the CL_S method. Note that the CL_S method is implemented separately for each value μ', extending that acceptance region horizontally. This effect occurs primarily for smaller values of μ where discrimination between the signal and background-only models is poorest. For example 2, the extended acceptance regions result in the absence of empty confidence regions, although in general, empty confidence regions may occur when using the CL_S method.

Figure 6.17 (c) shows the background-only model total probability of Q_{D0} being rejected from the acceptance region for each value of μ (after applying the CL_S method). The location where this probability is $\sim$0.977 occurs around $\mu' = 3$ and is usually called the "$+2\,\sigma$ upper limit" in particle physics (although this is not an expected value, but rather a quantile). This means that $\sim$97.7% of the outcomes (by probability using the background-only model) would not be in the acceptance region for this value of μ. At $\mu'\sim$2.1, the probability is $\sim$0.841 giving the $+1\,\sigma$ upper limit. The median upper limit occurs at $\mu\sim$1.5, the $-1\,\sigma$ upper limit at $\mu\sim$1.2, and the $-2\,\sigma$ upper limit at $\mu\sim$0.9. In this manner, the full Neyman construction may also be used to provide the expected upper limits often provided in particle physics.

[39] In theory, the CL_S method may be applied to any test statistic, but disconnected confidence intervals arise when applied to a test statistic that produces two-sided confidence intervals such as with Q_{FC}.

The values of the expected limits shown in figure 6.17 (c) were double-checked against those computed using the expected limits setting of the OpTHyLiC software package using 100,000 Monte Carlo simulations,[40] and were in good agreement. As a further check, figure 6.17 (d) shows the confidence intervals computed using the observed limit setting of OpTHyLiC. This was constructed by running OpTHyLiC for each of the 431 enumerated possible outcomes along the x-axis (with 10,000-100,000 Monte Carlo simulations each), and plotting each result on the y-axis. When compared to the result of the full Neyman construction shown in (b), the results are nearly identical.

Lastly, figures 6.17 (e) and (f) show the test statistic acceptance ventiles for Q_{D0} without the CL_S method and with it, respectively. The ventiles of figure (f) are determined by the CL_S critical values Q^*_{cls} for α_{cls} ranging from 0 to 1 in steps of 0.05. The insets of these figures demonstrate how the acceptance regions are widened to the left for lower values of μ.

[40] *OpTHyLiC: an Optimized Tool for Hybrid Limits Computation, v.2.00* [244].

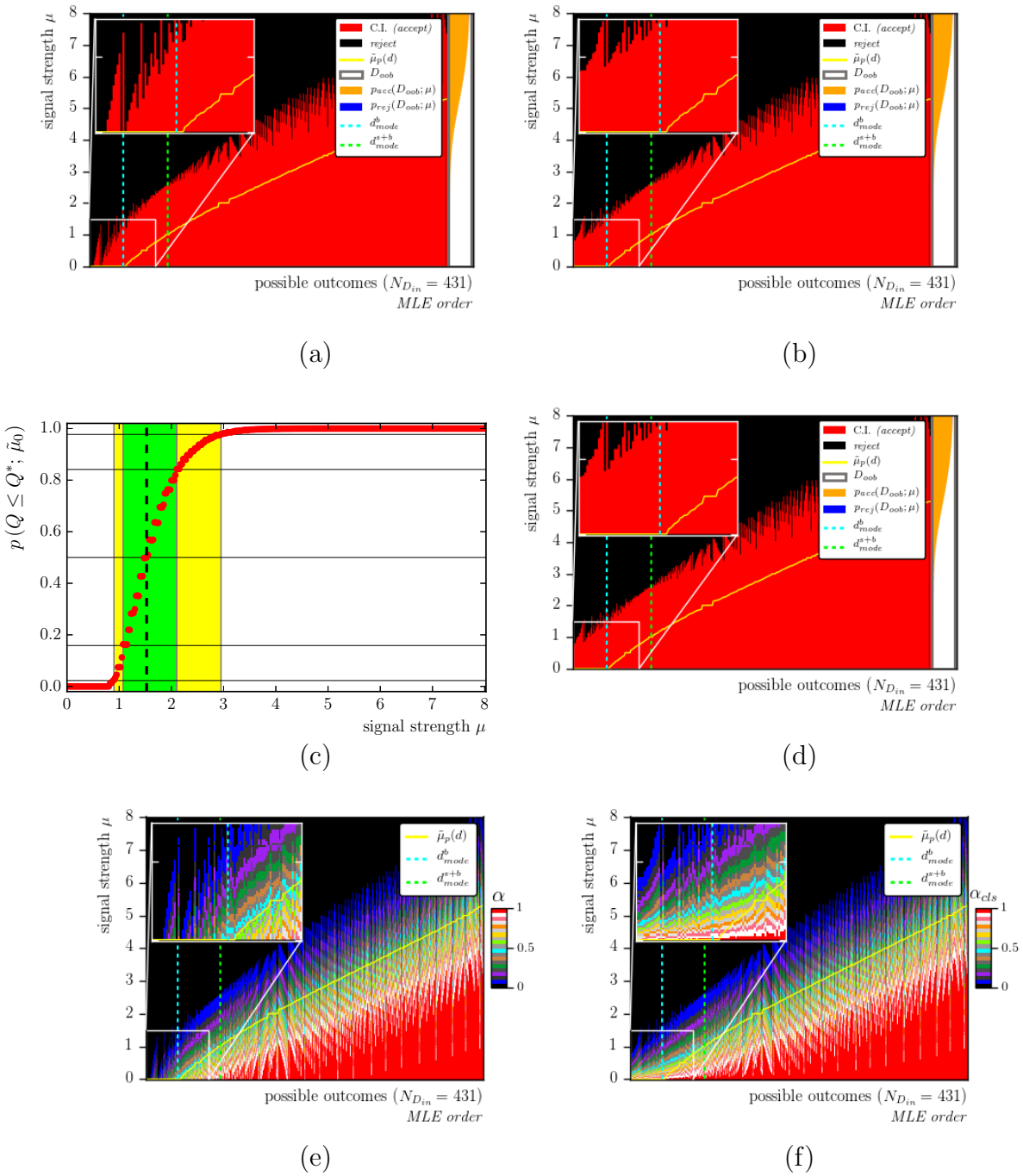


Figure 6.17: The Neyman construction for example 2 using the D0 test statistic, $Q_{D0}(d, \mu)$: (a) 95% confidence intervals without the CL_S method, (b) 95% confidence intervals with the CL_S method, (c) expected upper limits with the CL_S method, (d) 95% confidence intervals with the CL_S method using OpThyLiC [244], (e) acceptance ventiles without the CL_S method, and (f) with the CL_S method.

6.3. The Arizona Test Statistics

The Arizona test statistics are used to address possible model misspecification and to improve the properties of the resultant confidence intervals. The Arizona test statistics apply two multiplicative factors, Q_{inc} and Q_{rep} (defined in section 6.3.2), to the three common test statistics, Q_{FC}, Q_{D0} and Q_{AT}, yielding the three Arizona test statistics

$$\begin{aligned} Q_{AZ1} &= Q_{rep} \cdot Q_{inc} \cdot Q_{FC}, \\ Q_{AZ2} &= Q_{rep} \cdot Q_{inc} \cdot Q_{D0}, \text{ and} \\ Q_{AZ3} &= Q_{rep} \cdot Q_{inc} \cdot Q_{AT}. \end{aligned}$$

Q_{inc} is a function that increases the length of the confidence intervals for outcomes that are inconsistent with both the signal model and the background-only models. The increased confidence interval length for such outcomes reflects uncertainty due to possible model inconsistency. Q_{rep} is a function that decreases the expected distance of μ_{true} to the confidence interval when μ_{true} is not contained within the confidence interval. This reduces the risk that the reported confidence intervals may not be near μ_{true} (in addition to the standard accept vs. reject interpretation of confidence intervals).

This section proceeds as follows.

- Section 6.3.1 demonstrates a motivating example.
- Section 6.3.2 defines and examines Q_{inc} and Q_{rep},
- Section 6.3.3 examines the Neyman construction for the Arizona test statistics.
- Section 6.3.4 measures properties of the resultant confidence intervals.
- Section 6.3.5 shows the results of three additional examples.

A note on terminology. The term *local ordering rule* was defined in section 6.2.3 as the function $Q(d, \mu)$ used for ordering outcomes in the Neyman confidence region construction (NCRC) when determining the acceptance region for each value of μ. The term *test statistic* was defined in section 6.2.3.1 as the function $Q(d, \mu)$ used for ordering outcomes in Neyman-Pearson hypothesis testing (NPHT) when determining the critical/rejection region of the null hypothesis (the outcomes which for which the alternative hypothesis, H_A, is accepted and the null, H_N, is rejected). The term test statistic will now be used to refer to $Q(d, \mu)$ for both NPHT and NCRC as a matter of convenience.[41]

[41] The test statistic of NPHT evaluates the *comparative* evidence provided by an outcome for H_N vs. H_A. A local ordering rule in NCRC evaluates the *magnitude* of evidence provided by an outcome for supporting $H_{\mu'}$, the signal model for a given value of μ'. One may choose to evaluate the support for $H_{\mu'}$ in NCRC by comparing the relative support for some alternative model hypothesis, $H_{\mu''}$. (When a likelihood ratio is used as the local ordering rule, often μ'' is the maximum of a likelihood function used for the alternative of the comparison.) By using a comparison, one is making the claim that the outcomes providing the strongest evidence for $H_{\mu'}$ are those that favor μ' and simultaneously disfavor μ'' (and vice versa). The common test statistics, Q_{FC}, Q_{D0} and Q_{AT}, used as local ordering rules in NCRC are still associated with a test in the Neyman-Pearson sense (in NPHT). Because the Arizona test statistics modify the common test statistics, they are no longer associated with a proper Neyman-Pearson test. However, the Arizona test statistics retain the aspect of ordering outcomes based on comparative evidence so in this sense are "test statistics".

6.3.1. Motivation for the Arizona Test Statistics

So far this chapter has presented a method for visualizing how test statistics may be used in the NCRC used for a signal search. This included a description of the multivariate Poisson model used to make inferences about signal existence, a method for determining the most important finite subset of possible discrete outcomes, and the global MLE ordering of those outcomes that enable useful figures to be made (especially when using test statistics based on likelihood ratios). With this machinery in place, new methods can be described, namely the Arizona test statistics and the confidence-acceptance profiles (defined and discussed in section 6.4).

The NCRC is extremely flexible because it gives the experimenter the subjective freedom to choose any test statistic. The actual real-world observation may surprise the experimenter, but none of the confidence regions of any possible outcome should be surprising since the experimenter predetermines these through the choice of test statistic. Any confidence regions that seems improper to the experimenter is cause for reexamination of the test statistic being implemented.

Generally, test statistic values should be large for outcomes that strongly support the null hypothesis, and small for outcomes that strongly support the alternative hypothesis, but the design of the test statistic may depend on other criteria as well. One must consider what test statistic values to assign to possible outcomes that do not support *any* of the available models, so that an additional criteria for test statistic design is the treatment of outcomes that are inconsistent with all models.

These test statistic design criteria are depicted in figure 6.18. Using these criteria, the desired test statistic, Q, would order outcomes so the most signal-like are accepted first (high Q), and the most like alternative hypothesis are rejected first (low Q). Outcomes that end up with intermediate Q are either consistent with both H_N and H_A (poor discrimination), or else inconsistent with either (possible model misspecification). At the 95% confidence level, outcomes with intermediate Q are likely to be in the acceptance region. Therefore, if μ' is contained in a 95% confidence interval, it is either because μ' is likely to be μ_{true}, or else because the outcome does not provide enough evidence to claim that μ' is unlikely to be μ_{true}.

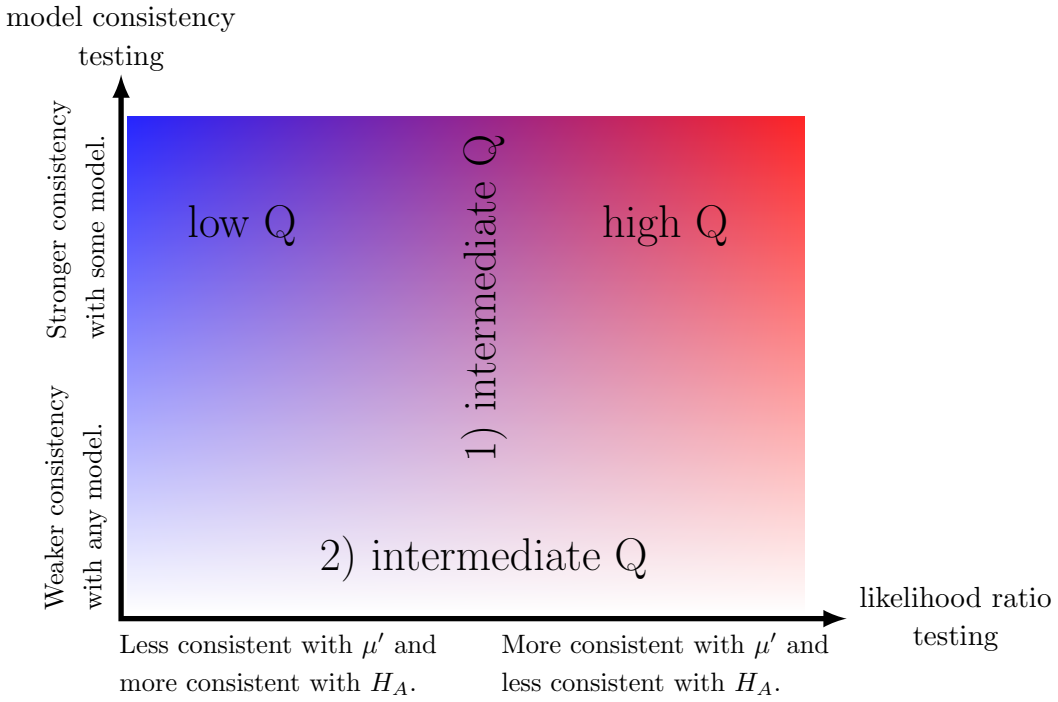


Figure 6.18: One possibility for desirable test statistic criteria to use in the Neyman construction. The test statistic should return high values for outcomes if they are most consistent with null hypothesis models, $H_N : \mu = \mu'$, low values if most consistent with alternative hypothesis models, H_A, and intermediate values if either 1) the outcomes are consistent with both H_N and H_A or else 2) the outcomes are inconsistent with both H_N and H_A.

Recall that example 2 was a bivariate Poisson model (2 bins) with expected background of $[4, 3]$ (counts per bin), and expected signal of $[1, 3]$. Ignoring the enumeration method of the previous section, examine the 132 possible outcomes $[d_1, d_2]$, where $d_1 \in [0, 10]$ and $d_2 \in [0, 11]$.[42] Figure 6.19 investigates some possible experimenter's beliefs regarding how the 95% confidence intervals should appear for each of these outcomes. Note, the top x-axis delineates all 132 of the outcomes, plotted using the global MLE order, and the bottom x-axis denotes the three notable possible outcomes, $d = [0, 0]$, $d^b_{mode} = [4, 3]$, and $d^{s+b}_{mode} = [5, 6]$.

First, a red line is drawn to represent the strength of the experimenter's subjective belief that the signal exists based on each possible outcome. By design, this belief function follows the physical MLEs for each outcome, $\tilde{\mu}_p$, shown with red crosses. In other words, an outcome for which the likelihood function has a larger MLE, $\tilde{\mu}_p$, is believed to provide stronger evidence for signal existence.[43]

Second, a blue line is drawn to represent the strength of the experimenter's subjective belief that an outcome is inconsistent with both the signal model for any value of μ and the background-only model. In other words, this belief function measures the experimenter's suspicion that the real-world experiment is not correctly modeled based on observing a possible outcome. For instance, the possible outcome $[0, 0]$ is inconsistent with the signal model for any $\mu > 0$. However, it is also inconsistent with the background-only model that predicts $[4, 3]$. Therefore, if this outcome were observed, then it would raise suspicion that the multivariate Poisson model with the given expected background and signal yields is inconsistent in modeling the real-world experiment.

The "intuition" that the upper limit for the outcome $d = 0$ should *increase* with increasing expected background yield, b, is discussed in [290]. The inconsistency belief function (blue line of figure 6.19) is meant to model this intuition (this subjective belief of the experimenter). The conclusion of [290] is that the experimenter's intuition is incorrect, and that from the likelihood principle (Bayesian inference), the upper limit for $d = 0$ should be independent of the expected number of background events, b. However, expecting 3 background events and measuring zero is much more plausible than expecting 100 and measuring zero. In the second case, one should consider that the statistical model (that of expecting 100 background events) is inconsistent with the real-world experiment.[44]

In figure 6.19, the blue crosses show $\log[p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)]$, the logarithm of the ratio of the likelihood of the unrestricted MLE, $\tilde{\mu}_u$, to the likelihood of the physical MLE, $\tilde{\mu}_p$, for each possible outcome. For outcomes to the right of d^b_{mode}, this value is zero since $\tilde{\mu}_u = \tilde{\mu}_p$. This ratio grows when moving to the left of d^b_{mode} as the outcomes become less consistent with the background-only model, and the outcome $[0, 0]$ has the highest value.

The maximum, $\tilde{\mu}_p$, of the likelihood function, $p(d; \mu)$, for outcome d is used as an objective proxy to quantify the subjective belief that d provides evidence for values of μ near $\tilde{\mu}_p$ and against values far from $\tilde{\mu}_p$. Similarly, the ratio $p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)$ may be used as an objective proxy to quantify the subjective belief in model inconsistency. This motivates the definition of the model inconsistency function, $Q_{inc}(d) \equiv p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)$, that is used

[42] These ranges were chosen merely for convenience in making informative figures.

[43] Aside from the Neyman-Pearson lemma that proves that likelihood ratios give test statistics with the greatest discrimination power, this subjective belief that the likelihood function provides a measure of evidence for different values of μ gives further motivation for the use of likelihood ratios as test statistics.

[44] In which case the experimenter may wish to report an inference result reflecting this inconsistency.

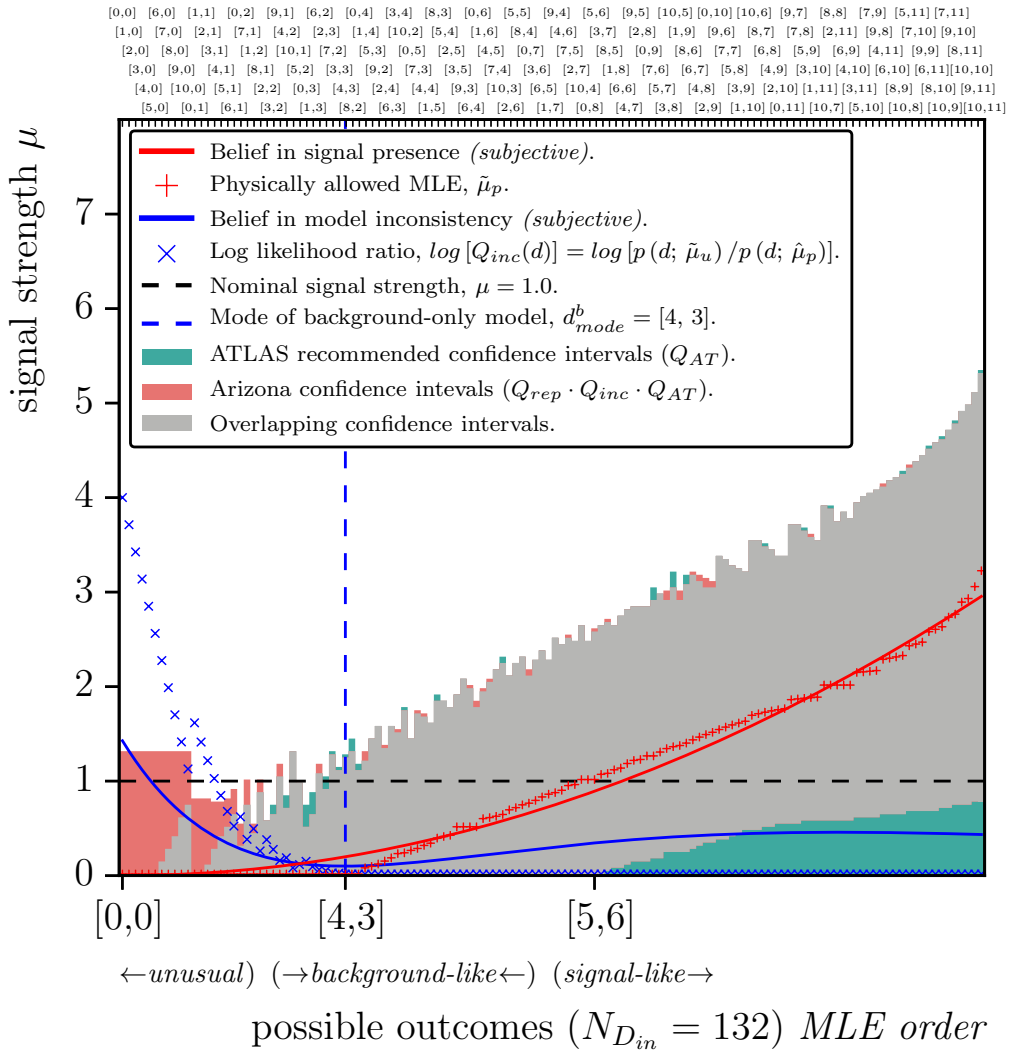


Figure 6.19: Example 2 – Motivation for the Arizona test statistics. The signal strength, μ, is shown on the y-axis, and 132 selected possible outcomes on the x-axis. The experimenter's subjective belief in signal existence given by each outcome (red line) is modeled with the physically allowed MLEs, $\tilde{\mu}_p$ (red crosses). The experimenter's subjective belief in model inconsistency (blue line) is modeled with the inconsistency function, $Q_{inc}(d)$ (blue crosses). The nominal signal strength, $\mu = 1$, is shown with a horizontal dashed black line. Outcomes with upper limits below $\mu = 1$ provide evidence against signal existence. The mode outcome of the background-only distribution, d^b_{mode}, is shown with a vertical dashed blue line. Most outcomes far from d^b_{mode} are inconsistent with the background-only hypothesis. The teal colored area shows the confidence intervals obtained using the ATLAS recommended test statistic, Q_{AT}. The mauve colored area shows those obtained with the modified ATLAS test statistic, Q_{AZ3} (with the gray area showing where these coincide). Confidence intervals obtained using Q_{AT} exclude signal existence furthest to the left where model inconsistency is believed to be high. (Please see footnotes 45 and 46.)

to modify the commonly used test statistics.

One can make the following two pragmatic claims regarding statistical inference.

1. The confidence interval for a given possible outcome is meant to describe the evidence given by an outcome that supports some values of μ (within the confidence interval), and not supporting others (outside the confidence interval). To avoid falsely rejecting signal existence, it may be useful for the length of the confidence interval to represent the level of uncertainty regarding the evidence provided by an outcome, longer confidence intervals for higher uncertainty. This is how many scientists *interpret* confidence intervals, with short confidence intervals often interpreted as reflecting a high level of precision in the measurement. With some modifications to the test statistics commonly used as local ordering rules, the Neyman construction is flexible enough to provide confidence intervals with lengths that model this uncertainty. The inconsistency belief function (blue line of figure 6.19) is used to demonstrate how this may be accomplished.
2. The experimenter must understand how any possible real-world outcome should be interpreted, and design the statistical inference procedure accordingly. Given any two possible outcomes, $d_a = [d_{a,1}, d_{a,2}]$ and $d_b = [d_{b,1}, d_{b,2}]$, the experimenter should be able to state which provides stronger evidence for signal existence. This means that the experimenter should be able to state which of the two outcomes should provide a confidence interval with a higher upper limit, which should provide a lower limit, and possibly, which should have a longer confidence interval. Again, the Neyman construction is flexible enough to allow the experimenter's best judgments to be included.[45]

Decisions need to be made regarding signal existence based on real-world outcomes. In the case of particle physics, if the outcome indicates that the signal does not exist, then resources can be devoted to other tasks. If the outcome indicates that the signal is likely to exist, even if the evidence for existence is not definitive, then experimental resources may

[45]It cannot be emphasized enough that the selection of *any* local ordering rule is a subjective decision based on the experimenter's beliefs. The use of a likelihood ratio implies the belief that the likelihood function provides an "evidence function" for evaluating the different values of the parameter of interest (given a particular possible outcome). This use of "vertical thinking" was not the original intention of Neyman (explained next). Let's compare two philosophies of implementing the Neyman construction by viewing figure 6.4 (a), the 95% confidence intervals using a Garwood-like method, and figure 6.12 (a), using Q_{D0} (pages 272 and 288). Both show the parameter of interest, μ, on the y-axis, and the possible outcomes, d, on the x-axis. For Garwood, the horizontal acceptance regions for a particular μ' were determined by *only* examining the horizontal row for μ'. "Evidence" used to determine the acceptance or rejection of each d' was determined only using $p(d'; \mu')$, i.e., solely horizontal thinking was used to construct the acceptance regions, and once this procedure was completed for all values of μ, the final interpretation was made by tilting one's head 90°, and viewing the result vertically (i.e., viewing the vertical confidence intervals). For Q_{D0}, evidence for whether or not d' was included in the acceptance region of μ' was determined with vertical thinking, i.e., comparing $p(d'; \mu')$ to $p(d'; \mu'')$ for a different value μ'' (Q_{D0} compares to $\mu'' = 0$, Q_{FC} to all values $0 \leq \mu'' < \infty$, and Q_{AT} to all values $0 \leq \mu'' \leq \mu'$). One could claim that Garwood is closer in spirit to a frequentist philosophy because it assumes μ' is the true value, and bases all evidence solely on $p(d; \mu')$, while $Q_{D0,FC,AT}$ attempt to align with the likelihood principle by using the shape of the likelihood function *during* construction. Both provide equally valid confidence intervals with 95% coverage, but each is based on a radically different philosophy. The Arizona test statistics also use vertical thinking during construction, they simply look to evaluate more vertical information than just the likelihood function.

remain in place to continue the search as new data is obtained. If the observed outcome, d_{obs}, is ambiguous, the statistical inference result should reflect this uncertainty.

A rational experimenter should desire confidence intervals with upper limits that appear something like the sum of the red and blue lines in figure 6.19. If the outcome d aligns with a stronger belief in signal existence, then the confidence intervals should include larger values of μ, and possibly not include smaller values. If the outcome suggests model inconsistency, then the confidence interval should contain both small and intermediate values of μ. In other words, the experimenter should require that each confidence interval for each possible outcome reflects the experimenter's best judgment regarding signal existence.

Figure 6.19 also compares the results of the Neyman construction using the ATLAS test statistic, $Q_{AT}(d, \mu') = p(d;\, \mu')/p(d;\, \tilde{\mu}_b)$, shown with the teal colored area, and the Arizona test statistic, $Q_{AZ3}(d, \mu')$, shown with the mauve colored area. The gray area shows the overlapping confidence regions between the two methods, which agree for the majority of confidence regions coordinates, (d, μ).

Both approaches provide intervals of small length for outcomes midway between [0,0] and [4,3] where the outcomes are more consistent with the background-only model than the signal model. The major difference is for outcomes furthest to the left where Q_{AZ3} produces confidence intervals of increasing length while those from Q_{AT} are decreasing. For the outcome $d = [0, 0]$, the Q_{AZ3} confidence interval extends from $\mu = 0$ to slightly past $\mu = 1$ and so this interval reflects the ambiguity of measuring $[0, 0]$. The Q_{AT} confidence interval for $[0, 0]$ is the empty set, a difficult result to interpret or use in decision making.[46]

Another difference is that for outcomes to the right of $d^{s+b}_{mode} = [5, 6]$, the Q_{AZ3} confidence intervals are two-sided, unlike the one-sided Q_{AT} upper limits. This demonstrates that a test statistic producing one-sided confidence intervals may produce two-sided confidence intervals if modified. Aside from the outcomes furthest to the left, the upper limits for both approaches are nearly always identical, and so either could be used to provide upper limits if only those were needed in decision making.[47] Both test statistics are equally valid from a frequentist perspective with both providing exact 95% coverage.[48]

It is useful to view the Neyman construction in another way. Figure 6.20 shows the

[46] The empty set is interpreted by some as a useful indicator of mismodeling (private communications). Aside from probable flip-flopping that would invalidate the principles of the Neyman construction, the more troublesome aspect is that there are usually very short confidence intervals for outcomes nearby those producing empty sets (essentially, the confidence interval shrinks to zero length (empty set) as one travels leftward along the x-axis of figure 6.19). How does one reconcile a very short confidence interval (that will likely be misinterpreted as a high precision measurement) being adjacent to an outcome producing an empty set? Is the transition from accurate modeling to mismodeling so discrete? (Almost certainly not.) Then should all shorter confidence intervals be viewed as indicating greater mismodeling? (Probably not, unless these outcomes are nearby empty set outcomes.) The solution proposed here is to use the amazing flexibility of the Neyman construction to capture the likelihood of mismodeling within the local ordering rule.

[47] How can the bulk of confidence intervals be nearly identical for the central outcomes of figure 6.19? This is because the outcomes that are switched from rejected to accepted (and vice versa) have *extremely* tiny probability masses, and so do not much affect the confidence intervals for outcomes of much higher probability masses (see figure 6.5 (b) on page 273). So why bother switching the acceptance status of these outcomes if they have such tiny probability masses? Because they only have tiny probability masses if the modeling is accurate, and any unexpected surprises may be factored into the Neyman construction *a priori*.

[48] Exact coverage except for the the usual caveats of discrete statistical models.

outcomes of the bivariate Poisson model of example 2, with the possible outcome count of the first bin on the x-axis and the possible outcome count of the second bin on the y-axis (so that each cell represents a unique possible outcome). The blue arrow represents the expected background yields, $d^b_{mode} = [4, 3]$, and the red arrow the expected signal yields, $[1, 3]$, so that their sum is $d^{s+b}_{mode} = [5, 6]$.[49] Two probabilities are listed in each cell. The top probability is the signal model probability *at the nominal signal strength*, $p(d; \mu = 1.0)$, (this is because figure 6.20 is demonstrating the construction of the acceptance region for the specific value $\mu = 1.0$). The bottom probability is the background-only model probability, $p(d; \tilde{\mu}_0)$. Outcomes more consistent with the signal model are shown in red, and those more consistent with the background-only model in blue. Outcomes that are consistent with both the signal model and the background-only model are shown in violet, and those outcomes that are inconsistent with either model are shown in white.[50]

During the Neyman construction, the $\mu = 1.0$ acceptance region at 95% confidence level is the set of outcomes that have a sum of signal model probabilities of 0.95. One would expect the $\mu = 1.0$ acceptance region to contain outcomes near the red arrow head, and reject those either far from the red arrow head or else close to the blue arrow head. Figure 6.20 is useful for examining which outcomes are accepted or rejected from the $\mu = 1.0$ acceptance region using different test statistics.

Outcomes with upper limits below $\mu = 1.0$ imply signal nonexistence at the nominal signal strength (the nominal signal strength corresponds to the expected signal yields predicted by the signal theory). Outcomes with upper limits below $\mu = 1.0$ are not contained in the $\mu = 1.0$ acceptance region. The acceptance region for $\mu = 1.0$ is thus useful to examine because the confidence intervals for those outcomes not contained in the $\mu = 1.0$ acceptance region imply signal nonexistence.

The Neyman construction of the $\mu = 1.0$ acceptance region is computed using the three commonly used test statistics, Q_{FC}, Q_{D0} and Q_{AT}, and the three Arizona test statistics, Q_{AZ1}, Q_{AZ2}, and Q_{AZ3} (that will be defined more explicitly in section 6.3). The $\mathrm{CL_S}$ method is often used with Q_{D0} and Q_{AT} in particle searches, and so $\mu = 1.0$ acceptance regions for these two methods are also constructed. These eight different methods for determining the $\mu = 1.0$ acceptance region are compared in order to determine which outcomes imply signal nonexistence according to each of these methods.

The results are shown in figure 6.21. If an outcome is rejected from the acceptance region for a particular method, that method is printed in the cell corresponding to that rejected outcome. For example, the bottom left cell of $d = [0, 0]$ lists "FC", "D0", and "AT", indicating that the acceptance regions constructed using Q_{FC}, Q_{D0}, and Q_{AT} exclude $[0, 0]$. This cell also lists "D0*" and "AT*", indicating that the acceptance regions for Q_{D0}, and Q_{AT} implementing the $\mathrm{CL_S}$ method also exclude $[0, 0]$.

The majority of outcomes that are signal-like (red cells) are accepted by all methods (no

[49] Figure 6.20 is made to show the construction of the acceptance region for $\mu = 1.0$. For a different value μ', the length of the red arrow would be adjusted to $[1 \cdot \mu', 3 \cdot \mu']$, so that the concepts of the following discussion will still apply. Note the utility of this visualization: outcomes near the head of the blue arrow are the most background-like, while those near the head of the red arrow are the most signal-like at strength μ' (or rather, the most background plus signal-like). Outcomes between the two arrow heads are consistent with both models, and those far from both arrow heads are inconsistent with either model.

[50] This colorization is only qualitative, and no color map is provided.

text written in the cells). Meanwhile, most methods reject the background-like outcomes that are distant from the signal-like outcomes (blue cells in lower left), while accepting the background-like outcomes (blue) that border the signal-like (red). Only the outcome cells furthest to the lower left show rejection by the two CL_S methods (showing the text "D0*" or "AT*"). This highlights one possible difficulty with the CL_S method. It extends the acceptance region leftward and downward so that the most background-like outcomes are accepted, while rejecting those outcomes on the extreme bottom left where the experimenter is less certain of model consistency.

The largest difference between the common methods and the Arizona test statistics can be seen in the first seven cells along the bottom row where $d_2 = 0$ and $d_1 < 7$. These show that all five common methods reject these outcomes from the acceptance region of $\mu = 1$ while none of the Arizona test statistics reject these outcomes. The MLE global order placed these outcomes farthest to the left of figure 6.19, which shows that $\mu = 1$ is in the confidence intervals for these outcomes when using the Arizona test statistics. These outcomes are included in the $\mu = 1$ acceptance region because the inconsistency function $Q_{inc}(d) = p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)$ was used as to indicate model inconsistency, and Q_{inc} returns the highest values for those outcomes.

This example demonstrates that the Arizona test statistics may be used to provide confidence intervals that are more aligned with the experimenter's beliefs, with longer confidence intervals for outcomes furthest to the left in figure 6.19 capturing the experimenter's uncertainty regarding modeling consistency for these outcomes.

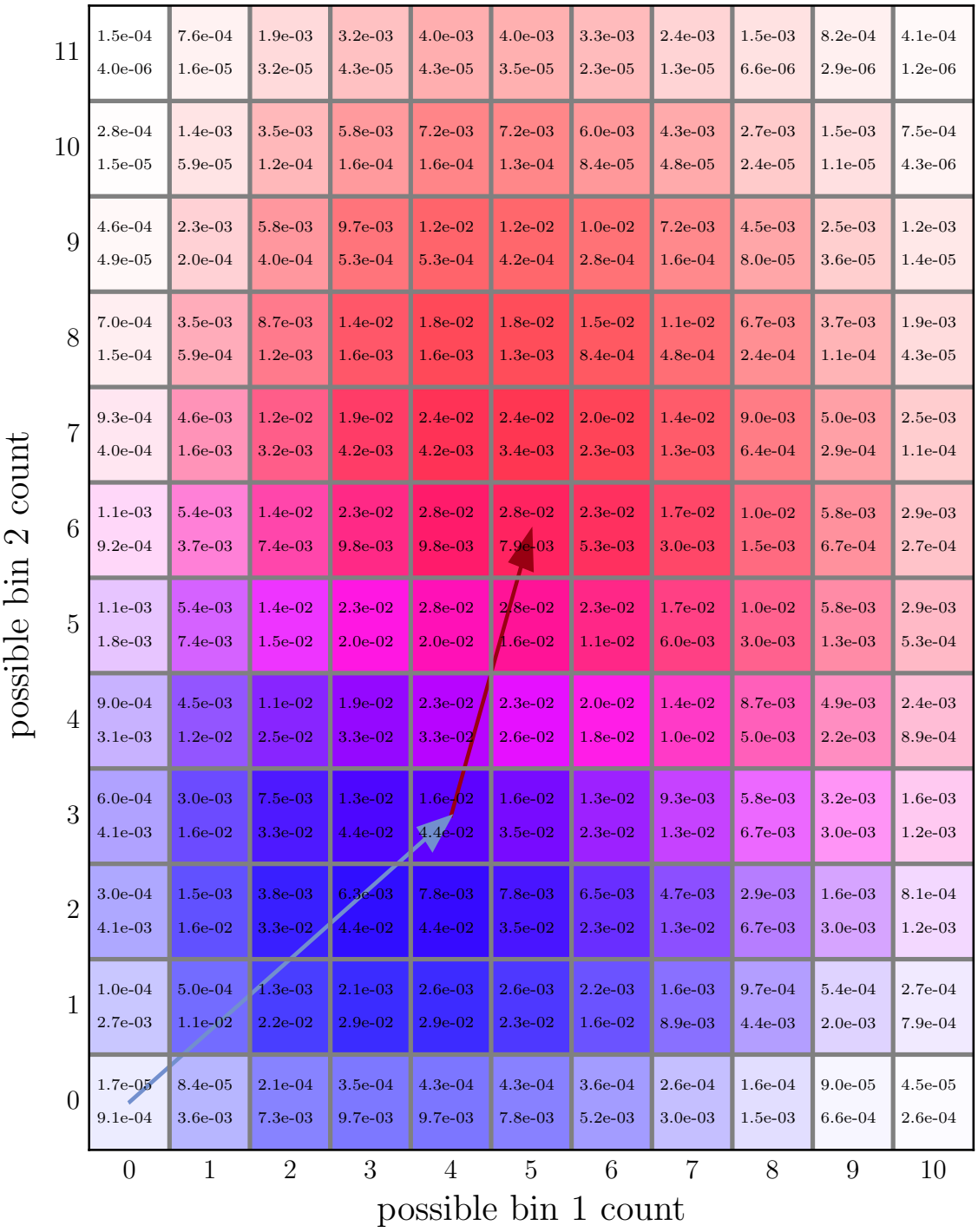


Figure 6.20: Example 2 – Model probabilities for $\mu = 1$. The blue arrow represents $d^b_{mode} = [4, 3]$ and the red arrow $[1, 3]$ so that their sum is $d^{s+b}_{mode} = [5, 6]$. The top number of each cell is the signal model probability, $p(d; \mu = 1)$, and the bottom number is the background-only model probability, $p(d; \tilde{\mu}_0)$. Red cells indicate signal-like outcomes, blue cells background-like, violet when like both, and white when like neither.

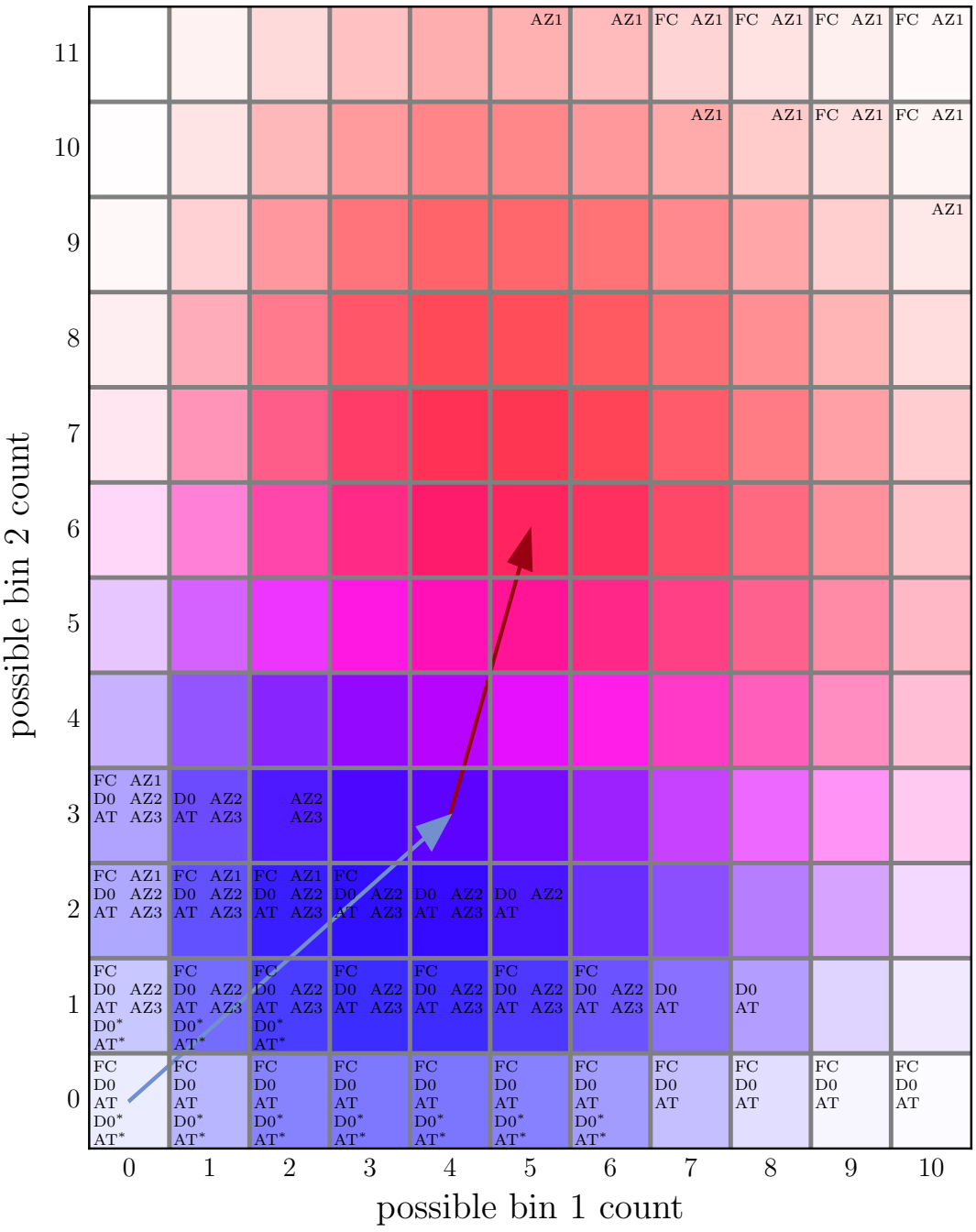


Figure 6.21: Example 2 – Comparing the $\mu = 1.0$ acceptance regions for several methods. If a method rejects an outcome, then the confidence interval for that outcome does not contain $\mu = 1.0$, which implies signal nonexistence. A cell without text represents an outcome that is accepted by all methods. Otherwise, the methods that reject the outcome are printed in its cell: either the common test statistics, Q_{FC}, Q_{D0} or Q_{AT}, or the Arizona test statistics, Q_{AZ1}, Q_{AZ2} or Q_{AZ3}. A star indicates the use of the $\mathrm{CL_S}$ method in conjunction with either Q_{D0} or Q_{AT}.

6.3.2. The Test Statistic Modification Functions, Q_{inc} and Q_{rep}

In the previous section, the inconsistency function defined as the ratio of two likelihoods, $Q_{inc}(d) = p(d; \tilde{\mu}_u)/p(d; \tilde{\mu}_p)$, was selected to quantify the amount of model inconsistency for each outcome. Appendix section D.1.1 provides details as to why this ratio was selected. The important feature of $Q_{inc}(d)$ is that it is at its minimum value of 1.0 for outcomes with likelihood functions having nonnegative MLEs, and grows for outcomes that have likelihood functions with increasingly negative MLEs.[51] An outcome with a positive MLE, $\tilde{\mu}_u > 0$, provides evidence for the signal model with values of μ near $\tilde{\mu}_u$. An outcome with an MLE near zero, $\tilde{\mu}_u \sim 0$, supports the background-only model. A negative MLE does not provide evidence for either model.

The inconsistency function can be incorporated into any test statistic, Q, by simple multiplication, $Q' = Q_{inc} \cdot Q$.[52] However, this results in some undesirable features of the resultant confidence intervals when used with test statistics Q that would otherwise produce one-sided upper limits. For example, figure 6.22 shows the result of using $Q_{inc} \cdot Q_{D0}$ as the test statistic. For small values of μ, the inclusion of Q_{inc} causes $Q_{inc} \cdot Q_{D0}$ to have the lowest values for outcomes near d^b_{mode} (so that these outcomes are the most strongly rejected from the acceptance regions of these small μ). This creates a "hole" in the confidence intervals at small values of μ for outcomes near d^b_{mode} that may not accurately represent the experimenter's beliefs.

It may be more appropriate for the acceptance regions of these lower-valued μ to first reject the outcomes furthest to the right. Those outcome furthest to the right provide supporting evidence for large values of μ and not small μ. Therefore, multiplying Q_{inc} to a test statistic is only sensible if that test statistic produces two-sided confidence intervals. The reporting risk function, Q_{rep}, that is discussed next, facilitates the use of Q_{inc} by yielding two-sided confidence intervals. For example, using $Q_{rep} \cdot Q_{D0}$ as the test statistic generates two-sided confidence intervals so that confidence intervals constructed using $Q_{rep} \cdot Q_{inc} \cdot Q_{D0}$ do not have the hole near d^b_{mode} as shown in figure 6.22.

The use of likelihood ratio test statistics as local ordering rules in the Neyman construction is motivated by 1) the Neyman-Pearson lemma that proves these have the greatest discrimination power, and 2) the belief that the likelihood function is useful for inference.[53] A test statistic with large discrimination power is desirable because it generates shorter, more accurate confidence intervals, where accuracy is defined as confidence intervals that reject incorrect values of μ with higher probability (but not in any way more probable to contain μ_{true} than any other confidence intervals for the same confidence level).

[51] Recall that for a given outcome, d', the maximum of its likelihood function, $p(d'; \mu)$, is the value of μ which maximizes $p(d'; \mu)$. If the MLE, $\tilde{\mu}_u$, is unrestricted, then it is allowed to take negative values. However, negative MLE values are unphysical since μ models the signal strength which must be nonnegative (although there are some particle physics searches where signal presence may lead to expected yields below those predicted by the background-only model).

[52] Harrison Prosper has argued against modifying likelihood ratio test statistics because it allows the experimenter to construct confidence intervals that behave according to the experimenter's wishes (discussion section of [291]). The argument here is that experimenters *should* modify test statistics to achieve confidence regions with the most sensible behavior according to their best judgments (but before unblinding, of course).

[53] The resultant confidence intervals of the Neyman construction often do not satisfy the likelihood principle. Nevertheless, it is often useful to assume that larger values of the likelihood function are indicative of the location of μ_{true}, which is in the spirit of the likelihood principle.

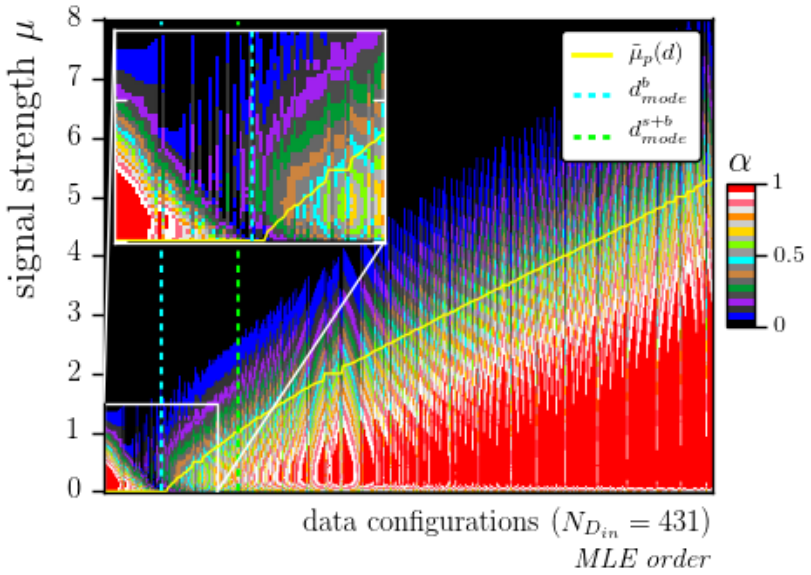


Figure 6.22: Example 2 – The acceptance ventiles using $Q = Q_{inc} \cdot Q_{D0} = \frac{p(d;\tilde{\mu}_u)}{p(d;\tilde{\mu}_p)} \frac{p(d;\mu)}{p(d;\tilde{\mu}_0)}$. There is a "hole" in the confidence intervals at small μ near d^b_{mode} (dashed cyan) that may be undesirable.

One may reframe the selection of the optimal test statistic in terms of risk minimization. A test statistic that is a likelihood ratio may be shown to be that which minimizes a particular risk function, specifically, the power loss risk function, $R^{\alpha}_{pow}(\mu', \mu'')$, defined in equation 6.2.21 on page 281. Here, μ' is the null hypothesis and μ'' the alternative hypothesis. $R^{\alpha}_{pow}(\mu', \mu'')$ is larger when the acceptance region of μ' contains outcomes d' that have a larger total probability using the alternative hypothesis, $p(d'; \mu'')$. In other words, minimizing $R^{\alpha}_{pow}(\mu', \mu'')$ is equivalent to assuming that $\mu' = \mu_{true}$ so that μ'' is incorrect, and finding the acceptance region for μ' that minimizes the amount of the incorrect model distribution (using μ'') contained in the acceptance region.[54]

Is the power loss risk, $R^{\alpha}_{pow}(\mu', \mu'')$, the most sensible risk to minimize in a signal search (and thus is the likelihood ratio the optimal test statistic to use)? Perhaps, but the selection of which risk function to use in determining an optimal test statistic must be recognized as a subjective decision made by the experimenter. A different risk function will be considered: the reporting risk function, $R^{\alpha}_{rep}(\mu', \mu)$, that is defined below. This risk function is motivated by the desire to reduce the distance from μ_{true} to the closest end of the confidence interval (in the case that μ_{true} lies outside the confidence interval). Selecting a test statistic for the Neyman construction that minimizes $R_{rep}(\mu', \mu)$ would allow the experimenter to claim that if the reported confidence interval did not contain μ_{true} (a reporting error), then μ_{true} is still

[54] Even though the concept of maximizing power (minimizing $R^{\alpha}_{pow}(\mu', \mu'')$) originates in NPHT, it is now being applied to the Neyman construction. Thus, this procedure is performed separately in determining the optimal acceptance region for each μ' by minimizing $R^{\alpha}_{pow}(\mu', \mu'')$. Note that μ'' may be the MLE of a likelihood function (giving Q_{FC} as the optimal test statistic), represent the background-only model (giving Q_{D0}), or any sensible alternative model, so that selection of the alternative hypothesis to use is also a choice of the experimenter.

expected to lie close to the confidence interval. A large reporting risk is problematic if one seeks to use confidence intervals to make decisions beyond the simple acceptance/rejection that is implied by the Neyman construction.

Recall that the Neyman-Pearson lemma proves that using likelihood ratios as test statistics results in confidence intervals with the shortest expected length. As such, maximizing discrimination power may lead to large reporting risk. On the other hand, longer confidence intervals would be more likely to contain μ_{true} and thus reduce reporting risk at the expense of discrimination power. There is clearly a tradeoff between these two types of risk.

One may consider the Neyman construction as equivalent to the determination of the rejection indicator function, $\phi^{\alpha}_{rej}(d, \mu)$, that is defined by equation 6.2.19 on page 277. Recall that $\phi^{\alpha}_{rej}(d, \mu)$ simultaneously gives both the acceptance regions (accepted outcomes d given a value of μ) and the confidence intervals (accepted values of μ given an outcome d; see figure 6.9 on page 278). The minimization of risk in the presence of a constraint may be used to determine the optimal rejection indicator function. After the confidence level is set, $(1 - \alpha)$, the Neyman construction requires that the type I risk be set to α so that $R^{\alpha}_{size}(\mu') = \alpha$ for all values of μ'.[55] This provides the constraint that applies when any other risk function is to be minimized, such as R^{α}_{pow} or R^{α}_{rep}.

In a manner similar to $R^{\alpha}_{pow}(\mu', \mu'')$, reporting risk may be defined as

$$R^{\alpha}_{rep}(\mu', \mu) \equiv E_{\mu'}[\delta(\mu', CI(d))] = \sum_{d'} p(d'; \mu') \min_{\mu''} \left| \mu' - \mu'' \left(1 - \phi^{\alpha}_{rej}(d', \mu'')\right) \right|, \tag{6.3.1}$$

where the minimum distance of μ' to the confidence interval $CI(d)$ of outcome d is given as $\min_{\mu''} \left| \mu' - \mu'' \left(1 - \phi^{\alpha}_{rej}(d', \mu'')\right) \right|$, which is 0 if $\phi^{\alpha}_{rej}(d', \mu') = 0$ (μ' in $CI(d)$). Therefore, the reduction of this risk for each μ' implies the reduction of the expected distance from μ' to the confidence interval, and since this applies to each μ', it therefore applies to μ_{true}.

It can be shown that minimizing $R^{\alpha}_{pow}(\mu', \mu'')$ subject to the constraint of $R^{\alpha}_{size}(\mu')$ gives an optimal rejection indicator function that corresponds to the likelihood ratio as the optimal test statistic. However, it is unknown what test statistic would minimize $R^{\alpha}_{rep}(\mu', \mu)$ in conjunction with the type I risk constraint. Therefore, experimentation with various modifications to the common test statistics was attempted, and it was observed that inclusion of a new function resulted in reduced reporting risk. This is the reporting risk modification function,

$$Q_{rep}(d, \mu) = \frac{1}{0.1 + |\mu - \tilde{\mu}_p|}, \tag{6.3.2}$$

where $\tilde{\mu}_p$ is the physical MLE (≥ 0), and 0.1 is a regularization parameter. Appendix section D.1.2 provides a more detailed discussion of the motivation for selecting $Q_{rep}(d, \mu')$.

[55] Recall from equation 6.2.13 on page 262 that $R^{\alpha}_{size}(\mu') = \sum_{d'} p(d'; \mu')\, \phi^{\alpha}_{rej}(d', \mu')$. Type I risk, $R^{\alpha}_{size}(\mu')$, was defined for NPHT, but it is an equivalent constraint for the acceptance region of each μ' to ensure $(1-\alpha)$ confidence in the Neyman construction. The usual caveat applies that due to the discrete Poisson model, $R^{\alpha}_{size}(\mu')$ may be slightly less than α.

6.3.3. Defining the Arizona Test Statistics

Many modifications of the three common test statistics, Q_{FC}, Q_{D0}, and Q_{AT}, were examined, but only three are discussed here to maintain clarity of presentation. These incorporate both the model inconsistency measure function, Q_{inc}, and the reporting risk modification function, Q_{rep}, to provide the following as the principal Arizona test statistics:

$$Q_{AZ1}(d,\mu) \equiv Q_{rep}(d,\mu)\cdot Q_{inc}(d)\cdot Q_{FC}(d,\mu) = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}\cdot\frac{p(d;\,\tilde{\mu}_u)}{p(d;\,\tilde{\mu}_p)}\cdot\frac{p(d;\,\mu)}{p(d;\,\tilde{\mu}_p)}, \quad (6.3.3)$$

$$Q_{AZ2}(d,\mu) \equiv Q_{rep}(d,\mu)\cdot Q_{inc}(d)\cdot Q_{D0}(d,\mu) = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}\cdot\frac{p(d;\,\tilde{\mu}_u)}{p(d;\,\tilde{\mu}_p)}\cdot\frac{p(d;\,\mu)}{p(d;\,\tilde{\mu}_0)}, \quad (6.3.4)$$

$$Q_{AZ3}(d,\mu) \equiv Q_{rep}(d,\mu)\cdot Q_{inc}(d)\cdot Q_{AT}(d,\mu) = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}\cdot\frac{p(d;\,\tilde{\mu}_u)}{p(d;\,\tilde{\mu}_p)}\cdot\frac{p(d;\,\mu)}{p(d;\,\tilde{\mu}_b)}. \quad (6.3.5)$$

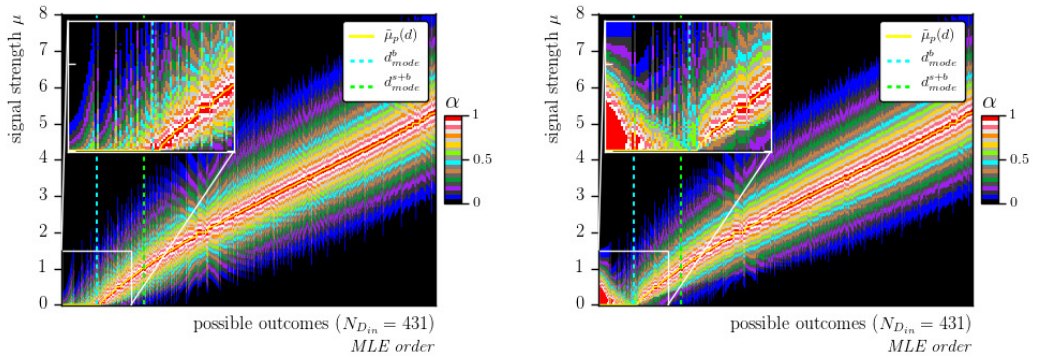
Figures 6.23 (a)–(f) show the acceptance ventiles for example 2 using the Arizona and common test statistics.[56] One difference between the common methods (left side figures) and the Arizona (right side) is the appearance of confidence intervals of increasing length for outcomes proceeding leftward from d^b_{mode}. This results from the inclusion of the model inconsistency measure, Q_{inc}. If the observed outcome were to be one of these outcomes leftward from d^b_{mode}, then the longer confidence intervals produced by the Arizona test statistics would reflect the experimenter's doubt that the experiment was correctly modeled by the multivariate Poisson distribution, i.e., that there was shape mismodeling in which a mistake had been made in the simulations used to compute the expected background and signal yields for each bin of the discriminating variable.

The insets of figures 6.23 (b), (d) and (f) show that the highest values of the acceptance ventiles for these outcomes further to the left occur for small values of μ and decrease as μ increases (starting with red near $\mu = 0$ and progressing upward to blue at larger μ). This may be interpreted as an inference statement indicating that the evidence provided by any of these outcomes more strongly supports smaller values of μ than larger values.[57]

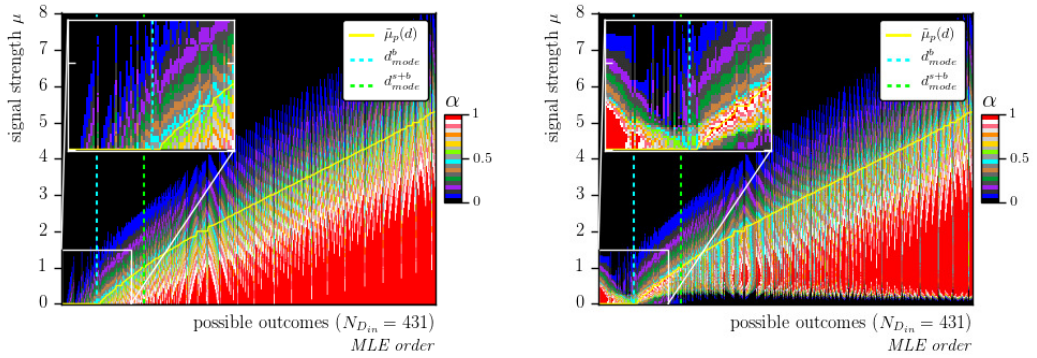
Figures 6.23 (c)–(f) also show that the one-sided confidence intervals of Q_{D0} and Q_{AT} become two-sided for outcomes to the right of d^{s+b}_{mode} when using Q_{AZ2} and Q_{AZ3}. A one-sided

[56] Recall that example 2 is a bivariate Poisson model (2 bins) with expected background of $[4, 3]$ (counts per bin), and expected signal of $[1, 3]$. The multivariate Poisson model is $p(d;\,\mu) = \prod_i (\mu s_i + b_i)^{d_i}/d_i! \cdot e^{\mu s_i + b_i}$ where b_i is the expected background yield in the i^{th} bin of the discriminating variable, s_i is the expected signal yield at nominal signal strength ($\mu = 1.0$), d_i is the outcome count (a whole number), and the signal strength parameter μ is the parameter of interest to be inferred.

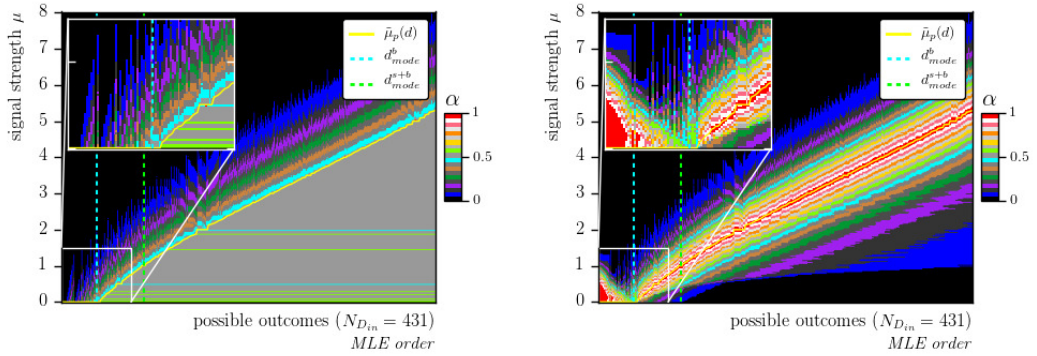
[57] Section 6.4.7 discusses the proper frequentist interpretation for comparing the evidence provided by an outcome for different values of μ using their differing acceptance ventiles. This is done by constructing significance ratio functions that provide a frequentist method for evaluating the strength of evidence provided by an outcome for each *separate* value of μ (i.e., a frequentist method for comparing the evidence for different values of μ beyond the simple accept/reject interpretation of a confidence interval). However, section 6.4.7 shows that it is *qualitatively* true that there is stronger supporting evidence for a value of μ when that value of μ has a higher acceptance ventile (although it is always preferable to construct the significance ratio functions to obtain a *quantitative* measure of evidence for each value of μ).



(a) Q_{FC} (b) Q_{AZ1}



(c) Q_{D0} (d) Q_{AZ2}



(e) Q_{AT} (f) Q_{AZ3}

Figure 6.23: Example 2 – Comparison of the acceptance ventiles of the Neyman construction using the common test statistics: (a) Q_{FC}, (c) Q_{D0}, (e) Q_{AT}, and the Arizona test statistics: (b) Q_{AZ1}, (d) Q_{AZ2}, (f) Q_{AZ3}. The Arizona test statistics produce longer confidence intervals for outcomes that may indicate modeling inconsistency (those furthest to the left), and also always produce two-sided confidence intervals (unlike Q_{D0} and Q_{AT}).

confidence interval for the signal strength μ is useful because it is simple to report and interpret; it extends from $\mu = 0$ to the value of the upper limit, with no lower limit defined. However, Feldman and Cousins assert that most scientists would adjust their inference methodology if the observed outcome gave evidence of signal existence (i.e., appeared further to the right of d_{mode}^{s+b}) and would instead report two-sided confidence intervals [272]. This *possibility* of "flip-flopping" invalidates the coverage properties of the Neyman construction *even if* the actual observed outcome did not give evidence of signal existence. The upper limits of Q_{AZ2} and Q_{AZ3} are quite similar to those of Q_{D0} and Q_{AT} (except for the outcomes furthest to the left that are adjusted for model inconsistency) while avoiding the danger of flip-flopping by providing two-sided limits for outcomes that support signal existence with moderate signal strength, μ.

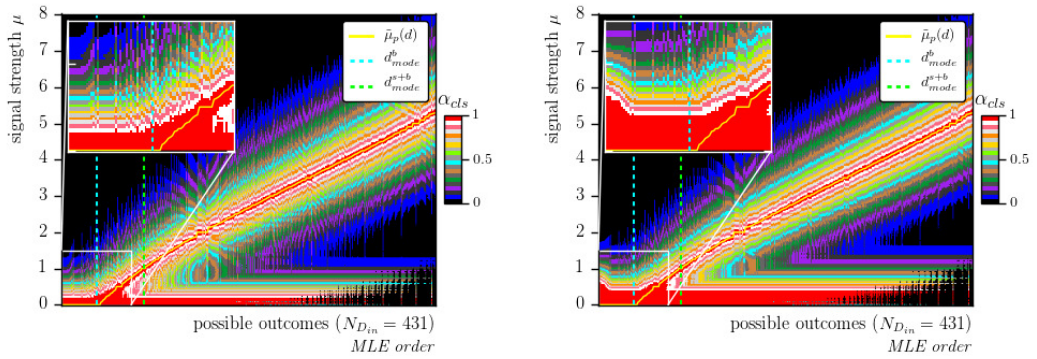
The structure of the acceptance ventiles for Q_{AZ2} appears similar to that of Q_{D0}, while that of Q_{AZ3} appears quite different from Q_{AT}. This is due to the definition of Q_{AT} where for many outcomes, many values of μ reach the maximum value of $Q_{AT}(d, \mu) = 1.0$. This results in the large gray area shown in figure 6.23 (e) where for a given value of μ (horizontal row), many outcomes are "tied" at the maximum so that they are all in the highest quantile together. The inclusion of the Q_{rep} and Q_{inc} terms with Q_{AT} (giving Q_{AZ3}) breaks the tie between these outcomes, and forces an ordering between them so that they are no longer tied at the maximum value. This is useful if one seeks to obtain comparative evidence for differing values of μ as is done in section 6.4.7.

Figures 6.24 (a)–(f) show how these results are modified when the CL_S method is implemented. The CL_S method purposefully overcovers acceptance regions for small values of μ where discrimination between the signal and background-only model is weak, and this is demonstrated by the structures of the resultant acceptance ventiles with higher values of the acceptance ventiles appearing for small values of μ.

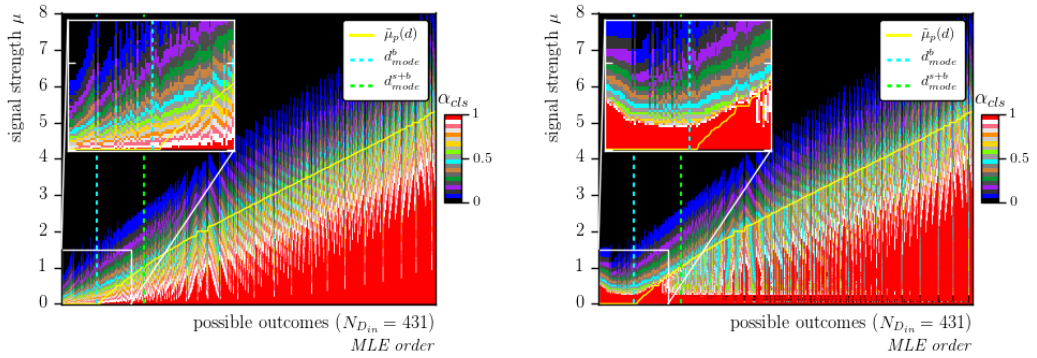
The CL_S method is typically applied to test statistics that produce one-sided upper limits, namely when using Q_{D0} and Q_{AT}. When the CL_S method is applied to test statistics that produce two-sided intervals, the overcoverage at low μ often results in disjoint confidence regions, where the confidence regions for some of the possible outcomes (vertical columns to the right of d_{mode}^{s+b}) are unions of two disjoint intervals. This can be seen in figures 6.24 (a), (b), (d) and (f), and is most pronounced for Q_{FC} and Q_{AZ1} (figures (a) and (b)).

The disjoint confidence regions produced by the CL_S method for some test statistics may be undesirable. Nevertheless, these disjoint confidence regions have a reasonable interpretation: the upper confidence interval for larger values of μ shows the most likely values for μ_{true}, if the signal were to exist with larger signal strength, while the lower confidence interval for smaller values of μ indicates that the lack of signal-background discrimination power prevents these smaller values from being ruled out as candidates for μ_{true}. This interpretation is in line with the ultimate goal of the CL_S method, to prevent false exclusion of values of μ from the confidence region when exclusion is unjustified.[58]

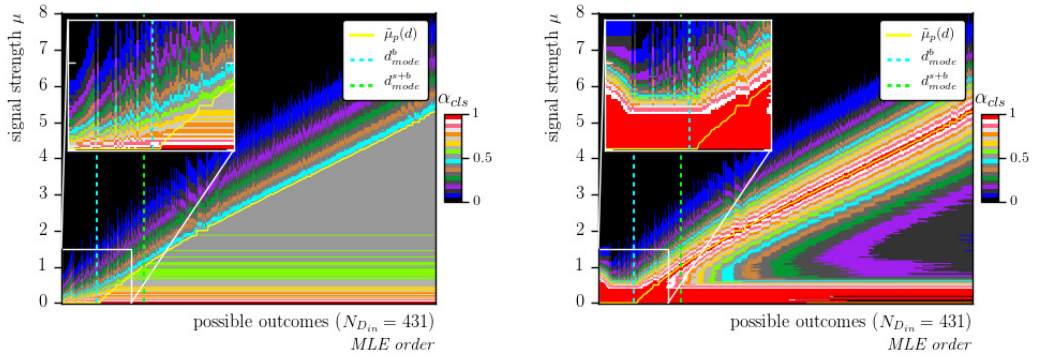
[58] The theoretical justification for the CL_S method has been debated [288]. However, section 6.4.6 shows that the signal significance ratio function automatically provides the CL_S confidence interval at any desired CL_S confidence level. The signal significance ratio function is a measure of evidence against signal existence for each *separate* value of μ (i.e., what values of μ are considered "ruled out" by the evidence provided by the observed outcome). Therefore, the CL_S method may be justified by the evidence-based approach discussed in section 6.4.6.



(a) Q_{FC}+CL$_S$ (b) Q_{AZ1}+CL$_S$



(c) Q_{D0}+CL$_S$ (d) Q_{AZ2}+CL$_S$



(e) Q_{AT}+CL$_S$ (f) Q_{AZ3}+CL$_S$

Figure 6.24: Example 2 – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics implementing the CL$_S$ method: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}. Note the appearance of some disconnected confidence regions in (a), (b), (d), and (e) (where the vertical columns for outcomes further to the right show that the highest quantiles are reached for two different values of μ, one close to $\mu = 0$ and one for larger μ).

6.3.4. Performance Metrics of the Neyman Construction

The experimenter cannot obtain arbitrary confidence regions for a given confidence level since the correct coverage probabilities must be maintained. But the experimenter does have control over the ordering of outcomes in determining acceptance regions. In principle, the experimenter could determine *by hand* the local ordering of the outcomes according to their best judgment and views of appropriate criteria. For example, consider how one should order the 132 outcomes of figure 6.20 (page 307), from the outcomes providing the least evidence for $\mu = 1$ to those providing the greatest. After ordering by hand according to their best judgment, the Neyman construction might perhaps deliver confidence regions most in line with the experimenter's understanding of the experiment. However, this is unfeasible since the set of possible outcomes is often infinite (such as with the multivariate Poisson model), and the ordering must be performed independently for each value of μ. Therefore, it is incumbent on the experimenter to formulate a test statistic to *represent* their best judgment. If the experimenter is unable to quantify their best judgment, then they are likely to select a test statistic based on the properties of the confidence regions that it provides, and where it is usually preferable that the test statistic produces a single confidence interval rather than a disconnected confidence region.[59]

Three categories of performance metrics are considered. The first relate to general properties of the confidence intervals, such as if they are empty or if they contain $\mu = 0$. The second examine the endpoints of the confidence intervals, such as the expected upper limit and its variance. The third type of performance metric is the computation of two measures of risk. Risk is the expected value of a loss function, which is subjectively defined by the experimenter according to some desired criteria regarding the confidence intervals. The numerical value of a risk measure provides a concise summary of how well the resultant confidence intervals reduce the loss function defined by the experimenter. These three categories of performance metrics provide quantitative summaries of the results of the Neyman construction using different test statistics, and deliver additional information beyond visually inspecting figures showing confidence intervals or acceptance ventiles, such as in figures 6.23 (a)–(f) (page 313).

One difficulty in designing frequentist measures of performance is that no Bayesian prior distribution may be used. For instance, the expected length between the upper and lower limits of the confidence intervals is defined using only one particular model probability distribution, $E_{\mu'}[u(d) - l(d)] = \sum_{d'}[(u(d') - l(d')]\, p(d'; \mu')$, so that one must select the specific value μ' to define this expectation. Any attempt to average such an expectation over different values of μ implicitly assumes a flat Bayesian prior distribution for μ. This is due to the frequentist assumption that μ is not a random variable, only that there exists one true value, μ_{true}, that is unknown.

[59] One cannot typically report a disconnected confidence region in practice. Therefore, if a confidence region is found to be a union of disconnected sets, then this confidence region may be expanded to be the smallest interval that contains all values of μ that are in the original confidence region. In the following computations of various confidence interval properties, only these singly connected confidence intervals are used. However, this enlargement of the confidence region is only needed for some of the test statistics when implementing the CL_S method (see figure 6.24).

Rather than appeal to Bayesian arguments, the background-only model, $p(d;\, \tilde{\mu}_0)$, and the nominal signal model, $p(d;\, \mu = 1)$, are often used here to evaluate the performance of the Neyman construction.[60] Additionally, the confidence intervals are only constructed for a finite set of observations, $d \in D_{in}$, the "in-bounds" enumerated outcomes deemed to be the most important (see section 6.2.2.2), but not for the remaining infinite "out-of-bounds" outcomes, $d \in D_{oob}$. The total probability of an outcome being contained in D_{oob} grows with increasing μ (see figure 6.3 (a) on page 270), but the enumerated outcomes of D_{in} are selected here so that the total probability of all the possible outcomes in D_{oob} is insignificant for the background and nominal signal models (typically less than 1%).[61]

Table 6.4 shows the results of the first four performance metrics of the Neyman construction at a 95% confidence level (these metrics are defined below). These results are for the bivariate Poisson model of example 2 that has expected background yields of $[4, 3]$ and expected signal of $[1, 3]$. The results for three additional multivariate Poisson models are provided in the next section (page 328).

The first six rows of table 6.4 show the results for the common and Arizona test statistics that are of primary interest. The following three rows are provided for reference and show the results for the upper and central limits discussed in section 6.2.2.1 (which use global ordering rather than test statistics), and the Q_{MP} test statistic discussed in section 6.2.3.4. The last six rows show the results using the common and Arizona test statistics while also implementing the $\mathrm{CL_S}$ method. These four performance metrics regard the general properties of the Neyman construction and are as follows.

1. $\mathbf{p_{acc}^{\mu=1}}$ — The frequentist coverage of the acceptance region for $\mu = 1.0$, i.e., the nominal signal model, $p(d;\, \mu = 1)$. This is given by

$$p_{acc}^{\mu=1} \equiv \sum_{d'} \left[1 - \phi_{rej}^{\alpha}(d';\mu = 1)\right]\, p(d';\, \mu = 1), \tag{6.3.6}$$

   where $\alpha = 0.05$ for a 95% confidence level. In the absence of nuisance parameters, the Neyman construction guarantees that there is no undercoverage for the acceptance region of any μ. However, there is often a small amount of overcoverage due to the discrete (non-continuous) integer values of the possible outcomes of the multivariate Poisson model. This can be seen in the first nine rows of table 6.4, for which values of $p_{acc}^{\mu=1}$ are often slightly larger then 0.950. The last six rows show that overcoverage is quite severe with nearly complete overcoverage of the $\mu = 1.0$ acceptance regions when employing the $\mathrm{CL_S}$ method. (Note that the overcoverage of the nominal $\mu = 1.0$ acceptance region due to the $\mathrm{CL_S}$ method will not be as severe in the three additional examples discussed in the next section).

[60] Most new theories being examined in experimental particle physics are expected to be incorrect, i.e., signal nonexistence is expected to occur more frequently than the rare discoveries of new particles, and so using the background-only model to evaluate performance is useful because background-only expectations will most frequently describe the reality of the searches.

[61] Therefore, the performance metrics are normalized over D_{in} so that, for instance, expected length is given as $\sum_{d' \in D_{in}} [u(d') - l(d')]\, p(d';\, \mu') / \sum_{d' \in D_{in}} p(d';\, \mu')$

2. $\mathbf{p}^{\tilde{\mu}_0}_{\mathbf{emp}}$ — The probability of obtaining an empty confidence region assuming signal nonexistence by using the background-only model, $p(d;\,\tilde{\mu}_0)$, and defined as

$$p^{\tilde{\mu}_0}_{emp} \equiv \frac{\sum_{d'\in D_{in}} I_{empty}\cdot p(d';\,\tilde{\mu}_0)}{\sum_{d'\in D_{in}} p(d';\,\tilde{\mu}_0)}, \tag{6.3.7}$$

where I_{empty} is 1 when the confidence interval for d' is empty and 0 otherwise. Empty confidence regions are difficult to interpret and possibly difficult to justify when reporting.[62] Table 6.4 shows that the Q_{D0} and Q_{AT} test statistics produce confidence regions (at a 95% confidence level) that are expected to be empty with a nontrivial probability of ∼5% (given signal nonexistence) for this particular bivariate Poisson model of example 2. If the experimenter fears obtaining an empty confidence region as the inference result, then the use of Q_{D0} or Q_{AT} may be undesirable.

3. $\mathbf{p}^{\tilde{\mu}_0}_{\mathbf{mle}}$ — The probability that the confidence interval contains the physical MLE, $\tilde{\mu}_p$, using the background-only model,

$$p^{\tilde{\mu}_0}_{mle} \equiv \frac{\sum_{d'\in D_{in}} I_{\tilde{\mu}_p}\cdot p(d';\,\tilde{\mu}_0)}{\sum_{d'\in D_{in}} p(d';\,\tilde{\mu}_0)}, \tag{6.3.8}$$

where $I_{\tilde{\mu}_p}$ is 1 when the confidence interval for d' contains $\tilde{\mu}_p$ and 0 otherwise. The weak likelihood principle implies that $\tilde{\mu}_p$ is the most likely value for μ_{true}, and it is often used as a point estimate of μ_{true}. Therefore, it may be desirable for all confidence intervals to contain $\tilde{\mu}_p$. Table 6.4 shows that confidence intervals produced by Q_{D0} or Q_{AT} have a 5% probability (given signal nonexistence) of not containing $\tilde{\mu}_p$.[63]

4. $\mathbf{p}^{\tilde{\mu}_0}_{\mathbf{0}}$ — The probability that the confidence interval contains the value $\mu = 0$ using the background-only model,[64]

$$p^{\tilde{\mu}_0}_{0} \equiv \frac{\sum_{d'\in D_{in}} I_0\cdot p(d';\,\tilde{\mu}_0)}{\sum_{d'\in D_{in}} p(d';\,\tilde{\mu}_0)}, \tag{6.3.9}$$

where I_0 is 1 if the confidence and 0 otherwise. The confidence interval may not contain $\mu = 0$ if it is two-sided with a positive lower limit or else if it is empty. Table 6.4 shows that for example 2, all methods produce confidence intervals that contain $\mu = 0$ with ∼95% probability (given signal nonexistence) unless implementing the $\mathrm{CL_S}$ method.

[62] Additionally, if an experimenter obtaining an empty confidence region would be tempted to resort to an alternate method of inference, then this *possibility* of flip-flopping will invalidate the frequentist interpretation of the Neyman construction, even if an empty confidence region is *not* obtained with the actual observed outcome.

[63] Note that an empty confidence region cannot contain $\tilde{\mu}_p$ even if $\tilde{\mu}_p = 0$.

[64] Technically, $\mu = \mu_{min}$, where μ_{min} is the smallest numerical value of μ used in construction. For some likelihood ratio test statistics, $\mu = 0$ cannot be used as the null hypothesis (the numerator of the likelihood ratio) because it is already used as the alternative hypothesis (denominator).

	$p_{acc}^{\mu=1}$	$p_{emp}^{\tilde{\mu}_0}$	$p_{mle}^{\tilde{\mu}_0}$	$p_0^{\tilde{\mu}_0}$
Q_{FC}	0.950	0.000	1.00	0.96
Q_{AZ1}	0.950	0.000	1.00	0.96
Q_{D0}	0.953	0.053	0.95	0.95
Q_{AZ2}	0.950	0.000	1.00	0.96
Q_{AT}	0.953	0.053	0.95	0.95
Q_{AZ3}	0.950	0.000	1.00	0.96
upper	0.952	0.053	0.95	0.95
central	0.950	0.022	0.98	0.96
Q_{MP}	0.951	0.017	0.98	0.95
Q_{FC}+$\mathrm{CL_S}$	1.000	0.000	1.00	1.00
Q_{AZ1}+$\mathrm{CL_S}$	1.000	0.000	1.00	1.00
Q_{D0}+$\mathrm{CL_S}$	0.996	0.000	1.00	1.00
Q_{AZ2}+$\mathrm{CL_S}$	1.000	0.000	1.00	1.00
Q_{AT}+$\mathrm{CL_S}$	0.996	0.000	1.00	1.00
Q_{AZ3}+$\mathrm{CL_S}$	1.000	0.000	1.00	1.00

Table 6.4: Example 2 – Comparison of four performance metrics describing the general properties of the Neyman construction for several test statistics using a 95% confidence level. The first six rows show the results for the common and Arizona test statistics and are of primary interest. The "upper" and "central" methods are constructed using global ordering rather than test statistics and provide the one-sided confidence intervals in the style of Garwood [273] and two-sided of Ricker [274], respectively (discussed in section 6.2.2.1). The Q_{MP} test statistic locally orders outcomes by probability, $Q_{MP}(d, \mu) = p(d; \mu)$, so that Q_{MP} is essentially a Fisher significance test for each value of μ (discussed in section 6.2.3.4). The last six rows show the results after implementation of the $\mathrm{CL_S}$ method.

Table 6.5 shows the results for the five performance metrics describing the properties of the upper and lower limits of the resultant confidence intervals of the Neyman construction. These are defined as follows.

1. $\mathbf{u(0)}$ — The upper confidence limit for the outcome $d_i = 0$, a zero count outcome in each i^{th} bin of the discriminating variable. In the bivariate Poisson model of example 2, the background is known to exist and has predicted background yields of $[4, 3]$. According to the background-only model, the observation $[0, 0]$ has a probability of $p([0, 0]; \tilde{\mu}_0) = e^{-(4+3)} \approx 0.09\%$, and thus represents a rather extreme deviation from background-only expectations (and even more extreme in the case of signal existence). Such an observation certainly does not provide evidence for signal existence, but it also may indicate an error in the modeling of the background.[65]

   Table 6.5 shows that the upper limit for the outcome $[0, 0]$ is either small or nonexistent (in the case of an empty confidence region) for the common test statistics, Q_{FC}, Q_{D0} and Q_{AT}, while those for the Arizona test statistics are much larger and include the nominal signal model, $\mu = 1.0$.[66] The longer confidence intervals for the outcome $[0, 0]$ using the Arizona test statistics may be interpreted as representing larger experimenter uncertainty in the background modeling if $[0, 0]$ were the observed outcome.

   Interestingly, when the CL_S method is implemented, $u(0)$ for the common test statistics increases toward the higher values obtained by the non-CL_S Arizona test statistics. The result of the CL_S method is overcoverage for the low count outcomes located to the left of d^b_{mode} in figures6.24 (a), (c) and (e). However, the length of these CL_S confidence intervals for the common methods decreases moving leftward from d^b_{mode} so that the outcome $[0, 0]$ approximately has the smallest upper limit of all possible outcomes. This is quite different from the non-CL_S Arizona test statistics' confidence intervals shown in figures 6.23 (b), (d) and (f), where there is an upward trend in the upper limit when moving leftward toward $[0, 0]$. Therefore, even though the CL_S common test statistics have similar values for $u(0)$ as the non-CL_S Arizona test statistics, the behavior of the upper limits near the outcome $[0, 0]$ are quite different. In general, the CL_S method applied to the common test statistics causes an increase in upper limits for all the outcomes between $[0, 0]$ and d^b_{mode} while the non-CL_S Arizona test statistics increases the upper limits mainly for the outcomes nearer to $[0, 0]$, the outcomes that most strongly imply mismodeling of the background.

[65] If an experimenter obtaining $[0, 0]$ as an outcome would be tempted to flip-flop to an alternate method of inference (perhaps by remodeling the background), then the frequentist interpretation of the Neyman construction is invalidated, even if $[0, 0]$ was not actually obtained. Therefore, the experimenter must select a test statistic that provides a confidence interval for $[0, 0]$ that best aligns with the experimenter's interpretation of that observation (and should probably report this first inference result before remodeling the background). A larger upper limit (with $\mu = 0$ for the lower limit), may be desirable to represent the experimenter's uncertainty in the accuracy of background modeling.

[66] Note that for all these methods the lower limit for the outcome $[0, 0]$ is $\mu = 0$ (or else is nonexistent in the case of empty confidence regions).

2. $\mathbf{E}_{\tilde{\boldsymbol{\mu}}_\mathbf{0}}[\mathbf{u}]$ — The expected upper limit using the background-only model,

$$E_{\tilde{\mu}_0}[u] = \frac{\sum_{d' \in D_{in}} u(d')\, p(d';\, \tilde{\mu}_0)}{\sum_{d' \in D_{in}} p(d';\, \tilde{\mu}_0)}, \tag{6.3.10}$$

where $u(d')$ is the upper limit of the confidence interval for outcome d', and is set to 0 if the confidence interval is empty. Table 6.5 shows that the expected upper limits for the Arizona test statistics are moderately increased by 4–10% above those for the corresponding common test statistics. This indicates that in the event of signal nonexistence, the Arizona test statistics are not expected to deliver confidence regions with upper bounds as small as are obtained with the common test statistics. Recall that all of these test statistics provide confidence intervals that are exact with 95% confidence,[67] and so all are equally valid Neyman constructions. A confidence interval with a smaller upper bound is useful since it will exclude more values of the signal strength parameter, μ, and thus provide a stronger inference statement against signal existence. However, the upper bound must be justified, which depends on whether or not the test statistic properly orders the possible outcomes during acceptance region construction according to the best judgment of the experimenter.

3. $\mathbf{E}_{\tilde{\boldsymbol{\mu}}_\mathbf{0}}[\mathbf{l}]$ — The expected lower limit using the background-only model,

$$E_{\tilde{\mu}_0}[l] = \frac{\sum_{d' \in D_{in}} l(d')\, p(d';\, \tilde{\mu}_0)}{\sum_{d' \in D_{in}} p(d';\, \tilde{\mu}_0)}, \tag{6.3.11}$$

where $l(d')$ is the lower limit of the confidence interval for outcome d', and is set to 0 if the confidence interval is empty. Table 6.5 shows that even for test statistics producing two-sided confidence intervals, such as Q_{FC} and the Arizona test statistics, the lower limit is expected to be near zero in the case of signal nonexistence. One may always obtain the expected confidence interval length according to the background-only model by subtracting the expected lower limit from the expected upper limit.

4. $\mathbf{u}_\mathbf{med}$ — The median upper limit defined as the value of μ for which the critical value rejects 50% of the outcomes (by probability using the background-only model). The median upper limit is explained in section 6.2.4 using figure 6.17(c) on page 297, and is more commonly reported in experimental particle physics searches than the expected upper limit.[68] The median upper limit is usually similar in value to the expected upper limit.

[67] There may be a small amount of numerical overcoverage due to the discrete outcomes of the multivariate Poisson model.

[68] The u_{med} and $\Delta\sigma_u$ metrics were not computed for the upper and central methods discussed in section 6.2.2.2, as can be seen in table 6.5. The software relied on the use of test statistics to compute these values, and the upper and central methods do not use test statistics.

5. $\boldsymbol{\Delta(\sigma_u)}$ — The expected upper limit uncertainty defined as the difference between the two values μ^+ and μ^-, where μ^+ is the value of μ for which its critical value (determined during the acceptance region construction for μ^+) rejects ∼84% of the outcomes by probability using the background-only model (and similarly, μ^- rejects ∼16%). Essentially, μ^+ is the value of μ for which the test power is ∼84% if one considers the background-only model, $p(d; \tilde{\mu}_0)$, as the alternate hypothesis in a Neyman-Pearson hypothesis test (with $p(d; \mu)$ the null hypothesis).[69] In experimental particle physics searches, μ^+ and μ^- are referred to as the ±1 standard deviations of the median upper limit, u_{med}. The width of the green region in figure 6.17(c) provides an example (page 297).

   Once the measurement is made, and d_{obs} is the observed outcome, then the confidence interval for d_{obs} is reported. If the signal does not exist, then the upper limit of this observed confidence interval, u_{obs}, is expected to lie near u_{med}. If $u_{obs} \sim u_{med}$, then the upper limit provided by d_{obs} is consistent with the background-only model, and provides no evidence for signal existence. If $u_{obs} \ll \mu^-$, then there is no evidence for signal existence, but also this may indicate mismodeling of the background since the further u_{obs} is from u_{med}, the less consistent d_{obs} is with background-only expectations. If $u_{obs} \gg \mu^+$, then this may indicate evidence of signal existence, or else may indicate possible background mismodeling.

   A larger value of $\Delta\sigma_u$ indicates greater the uncertainty in the median upper limit. The larger this uncertainty is, the more difficult it will be to draw conclusions regarding signal existence and background mismodeling from the observed upper limit, u_{obs}. This is because a larger $\Delta\sigma_u$ allows u_{obs} to be found in a larger range of values of μ thereby reducing the range of $\mu > \mu^+$ where one might conclude signal existence, and reducing the range of $\mu < \mu^-$ which may indicate background mismodeling.

   Table 6.5 shows that despite having slightly larger median upper limits, the Arizona test statistics have substantially smaller $\Delta\sigma_u$ (14–19% smaller in the first six rows, 18–28% smaller in the last six). The median upper limit uncertainty, $\Delta\sigma_u$, is determined by how rapidly the signal-background discrimination power changes near $\mu = u_{med}$. The more rapidly the discrimination power is increasing with increasing μ, the smaller $\Delta\sigma_u$ will be. Therefore, the inclusion of the Q_{inc} and Q_{rep} terms used to construct the Arizona test statistics from the common test statistics causes a more rapid increase in signal-background discrimination power with increasing μ near $\mu = u_{med}$.

[69] The test power using the background-only model as the alternate hypothesis is typically what is being referred to when one speaks of "signal-background discrimination power".

	$u(0)$	$E_{\tilde{\mu}_0}[u]$	$E_{\tilde{\mu}_0}[l]$	u_{med}	$\Delta(\sigma_u)$
Q_{FC}	0.312	1.54	0.01	1.47	1.57
Q_{AZ1}	1.487	1.69	0.01	1.57	1.27
Q_{D0}	empty	1.27	0.00	1.22	1.57
Q_{AZ2}	1.337	1.35	0.01	1.30	1.35
Q_{AT}	empty	1.27	0.00	1.22	1.65
Q_{AZ3}	1.337	1.38	0.01	1.30	1.37
upper	empty	1.27	0.00	na	na
central	empty	1.56	0.00	na	na
Q_{MP}	empty	2.01	0.01	1.85	1.75
Q_{FC}+$\mathrm{CL_S}$	1.312	1.86	0.00	1.75	1.07
Q_{AZ1}+$\mathrm{CL_S}$	1.712	1.99	0.00	1.80	0.77
Q_{D0}+$\mathrm{CL_S}$	0.787	1.60	0.00	1.52	1.02
Q_{AZ2}+$\mathrm{CL_S}$	1.487	1.67	0.00	1.57	0.82
Q_{AT}+$\mathrm{CL_S}$	0.787	1.61	0.00	1.52	1.10
Q_{AZ3}+$\mathrm{CL_S}$	1.487	1.67	0.00	1.57	0.90

Table 6.5: Example 2 – Comparison of five performance metrics describing the properties of the limits (endpoints, upper and lower) of the confidence intervals for several test statistics using a 95% confidence level. The upper limit is not defined for an empty confidence region and is denoted by "empty" (and note that for all these methods, the lower limit for the outcome $[0, 0]$ is $\mu = 0$, or else is nonexistent in the case of empty confidence regions). The median upper limit and its variance were not computed for the upper and central methods, denoted by "na".

Lastly, table 6.6 shows the results for two measures of risk: the discrimination risk and the reporting risk, which are defined below. Risk is the expectation of a loss function, and a frequentist expectation requires the selection of a probability model, a particular value of μ, in order to compute this expectation. The background-only model, $p(d;\, \tilde{\mu}_0)$, is used to compute the first risk, and the nominal signal model, $p(d;\, \mu = 1.0)$, is used to compute the second. The reasons for these selections are explained in the following definitions.

1. $\mathbf{R}^{\tilde{\mu}_0}_{\mathbf{disc}}$ — The discrimination risk is the background-only model probability of obtaining a confidence interval that contains both $\mu = 1$ and $\mu = 0$ (at a 95% confidence level). Obtaining such a confidence interval as the final inference result does not provide an interpretation that discriminates between the background-only model (signal nonexistence) and the nominal signal model (existence with strength $\mu = 1$). Discrimination risk is useful because it gives the frequentist probability that the observed confidence interval will not exclude the nominal signal model if the signal does not exist.

   Discrimination risk is defined as

$$R^{\tilde{\mu}_0}_{disc} \equiv \frac{\sum_{d' \in D_{in}} p(d';\tilde{\mu}_0)\ \left[1 - \phi^{0.05}_{rej}(d', \mu = 1)\right]\ \left[1 - \phi^{0.05}_{rej}(d', \tilde{\mu}_0)\right]}{\sum_{d' \in D_{in}} p(d';\tilde{\mu}_0)}, \tag{6.3.12}$$

   where $\left[1 - \phi^{0.05}_{rej}(d', \mu = 1)\right]$ indicates that d' lies within the $\mu = 1$ acceptance region (so that $\mu = 1$ is in the confidence interval for d'), and $\left[1 - \phi^{0.05}_{rej}(d', \tilde{\mu}_0)\right]$ indicates that $\mu = 0$ is also in the confidence interval for d'. This implies a loss function of

$$L^{0.05}_{disc}(d;\, \mu = 1, \tilde{\mu}_0) = \left[1 - \phi^{0.05}_{rej}(d', \mu = 1)\right]\ \left[1 - \phi^{0.05}_{rej}(d', \tilde{\mu}_0)\right], \tag{6.3.13}$$

   so that a loss of 1 occurs for d' when the confidence interval for d' includes both $\mu = 0$ and $\mu = 1$.[70]

   A zero discrimination risk of $R^{\tilde{\mu}_0}_{disc} = 0$ indicates that no possible outcome (in D_{in}) has a confidence interval that simultaneously contains both $\mu = 0$ and $\mu = 1$ so that there is no risk of an ambiguous inference result. Total discrimination risk of $R^{\tilde{\mu}_0}_{disc} = 1$ indicates that all possible outcomes produce confidence intervals that do not discriminate between the background-only and nominal signal models.

   Note that the discrimination risk is similar to the power loss risk defined in equation 6.2.21 (page 281) when one uses a 95% confidence level (and thus $\alpha = 0.05$) with $\mu' = 1.0$ and $\mu'' = \tilde{\mu}_0$. The only difference is the addition of the second bracketed term in equation 6.3.12, indicating the inclusion of d' in the $\mu = 0$ acceptance region.[71] Power loss risk is the sum of the background-model probabilities of outcomes contained in the $\mu = 1$ acceptance region (i.e., one minus test power, $1 - \beta$). Discrimination risk

[70] $R^{\tilde{\mu}_0}_{disc}$ does not include empty confidence intervals, but $p^{\tilde{\mu}_0}_{emp}$ may be added to it if desired. $R^{\tilde{\mu}_0}_{disc}$ also does not include confidence intervals containing neither $\mu = 0$ nor $\mu = 1$, but such confidence intervals have lower limits above $\mu = 1$ and are thus have an unambiguous interpretation of signal existence, i.e., no risk of not discriminating between signal nonexistence and existence.

[71] $R^{\tilde{\mu}_0}_{disc}$ also differs from equation 6.2.21 due to the normalization over D_{in} in the denominator, but this causes only minor adjustment since D_{in} is constructed to contain more than 99% of the background-only model.

is the sum of the background-model probabilities of outcomes contained in the $\mu = 1$ acceptance region that are also in the $\mu = 0$ acceptance region. For test statistics producing one-sided upper limits, such as Q_{D0} and Q_{AT}, these risks will be the same since any confidence interval containing $\mu = 1$ will always have $\mu = 0$ as its fixed lower end point. Therefore, $\left(1 - R_{disc}^{\tilde{\mu}_0}\right)$ gives the test power, β, of excluding the background-only model from the $\mu = 1$ acceptance region for test statistics producing one-sided upper limits.[72] For other test statistics producing two-sided confidence intervals, the discrimination risk will be slightly lower than the power loss risk since the summation in the numerator of equation 6.3.12 will exclude the outcomes d' with confidence intervals not containing $\mu = 0$.

The first six rows of table 6.6 show that about 60% of resultant confidence intervals contain both $\mu = 0$ and $\mu = 1$ when using Q_{D0}, Q_{AZ2}, Q_{AT}, or Q_{AZ3} as the test statistic for example 2. This increases to 67% for Q_{FC} and to 90% for Q_{AZ1}. The last six rows show that implementing the CL_S method greatly increases the discrimination risk (for example 2).

2. $\mathbf{R_{rep}^{\mu=1}}$ — The risk of reporting a confidence interval (at a 95% confidence level) that is distant from $\mu = 1$ when in fact $\mu_{true} = 1$. Reporting risk is the expected distance of $\mu = 1$ to the confidence interval evaluated using the nominal signal model.[73] It is defined as

$$R_{rep}^{\mu=1} \equiv \frac{\sum_{d' \in D_{in}} p(d';\, \mu = 1)\, \delta(\mu = 1, CI(d'))}{\sum_{d' \in D_{in}} p(d';\, \mu = 1)}, \tag{6.3.14}$$

where $CI(d')$ is the confidence interval for outcome d', and $\delta(\mu = 1, CI(d'))$ is the shortest distance from $\mu = 1$ to the nearest limit (upper or lower) of $CI(d')$, and is 0 if $\mu = 1$ is contained in $CI(d')$. The corresponding loss function is simply this distance function, $\delta(\mu = 1, CI(d'))$.[74]

The nominal signal strength, $\mu = 1$, is selected to evaluate reporting risk since this is the predicted value according to the new theory. It is possible for the signal to exist with $\mu_{true} \neq 1$ so that the new theory would need modification, but $\mu = 1$ is a sensible selection to make before any actual measurement occurs. If the new theory is correct so that $\mu_{true} = 1$, then reporting risk describes how far $\mu = 1$ is expected to fall outside the reported confidence interval of the observed outcome. It is desirable to produce confidence intervals that reduce reporting risk because if the signal exists at the nominal strength predicted by the new theory, $\mu_{true} = 1$, then the reported confidence interval is expected to lie close to this true value. Even if the reported confidence interval does not contain $\mu = 1$, its upper or lower limit should be expected to be close to $\mu = 1$ when it is the case that $\mu_{true} = 1$.

[72] One may view the Neyman construction of the $\mu = 1.0$ acceptance region as a Neyman-Pearson hypothesis test (NPHT) with the nominal signal model, $p(d; \mu = 1)$, as the null hypothesis and the background-only model, $p(d; \tilde{\mu}_0)$, as the alternate hypothesis.

[73] A similar risk function is briefly discussed in [292].

[74] An alternate definition of distance using "relative error", $\delta(\mu, CI(d))/\mu$, may be better motivated for some experiments.

Table 6.6 shows that using Q_{AZ1} rather than Q_{FC} lowers the reporting risk by 21%, and using Q_{AZ3} rather than Q_{AT} lowers it by 27%. The last six rows show that the overcoverage due to the CL_S method lowers the reporting risk, substantially for this scenario of example 2 (but three additional examples are examined in the next section).

These two risks quantify expectations of the final reported confidence interval that align with two important goals of inference. If the signal does not exist, then it is desirable to obtain a confidence interval that contains $\mu = 0$ but not $\mu = 1$ so that signal existence at the nominal strength is excluded. The smaller the discrimination risk, $R_{disc}^{\tilde{\mu}_0}$, the greater the probability of achieving this goal. If the signal exists at the nominal strength, then it is desirable that the reported confidence interval does not lie far from $\mu = 1$. The smaller the reporting risk, $R_{rep}^{\mu=1}$, the more the reported confidence interval is expected to accurately represent the location of $\mu_{true} = 1$.[75]

[75] Examine $R_{rep}^{\mu=1} = 0.0154$ for Q_{AT}, and assume the signal exists with $\mu_{true} = 1.0$. Q_{AT} produces one-sided upper limits, $u(d)$, so that if μ_{true} is not contained in the confidence interval for outcome d', then $u(d') < \mu_{true}$. $R_{rep}^{\mu=1} = 0.0154$ implies that the expected distance from μ_{true} *downward* to the upper limit is 0.0154. But this does not mean that $E_{\mu=1}[u] = 0.9846$ (the expected upper limit given $\mu = 1$). 95% (by outcome probability mass) of the confidence intervals contain μ_{true} and thus contribute 0.0 to $R_{rep}^{\mu=1}$. This implies that the $\mu_{true} - u(d')$ is expected to be $0.0154/0.05 = 0.308$ for the 5% of confidence intervals that do not contain μ_{true}. This large distance shrinks to 0.178 using Q_{AZ1}.

	$R_{disc}^{\tilde{\mu}_0}$	$R_{rep}^{\mu=1}$
Q_{FC}	0.672	0.0113
Q_{AZ1}	0.898	0.0089
Q_{D0}	0.608	0.0154
Q_{AZ2}	0.593	0.0152
Q_{AT}	0.608	0.0154
Q_{AZ3}	0.593	0.0112
upper	0.605	0.0156
central	0.729	0.0139
Q_{MP}	0.862	0.0238
Q_{FC}+$\mathrm{CL_S}$	1.000	0.0000
Q_{AZ1}+$\mathrm{CL_S}$	1.000	0.0000
Q_{D0}+$\mathrm{CL_S}$	0.920	0.0003
Q_{AZ2}+$\mathrm{CL_S}$	1.000	0.0000
Q_{AT}+$\mathrm{CL_S}$	0.920	0.0003
Q_{AZ3}+$\mathrm{CL_S}$	1.000	0.0000

Table 6.6: Example 2 – Comparison of two measures of risk for the confidence intervals of several test statistics using a 95% confidence level. The discrimination risk, $R_{disc}^{\tilde{\mu}_0}$, is the background-only model probability of obtaining a confidence interval that contains both $\mu = 0$ and $\mu = 1$ so that the final inference result does not allow the exclusion of the nominal signal model. The reporting risk, $R_{rep}^{\mu=1}$, is the nominal signal model expected distance of the reported confidence interval to $\mu = 1$.

6.3.5. Additional Examples

The results of the Arizona test statistics using example 2 demonstrated that the acceptance ventile structure met the goal of addressing model inconsistency and that the reporting risk was lowered when Q_{inc} and Q_{rep} were incorporated into the common test statistics. Example 2 was a bivariate Poisson model with only two bins for the discriminating variable used for inference. Three additional examples were analyzed in order to examine the robustness of the results obtained with example 2.

Table 6.7 provides the expected signal and background counts for three hypothetical signal searches, examples 3a (small), 3b (medium), and 3c (large). (Expected counts are usually fractional (non-integer) in a particle physics search). Example 3a has the smallest expected number of background counts distributed over a discriminating variable grouped into five bins. 3,187 enumerated outcomes, $d \in D_{in}$, are needed in order to guarantee that all outcomes with likelihood functions obtaining their maximum value at less than $\mu = 1.6$ were included, ($\tilde{\mu}_u \leq 1.6$ for all $d \in D_{in}$ and $\tilde{\mu}_u > 1.6$ for all $d \in D_{oob}$, where $\tilde{\mu}_u$ is the unrestricted maximum of each likelihood function). The value of 1.6 was chosen so that the total background-only model probability of all unenumerated outcomes is less than 1%, $p(d \in D_{oob}; \tilde{\mu}_0) < 0.01$, and less than 2.5% for the nominal signal model, $p(d \in D_{oob}; \mu = 1) < 0.025$. Table 6.7 lists these values for examples 3a, 3b and 3c, as well repeating those for example 2 that was use extensively in the previous analyses.

The values for these three examples are selected to represent the range of typical signal searches where a low background yield is expected, with example 3a having the smallest expected background (an expected total background count of 4.2), 3b having the "medium" amount of background (a count of 13), and 3c having the largest (a count of 47). The number of bins used for each example is constrained by the increasing number of enumerated outcomes ($d \in D_{in}$) needed as the expected background counts increases, which increases the computational complexity of determining the confidence interval for each possible outcome in D_{in}, and doing so for each test statistic being examined.[76]

The remainder of this section examines the resultant acceptance ventiles for these three examples using each of the various test statistics in the Neyman construction, followed by the an analysis of the performance metrics obtained. These results allow for a better informed discussion of which test statistics are most useful in meeting various possible experimenters' goals for a final inference result.

[76] The Neyman construction is computationally performed on a two dimensional grid of values of μ versus possible outcomes d, and must retain the acceptance quantile for each grid point, (μ', d'), as well as other various values so that the memory requirements are often quite large.

	value	bin 1	bin 2	bin 3	bin 4	bin 5
Example 3a (small)						
expected background, b_i		1.50	0.90	0.70	0.60	0.50
expected signal, s_i		1.70	1.05	0.53	0.53	0.53
# enumerated outcomes	3,187					
maximum enumerated $\tilde{\mu}_u$	$\sim$1.6					
$p(d \in D_{oob};\, \tilde{\mu}_0)$	< 0.01					
$p(d \in D_{oob};\, \mu = 1)$	< 0.025					
d^b_{mode}	[1, 0, 0, 0, 0]					
d^{s+b}_{mode}	[3, 1, 1, 1, 1]					
Example 3b (medium)						
expected background, b_i		6.50	4.00	2.00	0.50	na
expected signal, s_i		3.10	2.40	1.20	0.40	na
# enumerated outcomes	15,717					
maximum enumerated $\tilde{\mu}_u$	$\sim$1.7					
$p(d \in D_{oob};\, \tilde{\mu}_0)$	$< 6 \times 10^{-3}$					
$p(d \in D_{oob};\, \mu = 1)$	< 0.05					
d^b_{mode}	[6, 4, 2, 0]					
d^{s+b}_{mode}	[9, 6, 3, 0]					
Example 3c (large)						
expected background, b_i		30.5	15.5	1.0	na	na
expected signal, s_i		6.7	4.3	0.6	na	na
# enumerated outcomes	33,305					
maximum enumerated $\tilde{\mu}_u$	$\sim$1.9					
$p(d \in D_{oob};\, \tilde{\mu}_0)$	$< 5 \times 10^{-3}$					
$p(d \in D_{oob};\, \mu = 1)$	< 0.15					
d^b_{mode}	[30, 15, 1]					
d^{s+b}_{mode}	[37, 19, 1]					
Example 2 (bivariate)						
expected background, b_i		4	3	na	na	na
expected signal, s_i		1	3	na	na	na
# enumerated outcomes	431					
maximum enumerated $\tilde{\mu}_u$	$\sim$5.3					
$p(d \in D_{oob};\, \tilde{\mu}_0)$	$< 10^{-9}$					
$p(d \in D_{oob};\, \mu = 1)$	$< 3 \times 10^{-6}$					
d^b_{mode}	[4, 3]					
d^{s+b}_{mode}	[5, 6]					

Table 6.7: Other examples – Expected background and signal counts for a binned discriminating variable in three hypothetical signal searches. Other properties are also included, as well as a summary of example 2 that was used previously.

6.3.5.1. Comparison of the Acceptance Ventiles

Figures 6.25–6.27 show the acceptance ventiles obtained using the common and Arizona test statistics for examples 3a–3c, respectively, with the left subfigures showing the results for the common test statistics, Q_{FC}, Q_{D0} and Q_{AT}, and the right showing those for the three Arizona test statistics.[77]

Showing acceptance ventiles provides the confidence intervals at all confidence levels, $(1 - \alpha)$, in increments of 5% confidence level. Recall that the acceptance ventiles are constructed for each particular value of μ, i.e., for each *horizontal row* of these figures. The Neyman construction for μ' uses a test statistic to locally order the outcomes from those least providing evidence for μ' (small test statistic values) to those providing the most evidence for μ' (large values). The distribution of the test statistic using the signal model, $p(d; \mu')$, is then split into twenty quantiles (ventiles) so that the lowest ventile (shown with a black area) contains the 5% of outcomes (by probability mass) providing the least evidential support for μ', and the highest ventile the most support (shown with a red area). One then examines these acceptance ventiles for a particular outcome d', the *vertical column* corresponding to d', in order to determine the confidence intervals for d'. This is accomplished by viewing the colored areas of these figures as stacked upon one another: the black area covers the entire field, the dark blue lies on top of black, and so on with the red area lying on top of all others. Thus the 95% confidence level intervals ($\alpha = 0.05$) are given by the dark blue areas (extending underneath any the other non-black colors). The 5% confidence intervals ($\alpha = 0.95$) are simply the areas shown in red.

The acceptance ventiles shown in figures 6.25–6.27 demonstrate the following principal results.

1. For all three examples, the test statistics Q_{D0} and Q_{AT} produce empty set confidence regions for outcomes appearing furthest to the left, those outcomes having likelihood functions with the most negative unrestricted MLEs, $\tilde{\mu}_u$.[78] An empty set as an inference result is difficult to interpret, and may be problematic to report. If the experimenter would "flip-flop" to an alternate method of inference after obtaining an empty set result, then the frequentist properties of the Neyman construction are no longer valid, even if the actual result was not an empty set (i.e., the *possibility* of flip-flopping invalidates the frequentist interpretation). Unless the experimenter is willing to report the empty set as the final inference result, the selection of Q_{D0} or Q_{AT} as the test statistic may be avoided to ensure no risk of obtaining an empty set.

[77] As was done for example 2, the expected limits and 95% confidence intervals for Q_{D0} using the CL_S method that resulted from the full Neyman construction were checked against those computed using the OpThyLiC software package [244]. This required over 50,000 separate observed limit computations with OpThyLiC in addition to the three expected limit computations. The agreement between the Neyman construction and OpThyLiC was nearly perfect for all three examples. Note that OpThyLiC failed to return confidence intervals for outcomes furthest to the left when a 95% CL_S confidence level could not be reached for any small value of μ (the reason for this is discussed in section 6.2.4). It is doubtful that an experimenter would report this failure as an empty confidence region, but would rather redesign the experiment to obtain non-empty confidence intervals, i.e., flip-flop to an alternate inference method.

[78] It is difficult to view these empty regions in example 3a (figure 6.25), but they appear in more than 4% of the outcomes (by probability mass) using $p(d; \tilde{\mu}_0)$.

2. For all three examples, the common test statistics, Q_{FC}, Q_{D0} and Q_{AT}, produce confidence intervals with decreasing upper limits moving leftward in the possible outcomes along the x-axes.[79] Meanwhile, the upper limits of the confidence intervals produced by the Arizona test statistics are decreasing moving leftward past d^b_{mode} (the mode outcome of the background-only model), and then begin to increase for outcomes approaching the null outcome, $d_i = 0$ for each i^{th} bin (always plotted furthest to the left).

   For a particle physics search, the possible outcomes further to the left of d^b_{mode} are highly improbable whether or not the signal exists,[80] and obtaining such an outcome would cause suspicion that the modeling was inconsistent with the actual experiment (i.e., the simulated estimates of the expected background yields, b_i in equation 6.2.4 on page 258, may be incorrect). With the common test statistics, the confidence intervals become increasingly shorter and have decreasing upper limits as the outcomes become less consistent with any model (background-only or signal). This means that these inconsistent outcomes have short confidence intervals, $[0, \mu']$, where $\mu' << 1$, and unfortunately these may be misinterpreted as inference results providing strong evidence against signal existence.[81] It may be more useful for the confidence interval to grow in length as the outcomes become more inconsistent with any model, and to possibly contain both $\mu = 0$ and $\mu = 1$ to indicate that the evidence provided by the outcome does not support the exclusion of either the background-only or signal models. The confidence intervals generated by the Arizona test statistics provide this interpretation for these outcomes that are inconsistent with either the background-only or signal models.

3. In all of these figures showing acceptance ventiles (6.25–6.27), the confidence intervals corresponding to decreasing confidence levels (in increments of 5% confidence level with confidence level expressed as $(1 - \alpha)$) appear nested or stacked upon one another with the dark blue areas underneath showing the 95% confidence intervals ($\alpha = 0.05$), to the red areas on the top showing the 5% confidence intervals ($\alpha = 0.95$). A strict application of the Neyman construction requires the confidence level to be set at a desired level so that there is no interpretation for examining the nesting properties of the confidence intervals at different confidence levels. Nevertheless, examining the structure of the acceptance ventiles may be useful in the design or selection of the test statistics to be used for the Neyman construction.

---

[79] Though difficult to see with so many outcomes plotted along the x-axes, the values of the upper (or lower) limits do not vary smoothly from one possible outcome to the next, although these values do follow a general trend. This variation is due to the differing "widths" of the likelihood functions for adjacent outcomes, where smaller widths lead to shorter confidence intervals (when using likelihood ratios for test statistics). This is demonstrated more clearly with example 2 in figures 6.6–6.7 (page 274). The smaller width of a likelihood function is associated with smaller probability mass (figure 6.5 on page 273) for the multivariate Poisson model.

[80] The outcomes d' further to the left of d^b_{mode} typically have small probability mass both using the background-only model, $p(d'; \tilde{\mu}_0)$, and the signal models, $p(d'; \mu)$ for all $\mu > 0$.

[81] Additionally, the *possibility* that the experimenter would flip-flop to an alternate inference method if obtaining such an "unjustified" short confidence interval would invalidate the frequentist properties regardless of the actual outcome.

If the experimenter believes that location of the maximum of the likelihood function, the physical MLE $\tilde{\mu}_p$, represents the most likely inferred value of μ, then it may be sensible for the nested confidence intervals to increasingly surround $\tilde{\mu}_p$. In other words, the experimenter may desire for a lower confidence level interval to more tightly surround the physical MLE (shown with a yellow line) than a higher confidence level interval, so that for example, any red areas should immediately surround the yellow line (vertically) while any blue areas lie farther (vertically) from the yellow line. This would certainly be true of the nesting of any Bayesian credibility intervals if one were using a flat prior distribution, and in this sense is consistent with the likelihood principle (discussed in section 6.2 on page 255).[82]

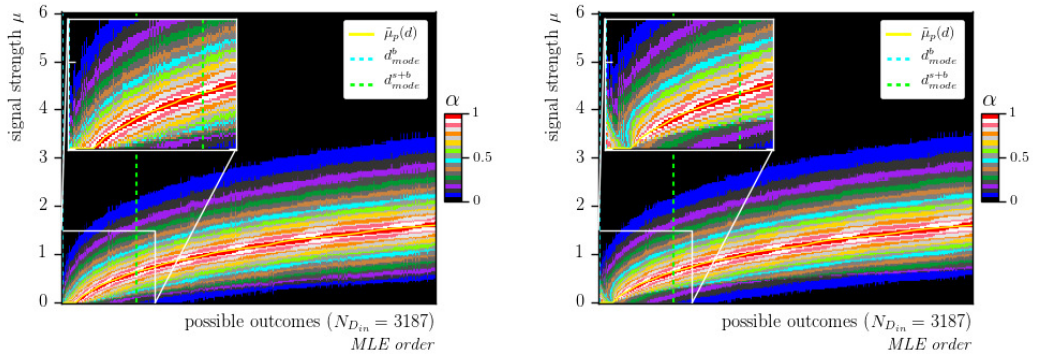
Figures 6.25–6.27 indicate that the test statistics Q_{FC}, Q_{AZ1} and Q_{AZ3} produce confidence intervals with this nesting property of surrounding $\tilde{\mu}_p$ while those of Q_{D0}, Q_{AZ2} and Q_{AT} do not. Note that the outcomes leftward from d^b_{mode} have a zero physical MLE, $\tilde{\mu}_p = 0$, (while the unrestricted MLE is negative, $\tilde{\mu}_u < 0$) so that the yellow line appears at $\mu = 0$ for these outcomes. Examine example 3c using Q_{FC}, shown in figure 6.27 (a). Starting at $\mu = 0$ for the outcomes leftward from d^b_{mode} (see the inset of the figure), the nested confidence intervals increase in confidence level as μ increases, eventually terminating in the dark blue areas of 95% confidence level. For Q_{FC}, these begin at $\mu = 0$ at fairly large confidence levels, from violet areas (90% confidence level) to approximately orange (15%) depending on the distance from d^b_{mode}. Meanwhile, the confidence intervals for these outcomes using Q_{AZ1} and Q_{AZ3} (figures 6.27 (b) and (f)) often begin at $\mu = 0$ with red areas, the lowest confidence level shown of 5%. All three test statistics, Q_{FC}, Q_{AZ1} and Q_{AZ3}, produce confidence intervals that satisfy this nesting property of surrounding $\tilde{\mu}_p$, but Q_{FC} begins at $\mu = 0$ with a much higher confidence level.

The resultant acceptance ventiles of these three examples shown in figures 6.25–6.27 indicate that if one seeks to avoid empty set confidence regions, Q_{D0} and Q_{AT} should not be used. If one further wishes to generate longer confidence intervals to represent the possibility of modeling inconsistency, Q_{FC} should also be removed from consideration, leaving the three Arizona test statistics. If one also desires that the nesting of the confidence intervals at decreasing confidence level increasingly surround the physical MLE, then only Q_{AZ1} and Q_{AZ3} provide confidence intervals that simultaneously meet these three requirements.[83,84]

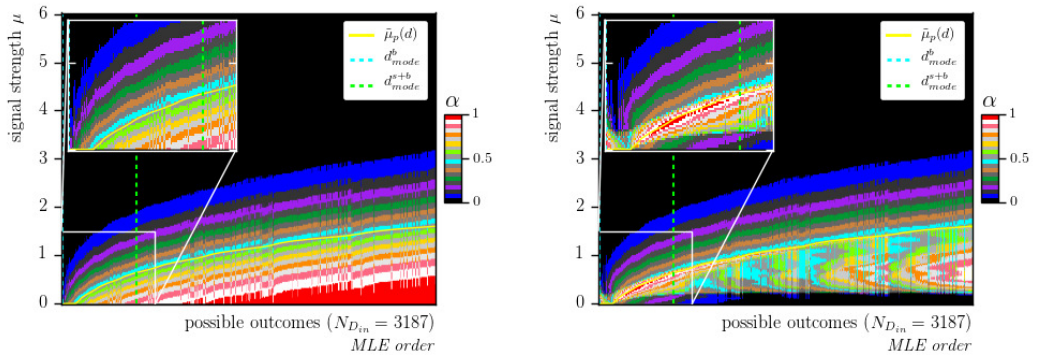
[82] The purpose of section 6.4 is to demonstrate how the acceptance quantiles may be used to construct significance ratio functions that quantify the evidence for *different* values of μ, rather than the simple accept vs. reject interpretation of a confidence interval.

[83] The acceptance ventiles for the most probable test statistics, Q_{MP} (see page 286), are provided in appendix section D.1.3 for reference (page 536).

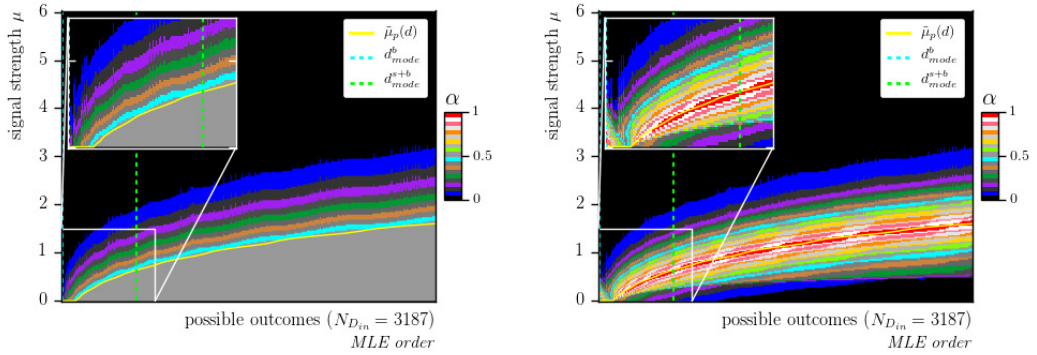
[84] The acceptance ventiles for the common and Arizona test statistics also implementing the CL_S method are provided in appendix section D.1.4 for reference (page 538).



(a) Q_{FC} (b) Q_{AZ1}

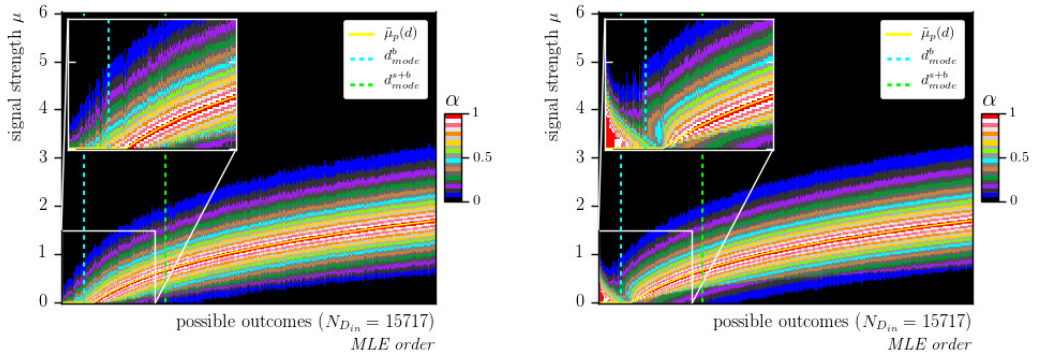


(c) Q_{D0} (d) Q_{AZ2}

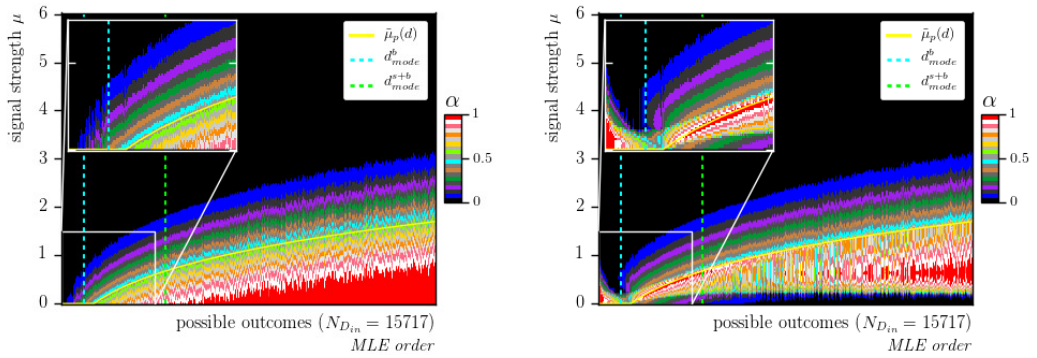


(e) Q_{AT} (f) Q_{AZ3}

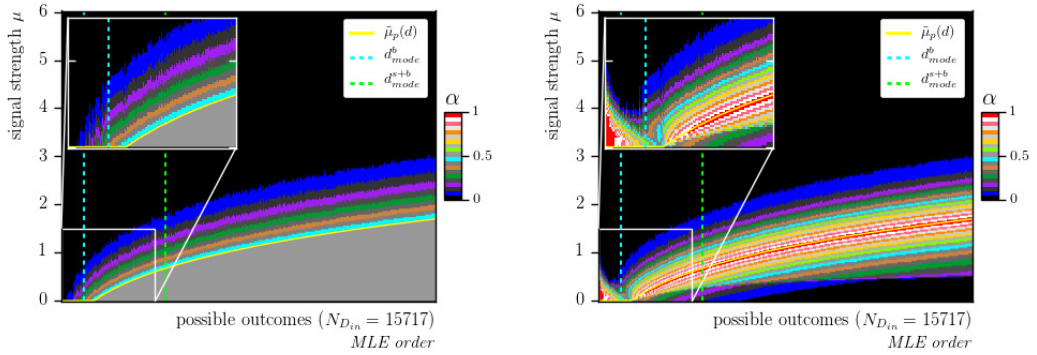
Figure 6.25: Example 3a (small) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}. Empty set confidence regions are produced by Q_{D0} and Q_{AT} for outcomes furthest to the left. For outcomes proceeding leftward from d^b_{mode}, the upper limits produced by the Arizona test statistics switch from decreasing to increasing.



