Studies of the misidentification probability of electrons as photons with the CMS experiment in data recorded in 2015

Masterthesis

von

Ralf Meyer

vorgelegt der Fakultät für Mathematik, Informatik
und Naturwissenschaften der
RWTH Aachen

im November 2016

angefertigt im
1. Physikalischen Institut B
bei Prof. Dr. Lutz Feld

Contents

1 Introduction **6**

2 The Standard Model of particle physics **8**

2.1 Overview . . . 8
2.2 Units and notation . . . 9
2.3 Particle content of the Standard Model . . . 9
2.4 Quantum electrodynamics and the weak interaction . . . 11
2.5 Quantum chromodynamics . . . 12
2.6 Higgs boson . . . 13
2.7 Z-boson resonance . . . 13
2.8 Open questions . . . 14

3 Experimental setup **17**

3.1 Large Hadron Collider . . . 17
3.2 Compact Muon Solenoid . . . 18
3.2.1 Coordinate system . . . 19
3.2.2 Tracker system . . . 21
3.2.3 Electromagnetic calorimeter . . . 22
3.2.4 Hadronic calorimeter . . . 24
3.2.5 Superconducting solenoid . . . 24
3.2.6 Muon chambers . . . 24
3.2.7 Trigger system . . . 25

4 Motivation and analysis strategy **26**

4.1 Overview . . . 26
4.2 Definition of the fake rate . . . 26
4.3 Tag and probe on the Z-boson resonance . . . 27
4.4 Analysis strategy and the template method . . . 31

5 Object definition 33
5.1 Particle-Flow Algorithm 33
5.2 Electrons and photons 34
5.3 Muons 36

6 Event selection 37
6.1 Data samples 37
6.1.1 Data 37
6.1.2 Simulation 37
6.2 Trigger 39
6.3 Tag and probe 41
6.4 Construction of the templates 46
6.4.1 Signal 46
6.4.2 Background 47

7 Fake rate estimation 49
7.1 Estimation of probability density functions from templates 49
7.1.1 Histogram based pdf 49
7.1.2 Gaussian kernel estimation 52
7.1.3 Signal and background shape 54
7.2 Extended maximum likelihood fit 57
7.3 Parametrized Model 60

8 Systematic uncertainties and results 63
8.1 Systematic uncertainties 63
8.1.1 Mismodelling of the templates 63
8.1.2 Dependencies on event observables 64
8.2 Closure on simulation 67
8.3 Summary and outlook 69

Appendices 70

A Parametrized functions 71
A.1 Double sided crystal ball function 71
A.2 CMS Shape 72

B Fit results for the dependecies 73

Bibliography **92**

1 Introduction

One of the most considered and discussed questions since the beginning of human thinking is this: What are the things that we see made of? As a consequence of this question another question arises, asking how the things that we see hold together or, more general, how the contents of the things interact with each other. The field of natural science that queries these fundamental questions is called particle physics, referring to the particle structure of matter. Potentially the most important step on the road from classical ideas to the modern picture has been the development of the concept of atoms and its acceptance in the scientific world. In his famous lectures on physics the Nobel laureate Richard Feynman opens with the question, what statement would contain the most information in the fewest words if all scientific knowledge where destroyed and only one sentence passed on to the next generations. His answer imposes with simplicity. *All things are made of atoms* [1].

Nevertheless, somehow ironically when the existence of atoms was confirmed by several observations in the beginning af the 20th century they already had lost their initially imagined property that they where named after in the first place, the indivisibility (Greek *átomos*). As it turned out there had to be more smaller substructures and in some way this was the beginning of modern particle physics. It's worth to notice that Feynmans statement was given in 1964 when physicists already made a few more steps towards the subcontent of atoms. Presumably he chose that sentence because the road to the acceptance of atoms has been a path of trial and tribulation in relation to the steps that came afterwards.

However, the continual research in experimental and theoretical physics developed two fundamental models to describe the kinds of interactions that are known today. Namely these are Einsteins theory of General Relativity to describe the effects of gravity, and the so called Standard Model of particle physics for the electromagnetic, weak and strong forces. The basic ideas, concepts, foundations and also open questions of the latter one are described in Chapter 2 of this thesis.