(a) Q_{FC} (b) Q_{AZ1}

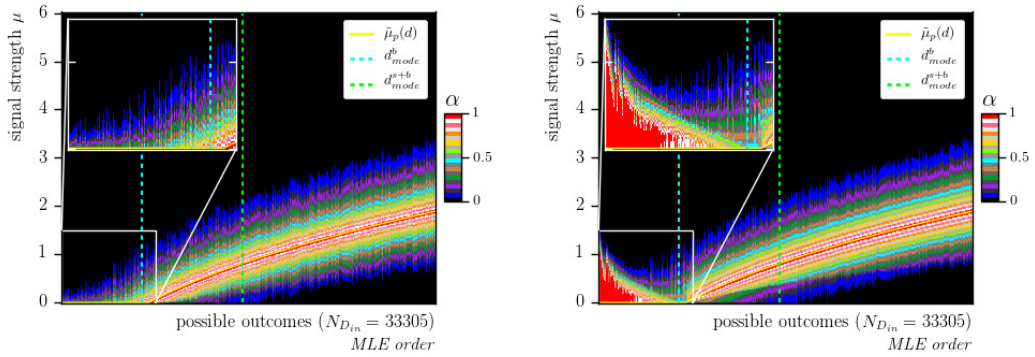


(c) Q_{D0} (d) Q_{AZ2}

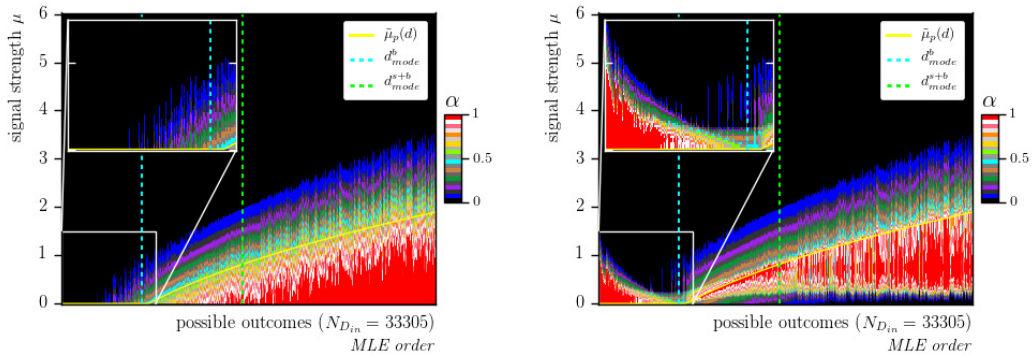


(e) Q_{AT} (f) Q_{AZ3}

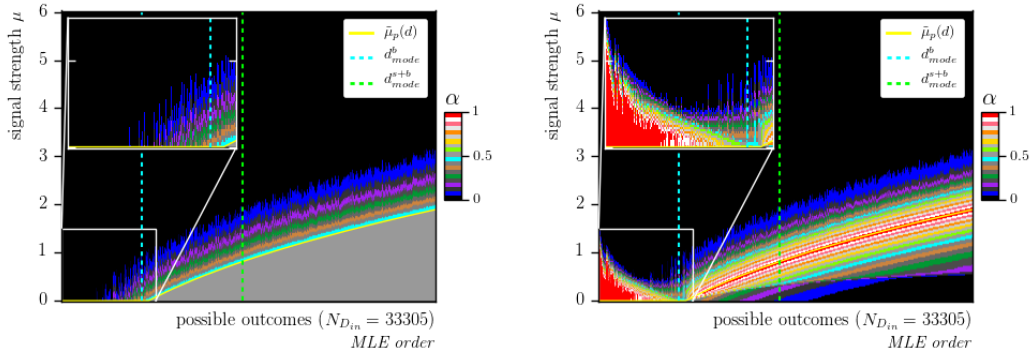
Figure 6.26: Example 3b (medium) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}. Empty set confidence regions are produced by Q_{D0} and Q_{AT} for outcomes furthest to the left. For outcomes proceeding leftward from d^b_{mode}, the upper limits produced by the Arizona test statistics switch from decreasing to increasing.



(a) Q_{FC} (b) Q_{AZ1}



(c) Q_{D0} (d) Q_{AZ2}



(e) Q_{AT} (f) Q_{AZ3}

Figure 6.27: Example 3c (large) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}. Empty set confidence regions are produced by Q_{D0} and Q_{AT} for outcomes furthest to the left. For outcomes proceeding leftward from d^b_{mode}, the upper limits produced by the Arizona test statistics switch from decreasing to increasing.

6.3.5.2. Comparison of the Performance Metrics

Examined next is a comparison of the performance metrics of the common and Arizona test statistics when applied to the three examples, 3a–3c of table 6.7 (page 329). The performance metrics are useful because in addition to examining the acceptance ventiles, the experimenter may wish to compare various quantifiable properties when choosing a test statistic for the Neyman construction. The performance metrics described in section 6.3.4 are summarized below in table 6.8 for the reader's convenience. The performance metrics of the upper and central methods, which utilize global ordering rules (page 272), and the most probable test statistic, Q_{MP} (page 286), are provided for reference.

Recall that these three examples were designed to simulate a signal search in the "low count" regime often observed in particle physics searches, with example 3c having the largest expected background of 47 counts (distributed across three bins of the discriminating variable). Small expected background yields with even smaller expected nominal signal yields causes large coefficient of variance in the multivariate Poisson distributions (the standard deviation of μ divided by the mean of μ), which leads to longer confidence intervals that make it difficult to arrive at an inference conclusion that definitively indicates signal existence or nonexistence. The point of using such examples is to investigate the test statistics' properties when the result of inference is likely to be ambiguous and difficult to interpret. Nevertheless, the results of these examples should extend to experiments with larger expected backgrounds.

metric	description
$p_{acc}^{\mu=1}$	The coverage of the nominal signal model, $\sum_{d'} p(d';\, \mu = 1)$.
$p_{emp}^{\tilde{\mu}_0}$	The probability of empty set confidence regions using $p(d;\, \tilde{\mu}_0)$.
$p_{mle}^{\tilde{\mu}_0}$	The probability the confidence interval (CI) contains $\tilde{\mu}_p$ using $p(d;\, \tilde{\mu}_0)$.
$p_0^{\tilde{\mu}_0}$	The probability the CI contains $\mu = 0$ using $p(d;\, \tilde{\mu}_0)$.
$u(0)$	The upper confidence limit for the outcome $d_i = 0$ for all i^{th} bins.
$E_{\tilde{\mu}_0}[u]$	The expected upper limit using $p(d;\, \tilde{\mu}_0)$.
$E_{\tilde{\mu}_0}[l]$	The expected lower limit using $p(d;\, \tilde{\mu}_0)$.
u_{med}	The median upper limit using $p(d;\, \tilde{\mu}_0)$.
$\Delta(\sigma_u)$	The expected upper limit uncertainty using $p(d;\, \tilde{\mu}_0)$.
$R_{disc}^{\tilde{\mu}_0}$	Discrimination risk: probability the CI contains both $\mu = 0$ and $\mu = 1$ using $p(d;\, \tilde{\mu}_0)$.
$R_{rep}^{\mu=1}$	Reporting risk: expected distance of the CI to $\mu = 1$ using $p(d;\, \mu = 1)$.

Table 6.8: Summary of the performance metrics' definitions from section 6.3.4 provided for convenience.

Table 6.9 shows the resultant performance metrics, which demonstrate the following principal results.

1. The test statistics Q_{D0} and Q_{AT} are expected to produce empty set confidence regions (at 95% confidence level) with a background-only model probability of ∼4–6% in these three examples (viewing the table column for $p_{emp}^{\tilde{\mu}_0}$). The Arizona test statistics and

Q_{FC} never produce empty set confidence regions at 95% confidence.[85]

2. The upper limit at 95% confidence for the null outcome, $d_i = 0$ for each i^{th} bin, is small for Q_{FC} while it is greater than the nominal signal strength $\mu = 1$ for the Arizona test statistics (viewing the table column for $u(0)$). The lower limit for these is $\mu = 0$ so that the short confidence interval of the null outcome for Q_{FC} is interpreted as providing strong evidence against signal existence. The long confidence intervals of the null outcome for the Arizona test statistics are interpreted as yielding no information as to whether or not the signal exists. Which of these two choices provides the correct interpretation of obtaining the null outcome depends on whether or not the experimenter believes the null outcome is consistent with the background modeling.

3. The expected upper limit, $E_{\tilde{\mu}_0}[u]$, and the median upper limit, u_{med}, (both at 95% confidence level) are slightly larger, 0–10%, for each Arizona test statistic than its corresponding common test statistic. However, the uncertainty on the median upper limit, $\Delta(\sigma_u)$, is much smaller, 13–18% for each Arizona test statistic. These three measures use the background-only model to compute these expectations so that a lower expected upper limit is desirable because a lower observed upper limit provides a stronger inference statement against signal existence. The upper limit uncertainty is important because an observed upper limit that deviates significantly from expectations may indicate signal existence (upward deviation of the observed upper limit above the median upper limit) or background mismodeling (downward deviation). A smaller upper limit uncertainty is beneficial because it constrains the range over which the upper limit may deviate from expectations, and thus more readily allow for either of these two conclusions from the final confidence interval.

4. If the signal does not exist, then one desires that the observed 95% confidence interval will include $\mu = 0$, but not include $\mu = 1$ so that the observed confidence interval will indicate exclusion of the nominal signal model. The discrimination risk is the background-only model probability that the observed confidence interval will contain both $\mu = 0$ and $\mu = 1$, so that a lower value of $R^{\tilde{\mu}_0}_{disc}$ is desirable. If the signal exists at the nominal strength, $\mu = 1$, then one desires that the observed confidence interval contains $\mu = 1$, but if it does not, then $\mu = 1$ should at least be close to the upper or lower limit of the observed confidence interval.[86] The reporting risk gives the nominal signal model expected distance of $\mu = 1$ to the observed confidence interval, so that a lower value of $R^{\mu=1}_{rep}$ is desirable.
   Table 6.9 shows that Q_{AZ1} has 9–19% higher discrimination risk than Q_{FC}, but that Q_{AZ1} provides more than a 50% reduction in reporting risk from that of Q_{FC} (viewing

[85] It should be noted that all these methods may produce empty set confidence regions at some lower confidence levels. For example, any vertical column in figures 6.25–6.27 with no acceptance ventile attaining $\alpha = 0.95$ (no red area in that column) has an empty set confidence region at 5% confidence level.

[86] The simple accept vs. reject interpretation of the Neyman construction would imply that if $\mu = 1$ is not within the observed confidence interval, then it is rejected without any further interpretation. However, section 6.4 (page 341) discusses how one may use the acceptance ventiles in a frequentist method of comparing the evidence for *different* values of μ, and this provides additional motivation for seeking to reduce the reporting risk, $R^{\mu=1}_{rep}$, because the final inference result will allow more than the accept vs. reject interpretation.

	$p_{acc}^{\mu=1}$	$p_{emp}^{\tilde{\mu}_0}$	$p_{mle}^{\tilde{\mu}_0}$	$p_0^{\tilde{\mu}_0}$	$u(0)$	$E_{\tilde{\mu}_0}[u]$	$E_{\tilde{\mu}_0}[l]$	u_{med}	$\Delta(\sigma_u)$	$R_{disc}^{\tilde{\mu}_0}$	$R_{rep}^{\mu=1}$
Example 3a (small)											
Q_{FC}	0.950	0.000	1.00	0.96	0.312	1.27	0.01	1.20	1.25	0.563	0.0080
Q_{AZ1}	0.951	0.000	1.00	0.96	1.337	1.39	0.01	1.25	1.02	0.670	0.0025
Q_{D0}	0.951	0.042	0.96	0.96	empty	1.03	0.00	1.02	1.27	0.503	0.0190
Q_{AZ2}	0.950	0.000	1.00	0.96	1.162	1.11	0.01	1.05	1.10	0.490	0.0191
Q_{AT}	0.951	0.042	0.96	0.96	empty	1.03	0.00	1.02	1.27	0.503	0.0190
Q_{AZ3}	0.950	0.000	1.00	0.96	1.162	1.13	0.01	1.05	1.07	0.490	0.0166
upper	0.951	0.052	0.95	0.95	empty	1.03	0.00	na	na	0.502	0.0191
central	0.950	0.026	0.97	0.96	empty	1.27	0.00	na	na	0.626	0.0087
Q_{MP}	0.950	0.001	1.00	0.96	1.662	2.91	0.01	2.85	1.45	0.962	0.0055
Example 3b (medium)											
Q_{FC}	0.950	0.000	1.00	0.96	0.188	1.21	0.01	1.17	1.22	0.555	0.0067
Q_{AZ1}	0.950	0.000	1.00	0.96	1.512	1.31	0.01	1.20	1.00	0.603	0.0031
Q_{D0}	0.950	0.056	0.94	0.94	empty	0.99	0.00	0.97	1.25	0.476	0.0166
Q_{AZ2}	0.950	0.000	1.00	0.96	1.387	1.05	0.01	0.97	1.07	0.440	0.0167
Q_{AT}	0.950	0.056	0.94	0.94	empty	0.99	0.00	0.97	1.25	0.476	0.0165
Q_{AZ3}	0.950	0.000	1.00	0.96	1.412	1.07	0.01	0.97	1.05	0.440	0.0149
upper	0.950	0.052	0.95	0.95	empty	0.99	0.00	na	na	0.474	0.0164
central	0.950	0.027	0.97	0.96	empty	1.20	0.00	na	na	0.593	0.0076
Q_{MP}	0.950	0.015	0.98	0.96	empty	2.01	0.01	2.00	1.47	0.888	0.0146
Example 3c (large)											
Q_{FC}	0.950	0.000	1.00	0.96	0.113	1.29	0.01	1.25	1.22	0.603	0.0058
Q_{AZ1}	0.950	0.000	1.00	0.96	1.812	1.38	0.01	1.25	1.02	0.660	0.0022
Q_{D0}	0.950	0.056	0.94	0.94	empty	1.05	0.00	1.05	1.30	0.514	0.0158
Q_{AZ2}	0.950	0.000	1.00	0.96	1.712	1.11	0.01	1.05	1.12	0.475	0.0161
Q_{AT}	0.950	0.056	0.94	0.94	empty	1.05	0.00	1.05	1.30	0.514	0.0158
Q_{AZ3}	0.950	0.000	1.00	0.96	1.762	1.14	0.01	1.05	1.07	0.475	0.0136
upper	0.950	0.054	0.95	0.95	empty	1.05	0.00	na	na	0.519	0.0159
central	0.950	0.027	0.97	0.95	empty	1.26	0.00	na	na	0.630	0.0073
Q_{MP}	0.950	0.024	0.98	0.96	empty	1.74	0.00	1.77	1.47	0.826	0.0170

Table 6.9: Other examples – Comparison of the performance metrics of the 95% confidence level intervals produced by the common and Arizona test statistics when applied to examples 3a–3c. The upper limit is not defined for an empty confidence region and is denoted by "empty". Also shown for reference are the upper, central and Q_{MP} methods. The median upper limit and its variance were not computed for the upper and central methods, denoted by "na".

the table columns for $R_{disc}^{\tilde{\mu}_0}$ and $R_{rep}^{\mu=1}$). The reporting risk for Q_{AZ2} and Q_{D0} are similar in value, but Q_{AZ2} has a 3–8% lower discrimination risk. Q_{AZ3} has 3–8% lower discrimination risk than Q_{AT}, and has 10–14% lower reporting risk.

If an experimenter were selecting a test statistic based on discrimination and reporting risk alone, then the results of these three examples imply that Q_{AZ1} and Q_{AZ3} are the most useful test statistics for reducing these two risks. If minimal reporting risk is desired at the expense of larger discrimination risk, then Q_{AZ1} is the best option. For best all around performance, Q_{AZ3} seems the most appropriate choice.

Table 6.10 summarizes these findings in lists of pros and cons for each test statistic. The last item in each list shows that there is a trade-off between obtaining reduced discrimination risk with shorter expected confidence intervals, and reduced reporting risk with longer ones. Q_{AZ3} appears to have the most pros and fewest cons if reduced discrimination risk is the primary concern, while Q_{AZ1} if reduced reporting risk is desired. These are the principal results of this section.

	Pros	Cons
Q_{FC}	• No empty set CIs*. • Nested CIs surround $\tilde{\mu}_p$. • Small $R_{rep}^{\mu=1}$.	• Large $E_{\tilde{\mu}_0}[u]$, u_{med} and $\Delta(\sigma_u)$. • Small $u(0)$. • Large $R_{disc}^{\tilde{\mu}_0}$.
Q_{D0} & Q_{AT}	• Smallest $E_{\tilde{\mu}_0}[u]$ and u_{med}. • Small $R_{disc}^{\tilde{\mu}_0}$.	• Large $\Delta(\sigma_u)$. • Empty set CIs. • Empty $u(0)$. • Nested CIs do not surround $\tilde{\mu}_p$. • Largest $R_{rep}^{\mu=1}$.
Q_{AZ1}	• No empty set CIs*. • Smallest $\Delta(\sigma_u)$. • Largest $u(0)$. • Nested CIs surround $\tilde{\mu}_p$. • Smallest $R_{rep}^{\mu=1}$.	• Largest $E_{\tilde{\mu}_0}[u]$ and u_{med}. • Largest $R_{disc}^{\tilde{\mu}_0}$.
Q_{AZ2}	• No empty set CIs*. • Small $\Delta(\sigma_u)$. • Large $u(0)$. • Smallest $R_{disc}^{\tilde{\mu}_0}$.	• Nested CIs do not surround $\tilde{\mu}_p$. • Largest $R_{rep}^{\mu=1}$.
Q_{AZ3}	• No empty set CIs*. • Small $\Delta(\sigma_u)$. • Large $u(0)$. • Nested CIs surround $\tilde{\mu}_p$. • Smallest $R_{disc}^{\tilde{\mu}_0}$.	• Large $R_{rep}^{\mu=1}$.

Table 6.10: Pros and cons of the common and Arizona test statistics. Confidence intervals (CIs) are computed at 95% confidence level. A small upper limit uncertainty, $\Delta(\sigma_u)$, strengthens the possible conclusions of final inference result (item 3 on page 337). The physical MLE, $\tilde{\mu}_p$, is the location of the maximum of a likelihood function restricted to be nonnegative. The utility of nested CIs that surround $\tilde{\mu}_p$ is discussed on page 331. The null outcome upper limit, $u(0)$, should be large if one desires a longer CI to represent possible background modeling inconsistency. *"No empty set CIs" is only true for reasonably large confidence levels (see footnote 85).

6.4. Inference using Confidence-Acceptance Profiles

"Maybe Fisher's biggest blunder will become a big hit in the 21st century!"
– Bradley Efron [293]

Ronald Fisher felt that the use of priors in Bayesian inference was an assumption of "baseless character" [294], and this motivated his development of *fiducial distributions* as a method of frequentist interval estimation [295]. Fisher never provided a complete definition of fiducial distributions [296], but the essential feature was to construct a probability distribution for the parameter of interest (i.e., μ is viewed as a random variable [268]) that is similar to the posterior distribution obtained with Bayesian methods, but without using any prior distribution [297].[87] Jerzy Neyman was able to use Fisher's ideas on fiducial inference to develop his own method of confidence that provided a purely frequentist interpretation [268]. Fisher's fiducial method was attacked because its results differed from those of Neyman's confidence method, the existence of some counter examples [300], and furthermore because of the ambiguous probabilistic interpretation Fisher gave to his fiducial distributions [299].

The Neyman construction at a specific confidence level provides a confidence region for the values of the parameter of interest, μ, for each possible outcome. Each confidence region defines a set of decision rules of whether or not to accept or reject each value of μ. The final inference result is a particular confidence region corresponding to the actual measured outcome, d_{obs}. However, this confidence region does not provide any further information as to which values of μ within the confidence region are more plausible to be μ_{true}.

Worse still, many researchers commonly misinterpret the definition of confidence regions [301], such as the misconception that the 95% frequentist confidence region contains μ_{true} with 95% probability. This misunderstanding is related to the misuse of p-values obtained from frequentist significance testing [302], which has led the American Statistical Association to warn that no single index should substitute for scientific reasoning [303, 304]. This may explain the modern trend toward Bayesian methods which provide posterior distributions that are more easily understood even if more difficult to justify.

The goal of this section is to investigate how the Neyman confidence region construction (NCRC) may be extended to give decision makers a frequentist inference result that describes the strength of evidence for each separate value of μ rather than a simple accept vs. reject binary classification of the different values of μ. This is shown to be possible by using *confidence-acceptance profiles* (CAPS) to construct *signal and background significance ratio functions* (SSR) and (BSR).

In section 6.2.3.2, concepts of Neyman-Pearson hypothesis testing (NPHT) were incorporated into NCRC, specifically, a test statistic from NPHT was used as the local ordering rule in NCRC. In this section, a further extension to NCRC is made by incorporating concepts from Fisher significance testing (FST) into NCRC. Specifically, this section shows

[87] If $a(q,\mu) \equiv p(q < q_{obs};\mu)$, the quantile of the observed test statistic value, q_{obs},then the fiducial distribution of μ is defined as $-\frac{\partial}{\partial\mu}a(q,\mu)$ (where Fisher in [298] makes certain monotonicity assumptions). Also, Fisher on page 533 of [298] implicitly provides the definition of confidence intervals (the confidence principle), which Aldrich seems to misinterpret (section 5 of [299]).

that the test statistic incorporated into NCRC gives rise to signal and background CAPs. Then with FST concepts, the signal and background CAPs can be used to construct the SSR and BSR functions.

Section 6.4 proceeds as follows.

- Section 6.4.1 demonstrates a motivating example wherein three different signal searches yield identical observed confidence intervals, but have drastically differing CAPs. The question for decision makers is whether the differing CAPs provide additional information: does any one search demonstrate more evidence for signal existence than the others (despite identical observed confidence intervals)?
- Section 6.4.2 defines and examines the signal CAP, which is obtained using the signal model, $p(d; \mu)$.
- Section 6.4.3 defines and examines the background CAP, which is obtained using the background-only model, $p(d; \tilde{\mu}_0)$.
- Section 6.4.4 explains how FST, NPHT and NCRC are to be brought together.
- Section 6.4.5 defines and examines the BSR function. The BSR function provides a measure of evidence for each separate value of μ for signal *existence*.
- Section 6.4.6 defines and examines the SSR function. The SSR function provides a measure of evidence for each separate value of μ for signal *nonexistence*.
- Section 6.4.7 demonstrates how a frequentist interpretation of the BSR and SSR functions is obtained.
- Section 6.4.8 discusses the usefulness of scientific *eduction*, a weaker form of frequentist induction that allows for more nuanced decision making.
- Section 6.4.9 shows how CAPs can be used in meta-analysis so that multiple experimental results may be combined, or so that eduction may be easily updated as new data arrives.

6.4.1. Motivation for Confidence-Acceptance Profiles

In particle physics, the search for an undiscovered particle is frequently performed for several possible invariant masses at which the theorized particle may exist, each as a separate signal search. Figure 6.28 shows a hypothetical example of the result of such a search: the upper limits for each possible invariant mass, produced at a 90% confidence level.[88]

For each possible mass on the x-axis, a 90% confidence interval extends vertically from $\mu = 0$ to the upper limit. The thick black line is the observed upper limit that is obtained from the (hypothetical) measured data, d_{obs}. The dashed black line represents the median upper limit, u_{med}, that would be expected if the signal did not exist. The green and yellow areas represent the $\pm 1\,\sigma_u$ and $\pm 2\,\sigma_u$ uncertainties in u_{med}, respectively. In this example, there appears to be an unexpected "bump" at the mass m_X with an observed upper limit above the expectations of the background-only model. This may indicate evidence of signal existence.

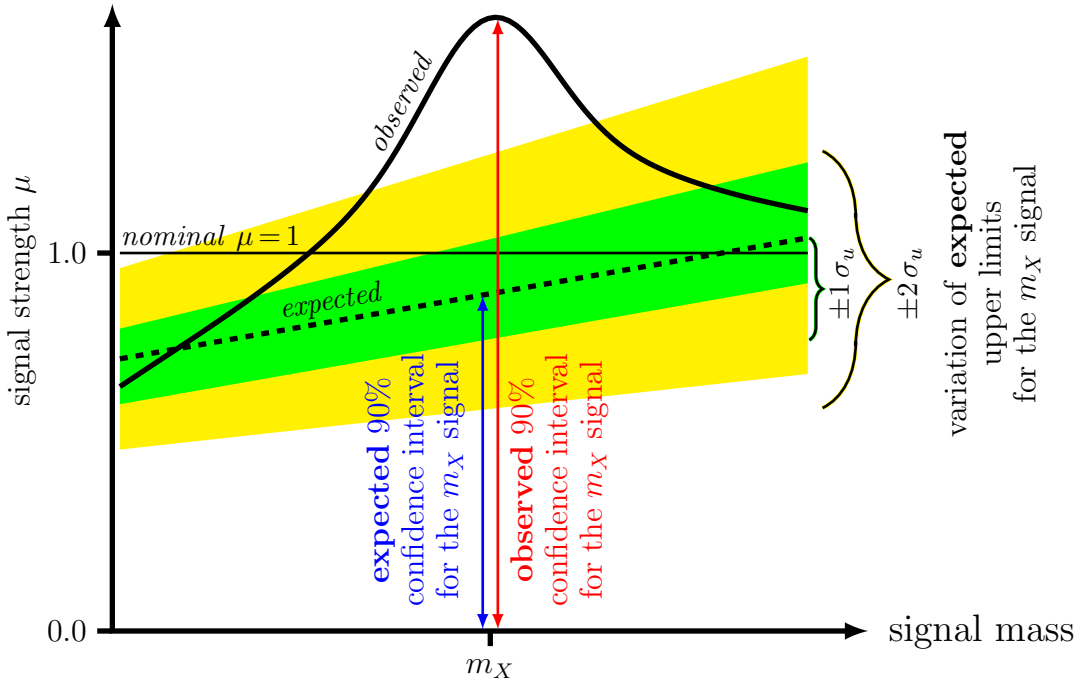


Figure 6.28: The 90% confidence intervals of a hypothetical signal search. The thick black line is the observed upper limit. The dashed black line is the median upper limit expected in the case of signal nonexistence, with green and yellow areas show its uncertainties. A "bump" at m_x indicates evidence of possible signal existence.

[88] This hypothetical example is motivated by an actual result from the combined CDF and D0 collaborations regarding the possible existence of the Higgs particle (figure 4 of [283]). It is also motivated by the excess of events in the diphoton mass spectrum at 750 GeV observed at the ATLAS detector (figure 3 of [305]) and CMS detector (figure 6 of [306]). Many theoreticians constructed new theories to explain this excess at 750 GeV, beginning with its initial discovery in December, 2011, until additional data revealed that it had merely been a statistical fluctuation in August, 2016 (page 27 of [307] and page 16 of [308]).

Imagine that three different particle searches were carried out, and that each returned the same inference result as shown in figures 6.29 (a)–(c). Now assume that decisions must be made based on the results of these three searches, particularly the unexpectedly large observed upper limits at the central invariant masses of each experiment that may indicate signal existence at those masses. These decisions might include: managers ranking these three searches for human resource allocation, possible termination of a search, an experimentalist deciding with which search team to work, a theoretician deciding which theory to study, a funding agency deciding financial resource allocation, etc. Because the results are identical for the three searches, these decisions must either rely on other factors unrelated to the three experimental outcomes, or else upon non-frequentist inference.

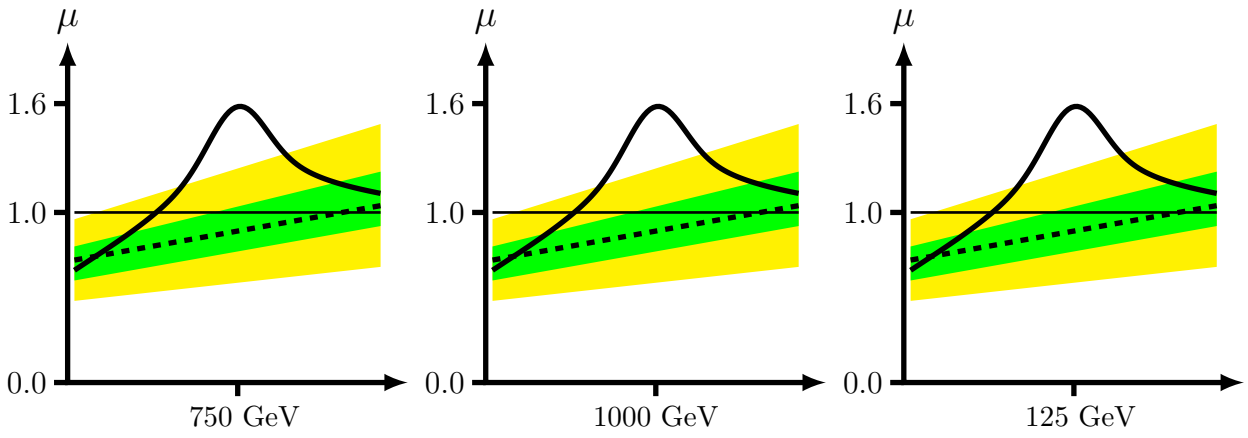
Now imagine that the observed acceptance deciles for these three experiments are also given as shown in figures 6.30 (a)–(c).[89] These only show the observed acceptance deciles for the invariant mass of interest, i.e., where the unexpected bump occurs in each experiment. The vertical boundary of the blue acceptance decile with the black gives the upper limit of the observed confidence interval at 90% confidence level.[90] This corresponds to the vertical location of the observed upper limits at the bumps in figures 6.29 (a)–(c), which is repeated in figures 6.30 (a)–(c) for reference.

It is apparent that the acceptance deciles are quite different for these three examples. For experiment A, the largest valued deciles occur near $\mu = 0$, with the highest (red) for the smallest signal strengths, $\mu < 0.15$. For experiment B, the smaller values of μ have higher acceptance deciles than the larger values of μ, but the deciles themselves are all below $\alpha = 40\%$. Lastly, the acceptance deciles for experiment C reach their highest levels at $\mu = 1.05$, and decrease for both higher and lower values of μ.

The central question of this section is whether these differing acceptance deciles provide additional frequentist information with which to make decisions based on experimental outcomes. The answer will be shown to be *yes*: that experiment C provides the strongest evidence for signal existence, experiment A the strongest evidence for signal nonexistence, and that experiment B provides weak evidence for either hypothesis. This is done by defining *confidence-acceptance profiles*, which are then used to derive functions that evaluate the evidence both for and against signal existence for each *separate* value of μ. These are the *signal and background significance functions*. A frequentist interpretation is then provided for these functions so that subjective beliefs are not needed in evaluating the evidence for each value of μ.

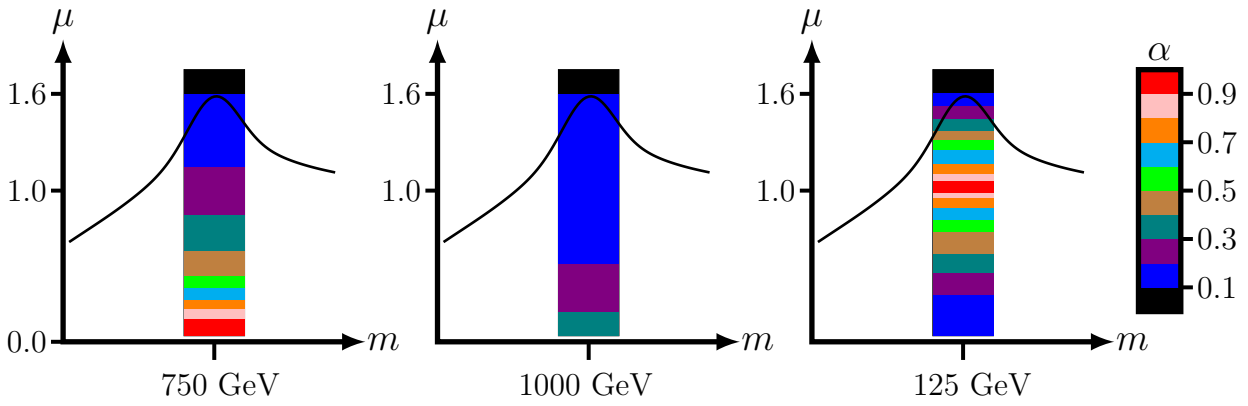
[89] Recall from section 6.2.3.4 that acceptance quantiles are constructed separately for each value of μ during the Neyman construction of the acceptance region for each μ. For a given value μ', the test statistic, $Q(d, \mu')$, is a function that returns higher values for outcomes d providing more evidence for signal existence with strength μ'. The distribution of $Q(d, \mu')$ is made from Monte Carlo sampling of the possible outcomes using the signal model, $p(d; \mu')$. This distribution may be divided into quantiles, with the lowest quantile containing the test statistic values of the outcomes providing the least evidence of signal existence, and the highest quantile containing those providing the most. This is done for each value of μ so that there is a unique *acceptance quantile* for a given μ' and d': the quantile that the test statistic value of d' was contained within when constructing the acceptance region of μ'. Figures 6.30 (a)–(c) show the acceptance quantiles (with decile precision) for the observed outcome, d_{obs}, for each experiment: the *observed acceptance deciles*.

[90] Recall the second paragraph of section 6.3.5.1 on page 330.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.29: Three hypothetical signal searches (for three different theorized particles) that result in identical 90% confidence intervals, each showing a "bump" at a particular invariant mass where the observed upper limit is above expectations. There is no way to distinguish which experiment provides stronger evidence for signal existence for the mass at the bump.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.30: The observed acceptance deciles at the masses with the unexpected bumps in their upper limits are shown for the three hypothetical signal searches sharing identical 90% confidence intervals. A color map for the deciles is shown on the far right, where the confidence level is given by $(1-\alpha)$. The observed upper limits from figures 6.29 (a)–(c) are shown for reference. The goal is to prove that experiment C provides the greatest evidence of signal existence at the bump, experiment A the greatest evidence against signal existence, and that experiment B gives ambiguous evidence.

6.4.2. The Signal Confidence-Acceptance Profiles

In recent years, there has been a re-examination of Fisher's approach with the development of confidence distributions [309, 310]. The confidence distribution was initially proposed by David Cox [252] as a more flexible approach to summarizing evidence via confidence, which was motivated by the claim that even if μ_{true} was not in the confidence interval for the observed outcome, $CI_\alpha(d_{obs})$, it would not be too far outside it [311].

A confidence distribution for outcome d' is a purely frequentist function of the parameter of interest, $H(d', \mu) = \alpha$, that has the mathematical properties of a cumulative distribution, $H(d', -\infty) = 0$ and $H(d', +\infty) = 1$ [310]. A confidence distribution does not have a probabilistic interpretation as μ cannot be viewed as a random variable in frequentist statistics. However, the confidence distribution may be considered a distribution estimator for μ so that one may obtain random variable estimates of $\hat{\mu}$ from the confidence distribution, $H(d', \hat{\mu})$.

The main advantage of an inference method that provides a distribution instead of a simple interval is that the distribution may be easily used to compute common statistics (mean, median, p-values, etc.). These may be derived from the confidence density, $h(d, \mu) = dH/d\mu$.

The confidence distribution is defined so that if $\mu = \mu_{true}$, then $H(d, \mu_{true})$ has a uniform distribution, $U[0, 1]$, as a function of d. This property allows confidence intervals to be computed at any confidence level, $(1 - \alpha)$, from the inverse function, $H^{-1}(d, \mu) = \alpha$, as

$$CI_\alpha(d) = \left[\, H^{-1}(d, \alpha/2),\ H^{-1}(d, 1 - \alpha/2) \,\right] \tag{6.4.1}$$

(or else $[-\infty,\ H^{-1}(d, 1 - \alpha)\,]$ for upper limits and $[\, H^{-1}(d, \alpha),\ +\infty\,]$ for lower). Therefore, having the confidence distribution as a final inference result gives decision makers the ability to examine confidence intervals at any level, and also to manipulate posterior-like distributions if desired.

However, a function with the properties of a cumulative distribution that provides a simple and easily interpreted summary of the data is not always obtainable [311]. In fact, one can see from figure 6.25 (a) on page 333 that the Feldman-Cousins test statistic does not always provide a 5%-confidence interval for outcomes furthest to the left (i.e., no red area vertically above those outcomes). Deriving H for these outcomes d' would result in $H(d', +\infty) = c < 1$. One could normalize in this situation, but that would destroy the confidence interpretation. The definition of the confidence distribution would have to be altered in some manner to accommodate this. Additionally, using $h(d, \mu)$ to provide statistical measure of μ may not provide the correct repetition frequency [257]. For instance, if one used $h(d_{obs}, \mu)$ to compute the mean $\bar{\mu}$ for use as a point estimate, what interpretation should be made if $\bar{\mu}$ is outside the confidence region at the desired confidence level?

Fiducial and confidence distributions seem to be of limited appeal to the broader statistical community due to these and possibly other issues. However, the figures showing acceptance quantiles in the previous sections of this chapter were found to greatly aid in the examination (and prediction) of inference results using different test statistics, where the goal was to obtain confidence intervals with specific properties. This leads one to the conclusion that more inference information is more useful than less information, so long as

this information is interpreted correctly.

Definition – The Signal Confidence-Acceptance Profile

Let the function $S(d, \mu)$ be the acceptance quantiles generated during the Neyman confidence region construction (NCRC) so that figures such as 6.26 (a)–(f) on page 334 shows examples of $S(d, \mu)$ plotted with ventile precision. Let the *signal confidence-acceptance profile* (SCAP) for an outcome d' and all values of μ be nothing more than the acceptance quantiles, $S(d', \mu)$, that were derived during the Neyman construction:

$$S(d', \mu) = \sum_{d} p(d;\, \mu), \text{ for } d \ni Q(d, \mu) \leq Q(d', \mu), \tag{6.4.2}$$

Stated another way, the SCAP gives the signal model test statistic cumulative distribution, $p(q \leq Q(d', \mu);\, \mu)$, that is found as one determines the acceptance region for μ during NCRC. Thus, the observed SCAP, $S(d_{obs}, \mu)$, is shown in each of figures 6.30 (a)–(c) with decile precision.

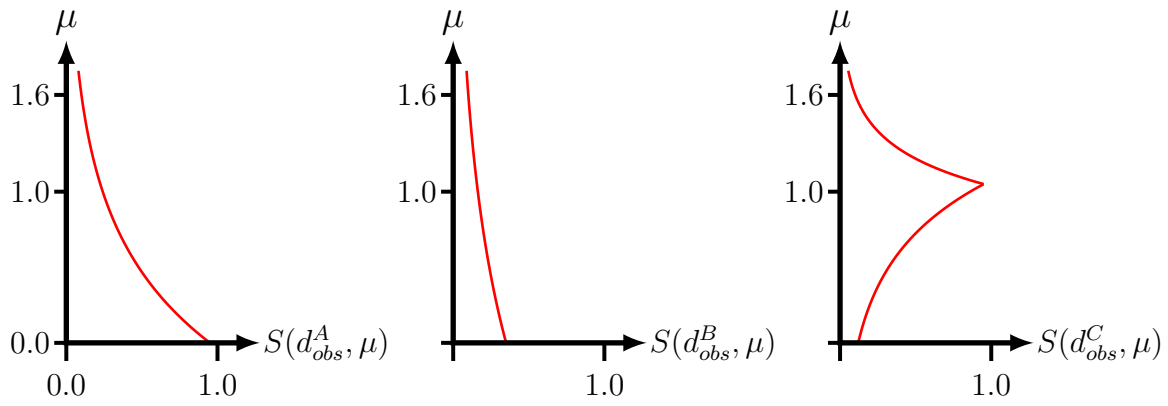
Clearly an SCAP, $S(d', \mu)$, for outcome d' can be used to provide confidence regions at any specified confidence level, $(1 - \alpha)$, with the possibility of returning the empty set if $S(d', \mu) < \alpha$ for all μ. But does the SCAP provide any further information that could be useful to a frequentist? The results of the three hypothetical experiments of figures 6.30 (a)–(c) may be used to illustrate that this is so.

Figures 6.31 (a)–(c) show the observed SCAPs, $S(d_{obs}, \mu)$, for the three experiments with full precision instead of the decile precision shown in figures 6.30 (a)–(c). These examples represent three different experiments and so have three different observed outcomes, d^A_{obs}, d^B_{obs} and d^C_{obs}. These SCAPs share the property that $S(d_{obs}, \mu) \to 0$ as $\mu \to +\infty$ because it is assumed that after some point, d_{obs} provides diminishing evidence for increasing values of μ (discussed more fully in appendix section D.2). These also show that there is a large difference in the value $S(d^C_{obs}, \mu = 1)$ from $S(d^A_{obs}, \mu = 1)$ and $S(d^B_{obs}, \mu = 1)$.

Examine figure 6.31 (c) where the observed SCAP reaches it highest values at $\mu = 1.05$. This does not mean that $\mu_{true} = 1.05$ with higher probability since μ_{true} is not a random variable. In fact, if $\mu_{true} = 1.05$, then $S(d, 1.05)$ has a uniform distribution, $U[0, 1]$, over d. This is a simple property of quantiles. For example, the probability of obtaining a value of $S(d, 1.05)$ in any of its *quintiles* is 20%, and the probability of $S(d, 1.05)$ to be at or below the value 0.42 is 42% (if $\mu_{true} = 1.05$).

In other words, if $\mu_{true} = 1.05$ so that one knew that $p(d;\, \mu = 1.05)$ correctly describes the experiment, then the probability of obtaining an observed outcome d^C_{obs} that has $S(d^A_{obs}, 1.05)$ in the highest decile (as is shown with the red area near $\mu = 1.05$ of figure 6.30 (c)) is 10% and is equal to the probability of obtaining an observed outcome d^C_{obs} such that $S(d^A_{obs}, 1.05)$ is in the lowest decile. This means that if $\mu_{true} = 1.05$ and before d^C_{obs} had been revealed, then it was equally probable to obtain a red area or black area near $\mu = 1.05$ (or any other decile color). Clearly one must be cautious not to interpret SCAPs as directly providing evidence for the location of μ_{true}.

However, if $\mu_{true} \neq 1.05$, then the distribution of the test statistic for $\mu = 1.05$, $Q(d, 1.05)$, is no longer uniform, and is shifted to lower values. This means that it would be less probable



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.31: The observed signal confidence-acceptance profiles (SCAPS), $S(d_{obs}, \mu)$, corresponding to the three hypothetical experiments shown in figures 6.30 (a)–(c), which also showed the observed SCAPs, but only with decile precision.

for $Q(d, 1.05)$ to reach high values, and thus more improbable for figure 6.30 (c) to reach the highest decile at $\mu = 1.05$. The red area at $\mu = 1.05$ is improbable when $\mu_{true} \neq 1.05$, and becomes more improbable as $|\mu_{true} - 1.05|$ increases.

With regard to figures such as 6.12 (b) on page 288 showing the acceptance ventiles for all possible outcomes (now called the signal CAPs with ventile precision), the result of experiment A of figure 6.31 typically occurs when $d_{obs} \sim d^b_{mode}$, the experiment B result occurs when $d_{obs} \ll d^b_{mode}$ (further to the left in an acceptance ventile figure), and experiment C when $d_{obs} \sim d^{s+b}_{mode}$.

Definition – The Frequentist Likelihood Property

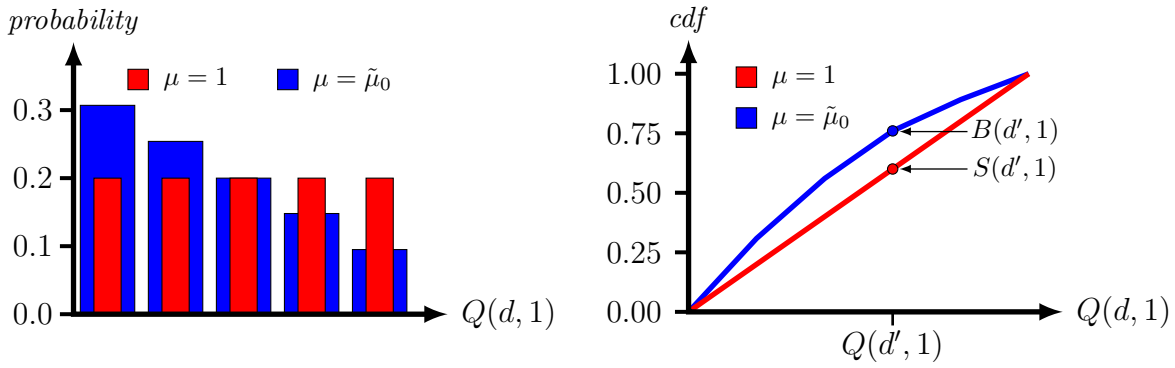
It will be useful at times to make assumptions regarding the structure of the SCAPs in relation to the physical MLEs, $\tilde{\mu}_p$. The *frequentist likelihood property* (FLP) is now defined to describe the result of the NCRC wherein the confidence intervals are increasingly clustered around the maximum of the likelihood function as the confidence level decreases (i.e., red acceptance ventiles are closer to the yellow line than other ventile colors). Figures 6.25–6.27 demonstrate that Q_{FC}, Q_{AZ1} and Q_{AZ3} satisfy the FLP while Q_{D0}, Q_{AT} and Q_{AZ2} do not.

More specifically, the FLP is satisfied for a given outcome d' if $S(d', \tilde{\mu}_p) > S(d', \mu)$ whenever $\mu \neq \tilde{\mu}_p$ (i.e., $S(d, \mu)$ reaches its maximum at the MLE), and also if given two values $\mu' > \tilde{\mu}_p$ and $\mu'' > \tilde{\mu}_p$ with $\mu' > \mu''$, then $S(d', \mu') < S(d', \mu'')$. And similarly if $\mu' < \tilde{\mu}_p$ and $\mu'' < \tilde{\mu}_p$ with $\mu' < \mu''$, then $S(d', \mu') < S(d', \mu'')$. It is assumed that the test statistics used to provide the results of the three hypothetical experiments satisfy the FLP. This assumption is needed to guarantee certain asymptotic properties of the SCAPs, and is discussed further in appendix section D.2.

6.4.3. The Background Confidence-Acceptance Profile

In figure 6.30 (c), it was improbable for d_{obs} to have attained the high acceptance decile at $\mu = 1$ (red area) if μ_{true} is distant from 1.[91] Therefore, in order to interpret the meaning of an observed SCAP value, $S(d_{obs}, \mu')$, one must consider not just the signal model, $p(d;\, \mu')$, but also some suitable alternative model(s). In the case of a signal search, the background-only model, $p(d;\, , \tilde{\mu}_0)$, is a sensible alternative.

In order to better understand the implication of the differences between figures 6.31 (a)–(c) at $\mu = 1$, it is useful to examine the hypothetical probability distributions of the test statistic, $Q(d, \mu = 1)$, which is shown in figure 6.32 (a). The following arguments that are constructed using this hypothetical distribution could apply to any of the three experiments, A, B or C. Figure 6.32 (a) shows that for $\mu = 1$, the test statistic provides some moderate discrimination between the signal and background-only models, with greater background-model probability (blue) at lower values of $Q(d, \mu = 1)$ than that of the signal model (red).[92] Figure 6.32 (b) shows the cumulative distributions of $Q(d, \mu = 1)$ obtained from integrating the test statistic distributions, beginning at the lowest values of $Q(d, \mu = 1)$ (integrating the left tails), and this further illustrates that more background-only model probability mass is located at lower $Q(d, \mu = 1)$ than for the signal model.



(a) Experiments A, B or C

(b) Experiments A, B or C

Figure 6.32: The hypothetical test statistic distributions for $\mu = 1$, binned into quintiles according to $p(d;\, \mu = 1)$: (a) quintile probabilities, and (b) cumulative distribution functions (cdf). The background CAP, $B(d', \mu = 1)$, and signal CAP, $S(d', \mu = 1)$, at an arbitrary outcome d' are shown for reference.

[91] The peak in the observed SCAP at $\mu = 1.05$ in figure 6.31 (c), where $S\left(d^C_{obs}, \mu = 1.05\right)$ achieves its maximum value, is *not* meaningful if $\mu_{true} = 1.05$ because if $\mu_{true} = 1.05$, then $S(d, \mu = 1.05)$ has a uniform distribution over d so that $S(d, \mu = 1.05)$ had equal probability to attain any value between 0 and 1 before d_{obs} was revealed. However, if $\mu_{true} \neq 1.05$, then it is improbable for $S(d, \mu = 1.05)$ to reach such a high value. This is because the values of $S(d, \mu = 1.05)$ are determined from the test statistic distribution using the *particular* model $p(d;\, \mu = 1.05)$ while the actual observed data is generated by $p(d;\, \mu_{true})$. If the test statistic, $Q(d, \mu = 1.05)$, is designed to return high values only for outcomes that provide strong evidence for $\mu = 1.05$, then the true test statistic distribution that is sampled and generated by $p(d;\, \mu_{true})$ will have smaller probability masses at higher values of $Q(d, \mu = 1.05)$ than the distribution generated by $p(d;\, \mu = 1.05)$.

[92] Technically, only the use of $Q_{D0}(d, \mu) = p(d;\, \mu)/p(d;\, \tilde{\mu}_0)$ guarantees that for any value $Q(d, \mu)$, the background-only model has greater total probability for all $q \leq Q(d, \mu)$ than the signal model. However, this was nearly always true with all the test statistics examined here.

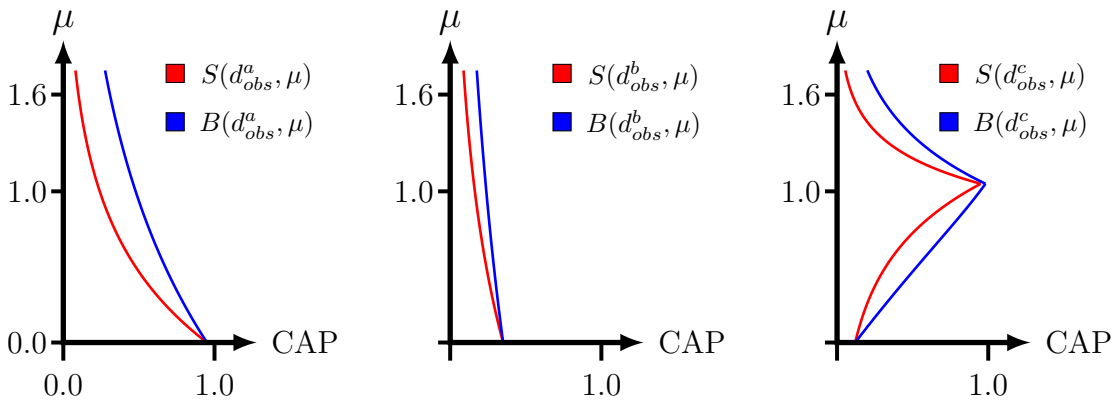
Definition – The Background Confidence-Acceptance Profile

Define the background confidence-acceptance profile (BCAP), $B(d, \mu)$, in a similar manner to the SCAP,

$$B(d', \mu) = \sum_{d} p(d; \tilde{\mu}_0), \text{ for } d \ni Q(d, \mu) \leq Q(d', \mu)., \tag{6.4.3}$$

which simply replaces the signal model of equation 6.4.2 with the background-only model. However, note that $B(d', \mu)$ is still a function of μ because the BCAP gives the background-only model test statistic cumulative distribution, $p(q \leq Q(d', \mu); \tilde{\mu}_0)$, that is found as one determines the acceptance region for μ during NCRC. For the $\mu = 1$ cumulative distributions shown in figure 6.32 (b), it is clear that $B(d, \mu = 1) \geq S(d, \mu = 1)$ for all outcomes d because more background-only model probability mass is located at lower test statistic values.

Figures 6.33 (a)–(c) show the observed SCAP and BCAP functions for the three hypothetical experiments.[93] The value of the SCAP is always smaller than that of the BCAP because more background-only model probability mass is located at lower test statistic values and the CAP values are obtained from integrating the test statistic distributions beginning with the smallest test statistic value (integrating the left tails).



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.33: The observed confidence-acceptance profiles (CAPs), $S(d_{obs}, \mu)$ and $B(d_{obs}, \mu)$, which measure the probability of the test statistic, $Q(d, \mu)$, having the value $Q(d_{obs}, \mu)$ or less according to the signal model (red) and background-only model (blue), respectively.

[93] A note on the asymptotic behavior. As μ approaches zero, there is no discrimination between the signal and background-only models so that $S(d, \mu) \sim B(d, \mu)$ as $\mu \to 0$ (for any d). It was already discussed that $S(d, \mu) \to 0$ as $\mu \to +\infty$. Furthermore, under certain assumptions (satisfying the FLP and Gaussian-like models), the BCAP has the property $B(d, \mu,) \to \frac{1}{2} B(d, \mu = 0)$ as $\mu \to +\infty$. The difficult explanation for this is provided in appendix section D.2.

6.4.4. Relating the Three methods: FST, NPHT and NCRC

> *"This concept* [of statistical evidence] *is not part of the Neyman-Pearson theory of tests and confidence region estimation, which denies any role to concepts of statistical evidence, as Neyman consistently insists."*
>
> – Allan Birnbaum [312]

The above quotation illustrates Birnbaum's frustration with the formulation of Neyman-Pearson hypothesis tests (NPHTs) and the Neyman confidence region construction (NCRC) that disallow an evidential interpretation of the final inference result. In NPHT and NCRC, the final inference result is simply to accept or reject each value of μ, with no other method to compare the evidence for differing values of μ.

For a given value μ', the SCAP and BCAP ($S(d, \mu')$ and $B(d, \mu')$) are functions of the possible outcomes, d. Once the observed outcome, d_{obs}, is obtained, the observed SCAP value and the observed BCAP value are $S(d_{obs}, \mu')$ and $B(d_{obs}, \mu')$, respectively. These observed CAP values are equivalent to Fisher significance test (FST) p-values of the test statistic distributions (left tail probabilities). A smaller observed CAP value occurs when $Q(d_{obs}, \mu')$ is small, indicating that d_{obs} was less consistent with signal existence at strength μ'. A small value of $S(d_{obs}, \mu')$ indicates a significant deviation from the signal model expectations.

However, from the viewpoint of signal nonexistence, one expects to measure smaller values of the test statistic. Significant deviations from the background-only model are *higher* values of $Q(d, \mu')$ that are more signal-like.[94] Subtracting the CAPs from one, $1 - S(d, \mu')$ and $1 - B(d, \mu')$, gives the result of integrating over values of $Q(d, \mu') > Q(d_{obs}, \mu')$ (integrating the upper tails of the test statistic distributions). A small value of $B(d_{obs}, \mu')$ indicates a significant deviation from the background-only model expectations.

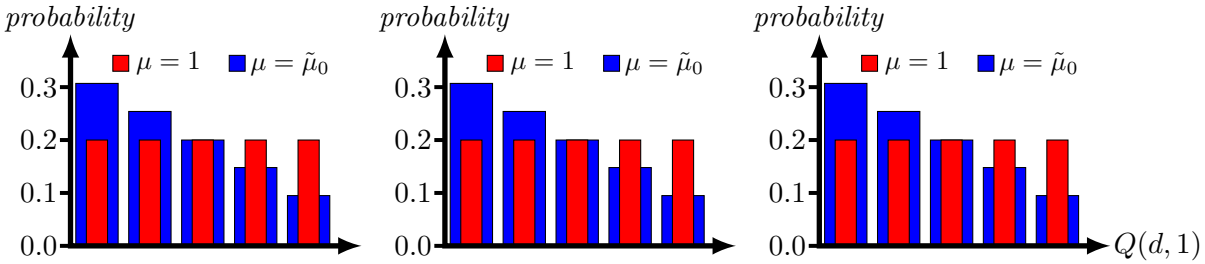
In other words, the test statistic, $Q(d, \mu')$, is now performing three tasks for a given value μ'. First, it provides the local ordering rule, $Q(d, \mu')$, that determines acceptance regions for μ' in NCRC. Second, it provides an NPHT test statistic, $Q(d, \mu')$, to compare the signal model, $p(d; \mu')$, to any desired alternative model(s), which quantifies the discrimination power between the signal model and the alternative. Third, it provides an FST test statistic, $Q(d_{obs}, \mu')$, that quantifies (via CAPs) how significant the deviation of d_{obs} is from the expectations of either the perspective of signal existence (at strength μ') or the perspective of the alternative model.

Because the background-only model is the natural alternative model to examine in evaluating the evidence for signal existence, one may wonder why $Q_{D0}(d, \mu') = \frac{p(d; \mu')}{p(d; \tilde{\mu}_0)}$ is not always used. After all, the Neyman-Pearson lemma gives that Q_{D0} provides the most powerful signal-background discrimination. However, recall from section 6.3 that the experimenter must design the local ordering rule for NCRC according to their best judgment: ranking outcomes in order of increasing evidence supporting μ', and this may involve criteria other than consistency with the background-only model. In other words, the experimenter may give up some signal-background discrimination to obtain confidence intervals that are most aligned with the experimenter's understanding of the experiment.

[94] If a test statistic, Q, other than $Q_{D0}(d, \mu) = p(d; \mu)/p(d; \tilde{\mu}_0)$ is used, low values of Q do not necessarily imply more background-like outcomes. However, the test statistics explored here provide good signal-background separation when μ is not too close to zero.

6.4.5. The Background Significance Ratio Function

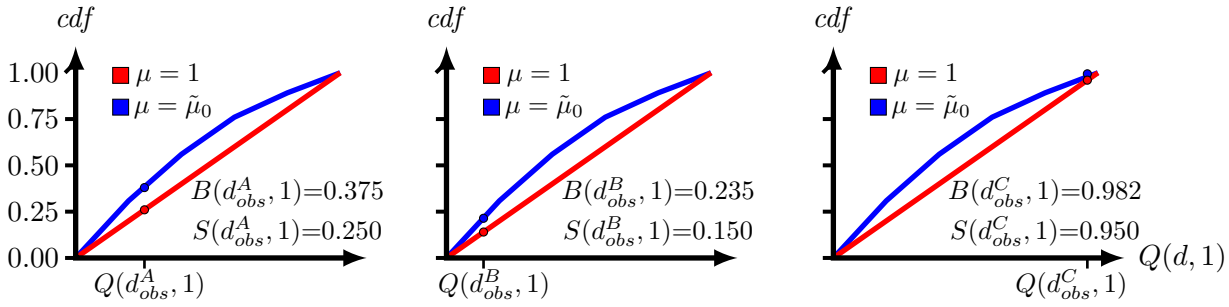
Imagine that these three experiments had identical test statistic distributions when performing the Neyman construction for the $\mu = 1$ acceptance region, as is shown in figures 6.34 (a)–(c), even though they are searches for different signals.[95] Again, note that the signal model probabilities (red) are 0.2 for each quintile bin because this is the definition of a quintile. However, the background-only model probabilities (blue) have less probability mass at higher test statistic values.



(a) Experiment A (b) Experiment C (c) Experiment C

Figure 6.34: The hypothetical test statistic distributions for $\mu = 1$, binned into quintiles according to $p(d; \mu = 1)$. (Repeating figure 6.32 (a) for each experiment for clarity of presentation.)

Similarly, figures 6.35 (a)–(c) show the cdfs of the test statistic distributions for $\mu = 1$ in each of the three experiments. These are equivalent to the BCAP and SCAP (as functions of d), $B(d, \mu = 1)$ and $S(d, \mu = 1)$, because the CAPs are defined as left tail probability sums of the test statistic distributions. The observed CAP values for $\mu = 1$ are obtained from figures 6.33 (a)–(c), which are different among the experiments.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.35: The cdfs of the test statistic distributions for $\mu = 1$ for the three experiments. (Repeating figure 6.32 (b) for each experiment.) The observed CAP values are shown with small circles and provided as text.

[95] This is not likely to happen in practice, but allows for a direct comparison between the three examples that elucidates the frequentist properties of the CAPs.

Before revealing an observed outcome, the probability at $\mu = 1$ to observe a particular SCAP value *lower* than some threshold may be computed, but the result will depend on which model is used. If the signal model, $p(d;\, \mu = 1)$, is used, then the computation is trivial:

$$P\left[S(d, \mu = 1) \leq S(d', \mu = 1);\, \mu = 1\right] = S(d', \mu = 1), \tag{6.4.4}$$

where any possible outcome d' may be used to determine the threshold value, $S(d', \mu = 1)$. This is merely a property of quantiles; the probability to be at or below the 25^{th} percentile is 25%. This is equivalent to the left tail probability of the distribution of $Q(d, \mu = 1)$ under the signal model.

If the background-only model is used, $p(d;\, \tilde{\mu}_0)$, then this probability is given by

$$P\left[S(d, \mu = 1) \leq S(d', \mu = 1);\, \mu = \tilde{\mu}_0\right] = B(d', \mu = 1), \tag{6.4.5}$$

because the BCAP gives the left tail probability of the distribution of $Q(d, \mu = 1)$ when the background-only model is used. Similarly,

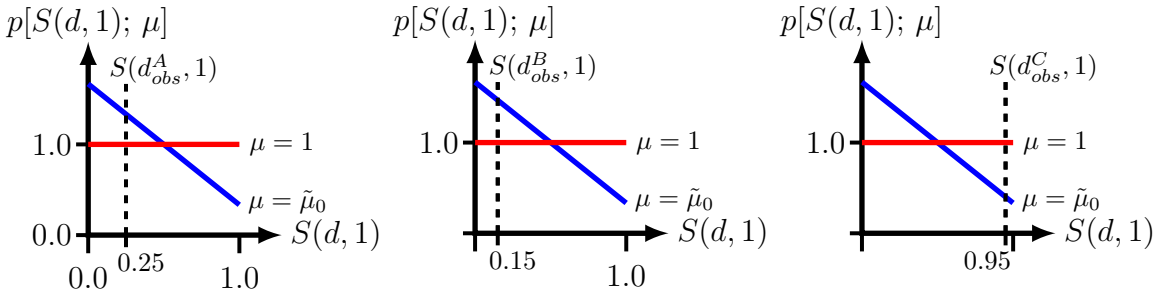
$$P\left[S(d, \mu = 1) > S(d', \mu = 1);\, \mu = 1\right] = 1 - S(d', \mu = 1), \tag{6.4.6}$$

and

$$P\left[S(d, \mu = 1) > S(d', \mu = 1);\, \mu = \tilde{\mu}_0\right] = 1 - B(d', \mu = 1). \tag{6.4.7}$$

Figures 6.36 (a)–(c) show the probability distributions of the SCAPs under both the signal (red) and background-only (blue) models. The signal model distribution is constant (uniform) because it is equally likely to observe the test statistic in any of its quantiles.

Integrating these distributions gives the tail probabilities. For example, in experiment A, the probability of obtaining a value at or below $S(d^A_{obs}, \mu = 1)$ is 0.25 according to the signal model, while it is 0.375 according to the background-only model.

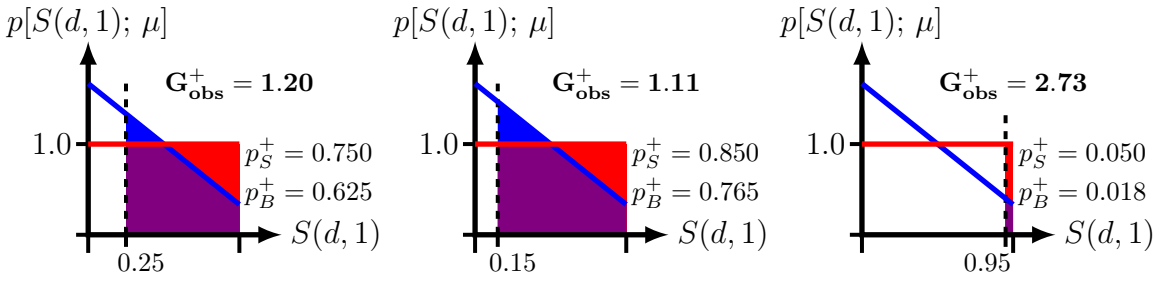


(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.36: The probability distributions of the SCAPs at $\mu = 1$ according to the signal model, $p[S(d, 1);\, \mu = 1]$ (red lines), and according to the background-only model, $p[S(d, 1);\, \tilde{\mu}_0]$ (blue lines). The observed SCAP for each experiment is shown with a dashed line.

If the test statistic, $Q(d, \mu')$, has been designed so that outcomes with large values of $Q(d, \mu')$ provide the strongest evidence for the signal existing with strength μ' (perhaps according to the criteria described in figure 6.18 on page 300), then a larger value of $S(d_{obs}, \mu')$ is more consistent with the signal model than with the background-only model. This is because a larger $S(d_{obs}, \mu')$ indicates that $Q(d_{obs}, \mu')$ has a high value where the background-only model has lower probability mass than the signal model. A larger value of $Q(d_{obs}, \mu')$ is thus a deviation of the observed outcome from the background-like possible outcomes.

A p-value provided by a Fisher significance test (FST) answers the question, "What is the probability to have observed a deviation this large or larger according to one particular model?". One may now use the SCAP distributions to answer a similar question, "What is the probability to have observed a value larger than $S(d_{obs}, \mu')$ according to the signal model compared to that of the background-only model?" The answer to this question with $\mu' = 1$ is shown for each experiment in figures 6.37 (a)–(c). This is done by finding the upper tail probabilities for both the signal and background-only models, p_S^+ and p_B^+, respectively, and taking their ratio, $G_{obs}^+ = p_S^+/p_B^+$.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.37: The probability to have obtained outcomes yielding higher SCAP values than the observed SCAP value, $S(d_{obs}, \mu = 1)$, after performing the Neyman construction for the $\mu = 1$ acceptance region. The red area shows this probability using the signal model, $p(d; \mu = 1)$, giving $p_S^+ = 1 - S(d_{obs}, \mu = 1)$. The blue area shows this probability using the background-only model, $p(d; \mu = \tilde{\mu}_0)$, giving $p_B^+ = 1 - B(d_{obs}, \mu = 1)$. (The violet area shows overlapping regions of integration.) The ratio of these two upper tail probabilities provides the observed background significance ratio (BSR) value, $G^+(d_{obs}, \mu = 1) = p_S^+/p_B^+$ (labeled G_{obs}^+ in figures).

Note that p_S^+ and p_B^+ could be computed directly from the upper tails of the test statistic distributions, $Q(d, \mu = 1)$, using the signal and background-only models after observing $Q(d_{obs}, \mu = 1)$. The SCAP distributions are used instead with the goal of making sense of the resultant SCAPs of the Neyman construction that have been shown as acceptance ventiles throughout this chapter (as opposed to showing confidence intervals at a single confidence level).

The ratio of two probabilities, each generated by a different model, is a *likelihood ratio*.[96] Thus, $G^+(d_{obs}, \mu = 1)$ is the likelihood ratio comparing the significance of the deviation of

[96] The ratio of two probabilities, each generated by the same model, is an *odds ratio*.

the observed outcome d_{obs} from the background-like possible outcomes according to the signal model vs. the background-only model. For the signal strength $\mu = 1$, experiments A and B have observed background significance ratios (BSRs) that are $\sim$1, while that of experiment C is $\sim$3. This indicates that for experiments A and B, the deviation significance is nearly equally likely with either the signal or background-only models. For experiment C, such a deviation is about three times more likely with the signal model than with the background-only model, and may be considered stronger evidence for signal existence at $\mu = 1$. This does *not* mean that experiment C indicates a higher *probability* for μ_{true} to be equal to 1. Because a likelihood ratio is obtained from different models, it does not have a frequentist interpretation, which requires the use of a single model. However, such an interpretation is provided in section 6.4.7 (page 363).

Definition – The Background Significance Ratio Function

Given a particular outcome d', the background significance ratio (BSR) function (of the parameter values μ) is given by the ratio of the two upper tail probabilities of the test statistic distributions (or equivalently, of the SCAP distributions) using the signal and background-only models,

$$\begin{aligned} G^{+}(d', \mu) &= \frac{p\left[Q(d, \mu) > Q(d', \mu);\, \mu\right]}{p\left[Q(d, \mu) > Q(d', \mu);\, \tilde{\mu}_0\right]} \\ &= \frac{p\left[S(d, \mu) > S(d', \mu);\, \mu\right]}{p\left[S(d, \mu) > S(d', \mu);\, \tilde{\mu}_0\right]} \\ &= \frac{1 - S(d', \mu)}{1 - B(d', \mu)}, \end{aligned} \tag{6.4.8}$$

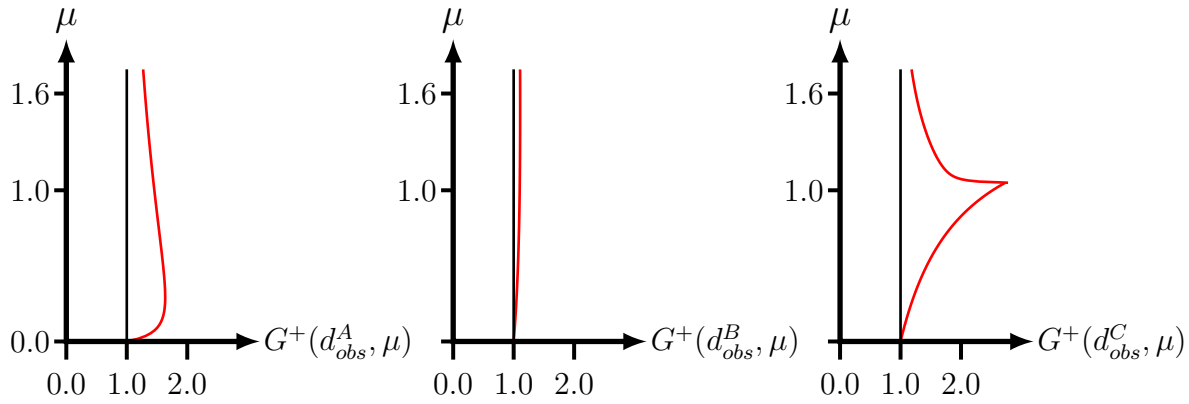
where the observed BSR function is given by $G^{+}(d_{obs}, \mu)$. The "+" superscript indicates the use of upper tail probabilities. Note that $G^{+}(d', \mu = 0) = 1$ since the signal model with strength $\mu = 0$ is equivalent to the background-only model (in the absence of nuisance parameters). Also, $G^{+}(d, \mu) \geq 1$ because $B(d, \mu) \geq S(d, \mu)$ due to the background-only model having greater probability mass than the signal model at lower test statistic values.

The numerator and denominator of the BSR function each measure the significance of the deviation of outcome d' from background-like possible outcomes. Qualitatively, the evidence supporting the inference $\mu_{true} = 1$ is more convincing if the numerator is large so that d_{obs} is not an extreme occurrence according to the signal model. Evidence for signal existence is also more convincing if the denominator is small so that d_{obs} is an extreme occurrence according to the background-only model.

Alternatively, one may note for a given value μ' that $B(d_{obs}, \mu')$ is equal to the observed test power [313] of the test statistic, $(1 - \beta_{obs})$, and that $S(d_{obs}, \mu')$ equals the observed test size, α_{obs}, when one views the Neyman construction of the acceptance region of μ' as a hypothesis test between the signal model (null hypothesis) and the background-only model (alternative), and using the observed test statistic value as the critical value,[97] $Q^{*}_{obs}(\mu) \equiv Q(d_{obs}, \mu)$.

[97] See Mayo and Spanos for a similar discussion of using the observed test statistic value as the critical value in quantifying evidence [260].

Thus far, the value of the observed BSR function at $\mu = 1$ has allowed for the comparison of the evidence for $\mu = 1$ provided by the three different experiments. Using the CAPs shown in figures 6.33 (a)–(c), this can be done for each value of μ, and the result is shown in figures 6.38 (a)–(c).



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.38: The observed background significance ratio functions $G^+(d_{obs}, \mu)$ (red lines) of the three hypothetical experiments. A vertical line denotes the value 1.0 for which $G^+(d_{obs}, \mu)$ always greater than or equal to (because $B(d, \mu) \geq S(d, \mu)$ as demonstrated in figures 6.33 (a)–(c)). Recall that the observed confidence interval for all three experiments is $\mu \in [\,0,\, 1.6\,]$.

For any choice of test statistic, there is very little signal-background discrimination for small values of μ where the test statistic distributions for the signal and background-only models are very similar. This means that tail probabilities of both distributions will be nearly equal at small μ and cause $G^+(d, \mu)$ to be near 1.0. In other words, there can be no strong evidence for signal existence over nonexistence at very small values of μ.

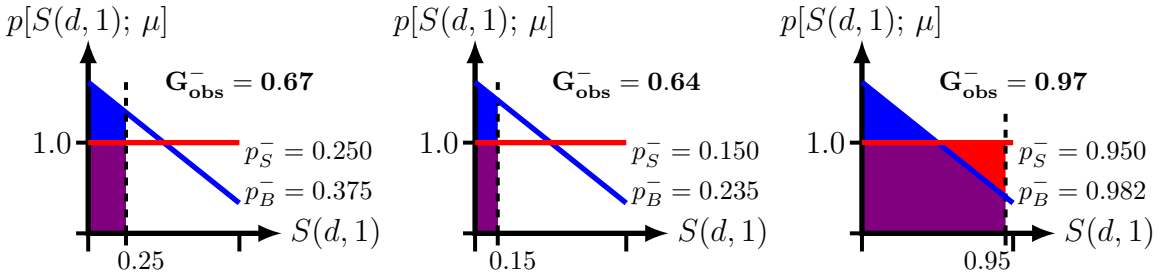
In the case of experiment A, the strongest evidence of signal existence vs. nonexistence is found at strength $\mu = 0.28$ and decreasing for larger μ. This may be interpreted as d_{obs} providing stronger evidence supporting smaller values of μ than for the nominal signal strength, $\mu = 1$, predicted by the new theory. The observed BSR function of experiment B does not seem to support any particular value of μ over any other value. There is a slow increase from 1.0 at $\mu = 0$ toward an asymptotic value of 1.08, but the significance of $Q(d^B_{obs}, \mu = 1$ is nearly equally likely with both the signal and background-only models. The observed BSR function of experiment C peaks with a value of 2.73 at $\mu = 1.05$, which is close to the nominal signal strength, $\mu = 1$. If one was forced to take a bet on which experiment had true signal existence at its nominal strength (against nonexistence), then one should be most comfortable choosing experiment C because it provides the strongest evidence for existence with $\mu = 1$.

6.4.6. The Signal Significance Ratio Function

The BSR function, $G^+(d, \mu)$, was constructed from the ratio of the upper tail probabilities of the SCAP distributions (or alternatively, the test statistic distributions). The observed BSR function measures the evidence for signal *existence* by examining the significance of the deviation of the observed outcome from the background-like possible outcomes. The greater the value of $G^+(d_{obs}, \mu)$, then the greater the likelihood for the signal model than the likelihood of the background-only model to generate such a deviation.[98] Thus, one takes the perspective of signal nonexistence, and examines the deviation from this perspective as evidence for signal existence.

One may take the opposite perspective, that of signal existence with strength μ', so that low values of the observed test statistic value, $Q(d_{obs}, \mu')$, represent a deviation from the signal-like possible outcomes. A more significant deviation implies stronger evidence for signal *nonexistence*. In a similar manner, the signal significance ratio (SSR) function may be defined in terms of lower tail probabilities of the SCAP distributions (or alternatively, of the test statistic distributions).

Figures 6.39 (a)–(c) demonstrate the computation of the observed SSR function values at $\mu = 1$ for the three experiments. This shows for experiments A and B that the signal model likelihood of the deviation of $Q(d_{obs}, \mu = 1)$ is roughly two thirds of the background-only model likelihood, while the likelihoods are nearly equal in experiment C. Therefore, the observed outcomes in experiments A and B provide stronger evidence against signal existence at $\mu = 1$ than does d^C_{obs}.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.39: The probability to have obtained outcomes yielding lower SCAP values than the observed SCAP value, $S(d_{obs}, \mu = 1)$, after performing the Neyman construction for the $\mu = 1$ acceptance region. The red area shows this probability using the signal model, $p(d; \mu = 1)$, giving $p^-_S = S(d_{obs}, \mu = 1)$. The blue area shows this probability using the background-only model, $p(d; \mu = \tilde{\mu}_0)$, giving $p^-_B = B(d_{obs}, \mu = 1)$. (The violet area shows overlapping regions of integration.) The ratio of these two lower tail probabilities provides the observed signal significance ratio (SSR) value, $G^-(d_{obs}, \mu = 1) = p^-_S / p^-_B$ (labeled G^-_{obs} in figures).

[98] The term likelihood is typically applied to a single possible outcome. Each of the numerator and denominator of $G^+(d_{obs}, \mu)$ are p-values from a Fisher significance test (tail probabilities), using either the signal model or background-only model, respectively. One may consider a p-value to be a likelihood in the sense that it is the sum of the likelihoods of outcomes in a distribution tail.

Definition – The Signal Significance Ratio Function

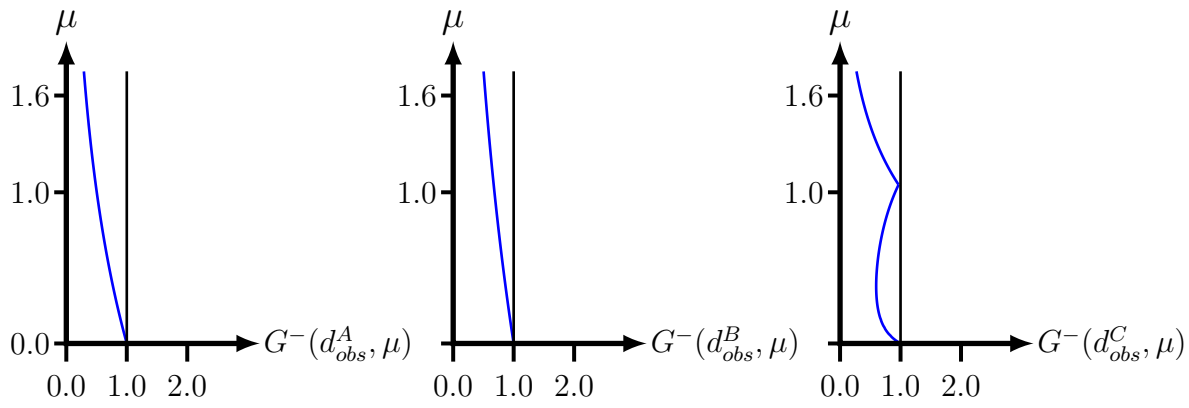
Given an particular outcome d', the signal significance ratio (SSR) function (of the parameter values μ) is given by the ratio of the two lower tail probabilities of the test statistic distributions (or equivalently of the SCAP distributions) using the signal and background-only models,

$$G^{-}(d',\mu) = \frac{p\left[Q(d,\mu) \leq Q(d',\mu);\, \mu\right]}{p\left[Q(d,\mu) \leq Q(d',\mu);\, \tilde{\mu}_0\right]} = \frac{p\left[S(d,\mu) \leq S(d',\mu);\, \mu\right]}{p\left[S(d,\mu) \leq S(d',\mu);\, \tilde{\mu}_0\right]} = \frac{S(d',\mu)}{B(d',\mu)}, \qquad (6.4.9)$$

where the observed SSR function is given by $G^{-}(d_{obs},\mu)$. The "$-$" superscript indicates the use of lower tail probabilities. Note that $G^{-}(d,\mu) \leq 1$ because $B(d,\mu) \geq S(d,\mu)$.[99]

Figures 6.40 (a)–(c) show the observed SSR function for each of the three experiments. For a given value μ', a smaller value of $G^{-}(d_{obs},\mu')$ indicates that $Q(d_{obs},\mu')$ is small and therefore less signal-like and more background-like. As with the BSR function, there is very little signal-background discrimination for small values of μ when the signal and background-only model test statistic distributions are very similar, and so $G^{-}(d,\mu)\sim 1$. There can not be strong evidence for signal nonexistence at very small values of μ.

For experiment A, the observed SSR function begins at $G^{-}(d^{A}_{obs},\mu=1)=1$ and decreases steadily with increasing μ. This may be interpreted as d^{A}_{obs} providing increasing evidential support for signal nonexistence at increasing signal strength μ. For experiment B, $G^{-}(d^{B}_{obs},\mu)$ decreases more slowly than in experiment A. For a given value μ', the evidence provided by d^{B}_{obs} for signal nonexistence at μ' is not as strong as that provided by d^{A}_{obs}. For experiment C, the evidence provided by d^{C}_{obs} for nonexistence is stronger for values near $\mu=0.4$ than with the other experiments, while there is no evidence for nonexistence around $\mu=1$.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.40: The observed signal significance ratio (SSR) functions, $G^{-}(d_{obs},\mu)$ (blue), of the three hypothetical experiments. A vertical line denotes the value 1.0 for which the SSR function is always smaller. Recall that the observed confidence interval for all three experiments is $\mu \in [\,0,\,1.6\,]$.

[99] This is due to the background-only model having greater probability mass than the signal model at lower test statistic values (or lower SCAP values).

6.4.6.1. The SSR Function and the CL_S Confidence Interval

It will now be shown that the SSR function may be used to derive the CL_S confidence interval at any desired CL_S confidence level. This implies that the confidence intervals obtained using the CL_S method may be explained as being based on the strength of evidence for signal *nonexistence*. Specifically, the CL_S confidence interval only contains values of μ that have values of the SSR function above a threshold, $G^-(d,\mu) \geq \alpha_{cls}$, and excludes values where $G^-(d,\mu) < \alpha_{cls}$.

Recall from section 6.2.4 (page 295) that the CL_S method constructs a confidence-like interval for a given CL_S confidence level, $(1-\alpha_{cls})$. For a given value μ', this is done by enlarging the acceptance region for μ' (overcovering), with greater overcovering when there is less signal-background discrimination by the test statistic.

This is achieved by viewing the Neyman construction for μ' as a Neyman-Pearson hypothesis test (NPHT) with the signal model, $p(d;\mu')$, as the null hypothesis and the background-only model, $p(d;\tilde{\mu}_0)$, as the alternative hypothesis, and then lowering the critical value to $Q^*_{cls}(\mu')$ to accept more possible outcome (as described in equation 6.2.30 on page 291). The new critical value is found by solving for $Q^*_{cls}(\mu')$ that gives $\alpha_{cls} = \alpha^*/(1-\beta^*)$, where α^* is the test size using the critical value $Q^*_{cls}(\mu')$ (the left tail of the test statistic distribution integrated from $Q^*_{cls}(\mu')$ and using the signal model), and $(1-\beta^*)$ is the test power (the left tail probability using the background-only model).

Once d_{obs} is observed, one then compares (for a given μ') the observed test size, α_{obs}, and observed test power, $(1-\beta_{obs})$, to the CL_S confidence level, α_{cls}.[100] If $\alpha_{cls} \leq \alpha_{obs}/(1-\beta_{obs})$, then μ' is accepted in the CL_S confidence interval, otherwise it is rejected.

The observed test size and power are equivalent to the observed SCAP and BCAP since they are defined by the same left tail probabilities: $\alpha_{obs} = S(d_{obs},\mu')$ and $(1-\beta_{obs}) = B(d_{obs},\mu')$. Therefore,

$$G^-(d_{obs},\mu') = \frac{S(d_{obs},\mu')}{B(d_{obs},\mu')} = \frac{\alpha_{obs}}{(1-\beta_{obs})} \tag{6.4.10}$$

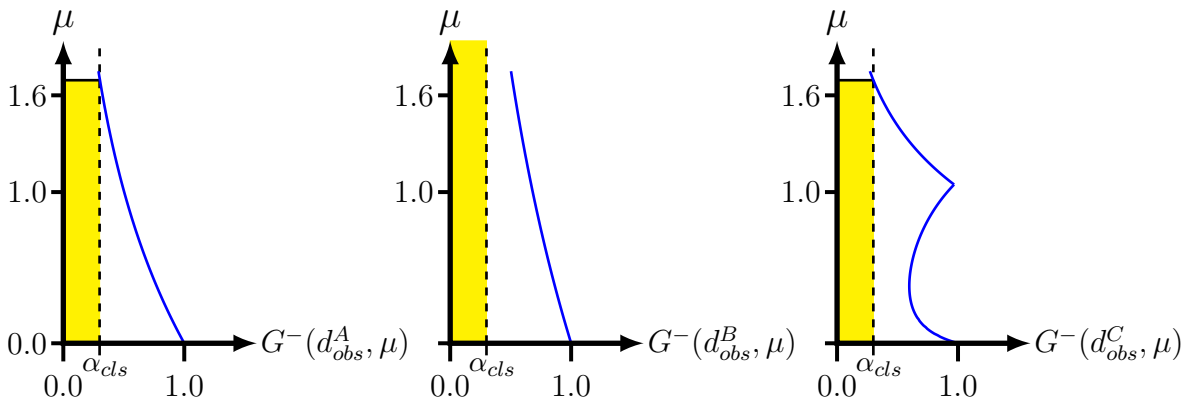
so that the observed SSR function for μ' may be used to determine if μ' is accepted in the CL_S confidence interval according to $\alpha_{cls} \leq G^-(d_{obs},\mu')$.

Figures 6.41 (a)–(c) demonstrate the use of the observed SSR function to obtain CL_S confidence intervals in the three experiments. These show the 70% CL_S confidence level by setting $\alpha_{cls} = 0.3$. Values of μ that are greater than $G^-(d_{obs},\mu)$ are accepted, and otherwise rejected. The vertical span of the yellow area shows the accepted values of μ within the 70% CL_S confidence interval (and show that the observed CL_S upper limit is larger than that of the non-CL_S method of 1.6).

For both experiments A and C, $G^-(d_{obs},\mu = 1.74) = 0.3$ so that the SSR function is smaller than $\alpha_{cls} = 0.3$ for $\mu > 1.74$. This gives the same 70% CL_S confidence interval of $[\,0,\,1.74\,]$ for both experiments. For experiment C, $G^-(d_{obs},\mu = 3.94) = 0.3$ giving the interval $[\,0,\,3.94\,]$ with more than twice the upper limit of the other two experiments.

[100] The left tail probability of the test statistic distribution integrating from $Q(d_{obs},\mu')$ gives α_{obs} when using the signal model, $p(d;\mu')$, and gives $(1-\beta_{obs})$ when using the background-only model, $p(d;\tilde{\mu}_0)$.

The motivation for the use of significance ratio functions began with the three hypothetical experiments having identical 90% confidence intervals (figure 6.29 on page 345), and the desire to quantify the evidence for each separate value of μ in a meaningful way. This same argument may be made for experiments using the CL_S method. The same 70% CL_S confidence intervals of examples A and C illustrate that it is possible for the CL_S method to deliver identical confidence intervals for two different experiments, and thus also motivates the need for the significance ratio functions.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.41: The CL_S confidence intervals at 70% CL_S confidence level (so that $\alpha_{cls} = 0.3$, shown with a dashed vertical line) for the three hypothetical experiments. The vertical span of the yellow areas show the range of μ values included in the CL_S confidence intervals: $[0, 1.74]$ for A and C, and $[0, 3.94]$ for B (yellow area extends vertically beyond the visible range).

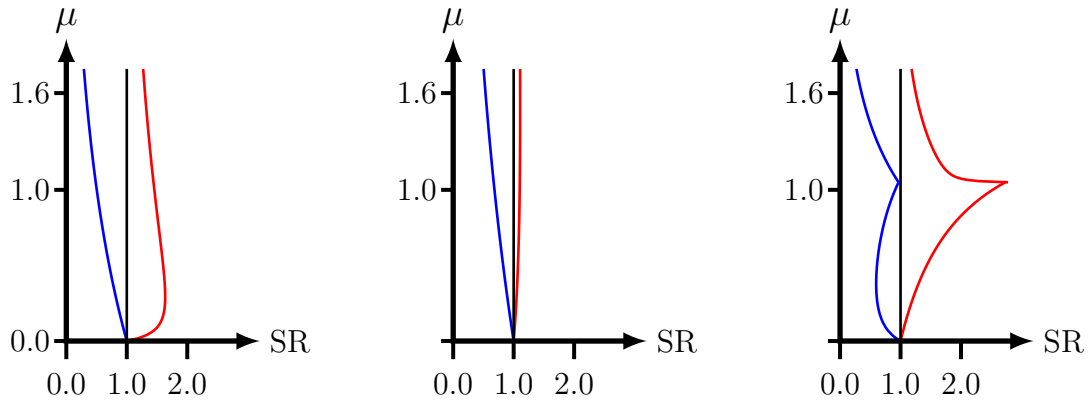
6.4.6.2. Combining the BSR and SSR Functions

The observed significance ratio functions quantify the strengths of two kinds of evidence provided by the observed outcome: evidence supporting signal existence at each value of μ with the BSR function, and evidence supporting signal nonexistence at each value of μ with the SSR function. By integrating over the upper tails of the signal and background-only model test statistic distributions, the BSR function compares how significantly the observed outcome deviates from the background-like outcomes. By integrating over the lower tails, the SSR function compares the deviation from the signal-like outcomes. For a given value μ', a large value of $G^+(d_{obs}, \mu')$ indicates stronger evidence of signal existence while a smaller value of $G^-(d_{obs}, \mu')$ indicates stronger evidence of signal nonexistence. Because $G^-(d, \mu) \leq 1$ and $G^+(d, \mu) \geq 1$, both types of evidence may be summarized simultaneously as is shown in figures 6.42 (a)–(c).

Figure 6.42 (a) shows that the evidence for signal nonexistence is stronger with increasing μ (blue line), and that the evidence for signal existence is strongest for small μ (red line). Taken together, these imply that if the signal does exist, then the evidence points to smaller values of μ below the nominal strength predicted by the new theory ($\mu = 1$) in experiment A.

Figure 6.42 (b) shows weakly increasing evidence for signal nonexistence with increasing μ (blue line), and that the evidence for signal existence is very weakly increasing with μ (red line). These simply indicate an overall lack of strong evidence for any conclusion to be drawn from experiment B.

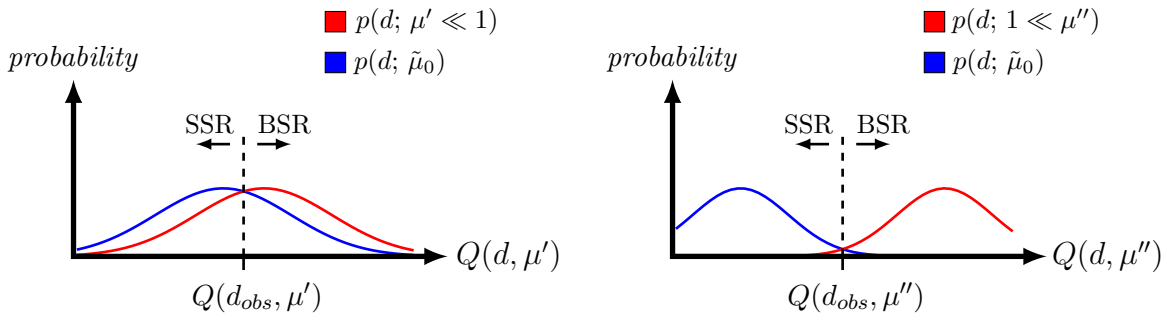
Figure 6.42 (c) shows both the weakest evidence for nonexistence and the strongest evidence for existence occurs at $\mu = 1$. Thus, experiment C demonstrates the strongest evidence for signal existence at the nominal strength of the three experiments (despite all three having identical 90% confidence intervals for μ as shown in figure 6.29 on page 345).



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.42: The observed background significance ratio (BSR) functions, $G^+(d_{obs}, \mu)$ (red lines), and the observed signal significance ratio (SSR) functions, $G^-(d_{obs}, \mu)$ (blue lines). The x-axis is the significance ratio (SR) for either of the functions. A vertical line denotes the value $SR = 1.0$, which separates the two functions.

It should be noted that it is possible to obtain strong evidence for both signal existence and nonexistence simultaneously for the same value of μ. This occurs when the test statistic provides large signal-background discrimination and the observed outcome strongly deviates from both the background-like outcomes and the signal-like outcomes. This situation is demonstrated in figure 6.43 (b) where the observed test statistic value lies in the extreme right tail of the background-only model test statistic distribution (blue line), and the extreme left tail of the signal model distribution. In this case, the value of the SSR function is nearly zero, indicating that d_{obs} is not at all signal-like, while the value of the BSR function will be very large, indicating that d_{obs} is not at all background-like. An outcome that is strongly neither signal-like nor background-like may indicate possible mismodeling of the experiment.



(a) Low discrimination (small μ'). (b) High discrimination (large μ'').

Figure 6.43: Two example comparing the significance ratio function values when the test statistic distributions have low vs. high signal-background discrimination. Each figure shows the test statistic distributions for the signal model (red line) and background-only model (blue line), each at a unique value of μ. (These figures are provided merely for explanatory demonstration.) In (a) there is very little signal-background discrimination because μ' is much smaller than 1. In (b) there is very large signal-background discrimination because μ'' is much greater than 1. In both (a) and (b), assume that the observed value of the test statistic lies between the two distributions (dashed line). The value of the SSR function is obtained from the left tail probabilities while that of the BSR function is from the right tails (shown with arrows). In (a), the SSR function value is slightly less than 1 (left red tail divided by slightly larger left blue tail), while the BSR function value is slightly greater than 1 (implying weak evidence for both existence and nonexistence). In (b), the SSR function value is nearly 0 (left red tail nearly equal to 0 divided by left blue tail nearly equal to 1), while the BSR function value is very large (implying strong evidence for both existence and nonexistence), and may indicate possible mismodeling.

6.4.7. Frequentist Interpretation of Significance Ratio Functions

Inductive reasoning from observations can never guarantee truth, but it does provide evidence in the search for truth. If the final inference result is simply a confidence interval at some specified confidence level, $(1-\alpha)$, then the only possible evaluation of a particular value μ' is whether or not it lies within the confidence interval: accept μ' or else reject it. The confidence interval does not provide a way to compare the evidence for differing values of μ beyond whether each is accepted or rejected. However, the Neyman confidence region construction (NCRC) is useful because it gives the long-run frequency $(1-\alpha)$ that confidence intervals constructed in this manner will correctly contain μ_{true}. The goal of this section is to provide such a useful frequentist interpretation to the significance ratio functions that describe the evidence for each value of μ that is provided by the observed outcome.

Birnbaum proposed that in Neyman-Pearson hypothesis testing (NPHT), the test size relative to test power, $\alpha/(1-\beta)$, could be used as a measure of the level of evidence provided by the test [264, 314]. If H_N is rejected (the observed value of the test statistic is smaller than the critical value), then smaller values of $\alpha/(1-\beta)$ indicate stronger evidence for H_A as opposed to H_N. This is because smaller $\alpha/(1-\beta)$ indicates that proportionately less of the distribution for H_N is contained in the region where H_A is accepted. This allowed Birnbaum to compare the level of evidence between two experiments by comparing the ratio of the test size to test power of each experiment.

One may apply Birnbaum's theory of evidence to the NCRC method by treating the construction of each acceptance region for each value μ' as a separate NPHT, where H_N is the signal model with strength μ', and H_A is the background-only model. The critical value, $Q^*(\mu')$ determines the test size, $\alpha_{\mu'}$, which is the lower tail probability of the signal model integrated from $Q^*(\mu')$, and the test power, $(1-\beta_{\mu'})$, the lower tail of the background-only model. For two different values, μ' and μ'', the acceptance region of each may be said to provide equal evidence if the critical value of each is chosen so that $\alpha_{\mu'}/(1-\beta_{\mu'}) = \alpha_{\mu''}/(1-\beta_{\mu''})$. This is in fact the logic of the $\mathrm{CL_S}$ method, which determines the critical value for each acceptance region of each value of μ by setting $\alpha_\mu/(1-\beta_\mu) = \alpha_{cls}$, i.e., all acceptance regions are given an equal level of evidence.

However, Birnbaum's theory does not differentiate the strength of evidence supplied by two possible observed outcomes, $d_{obs,1}$ and $d_{obs,2}$, of a single experiment where both observed outcomes would have resulted in rejection of H_N. The test size and test power of the NPHT are determined before any observation is made, and so the level of evidence is constant for any possible observed outcome falling within the rejection region.

The use of the actual observed outcome in determining the strength of evidence is central to the severe testing described in [260] (see also [315]). The severe test uses the observed test statistic value, $Q(d_{obs})$, to determine the strength of evidence provided by d_{obs}. The strength of evidence is determined by $\alpha_{obs}/(1-\beta_{obs})$, where the observed test size, α_{obs}, is the left tail probability of test statistic under H_N integrated from $Q(d_{obs})$, and the observed test power, $(1-\beta_{obs})$, is the left tail under H_A. If $\alpha_{obs,1}/(1-\beta_{obs,1}) < \alpha_{obs,2}/(1-\beta_{obs,2})$, then $d_{obs,1}$ provides stronger evidence for rejecting H_N. In other words, the farther $Q(d_{obs})$ is below the critical value, the stronger the evidence against H_N. This concept appears frequently in binary classification machine learning where a smaller classifier output value is considered more background-like.

The observed SSR function for NCRC is similar to the severe test that is applied to NPHT. After observing d_{obs}, the value $G^{-}(d_{obs}, \mu')$ is given by $\alpha_{obs}/(1-\beta_{obs})$ for a given μ' as described by equation 6.4.10.[101]

Although the severe test allows the comparison of the strength of evidence provided by different outcomes, it does not provide a frequentist interpretation. If one finds a small value of $\alpha_{obs}/(1-\beta_{obs})$ implying strong evidence against H_N, then one might wonder if such a small value was simply obtained by chance. Similarly, if one finds a small value of $G^{-}(d_{obs}, \mu')$ for the value μ', which indicates strong evidence for signal nonexistence at strength μ', then one might also wonder if such a small value was simply obtained by chance.

By examining the observed BSR and SSR functions, one obtains the strength of evidence for signal nonexistence and for signal existence, respectively, based on the observed outcome, d_{obs}. However, these statements of evidence are devoid of a proper frequentist interpretation unless one is able to examine their long run frequency properties, i.e., what the observed BSR and FSR functions would look like upon repeated sampling from the different model populations. Whereas the focus of Bayesian inference is on the single outcome, d_{obs}, frequentist inference considers the long run behavior of any statistic. For a frequentist, the long run behavior of the observed BSR and SSR functions are of primary importance, not just their values for the particular observed outcome.

Because the observed SSR function, $G^{-}(d_{obs}, \mu')$, gives the strength of evidence for signal nonexistence with strength μ', one must be cautious about falsely claiming nonexistence at strength μ'. Strong evidence indicated with a small value of $G^{-}(d_{obs}, \mu')$ is only meaningful if obtaining that value would be improbable if the signal did in fact exist with strength μ'. Therefore, one should analyze the long run frequency properties of $G^{-}(d, \mu')$ using the signal model, $p(d; \mu')$. This allows one to answer the question, "If the signal exists with strength μ', how *probable* is it to obtain a particular value for $G^{-}(d, \mu')$?"

The frequentist evaluation of $G^{-}(d, \mu')$ is performed using only one model, the signal model, $p(d; \mu')$, in order to examine the probability of falsely obtaining strong evidence for signal nonexistence. The signal model provides the probability distribution of $G^{-}(d, \mu')$, which one may use to compute useful statistics: central values, intervals, quantiles, etc. In keeping with the spirit of the NCRC method with its use of acceptance regions, and because there is only one model being evaluated, a Fisher significance test (FST) is used to interpret the significance of $G^{-}(d_{obs}, \mu')$. This FST is done using a predetermined significance level, α_{FST}, so that a p-value below α_{FST} is considered a significant deviation from the model (as will be illustrated below). If the Neyman construction of the confidence intervals for the values of μ is performed with a confidence level of $(1-\alpha)$, then it may be appropriate to use $\alpha_{FST} = \alpha$ for the FST when determining the significance of $G^{-}(d_{obs}, \mu')$ for any particular value μ'.

A demonstration of this using a significance level of $\alpha_{FST} = 0.10$ is shown in figure 6.44. For the value μ', an example probability distribution, $p\left(G^{-}(d, \mu'); \mu'\right)$ (the probability of

[101] If one uses the observed test statistic value as the critical value, then the observed test size, α_{obs}, and the observed test power, $(1-\beta_{obs})$, do not equal the test size, α that must be predetermined before measurement, nor the test power, β, that follows from α. The observed test power is reported in many fields of study, and it is argued in [313] that observed test power is misleading. However, the observed test size and observed test power here are used to allow comparison of different values of μ only as supplemental evidence to the observed confidence interval, not to replace it.

$G^-(d, \mu')$ using the signal model at strength μ') is shown with a Gaussian-shaped curve. The critical value, $G^-_{low}(\mu')$, is determined by finding the value of $G^-(d, \mu')$ that gives an upper tail probability of 0.90 (pink area). If the signal exists with strength μ', then the long run frequency of experiments obtaining $G^-(d_{obs}, \mu')$ smaller than $G^-_{low}(\mu')$ is only 0.10. If $d_{obs,1}$ is observed, then the value $G^-(d_{obs,1}, \mu')$ is small and is also less than $G^-_{low}(\mu')$. This may be interpreted as strong evidence for signal nonexistence at strength μ' that is also significant. If $d_{obs,2}$ is observed instead, then even if $G^-(d_{obs,2}, \mu')$ is small, the evidence for signal nonexistence at strength μ' cannot be claimed to be significant because the value $G^-(d_{obs,2}, \mu')$ is consistent with the expectations of the signal model.

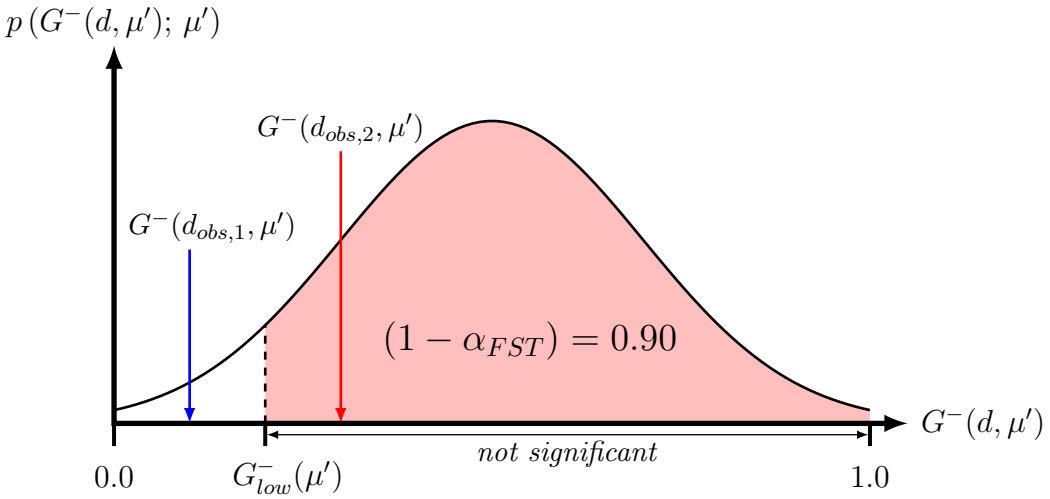
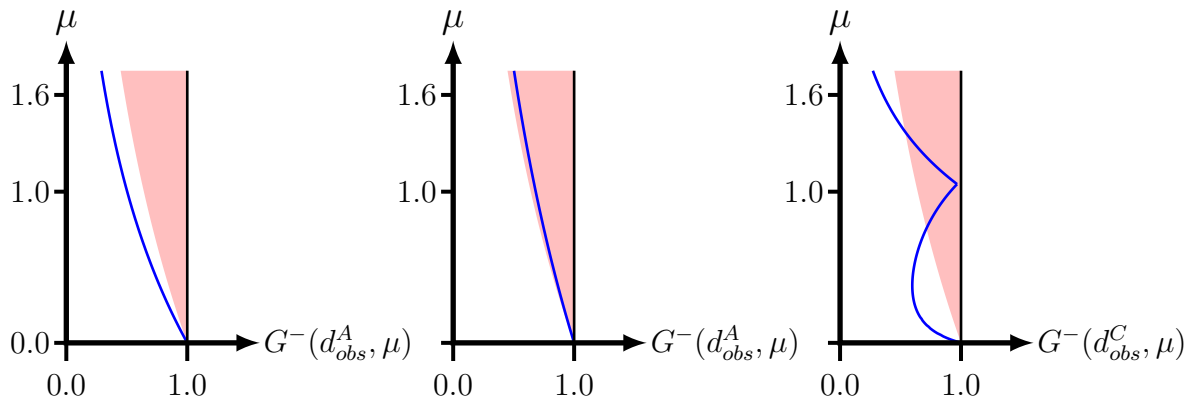


Figure 6.44: The Fisher significance test for the signal significance ratio function at the value μ'. The critical value is given by $G^-_{low}(\mu')$ (dashed line), which gives an upper tail probability of 0.90 (pink area). (This figure is provided merely for explanatory demonstration.) If $d_{obs,1}$ is observed, then $G^-(d_{obs,1}, \mu')$ falls below the critical value and is significant evidence for signal nonexistence at strength μ'. If $d_{obs,2}$ is observed, then $G^-(d_{obs,2}, \mu')$ does not provide significant evidence of nonexistence, even if $G^-(d_{obs,2}, \mu')$ is small.

This procedure may be performed for each value of μ in order to provide a frequentist interpretation of the entire observed SSR function, $G^-(d_{obs}, \mu)$. Figures 6.45 (a)–(c) show the results for the three experiments. In experiment A, the observed SSR function lies below the critical values for all but the smallest values of μ. This indicates that the evidence for signal nonexistence is significant for nearly all μ. In experiment B, $G^-(d^B_{obs}, \mu)$ is never significant for the values of μ shown in the figure. At a 10% significance level, one may not claim signal nonexistence for any value of μ based on the evidence provided by d^B_{obs}. In experiment C, $G^-(d^C_{obs}, \mu)$ is significant for small μ and large μ, but not near $\mu = 1$. The evidence provided by d^C_{obs} is stronger for signal nonexistence for small and large μ, but is weak and insignificant near $\mu = 1$.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.45: The Fisher significance tests at each value of μ for the signal significance ratio function. The blue line shows the observed SSR function, $G^-(d_{obs}, \mu)$. For a given value μ', the horizontal span of the pink area delineates the values of $G^-(d, \mu')$ that are not significant if observed (the left edge of the pink area shows the critical values, $G^-_{low}(\mu')$). For a given μ', if $G^-(d_{obs}, \mu')$ is less than $G^-_{low}(\mu')$, then the evidence for signal nonexistence is significant (with the blue line outside of the pink area). A vertical line denotes the value 1.0 for which $G^-(d_{obs}, \mu)$ always less than or equal to.

Because the observed BSR function, $G^+(d_{obs}, \mu')$, gives the strength of evidence for signal existence with strength μ', one must be cautious about falsely claiming signal existence at strength μ'. Strong evidence indicated with a large value of $G^+(d_{obs}, \mu')$ is only meaningful if obtaining that value would be improbable if the signal did not exist. Therefore, one should analyze the long run frequency properties of $G^+(d, \mu')$ using the background-only model, $p(d; \tilde{\mu}_0)$. This allows one to answer the question, "If the signal does not exist, how *probable* is it to obtain a particular value for $G^+(d, \mu')$?"

Figure 6.46 shows the FST used to evaluate the observed BSR function at the value μ'. This shows an example probability distribution, $p\left(G^+(d, \mu'); \tilde{\mu}_0\right)$ (the probability of $G^+(d, \mu')$ using the background-only model) as a Gaussian-shaped curve. The critical value, $G^+_{high}(\mu')$, is determined by finding the value of $G^+(d, \mu')$ that gives a lower tail probability of 0.90 (cyan area). If the signal does not exist, then the long run frequency of experiments obtaining $G^+(d_{obs}, \mu')$ larger than $G^+_{high}(\mu')$ is only 0.10. For this value μ', observing $d_{obs,1}$ gives insignificant evidence for signal existence, even if $G^+(d_{obs,1}, \mu')$ is large, while observing $d_{obs,2}$ provides significant evidence.

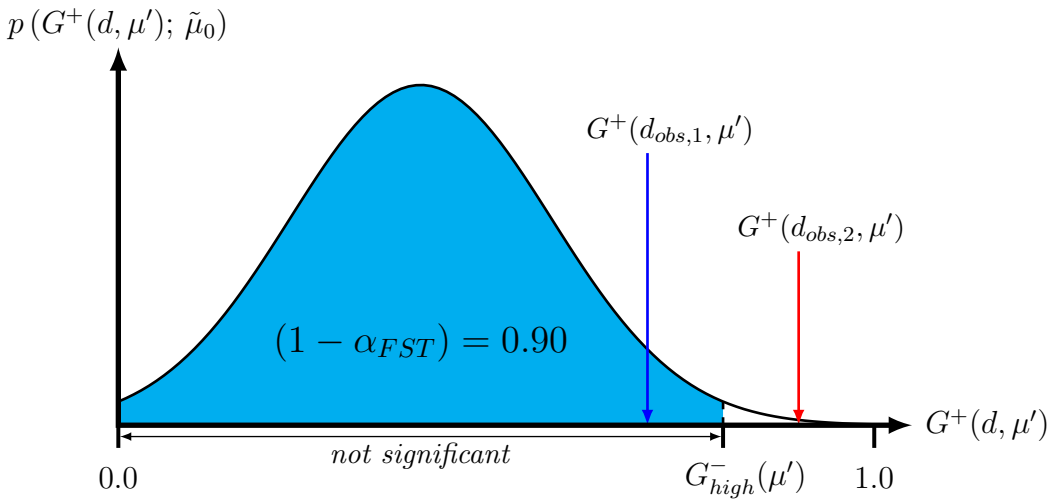
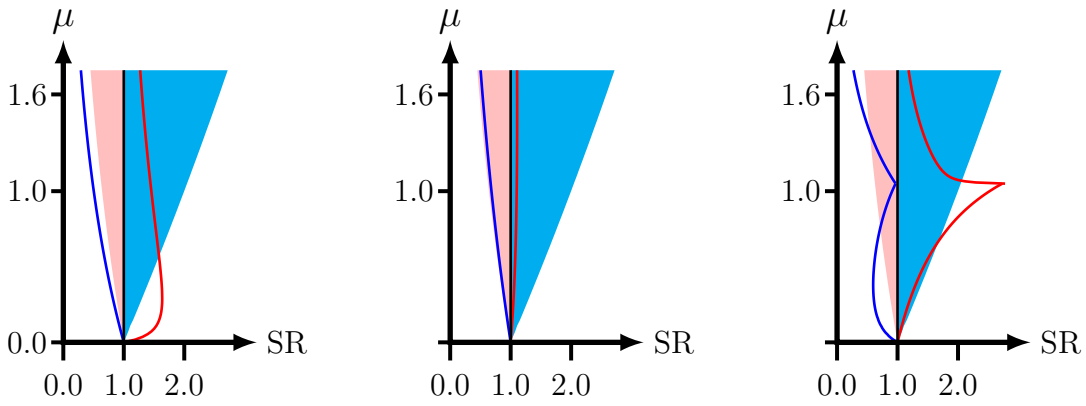


Figure 6.46: The Fisher significance test for the background significance ratio function at the value μ'. The critical value is given by $G^+_{high}(\mu')$ (dashed line), which gives a lower tail probability of 0.90 (cyan area). (This figure is provided merely for explanatory demonstration.) If $d_{obs,2}$ is observed, then $G^+(d_{obs,2},\mu')$ falls above the critical value and is significant evidence for signal existence at strength μ'. If $d_{obs,1}$ is observed, then $G^+(d_{obs,1},\mu')$ does not provide significant evidence of nonexistence, even if $G^+(d_{obs,1},\mu')$ is large.

Figures 6.47 (a)–(c) show the results of the FSTs at all values of μ for the three experiments. Because $G^-(d,\mu) \leq 1$ and $1 \leq G^+(d,\mu)$, the results for the SSR and BSR functions may be combined into a single summary.



(a) Experiment A (b) Experiment B (c) Experiment C

Figure 6.47: The Fisher significance tests at each value of μ for the significance ratio functions. The blue line shows the observed SSR function, $G^-(d_{obs},\mu)$, and pink area the range of insignificant values. The red line shows the observed BSR function, $G^+(d_{obs},\mu)$, and the cyan area the range of insignificant BSR function values. A vertical line at 1.0 separates the two types of evidence: evidence for signal nonexistence on the left, and for existence on the right.

The results shown in figures 6.47 (a)–(c) summarize the evidence for both signal nonexistence and signal existence at each value of μ for the three hypothetical experiments. The final analysis shows that d^A_{obs} provides significant evidence of signal existence only for small μ, and evidence of nonexistence for larger μ, including the nominal signal strength $\mu = 1$. Meanwhile, d^B_{obs} provides weak and insignificant evidence for either conclusion for all values of μ. Lastly, d^C_{obs} provides strong and significant evidence of signal nonexistence for small and large values of μ, and strong and significant evidence for signal existence at the nominal signal strength.

While all three experiments share the same 90% confidence interval for μ, they each have different confidence-acceptance profiles (CAPs). Although these CAPs are only suggestive of the evidence for different μ (much as fiducial and confidence distributions are suggestive), it is the significance ratio (SR) functions that may be derived from the CAPs that give a quantitative measure of evidence. Furthermore, the simplicity of the SR functions allows for the computation of the long run frequency of their values, which provides a frequentist interpretation of their significance.[102]

[102] One may wonder why α_{obs}/β_{obs} is not used as a measure of evidence, especially since α_{obs} (left tail) quantifies a deviation from signal expectations (not $1 - \alpha_{obs}$) and β_{obs} (upper tail) quantifies a deviation from background-only expectations (not $1 - \beta_{obs}$). However, α_{obs}/β_{obs} does not have desirable frequentist properties: the signal and background-only model expectations overlap substantially. (I.e., if α_{obs}/β_{obs} was plotted as in figure 6.47, the pink and cyan areas would greatly overlap.)

6.4.8. The Utility of Statistical Eduction via Confidence

In his critique of the strict interpretation of the Neyman construction of confidence regions [314], Birnbaum argued that Neyman and Pearson had been somewhat disingenuous over the years in the strict accept vs. reject interpretation of confidence regions, but that instead one could use the Neyman construction in some heuristic sense to provide levels of evidential support.[103] Egon Pearson admitted that science had suffered from the often "absurd extent" to which people applied the logic of the strict interpretation of confidence [314]. Birnbaum conjectured that the lack of a clearly defined evidential interpretation had led many scientists to invent their own interpretations. Perhaps this explains the poor understanding of confidence intervals reported in [301], where researchers and students performed equally poorly when quizzed about the correct interpretation of a single observed confidence interval.[104]

The use of likelihood offers one solution to evaluating evidence for different values of μ. The consensus recommendation produced at the CERN Workshop on Confidence Limits in 2000 [318] was that the observed likelihood function should be published.[105] However, a single likelihood function does not provide the information needed for frequentist inference using the Neyman construction.

If a test statistic is used that produces confidence intervals that satisfy the frequentist likelihood property (FLP, defined on page 348), such as Q_{AZ1}, then an evidential statement is available for a single observed confidence interval: the values of the parameter of interest within the reported confidence interval are the most *likely*.[106]

However, likelihood is a weaker form of probability [294], and it is probability statements that are truly desired in decision making.[107] An inference result in the form of observed SSR and BSR functions obtained via the Neyman construction *educes* such a probability statement from the observed outcome.[108]

---

[103] The strict interpretation allows for probabilistic statements to be made about the *process* of constructing confidence regions, but when a single confidence region is selected based on the actual measurement, all that may be said is that the values of the parameter of interest are grouped into "accepted" or "rejected" categories with no probabilistic statement available to describe these two groups. Also see [316].

[104] Essentially, there is no correct probability interpretation of a single observed confidence interval per the strict interpretation (i.e., without Bayesian assumptions [317]). This led one statistician collaborating with particle physicists to complain that perhaps an observed confidence interval is to be reported "simply as a ritual activity" [271].

[105] The combined CDF and D0 Higgs search reported in [283] provides a result using the test statistic equivalent to $Q = p(d;\, \mu = 0)/p(d;\, \mu = 1)$ that shows the observed value of Q_{obs}, the expected value assuming signal existence, $E[Q_1]$, and its distribution assuming signal nonexistence, $p(Q_0)$. The existence of such results demonstrates the need in particle physics for more than simple confidence intervals.

[106] One caveat is that satisfying the FLP does not guarantee that the values of μ at one end of the confidence interval have equal likelihood to those at the other end. This means it is possible for the values of μ just outside one end of a confidence interval may have higher likelihood than those just within the other end. Nevertheless, the previous statement regarding the FLP is correct in spirit.

[107] For instance, see [319] where it is recommended that both frequentist and Bayesian methods are useful in particle physics, and [320] where the deficiencies of confidence intervals used in particle physics is discussed.

[108] Compare the definitions of *induce* and *educe*. Induce, from the Latin "to lead in", means to bring about a change. Educe, "to lead out", means to bring out a latent property. Related to statistical inference, an induced result indicates a substantial change in understanding of the causation of the observed data, while an educed result suggests weighing the evidence of different causes without reaching a definitive conclusion.

The three examples of figures 6.30 (a)–(c) were examined to provide motivation for the development of evidential confidence methods in future particle physics experimentation. It is usually the case that hints of discovery long precede definitive discovery of a new particle. Providing more detailed information regarding what these hints imply may help guide future efforts.

The process of science is essentially a cycle of inductive reasoning and deductive reasoning. In the inductive stage, real-world data is obtained and the most plausible theory induced from it. Then in the deductive stage, the logical consequences of the theory are deduced so that predictions are established for possible falsification in the next stage of induction. This is what occurs over the long term of scientific progress.

However, it is often the case in the short term that decisions must be made before enough data is collected to construct a definitive inductive result. The seldom used term "eduction" implies a weaker form of induction. It is the inductive process of obtaining weak evidence from data without drawing definitive conclusions from the data. In fields where data collection and analysis occur over a long time period, eduction may be useful in helping make timely decisions based on limited data. Such a scheme is illustrated in figure 6.48.

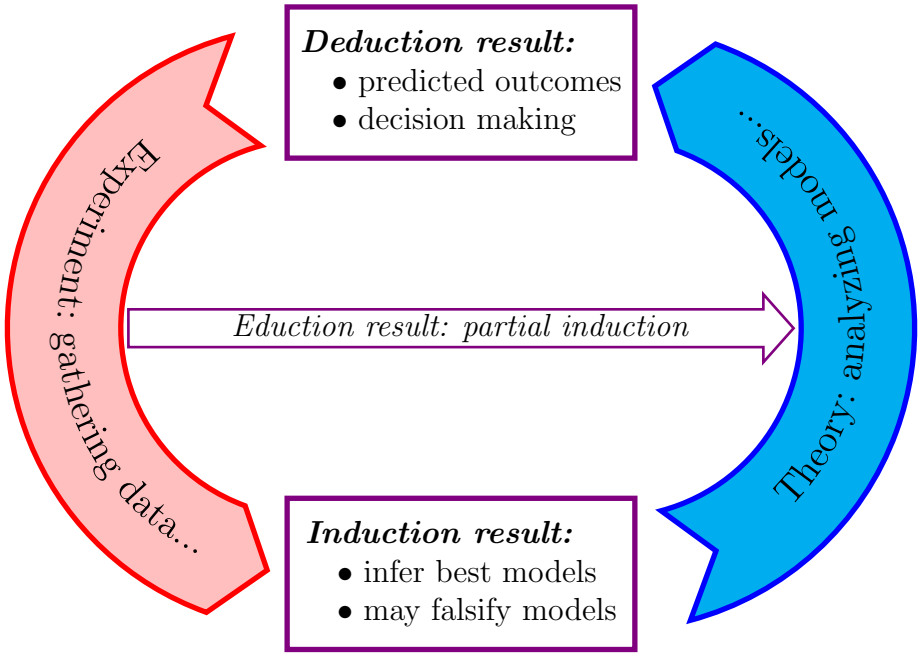


Figure 6.48: The scientific process may be viewed as a cycle of experiment (induction) and theory (deduction), where the results of one inform the other. Theory provides model predictions to be tested by experimentation. Deductive predictions based on models are also what inform decision making. Experimentation involves collecting enough measured data with which to make definitive inferences, which either enhance or falsify current models. When a long period of time is needed to gather enough data, a weaker inference result may be educed from incomplete data to inform decision makers in the present.

In 2011, the search for the Higgs boson was terminated at the Tevatron accelerator in Illinois. There was some evidence of the presence of the Higgs with an invariant mass in the neighborhood of 125 GeV in the data collected there [283]. Figure 6.29 (a) (page 345) is qualitatively similar to the results those reported results. Even if the Tevatron had remained in operation, it is unlikely that the Higgs could have been discovered there before discovery was achieved at the LHC. Nevertheless, it may have been useful and informative to examine purely frequentist distribution-like functions made from that data, such as confidence-acceptance profiles and significance ratio functions, even if simply to better inform the efforts at the LHC.

Figures 6.47 (a)–(c) demonstrate that it is possible to provide frequentist interpretations of the observed SSR and BSR functions in order to evaluate the strength of evidence for both the signal nonexistence and existence for each value of μ. Thus, frequentist statistics can provide distribution-like functions that quantify the strength of evidence for different values of μ. In other words, frequentist statistics can provide more inferential information than simple accept vs. reject statements.[109] The brief survey provided here demonstrating the utility of confidence-acceptance profiles and significance ratio functions deserves further investigation, such as examining bias and efficiency properties, especially when recognizing the limitations of current confidence methods [320].[110]

[109] The SSR and BSR functions developed here by extending the NCRC method are an attempt to bridge the gap between inductive inference and inductive behavior [268, 321], a source of disagreement between Fisher and Neyman. This are similar in spirit to the severe testing of [260], but this work also gives a frequentist interpretation by providing statements of significance for the observed SSR and BSR functions.

[110] Although these methods have been developed here with the multivariate Poisson model used for signal searches, they could easily be applied to other models used for parametric inference. The statistical testing of medical treatment clinical trials is one such example. In clinical trials, one must balance the goal of determining the efficacy of a new treatment while also determining the level of safety provided by the treatment [322]. During a clinical trial, one might use frequentist eduction to examine the current evidence of safety and efficacy before a definitive induction is available. The use of Bayesian parametric inference in medical research has grown in recent years, perhaps due the ease of interpretation of the resultant posterior distribution [323, 324]. Weibul, exponential, log normal, and gamma distributions are often used in parametric survival analysis, and the methods developed here may be extended to these statistical models to provide comparative evidence for the different parameter values of the survival and hazard functions, which should aid in decision making.

6.4.9. Combining Confidence-Acceptance Profiles via Meta-analysis

One advantage of Bayesian inference is the ease with which additional information may be incorporated. The resultant posterior distribution of one analysis may become the prior distribution for the next. For frequentist interval estimation, the ideal way to combine the confidence results of two independent experiments measuring the same parameter, μ, is to construct a total statistical model by multiplying the separate models, and reperform the Neyman construction using both sets of the original data. However, the original data may not be available so that a *meta-analysis* must be conducted.

Methods for combining confidence distributions have been developed [325, 326], but this requires that both experiments are able to provide confidence distributions which is not always possible (see section 6.4.2 on page 346). However, it is always possible to construct confidence-acceptance profiles (CAPs). This section explains how one may combine the CAPs of two or more experiments via meta-analysis, without the original data nor any knowledge of the underlying statistical models used in the experiments.[111] The notation $A(d, \mu)$ is used here to represent either the signal CAP, $S(d, \mu)$, or background CAP, $B(d, \mu)$, as the formulation applies equally to either.

If only the observed CAP is provided for each experiment, then very little information is given describing the distributions of the test statistics, $Q(d, \mu)$, for each value of μ. Only the quantile of the observed test statistic, $A(d_{obs}, \mu)$ is given for each μ. However, recall from section 6.2.3.1 (page 279) that the test statistic is actually a local ordering rule used in the Neyman construction that ideally is designed to return high values for outcomes that provide the most evidence for μ. The CAP, $A(d_{obs}, \mu)$, thus shows within which quantile d_{obs} was located for each value of μ. For example, if $A(d_{obs}, \mu') = 0.82$, then it is known that for μ', 82% of the outcomes (by probability mass) had test statistic values equal to or smaller than $Q(d_{obs}, \mu')$.

In the absence of any additional knowledge, the observed CAP provides a measure of the strength of evidence given by the original test statistic. Define the meta-analysis test statistic simply as $Q_{meta}(d, \mu) = A(d, \mu)$. In other words, the meta-analysis test statistic value of d is the quantile rank of the original test statistic value. Essentially, the meta-analysis test statistic replaces the original test statistic, but preserves the same local ordering.

With Q_{meta}, one immediately knows the quantile rank of the meta-analysis test statistic for any outcome d', $p(q \leq Q_{meta}(d', \mu); \mu) = A(d', \mu)$. For example, when one is performing the Neyman construction for the value μ', if d' had a test statistic value, $Q(d', \mu')$, that was in the 56^{th} percentile of the test statistic distribution, then $Q_{meta}(d', \mu')$ will be in the 56^{th} percentile of the meta-analysis test statistic distribution. Knowledge of the span of possible outcomes d and even of the original statistical model is unnecessary because the observed CAP provides the test statistic quantile of d_{obs}, and thus one knows the quantile rank of the observed outcome test statistic, $p(q \leq Q_{meta}(d_{obs}, \mu); \mu) = A(d_{obs}, \mu)$.

With the meta-analysis test statistic defined, one can now consider combining the results of multiple experiments. If $\mu' = \mu_{true}$, then $A(d, \mu')$ is a random variable from a uniform distribution, $U[0, 1]$. This is a simple property of quantiles. For example, the probability of obtaining a value of $A(d, \mu')$ in any *quintile* is 20%, and the probability of $A(d, \mu')$ to be at

[111] A similar approach is given in [327], however they make the assumption that the combined confidence level must always be less than any component confidence level. A much different approach is taken here.

or below the value 0.42 is 42%. Therefore, combining two CAPs, $A_1(d_1, \mu')$ and $A_2(d_2, \mu')$, where d_1 is a possible outcome of the first experiment and d_2 a possible outcome of the second, is equivalent to combining two random variables each from a uniform distribution, $U[0, 1]$.

For values of the CAPs drawn at random, $a_1 \in A_1(d_1, \mu')$ and $a_2 \in A_2(d_2, \mu')$, the probability distribution of their average, $a = \frac{1}{2}(a_1 + a_2)$, is given by

$$p(a) = \begin{cases} 4a & \text{if } 0 \leq a \leq 0.5 \\ 4 - 4a & \text{if } 0.5 < a \leq 1.0 \end{cases}, \tag{6.4.11}$$

so that the cumulative distribution of a is

$$p(a' \leq a) = \begin{cases} 2a^2 & \text{if } 0 \leq a \leq 0.5 \\ 4a - 2a^2 - 1 & \text{if } 0.5 < a \leq 1.0 \end{cases}. \tag{6.4.12}$$

These are shown in figures 6.49 (a) and (b).

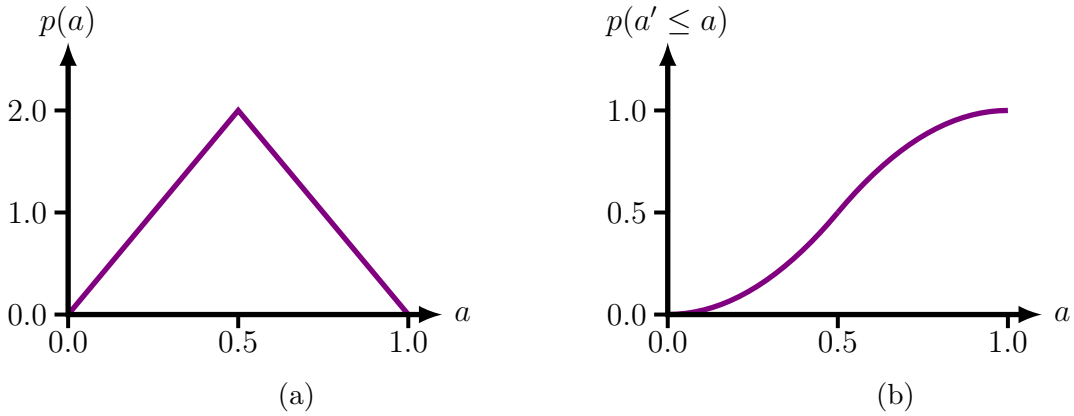


Figure 6.49: The probability distribution (a) and cumulative distribution (b) of the meta-analysis test statistic used for the combination of two experimental results.

Because the cumulative distribution of a is equivalent to the quantile of a, the meta-analysis CAP may be defined as this cumulative distribution, $A_{meta}(\mu) = p(a' \leq a)$. Therefore, given $a_{obs,1} = A_1(d_{obs,1}, \mu)$ and $a_{obs,2} = A_2(d_{obs,2}, \mu)$, one immediately knows the resultant CAP of the meta-experiment for any value of μ,

$$A_{meta}(\mu) = \begin{cases} 2a_{obs}^2 & \text{if } 0 \leq a_{obs} \leq 0.5 \\ 4a_{obs} - 2a_{obs}^2 - 1 & \text{if } 0.5 < a_{obs} \leq 1.0 \end{cases}, \tag{6.4.13}$$

where

$$a_{obs} = \frac{1}{2}\left(A_1(d_{obs,1}, \mu) + A_2(d_{obs,2}, \mu)\right). \tag{6.4.14}$$

Therefore, from the average of two CAPs of two independent experiments, one obtains a combined CAP for the meta-experiment.

The averaging of CAP values is justified by the design of the underlying test statistics,

$Q_1(d_{obs,1}, \mu)$ and $Q_2(d_{obs,1}, \mu)$, used in the Neyman constructions of the two experiments. High values of $Q_i(d_{obs,i}, \mu)$ provide stronger evidence of signal existence, and these results in larger values for $A_i(d_{obs,i}, \mu)$. If both experiments indicate a medium level of evidence, $A_1(d_{obs,1}, \mu) = 0.50$ and $A_1(d_{obs,2}, \mu) = 0.50$, then the resultant meta-analysis should reflect this, $A_{meta}(\mu) = 0.50$. However, if the first experiment provides strong evidence while the second weak, $A_1(d_{obs,1}, \mu) = 0.90$ and $A_1(d_{obs,2}, \mu) = 0.10$, then one must resolve this conflicting evidence. The simplest resolution is for the conflicting evidence to offset one another so that one still has $A_{meta}(\mu) = 0.50$. This is reflected in the probability distribution of the meta-analysis test statistic shown in figure 6.49 (a) where the largest densities are near $a = 0.5$. The value $a = 0.5$ is obtained whenever $A_1(d_{obs,1}, \mu) = x$ and $A_1(d_{obs,2}, \mu) = (1 - x)$ for $x \in [0, 1]$.

However, for a_{obs} to reach a value of $\sim$0.0 or $\sim$1.0, the evidence from both experiments must be in agreement, and this is much less probable if $\mu = \mu_{true}$ since $a_{obs,1}$ and $a_{obs,2}$ are each drawn from uniform distributions. The averaging of CAP values is useful because when $\mu \neq \mu_{true}$, the distributions of the test statistics of the experiments, $Q(d, \mu)$, are shifted to lower values so that it is more probable to receive smaller values for $A(d, \mu)$. Therefore, when $\mu \neq \mu_{true}$, one expects to obtain smaller values of $A_{meta}(\mu)$ than would be implied by the density of figure 6.49 (a), and it would be more improbable to obtain large values of $A_{meta}(\mu)$.

For N_{meta} experiments to be combined, the probability distribution of their average is more complicated. The exact density is derived with equation 20 of [328], but is approximately Gaussian with as little as $N_{meta} = 3$. Also note that the weighted sum, $w_1 A_1(d_1, \mu) + w_2 A_2(d_2, \mu')$, retains the probability distribution of equation 6.4.11. This allows one to weight the experiments according to some criteria. Therefore, the most general meta-analysis CAP is defined as

$$A_{meta}(\mu) = \sum_{i=1}^{N_{meta}} w_i(\mu) A_i(d_{obs,i}, \mu), \tag{6.4.15}$$

where $\sum_i w_i(\mu) = 1$ are the weights assigned to each of the combined experiments, and where these weights may also be functions of μ.

6.4.9.1. Combining CAPs: Repeated Trials

Some examples are provided to illustrate the use of CAPs in meta-analysis. Two signal models are investigated. The first is the discrete univariate Poisson signal model,

$$p_{pois}(d;\,\mu) = \frac{(\mu\, s + b)^d}{d!} e^{-(\mu\, s + b)}, \tag{6.4.16}$$

which has increasing variance with increasing μ, and the second is a continuous univariate truncated Gaussian signal model,

$$p_{gaus}(d;\,\mu) = \frac{1}{\mathcal{N}(\mu)\sqrt{2\pi}} e^{-\frac{1}{2}(\mu\, s + b - d)^2}, \tag{6.4.17}$$

which has constant unit variance. Both models are restricted to nonnegative outcomes d so that $\mathcal{N}(\mu)$ is the normalization constant due to truncation (that depends on μ). The signal strength parameter is restricted to be physical, $\mu \geq 0$. Three test statistics were used for the examples, two that provide two-sided confidence intervals,[112] Q_{FC} and $Q_{AZ} \equiv Q_{inc} \cdot Q_{FC}$ and one the provides one-sided upper limits Q_{D0}.

If an experiment is repeated many times, one expects the combined evidence of the multiple trials to provide stronger evidence for the true underlying process and less for the false. An example is provided for both the Poisson and Gaussian models by using $b = 5.0$ as the expected background count and $s = 4.0$ as the expected signal count at nominal signal strength, $\mu = 1.0$.

Figures 6.50 (a)–(f) show the signal CAPs produced for both models using the three test statistics. Rather than show CAPs with decile or ventile precision, they are shown in terms of σ-probabilities which correspond to the one-tailed p-values of a standard normal distribution. The x-axis is labeled d_1 for the possible outcomes indicating that this is the result for a single experiment. Note that CAPs could possibly be used for discovery purposes if an outcome d_{obs} provided a 5σ confidence interval that did not include $\mu = 0$. The results for the Poisson model (left column figures) show that confidence intervals widen with increasing μ due to increasing model variance while this is not the case for the Gaussian model with constant variance (right column).

[112] Experimentation not reported here shows that while Q_{rep} is necessary for inclusion with Q_{D0} and Q_{AT} to achieve desirable confidence interval properties, this is not necessarily so with Q_{FC}.

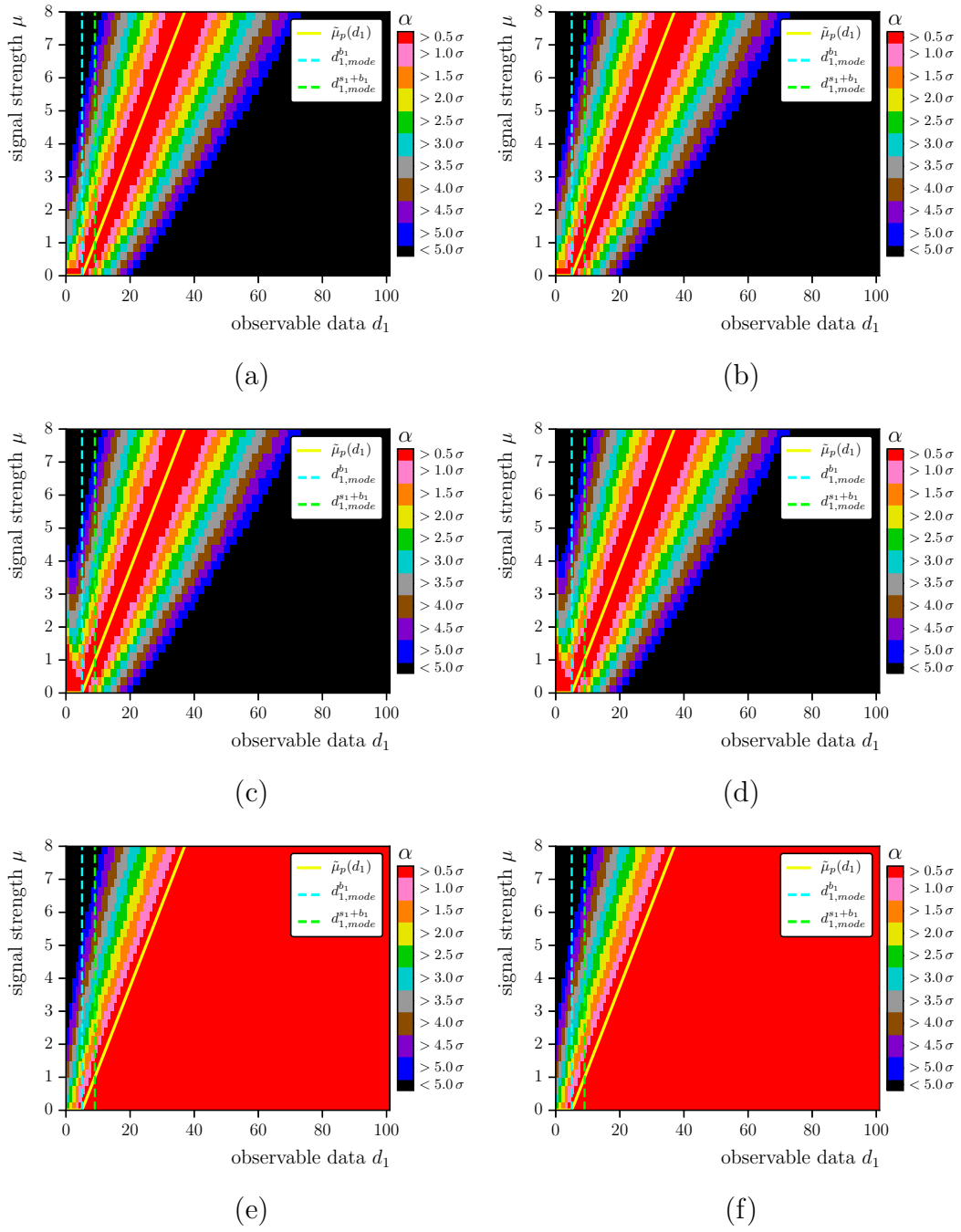


Figure 6.50: Meta-analysis – The signal confidence-acceptance profiles (SCAPs) using $b = 5$ and $s = 4$ as the expected values for both the Poisson and Gaussian models. Rather than show SCAPs with ventile precision, they are shown in terms of σ-probabilities (corresponding to the one-tailed p-values of a standard normal distribution). Using Q_{FC}, (a) Poisson and (b) Gaussian. Using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$, (c) Poisson and (d) Gaussian. Using Q_{D0}, (e) Poisson and (f) Gaussian.

Ten random samples are sequentially taken from each model distribution so that ten observed signal CAPs are provided as ten separate experiments, i.e., ten random vertical sections are obtained from each of figures 6.50 (a)–(f). Rather than combine these ten results using an ideal model, which would only be possible if the original data was known, they are combined via meta-analysis as per equation 6.4.15 (unweighted). This procedure is done once assuming signal nonexistence ($\mu = 0$), and then repeated assuming signal existence with nominal signal strength ($\mu = 1.0$).

Figures 6.51 (a)–(d) show the meta-analysis results of combining the observed signal confidence-acceptance profiles (SCAPs) from repeated experiments with the Poisson model using Q_{FC} and $Q_{AZ} = Q_{inc} \cdot Q_{FC}$. Because these are derived from the ten *observed* outcomes, $d_{obs,1}$ through $d_{obs,10}$, the possible outcomes are no longer plotted as a separate axis (the ten randomly sampled observed counts are provided in the top left of each figure). The resultant observed SCAP from the first experiment is shown with a solid gray line, then the second experiment is combined with the first resulting in the meta-analysis SCAP shown with a solid black line, and so on until all 10 experimental results are combined into a final meta-analysis SCAP shown with a solid red line. The x-axis shows the range of μ and the y-axis shows one minus the confidence level. One may obtain the upper and lower limits of the confidence interval of a given SCAP by selecting any desired value for α, and finding where that SCAP intersects this value (so that a confidence interval is the horizontal span between the lower and upper limits). For example, a red dashed line denotes $\alpha = 0.05$ so that the final meta-analysis SCAP (solid red line) in figure 6.51 (a) has the 95% confidence interval $[\,0,\, 0.8\,]$ (the red solid line intersects the dashed red line at 0.8 for the upper limit). Gray horizontal dashed lines denoting the σ-probabilities are provided for convenience. A vertical gray line at $\mu = 1.0$ denotes the nominal signal strength, where it is often useful to know the confidence level at which $\mu = 1.0$ is excluded from a confidence interval (when the theory predicting signal existence at the nominal strength is rejected).

Figure 6.51 (a) shows the result using Q_{FC} assuming $\mu_{true} = 0.0$. This shows that when $\mu_{true} = 0$ and with only one trial (solid gray line), the resultant confidence intervals are quite long (horizontally). (The values obtained in random sampling are listed in the top left so that $d_{obs,1} = 7$, which is above the background-only expectation of 5.) The gray line indicating the SCAP of experiment 1 gives the 95% confidence interval $[\,0,\, 2.25\,]$. When this result is combined via meta-analysis with $d_{obs,2} = 4$ (solid black line), the 95% confidence interval shrinks to $[\,0,\, 1.4\,]$. After combining 5 trials (lime line), the 95% confidence interval excludes $\mu = 1.0$. After all ten trials (solid red line), $\mu = 1$ is excluded with (1-3σ) confidence.

Meanwhile, figure 6.51 (b) shows the result using Q_{FC} when $\mu_{true} = 1.0$. After 10 trials (solid red line), the 95% confidence interval is $[\,0.4,\, 2.2\,]$, correctly including $\mu = 1$, while $\mu = 0$ is excluded with more than (1-5σ) confidence (the typical level needed to claim discovery in experimental particle physics).

Figures 6.51 (c) and (d) show the results using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$ which are similar to those of Q_{FC}.

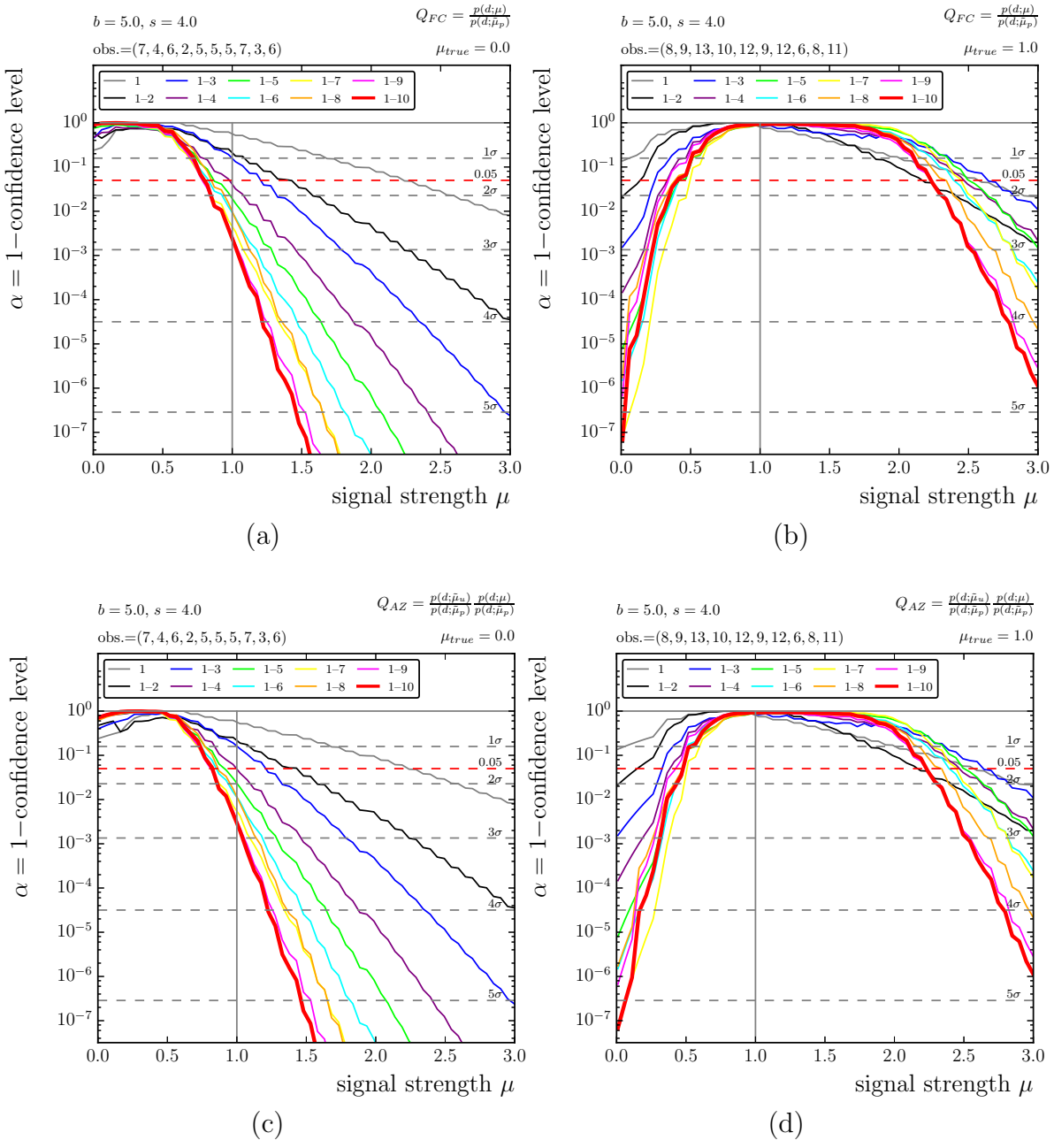


Figure 6.51: Meta-analysis – Progressively combining the SCAPs of 10 repeated samples from the Poisson model. A horizontal dashed red line indicated the 95% confidence level, and horizontal dashed gray lines show other useful confidence levels. Using Q_{FC}, (a) sampling observed outcomes when $\mu_{true} = 0$ and (b) sampling when $\mu_{true} = 1.0$. Using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$, (c) $\mu_{true} = 0$ and (d) $\mu_{true} = 1.0$.

Figures 6.52 (a)–(d) show the results for the continuous Gaussian model using Q_{FC} and Q_{AZ}. The variance of the Gaussian model is fixed at 1.0, which is smaller than the Poisson model variance (minimum of 4 when $\mu = 0$). This causes confidence intervals to be much shorter for the Gaussian model than the Poisson. Note that the Gaussian model is not well-suited for a typical particle search, but is examined here to demonstrate the utility of meta-analysis using CAPs for an alternate signal model.

Figure 6.52 (a) shows the results of meta-analysis combination of results obtained using Q_{FC} when $\mu_{true} = 0.0$. After ten trials (solid red line), the 95% confidence interval is $[0, 0.2]$, with exclusion of $\mu = 1$ well above the (1-5σ) confidence.

Figure 6.52 (b) shows the results using Q_{FC} when $\mu_{true} = 1.0$. After ten trials (solid red line), the meta-analysis SCAP, $A_{meta}(\mu)$, converges around μ_{true} to $[0.75, 1.25]$. However, the first two trials, $d_{obs,1} = 11.2$ and $d_{obs,2} = 7.7$, provide conflicting evidence for different values of μ, resulting in that meta-analysis SCAP (solid black line) with two well-defined maxima. The 95% confidence region for these two trials is disjoint, $[0.3, 0.9] \cup [1.3, 1.9]$, which excludes $\mu_{true} = 1.0$. (A comparison of the meta-analysis combination to the ideal combination will be provided shortly.) Also, $A_{meta}(\mu)$ does not always reach a maximum of 1.0 ("10^0" on the y-axis) for the Gaussian model, and experimentation has shown that it is possible for the maximum of $A_{meta}(\mu)$ to be quite low if the observed outcomes convey strongly conflicting evidence, although this is rare.

Figures 6.52 (c) and (d) show the results using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$ which are similar to those of Q_{FC}.

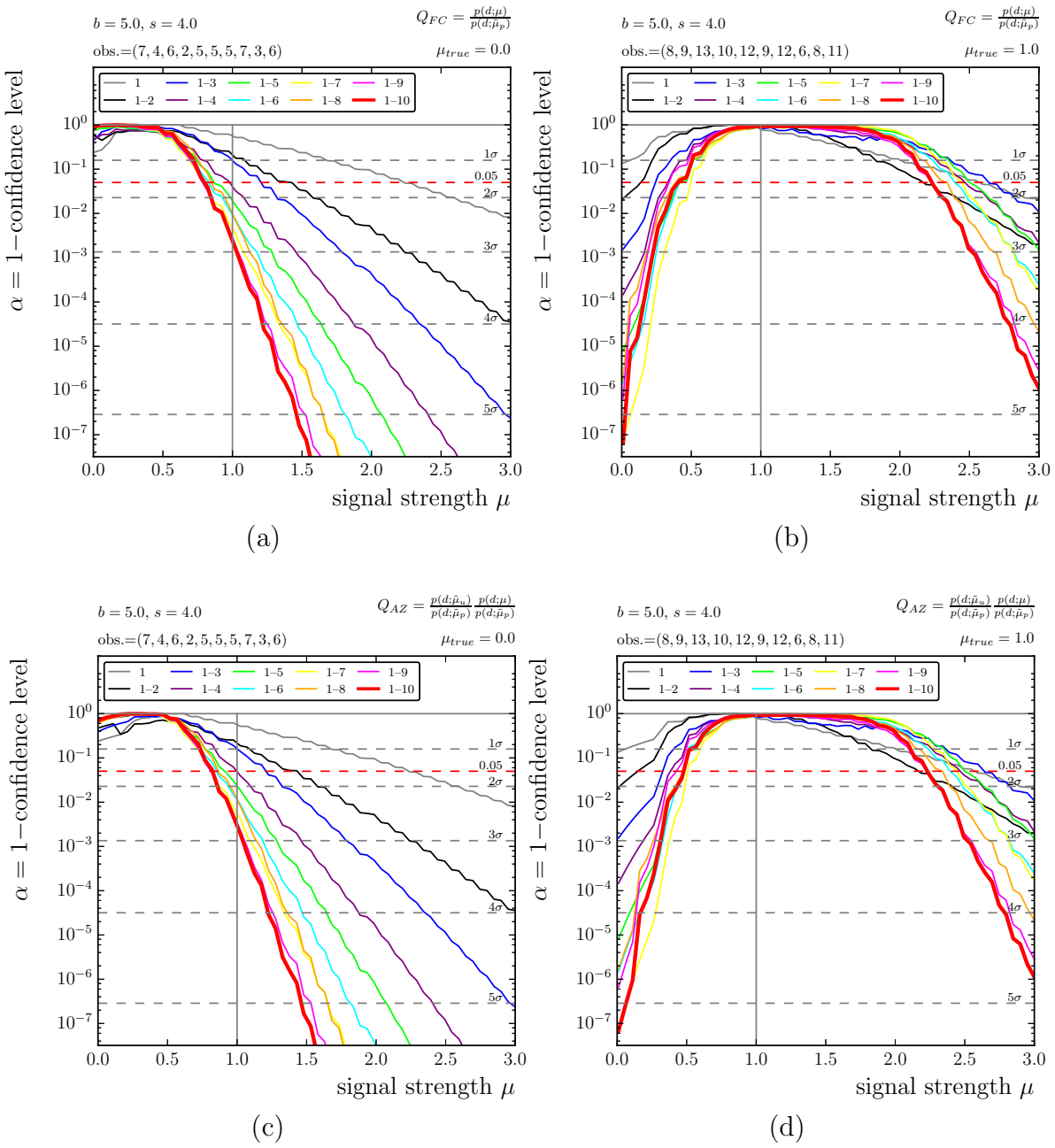


Figure 6.52: Meta-analysis – Progressively combining the SCAPs of 10 repeated samples from the Gaussian model. Using Q_{FC}, (a) sampling observed outcomes when $\mu_{true} = 0$ and (b) sampling when $\mu_{true} = 1.0$. Using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$, (c) $\mu_{true} = 0$ and (d) $\mu_{true} = 1.0$.

Figures 6.53 (a)–(d) show the results of meta-analysis combination when Q_{D0} is used. Because Q_{D0} produces one-sided upper limits with the smallest μ always having larger acceptance quantile, the SCAPs are always decreasing with increasing μ.

Figure 6.53 (a) shows that when $\mu_{true} = 0$, the Poisson model excludes $\mu = 1.0$ at 95% confidence after three trials (blue line). Figure 6.53 (b) shows that when $\mu_{true} = 1.0$, the Poisson model does not exclude $\mu = 1.0$, but has an upper limit that generally decreases toward $\mu = 1.0$ with the number of trials.

Figures 6.53 (c) and (d) show the results for the Gaussian model. For the case of $\mu_{true} = 0$, figure 6.53 (c) shows that the maximum of $A_{meta}(\mu)$ after ten trials does not reach the 90% confidence level (10^{-1} on y-axis).

Figures 6.52 (c) and (d) show the results using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$ which are similar to those of Q_{FC}.

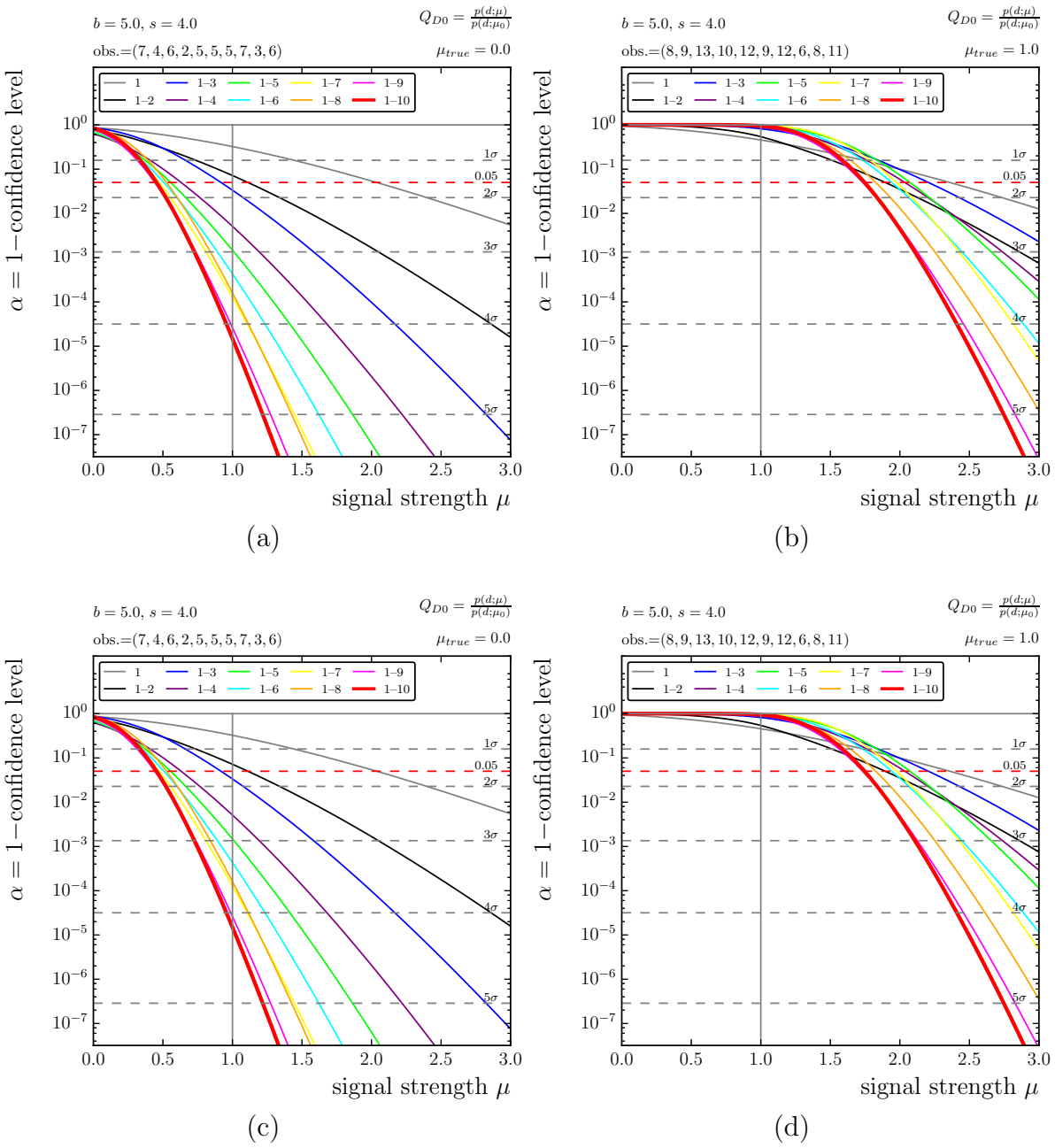


Figure 6.53: Meta-analysis – Progressively combining the SCAPs of 10 repeated samples from the Poisson model using Q_{D0}. Poisson model with (a) sampling observed outcomes when $\mu_{true} = 0$ and (b) sampling when $\mu_{true} = 1.0$. Gaussian model with (c) $\mu_{true} = 0$ and (d) $\mu_{true} = 1.0$.

6.4.9.2. Combining CAPs: Comparison to Ideal

The result of combining three experiments via meta-analysis is now compared to that which would be ideally obtained if the original data was available. This is done to examine how closely the final meta-analysis SCAP of three different experiments follows the SCAP obtained from the optimal combination achieved by multiplying the three models to form a combined model, then performing the Neyman construction, and finally reporting the observed SCAP obtained from the simultaneous measurement of d_1, d_2 and d_3.

Figures 6.54 (a) and (b) show these results for the univariate Poisson signal model using Q_{FC} with $\mu_{true} = 0$ (left panel) and $\mu_{true} = 1.0$ (right panel). For both (a) and (b), the expected background and signal counts of the three different experiments are the same (listed at the top of each): for experiment 1 the expected background is $b_1 = 5.0$ and signal is $s_1 = 4.0$, for experiment 2, $b_2 = 10.0$ and $s_2 = 3.0$, and for experiment 3, $b_3 = 6.0$ and $s_3 = 3.0$. Using a simple definition of signal significance, $g \equiv s/\sqrt{s+b}$ as a measure of signal-background discrimination power, these expected values indicate that experiment 1 has the greatest signal-background discrimination and experiment 2 the least: $g_1 = 1.3$, $g_2 = 0.8$, and $g_3 = 1.0$.

The top panel of figure 6.54 (a) shows the result if the observed outcomes are exactly equal to the expected background-only counts of each experiment when assuming $\mu_{true} = 0$: $d_1 = 5$, $d_2 = 10$ and $d_3 = 6$. The SCAPs this produces for each separate experiment are shown with blue, violet, and black lines, respectively. Because experiment 1 has greater signal-background discrimination, it provides the shortest 95% confidence interval, [0, 1.6], while experiment 2 yields [0, 2.9] and experiment 3 yields [0, 2.2]. A solid red line shows the result of combining these three results via the meta-analysis of equation 6.4.15 when the weights are set to 1.0 (equal weighting of experiments). This has a 95% confidence interval of [0, 1.2]. The dashed black line shows the ideal SCAP from a combined trivariate Poisson model, which produces slightly shorter confidence intervals than the meta-analysis (solid red line). The ideal 95% confidence interval is [0, 1.0]. Recall that the purpose of meta-analysis is to allow the combination of experiments when *only* the observed SCAPs are known.

One could argue that experiment 1 should be given greater weight and that experiment 2 the least weight due to differences in signal-background discrimination power. The discrimination power depends on the value of μ being examined and so the local significance is defined as $g_\mu \equiv \mu\, s/\sqrt{\mu\, s + b}$. For each value of μ, the experiments are also weighted by their local significance (normalized to 1.0) per equation 6.4.15, and this result is shown with a red dotted line. The result of this local significance weighting is nearly identical to the unweighted result (solid red line).

The second and third panels of figure 6.54 (a) show two separate trials of random samples of observed counts. Taken together, the three panels of figure 6.54 (a) show that in general the meta-analysis combination of SCAPs may approximate the ideal combination, except that in the third panel a small observed count in experiment 1 (blue line) contributes more strongly to the ideal combination than it does to the meta-analysis combination.

Figure 6.54 (b) shows similar results when $\mu_{true} = 1.0$. In all three panels, the ideal combination provides shorter 95% confidence intervals than the meta-analysis, which is sensible since the ideal combination has access to greater information regarding the experiments.

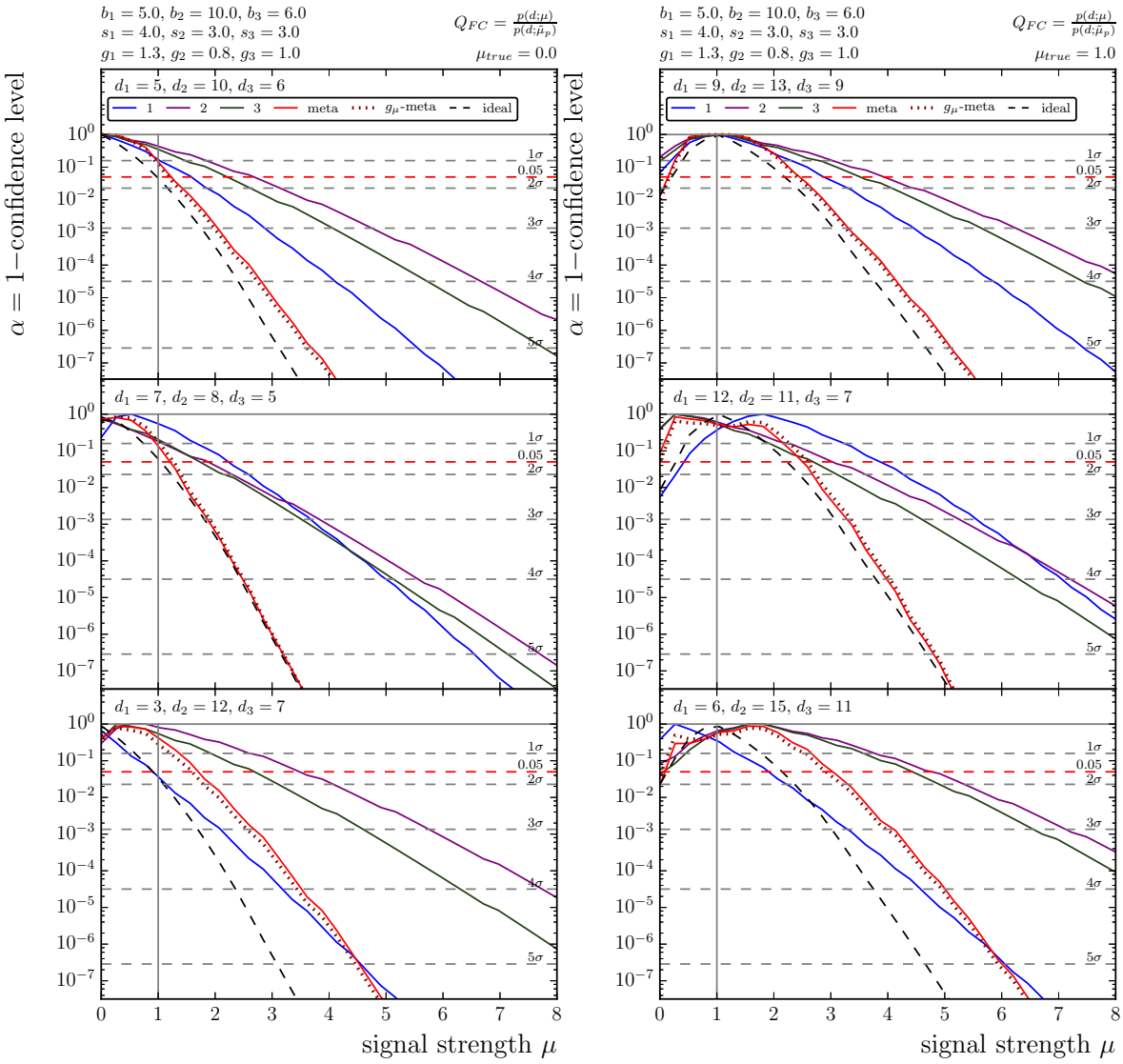


Figure 6.54: Meta-analysis – Combining three experiments from the Poisson model using Q_{FC}: (a) random sampling using $\mu_{true} = 0$ and (b) using $\mu_{true} = 1.0$. The top panel of (a) shows the result when the observed outcome counts are exactly equal to the expected background, while the top panel of (b) shows the result when the observed counts are exactly equal to the expected background plus the expected signal at nominal strength ($\mu = 1$). The lower two panels of each show two results of random sampling for the observed outcome counts. The solid red line shows the meta-analysis SCAP (the dotted red line is the significance-weighted meta-analysis SCAP) and the dashed black line shows the ideally combined observed SCAP.

Figures 6.55 (a) an (b) show the results for the univariate Gaussian signal model using Q_{FC}. (Results using $Q_{AZ} = Q_{inc} \cdot Q_{FC}$ were nearly identical to Q_{FC} for both the Poisson and Gaussian models and so are not provided.) These results are similar to those for the Poisson model. However, the second panel of figure 6.56 (b) shows that unlike the ideally combined SCAP with 95% confidence interval $[0.55, 1.25]$, the meta-analysis SCAP has two maxima and produces the disjoint confidence interval $[0.2, 0.8] \cup [1.55, 1.75]$. This is because for this random sampling of observed outcomes, experiment 1 provides evidence for large μ while experiment 3 for small (with experiment 2 in between). It is clear that the ideal combination will always returns values of μ that best fit the independent channels simultaneously, which results in a sort of averaging. Perhaps when independent channels provide conflicting evidence, the advice of the American Statistical Association not to use any index as a substitute for scientific reasoning should apply [303]. In this case, the conflicting evidence of experiments 1 and 3 is reflected in the meta-analysis SCAP, but not in the ideal combination.

Figures 6.56 (a) and (b) show the results for the Poisson model using Q_{D0}. For these upper limits, the meta-analysis SCAP well approximates that of the ideal combination, with minor differences when one experiment provides a much shorter upper limit than the other two.

A confidence-acceptance profile (CAP) provides more inferential information than a single confidence interval at a specified confidence level. This section shows that if observed CAPs are provided as the results of multiple experiments measuring the same parameter of interest, μ, then the CAPs of the independent experiments may be combined in a straightforward manner to provide a meta-analysis CAP that summarizes the combined information. Section 6.4.7 discussed how CAPs obtained during the Neyman construction offer the advantage of providing significance ratio functions that allow for the evaluation of the evidence for each *separate* value of μ (and which have a frequentist interpretation). This section shows that multiple sources of evidence may be easily combined via meta-analysis *without* reperforming the Neyman construction.

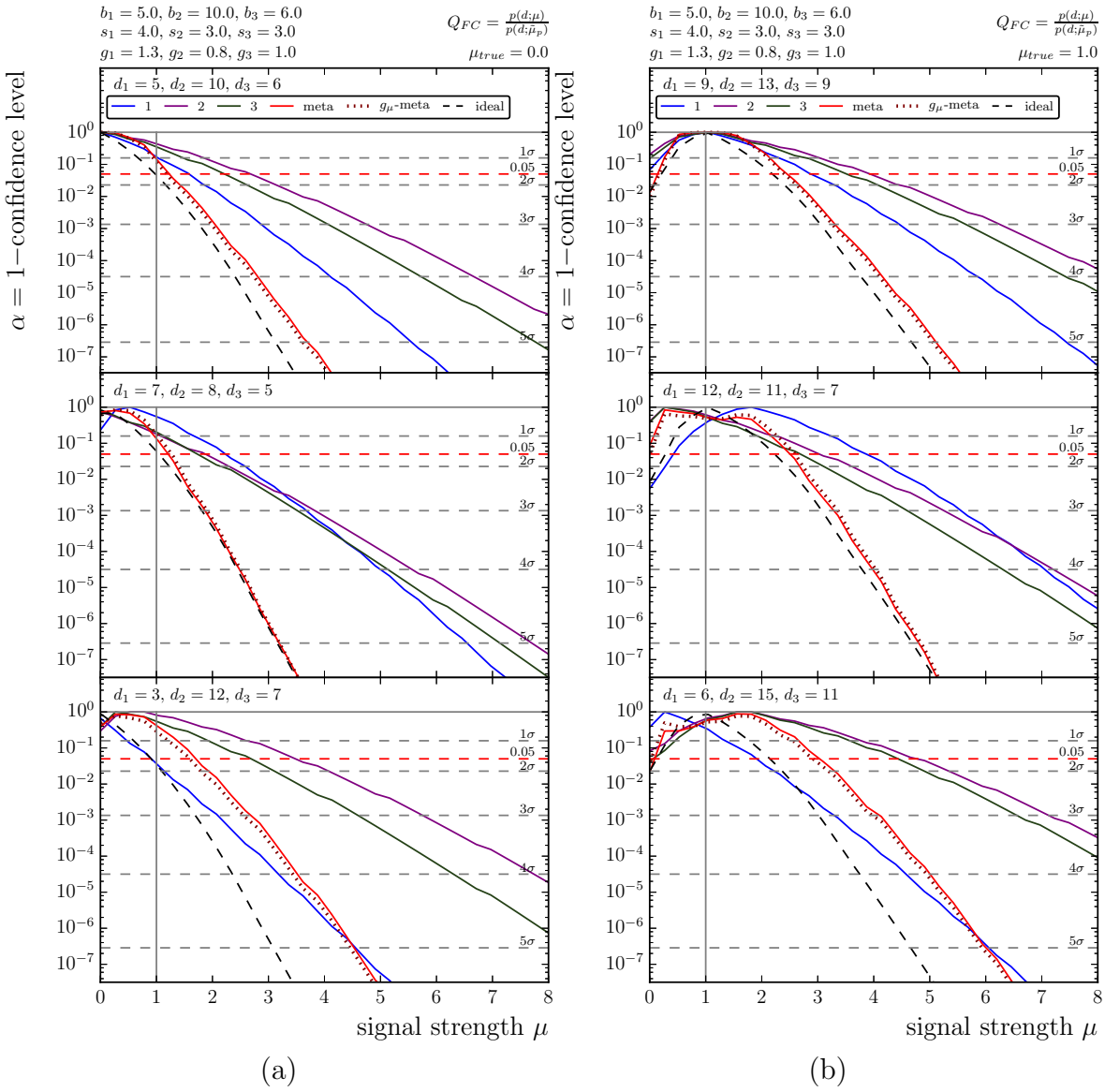


Figure 6.55: Meta-analysis – Combining three experiments from the Gaussian model using Q_{FC}: (a) $\mu_{true} = 0$ and (b) $\mu_{true} = 1.0$. The top panel of (a) shows the result when the observed outcome counts are exactly equal to the expected background, while the top panel of (b) shows the result when the observed counts are exactly equal to the expected background plus the expected signal at nominal strength ($\mu = 1$). The lower two panels of each show two results of random sampling for the observed outcome counts. The solid red line shows the meta-analysis SCAP (the dotted red line is the significance-weighted meta-analysis SCAP) and the dashed black line shows the ideally combined observed SCAP.

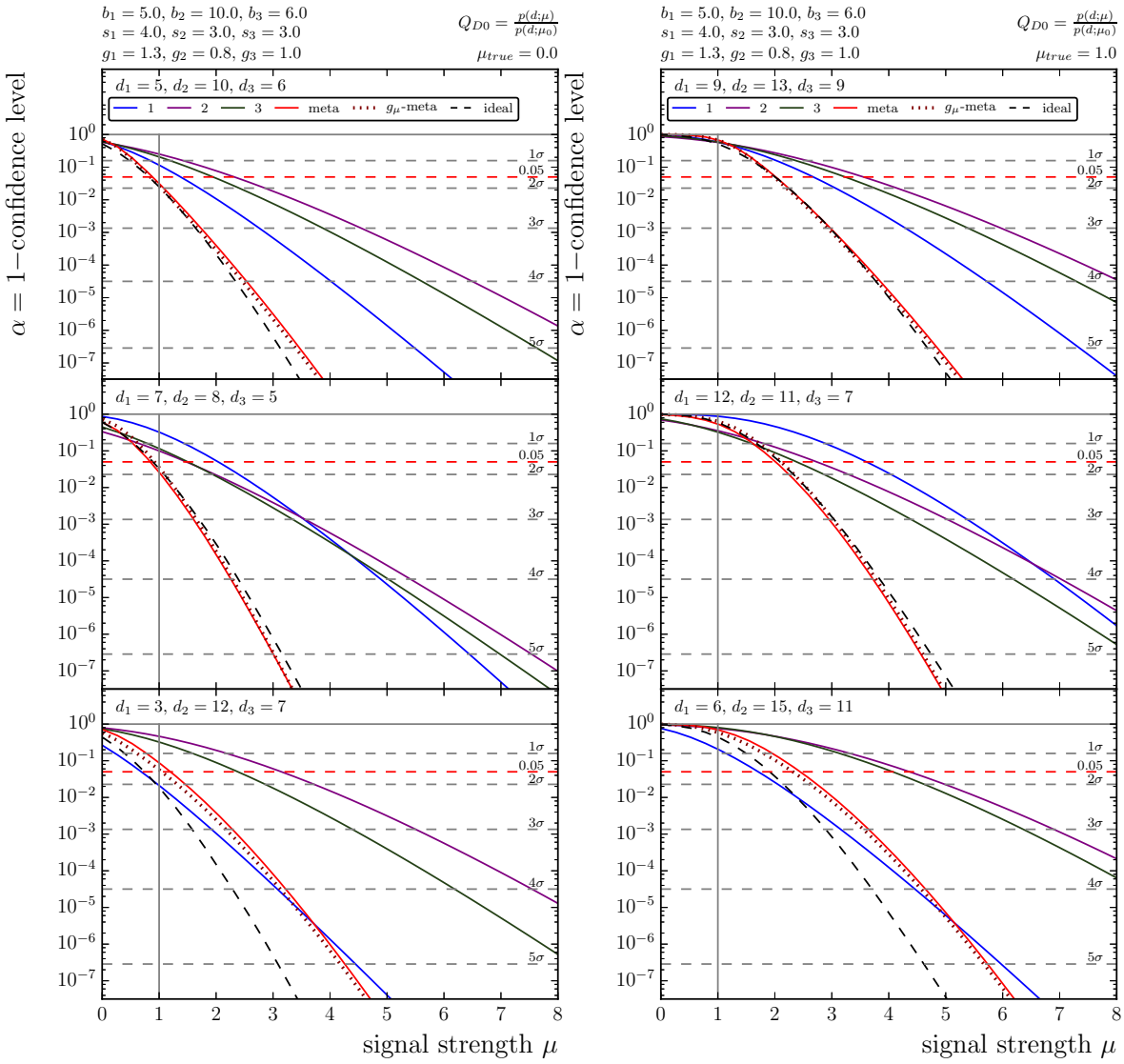


Figure 6.56: Meta-analysis – Combining three experiments from the Poisson model using Q_{D0}: (a) $\mu_{true} = 0$ and (b) $\mu_{true} = 1.0$. The top panel of (a) shows the result when the observed outcome counts are exactly equal to the expected background, while the top panel of (b) shows the result when the observed counts are exactly equal to the expected background plus the expected signal at nominal strength ($\mu = 1$). The lower two panels of each show two results of random sampling for the observed outcome counts. The solid red line shows the meta-analysis SCAP (the dotted red line is the significance-weighted meta-analysis SCAP) and the dashed black line shows the ideally combined observed SCAP.

6.5. Test Statistics in the Presence of Nuisance Parameters

Estimated systematic uncertainties are incorporated into the multivariate Poisson model with the inclusion of nuisance parameters that have Bayesian prior probability distributions as described in section 6.2.1 (page 258). These estimated uncertainties are derived from supplemental measurements so that the nuisance parameter priors may be considered empirically justified [248].

In the case of a single nuisance parameter, ν, the Neyman construction results in two dimensional confidence regions formed by the coordinate values, (ν, μ), that are accepted. This is illustrated in the example shown in figure 6.57 (a), where pink colored areas show the confidence region for each possible outcomes (at some confidence level $(1-\alpha)$). The observed confidence region for d_{obs} is shown with a red area.

Figure 6.57 (b) shows that if the experimenter wishes to report a confidence *interval* solely for the parameter of interest, μ, then the projection of the observed confidence region onto the μ-axis provides the only confidence interval that guarantees the desired $(1-\alpha)$ confidence level. In other words, the upper and lower limits of the projected confidence interval must capture all accepted values of μ in the confidence region since ν_{true} may be any value. This may lead to severe overcoverage, particularly in a particle search where there are usually many dozens of nuisance parameters.

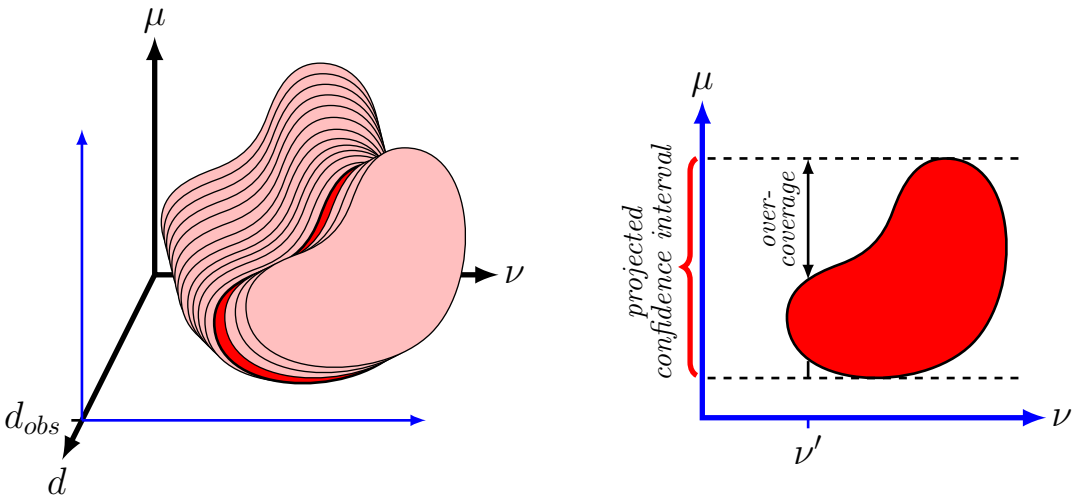
An alternative method is to assign a Bayesian prior probability distribution, $\pi(\nu)$, for each of the nuisance parameters. These prior distributions may then be used during the Neyman construction of the acceptance region of each value of μ by performing *Bayesian marginalization* over the nuisance parameters [329].[113] This procedure is demonstrated with the example shown in figures 6.58 (a) and (b).

Figure 6.58 (a) shows the test statistic distributions for the value μ'. If $Q(d, \mu')$ is the test statistic used to construct the acceptance region of the value μ', then $Q(d, \mu')$ is only a function of the possible outcomes, d, and is independent of ν. However, the probability distribution of $Q(d, \mu')$ does depend on the value of ν because ν alters the model parameters s_i and b_i in the multivariate Poisson model, and thus alters the probability for each outcome, $p(d|\,\mu', \nu)$ (which in turn gives the test statistic distribution, $p(q|\,\mu', \nu)$, where q is in the range of $Q(d, \mu')$).[114]

Nevertheless, figure 6.58 (b) shows that the ν-independent probability distribution of $Q(d, \mu')$ may be found by Bayesian marginalization, where $Q(d, \mu')$ is averaged over the possible values of ν, weighted by the prior, $\pi(\nu)$. The Bayesian marginalized test statistic distribution, $p(q;\, \mu')$, is then used to locate the critical value $Q^*(\mu')$ that determines the acceptance region for μ' (red area).

[113] For a single nuisance parameter when not using Bayesian marginalization, each acceptance region (of possible outcomes, d,) is determined for *each* particular coordinate pair, (ν', μ'), eventually resulting in two dimensional confidence regions for each outcome d'. When using Bayesian marginalization, each acceptance region is determined for each value μ', resulting in one dimensional confidence intervals for each outcome d'.

[114] This probability could be written as $p(d|\,\nu;\, \mu')$ because ν is treated as a random variable so that the probability of d is conditional on ν, while μ remains an unknown parameter value.



(a) Neyman construction.

(b) Observed confidence region.

Figure 6.57: Projection is a frequentist method of obtaining a confidence interval for the parameter of interest, μ, in the presence of nuisance parameters, the example here having a single nuisance parameter, ν. (These figures are provided merely for explanatory demonstration.)

Figure (a): The Neyman construction at confidence level $(1 - \alpha)$ gives a two dimensional confidence region (in ν–μ space) for each possible outcome, d, (shown with pink areas). When viewed in all three dimensions, the result is a *confidence tube* extending outward in the direction of the d-axis. The confidence region for the observed outcome, d_{obs}, is shown in red. Blue arrows denote the μ-axis and ν-axis relative to d_{obs}.

Figure (b): The experimenter may wish to only report a confidence interval for μ rather than a two dimensional confidence region of coordinate values (ν, μ). Because ν_{true} is unknown, only the full projection of the two dimensional confidence region onto the μ-axis can guarantee coverage at the $(1 - \alpha)$ confidence level for *all* possible values of ν. Projection may lead to severe overcoverage. For example, the range of accepted values of μ for the value ν' is much shorter than the projected confidence interval.

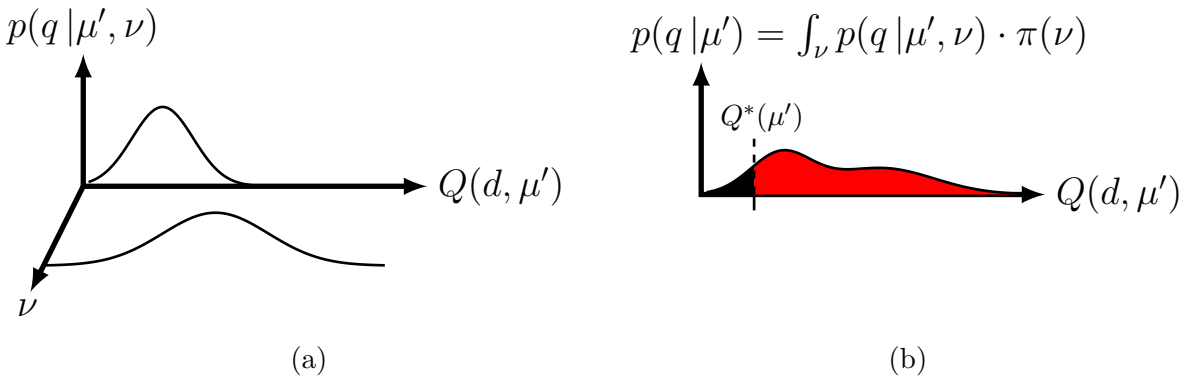


Figure 6.58: An example demonstrating how Bayesian marginalization provides one dimensional acceptance regions for the parameter of interest *during* the Neyman construction. (These figures are provided merely for explanatory demonstration.)

Figure (a) For a given μ', the probability distribution for the test statistic, $p(q|\mu',\nu)$, depends on both d and ν (with $q = Q(d,\mu')$). Note that if $q' = Q(d',\mu')$ and also $q' = Q(d'',\mu')$, the same test statistic value q' for two different outcomes, d' and d'', then $p(q'|\,\mu',\nu) = p(d';\,\mu',\nu) + p(d'';\,\mu',\nu)$.

Figure (b) Marginalizing over ν (averaging over ν weighted by $\pi(\nu)$) removed the ν-dependence of the test statistic distribution allowing the determination of a one dimensional acceptance region for μ' (as defined by the critical value, $Q^*(\mu')$).

Frequentist coverage with $(1-\alpha)$ confidence is no longer guaranteed when using Bayesian marginalization. This is because ν_{true} exists with a single value, and in order to guarantee $(1-\alpha)$ coverage, one must construct a confidence region with $(1-\alpha)$ acceptance for *each* possible value of ν. This is no longer true after averaging over the ν, and a coverage study may be needed to understand the coverage properties of the marginalized acceptance regions.

Using a test statistic, $Q(d,\mu)$, that is independent of ν is common in a particle search (with either the nominal or profiled methods described in the next section). Once the critical values are determined for each value of μ using Bayesian marginalization, then the acceptance or rejection of a value μ' in the confidence interval for an outcome d' is found simply by comparing $Q(d',\mu')$ to $Q^*(\mu')$ (specifically by accepting μ' into the confidence interval for d' if $Q(d',\mu') \geq Q^*(\mu')$).

The test statistic itself need not be independent of ν, and may be written more generally as $Q(d,\mu',\nu)$. This does not alter the Bayesian marginalization process because one still obtains a test statistic distribution that is independent of ν after the marginalization. However, the acceptance of a value μ' within the confidence interval of an outcome d' cannot be determined immediately because $Q(d',\mu',\nu)$ cannot be compared to $Q^*(\mu')$ since ν is unknown. Therefore an additional Bayesian marginalization step is required to determine the confidence region for a given outcome d'.

This second marginalization step is demonstrated in figures 6.58 (a) and (b). Figure 6.58 (a) shows the test statistic for the outcome d' and particular value μ' as a function of ν. In other words, for the specified d' and μ', the test statistic does not provide a single value to compare to the critical value, $Q^*(\mu')$, but rather a range of values.

Figure 6.58 (b) shows that a single value for d' and μ' may be obtained by marginalizing over the test statistic itself: $Q(d', \mu') = \int_\nu Q(d', \mu', \nu) \cdot \pi(\nu)$. This provides the single value, $Q(d', \mu')$ that may be compared to $Q^*(\mu')$ in order to determine the acceptance of μ' in the confidence interval of d'.

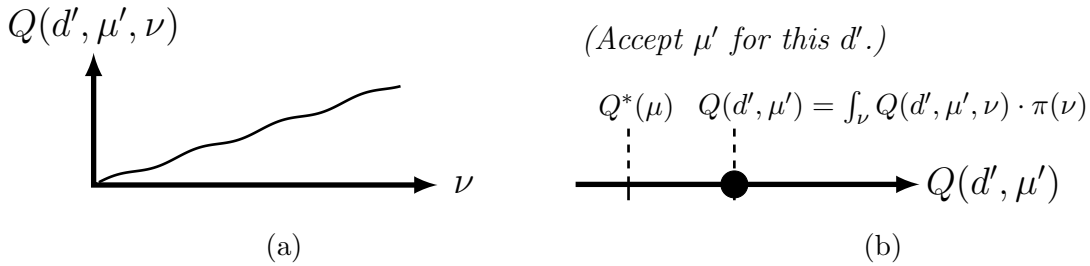


Figure 6.59: An example demonstrating how Bayesian marginalization after the Neyman construction must be performed if the test statistic is an explicit function of ν. (These figures are provided merely for explanatory demonstration.) (a) For a given d' and μ', the value of Q is not uniquely defined. (b) Marginalizing over ν provides a single value, $Q(d', \mu')$ that can be compared to the critical value $Q^*(\mu'$ in order to determine acceptance/rejection of μ'.

6.5.1. Defining the Test Statistics Including Nuisance Parameters

The most general form of the test statistic is $Q(d, \mu, \nu)$, a function of outcomes d, the parameter of interest μ, and all nuisance parameters ν.[115] Three options for defining the test statistic are examined here, two of which are commonly used in particle physics searches.

- $\mathbf{Q^N(d, \boldsymbol{\mu})}$ **,** ***Nominal*** – The simplest approach is to define the test statistic using the nominal values of the nuisance parameters, where $\nu = 0$ in particle physics typically corresponds to the nominal simulated estimates for b_i and s_i in the absence of systematic uncertainties. The nominal method can be applied to any of the previously discussed test statistics and likelihood ratios so that, for example,

$$Q^N_{D0}(d, \mu) = \frac{p(d;\, \mu, \nu = 0)}{p(d;\, \tilde{\mu}^N_0, \nu = 0)}, \tag{6.5.1}$$

  where $\tilde{\mu}^N_0$ represents the background-only model when the nominal values for the nuisance parameters are used. Q^N_{D0} is the test statistic is used in the OpThyLiC software [244] and in the MCLimit software [245] for unprofiled nuisance parameters.

- $\mathbf{Q^P(d, \boldsymbol{\mu})}$ **,** ***Profiled*** – Profiling may be used to determine the most likely values of the nuisance parameters for a given outcome and value of μ. Profiling may be performed on any of the previously discussed test statistics so that, for instance,[116]

$$Q^P_{D0}(d', \mu') = \frac{p(d';\, \mu',\, \tilde{\nu}^P_{\mu'})}{p(d';\, [\tilde{\mu}^P_0,\, \tilde{\nu}^P_0])}. \tag{6.5.2}$$

  The value $\tilde{\nu}^P_{\mu'}$ is that which maximizes $p(d';\, \mu',\, \nu)$ for given μ' and d'. The term $\left[\tilde{\mu}^P_0,\, \tilde{\nu}^P_0\right]$ is written to denote that these values are found simultaneously as those which maximize $p(d';\, \mu,\, \nu)$ for a given d' (the "0" subscripts indicate that these are for the $D0$ test statistic). This is the method used in the MCLimit software [245] for profiled nuisance parameters.[117] Another example is applying profiling to the Feldman-Cousins test statistic which yields

$$Q^P_{FC}(d', \mu') = \frac{p(d';\, \mu',\, \tilde{\nu}^P_{\mu'})}{p(d';\, \left[\tilde{\mu}^P_p,\, \tilde{\nu}^P_p\right])}, \tag{6.5.3}$$

  where $\left[\tilde{\mu}^P_p,\, \tilde{\nu}^P_p\right]$ indicates that these MLEs are found simultaneously as those that maximize $p(d;\, \mu, \nu)$ with the restriction that $0 \leq \mu$.

---

[115] Vector notation is usually suppressed here so that ν represents the vector of all nuisance parameters, $\nu = [\nu_1, \nu_2, ...]$. The "value of ν" or "a particular value ν'" when written indicates specified values for each vector component of ν.

[116] For a fuller discussion of this method of profiling, see the chapter *Likelihood Ratio Test and Test Efficiency* in [263]. For specific applications in particle physics, [330] discusses the situations when profiling should not be implemented. Also see [331].

[117] One study has found that upper limits computed both with and without profiling were observed to be nearly identical [154]. Also, each of the 199 upper limit results reported in appendix section C.4 (page 512) were found to be nearly identical whether implementing Q^N_{D0} or else Q^P_{D0} using the MCLimit software.

- $\mathbf{Q^F(d, \boldsymbol{\mu}, \boldsymbol{\nu})}$ **,** ***Free*** – The value of ν need not be set to any specific value if the test statistic is given as an explicit function of ν. For a given value ν', the Q_{D0} test statistic becomes

$$Q^F_{D0}(d', \mu', \nu') = \frac{p(d';\, \mu',\, \nu')}{p(d';\, \tilde{\mu}^F_0,\, \nu')}, \tag{6.5.4}$$

where $\tilde{\mu}^F_0$ provides the maximum likelihood of the background-only model, $p(d';\, \tilde{\mu}_0,\, \nu')$, given the outcome d' and value μ'. When applied to the Feldman-Cousins test statistic,

$$Q^F_{FC}(d', \mu', \nu') = \frac{p(d';\, \mu',\, \nu')}{p(d';\, \tilde{\mu}^F_p,\, \nu')}, \tag{6.5.5}$$

where $\tilde{\mu}^F_p$ maximizes $p(d;\, \mu, \nu')$ with the restriction that $0 \leq \mu$. When the free method is used, the additional marginalization step demonstrated in figures 6.58 (a) and (b) is necessary to determine the confidence regions.

The profiling of ν may be extremely computationally expensive when many nuisance parameters are present. Also, the values $\left[\tilde{\mu}^P_X,\, \tilde{\nu}^P_X\right]$ (where X denotes the type of test statistic), are usually not the same [289]. For Q^P_{FC}, the hypothesis test this corresponds to is $H_N : \mu = \mu', \nu = \tilde{\nu}^P_{\mu'}$ and $H_A : 0 \leq \mu \neq \mu', \nu = \tilde{\nu}^P_p$ so that one is comparing the most likely model value of ν given μ' against the most likely model value of ν given the most likely signal model. In other words, one is using different values of ν for the numerator and denominator.

From the frequentist perspective, there exists only one value ν_{true}, and so it may be more sensible to keep the value of ν the same for both H_N and H_A. Rather set ν to any particular value, it may be sensible to examine the free likelihood ratio, Q^F_{FC}. This would correspond to a series of hypothesis tests where if $\nu' = \nu_{true}$, then $H_N : \mu = \mu' \,|\, \nu'$ and $H_A : 0 \leq \mu \neq \mu' \,|\, \nu'$. In other words, as each ν' is obtained from Monte Carlo sampling, ν' is treated as if it were the true value to obtain the test statistic value. It is often the case that for two different values, ν' and ν'', the test statistic values for outcome d' and value μ' are quite different: $Q(d', \mu', \nu')$ may be small while $Q(d', \mu', \nu'')$ is large. This means that outcome d' provides weak evidence for μ' if $\nu_{true} = \nu'$, but provides strong evidence for μ' if $\nu_{true} = \nu''$. The free likelihood ratio attempts to address these differences.

With three methods for incorporating nuisance parameters into test statistics, and with the three common and three Arizona test statistics, this provides 18 test statistics to be studied. These are defined in tables 6.11–6.13. There are several maximum likelihood values used in these test statistics that have the following definitions.

- The background-only model, where the nuisance parameters, ν, are do not include those that solely apply to the signal estimates.
  - $\tilde{\boldsymbol{\mu}}^{\mathbf{N}}_{\mathbf{0}}$ – A function of d obtained by determining b_i when the nominal nuisance parameter values are used.
  - $\tilde{\boldsymbol{\mu}}^{\mathbf{P}}_{\mathbf{0}}$ – A function of d obtained by determining b_i when the profiled nuisance parameter values, $\tilde{\nu}^P_0$, are used, those that maximize $p(d;\, \tilde{\mu}_0, \nu)$

- $\tilde{\boldsymbol{\mu}}_{\mathbf{0}}^{\mathbf{F}}$ – A function of both d and ν obtained by determining b_i separately for each value of ν.

- The unrestricted MLE, $\tilde{\mu}_u$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{u}}^{\mathbf{N}}$ – A function of d obtained by maximizing $p(d;\, \mu, \nu = 0)$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{u}}^{\mathbf{P}}$ – A function of d found in conjunction with $\tilde{\nu}_u^P$ by maximizing $p(d;\, \mu, \nu)$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{u}}^{\mathbf{F}}$ – A function of d and ν obtained for a given ν' by maximizing $p(d;\, \mu, \nu')$.

- The physical MLE, $\tilde{\mu}_p$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{p}}^{\mathbf{N}}$ – A function of d found by maximizing $p(d;\, \mu, \nu = 0)$ with the restriction $0 \leq \mu$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{p}}^{\mathbf{P}}$ – A function of d found in conjunction with $\tilde{\nu}_p^P$ by maximizing $p(d;\, \mu, \nu)$ with the restriction $0 \leq \mu$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{p}}^{\mathbf{F}}$ – A function of d and ν obtained for a given ν' by maximizing $p(d;\, \mu, \nu')$ with the restriction $0 \leq \mu$.

- The bounded MLE, $\tilde{\mu}_b$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{b}}^{\mathbf{N}}$ – A function of d and μ' where given a specific value μ', $p(d;\, \mu, \nu = 0)$ is maximized with the restriction $0 \leq \mu \leq \mu'$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{b}}^{\mathbf{P}}$ – A function of d and μ' found in conjunction with $\tilde{\nu}_b^P$, where given a specific value μ', $p(d;\, \mu, \nu)$ is maximized with the restriction $0 \leq \mu \leq \mu'$.
  - $\tilde{\boldsymbol{\mu}}_{\mathbf{b}}^{\mathbf{F}}$ – A function of d, μ' and ν obtained for given μ' and ν' by maximizing $p(d;\, \mu, \nu')$ with the restriction $0 \leq \mu \leq \mu'$.

- Other MLE.
  - $\tilde{\boldsymbol{\nu}}_{\mu'}^{\mathbf{P}}$ – A function of d and μ' found by maximizing $p(d;\, \mu', \nu)$.

$Q_X^N(d,\mu')$	definition (nominal)
Q_{FC}^N	$\frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)}$
Q_{D0}^N	$\frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_0^N,\nu{=}0)}$
Q_{AT}^N	$\frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_b^N,\nu{=}0)}$
Q_{rep}^N	$\frac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^N\right\rvert}$
Q_{inc}^N	$\frac{p(d;\tilde{\mu}_u^N,\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)}$
Q_{AZ1}^N	$\frac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^N\right\rvert} \cdot \frac{p(d;\tilde{\mu}_u^N,\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)} \cdot \frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)}$
Q_{AZ2}^N	$\frac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^N\right\rvert} \cdot \frac{p(d;\tilde{\mu}_u^N,\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)} \cdot \frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_0^N,\nu{=}0)}$
Q_{AZ3}^N	$\frac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^N\right\rvert} \cdot \frac{p(d;\tilde{\mu}_u^N,\nu{=}0)}{p(d;\tilde{\mu}_p^N,\nu{=}0)} \cdot \frac{p(d;\mu',\nu{=}0)}{p(d;\tilde{\mu}_b^N,\nu{=}0)}$

Table 6.11: Definitions of test statistics, $Q_X^N(d,\mu')$, that are used to determine the acceptance/rejection regions for the value μ' using the nominal method treatment of nuisance parameters.

$Q_X^P(d,\mu')$	definition (profile)
Q_{FC}^P	$\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}$
Q_{D0}^P	$\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_0^P,\tilde{\nu}_0^P\right])}$
Q_{AT}^P	$\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_b^P,\tilde{\nu}_b^P\right])}$
Q_{rep}^P	$\dfrac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^P\right\rvert}$
Q_{inc}^P	$\dfrac{p(d;\left[\tilde{\mu}_u^P,\tilde{\nu}_u^P\right])}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}$
Q_{AZ1}^P	$\dfrac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^P\right\rvert}\cdot\dfrac{p(d;\left[\tilde{\mu}_u^P,\tilde{\nu}_u^P\right])}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}\cdot\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}$
Q_{AZ2}^P	$\dfrac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^P\right\rvert}\cdot\dfrac{p(d;\left[\tilde{\mu}_u^P,\tilde{\nu}_u^P\right])}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}\cdot\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_0^P,\tilde{\nu}_0^P\right])}$
Q_{AZ3}^P	$\dfrac{1}{0.1+\left\lvert\mu'-\tilde{\mu}_p^P\right\rvert}\cdot\dfrac{p(d;\left[\tilde{\mu}_u^P,\tilde{\nu}_u^P\right])}{p(d;\left[\tilde{\mu}_p^P,\tilde{\nu}_p^P\right])}\cdot\dfrac{p(d;\mu',\tilde{\nu}^P_{\mu'})}{p(d;\left[\tilde{\mu}_b^P,\tilde{\nu}_b^P\right])}$

Table 6.12: Definitions of test statistics, $Q_X^P(d,\mu')$, that are used to determine the acceptance/rejection regions for the value μ' using the profiled method treatment of nuisance parameters.

$Q_X^F(d,\mu',\nu')$	definition (free)
Q_{FC}^F	$\frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_p^F,\nu')}$
Q_{D0}^F	$\frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_0^F,\nu')}$
Q_{AT}^F	$\frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_b^F,\nu')}$
Q_{rep}^F	$\frac{1}{0.1+\left\|\mu'-\tilde{\mu}_p^F\right\|}$
Q_{inc}^F	$\frac{p(d;\tilde{\mu}_u^F,\nu')}{p(d;\tilde{\mu}_p^F,\nu')}$
Q_{AZ1}^F	$\frac{1}{0.1+\left\|\mu'-\tilde{\mu}_p^F\right\|} \cdot \frac{p(d;\tilde{\mu}_u^F,\nu')}{p(d;\tilde{\mu}_p^F,\nu')} \cdot \frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_p^F,\nu')}$
Q_{AZ2}^F	$\frac{1}{0.1+\left\|\mu'-\tilde{\mu}_p^F\right\|} \cdot \frac{p(d;\tilde{\mu}_u^F,\nu')}{p(d;\tilde{\mu}_p^F,\nu')} \cdot \frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_0^F,\nu')}$
Q_{AZ3}^F	$\frac{1}{0.1+\left\|\mu'-\tilde{\mu}_p^F\right\|} \cdot \frac{p(d;\tilde{\mu}_u^F,\nu')}{p(d;\tilde{\mu}_p^F,\nu')} \cdot \frac{p(d;\mu',\nu')}{p(d;\tilde{\mu}_b^F,\nu')}$

Table 6.13: Definitions of test statistics, $Q_X^F(d,\mu',\nu')$, that are used to determine the acceptance/rejection regions for the value μ' using the free method treatment of nuisance parameters, which is also a function of the value ν' that is sampled during Bayesian marginalization.

6.5.2. Effect of Nuisance Parameters on Likelihood functions

Profiling the nuisance parameters to obtain the MLE values $\tilde{\mu}_p^P$ and $\tilde{\nu}_p^P$ is achieved by maximizing $p(d'; \mu, \nu)$ over both μ and ν simultaneously for a given outcome d'. As discussed in section 6.2.1, the effect of the the nuisance parameter is to adjust the expected signal and background yields (per bin of the discriminating variable). If s_i and b_i are the nominal expected signal and background yields per bin without systematic uncertainty, then for the nuisance parameter value ν', the adjusted yields are given by $s_i^* = f_i^s(\nu')\, s_i$ for the adjusted signal expected yield and $b_i^* = f_i^b(\nu')\, b_i$ for background, and where $\nu' = 0$ corresponds to no adjustment, $f_i^s(0) = 1$ and $f_i^b(0) = 1$.[118] For example, if there are 10 expected background events expected in the first bin, $b_1 = 10$, and for the value ν' the adjustment function is $f_1^b(\nu') = 1.5$, then the adjusted expected number of background events is $b_1^* = 15$.

Given a particular outcome, d', with d_i' for each bin of the discriminating variable, the multivariate Poisson distribution, $p(d'; \mu, \nu)$, has a global maximum when $\mu^*\, s_i^* + b_i^* = d_i'$ for each bin. Therefore, if either d_i' is already close to the nominal expected yields or else if the systematic uncertainty is large, it may be possible to find ν' such that $s_i^* = f_i^s(\nu')\, s_i$ and $b_i^* = f_i^b(\nu')\, b_i$. Therefore, profiling ν for a given outcome d' is equivalent to allowing the expected signal and background yields to fluctuate so that $\tilde{\mu}_p^P\, f_i^s\left(\tilde{\nu}_p^P\right)\, s_i + f_i^b\left(\tilde{\nu}_p^P\right)\, b_i$ is as close to d_i' as possible.

Because the nuisance parameter prior distributions are typically Gaussian, they have no boundary and may fluctuate arbitrarily far from the nominal value, which is typically $\nu = 0$. However, the systematic uncertainties that the nuisance parameters model are empirically determined from auxiliary measurements. Allowing very large deviations in ν would violate the spirit of these measurements. For example, if a one standard deviation change in ν represents a 2% deviation in the measured luminosity, then 50 standard deviations implies a 100% deviation. This would be considered physically impossible despite being mathematically possible according to the Gaussian prior describing the uncertainty in luminosity. Therefore, in a particle search, one usually restricts $\tilde{\nu}_p^P$ to lie within a limited range about the nominal value. This is typically two standard deviations above and below the nominal value.

In order to examine the effects of nuisance parameters, a representative set of systematic uncertainties is applied. A systematic uncertainty for the expected yield of each bin (for both signal and background counts, b_i and s_i) is modeled with a Gaussian distribution with a standard deviation of 5% of the nominal value. For example, if $b_1 = 10$, then an upward fluctuation by one standard deviation results in $b_1^* = 10.5$. Each bin systematic uncertainty is independent, and for a particle search, is usually due to the volume of the simulated data that is available (often referred to as "statistical uncertainty" in particle searches). The 5% standard deviation was approximately the amount observed in the vector-like quark search described in the previous chapter.

Additionally, a single global systematic uncertainty is applied that affects all bins equally (completely correlated between bins). The prior distribution for this is also modeled with a Gaussian distribution that has a standard deviation of 10% of each bin's nominal values.

[118] In equation 6.2.10, the nuisance parameters were separated by type while here ν generically refers to all nuisance parameters, so that simpler function notation is used defined by that of equation 6.2.10: $f_i^s(\nu) \equiv \nu_i^s g_i^s(\boldsymbol{\nu})$ and $f_i^b(\nu) \equiv \nu_i^b g_i^b(\boldsymbol{\nu})$.

For example, if $b_1 = 10$ and $b_2 = 20$, then an upward fluctuation of one standard deviation results in $b_1^* = 11$ and $b_2^* = 22$. For the bivariate Poisson model of example 2, this results in five nuisance parameters: two for the signal bins, two for the background bins, and one for the global systematic uncertainty.

Profiling many parameters for each possible outcome is computationally expensive. Maximization here is performed over a grid of the nuisance parameters, ν, with spacing of 0.1 standard deviations for each ν, and is performed using a simple line search over this grid, while searching over a continuous set of μ to within a tolerance of $\frac{\Delta\mu}{10}$, where $\Delta\mu$ is the original grid spacing of μ. The grid spacing of μ is increased from the value 0.025 used in the results of the previous sections of this chapter to 0.10 in order to decrease the time of computation. The results of this larger grid spacing do not much alter the previous results. The same number of Monte Carlo simulated outcomes as previously used (100,000) is used to construct the test statistic distributions for each value of μ.

It is instructive to examine how increasing the range of permissible values of ν explored in profiling affects the location of the maxima of the likelihood functions. Figures 6.60 (a)–(f) demonstrate how increasing the range of permissible values of ν affects the profiled likelihood functions for example 2 (the bivariate Poisson model described in table 6.7 on page 329). Figure 6.60 (a) shows the likelihood function, $p(d;\, \mu, \nu = 0)$, for each outcome when ν is restricted to its nominal value (zero for each of the five nuisance parameter). In this case, the MLE values $\tilde{\mu}_p^P$ (yellow circles) have the same values as $\tilde{\mu}_p$ in the model without nuisance parameters that was shown in figure 6.5 (b) (page 273).

Figure 6.60 (b) shows what happens when the MLE values of the nuisance parameters, $\tilde{\nu}_p^P$, are restricted to lie within one standard deviation of their nominal values (i.e., $\pm 5\%$ for each bin and $\pm 10\%$ for the global). The signal strength MLE, $\tilde{\mu}_p^P$, which is solved for simultaneously with $\tilde{\nu}_p^P$, now either increases or decreases from $\tilde{\mu}_p$. This is because the range of explored values of ν allow $\tilde{\mu}_p^P\, f_i^s\left(\tilde{\nu}_p^P\right)\, s_i + f_i^b\left(\tilde{\nu}_p^P\right)\, b_i$ to get closer to the global maximum of the likelihood function of each outcome, d.[119]

Figures 6.60 (c)–(f) show the same as (b) for two through five standard deviations, respectively. As profiling for ν occurs over a larger range of permissible values of ν, the spread in $\tilde{\mu}_p^P$ becomes greater (vertical spread of yellow circles). For the commonly used $\pm 2\,\sigma$ boundary values shown in (c), the spread for outcomes near d_{mode}^{s+b} is appreciable, $[0.55, 1.45]$.

Also, the widths of the profiled likelihood functions also increase (vertical height of gray bands) with the larger range of permissible values of ν. The reason for this was discussed in section 6.2.2.2 where it was observed that the likelihood function width was greater for likelihood functions with larger maximum values. Profiling essentially gives a larger maximum value for the likelihood function, $p(d;\, \tilde{\mu}_p^P, \tilde{\nu}_p^P) > p(d;\, \tilde{\mu}_p, \nu = 0)$, and so the profiled likelihood function has greater width than the unprofiled. A wider likelihood function for outcome d' is correlated with a longer confidence interval for d', so that increasing the range of permissible ν should result in longer confidence intervals.

Figures 6.61 (a)–(f) show these results when applied to example 3b (the multivariate Poisson model described in table 6.7), which are qualitatively similar to those of example 2.

[119] These figures specifically show the likelihood function of each outcome, d, that corresponds to the profiled nuisance parameter: $p(d;\, \mu, \tilde{\nu}_p^P)$.

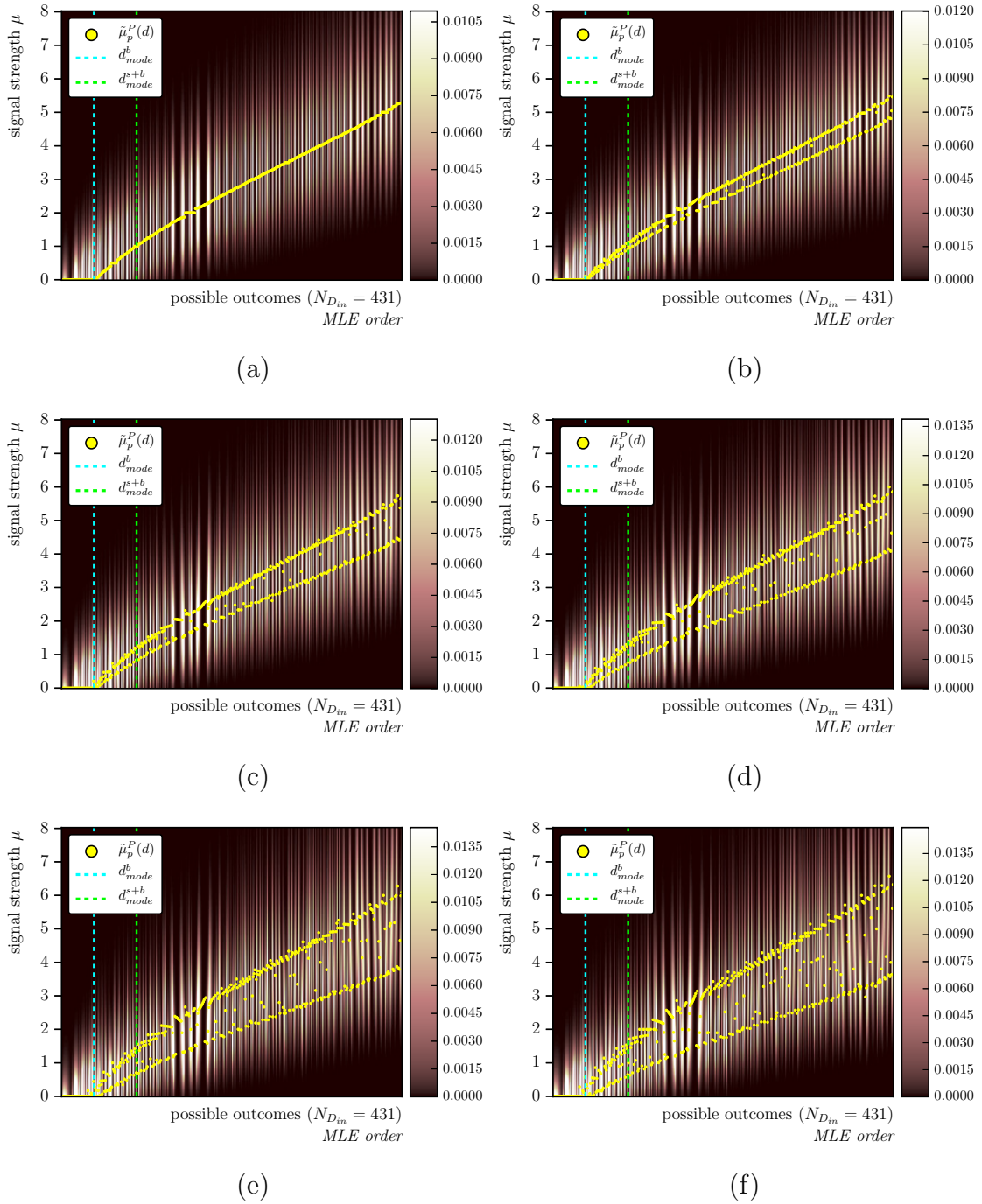


Figure 6.60: Example 2 (bivariate) – The impact of profiling nuisance parameters, ν, on the profiled likelihood functions, $p(d; \mu, \tilde{\nu}_p^P)$. The model $p(d; \mu, \nu)$ is maximized to find $\tilde{\mu}_p^P$ (yellow circles) and $\tilde{\nu}_p^P$ but only allowing $\tilde{\nu}_p^P$ in the range $[-x\sigma, +x\sigma]$ of its Gaussian distribution. The plot of the likelihood function is made by then allowing μ to deviate from $\tilde{\mu}_p^P$. (a) $x = 0$ (equivalent to $\nu = 0$), (b) $x = 1$, (c) $x = 2$, (d) $x = 3$, (e) $x = 4$, (f) $x = 5$.

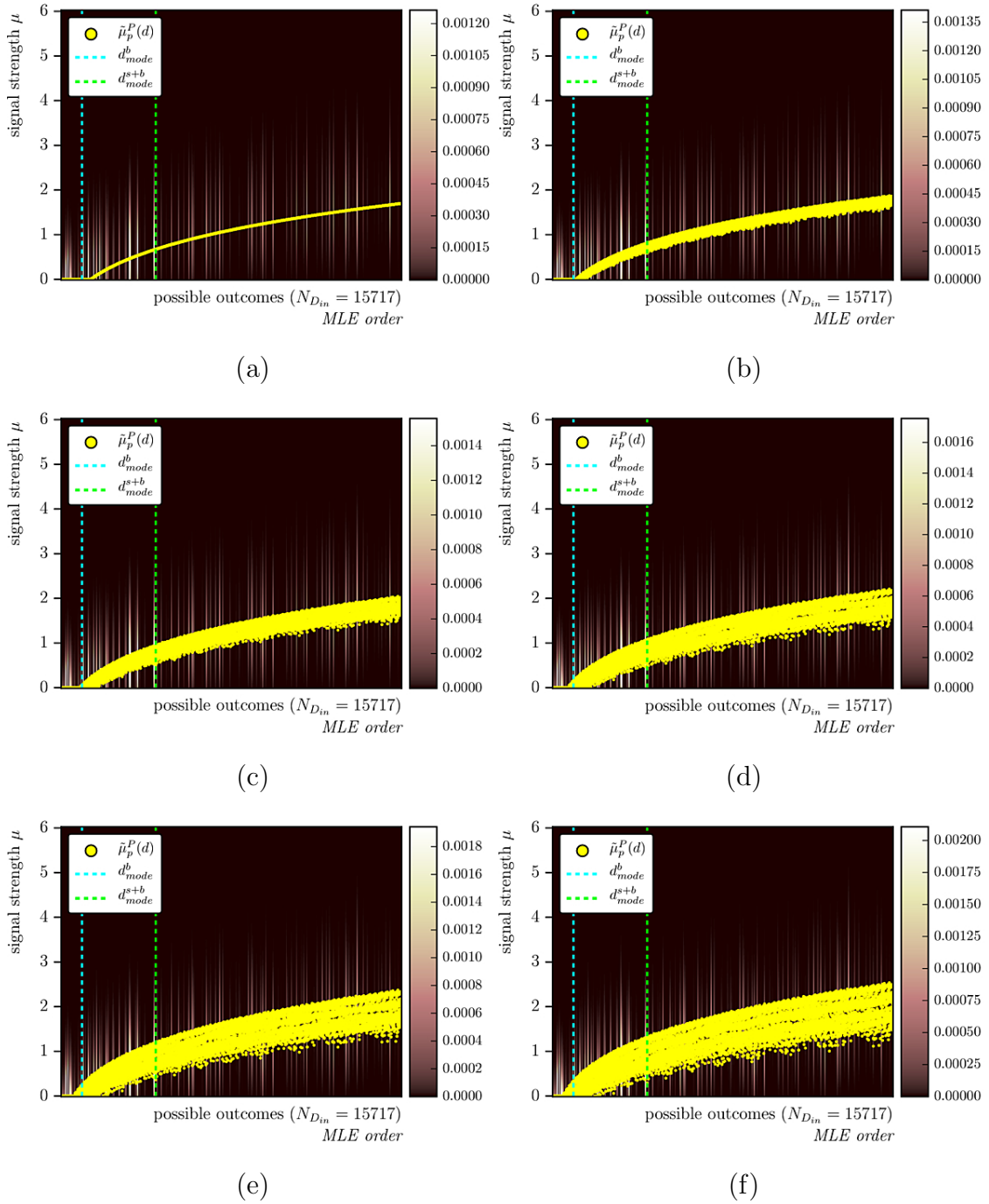


Figure 6.61: Example 3b (medium) – The impact of profiling nuisance parameters, ν, on the profiled likelihood functions. The model $p(d;\, \mu, \nu)$ is maximized to find $\tilde{\mu}_p^P$ (yellow circles) and $\tilde{\nu}_p^P$ but only allowing $\tilde{\nu}_p^P$ in the range $[-x\sigma,\, +x\sigma]$ of its Gaussian distribution. The plot of the likelihood function is made by then allowing μ to deviate from $\tilde{\mu}_p^P$ and plotting $p(d;\, \mu, \tilde{\nu}_p^P)$. (a) $x = 0$ (equivalent to $\nu = 0$), (b) $x = 1$, (c) $x = 2$, (d) $x = 3$, (e) $x = 4$, (f) $x = 5$.

6.5.3. Effect of Nuisance Parameters on Acceptance Regions

In this section, the impact of nuisance parameters on determining acceptance regions is examined using example 3b (medium). Example 3b is a four-binned multivariate Poisson model (see table 6.7 on page 329). There is a nuisance parameter for each bin to describe the uncertainty from limited simulated data for both the signal and background. This gives eight nuisance parameters. There are additional nuisance parameters for each type of systematic measurement uncertainty. Assuming only one type of systematic uncertainty gives nine total nuisance parameters.

In order to examine the effects of increasing systematic uncertainty, these nine nuisance parameters are assigned Gaussian prior distributions having equal variance, σ_ν^2. The acceptance region for $\mu = 1$ is constructed for four examples with increasing σ_ν to examine the effect of increasing uncertainty: $\sigma_\nu = 0.001\%$, 5%, 10%, and 25%. For example, when all nuisance parameters have 25% uncertainty and the value of the expected background count the first bin is 10, $b_1 = 10$, then b_1 has a $\pm 25\%$ fluctuation about its nominal value due to limited simulated data uncertainty, $b_1 = 10 \pm 2.5$, and an additional, uncorrelated $\pm 25\%$ fluctuation due to the single systematic uncertainty, $b_1 = 10 \pm 2.5$, giving a total uncertainty of $b_1 = 10 \pm 3.54$.

Figure 6.62 shows the distributions of the test statistics used to construct the $\mu = 1$ acceptance regions with the $D0$ test statistic, Q_{D0}. Each column corresponds to the nuisance parameter method, Q_{D0}^N (column 1), Q_{D0}^P (column 2) and Q_{D0}^F (column 3), and each row a value of nuisance parameter uncertainty, $\sigma_\nu = 0.001\%$ (row 1), 5% (row 2), 10% (row 3), and 25% (row 4). The first row for of figure 6.62 shows that for minuscule uncertainty, all three methods, nominal, profile and free, yield the same distributions for the signal and background-only models (and these are the same as those obtained using Q_{D0} without including nuisance parameters).

As mentioned previously, 100,000 Monte Carlo simulated outcomes are always used to generate each distribution of Q when constructing an acceptance region. Profiling for maximum likelihood with so many nuisance parameters is computationally demanding. The results here are only approximately profiled by constructing an equally-spaced grid for each nuisance parameter that ranges from -2.0 to 2.0 in steps of 0.2. This means that the profiled nuisance parameter was restricted to be within $\pm 2\sigma_\nu$ of its nominal value (zero), and examined at steps of $0.2\sigma_\nu$. The $\pm 2\sigma_\nu$ restriction is justified in particle physics because the Bayesian priors of the nuisance parameters are empirically determined, and greater than $2\sigma_\nu$ deviations from nominal values would be cause for suspicion. This is a necessary restriction because many of the enumerated outcomes, $d \in D_{in}$, in the various examples would reach maximum likelihood for absurdly large deviations from nominal nuisance parameter values.[120]

As σ_ν is increased, the ability to discriminate between the signal and background-only models decreases due to the increased spread in the test statistic distributions causing more overlap between signal and background-only models. This is quantified by examining how much of the background-only model probability is excluded by the 95% confidence level critical value, $p_{rej}^{\tilde{\mu}_0}$ (note that $p_{rej}^{\mu=1} \simeq 0.05$ is always guaranteed). For Q_{D0}^N (column 1), the

[120] An example of particle physics coverage studies with nuisance parameter values that are restricted to some range and not allowed to assume any possible value is found in [270].

background rejection begins at 0.522, but decreases to 0.512 with $\sigma_\nu = 0.05$, then to 0.462 for $\sigma_\nu = 0.1$, and finally 0.319 for $\sigma_\nu = 0.25$.

This loss of discrimination is more pronounced with profiling, Q^P_{D0} (column 2), which decreases to $p^{\tilde{\mu}_0}_{rej} = 0.261$ for $\sigma_\nu = 0.25$. Meanwhile, there appears to be no discernible change in discrimination power for Q^F_{D0} (column 2), which maintains nearly constant $p^{\tilde{\mu}_0}_{rej}$. Because these values are difficult to view in these figures, table 6.14 tabulates $p^{\tilde{\mu}_0}_{rej}$ for all methods. Q^Y_{FC} has less discriminating power (with Y any of the three nuisance parameter methods, N, P, or F), which is expected from the discussion in section 6.3.5 since the Feldman-Cousins test statistic produces smaller reporting risk at the expense of higher power loss risk, i.e., its confidence regions are more clustered around the physical MLE, $\tilde{\mu}_p$.

The general trend shown in table 6.14 is that Q^F_X shows the least loss of discrimination power with increasing σ_ν, while Q^P_X shows the most loss of discrimination (where X is any of the six test statistics, common or Arizona). This would imply that Q^F is a better method for defining test statistics in the presence of nuisance parameters. However, because Bayesian marginalization is used to construct acceptance regions, the final determination of their performance must be based on their coverage properties, i.e., how closely each meets the stated level of confidence level in the presence of nuisance parameters. Coverage properties are evaluated in the next section. Figures D.9–D.13 in appendix section D.3 provide the $\mu = 1$ signal and background-only distributions for the other five test statistics: Q_{FC}, Q_{AT}, Q_{AZ1}, Q_{AZ2}, and Q_{AZ3}, respectively. These exhibit the same trends as Q_{D0} (figure 6.62).

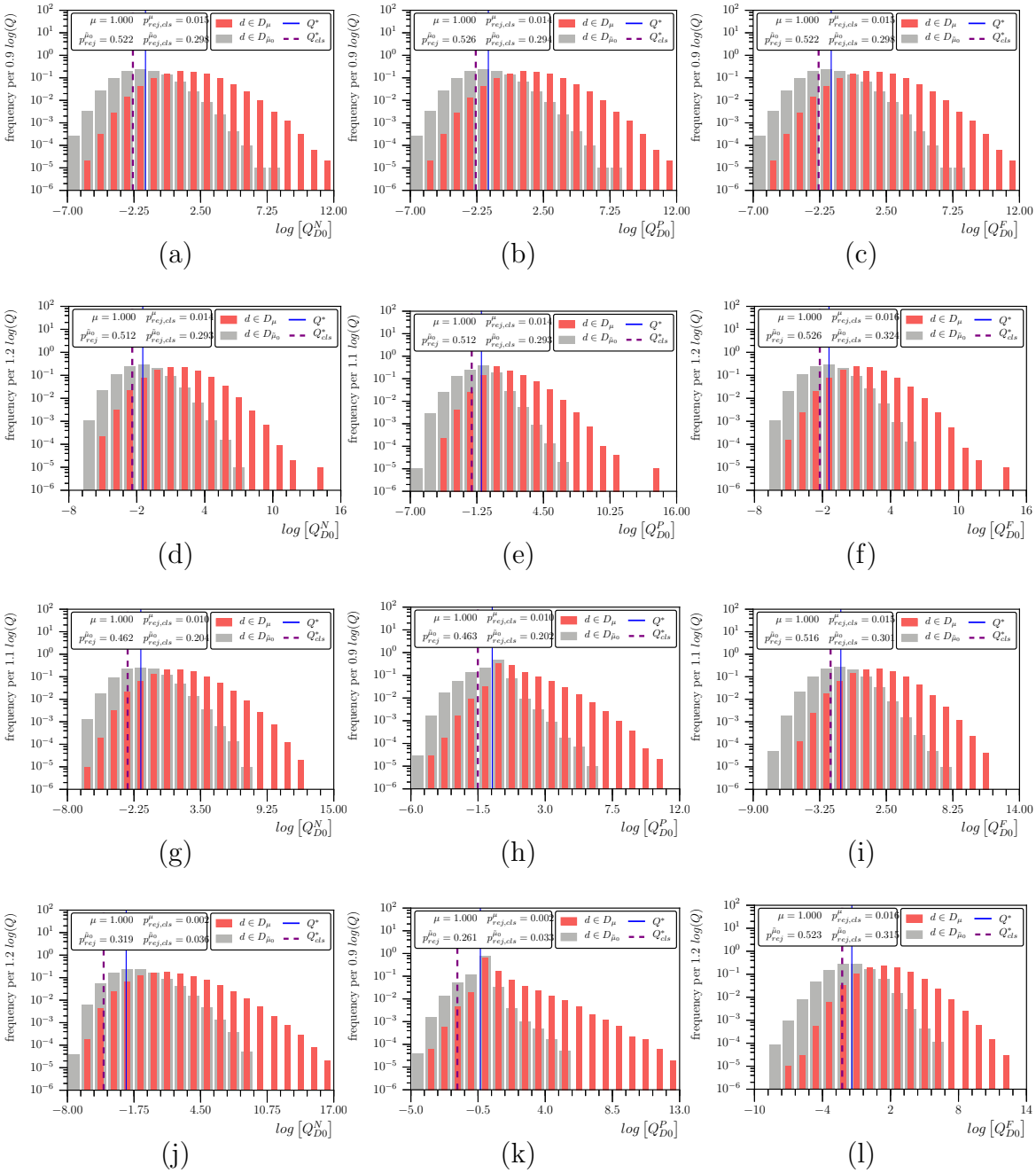


Figure 6.62: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{D0}, (b) Q^P_{D0}, and (c) Q^F_{D0}. With 5% uncertainty using (d) Q^N_{D0}, (e) Q^P_{D0}, and (f) Q^F_{D0}. With 10% uncertainty using (g) Q^N_{D0}, (h) Q^P_{D0}, and (i) Q^F_{D0}. With 25% uncertainty using (j) Q^N_{D0}, (k) Q^P_{D0}, and (l) Q^F_{D0}.

	$\sigma_\nu = 0.001\%$	$\sigma_\nu = 5\%$	$\sigma_\nu = 10\%$	$\sigma_\nu = 25\%$
		Feldman-Cousins type		
Q^N_{FC}	0.398	0.385	0.329	0.146
Q^N_{AZ1}	0.346	0.333	0.238	0.001
Q^P_{FC}	0.396	0.404	0.357	0.210
Q^P_{AZ1}	0.345	0.359	0.318	0.163
Q^F_{FC}	0.395	0.404	0.391	0.401
Q^F_{AZ1}	0.367	0.379	0.361	0.358
		D0 type		
Q^N_{D0}	0.522	0.512	0.462	0.319
Q^N_{AZ2}	0.518	0.505	0.463	0.278
Q^P_{D0}	0.526	0.512	0.463	0.261
Q^P_{AZ2}	0.522	0.508	0.456	0.203
Q^F_{D0}	0.522	0.526	0.516	0.523
Q^F_{AZ2}	0.518	0.522	0.511	0.513
		ATLAS type		
Q^N_{AT}	0.525	0.512	0.462	0.319
Q^N_{AZ3}	0.521	0.505	0.453	0.278
Q^P_{AT}	0.526	0.510	0.468	0.260
Q^P_{AZ3}	0.522	0.505	0.407	0.177
Q^F_{AT}	0.522	0.526	0.516	0.523
Q^F_{AZ3}	0.518	0.522	0.511	0.512

Table 6.14: Example 3b (medium) – The discrimination between the $\mu = 1$ signal and background-only models as measured by the total probability of the background-only model below the 95% confidence critical value, $p^{\tilde{\mu}_0}_{rej}$, in the presence of increasing systematic uncertainty.

6.5.4. Effect of Nuisance Parameters on Confidence-Acceptance Profiles

The bivariate Poisson model of example 2 is examined with the systematic uncertainties described in section 6.5.2: 5% bins systematic uncertainty for both signal and background expected yields, and a 10% global systematic uncertainty for both for signal and background. These are modeled with five nuisance parameters, ν, as described in section 6.2.1, and incorporated into the test statistics with each of the three methods described in section 6.5.1: the nominal method, Q^N, the profiling method, Q^P, and the full method, Q^F. The number of enumerated possible outcomes for example 2 was reduced from the previous value of 431 to 149 in order to reduce computational time. These 149 outcomes still comprise more than 99.9% of the possible outcomes using the background-only model probability, and 94.8% using the nominal signal model ($\mu = 1.0$).

Figures 6.63 (a)–(l) show the resultant confidence-acceptance profiles (CAPs) after incorporating ν for the test statistics common to particle physics, Q_{FC}, Q_{D0} and Q_{AT}, in the first, second and third columns of subfigures, respectively. The top row of subfigures, (a)–(c), shows the results of the Neyman construction without ν for reference.

The second row of subfigures, (d)–(f), shows the three results for the nominal method, Q^N_{FC}, Q^N_{D0} and Q^N_{AT}, and shows very little change from the first row. Careful inspection reveals that the 95% confidence intervals are slightly longer after including ν (boundary of blue area with black has moved slightly into the black region when compared to the first row results).

The third row of subfigures, (g)–(i), shows the results of profiling when the nuisance parameters are restricted to remain within $\pm 2\,\sigma$ of their nominal values. In figure (g), Q^P_{FC}, the red areas for the possible outcome denoting $\alpha > 0.95$ are shifted either up or down from their positions in figure (a). This is due to the spread in $\tilde{\mu}^P_p$ discussed in section 6.5.2.

The last row of subfigures, (j)–(l), shows the results of the full method with 1,000 Monte Carlo samples used for the second stage of marginalization to determine the test statistic value for each outcome, $Q^F(d,\mu) = \frac{1}{1000}\sum_i Q(d,\mu,\nu_i)$. For Q^F_{FC} (j), the large negative skewness of the signal test statistic distribution (shown in figure D.9) causes this averaging over test statistic values to fail to reach the higher ventiles. Though the full method still provides similar 95% confidence intervals, this effect of averaging invalidates any probabilistic interpretation of the CAPs. For Q_{AT} (c), coordinates (d,μ) below $\tilde{\mu}_p$ (yellow line) have the maximum test statistic value, $Q_{AT}(d,\mu) = 1$. For Q^F_{AT} (l), coordinates (d,μ) below but near to $\tilde{\mu}_p$ will sometimes have $Q^F_{AT}(d,\mu) < 1$, which partially breaks the degeneracy in the test statistic for coordinates below $\tilde{\mu}_p$.

Figures 6.64 (a)–(l) show the CAPs for the Arizona test statistics, Q_{AZ1}, Q_{AZ2} and Q_{AZ3}, in the first, second and third columns of subfigures, respectively. The inclusion of the $Q_{rep}(d,\mu) = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}$ term in all three test statistics causes them all to generate two-sided confidence intervals (discussed in section 6.3.2). As with the common test statistics shown in figure 6.63, there is little different in the structure of the CAPs between the nominal (without ν) of the first row, the nominal method in the second row, and the profile method of the third. The exception is that the profile method causes the up or down shifting of the CAPs.

The full method shown in the last row, (j)–(l), again demonstrates the difficulty in averaging over test statistic values. Unlike Q^F_{FC} shown in figure 6.63 (j), where negative

skewness caused a decrease in the average value, the inclusion of the Q_{rep} term causes an increase in the average value that leads to overestimating the acceptance ventile for coordinates near $\tilde{\mu}_p$. Again, the full method provides comparable 95% confidence intervals to the other methods. However, if one wishes to make full inferential use of the evidential information provided by CAPs, the full method cannot be chosen due to the instability in the averaging process.

The CAPs for these test statistics after implementation of the $\mathrm{CL_S}$ method are provided in appendix section D.3.2 for reference.

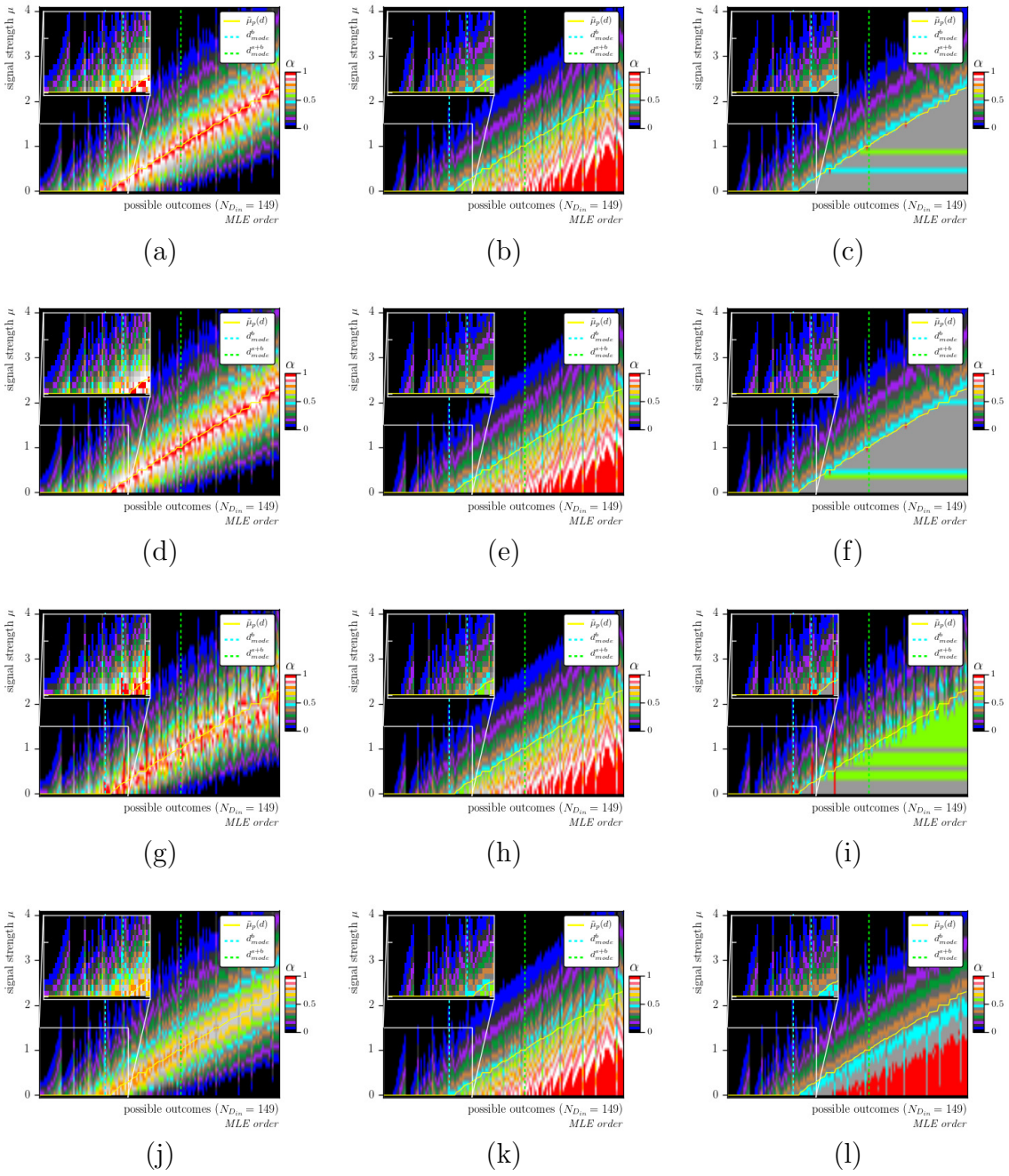


Figure 6.63: Example 2 (bivariate) – The confidence-acceptance profiles when nuisance parameters, ν, are included (shown with ventile precision). The first column shows results for Q_{FC}, the second for Q_{D0}, and third for Q_{AT}. The first row shows the results for Q without including ν, the second for the nominal method, Q^N, the third for profiling, Q^P, and fourth for full, Q^F.

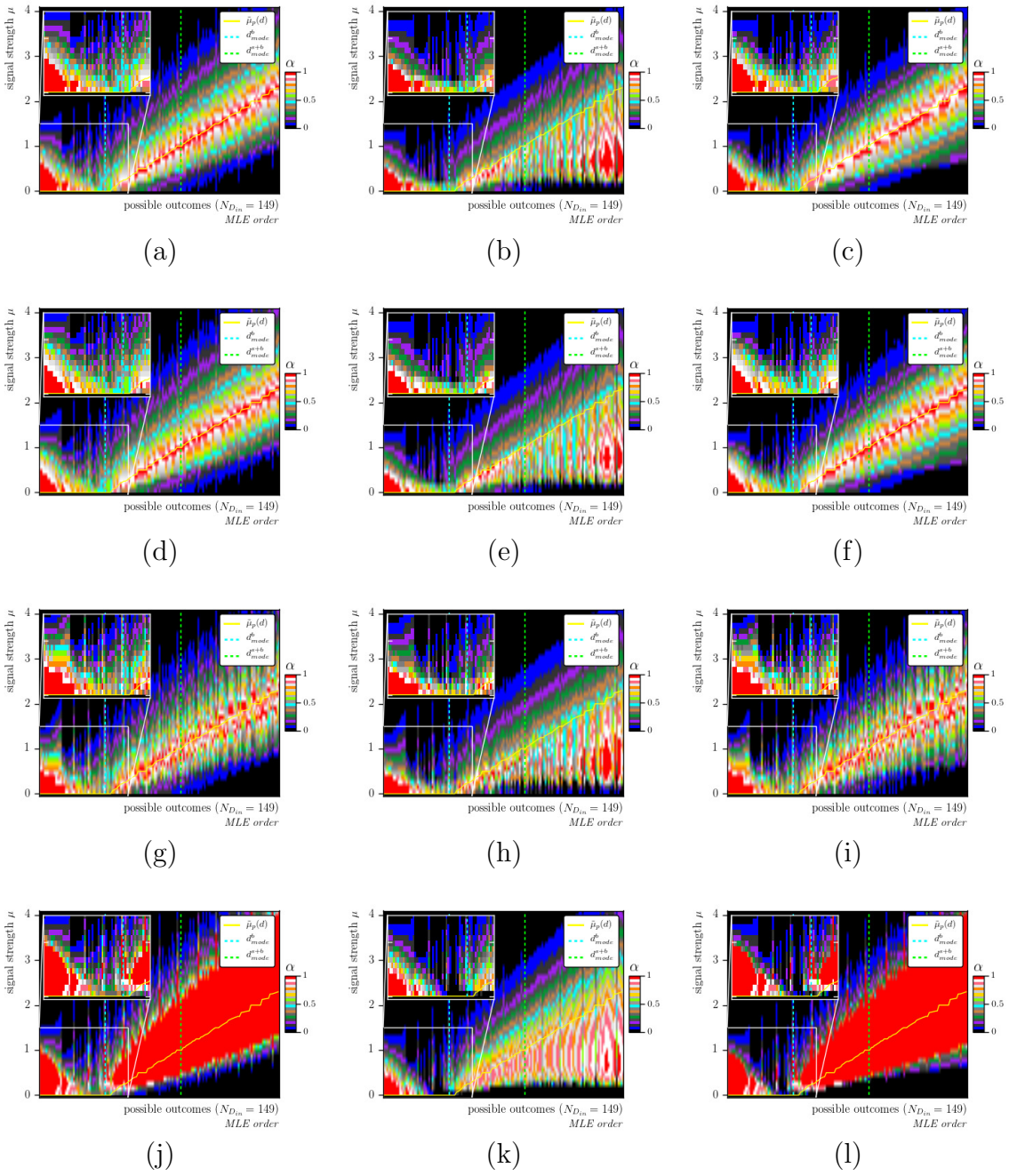


Figure 6.64: Example 2 (bivariate) – The confidence-acceptance profiles when nuisance parameters, ν, are included (shown with ventile precision). The first column shows results for Q_{AZ1}, the second for Q_{AZ2}, and third for Q_{AZ3}. The first row shows the results for Q without including ν, the second for the nominal method, Q^N, the third for profiling, Q^P, and fourth for full, Q^F.

6.5.5. Effect of Nuisance Parameters on Coverage

The validity of the stated confidence level, $(1-\alpha)$, delivered by the Neyman construction in the presence of nuisance parameters is evaluated by coverage studies. A confidence level of 95%, for instance, implies that for a given value of the parameter of interest, μ', the acceptance region for μ' contains at least 95% of the possible outcomes (by probability mass). This allows one to state that given any particular outcome, d', there is only no more than a 5% chance that d' lies in the rejection region for μ_{true}. This in turn implies that as experiments are performed many times, the long run frequency with which each μ_{true} lies outside the reported confidence intervals is 0.05. A coverage study examines changes to the stated confidence level when nuisance parameters are included in the statistical model.

Recall from section 6.5.1 that the test statistic $Q(d, \mu')$ used to determine the acceptance region for a particular value μ' of the parameter of interest may be defined in multiple ways with regard to the nuisance parameters ν.[121] For the nominal method, $Q^N(d, \mu')$, the nuisance parameter values, ν, are set to their nominal values (typically $\nu_i = 0$), while for the profiled method, $Q^P(d, \mu')$, they are set to their most likely values, $\nu = \tilde{\nu}^P$. The test statistic distribution is then determined by randomly sampling values of ν from their Bayesian prior probability distributions (Bayesian marginalization). For the full method, $Q^F(d, \mu', \nu')$, the test statistic value is determined by each specific set of nuisance parameter values, ν', that are sampled during marginalization. For any method, a critical value, $Q^*_\alpha(\mu')$, determines the acceptance region for that value μ' at the confidence level $(1-\alpha)$.

A difficulty with Bayesian marginalization arises in that each nuisance parameter in reality has only one unknown value, ν_{true}. For example, the measured luminosity may be stated as 20.3 ± 0.8 fb^{-1}, which implies a Gaussian spread of possible values due to measurement uncertainty, but the actual luminosity equals only one true value, even if unknown. If the true luminosity is 18.7 fb^{-1}, then the acceptance region determined for μ' via Bayesian marginalization over ν might contain less than $(1-\alpha)$ of the possible outcomes (by probability mass). This would cause the long run frequency with which each μ_{true} lies outside the reported confidence intervals to be larger than α.

Each nuisance parameter has a lower physical bound that prevents negative counts from being obtained. In the example of luminosity for instance, ν must be greater than -25.4 since one cannot have negative luminosity. If $\nu_{true} = -20$, then the coverage of the acceptance region for a value μ' is likely to be quite different from the stated $(1-\alpha)$. However, although such an extreme deviation of the nuisance parameter from its nominal value is mathematically permissible, it does not accurately represent the experimenter's knowledge of the systematic uncertainty that the nuisance parameter models. The restriction of $\nu \in [-2, 2]$ provides a more sensible range of allowed nuisance parameter values with which to investigate coverage properties since more extreme deviations would indicate a poor understanding of the systematic uncertainties of the experiment.[122] In other words, the coverage study performed here is a "2σ coverage study" because the nuisance parameter values examined lie within two standard deviations of their nominal values.

[121] Vector notation is suppressed with ν representing the vector of all nuisance parameters, $\nu = [\nu_i]$ with ν_i modeling the i^{th} systematic uncertainty.

[122] Or in the case of more than one nuisance parameter where $\nu = \{\nu_i\}$, the restriction of $\nu_i \in [-2, 2]$ for each i^{th} nuisance parameter.

Furthermore, a coverage study examines the coverage properties of each value of ν that may be ν_{true}. For example, the coverage of the acceptance region for a particular value μ' may be 97% for the value ν', but only 92% for ν''. In this case, if $\nu_{true} = \nu'$, then a stated confidence level of 95% is achieved (by overcovering), but if $\nu_{true} = \nu''$, then it is not (undercovers). It is useful to know what percent of the values of ν provide the stated level of confidence (or greater). If a 95% confidence level is desired, but only 50% of the possible values of ν generate acceptance regions for μ' that contain 95% of the outcomes or more (by probability mass), then the stated confidence level may not accurately reflect the true coverage (that belonging to whichever value is ν_{true}).

The coverage study thus proceeds by sampling many possible values of ν from a uniform distribution in order to find the percentage reaching the stated level of confidence (and thus are *not* sampled from their Gaussian prior distributions). If one permits values of ν deep within the Gaussian tails, then these extreme values of ν typically have very small coverage of the acceptance regions. This will cause the percent of ν values reaching the stated level of confidence to be very small, and imply that the stated level of confidence is entirely incorrect. However, extreme values of ν are unrealistic, and hence the use of the 2σ coverage study. Such a coverage study is not definitive since it is always possible for ν_{true} to lie more than two standard deviations from its nominal value, but it does describe the general coverage properties that may be expected if the modeling of systematic uncertainties is accurate.

Monte Carlo simulation is used in two steps in order to determine the coverage properties of the acceptance region for a given value μ'. First, a nuisance parameter value, ν', is randomly sampled to give the generative model $p(d;\, \mu', \nu')$.[123] This model is then treated as if it were the true generating distribution, $p(d;\, \mu_{true}, \nu_{true})$. This distribution is then used in a second step of simulation to randomly generate the possible outcomes, d, and this gives the distribution of the test statistic, $p(q;\, \mu', \nu')$. The Neyman construction with Bayesian marginalization has already been performed so that one already knows the critical value, $Q^*_\alpha(\mu')$, that determines the acceptance region of μ' (i.e., the coverage study is performed after the original Neyman construction). Hence the percent coverage of μ' for this sampled ν' is given by

$$K_\alpha(\mu', \nu') \equiv 100\% \cdot \sum_{q > Q^*_\alpha(\mu')} p(q;\, \mu', \nu'). \tag{6.5.6}$$

The first step of Monte Carlo sampling is repeated many times so that a distribution of $K_\alpha(\mu', \nu'_j)$ is generated, where ν'_j is the j^{th} sampled set of nuisance parameter values. From this distribution, one may obtain various percent coverage statistics such as mean, median, etc.

The coverage study performed here uses the following specific details. The bivariate Poisson model of example 2 is used (with expected background and signal yields of $b_i = [\,4, 3\,]$ and $s_i = [\,1, 3\,]$ given in table 6.7 on page 329), which demonstrates the simplest case of a multivariate Poisson distribution. A relative $\pm 5\%$ systematic uncertainty is assumed for each

[123] For the coverage study, the set of nuisance parameter values are sampled from the uniform distribution, $\nu_i \in [-2, 2]$, and *not* the Bayesian prior distribution for ν_i that is used in the original Neyman construction to determine the critical value, $Q^*_\alpha(\mu')$. In frequentist statistics, ν is not a random variable, and one seeks to evaluate what the percent coverage would be if $\nu_{i,true}$ was any of the allowed values in the range $[-2, 2]$.

of the two discriminating variable bins for both background and signal expected yields. This is modeled by four nuisance parameters with identical independent Gaussian Bayesian prior distributions, $\mathcal{N}(0, 0.05^2)$. An additional $\pm 10\%$ global systematic systematic uncertainty is applied to both background and signal expected yields, modeled by an additional fifth independent nuisance parameter with $\mathcal{N}(0, 0.10^2)$.[124] These systematic uncertainties are similar in magnitude to those often observed in particle searches.

[124] Global systematic uncertainties are usually of primary interest in a particle search, and there are typically dozens of them (luminosity, jet energy resolution, etc.). Sometimes the Neyman construction is performed with the greatest global systematic uncertainty in order to obtain approximate results for intermediate analysis (but before unblinding). This motivates the use of a single global systematic uncertainty here.

6.5.5.1. Coverage Properties of Q^N_{FC}

The 95% confidence level coverage properties of ten values of μ are examined, from $\mu' = 0.2$ to 2.0 in steps of 0.2. This reduces the computational expense while providing the coverage properties for the range of μ of primary interest. For each μ', 1,000 sets of sampled nuisance parameter values of $\nu = \{\nu_i\}$ are generated from the uniform distribution discussed above. For each sampled ν', 1,000 possible outcomes d are simulated and provide the percent coverage, $K_{0.05}(\mu', \nu')$. Several statistics of the distribution of $K_{0.05}(\mu', \nu)$ for the 1,000 sampled ν are reported.

Figure 6.65 shows the result for the Feldman-Cousins test statistic using the nominal method, Q^N_{FC}. Recall from section 6.5.1 that the nominal method defines the test statistic using the nominal values of ν, typically $\nu_i = 0$, so that $Q^N_{FC}(d, \mu) = p(d;\, \mu, \nu = 0)/p(d;\, \tilde{\mu}_p, \nu = 0)$. A heat map is used to present the plotted values. For example, a value of 0.5 appears as a white area. The x-axis shows the values of μ for which the coverage properties were examined. The y-axis shows the percent coverage from 90% to 100% in steps of 1%, with coverage lower than 90% grouped together. For example, if 500 of the 1,000 simulated nuisance parameters attained 92% coverage of the acceptance region for $\mu = 0.2$, then a value of $500/1000 = 0.5$ would be placed at the coordinates $(\,0.2,\, [92, 93]\,)$, which would appear there as a white rectangle. A green line at 95% denotes the stated confidence level. In the absence of nuisance parameters, coverage would be exact and a value of $1000/1000 = 1.0$ would be placed at $y = [95 - 96)$ for each value of μ.

For each value of μ examined, the frequency achieved by the 1,000 sampled percent coverages, $K_{0.05}(\mu, \nu)$, is shown with a heat map in the vertical column above that value of μ. For example, the orange cell in the bottom row above $\mu = 1.0$ indicates that for the distribution of $K_{0.05}(\mu = 1.0, \nu)$, approximately 25% of the sampled ν ($\sim$250 of 1,000) provide less than 90% coverage of $\mu = 1.0$. In other words, for each of those $\sim$250 sampled ν, less than 90% (by probability mass) of the test statistic values of the 1,000 simulated outcomes lie in the $\mu = 1.0$ acceptance region. If magnitudes of the systematic uncertainties were decreased toward zero, the distribution of $K_{0.05}(\mu, \nu)$ would become increasingly clustered around the stated confidence level of 95% until providing the exact 95% coverage.