To measure interactions on the smallest scales, a lot of energy is needed. In modern particle physics this leads to large experiments, operated by international collaborations. One of the largest research centers is the European Organization for Nuclear Research

(CERN) which is located in Geneva, Switzerland. It operates the Large Hadron Collider which is one of the worlds most complex experimental facilities and the most powerful particle collider ever built. In the year 2012 the Large Hadron Collider experiments ATLAS and CMS observed a new particle that turned out to be the last predicted component of the Standard Model, the famous Higgs boson. The presented thesis is based on data recorded by the CMS experiment which is described in Chapter 3.

Beside the measurement of known quantities with improved precision, one of the main tasks of the LHC after the Higgs observation is the search for physics beyond the Standard Model. For this it is nescessary to understand the known physical processes very well to give precise predictions of the expected Standard Model contributions.

One of the contributions to event signatures is the misidentification of physical objects in the detector. In this thesis the probability of an electron to be reconstructed as a photon is studied. The motivation for the presented analysis is given in Chapter 4. The technical details of the analysis are then described in the Chapters 5, 6 and 7. A summary of the results and an outlook is given in Chapter 8.

2 The Standard Model of particle physics

2.1 Overview

Particle physics, or high-energy physics, pursues the ultimate structure of matter. It is concerned with the fundamental constituents of the Universe, the *elementary particles*, and the interactions between them, the *forces*.

The modern theoretical concept to describe both developed from the quantum mechanical view of matter as states in a complex vector space and is based on fields. Each elementary particles state is described as a discrete excitation of the underlying field. The interactions between the particles are mediated by force-carrying particles known as gauge bosons, which are themselves again excitations of the corresponding gauge fields.

In this picture everything is described by Quantum Field Theories and the equations of motion arise from the principal of least action where the dynamic of a system is expressed by its lagrangian density. The field excitations, interpreted as particles, are segmented in two classes based on their spin which is expressed in terms of the reduced Planck constant $\hbar$. Particles with half-integer spin are called *fermions* and build the matter that we know, while particles with integer spin are called *bosons* which are the mediators of the forces.

The entirety of theories to describe the dynamics of the known particles and their interactions are summarized in the framework of the Standard Model of particle physics which is simply referred to as the Standard Model or SM. In this section its main aspects will be described following Ref. [2–4].

2.2 Units and notation

The commonly used unit system in particle physics is called natural units and is based on the fundamental constants of quantum mechanics and special relativity. In natural units, the familiar dimensions of meter for the length, kilogram for the mass and second for the time are replaced by the reduced Planck constant $\hbar$, the speed of light in vacuum c and the dimension of energy, commonly expressed in terms of the approximately rest mass energy of the proton 1 GeV.

This replacement can be simplified by choosing

$$\hbar = c = 1 \ . \tag{2.2.1}$$

Natural units provide a well motivated basis for expressing quantities in powers of energy and simplify commonly occuring equations and are mostly used throughout this work. Cross sections and integrated luminosities are expressed in powers of barn where $1\ \mathrm{b} \equiv 10^{-24}\ \mathrm{cm}^{-2}$.

2.3 Particle content of the Standard Model

The Standard Model contains twelve fundamental fermions listed in Table 2.1 where each of the fermions has a corresponding anti-fermion with an inverse electrical charge. These are commonly organized in three *generations*, where the first generation includes the electron, its neutrino, the up-quark and the down-quark. For each of these particles there are two copies which only differ in their masses and are known as the second and third generation. The properties of these copies are the same in the sense that they possess the same fundamental interactions. The generations are also called *families*.

Table 2.1: Standard Model fermions divided into their three generations and splitted into leptons and quarks. The masses of the particles are rounded experimental values or, in case of the neutrinos, upper limits.

	lepton		charge [e]	mass [GeV]	quark		charge [e]	mass [GeV]
first	electron	(e^-)	-1	0.0005	down	(d)	$-1/3$	0.003
	neutrino	(ν_e)	0	$< 10^{-9}$	up	(u)	$+2/3$	0.005
second	muon	(μ^-)	-1	0.106	strange	(s)	$-1/3$	0.1
	neutrino	(ν_μ)	0	$< 10^{-9}$	charm	(c)	$+2/3$	1.3
third	tau	(τ^-)	-1	1.78	bottom	(b)	$-1/3$	4.5
	neutrino	(ν_τ)	0	$< 10^{-9}$	top	(t)	$+2/3$	174

Table 2.2: The interactions in which the particles participate.