In addition to the heat map showing the distribution of $K_{0.05}(\mu, \nu)$, the mean percent coverage for each μ is shown with a blue circle, $\bar{K}_{0.05}(\mu) = \frac{1}{1000} \sum_{j=1}^{1000} K_{0.05}(\mu, \nu_j)$. Also, a vertical blue line stretches between the 25^{th} and 75^{th} quantiles of $K_{0.05}(\mu, \nu)$ (with horizontal blue hashes marking the end points), and an intermediate horizontal blue hash shows the median value of $K_{0.05}(\mu, \nu)$. A median percent coverage larger than the mean percent coverage implies that the distribution of $K_{0.05}(\mu, \nu)$ is skewed toward lower values (and vice versa).

Two values are printed at the top of the vertical column for each value of μ. The first is the frequency (as a percentage) that $K_{0.05}(\mu, \nu)$ lies above the stated confidence level of 95%. The second (in parentheses) is the percent coverage difference between the 25^{th} and 75^{th} quantiles of $K_{0.05}(\mu, \nu)$, which provides a measure of the spread of the distribution of $K_{0.05}(\mu, \nu)$.

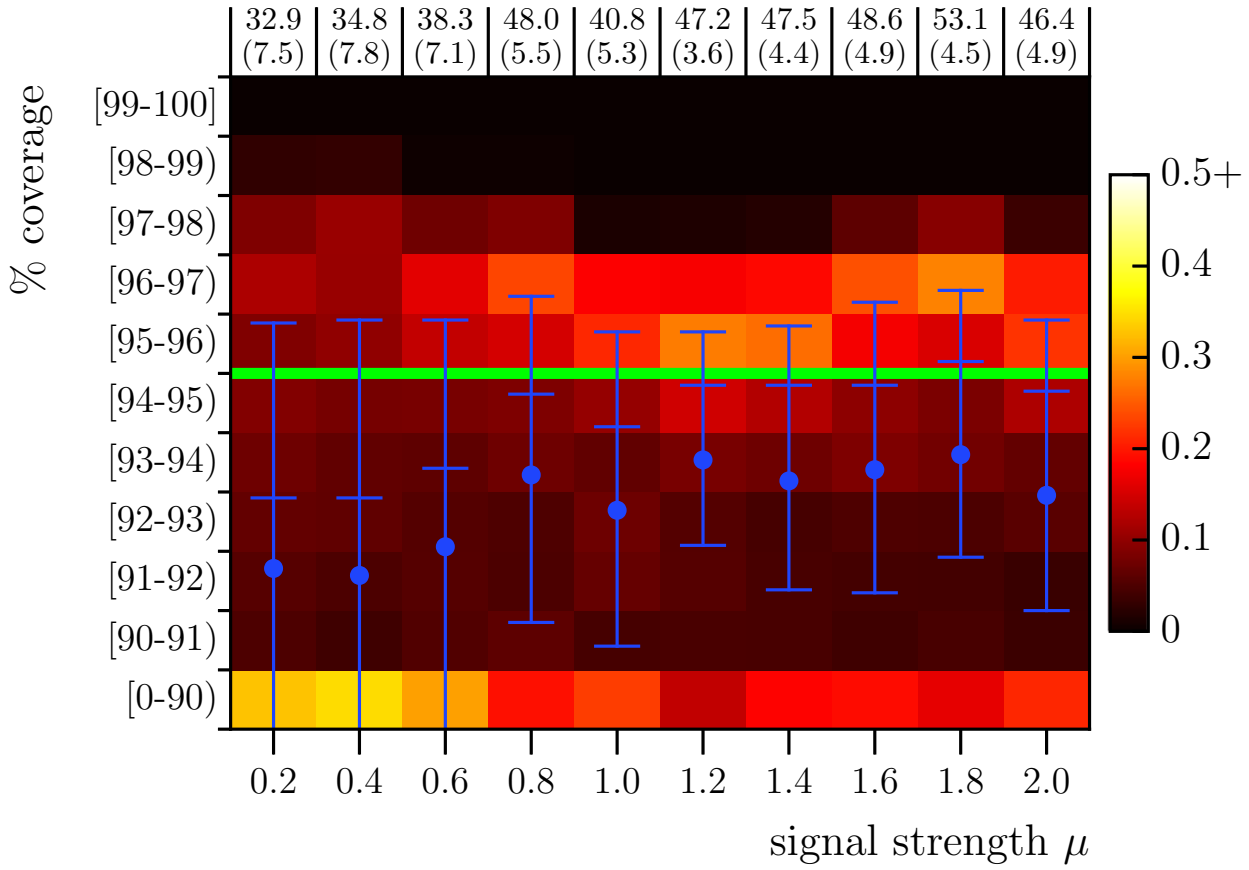


Figure 6.65: Example 2 (bivariate) – The 2σ coverage study for the Q^N_{FC} test statistic. The y-axis measures the percent coverage with a horizontal green line delineating the stated confidence level, 95%. Each column along the x-axis gives the coverage properties for that value of μ. For each μ, 1,000 values of ν are randomly sampled and the percent coverage of each is obtained to generate the distribution of the percent coverage for that value of μ. This distribution of the 1,000 percent coverages is shown with a heat map giving the frequency (out of 1,000) with which each percent coverage was observed (for each μ-column). A small spread near 95% would be ideal. A blue circle gives the mean percent coverage of the 1,000 sampled percent coverages. A vertical blue line denotes the 25% through 75% quantiles of the 1,000 sampled percent coverages, with an intermediate horizontal blue hash showing the median (usually near the mean blue circle). Above each column are two values: the first is the total frequency of the 1,000 above 95% coverage, and the second in parentheses gives the difference between the 75% and 25% percentiles (i.e., a measure of spread). Ideally, 100% of the percent coverages would be above the stated confidence level of 95% with minimum spread.

The 2σ coverage study for Q_{FC}^N shown in figure 6.65 indicates that there is undercoverage for the majority of the nuisance parameter values sampled from their uniform 2σ distributions. The frequency of ν achieving 95% coverage (top value above each column) is less than half for all values of μ except for $\mu = 1.8$, which has 53.1% (531 of 1,000) reaching at least 95% coverage. This is particularly pronounced for the lower values of μ, with less than a third (32.9%) of the sampled ν reaching at least 95% coverage. For $\mu = 0.4$, the heat map for the cell in the bottom row (yellow) indicates that ∼40% of the sampled ν do not even reach 90% coverage.

For all μ, the mean percent coverage (blue circle) is at least 1% lower than the median percent coverage (intermediate horizontal blue hash) indicating that the distribution of percent coverage is skewed toward lower values. Finally, the second value listed at the top of each column (in parentheses) shows that the spread of the percent coverage distribution is quite large for most μ, especially for lower values of μ with spreads over 7%. A small spread in percent coverage is useful because it may allow the experimenter to simply restate the confidence level. For example, if the percent coverage for each value of μ was tightly clustered around 90%, then one could consider restating the confidence level as 90% rather than the initial 95%. This may not be an option with the large spread for Q_{FC}^N.

6.5.5.2. Coverage Properties of Q_{FC}^N with the CL_S Method

Figure 6.66 shows the 2σ coverage study for Q_{FC}^N after implementing the CL_S method described in sections 6.2.3.4 and 6.2.4 (see page 289) that purposefully overcovers acceptance regions when the signal-background discrimination power is small (i.e., when the evidence for signal nonexistence is weak, see section 6.4.6). The effect of the CL_S method is dramatic for smaller μ, yielding complete overcoverage (100% percent coverage for 100% of the sampled ν).

As μ increases above 1.2, the mean of the percent coverage distribution decreases from 100%, and the distribution increasingly widens and becomes more skewed to lower values. This behavior was observed for all test statistics and nuisance parameter methods (see appendix section D.3.3).

The CL_S method causes complete overcoverage for small μ when there is little signal-background discrimination, while it has no effect on the acceptance for large μ where discrimination is large. Therefore, for large enough μ, the coverage properties will be the same whether or not the CL_S method is implemented.

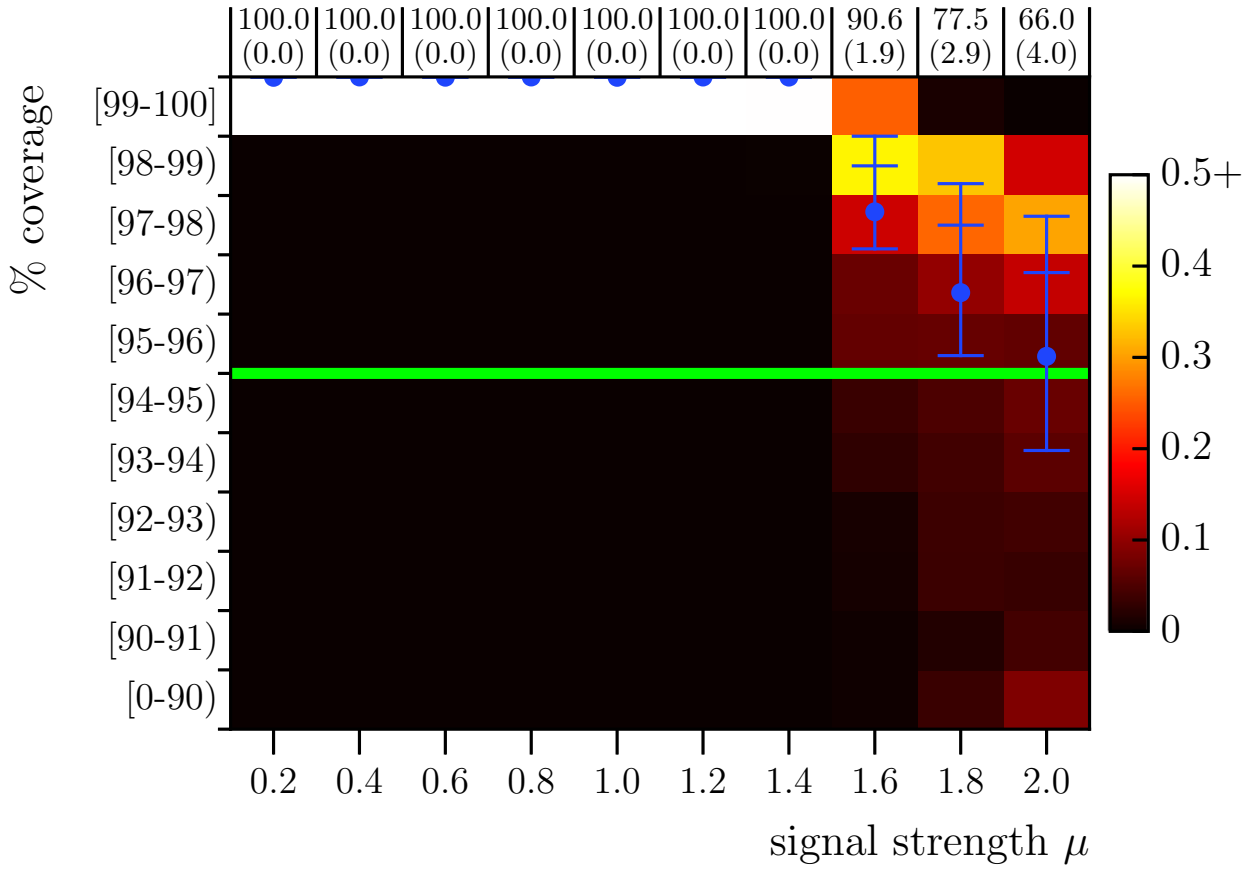


Figure 6.66: Example 2 (bivariate) – The 2σ coverage study for the Q^N_{FC} test statistic also implementing the $\mathrm{CL_S}$ method. The y-axis measures the percent coverage with a horizontal green line delineating the stated confidence level, 95%. Each column along the x-axis gives the coverage properties for that value of μ. For each μ, 1,000 values of ν are randomly sampled and the percent coverage of each is obtained to generate the distribution of the percent coverage for that value of μ. This distribution of the 1,000 percent coverages is shown with a heat map giving the frequency (out of 1,000) with which each percent coverage was observed (for each μ-column). A small spread near 95% would be ideal. A blue circle gives the mean percent coverage of the 1,000 sampled percent coverages. A vertical blue line denotes the 25% through 75% quantiles of the 1,000 sampled percent coverages, with an intermediate horizontal blue hash showing the median (usually near the mean blue circle). Above each column are two values: the first is the total frequency of the 1,000 above 95% coverage, and the second in parentheses gives the difference between the 75% and 25% percentiles (i.e., a measure of spread). Ideally, 100% of the percent coverages would be above the stated confidence level of 95% with minimum spread.

6.5.5.3. Coverage Properties of Q^P_{FC} and Q^F_{FC}

Figures 6.67 (a) and (b) show the 2σ coverage studies for the profiled Feldman-Cousins test statistic, Q^P_{FC}, and the full method, Q^F_{FC}, respectively. The percent coverage distributions of each μ for Q^P_{FC} shown in figure 6.67 (a) are similar to those of Q^N_{FC} shown in figure 6.65, and this similarity between the null and profile methods was observed for the other test statistics, Q_{D0}, Q_{AT}, Q_{AZ1}, Q_{AZ2}, and Q_{AZ3} (reported in appendix section D.3.3).

The coverage study result for the full method shown in figure 6.67 (b) exhibits intriguing coverage properties. Here the percent coverage distributions of each μ are tightly clustered around the stated confidence level of 95%, with percent coverage means very near to 95%. This is highly desirable behavior since one wishes that the stated confidence level applies to any possible value of ν_{true}.

However, the second stage of Bayesian marginalization is sensitive to the shape of the test statistic distribution, and the shape is different for the different test statistics one might choose to use. Recall figure 6.59 (b) (page 391) that shows that if the test statistic is a function of ν, then one must average $Q(d', \mu', \nu)$ over ν to obtain $Q(d', \mu')$ in order to determine if possible outcome d' is in the acceptance region of μ'. The weighting by the Bayesian prior distribution is used to construct this average. Also, figure 6.63 (j) (page 408) shows that the CAPs of Q^F_{FC} do not often reach the 90^{th} quantile (very little red area surrounding $\tilde{\mu}_p$). This is because the average of $Q_{FC}(d', \mu', \nu)$ over many ν for the given outcome d' is usually less than the highest values of $Q_{FC}(d, \mu', \nu')$ attained by some other outcomes d with certain values of ν'. Meanwhile, figures 6.64 (j)–(l) (page 409) show that the shapes of the Arizona test statistics' distributions cause the opposite effect with rather large values for the average (large red areas indicating that the test statistic averages of many outcomes reach the 90^{th} quantile of the acceptance regions for many μ).

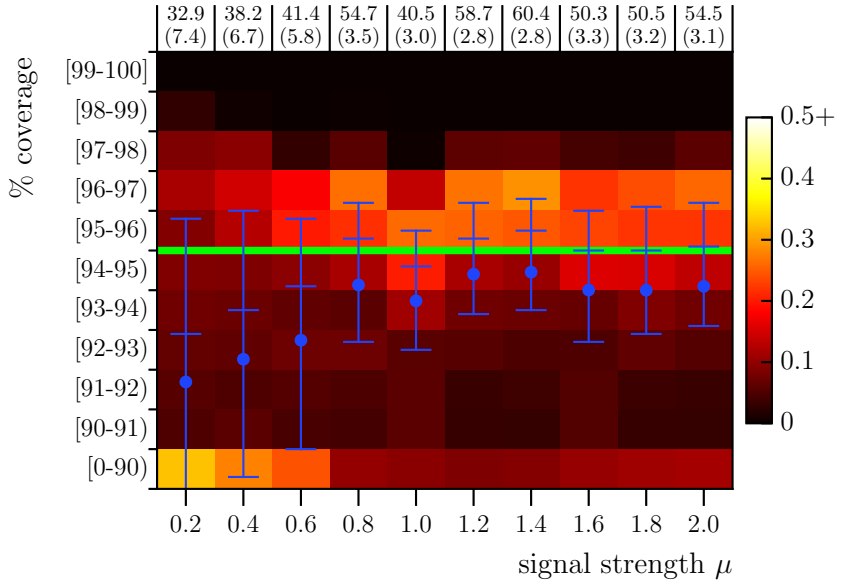
Therefore, the second stage of Bayesian marginalization when using the full method does not necessarily provide accurate quantiles for the acceptance region of each μ. For example, the highest acceptance ventile for the $\mu = 1.0$ acceptance region (horizontal row) using Q^F_{FC} shown in figure 6.63 (j) does not appear to contain 5% of the outcomes (by probability) because there is very little red area shown in that horizontal row. This would seemingly violate the Neyman construction method, which relies on the acceptance quantiles having a uniform probability distribution.[125] The discrepancy is due to determining critical values for the ν-dependent test statistic, $Q^F_{FC}(d, \mu, \nu)$, and then using these to determine the acceptance quantile for the ν-averaged test statistic, $Q^F_{FC}(d, \mu)$.

Despite this discrepancy, the coverage of the acceptance region of each μ for *nearly every* specific sampled value of ν appears to be very close to the stated confidence level. At this point, the experimenter must question the ultimate goal of the inference procedure.

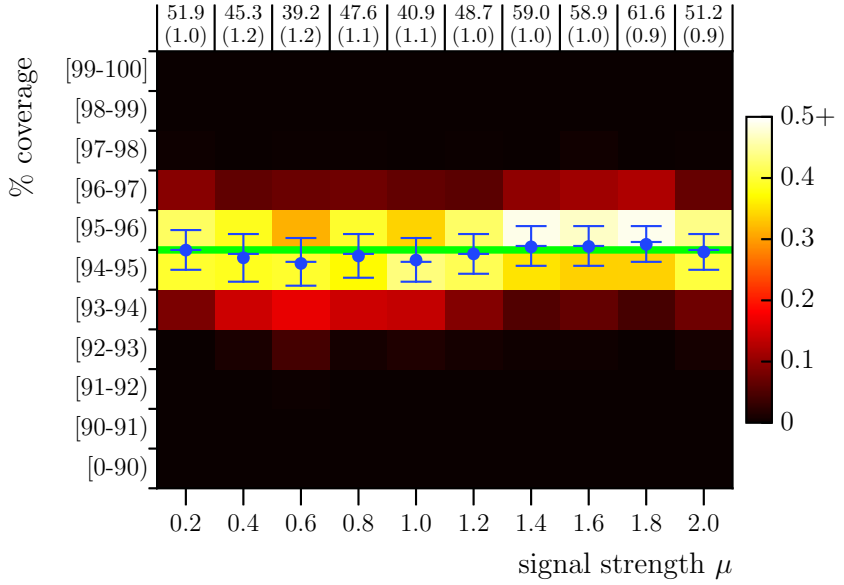
1. Is one searching for the confidence interval that accurately provides 95% coverage of μ after averaging over the possible values of the nuisance parameters?
2. Or is one searching for the confidence interval that accurately provides 95% coverage of μ for each particular ν that may happen to be ν_{true}?

The second seems more appropriate for frequentist analysis. Therefore, the full method used in the treatment of nuisance parameters seems to provide proper frequentist coverage if one

[125] This is a simple property of quantiles. For example, the probability to be in the upper ventile should be 5%.



(a)



(b)

Figure 6.67: Example 2 (bivariate) – The 2σ coverage studies for (a) the profiling method, Q^P_{FC}, and (b) the full method, Q^F_{FC}. The result of profiling with Q^P_{FC} is similar to that of the nominal method, Q^N_{FC}, of figure 6.65. Use of the free method, Q^F_{FC}, results in a coverage distribution with a narrow spread around the stated confidence level.

considers each possible ν separately rather than in aggregate via Bayesian marginalization *even though* Bayesian marginalization is used in the full method.

6.5.5.4. Coverage Properties of Q^N_{D0}, Q^P_{D0} and Q^F_{D0}

Figures 6.68 (a)–(c) show the coverage studies for the D0 test statistic using the three nuisance parameter methods, Q^N_{D0}, Q^P_{D0} and Q^F_{D0}, respectively. Figures 6.68 (a) and (b) demonstrate significant overcoverage when using Q^N_{D0} or Q^P_{D0} with the percent coverage distributions located mostly above the stated confidence level (green line). In fact, the vertical blue lines showing the 25^{th} to 75^{th} quantiles lie entirely above 95%. This is the opposite result from the Feldman-Cousins test statistic that exhibited significant undercoverage when using the nominal and profiled methods, Q^N_{FC} and Q^P_{FC} (figures 6.65 and 6.67 (a)). Also, the percent confidence distributions for Q^N_{D0} and Q^P_{D0} have smaller spread than Q^N_{FC} or Q^P_{FC}, which would suggest that coverage for the D0 test statistic is better than Feldman-Cousins when nuisance parameters are included in the statistical model. Figure 6.68 (c) again demonstrates very accurate coverage when using the full method, Q^F_{D0}.

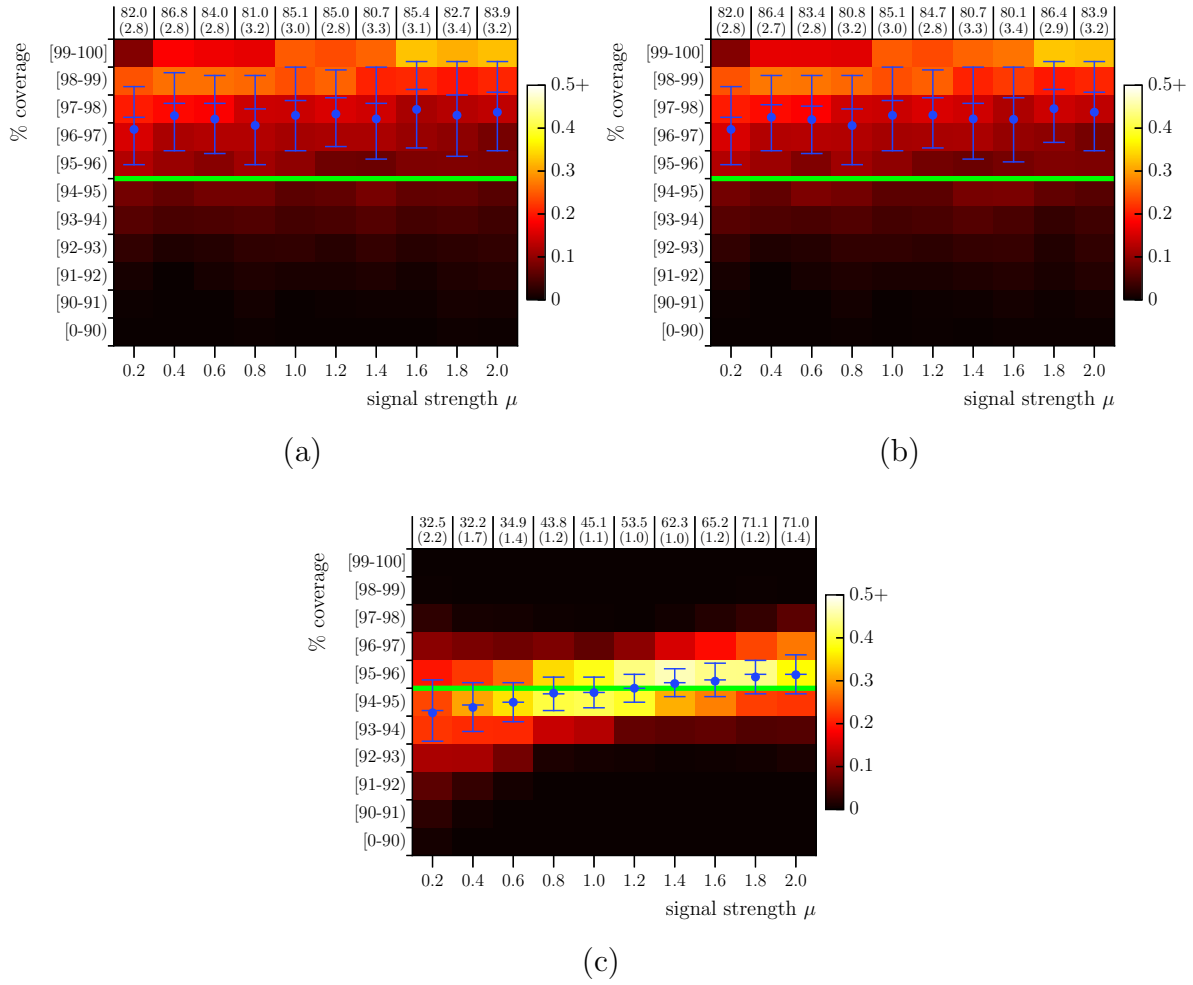


Figure 6.68: Example 2 (bivariate) – The 2σ coverage studies for (a) the nominal method, Q^N_{D0}, (b) the profiling method, Q^P_{D0}, and (c) the full method, Q^F_{D0}. Unlike the results using Q^N_{FC} and Q^P_{FC} that exhibited significant undercoverage, figures (a) and (b) show significant overcoverage. As with Q^F_{FC}, (c) shows that the full method using Q^F_{D0} provides more accurate coverage.

6.5.5.5. Summary of Coverage Properties

The remaining coverage studies for the other test statistics are provided in appendix section D.3.3 (page 544).[126] The results for all of the test statistics are summarized in table 6.15. These results show that when systematic uncertainties (modeled by nuisance parameters) are a consideration, the experimenter may choose the test statistic and method of nuisance parameter treatment based on their coverage properties. When undercoverage for any value of μ is the primary concern, $Q_{FC}^{N\,or\,P}$ and $Q_{AZ1}^{N\,or\,P}$ should be avoided. If undercoverage for small μ can be tolerated, Q_{AZ2}^{F} and $Q_{AZ3}^{N,P\ or\ F}$ may be acceptable. The test statistics $Q_{D0\ or\ AT}^{N\,or\,P}$ and $Q_{AZ2}^{N\,or\,P}$ always overcover, but have moderate spread. The methods that provide coverage best approximating the stated confidence level with small spread are $Q_{FC,\,D0\ or\ AT}^{F}$ and Q_{AZ1}^{F}.

In [288], David Cox and Nancy Reid complained that the use of the CL_S method was unlikely to have satisfactory statistical properties. In section 6.4.6.1 (page 359), it was shown that the CL_S method could be explained as estimating intervals based on the signal significance ratio function. The results of this section show that the CL_S method by design always overcovers lower values of μ with 100% coverage, and decreases toward the non-CL_S coverage result as μ increases. Therefore, the CL_S method may be useful to correct methods that undercover for lower values of μ, but otherwise provide accurate coverage for higher values, such as Q_{AZ2}^{F} and $Q_{AZ3}^{N,P\ or\ F}$.

[126] The conclusions drawn from these coverage studies are solely based on the bivariate Poisson model of example 2 and for only one example set of systematic uncertainties.

test statistic	spread of percent coverage distributions	location of percent coverage distributions
	common test statistics	
$Q_{FC}^{N\ or\ P}$	large	low → undercoverage
$Q_{D0\ or\ AT}^{N\ or\ P}$	moderate	high → overcoverage
$Q_{FC,\,D0\ or\ AT}^{F}$	small	centered at stated confidence level
	Arizona test statistics	
$Q_{AZ1}^{N\ or\ P}$	large	low → undercoverage
Q_{AZ1}^{F}	small	centered at stated confidence level
$Q_{AZ2}^{N\ or\ P}$	moderate	high → overcoverage
Q_{AZ2}^{F}	small μ: large large μ: small	small μ: low → undercoverage large μ: centered at stated confidence level
$Q_{AZ3}^{N\ or\ P}$	small μ: large large μ: moderate	small μ: low → undercoverage large μ: high → overcoverage
Q_{AZ3}^{F}	small μ: large large μ: small	small μ: low → undercoverage large μ: centered at stated confidence level
	CL_S method	
any Q	small μ: small	small μ: 100%, total overcoverage
	increasing μ: reversion to non-CL_S result	

Table 6.15: Example 2 (bivariate) – Summary of the results of the 2σ coverage studies (most results presented in appendix section D.3.3 on page 544). The spread of the percent coverage distribution for a given value of μ is measured by the difference between the 75^{th} quantile percent coverage and the 25^{th} quantile (smaller is desirable). The location of the percent coverage distribution for a given μ is given by either the mean of the distribution. The spreads and locations for the percent coverage distributions over the range of examined μ are usually consistent, but it is noted in the table when there is μ-dependence.

6.6. Conclusion to Statistical Inference in a Signal Search

Many concepts were explored in this chapter.

1. Frequentist inference using the Neyman confidence region construction (NCRC).
2. Properties of the confidence intervals produced by various test statistics including the evaluation of different types of risk.
3. The utility of confidence-acceptance profiles (CAPs) produced by the NCRC in providing inference statements that give frequentist evidential support for each value of the parameter of interest, μ, rather than simply dividing the values of μ into two sets: accept or else reject. And also the utility of CAPs for frequentist-based meta-analysis.
4. The analysis of three Bayesian marginalization methods for the treatment of nuisance parameters when using the NCRC.

A brief summary discussion for each of these is provided in the following subsections.

6.6.1. Classical Inference to Explain, After Machine Learning Predicts

In 2001, Leo Breiman challenged the academic statistics community to rely more on algorithmic modeling and less on data modeling [225]. By this, he meant that generative statistical models, such as the multivariate Poisson model studied here, were often not likely to accurately capture the real world phenomena they were designed to model. He opined that models constructed from machine learning algorithms, although usually incomprehensible to human intelligence, provide more accurate predictions of future measured data than is typical of statistical models designed by humans. At the heart of this discussion is whether or not the goal of statistical inference is to formulate the most accurate *predictive* model (the "black box" obtained with machine learning) or to inductively infer the best *explanatory* model (the statistical models designed and analyzed by humans).

A particle physics search often utilizes both methods of inference: the predictive conditional models derived from data with machine learning, and the explanatory generative statistical models designed by experimenters. First, a dataset obtained by the ATLAS experiment at the Large Hadron Collider (either through actual measurement or else simulation) is ultimately an enormous collection of collision events, e_i (for the i^{th} event), with hundreds of discriminating variables available for each event, $\mathbf{x}_i = x_{i,j}$ (for the j^{th} discriminating variable). The goal of the particle search is determine if some of these events arise from new signal processes (theories beyond the Standard Model) or else are solely due to the Standard Model background processes. Thus, one wishes to predict if e_i is a signal event or else if e_i is a background event. This is accomplished by using simulated data to construct a machine learning predictor function, $f_{ML}(\mathbf{x_{i,j}}) = y_i$, which gives the classifier response value y_i (see figure 5.7 (a) on page 136). This may be viewed as constructing the conditional model, $p(y \,|\, \mathbf{x})$.

Once the classifier response value of each event is predicted by the (incomprehensible) machine learning model, $f_{ML}(\mathbf{x_i}) = y_i$, one must then make sense of these predictions to determine the validity of any claims of signal existence or signal nonexistence. In other words, one wishes to *explain* these predictions. This is accomplished with a generative parametric statistical model that describes the probability of obtaining the machine learning

prediction given a signal strength parameter, $p(y;\mu)$. The spectrum of y is then binned so that the counts in each bin, $d = [d_1, d_2, \ldots, d_n]$ (for n bins), become the final variables used to model the possible outcomes of the experiment, $p(d;\mu)$. The multivariate Poisson distribution models these bin counts, and one uses it in the Neyman construction to provide an inference result for the signal strength.[127]

Ultimately, chapter 5 concerns the exploration of novel machine learning methods for a signal search in particle physics. This chapter explores novel methods of frequentist inference that are performed after machine learning has made its predictions in order to explain the evidence for signal existence and for nonexistence provided by experimental measurements.

6.6.2. Selection of Test Statistics

The central feature of the Neyman confidence region construction (NCRC) method is the use of a local ordering rule, $Q(d,\mu)$, that defines a measure of the strength of evidence supporting the signal model with strength μ for different possible outcomes d.[128] The selection of the local ordering rule is often achieved by borrowing a test statistic, $Q(d,\mu)$, from Neyman-Pearson hypothesis testing (NPHT). A test statistic in NPHT measures the strength of evidence provided by d for the null hypothesis against the strength of evidence for the competing alternative hypothesis. Likelihood ratios are often used as the test statistic in NPHT since the Neyman-Pearson lemma guarantees that these provide the strongest discrimination power between the null and alternative hypothesis.

There are three likelihood ratio test statistics commonly used in particle physics as local ordering rules in NCRC: the Feldman-Cousins test statistic, Q_{FC}, the DZero recommended test statistic, Q_{D0}, and the ATLAS recommended test statistic, Q_{AT}. That there are three different commonly used test statistics must imply that there is no "correct" choice of test statistic, but rather the test statistic is selected for the particular properties of the confidence intervals it produces for μ. This fact motivated two modifications to the common test statistics producing the Arizona test statistics in order to achieve the following two desirable properties. First, confidence intervals should be longer for outcomes that may indicate mismodeling in order to reflect the experimenter's uncertainty of the modeling quality. Second, the confidence intervals should be expected to either include μ_{true}, or else be close to μ_{true}.

Longer confidence intervals for suspicious outcomes:

A likelihood ratio test statistic provides a scale such that outcomes with high values of $Q(d,\mu)$ provide strong evidence for the signal model, and weak evidence for the competing alternative model, where the specific alternative is determined by the choice of the likelihood ratio. Outcomes with intermediate values of $Q(d,\mu)$ suggest some evidence for both models.

However, there may be outcomes that provide little evidence of either model. In the

[127] Recall that the final discriminating variable spectrum is binned to facilitate the modeling of systematic uncertainties.

[128] The term *local ordering rule* is used to differentiate the ordering rules needed to determine the acceptance region for each separate value of μ from the *global ordering* that one must select to define the x-axis in order to make sensible figures.

case of the common test statistics used in particle searches, these outcomes typically have bin counts that are considerably less than the expected bin counts of the background-only model. In other words, these outcomes would be considered significant deviations from the background-only model if one performed a Fisher significance test (FST). If any of these outcomes were actually measured, the experimenter would be justified to be suspicious that something was wrong with the experimental design, i.e., that all the statistical models available in the NCRC were inconsistent with the real-world processes being measured.

Unfortunately, even though these outcomes are inconsistent with both the signal and alternative models of Q_{FC}, Q_{D0} and Q_{AT}, they are more inconsistent with the signal model.[129] This causes the likelihood ratio to be quite low for these inconsistent outcomes so that they *appear* to provide strong evidence for the alternative models. The Arizona test statistics include the adjustment factor Q_{inc} that causes these outcomes to achieve higher test statistic values, and this causes the confidence intervals for those outcomes to become longer. This result seems more aligned with the experimenter's knowledge of the experiment so that these longer confidence intervals better convey the experimenter's uncertainty regarding these inconsistent outcomes.[130]

Confidence intervals should be close to μ_{true} when they do not contain μ_{true}:

The reliance on likelihood ratios as test statistics is natural since these are well-studied and their properties well-understood. Namely, the experimenter can always justify their use by appealing to the Neyman-Pearson lemma. When the Neyman-Pearson lemma is applied to the NCRC (as opposed to NPHT), the resultant confidence intervals may be uniformly most accurate (UMA), and thus have the smallest probability of containing *incorrect* values of μ. However, this does not imply that these UMA confidence intervals provide greater confidence in containing μ_{true} than non-UMA confidence intervals. In fact, the confidence of μ_{true} to lie with the observed confidence interval is simply given by the confidence level used, independent of the choice of test statistic.[131]

One can recast the selection of test statistic in terms of risk minimization. Using a likelihood ratio test statistic in NCRC minimizes the loss of power risk (defined in section 6.3.4 on page 316). In other words, if one selects the loss of power risk function as the risk to be minimized, then a likelihood ratio is the correct choice of test statistic. However, the reporting risk defined in section 6.3.4 may be more appropriate when one desires confidence intervals that are as close as possible to μ_{true}, even if not containing μ_{true}.

[129] The alternative model for Q_{FC} is $p(d;\tilde{\mu}_p)$, for Q_{D0} is $p(d;\tilde{\mu}_0)$, and for Q_{AT} is $p(d;\tilde{\mu}_b)$ as described in section 6.2.3.4 on page 285. The term "inconsistent" here indicates outcomes with very small model probabilities.

[130] "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful," – George Box [332]. If the experimenter is less certain of the utility of the model for certain possible outcomes, then this uncertainty should be incorporated into the inference procedure since it is known *a priori*. It is illogical to ignore doubts of the model for certain possible outcomes, construct the confidence regions for all possible outcomes, and then obtain an unsatisfying result in the event of an unexpected observed outcome. If the logic of the confidence principle is to be strictly adhered to, then the inference procedure must account for the experimenter's *a priori* beliefs as well as the experimenter's possible *pos hoc* behavior.

[131] Recall that the term "test statistic" is being used to refer to both NPHT and the local ordering rule of NCRC.

This is because the experimenter may wish to reduce the risk that the reported confidence interval lies far from μ_{true} when it does not contain μ_{true}. The inclusion of the adjustment factor Q_{rep} in the Arizona test statistics was observed to reduce the reporting risk in several examples of the multivariate Poisson model used for signal searches.

6.6.3. Confidence-Acceptance Profiles and Frequentist Evidential Support

Compared to the Neyman construction, Bayesian inference has the advantage of an easily interpreted posterior probability distribution for μ as its result. The posterior distribution provides a way to compare the strength of evidence for different values of μ. The result of the Neyman construction is simply two sets of values for μ, those that are accepted and those that are rejected so that no evidential support for different values of μ is available. Aside from the problem of selecting a prior distribution, the disadvantage of Bayesian inference is the lack of a frequentist interpretation. This is because Bayesian inference is completely conditioned on the observed outcome, while frequentist inference considers the sampling space of all possible outcomes in determining the meaning of the observed outcome.

Ronald Fisher's fiducial method of inference offered a fiducial probability distribution for μ as its result. The fiducial distribution was similar to the Bayesian posterior, but did not require the use of a prior and instead relied on the sampling space of possible outcomes. However, Fisher was unable to provide a purely frequentist interpretation of the fiducial distribution.[132] Modern developments in theoretical statistics have considered confidence distributions that could be used to provide evidential support for different values of μ, but these too suffer from a lack of purely frequentist interpretation and are not always guaranteed to exist for all possible outcomes.

The confidence-acceptance profile (CAP) defined and examined in section 6.4 is simply the collection of confidence intervals at all confidence levels.[133] CAPs by themselves do not provide evidential support for different values of μ (although one may be tempted to interpret them as such since they are functions of μ), but rather one may derive special functions from the CAPs that do provide a way to compare the evidence for differing μ. These are the background significance ratio (BSR) function and the signal significance ratio (SSR) function. For each value of μ, the BSR function provides evidence for signal existence while the SSR function provides evidence for signal nonexistence. The definitions of the BSR and SSR functions utilize concepts from the three different methods of frequentist inference: Fisher's significance testing, Neyman-Pearson hypothesis testing and the Neyman construction, as well as concepts from Birnbaum's theory of evidence and Mayo's theory of severe testing. The main advantage of the BSR and SSR functions is that one may provide a proper frequentist interpretation for their observed values.[134]

[132] The well-documented animosity between Fisher and Neyman was motivated in part on their disagreement regarding the role of inference. Fisher believed that evidential support was possible with an *inductive inference* statement. Neyman disagreed and felt that only statements of acceptance/rejection were possible as a process of *inductive behavior*.

[133] The confidence-acceptance profile is not in any way a probability distribution, and thus the use of the term "profile" to distinguish it.

[134] An advantage of Bayesian inference is that the posterior distribution is easily updated as new

6.6.4. Bayesian Methods for Nuisance Parameter Treatment

Nuisance parameters are included in the statistical model to describe the sources of systematic uncertainties present in determining the expected signal and background yields in each of the bins of the discriminating variable, $s = [s_1, s_2, \ldots, s_n]$ and $b = [b_1, b_2, \ldots, b_n]$, respectively. The large number of sources of systematic uncertainty in a particle search gives rise to a large number of nuisance parameters, $\nu = [\nu_1, \nu_2, \ldots, \nu_k]$ (for k systematic uncertainties).

Because the systematic uncertainties in a particle search are estimated from additional empirical measurements, Bayesian prior distributions for the ν are constructed from the results of these measurements. This allows for Bayesian marginalization over the possible values of the ν to be made during the Neyman construction so that confidence intervals only regarding μ may be determined. Although this hybrid frequentist-Bayesian approach is sensible, there is always concern that the approach will invalidate the purely frequentist interpretation of the confidence intervals, i.e., that the acceptance regions for each μ may contain less than $(1 - \alpha)$ of the possible outcomes (by probability mass) when the stated confidence level is $(1 - \alpha)$. Therefore, studies are needed to examine the coverage properties of the Neyman construction when performing Bayesian marginalization over the ν.

Three methods of Bayesian marginalization for the treatment of nuisance parameters are the nominal method, Q^N and the profiling method, Q^P, which are commonly used in particle searches, and the full method, Q^F, which is exploratory. The results of the 2σ-coverage studies shown in table 6.15 (page 422) show similar coverage properties between Q^N and Q^P, while the use of Q^F seems to offer superior coverage performance in general. These studies also show that there are different coverage properties for the six test statistics examined, the three common and three Arizona. For example, Q^N_{D0} was observed to provide more overcoverage above the stated confidence level of 95%, especially when the $\mathrm{CL_S}$ method was implemented.[135] This suggests that for a signal search where a large amount of systematic uncertainty is present, the particular coverage properties of each test statistic may be an important factor in the selection of which test statistic to use.

information arrives from additional experimentation. Section 6.4.9 shows how the CAPs obtained from several experiments may be combined via meta-analysis. This method of frequentist updating may be useful when it is not possible to reperform the Neyman construction with the original datasets from the original experiments.

[135] This was the method used in chapter 5, where multi-stage machine learning was used in the search for vector-like quarks.

Chapter 7
Conclusion

Machine learning is useful for *predicting* possible outcomes based on complex datasets. Parametric statistical inference is useful for *explaining* observed outcomes using simple parametrized models. The powerful descriminative models created by machine learning are unintelligible to humans, while the simple generative models used in parametric inference are designed by humans. This work utilized both approaches in the search for vector-like quarks: machine learning to predict which of the simulated collision events were most signal-like, and then parametric inference to make sense of the measured collision events. A search strategy using machine learning regression was found in most cases to provide a large improvement over the standard square cuts search strategies by halving the upper limits for the largest signal masses. Perhaps more importantly, several innovations in methodology were developed.

7.1. Multi-stage Machine Learning

The motivation for multi-stage machine learning was a lengthy series of initial failures[1] to improve upon the results of the square cuts search using single-stage machine learning. It was hypothesized that multiple stages might act as a kind of deep learning method, where the first stage of learning extracted and condensed the information contained in the "raw" input variables into fewer, more powerful input variables for the second stage of learning.

The use of partial randomization in the selection of input variables at each splitting node during decision tree construction was allowed for the retention of many useful input variables. This was the primary motivation for selecting boosted decision trees for use as the stage 1 learning algorithm. The first step of machine learning requires selection of the set of input variables to use for training, and this task often requires considerable effort. For example, if one has 14 input variables, then there are $2^{14} - 1 = 16,383$ possible sets from which to choose. Although several input variables were rejected from the final set used here,[2] the remaining input variables were selected mainly due to the unique information each contributed (see table 5.4). The use of randomized forests was believed to mitigate any effects from pairwise linear correlations, slight (undetected) mismodeling, or large systematic uncertainties. This was because randomization made it unlikely that two highly correlated variables were used in the same decision tree, and also prevented any one particular input variable from contributing much more often than any other input variable (see table 5.13).

The search for a hypothesized particle is really many signal searches, one for each invariant mass that the particle may have. The used of two-stage learning allowed the first

[1] The failed attempts using single-stage machine learning were too numerous to report here.

[2] The rejection of several alternate input variables was not reported here. These were variables that were observed to rarely be selected for use in decision trees, or else were found to often cause overtraining or mismodeling issues.

stage to target each specific signal mass, giving a unique classification predictor function for each mass. The second stage then used these separate classifiers as its input variables, and combined them into a single classifier.[3] The use of two stages of learning also allowed for a natural partitioning of the data: the training and validation data partitions of the first stage were switched for the second stage.

A learning algorithm has several hyper-parameters that must be tuned, but with 59 signal masses to train at stage 1, tuning by hand was prohibitive. Therefore, the process of hyper-parameter optimization was automated so that a grid search for optimal hyper-parameters could be performed. This required designing a procedure to maximize signal-background discrimination while simulataneously checking for overtraining. Several statistical tests comparing training data distributions to validation data distributions were used to check for overtraining, where a small p-value returned by a test indicated likely overtraining. These tests were the Kolmogorov-Smirnov, Anderson-Darling, chi-squared, and a new autoregressive test (developed here to check for small "shape shift" errors), where the use of multiple test was desirable since each detected different types of overtraining. The resultant eight p-values (four for background distributions and four for signal) were highly correlated, and so were combined into a single p-value using both Brown's and Fisher's methods. This single p-value simplified the automated grid search.

After completing the stage 1, it was observed that the use of powerful learning algorithms (boosted decision trees, multilayer perceptrons, and support vector machines) led to significant overtraining. Instead, simple logistic regression (logit classification) was found to combine the individual stage 1 results into a single classification predictor function without overtraining.

This two-stage classification strategy did not provide any improvement over the square cuts strategy[4], and so a third stage of machine learning regression was implemented. Classification had been used in the first two stages in order to identify optimal *signal regions*, but the same $m(Zb)$ discriminating variable was used for statistical inference. Machine learning regression provided a unique approach to constructing a *discriminating variable*, the use of which had not been observed in the particle physics literature. Boosted decision tree regression as well as regression techniques more common to econometrics were implemented. The use of the stage 3 regressand as the discriminating variable for statistical inference had the effect of lowering the upper limits of the T_{pair}, B_{pair} and T_{single} signal types by as much as 50% for the largest signal masses (figure 5.52 on page 226 demonstrates why this result was important). Based on these results, the use of machine learning regression may be useful in extending the reach of particle searches to higher signal masses.

Figure 7.1 summarizes the methodology of the machine learning search.

[3] Technically four stage 2 resultant classifiers, one for each of the signal types, T_{pair}, B_{pair}, T_{single}, and B_{single}.

[4] This null result was not surprising given the many failures witnessed with single-stage learning, and after finding that only the less powerful logit classification worked as the stage 2 learning algorithm.

Problem 1: How to proceed after single-stage machine learning failed?

Solution: Implement multi-stage machine learning methods.

- Many stage 1 classifiers were created, one for each specific signal mass.
- A stage 2 classifier combined the mass-specific classifiers into a single tool.
- Two-stage learning gave a natural partitioning of the datasets.

Problem 2: Which learning algorithm to use for stage 1 classification?

Solution: Partially randomized gradient-boosted decision trees (BDTs).

- BDTs provided comparable signal-background discrimination to multilayer perceptrons and support vector machines.
- Randomization reduced the effort expended on input variable selection.
- Randomized BDTs were less sensitive to any one input variable (partially mitigated potential mismodeling).
- Training BDTs was much faster than other algorithms, allowing a grid search for optimal hyper-parameters to be performed quickly (for each of 59 classifiers).

Problem 3: How to automate the search for optimal hyper-parameters?

Solution: Performed several statistical tests for overtraining and combined the results.

- Obtained p-values from the Kolmogorov-Smirnov, Anderson-Darling and chi-squared tests, as well as the new AR(1) shape shift test.
- The correlated p-values were combined using Brown's method (after estimating a p-value covariance matrix).
- The uncorrelated p-values were combined with Fisher's method, giving a single p-value to quantify overtraining.
- Finally, the signal-background discrimination power was maximized under the constraint of no overtraining.

Figure 7.1: A summary the machine learning methodologies implemented in the search for vector-like quarks. *(Continues on next page.)*

Problem 4: Which learning algorithm for stage 2 classification?

Solution: Logistic regression (logit classification).

- This less powerful learning algorithm did not result in overtraining.

Problem 5: How to improve results after two-stage classification was unsuccessful?

Solution: Performed machine learning regression to obtain a more powerful discriminating variable.

- Only gradient-boosted regression trees and econometric style regressions gave useful results (but not adaptively-boosted regression trees, multilayer perceptron regression, nor support vector machine regression).

Final Result

The upper limits for large signal masses were reduced by as much as 50%.

Figure 7.1: – *Continued from previous page.*

7.2. Other Machine Learning Results

Unsupervised machine learning is used to find patterns in unlabeled data (data that is not labeled as signal or background). In appendix section B.2, the k-means and Gaussian mixture models were trained on simulated background-only data to find clusters in the background data. Data insufficiency prevented the Gaussian mixture model from achieving reliable results, but the simpler k-means algorithm providing results with interesting properties. When the clustering results were applied to the simulated signal data, many clusters were found to have excellent signal significance. This simple method could be useful in identifying signal regions for statistical inference, as well as finding control regions for data driven corrections to the simulated data.

Inspired by the 2014 "Higgs Boson Machine Learning Challenge" [333], a new learning algorithm was invented called the *Partition Trees Learning Algorithm* (PTLA), which is discussed in appendix A. The PTLA was developed as a lightweight, tree-based classifier that utilized bitwise operations for rapid training and application. Its use was demonstrated on three different datasets using supervised and semisupervised learning. The structure of the PTLA make it ideally suited for implementation on programmable logic devices,[5] such as the field programmable gate arrays commonly used in particle detectors. An interesting feature of the PTLA is its production of more complex structures compared to those of boosted decision trees, and a method for exploiting this structure using instance-based learning was demonstrated.

[5]Loosly speaking, the bitwise operations used by the PTLA could be encoded as *hardware* on such a device, which would greatly increase its operational speed.

7.3. The Neyman Construction in Signal Searches

The Neyman construction of confidence intervals was the statistical inference procedure used to compute the confidence intervals (upper limits) on the measured cross sections of the vector-like quarks. Specifically, this was done by applying the DZero recommended test statistic to the multivariate Poisson model. The methodology of the Neyman construction was investigated and it was determined that confidence intervals computed this way could lead to empty set solutions for certain possible outcomes of the measured data, as well as short confidence intervals for other outcomes with small upper limits that might not be justified (and were likely to be misinterpreted as highly precise cross section measurements). These problems were shown to exist for two other commonly used test statistics in particle physics, the Feldman-Cousins and ATLAS recommended test statistics.

A possible solution to these problems was found by modifying the test statistics to account for the possible mismodeling of the experiment so that longer confidence intervals with larger upper limits were produced for outcomes that were more likely to indicate mismodeling. These modified test statistics are the proposed Arizona test statistics, which were also shown to reduce the risk of reporting a confidence interval with endpoints far from the true value of the cross section.

The advantage of the Neyman construction is that the resultant confidence interval provides a frequentist interpretation: that confidence intervals constructed in such a manner will contain the true cross section at a frequency equal to the confidence level.[6] However, unlike the posterior distribution of Bayesian inference, the Neyman construction does not provide a method for comparing the evidence for different possible cross sections within the confidence interval. The values of the cross section lying within the confidence are equally accepted, while the values outside are all equally rejected.

A method for examining the evidence for specific values of the cross section was developed. This was done by beginning with the confidence intervals at all levels of confidence, the confidence-acceptance profiles, and deriving a signal significance ratio (SSR) function and a background significance ratio (BSR) function. The SSR function provided a succinct summary of the evidence against signal existence, and the BSR function showed the evidence for signal existence at each possible value of the cross section. These functions were shown to have a proper frequentist interpretation, and were proposed as a method of providing more information regarding the result of an experiment so that more nuanced decision making would be possible.

In addition to deriving the significance ratio functions, a method of meta-analysis using the confidence-acceptance profiles was developed so that results from multiple experiments could be easily combined. Finally, three methods for the treatment of nuisance parameters were investigated, the null and profiling methods currently used in particle searches, as well as a new "full" method. The full method was observed to provide better coverage properties than the null and profiling methods.

[6] Note that this is not equivalent to a statement regarding the probability of the confidence interval to contain the true cross section.

Appendix A
The Partition Trees Learning Algorithm

The results of stage 1 machine learning with gradient boosted decision trees (BDTs) in section 5.5 motivated the development of a new binary classification learning algorithm, the Partition Trees learning algorithm (PTLA). The design of the new algorithm was guided by the desire to meet the following criteria:

1. *Lightweight* – Memory usage should be minimized and computational speed maximized. In section 5.5, the use of BDTs for stage 1 learning over neural networks was in part driven by these practical considerations. More specifically, the use of integer arithmetic and bitwise operations over floating point arithmetic should be a primary focus so that the PTLA will be able to run more quickly on certain limited instruction processors, such as graphical processing units (GPUs), and especially on field-programmable gate arrays (FPGAs).
2. *Insensitivity to feature selection* – When a learning algorithm is sensitive to feature selection, the selection of the best subset of features to use often requires a lengthy search. Given N_F features as a starting point, one may consider feature selection as N_F Boolean hyper-parameters yielding 2^{N_F}–1 possible hyper-parameter combinations to explore. Common tools such as principal component analysis may reduce the difficulty of feature selection, but results may be obtained more quickly if feature selection can be avoided.
3. *Highly parallelizable* – The use of boosting for decision trees requires that each subsequent tree has access to the results of the previous trees. However, random forests of decision trees, which may be trained in parallel, often perform as well as BDTs, and are more resistant to overtraining and feature selection dependencies [334]. Parallel training is desirable because it may be performed rapidly on FPGAs and GPUs.
4. *Native handling of missing or null-valued features* – If a certain datum has a missing or null feature value, the PTLA should still be able to utilize the remaining information contained in this datum's other features without making any assumptions about the missing value. This simplifies data xxxpre-processing.
5. *Insensitivity to outliers, noise and misclassification* – It was observed in section 5.5 that decision trees of shallow depth were required in order to reduce overtraining on the noise contained in the training data features. Additionally, boosting algorithms are often sensitive to misclassification errors contained in the training data [335]. The PTLA was designed to reduce these effects.
6. *Insensitivity to data imbalance* – A larger amount of one class versus another should not prevent the PTLA from providing good class separation.
7. *Semi-supervised and instance-based learning* – The option to use more advanced machine learning techniques should be easily incorporated into the PTLA.

A.1. Partition Trees

The PTLA begins with the selection of a $MaxBit$, and transforms all data features into integers by evenly dividing the training data into 2^{MaxBit} bins, each of which is labeled by an unsigned integer represented in binary format. Missing or null-valued data is assigned the value of 0. Figure A.1 illustrates the integer assignments for an example feature that contains some missing values using $MaxBit$=4. The binary representation of a feature value is termed *P-byte* with *P-level* denoting the level of granularity with which the P-byte is to be viewed, i.e., the number of bits to be read left to right from the P-byte. Note that when $MaxBit$=8, each feature value for each datum may be stored as an unsigned 8-bit integer. Representing the input features as integer-valued bin assignments in this manner 1) minimizes the memory footprint of the data, 2) allows a data feature to be viewed at any desired level of granularity, and 3) may reduce the impact from outliers.

This partitioning method works equivalently for floating point-valued features and integer-valued features. Categorical features may also be partitioned by first labeling each

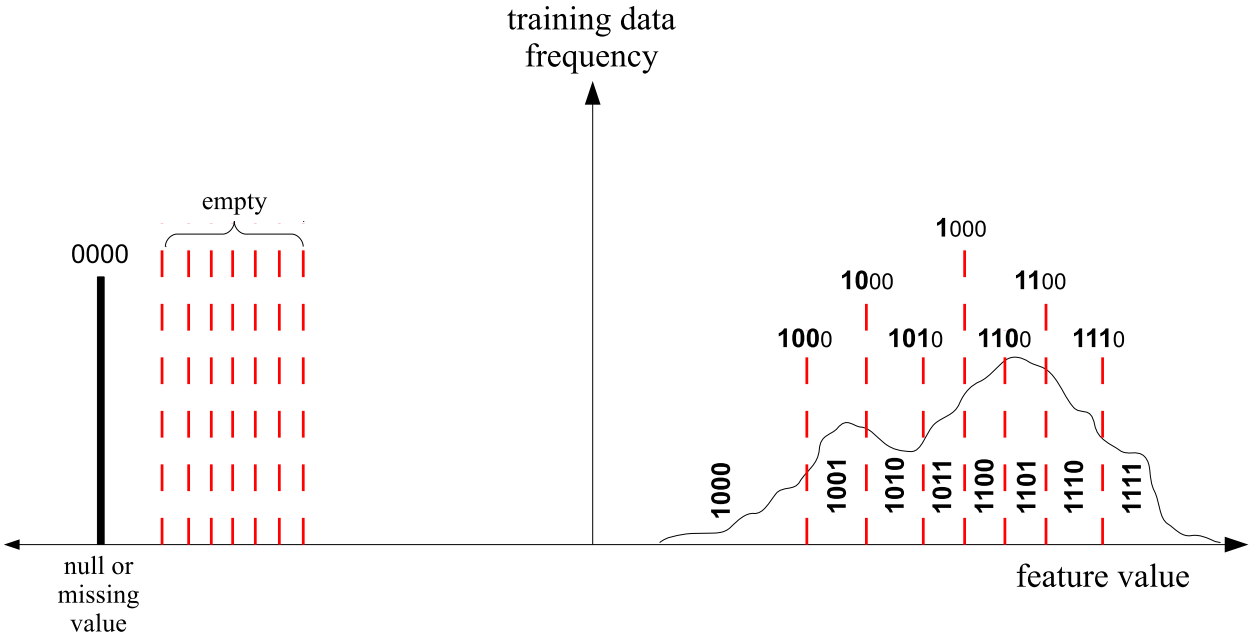


Figure A.1: Illustration of the partitioning of an example feature using $MaxBit$=4. Each datum falls into a bin represented by a 4-bit unsigned integer P-byte. Missing or null-valued data is assigned to the first bin labeled 0000. The remaining well-valued data is subsequently evenly divided into 2^3=8 bins, 1000 through 1111, with bin edges shown with dashed red lines. The location of the bin edges are only determined from training data, and bin edge values are taken as the average feature value of the adjacent two data points separated by the edge. Each bit from left to right represents the P-level of granularity with which the data is to be viewed (shown in bold faced type). At *P-level*=1, the data has either missing feature values (**0**000) or is well-valued (**1**000). At *P-level*=2 for well-valued data, 50% of this data has lower feature values (**10**00), and 50% has higher feature values (**11**00), and so on.

category by its frequency rank in the training data. Once the input features have been partitioned, the original feature data may be removed from memory.

The partitioning of the data into a binary representation allows the data to be split into decision paths according to increasing P-level. Figure A.2 illustrates this process if only a single input feature were used, and provides a maximum of 9 possible leaf nodes for maximum feature granularity. Even with only a single feature available, this already differentiates the PTLA from other decision tree algorithms such as the CART and C4.5 algorithms. With those algorithms, the selection of subnodes involves a greedy search for the splitting value that provides the best separation of classes between two subnodes. The default number of split values that are searched with the TMVA software used in chapter 5 is 20. This means that for the selection of the splitting value to use for a node, 20 separate hypotheses are examined (plus 1 for the option to not split). This is done for each candidate feature examined so that a large number of hypotheses are explored at each node, even when using random forests. For a single node of the partition tree, however, only two hypotheses are examined per candidate feature: whether or not to increase the granularity of the feature data by increasing the P-level by one unit (examining the next bit of the P-bytes). Moreover, the splatting values (bin edges) of the PTLA are predetermined before tree construction begins.

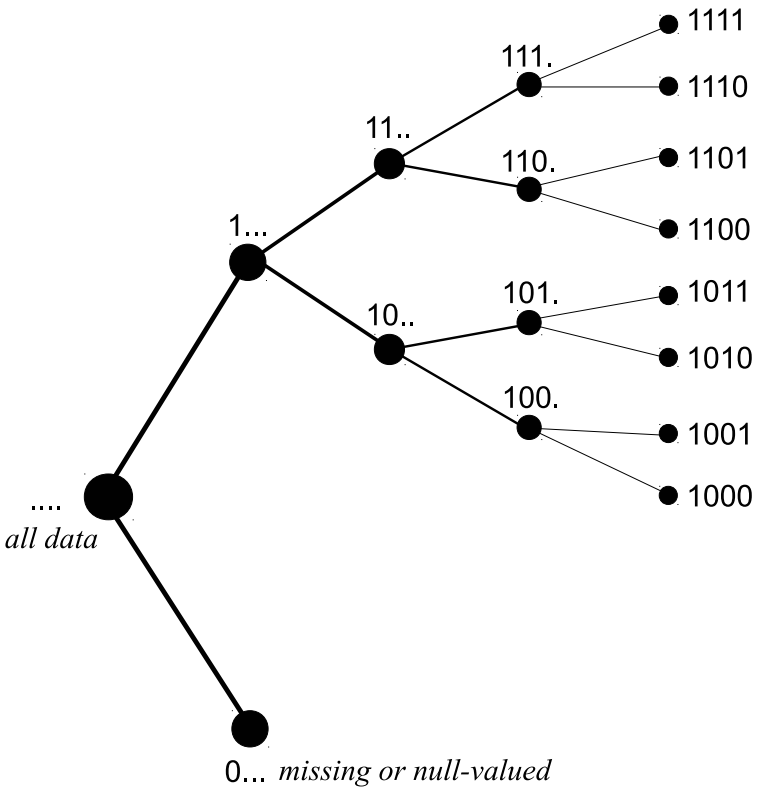
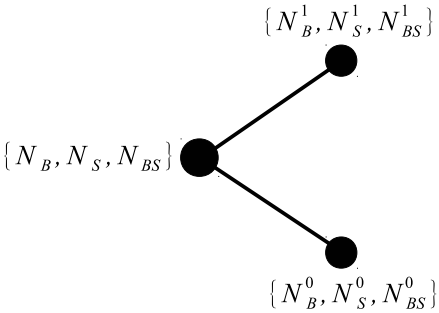


Figure A.2: Illustration of the maximum possible splitting of the dataset by a partition tree using a single feature with $MaxBit = 4$. The partition tree is read from left to right with all data beginning at the base node, labeled "...." where periods denote insignificant bits. In this case, the tree may reach a possible depth of 4, the maximum number of significant bits representing the data, with 9 final leaf nodes possible.

The results of the grid search for optimal hyper-parameters in section 5.5.2.3 demonstrated that only shallow decision trees with depths no greater than 3 could be used in order to prevent overtraining. By examining much fewer hypotheses at each decision node using predetermined bin edges, the partition tree is able to reach much greater tree depths without overtraining. This is the primary utility of the PTLA.

The partition tree uses standard separation criteria to select the optimal feature to use for splitting during partition tree construction, such as maximizing the reduction of the Gini impurity measure Δg shown in figure A.3. However, it was often observed that this led to "shaving off" a small portion of the data into one of the subnodes, and increased the likelihood of overtraining for the entire partition tree. Therefore, the Gini impurity was modified by an *ad hoc* factor r_{shave} that gave a lower score to features that led to data shaving. Note that this scoring operation may be performed entirely with integer arithmetic if one multiplies through by $0.3\,(N_{BS})^3\,N_{BS}^0 N_{BS}^1$. Node splitting continues until a minimum node size is obtained, and when a feature P-level has reached $MaxBit$, the maximum granularity available for that feature, it cannot be used as a candidate for splitting in the later subnodes in that path of the partition tree.



$$\Delta g = N_{BS}\frac{N_S}{N_{BS}}\left(1-\frac{N_S}{N_{BS}}\right) - N_{BS}^1\frac{N_S^1}{N_{BS}^1}\left(1-\frac{N_S^1}{N_{BS}^1}\right) - N_{BS}^0\frac{N_S^0}{N_{BS}^0}\left(1-\frac{N_S^0}{N_{BS}^0}\right)$$

$$r = \frac{N_{BS}^1}{N_{BS}}\left(1-\frac{N_{BS}^1}{N_{BS}}\right) \quad \rightarrow \quad r_{shave} = min\left(1,\ \frac{r}{0.3}\right)$$

$$s_{split} = \Delta g \cdot r_{shave}$$

Figure A.3: The selection of the feature to be used for splitting is determined by the splitting score s_{split} which measures the decrease in Gini impurity (maximizing Δg), while penalizing features that tend to split a small portion of the data into one of the two subnodes with r_{shave}. The hyper-parameter 0.3 was determined through experimentation, and may be adjusted. (Note that $N_{BS} \equiv N_B + N_S$.)

Once the partition tree has been constructed, a score is obtained for each leaf node, (i), by computing the frequency of the signal class in that node, $f_S^{(i)} = \frac{N_S^{(i)}}{N_B^{(i)}+N_S^{(i)}}$. However, some leaf nodes will contain a larger amount of training data than others, providing a more precise measure of $f_S^{(i)}$. Therefore, the final leaf node score is determined as $f_S^{(i)} - \sigma_{f_S^{(i)}}$, where $\sigma_{f_S^{(i)}}$ is the numerically estimated downward deviation of $f_S^{(i)}$ determined by Poisson statistics. Note that this operation and node score may be approximated entirely with integer arithmetic.

As with the BDTs used in section 5.5, the likelihood of overtraining is greatly reduced by randomly choosing a subset of features at each node during splitting, as well as the use of bagging. Figure A.4 provides the pseudocode for the construction of a single partition tree.

Pseudocode for the construction of a single partition tree.

1. Set hyper-parameters:
   - $MaxBit$ – Determines the maximum precision (granularity) with which the data can be viewed, or alternatively, the memory footprint of the data.
   - $MinSize$ – The minimum amount of training data required to allow the creation of a subnode.
   - r_{shave} – The penalty factor applied to nodes that send only a small portion of the data to one of the two subnodes.
   - N_{RF} – The number of randomly chosen features to provide to each node in determining splitting.
   - $BagFraction$ – The fraction of training data sampled with replacement for use in partition tree construction.
2. Determine the partitioning (bin edge values) for each feature using only the training data, and then transform all data into unsigned integer representation (P-bytes). Remove the original feature data from memory.
3. Maintain a container of node inputs (useful for parallelization). A node input is a pointer to the global P-byte data, an unordered container (set) of unsigned integers representing the indices of the data contained in a node, and the current P-levels of the features at this node..
4. Add the top node to this container: a set of all data indices and each feature P-level set to 0 (or 1 if a feature has no missing data).
5. Loop over this container of node inputs. When finished, all data will have been assigned a score.
   (a) Randomly choose N_{RF} candidate features to examine.
   (b) For each feature:
      i. If this candidate feature's P-level=$MaxBit$, return a splitting score of 0.
      ii. Compare the next bit of the P-bytes to 0 or 1 for each of the data in the node, placing the index of each into one of two sets depending on the result of this comparison.
      iii. Compute a splitting score. If lower than the maximum, discard these sets and continue with the next candidate feature, otherwise retain these sets, discard the previous, and update the maximum splitting score.
   (c) If all features' splitting scores are 0, this becomes a terminal node. Set the score of each datum in this node to $f_S - \sigma_{f_S}$.
   (d) If there are two valid subnode, add these as two new node inputs with the appropriate P-level increased by 1.
   (e) If there is only one valid subnode, then this is a partial terminal node. Add the valid subnode as a new node input with the appropriate P-level increased by 1, and set the scores of the data in the non-valid subnode to $f_S - \sigma_{f_S}$ computed using this node (and not the non-valid subnode).

Figure A.4: The pseudocode for one possible implementation of the construction of a partition tree. If new data is expected, the partitioning bin edges must be saved, as well as the tree structure which can be done minimally with a few small ordered containers of unsigned integers.

A.2. Partition Tree Ensembles and Semi-supervised Learning

An unboosted forest of partition trees may be constructed, with each tree being differentiated from the others by the use of bagging and random forests (see section 5.2.1). The ensemble score for the partition tree forest is a simple weighted sum of tree scores, where an individual partition tree weight is determined *ad hoc* as the maximum of some output parameter of interest obtained with a simple decision rule (cut on classification score). For particle physics searches, the output parameter of interest is typically the significance, $\frac{N_S^{pass}}{\sqrt{N_S^{pass}+N_B^{pass}}}$ where "pass" implies passing the decision rule. In many other applications, the output parameter of interest is the classification accuracy of the learning algorithm, $\frac{N_S^{pass}+N_B^{fail}}{N_S+N_B}$. This ensemble scoring could be improved upon, such as with the Bayesian model combination algorithm [336], but this simple approach can be performed with parallel tree construction and minimal computational expense.

Three datasets were examined with the PTLA. The first is the dataset for the vector-like top quark with 650 GeV/c^2 signal mass and the accompanying Standard Model background data that were used in the searches in chapters 4 and 5. This is referred to here as the "vlq dataset". This dataset contains 14 features with 7545 labeled events. For consistency, the division of this dataset into training and validation samples is the same as that used in chapter 5. The second dataset is taken from the Kaggle Higgs Boson Machine Learning Challenge [333], for which the second place winner implemented boosted decision trees [337]. This dataset contains 30 features, 12 of which contain null values, with 250,000 labeled events, as well as 550,000 unlabeled test data events, the "higgs dataset". The final dataset is the Wisconsin breast cancer database [338, 339] that has 30 features with 568 labeled events, 62.9% with the malignant diagnosis and 37.1% with the benign diagnosis, the "cancer dataset". Table A.1 summarizes the three dataset characteristics.

data set	$N_{features}$	N_{null}	$N_{labeled}$	% signal	% train	% validation	$N_{unlabeled}$
vlq	14	0	7545	10.5	49.5	50.5	0
higgs	30	12	250,000	34.3	90.0	10.0	550,000
cancer	30	0	568	62.9	79.8	20.2	0

Table A.1: Summary of the three dataset properties examined with the PTLA.

In many real world tasks, unlabeled data is often available in large quantities while labeled data may be expensive to obtain. In these situations, semi-supervised learning (SSL) is used where the unlabeled data is included during the learning process to enhance classification performance [340]. The large amount of unlabeled data available with the higgs dataset motivated the incorporation of SSL into the PTLA algorithm. In this case, it is assumed the full 30 dimensional probability density of the feature space of the unlabeled test data accurately reflects that of the labeled data, and that the proportion of signal events "hidden" in the test data is the same as that contained in the labeled data.

SSL may be easily incorporated into the PTLA node selection by two methods as shown in figure A.5. As each feature is examined for use in node splitting, it may be rejected by

comparing the ratio of training data sent to one subnode R^{split}_{train} to the ratio of test data sent to the same subnode R^{split}_{test}. If these values significantly disagree, then this candidate feature is given a splitting score of 0. The number of standard deviations for rejection is an additional hyper-parameter with values between 1.0 and 2.0 observed to perform best. The second method is to reject individual subnodes when the ratio of training data to testing data deviates significantly from the global ratio of training data to testing data. In cases where no unlabeled test data is available and the use of bagging is implemented, the unbagged portion of training data with labels removed may proxy as a test set.

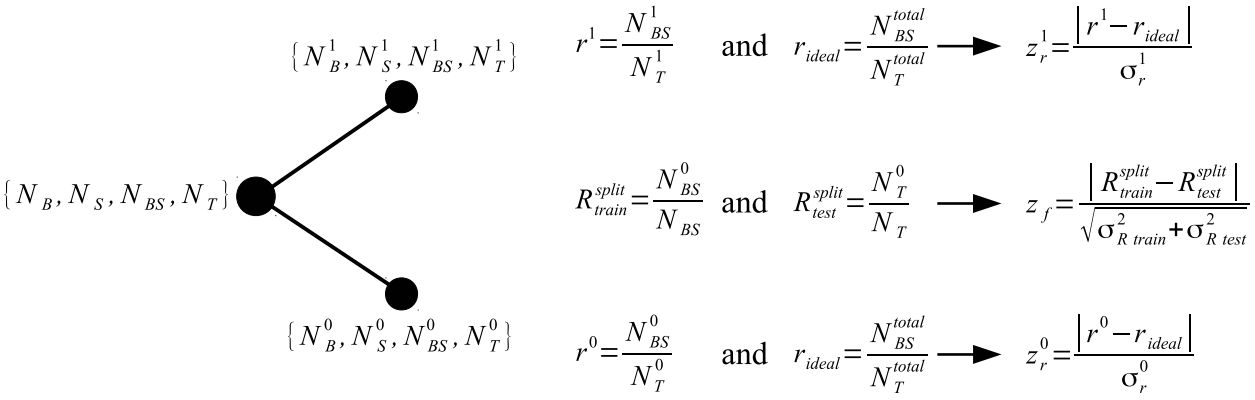


Figure A.5: Unlabeled test data may be used for SSL in two simple ways. First, the ratio of training events separated into one of the subnodes should be similar to the ratio of test data separated into that same subnode. Numerical estimation of the variances of these ratios provides an approximate z-score that can be used to reject the use of a feature for node splitting. The second method is performed for each of the subnodes separately. When the ratio of training data to test data of a subnode deviates too much from the global ratio computed from all initial data, then this subnode may be rejected.

As a forest of partition trees is grown, SSL also allows the rejection of overtrained partition trees. This may be done by determining the optimal decision rule for the output parameter of interest (such as maximizing significance with a cut on the classification score), and examining the fraction of training data passing the rule to the fraction of test data passing the rule. However, this global rejection scheme was not observed to improve performance significantly. Alternatively, when bagging is implemented, the unbagged data (including labels) may be used to determine an alternate optimal decision rule. If the output parameter of interest as predicted by the bagged data differs significantly from the unbagged prediction, the partition tree may be removed from the forest. This method proved more reliable than the previous with values of 2.5–5% in maximum deviation observed to best in removing overtrained partition trees.

Because the PTLA uses random forests of partition trees where only a few features are examined during node splitting, the partition tree leaf nodes tend not to reach high P-levels (do not reach high levels of granularity of any particular feature) in situations of small data volume with a large number of features. Higher P-levels can be reached by using the same splitting feature of a parent node as the splitting feature for its subnodes. At the extreme,

this would create each partition tree from a single feature, but this can be performed with some level of randomness. Values of 25–50% for the probability to reuse the parent node's splitting feature were tested during the performance studies of the following section. Note that including a probability to reuse a splitting feature during subnode creation reduces the complexity of a tree to some extent (less hypotheses to explore), and was usually observed to have the effect of reducing the amount of overtraining.

A.3. Performance of Random Forests of Partition Trees

The PTLA was applied to each of the three datasets with minimal hyper-parameter tuning. In all three cases, SSL was implemented using the unbagged training data, while the unlabeled test data contained in the higgs dataset was used in addition to the unbagged data for SSL.

A.3.1. PTLA Performance with the VLQ Dataset

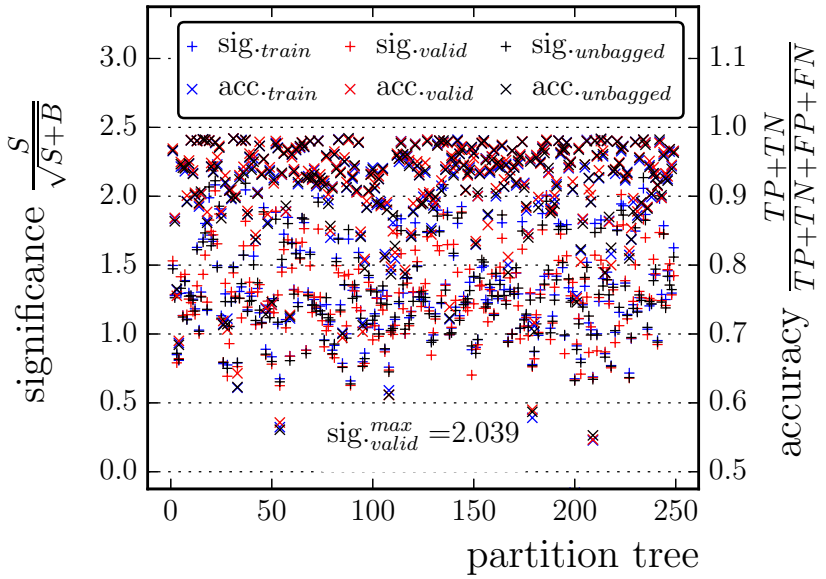
The performance of the PTLA on the vlq dataset is shown in figure A.6. Figure (a) shows the maximum significance that was achieved for each individual partition tree, as well as the classification accuracy at that significance value. The maximum significance for any particular tree ranged between 0.6 to 2.04 with the training data significance an average of 4.0% higher than that of validation. The performance of the unbagged training data for a given partition tree is also shown for reference.

Figure (b) shows the ensemble performance as the forest was grown (note the changes in both y-axes from figure (a)). The ensemble significance approached 2.25 after about 50 partition trees. Very little tuning of the hyper-parameters was attempted to reach this result (~ 5 attempts), with final selections shown in table A.2. This ensemble performance is comparable to that achieved in the stage 1 learning of section 5.5.2.4 that obtained a maximum significance of 2.3 after an extensive grid search for optimal hyper-parameters.

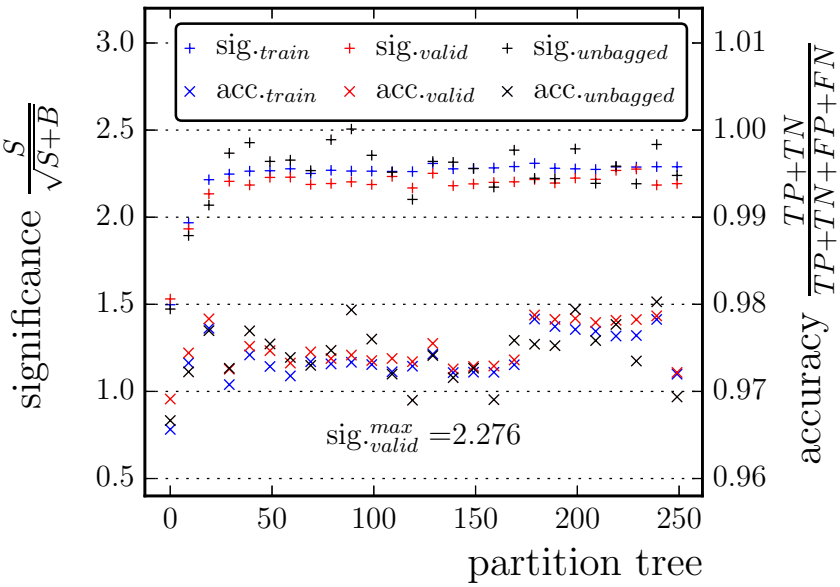
hyper-parameter	value	description
$MaxBit$	7	Maximum number of bits to use in each feature's P-byte.
N_{trees}	250	Maximum number of trees to grow.
N_{RF}	2	Number of randomly selected features to provide to each node.
$MinSize$	20	Minimum amount of training data needed to form a subnode.
$f_{bagging}$	50%	Fraction of bagged training to use.
$useSSLunbagged$	True	Use the unbagged training data for SSL.
N_σ	1.5	Number of standard deviation used to reject node z-scores (fig. A.5).
p_{parent}	25%	Probability of reusing the same feature for subnode splitting.
$MaxDeviation$	5%	Maximum deviation between bagged and unbagged predictions.

Table A.2: The hyper-parameters used by the PTLA for the vlq dataset.

Figure A.7 shows the first partition tree constructed in the forest. The optimized gradient boosted decision tree forest of §5.5.2.4 contained a maximum of 15 nodes, while here the partition tree contained over 100 nodes. In this particular tree, the first feature, $H_T(jets)$, was observed to be the most frequently selected feature used for node splitting. Note that



(a)



(b)

Figure A.6: The performance of the PTLA using the vlq dataset. The maximum significance was determined by the optimization of a decision rule (a cut on the classifier response score). The classification accuracy corresponding to this decision rule is also shown for reference. Classification accuracy is defined as the ratio of correctly predicted results (true positive and true negative) to all data. (a) The performance of individual partition trees in the ensemble (after rejection of some trees by semi-supervised learning). (b) The ensemble performance as partition trees were added.

$H_T(jets)$ provided much of the separation power of the standard square cuts analysis of chapter 4. Figure A.8 shows that maximum level of granularity for any of the features for a given leaf node was most often 4 or 5. This would be equivalent to using feature values partitioned into only 8 or 16 bins.

Figures A.9 (a)–(d) examine several distributions for the PTLA forest of 250 partition trees. Figure (a) shows the frequency spectrum of the response scores and indicated good agreement between training and validation data performance. Figure (b) shows the "reversed cumulative yield" distributions, the weighted yields of data remaining above the given response score. Visual inspection of reversed cumulative yield distributions was observed to be useful for the detection of overtraining. In this case, the training and validation data performance were in good agreement. Figure (c) provides the ROC curve for both training and validation data, and figure (d) plots the significance achieved for data remaining above a given response score. Overall the PTLA performed nearly as well as the optimized gradient boosted decision trees of section 5.5.

A.3.2. PTLA Performance with the Higgs Dataset

The performance of the PTLA on the higgs dataset is shown in figure A.10. Figure (a) shows that the maximum significance for any particular tree ranged between 1.5 to 2.5 with the training data significance an average of 3.5% higher than that of validation. Figure (b) shows the ensemble performance as the forest was grown. The ensemble significance approached 3.2 after about 40 trees. Very little tuning of the hyper-parameters was attempted to reach this result ($\sim$ 5 attempts), and are shown in table A.3. This ensemble performance compares well with the 3.2 significance achieved with an adaptively boosted decision tree implemented with TMVA using 800 trees with a maximum tree depth of 3 (i.e., using the default values listed in [159]).

hyper-parameter	value	description
$MaxBit$	7	Maximum number of bits to use in each feature's P-byte.
N_{trees}	250	Maximum number of trees to grow.
N_{RF}	4	Number of randomly selected features to provide to each node.
$MinSize$	20	Minimum amount of training data needed to form a subnode.
$f_{bagging}$	30%	Fraction of bagged training to use.
$useSSLtest$	True	Use the unlabeled test data for SSL.
$useSSLunbagged$	True	Use the unbagged training data for SSL.
N_σ	2.0	Number of standard deviation used to reject node z-scores (fig. A.5).
p_{parent}	25%	Probability of reusing the same feature for subnode splitting.
$MaxDeviation$	5%	Maximum deviation between bagged and unbagged predictions.

Table A.3: The hyper-parameters used by the PTLA for the higgs dataset.

Figure A.11 shows the high level of complexity achieved by a representative partition tree from the forest. The BDTs implemented with TMVA contained a maximum of 15 nodes, while this partition tree contained over 4,000 nodes. Figure A.12 shows the maximum P-levels attained by this partition tree's nodes, shows that increasing the probability of reusing the same splitting feature allows more nodes to reach the maximum level of granularity for one

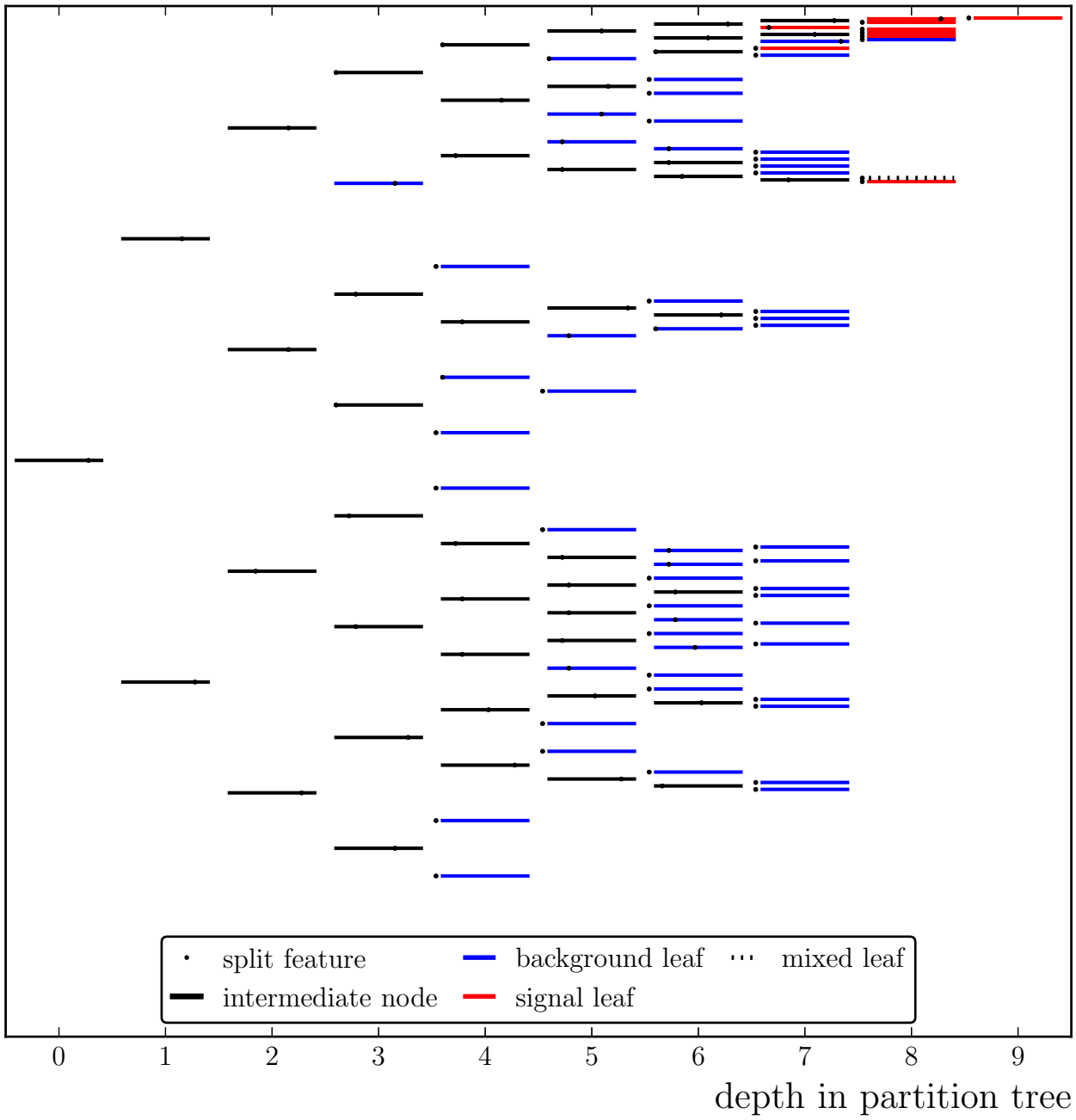


Figure A.7: The first partition tree constructed in the PTLA forest using the vlq dataset. There are over 100 nodes represented by horizontal lines. The position of a shaded circle along each node line represents the feature used for the splitting of that node (14 possible positions along each horizontal node line).

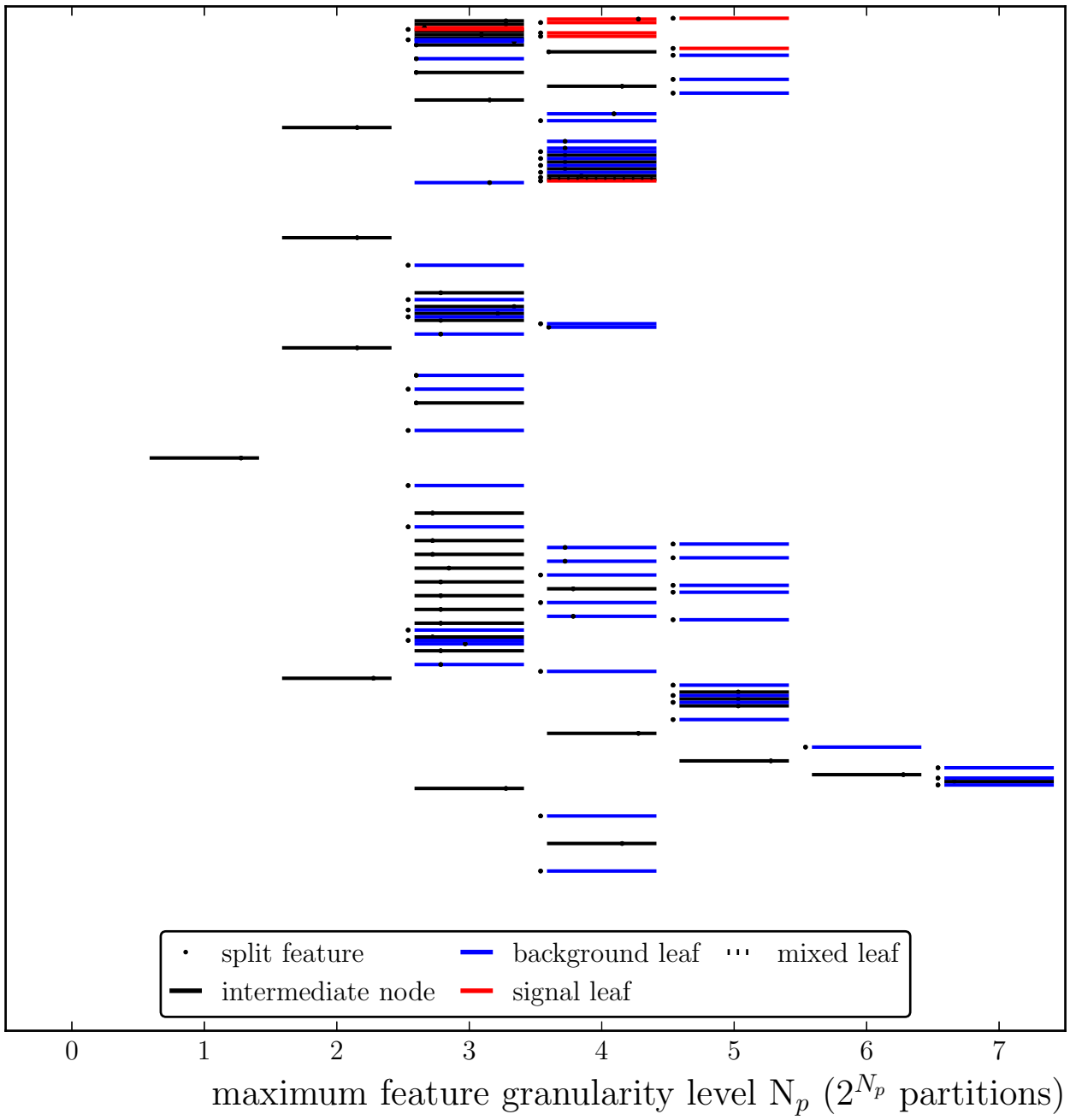


Figure A.8: The maximum P-levels attained by each of the nodes of the first partition tree constructed in the PTLA forest using the vlq dataset.

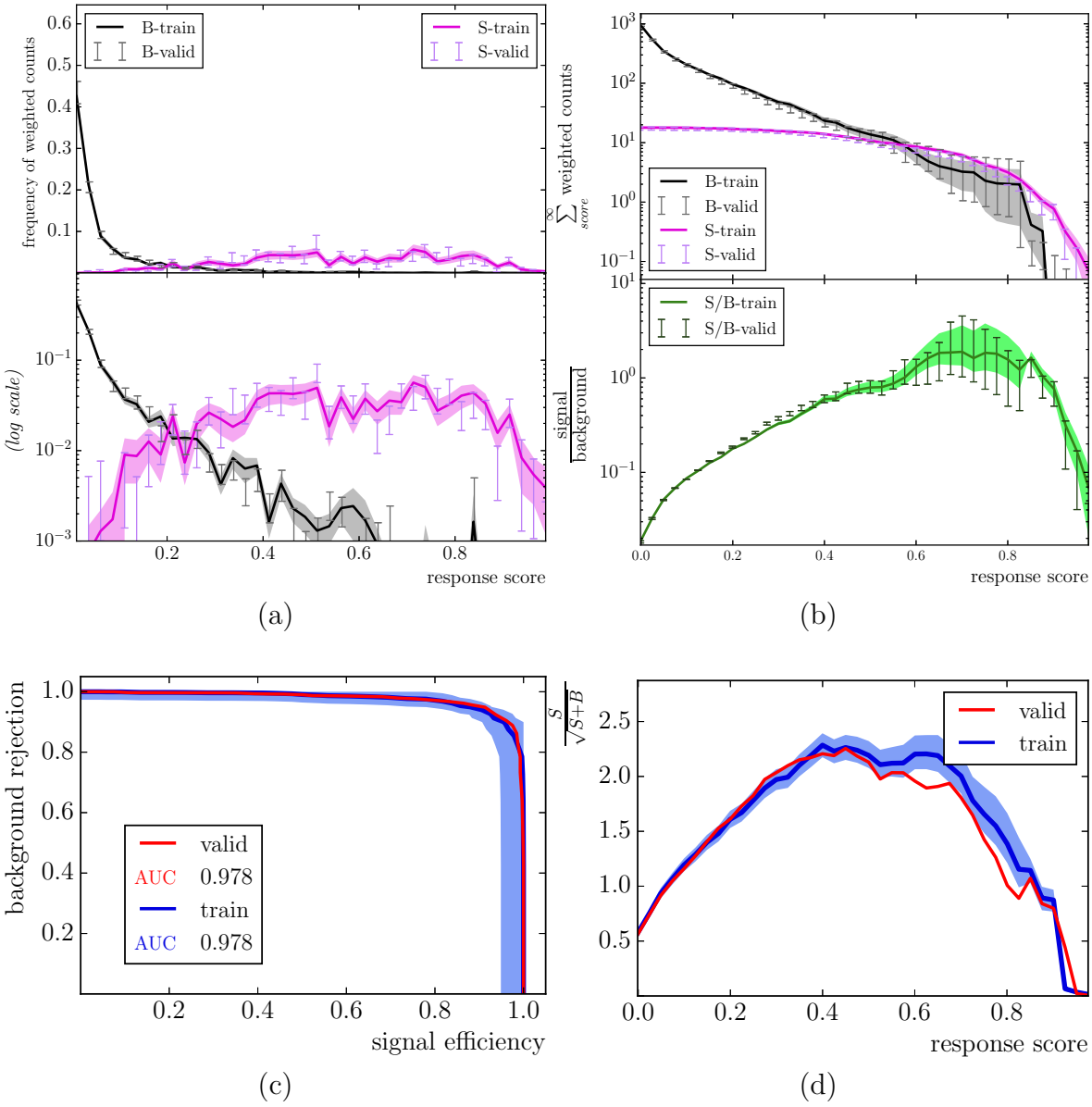
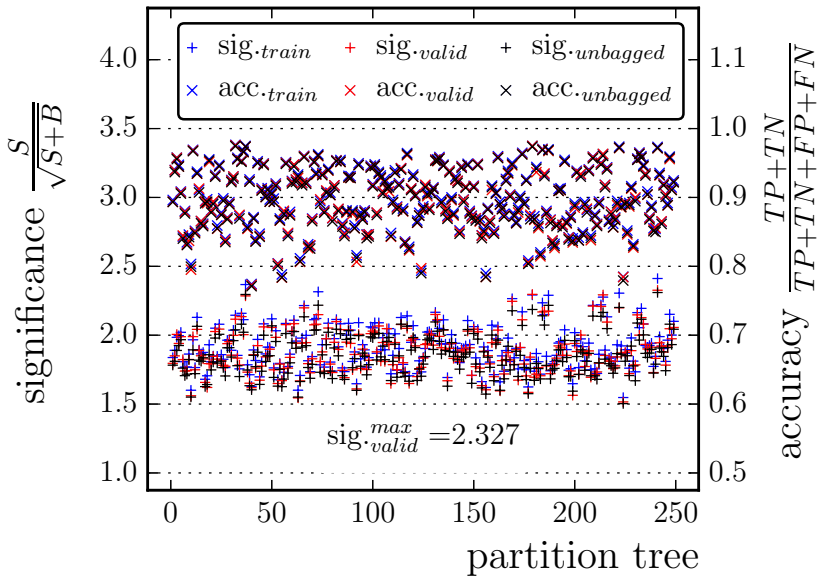
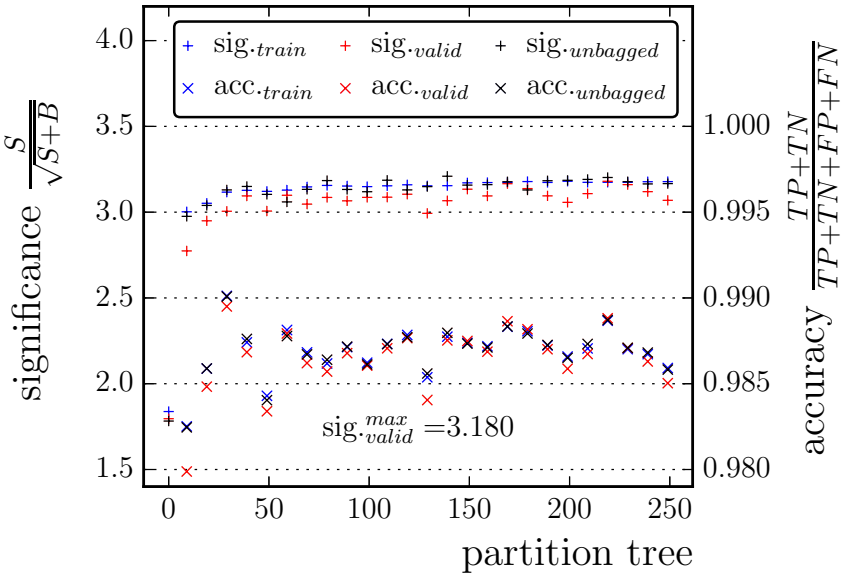


Figure A.9: Pertinent distributions of the PTLA ensemble using the vlq dataset. (a) The frequency distributions of the classification response score. (b) The reversed cumulative yield distributions (weighted) showing the amount of data remaining above a given response score. (c) The ROC curve for both training and validation data. (d) The significance achieved for both training and validation data remaining above a given response score.



(a)



(b)

Figure A.10: The performance of the PTLA using the higgs dataset. The maximum significance was determined by the optimization of a decision rule. (a) The performance of individual partition trees in the ensemble (after rejection of some trees by semi-supervised learning). (b) The ensemble performance as partition trees were added.

or more of their splitting features.

Figures A.13 (a)–(d) examine the performance of the first partition tree of the PTLA forest where the classification response score is determined solely by this single tree. This may be compared to figure A.14 that provides the performance of the entire ensemble of 250 partition trees. The performance of a single partition tree is much weaker than the ensemble, reaching a maximum significance of only ∼1.8 as seen in figure (d), while the ensemble reaches a significance of 3.2. The single tree shows a small amount of overtraining as detected in figure (b), while the ensemble shows very little. Nevertheless, it was surprising that the single tree has such good agreement between training and validation data despite containing over 4,000 decision nodes. The implementation of semi-supervised learning aided in the prevention of overtraining during tree construction.

A.3.3. PTLA Performance with the Cancer Dataset

The performance of the PTLA on the cancer dataset is shown in figure A.15. Figure (a) shows the classification accuracy of the individual partition trees. These fluctuated between 70 to 90% accuracy, and often showed some amount of disagreement between the training and validation accuracies. On average the training accuracy was 2.2% higher than validation. Figure (b) shows the ensemble performance as the forest was grown. After about 200 partition trees, the ensemble had good agreement between the training and validation data with ∼93% accuracy. Although classification accuracies over 99% have been reported [341], this result was obtained with minimal tuning and without any feature selection, and compares reasonably well with early work on the C4.5 algorithm which reached ∼95% accuracy [342].

The hyper-parameters used to generate these results are shown in table A.4, with only a few attempts at tuning by hand performed (∼5 attempts). The first partition tree of the PTLA forest had 36 nodes and is shown in figure A.16 for reference.

Figure A.18 shows the ensemble performance after 250 partition trees. Figures (a) and (b) show some disagreement between training and validation performance. However, this was only after a single random selection of a validation dataset containing 115 data events. To further investigate this disagreement, 100 bootstrap rounds were performed, each with a random selection of ∼20% of the total data for a validation set. The results shown in figures A.18 (c) and (d) show that there is minimal overtraining when averaged over the bootstrap rounds.

hyper-parameter	value	description
$MaxBit$	7	Maximum number of bits to use in each feature's P-byte.
N_{trees}	250	Maximum number of trees to grow.
N_{RF}	3	Number of randomly selected features to provide to each node.
$MinSize$	10	Minimum amount of training data needed to form a subnode.
$f_{bagging}$	50%	Fraction of bagged training to use.
$useSSLunbagged$	True	Use the unbagged training data for SSL.
N_σ	1.5	Number of standard deviation used to reject node z-scores (fig. A.5).
p_{parent}	50%	Probability of reusing the same feature for subnode splitting.
$MaxDeviation$	2.5%	Maximum deviation between bagged and unbagged predictions.

Table A.4: The hyper-parameters used by the PTLA for the cancer dataset.

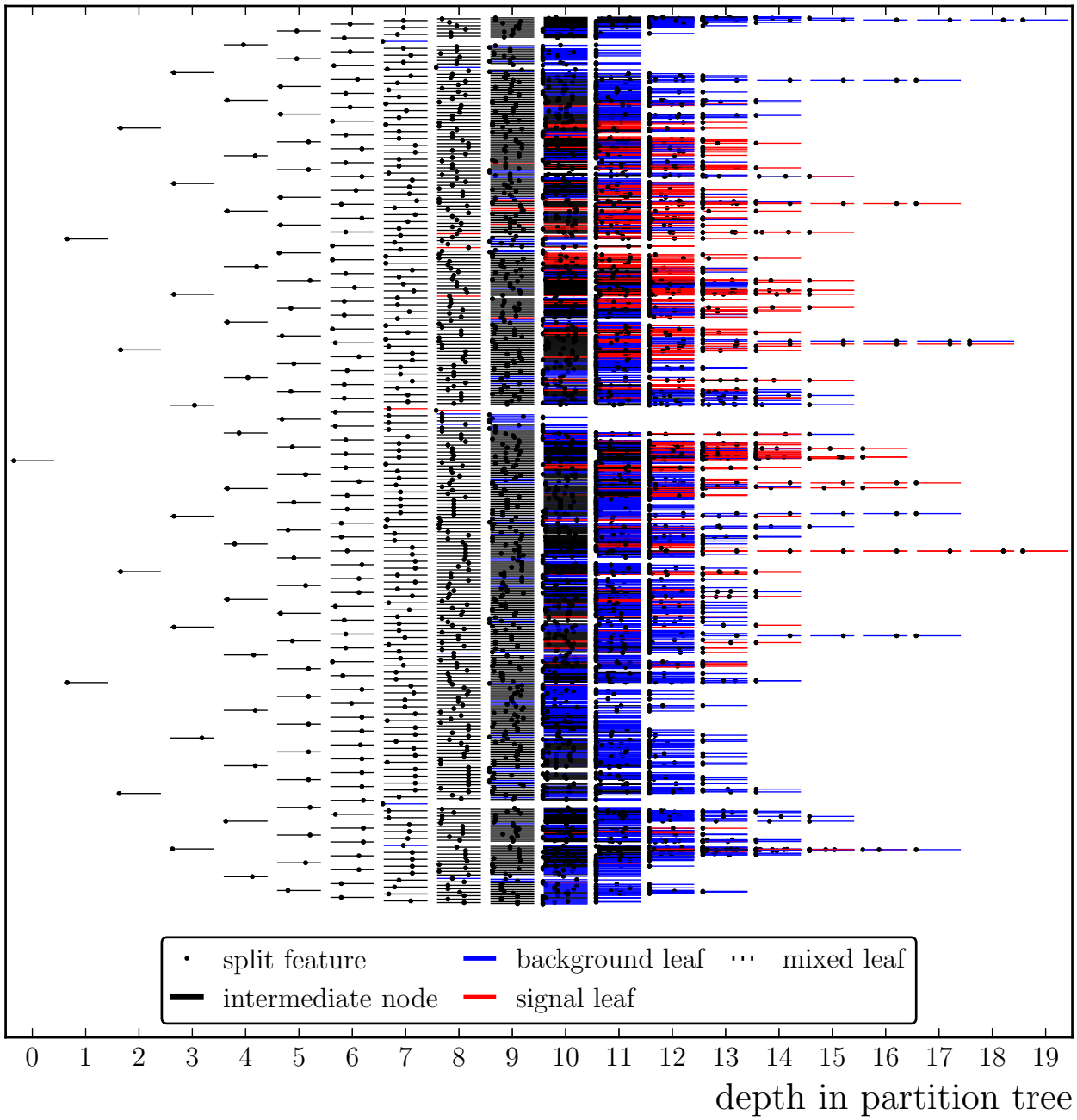


Figure A.11: The first partition tree constructed in the PTLA forest using the higgs dataset. There are over 4,000 nodes represented by horizontal lines. The position of a shaded circle along each node line represents the feature used for the splitting of that node (30 possible positions along each horizontal node line).

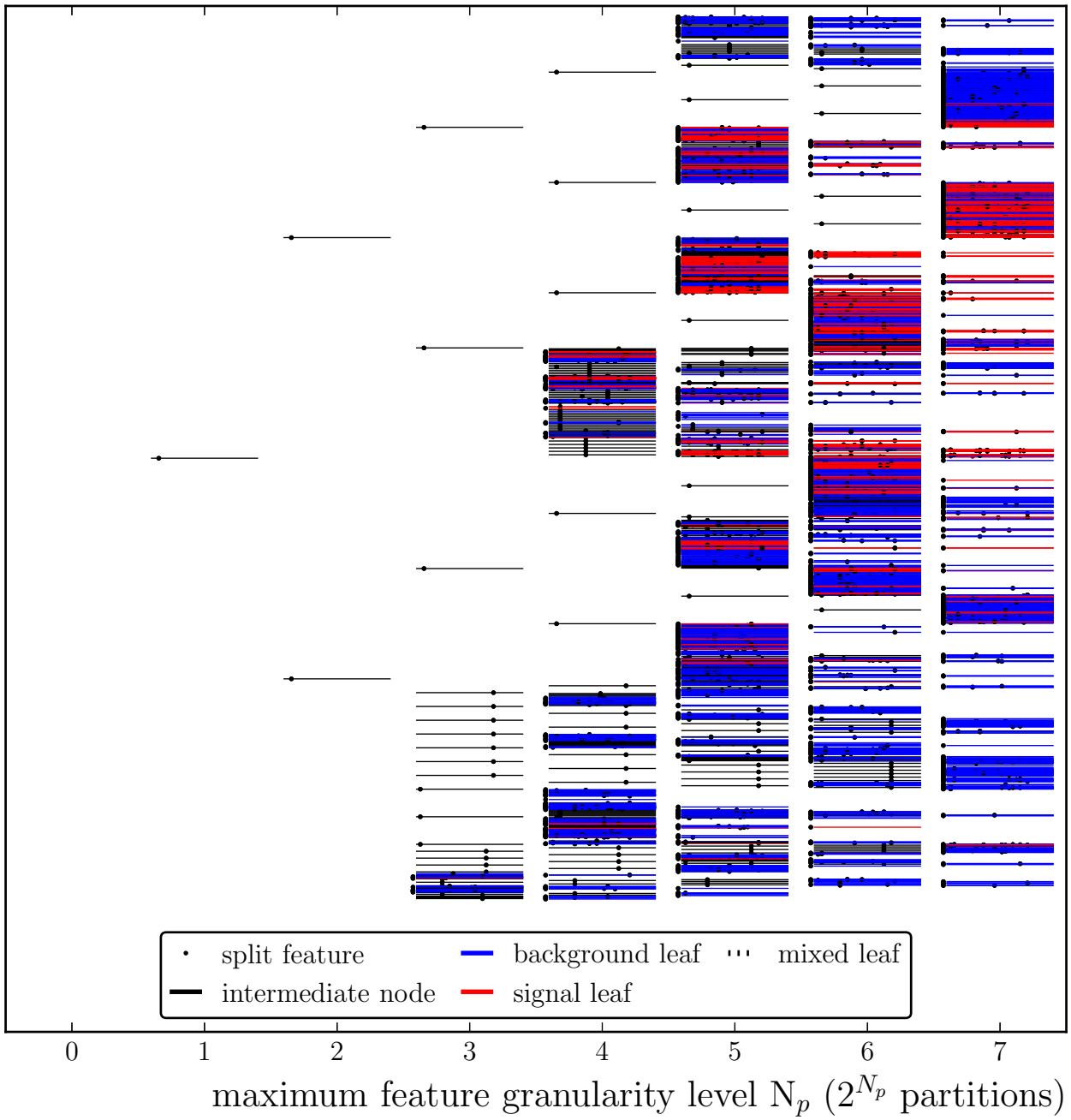


Figure A.12: The maximum P-levels of the first partition tree constructed in the PTLA forest using the higgs dataset. Many nodes reach a maximum level of granularity for one or more of their splitting features due to including a probability for reusing the same feature for subnode splitting as the previous parent node.

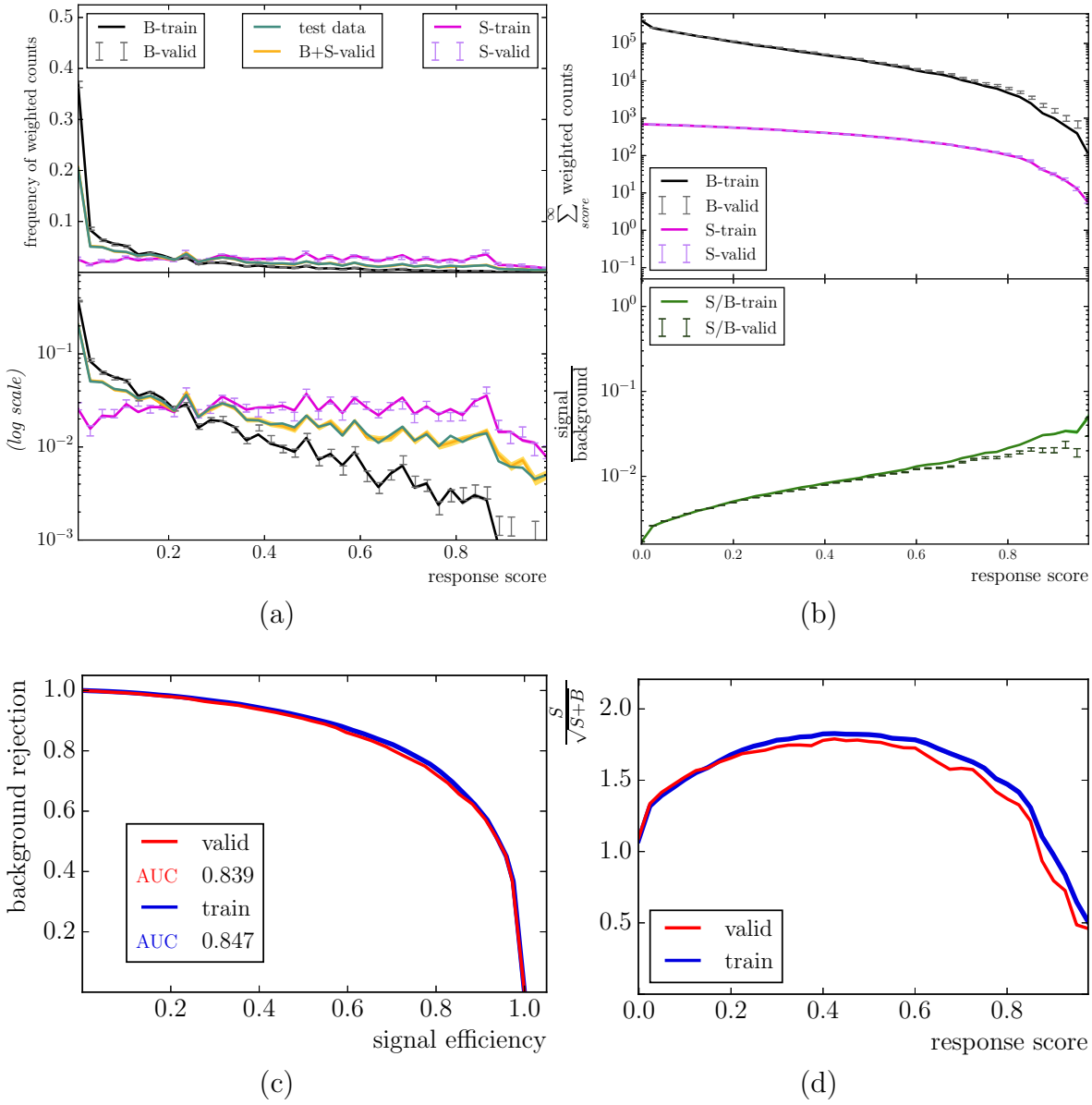


Figure A.13: Pertinent distributions of the first partition tree in the 250 tree PTLA forest using the higgs dataset. (a) The frequency distributions of the classification response score, including a comparison of the unlabeled test data to the combined signal and background validation data. (b) The reverse cumulative yield distributions (weighted), (c) The ROC curve, and (d) the significance distributions for data remaining above a given response score.

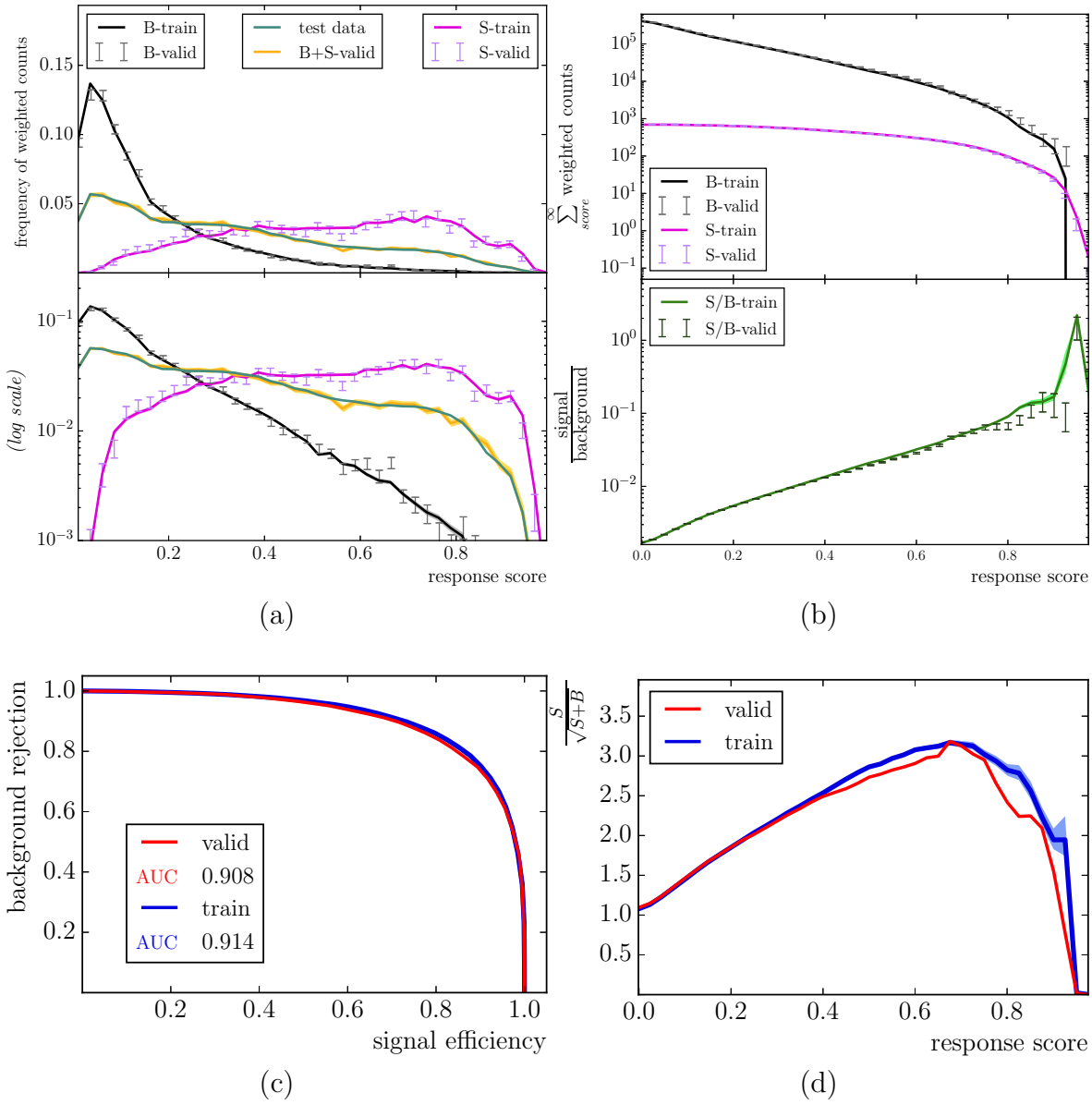
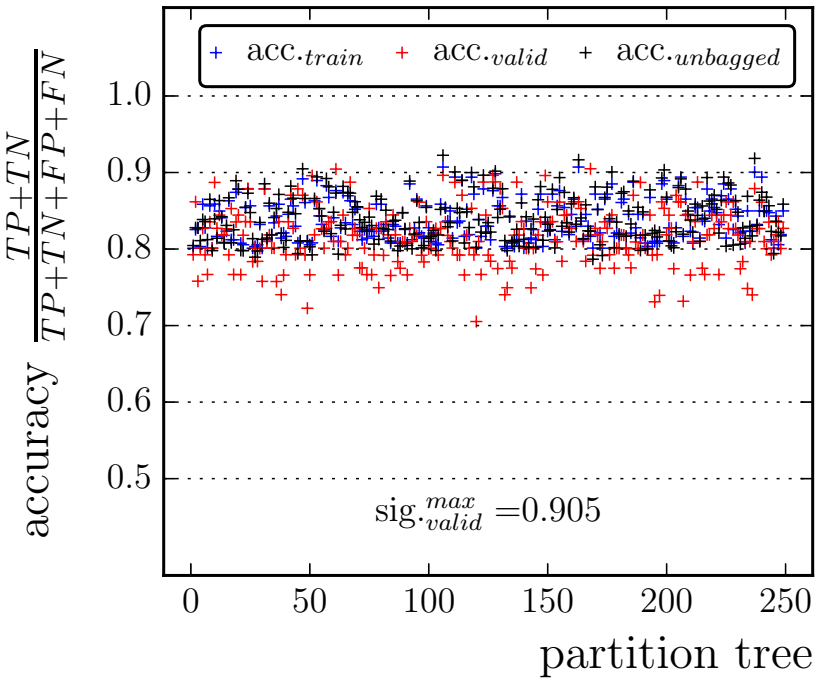
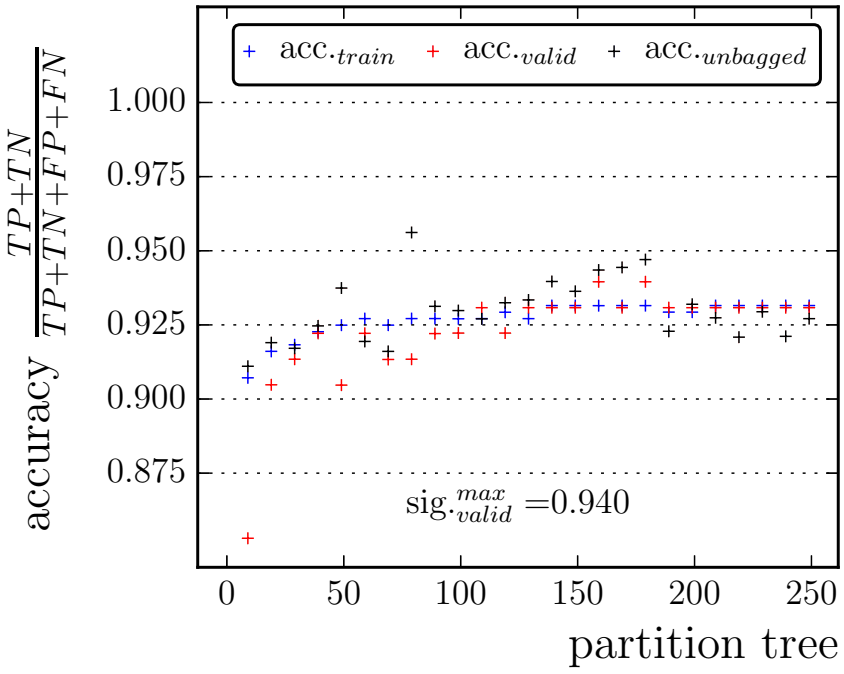


Figure A.14: Pertinent distributions of the PTLA ensemble using the higgs dataset. (a) The frequency distributions of the classification response score. (b) The reverse cumulative yield distributions (weighted), (c) the ROC curve, and (d) the significance distributions for data remaining above a given response score.



(a)



(b)

Figure A.15: The performance of the PTLA ensemble using the cancer dataset. (a) The classification accuracy for individual partition trees in the ensemble, and (b) the ensemble performance as partition trees were added.

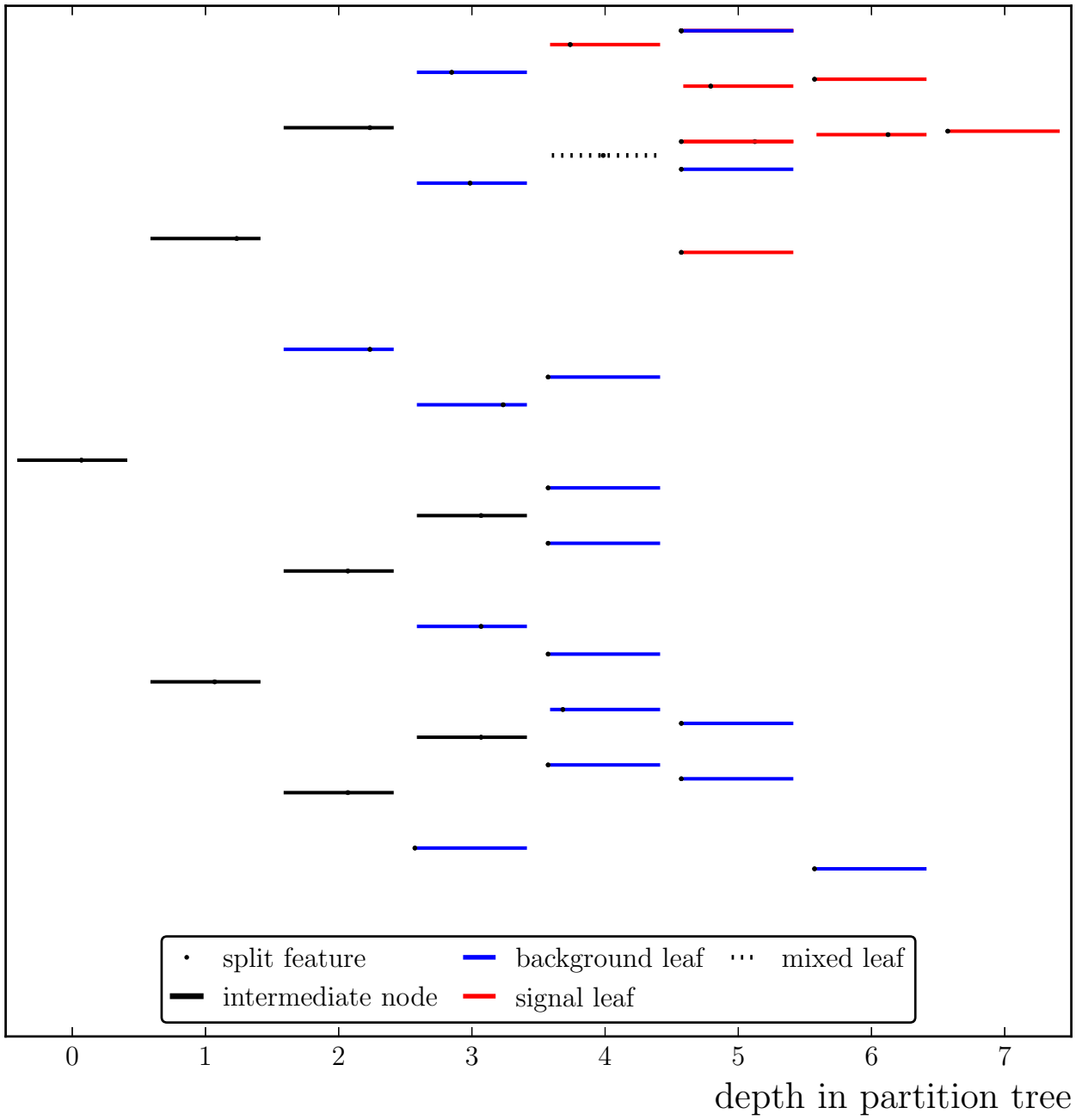


Figure A.16: The first partition tree constructed in the PTLA forest using the cancer dataset. There were 36 nodes in this partition tree, about average for the partition trees of this forest. The position of a shaded circle along each node line represents the feature used for the splitting of that node (30 possible positions along each horizontal node line).

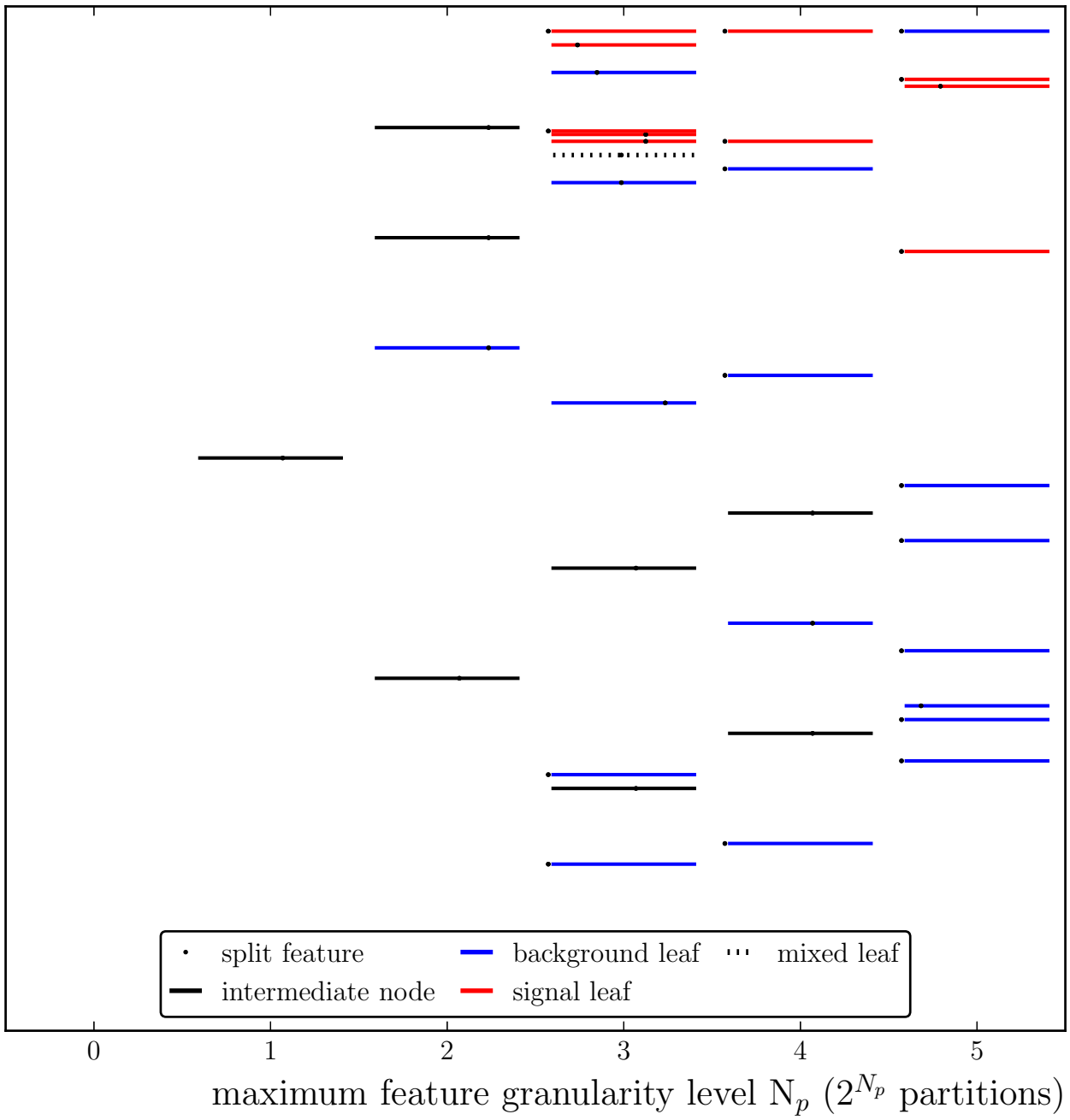


Figure A.17: The maximum P-levels attained by each of the nodes of the first partition tree constructed in the PTLA forest using the cancer dataset.

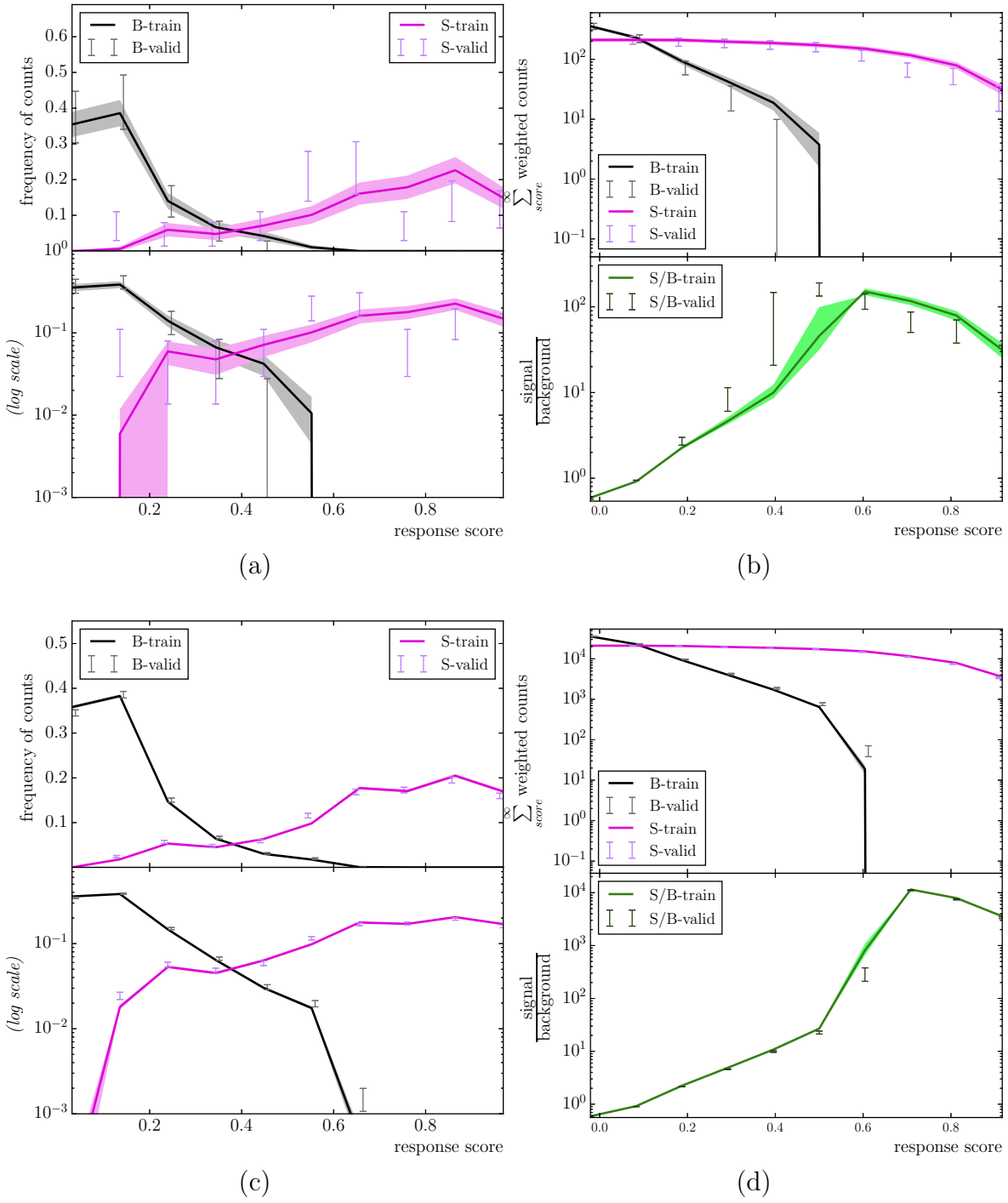


Figure A.18: Pertinent distributions of the PTLA ensemble of 250 partition trees using the cancer dataset. (a) The frequency distributions of the classification response score, and (b) the reverse cumulative yield distributions (weighted). (c)–(d) The same distributions after 100 bootstrap rounds of random selection of the validation set.

A.4. Instance-based Learning

In §A.3, the PTLA was observed to provide results on par with BDTs. However, each partition tree in the ensemble contained a richer structure than that typically observed with BDTs. The complex structure of the partition tree provides a measure of distance between data pairs. A simple approach is to define the distance between a pair of data points as the number of nodes that each datum must cross to reach a common node (each datum provides a node count, select the greater of the two as demonstrated in figure A.19).

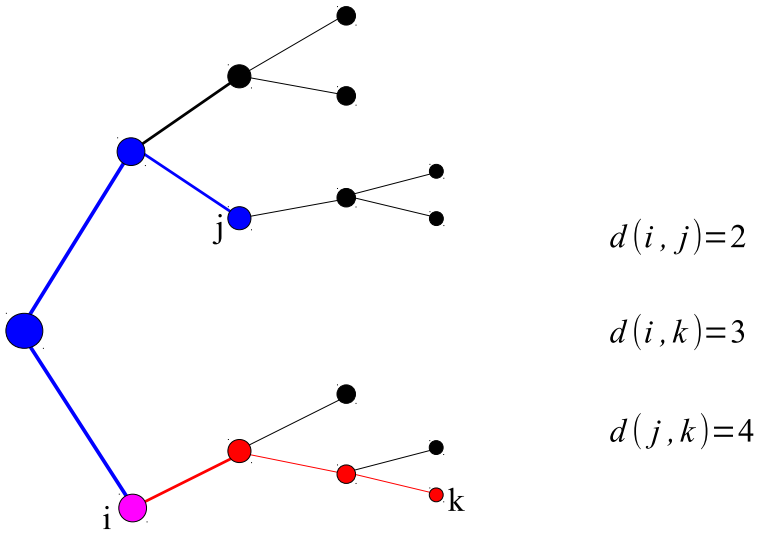


Figure A.19: Illustration of how a partition tree provides a measure of distance between data pairs located in different nodes.

If the distances between each pair of a datum and a labeled datum were recorded during the construction of many partition trees, an estimate of the average distances of any datum from each of the labeled datum may be obtained. It is then possible to infer the labels of the unlabeled data by their proximity to neighboring labeled data. This is a type of instance-based learning where predictions on test data are made based on particular instances in the training data.

Many methodological approaches may be implemented for machine learning. Figure A.20 illustrates some of these methodologies, beginning with purely inductive learning in figure (a) where training data is used to construct a general model of the features for making future predictions. Figure (b) shows the semi-supervised learning approach that was used in the PTLA results of the previous section where unlabeled data was used to aid in the construction of the general model. Figure (c) describes transduction which is the inference from particular to particular [343]. A purely transductive approach does not allow for the classification of any new unlabeled data added after learning has occurred. The instance-based learning shown in figure (d) is similar to transduction, but constructs specific model that relates unlabeled data to the specific instances of the labeled data, and can be used on data added after learning has occurred. This is often accomplished by defining a distance metric on the feature space, such as with the k-nearest neighbor algorithm.

Figure A.20(d) describes the approach that was implemented for the PTLA with the

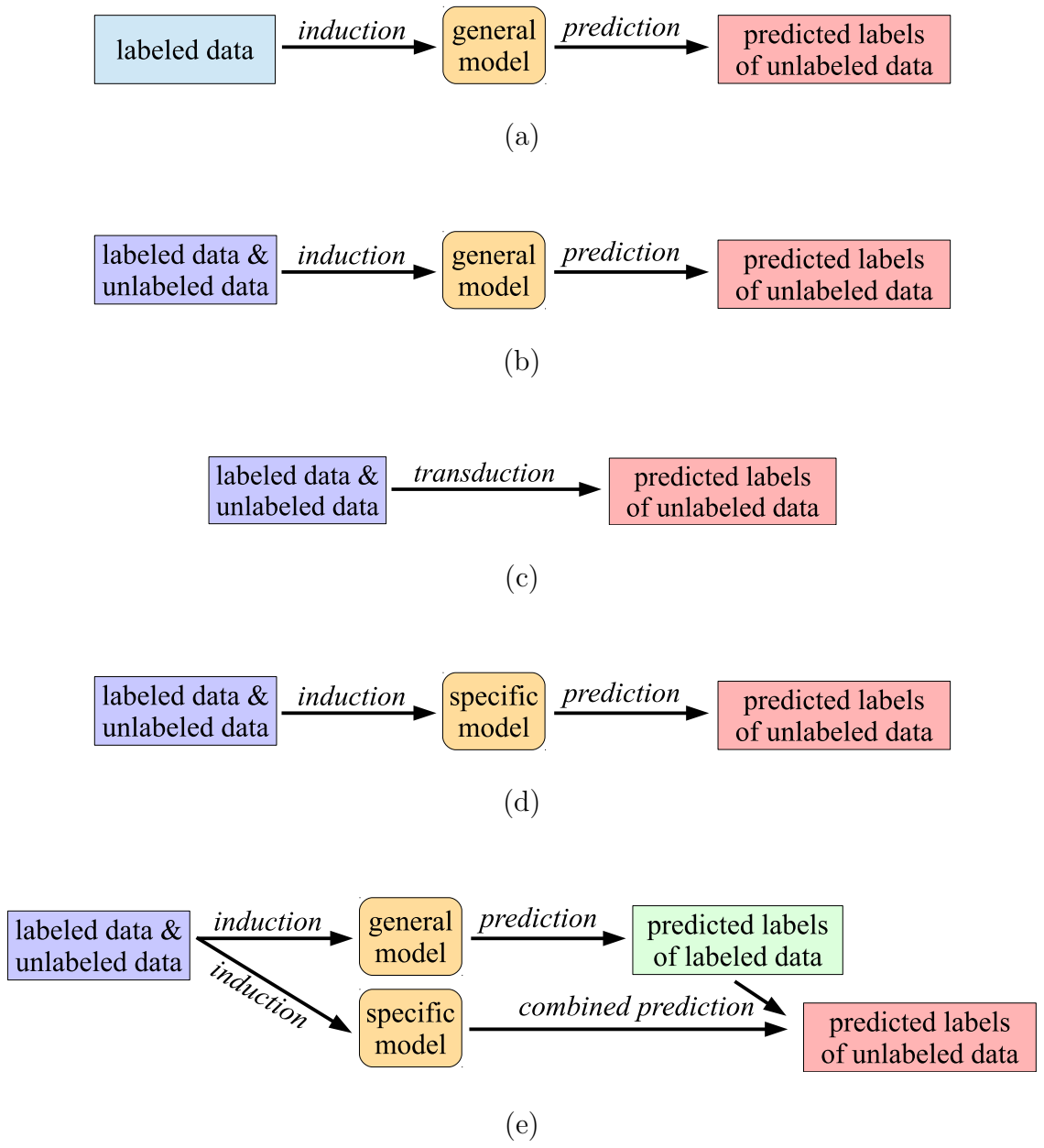


Figure A.20: Some of the methodologies used in machine learning: (a) induction provides a general model that may be used to make predictions for any data, (b) semi-supervised learning uses the unlabeled data in model construction, (c) transduction is the principle of making particular predictions purely from particular instances without the use of any inferred model [340, 343], (d) instance-based learning constructs a specific model, usually based on some measure of distance between data points, and (e) combining semi-supervised learning with instance-based learning.

distance metric described in figure A.19. In this approach, the PTLA is implemented as before with a forest of partition trees in order to construct a general model for prediction, while at the same time keeping track of the distances between all of the data and the labeled data. Theses distances are then used to construct a specific model that relates each datum to its labeled data neighbors.

There are many possible ways to implement the use the PTLA in constructing a specific distance-based model. A simple approach applied here is to determine the nearest labeled datum neighbor l_i for every data point x_j. The classification score for x_j is then simply assigned that of l_j that was determined from the general model (the ensemble PTLA scores described in section A.2), or the average of the closest labeled neighbor scores when more than one labeled data has equal distance to x_j. This approach is depicted in figure A.21 where a distance matrix D is constructed from the nearest neighbor information and is multiplied by the matrix of labeled data ensemble scores E to provide a final classification score S. This process can be performed iteratively.

$$D^{n_{iterations}} \cdot E_{labeled\ ensemble\ scores} = S_{final\ scores}$$

$$N_{all\ data}\left\{\overbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ ½ & 0 & ½ & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ & & \cdots & & & \\ ½ & 0 & 0 & 0 & 0 & ½ \end{bmatrix}}^{\wedge n_{iterations}}\right. \bullet \begin{bmatrix} e_0 \\ e_1 \\ e_2 \\ \cdots \\ e_{N_l} \end{bmatrix} \left.\right\} N_{labeled} = \begin{bmatrix} s_0 \\ s_1 \\ s_2 \\ \\ \cdots \\ \\ s_{N_{all}} \end{bmatrix} \left.\right\} N_{all\ data}$$

$$N_{labeled}$$

Figure A.21: Illustration of how nearest labeled neighbors determined by the PTLA can be combined with the ensemble classification scoring to provide a final classification score.

This procedure was applied to the vlq and cancer datasets with the result shown in figure A.22. Figures (a) and (b) show the results on the vlq and cancer datasets without iteration, while (c) and (d) show the results after 10 iterations. For the vlq dataset, the instance-based scoring does not significantly alter the results from the ensemble scoring shown in figure A.6 b, with some degradation in training-validation agreement after 10 iterations. For the cancer dataset, the instance-based scoring appears to improve the performance on the out-of-sample validation data.

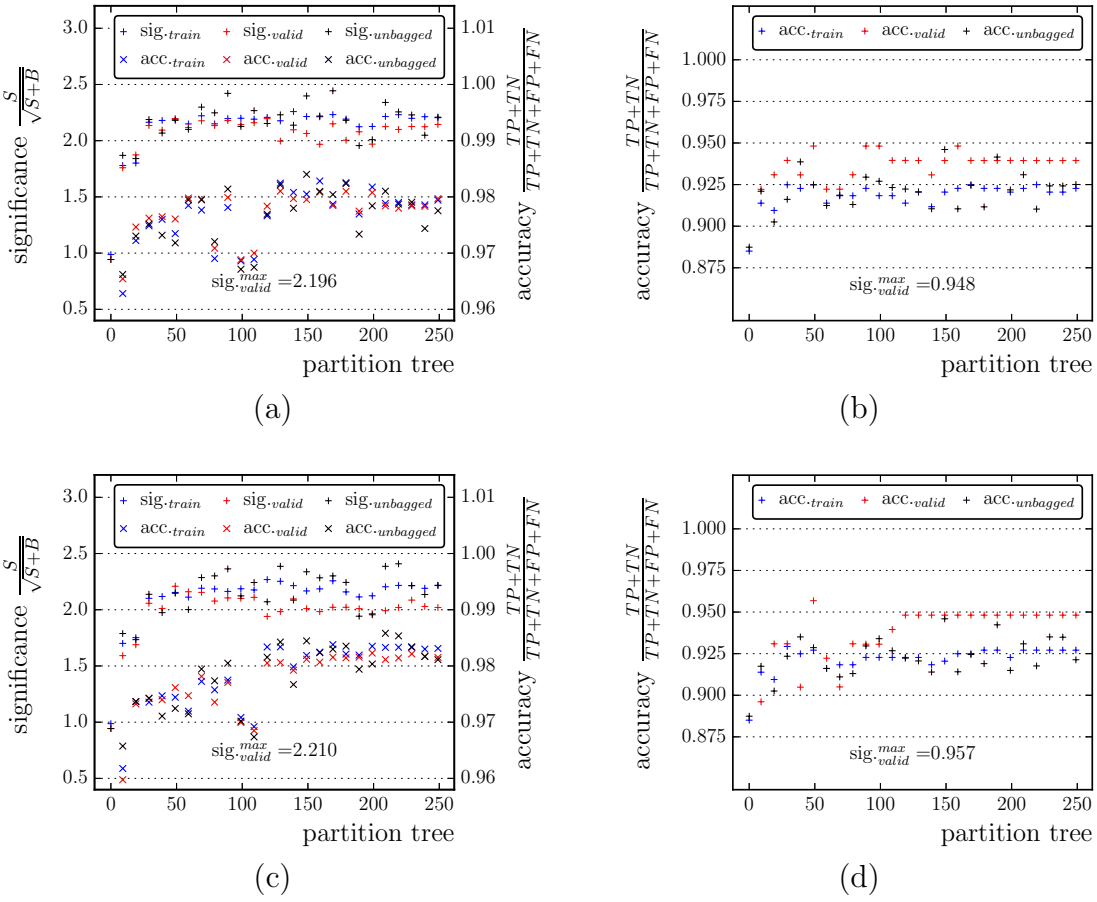


Figure A.22: The performance of the PTLA on the Wisconsin breast cancer dataset. Classification accuracy is defined as the ratio of correctly predicted results (true positive and true negative) to all data. (a) The classification accuracy for individual partition trees in the ensemble (after rejection by semi-supervised learning using the unbagged data). (b) The ensemble performance as partition trees are added.

A.5. Summary of the Partition Trees Learning Algorithm

The PTLA was observed to perform about as well as traditional BDTs on three datasets, the vlq dataset with ∼8,000 events, the higgs dataset with 800,000, and the cancer dataset with ∼600. There are several advantages of the PTLA over BDTs. The PTLA is highly parallelizable since partition trees can be grown independently, making it a good candidate for use on FPGAs and GPUs, and the partitioning of features allows for simple handling of missing and null-valued data. The PTLA is so simple that one could design the *physical structure* of an FPGA to implement it (perhaps with a tree-based architecture [344, 345]).

The PTLA has a minimal memory footprint since data may be stored as unsigned 8-bit integers (and possibly smaller), and it is a fast algorithm since it relies on integer arithmetic and bitwise operations. These characteristics make the PTLA a good candidate for use on limited instruction set processors. On a small processor, if the feature data were already a collection of integer values, then reading the bits of these integers (left to right) may itself provide the partitioning scheme (although this would no longer equally divide the dataset among the bins). It would be interesting to examine the performance of such a "microscale" implementation.

The PTLA also provides a richer tree structure than BDTs, which may allow for instance-based learning methodologies. The results of the simple approach investigated in the previous section could likely be improved upon with different distance measures, by utilizing more of the distance information rather than simply using nearest neighbors, or by use of some other form of graph-based learning.

In addition to semi-supervised and instance-base learning, other machine learning methodologies are possible. *Self-training* is a type of semi-supervised learning procedure where an initial classifier is constructed using the labeled data that provides initial classification predictions for the unlabeled data. A portion of the unlabeled data with the most certain classification predictions is then added to the labeled dataset using the predicted labels, and another round of machine learning is conducted. This process can be performed iteratively. Some self-training was attempted on the higgs dataset, but initial results did not provide much improvement in performance.

Co-training is the process of selecting uncorrelated subsets of features and constructing separate classifiers from each subset. The separate classifier predictions are then used to reinforce each other where strong agreement for a prediction by the separate classifiers increases the certainty in the prediction. An approximate co-training approach was attempted by simply providing each partition tree of the forest with a randomly selected subset of features for use in the construction of that tree. This approach did not alter the performance of the PTLA, neither improving nor degrading performance.

In conclusion, this initial examination of the PTLA demonstrated comparable performance to standard BDTs. The PTLA has certain design characteristics that make it a good candidate for use on lightweight processors or in highly parallel computing environments. This warrants further investigation into the PTLA, both its theoretical properties and possible alternative implementations.

Appendix B
Other Machine Learning Results

B.1. The Cross-Evaluation Partitioning Scheme

Although not investigated in this work, it is possible to combine the cross-evaluation and two-stage partitioning schemes to achieve superior separation without loss of power due to data insufficiency (see section 7.1.1 of [201]). This is shown in table B.1.

One limitation for the implementation of this partitioning scheme when applied to an analysis performing spectral hypothesis testing is the expected yield of measured data in the final signal region. If this is already low so that it is difficult to produce a reasonably well-behaved spectrum, then splitting the measured data in half may degrade the quality of each of the resultant two separate spectra, and thus reduce the power of hypothesis tests.

For example, if only ∼10 measured data events are expected in the final signal region as determined by simulated background data (assuming the non-existence of signal), then a splitting will produce two independent signal regions each with ∼5 expected events. Wide binning of the spectra will be required in order for such spectra to be useful in hypothesis testing, which will degrade sensitivity to individual signal masses (assuming a mass spectrum is being tested). Furthermore, even if ∼10 events are obtained from measured data, it is unlikely that a random splitting measured data will produce two spectra with ∼5 events each. This may result in apparent mismodeling in the final signal region and invalidate the results of hypothesis testing. Due to these reasons, the use of cross-evaluation must be decided at the beginning of an analysis, effectively creating two separate searches, and thus doubling the amount of validation of simulated data quality needed.

	test 1		test 2	
	stage 1	stage 2	stage 1	stage 2
partition 1	train	validate[1]	validate[2] & test	
partition 2	validate[1]	train	validate[2] & test	
partition 3	validate[2] & test		train	validate[1]
partition 4	validate[2] & test		validate[1]	train

Table B.1: Possible data partitioning for two-stage machine learning implementing cross-evaluation[1].

[1] Validate[1] refers to validation of the learning algorithm (checking for overtraining) while validate[2] refers to validation of the modeling accuracy of simulated data compared to measured data.

B.2. Unsupervised Machine Learning to Identify Signal Regions

In order to better understand the performance of stage 2 classification, unsupervised clustering was conducted on the simulated background using all the features from stage 1, along with $H_T(jets)$, $H_T(jets + Z)$, $N(forward\ jets)$, and the testing spectrum $m(Zb)$ (a total of 18 features, which were standardized for use in learning). This analysis was performed after all hypothesis testing was concluded and was therefore an unblinded study. The goal of this study was to examine the regions of the feature space that were likely to have been the most powerful for the hypothesis testing.

The k-means clustering algorithm was used with the number of clusters k determined by simulated modeling data agreement with measured data, by the silhouette score, and by the p-value for Pearson's chi-squared test p_{χ^2}. The k-means clusters were determined from the background training sample, and results were then applied to all other samples for examination of performance.

Figures B.1(a)–(d) show the results of unsupervised clustering for the four signal types. Only the results for the five clusters with maximum $G - \sigma_G$ are shown since these represent regions that most affect hypothesis testing. The upper halves of each figure shows the "violin" plots of the (unweighted) density distributions for the stage 2 classifier response value of the testing background sample (solid gray), of the testing signal sample with all signal masses combined (solid red), and of the measured data (black line). These distributions were constructed using Gaussian kernel smoothing to aid in visual inspection. The training samples for background and signal are shown with blue dotted lines. A green dashed line shows the decision rule (cut on stage 2 classifier value that was used).

The lower halves of figures B.1(a)–(d) show the predicted and measured yields after applying the decision rule. The gray bar shows the predicted background yield of the testing partition with a gray box showing statistical error, the measured data is shown with a black cross with height measuring statistical error. Red bars show the predicted signal yields for each of the signal masses (in increasing mass). Blue diamonds show the predicted yields of the training partitions.

Accurate simulation modeling was the primary determinant in selection of k. This was performed by visual inspection for similarity of density distributions between the background and measured data below the classifier decision rule (in the upper halves, comparing the solid gray to black line below the dashed green line: a control region). Modeling agreement was rarely perfect, but generally good, with much of the visual disagreement arising from the density smoothing and the use of unweighted events. The modeling quality of the stage 2 classifier value was then determined by examining the predicted yields between background and measured data (in the lower halves, gray bar and black cross), assuming the non-existence of the signals (as supported by the hypothesis test results of section 5.10).

The silhouette score measures how well the data is separated in the learned clusters, and is defined by [346]

$$\text{silhouette} = \text{ave}\left[\frac{b_i - a_i}{\max[a_i, b_i]}\right], \tag{B.2.1}$$

where a_i is the average Euclidean distance between each data sample i, and the other data

samples belonging to the same cluster, and b_i is the average distance between i and the data samples belonging to the nearest neighboring cluster. The silhouette score ranges from -1 to 1, with a large silhouette indicating good separation between clusters. The silhouette score was observed to slowly decrease with increasing k and so was of limited use in selection of k.

Pearson's categorical test compared training and testing background sample frequencies in each cluster. If the frequency of training events occurring in a particular cluster was different than the frequency of testing events, then it was likely that overtraining had occurred. The p-values of this test were observed to never reject training-testing frequency consistency (at 10% confidence) over a wide range of k, and so was of limited use in selection of k. This was found to be robust by also checking the p-values from log-likelihood and Cressie-Read tests.

The k-means unsupervised learning also provided an additional check for overtraining of the stage 2 classifier. Visual inspection training versus testing partitions (both with unweighted densities in the upper halves, and predicted yields of the lower halves) generally showed excellent agreement for most signal masses. By using unsupervised learning to find hidden structure only within the background, and then selecting clusters rich in signal, important "cluster control regions" were formed that were useful in evaluating the stage 2 classifier performance. Tables B.2(a)–(d) shows the cluster centers for the selected clusters shown in figuresB.1(a)–(d), respectively, and were used to examine the properties of these control regions. For example, the lower half of figure B.1(c) shows that in cluster 6, there is training-testing partition disagreement in the lower signal masses. The last column of table B.2(c) shows the location of this cluster. Events belonging to this cluster had a highly energetic Z boson (high $p_T(Z)$ and low $\Delta R(l_1, l_2)$), a very high $H_T(jets)$, and would appear near 750 GeV of the $m(Zb)$ spectrum during hypothesis testing. Since it was not expected for this region of $m(Zb)$ to be useful for the lower signal masses during hypothesis testing, this disagreement is unimportant.

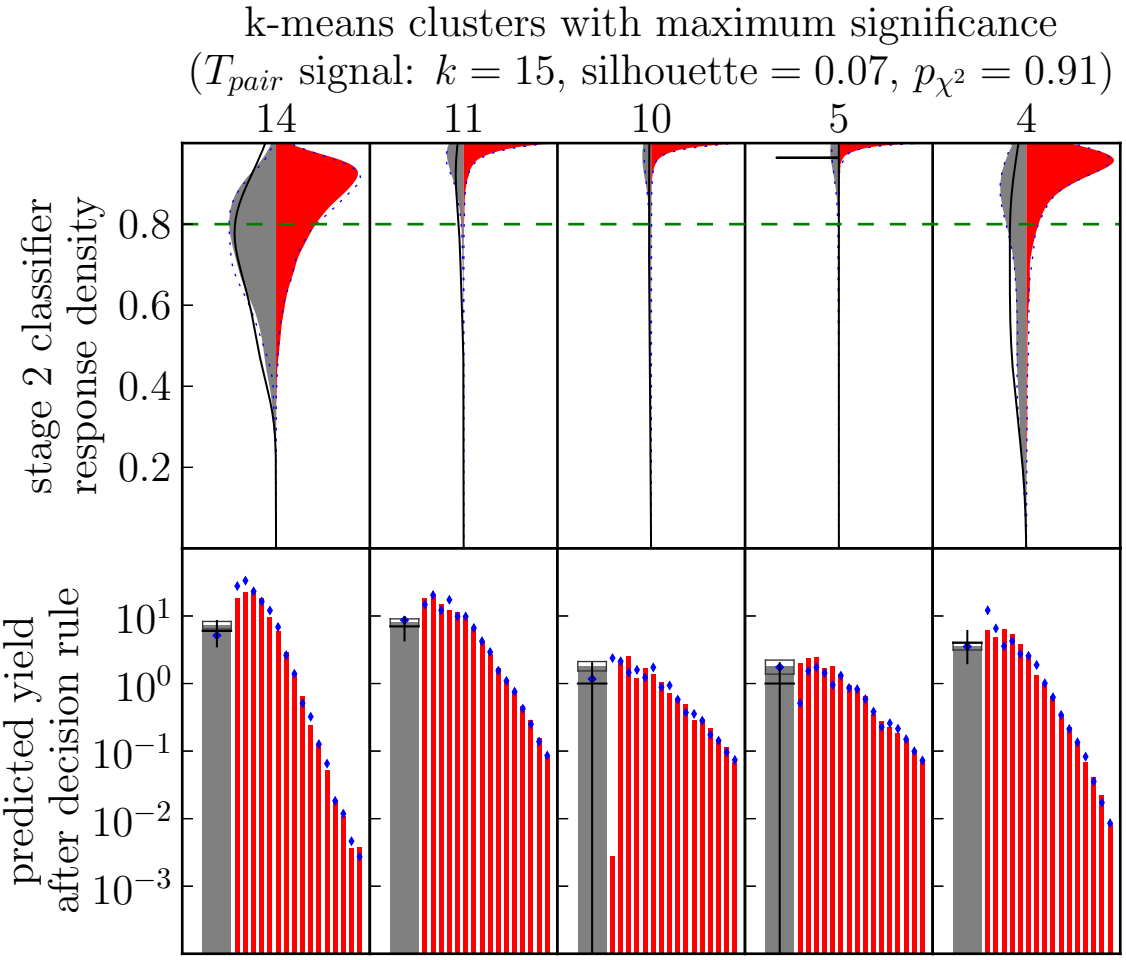
The last row of tableB.2(a) indicated that the use of $N(forward\ jets) \geq 1$ could likely be replaced with a $N(forward\ jets) = 0$ cut for the T_{pair} signal. This is because the 5 clusters with highest significance showed little or no presence of forward jets. However, tables B.2(b) and (d) show that for the B_{pair} and B_{single} signals, clusters formed around both high and low values of $N(forward\ jets) \geq 1$. (The T_{single} signal decision rule included a forward jet requirement so all clusters listed in table B.2(c) are centered at 1.)

The use of k-means clustering was motivated by its simplicity. The high correlation between several of the features does not satisfy the assumption of diagonal covariance structure of the features (for example, see figure 5.6). A Gaussian mixture model (GMM) that allows for correlation between features was also implemented. The main difference of GMM clustering from k-means is the use of the Mahalanobis metric instead of Euclidean. The Mahalanobis distance between a data event with feature values $\vec{x}$ and a cluster c with feature means $\vec{m}_c$ is given by

$$D_M\left(\vec{x}, \vec{m}_c\right) = \sqrt{\left(\vec{x} - \vec{m}_c\right)^T \hat{V}_c^{-1} \left(\vec{x} - \vec{m}_c\right)}, \tag{B.2.2}$$

where V_c is the feature covariance matrix for cluster c. The use of feature correlations in the distance metric allowed for the construction of non-spherical clusters. However, data

sparsity prevented the accurate estimation of the covariance matrices for the clusters, and nearly always led to overtraining, even after optimizing c using the Akaike information criterion (AIC). Because only a few of the features were correlated while the feature volume was very large (18 dimensional), k-means clustering was determined to provide the more robust results.



(a) T_{pair}

Figure B.1: K-means unsupervised clustering is trained on only the background data, and then the resultant clusters are applied to all other data. The five clusters with highest signal significance are shown. The upper half shows the Gaussian kernel smoothed frequency distributions (with smoothing factor $(.75n)^{-\frac{1}{5}}$ [347]: testing background data (solid gray areas), measured data (solid black lines), testing signal data with all signal masses combined (solid red areas), and training data (dotted blue lines), with the stage 2 decision rule (dashed green line). The lower half shows the predicted yields after applying the decision rule: testing background data (gray bars), measured data (black cross), testing signal data by increasing signal mass (red bars, specific masses unlabeled), and training data (blue diamonds).

(Continues on next three pages.)

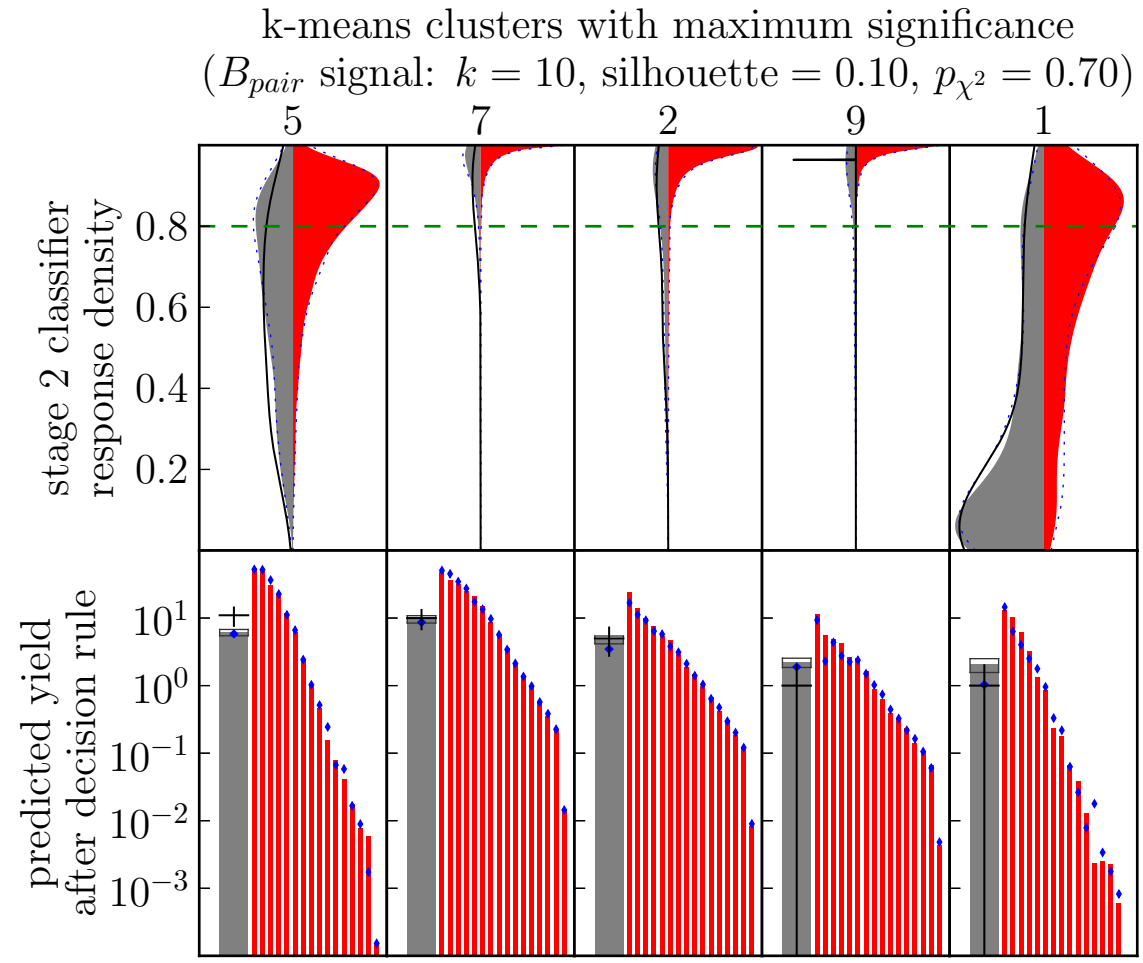


Figure B.1: (b) B_{pair}

(Continues on next two pages.)

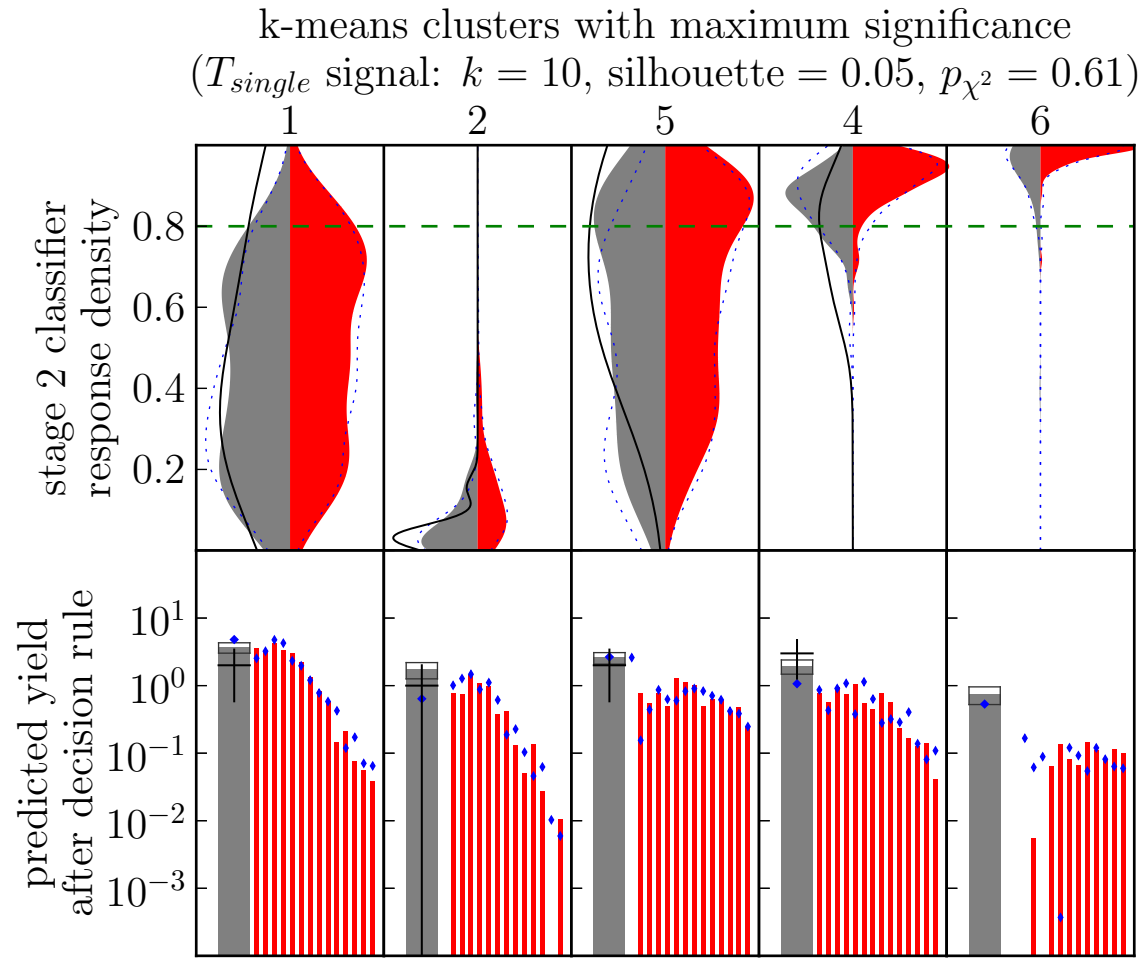


Figure B.1: (c) T_{single}

(Continues on next page.)

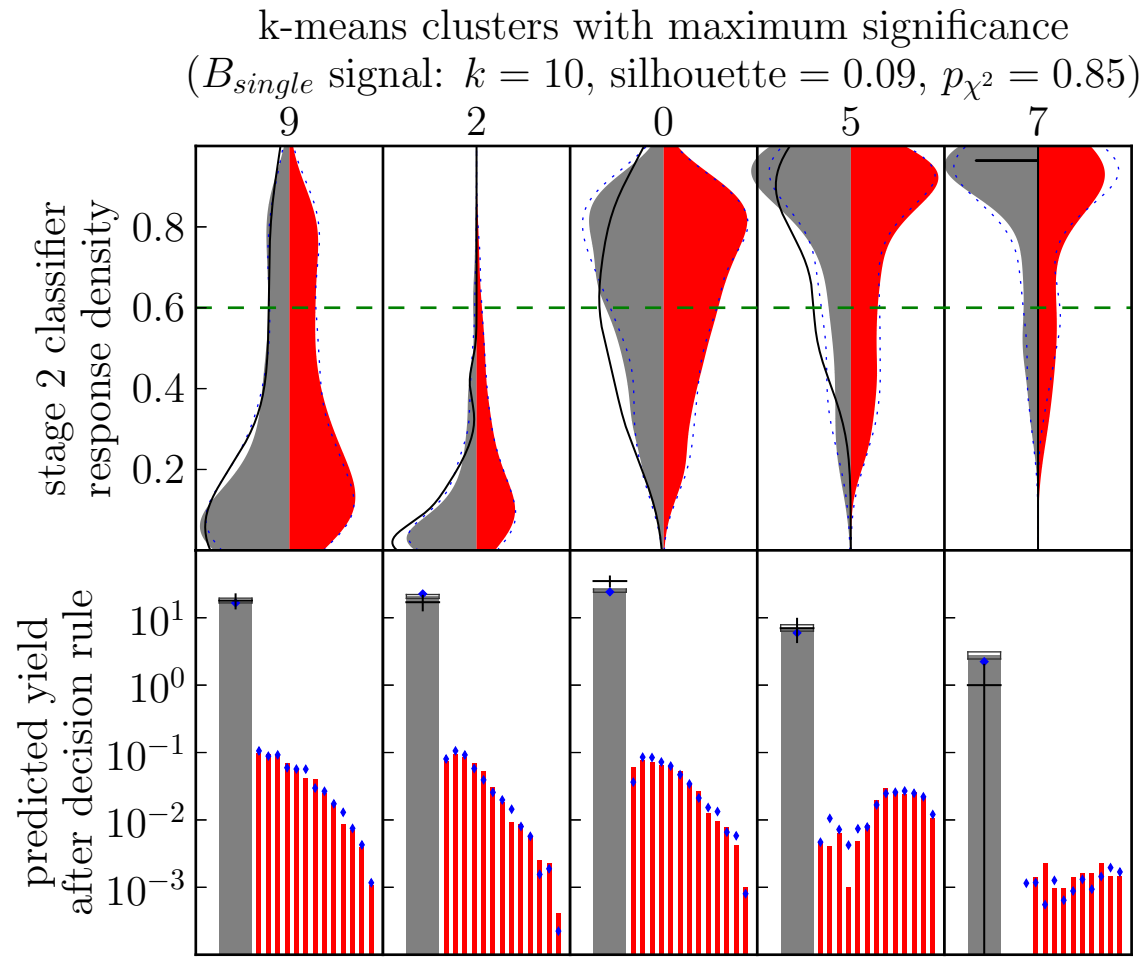


Figure B.1: (d) B_{single}

(Last page.)

	cluster ($k = 15$)				
	14	11	10	5	4
$p_T(Z)$	243	238	598	599	439
$\|y(Z)\|$	0.69	0.69	0.62	0.69	0.65
$\Delta R(l_1, l_2)$	0.86	0.91	0.42	0.42	0.50
$\Delta\phi(l_1, l_2)$	0.53	0.64	0.26	0.28	0.31
$\Delta R(Z, b_1)$	2.50	2.02	3.02	2.11	2.87
$\phi(Z, T_T)$	1.52	1.60	1.51	1.56	1.70
$H_T^Z(jets)$	370	677	359	345	237
$T_T - T_M$	0.34	0.55	0.53	0.60	0.49
S_T	0.67	0.49	0.57	0.55	0.53
$p_T(b_1)$	164	217	473	133	173
$\Delta R(b_1, b_2)$	2.15	2.04	1.67	1.64	1.41
$p_T(a_1)$	164	353	137	596	194
$\Delta R(b_1, a_1)$	2.15	2.46	1.84	2.08	1.32
$\Delta R(b_2, a_1)$	2.14	2.06	2.10	1.96	1.52
$m(Zb)$	429	421	1205	520	601
$H_T(jets)$	562	937	798	922	556
$H_T(jets + Z)$	813	1186	1399	1525	1000
$N(forward\ jets) \geq 1$	0.00	0.06	0.06	0.04	0.06

(a)

	cluster ($k = 10$)				
	5	7	2	9	1
$p_T(Z)$	292	258	506	570	220
$\|y(Z)\|$	0.73	0.69	0.73	0.71	0.82
$\Delta R(l_1, l_2)$	0.73	0.87	0.49	0.44	0.96
$\Delta\phi(l_1, l_2)$	0.42	0.59	0.31	0.30	0.66
$\Delta R(Z, b_1)$	2.75	2.16	3.14	2.33	2.71
$\phi(Z, T_T)$	1.54	1.55	1.46	1.56	1.49
$H_T^Z(jets)$	285	678	272	299	185
$T_T - T_M$	0.33	0.54	0.57	0.60	0.44
S_T	0.69	0.49	0.51	0.52	0.54
$p_T(b_1)$	161	233	390	116	128
$\Delta R(b_1, b_2)$	1.83	1.99	1.74	1.63	1.86
$p_T(a_1)$	141	354	114	515	103
$\Delta R(b_1, a_1)$	1.83	2.39	1.91	1.84	2.05
$\Delta R(b_2, a_1)$	1.90	2.13	2.05	1.84	2.22
$m(Zb)$	484	483	1051	506	386
$H_T(jets)$	489	954	678	818	353
$H_T(jets + Z)$	787	1221	1189	1392	584
$N(forward\ jets) \geq 1$	0.01	0.06	0.06	0.05	1.00

(b)

Table B.2: —*continues on next page.*

	cluster ($k = 10$)				
	1	2	5	4	6
$p_T(Z)$	242	195	349	282	524
$\|y(Z)\|$	0.51	0.74	0.87	0.68	0.82
$\Delta R(l_1, l_2)$	0.83	1.02	0.65	0.77	0.47
$\Delta\phi(l_1, l_2)$	0.50	0.54	0.44	0.50	0.37
$\Delta R(Z, b_1)$	2.60	2.89	3.17	2.26	2.75
$\phi(Z, T_T)$	0.99	1.00	1.08	1.76	1.19
$H_T^Z(jets)$	205	100	166	440	496
$T_T - T_M$	0.30	0.55	0.58	0.29	0.44
S_T	0.68	0.40	0.55	0.69	0.50
$p_T(b_1)$	96	104	244	171	226
$\Delta R(b_1, b_2)$	1.92	1.24	1.86	2.28	1.49
$p_T(a_1)$	137	59	86	178	307
$\Delta R(b_1, a_1)$	1.77	1.33	1.55	1.93	1.45
$\Delta R(b_2, a_1)$	1.98	1.37	2.34	2.44	1.80
$m(Zb)$	334	342	707	438	752
$H_T(jets)$	370	229	452	650	822
$H_T(jets + Z)$	620	432	808	940	1351
$N(forward\ jets) \geq 1$	1.00	1.00	1.00	1.00	1.00

(c)

	cluster ($k = 10$)				
	9	2	0	5	7
$p_T(Z)$	221	204	287	507	559
$\|y(Z)\|$	0.83	0.74	0.68	0.74	0.72
$\Delta R(l_1, l_2)$	0.95	1.01	0.74	0.49	0.44
$\Delta\phi(l_1, l_2)$	0.67	0.71	0.41	0.31	0.30
$\Delta R(Z, b_1)$	2.72	3.14	2.76	3.14	2.41
$\phi(Z, T_T)$	1.52	1.42	1.51	1.49	1.59
$H_T^Z(jets)$	187	116	287	271	288
$T_T - T_M$	0.44	0.68	0.36	0.57	0.58
S_T	0.53	0.29	0.67	0.51	0.53
$p_T(b_1)$	130	166	173	393	112
$\Delta R(b_1, b_2)$	1.83	2.05	1.96	1.71	1.59
$p_T(a_1)$	103	77	134	113	474
$\Delta R(b_1, a_1)$	2.08	2.37	1.84	1.93	1.81
$\Delta R(b_2, a_1)$	2.25	2.93	1.98	2.07	1.78
$m(Zb)$	390	454	493	1056	513
$H_T(jets)$	357	334	496	677	772
$H_T(jets + Z)$	589	549	789	1189	1335
$N(forward\ jets) \geq 1$	1.00	0.00	0.00	0.06	0.06

(d)

Table B.2: Central values of the features for the k-means clusters providing maximum signal significance: (a) T_{pair}, (b) B_{pair}, (c) T_{single}, and (d) B_{single}.

B.3. Decision Tree Regression: Gradient vs. Adaptive Boosting

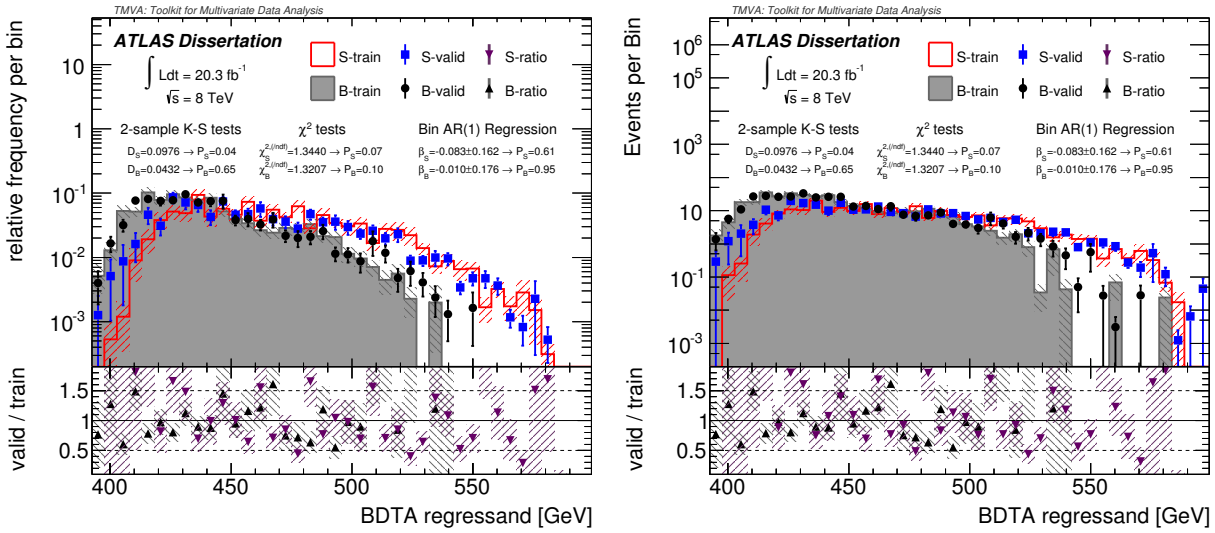
Figure B.2 shows the result of decision tree regression using adaptive boosting. Unlike the successful decision tree regression using gradient boosting discussed in section 5.7.2, adaptive boosting demonstrated poor signal-background separation, as well as severely contracting the signal mass predictions to incorrect values. Figures B.2(a)–(c) show these poor regression results. Attempts to correct this problem via hyper-parameter tuning were unsuccessful.

It is puzzling that similar boosting procedures should provide such differing regression results, accurate for gradient boosting while poor for adaptive boosting. A short investigation was conducted by applying the adaptive and gradient boosting procedures to OLS regression, for which an exact solution is known, and applied to the particle physics data used in the search for vector-like quarks. Each of the regression fit parameters were initialized to the negative sign of the OLS solution, $\beta_0 = -\text{sign}\left[\beta_{OLS}\right]$, far from the true quadratic loss minimum, and boosting proceeded for 1000 iterations.

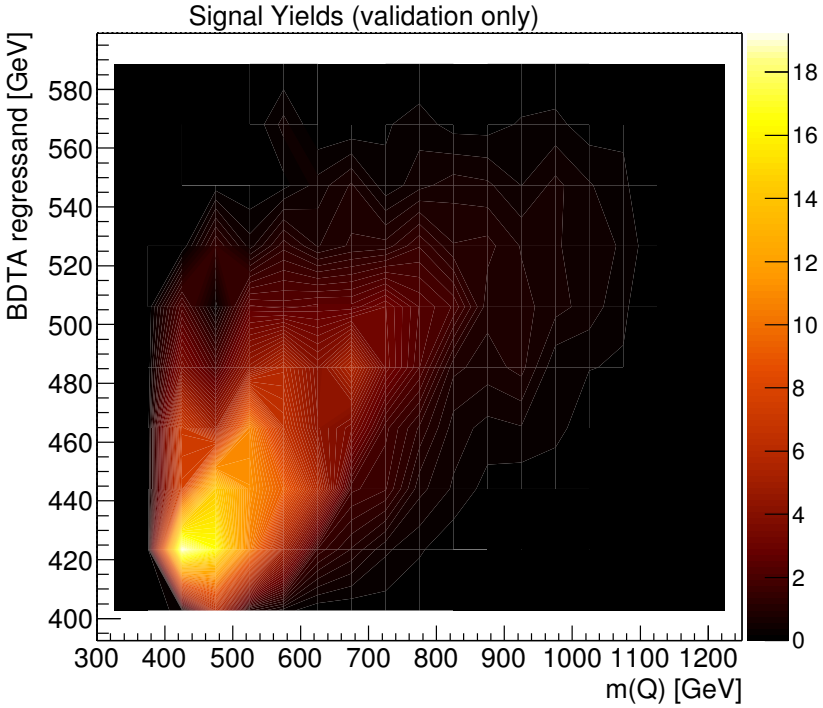
Figures B.3(a)–(d) show the evolution of the adaptive boosted fit parameters for four time steps. The fit parameters of adaptive boosting rapidly evolved away from the initial (highly incorrect) values, then slowly converged to a non-minimal solution. Figure B.4 shows the evolution of the adjusted R^2 difference vs. average fit parameter difference from the correct solution. As the iterations proceeded, the training and validation samples converged so that there was no overfitting, but the quality of the fit did not as determined by the coefficient of determination $\bar{R}^2$.

Figures B.5(a)–(d) and B.6 show the evolution of the gradient boosted fit parameters. For gradient boosting, fit parameters rapidly descend away from initial values and continue to improve fit quality, albeit at an increasingly slow rate. As with adaptive boosting, the fit parameters do not converge to the correct OLS solution (within 1000 iterations), but nevertheless provide increasingly higher quality fits.

These results are best conceptualized with the analogy of a person hiking in a mountainous terrain, one with many peaks and valleys, who seeks to descend as far as possible. The lowest point at the bottom of some particular valley represents the true OLS solution, and the person begins the hike at some location far away from the lowest point. With gradient descent, the person quickly reaches *some* valley, regardless of the complexity of the terrain, and then slowly walks down that valley toward the lowest elevation that can be obtained within it. This may not be the lowest point in all the terrain, but may be "good enough" for the person's purposes. The results of this quick study confirm that although gradient boosting may not lead to the absolute minimum of some loss function, it is still likely to find a near minimum and produce a useful result (when applied to this particular data set).



(a) (b)



(c)

Figure B.2: Stage 3 BDTA regression spectra for the T_{single} signal (before cut on stage 2 classifier response value): (a) frequency spectra for all signals combined, (b) the predicted yield spectra with all signals combined, and (c) a heat map showing predicted signal mass vs. actual signal mass.

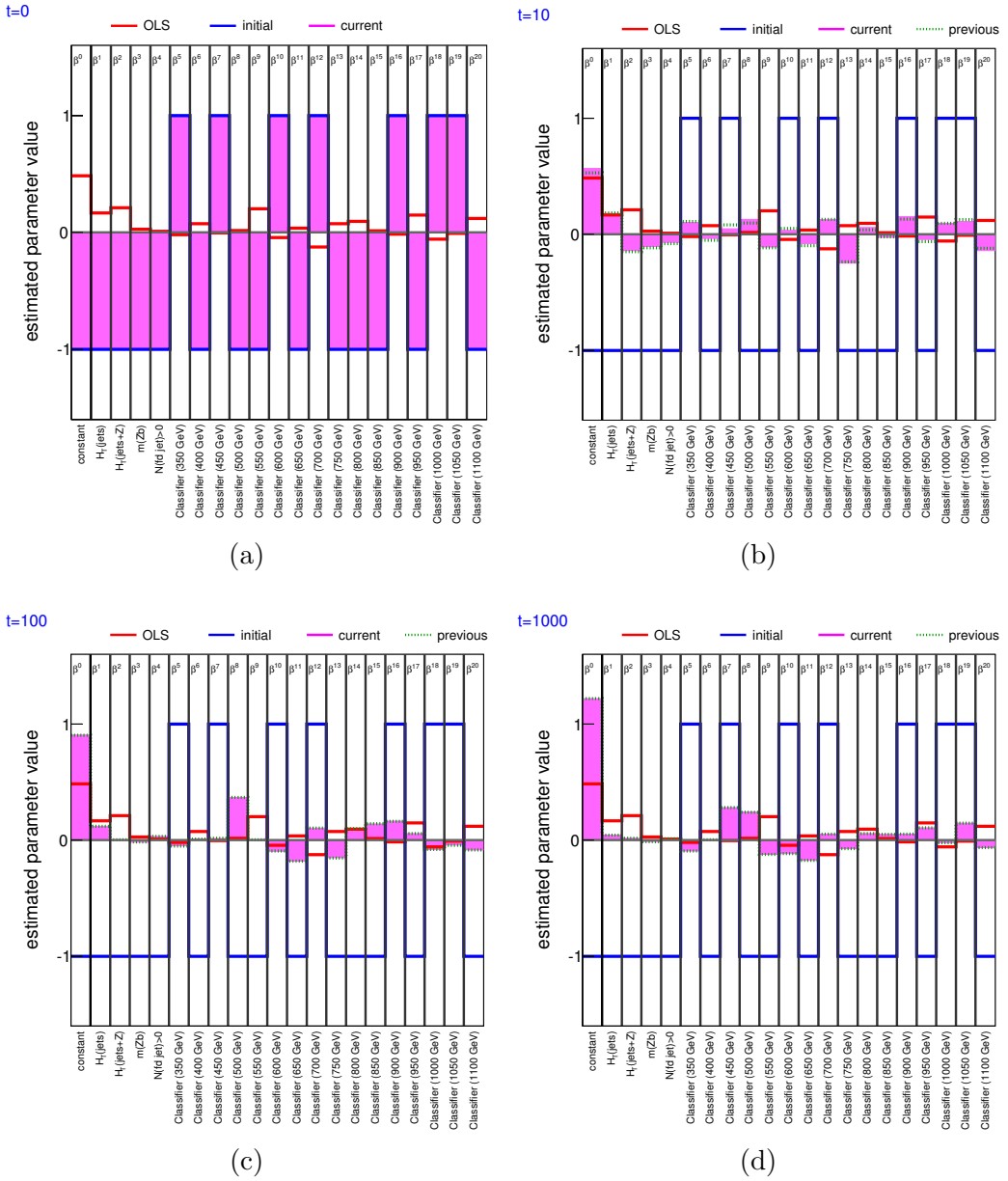


Figure B.3: Linear regression using adaptive boosting for the T_{pair} signal at time steps (a) 0, (b) 10, (c) 100, and (d) 1000.

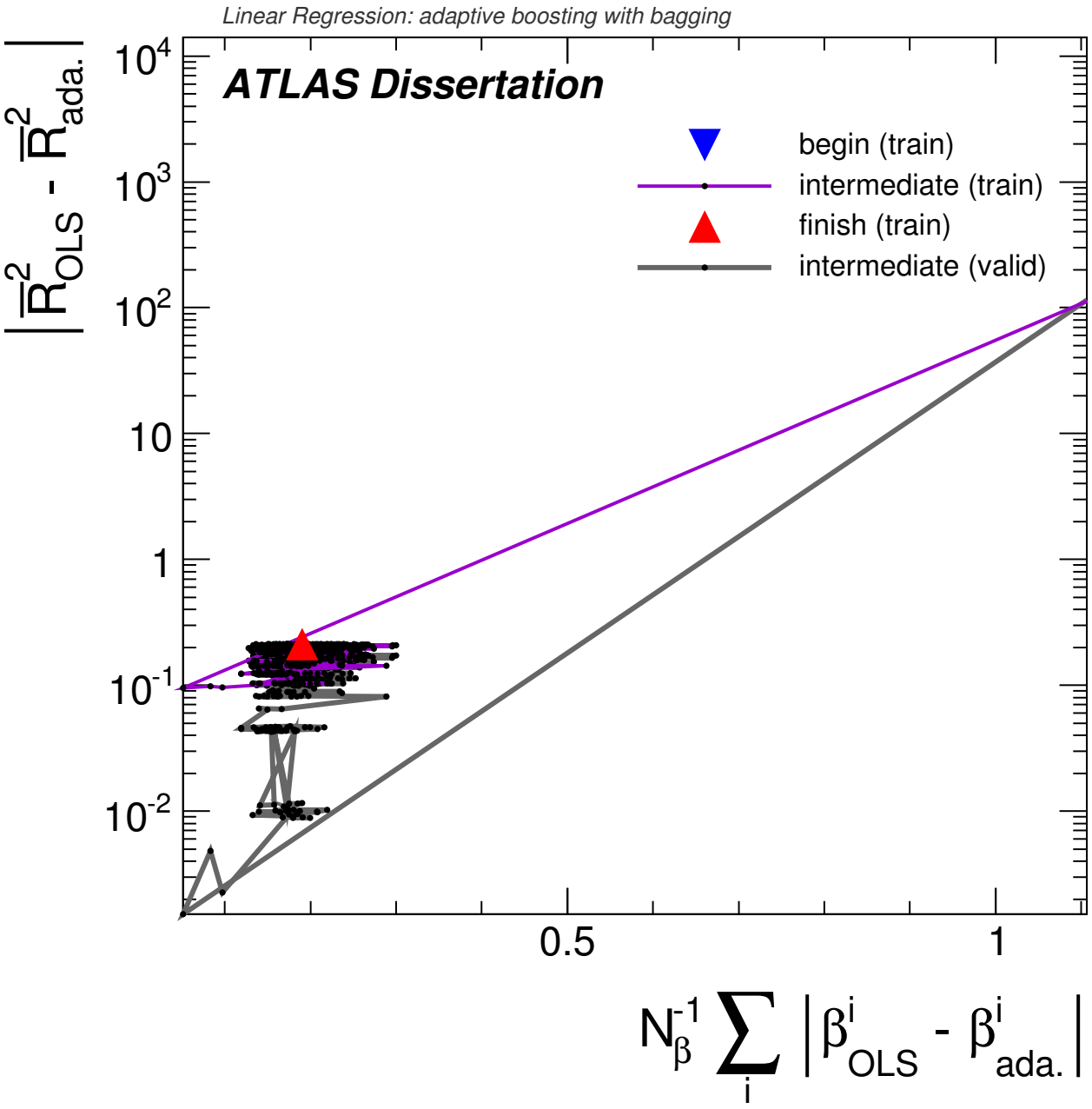


Figure B.4: Analysis of linear regression using adaptive boosting for the T_{pair} signal.

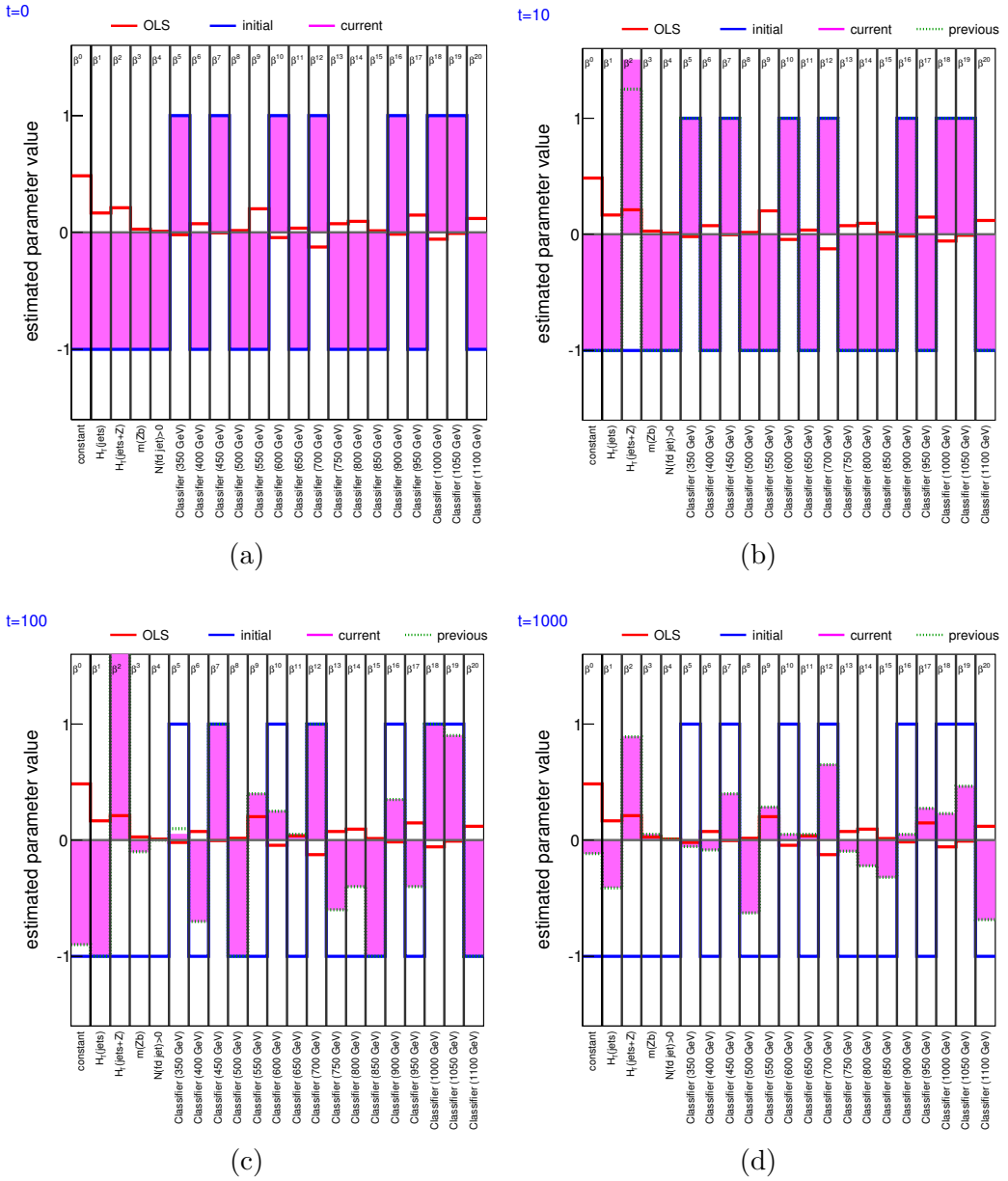


Figure B.5: Linear regression using gradient boosting for the T_{pair} signal at time steps (a) 0, (b) 10, (c) 100, and (d) 1000.

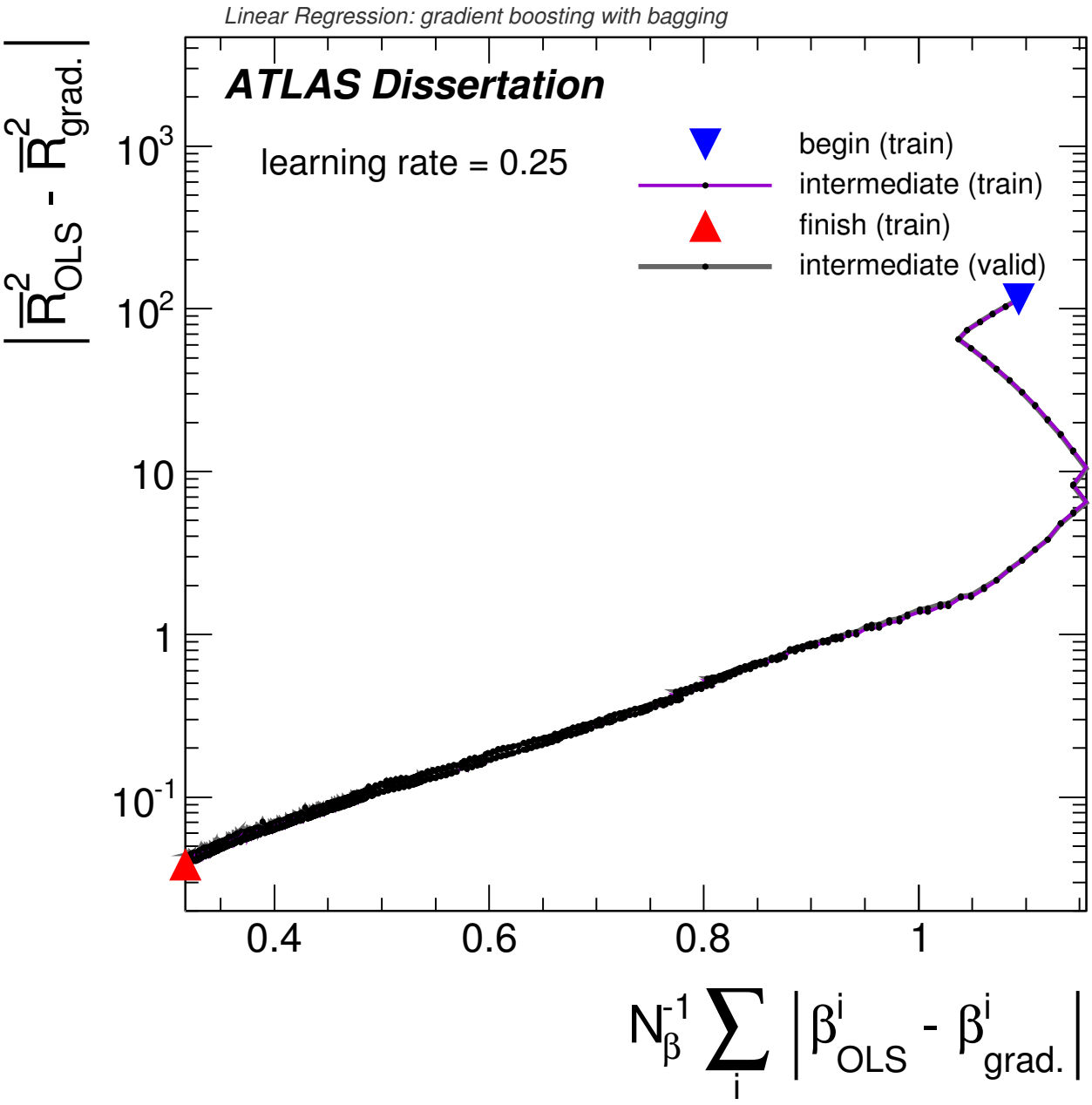


Figure B.6: Analysis of linear regression using gradient boosting for the T_{pair} signal.

Appendix C
The Search Using Machine Learning (Addendum)

C.1. Stage 2 Classification with Other Methods

Learning algorithms other than logistic regression were investigated for use in stage 2 classification: support vector machines, multilayer perceptrons and boosted decision trees. These were observed to underperform logit classification by either having more overtraining or by introducing bias into the simulated data.

C.1.1. Support Vector Machines

SVMs provided good signal separation but were hindered by background overtraining. TMVA provides SVMs with a Gaussian kernel with a single parameter σ_{SVM}, and a cost parameter C_{SVM} for non-separable data. With only two parameters to set, a quick scan was conducted and little variation in performance observed over a wide range of parameters. Figures C.1 (a)–(d) show the background and signal SVM classifier response frequency distributions. The test statistics indicated the presence of background overtraining, especially at high response value. This was problematic because modeling accuracy of simulated data was later found to degrade after implementing a decision rule, and prevented the use of the SVM classifier for final hypothesis testing. It was important that classifiers exhibit no background overtraining at high response value because it was only these "outlier" background events that would be used after implementing a decision rule. It was likely that SVMs, although providing reasonable classification performance overall, struggled to correctly classify the small amount of background present in extreme signal-like regions.

C.1.2. Multilayer Perceptrons

MLPs have a many parameters to select, and an exhaustive search was prohibitive. Several selections of neural network structure were attempted, but results were consistently inferior to logit classification. Figures C.2 (a)–(d) show the results for an example MLP classification using two hidden layers with N_f and $N)f-1$ nodes where N_f was the number of features (the stage 1 BDT classifiers and three additional features, $H_T(jets)$, $H_T(jets+Z)$ and $N(forward\ jets) \geq 1$. These figures show the presence of a large amount of background overtraining (least in the B_{single} signal shown in figure C.2 (d)), and was a common problem to all MLP attempts. The learned neural networks are shown in figures C.3.

Note that the stage 2 logit classification that was successful was performed using an MLP with a single node. This suggested that the "deeper" the neural network implemented, the greater the overtraining. The stage 1 BDTs were used to model a complex target function, and then used as input features for stage 2. After stage 1 BDTs had extracted the complex information contained in the initial 14 features, the stage 2 classification sought to model a much less complex target function. However, there was likely a large amount of statistical noise accompanying the stage 1 BDT classifiers as well as the H_T variables which are known

to be noisy. To model a low complexity target function in a noisy environment, less complex learning algorithms were observed to perform better than complex ones [348].

C.1.3. Boosted Decision Trees

The negative results of the stage 2 MLPs suggested the use of low complexity BDTs. The best results were found using a 500 tree forest, each tree with only a maximum depth of 2, and using highly randomized forests with only 2 randomly selected features available to each node (and with 50% bagging, gradient boosting with 0.05 shrinkage). The results were comparable with logit classification for the T_{single} and B_{single} signals, while the CDFs shown in figures C.4 (a)–(b) indicate some overtraining for the T_{pair} and B_{pair} signals. Later analysis of the stage 2 BDT classifiers for the T_{single} and B_{single} signals showed slightly more modeling inaccuracy than with logit classification, and so logit classification remained the selected stage 2 classification method.

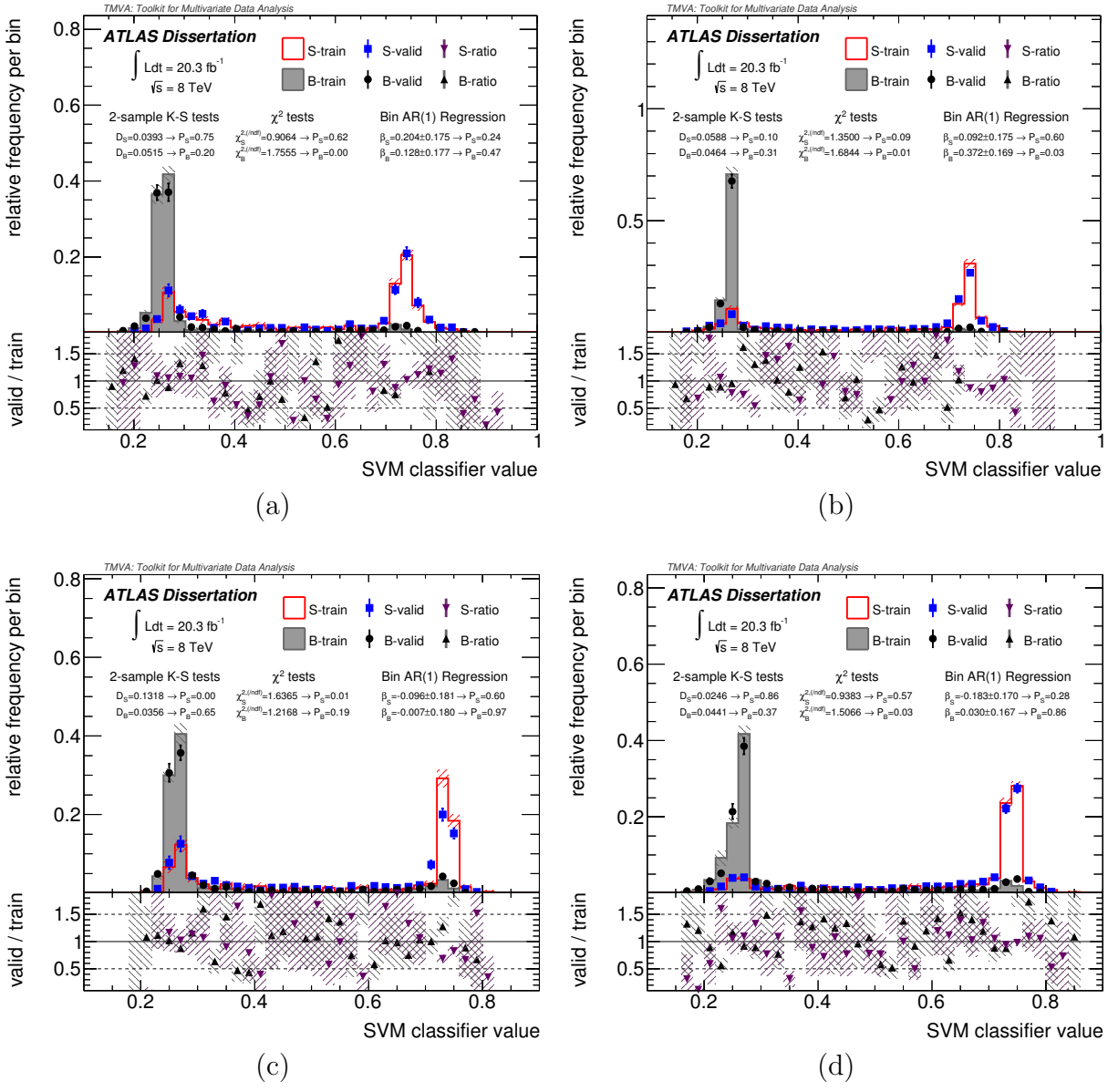


Figure C.1: SVM classifier response background and signal frequency distributions (training and validation partitions): (a) T_{pair} signal, (b) B_{pair} signal, (c) T_{single} signal, and (d) B_{single} signal.

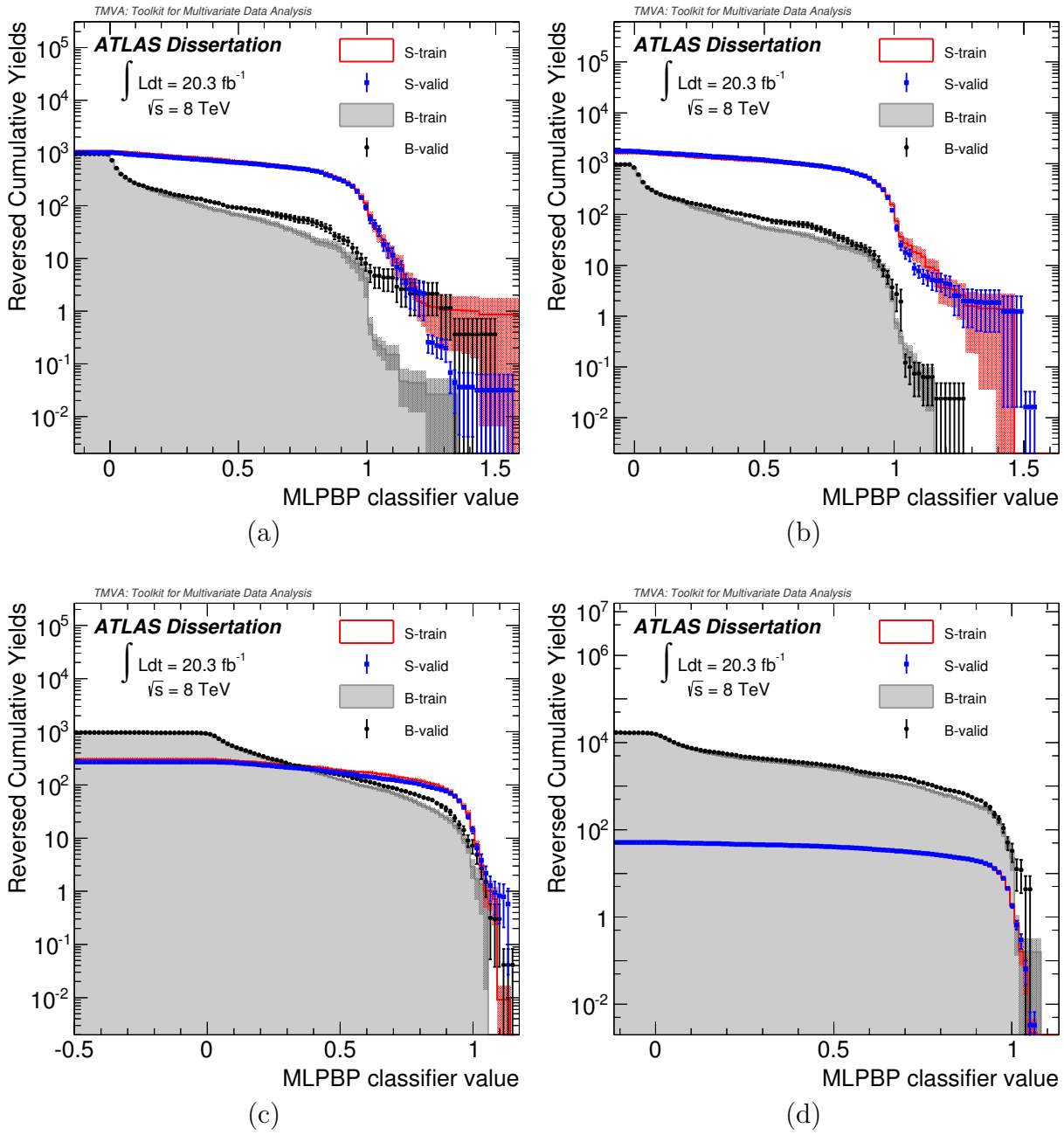


Figure C.2: MLP classifier response background and signal cumulative yield distributions (training and validation partitions, using two hidden layers with N and $N-1$ nodes and back propagation (MLPBP)): (a) T_{pair} signal, (b) B_{pair} signal, (c) T_{single} signal, and (d) B_{single} signal.

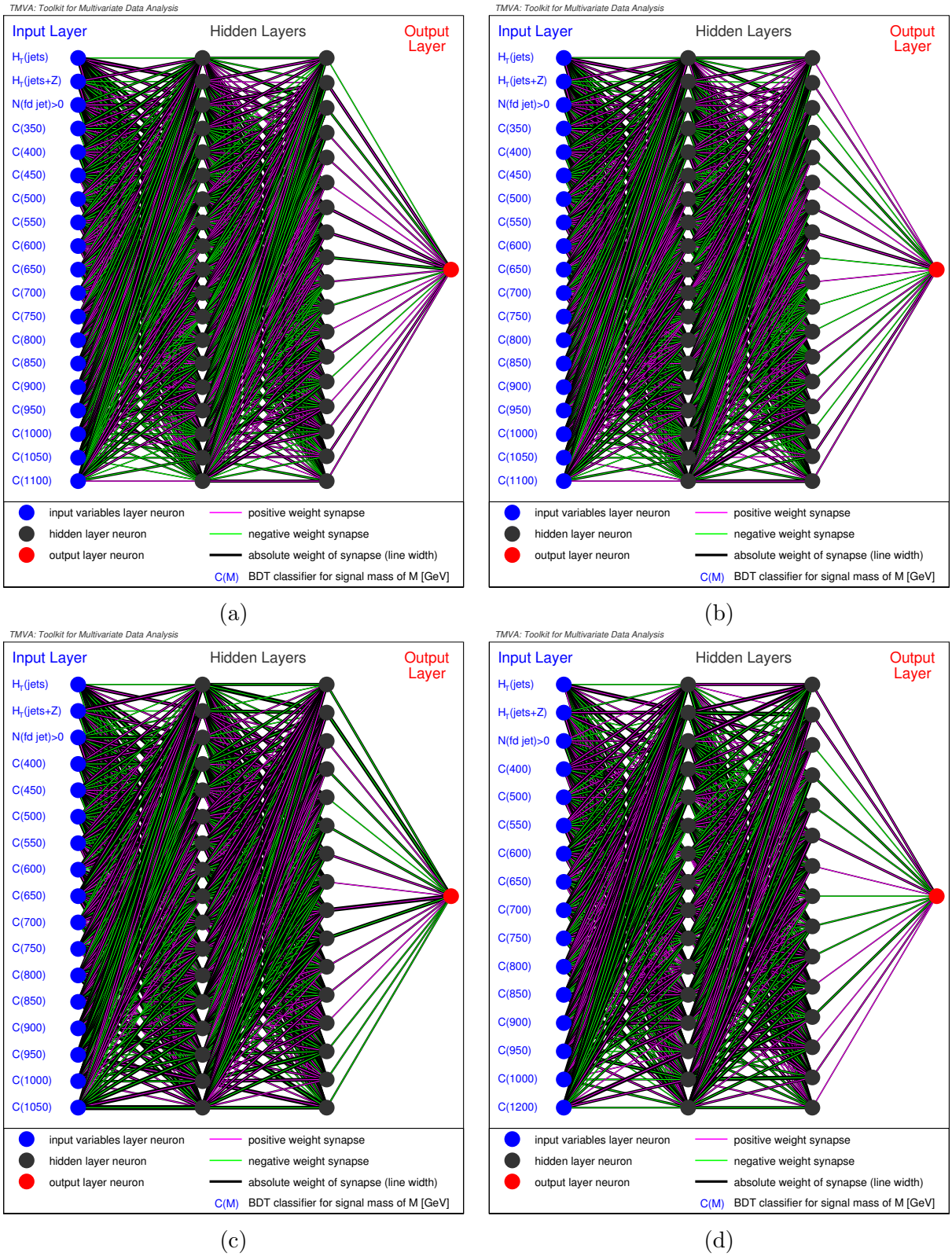


Figure C.3: The neural networks for the stage 2 multilayer perceptrons for the (a) T_{pair} signal, (b) B_{pair} signal, (c) T_{single} signal, and (d) B_{single} signal.

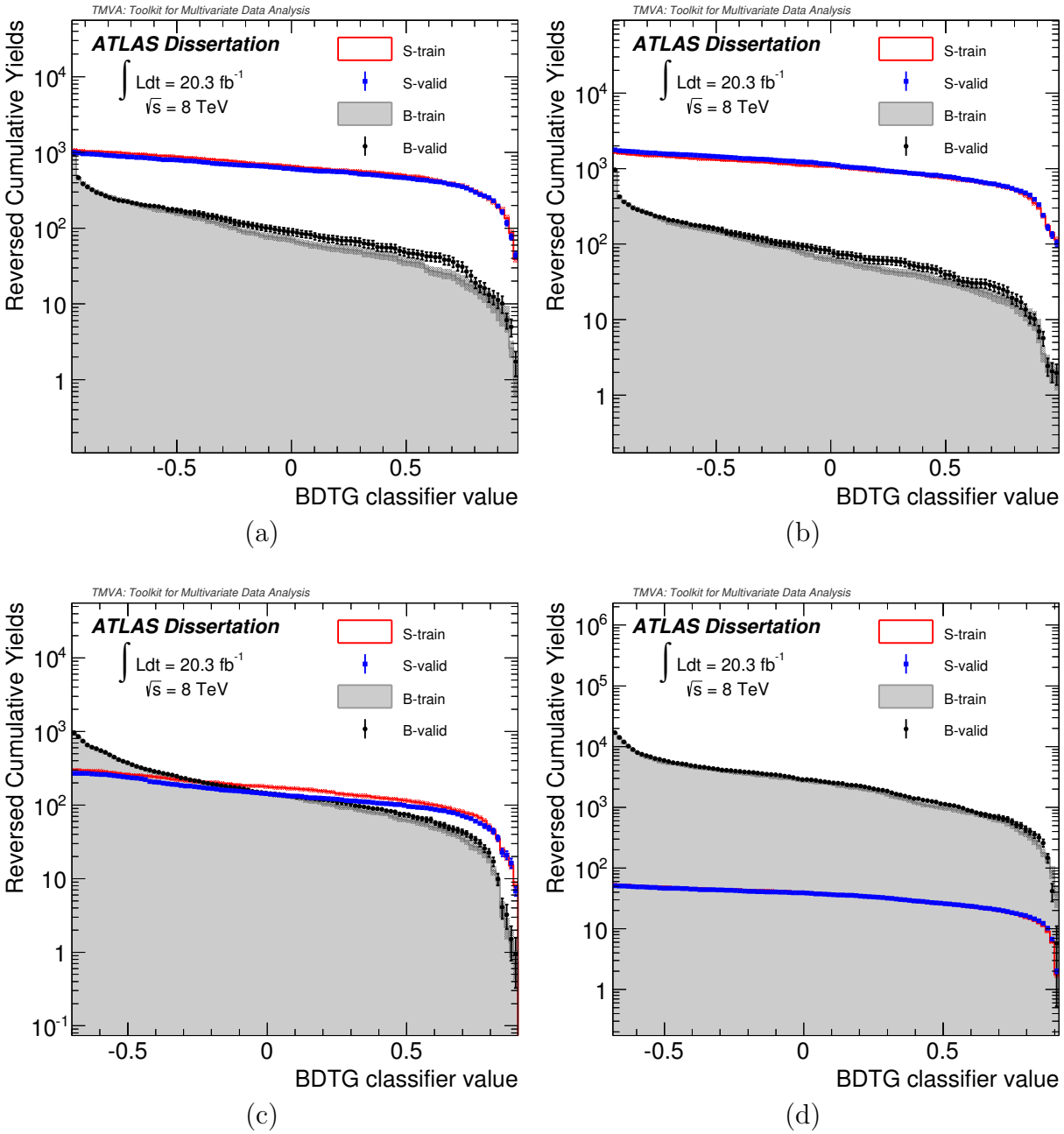


Figure C.4: BDT classifier response background and signal cumulative yield distributions (training and validation partitions, using gradient boosting (BDTG)): (a) T_{pair} signal, (b) B_{pair} signal, (c) T_{single} signal, and (d) B_{single} signal.

C.2. Stage 3 Linear Regression: All Tables

The following tables provide the linear regression results for the three signals that utilized OLS regression for the stage 3 regression analysis: T_{pair}, B_{pair} and B_{single} (linear regression was not the selected stage 3 regression method for the T_{single} signal). First are the three tables showing the covariate summary statistics, followed by the linear parameter results for the models including quadratic feature interactions (for T_{pair} and B_{pair}) and without interactions (B_{single}), and lastly the quadratic parameter results (for T_{pair} and B_{pair}).

covariates **y** and **X**	data sample	mean	standard deviation	minimum value	maximum value
$m\,(Q)$ GeV/c^2	train	617.79	356.44	0.00	1100.00
	valid	620.88	357.11	0.00	1100.00
1. $H_T(jets)$ GeV/c	train	878.75	420.32	97.69	2807.12
	valid	880.75	409.25	110.33	3927.52
2. $H_T(jets + Z)$ GeV/c	train	1237.15	512.47	278.94	3961.74
	valid	1237.75	494.60	280.80	4140.85
3. $m\,(Zb)$ GeV/c^2	train	578.11	280.45	113.68	3103.37
	valid	586.63	279.41	101.95	2370.12
4. $N(fd\ jet) > 0$ [T or F]	train	0.08	0.28	0.00	1.00
	valid	0.09	0.28	0.00	1.00
5. Classifier (350 GeV/c^2)	train	0.00	0.05	-0.29	0.04
	valid	0.00	0.05	-0.29	0.04
6. Classifier (400 GeV/c^2)	train	0.03	0.07	-0.26	0.07
	valid	0.02	0.07	-0.26	0.07
7. Classifier (450 GeV/c^2)	train	0.03	0.08	-0.31	0.09
	valid	0.03	0.08	-0.31	0.09
8. Classifier (500 GeV/c^2)	train	0.08	0.15	-0.55	0.19
	valid	0.08	0.15	-0.55	0.19
9. Classifier (550 GeV/c^2)	train	0.57	0.59	-1.00	0.98
	valid	0.56	0.60	-1.00	0.98
10. Classifier (600 GeV/c^2)	train	0.12	0.20	-0.58	0.28
	valid	0.12	0.20	-0.59	0.28
11. Classifier (650 GeV/c^2)	train	0.49	0.65	-1.00	0.97
	valid	0.49	0.65	-1.00	0.97
12. Classifier (700 GeV/c^2)	train	0.46	0.53	-0.96	0.92
	valid	0.46	0.53	-0.94	0.92
13. Classifier (750 GeV/c^2)	train	0.39	0.60	-0.99	0.87
	valid	0.40	0.60	-0.99	0.87
14. Classifier (800 GeV/c^2)	train	0.27	0.59	-0.99	0.75
	valid	0.28	0.59	-0.99	0.75
15. Classifier (850 GeV/c^2)	train	0.27	0.59	-0.99	0.78
	valid	0.28	0.59	-0.99	0.78
16. Classifier (900 GeV/c^2)	train	0.07	0.29	-0.71	0.33
	valid	0.07	0.29	-0.71	0.33
17. Classifier (950 GeV/c^2)	train	0.32	0.74	-1.00	0.96
	valid	0.33	0.73	-1.00	0.96
18. Classifier (1000 GeV/c^2)	train	0.25	0.62	-0.99	0.89
	valid	0.26	0.62	-0.99	0.90
19. Classifier (1050 GeV/c^2)	train	0.00	0.24	-0.59	0.27
	valid	0.01	0.24	-0.59	0.27
20. Classifier (1100 GeV/c^2)	train	0.08	0.49	-0.92	0.68
	valid	0.09	0.49	-0.92	0.68
event weights (**w**)	train	0.14	0.41	7.997e-10	8.072e+00
	valid	0.13	0.39	3.104e-09	5.516e+00

	N_{bkg}	$\sum_{bkg}(w_i)$	N_{sig}	$\sum_{sig}(w_i)$
train	1235	246.27	5078	642.40
valid	1203	240.54	4878	549.18

Table C.1: Covariates summary statistics for the T_{pair} signal: in-sample simulated data for training ("train") and out-of-sample simulated data for validation ("valid").

covariates **y** and **X**	data sample	mean	standard deviation	minimum value	maximum value
$m\,(Q)$ GeV/c^2	train	622.12	347.78	0.00	1100.00
	valid	617.36	345.90	0.00	1100.00
1. $H_T(jets)$ GeV/c	train	892.34	421.28	97.69	2818.54
	valid	886.17	420.71	106.87	3001.07
2. $H_T(jets + Z)$ GeV/c	train	1260.23	515.83	278.94	3281.71
	valid	1250.31	511.08	278.75	3548.52
3. $m\,(Zb)$ GeV/c^2	train	666.42	307.48	104.44	2680.28
	valid	666.29	307.28	101.95	2704.79
4. $N(fd\ jet) > 0$ [T or F]	train	0.09	0.29	0.00	1.00
	valid	0.10	0.30	0.00	1.00
5. Classifier (350 GeV/c^2)	train	0.11	0.10	-0.39	0.18
	valid	0.11	0.10	-0.37	0.18
6. Classifier (400 GeV/c^2)	train	0.61	0.49	-1.00	0.97
	valid	0.61	0.49	-0.99	0.97
7. Classifier (450 GeV/c^2)	train	0.51	0.46	-0.97	0.82
	valid	0.51	0.46	-0.97	0.83
8. Classifier (500 GeV/c^2)	train	0.63	0.58	-1.00	0.99
	valid	0.62	0.59	-1.00	0.99
9. Classifier (550 GeV/c^2)	train	0.52	0.52	-0.98	0.92
	valid	0.52	0.52	-0.98	0.91
10. Classifier (600 GeV/c^2)	train	0.13	0.29	-0.74	0.37
	valid	0.13	0.29	-0.74	0.37
11. Classifier (650 GeV/c^2)	train	0.54	0.69	-1.00	1.00
	valid	0.54	0.69	-1.00	1.00
12. Classifier (700 GeV/c^2)	train	0.53	0.72	-1.00	1.00
	valid	0.52	0.72	-1.00	1.00
13. Classifier (750 GeV/c^2)	train	0.47	0.74	-1.00	0.99
	valid	0.46	0.74	-1.00	0.99
14. Classifier (800 GeV/c^2)	train	0.36	0.61	-0.98	0.92
	valid	0.35	0.61	-0.97	0.92
15. Classifier (850 GeV/c^2)	train	0.16	0.40	-0.83	0.60
	valid	0.16	0.40	-0.81	0.60
16. Classifier (900 GeV/c^2)	train	0.18	0.44	-0.86	0.66
	valid	0.17	0.44	-0.85	0.66
17. Classifier (950 GeV/c^2)	train	0.28	0.70	-0.99	0.92
	valid	0.27	0.70	-0.99	0.92
18. Classifier (1000 GeV/c^2)	train	0.24	0.67	-0.99	0.86
	valid	0.23	0.67	-0.99	0.86
19. Classifier (1050 GeV/c^2)	train	0.17	0.59	-0.98	0.72
	valid	0.16	0.59	-0.98	0.72
20. Classifier (1100 GeV/c^2)	train	0.17	0.61	-0.98	0.83
	valid	0.16	0.61	-0.98	0.83
event weights (**w**)	train	0.17	0.49	6.073e-11	6.595e+00
	valid	0.19	0.52	2.722e-10	6.341e+00
		N_{bkg}	$\sum_{bkg}(w_i)$	N_{sig}	$\sum_{sig}(w_i)$
	train	1617	349.06	7635	1263.82
	valid	1588	343.63	7496	1348.98

Table C.2: Covariates summary statistics for the B_{pair} signal: in-sample simulated data for training ("train") and out-of-sample simulated data for validation ("valid").

covariates **y** and **X**	data sample	mean	standard deviation	minimum value	maximum value
$m\,(Q)$ GeV/c^2	train	468.74	402.88	0.00	1200.00
	valid	476.64	404.65	0.00	1200.00
1. $H_T(jets)$ GeV/c	train	455.94	198.06	97.69	1639.36
	valid	458.30	198.85	103.08	1961.07
2. $H_T(jets + Z)$ GeV/c	train	754.55	273.86	278.94	2611.86
	valid	759.44	276.09	278.75	2293.49
3. $m\,(Zb)$ GeV/c^2	train	564.67	246.63	95.00	2015.33
	valid	573.05	249.47	101.95	1829.41
4. $N(fd\ jet) > 0$ [T or F]	train	0.23	0.42	0.00	1.00
	valid	0.23	0.42	0.00	1.00
5. Classifier (400 GeV/c^2)	train	0.10	0.33	-0.89	0.61
	valid	0.11	0.33	-0.88	0.63
6. Classifier (500 GeV/c^2)	train	0.02	0.04	-0.15	0.09
	valid	0.02	0.04	-0.18	0.09
7. Classifier (550 GeV/c^2)	train	0.06	0.09	-0.36	0.20
	valid	0.06	0.09	-0.31	0.21
8. Classifier (600 GeV/c^2)	train	0.03	0.13	-0.42	0.17
	valid	0.03	0.13	-0.43	0.17
9. Classifier (650 GeV/c^2)	train	0.11	0.39	-0.86	0.52
	valid	0.12	0.39	-0.88	0.52
10. Classifier (700 GeV/c^2)	train	0.05	0.29	-0.79	0.43
	valid	0.06	0.28	-0.79	0.44
11. Classifier (750 GeV/c^2)	train	-0.02	0.10	-0.32	0.09
	valid	-0.01	0.10	-0.31	0.09
12. Classifier (800 GeV/c^2)	train	-0.05	0.31	-0.68	0.25
	valid	-0.04	0.31	-0.68	0.25
13. Classifier (850 GeV/c^2)	train	-0.04	0.44	-0.87	0.50
	valid	-0.03	0.44	-0.86	0.50
14. Classifier (900 GeV/c^2)	train	0.06	0.52	-0.93	0.73
	valid	0.08	0.52	-0.94	0.73
15. Classifier (950 GeV/c^2)	train	-0.03	0.72	-1.00	0.87
	valid	-0.01	0.72	-1.00	0.88
16. Classifier (1000 GeV/c^2)	train	-0.19	0.34	-0.78	0.24
	valid	-0.19	0.35	-0.77	0.24
17. Classifier (1200 GeV/c^2)	train	-0.15	0.79	-1.00	0.97
	valid	-0.13	0.79	-1.00	0.97
event weights (**w**)	train	0.093	0.2	6.208e-08	3.428e+00
	valid	0.092	0.2	1.006e-07	4.448e+00
		N_{bkg}	$\sum_{bkg}(w_i)$	N_{sig}	$\sum_{sig}(w_i)$
	train	1617	349.06	2605	44.95
	valid	1588	343.63	2637	45.88

Table C.3: Covariates summary statistics for the B_{single} signal: in-sample simulated data for training ("train") and out-of-sample simulated data for validation ("valid").

independent covariate $\mathbf{X}$	$\boldsymbol{\beta}_{OLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\boldsymbol{\beta}_{WLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\boldsymbol{\beta}_{GLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2)$	$\boldsymbol{\beta}_{WGLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
constant	-0.08 ±0.270 (±0.329)	-0.54 ±0.255 (±0.333)	0.33 ±0.118	0.26 ±0.114
1. $H_T(jets)$	0.41 ±0.252 (±0.333)	-0.11 ±0.249 (±0.335)	0.59 ±0.111	0.16 ±0.114
2. $H_T(jets + Z)$	0.70 ±0.301 (±0.396)	1.32 ±0.295 (±0.390)	0.28 ±0.133	0.66 ±0.134
3. $m(Zb)$	0.01 ±0.059 (±0.071)	0.01 ±0.057 (±0.075)	0.01 ±0.025	-0.01 ±0.024
4. $N(fd\ jet) > 0$	-0.04 ±0.048 (±0.052)	0.02 ±0.052 (±0.056)	0.01 ±0.021	0.02 ±0.022
5. Classifier (350)	-0.05 ±0.117 (±0.126)	-0.25 ±0.111 (±0.135)	-0.14 ±0.041	-0.25 ±0.038
6. Classifier (400)	-0.44 ±0.233 (±0.265)	-0.06 ±0.220 (±0.304)	-0.35 ±0.097	0.05 ±0.095
7. Classifier (450)	-0.25 ±0.306 (±0.381)	-0.71 ±0.307 (±0.406)	0.17 ±0.127	-0.43 ±0.125
8. Classifier (500)	-0.57 ±0.317 (±0.368)	-1.07 ±0.293 (±0.363)	-0.39 ±0.110	-0.43 ±0.097
9. Classifier (550)	0.77 ±0.263 (±0.363)	0.61 ±0.253 (±0.440)	0.25 ±0.146	0.21 ±0.147
10. Classifier (600)	0.03 ±0.163 (±0.194)	-0.30 ±0.160 (±0.213)	-0.02 ±0.064	-0.38 ±0.062
11. Classifier (650)	0.03 ±0.343 (±0.492)	0.31 ±0.333 (±0.543)	0.07 ±0.188	0.36 ±0.196
12. Classifier (700)	-1.04 ±0.411 (±0.516)	-0.85 ±0.408 (±0.569)	-0.91 ±0.182	-0.65 ±0.188
13. Classifier (750)	0.30 ±0.429 (±0.562)	0.09 ±0.424 (±0.619)	0.08 ±0.206	-0.43 ±0.215
14. Classifier (800)	0.61 ±0.371 (±0.545)	1.61 ±0.364 (±0.590)	0.60 ±0.193	1.02 ±0.211
15. Classifier (850)	0.07 ±0.437 (±0.575)	0.50 ±0.440 (±0.699)	0.06 ±0.210	0.03 ±0.228
16. Classifier (900)	-0.21 ±0.210 (±0.239)	-0.42 ±0.207 (±0.258)	-0.33 ±0.087	-0.22 ±0.087
17. Classifier (950)	0.08 ±0.356 (±0.454)	-0.45 ±0.355 (±0.545)	0.45 ±0.166	0.46 ±0.178
18. Classifier (1000)	-0.29 ±0.355 (±0.464)	-0.88 ±0.359 (±0.523)	-0.07 ±0.157	-1.12 ±0.167
19. Classifier (1050)	0.22 ±0.297 (±0.328)	0.79 ±0.295 (±0.368)	0.02 ±0.126	0.80 ±0.133
20. Classifier (1100)	1.15 ±0.286 (±0.331)	1.33 ±0.288 (±0.377)	0.86 ±0.119	1.28 ±0.128
$\hat{\sigma}$	0.59	0.53	n/a	n/a
Breusch-Pagan (p-value)	802 (< 0.01)	1165 (< 0.01)	n/a	n/a
$\bar{R}^2_{train,unwtd.}$	0.65	0.61	0.92	0.91
$\bar{R}^2_{valid,unwtd.}$	0.62	0.59	0.60	0.57
$\bar{R}^2_{train,wtd.}$	0.11	0.03	0.81	0.80
$\bar{R}^2_{valid,wtd.}$	0.04	-0.03	-0.00	-0.07
Unscaled $\boldsymbol{\beta}$ estimates:				
constant	-27 ±96.4 (±117.2)	-193 ±90.9 (±118.8)	116 ±41.9	93 ±40.7
1. $H_T(jets)$ GeV/c	0.35 ±0.214 (±0.282)	-0.09 ±0.212 (±0.284)	0.50 ±0.094	0.13 ±0.10
2. $H_T(jets + Z)$ GeV/c	0.49 ±0.210 (±0.275)	0.92 ±0.205 (±0.271)	0.19 ±0.092	0.46 ±0.09
3. $m(Zb)$ GeV/c^2	0.01 ±0.075 (±0.090)	0.01 ±0.073 (±0.096)	0.01 ±0.032	-0.01 ±0.03
4. $N(fd\ jet) > 0$ [T or F]	-48 ±61.5 (±66.2)	30 ±66.0 (±71.5)	7 ±26.7	22 ±28.6
5. Classifier (350)	-380 ±817.0 (±881.6)	-1751 ±778.8 (±943.6)	-992 ±289.5	-1751 ±268.4
6. Classifier (400)	-2333 ±1240.0 (±1408.5)	-315 ±1170.4 (±1617.5)	-1882 ±517.3	254 ±503.3
7. Classifier (450)	-1101 ±1367.9 (±1701.3)	-3182 ±1370.6 (±1815.2)	765 ±566.4	-1919 ±559.0
8. Classifier (500)	-1399 ±774.0 (±898.0)	-2617 ±714.8 (±885.8)	-961 ±269.0	-1054 ±237.2
9. Classifier (550)	465 ±158.0 (±218.3)	365 ±152.3 (±264.2)	151 ±87.5	129 ±88.6
10. Classifier (600)	62 ±294.0 (±349.0)	-545 ±287.3 (±383.2)	-35 ±114.5	-678 ±111.0
11. Classifier (650)	14 ±188.1 (±269.7)	170 ±182.7 (±297.6)	40 ±102.8	199 ±107.4
12. Classifier (700)	-697 ±276.0 (±346.5)	-569 ±274.1 (±381.5)	-607 ±122.4	-437 ±126.3
13. Classifier (750)	179 ±253.5 (±332.3)	52 ±250.6 (±365.8)	46 ±121.5	-253 ±126.8
14. Classifier (800)	371 ±223.6 (±328.5)	970 ±219.1 (±355.6)	360 ±116.5	614 ±127.0
15. Classifier (850)	41 ±263.4 (±346.5)	303 ±265.3 (±421.3)	39 ±126.8	20 ±137.4
16. Classifier (900)	-255 ±257.6 (±293.1)	-515 ±254.5 (±316.3)	-408 ±106.2	-276 ±106.5
17. Classifier (950)	39 ±171.1 (±218.4)	-216 ±171.0 (±262.3)	217 ±80.0	222 ±85.6
18. Classifier (1000)	-164 ±203.1 (±265.3)	-503 ±205.5 (±299.0)	-41 ±89.7	-641 ±95.7
19. Classifier (1050)	324 ±445.1 (±492.0)	1185 ±442.5 (±550.8)	33 ±189.0	1203 ±199.3
20. Classifier (1100)	838 ±207.8 (±240.0)	967 ±209.2 (±274.0)	624 ±86.1	928 ±92.8

Table C.4: Linear regression parameter estimates for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{XBX} + \mathbf{u}$ with alternate models of disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{model})$ for the T_{pair} signal. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]), and the Breusch-Pagan test for heteroscedasticity is provided [179] along with adjusted R^2 goodness of fit statistics.

independent covariate $\mathbf{X}$	$\boldsymbol{\beta}_{OLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\boldsymbol{\beta}_{WLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\boldsymbol{\beta}_{GLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2)$	$\boldsymbol{\beta}_{WGLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
constant	0.53 ±0.188 (±0.248)	0.00 ±0.154 (±0.190)	0.33 ±0.065	-0.05 ±0.060
1. $H_T(jets)$	0.92 ±0.151 (±0.178)	0.63 ±0.124 (±0.142)	0.20 ±0.050	0.19 ±0.044
2. $H_T(jets + Z)$	-0.18 ±0.175 (±0.214)	0.30 ±0.140 (±0.156)	0.44 ±0.056	0.63 ±0.049
3. $m\,(Zb)$	0.12 ±0.042 (±0.053)	0.04 ±0.039 (±0.053)	0.13 ±0.016	0.04 ±0.017
4. $N(fd\ jet) > 0$	0.01 ±0.031 (±0.033)	0.01 ±0.034 (±0.042)	-0.00 ±0.011	-0.04 ±0.014
5. Classifier (350)	-0.26 ±0.120 (±0.131)	-0.15 ±0.119 (±0.151)	-0.17 ±0.048	0.07 ±0.047
6. Classifier (400)	0.27 ±0.154 (±0.211)	0.79 ±0.149 (±0.231)	0.01 ±0.070	0.01 ±0.075
7. Classifier (450)	-0.77 ±0.291 (±0.413)	-1.40 ±0.273 (±0.409)	-1.21 ±0.141	-1.53 ±0.145
8. Classifier (500)	0.48 ±0.211 (±0.353)	0.30 ±0.221 (±0.345)	1.10 ±0.121	0.73 ±0.138
9. Classifier (550)	-0.65 ±0.252 (±0.332)	-0.90 ±0.249 (±0.351)	-0.34 ±0.125	-0.37 ±0.134
10. Classifier (600)	0.18 ±0.167 (±0.207)	0.07 ±0.167 (±0.236)	0.01 ±0.070	-0.09 ±0.079
11. Classifier (650)	-0.50 ±0.333 (±0.490)	0.03 ±0.316 (±0.490)	0.20 ±0.177	0.03 ±0.193
12. Classifier (700)	0.08 ±0.210 (±0.305)	-0.24 ±0.201 (±0.310)	-0.29 ±0.109	-0.13 ±0.114
13. Classifier (750)	1.04 ±0.286 (±0.438)	0.53 ±0.290 (±0.476)	0.83 ±0.153	0.74 ±0.175
14. Classifier (800)	-0.33 ±0.289 (±0.392)	-0.10 ±0.288 (±0.417)	-0.10 ±0.124	0.44 ±0.140
15. Classifier (850)	0.52 ±0.266 (±0.311)	0.74 ±0.272 (±0.357)	0.04 ±0.108	0.14 ±0.124
16. Classifier (900)	0.06 ±0.216 (±0.242)	-0.12 ±0.220 (±0.277)	0.12 ±0.086	-0.46 ±0.101
17. Classifier (950)	0.59 ±0.315 (±0.417)	0.43 ±0.313 (±0.474)	0.28 ±0.155	-0.30 ±0.168
18. Classifier (1000)	-0.37 ±0.388 (±0.506)	-0.11 ±0.384 (±0.619)	-0.01 ±0.184	0.58 ±0.201
19. Classifier (1050)	0.25 ±0.287 (±0.347)	1.09 ±0.281 (±0.396)	0.03 ±0.127	0.33 ±0.140
20. Classifier (1100)	-0.08 ±0.293 (±0.343)	-0.52 ±0.284 (±0.387)	-0.05 ±0.118	0.15 ±0.129
$\hat{\sigma}$	0.56	0.51	NA	NA
Breusch-Pagan (p-value)	1036 (< 0.01)	1850 (< 0.01)	NA	NA
$\bar{R}^2_{train,unwtd.}$	0.69	0.64	0.90	0.89
$\bar{R}^2_{valid,unwtd.}$	0.67	0.62	0.66	0.61
$\bar{R}^2_{train,wtd.}$	-0.04	-0.25	0.66	0.62
$\bar{R}^2_{valid,wtd.}$	-0.12	-0.42	-0.16	-0.44
Unscaled $\boldsymbol{\beta}$ estimates:				
constant	184 ±65.3 (±86.2)	1 ±53.4 (±66.1)	114 ±22.7	-18 ±20.8
1. $H_T(jets)$ GeV/c	0.76 ±0.125 (±0.147)	0.52 ±0.102 (±0.117)	0.17 ±0.041	0.15 ±0.04
2. $H_T(jets + Z)$ GeV/c	-0.12 ±0.118 (±0.144)	0.20 ±0.095 (±0.105)	0.30 ±0.038	0.43 ±0.03
3. $m\,(Zb)$ GeV/c^2	0.14 ±0.047 (±0.060)	0.04 ±0.044 (±0.060)	0.15 ±0.018	0.05 ±0.02
4. $N(fd\ jet) > 0$ [T or F]	18 ±37.8 (±39.7)	6 ±40.5 (±51.2)	-2 ±13.9	-54 ±17.0
5. Classifier (350)	-955 ±435.6 (±475.4)	-561 ±433.0 (±547.3)	-605 ±175.1	258 ±172.3
6. Classifier (400)	191 ±109.0 (±149.5)	558 ±105.3 (±163.6)	5 ±49.9	8 ±53.3
7. Classifier (450)	-585 ±220.4 (±312.8)	-1060 ±206.4 (±309.6)	-915 ±106.8	-1159 ±110.0
8. Classifier (500)	285 ±126.0 (±210.9)	181 ±131.9 (±205.9)	657 ±72.3	438 ±82.4
9. Classifier (550)	-432 ±168.8 (±222.2)	-604 ±166.7 (±234.9)	-230 ±83.3	-246 ±89.7
10. Classifier (600)	222 ±200.4 (±248.4)	79 ±201.0 (±283.0)	14 ±84.2	-110 ±95.3
11. Classifier (650)	-252 ±168.3 (±247.6)	17 ±159.9 (±247.5)	101 ±89.6	17 ±97.5
12. Classifier (700)	39 ±101.5 (±147.1)	-116 ±97.2 (±149.6)	-141 ±52.8	-63 ±54.8
13. Classifier (750)	490 ±134.1 (±205.4)	249 ±136.1 (±223.4)	388 ±71.7	346 ±82.1
14. Classifier (800)	-187 ±163.8 (±222.7)	-55 ±163.6 (±236.5)	-55 ±70.6	251 ±79.2
15. Classifier (850)	455 ±232.9 (±271.6)	645 ±237.7 (±312.3)	39 ±94.8	126 ±108.6
16. Classifier (900)	48 ±172.3 (±193.0)	-93 ±175.4 (±221.5)	92 ±68.8	-367 ±80.4
17. Classifier (950)	295 ±157.1 (±207.9)	216 ±156.1 (±236.4)	142 ±77.1	-149 ±83.7
18. Classifier (1000)	-191 ±203.0 (±264.3)	-59 ±200.8 (±323.6)	-4 ±96.0	304 ±105.2
19. Classifier (1050)	150 ±169.3 (±204.7)	642 ±165.6 (±233.4)	20 ±75.0	193 ±82.4
20. Classifier (1100)	-48 ±167.7 (±195.9)	-300 ±162.5 (±221.6)	-27 ±67.2	83 ±73.6

Table C.5: Linear regression parameter estimates for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \frac{1}{2}\mathbf{XBX} + \mathbf{u}$ with alternate models of disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{model})$ for the B_{pair} signal. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]), and the Breusch-Pagan test for heteroscedasticity is provided [179] along with adjusted R^2 goodness of fit statistics.

independent covariate $\mathbf{X}$	$\boldsymbol{\beta}_{OLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\boldsymbol{\beta}_{WLS}$ $\boldsymbol{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\boldsymbol{\beta}_{GLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2)$	$\boldsymbol{\beta}_{WGLS}$ $\boldsymbol{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
constant	0.84 ±0.076 (±0.106)	0.05 ±0.082 (±0.107)	0.12 ±0.043	-0.37 ±0.047
1. $H_T(jets)$	0.27 ±0.047 (±0.061)	-0.13 ±0.049 (±0.064)	0.00 ±0.026	-0.24 ±0.028
2. $H_T(jets + Z)$	-0.44 ±0.057 (±0.076)	0.18 ±0.062 (±0.084)	-0.03 ±0.033	0.36 ±0.037
3. $m(Zb)$	0.33 ±0.020 (±0.031)	0.31 ±0.019 (±0.031)	0.44 ±0.010	0.41 ±0.010
4. $N(fd\ jet) > 0$	0.17 ±0.011 (±0.011)	0.16 ±0.011 (±0.011)	0.12 ±0.004	0.11 ±0.004
5. Classifier (400)	0.01 ±0.019 (±0.020)	0.04 ±0.016 (±0.017)	-0.02 ±0.008	0.01 ±0.007
6. Classifier (500)	0.04 ±0.018 (±0.017)	0.01 ±0.015 (±0.015)	0.02 ±0.007	-0.02 ±0.007
7. Classifier (550)	0.00 ±0.037 (±0.035)	-0.02 ±0.030 (±0.027)	-0.01 ±0.014	-0.03 ±0.012
8. Classifier (600)	0.09 ±0.048 (±0.047)	0.02 ±0.040 (±0.041)	0.07 ±0.020	0.05 ±0.018
9. Classifier (650)	0.22 ±0.055 (±0.053)	0.14 ±0.045 (±0.045)	0.21 ±0.022	0.12 ±0.020
10. Classifier (700)	0.10 ±0.050 (±0.050)	0.14 ±0.041 (±0.043)	0.12 ±0.021	0.14 ±0.019
11. Classifier (750)	-0.31 ±0.067 (±0.069)	-0.29 ±0.055 (±0.057)	-0.29 ±0.028	-0.29 ±0.025
12. Classifier (800)	0.06 ±0.036 (±0.037)	0.03 ±0.030 (±0.034)	0.06 ±0.017	0.04 ±0.016
13. Classifier (850)	-0.14 ±0.080 (±0.081)	-0.21 ±0.070 (±0.076)	-0.09 ±0.035	-0.12 ±0.033
14. Classifier (900)	0.14 ±0.069 (±0.073)	0.26 ±0.061 (±0.068)	0.06 ±0.032	0.17 ±0.031
15. Classifier (950)	0.06 ±0.050 (±0.052)	0.19 ±0.047 (±0.053)	0.08 ±0.023	0.23 ±0.024
16. Classifier (1000)	0.17 ±0.033 (±0.034)	0.05 ±0.030 (±0.035)	0.08 ±0.014	-0.00 ±0.014
17. Classifier (1200)	0.15 ±0.037 (±0.041)	0.08 ±0.035 (±0.042)	0.06 ±0.017	-0.03 ±0.018
$\hat{\sigma}$	0.72	0.65	NA	NA
Breusch-Pagan (p-value)	568 (< 0.01)	400 (< 0.01)	NA	NA
$\bar{R}^2_{train,unwtd.}$	0.48	0.44	0.60	0.58
$\bar{R}^2_{valid,unwtd.}$	0.47	0.45	0.46	0.42
$\bar{R}^2_{train,wtd.}$	-1.11	-0.89	-0.66	-0.45
$\bar{R}^2_{valid,wtd.}$	-1.44	-1.19	-1.64	-1.41
Unscaled $\boldsymbol{\beta}$ estimates:				
constant	337 ±30.8 (±42.6)	19 ±33.1 (±42.9)	49 ±17.5	-149 ±18.8
1. $H_T(jets)$ GeV/c	0.55 ±0.096 (±0.123)	-0.27 ±0.100 (±0.131)	0.01 ±0.053	-0.50 ±0.06
2. $H_T(jets + Z)$ GeV/c	-0.64 ±0.084 (±0.112)	0.26 ±0.092 (±0.123)	-0.04 ±0.049	0.53 ±0.05
3. $m(Zb)$ GeV/c^2	0.54 ±0.032 (±0.051)	0.51 ±0.031 (±0.051)	0.71 ±0.016	0.67 ±0.02
4. $N(fd\ jet) > 0$ [T or F]	168 ±10.9 (±10.6)	154 ±10.4 (±11.1)	119 ±4.0	106 ±4.1
5. Classifier (400)	13 ±23.3 (±24.5)	46 ±19.3 (±19.9)	-20 ±9.5	12 ±8.2
6. Classifier (500)	354 ±174.5 (±169.6)	70 ±148.9 (±146.4)	148 ±70.4	-173 ±63.5
7. Classifier (550)	9 ±160.8 (±153.6)	-74 ±132.3 (±120.2)	-59 ±62.7	-114 ±53.9
8. Classifier (600)	266 ±145.2 (±142.2)	69 ±121.7 (±122.4)	214 ±60.5	137 ±54.6
9. Classifier (650)	233 ±57.1 (±55.2)	144 ±47.3 (±47.1)	225 ±23.3	131 ±20.7
10. Classifier (700)	136 ±70.8 (±70.4)	201 ±58.2 (±60.9)	167 ±29.8	194 ±26.5
11. Classifier (750)	-1252 ±275.0 (±282.0)	-1166 ±224.8 (±233.8)	-1168 ±115.3	-1168 ±100.5
12. Classifier (800)	79 ±47.2 (±48.5)	35 ±40.0 (±44.2)	81 ±22.0	56 ±20.5
13. Classifier (850)	-128 ±72.3 (±73.2)	-193 ±63.9 (±69.2)	-80 ±31.6	-112 ±30.4
14. Classifier (900)	107 ±53.4 (±57.0)	206 ±47.1 (±53.0)	45 ±25.0	136 ±23.8
15. Classifier (950)	33 ±28.0 (±29.0)	106 ±26.0 (±29.7)	43 ±13.1	128 ±13.3
16. Classifier (1000)	204 ±38.2 (±40.4)	63 ±35.6 (±40.5)	93 ±16.1	-1 ±16.3
17. Classifier (1200)	79 ±18.9 (±20.7)	39 ±18.0 (±21.7)	31 ±8.9	-15 ±9.2

Table C.6: Linear regression parameter estimates for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ with alternate models of disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{model})$ for the B_{single} signal. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]), and the Breusch-Pagan test for heteroscedasticity is provided [179] along with adjusted R^2 goodness of fit statistics.