					strong	electromagnetic	weak
quarks	down-type	d	s	b	✓	✓	✓
	up-type	u	c	t	✓	✓	✓
leptons	charged	e^-	μ^-	τ^-		✓	✓
	neutrinos	ν_e	ν_μ	ν_τ			✓

Table 2.3: Known forces and their gauge bosons. The relative strengths are approximate indicative values for two particles at a distance of 10^{-15} m which is roughly the radius of a proton. The forces depend on the considered distance and energy scale.

force	strength	gauge boson		spin	mass [GeV]
strong	1	gluon	g	1	0
electromagnetic	10^{-3}	photon	γ	1	0
weak	10^{-8}	W	$W^\pm$	1	80.4
	10^{-8}	Z	Z	1	91.2
gravitational	10^{-37}	graviton(?)	G	2	0

Besides the classification of particles by their spin there exists further quantum numbers mainly determining the kind of interaction where the particle participates. Table 2.2 gives an overview of this. Each force is exchanged by force carrying gauge bosons which are listed in Table 2.3. For each type of interaction there is an associated coupling strength determined by a quantum number that is called the *charge* of the interaction, where a particle only couples to the corresponding interaction if it carries the charge. The coupling strength can be interpreted as a quantity for the probability that a fermion emitts or absorbs the boson corresponding to the interaction type. In terms of electromagnetism the coupling strength is the electrical elementary charge e commonly expressed by the fine-structure constant $\alpha_{\text{em}} \propto e^2$.

It should be noted that gravity is *not* included in the Standard Model and a possible quantum gravitational theory is one of the main tasks of current theoretical research. Therefore the proposed gauge boson of gravity, commonly called graviton, is a hypothetical particle.

The further details of the underlying theoretical mechanisms of the strong, electromagnetic and weak interaction are given in the next two sections.

2.4 Quantum electrodynamics and the weak interaction

The mathematical framework for treating the quantum dynamics of electrical charged Dirac particles (which means their dynamics follow the Dirac equation of relativistic quantum mechanics) with the quantized electromagnetic field is Quantum Electrodynamics (QED). It is an abelian gauge theory and is based on the $U(1)$ symmetry. The interaction is exchanged by *virtual* photons, where virtual means that it is a mathematical construct which represents the sum of all time-ordered possibilities of interaction and, if appropriate, the sum over the possible polarisation states of the exchange photon.

Therefore virtual particles in general does not satisfy Einsteins energy-momentum relationship $p_\mu p^\mu = m^2$ and are referred to as *off-shell* (off the mass shell).

All twelve SM fermions carry a quantum number called weak isospin I which is the charge of the weak force. The corresponding gauge group is $SU(2)$ and in its terminology the weak force carriers constitute an isospin triplet. The SM leptons can be classified by their isospin component where the left-handed fermions constitute doublets and the right-handed particles belong to $SU(2)$ singlets. Due to the left-handed-only chiral particle state composition of the doublets the symmetry group is also denoted as $SU(2)_L$. The requirement of gauge invariance under $SU(2)$ transformations can be satisfied by introducing three new gauge fields W_1, W_2, W_3.

Since the experimental observation of Z boson couplings to left- and right-handed chiral states is in contradiction to the prediction that the weak neutral-current could only couple to left-handed particles and right-handed anti-particles, Glashow, Salam and Weinberg proposed a new symmetry [5–7]. In this model the $U(1)$ symmetry of QED is replaced with a new $U(1)_\gamma$ local gauge symmetry that gives rise to a new gauge field B that couples to a new kind of charge referred to as *weak hypercharge* Y. In the unified electroweak model the photon and the Z boson can be written as linear combinations of the B field and the neutral W_3 of the weak interaction:

$$\begin{pmatrix} \gamma \\ Z \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \begin{pmatrix} B \\ W_3 \end{pmatrix} \tag{2.4.2}$$

Here θ_W denotes the electroweak mixing angle, also called Weinberg angle. The underlying gauge symmetry then is the $U(1)_\gamma$ of weak hypercharge and $SU(2)_L$ of the weak interaction, written as $U(1)_\gamma \times SU(2)_L$.