covariate products $X_iX_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
1×2	-0.10 ±0.013 (±0.014)	-0.12 ±0.014 (±0.018)	-0.10 ±0.005	-0.10 ±0.006
1×3	0.13 ±0.028 (±0.031)	0.20 ±0.033 (±0.038)	0.12 ±0.011	0.17 ±0.014
1×4	-0.04 ±0.039 (±0.036)	0.04 ±0.048 (±0.053)	-0.05 ±0.014	0.02 ±0.020
1×5	0.08 ±0.087 (±0.098)	0.04 ±0.092 (±0.119)	0.10 ±0.035	-0.05 ±0.036
1×6	0.09 ±0.194 (±0.250)	0.17 ±0.200 (±0.276)	0.46 ±0.089	0.66 ±0.092
1×7	-0.06 ±0.244 (±0.304)	-0.41 ±0.263 (±0.348)	0.12 ±0.110	-0.20 ±0.120
1×8	-0.19 ±0.290 (±0.365)	0.09 ±0.285 (±0.367)	-0.17 ±0.105	-0.14 ±0.096
1×9	-0.35 ±0.227 (±0.329)	-0.64 ±0.244 (±0.419)	-0.47 ±0.126	-0.57 ±0.145
1×10	0.34 ±0.137 (±0.173)	0.03 ±0.144 (±0.191)	0.43 ±0.056	0.01 ±0.061
1×11	-0.11 ±0.281 (±0.404)	0.48 ±0.292 (±0.450)	-0.59 ±0.166	0.07 ±0.186
1×12	-0.60 ±0.395 (±0.537)	-0.21 ±0.410 (±0.606)	-0.77 ±0.185	-0.23 ±0.202
1×13	0.86 ±0.341 (±0.471)	0.81 ±0.366 (±0.563)	0.98 ±0.168	0.94 ±0.192
1×14	-0.78 ±0.293 (±0.468)	-0.32 ±0.306 (±0.505)	-0.93 ±0.164	-0.70 ±0.183
1×15	0.18 ±0.359 (±0.513)	0.45 ±0.408 (±0.706)	0.28 ±0.190	0.19 ±0.234
1×16	-0.04 ±0.168 (±0.210)	-0.14 ±0.180 (±0.231)	-0.07 ±0.074	-0.32 ±0.083
1×17	0.68 ±0.292 (±0.396)	-0.13 ±0.307 (±0.526)	0.76 ±0.148	0.64 ±0.167
1×18	-0.78 ±0.277 (±0.423)	-1.06 ±0.316 (±0.509)	-0.43 ±0.135	-1.26 ±0.157
1×19	-0.36 ±0.242 (±0.268)	-0.06 ±0.268 (±0.351)	-0.46 ±0.103	0.28 ±0.124
1×20	0.92 ±0.234 (±0.272)	0.87 ±0.259 (±0.347)	0.75 ±0.102	0.61 ±0.122
2×3	-0.10 ±0.024 (±0.028)	-0.17 ±0.029 (±0.032)	-0.10 ±0.009	-0.14 ±0.011
2×4	0.05 ±0.044 (±0.037)	-0.03 ±0.052 (±0.053)	0.04 ±0.015	-0.02 ±0.020
2×5	-0.03 ±0.102 (±0.116)	0.05 ±0.106 (±0.136)	-0.06 ±0.041	0.11 ±0.041
2×6	-0.01 ±0.221 (±0.279)	-0.06 ±0.227 (±0.315)	-0.36 ±0.102	-0.49 ±0.107
2×7	0.02 ±0.289 (±0.357)	0.52 ±0.309 (±0.397)	-0.16 ±0.129	0.30 ±0.139
2×8	0.47 ±0.346 (±0.430)	0.50 ±0.340 (±0.434)	0.32 ±0.124	0.29 ±0.113
2×9	0.32 ±0.266 (±0.399)	0.45 ±0.284 (±0.506)	0.46 ±0.151	0.41 ±0.171
2×10	-0.36 ±0.163 (±0.206)	-0.00 ±0.170 (±0.225)	-0.43 ±0.066	0.05 ±0.070
2×11	-0.00 ±0.337 (±0.488)	-0.73 ±0.346 (±0.536)	0.49 ±0.203	-0.12 ±0.224
2×12	0.90 ±0.480 (±0.652)	0.41 ±0.502 (±0.735)	1.24 ±0.222	0.46 ±0.242
2×13	-1.22 ±0.411 (±0.569)	-1.02 ±0.438 (±0.662)	-1.27 ±0.206	-1.02 ±0.230
2×14	0.60 ±0.341 (±0.559)	-0.00 ±0.353 (±0.593)	0.76 ±0.198	0.59 ±0.222
2×15	-0.12 ±0.432 (±0.616)	-0.21 ±0.491 (±0.847)	-0.18 ±0.231	0.02 ±0.282
2×16	0.11 ±0.209 (±0.260)	0.14 ±0.220 (±0.274)	0.16 ±0.091	0.35 ±0.100
2×17	-0.91 ±0.347 (±0.483)	0.01 ±0.363 (±0.638)	-0.94 ±0.182	-0.86 ±0.205
2×18	1.02 ±0.331 (±0.518)	1.16 ±0.379 (±0.614)	0.46 ±0.164	1.24 ±0.190
2×19	0.35 ±0.282 (±0.314)	-0.19 ±0.308 (±0.402)	0.44 ±0.121	-0.46 ±0.143
2×20	-1.16 ±0.280 (±0.324)	-1.04 ±0.308 (±0.411)	-0.91 ±0.121	-0.71 ±0.144
3×4	0.02 ±0.012 (±0.013)	-0.00 ±0.012 (±0.014)	0.02 ±0.004	0.00 ±0.005
3×5	-0.01 ±0.024 (±0.027)	0.02 ±0.021 (±0.024)	0.00 ±0.009	0.01 ±0.007
3×6	0.09 ±0.047 (±0.057)	0.05 ±0.039 (±0.043)	0.04 ±0.017	-0.03 ±0.012
3×7	0.10 ±0.061 (±0.075)	-0.04 ±0.052 (±0.057)	0.02 ±0.025	-0.00 ±0.022
3×8	-0.01 ±0.060 (±0.067)	-0.04 ±0.047 (±0.049)	0.01 ±0.018	-0.01 ±0.014
3×9	-0.25 ±0.053 (±0.077)	-0.14 ±0.047 (±0.073)	-0.06 ±0.028	0.09 ±0.027
3×10	0.01 ±0.036 (±0.045)	0.04 ±0.032 (±0.040)	0.03 ±0.014	0.01 ±0.012
3×11	0.05 ±0.075 (±0.106)	0.09 ±0.068 (±0.108)	-0.04 ±0.039	-0.14 ±0.040
3×12	-0.07 ±0.090 (±0.124)	0.03 ±0.084 (±0.132)	-0.12 ±0.042	-0.08 ±0.040
3×13	0.38 ±0.093 (±0.126)	0.30 ±0.090 (±0.131)	0.41 ±0.046	0.29 ±0.047
3×14	-0.15 ±0.078 (±0.107)	-0.13 ±0.076 (±0.119)	-0.18 ±0.040	-0.09 ±0.044
3×15	-0.09 ±0.095 (±0.128)	-0.20 ±0.093 (±0.147)	-0.16 ±0.049	-0.01 ±0.051
3×16	-0.01 ±0.050 (±0.060)	0.04 ±0.046 (±0.059)	-0.01 ±0.021	-0.02 ±0.020
3×17	0.16 ±0.075 (±0.097)	-0.00 ±0.076 (±0.116)	0.13 ±0.035	-0.07 ±0.039
3×18	-0.10 ±0.080 (±0.105)	-0.06 ±0.082 (±0.122)	0.02 ±0.036	0.08 ±0.038
3×19	-0.03 ±0.075 (±0.086)	0.05 ±0.078 (±0.098)	-0.03 ±0.031	0.02 ±0.032
3×20	-0.05 ±0.067 (±0.081)	-0.02 ±0.070 (±0.102)	-0.04 ±0.028	-0.05 ±0.031
4×5	-0.03 ±0.017 (±0.019)	-0.01 ±0.013 (±0.013)	-0.00 ±0.007	0.00 ±0.005
4×6	0.05 ±0.031 (±0.035)	0.01 ±0.025 (±0.029)	0.01 ±0.015	-0.00 ±0.011
4×7	-0.01 ±0.039 (±0.048)	-0.02 ±0.035 (±0.047)	0.02 ±0.021	0.00 ±0.017
4×8	-0.01 ±0.029 (±0.033)	-0.01 ±0.021 (±0.024)	-0.02 ±0.012	-0.02 ±0.008
4×9	-0.03 ±0.030 (±0.038)	-0.00 ±0.025 (±0.034)	-0.02 ±0.018	-0.02 ±0.016
4×10	-0.01 ±0.019 (±0.023)	-0.00 ±0.016 (±0.020)	0.01 ±0.009	0.02 ±0.008
4×11	0.03 ±0.041 (±0.052)	0.05 ±0.039 (±0.060)	-0.07 ±0.023	-0.03 ±0.024
4×12	0.04 ±0.046 (±0.059)	0.09 ±0.045 (±0.063)	-0.01 ±0.023	0.05 ±0.022

Table C.7—*continues on next page*

covariate products $X_i X_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
4×13	-0.05 ±0.053 (±0.074)	-0.06 ±0.054 (±0.086)	0.07 ±0.028	0.01 ±0.030
4×14	0.04 ±0.047 (±0.062)	0.02 ±0.050 (±0.076)	0.04 ±0.024	0.02 ±0.028
4×15	-0.03 ±0.054 (±0.066)	-0.15 ±0.056 (±0.081)	-0.08 ±0.027	-0.12 ±0.031
4×16	-0.01 ±0.027 (±0.036)	0.03 ±0.025 (±0.037)	0.02 ±0.013	0.02 ±0.013
4×17	0.02 ±0.038 (±0.049)	0.08 ±0.036 (±0.057)	0.08 ±0.017	0.09 ±0.019
4×18	-0.08 ±0.041 (±0.049)	-0.08 ±0.040 (±0.052)	-0.06 ±0.018	-0.04 ±0.019
4×19	0.03 ±0.046 (±0.052)	0.07 ±0.048 (±0.062)	0.02 ±0.021	0.04 ±0.022
4×20	0.00 ±0.040 (±0.046)	-0.01 ±0.044 (±0.054)	-0.03 ±0.017	-0.03 ±0.020
5×6	-0.02 ±0.042 (±0.042)	0.05 ±0.027 (±0.023)	-0.02 ±0.009	-0.01 ±0.006
5×7	0.03 ±0.052 (±0.051)	-0.03 ±0.035 (±0.031)	0.02 ±0.013	0.00 ±0.008
5×8	-0.01 ±0.047 (±0.043)	-0.02 ±0.029 (±0.024)	0.02 ±0.009	-0.01 ±0.006
5×9	-0.02 ±0.063 (±0.077)	-0.16 ±0.046 (±0.054)	-0.07 ±0.027	-0.10 ±0.020
5×10	-0.03 ±0.036 (±0.036)	0.00 ±0.023 (±0.021)	-0.00 ±0.009	-0.01 ±0.005
5×11	0.10 ±0.088 (±0.112)	0.18 ±0.078 (±0.104)	-0.01 ±0.041	-0.00 ±0.038
5×12	0.00 ±0.102 (±0.113)	-0.02 ±0.083 (±0.096)	0.11 ±0.037	0.14 ±0.029
5×13	-0.00 ±0.112 (±0.143)	-0.08 ±0.097 (±0.125)	-0.07 ±0.049	-0.15 ±0.042
5×14	-0.18 ±0.103 (±0.133)	0.11 ±0.093 (±0.128)	-0.09 ±0.050	0.06 ±0.046
5×15	-0.04 ±0.114 (±0.149)	-0.01 ±0.098 (±0.125)	-0.02 ±0.049	0.06 ±0.040
5×16	0.02 ±0.055 (±0.058)	0.01 ±0.040 (±0.043)	0.01 ±0.017	0.00 ±0.011
5×17	-0.05 ±0.098 (±0.126)	-0.26 ±0.091 (±0.125)	0.17 ±0.042	0.06 ±0.041
5×18	0.07 ±0.092 (±0.108)	-0.01 ±0.081 (±0.098)	-0.05 ±0.036	-0.10 ±0.031
5×19	0.13 ±0.108 (±0.113)	0.01 ±0.081 (±0.085)	-0.05 ±0.030	-0.06 ±0.022
5×20	-0.06 ±0.099 (±0.108)	0.09 ±0.089 (±0.101)	-0.01 ±0.034	0.03 ±0.029
6×7	0.00 ±0.074 (±0.084)	-0.04 ±0.049 (±0.049)	0.00 ±0.024	-0.01 ±0.015
6×8	0.01 ±0.084 (±0.081)	-0.02 ±0.055 (±0.050)	-0.06 ±0.021	-0.02 ±0.015
6×9	-0.19 ±0.089 (±0.115)	-0.18 ±0.071 (±0.089)	-0.03 ±0.043	-0.08 ±0.033
6×10	-0.01 ±0.067 (±0.073)	-0.00 ±0.046 (±0.045)	-0.06 ±0.019	0.01 ±0.013
6×11	-0.10 ±0.137 (±0.200)	-0.20 ±0.114 (±0.170)	-0.31 ±0.073	-0.22 ±0.067
6×12	0.29 ±0.185 (±0.221)	0.10 ±0.156 (±0.201)	0.51 ±0.080	0.04 ±0.065
6×13	0.15 ±0.192 (±0.252)	0.17 ±0.176 (±0.245)	0.11 ±0.099	0.26 ±0.089
6×14	-0.32 ±0.186 (±0.256)	0.33 ±0.173 (±0.250)	-0.27 ±0.098	0.24 ±0.097
6×15	0.33 ±0.217 (±0.298)	0.10 ±0.203 (±0.284)	0.10 ±0.111	-0.01 ±0.102
6×16	-0.06 ±0.102 (±0.122)	0.03 ±0.078 (±0.095)	-0.07 ±0.039	-0.01 ±0.028
6×17	-0.22 ±0.182 (±0.240)	-0.34 ±0.170 (±0.239)	-0.02 ±0.086	-0.35 ±0.085
6×18	0.24 ±0.169 (±0.217)	0.12 ±0.148 (±0.185)	0.30 ±0.078	0.27 ±0.067
6×19	-0.11 ±0.206 (±0.238)	-0.05 ±0.169 (±0.195)	-0.06 ±0.073	-0.24 ±0.054
6×20	-0.14 ±0.190 (±0.224)	-0.07 ±0.172 (±0.223)	-0.23 ±0.080	0.15 ±0.070
7×8	0.02 ±0.099 (±0.104)	0.05 ±0.072 (±0.075)	0.05 ±0.028	0.02 ±0.021
7×9	0.19 ±0.119 (±0.156)	0.34 ±0.095 (±0.120)	0.06 ±0.061	0.20 ±0.048
7×10	0.00 ±0.080 (±0.093)	-0.04 ±0.061 (±0.067)	0.02 ±0.029	-0.01 ±0.020
7×11	-0.25 ±0.165 (±0.234)	-0.33 ±0.143 (±0.223)	-0.10 ±0.089	-0.18 ±0.082
7×12	-0.08 ±0.231 (±0.298)	0.14 ±0.210 (±0.283)	-0.31 ±0.109	-0.05 ±0.095
7×13	-0.03 ±0.246 (±0.349)	-0.25 ±0.242 (±0.360)	-0.22 ±0.128	-0.44 ±0.126
7×14	-0.00 ±0.223 (±0.329)	-0.63 ±0.216 (±0.342)	0.07 ±0.126	-0.67 ±0.131
7×15	-0.17 ±0.293 (±0.418)	0.56 ±0.282 (±0.410)	0.06 ±0.150	0.56 ±0.145
7×16	0.04 ±0.135 (±0.174)	-0.01 ±0.108 (±0.134)	0.01 ±0.055	-0.02 ±0.042
7×17	0.27 ±0.225 (±0.305)	0.44 ±0.228 (±0.332)	0.12 ±0.112	0.31 ±0.119
7×18	0.06 ±0.230 (±0.306)	-0.28 ±0.218 (±0.295)	0.07 ±0.107	-0.12 ±0.101
7×19	0.01 ±0.264 (±0.323)	-0.07 ±0.234 (±0.282)	0.12 ±0.097	0.21 ±0.076
7×20	-0.11 ±0.242 (±0.304)	-0.08 ±0.236 (±0.318)	0.03 ±0.104	-0.02 ±0.099
8×9	-0.04 ±0.102 (±0.122)	-0.04 ±0.081 (±0.099)	0.05 ±0.049	0.08 ±0.038
8×10	-0.06 ±0.061 (±0.065)	-0.05 ±0.044 (±0.043)	0.01 ±0.017	-0.02 ±0.012
8×11	0.26 ±0.165 (±0.215)	0.25 ±0.137 (±0.188)	0.17 ±0.083	0.33 ±0.074
8×12	0.04 ±0.186 (±0.223)	-0.09 ±0.157 (±0.179)	-0.15 ±0.071	-0.12 ±0.054
8×13	-0.16 ±0.223 (±0.297)	0.07 ±0.204 (±0.267)	0.02 ±0.102	-0.05 ±0.089
8×14	0.06 ±0.226 (±0.298)	0.08 ±0.211 (±0.296)	0.22 ±0.111	0.30 ±0.106
8×15	0.00 ±0.275 (±0.359)	-0.29 ±0.240 (±0.331)	-0.04 ±0.121	-0.58 ±0.099
8×16	0.04 ±0.112 (±0.134)	-0.15 ±0.087 (±0.109)	0.02 ±0.035	0.00 ±0.027
8×17	-0.14 ±0.218 (±0.278)	-0.77 ±0.202 (±0.265)	-0.23 ±0.097	-0.27 ±0.089
8×18	-0.40 ±0.213 (±0.272)	-0.04 ±0.189 (±0.244)	-0.27 ±0.085	-0.03 ±0.071
8×19	0.16 ±0.205 (±0.223)	0.33 ±0.157 (±0.171)	-0.08 ±0.059	0.08 ±0.042
8×20	-0.08 ±0.226 (±0.248)	0.09 ±0.188 (±0.214)	0.08 ±0.074	0.06 ±0.059

Table C.7—*continues on next page*

covariate products $X_i X_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
9×10	-0.11 ±0.063 (±0.079)	-0.13 ±0.053 (±0.070)	-0.10 ±0.033	-0.19 ±0.029
9×11	0.03 ±0.104 (±0.148)	0.17 ±0.090 (±0.131)	0.24 ±0.056	0.19 ±0.053
9×12	-0.44 ±0.162 (±0.200)	-0.29 ±0.127 (±0.166)	-0.61 ±0.087	-0.39 ±0.070
9×13	0.30 ±0.176 (±0.235)	0.35 ±0.153 (±0.222)	0.37 ±0.102	0.37 ±0.093
9×14	0.34 ±0.153 (±0.209)	-0.07 ±0.135 (±0.203)	0.00 ±0.087	-0.07 ±0.082
9×15	0.07 ±0.193 (±0.265)	0.07 ±0.176 (±0.269)	0.35 ±0.115	0.22 ±0.105
9×16	-0.21 ±0.102 (±0.132)	-0.33 ±0.088 (±0.124)	-0.08 ±0.054	-0.18 ±0.048
9×17	-0.11 ±0.160 (±0.207)	0.21 ±0.153 (±0.223)	-0.44 ±0.089	-0.25 ±0.094
9×18	-0.09 ±0.168 (±0.220)	0.19 ±0.156 (±0.229)	-0.08 ±0.093	0.14 ±0.092
9×19	0.26 ±0.203 (±0.249)	0.17 ±0.171 (±0.234)	0.28 ±0.096	0.31 ±0.085
9×20	0.15 ±0.196 (±0.248)	-0.12 ±0.180 (±0.271)	0.09 ±0.100	-0.21 ±0.098
10×11	0.10 ±0.086 (±0.106)	0.16 ±0.080 (±0.121)	0.05 ±0.044	0.05 ±0.043
10×12	0.08 ±0.108 (±0.129)	0.13 ±0.099 (±0.131)	0.09 ±0.048	0.28 ±0.044
10×13	-0.11 ±0.113 (±0.137)	-0.25 ±0.108 (±0.160)	-0.03 ±0.053	-0.07 ±0.054
10×14	0.13 ±0.103 (±0.134)	0.09 ±0.103 (±0.158)	-0.08 ±0.047	-0.02 ±0.052
10×15	0.03 ±0.130 (±0.162)	0.21 ±0.127 (±0.185)	0.10 ±0.057	0.33 ±0.060
10×16	0.00 ±0.071 (±0.087)	0.09 ±0.061 (±0.080)	0.07 ±0.029	0.02 ±0.023
10×17	-0.01 ±0.104 (±0.128)	-0.06 ±0.104 (±0.143)	-0.02 ±0.044	-0.09 ±0.047
10×18	-0.06 ±0.108 (±0.130)	-0.09 ±0.107 (±0.151)	-0.08 ±0.045	-0.26 ±0.047
10×19	0.16 ±0.128 (±0.151)	0.06 ±0.115 (±0.146)	0.02 ±0.048	-0.04 ±0.039
10×20	-0.21 ±0.115 (±0.129)	-0.29 ±0.116 (±0.150)	-0.07 ±0.046	-0.19 ±0.046
11×12	0.25 ±0.199 (±0.267)	0.25 ±0.179 (±0.278)	0.30 ±0.102	0.14 ±0.098
11×13	0.04 ±0.204 (±0.257)	0.26 ±0.187 (±0.281)	-0.14 ±0.111	0.09 ±0.111
11×14	-0.21 ±0.191 (±0.263)	-0.05 ±0.173 (±0.256)	-0.12 ±0.104	0.04 ±0.102
11×15	-0.78 ±0.219 (±0.337)	-0.80 ±0.209 (±0.338)	-0.85 ±0.123	-0.90 ±0.124
11×16	0.27 ±0.132 (±0.165)	0.24 ±0.119 (±0.169)	0.19 ±0.067	0.39 ±0.064
11×17	0.46 ±0.191 (±0.254)	0.35 ±0.187 (±0.276)	0.44 ±0.100	0.39 ±0.108
11×18	-0.16 ±0.200 (±0.279)	0.07 ±0.183 (±0.273)	0.10 ±0.105	0.11 ±0.103
11×19	-0.32 ±0.259 (±0.346)	-0.71 ±0.242 (±0.350)	-0.21 ±0.124	-0.56 ±0.124
11×20	0.29 ±0.217 (±0.282)	0.30 ±0.209 (±0.309)	0.13 ±0.113	0.19 ±0.117
12×13	0.19 ±0.232 (±0.290)	-0.17 ±0.225 (±0.329)	0.16 ±0.111	0.29 ±0.115
12×14	-0.14 ±0.247 (±0.332)	-0.38 ±0.233 (±0.350)	0.07 ±0.123	-0.17 ±0.126
12×15	0.68 ±0.296 (±0.377)	0.60 ±0.278 (±0.412)	0.37 ±0.141	0.30 ±0.141
12×16	-0.12 ±0.158 (±0.195)	0.39 ±0.147 (±0.190)	0.03 ±0.070	0.04 ±0.065
12×17	-0.29 ±0.234 (±0.290)	0.27 ±0.224 (±0.318)	-0.55 ±0.107	0.01 ±0.109
12×18	0.04 ±0.234 (±0.297)	0.21 ±0.221 (±0.321)	0.05 ±0.106	0.45 ±0.106
12×19	-0.57 ±0.297 (±0.358)	-0.86 ±0.276 (±0.351)	-0.07 ±0.125	-0.64 ±0.113
12×20	-0.26 ±0.234 (±0.288)	-0.39 ±0.226 (±0.307)	-0.44 ±0.099	-0.42 ±0.097
13×14	-0.16 ±0.211 (±0.298)	0.03 ±0.205 (±0.324)	-0.03 ±0.114	-0.42 ±0.119
13×15	-0.48 ±0.236 (±0.317)	-0.34 ±0.241 (±0.347)	-0.53 ±0.118	-0.02 ±0.122
13×16	-0.13 ±0.148 (±0.185)	-0.08 ±0.143 (±0.193)	-0.09 ±0.074	-0.03 ±0.074
13×17	-0.03 ±0.231 (±0.308)	-0.22 ±0.234 (±0.363)	0.17 ±0.109	-0.14 ±0.118
13×18	0.09 ±0.240 (±0.315)	-0.14 ±0.240 (±0.356)	-0.05 ±0.113	0.13 ±0.122
13×19	0.56 ±0.319 (±0.383)	0.79 ±0.317 (±0.444)	0.27 ±0.144	0.25 ±0.155
13×20	0.10 ±0.260 (±0.330)	-0.13 ±0.267 (±0.378)	0.33 ±0.120	-0.16 ±0.134
14×15	0.63 ±0.248 (±0.374)	-0.17 ±0.256 (±0.430)	0.63 ±0.135	0.03 ±0.151
14×16	-0.06 ±0.152 (±0.195)	0.09 ±0.148 (±0.218)	-0.21 ±0.076	0.05 ±0.080
14×17	0.26 ±0.218 (±0.303)	0.52 ±0.224 (±0.350)	0.35 ±0.105	0.76 ±0.117
14×18	0.04 ±0.212 (±0.292)	0.32 ±0.212 (±0.328)	-0.31 ±0.109	-0.33 ±0.121
14×19	-0.35 ±0.322 (±0.434)	0.07 ±0.324 (±0.470)	0.05 ±0.156	0.69 ±0.170
14×20	0.17 ±0.244 (±0.327)	0.31 ±0.250 (±0.398)	-0.14 ±0.118	-0.19 ±0.134
15×16	0.12 ±0.169 (±0.216)	-0.04 ±0.168 (±0.257)	-0.11 ±0.081	-0.33 ±0.083
15×17	0.02 ±0.223 (±0.314)	-0.10 ±0.234 (±0.377)	0.26 ±0.104	-0.16 ±0.118
15×18	0.01 ±0.236 (±0.324)	0.23 ±0.235 (±0.379)	0.09 ±0.114	0.32 ±0.123
15×19	0.12 ±0.266 (±0.343)	0.76 ±0.266 (±0.424)	-0.29 ±0.126	0.39 ±0.132
15×20	-0.46 ±0.263 (±0.346)	-0.32 ±0.273 (±0.425)	-0.07 ±0.124	-0.21 ±0.136
16×17	-0.00 ±0.135 (±0.169)	-0.13 ±0.131 (±0.193)	0.18 ±0.057	0.07 ±0.058
16×18	0.28 ±0.150 (±0.185)	0.26 ±0.144 (±0.203)	0.27 ±0.065	0.13 ±0.067
16×19	-0.33 ±0.168 (±0.192)	-0.58 ±0.150 (±0.190)	-0.11 ±0.064	-0.22 ±0.054
16×20	0.10 ±0.146 (±0.172)	0.18 ±0.147 (±0.195)	-0.14 ±0.058	0.12 ±0.063
17×18	-0.02 ±0.176 (±0.230)	-0.16 ±0.167 (±0.267)	0.00 ±0.079	-0.13 ±0.082
17×19	-0.15 ±0.243 (±0.285)	-0.15 ±0.249 (±0.334)	-0.34 ±0.100	-0.58 ±0.111

Table C.7—*continues on next page*

covariate products $X_iX_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
17×20	0.08 ±0.208 (±0.247)	0.02 ±0.216 (±0.315)	0.03 ±0.090	0.50 ±0.100
18×19	-0.09 ±0.245 (±0.292)	-0.47 ±0.242 (±0.345)	-0.06 ±0.104	-0.37 ±0.109
18×20	-0.23 ±0.190 (±0.240)	-0.49 ±0.203 (±0.298)	-0.08 ±0.081	-0.52 ±0.091
19×20	0.74 ±0.201 (±0.235)	1.02 ±0.208 (±0.284)	0.65 ±0.082	1.05 ±0.090

Table C.7: Linear regression covariate products' parameter estimates for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + X_i\mathbf{B}_{i,j>i}X_j + \mathbf{u}$ with alternate models of disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \mathbf{\Sigma}_{model})$ for the T_{pair} signal. Covariates $\mathbf{X}_i$ have been scaled by $1/\hat{s}_{X_i}$ ($\mathbf{y}$ by $1/\hat{s_y}$) before the products X_iX_j are constructed. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]).

covariate products $X_iX_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
1×2	-0.08 ±0.011 (±0.011)	-0.13 ±0.012 (±0.015)	-0.08 ±0.004	-0.14 ±0.006
1×3	0.07 ±0.023 (±0.027)	0.13 ±0.026 (±0.033)	0.06 ±0.009	0.14 ±0.012
1×4	-0.02 ±0.024 (±0.023)	0.02 ±0.029 (±0.033)	0.02 ±0.008	0.02 ±0.011
1×5	0.17 ±0.100 (±0.122)	0.00 ±0.103 (±0.131)	0.08 ±0.044	0.05 ±0.044
1×6	0.13 ±0.115 (±0.161)	0.16 ±0.118 (±0.173)	0.06 ±0.057	0.14 ±0.061
1×7	-0.38 ±0.235 (±0.349)	-0.05 ±0.236 (±0.358)	-0.15 ±0.123	-0.16 ±0.129
1×8	-0.17 ±0.165 (±0.267)	-0.28 ±0.176 (±0.263)	-0.18 ±0.103	-0.23 ±0.115
1×9	-0.78 ±0.199 (±0.257)	-0.67 ±0.202 (±0.270)	-0.30 ±0.105	-0.30 ±0.113
1×10	-0.25 ±0.132 (±0.175)	-0.46 ±0.142 (±0.205)	-0.35 ±0.062	-0.44 ±0.071
1×11	0.46 ±0.257 (±0.366)	0.29 ±0.266 (±0.411)	0.97 ±0.146	0.98 ±0.166
1×12	-0.22 ±0.163 (±0.230)	0.14 ±0.167 (±0.260)	0.12 ±0.088	0.41 ±0.095
1×13	0.17 ±0.235 (±0.370)	0.30 ±0.255 (±0.443)	-0.63 ±0.138	-0.99 ±0.162
1×14	-0.10 ±0.227 (±0.295)	-0.52 ±0.249 (±0.347)	-0.23 ±0.108	-0.54 ±0.130
1×15	-0.12 ±0.205 (±0.256)	0.07 ±0.226 (±0.291)	-0.05 ±0.094	0.08 ±0.115
1×16	0.24 ±0.171 (±0.207)	-0.37 ±0.191 (±0.249)	0.43 ±0.078	-0.23 ±0.096
1×17	0.18 ±0.220 (±0.302)	-0.23 ±0.242 (±0.367)	0.14 ±0.115	0.04 ±0.142
1×18	0.24 ±0.268 (±0.382)	-0.41 ±0.297 (±0.504)	0.15 ±0.137	-0.20 ±0.170
1×19	0.23 ±0.196 (±0.267)	1.24 ±0.215 (±0.338)	0.11 ±0.097	0.86 ±0.118
1×20	-0.14 ±0.215 (±0.271)	0.70 ±0.237 (±0.356)	-0.17 ±0.101	0.70 ±0.123
2×3	-0.09 ±0.021 (±0.022)	-0.14 ±0.024 (±0.028)	-0.08 ±0.007	-0.13 ±0.010
2×4	-0.01 ±0.028 (±0.026)	-0.04 ±0.034 (±0.040)	-0.03 ±0.010	-0.01 ±0.014
2×5	-0.14 ±0.119 (±0.147)	-0.00 ±0.123 (±0.152)	-0.03 ±0.052	-0.06 ±0.050
2×6	-0.30 ±0.140 (±0.197)	-0.54 ±0.145 (±0.209)	-0.14 ±0.068	-0.20 ±0.072
2×7	0.47 ±0.284 (±0.435)	0.48 ±0.288 (±0.436)	0.38 ±0.148	0.59 ±0.154
2×8	0.18 ±0.191 (±0.332)	0.33 ±0.209 (±0.324)	-0.07 ±0.120	0.17 ±0.135
2×9	1.21 ±0.245 (±0.323)	1.18 ±0.249 (±0.339)	0.70 ±0.129	0.59 ±0.138
2×10	0.09 ±0.162 (±0.221)	0.41 ±0.173 (±0.249)	0.25 ±0.077	0.47 ±0.087
2×11	-0.32 ±0.310 (±0.465)	-0.25 ±0.322 (±0.505)	-0.84 ±0.178	-0.75 ±0.203
2×12	0.25 ±0.190 (±0.280)	-0.03 ±0.192 (±0.298)	-0.06 ±0.102	-0.43 ±0.109
2×13	-0.56 ±0.288 (±0.471)	-0.44 ±0.316 (±0.560)	0.30 ±0.171	0.63 ±0.203
2×14	0.19 ±0.275 (±0.369)	0.45 ±0.298 (±0.423)	0.29 ±0.129	0.34 ±0.153
2×15	-0.07 ±0.249 (±0.326)	-0.45 ±0.272 (±0.351)	0.03 ±0.116	-0.22 ±0.138
2×16	-0.29 ±0.209 (±0.259)	0.20 ±0.233 (±0.309)	-0.55 ±0.095	0.09 ±0.117
2×17	-0.32 ±0.271 (±0.381)	-0.04 ±0.296 (±0.456)	-0.14 ±0.144	-0.04 ±0.176
2×18	-0.12 ±0.331 (±0.496)	0.31 ±0.365 (±0.613)	-0.23 ±0.173	-0.22 ±0.212
2×19	-0.09 ±0.234 (±0.327)	-1.16 ±0.255 (±0.399)	-0.10 ±0.119	-0.74 ±0.145
2×20	0.15 ±0.261 (±0.339)	-0.29 ±0.286 (±0.410)	0.30 ±0.123	-0.19 ±0.148

Table C.8—*continues on next page*

covariate products $X_i X_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
3×4	0.03 ±0.010 (±0.011)	0.01 ±0.011 (±0.015)	0.02 ±0.004	0.00 ±0.005
3×5	0.07 ±0.029 (±0.040)	0.03 ±0.024 (±0.031)	0.01 ±0.012	-0.00 ±0.009
3×6	0.07 ±0.035 (±0.054)	0.10 ±0.033 (±0.053)	0.06 ±0.018	0.07 ±0.017
3×7	0.01 ±0.066 (±0.102)	0.07 ±0.055 (±0.077)	0.13 ±0.034	0.10 ±0.029
3×8	-0.07 ±0.046 (±0.077)	-0.10 ±0.042 (±0.070)	-0.07 ±0.028	-0.17 ±0.026
3×9	-0.18 ±0.060 (±0.097)	-0.12 ±0.053 (±0.092)	-0.22 ±0.032	-0.09 ±0.029
3×10	0.06 ±0.042 (±0.061)	-0.09 ±0.038 (±0.054)	0.05 ±0.020	-0.07 ±0.017
3×11	0.14 ±0.075 (±0.118)	-0.02 ±0.070 (±0.107)	0.03 ±0.043	0.01 ±0.043
3×12	-0.08 ±0.051 (±0.083)	0.02 ±0.047 (±0.093)	0.13 ±0.030	0.17 ±0.029
3×13	0.11 ±0.072 (±0.121)	0.22 ±0.069 (±0.127)	0.08 ±0.038	0.13 ±0.040
3×14	0.03 ±0.072 (±0.136)	-0.05 ±0.072 (±0.174)	-0.02 ±0.037	-0.20 ±0.040
3×15	0.07 ±0.072 (±0.094)	0.09 ±0.072 (±0.100)	0.04 ±0.033	0.05 ±0.035
3×16	0.09 ±0.059 (±0.094)	0.22 ±0.059 (±0.122)	0.11 ±0.029	0.33 ±0.031
3×17	0.11 ±0.075 (±0.111)	0.07 ±0.076 (±0.140)	-0.00 ±0.039	0.15 ±0.044
3×18	-0.24 ±0.094 (±0.181)	-0.20 ±0.088 (±0.250)	-0.22 ±0.050	-0.09 ±0.052
3×19	-0.17 ±0.074 (±0.108)	-0.13 ±0.075 (±0.138)	0.04 ±0.038	-0.03 ±0.043
3×20	-0.03 ±0.073 (±0.123)	-0.11 ±0.073 (±0.165)	-0.11 ±0.036	-0.34 ±0.039
4×5	-0.03 ±0.016 (±0.019)	-0.02 ±0.012 (±0.012)	-0.01 ±0.007	-0.00 ±0.005
4×6	-0.03 ±0.020 (±0.023)	-0.04 ±0.015 (±0.017)	-0.03 ±0.010	-0.03 ±0.008
4×7	0.10 ±0.034 (±0.043)	0.07 ±0.028 (±0.035)	0.06 ±0.017	0.01 ±0.014
4×8	-0.09 ±0.029 (±0.040)	-0.05 ±0.027 (±0.036)	-0.07 ±0.017	-0.02 ±0.017
4×9	0.04 ±0.030 (±0.036)	0.06 ±0.026 (±0.035)	0.04 ±0.018	0.08 ±0.016
4×10	-0.04 ±0.021 (±0.026)	-0.05 ±0.017 (±0.018)	-0.02 ±0.010	-0.03 ±0.008
4×11	0.12 ±0.035 (±0.047)	0.16 ±0.033 (±0.046)	0.11 ±0.021	0.13 ±0.022
4×12	0.00 ±0.023 (±0.031)	0.00 ±0.024 (±0.038)	-0.00 ±0.014	0.00 ±0.016
4×13	-0.12 ±0.032 (±0.043)	-0.21 ±0.033 (±0.045)	-0.07 ±0.016	-0.15 ±0.019
4×14	0.03 ±0.037 (±0.044)	0.09 ±0.036 (±0.048)	-0.01 ±0.019	0.05 ±0.020
4×15	-0.02 ±0.039 (±0.042)	-0.04 ±0.037 (±0.044)	-0.02 ±0.017	-0.04 ±0.017
4×16	0.03 ±0.036 (±0.039)	0.09 ±0.038 (±0.048)	0.00 ±0.016	0.05 ±0.019
4×17	0.06 ±0.040 (±0.047)	0.10 ±0.041 (±0.059)	-0.00 ±0.018	0.02 ±0.022
4×18	0.08 ±0.054 (±0.061)	0.19 ±0.057 (±0.085)	0.06 ±0.026	0.18 ±0.031
4×19	-0.12 ±0.038 (±0.047)	-0.25 ±0.040 (±0.061)	-0.08 ±0.018	-0.17 ±0.022
4×20	-0.02 ±0.044 (±0.050)	-0.11 ±0.043 (±0.057)	0.03 ±0.020	-0.09 ±0.023
5×6	0.08 ±0.039 (±0.045)	0.10 ±0.027 (±0.027)	0.03 ±0.017	0.02 ±0.010
5×7	0.03 ±0.056 (±0.069)	-0.00 ±0.039 (±0.042)	0.07 ±0.026	0.05 ±0.018
5×8	-0.17 ±0.058 (±0.074)	-0.13 ±0.043 (±0.050)	-0.14 ±0.033	-0.07 ±0.025
5×9	0.10 ±0.075 (±0.095)	0.07 ±0.061 (±0.076)	0.01 ±0.036	-0.03 ±0.025
5×10	0.03 ±0.049 (±0.058)	0.00 ±0.037 (±0.041)	-0.01 ±0.019	0.00 ±0.012
5×11	-0.10 ±0.087 (±0.114)	-0.02 ±0.080 (±0.119)	-0.12 ±0.050	-0.18 ±0.049
5×12	0.07 ±0.059 (±0.080)	0.01 ±0.058 (±0.100)	0.19 ±0.035	0.16 ±0.037
5×13	0.16 ±0.083 (±0.105)	0.25 ±0.079 (±0.130)	0.02 ±0.042	0.06 ±0.045
5×14	-0.01 ±0.111 (±0.141)	-0.12 ±0.104 (±0.146)	0.01 ±0.053	-0.12 ±0.049
5×15	-0.06 ±0.112 (±0.133)	-0.03 ±0.087 (±0.106)	0.02 ±0.045	-0.02 ±0.032
5×16	-0.09 ±0.094 (±0.108)	-0.20 ±0.079 (±0.092)	-0.08 ±0.039	-0.01 ±0.030
5×17	0.22 ±0.118 (±0.142)	0.16 ±0.120 (±0.181)	0.12 ±0.057	0.13 ±0.064
5×18	-0.39 ±0.161 (±0.196)	-0.38 ±0.163 (±0.242)	-0.06 ±0.077	-0.15 ±0.086
5×19	-0.18 ±0.115 (±0.142)	-0.11 ±0.115 (±0.174)	-0.16 ±0.054	0.04 ±0.058
5×20	0.22 ±0.133 (±0.160)	0.37 ±0.134 (±0.193)	0.04 ±0.060	0.15 ±0.063
6×7	0.01 ±0.066 (±0.080)	-0.06 ±0.047 (±0.050)	0.04 ±0.036	-0.06 ±0.024
6×8	-0.05 ±0.059 (±0.076)	0.03 ±0.047 (±0.058)	-0.10 ±0.036	-0.02 ±0.030
6×9	-0.15 ±0.065 (±0.083)	-0.04 ±0.052 (±0.064)	-0.02 ±0.039	0.06 ±0.032
6×10	0.09 ±0.053 (±0.066)	0.10 ±0.041 (±0.051)	0.03 ±0.028	0.05 ±0.021
6×11	0.15 ±0.098 (±0.128)	-0.13 ±0.088 (±0.135)	0.08 ±0.058	0.06 ±0.056
6×12	-0.08 ±0.066 (±0.084)	0.07 ±0.061 (±0.090)	0.01 ±0.041	-0.05 ±0.041
6×13	-0.10 ±0.099 (±0.137)	-0.11 ±0.090 (±0.144)	-0.11 ±0.056	-0.21 ±0.057
6×14	-0.07 ±0.117 (±0.148)	-0.19 ±0.102 (±0.137)	-0.06 ±0.065	-0.02 ±0.060
6×15	-0.10 ±0.117 (±0.142)	0.08 ±0.094 (±0.110)	-0.04 ±0.062	0.17 ±0.050
6×16	-0.04 ±0.107 (±0.135)	-0.29 ±0.087 (±0.114)	-0.03 ±0.058	-0.22 ±0.048
6×17	0.09 ±0.128 (±0.163)	0.31 ±0.128 (±0.214)	0.25 ±0.072	0.61 ±0.078
6×18	-0.47 ±0.179 (±0.245)	-0.65 ±0.173 (±0.269)	-0.56 ±0.102	-0.95 ±0.106
6×19	0.27 ±0.135 (±0.193)	0.40 ±0.125 (±0.184)	0.16 ±0.078	0.17 ±0.080
6×20	0.39 ±0.158 (±0.202)	0.66 ±0.145 (±0.233)	0.30 ±0.085	0.33 ±0.086

Table C.8—*continues on next page*

covariate products $X_i X_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
7×8	0.01 ±0.084 (±0.109)	0.10 ±0.069 (±0.090)	-0.03 ±0.053	0.01 ±0.043
7×9	0.11 ±0.120 (±0.150)	-0.06 ±0.089 (±0.102)	-0.05 ±0.067	-0.15 ±0.050
7×10	-0.24 ±0.083 (±0.103)	-0.24 ±0.062 (±0.075)	-0.10 ±0.044	-0.14 ±0.033
7×11	0.20 ±0.157 (±0.222)	0.43 ±0.136 (±0.200)	0.10 ±0.098	0.46 ±0.090
7×12	-0.21 ±0.113 (±0.154)	-0.37 ±0.107 (±0.168)	-0.12 ±0.072	-0.38 ±0.073
7×13	-0.07 ±0.169 (±0.230)	-0.34 ±0.155 (±0.234)	-0.07 ±0.096	-0.13 ±0.095
7×14	0.09 ±0.217 (±0.288)	0.38 ±0.183 (±0.257)	0.07 ±0.118	0.46 ±0.104
7×15	0.30 ±0.197 (±0.253)	0.41 ±0.148 (±0.187)	0.30 ±0.095	0.23 ±0.071
7×16	-0.21 ±0.175 (±0.233)	-0.07 ±0.142 (±0.177)	-0.06 ±0.091	-0.13 ±0.074
7×17	-0.27 ±0.238 (±0.334)	0.18 ±0.217 (±0.378)	-0.13 ±0.135	0.27 ±0.134
7×18	0.31 ±0.322 (±0.441)	-0.44 ±0.304 (±0.468)	-0.10 ±0.181	-0.91 ±0.184
7×19	-0.14 ±0.245 (±0.347)	-0.24 ±0.219 (±0.330)	-0.04 ±0.136	0.10 ±0.129
7×20	-0.19 ±0.285 (±0.383)	-0.20 ±0.247 (±0.370)	-0.35 ±0.145	-0.30 ±0.137
8×9	0.05 ±0.096 (±0.126)	-0.03 ±0.078 (±0.099)	0.16 ±0.060	0.21 ±0.052
8×10	0.20 ±0.067 (±0.087)	0.21 ±0.055 (±0.069)	0.20 ±0.042	0.13 ±0.036
8×11	-0.31 ±0.116 (±0.161)	-0.23 ±0.099 (±0.142)	-0.16 ±0.074	-0.18 ±0.069
8×12	0.07 ±0.080 (±0.111)	-0.03 ±0.075 (±0.114)	-0.05 ±0.053	-0.09 ±0.052
8×13	0.15 ±0.125 (±0.183)	0.26 ±0.110 (±0.178)	0.11 ±0.076	0.14 ±0.075
8×14	0.02 ±0.154 (±0.216)	0.08 ±0.142 (±0.194)	-0.06 ±0.094	-0.17 ±0.090
8×15	-0.30 ±0.162 (±0.212)	-0.67 ±0.127 (±0.164)	-0.42 ±0.094	-0.44 ±0.074
8×16	0.25 ±0.146 (±0.193)	0.49 ±0.125 (±0.166)	0.19 ±0.085	0.39 ±0.076
8×17	-0.20 ±0.169 (±0.232)	-0.23 ±0.164 (±0.271)	-0.11 ±0.104	-0.28 ±0.110
8×18	0.37 ±0.235 (±0.337)	0.79 ±0.218 (±0.349)	0.64 ±0.144	1.37 ±0.145
8×19	0.16 ±0.186 (±0.264)	-0.03 ±0.166 (±0.252)	0.11 ±0.113	-0.14 ±0.110
8×20	-0.17 ±0.215 (±0.299)	-0.48 ±0.197 (±0.297)	-0.07 ±0.124	-0.55 ±0.125
9×10	-0.07 ±0.083 (±0.108)	0.02 ±0.066 (±0.082)	-0.12 ±0.046	0.05 ±0.037
9×11	0.17 ±0.130 (±0.168)	0.13 ±0.114 (±0.163)	0.09 ±0.080	-0.22 ±0.076
9×12	0.07 ±0.096 (±0.128)	0.13 ±0.095 (±0.145)	-0.05 ±0.058	0.19 ±0.062
9×13	-0.33 ±0.138 (±0.185)	-0.35 ±0.127 (±0.179)	-0.23 ±0.076	-0.21 ±0.078
9×14	-0.56 ±0.165 (±0.213)	-0.47 ±0.139 (±0.184)	-0.36 ±0.094	-0.38 ±0.084
9×15	-0.06 ±0.187 (±0.238)	0.17 ±0.160 (±0.204)	-0.01 ±0.101	-0.01 ±0.088
9×16	0.17 ±0.166 (±0.211)	0.34 ±0.144 (±0.184)	0.17 ±0.090	0.19 ±0.082
9×17	-0.20 ±0.195 (±0.259)	-0.49 ±0.192 (±0.296)	-0.30 ±0.110	-0.72 ±0.117
9×18	0.87 ±0.250 (±0.352)	0.69 ±0.246 (±0.378)	0.79 ±0.142	1.07 ±0.146
9×19	-0.15 ±0.194 (±0.248)	-0.15 ±0.188 (±0.264)	-0.24 ±0.107	-0.22 ±0.111
9×20	-0.18 ±0.234 (±0.315)	-0.21 ±0.214 (±0.304)	-0.03 ±0.126	0.05 ±0.123
10×11	-0.11 ±0.096 (±0.122)	-0.25 ±0.086 (±0.130)	-0.08 ±0.059	-0.23 ±0.058
10×12	0.03 ±0.070 (±0.093)	0.19 ±0.066 (±0.106)	0.03 ±0.042	0.31 ±0.044
10×13	0.13 ±0.093 (±0.120)	0.15 ±0.087 (±0.140)	0.18 ±0.051	0.09 ±0.055
10×14	0.15 ±0.125 (±0.150)	0.12 ±0.113 (±0.147)	0.09 ±0.061	-0.08 ±0.060
10×15	-0.05 ±0.115 (±0.140)	-0.22 ±0.096 (±0.123)	-0.05 ±0.056	-0.21 ±0.049
10×16	0.15 ±0.113 (±0.136)	0.30 ±0.102 (±0.135)	0.19 ±0.054	0.24 ±0.049
10×17	0.29 ±0.128 (±0.164)	-0.09 ±0.124 (±0.194)	0.16 ±0.062	-0.28 ±0.069
10×18	-0.50 ±0.179 (±0.230)	-0.24 ±0.172 (±0.276)	-0.25 ±0.090	0.26 ±0.095
10×19	0.11 ±0.138 (±0.167)	0.40 ±0.134 (±0.197)	0.06 ±0.067	0.17 ±0.071
10×20	-0.10 ±0.142 (±0.175)	-0.41 ±0.133 (±0.198)	-0.26 ±0.069	-0.34 ±0.071
11×12	-0.08 ±0.089 (±0.119)	0.02 ±0.085 (±0.137)	-0.16 ±0.054	-0.15 ±0.057
11×13	-0.12 ±0.117 (±0.145)	-0.20 ±0.112 (±0.172)	-0.10 ±0.064	0.05 ±0.068
11×14	0.48 ±0.185 (±0.233)	0.48 ±0.169 (±0.237)	0.32 ±0.105	0.54 ±0.109
11×15	0.09 ±0.207 (±0.261)	0.24 ±0.187 (±0.270)	0.10 ±0.115	0.30 ±0.117
11×16	-0.01 ±0.178 (±0.235)	-0.43 ±0.161 (±0.223)	-0.05 ±0.101	-0.11 ±0.104
11×17	0.36 ±0.194 (±0.248)	0.71 ±0.196 (±0.297)	0.08 ±0.105	0.70 ±0.118
11×18	-0.28 ±0.277 (±0.354)	0.12 ±0.266 (±0.400)	-0.08 ±0.151	-0.01 ±0.164
11×19	-0.43 ±0.215 (±0.271)	-0.90 ±0.204 (±0.303)	-0.15 ±0.116	-0.97 ±0.123
11×20	-0.05 ±0.264 (±0.349)	0.24 ±0.245 (±0.379)	0.12 ±0.144	0.05 ±0.152
12×13	0.20 ±0.085 (±0.109)	0.19 ±0.081 (±0.120)	0.19 ±0.045	0.30 ±0.048
12×14	-0.04 ±0.129 (±0.162)	0.11 ±0.128 (±0.206)	0.02 ±0.066	0.09 ±0.072
12×15	0.09 ±0.138 (±0.172)	0.26 ±0.132 (±0.204)	0.08 ±0.073	0.01 ±0.080
12×16	-0.33 ±0.121 (±0.148)	-0.34 ±0.119 (±0.180)	-0.23 ±0.062	-0.24 ±0.069
12×17	0.06 ±0.123 (±0.161)	-0.31 ±0.122 (±0.216)	-0.02 ±0.063	-0.46 ±0.069
12×18	0.06 ±0.175 (±0.226)	0.10 ±0.171 (±0.283)	0.07 ±0.091	0.12 ±0.098
12×19	-0.20 ±0.135 (±0.172)	-0.41 ±0.134 (±0.221)	-0.08 ±0.066	-0.39 ±0.071

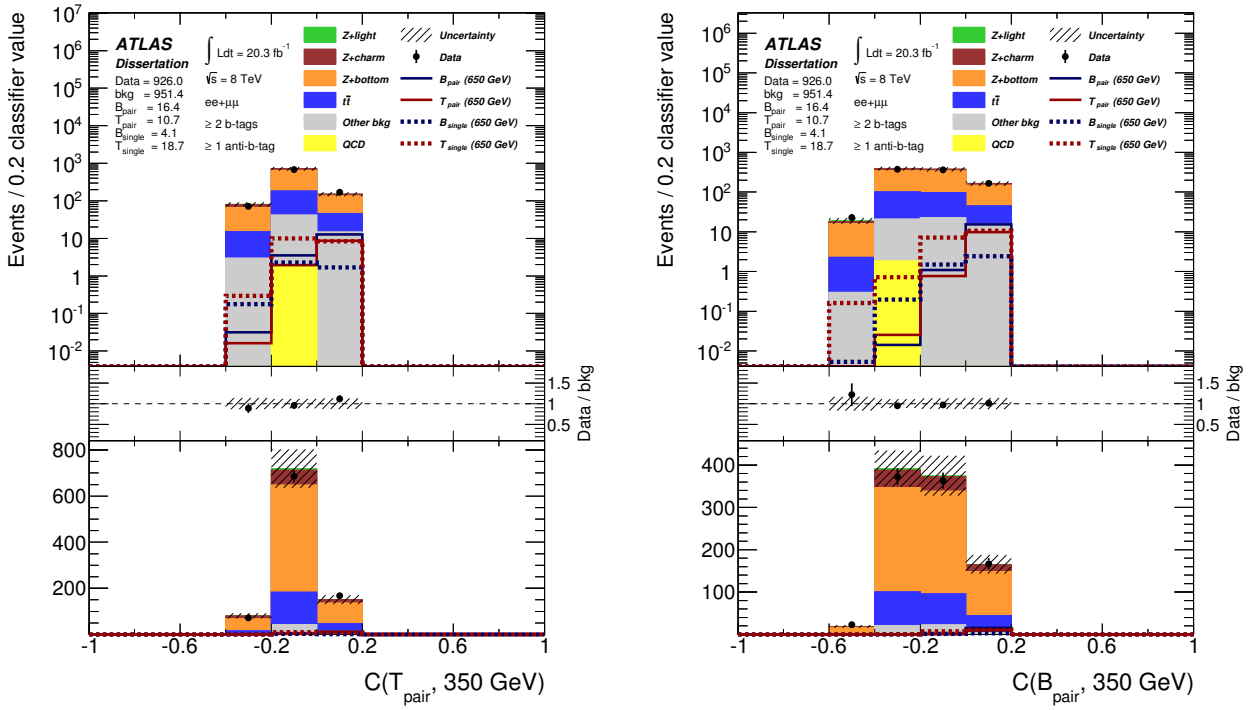
Table C.8—*continues on next page*

covariate products $X_i X_{j>i}$	$\mathbf{B}^{i,j}_{OLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2\mathbf{I}$	$\mathbf{B}^{i,j}_{WLS}$ $\mathbf{\Sigma} = \hat{\sigma}^2 diag(1/w_i)$	$\mathbf{B}^{i,j}_{GLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2)$	$\mathbf{B}^{i,j}_{WGLS}$ $\mathbf{\Sigma} = diag(\hat{\sigma}_i^2/w_i)$
12×20	0.10 ±0.171 (±0.226)	0.16 ±0.168 (±0.294)	0.02 ±0.087	0.46 ±0.097
13×14	0.07 ±0.170 (±0.207)	-0.02 ±0.161 (±0.245)	-0.10 ±0.087	-0.26 ±0.092
13×15	-0.19 ±0.189 (±0.228)	-0.20 ±0.179 (±0.259)	-0.06 ±0.094	0.05 ±0.103
13×16	0.21 ±0.159 (±0.192)	0.72 ±0.152 (±0.219)	0.10 ±0.078	0.38 ±0.084
13×17	-0.46 ±0.176 (±0.236)	-0.39 ±0.175 (±0.294)	-0.12 ±0.088	-0.23 ±0.099
13×18	0.20 ±0.233 (±0.311)	0.42 ±0.230 (±0.371)	-0.17 ±0.118	-0.31 ±0.127
13×19	0.43 ±0.185 (±0.234)	0.55 ±0.180 (±0.268)	0.31 ±0.090	0.86 ±0.099
13×20	-0.04 ±0.233 (±0.304)	-0.85 ±0.223 (±0.363)	0.11 ±0.113	-0.44 ±0.121
14×15	0.14 ±0.221 (±0.255)	-0.33 ±0.214 (±0.271)	0.33 ±0.104	0.12 ±0.112
14×16	-0.02 ±0.186 (±0.223)	-0.24 ±0.183 (±0.266)	-0.07 ±0.087	-0.22 ±0.097
14×17	-0.39 ±0.207 (±0.257)	-0.58 ±0.202 (±0.310)	-0.16 ±0.097	-0.19 ±0.108
14×18	0.35 ±0.273 (±0.341)	0.54 ±0.269 (±0.417)	0.17 ±0.129	-0.21 ±0.142
14×19	-0.10 ±0.183 (±0.215)	0.21 ±0.185 (±0.285)	-0.03 ±0.075	0.68 ±0.084
14×20	-0.37 ±0.231 (±0.297)	-0.25 ±0.232 (±0.387)	-0.26 ±0.112	-0.17 ±0.126
15×16	-0.11 ±0.165 (±0.177)	-0.24 ±0.150 (±0.174)	-0.13 ±0.073	-0.17 ±0.068
15×17	0.11 ±0.232 (±0.272)	-0.11 ±0.236 (±0.332)	-0.04 ±0.105	-0.29 ±0.122
15×18	-0.08 ±0.301 (±0.346)	0.18 ±0.307 (±0.424)	-0.20 ±0.133	-0.15 ±0.151
15×19	0.01 ±0.222 (±0.269)	-0.21 ±0.230 (±0.326)	-0.08 ±0.105	-0.31 ±0.120
15×20	0.28 ±0.236 (±0.278)	0.96 ±0.248 (±0.337)	0.16 ±0.106	0.81 ±0.125
16×17	-0.21 ±0.185 (±0.211)	0.18 ±0.198 (±0.278)	-0.26 ±0.085	0.22 ±0.101
16×18	-0.14 ±0.243 (±0.279)	-1.13 ±0.253 (±0.387)	-0.07 ±0.106	-0.90 ±0.128
16×19	0.09 ±0.187 (±0.213)	0.09 ±0.194 (±0.277)	-0.02 ±0.083	-0.14 ±0.099
16×20	0.27 ±0.184 (±0.204)	0.67 ±0.194 (±0.265)	0.41 ±0.086	0.55 ±0.102
17×18	-0.02 ±0.199 (±0.239)	-0.22 ±0.206 (±0.318)	0.19 ±0.088	0.01 ±0.103
17×19	0.13 ±0.181 (±0.217)	0.35 ±0.189 (±0.289)	0.04 ±0.081	0.16 ±0.095
17×20	0.50 ±0.234 (±0.282)	0.79 ±0.238 (±0.382)	0.31 ±0.108	0.47 ±0.128
18×19	0.13 ±0.234 (±0.259)	0.43 ±0.246 (±0.348)	0.15 ±0.101	0.75 ±0.118
18×20	-0.34 ±0.276 (±0.335)	-0.71 ±0.279 (±0.436)	-0.37 ±0.127	-0.07 ±0.145
19×20	-0.01 ±0.241 (±0.292)	-0.12 ±0.256 (±0.399)	0.04 ±0.112	-0.71 ±0.133

Table C.8: Linear regression covariate products' parameter estimates for the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + X_i\mathbf{B}_{i,j>i}X_j + \mathbf{u}$ with alternate models of disturbance covariance $\mathbf{u} \sim \mathcal{N}(0, \mathbf{\Sigma}_{model})$ for the B_{pair} signal. Covariates $\mathbf{X}_i$ have been scaled by $1/\hat{s}_{X_i}$ ($\mathbf{y}$ by $1/\hat{s_y}$) before the products X_iX_j are constructed. Heteroscedasticity-consistent errors are shown in parentheses (HC3 method [241]).

C.3. Blinded Modeling Validation: All Stage 1 Classifiers

This section provides additional modeling validation (data–background comparison plots) not shown in the main text of section 5.8, the validation results of all 59 signal mass classifiers shown in figures C.5 (1)–(59).

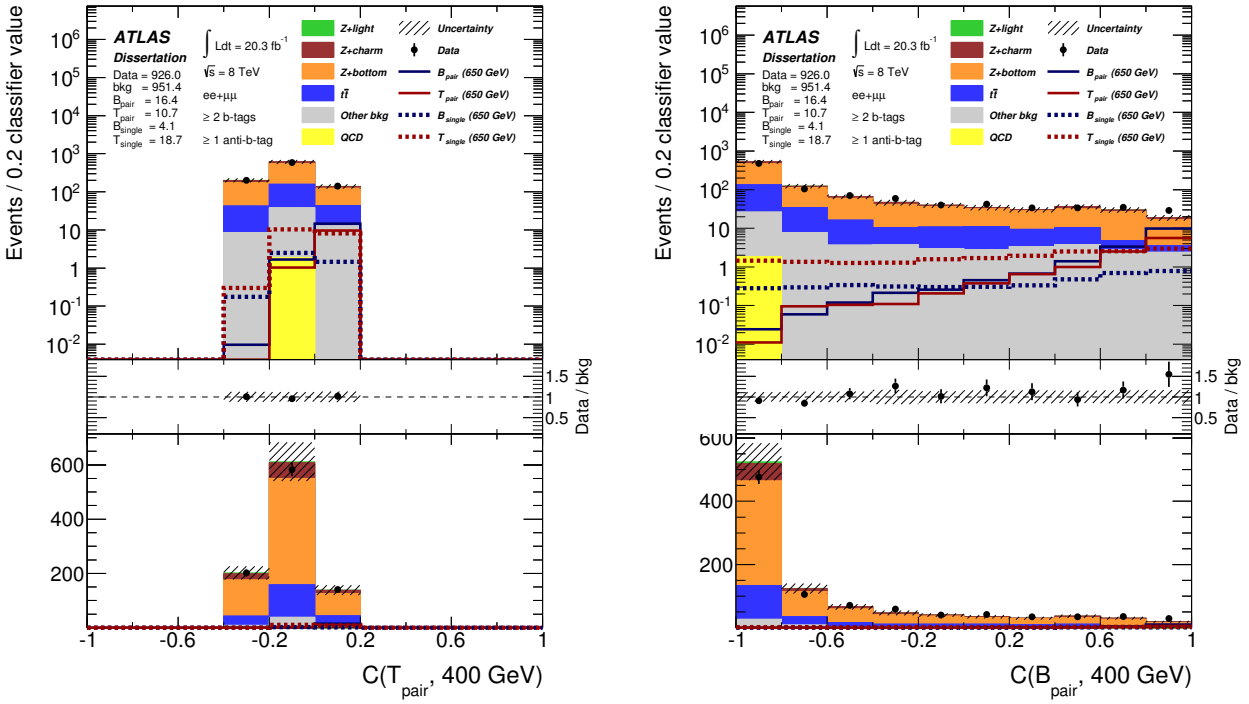


(1) (2)

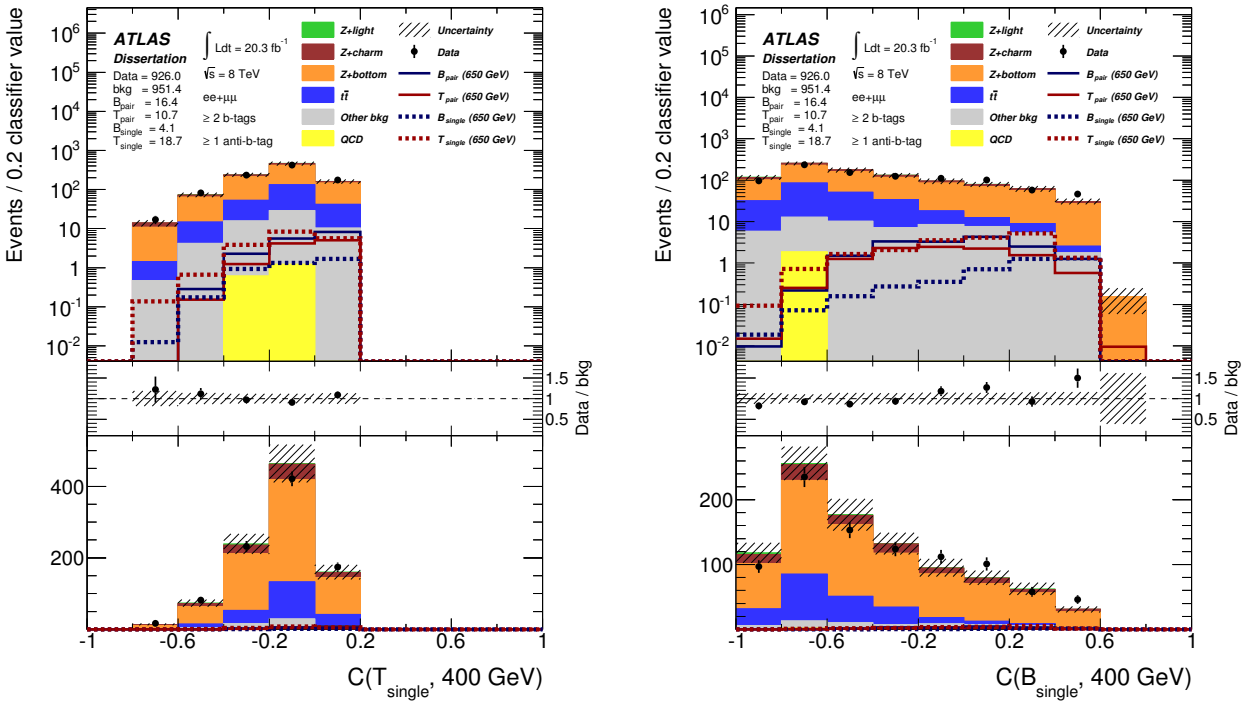
Figure C.5: Simulated data modeling validation of all stage 1 classifiers. The subfigures are:

(1) $m(T_{pair}) = 350$ [GeV], (2) $m(B_{pair}) = 350$ [GeV],
(3) $m(T_{pair}) = 400$ [GeV], (4) $m(B_{pair}) = 400$ [GeV], (5) $m(T_{single}) = 400$ [GeV], (6) $m(B_{single}) = 400$ [GeV],
(7) $m(T_{pair}) = 450$ [GeV], (8) $m(B_{pair}) = 450$ [GeV], (9) $m(T_{single}) = 450$ [GeV],
(10) $m(T_{pair}) = 500$ [GeV], (11) $m(B_{pair}) = 500$ [GeV], (12) $m(T_{single}) = 500$ [GeV], (13) $m(B_{single}) = 500$ [GeV],
(14) $m(T_{pair}) = 550$ [GeV], (15) $m(B_{pair}) = 550$ [GeV], (16) $m(T_{single}) = 550$ [GeV], (17) $m(B_{single}) = 550$ [GeV],
(18) $m(T_{pair}) = 600$ [GeV], (19) $m(B_{pair}) = 600$ [GeV], (20) $m(T_{single}) = 600$ [GeV], (21) $m(B_{single}) = 600$ [GeV],
(22) $m(T_{pair}) = 650$ [GeV], (23) $m(B_{pair}) = 650$ [GeV], (24) $m(T_{single}) = 650$ [GeV], (25) $m(B_{single}) = 650$ [GeV],
(26) $m(T_{pair}) = 700$ [GeV], (27) $m(B_{pair}) = 700$ [GeV], (28) $m(T_{single}) = 700$ [GeV], (29) $m(B_{single}) = 700$ [GeV],
(30) $m(T_{pair}) = 750$ [GeV], (31) $m(B_{pair}) = 750$ [GeV], (32) $m(T_{single}) = 750$ [GeV], (33) $m(B_{single}) = 750$ [GeV],
(34) $m(T_{pair}) = 800$ [GeV], (35) $m(B_{pair}) = 800$ [GeV], (36) $m(T_{single}) = 800$ [GeV], (37) $m(B_{single}) = 800$ [GeV],
(38) $m(T_{pair}) = 850$ [GeV], (39) $m(B_{pair}) = 850$ [GeV], (40) $m(T_{single}) = 850$ [GeV], (41) $m(B_{single}) = 850$ [GeV],
(42) $m(T_{pair}) = 900$ [GeV], (43) $m(B_{pair}) = 900$ [GeV], (44) $m(T_{single}) = 900$ [GeV], (45) $m(B_{single}) = 900$ [GeV],
(46) $m(T_{pair}) = 950$ [GeV], (47) $m(B_{pair}) = 950$ [GeV], (48) $m(T_{single}) = 950$ [GeV], (49) $m(B_{single}) = 950$ [GeV],
(50) $m(T_{pair}) = 1000$ [GeV], (51) $m(B_{pair}) = 1000$ [GeV], (52) $m(T_{single}) = 1000$ [GeV], (53) $m(B_{single}) = 1000$ [GeV],
(54) $m(T_{pair}) = 1050$ [GeV], (55) $m(B_{pair}) = 1050$ [GeV], (56) $m(T_{single}) = 1050$ [GeV], (57) $m(B_{single}) = 1200$ [GeV],
(58) $m(T_{pair}) = 1100$ [GeV], (59) $m(B_{pair}) = 1100$ [GeV].

Figure C.5 continues on the next 15 pages.

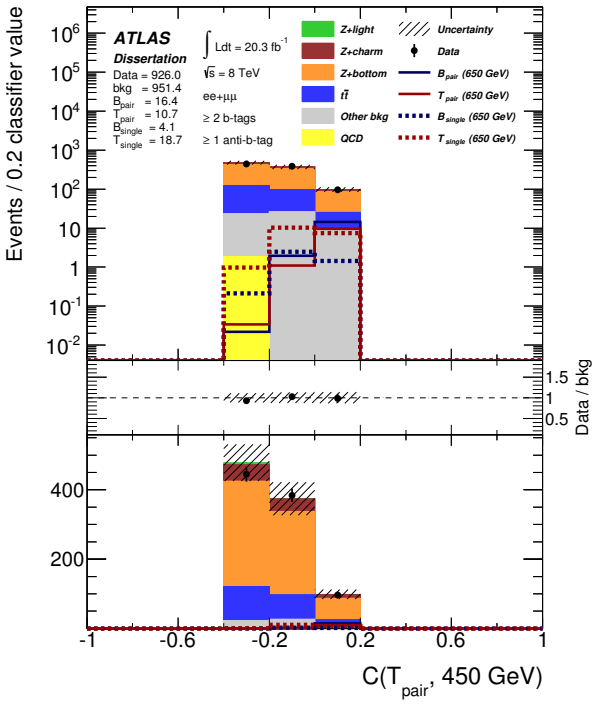


(3) (4)

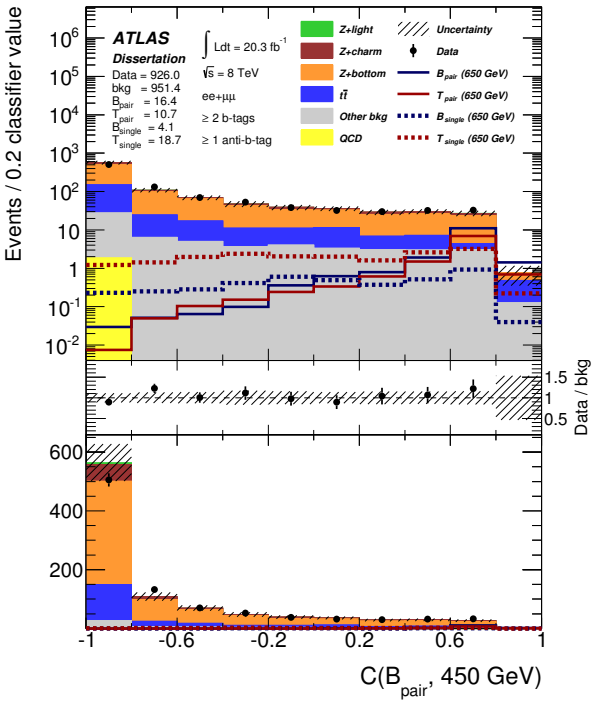


(5) (6)

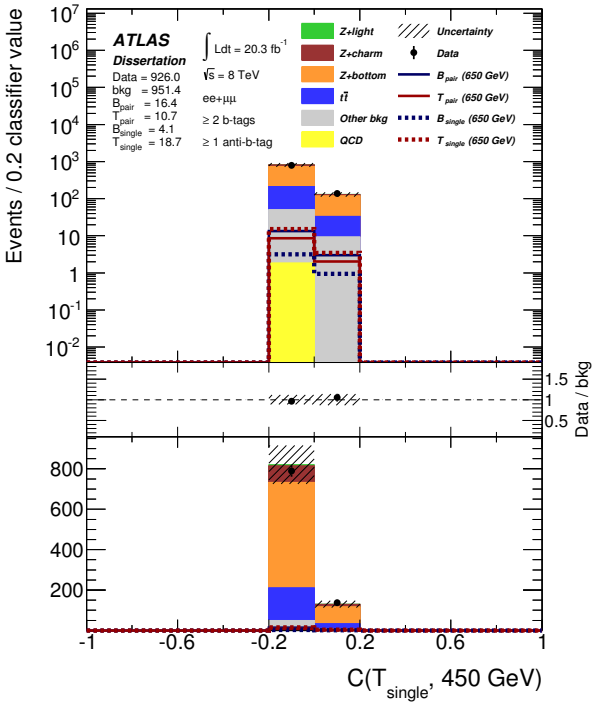
Figure C.5: – *Continues on next 14 pages.*



(7)

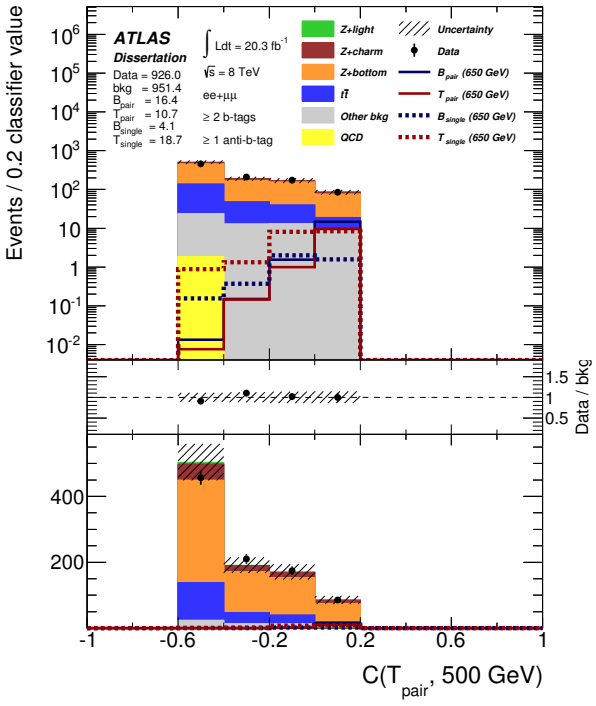


(8)

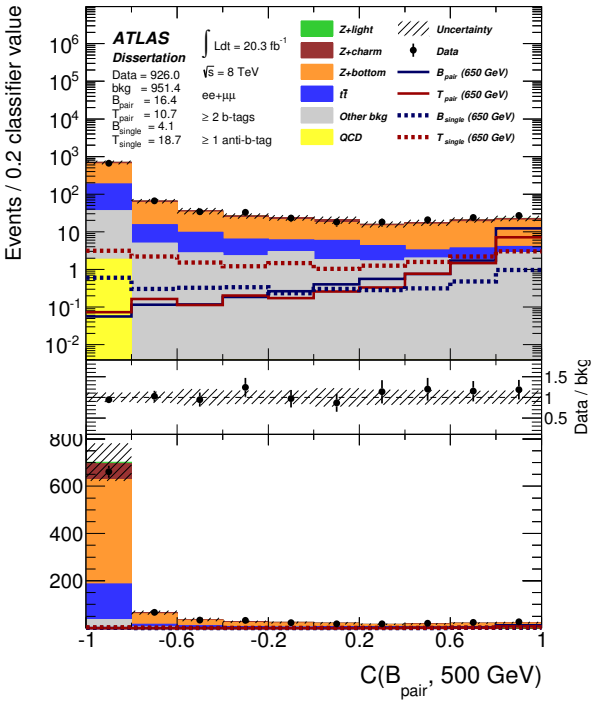


(9)

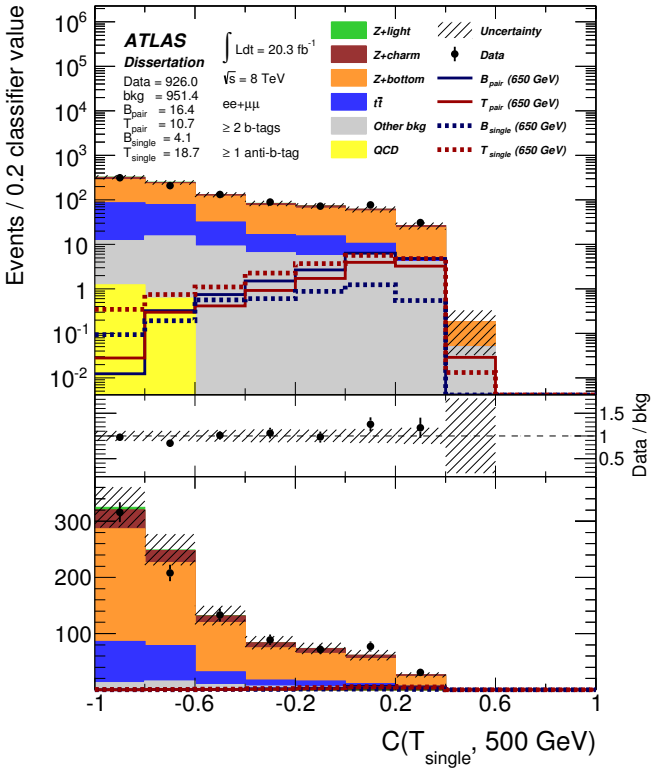
Figure C.5: – *Continued, continues on next 13 pages.*



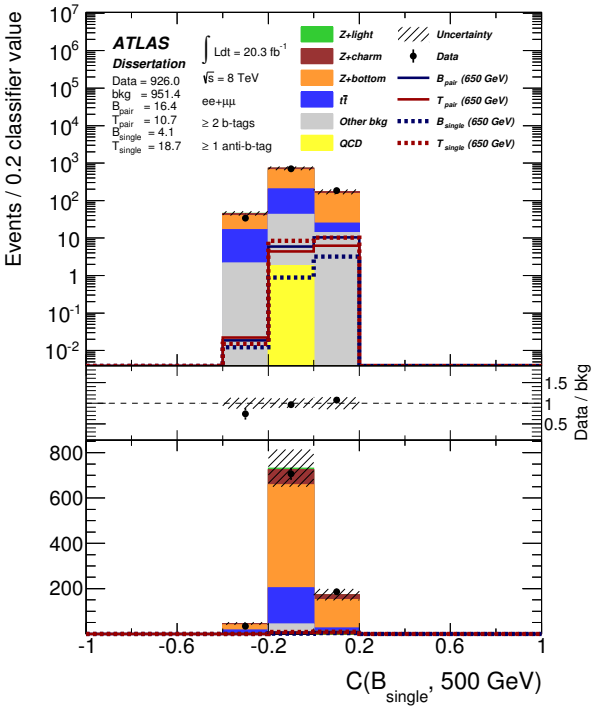
(10)



(11)

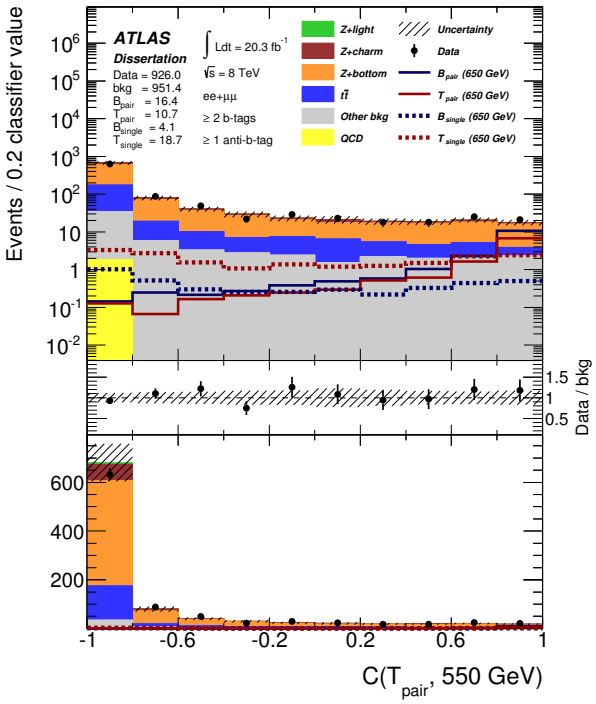


(12)

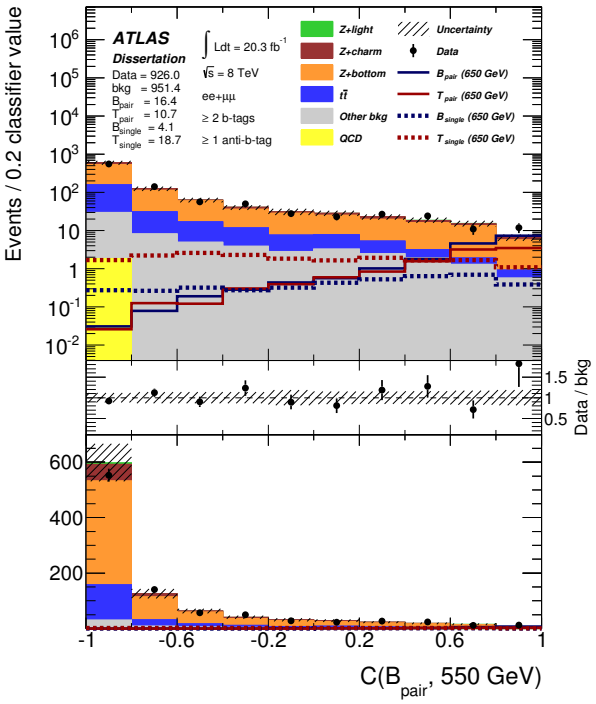


(13)

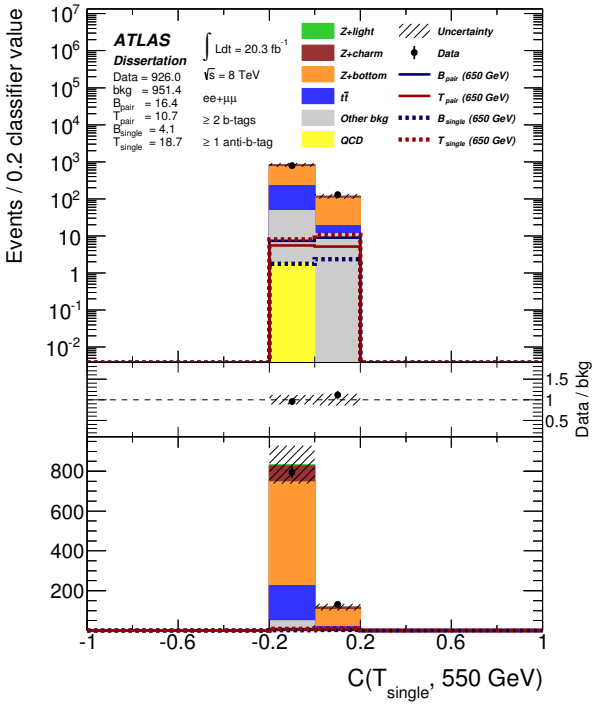
Figure C.5: – *Continued, continues on next 12 pages.*



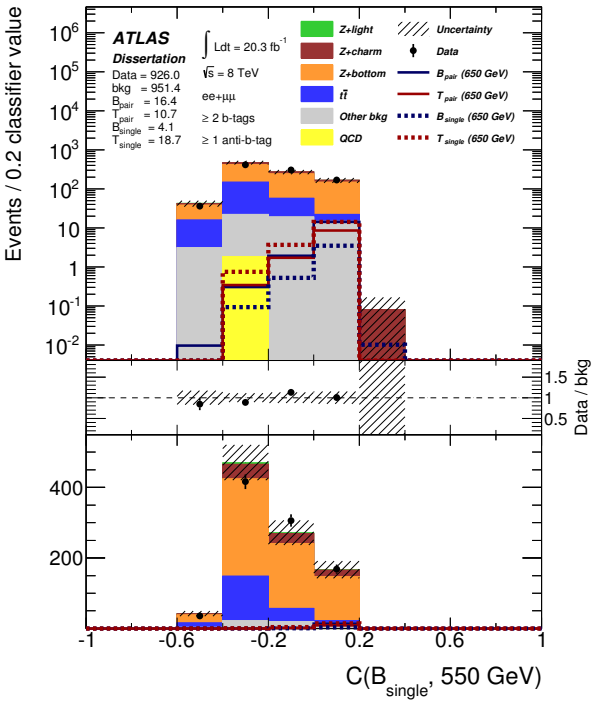
(14)



(15)

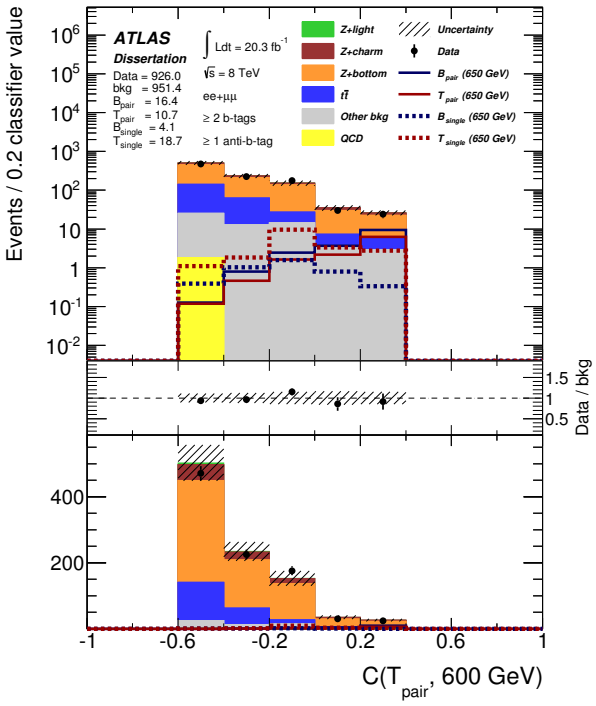


(16)

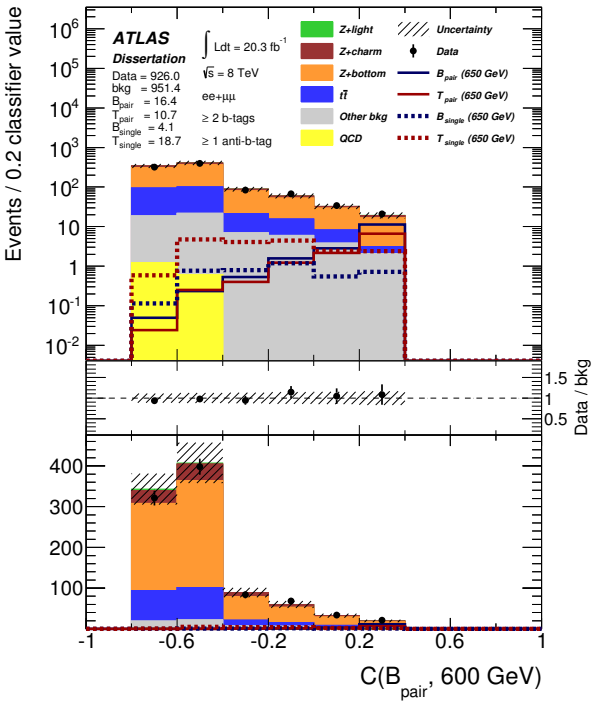


(17)

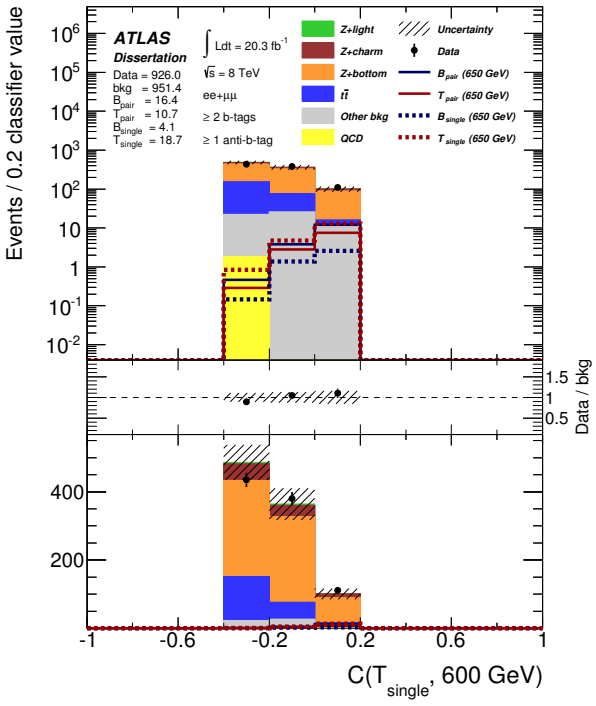
Figure C.5: – *Continued, continues on next 11 pages.*



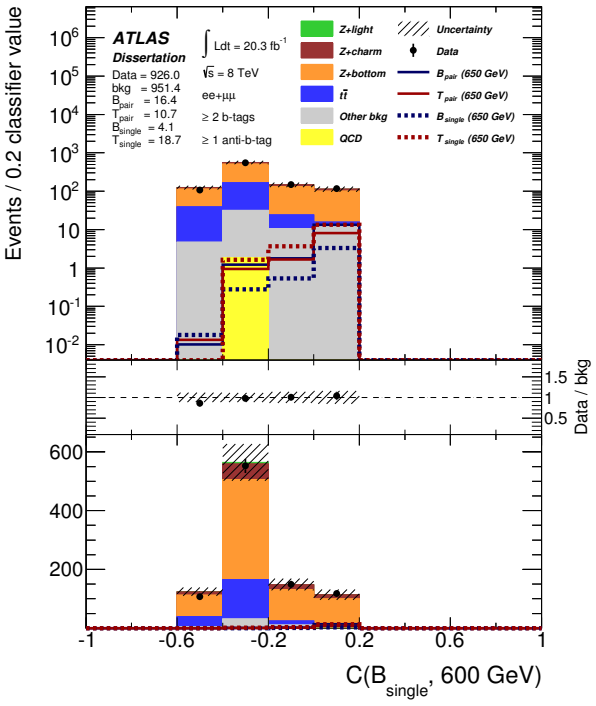
(18)



(19)

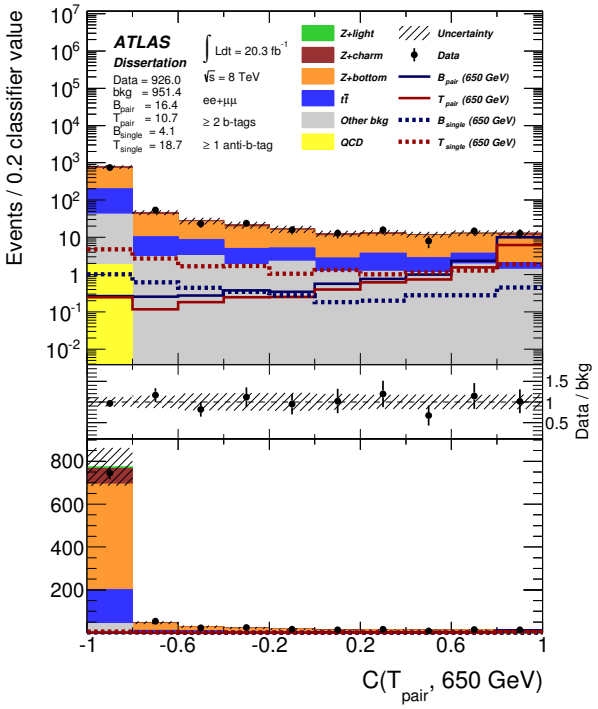


(20)

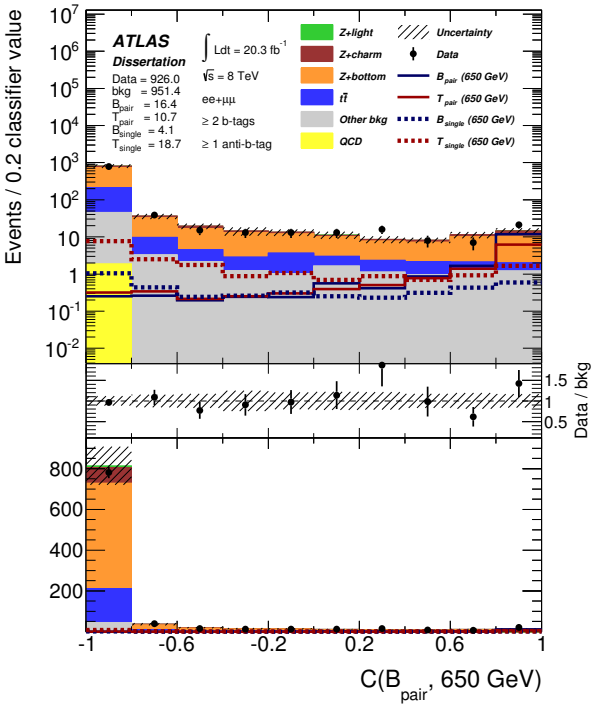


(21)

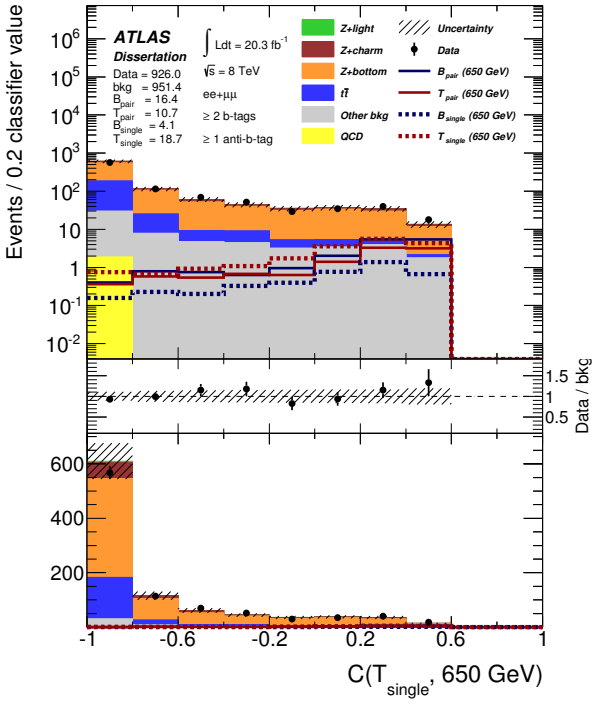
Figure C.5: – *Continued, continues on next 10 pages.*



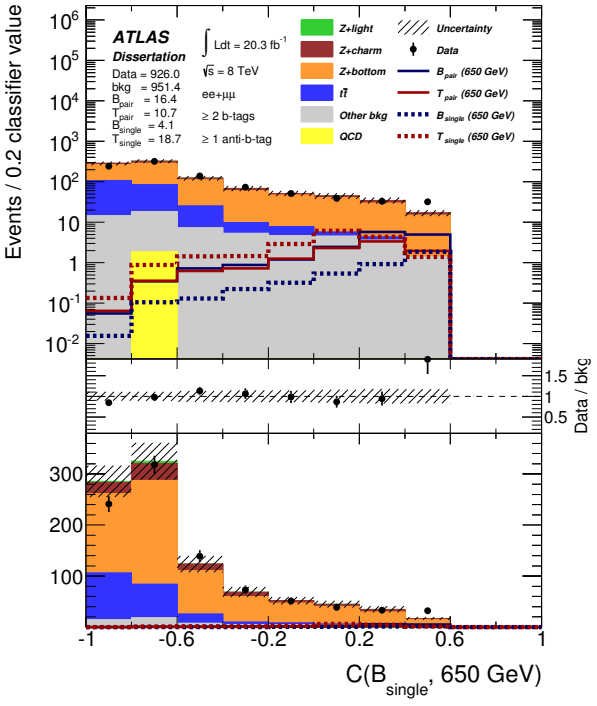
(22)



(23)

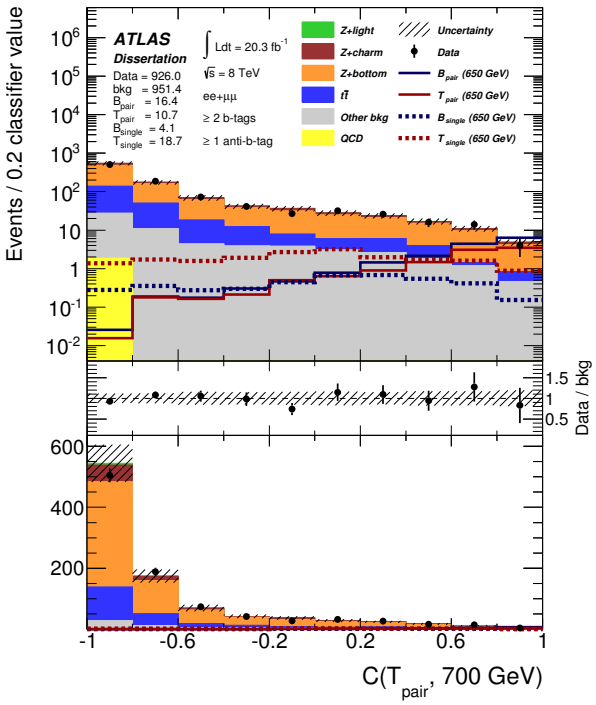


(24)

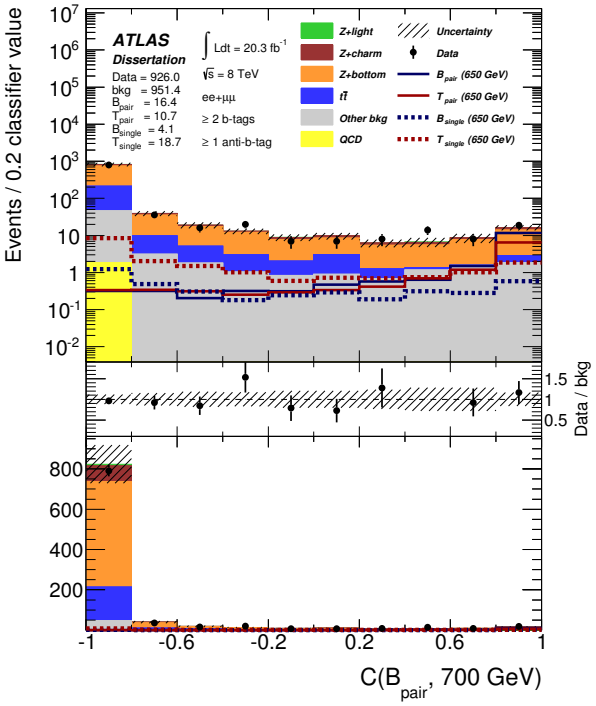


(25)

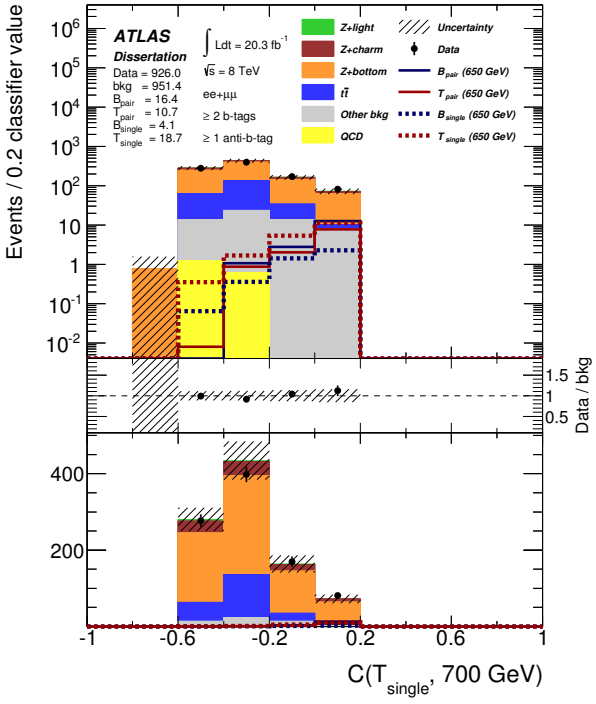
Figure C.5: – *Continued, continues on next 9 pages.*



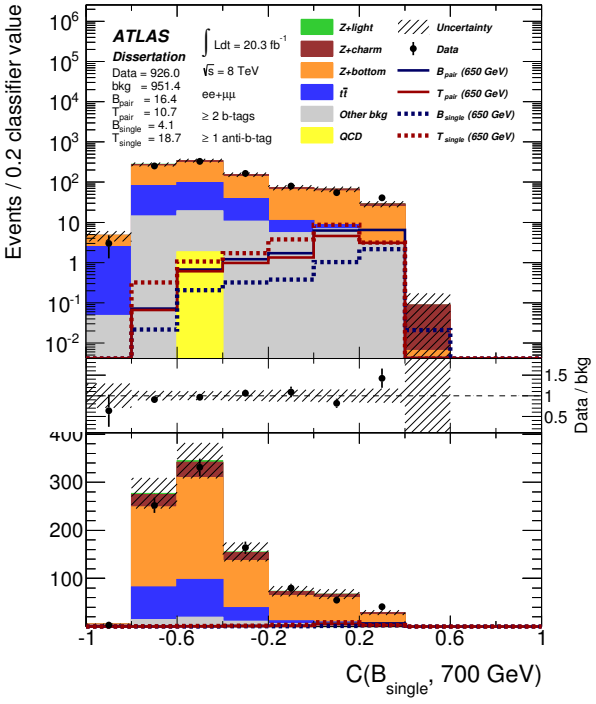
(26)



(27)

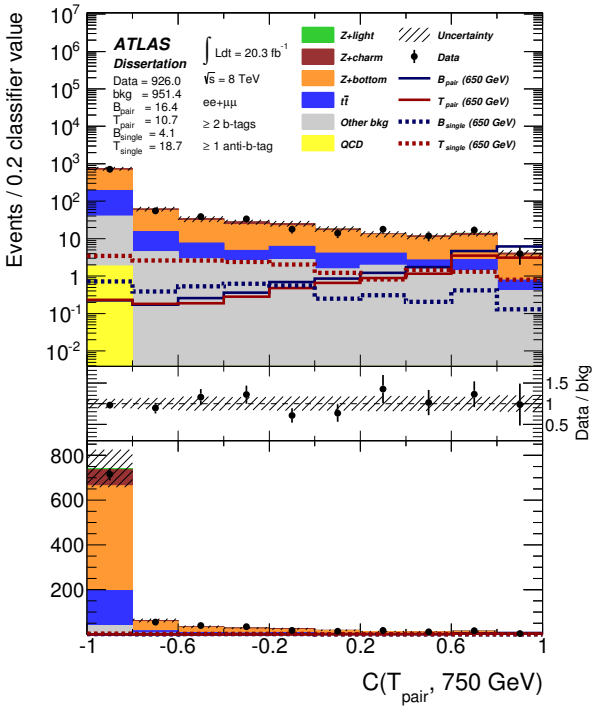


(28)

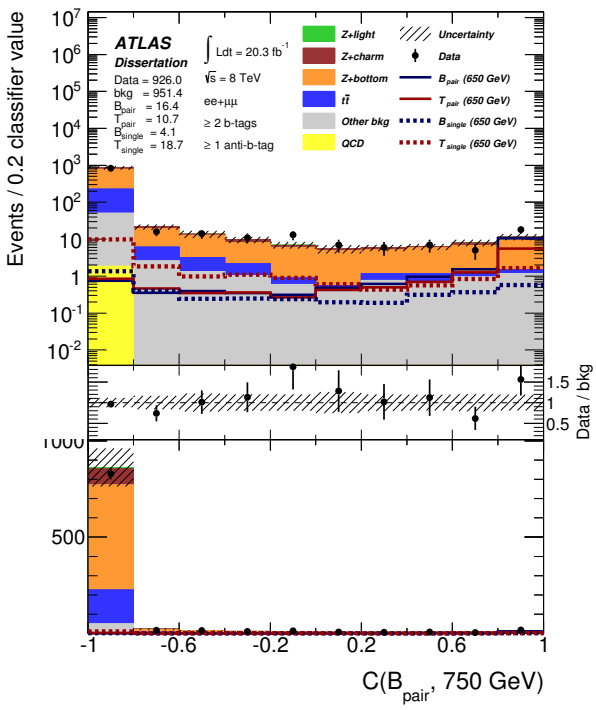


(29)

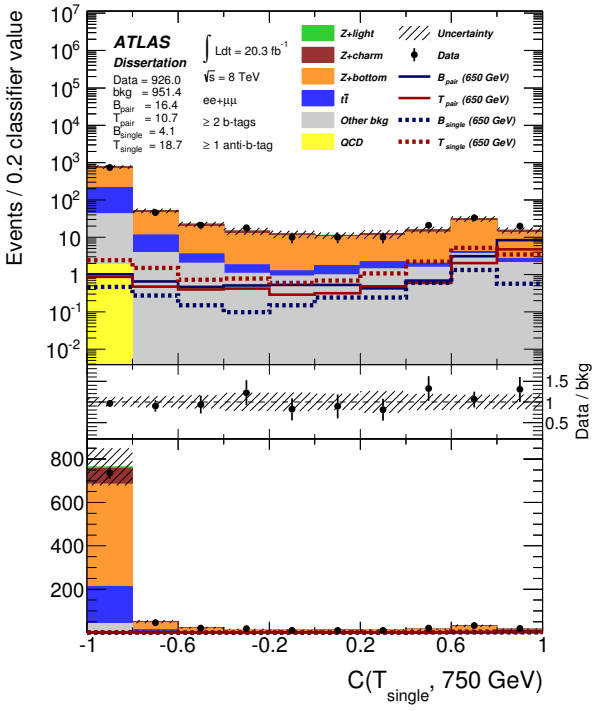
Figure C.5: – *Continued, continues on next 8 pages.*



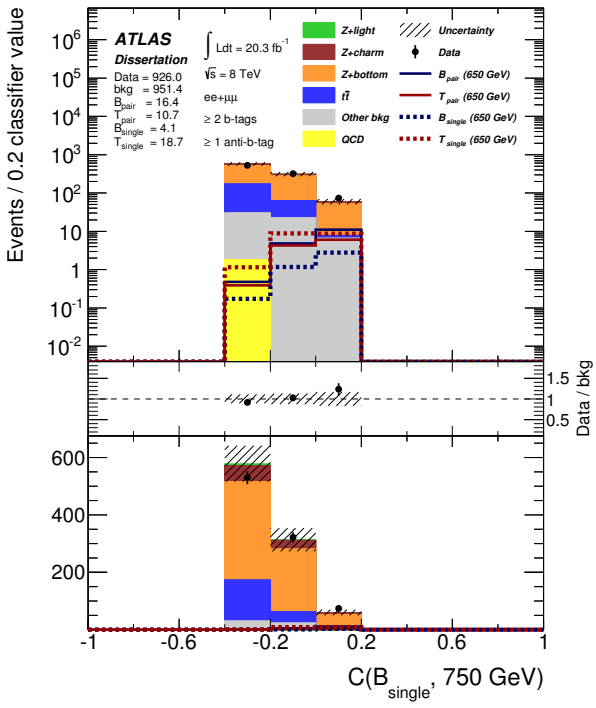
(30)



(31)

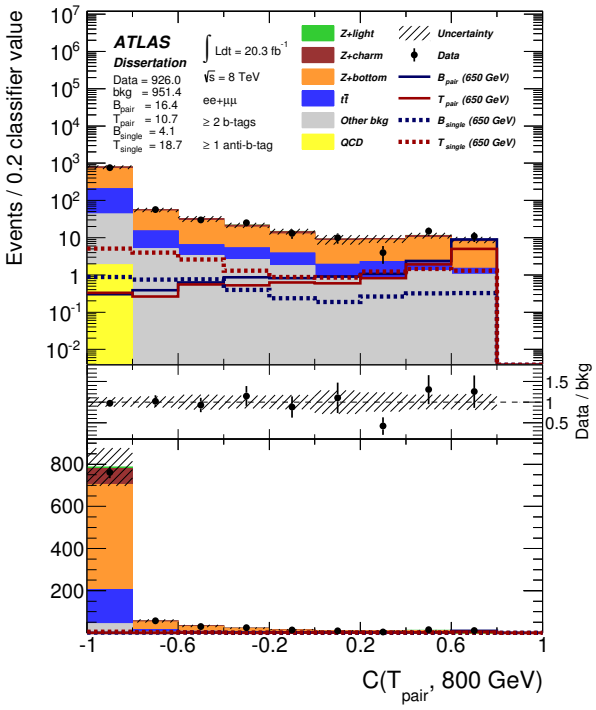


(32)

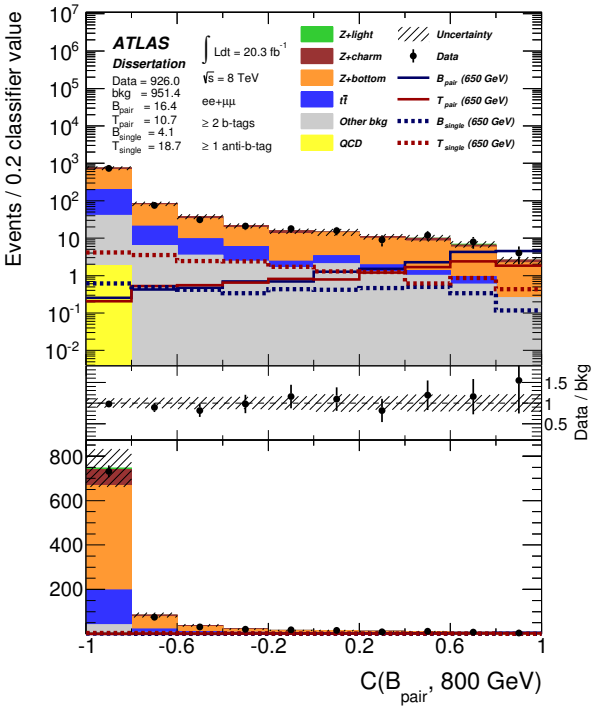


(33)

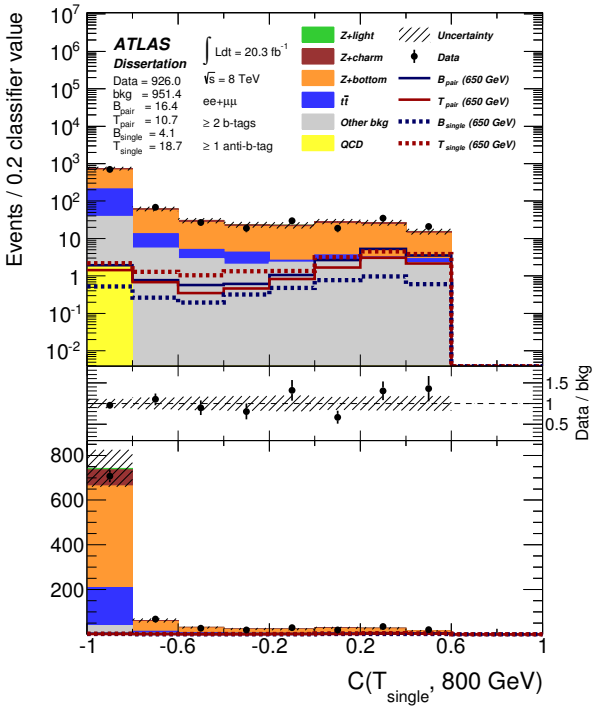
Figure C.5: – *Continued, continues on next 7 pages.*



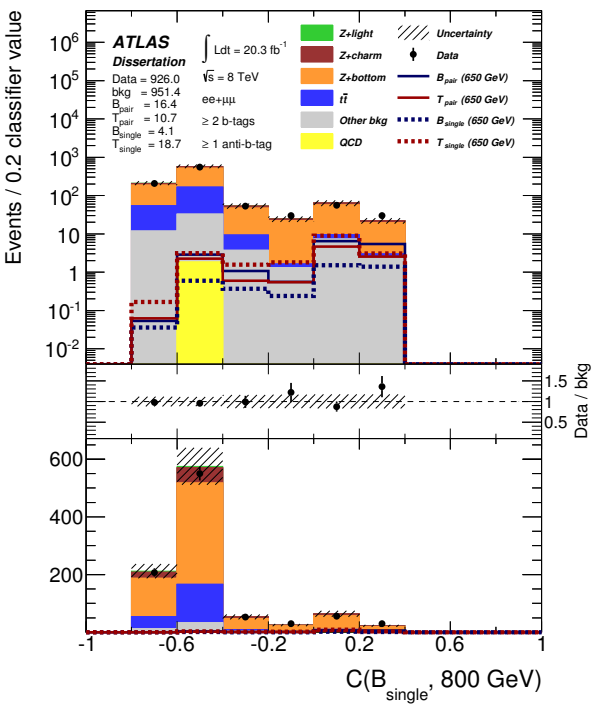
(34)



(35)

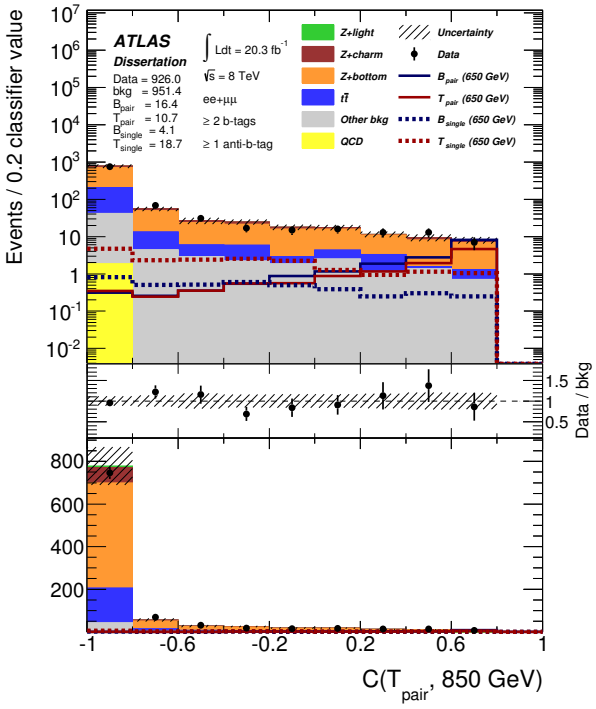


(36)

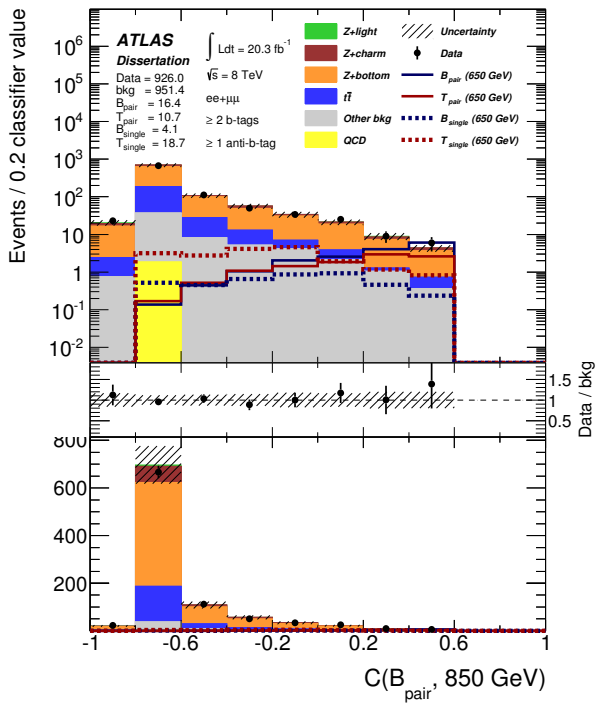


(37)

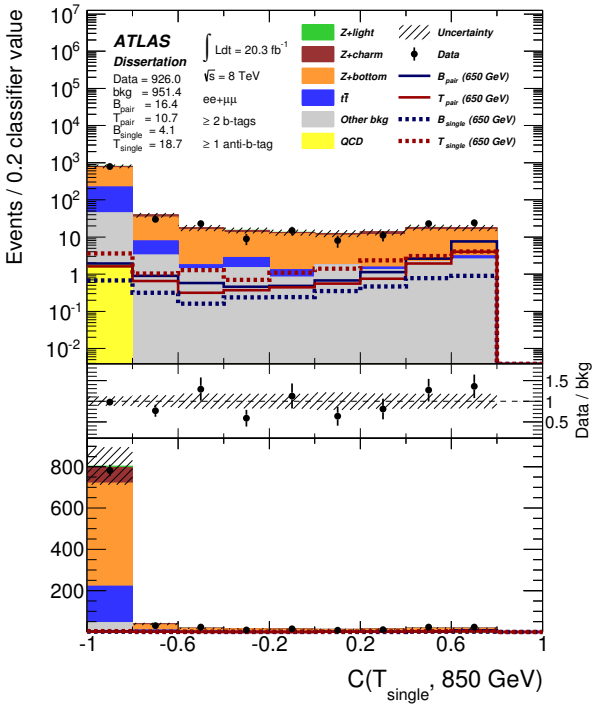
Figure C.5: – *Continued, continues on next 6 pages.*



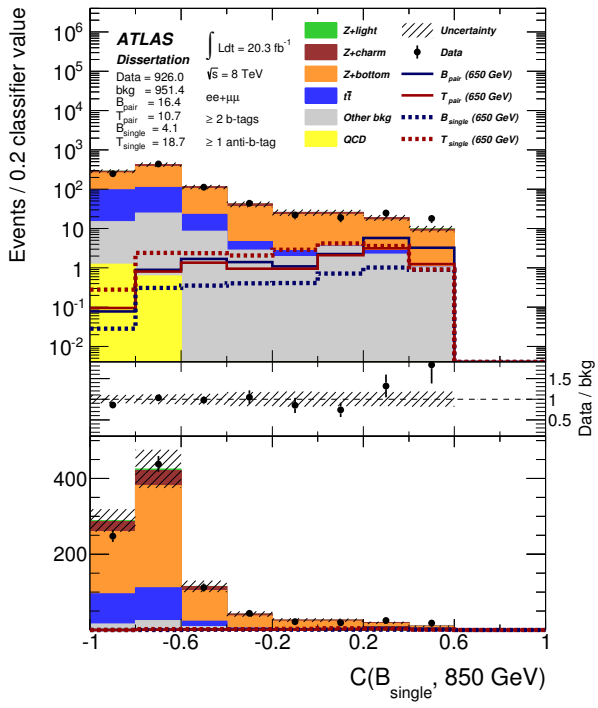
(38)



(39)

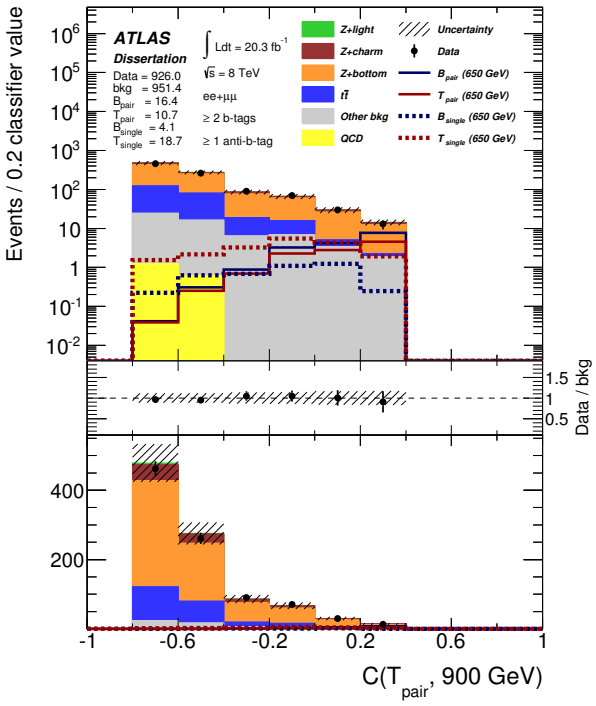


(40)

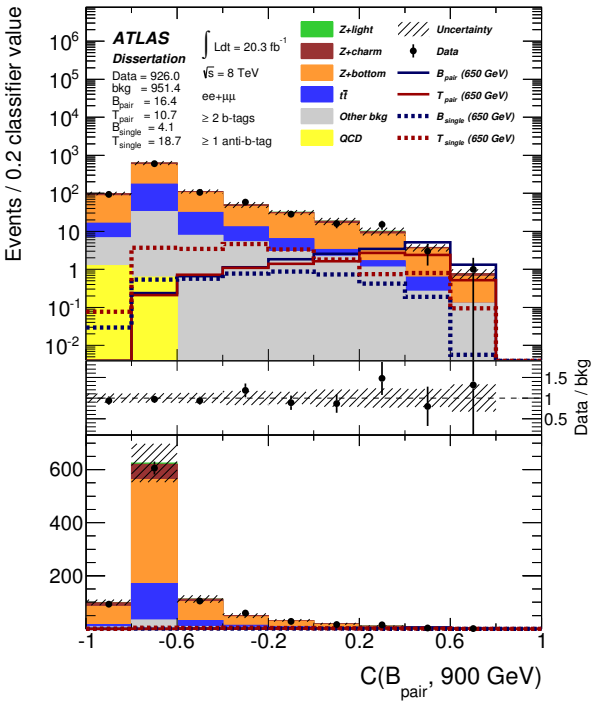


(41)

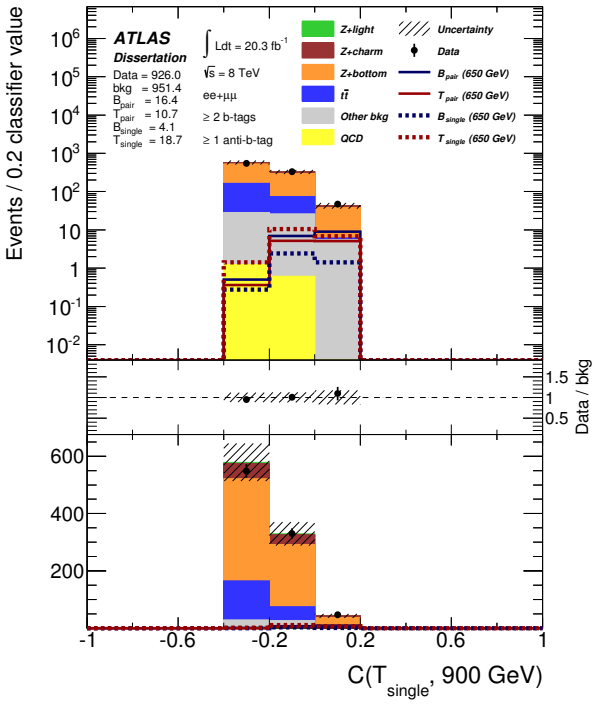
Figure C.5: – *Continued, continues on next 5 pages.*



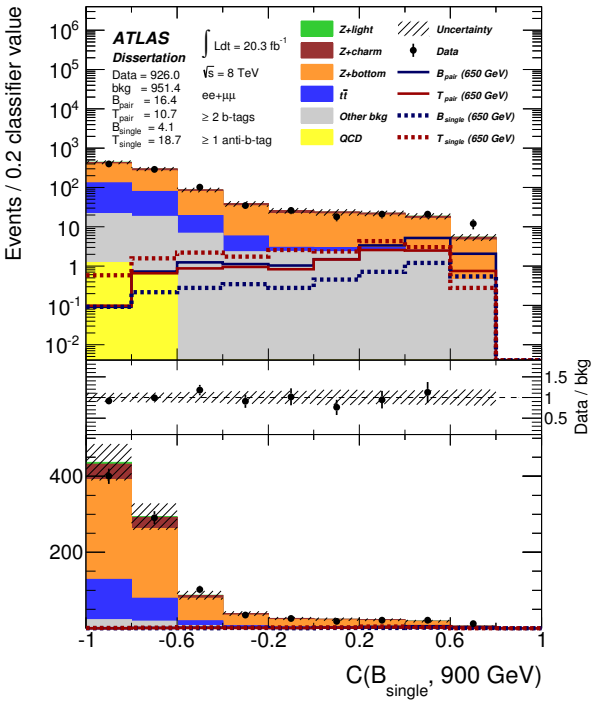
(42)



(43)

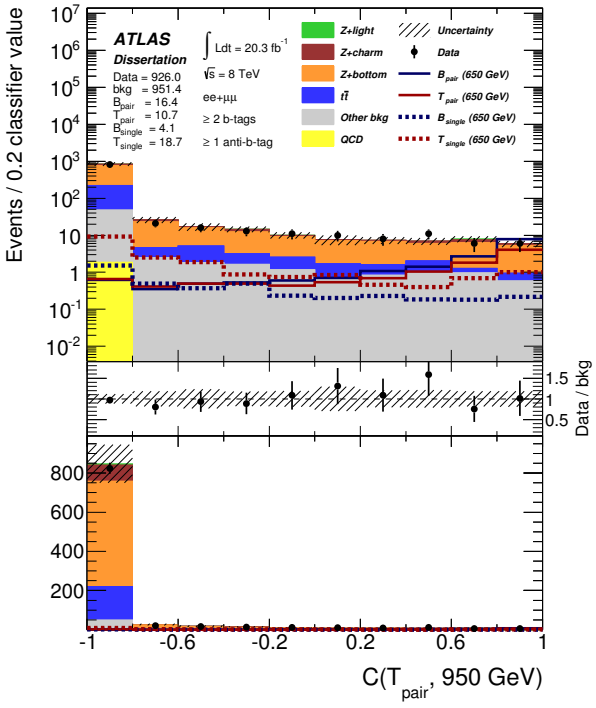


(44)

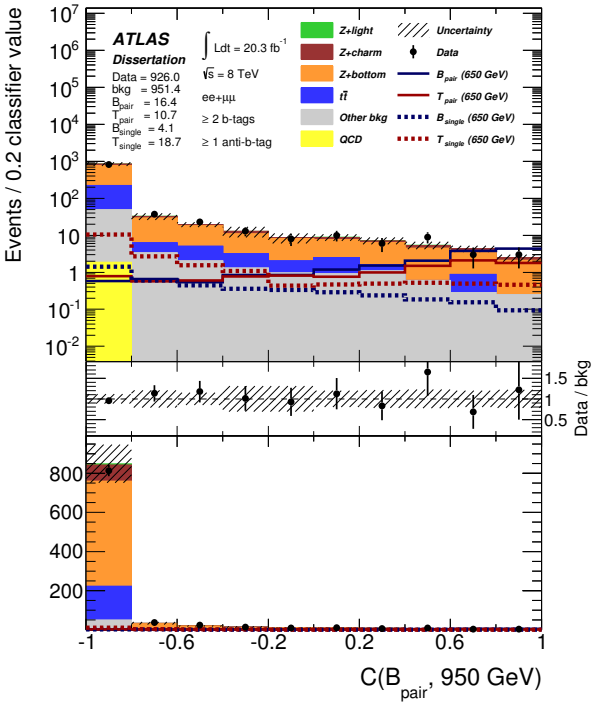


(45)

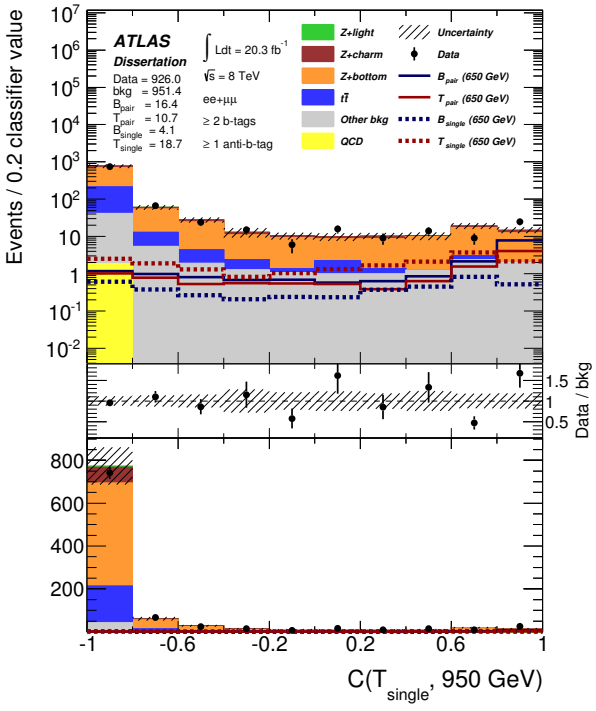
Figure C.5: – *Continued, continues on next 4 pages.*



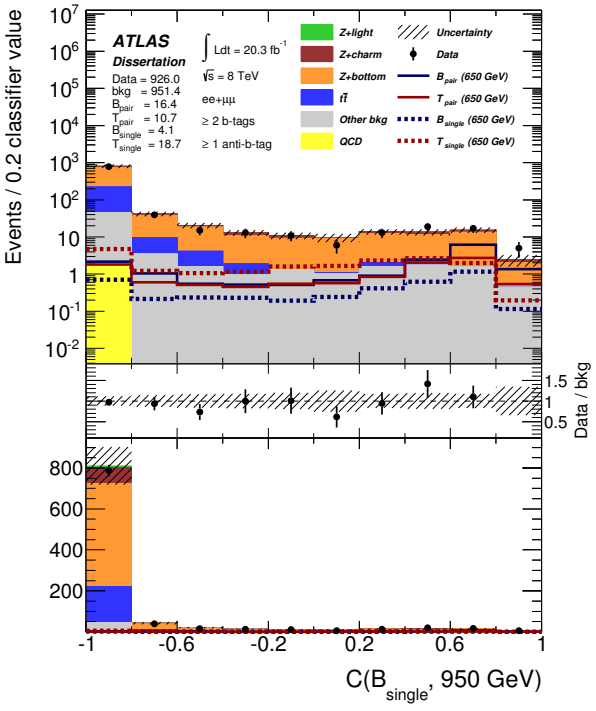
(46)



(47)

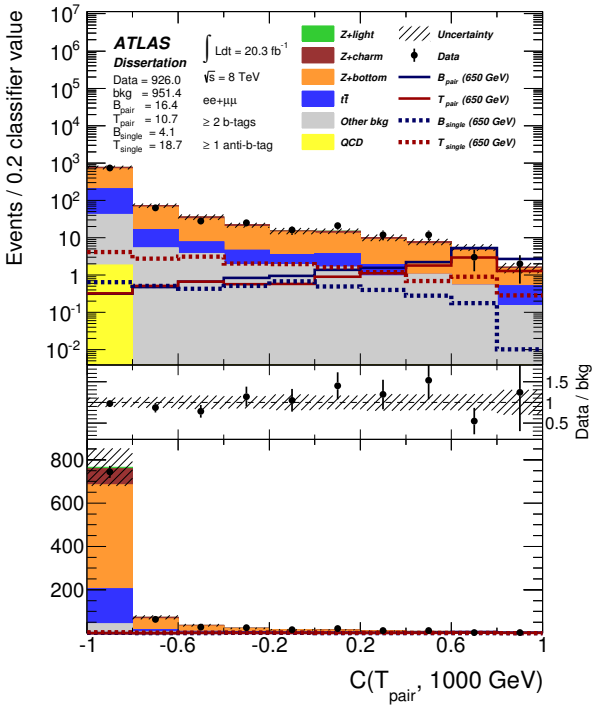


(48)

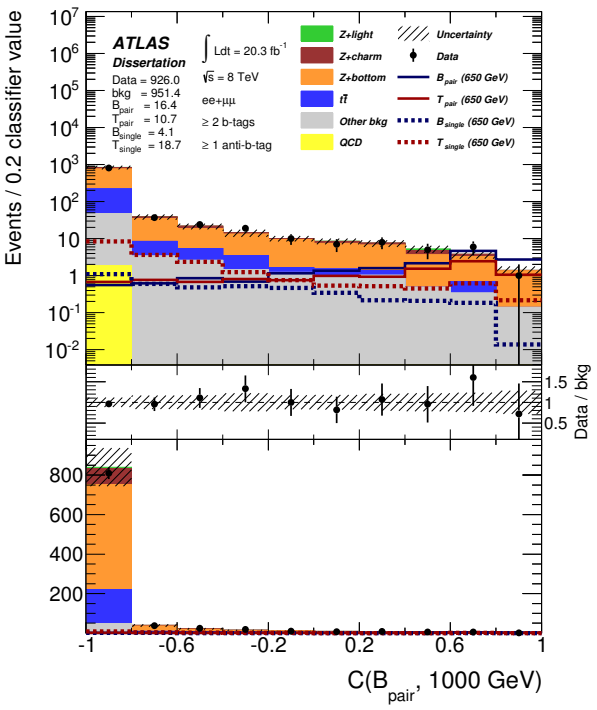


(49)

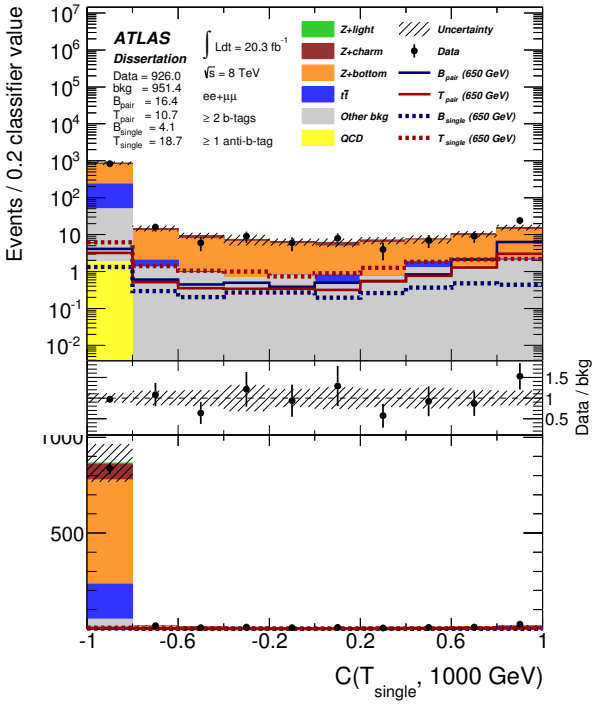
Figure C.5: – *Continued, continues on next 3 pages.*



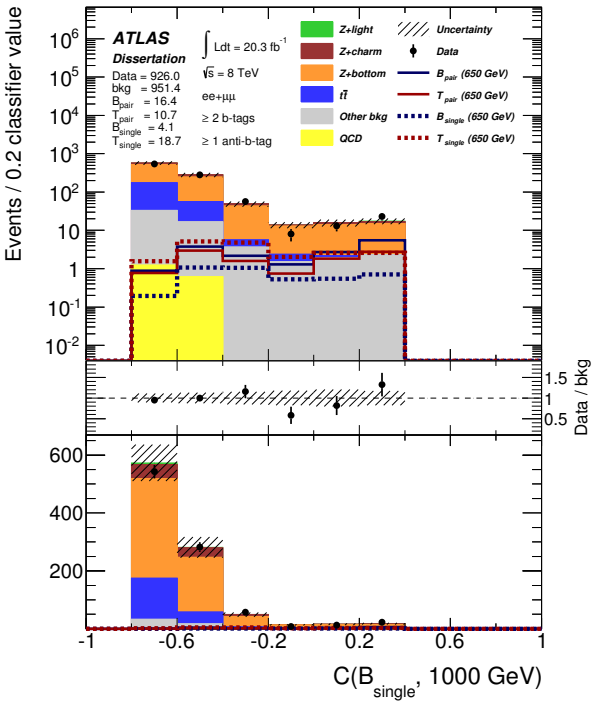
(50)



(51)

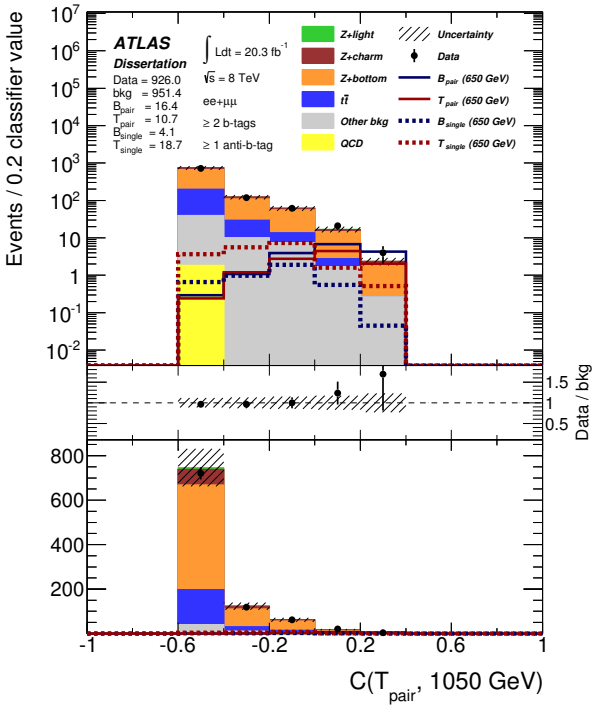


(52)

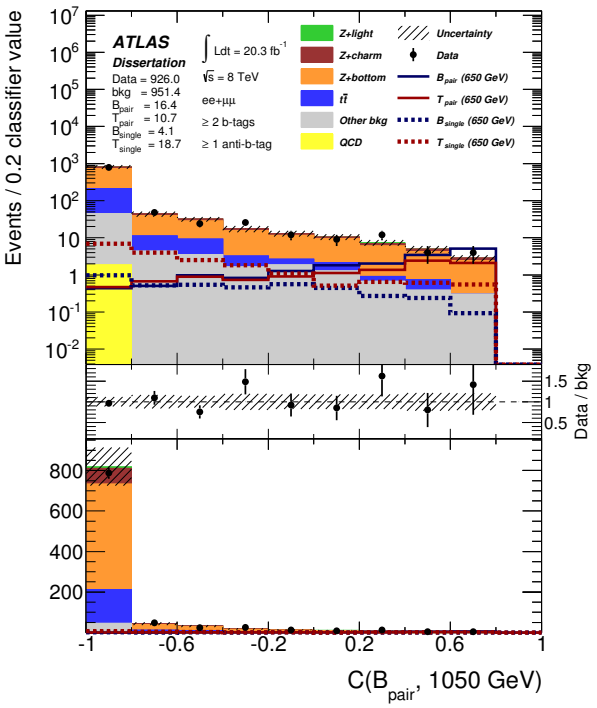


(53)

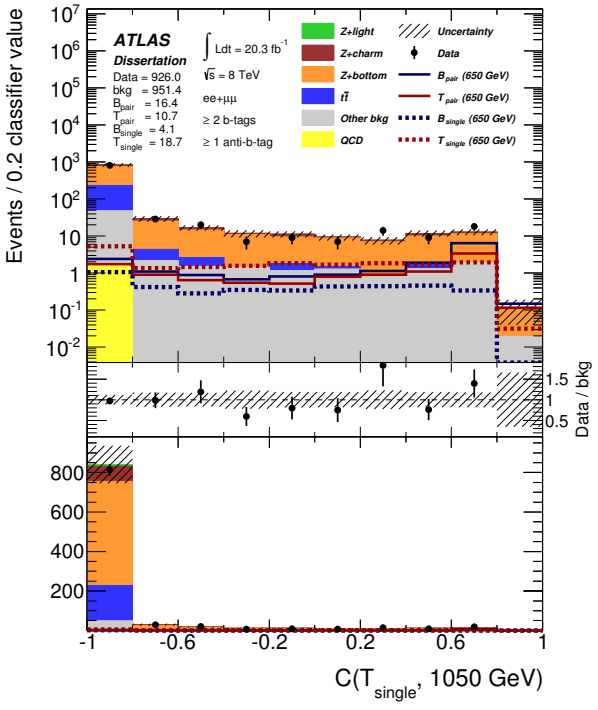
Figure C.5: – *Continued, continues on next 2 pages.*



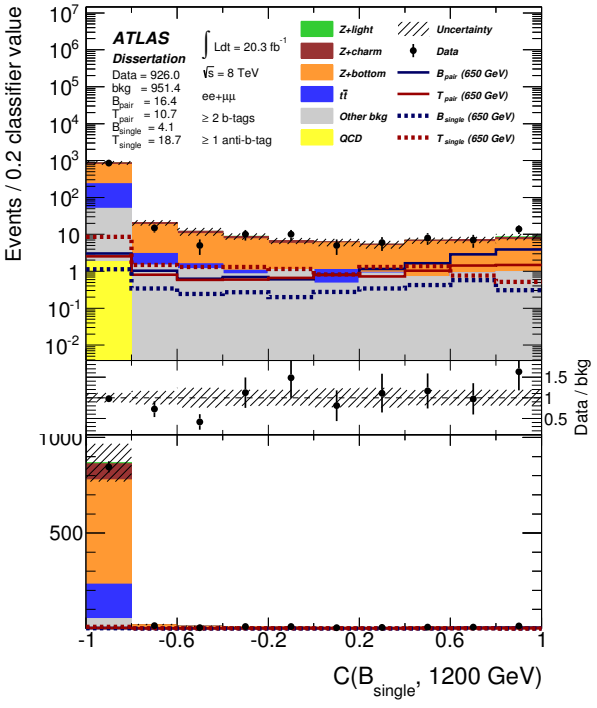
(54)



(55)

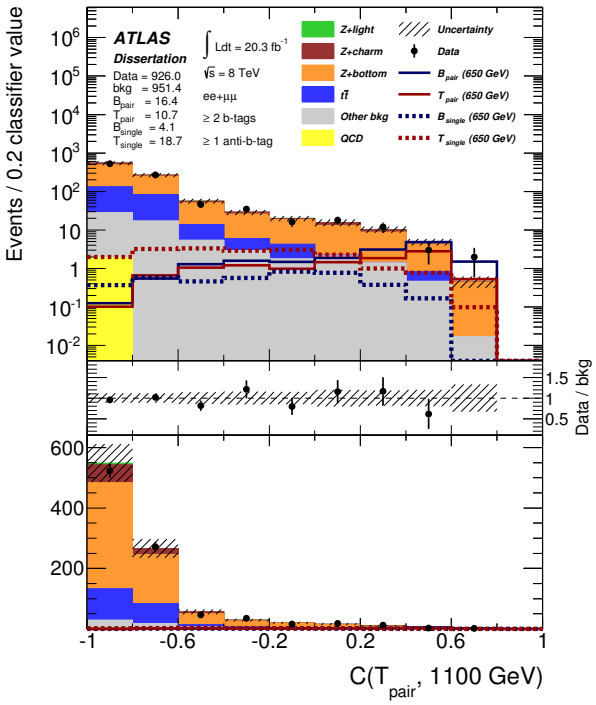


(56)

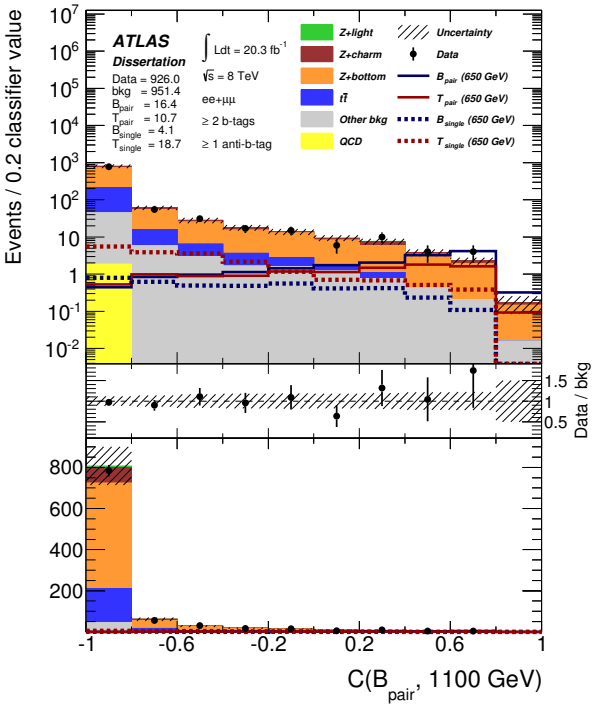


(57)

Figure C.5: – *Continued, continues on next page.*



(58)



(59)

Figure C.5: – *Continued, last page.*

C.4. Final Results – CL_S Upper Limits (Addendum)

For the vector-like quark searches, the Neyman construction of confidence regions was used to compute 95% confidence level upper limits (see section 6.2.2 on page 261). The discriminating variable spectra were modeled with the multivariate Poisson model described in section 6.2.1. The DZero test statistic (section 6.2.3.4 on page 285) and CL_S method (page 289 and section 6.2.4 on page 295) were used in the Neyman construction. Nuisance parameters modeling the systematic uncertainties were treated using the null method described in section 6.5 (page 388).

The 95% CL_S upper limits were computed in five ways for the square cuts searches:

a) ***All systematics, all simulated data, and no profiling.*** — These were the most complete set of limits, using all available information and the null treatment of nuisance parameters discussed in section 6.5 (page 388).
b) ***All systematics, testing partition data, and no profiling.*** — The limits were computed after restricting the search to the same testing partition that machine learning used.
c) ***No systematics, all simulated data, and no profiling.*** — The systematic uncertainties were removed from the limit computations (i.e., the "stat only limits" of particle searches).
d) ***No systematics, testing partition data, and no profiling.*** — Both the systematic uncertainties were removed, and the search was restricted to the machine learning testing partition.
e) ***All systematics, all simulated data, and using profiling.*** — These limits were also the most complete, but implemented the profiling treatment of nuisance parameters discussed in section 6.5 (page 388). In nearly all cases, these upper limits were identical to those obtained using the null treatment of a).[1]

The 95% CL_S upper limits were computed in three ways for the machine learning searches:

a) ***All systematics, testing partition data, and no profiling.*** — These were the most complete set of limits, using all available information and the null treatment of nuisance parameters.
b) ***No systematics, testing partition data, and no profiling.*** — The systematic uncertainties were removed from the limit computations (i.e., the "stat only limits" of particle searches).
c) ***All systematics, testing partition data, and using profiling.*** — These limits were also the most complete, but implemented the profiling treatment of nuisance parameters. In nearly all cases, these upper limits were identical to those obtained using the null treatment of a).

[1] One study has found that upper limits computed both with and without profiling were observed to be nearly identical [154].

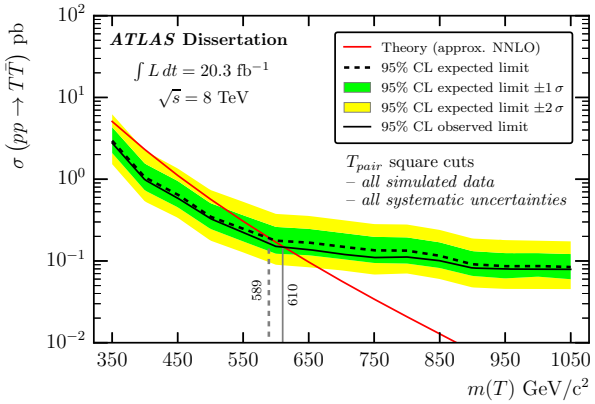
Table 5.18 that shows the four search strategies is repeated here for convenience.

	search strategy	=	signal region	+	discriminating variable
1)	traditional square cuts	=	square cuts	+	reconstructed $m(Zb)$
2)	two-stage machine learning	=	machine learning	+	reconstructed $m(Zb)$
3)*	two-stage machine learning	=	machine learning	+	reconstructed $m(Zt)$
4)	three-stage machine learning	=	machine learning	+	stage 3 regressand mass

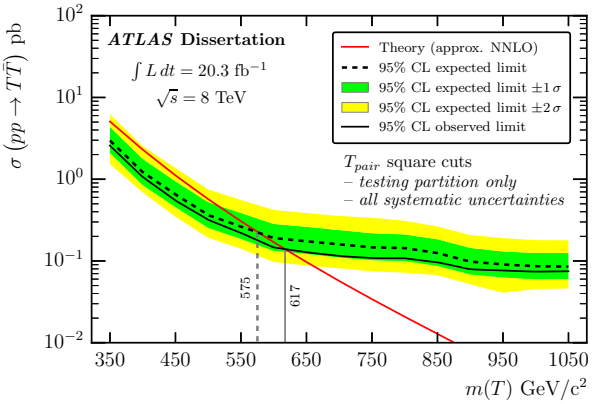
Table 5.18: The four search strategies. *(Repeated from page 204.)*

**Only used for the T_{single} signal type.*

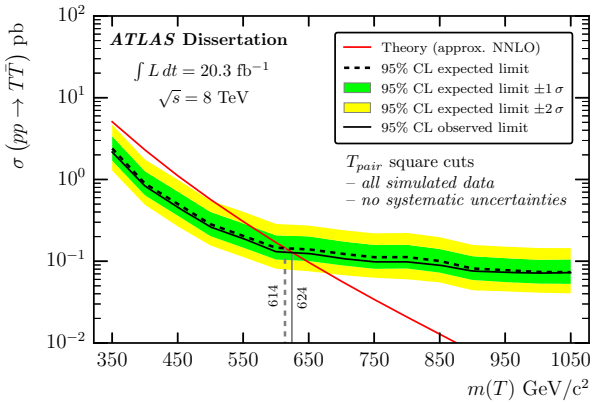
The figures showing the 95% CL_S upper limits are ordered as follows:
- (Page 514) — Figure C.6: T_{pair} square cuts.
- (Page 515) — Figure C.7: T_{pair} two-stage machine learning.
- (Page 516) — Figure C.8: T_{pair} three-stage machine learning.
- (Page 517) — Figure C.9: B_{pair} square cuts.
- (Page 518) — Figure C.10: B_{pair} two-stage machine learning.
- (Page 519) — Figure C.11: B_{pair} three-stage machine learning.
- (Page 520) — Figure C.12: T_{single} square cuts.
- (Page 521) — Figure C.13: T_{single} two-stage machine learning (using $m(Zb)$).
- (Page 522) — Figure C.14: T_{single} two-stage machine learning (using $m(Zt)$).
- (Page 523) — Figure C.15: T_{single} three-stage machine learning.
- (Page 524) — Figure C.16: B_{single} square cuts.
- (Page 525) — Figure C.17: B_{single} two-stage machine learning.
- (Page 526) — Figure C.18: B_{single} three-stage machine learning.



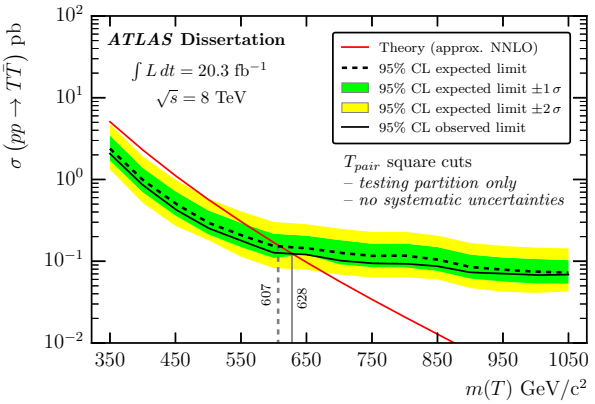
(a) all syst., all data, no profiling



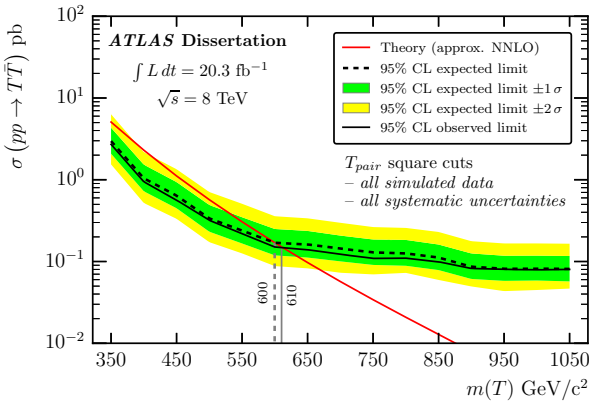
(b) all syst., test data, no profiling



(c) no syst., all data, no profiling

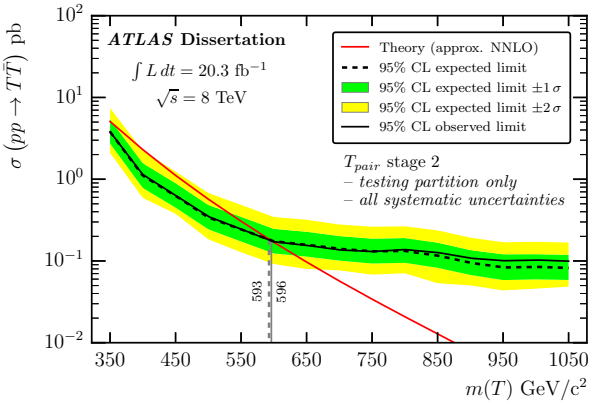


(d) no syst., test data, no profiling

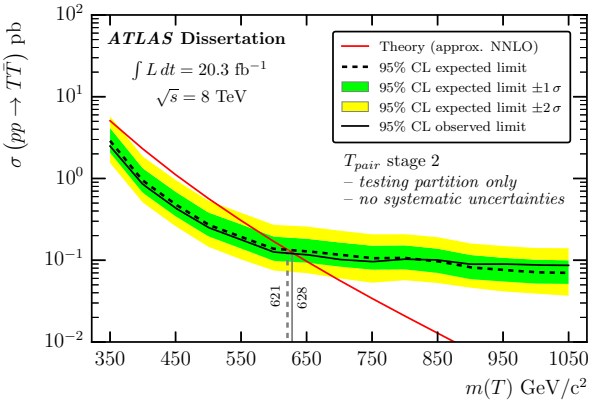


(e) all syst., all data, with profiling

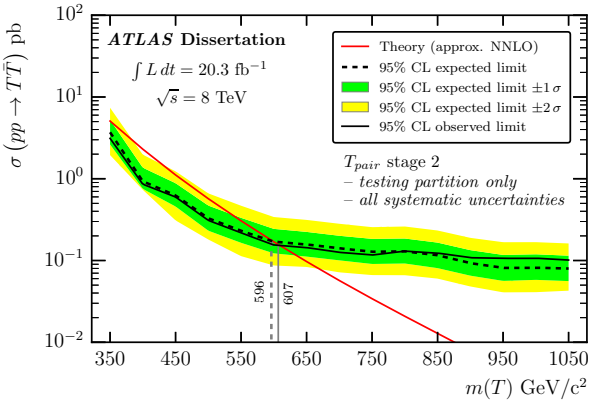
Figure C.6: The 95% CL_S upper limits of the T_{pair} square cuts search strategy = square cuts signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

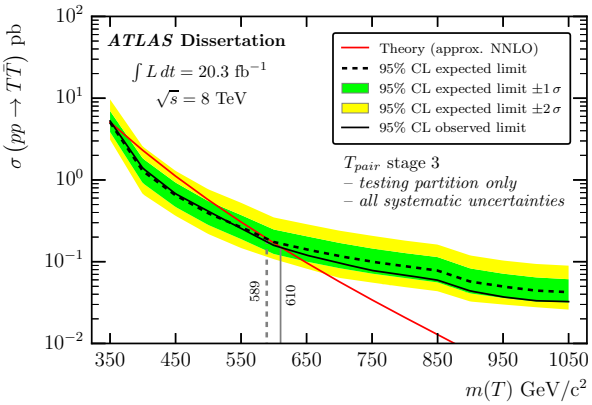


(b) no syst., test data, no profiling

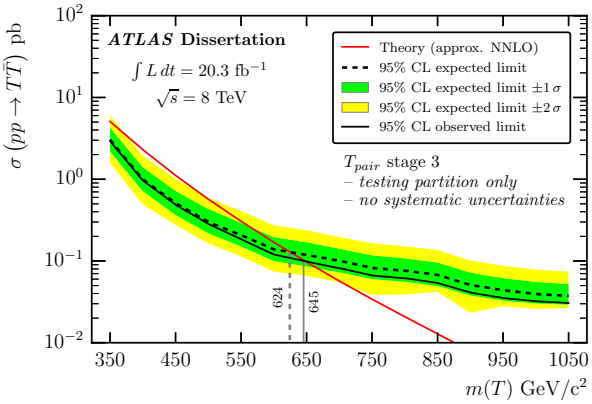


(c) all syst., test data, with profiling

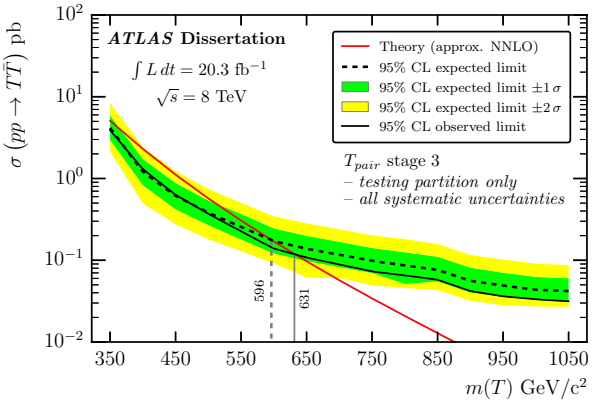
Figure C.7: The 95% CL_S upper limits of the T_{pair} stage 2 machine learning search strategy = machine learning signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

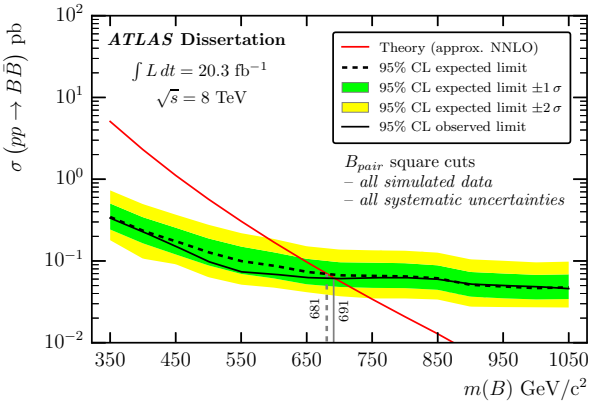


(b) no syst., test data, no profiling

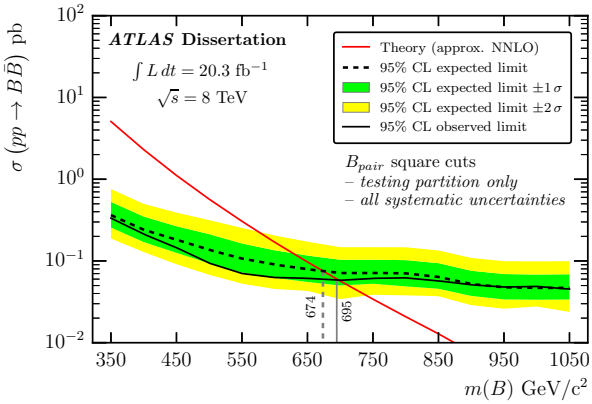


(c) all syst., test data, with profiling

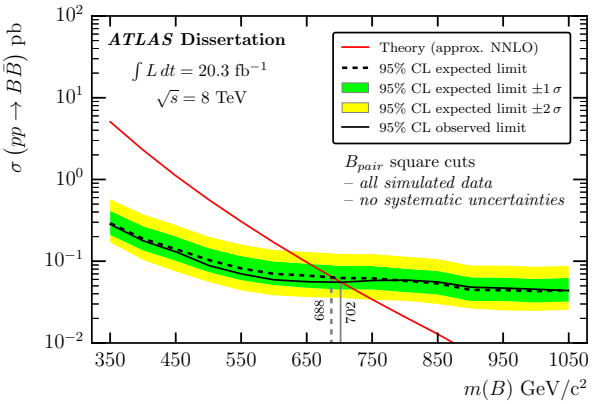
Figure C.8: The 95% CL_S upper limits of the T_{pair} stage 3 machine learning search strategy = machine learning signal region + stage 3 regressand discriminating variable.



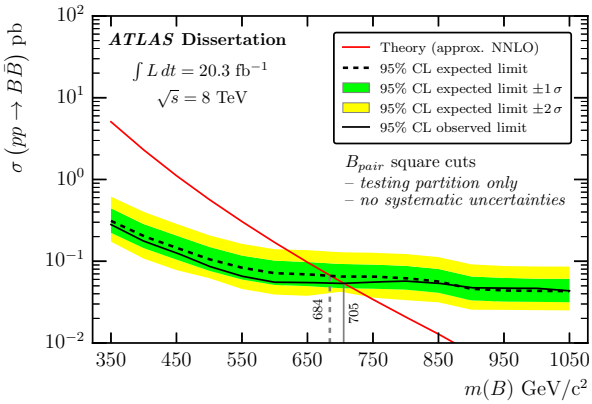
(a) all syst., all data, no profiling



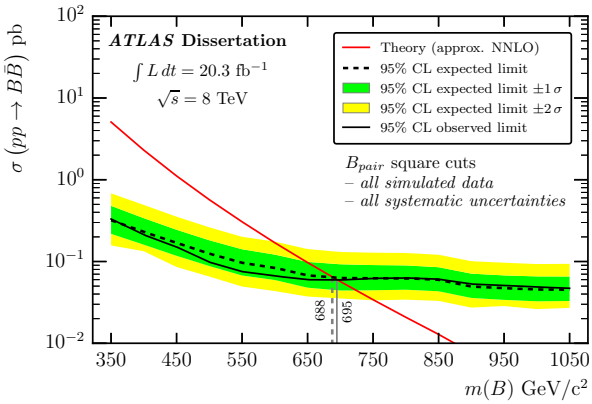
(b) all syst., test data, no profiling



(c) no syst., all data, no profiling

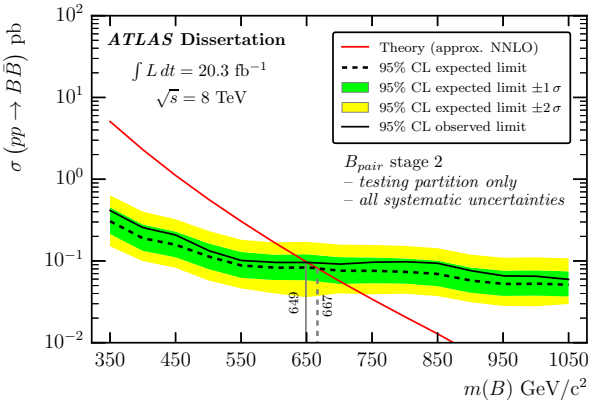


(d) no syst., test data, no profiling

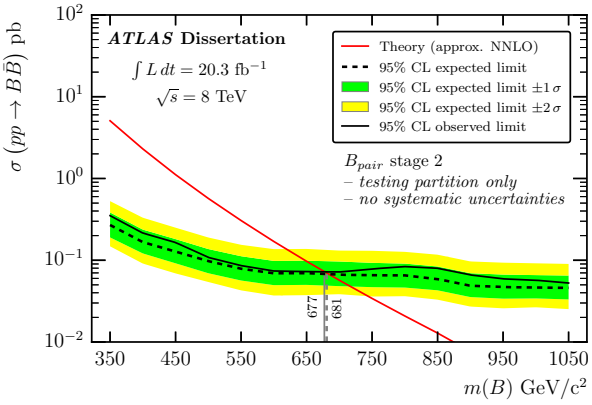


(e) all syst., all data, with profiling

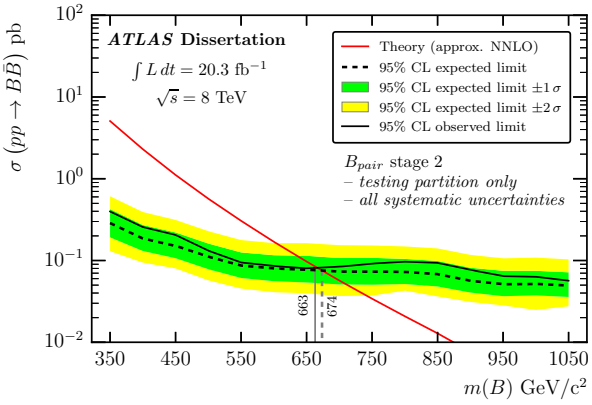
Figure C.9: The 95% CL_S upper limits of the B_{pair} square cuts search strategy = square cuts signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

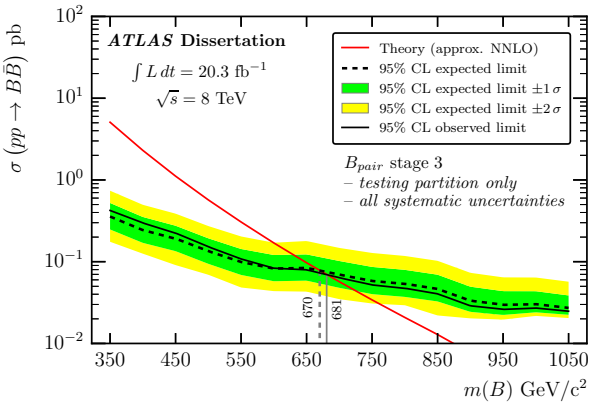


(b) no syst., test data, no profiling

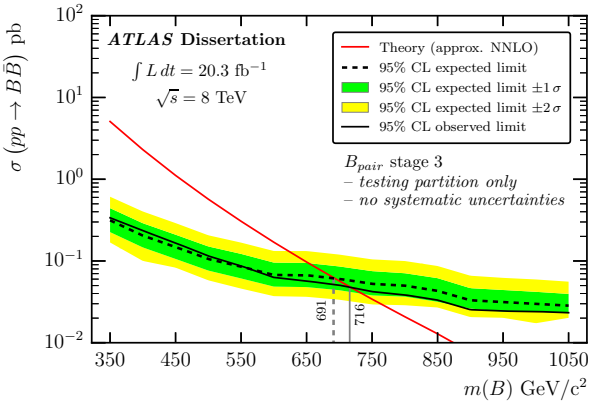


(c) all syst., test data, with profiling

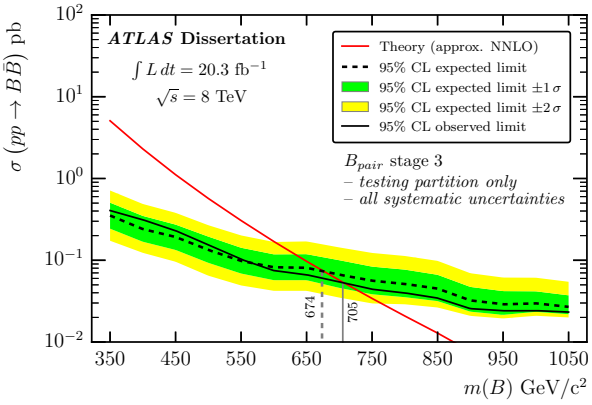
Figure C.10: The 95% CL_S upper limits of the B_{pair} stage 2 machine learning search strategy = machine learning signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

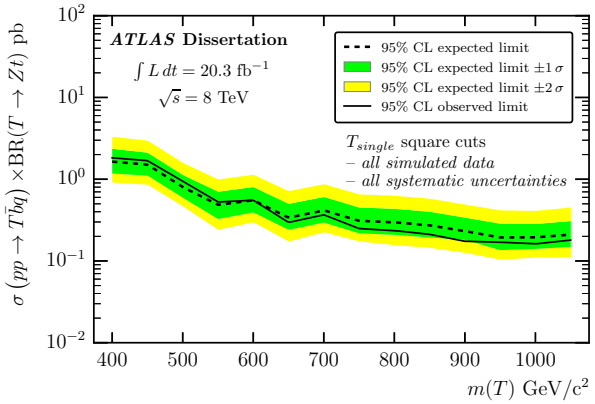


(b) no syst., test data, no profiling

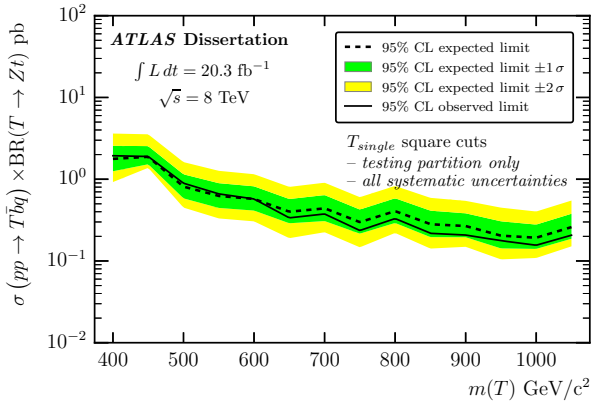


(c) all syst., all data, with profiling

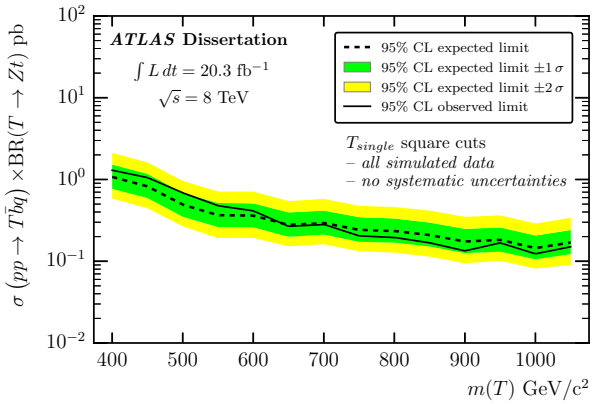
Figure C.11: The 95% CL_S upper limits of the B_{pair} stage 3 machine learning search strategy = machine learning signal region + stage 3 regressand discriminating variable.



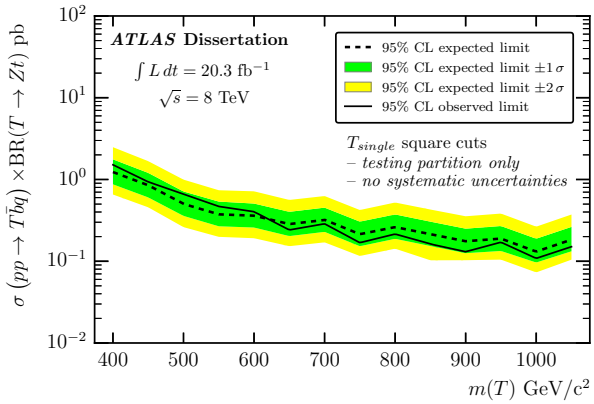
(a) all syst., all data, no profiling



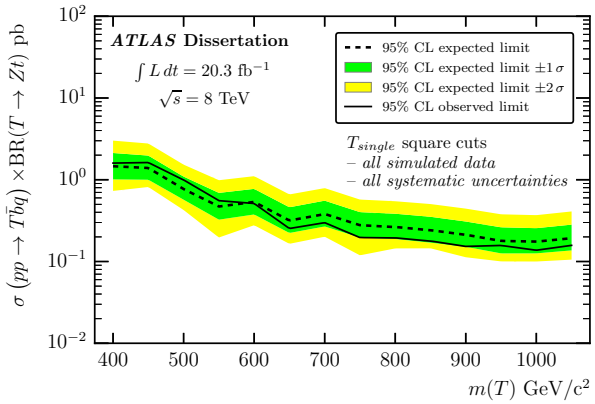
(b) all syst., test data, no profiling



(c) no syst., all data, no profiling

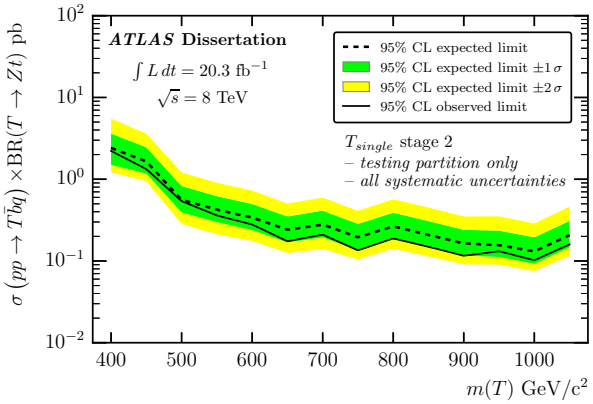


(d) no syst., test data, no profiling

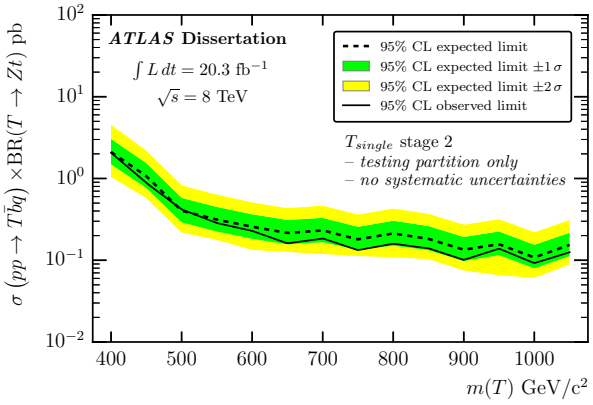


(e) all syst., all data, with profiling

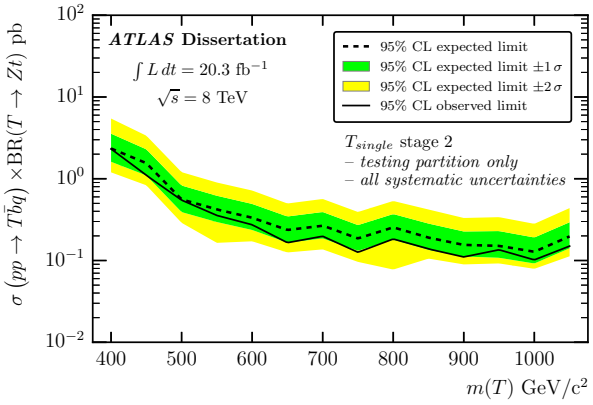
Figure C.12: The 95% CL_S upper limits of the T_{single} square cuts search strategy = square cuts signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

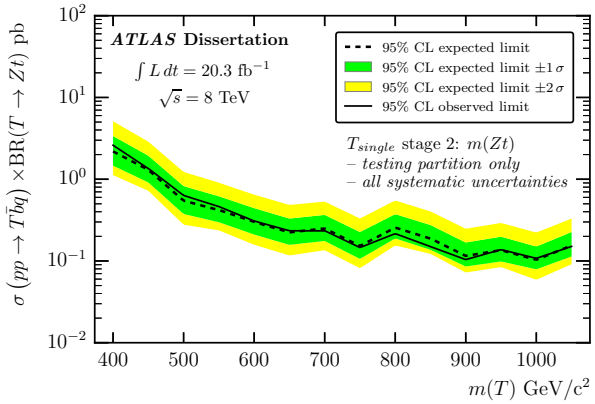


(b) no syst., test data, no profiling

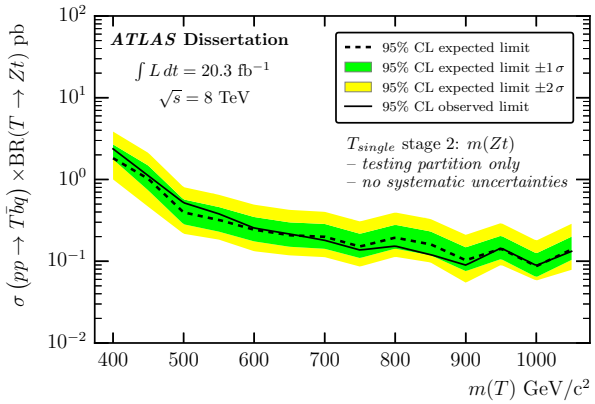


(c) all syst., test data, with profiling

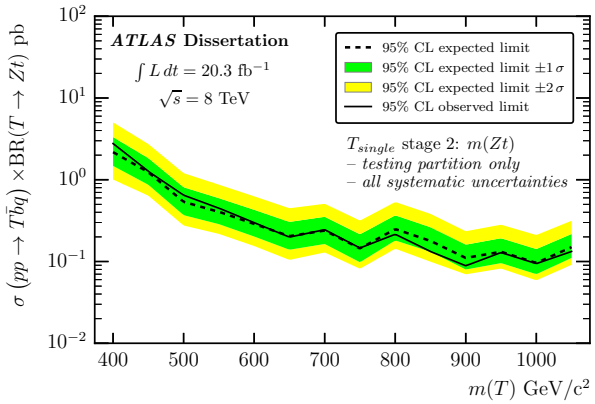
Figure C.13: The 95% CL_S upper limits of the T_{single} stage 2 machine learning search strategy = machine learning signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

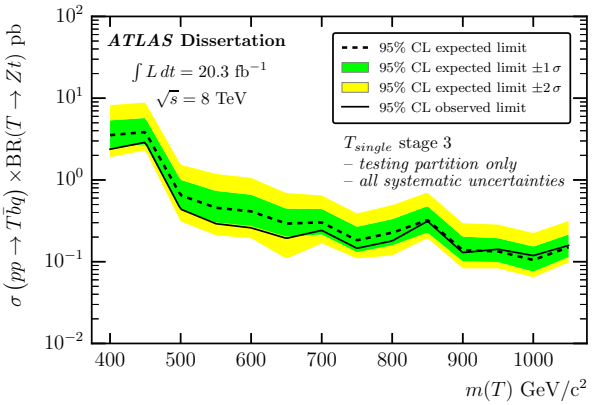


(b) no syst., test data, no profiling

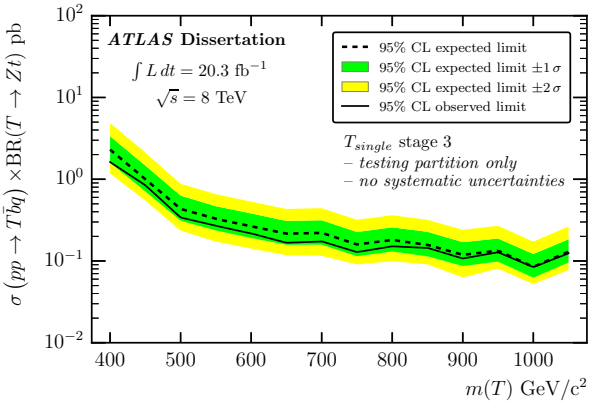


(c) all syst., test data, with profiling

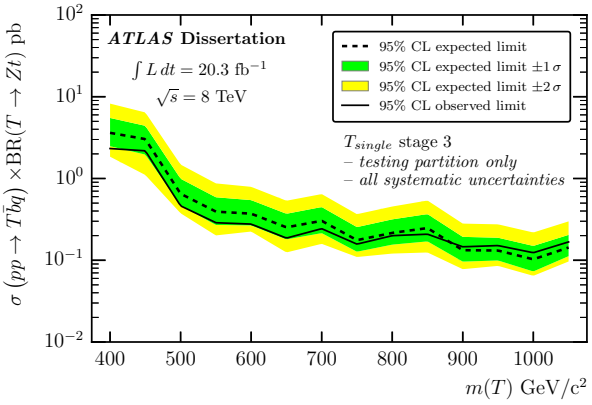
Figure C.14: The 95% CL_S upper limits of the T_{single} stage 2 machine learning search strategy = machine learning signal region + $m(Zt)$ discriminating variable.



(a) all syst., test data, no profiling

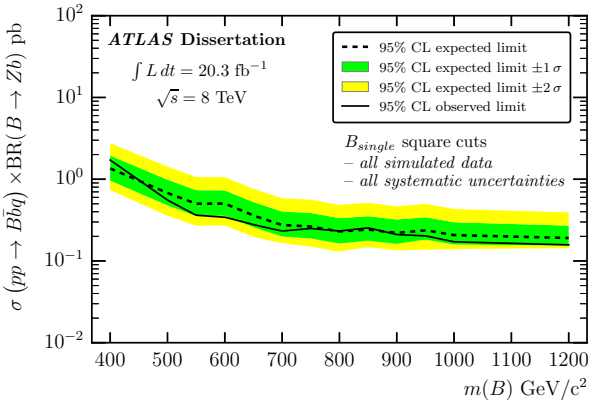


(b) no syst., test data, no profiling

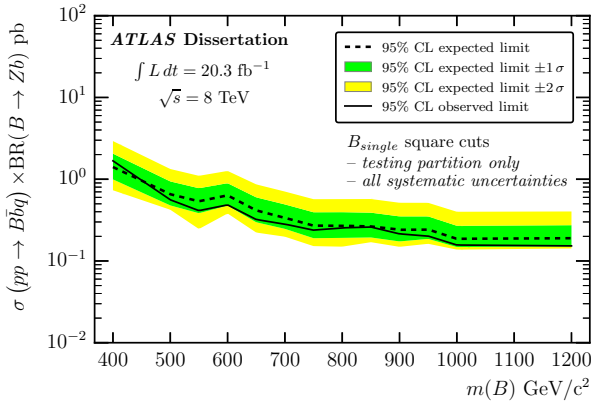


(c) all syst., test data, with profiling

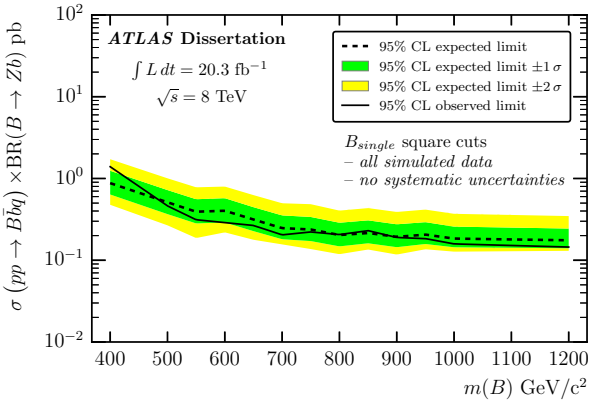
Figure C.15: The 95% CL_S upper limits of the T_{single} stage 3 machine learning search strategy = machine learning signal region + stage 3 regressand discriminating variable.



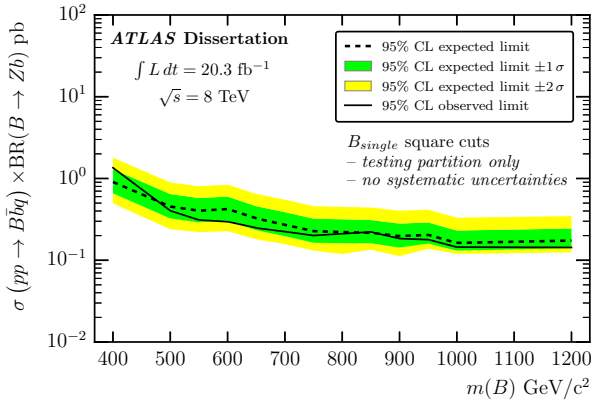
(a) all syst., all data, no profiling



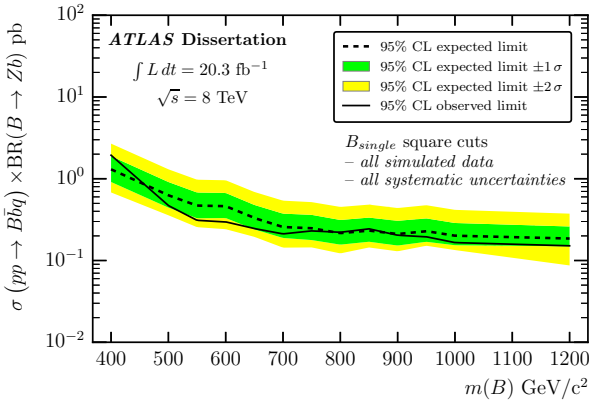
(b) all syst., test data, no profiling



(c) no syst., all data, no profiling

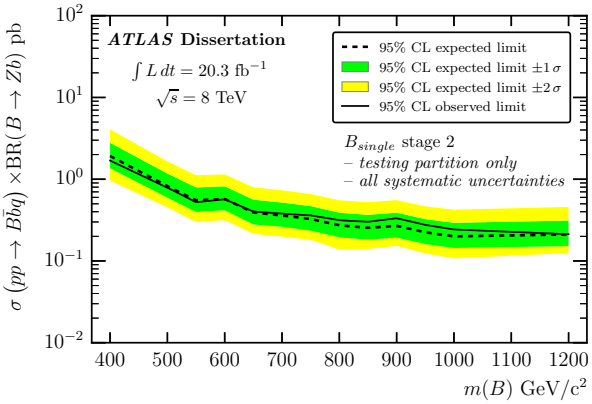


(d) no syst., test data, no profiling

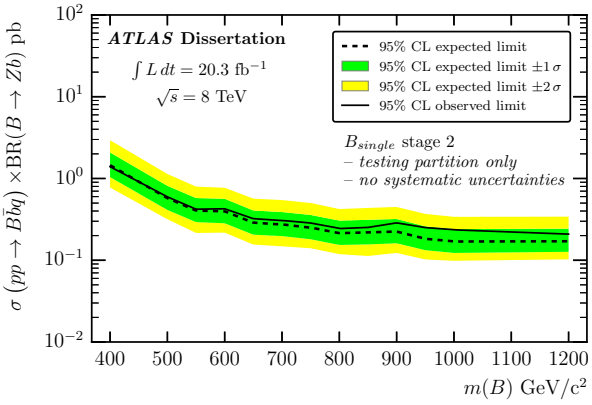


(e) all syst., all data, with profiling

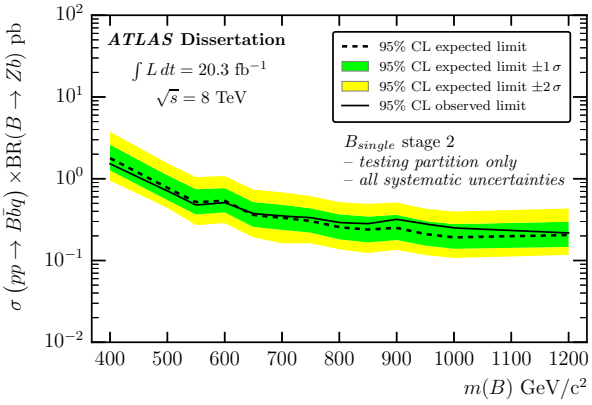
Figure C.16: The 95% CL_S upper limits of the B_{single} square cuts search strategy = square cuts signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling

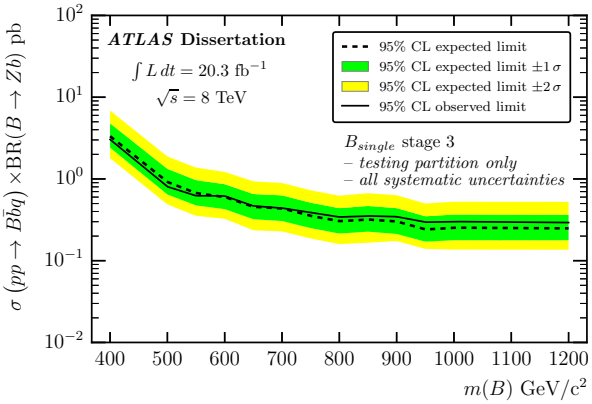


(b) no syst., test data, no profiling

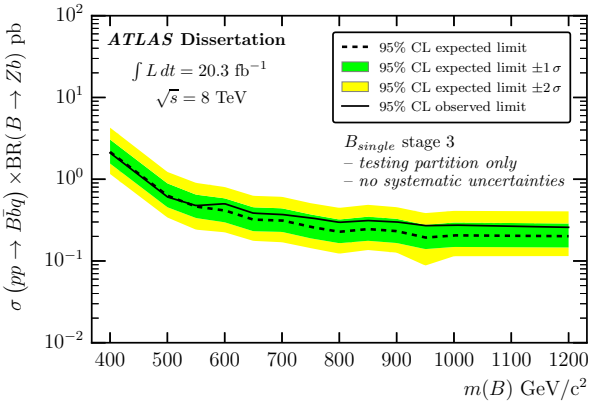


(c) all syst., test data, with profiling

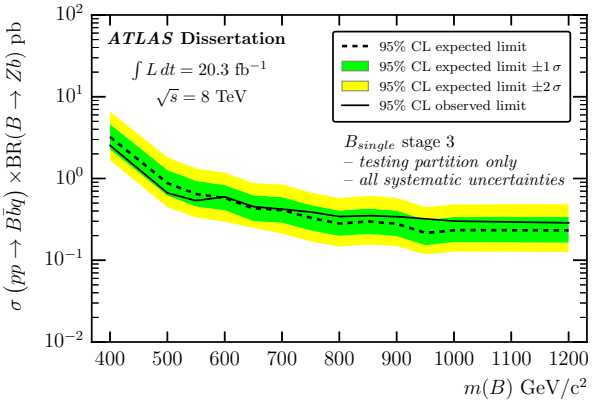
Figure C.17: The 95% CL_S upper limits of the B_{single} stage 2 machine learning search strategy = machine learning signal region + $m(Zb)$ discriminating variable.



(a) all syst., test data, no profiling



(b) no syst., test data, no profiling



(c) all syst., test data, with profiling

Figure C.18: The 95% CL_S upper limits of the B_{single} stage 3 machine learning search strategy = machine learning signal region + stage 3 regressand discriminating variable.

Appendix D
The Neyman Construction in Signal Searches (Addendum)

This appendix chapter supports some sections of chapter 6.

- Section D.1 supports 6.3: *The Arizona Test Statistics.*
- Section D.2 supports 6.4: *Inference using Confidence-Acceptance Profiles.*
- Section D.3 supports 6.5: *Test Statistics in the Presence of Nuisance Parameters.*

D.1. The Arizona Test Statistics (Addendum)

The following subsections support those in section 6.3 with

- subsection D.1.1 providing a deeper discussion of the inconsistency function, Q_{inc}, and
- subsection D.1.2 some details regarding the selection of the reporting risk function, Q_{rep}.

D.1.1. Incorporating the Model Inconsistency Measure, Q_{inc}

The ultimate goal is to modify the test statistic to better meet the experimenter's best judgment regarding the two dimensional criteria depicted in figure 6.18 on page 300. There are an infinite number of parameter values μ, where each requires determining an acceptance region of possible outcomes d, and there are an infinite number of possible outcomes d in the sample space that must be ordered for each selection of μ. Clearly the experimenter cannot order the outcomes by hand and must determine a local ordering rule that aligns with their best judgment as to the kind of evidence supporting different values of μ that is supplied by each outcome.

When performing machine learning, one has a multitude of input variables (features) from which to choose. For parametric inference (such as the Neyman construction) in a signal search using the multivariate Poisson model, one merely has the properties of the likelihood functions to use in discriminating between the outcomes (ordering them). Therefore, the experimenter is generally limited to a few key features of the likelihood function in determining the local ordering rule that aligns with their best judgment. This includes different types of maximum likelihood values (such as the unrestricted MLE, $\tilde{\mu}_u$), or the likelihood function at a particular value ($\mu = 0$ for the background-only model [1]). The width of the likelihood function was also found to be a useful "feature" in expressing the experimenter's best judgment, but those results are not reported here [2].

There are several options for constructing test statistics using likelihood ratios, depending on the types of MLEs chosen. The MLE types that were examined are summarized here.

- $\tilde{\mu}_u$ – the unrestricted MLE that is the maximum of the likelihood function, $\arg\max_\mu[p(d;\,\mu)]$.
- $\tilde{\mu}_p$ – the physical MLE given by $\tilde{\mu}_p = \max[0, \tilde{\mu}_u]$.
- $\tilde{\mu}_b$ – the (upper) bounded MLE given by $\tilde{\mu}_b = \min[\mu', \tilde{\mu}_p]$, where μ' is the specific value for which the acceptance region is being determined during the Neyman construction.
- $\tilde{\mu}_g$ – the generalized MLE given by $\tilde{\mu}_g = \max[\mu', \tilde{\mu}_p]$. While $\tilde{\mu}_b$ is bounded above by μ', $\tilde{\mu}_g$ is bounded below by μ'.
- $\tilde{\mu}_0$ – the background-only model, where $\tilde{\mu}_0 = 0$ when nuisance parameters are absent, and is not actually an MLE.

Figures D.1 (a)–(f) show the values of various MLEs (using the bivariate Poisson model

[1] The background-only model is usually written as $\mu = \tilde{\mu}_0$ to emphasize differences regarding nuisance parameters.

[2] The likelihood width is useful for identifying inconsistent outcomes that have *positive* unrestricted MLE, $\tilde{\mu}_u > 0$. These are outcomes with significant "shape" deviations in their bin counts from those expected with either the signal or background-only models. However, further investigation is needed to determine how to incorporate likelihood width into the test statistic modifications when nuisance parameters are present in the model. Likelihood function width was discussed in section 6.2.2.2 (and figures 6.6 (a)–(f)).

of example 2) which may be by used to modify a test statistic by multiplication or division. These MLEs are shown for three selected values of μ in order to view how they behave over a range of μ: $\mu' = 0.1$ (figures (a) and (d)), $\mu' = 0.5$ (figures (b) and (e)), and $\mu' = 1.0$ (figures (c) and (f)). Generally, the MLEs have the same form for different μ, with increasing μ vertically "stretching" the MLE values over a wider range.

The goal here is to take any of the three current test statistics, Q_{FC}, Q_{D0} or Q_{AT}, and multiply or divide them by one or more of these other MLEs in order to achieve the desired test statistic properties. Each of the current test statistics is already a simple ratio of likelihoods (such as $Q_{FC}(d,\mu) = p(d;\,\mu)/p(d;\,\tilde{\mu}_p)$) so that modification is equivalent to adjusting either the numerator or denominator of the original likelihood ratio test statistic.

Two possible numerators for the simple likelihood ratio are examined. The first is the simple likelihood function, $p(d;\,\mu')$, which for hypothesis testing corresponds to null hypothesis $H_N : \mu = \mu'$ (figures (a)–(c) on the left side). The second is a generalized likelihood function, $p(d;\,\tilde{\mu}_g)$, which for hypothesis testing corresponds to $H_N : \mu \geq \mu'$ (figures (d)–(f) on the right side). Four possible denominators for the simple likelihood ratio are investigated. These are the unrestricted likelihood $p(d;\,\tilde{\mu}_u)$ for $H_A : \mu \neq \mu'$, the physical likelihood $p(d;\,\tilde{\mu}_p)$ for $H_A : \mu \neq \mu'$ with $0 < \mu$, the bounded likelihood $p(d;\,\tilde{\mu}_b)$ for $H_A : 0 < \mu < \mu'$, and the background-only model $p(d;\,\tilde{\mu}_0)$ for $H_A : \mu \neq \tilde{\mu}_0$.

Background-only denominator — First examine the ratios $Q_{D0} = p(d;\,\mu')/p(d;\,\tilde{\mu}_0)$ (left hand figures) and $p(d;\,\tilde{\mu}_g)/p(d;\,\tilde{\mu}_0)$ (right hand figures), which is shown with dark green vertical hashes. These ratios are nearly always monotonically increasing in possible outcomes from left to right using the MLE global order. This is true regardless of the value of μ. This indicates that during the Neyman construction, using $p(d;\,\tilde{\mu}_0)$ as the denominator orders the outcomes from left to right so that those on the left are rejected first from the acceptance region. This results in one-sided upper limits once completed for all values of μ. This explains the structure of the confidence intervals shown previously in figures 6.12 (a) and (b).

Physical denominator — Next examine the ratio $Q_{FC} = p(d;\,\mu')/p(d;\,\tilde{\mu}_p)$ (left hand figures), which is shown with red crosses. The lowest ratio values occur with outcomes d further to the right. The ratio is concave with a maximum near d^b_{mode} for small μ (figure (a)), but shifting to the right with increasing μ (figures (b) and (c)). This indicates that the acceptance regions for small μ are primarily formed from outcomes on the left side for small μ, but shifts away from the leftmost outcomes as μ grows. This explains the structure of the two-sided confidence intervals shown previously in figure 6.14.

It is apparent that one may use the properties of the simple likelihood ratios shown in figures D.1 (a)–(f) to predict the structure of the confidence intervals that would result from their use in Neyman construction. Examine any use of $p(d;\,\tilde{\mu}_g)$ as the numerator (right hand figures). Regardless of denominator, these are nearly always monotonically increasing, which results in one-sided upper limits. As demonstrated in figure 6.22, modifying upper limit producing test statistics with Q_{inc} creates difficulties in the modified confidence interval structure. For this reason, generalized likelihood ratio test statistics using $p(d;\,\tilde{\mu}_g)$ were not used in the development of Arizona test statistics.

Bounded denominator — Next examine the ratios $Q_{AT} = p(d;\,\mu')/p(d;\,\tilde{\mu}_b)$ (left hand figures), which is shown with black tri-stars. These are weakly monotonically increasing with many outcomes on the right give the same value of 1.0. This explains the structure of the confidence intervals shown previously in figure 6.15.

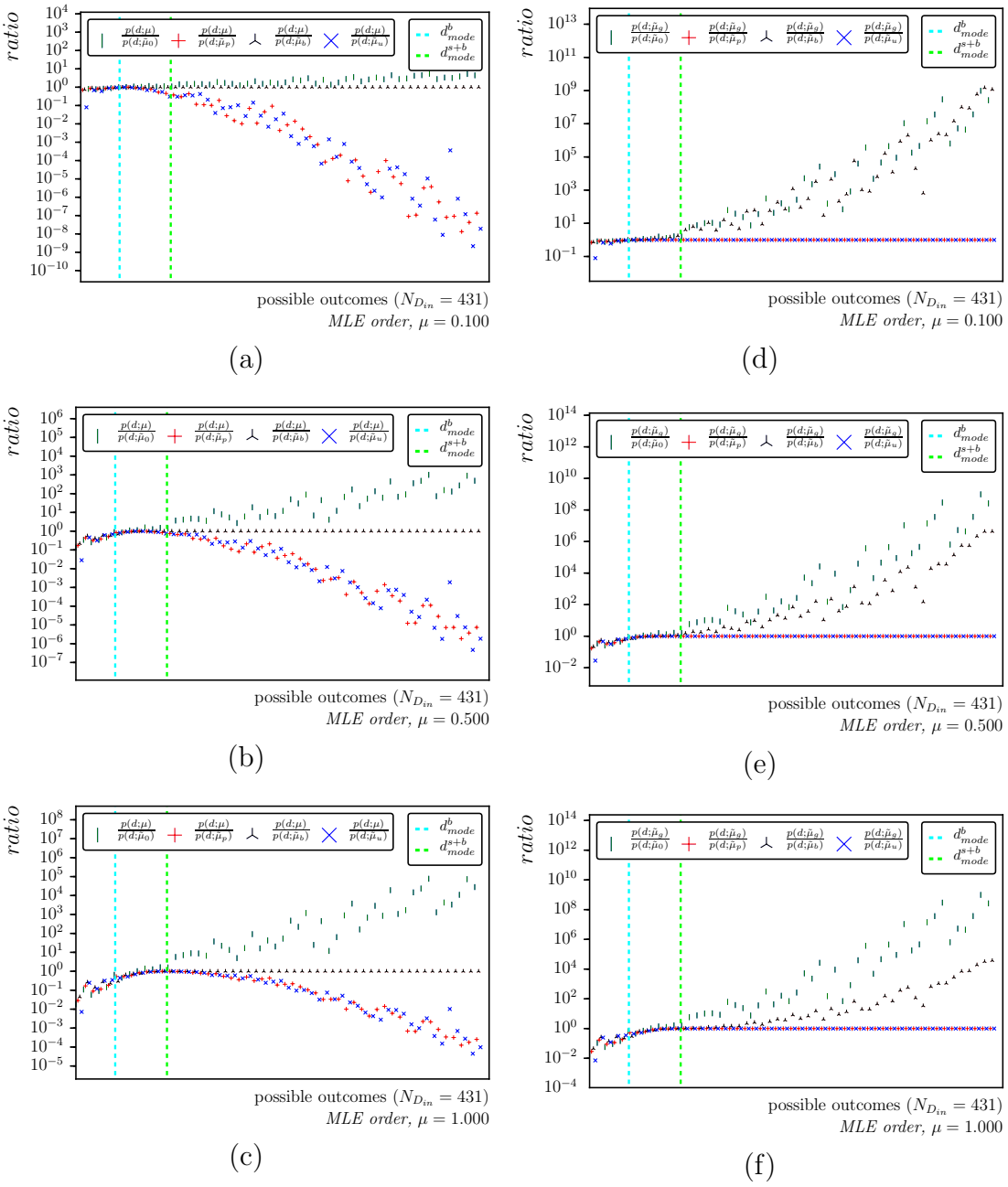


Figure D.1: Example 2 – Properties of some simple likelihood ratios. (a)–(c) Using $p(d; mu)$ for the numerator with (a) $\mu = 0.1$, (b) $\mu = 0.5$, and (c) $\mu = 1.0$. (d)–(f) Using $p(d; \tilde{\mu}_g)$ for the numerator with (d) $\mu = 0.1$, (e) $\mu = 0.5$, and (f) $\mu = 1.0$. Dark green vertical hashes show $p(d; \tilde{\mu}_0)$ for the denominator (D0), red crosses show $p(d; \tilde{\mu}_p)$ (Feldman-Cousins), black tri-stars show $p(d; \tilde{\mu}_b)$ (ATLAS), and blue crosses show $p(d; \tilde{\mu}_u)$. Dashed blue lines give the location of d^b_{mode} and dashed green d^{s+b}_{mode} for reference.

Unrestricted denominator — Finally, examine the ratios $p(d;\,\mu')/p(d;\,\tilde{\mu}_u)$ (left hand figures) shown with blue crosses. These ratios are nearly identical to the physical ratios, Q_{FC}, and produce similar confidence intervals. The only difference between these is smaller values for outcomes furthest to the left for the unrestricted than for Q_{FC}. This means that those outcomes are rejected more strongly for the unrestricted ratio than the for Q_{FC}. Because one goal is to modify a simple likelihood ratio test statistic so that it more weakly rejects outcomes furthest to the left, the unrestricted likelihood $p(d;\,\tilde{\mu}_u)$ was not used as the denominator for the simple likelihood ratios.

The structure of the test statistic determines the ordering of outcomes when forming acceptance regions, as demonstrated in figures D.1 (a)–(f) for example 2. From such plots one can approximately determine the structure of the resultant confidence regions. If the experimenter seeks a test statistic that represents their best judgment, then they should be able to describe what the ideal structure of the test statistic should be, i.e., sketch by hand how it should appear for different values of the parameter of interest. If the experimenter desires confidence intervals that are longer for outcomes that provide ambiguous evidence for both the signal and background-only models, i.e., are inconsistent with both models, then the test statistic should assign a higher value to those inconsistent outcomes. In the case of example 2, this means seeking a test statistic that increases the values for the leftmost outcomes from those shown in figures D.1 (a)–(f).

Because none of the simple likelihood ratios shown in figures D.1 (a)–(f) possess this feature, additional functions are examined in order to find a suitable and well-motivated method for achieving the desired test statistic. Figures D.2 (a)–(f) show the various ratios obtained by pairwise combining $p(d;\,\tilde{\mu}_0)$, $p(d;\,\tilde{\mu}_p)$, $p(d;\,\tilde{\mu}_b)$, and $p(d;\,\tilde{\mu}_u)$. These are shown for the two values of μ', 0.1 and 1.0, which alters the values of $p(d;\,\tilde{\mu}_b)$ due to the dependence of $\tilde{\mu}_b$ on its upper boundary, μ'.

If any of the ratios in figures D.2 (a)–(f) were used as the test statistic, they would result in one and two-sided confidence intervals similar to the common methods, but also not achieving the desired behavior. A quick scan reveals that only one ratio, $p(d;\,\tilde{\mu}_u)/p(d;\,\tilde{\mu}_p)$, has the property of having higher values for outcomes to the left of d^b_{mode} (and increasing leftward) while having a value of 1.0 for all other outcomes, as shown in figures D.1 (a) and (b). This indicates that by combining $p(d;\,\tilde{\mu}_u)/p(d;\,\tilde{\mu}_p)$ with one of the simple likelihood ratios from figures D.1 (a)–(f), one may possibly obtain the desired behavior.

Furthermore, the ratio $p(d;\,\tilde{\mu}_u)/p(d;\,\tilde{\mu}_p)$ has a useful interpretation. When $p(d;\,\tilde{\mu}_u \geq 0$, the ratio is 1. This occurs for outcomes on the right and implies that the signal model for *some* signal strength is consistent with the outcome. When $p(d;\,\tilde{\mu}_u < 0$, $p(d;\,\tilde{\mu}_p) = p(d;\,\mu = 0)$ so that the ratio is the unconstrained MLE probability to the probability of the background-only model. In other words, the larger the ratio is, the less consistent the outcome is with the background-only model, as well as the signal model.

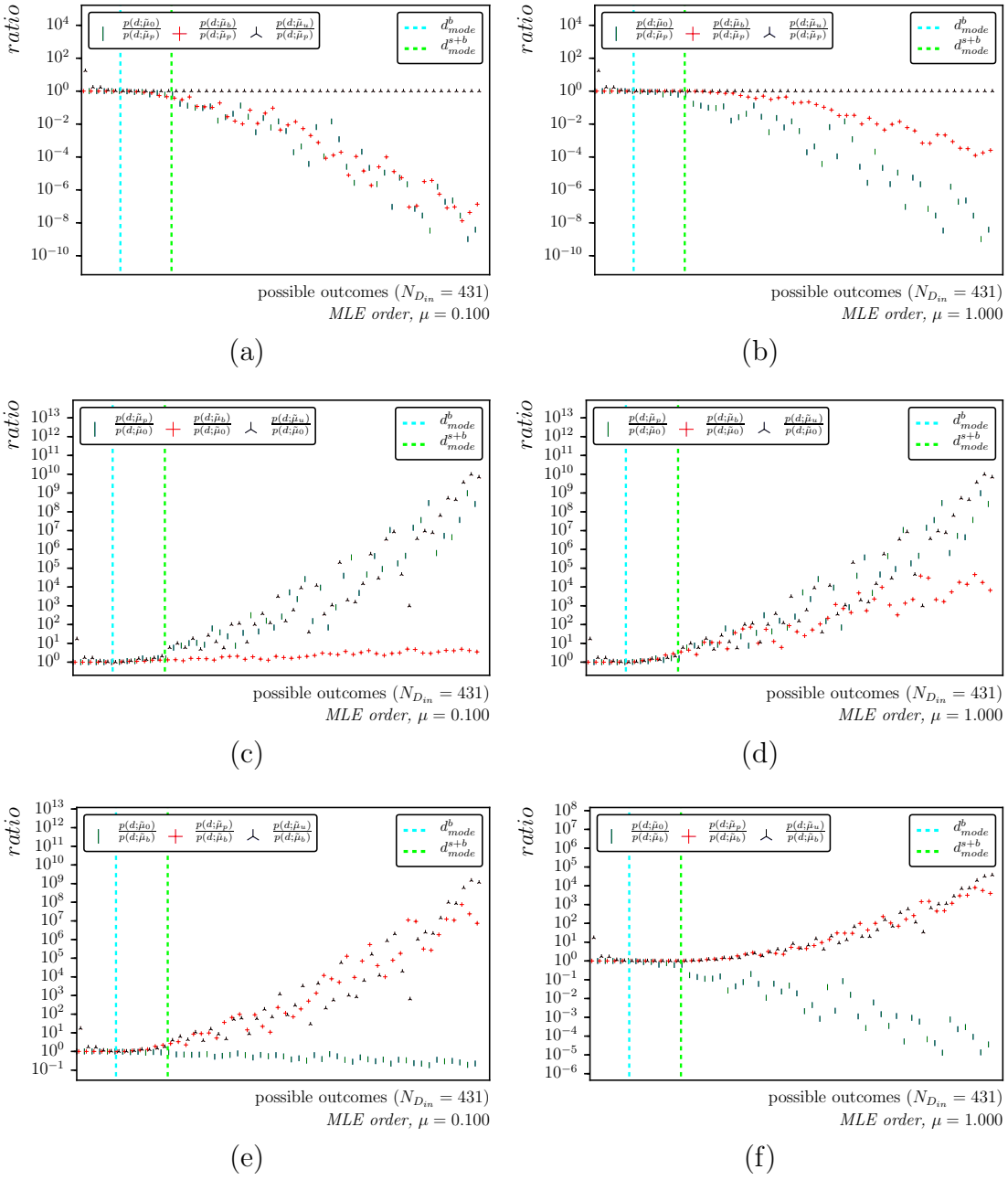


Figure D.2: Example 2 – Properties of some other likelihood ratios. (a)–(b) Using $p(d; \tilde{\mu}_0)$ for the denominator with (a) $\mu = 0.1$, (b) $\mu = 1.0$. (c)–(d) Using $p(d; \tilde{\mu}_p)$ for the denominator with (c) $\mu = 0.1$, (d) $\mu = 1.0$. (e)–(f) Using $p(d; \tilde{\mu}_b)$ for the denominator with (e) $\mu = 0.1$, (f) $\mu = 1.0$. One may obtain the plot of a ratio having $p(d; \tilde{\mu}_u)$ as a denominator by taking the reciprocal one of the ratios that is shown.

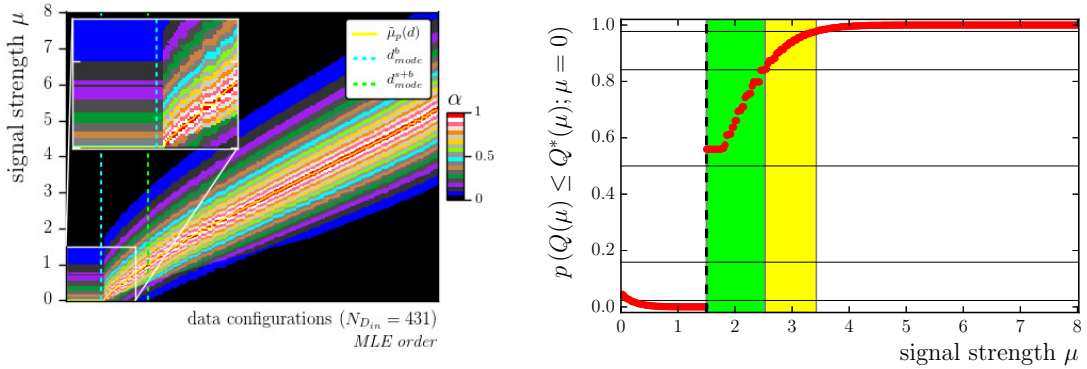
D.1.2. Reducing Reporting Risk with Q_{rep}

The second goal of the new test statistics is to reduce reporting risk, $R^{\alpha}_{rep}(\mu', \mu)$. Reporting risk was discussed in section 6.3.2 as the risk of producing a confidence interval such that if it does not contain μ_{true}, then the expected distance from μ_{true} to the closest end of the confidence interval may be large. The difficulty is that it is unknown what test statistic minimizes $R^{\alpha}_{rep}(\mu', \mu)$, so that one is left experimenting for test statistics that reduce $R^{\alpha}_{rep}(\mu', \mu)$ below the values achieved with the commonly used test statistics, Q_{FC}, Q_{D0} and Q_{AT}.

One possible way to minimize reporting risk would be to use the structure of the likelihood function. In section 6.2.2.2, the likelihood function for the multivariate Poisson distribution was seen to have a maximum value, $p(d; \tilde{\mu}_p)$ and to smoothly decrease from this value as μ increased in distance from $\tilde{\mu}_p$ (and not considering negative, unphysical values of μ). If one takes the likelihoodist approach to statistical inference [69], then the strongest evidence given an outcome d is for values of μ near $\tilde{\mu}_p$, and less evidence for values further from it.

One could consider a test statistic based on the distance of μ' to the MLE of the likelihood function [3], $\tilde{\mu}_p$, such as $Q_{rep}(d, \mu') = 1/|\mu' - \tilde{\mu}_p|$. While determining the acceptance region for each μ', this function would reject first the outcomes where μ' is far from the MLE of that outcome's likelihood function.

Figures D.3 (a) and (b) show the results of using $Q_{rep}(d, \mu) = 1/(0.1 + |\mu - \tilde{\mu}_p|)$ as the test statistic for example 2. The factor of 0.1 is inserted to regulate values of μ very close to $\tilde{\mu}_p$. These confidence intervals shown in figure D.3 (a) appear most similar to results obtained using Q_{FC} (figure 6.14 on page 292). Unlike Q_{FC}, confidence intervals for Q_{rep} for outcomes left of d^{b}_{mode} are identical because $|\mu - \tilde{\mu}_p|$ is the same for these. Also, confidence intervals of Q_{FC} are more sensitive to changes in the width of likelihood function, longer confidence intervals for outcomes with wider likelihood functions [4]. Additionally, figure D.3 (b) shows a discontinuity in the expected limits at the value of $\mu \sim 1.5$ where all the leftmost outcomes are simultaneously transition from being included to excluded.



(a) (b)

Figure D.3: Example 2 – Results of the Neyman construction using $Q_{rep}(d, \mu) = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}$: (a) acceptance ventiles and (b) expected limits.

[3] Recall that $\tilde{\mu}_p$ is a function of the outcomes d, as is any type of MLE.

[4] Likelihood function width was discussed in section 6.2.2.2 (and figures 6.6 (a)–(f)).

Table D.1 shows the performance metrics[5] of the confidence intervals obtained using $Q_{rep}(d,\mu) = 1/(0.1 + |\mu - \tilde{\mu}_p|)$ (and repeats those of some others from section 6.3.4 for ease of comparison). Like Q_{FC}, Q_{rep} does not produce empty regions. However, the expected upper limit for Q_{rep} is much higher, and both the discrimination risk and reporting risk are higher. The lower power of Q_{rep} to distinguish the nominal signal model from the background-only model is caused by the $\mu = 1$ acceptance region containing all the outcomes left of d^b_{mode}. Table D.2 provides a summary of the definitions of the performance metrics for convenience.

	$p^{\mu=1}_{acc}$	$p^{\tilde{\mu}_0}_{emp}$	$p^{\tilde{\mu}_0}_{mle}$	$p^{\tilde{\mu}_0}_{0}$	$u(0)$	$E_{\tilde{\mu}_0}[u]$	$E_{\tilde{\mu}_0}[l]$	u_{med}	$\Delta(\sigma_u)$	$R^{\tilde{\mu}_0}_{disc}$	$R^{\mu=1}_{rep}$
Q_{rep}	0.950	0.000	1.00	0.96	1.487	1.90	0.01	1.50	1.03	0.955	0.0141
Q_{FC}	0.950	0.000	1.00	0.96	0.312	1.54	0.01	1.47	1.57	0.672	0.0113
Q_{D0}	0.953	0.053	0.95	0.95	empty	1.27	0.00	1.22	1.57	0.608	0.0154
Q_{AT}	0.953	0.053	0.95	0.95	empty	1.27	0.00	1.22	1.65	0.608	0.0154

Table D.1: Example 2 – The performance of 95% confidence intervals made using $Q_{rep}(d,\mu') = \frac{1}{0.1+|\mu-\tilde{\mu}_p|}$. Results from some common methods are also provided for comparison.

metric	description
$p^{\mu=1}_{acc}$	The coverage of the nominal signal model, $\sum_{d'} p(d';\, \mu = 1)$.
$p^{\tilde{\mu}_0}_{emp}$	The probability of empty set confidence regions using $p(d;\, \tilde{\mu}_0)$.
$p^{\tilde{\mu}_0}_{mle}$	The probability the confidence interval (CI) contains $\tilde{\mu}_p$ using $p(d;\, \tilde{\mu}_0)$.
$p^{\tilde{\mu}_0}_{0}$	The probability the CI contains $\mu = 0$ using $p(d;\, \tilde{\mu}_0)$.
$u(0)$	The upper confidence limit for the outcome $d_i = 0$ for all i^{th} bins.
$E_{\tilde{\mu}_0}[u]$	The expected upper limit using $p(d;\, \tilde{\mu}_0)$.
$E_{\tilde{\mu}_0}[l]$	The expected lower limit using $p(d;\, \tilde{\mu}_0)$.
u_{med}	The median upper limit using $p(d;\, \tilde{\mu}_0)$.
$\Delta(\sigma_u)$	The expected upper limit uncertainty using $p(d;\, \tilde{\mu}_0)$.
$R^{\tilde{\mu}_0}_{disc}$	Discrimination risk: probability the CI contains both $\mu = 0$ and $\mu = 1$ using $p(d;\, \tilde{\mu}_0)$.
$R^{\mu=1}_{rep}$	Reporting risk: expected distance of the CI to $\mu = 1$ using $p(d;\, \mu = 1)$.

Table D.2: Summary of the performance metrics' definitions from section 6.3.4 provided for convenience.

[5] The performance metrics of confidence intervals are discussed in section 6.3.4.

Figure 6.3 (a) (page 270) shows why the reporting risk is higher for Q_{rep} than Q_{FC}. Examining the probability distribution for $\mu = 1$ (second horizontal row from the bottom of the figure), one sees that the highest probabilities generally occur for outcomes near d^{s+b}_{mode}, and the widths of the likelihood functions are greater for these outcomes. The Q_{FC} confidence intervals are correspondingly larger for these outcomes, and this lowers the reporting risk.

One other difference that is seen in table D.1 is that use of Q_{rep} results in the decrease of expected upper limit uncertainty, $\Delta(\sigma_\mu)$. Although this uncertainty reduction is desirable, the large increase in expected upper limit is not. However, it is shown in section 6.3.4 (page 316) that using $1/(0.1 + |\mu - \tilde{\mu}_p|)$ in conjunction with likelihood ratios in defining a test statistic usually provides the beneficial performance aspects of both reduced reporting risk and decreased upper limit uncertainty.

Combining $1/(0.1 + |\mu - \tilde{\mu}_p|)$ with a likelihood ratio alters the properties of the resultant confidence intervals. This is shown in figure D.4, where the test statistic $Q_{rep} \cdot Q_{AT}$ is used in the Neyman construction for example 2. These confidence intervals are two-sided for outcomes to the right of d^{s+b}_{mode}. As discussed in section 6.3.2, this property of transforming one-sided confidence intervals into two-sided is useful because it prevents the problem of the "hole" from forming in the confidence intervals (see figure 6.22, page 310).

Figure D.4 also shows that there are no empty confidence regions, and that the highest acceptance quantiles are clustered around the physical MLE, $\tilde{\mu}_p$ (yellow line). Compare this to the result using only Q_{AT} (figure 6.15 on page 293) which shows the presence of empty confidence regions for outcomes left of d^{b}_{mode}. Figure 6.15 also shows that there is a large gray area of constant acceptance quantile caused by so many outcomes above d^{s+b}_{mode} receiving the maximum Q_{AT} value of 1.0 for many lower values of μ. This may be an undesirable feature of using Q_{AT} if one wishes confidence intervals that are increasingly nested around the MLE for decreasing confidence level. Modifying Q_{AT} by multiplying by Q_{rep} achieves this.

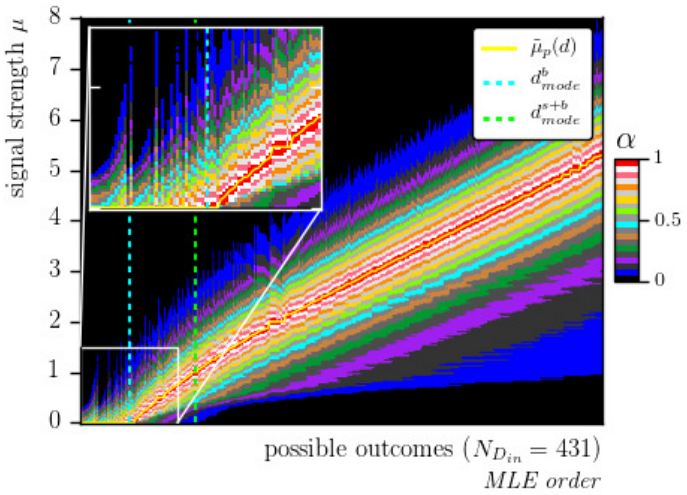


Figure D.4: Example 2 – Acceptance ventiles of the Neyman construction using $Q_{rep} \cdot Q_{AT}$.

D.1.3. Additional Examples: The Most Probable Test Statistic, Q_{MP}

The most probable test statistic, $Q_{MP}(d, \mu) = p(d; \mu)$ (discussed on page 286), is infrequently used in particle physics signal searches. It is based on Fisher significance testing rather than the on Neyman-Pearson hypothesis testing for the likelihood ratio test statistics. The following results show the properties of Q_{MP} applied to the four examples shown in table 6.7 on page 329. Table D.3 summarizes the performance metrics and is provided for convenience.

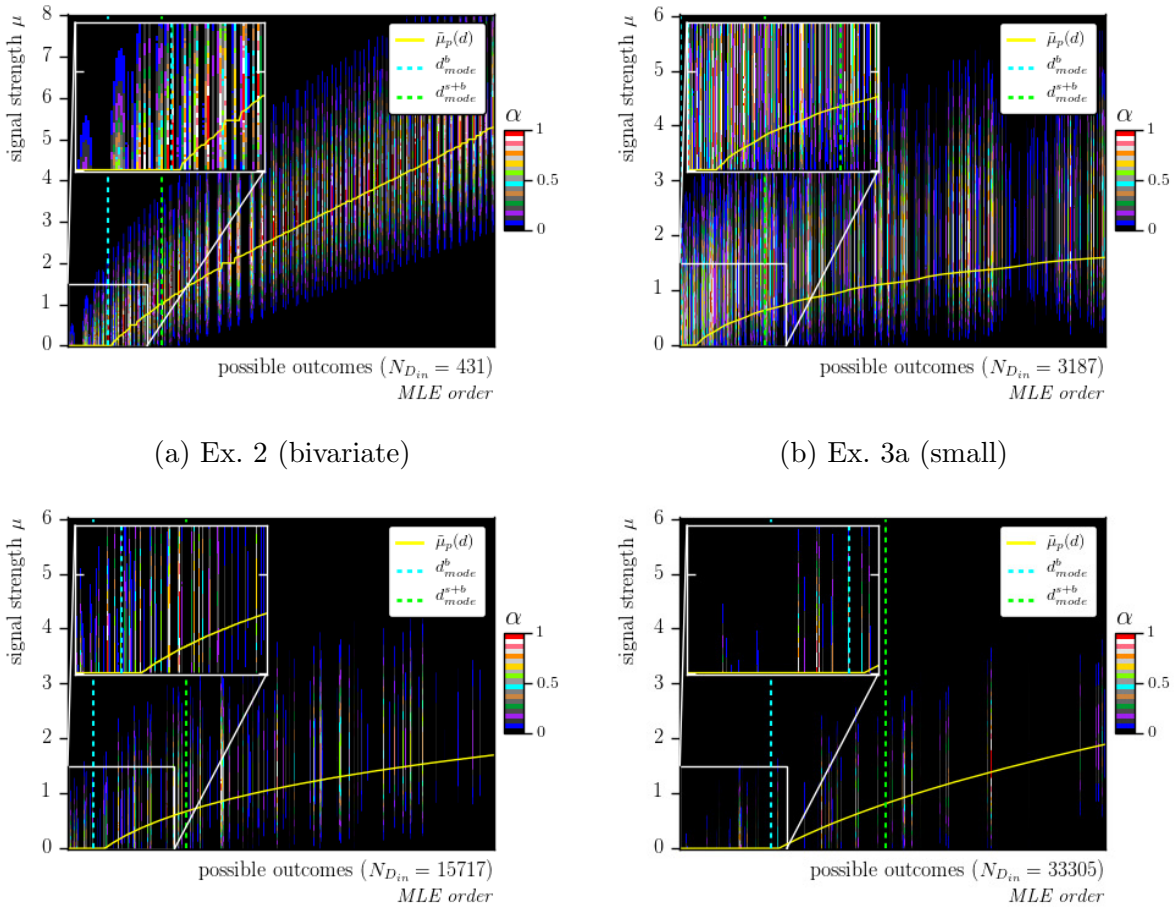
Figures D.5 (a)–(d) show the result of using significance, $Q_{MP}(d, \mu) = p(d; \mu)$, as the local ordering rule. As the examples progress from small to more complex, the number of enumerated outcomes grows (431 to 33,305) in order to include all outcomes with $\tilde{\mu}_p$ below some maximum value. Many of these outcomes have very small probabilities using either the background-only model or the signal model for any value of μ. By using Q_{MP}, only the more probable outcomes are included in acceptance regions, and so only the more probable outcomes have non-empty confidence intervals. This suggests that Q_{MP} is more appropriate when the experimenter is more certain of the modeling consistency, i.e., more certain that rare events *according to the models* are indeed truly rare [6].

Table D.4 shows that Q_{MP} has high discrimination loss risk, and this is sensible since Q_{MP} corresponds to significance testing where only a null hypothesis is examined, rather than hypothesis testing that compares a null hypothesis to an alternative hypothesis that would provide discrimination between models. Table 6.9 on page 338 tabulates the performance metrics of all the various test statistics. This shows Q_{MP} has significantly longer expected confidence interval lengths than the other test statistics, which severely limits the utility of Q_{MP}.

metric	description
$p_{acc}^{\mu=1}$	The coverage of the nominal signal model, $\sum_{d'} p(d'; \mu = 1)$.
$p_{emp}^{\tilde{\mu}_0}$	The probability of empty set confidence regions using $p(d; \tilde{\mu}_0)$.
$p_{mle}^{\tilde{\mu}_0}$	The probability the confidence interval (CI) contains $\tilde{\mu}_p$ using $p(d; \tilde{\mu}_0)$.
$p_0^{\tilde{\mu}_0}$	The probability the CI contains $\mu = 0$ using $p(d; \tilde{\mu}_0)$.
$u(0)$	The upper confidence limit for the outcome $d_i = 0$ for all i^{th} bins.
$E_{\tilde{\mu}_0}[u]$	The expected upper limit using $p(d; \tilde{\mu}_0)$.
$E_{\tilde{\mu}_0}[l]$	The expected lower limit using $p(d; \tilde{\mu}_0)$.
u_{med}	The median upper limit using $p(d; \tilde{\mu}_0)$.
$\Delta(\sigma_u)$	The expected upper limit uncertainty using $p(d; \tilde{\mu}_0)$.
$R_{disc}^{\tilde{\mu}_0}$	Discrimination risk: probability the CI contains both $\mu = 0$ and $\mu = 1$ using $p(d; \tilde{\mu}_0)$.
$R_{rep}^{\mu=1}$	Reporting risk: expected distance of the CI to $\mu = 1$ using $p(d; \mu = 1)$.

Table D.3: Summary of the performance metrics' definitions from section 6.3.4 provided for convenience.

[6] Similar results comparing Q_{MP} to Q_{FC} were obtained in [349].



(a) Ex. 2 (bivariate)

(b) Ex. 3a (small)

(c) Ex. 3b (medium)

(d) Ex. 3c (large)

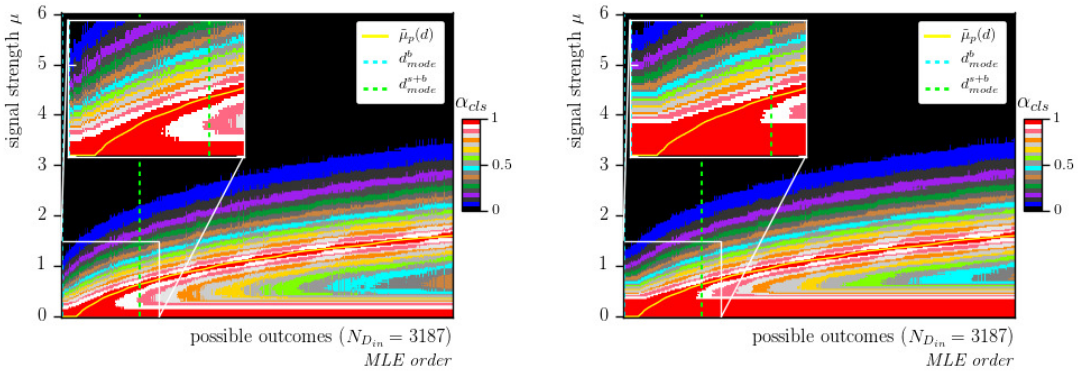
Figure D.5: Examination of the acceptance ventiles of the Neyman construction using Q_{MP}: (a) example 2 (bivariate), (b) example 3a (small), (c) example 3b (medium), and (d) example 3c (large).

$p^{\mu=1}_{acc}$	$p^{\tilde{\mu}_0}_{emp}$	$p^{\tilde{\mu}_0}_{mle}$	$p^{\tilde{\mu}_0}_0$	$u(0)$	$E_{\tilde{\mu}_0}[u]$	$E_{\tilde{\mu}_0}[l]$	u_{med}	$\Delta(\sigma_u)$	$R^{\tilde{\mu}_0}_{disc}$	$R^{\mu=1}_{rep}$
Example 2 (bivariate)										
0.951	0.017	0.98	0.95	empty	2.01	0.01	1.85	1.75	0.862	0.0238
Example 3a (small)										
0.950	0.001	1.00	0.96	1.662	2.91	0.01	2.85	1.45	0.962	0.0055
Example 3b (medium)										
0.950	0.015	0.98	0.96	empty	2.01	0.01	2.00	1.47	0.888	0.0146
Example 3c (large)										
0.950	0.024	0.98	0.96	empty	1.74	0.00	1.77	1.47	0.826	0.0170

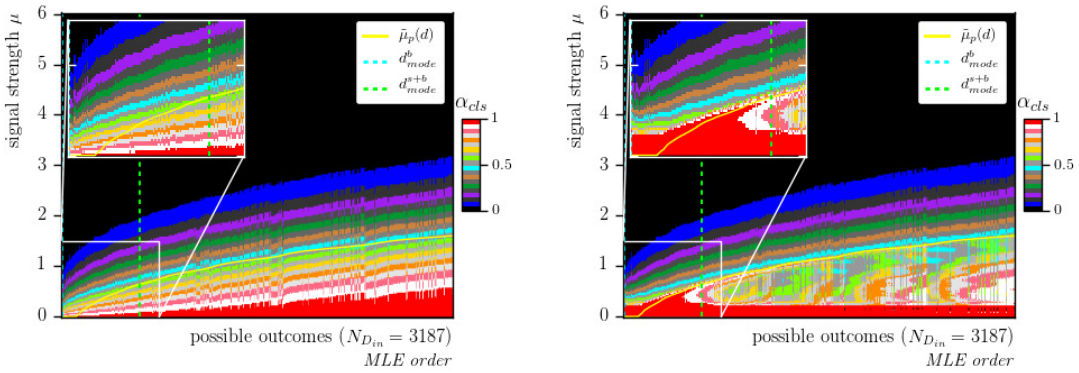
Table D.4: Other examples – Performance metrics of the most probable test statistic, $Q_{MP}(d, \mu) = p(d;\, \mu)$.

D.1.4. Additional Examples: The CL_S Method

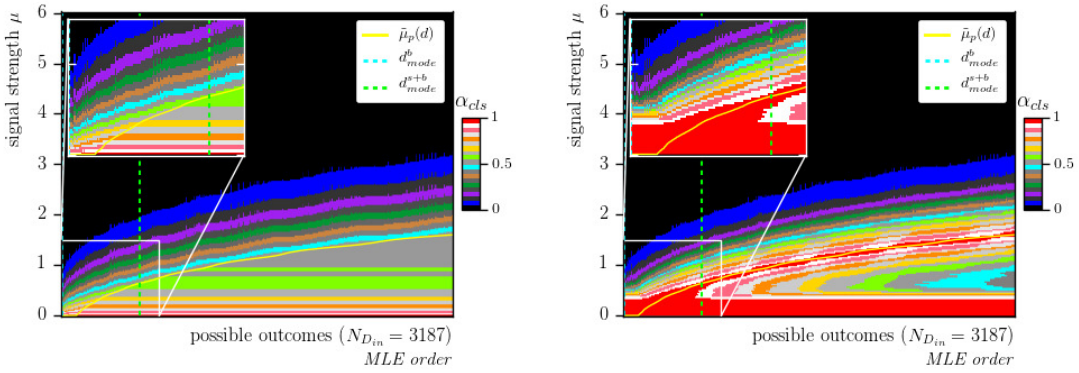
Figures D.6–D.8 show the acceptance ventiles of the three examples, 3a–3c, for the common and Arizona test statistics after also implementing the CL_S method.



(a) Q_{FC} (b) Q_{AZ1}

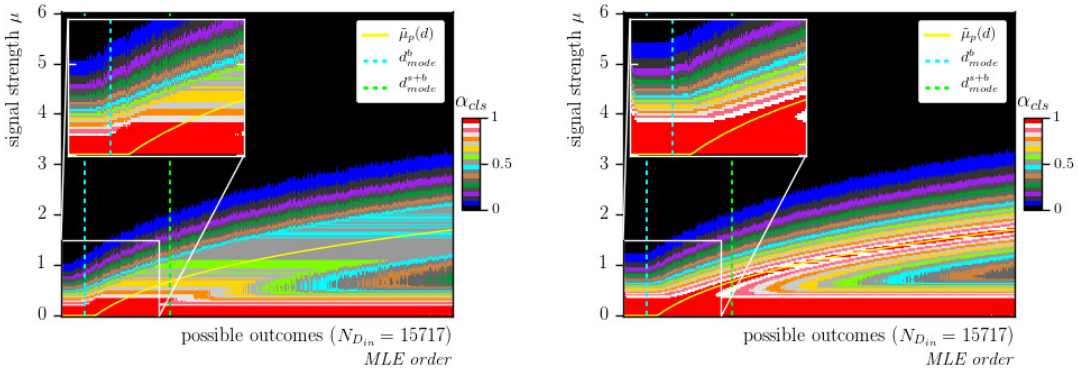


(c) Q_{D0} (d) Q_{AZ2}

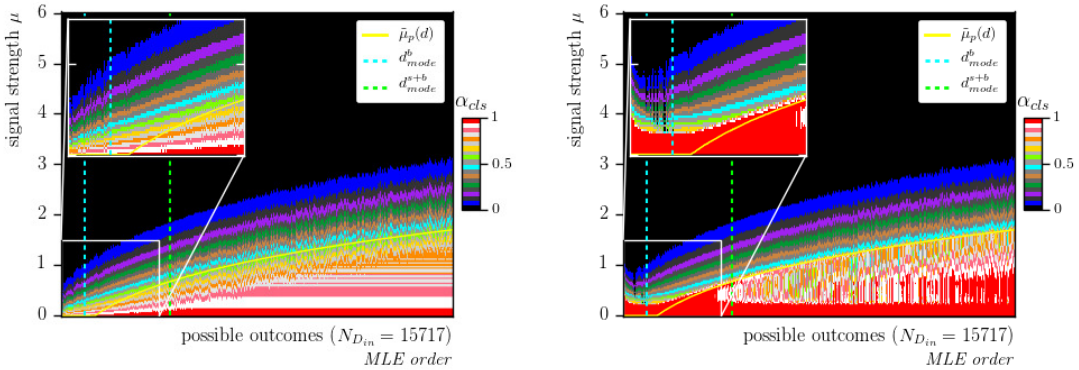


(e) Q_{AT} (f) Q_{AZ3}

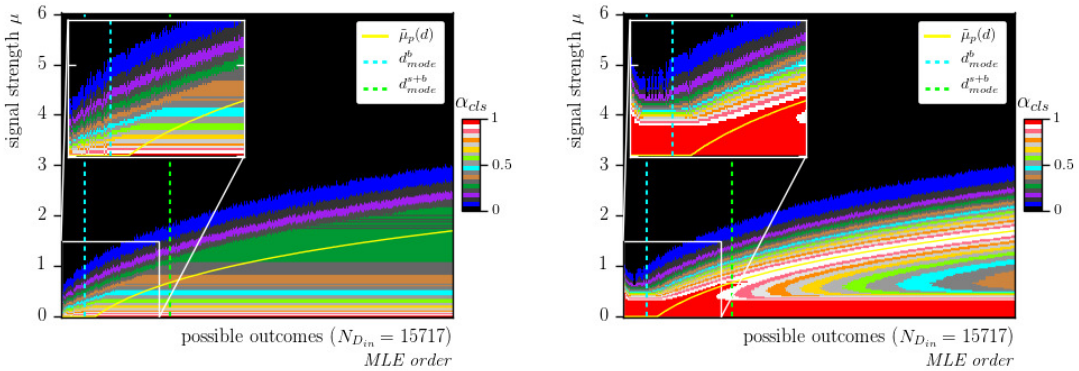
Figure D.6: Example 3a (small) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics and implementing the CL_S method: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}.



(a) Q_{FC} (b) Q_{AZ1}

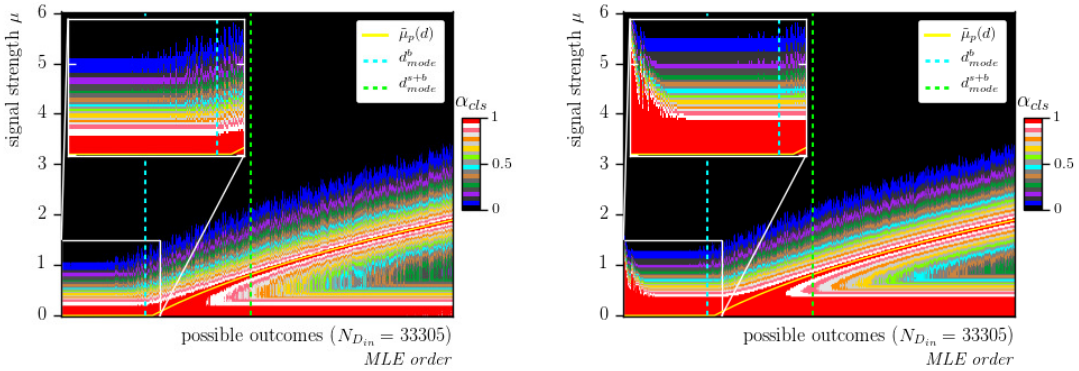


(c) Q_{D0} (d) Q_{AZ2}

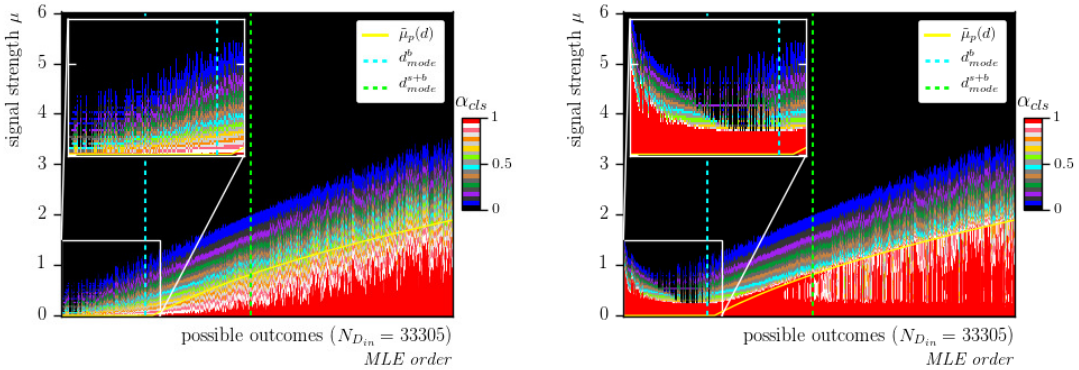


(e) Q_{AT} (f) Q_{AZ3}

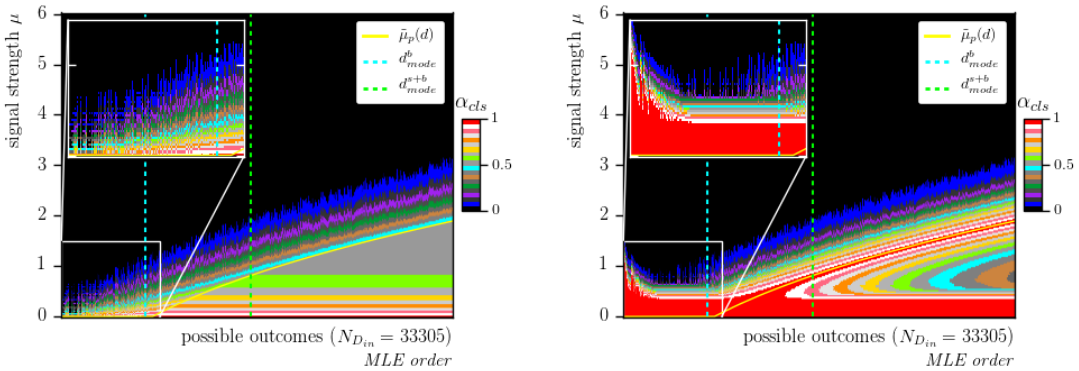
Figure D.7: Example 3b (medium) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics and implementing the CL_S method: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}.



(a) Q_{FC} (b) Q_{AZ1}



(c) Q_{D0} (d) Q_{AZ2}



(e) Q_{AT} (f) Q_{AZ3}

Figure D.8: Example 3c (large) – Comparison of the acceptance ventiles of the Neyman construction using the common and Arizona test statistics and implementing the CL_S method: (a) Q_{FC}, (b) Q_{AZ1}, (c) Q_{D0}, (d) Q_{AZ2}, (e) Q_{AT}, and (f) Q_{AZ3}.

D.2. Confidence-Acceptance Profile Properties (Addendum)

It is difficult to provide mathematical proofs of the CAP properties because even if one can generalize to a family of statistical models, the proofs will still be dependent on the form of the test statistic used in the Neyman construction. Therefore, the following heuristic arguments are provided as evidence that it is possible to construct such proofs, i.e., sketches of proofs. Furthermore, to avoid the complications of multivariate and discrete probability models, the following simple truncated Gaussian model will be used (designed to approximate a univariate Poisson signal model)

$$p(d;\,\mu) = \frac{C(\mu)}{\sqrt{2\pi(\mu\,s+b)}}\,exp\left[-\frac{1}{2}\frac{(d-\mu\,s-b)^2}{\mu\,s+b}\right], \tag{D.2.1}$$

where s and b are the positive fixed signal and background expected yields, μ is physical, $\mu \geq 0$, and outcomes are physical, $d \geq 0$, so that $C(\mu)$ is a normalization constant that depends on μ. For simplicity, the background-only model is given by $p(d;\,\tilde{\mu}_0) = p(d;\,\mu = 0)$.

For any value μ', the mode of the outcomes is $d_{mode}(\mu') = \mu'\,s + b$ with probability $p(d_{mode}(\mu'), \mu') = 0.4/\sqrt{\mu'\,s+b}$. For any outcome d', the physical MLE is $\tilde{\mu}_p(d') = max[0, (d'-b)/s]$. Note that $d_{mode}(\mu)$ is written with its functional dependence on μ and $\tilde{\mu}_p(d')$ on with its dependence on d'.

(1) FLP $\rightarrow$ acceptance regions increasingly clustered around $\mathbf{d_{mode}}$:

(Heuristic.) This is proven with a pivoting argument. The frequentist likelihood property (FLP) states that confidence intervals are increasing clustered around $\tilde{\mu}_p(d)$. Since $\tilde{\mu}_p(d) = (d-b)/s$ when $d \geq b$ (or else zero) and $d_{mode}(\mu) = \mu\,s + b$, it is clear that for $d \geq b$, $\tilde{\mu}_p(d) = (d_{mode}(\mu) - b)/s$. In other words, the maximum of the likelihood function, $p(d;\mu)$, (μ location) is also the maximum of the signal model pdf, $p(d;\mu)$, (d location).

Note that $S(d_{mode}(\mu), \tilde{\mu}_p(d))$ is greater than both $S(d_{mode}(\mu), \mu' \neq \tilde{\mu}_p(d))$ (μ' as variable) and $S(d' \neq d_{mode}(\mu), \tilde{\mu}_p(d))$ (d' as variable). The FLP provides that $S(d_{mode}(\mu), \mu')$ is decreasing as μ' gets more distant from $\tilde{\mu}_p(d)$. Let the outcomes be globally ordered using the MLE order. Now one needs to show that $S(d', \tilde{\mu}_p(d))$ is decreasing as d' becomes increasing distant from $d_{mode}(\mu)$ (although care must be taken for $d' < b$). This has been demonstrated in many acceptance ventile plots for the examples using Q_{FC}, Q_{AZ1} and Q_{AZ3}, the test statistics that satisfy the FLP.

(2) $\mathbf{S(d,\boldsymbol{\mu}) < B(d,\boldsymbol{\mu})}$, $\mathbf{G^+(d,\boldsymbol{\mu}) \geq 1}$ and $\mathbf{G^-(d,\boldsymbol{\mu}) \leq 1}$:

If the term $p(d;\,\mu)$ is defined to be in the numerator of the test statistic, Q, then it is usually the case that $Q(d', \tilde{\mu}_0) < Q(d', \mu)$ for any outcome d' and any $\mu > 0$. This means that the test statistic distribution for the signal model is shifted to higher values than the background-only model.

In turn, this implies that the left hand integral from $Q(d', \mu)$ for any outcome d' using the background-only distribution, given by $B(d, \mu)$, will be greater than that of the signal distribution, $S(d, \mu)$. Because the BSR function is defined as $(1-A)/(1-B)$, the BSR function is always greater than 1.0 for $\mu > 0$. Because the SSR function is defined as A/B, the SSR function is always less than 1.0 for $\mu > 0$.

(3) $\mathbf{S(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to 0}} B(d,\boldsymbol{\mu})}$, $\mathbf{G^{+}(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to 0}} 1}$, and $\mathbf{G^{-}(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to 0}} 1}$:

As $\mu \to 0$, the signal model pdf, $p(d;\,\mu)$, becomes increasingly similar to the background-only model pdf, $p(d;\,\tilde{\mu}_0)$, so that discrimination power decreases as the test statistic distributions of the two models grow increasingly similar. In that case, any integrations over the test statistic distributions will be nearly equal so that

$$S(d,\mu) \xrightarrow{\mu\to 0} B(d,\mu), \tag{D.2.2}$$

regardless of the outcome d. Therefore, the BSR function approaches 1.0 from above for decreasing μ, and the SSR function approaches 1.0 from below.

(4) $\mathbf{S(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to+\infty}} 0}$:

(Heuristic.) For any given outcome d_L, if μ is large enough so that $\mu > (d_L - b)/s = \tilde{\mu}_p(d_L)$, then $d_L < d_{mode}(\mu)$. Thus for large enough μ, d_L lies in the left tail of $p(d;\,\mu)$.

Assume that for large μ, $Q(d'_L,\mu) < Q(d_L,\mu)$ when $d'_L < d_L$. This is approximately true for test statistics defined with $p(d;\,\mu)$ in the numerator because $p(d'_L;\,\mu) < p(d_L;\,\mu)$. Thus the outcomes $d'_L \le d_L$ are included in the integral $S(d_L,\mu)$.

However, d'_L in the left tail of the distribution may not be the only outcomes with smaller test statistic values than d_L. For most test statistics, there will be typically a right tail of outcomes with low test statistic values. Roughly speaking, let $d_R(\mu) = 2d_{mode}(\mu) - d_L$ be the outcome to the right of $d_{mode}(\mu)$ that is equally distant from the mode as d_L. Then approximately for large μ, $Q(d'_R,\mu) > Q(d_R(\mu),\mu) \sim Q(d_L,\mu)$ for $d'_R \ge d_R(\mu)$. In other words, $S(d_L,\mu)$ is given by the sum of the tail integrals defined by d_L and $d_R(\mu)$. Note that d_L is fixed as the behavior of $S(d_L,\mu)$ is examined for increasing μ, but that $d_R(\mu)$ increases as μ increases.

As μ increases, the distance from d_L to the mode, $d_{mode}(\mu)$, increases (and thus the distance from $d_R(\mu)$ to the mode also increases). In other words, as μ increases, d_L is located farther into the left tail of $p(d;\,\mu)$ and $d_R(\mu)$ farther into the right tail. However, as μ increases, so too does the width of $p(d;\,\mu)$ so that it must be shown that the tail probabilities are still decreasing. The likelihood function provides the solution. When $\mu > \tilde{\mu}_p(d)$, the likelihood function $p(d;\,\mu)$ is decreasing with increasing μ at a rate $O\left(exp[-\mu\, s]/\sqrt{\mu\, s}\right)$ ("big O" notation), so that indeed these tail probabilities are all approaching zero, and thus so is $S(d_L,\mu)$ for any d_L.

(5) $\mathbf{B(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to+\infty}} \sim \tfrac{1}{2}B(d,\boldsymbol{\mu}=0)}$, $\mathbf{G^{+}(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to+\infty}} \sim 1/\left(1-\tfrac{1}{2}B(d,\boldsymbol{\mu})\right)}$ and $\mathbf{G^{-}(d,\boldsymbol{\mu}) \xrightarrow{\boldsymbol{\mu\to+\infty}} 0}$:

Begin with $\mu' \sim 0$. From (1) the acceptance regions are clustered about $d_{mode}(\mu')$ so that the smallest test statistic values are for outcomes in the left and right tails of $p(d;\,\mu')$ with roughly equal total probability in each tail.

Let any given outcome d_L define the boundaries of left and right tails in the same manner as in (4) so that $d_R(\mu')$ is also defined. Then the outcomes $d'_L \le d_L$ and $d'_R \ge d_R(\mu')$ are the outcomes included in the integral $B(d_L,\mu')$ for $\mu' \sim 0$. Define $p_{left} + p_{right}(\mu) \equiv B(d_L,\mu)$ to emphasize that while d_L is fixed, $d_R(\mu)$ is not so neither is $p_{right}(\mu)$.

Now examine what happens to the outcomes defined for μ', $d'_R \geq d_R(\mu')$, as μ increases. For $\mu > \mu'$, the mode is shifted to larger values, $d_{mode}(\mu) > d_{mode}(\mu')$. Therefore, the outcomes $d'_L \leq d_L$ are further from $d_{mode}(\mu)$ than they are from $d_{mode}(\mu')$ while the outcomes $d'_R \geq d_R(\mu')$ are closer to $d_{mode}(\mu)$ than to $d_{mode}(\mu' \sim 0)$. Thus the test statistic values for d'_L are decreasing, while d'_R are increasing. This means that some of the outcomes contributing to $p_{right}(\mu')$ are removed as μ increases giving $p_{right}(\mu'') < p_{right}(\mu')$.

Note that the outcome d_L is fixed and the behavior of $B(d_L, \mu)$ as μ increases is being examined. For a very large value μ'', d_L will be very distant from $d_{mode}(\mu'')$, i.e., it will define a more distant left tail for $p(d; \mu'')$. Let the location of the corresponding right tail for μ'' be given by $d_R(\mu'')$ in the same manner as (4). It must be the case that $d'_L \leq d_L$ and $d''_R \geq d_R(\mu'')$ are the outcomes included in the integral $B(d_L, \mu'')$ for very large μ''. Thus, the outcomes d'_L defined for $\mu' \sim 0$ are constant regardless of μ, but the outcomes d''_R is a smaller set than the outcome d'_R defined for $\mu' \sim 0$.

$B(d_L, \mu'')$ is the probability integration assuming the background-only model, $p(d; \tilde{\mu}_0)$. The integration over $d'_L \leq d_L$ gives the same value, p_{left} regardless of μ. However, for very large μ'', $d_R(\mu'')$ is located much farther into the right tail of the background-only model than was $d_R(\mu')$. As μ increases, the integration over the increasingly distant right tail decreases toward zero, $p_{right}(\mu) \xrightarrow{\mu \to +\infty} 0$. Therefore, only half of the integration given by $B(d_L, \mu' \sim 0)$ remains so that $B(d_L, \mu) \xrightarrow{\mu \to +\infty} \sim \frac{1}{2} B(d_L, \mu = 0)$.