The physical W bosons can be identified as linear combinations

$$W^\pm = \frac{1}{\sqrt{2}} \left(W_1 \mp iW_2\right) \ . \tag{2.4.3}$$

The weak interaction of the three generations of quarks is described in terms of the Cabbibo-Kobayashi-Maskawa (CKM) matrix that relates the mass eigenstates of the quarks to the weak eigenstates. A similar mechanism explains the composition of the neutrino mass states from weak eigenstates via the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix which leads to a phenomenon called *neutrino oscillation* where the observed neutrinos oscillate between the flavours.

Mass and charge of the mass eigenstates are shown in Table 2.1.

2.5 Quantum chromodynamics

Quarks are the only fermions that carry another quantum number called *colour charge* which comes in three states commonly denoted as *red*, *green* and *blue*. This is the coupling charge for the strong interaction and its quantum field theoretical treatment is Quantum Chromodynamics (QCD). The underlying symmetry is invariance under $SU(3)$ local phase transformations and the generators of the symmetry group can be identified as the eight Gell-Mann matrices. Each of the non-diagonal matrices corresponds to a boson state called *gluon* that exchanges the strong interaction and connect quark states of different color. Gluons carry one colour and one anti colour, quarks only carry either one colour or anti colour.

The experimental evidence that there are no free quarks is explained with the *color confinement* which basically states that coloured objects are always confined to colour singlet states and that objects with zero colour charge are the only ones that can propagate as free particles. It is believed that this behaviour originates from the gluon-gluon self interactions because the gluons carry colour themselves. The qualitative understanding of what happens if two quarks are tried to be seperated is that the virtual gluons, arising from the quark, interact via new virtual gluons and the colour charge based attraction squeezes the colour field between the quarks into a tube. The energy stored in the field is proportional to the seperation of the quarks and is measured to be ~ 1 GeV/fm.

As a consequence of the confinement the production of high energy quark anti-quark pairs which are back to back leads to a process that is known as hadronisation. The energy stored in the colour field tube increases with the quark separation and becomes sufficient to the energy needed to form new quark anti-quark pairs which then breaks the colour tube. This process continues until all quarks have sufficiently low energy to form colourless hadrons. The hadronisation is then observed as hadronic jets.

Due to the symmetry groups of the strong and electroweak interactions the Standard Model can be described as a renormalizable quantum field theory that is invariant under the $SU(3) \times SU(2)_L \times U(1)_\gamma$ symmetry group.

2.6 Higgs boson

The mechanism that is believed to provide the mass to the heavy gauge bosons of the electroweak interactions and to the fermions is known as the Higgs mechanism. Here a complex scalar field with a non-zero vacuum expectation value, commonly referred to as a *mexican hat potential*, is added to the SM lagrangian. In the Salam-Weinberg model, the Higgs mechanism is embedded in the $SU(2)_L \times U(1)_\gamma$ local gauge symmetry and the ground state of the Higgs potential is not symmetric with respect to this group. To provide the longitudinal degrees of freedom of the heavy electroweak gauge bosons $W^\pm$ and Z, three new particles called Goldstone bosons are required. After the symmetry breaking the Goldstone bosons mix to the electroweak gauge bosons and one massive scalar particle, where the latter one corresponds to the field excitation in the direction picked out by choice in physical vacuum. This scalar particle is known as the Higgs boson which has been observed by ATLAS and CMS in 2012 [8, 9].

Therefore, the complete particle content of the Standard Model has been observed.

2.7 Z-boson resonance

The QED production of a pair of leptons in hadron-hadron collisions such as proton-proton at the LHC from the annihilation of a quark and an anti-quark is commonly referred to as the Drell-Yan process. One possible intermediate particle is the Z boson in the process

$$q\bar{q} \to Z \to l^+ l^- \ . \tag{2.7.4}$$

For a particle described by a quantum mechanical wave function $\Psi(E)$ as a function of the energy E it can be easily shown that the probability, that the particle takes an energy value E follows a Breit-Wigner resonance formula

$$P(E) \propto |\Psi(E)|^2 = \frac{|\Psi(0)|^2}{2\pi} \frac{\hbar^2}{(E - E_0)^2 + (\Gamma/2)^2} \tag{2.7.5}$$

where the *decay width* Γ of the particle is a quantity for its lifetime τ by $\Gamma = \hbar/\tau$ [2]. The peak center energy E_0 is referred to as the particles rest mass. Thus by detecting the lepton pair originating from the Z decay and calculating the invariant mass of both, the mass distribution for the Z boson can be reconstructed. Figure 2.1 shows a sketch of the Breit-Wigner resonance shape and the leading order feynman graph of the production of a Z boson by quark-antiquark-annihilation and its subsequent dileptonic decay.