From this and (4), the BSR function must have $\sim 1/(1 - \frac{1}{2} B(d_L, \mu' \sim 0))$ as a limiting value, while since $B(d_L, \mu)$ approaches a constant positive value, the SSR function must have zero as a limiting value.

D.3. Additional Nuisance Parameter Results (Addendum)

The following subsections support those in section 6.5 with

- subsection D.3.1 supporting subsection 6.5.3: *Effect of Nuisance Parameters on Acceptance Regions*,
- subsection D.3.2 supporting subsection 6.5.4: *Effect of Nuisance Parameters on Confidence-Acceptance Profiles*,
- subsection D.3.3 supporting 6.5.5: subsection *Effect of Nuisance Parameters on Acceptance Regions*.

D.3.1. Effect of Nuisance Parameters on Acceptance Regions

Supplemental to section 6.5.3, figures D.9–D.13 provide the $\mu = 1$ signal and background-only distributions for the five test statistics: Q_{FC}, Q_{AT}, Q_{AZ1}, Q_{AZ2}, and Q_{AZ3}, respectively. The demonstrate the same properties that were discussed for Q_{D0} in section 6.5.3 (figure 6.62 on page 404).

D.3.2. Effect of Nuisance Parameters on Confidence-Acceptance Profiles

Supplemental to section 6.5.4, figures D.14 (a)–(l) show the confidence-acceptance profiles (CAPs) for the common test statistics after implementing the CL_S method, and figures D.15 (a)–(l) show the CAPs for the Arizona test statistics after implementing the CL_S method.

D.3.3. Effect of Nuisance Parameters on Coverage

Supplemental to section 6.5.5, figures D.16–D.21 show the 2σ coverage studies for the common and Arizona test statistics using all three nuisance parameter methods, nominal, profile, and full, and showing results both without and with the implementation of the CL_S method.

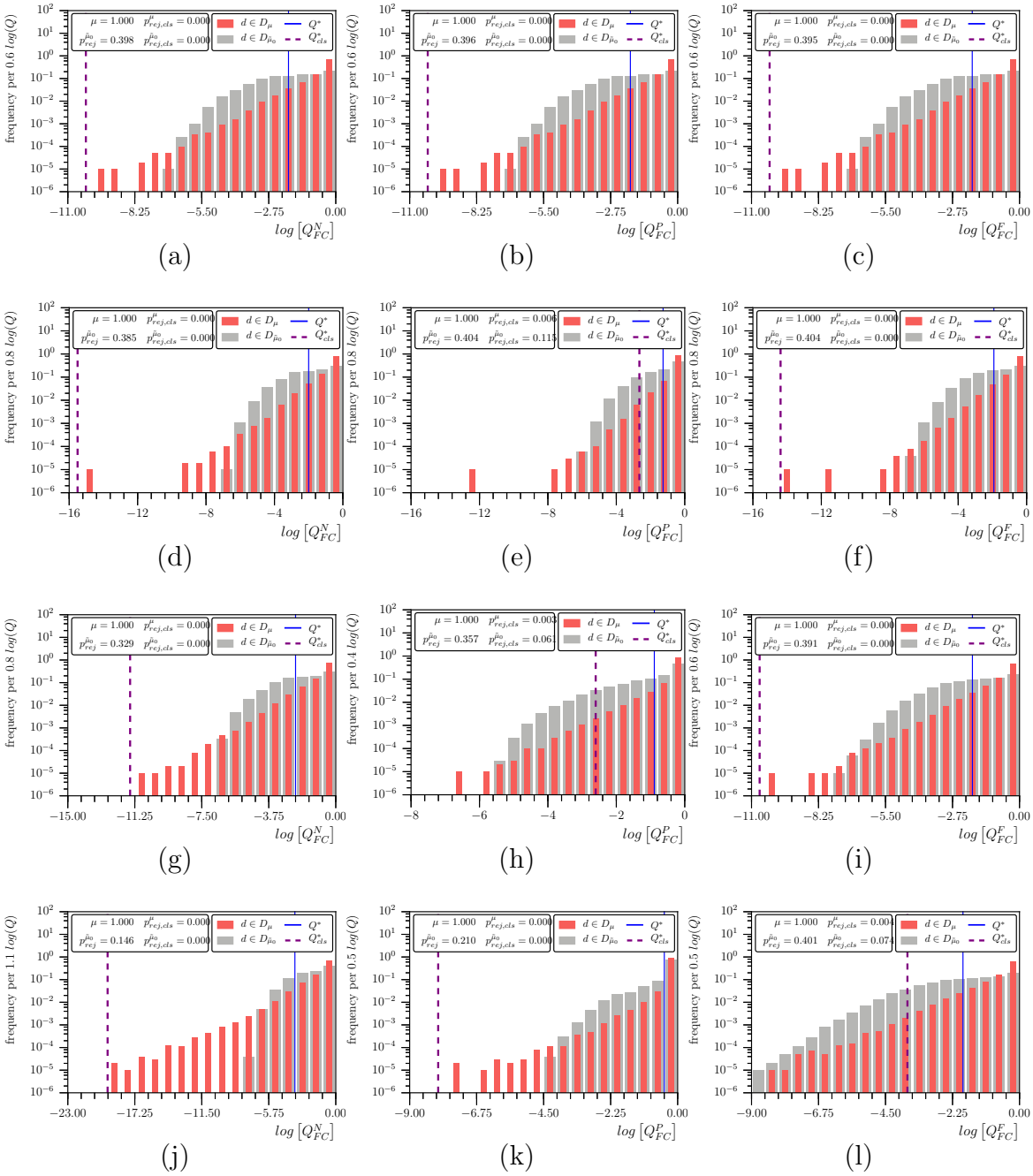


Figure D.9: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{FC}, (b) Q^P_{FC}, and (c) Q^F_{FC}. With 5% uncertainty using (d) Q^N_{FC}, (e) Q^P_{FC}, and (f) Q^F_{FC}. With 10% uncertainty using (g) Q^N_{FC}, (h) Q^P_{FC}, and (i) Q^F_{FC}. With 25% uncertainty using (j) Q^N_{FC}, (k) Q^P_{FC}, and (l) Q^F_{FC}.

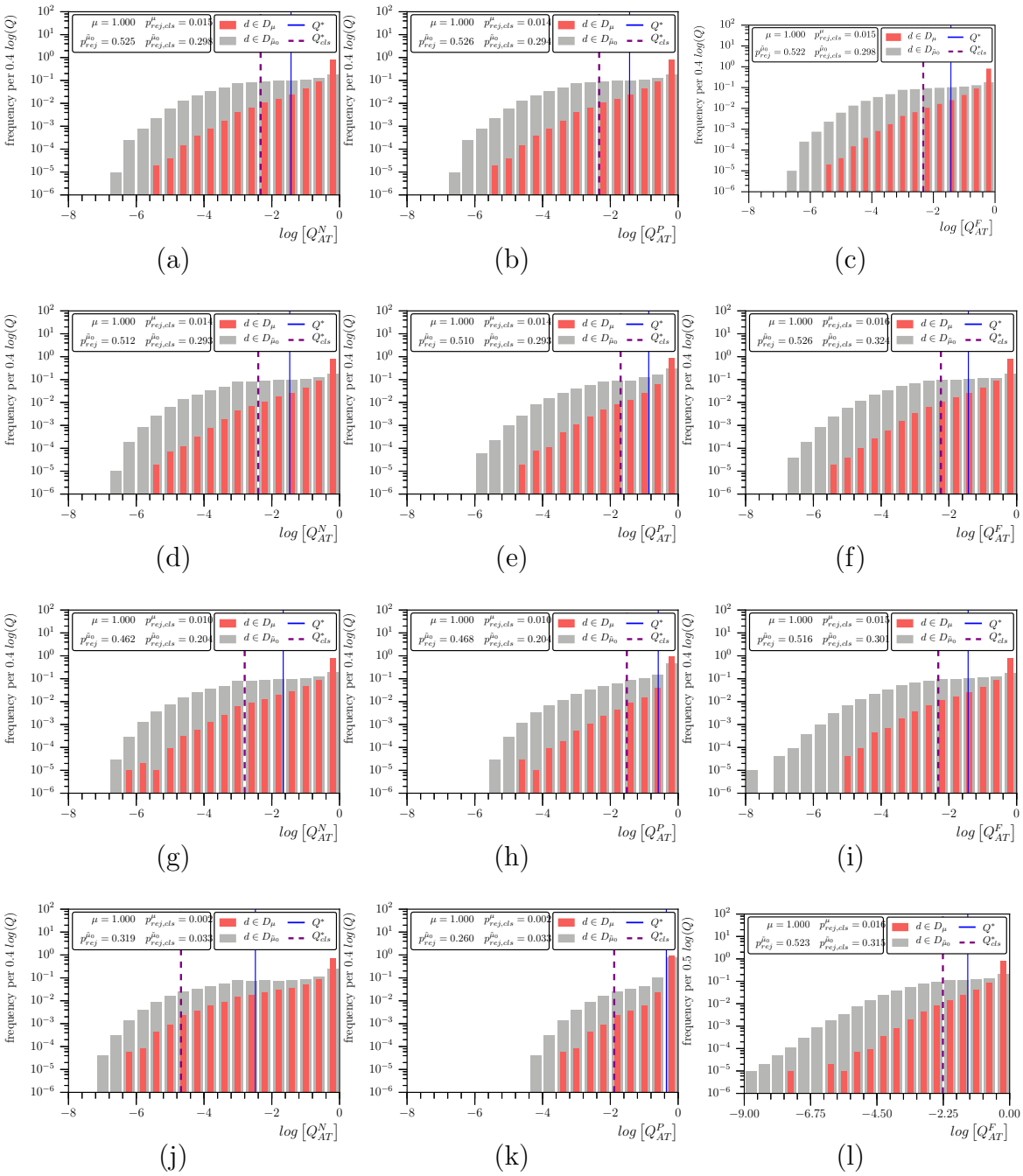


Figure D.10: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{AT}, (b) Q^P_{AT}, and (c) Q^F_{AT}. With 5% uncertainty using (d) Q^N_{AT}, (e) Q^P_{AT}, and (f) Q^F_{AT}. With 10% uncertainty using (g) Q^N_{AT}, (h) Q^P_{AT}, and (i) Q^F_{AT}. With 25% uncertainty using (j) Q^N_{AT}, (k) Q^P_{AT}, and (l) Q^F_{AT}.

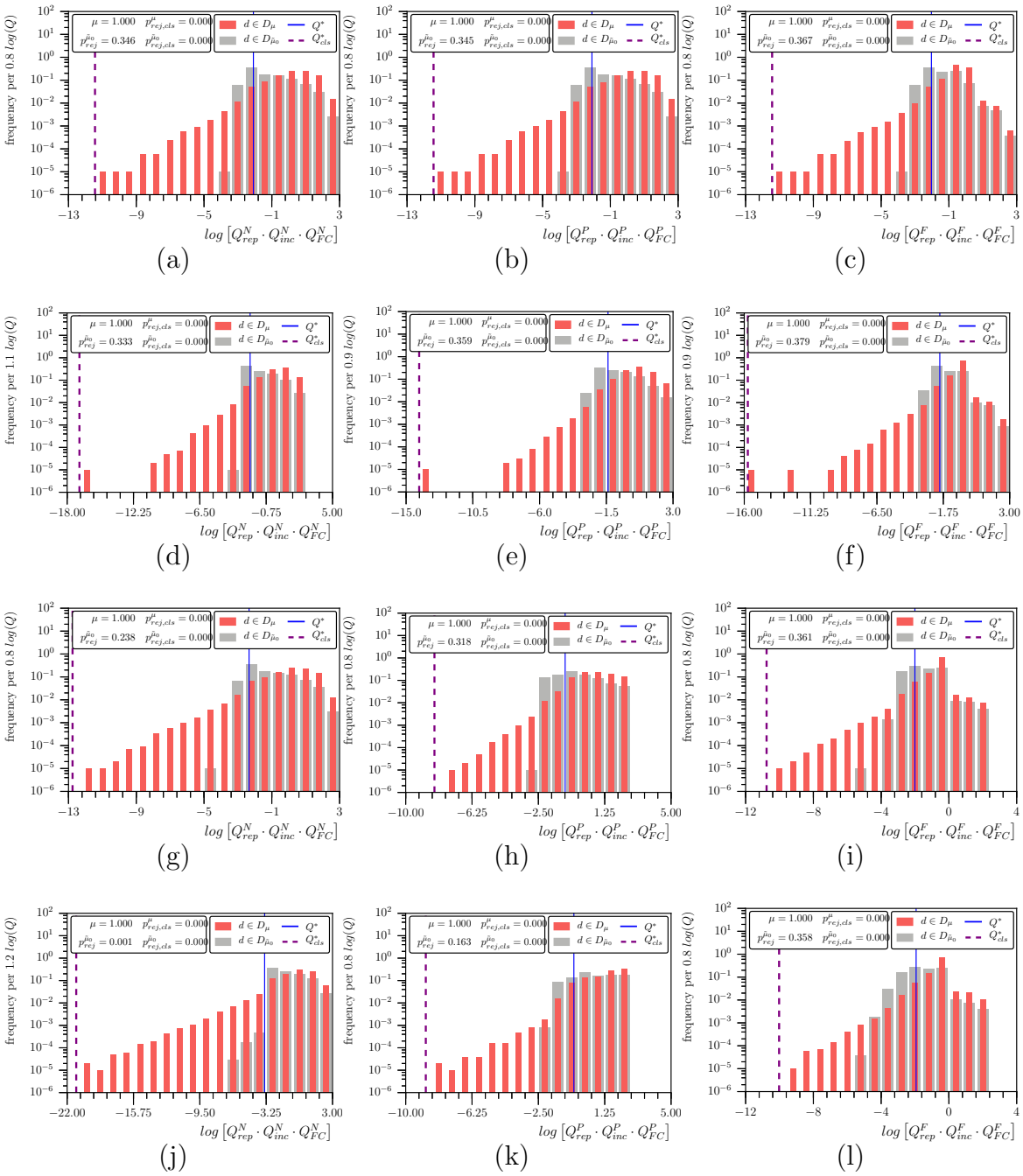


Figure D.11: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{AZ1}, (b) Q^P_{AZ1}, and (c) Q^F_{AZ1}. With 5% uncertainty using (d) Q^N_{AZ1}, (e) Q^P_{AZ1}, and (f) Q^F_{AZ1}. With 10% uncertainty using (g) Q^N_{AZ1}, (h) Q^P_{AZ1}, and (i) Q^F_{AZ1}. With 25% uncertainty using (j) Q^N_{AZ1}, (k) Q^P_{AZ1}, and (l) Q^F_{AZ1}.

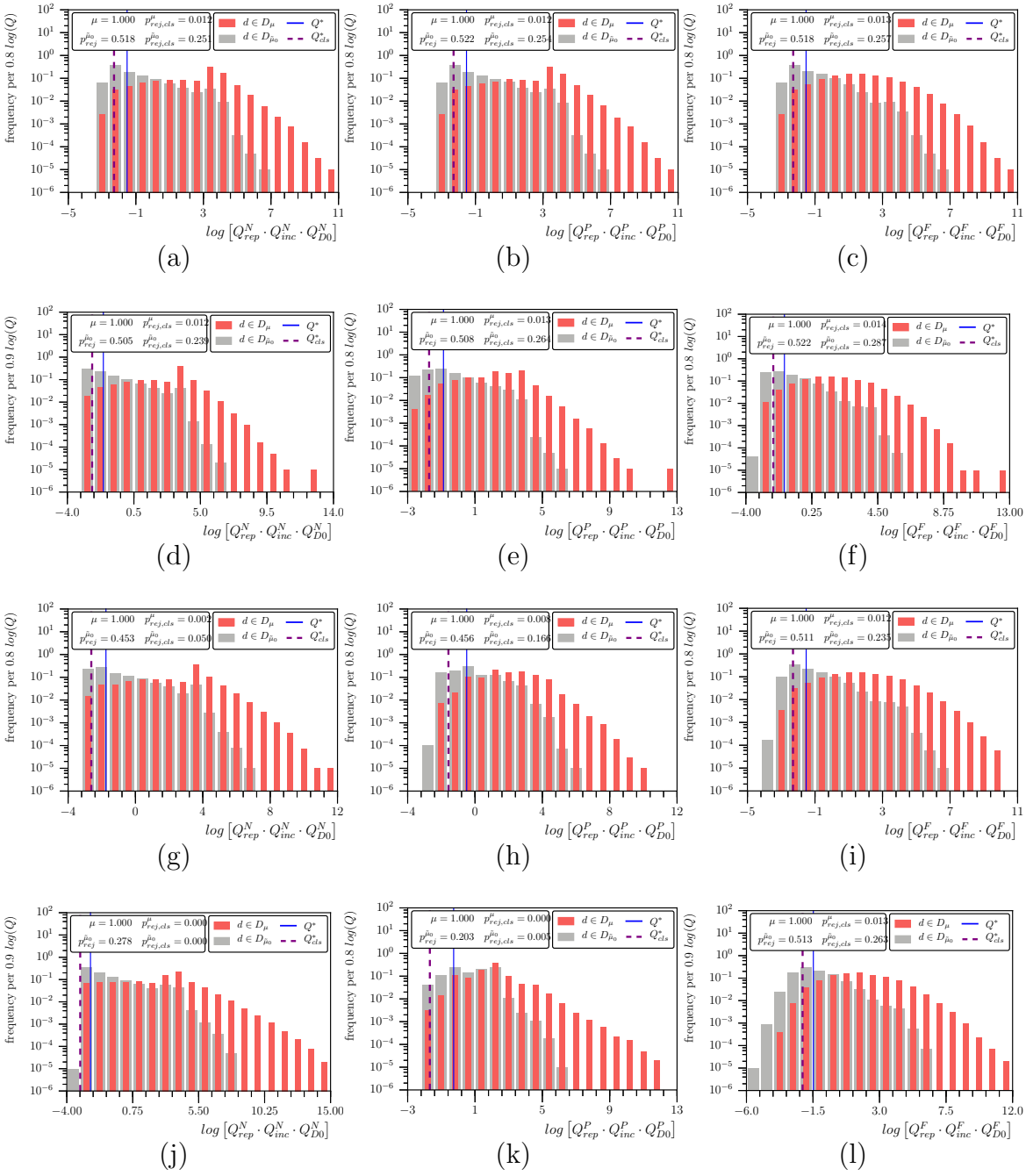


Figure D.12: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{AZ2}, (b) Q^P_{AZ2}, and (c) Q^F_{AZ2}. With 5% uncertainty using (d) Q^N_{AZ2}, (e) Q^P_{AZ2}, and (f) Q^F_{AZ2}. With 10% uncertainty using (g) Q^N_{AZ2}, (h) Q^P_{AZ2}, and (i) Q^F_{AZ2}. With 25% uncertainty using (j) Q^N_{AZ2}, (k) Q^P_{AZ2}, and (l) Q^F_{AZ2}.

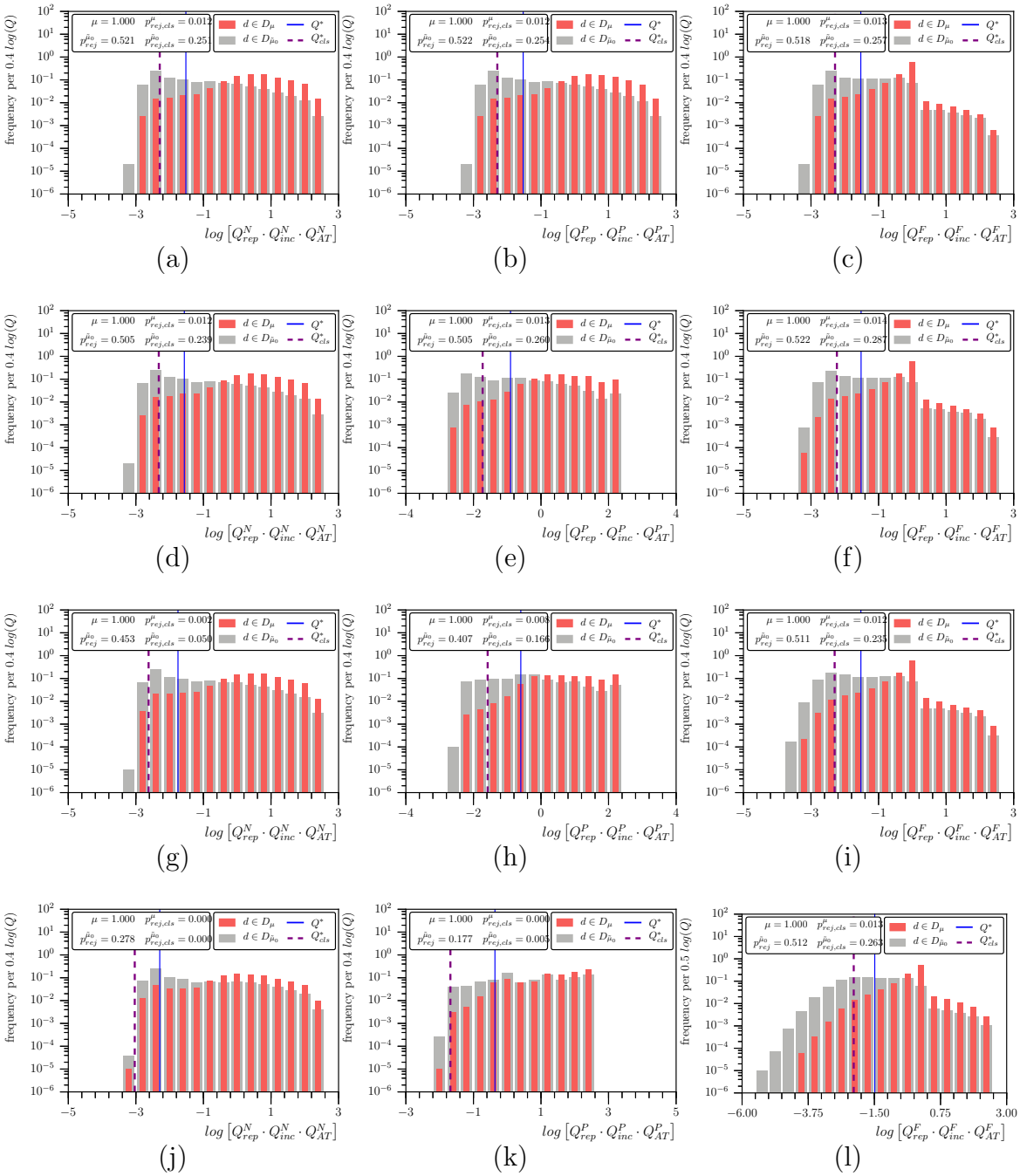


Figure D.13: Example 3b (medium) – The impact of nuisance parameters on the $\mu = 1.0$ acceptance regions with. With 0.001% uncertainty in all ν using (a) Q^N_{AZ3}, (b) Q^P_{AZ3}, and (c) Q^F_{AZ3}. With 5% uncertainty using (d) Q^N_{AZ3}, (e) Q^P_{AZ3}, and (f) Q^F_{AZ3}. With 10% uncertainty using (g) Q^N_{AZ3}, (h) Q^P_{AZ3}, and (i) Q^F_{AZ3}. With 25% uncertainty using (j) Q^N_{AZ3}, (k) Q^P_{AZ3}, and (l) Q^F_{AZ3}.

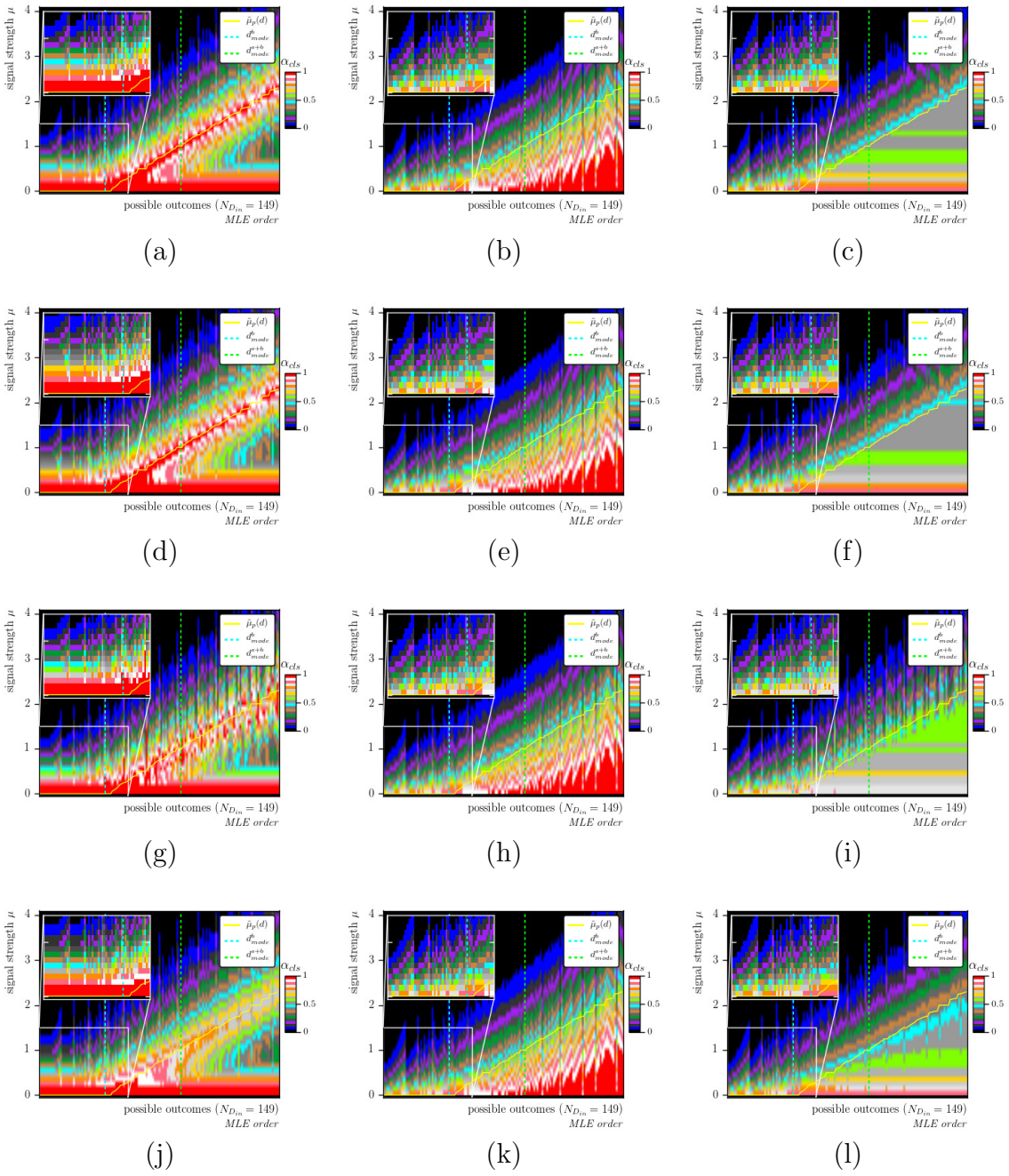


Figure D.14: Example 2 (bivariate) – The confidence-acceptance profiles when nuisance parameters, ν, are included (shown with ventile precision), and implementing the $\mathrm{CL_S}$ method. The first column shows results for Q_{FC}, the second for Q_{D0}, and third for Q_{AT}. The first row shows the results for Q without including ν, the second for the nominal method, Q^N, the third for profiling, Q^P, and fourth for full, Q^F.

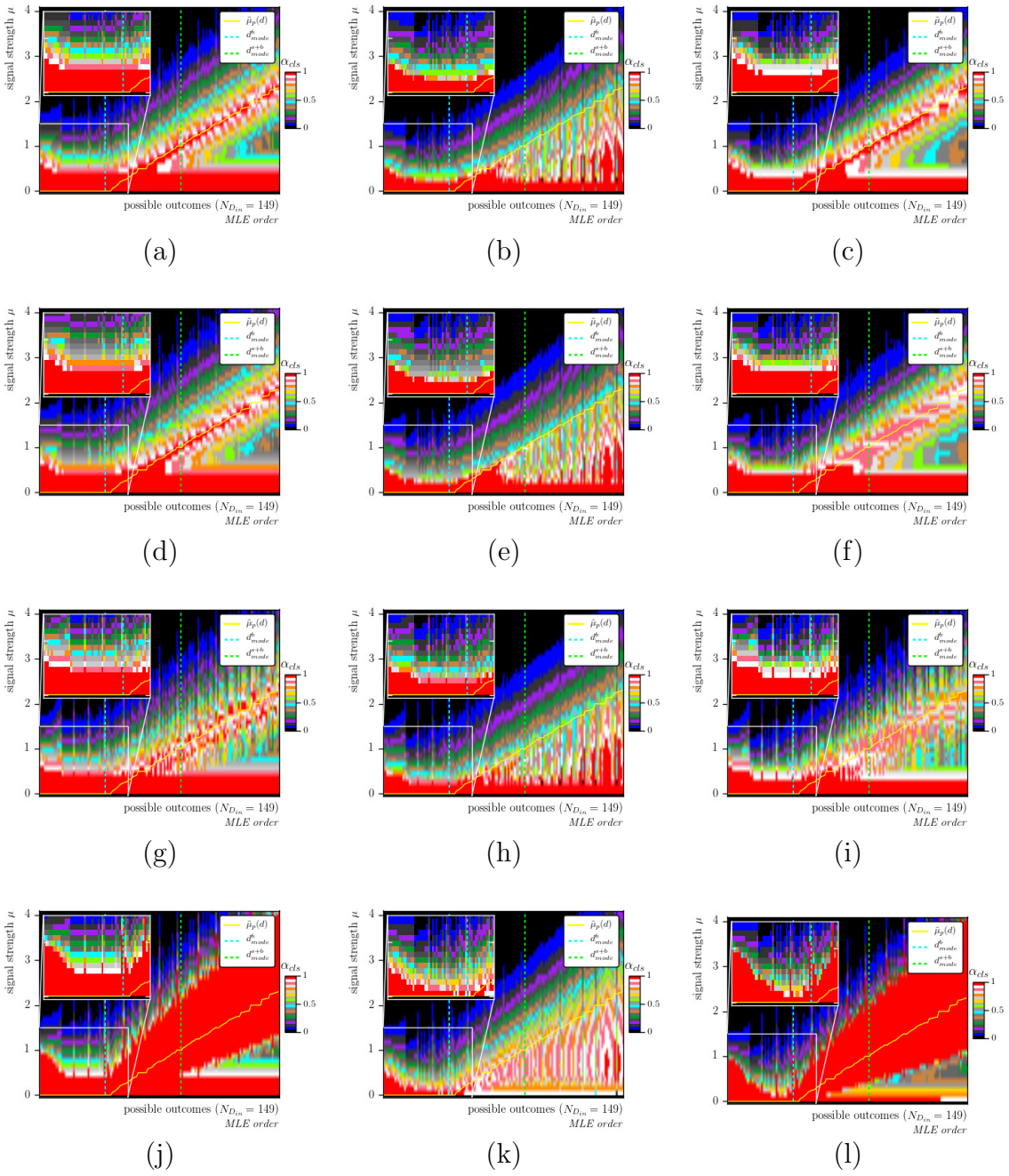


Figure D.15: Example 2 (bivariate) – The confidence-acceptance profiles when nuisance parameters, ν, are included (shown with ventile precision), and implementing the $\mathrm{CL_S}$ method. The first column shows results for Q_{AZ1}, the second for Q_{AZ2}, and third for Q_{AZ3}. The first row shows the results for Q without including ν, the second for the nominal method, Q^N, the third for profiling, Q^P, and fourth for full, Q^F.

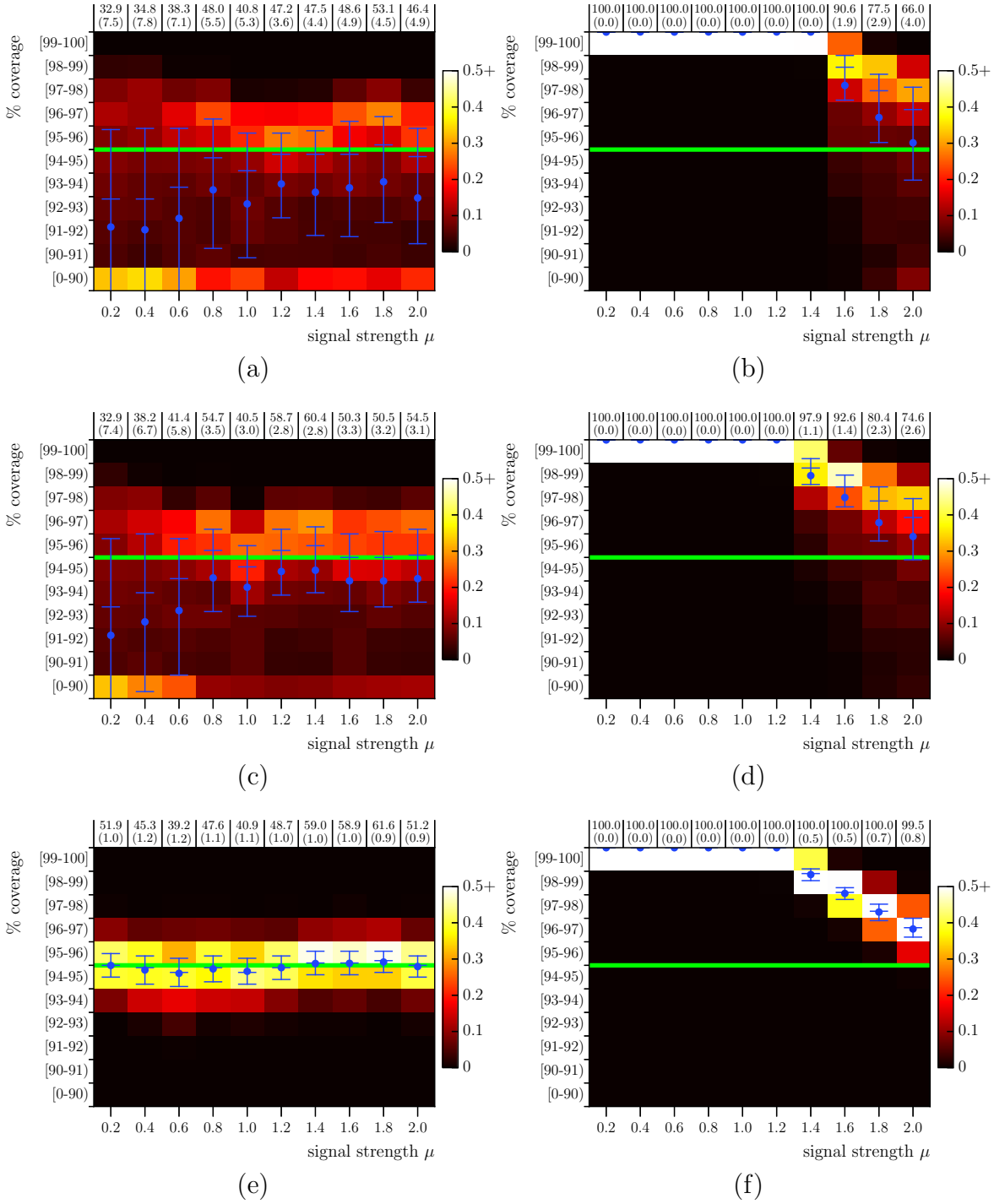


Figure D.16: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{FC} test statistic: (a) Q^N_{FC}, (b) Q^N_{FC} with CLs method, (c) Q^P_{FC}, (d) Q^P_{FC} with CLs method, (e) Q^F_{FC}, (f) Q^F_{FC} with CLs method.

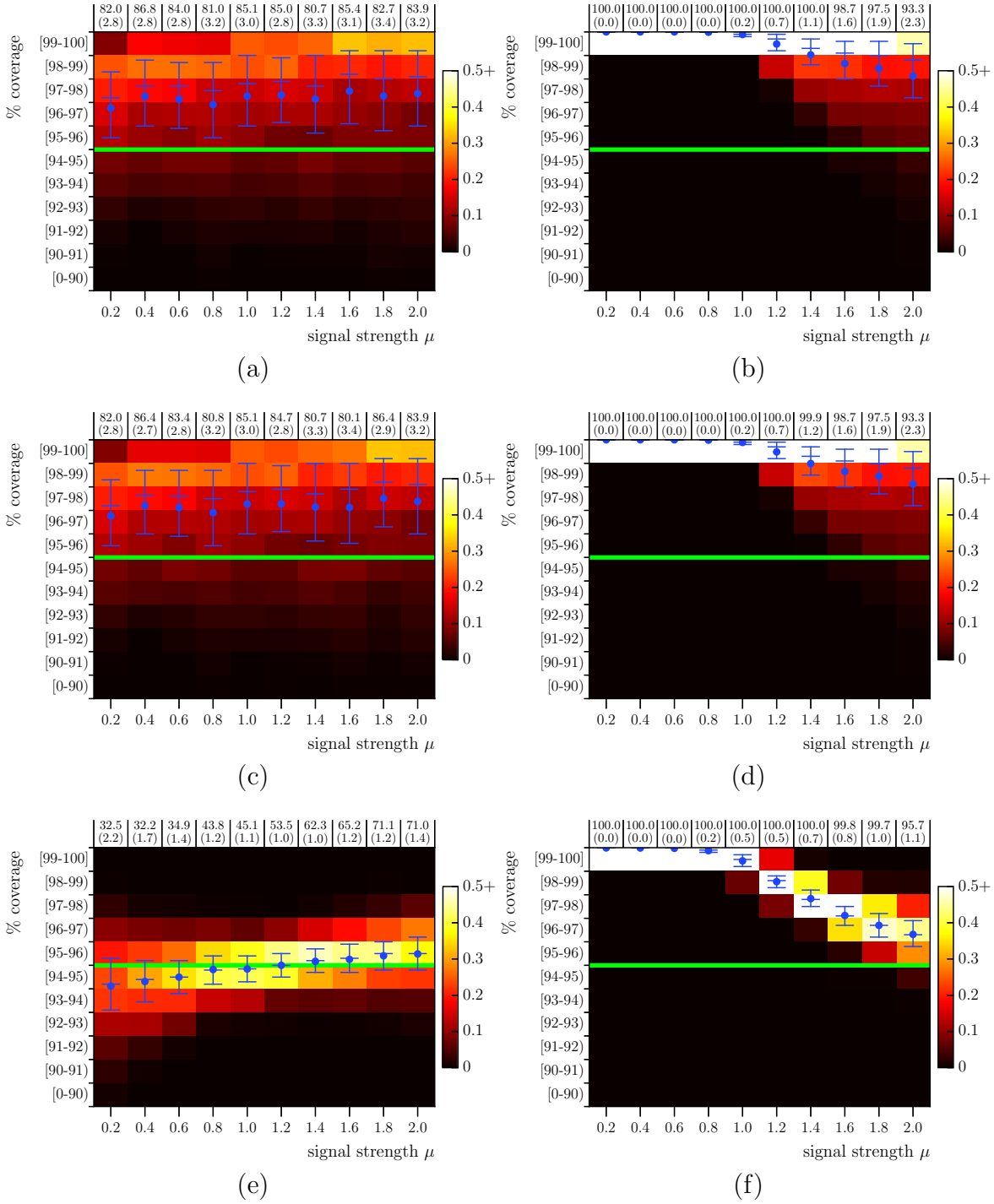


Figure D.17: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{D0} test statistic: (a) Q^N_{D0}, (b) Q^N_{D0} with $\mathrm{CL_S}$ method, (c) Q^P_{D0}, (d) Q^P_{D0} with $\mathrm{CL_S}$ method, (e) Q^F_{D0}, (f) Q^F_{D0} with $\mathrm{CL_S}$ method.

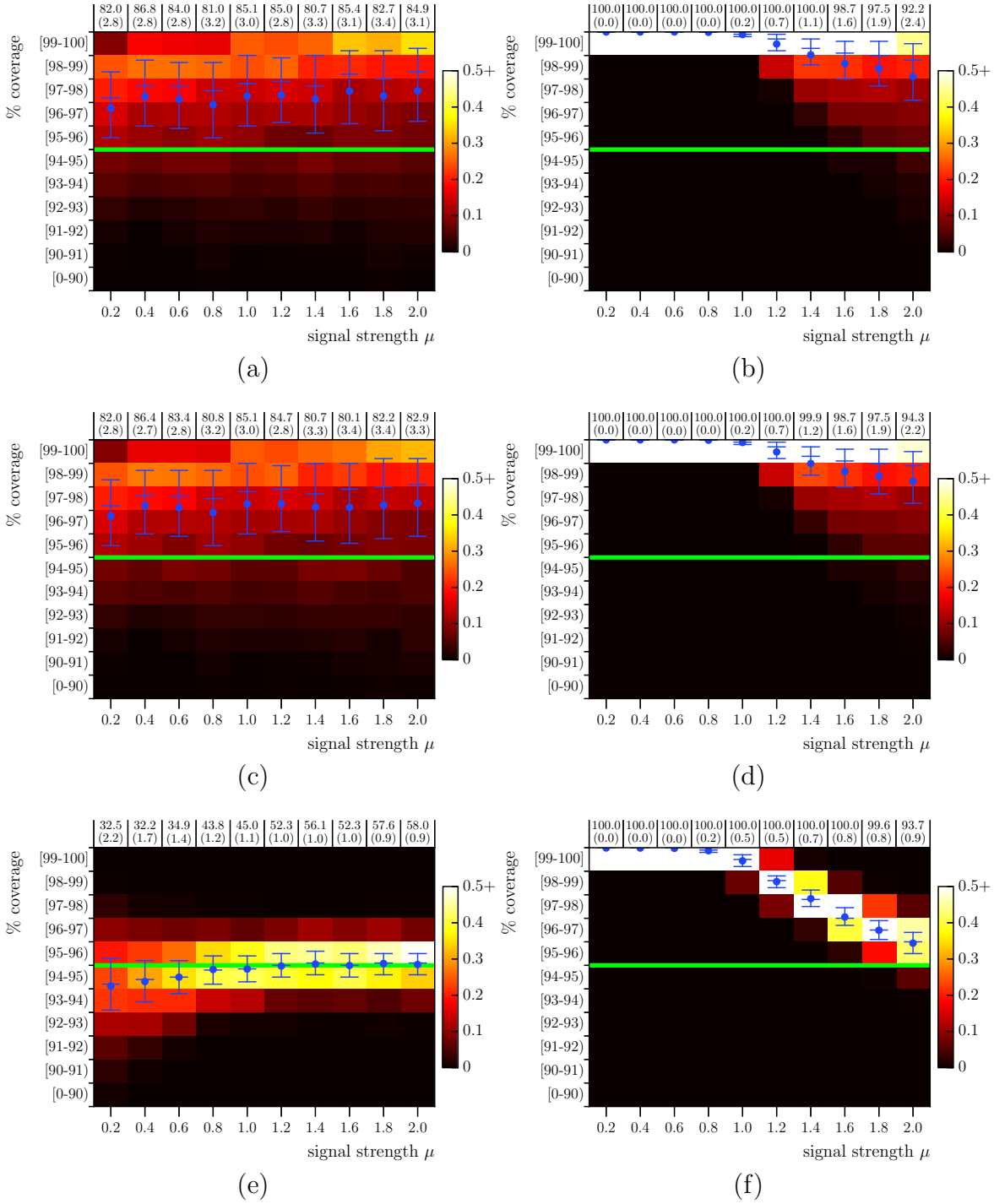


Figure D.18: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{AT} test statistic: (a) Q^N_{AT}, (b) Q^N_{AT} with $\mathrm{CL_S}$ method, (c) Q^P_{AT}, (d) Q^P_{AT} with $\mathrm{CL_S}$ method, (e) Q^F_{AT}, (f) Q^F_{AT} with $\mathrm{CL_S}$ method.

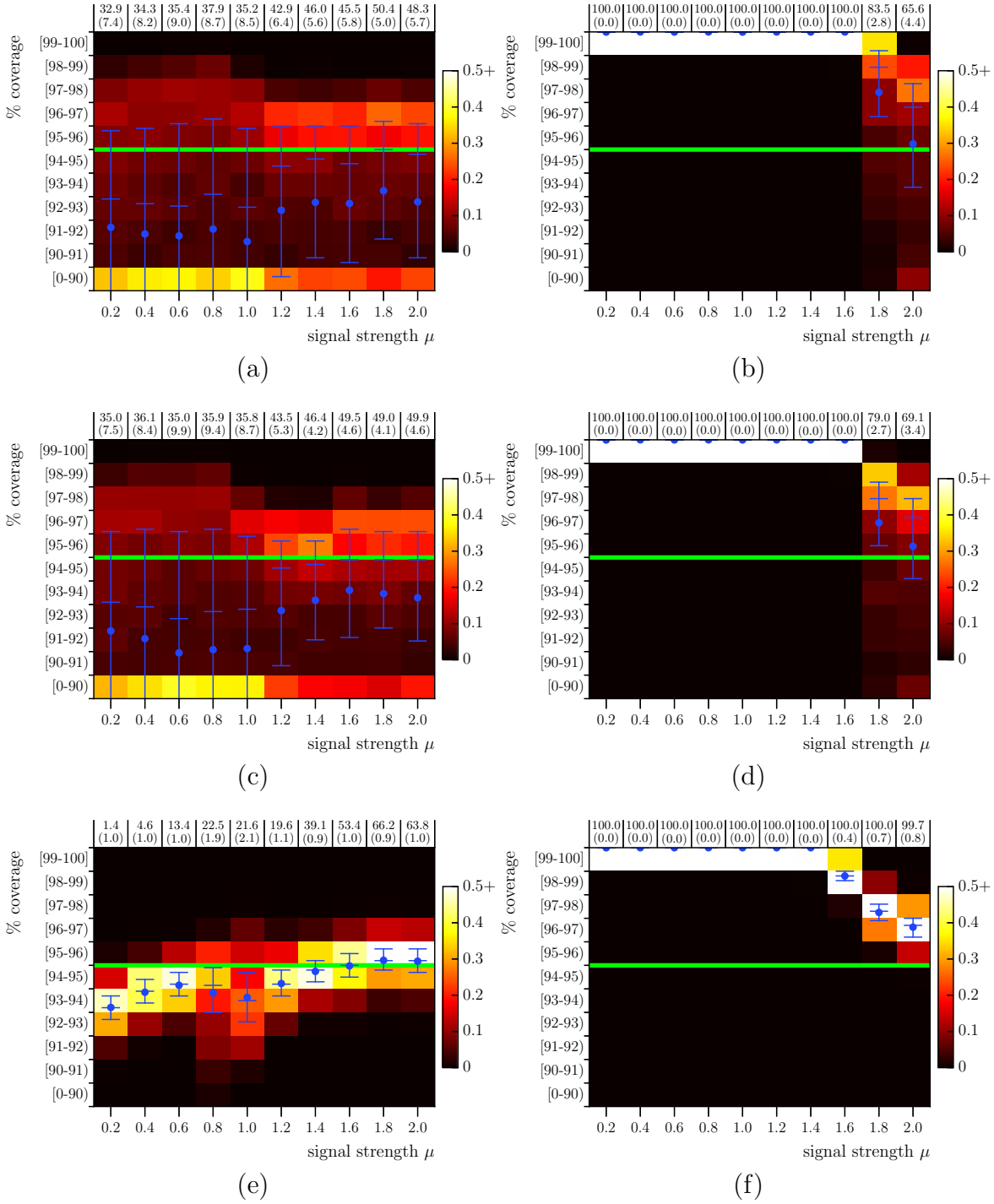


Figure D.19: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{AZ1} test statistic: (a) Q^N_{AZ1}, (b) Q^N_{AZ1} with $\mathrm{CL_S}$ method, (c) Q^P_{AZ1}, (d) Q^P_{AZ1} with $\mathrm{CL_S}$ method, (e) Q^F_{AZ1}, (f) Q^F_{AZ1} with $\mathrm{CL_S}$ method.

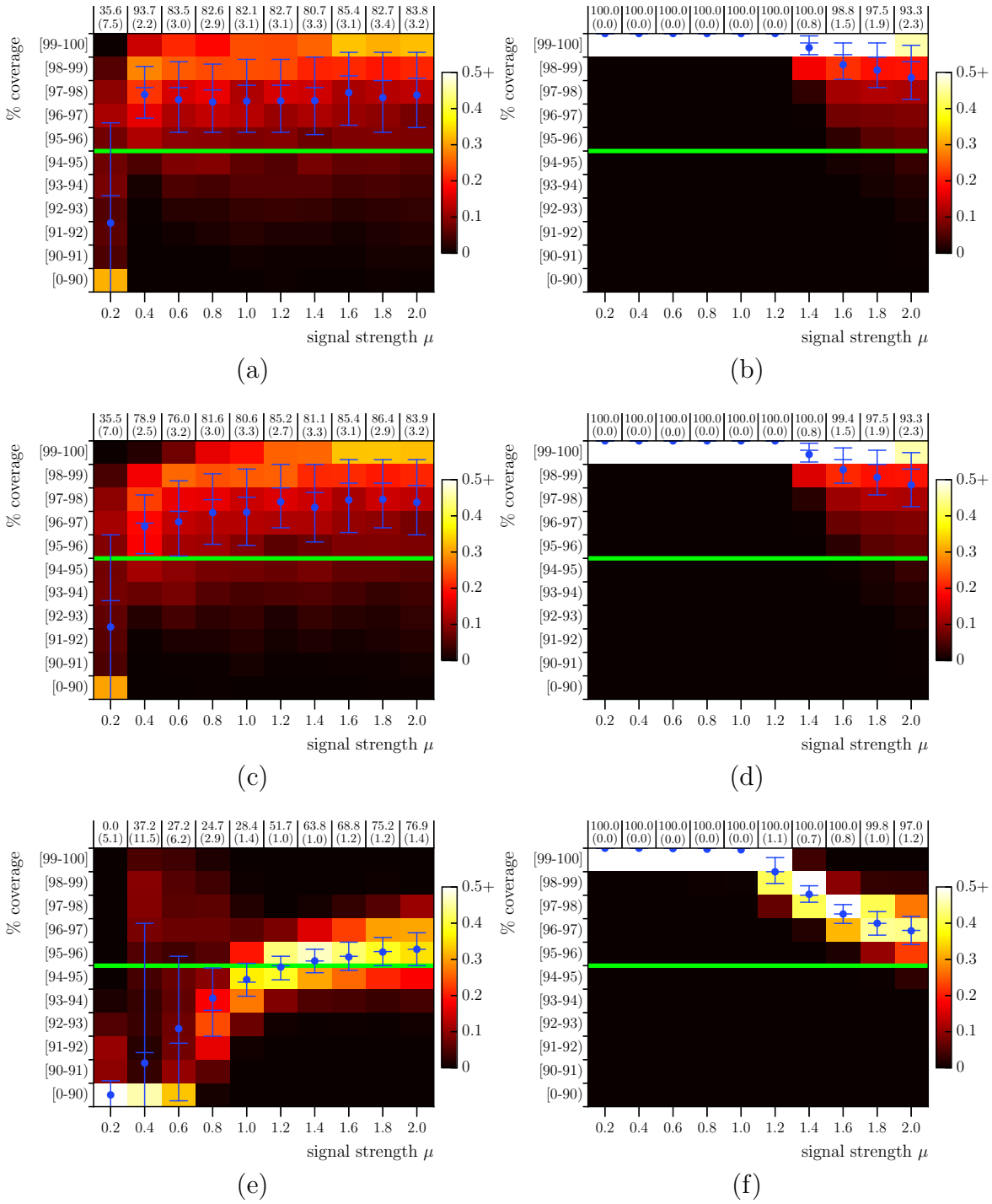


Figure D.20: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{AZ2} test statistic: (a) Q^N_{AZ2}, (b) Q^N_{AZ2} with $\mathrm{CL_S}$ method, (c) Q^P_{AZ2}, (d) Q^P_{AZ2} with $\mathrm{CL_S}$ method, (e) Q^F_{AZ2}, (f) Q^F_{AZ2} with $\mathrm{CL_S}$ method.

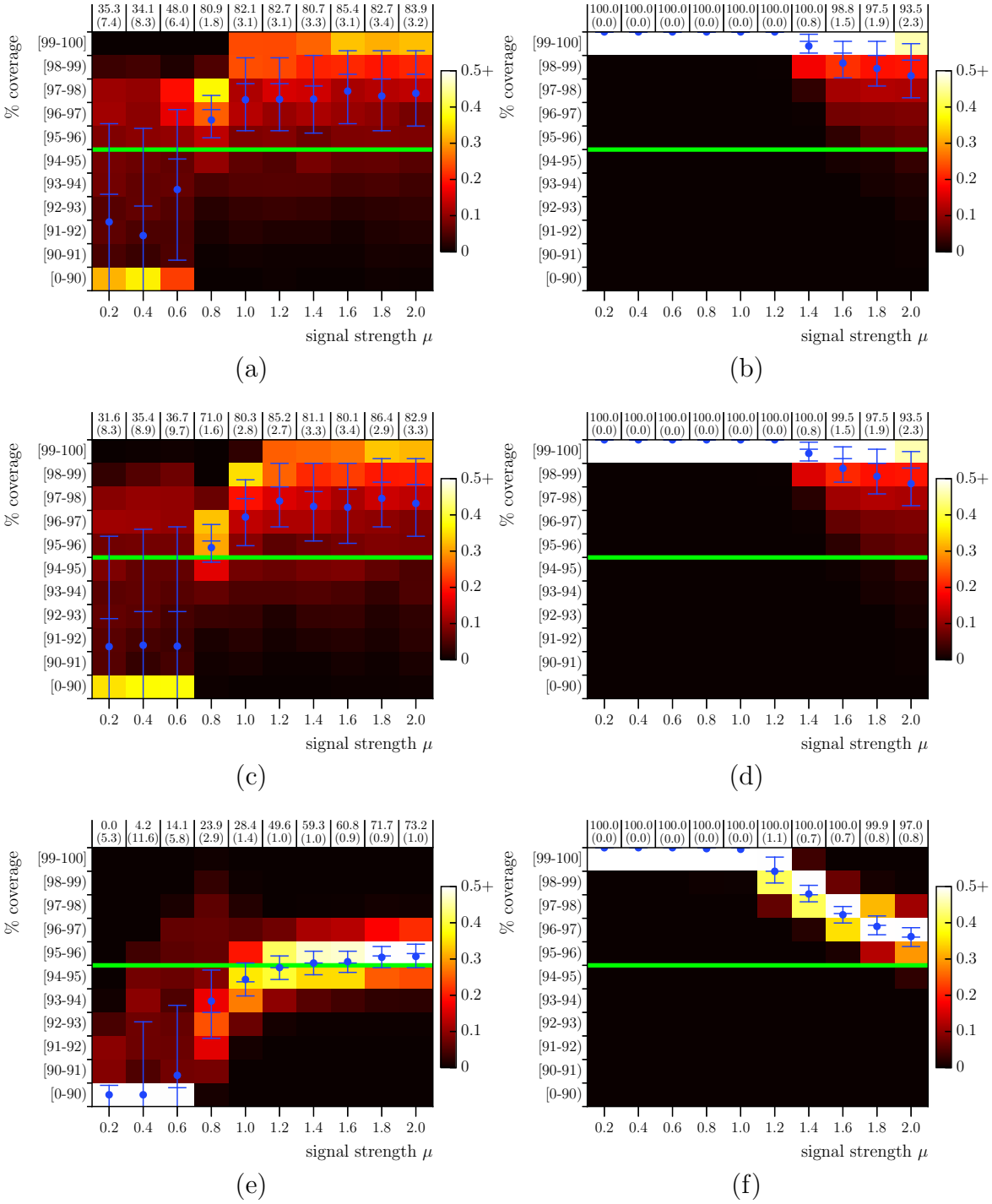


Figure D.21: Example 2 (bivariate) – The 2σ coverage study for the Q^X_{AZ3} test statistic: (a) Q^N_{AZ3}, (b) Q^N_{AZ3} with $\mathrm{CL_S}$ method, (c) Q^P_{AZ3}, (d) Q^P_{AZ3} with $\mathrm{CL_S}$ method, (e) Q^F_{AZ3}, (f) Q^F_{AZ3} with $\mathrm{CL_S}$ method.

References

[1] ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC." *Phys. Lett. B* **716** (2012), 1–29 — (***p. 26***).

[2] CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC." *Phys. Lett. B* **716** (2012), 30–61 — (***p. 26***).

[3] K. A. Olive et al. "Review of Particle Physics." *Chin. Phys. C* **38** 9 (2014), 090001 — (***pp. 26, 36, 42***).

[4] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Westview Press, 1995. ISBN: 9780201503975 — (***p. 26***).

[5] David Griffiths. *Introduction to Elementary Particles*. 2nd. Wiley-VCH, 2008. ISBN: 9783527406012 — (***pp. 26, 70***).

[6] Abraham Seiden. *Particle Physics: A Comprehensive Introduction*. Pearson Education, Inc., 2005. ISBN: 9780805387360 — (***p. 26***).

[7] Cliff P. Burgess and Guy D. Moore. *The Standard Model: A Primer*. Cambridge University Press, 2007. ISBN: 9780521860369 — (***p. 26***).

[8] W. N. Cottingham and D. A. Greenwood. *An Introduction to the Standard Model of Particle Physics*. Cambridge University Press, 1998. ISBN: 9780521581915 — (***p. 26***).

[9] T. Morii, C. S. Lim, and S. N. Mukherjee. *The Physics of the Standard Model and Beyond*. World Scientific Publishing, 2004. ISBN: 9789810245718 — (***pp. 26, 40***).

[10] K. Grotz and H. V. Klapdor. *The Weak Interaction in Nuclear, Particle and Astrophysics*. IOP Publishing Ltd., 1990. ISBN: 9780852743122 — (***p. 26***).

[11] Brian G. Wybourne. *Classical Groups for Physicists*. John Wiley & Sons, 1974. ISBN: 9780471965053 — (***p. 26***).

[12] Surender Verma. "Theoretical and Phenomenological Status of Neutrino Physics: A Brief Review." *Advances in High Energy Theory* 385968 (2015) — (***p. 27***).

[13] Marco Drewes. "Theoretical Status of Neutrino Physics." *PoS* (2015) — (***p. 27***).

[14] Jeffrey Goldstone, Abdus Salam, and Steven Weinberg. "Broken Symmetries." *Phys. Rev.* **127** (3 1962), 965–970 — (***p. 33***).

[15] Dan Green. *High P_T Physics at Hadron Colliders*. Cambridge University Press, 2005. ISBN: 9780521835097 — (***p. 41***).

[16] Michaelangelo L. Mangano. "Introduction to the theory of LHC collisions." *Theoretical Physics to Face the Challenge of LHC: Lecture Notes of the Les Houches Summer School: Volume 97, August 2011*. Ed. by Laurent Baulieu et al. Oxford University Press, 2015. ISBN: 9780198727965 — (***p. 41***).

[17] Bryan R. Webber. "Fragmentation and Hadronization." *Int. J. Mod. Phys. A* **15** (2000), 577–606 — (*p. 41*).

[18] Ignatios Antoniadis. "Mass hierarchy and physics beyond the Standard Model." *EPJ Web of Conferences* **71** (2014), 00007 — (*p. 40*).

[19] Walter D. Goldberger. "Brane Models and the Hierarchy Problem." ProQuest 275854876. Ph.D. Dissertation. California Institute of Technology, 2001 — (*p. 40*).

[20] Katherine Garrett and Gintaras Duda. "Dark Matter: A Primer." *Adv. Astron.* (2011), 968283 — (*p. 40*).

[21] Eldad Gildener. "Gauge-symmetry hierarchies." *Phys. Rev. D* **14** (1976), 1667–1672 — (*p. 42*).

[22] Leonard Susskind. "Dynamics of Spontaneous Symmetry Breaking in the Weinberg-Salam Theory." *Phys. Rev. D* **20** (1979), 2619–2625 — (*p. 42*).

[23] C.H.Llewellyn Smith and G.G. Ross. "The real gauge hierarchy problem." *Phys. Lett. B* **105** 1 (1981), 38–40. ISSN: 0370-2693 — (*p. 42*).

[24] Maarte C. Brak. "The Hierarchy Problem in the Standard Model and Little Higgs Theories." MA thesis. Utrecht University, 2004 — (*p. 42*).

[25] Stephen P. Martin. "Extra vectorlike matter and the lightest Higgs scalar boson mass in low-energy supersymmetry." *Phys. Rev. D* **81** (3 2010), 035004 — (*p. 42*).

[26] Stephen P. Martin and James D. Wells. "Implications of gauge-mediated supersymmetry breaking with vectorlike quarks and a ∼125 GeV Higgs boson." *Phys. Rev. D* **86** (3 2012), 035017 — (*p. 42*).

[27] Stephen P. Martin. "A Tumbling top quark condensate model." *Phys. Rev.* **D46** (1992), 2197–2202 — (*p. 42*).

[28] Andrea De Simone et al. "A first top partner hunter's guide." *J. High Energy Phys.* 4 (2013) — (*p. 42*).

[29] Tao Han et al. "Phenomenology of the little Higgs model." *Phys. Rev. D* **67** (9 2003), 095004 — (*p. 42*).

[30] Heather E. Logan. "Phenomenology of the littlest Higgs model." *Search for SUSY and unification. Proceedings, International Conference, 20 years of SUGRA,* 2003, pp. 589–605 — (*p. 42*).

[31] Lei Wang and Xiao-Fang Han. "The recent Higgs boson data and Higgs triplet model with vector-like quark." *Phys. Rev.* **D86** (2012), 095007 — (*p. 43*).

[32] Paolo Lodone. "Vector-like quarks in a 'composite' Higgs model." *J. High Energy Phys.* **12** (2008), 029 — (*p. 43*).

[33] Rabindra N. Mohapatra. "Gauged Flavor, Supersymmetry and Grand Unification." *AIP Conf. Proc.* **1467** (2012), 7–14 — (*p. 43*).

[34] Francisco del Aguila and Jose Santiago. "Signals from extra dimensions decoupled from the compactification scale." *J. High Energy Phys.* 03 (2002), 010 — (*p. 43*).

[35] Nima Arkani-Hamed, Andrew G. Cohen, and Howard Georgi. "Electroweak symmetry breaking from dimensional deconstruction." *Phys. Lett. B* **513** 1–2 (2001), 232–240 — (***p. 43***).

[36] Francisco del Aguila, Jose Santiago, and Manolo Perez-Victoria. "Observable contributions of new exotic quarks to quark mixing." *J. High Energy Phys.* 09 (2000), 011 — (***p. 43***).

[37] J. A. Aguilar-Saavedra et al. "Handbook of vectorlike quarks: Mixing and single production." *Phys. Rev. D* **88** (9 2013), 094010 — (***pp. 43 sq., 46***).

[38] F. del Aguila, M. Perez-Victoria, and Jose Santiago. "Effective description of quark mixing." *Phys. Lett.* **B492** (2000), 98–106 — (***p. 43***).

[39] F. Del Aguila and M.J. Bowick. "The possibility of new fermions with $\Delta I = 0$ mass." *Nucl. Phys. B* **224** 1 (1983), 107–136 — (***p. 43***).

[40] S. L. Glashow, J. Iliopoulos, and L. Maiani. "Weak Interactions with Lepton-Hadron Symmetry." *Phys. Rev. D* **2** (1970), 1285–1292 — (***p. 44***).

[41] ATLAS Collaboration. "Search for pair and single production of new heavy quarks that decay to a Z boson and a third-generation quark in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." *J. High Energy Phys.* **104** 11 (2014) — (***pp. 44 sq., 49 sq., 77, 103, 105 sq., 170, 224***).

[42] J. A. Aguilar-Saavedra. "Identifying top partners at LHC." *J. High Energy Phys.* **11** (2009), 030 — (***pp. 44, 74***).

[43] Robert Contino, Roberto Franceschini, and Michael Peskin. "A Module for Vectorlike Heavy Top Quarks." Nov. 2010. URL: `www.lhcnewphysics.org/wiki/images/6/60/Tprime.pdf` — (***p. 44***).

[44] J. A. Aguilar-Saavedra. "Mixing with vector-like quarks: constraints and expectations." *EPJ Web Conf.* **60** (2013), 16012 — (***p. 44***).

[45] ATLAS Collaboration. "Search for production of vector-like quark pairs and of four top quarks in the lepton-plus-jets final state in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." *J. High Energy Phys.* **08** (2015), 105 — (***pp. 44, 49 sq.***).

[46] Charalampos Anastasiou, Elisabetta Furlan, and Jose Santiago. "Realistic Composite Higgs Models." *Phys. Rev.* **D79** (2009), 075003 — (***p. 46***).

[47] Natascia Vignaroli. "Early discovery of top partners and test of the Higgs nature." *Phys. Rev.* **D86** (2012), 075017 — (***p. 46***).

[48] Roberto Contino and Geraldine Servant. "Discovering the top partners at the LHC using same-sign dilepton final states." *J. High Energy Phys.* **06** (2008), 026 — (***p. 46***).

[49] CMS Collaboration. "Search for a Vectorlike Quark with Charge 2/3 in $t + Z$ Events from pp Collisions at $\sqrt{s} = 7$ TeV." *Phys. Rev. Lett.* **107** (27 2011), 271802 — (***pp. 49 sq.***).

[50] ATLAS Collaboration. "Search for Pair Production of a Heavy Up-Type Quark Decaying to a W Boson and a b Quark in the lepton+jets Channel with the ATLAS Detector." *Phys. Rev. Lett.* **108** (26 2012), 261802 — (***pp. 49 sq.***).

[51] CMS Collaboration. "Search for heavy, top-like quark pair production in the dilepton final state in pp collisions at $\sqrt{s} = 7$ TeV." *Phys. Lett.* **B716** (2012), 103–121 — (***pp. 49 sq.***).

[52] ATLAS Collaboration. "Search for pair production of heavy top-like quarks decaying to a high-p_T W boson and a b quark in the lepton plus jets final state at $\sqrt{s} = 7$ TeV with the ATLAS detector." *Phys. Lett. B* **718** 4–5 (2013), 1284–1302 — (***pp. 49 sq.***).

[53] ATLAS Collaboration. "Search for Pair Production of a New b' Quark that Decays into a Z Boson and a Bottom Quark with the ATLAS Detector." *Phys. Rev. Lett.* **109** 7 (2012), 071801 — (***pp. 49 sq.***).

[54] Andre Milton Bach. "Search for Pair Production of a New b' Quark that Decays into a Z Boson and a Bottom Quark with the ATLAS Detector at the LHC." CERN-THESIS-2012-207. Ph.D. Dissertation. U.C. Berkeley, 2012 — (***pp. 49 sq., 121, 123***).

[55] CMS Collaboration. "Search for heavy quarks decaying into a top quark and a W or Z boson using lepton + jets events in pp collisions at $\sqrt{s} = 7$ TeV." *J. High Energy Phys.* **01** (2013), 154 — (***pp. 49 sq.***).

[56] CMS Collaboration. "Inclusive search for a vector-like T quark with charge $\frac{2}{3}$ in pp collisions at $\sqrt{s} = 8$ TeV." *Phys. Lett. B* **729** (2014), 149–171 — (***pp. 49 sq., 121, 123***).

[57] CMS Collaboration. "Search for vector-like T quarks decaying to top quarks and Higgs bosons in the all-hadronic channel using jet substructure." *J. High Energy Phys.* **06** (2015), 080 — (***pp. 49 sq.***).

[58] CMS Collaboration. "Search for pair-produced vector-like B quarks in proton-proton collisions at $\sqrt{s} = 8$ TeV." *Submitted to Phys. Rev. D* arXiv:1507.07129. (July 2015) — (***pp. 49 sq.***).

[59] ATLAS Collaboration. "Search for vector-like B quarks in events with one isolated lepton, missing transverse momentum and jets at $\sqrt{s} = 8$ TeV with the ATLAS detector." *Phys. Rev. D* **91** (2015), 112011 — (***pp. 49 sq., 121, 123***).

[60] ATLAS Collaboration. "Analysis of events with b-jets and a pair of leptons of the same charge in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." *J. High Energy Phys.* 10 (2015) — (***pp. 49 sq.***).

[61] Clare S. Bernard. "Standard Model and Exotic Physics with the Top Quark at ATLAS." ProQuest 1626364439. Ph.D. Dissertation. Boston University, 2014 — (***pp. 49, 77, 103, 106***).

[62] Juan Pedro Araque Espinosa. "Search for heavy fermions with the ATLAS experiment at the LHC collider." CERN-THESIS-2015-330. Ph.D. Dissertation. Minho U., 2015 — (***pp. 49, 77, 90, 103, 106***).

[63] J. P. Araque et al. *Search for pair production of new heavy quarks that decay to a Z boson and a third generation quark in pp collisions at* $\sqrt{s} = 8$ *TeV with the ATLAS detector*. Tech. rep. ATL-COM-PHYS-2014-001. 2014 — (***pp. 49, 77, 85, 91, 98, 103, 106***).

[64] ATLAS collaboration. *Search for pair and single production of new heavy quarks that decay to a Z boson and a third generation quark in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2014-036. 2014 — (***pp. 49, 77, 103, 106***).

[65] Oliver S. Bruning et al. "LHC Design Report Vol.1: The LHC Main Ring." CERN-2004-003-V1 (2004) — (***p. 53***).

[66] O. Buning et al. "LHC Design Report. 2. The LHC infrastructure and general services." CERN-2004-003-V-2 (2004) — (***p. 53***).

[67] M. Benedikt et al. "LHC Design Report. 3. The LHC injector chain." CERN-2004-003-V-3 (2004) — (***pp. 53, 55***).

[68] Mario Conte and W. MacKay William. *An Introduction to the Physics of Particle Accelerators*. World Scientific Publishing Co., 1991. ISBN: 9789810208138 — (***p. 53***).

[69] Donald A. Edwards and Michael J. Syphers. *An introduction to the Physics of High Energy Accelerators*. Wiley, 1993 — (***pp. 53, 275, 533***).

[70] Werner Herr and Bruno Muratori. "Concept of Luminosity." *Intermediate accelerator physics*. Ed. by D. Brandt. CERN-2006-002. CERN Accelerator School, 2006 — (***p. 54***).

[71] C. E. Hill. "Ion and electron sources." CERN-PS-94-36-HI (1994) — (***p. 54***).

[72] C. E. Hill et al. "Tests of the CERN proton linac performance for LHC–type beams." *Linac. Proceedings, 20th International Conference, Linac 2000, Monterey, USA, August 21-25, 2000. Vol. 1,2: C000821*. 2000, TUD17 — (***p. 54***).

[73] A. M. Lombardi. "The radio frequency quadrupole (RFQ)" (2006). CERN-2006-012.201 — (***p. 54***).

[74] NA61/SHINE Collaboration. "NA61/SHINE facility at the CERN SPS: beams and detector system." *J. Instr.* **9** (2014), P06005 — (***p. 54***).

[75] K. Hanke. "Past and present operation of the CERN PS Booster." *Int. J. Mod. Phys. A* **28** (2013), 1330019 — (***p. 55***).

[76] Jean-Paul Burnet et al. *Fifty years of the CERN Proton Synchrotron: Volume 1*. 2011 — (***p. 55***).

[77] A. Blas et al. "Studies on single batch transfer of the LHC type beams between the CERN PS Booster and the PS." CERN-ATS-2009-037 (2009) — (***p. 55***).

[78] Alessio Mereghetti. "Performance Evaluation of the SPS Scraping System in View of the High Luminosity LHC." CERN-THESIS-2015-116. Ph.D. Dissertation. Manchester U., 2015 — (***p. 55***).

[79] Mike Lamont. "Status of the LHC." *J. Phys. Conference Series* **455** 1 (2013), 012001 — (***p. 55***).

[80] P. Baudrenghein and T. Mastoridis. "Performances and Future Plans of the LHC RF." *Advanced Beam Dynamics Workshop on High-Intensity and High-Brightness Hadron Beams Proceedings*. THO1C04. 2012 — (***p. 55***).

[81] Lucio Rossi. "Superconductivity: its role, its success and its setbacks in the Large Hadron Collider of CERN." *Superconductor Sci. Tech.* **23** 3 (2010), 034001 — (***pp. 55 sq.***).

[82] B. J. Holzer. "Introduction to Transverse Beam Dynamics." *Superconductivity for Accelerators*. Ed. by R. Bailey. CERN-2014-005. CERN Accelerator School, 2013 — (***p. 56***).

[83] ATLAS Collaboration. "Improved luminosity determination in pp collisions at $sqrt(s)$ = 7 TeV using the ATLAS detector at the LHC." *Eur. Phys. J. C* **73** 8 (2013), 2518 — (***pp. 56, 62***).

[84] Kunihiro Nagano and the ATLAS Collaboration. "Algorithms, performance, development of the ATLAS High-Level trigger." *J. Phys. Conference Series* **513** 1 (2014), 012026 — (***pp. 56, 64***).

[85] ATLAS Collaboration. *ATLAS inner detector: Technical Design Report, 1*. CERN, 1997 — (***pp. 58 sq.***).

[86] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider." *J. Instr.* **3** (2008), S08003 — (***pp. 59 sqq.***).

[87] G. Aad et al. "ATLAS pixel detector electronics and sensors." *J. Instr.* **3** 07 (2008), P07007 — (***p. 59***).

[88] A. Abdesselam et al. "The barrel modules of the ATLAS semiconductor tracker." *Nucl. Instr. Meth. Phys. Res. A* **568** 2 (2006), 642–671 — (***p. 59***).

[89] The ATLAS TRT Collaboration. "The ATLAS Transition Radiation Tracker (TRT) proportional drift tube: design and performance." *J. Instr.* **3** 02 (2008), P02013 — (***p. 59***).

[90] Glenn F. Knoll. *Radiat. Detection Meas.* 4th ed. Wilely, 2010. ISBN: 9780470131480 — (***p. 59***).

[91] ATLAS Collaboration. *ATLAS central solenoid: Technical Design Report*. CERN, 1997 — (***p. 59***).

[92] ATLAS Collaboration. *ATLAS magnet system: Technical Design Report, 1*. CERN, 1997 — (***pp. 59, 63***).

[93] ATLAS Collaboration. *ATLAS liquid-argon calorimeter: Technical Design Report*. 1996 — (***p. 60***).

[94] ATLAS Collaboration. *ATLAS tile calorimeter: Technical Design Report*. 1996 — (***p. 62***).

[95] Z. Marshall. *Simulation of Pile-up in the ATLAS Experiment*. Tech. rep. ATL-SOFT-PROC-2013-030. 2013 — (***p. 62***).

[96] ATLAS Collaboration. "ATLAS muon spectrometer: Technical design report." CERN-LHCC-97-22, ATLAS-TDR-10 (1997) — (***p. 63***).

[97] S. Horvat et al. "Precision Drift Chambers for the ATLAS Muon Spectrometer." *Proceedings of the XXIII International Conference on Physics in Collision*. Ed. by S. Riemann and W. Lohmann. Deutsches Elektronensynchrotron, 2003. ISBN: 978839171411 — (**p. 63**).

[98] Adele Rimoldi. "The ATLAS muon trigger chamber system." *Nucl. Instr. Meth. Phys. Res. A* **409** 1–3 (1998), 669–674 — (**p. 64**).

[99] R. Santonico and R. Cardarelli. "Development of resistive plate counters." *Nucl. Instr. Meth. Phys. Res.* **187** 2 (1981), 377–380 — (**p. 64**).

[100] G. Aielli et al. "The RPC first level muon trigger in the barrel of the ATLAS experiment." *Nucl. Phys. B, Proc. Suppl.* **158** (2006), 11–15 — (**p. 64**).

[101] S. Tanaka et al. *Techniques developed for the ATLAS Thin Gap Chambers mass production in Japan*. Tech. rep. ATL-MUON-2004-015. 2004 — (**p. 64**).

[102] ATLAS Collaboration. *ATLAS Computing: technical design report*. CERN-LHCC-2005-022. 2005 — (**p. 64**).

[103] ATLAS Collaboration. "ATLAS trigger performance: Status report." CERN-LHCC-98-15 (1998) — (**p. 64**).

[104] Will Buttinger and the ATLAS Collaboration. "The ATLAS Level-1 Trigger System." *J. Phys. Conference Series* **396** 1 (2012), 012010 — (**p. 64**).

[105] Peter Jenni et al. *ATLAS high-level trigger, data-acquisition and controls: Technical Design Report*. CERN-LHCC-2003-022. 2003 — (**p. 65**).

[106] Kirill Prokofiev. "Reconstruction of primary vertices in pp collisions at energies of 900 GeV and 7 TeV with the ATLAS detector." *Hadron collider physics. Proceedings, 22nd Conference*. 2010 — (**p. 66**).

[107] ATLAS Collaboration. *Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s}$ =7 TeV in the ATLAS experiment*. Tech. rep. ATLAS-CONF-2010-069. 2010 — (**p. 66**).

[108] K. Grimm et al. *Methods to quantify the performance of the primary vertex reconstruction in the ATLAS experiment under high luminosity conditions*. Tech. rep. ATL-PHYS-PROC-2012-107. 2012 — (**p. 66**).

[109] Rick Field. "Min-Bias and the Underlying Event at the LHC." *Proceedings, 31st International Conference on Physics in collisions*. PIC2011. 2012 — (**p. 66**).

[110] ATLAS Collaboration. "Electron reconstruction and identification efficiency measurements with the ATLAS detector using the 2011 LHC proton-proton collision data." *Eur. Phys. J. C* **74** 7 (2014), 2941 — (**pp. 67, 95**).

[111] Maximilian Goblirsch-Kolb. "Muon Reconstruction Efficiency, Momentum Scale and Resolution in pp Collisions at 8 TeV with ATLAS." *Proceedings, 2nd Conference on Large Hadron Collider Physics Conference*. ATL-PHYS-PROC-2014-110. 2014 — (**pp. 68, 95**).

[112] Stefan Hoche. "Introduction to parton-shower event generators." *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders*. SLAC-PUB-16160. 2014 — (***pp. 69, 73 sq.***).

[113] Jason Veatch. "A Search For New Physics Using $t\bar{t}$ Resonances in the Lepton Plus Jets Channel at $\sqrt{s} = 8$ TeV in 20 fb^{-1} of pp Collision Data at the ATLAS Experiment." 9781339020129. Ph.D. Dissertation. University of Arizona, 2015 — (***p. 69***).

[114] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k$_t$ jet clustering algorithm." *J. High Energy Phys.* **04** (2008) — (***p. 69***).

[115] CMS Collaboration. "Jet and underlying event properties as a function of charged-particle multiplicity in proton–proton collisions at $\sqrt{s} = 7$ TeV." *Eur. Phys. J. C* **73** (12 2013), 2674 — (***p. 69***).

[116] ATLAS Collaboration. "The underlying event in jet events at 7 TeV with the ATLAS experiment." ATLAS-CONF-2012-164 (2012) — (***p. 69***).

[117] S. D. Ellis et al. "Jets in hadron-hadron collisions." *Prog. Part. Nucl. Phys.* **60** (2008) — (***p. 69***).

[118] ATLAS Collaboration. *Pile-up subtraction and suppression for jets in ATLAS*. Tech. rep. ATLAS-CONF-2013-083. 2013 — (***p. 69***).

[119] ATLAS Collaboration. *Calibration of the performance of b-tagging for c and light-flavour jets in the 2012 ATLAS data*. Tech. rep. ATLAS-CONF-2014-046. 2014 — (***p. 70***).

[120] ATLAS Collaboration. *Calibration of b-tagging using dileptonic top pair events in a combinatorial likelihood approach with the ATLAS experiment*. Tech. rep. ATLAS-CONF-2014-004. 2014 — (***pp. 70, 95***).

[121] ATLAS Collaboration. *Commissioning of the ATLAS high-performance b-tagging algorithms in the 7 TeV collision data*. Tech. rep. ATLAS-CONF-2011-102. 2011 — (***p. 70***).

[122] Ringaile Placakyte. "Parton Distribution Functions." *Proceedings, 31st International Conference on Physics in collisions*. 2011 — (***p. 71***).

[123] John M. Campbell, J. W. Huston, and W. J. Stirling. "Hard Interactions of Quarks and Gluons: A Primer for LHC Physics." *Rept. Prog. Phys.* **70** (2007), 89 — (***p. 71***).

[124] Johan Alwall et al. "MadGraph 5 : Going Beyond." *J. High Energy Phys.* **06** (2011), 128 — (***pp. 71, 73 sq.***).

[125] Arnulf Quadt. *Top Quark Physics at Hadron Colliders*. Springer, 2007. ISBN: 9783540710592 — (***p. 71***).

[126] Michelangelo L. Mangano et al. "ALPGEN, a generator for hard multiparton processes in hadronic collisions." *J. High Energy Phys.* **07** (2003) — (***pp. 72, 74***).

[127] GEANT4 Collaboration. "GEANT4: a simulation toolkit." *Nucl. Instrum. Meth. A* **506** (2003), 250–303 — (***pp. 72, 74***).

[128] T. Gleisberg et al. "Event generation with SHERPA 1.1." *J. High Energy Phys.* **02** (2009), 007 — (*pp. 72 sq.*).

[129] Hung-Liang Lai et al. "New parton distributions for collider physics." *Phys. Rev. D* **82** (2010), 074024 — (*pp. 72 sq.*).

[130] Stefano Catani et al. "Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO." *Phys. Rev. Lett.* **103** (2009), 082001 — (*p. 72*).

[131] ATLAS Collaboration. *ATLAS tunes of PYTHIA 6 and Pythia 8 for MC11*. Tech. rep. ATL-PHYS-PUB-2011-009. 2011 — (*p. 72*).

[132] John M. Campbell and R. Keith Ellis. "An Update on vector boson pair production at hadron colliders." *Phys. Rev. D* **60** (1999), 113006 — (*p. 73*).

[133] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. "PYTHIA 6.4 Physics and Manual." *J. High Energy Phys.* **05** (2006), 026 — (*pp. 73 sq.*).

[134] S. Kretzer et al. "CTEQ6 Parton Distributions with Heavy Quark Mass Effects." *Phys. Rev. D* **69** (2004), 114005 — (*pp. 73 sq.*).

[135] M. V. Garzelli et al. "t $\bar{t}$ W^{+-} and t $\bar{t}$ Z Hadroproduction at NLO accuracy in QCD with Parton Shower and Hadronization effects." *J. High Energy Phys.* **11** (2012), 056 — (*p. 73*).

[136] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method." *J. High Energy Phys.* **11** (2007), 070 — (*p. 73*).

[137] Stefano Frixione and Bryan R. Webber. "Matching NLO QCD computations and parton shower simulations." *J. High Energy Phys.* **06** (2002), 029 — (*p. 74*).

[138] Stefano Frixione et al. "Single-top hadroproduction in association with a W boson." *J. High Energy Phys.* **07** (2008), 029 — (*p. 74*).

[139] G. Corcella et al. "HERWIG 6.5 release note." CERN-TH-2002-270 (2002) — (*p. 74*).

[140] Borut Paul Kersevan and Elzbieta Richter-Was. "The Monte Carlo event generator AcerMC versions 2.0 to 3.8 with interfaces to PYTHIA 6.4, HERWIG 6.5 and ARIADNE 4.1." *Comput. Phys. Commun.* **184** (2013), 919–985 — (*p. 74*).

[141] Nikolaos Kidonakis. "Differential and total cross sections for top pair and single top production." *Proceedings, 20th International Workshop on Deep-Inelastic Scattering and Related Subjects*. DESY-PROC-2012-02. 2012, pp. 831–834 — (*p. 74*).

[142] J. A. Aguilar-Saavedra. *Protos - PROgram for TOp Simulations*. URL: `http://jaguilar.web.cern.ch/jaguilar/protos/` (visited on 10/22/2016) — (*p. 74*).

[143] J. A. Aguilar-Saavedra. "Single top quark production at LHC with anomalous Wtb couplings." *Nucl. Phys. B* **804** (2008), 160–192 — (*p. 74*).

[144] A. D. Martin et al. "Parton distributions for the LHC." *Eur. Phys. J.* **C63** (2009), 189–285 — (*p. 74*).

[145] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders." *Comput. Phys. Commun.* **185** (2014), 2930 — (**p. 74**).

[146] Torbjorn Sjostrand. "PYTHIA 8 Status Report." *Proceedings, workshop: HERA and the LHC workshop series on the implications of HERA for LHC physics*. 2008, pp. 726–732 — (**p. 74**).

[147] ATLAS Collaboration. "The ATLAS Simulation Infrastructure." *Eur. Phys. J. C* **70** (2010), 823–874 — (**p. 74**).

[148] ATLAS Collaboration. *The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim*. Tech. rep. ATL-PHYS-PUB-2010-013. 2010 — (**p. 74**).

[149] J. P. Araque et al. *Search for pair production of new heavy quarks that decay to a Z boson and a third generation quark in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*. Tech. rep. ATL-COM-PHYS-2013-609. 2013 — (**pp. 77, 103, 106**).

[150] ATLAS collaboration. *Search for pair production of new heavy quarks that decay to a Z boson and a third generation quark in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2013-056. 2013 — (**pp. 77, 103, 106**).

[151] A. Bundock et al. *Measurement of Zb and Zbb cross sections with 4.6 fb^{-1} of 7 TeV ATLAS data*. Tech. rep. ATL-COM-PHYS-2012-1504. 2012 — (**p. 89**).

[152] ATLAS Collaboration. "Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector." *Eur. Phys. J. C* **75** (2015), 17 — (**p. 95**).

[153] ATLAS Collaboration. "Jet energy resolution in proton-proton collisions at $\sqrt{s} = 7$ TeV recorded in 2010 with the ATLAS detector." *Eur. Phys. J. C* **73** 3 (2013), 2306 — (**p. 95**).

[154] J. S. Conway. "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra." *PHYSTAT Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*. Ed. by Harrison B. Prosper. 2011 — (**pp. 103, 392, 512**).

[155] Philip E. Tetlock. *Expert Political Judgment: How Good Is It? How Can We Know?* Princeton University Press, Inc., 2005. ISBN: 9781400830312 — (**p. 106**).

[156] E. Warren Buffett. *Letter to Shareholders*. Feb. 9, 2009. URL: `www.berkshirehathaway.com/letters/letters.html` (visited on 02/22/2014) — (**p. 106**).

[157] Ilya Narsky and Frank C. Porter. *Statistical Analysis Techniques In Particle Physics*. Wiley-VCH, 2014. ISBN: 9783527410866 — (**pp. 106, 110, 120, 141, 251**).

[158] Andreas Hoecker et al. "TMVA: Toolkit for Multivariate Data Analysis." *PoS* **ACAT** (2007), 040. arXiv: `physics/0703039` — (**pp. 106, 112, 114**).

[159] A. Hoecker et al. *TMVA—Toolkit for Multivariate Data Analysis with ROOT, Users Guide*. arXiv:physics/0703039. 2013 — (**pp. 106, 124, 442**).

[160] Rene Brun and Fons Rademakers. "ROOT—An object oriented data analysis framework." *Nucl. Instr. Meth. Phys. A* **389** 1–2 (1997). New Computing Techniques in Physics Research V, 81–86 — (***p. 106***).

[161] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python." *J. Machine Learning Res.* **12** (2011), 2825–2830 — (***p. 106***).

[162] Eric Jones, Travis Oliphant, Pearu Peterson, et al. *SciPy: Open source scientific tools for Python*. 2001–. URL: `www.scipy.org` (visited on 03/08/2015) — (***p. 106***).

[163] J. D. Hunter. "Matplotlib: A 2D graphics environment." *Comput. Sci. & Eng.* **9** 3 (2007), 90–95 — (***p. 106***).

[164] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012. ISBN: 9780262018029 — (***pp. 110 sqq., 116 sq.***).

[165] Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning From Data*. AMLBook, 2012. ISBN: 9781600490069 — (***pp. 110 sq., 115, 172***).

[166] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1995. ISBN: 9781475724424 — (***p. 110***).

[167] Patrick Stalph. *Analysis and Design of Machine Learning Techniques: Evolutionary Solutions for Regression, Prediction, and Control Problems*. Springer Viewag, 2014. ISBN: 9783658049362 — (***p. 110***).

[168] Jason Bell. *Machine Learning: Hands–On for Developers and Technical Professionals*. John Wiley & Sons, Inc., 2015. ISBN: 9781118889060 — (***p. 110***).

[169] Leo Breiman. "Bagging Predictors." *Machine Learning* **24** 2 (1996), 123–140 — (***p. 111***).

[170] Robert E. Schapire. "The strength of weak learnability." *Machine Learning* **5** 2 (1990), 197–227 — (***p. 113***).

[171] Yoav Freund and Robert E. Schapire. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *J. Computer and System Sci.* **55** 1 (1997), 119–139 — (***p. 114***).

[172] Robert E. Schapire and Yoav Freund. *Boosting: Foundations and Algorithms*. MIT Press, 2012. ISBN: 9780262017183 — (***p. 114***).

[173] Artur J. Ferreira and Mario A. T. Figueiredo. "Boosting Algorithms: A Review of Methods, Theory, and Applications." *Ensemble Machine Learning: Methods and Applications*. Ed. by Cha Zhang and Yunqian Ma. Springer, 2012. ISBN: 9781441993250 — (***p. 114***).

[174] Jerome H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine." *Ann. Stat.* **29** 5 (2001), 1189–1232 — (***p. 114***).

[175] Kurt Hornik. "Approximation capabilities of multilayer feedforward networks." *Neural Networks* **4** 2 (1991), 251–257 — (***p. 116***).

[176] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd. Springer-Verlag, 2009. ISBN: 9780387848570 — (***pp. 116, 120, 126***).

[177] Badi H. Baltagi. *Econometrics*. 5th. Springer-Verlag, 2011. ISBN: 9783642200588 — (**p. 118**).

[178] George A. F. Serber and Alan J. Lee. *Linear Regression Analysis*. 2nd. John Wiley & Sons, Inc., 2003. ISBN: 9780471415405 — (**p. 118**).

[179] T. S. Breusch and A. R. Pagan. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica* **47** 5 (1979), 1287–1294 — (**pp. 119, 176 sq., 486 sqq.**).

[180] James G MacKinnon and Halbert White. "Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties." *J. Econ.* **29** 3 (1985), 305–325 — (**p. 119**).

[181] J. Scott Long and Laurie H. Ervin. "Correcting for Heteroscedasticity with Heteroscedasticity Consistent Standard Errors in the Linear Regression Model: Small Sample Considerations." *Working Paper, Indiana University, Bloomington* (1998) — (**p. 119**).

[182] Ludwig Fahrmeir et al. *Regression: Models, Methods and Applications*. Springer-Verlag, 2013. ISBN: 9783642343322 — (**p. 120**).

[183] ATLAS Collaboration. "Search for a multi-Higgs-boson cascade in $W^+W^-b\bar{b}$ events with the ATLAS detector in pp collisions at $\sqrt{s} = 8$ TeV." *Phys. Rev. D* **89** 3 (2014), 032002 — (**pp. 121, 123**).

[184] Caleb Parnell Lampen. "Search for a Three Lepton and Missing Transverse Energy Signature of Supersymmetry Using a Multivariate Analysis." CERN-THESIS-2013-139. Ph.D. Dissertation. University of Arizona, 2013 — (**pp. 121, 123**).

[185] Tamara Vazquez Schroeder. "Measurement of the associated production of a vector boson (W, Z) and top quark pair in the opposite sign dilepton channel with pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." CERN-THESIS-2014-225. Ph.D. Dissertation. Georg-August University School of Science, 2014 — (**pp. 121, 123, 128**).

[186] ATLAS Collaboration. "Measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ channels in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector." *Phys. Rev. D* **90** 5 (2014), 052004 — (**pp. 121, 123**).

[187] Esch Hendrik. "Measurement of the top quark mass in topologies enhanced with single top-quarks produced in the t-channel using flavour-tagging and a neural network with ATLAS data at $\sqrt{s} = 8$ TeV." CERN-THESIS-2015-001. Ph.D. Dissertation. Technical University of Dortmund, 2015 — (**pp. 121, 123**).

[188] Alberto Zucchetta. "Searches for signatures of an extended Higgs sector in final states with leptons and Higgs to bb decays at CMS." CERN-THESIS-2015-025. Ph.D. Dissertation. University of Padua, 2015 — (**pp. 121, 123**).

[189] ATLAS Collaboration. "Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated $(W/Z)H$ production with the ATLAS detector." *J. High Energy Phys.* 1 (2015) — (**pp. 121, 123**).

[190] ATLAS Collaboration. "Comprehensive measurements of t-channel single top-quark production cross sections at $\sqrt{s} = 7$ TeV with the ATLAS detector." *Phys. Rev. D* **90** 11 (2014), 112006 — (***pp. 121, 123***).

[191] ATLAS Collaboration. "Search for s-channel single top-quark production in proton–proton collisions at with the ATLAS detector." *Phys. Lett. B* **740** (2015), 118–136 — (***pp. 121, 123***).

[192] ATLAS Collaboration. "Measurement of the cross section of high transverse momentum production in proton–proton collisions at with the ATLAS detector." *Phys. Lett. B* **738** (2014), 25–43 — (***pp. 121, 123***).

[193] ATLAS Collaboration. "Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector." *J. High Energy Phys.* **2015** 4 (2015) — (***pp. 121, 123***).

[194] Nils Ruthmann. "Search for Standard Model $H \to \tau^+ + \tau^-$ Decays in the Lepton-Hadron Final State in Proton-Proton Collisions with the ATLAS Detector at the LHC." CERN-THESIS-2014-238. Ph.D. Dissertation. University of Freiburg, 2014 — (***pp. 121, 123***).

[195] Keisuke Yoshihara. "Measurement of the Higgs boson couplings using the $WW^* \to l\nu l\nu$ final state." CERN-THESIS-2014-151. Ph.D. Dissertation. University of Tokyo, 2014 — (***pp. 121, 123***).

[196] Antonio Salvucci. "The Higgs Boson in the $H \to ZZ^{(*)} \to 4l$ decay channel with the ATLAS detector at the LHC." CERN-THESIS-2014-132. Ph.D. Dissertation. Radboud University, 2014 — (***pp. 121, 123***).

[197] CMS Collaboration. "Search for a W' boson decaying to a bottom quark and a top quark in pp collisions at $\sqrt{s} = 7$ TeV." *Phys. Lett. B* **718** 4–5 (2013), 1229–1251 — (***pp. 121, 123, 128***).

[198] Alexis Kalogeropoulos. "Search for direct stop pair production at the LHC with the CMS experiment." CERN-THESIS-2013-347. Ph.D. Dissertation. Vrije Universiteit Brussel, 2013 — (***pp. 121, 123***).

[199] Jakob Salfeld-Nebgen. "Search For The Higgs Boson Decaying Into τ-Leptons In The Di-Electron Channel." CERN-THESIS-2014-111. Ph.D. Dissertation. University of Hamberg, 2014 — (***pp. 121 sqq.***).

[200] CMS Collaboration. "Search for standard model production of four top quarks in the lepton + jets channel in pp collisions at $\sqrt{s} = 8$ TeV." *J. High Energy Phys.* 11 (2014) — (***pp. 121 sq.***).

[201] Michel Trottier-McDonald. "Evidence for the Standard Model Higgs Boson Decaying to One Semi-hadronically Decaying Tau and One Leptonically Decaying Tau at ATLAS." CERN-THESIS-2014-055. Ph.D. Dissertation. Simon Fraser University, 2014 — (***pp. 121, 123 sq., 461***).

[202] ATLAS Collaboration. "Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector." *Phys. Rev. D* **90** 11 (2014), 112015 — (***pp. 121, 123***).

[203] Jonathan Mark Stahlman. "First Measurements of the Differential Cross Sections of Higgs Boson Production and Decay in the Four Lepton Final State." CERN-THESIS-2014-175. Ph.D. Dissertation. University of Pennsylvania, 2014 — (***pp. 121, 123***).

[204] A. Buckley, A. Shilton, and M.J. White. "Fast supersymmetry phenomenology at the Large Hadron Collider using machine learning techniques." *Computer Physics Communications* **183** 4 (2012), 960–970 — (***pp. 121 sq.***).

[205] M. Wolter. "Multivariate analysis methods in physics." *Phys. Particles and Nuclei* **38** 2 (2007), 255–268 — (***p. 121***).

[206] Jiri Franc et al. *New Statistical Techniques in the Measurement of the inclusive Top Pair Production Cross Section*. 2014. arXiv: 1412.1076 [hep-ex] — (***p. 121***).

[207] Simon Viel. "Search for new neutral high-mass resonances decaying into muon pairs with the ATLAS detector." CERN-THESIS-2014-093. Ph.D. Dissertation. University of British Columbia, 2014 — (***p. 121***).

[208] Kun Liu. "Observation of the Higgs particle in $\gamma\gamma$ events and search for the Higgs particle in $Z\gamma$ events at ATLAS." CERN-THESIS-2014-146. Ph.D. Dissertation. Pierre and Marie Curie University, 2014 — (***p. 121***).

[209] Christopher Palmer. "Discovery of the Higgs Boson Decaying to Two Photons." CERN-THESIS-2014-133. Ph.D. Dissertation. University of California, San Diego, 2014 — (***pp. 121 sq.***).

[210] Byron P. Roe et al. "Boosted decision trees as an alternative to artificial neural networks for particle identification." *Nucl. Instr. Meth. Phys. Res. A* **543** 2–3 (2005), 577–584 — (***p. 121***).

[211] Hai-Jun Yang, Byron P. Roe, and Ji Zhu. "Studies of stability and robustness for artificial neural networks and boosted decision trees." *Nucl. Instr. Meth. Phys. Res. A* **574** 2 (2007), 342–349 — (***p. 121***).

[212] M. Hansen, R. Dubayah, and R. DeFries. "Classification trees: an alternative to traditional land cover classifiers." *Int. J. Remote Sensing* **17** 5 (1996), 1075–1081 — (***p. 121***).

[213] ATLAS Collaboration. *Calibration of the performance of b-tagging for c and light-flavour jets in the 2012 ATLAS data*. Tech. rep. ATLAS-CONF-2014-046. 2014 — (***p. 122***).

[214] Caterina Vernieri. "Search for resonances decaying to $H(bb^-)$ pairs with the CMS experiment at the LHC." CERN-THESIS-2014-161. Ph.D. Dissertation. Scuola Normale Superiore, 2014 — (***p. 122***).

[215] CMS Collaboration. "Measurement of the properties of a Higgs boson in the four-lepton final state." *Phys. Rev. D* **89** 9 (2014), 092007 — (***p. 122***).

[216] CMS Collaboration. "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks." *Phys. Rev. D* **89** 1 (2014), 012003 — (***p. 122***).

[217] Matthew Eric Quenneville. *Boosted Regression Tree Based Invariant Mass Reconstruction of τ Lepton Pairs at ATLAS*. Undergraduate Project CERN-STUDENTS-Note-2014-236. 2014 — (***p. 122***).

[218] Natascha Sylvia Hedrich. *Boosted Regression Trees in the $H \to \tau\tau$ decay channel*. Undergraduate Project CERN-STUDENTS-Note-2013-066. 2013 — (***p. 122***).

[219] Rich Caruana and Alexandru Niculescu-Mizil. "An Empirical Comparison of Supervised Learning Algorithms." *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. New York, NY, USA: ACM, 2006, pp. 161–168. ISBN: 1595933832 — (***p. 126***).

[220] Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. "An Empirical Evaluation of Supervised Learning in High Dimensions." *Proceedings of the 25th International Conference on Machine Learning*. ICML '08. New York, NY, USA: ACM, 2008, pp. 96–103. ISBN: 9781605582054 — (***p. 126***).

[221] Saeed Abu-Nimeh et al. "A Comparison of Machine Learning Techniques for Phishing Detection." *Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit*. eCrime '07. New York, NY, USA: ACM, 2007, pp. 60–69. ISBN: 9781595939395 — (***p. 126***).

[222] Nigel Williams, Sebastian Zander, and Grenville Armitage. *Evaluating Machine Learning Algorithms for Automated Network Application Identification*. Tech. rep. 060410B. Melbourne, Australia: Centre for Advanced Internet Architectures, Swinburne University of Technology, Oct. 2006. URL: `caia.swin.edu.au/reports/060410B` — (***p. 126***).

[223] M. Gashler, C. Giraud-Carrier, and T. Martinez. "Decision Tree Ensemble: Small Heterogeneous Is Better Than Large Homogeneous." *Machine Learning and Applications, 2008. ICMLA '08. Seventh International Conference*. 2008, pp. 900–905 — (***p. 126***).

[224] R.S DeFries and Jonathan Cheung-Wai Chan. "Multiple Criteria for Evaluating Machine Learning Algorithms for Land Cover Classification from Satellite Data." *Remote Sensing Environment* **74** 3 (2000), 503–515 — (***p. 126***).

[225] Leo Breiman. "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)." *Stat. Sci.* **16** 3 (2001), 199–231 — (***pp. 126, 423***).

[226] Leo Breiman. "Heuristics of instability and stabilization in model selection." *Ann. Stat.* **24** 6 (1996), 2350–2383 — (***p. 126***).

[227] Andrea Banfi, GavinP. Salam, and Giulia Zanderighi. "Phenomenology of event shapes at hadron colliders." *J. High Energy Phys.* 6, 38 (2010) — (***p. 126***).

[228] ATLAS Collaboration. "Measurement of charged-particle event shape variables in inclusive $\sqrt{s} = 7$ TeV proton-proton interactions with the ATLAS detector." *Phys. Rev. D* **88** (3 2013), 032004 — (***p. 126***).

[229] Donald E. Knuth. "Computer Science and Its Relation to Mathematics." *Am. Math. Mon.* **81** 4 (1974), 323–343 — (***p. 135***).

[230] F. Sforza et al. "Rejection of multi-jet background in a hadron collider environment through a SVM classifier." *Nuclear Science Symposium and Medical Imaging Conference*. 2011, pp. 1404–1408 — (***p. 135***).

[231] James Bergstra and Yoshua Bengio. "Random Search for Hyper-parameter Optimization." *J. Machine Learning Res.* **13** 1 (2012), 281–305 — (***p. 135***).

[232] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics." *Eur. Phys. J. C* **71** 2 (2011), 1554 — (***pp. 138, 238***).

[233] Claire Adam-Bourdarios et al. *Learning to discover: the Higgs boson machine learning challenge*. v.1.8. URL: `www.kaggle.com/c/higgs-boson/data` (visited on 07/21/2014) — (***p. 138***).

[234] Wojciech Kotlowski. "Consistent optimization of AMS by logistic loss minimization." *CoRR* **abs/1412.2106** (2014). URL: `arxiv.org/abs/1412.2106` — (***p. 138***).

[235] N. D. Gagunashvili. "Comparison of weighted and unweighted histograms" (May 2006). eprint: `arXiv:physics/0605123` — (***p. 140***).

[236] Nornadiah Mohd Razali and Yap Bee Way. "Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests." *J. Stat. Modeling Anal.* **2** 1 (2011), 21–33 — (***p. 141***).

[237] M. C. Whitlock. "Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach." *J. Evol. Biol.* **18** 5 (2005), 1368–1373 — (***p. 142***).

[238] Morton B. Brown. "400: A Method for Combining Non-Independent, One-Sided Tests of Significance." *Biometrics* **31** 4 (1975), 987–992 — (***p. 142***).

[239] James T. Kost and Michael P. McDermott. "Combining dependent P-values." *Stat. Prob. Lett.* **60** 2 (2002), 183–190 — (***p. 142***).

[240] Pierre Geurts, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." *Machine Learning* **63** 1 (2006), 3–42 — (***p. 148***).

[241] J. Scott Long and Laurie H. Ervin. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *Am. Stat.* **54** 3 (2000), 217–224 — (***pp. 173, 176 sq., 486 sqq., 492, 495***).

[242] Joel Heinrich and Louis Lyons. "Systematic Errors." *Annu. Rev. of Nucl. and Part. Sci.* **57** 1 (2007), 145–169 — (***p. 216***).

[243] Helge Voss. *What to expect from TMVA, the Toolkit for Multivariate Analysis in ROOT*. ROOT Users Workshop 2013. URL: `tmva.sourceforge.net/talks.shtml` (visited on 11/24/2015) — (***p. 216***).

[244] Emmanuel Busato, David Calvet, and Timothee Theveneaux-Pelzer. "OpTHyLiC: an Optimised Tool for Hybrid Limits Computation" (2015). arXiv: `1502.02610 [hep-ex]` — (***pp. 216, 220, 259 sq., 296 sq., 330, 392***).

[245] ATLAS Collaboration: HQT Working Subgroup. *MCLimit Code*. URL: svn.cern.ch / reps / atlasphys / Physics / Exotic / Analysis / HeavyQuarks / LimitsCode / mclimitCode / tags / mclimitCode - 01 - 02 - 01 (visited on 12/03/2015) — (***pp. 217, 392***).

[246] J. K. Lindsey. *Parametric Statistical Inference*. Oxford University Press, 1996. ISBN: 0198523599 — (***pp. 249, 251***).

[247] E. L. Lehmann. *Testing Statistical Hypotheses*. John Wilely & Sons, 1959 — (***pp. 249, 251***).

[248] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006. ISBN: 9780521685672 — (***pp. 249, 251, 388***).

[249] Min-ge Xie. "Rejoinder." *Int. Stat. Rev.* **81** 1 (2013), 68–77 — (***p. 250***).

[250] J. Neyman and E. S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Phil. Trans. Roy. Soc. London A* **231** 694-706 (1933), 289–337 — (***pp. 250, 281 sq., 285***).

[251] Peter J. Bickel and Kjell A. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day, Inc., 1977. ISBN: 0816207844 — (***pp. 250 sq.***).

[252] D. R. Cox. "Some Problems Connected with Statistical Inference." *Ann. Math. Stat.* **29** 2 (1958), 357–372 — (***pp. 250, 346***).

[253] Luca Lista. *Statistical Methods for Data Analysis in Particle Physics*. Springer International Publishing, 2016. ISBN: 9783319201764 — (***pp. 251, 255, 285***).

[254] Olaf Behnke et al., eds. *Data Analysis in High Energy Physics*. Wiley-VCH Verlag GmbH & Co. KGaA, 2013. ISBN: 9783527653416 — (***p. 251***).

[255] Gerhard Bohm and Gunter Zech. *Introduction to Statistics and Data Analysis for Physicists*. Verlag, 2010. ISBN: 9783935702416 — (***pp. 251, 275***).

[256] David J. Olive. *Statistical Theory and Inference*. Springer, 2014. ISBN: 9783319049717 — (***p. 251***).

[257] Donald A. S. Fraser. "Discussion." *Int. Stat. Rev.* **81** 1 (2013), 42–48 — (***pp. 254, 346***).

[258] Mr. Bayes and Mr. Price. "An Essay towards Solving a Problem in the Doctrine of Chances." *Phil. Trans. Roy. Soc.* **53** (1763), 370–418 — (***p. 254***).

[259] R. D. Cousins. "Comments on methods for setting confidence limits." *CERN Workshop on confidence limits: Proceedings*. Ed. by F. James, L. Lyons, and Y. Perrin. CERN-OPEN-2000-203. 2000, pp. 49–62 — (***pp. 254, 260***).

[260] Deborah G. Mayo and Aris Spanos. "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction." *Br. J. Phil. Sci.* **57** 2 (2006), 323–357 — (***pp. 255, 355, 363, 371***).

[261] James O. Berger et al. "The Likelihood Principle." *Lecture Notes-Monograph Series* **6** (1988), iii–199 — (***p. 255***).

[262] Ronald N. Giere. "Allan Birnbaum's Conception of Statistical Evidence." *Synthese* **36** 1 (1977), 5–13 — (***p. 255***).

[263] Alan Stuart, J. Keith Ord, and Steven Arnold. *Kendall's Advanced Theory of Statistics*. 6th. Vol. 2A. Oxford University Press, Inc., 1999. ISBN: 9780340662301 — (***pp. 255, 257, 392***).

[264] Allan Birnbaum. "On the Foundations of Statistical Inference." *J. Am. Stat. Assoc.* **57** 298 (1962), 269–306 — (***pp. 255, 363***).

[265] Deborah G. Mayo. *Error and the Growth of Experimental Knowledge*. University of Chicago Press, 1996. ISBN: 9780226511979 — (***p. 255***).

[266] Deborah G. Mayo. "On the Birnbaum Argument for the Strong Likelihood Principle." *Stat. Sci.* **29** 2 (2014), 227–239 — (***p. 255***).

[267] Glen Cowan. "Review of Particle Physics: 38. Statistics." *Chin. Phys. C* **38** 9 (2014), 472 — (***pp. 255, 261***).

[268] Erich. L. Lehmann. *Fisher, Neyman, and the Creation of Classical Statistics*. Springer, 2011. ISBN: 9781441994998 — (***pp. 257, 341, 371***).

[269] J. Neyman. "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability." *Phil. Trans. Roy. Soc. London A.* **236** 767 (1937), 333–380 — (***p. 261***).

[270] Luc Demortier. "Interval Estimation." *Data Analysis in High Energy Physics*. Ed. by Olaf Behnke et al. Wiley-VCH Verlag GmbH & Co. KGaA, 2013, pp. 107–151. ISBN: 9783527653416 — (***pp. 261, 402***).

[271] Louis Lyons, Harrison Prosper, and Albert de Roeck, eds. *Computing Likelihood Functions for High-Energy Physics Experiments when Distributions are Defined by Simulators with Nuisance Parameters* — (***pp. 261, 292, 369***).

[272] Gary J. Feldman and Robert D. Cousins. "A Unified approach to the classical statistical analysis of small signals." *Phys. Rev. D* **57** (1998), 3873–3889 — (***pp. 261, 265, 285 sq., 292, 314***).

[273] F. Garwood. "Fiducial Limits for the Poisson Distribution." *Biometrika* **28** 3-4 (1936), 437–442 — (***pp. 263 sq., 269, 272, 319***).

[274] William E. Ricker. "The Concept of Confidence or Fiducial Limits Applied to the Poisson Frequency Distribution." *J. Am. Statistical Assoc.* **32** 198 (1937), 349–356 — (***pp. 264 sq., 269, 272, 319***).

[275] Edwin L. Crow and Robert S. Gardner. "Confidence Intervals for the Expectation of a Poisson Variable." *Biometrika* **46** 3-4 (1959), 441–453 — (***p. 263***).

[276] Christian Robert George Casella. "Refining Poisson Confidence Intervals." *Can. J. Stat.* **17** 1 (1989), 45–57 — (***p. 263***).

[277] Gioacchino Ranucci. *Binomial and ratio-of-Poisson-means frequentist confidence intervals applied to the error evaluation of cut efficiencies*. arXiv:0901.4845. 2009 — (***p. 282***).

[278] E. L. Lehmann. "The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?" *Selected Works of E. L. Lehmann*. Ed. by Javier Rojo. Springer, 2012, pp. 201–208. ISBN: 9781461414117 — (***p. 282***).

[279] V. Buescher et al. *Recommendation of the ad-hoc committee on limit-setting procedures to be used by D0 in Run II*. D0 Note 4629. Oct. 2004 — (***p. 285***).

[280] ATLAS Statistics Forum. *Frequentist Limit Recommendation v.1.6*. June 2011 — (***pp. 285, 295***).

[281] Kristian Gregersen and Jorgen Beck Hansen. *Frequentist Limit Setting in Effective Field Theories*. ATLAS Statistics Forum. Apr. 2016 — (***p. 285***).

[282] Thomas Junk. "Confidence level computation for combining searches with small statistics." *Nucl. Instr. Meth. A* **434** (1999), 435–443 — (***pp. 289, 295***).

[283] CDF Collaboration and D0 Collaboration. "Evidence for a Particle Produced in Association with Weak Bosons and Decaying to a Bottom-Antibottom Quark Pair in Higgs Boson Searches at the Tevatron." *Phys. Rev. Lett.* **109** (7 2012), 071804 — (***pp. 291, 343, 369, 371***).

[284] Louis Lyons. "Open statistical issues in Particle Physics." *Ann. Appl. Statistics* **2** 3 (2008), 887–915 — (***p. 295***).

[285] Glen Cowan. "Statistics for Searches at the LHC." *Proceedings, 69th Scottish Universities Summer School in Physics : LHC Phenomenology (SUSSP69)*. 2013, pp. 321–355. arXiv: 1307.2487 — (***p. 295***).

[286] A. L. Read. "Modified frequentist analysis of search results (the CL_S method)." CERN-OPEN-2000-205 (2000) — (***p. 295***).

[287] Alexander L. Read. "Presentation of search results: the CL_S technique." *J. Phys. G* **28** (2002), 2693–2704 — (***p. 295***).

[288] Louis Lyons, Harrison Prosper, and Albert de Roeck, eds. *The Wish-lists: Some Comments*, pp. 119–122 — (***pp. 295, 314, 421***).

[289] Tom Junk. "Sensitivity, Exclusion and Discovery with Small Signals, Large Backgrounds, and Large Systematics, v.2.4" (2007). CDF-8128 — (***pp. 295, 393***).

[290] P. Astone and G. Pizzella. "Upper limits in the case that zero events are observed: an intuitive solution to the background dependence puzzle." *CERN Workshop on confidence limits: Proceedings*. Ed. by F. James, L. Lyons, and Y. Perrin. CERN-OPEN-2000-203. 2000, pp. 199–206 — (***p. 301***).

[291] Giovanni Punzi. "Stronger Classical Confidence Limits." *CERN Workshop on confidence limits: Proceedings*. Ed. by F. James, L. Lyons, and Y. Perrin. CERN-OPEN-2000-203. 2000, pp. 204–218 — (***p. 309***).

[292] J. Kiefer. "Conditional Confidence Statements and Confidence Estimators." *J. Am. Stat. Assoc.* **72** 360 (1977), 789–808 — (***p. 325***).

[293] Bradley Efron. "R. A. Fisher in the 21st century." *Stat. Sci.* **13** 2 (1998), 95–122 — (***p. 341***).

[294] R. A. Fisher. "On the Mathematical Foundations of Theoretical Statistics." *Phil. Trans. Roy. Soc. London A* **222** 594-604 (1922), 309–368 — (***pp. 341, 369***).

[295] R. A. Fisher. "Theory of Statistical Estimation." *Proc. Cambridge Phil. Trans.* **22** 594-604 (1925), 700–725 — (***p. 341***).

[296] David Hinkley. "Theory of Statistical Estimation: The 1925 Paper." *R.A. Fisher: An Appreciation*. Ed. by Stephen E. Fienberg and David V. Hinkley. Springer, 1980, pp. 85–94. ISBN: 9781461260790 — (***p. 341***).

[297] R. A. Fisher. "The Fiducial Argument in Statistical Inference." *Ann. Eug.* **6** 4 (1935), 391–398 — (***p. 341***).

[298] R. A. Fisher. "Inverse Probability." *Math. Proc. Cambridge Phil. Soc.* **26** 4 (1930), 528–535 — (***p. 341***).

[299] John Aldrich. "Fisher's "Inverse Probability" of 1930." *Int. Stat. Rev.* **68** 2 (2000), 155–172 — (***p. 341***).

[300] Robert Buehler. "Fiducial Inference." *R.A. Fisher: An Appreciation*. Ed. by Stephen E. Fienberg and David V. Hinkley. Springer, 1980, pp. 109–118. ISBN: 9781461260790 — (***p. 341***).

[301] Rink Hoekstra et al. "Robust misinterpretation of confidence intervals." *Psychonomic Bull. & Rev.* **21** 5 (2014), 1157–1164 — (***pp. 341, 369***).

[302] Steven Goodman. "A Dirty Dozen: Twelve P-Value Misconceptions." *Semin. Hematology* **45** 3 (2008), 135–140 — (***p. 341***).

[303] Ronald L. Wasserstein and Nicole A. Lazar. "The ASA's statement on p-values: context, process, and purpose." *forthcoming in Am. Stat.* (2016) — (***pp. 341, 385***).

[304] Monya Baker. "Statisticians issue warning over misuse of P values." *Nature* **531** (2016), 151 — (***p. 341***).

[305] ATLAS Collaboration. *Search for resonances decaying to photon pairs in 3.2 fb^{-1} of pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*. Tech. rep. ATLAS-CONF-2015-081. Dec. 2015 — (***p. 343***).

[306] CMS Collaboration. *Search for new physics in high mass diphoton events in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. Tech. rep. CMS-PAS-EXO-15-004. Dec. 2015 — (***p. 343***).

[307] Dave Charlton. *Highlights from the LHC (I): Selected results from ATLAS and LHCb*. Presented at ICHEP. Aug. 2016 — (***p. 343***).

[308] T. Camporesi. *LHC results highlights: CMS highlights, ALICE highlights*. Presented at ICHEP. Aug. 2016 — (***p. 343***).

[309] Tore Schweder and Nils Lid Hjort. "Confidence and Likelihood." *Scand. J. Stat.* **29** 2 (2002), 309–332 — (***p. 346***).

[310] Min-ge Xie and Kesar Singh. "Confidence Distribution, the Frequentist Distribution Estimator of a Parameter: A Review." *Int. Stat. Rev.* **81** 1 (2013), 3–39 — (***p. 346***).

[311] David R. Cox. "Discussion." *Int. Stat. Rev.* **81** 1 (2013), 40–41 — (***p. 346***).

[312] Allan Birnbaum. "Statistical Methods in Scientific Inference." *Nature* **225** 225 (1970), 1033 — (***p. 351***).

[313] Daniel J. O'Keefe. "Brief Report: Post Hoc Power, Observed Power, A Priori Power, Retrospective Power, Prospective Power, Achieved Power: Sorting Out Appropriate Uses of Statistical Power Analyses." *Commun. Methods Meas.* **1** 4 (2007), 291–299 — (***pp. 355, 364***).

[314] Allan Birnbaum. "The Neyman-Pearson theory as decision theory, and as inference theory; with a criticism of the Lindley-savage argument for Bayesian theory." *Synthese* **36** 1 (1977), 19–49 — (***pp. 363, 369***).

[315] M. J. Bayarri et al. "Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses." *J. Math. Psych.* (2016) — (***p. 363***).

[316] Giulio D'Agostini. "Confidence limits: what is the problem? Is there *the* solution?" hep-ex/0002055 (2000) — (***p. 369***).

[317] Wojtek J. Krzanowski. *Statistical Principles and Techniques in Scientific and Social Investigations*. Oxford, 2007. ISBN: 9780199213108 — (***p. 369***).

[318] G. Cowan et al. "Panel Discussion." *CERN Workshop on confidence limits: Proceedings*. Ed. by F. James, L. Lyons, and Y. Perrin. CERN-OPEN-2000-203. 2000, pp. 271–288 — (***p. 369***).

[319] Louis Lyons. "Bayes and Frequentism: a particle physicist's perspective." *Contemporary Phys.* **54** 1 (2013), 1–16 — (***p. 369***).

[320] Steven D. Biller and Scott M. Oser. "Another look at confidence intervals: Proposal for a more relevant and transparent approach." *Nucl. Instr. Meth. Phys. Res. A* **774** (2015), 103–119 — (***pp. 369, 371***).

[321] E. L. Lehmann. "The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two?" *J. Am. Stat. Assoc.* **88** 424 (1993), 1242–1249 — (***p. 371***).

[322] Ton J. Cleophas et al. *Statistics Applied to Clinical Trials, 4^{th} Ed.* Springer, 2009. ISBN: 9781402095221 — (***p. 371***).

[323] Bradley P. Carlin et al. *Case Study Comparing Bayesian and Frequentist Approaches for Multiple Treatment Comparisons*. Tech. rep. prepared for the U.S. Department of Health and Human Services. Minnesota Evidence-based Practice Center, 2013 — (***p. 371***).

[324] Vinaitheerthan Renganathan. "Overview of Frequentist and Bayesian Approach to Survival Analysis." *App. Med. Info.* **38** 1 (2016), 25–38 — (***p. 371***).

[325] Jan Hannig and Min-ge Xie. "A note on Dempster-Shafer recombination of confidence distributions." *Elec. J. Stat.* **6** (2012), 1943–1966 — (***p. 372***).

[326] William E. Strawderman Kesar Singh Minge Xie. "Combining Information from Independent Sources through Confidence Distributions." *Ann. Stat.* **33** 1 (2005), 159–183 — (***p. 372***).

[327] P. Janot and F. Le Diberder. "Optimally combined confidence limits." *Nucl. Instr. Meth.* **A411** (1998), 449–466 — (***p. 372***).

[328] Bruce Denardo and Andrés Larraza. "A geometrical approach to sums of uniform random variables." *Am. J. Phys.* **57** 8 (1989), 749–755 — (*p. 374*).

[329] Robert D. Cousins and Virgil L. Highland. "Incorporating systematic uncertainties into an upper limit." *Nucl. Instr. Meth. A* **320** 1 (1992), 331–335 — (*p. 388*).

[330] ATLAS Collaboration: Top Working Group. *Top Profiling Checks*. URL: `twiki.cern.ch/twiki/bin/view/AtlasProtected/TopProfilingChecks` (visited on 12/03/2015) — (*p. 392*).

[331] Glen Cowan. "Marginal vs. Profile Likelihood for systematic uncertainties." Draft 1.2. May 2011 — (*p. 392*).

[332] George E. P. Box and Norman R. Draper. "Empirical Model-Building and Response Surfaces." John Wiley & Sons, 1987, p. 74. ISBN: 0471810339 — (*p. 425*).

[333] Kaggle and the ATLAS Collaboration. *Higgs Boson Machine Learning Challenge*. May 12, 2014. URL: `www.kaggle.com/c/higgs-boson` (visited on 07/21/2015) — (*pp. 431, 438*).

[334] A. Mohan, Z. Chen, and K.Q. Weinberger. "Web-Search Ranking with Initialized Gradient Boosted Regression Trees." *JMLR Workshop and Conference Proceedings: Proceedings of the Yahoo! Learning to Rank Challenge* **14** (June 2011), 77–89 — (*p. 433*).

[335] Philip M. Long and Rocco A. Servedio. "Random classification noise defeats all convex potential boosters." *Machine Learning* **78** 3 (2010), 287–304 — (*p. 433*).

[336] K. Monteith et al. "Turning Bayesian model averaging into Bayesian model combination." *The 2011 International Joint Conference on Neural Networks (IJCNN)*. 2011, pp. 2657–2663 — (*p. 438*).

[337] Tim Salimans. *My second place solution to the Higgs Boson Machine Learning Challenge*. URL: `github.com/TimSalimans/HiggsML` (visited on 12/11/2015) — (*p. 438*).

[338] Dr. William H. Wolberg, Olvi Mangasarian, and the UCI Machine Learning Repository. *Breast Cancer Wisconsin (Diagnostic) Data Set*. July 15, 1992. URL: `archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/` (visited on 10/21/2015) — (*p. 438*).

[339] O. L. Mangasarian and W. H. Wolberg. "Cancer diagnosis via linear programming." *SIAM News* **23** 5 (1990), 1–18 — (*p. 438*).

[340] Xiaojin Zhu and Andrew B. Goldberg. *Introduction to Semi-supervised Machine Learning*. Morgan and Claypool, 2009. ISBN: 9781598295474 — (*pp. 438, 457*).

[341] Mehmet Fatih Akay. "Support vector machines combined with feature selection for breast cancer diagnosis." *Expert Systems Appl.* **36** 2, Part 2 (2009), 3240–3247 — (*p. 447*).

[342] J. R. Quinlan. "Improved Use of Continuous Attributes in C4.5." *J. Artificial Intelligence* **4** (1996), 77–90 — (*p. 447*).

[343] A. Gammerman, V. Vovk, and V. Vapnik. "Learning by Transduction." *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. 1998, pp. 148–155 — (***pp. 456 sq.***).

[344] Vinod Pangracious, Zied Marrakchi, and Habib Mehrez. *Three-Dimensional Design Methodologies for Tree-based FPGA Architecture*. Springer, 2015. ISBN: 9783319191737 — (***p. 460***).

[345] Umer Farooq, Zied Marrakchi, and Habib Mehrez. *Tree-based Heterogeneous FPGA Architectures: Application Specific Exploration and Optimization*. Springer, 2012. ISBN: 9781461435938 — (***p. 460***).

[346] Peter J. Rousseeuw. "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *J. Comput. App. Math.* **20** (1987), 53–65 — (***p. 462***).

[347] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, Ltd., 1986. ISBN: 0412246201 — (***p. 464***).

[348] Pedro Domingos. "A Few Useful Things to Know About Machine Learning." *Commun. ACM* **55** 10 (2012), 78–87 — (***p. 477***).

[349] Louis Lyons and Luc Demortier. *Importance of ordering rule for Neyman construction*. CDF/MEMO/STATISTICS/PUBLIC. Nov. 2003 — (***p. 536***).