Since QED processes are symmetric with respect to the lepton flavour the probability of the Z boson to decay in e^+e^+, $\mu^+\mu^-$ or $\tau^+\tau^-$ is equal.

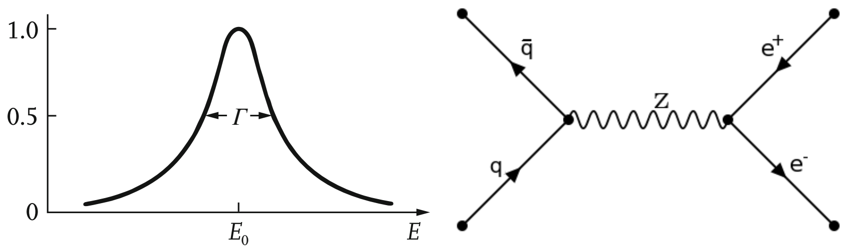


Figure 2.1: Left: Breit-Wigner resonance shape for the mass distribution of unstable particles, taken from [2]. Right: the leading order s-channel feynman graph for the process from Eq. 2.7.4.

2.8 Open questions

The Standard Model is able to describe most of the known physics in the universe with a high precision. It is the most fundamental theoretical model which is able to make clear predictions that has been developed so far. On scales above the subatomic structure the usage of many different and more classic approximative theories is more convenient or even nescessary due to the impracticality of doing calculations based on the most fundamental level. However, the Standard Model provides the fundamentals. Nevertheless, there are still many open questions and problems which the Standard Model in its current status is not able to solve. The outstanding topics and its possible solutions that are most relevant in current research will be shortly listed here.

Dark Matter A number of astronomical observations suggest that a significant fraction of the mass in the universe is not bound up to the luminous stars. The strongest hint for this comes from the measured velocity distributions of stars in rotating spiral galaxies, where classical mechanics would predict a decrease of the velocities with the radius by $r^{-1/2}$, assuming that most of the mass in the galaxy is concentrated in its center. Instead the observed distributions only decrease slowly with the radius. From this it can be concluded that there has to be a significant non-luminous mass component known as dark matter which is not described by the Standard Model.

Supersymmetry A popular extension of the SM is Supersymmetry (SUSY). Here a new symmetry is introduced where each SM particle has a super-partner which differs by half a unit in its spin. Different SUSY models predict different behaviours and couplings of the super-partners to the SM particles. In many of these models the lightest SUSY particle is assumed to be stable and weakly interacting and therefore would be a candidate for the dark matter. The main motivation for SUSY is the so called *hierarchy problem*. It basically means that due to the interaction with the fermions and bosons the self energy of the Higgs boson is influenced by quantum loop corrections. On large energy scales of about $\Lambda \gtrsim 10^{16} - 10^{19}$ GeV these corrections become very large and the Higgs mass would differ from the elektroweak scale. In unbroken SUSY this problem is naturally solved since each SM particle loop would be balanced by a loop coupled to the corresponding super-partner. The search for SUSY is one of the main tasks of the LHC experiments. After the Run 1 no significant hint for SUSY has been found [10].

Unification of forces In the Standard Model, the coupling strenghts of the three forces depend on the considered energy scale. On the electroweak scale they are roughly related as $\alpha_{\mathrm{em}}^{-1} : \alpha_{\mathrm{w}}^{-1} : \alpha_{\mathrm{s}}^{-1} \simeq 128 : 30 : 9$. The running of the coupling strengths indicates that they might converge on large energy scales. This is the motivation for Grand Unified Theories (GUT) where all the known forces are just low energy manifestations of some larger, unified theory.

Nature of the Higgs boson Within the Standard Model, the Higgs boson is unique since it is the only scalar particle. The choice of a single Higgs doublet in the SM is the most economical choice, but it is not the only possibility. For example supersymmetric extensions require at least two doublets which have different vacuum expectation values and give the masses to the fermions and the upper and lower components of the weak isospin doublets. If this is the case, the couplings of the light Higgs boson to the fermions will slighty differ from the SM predictions. Precise studies of the branching ratios of the 125 GeV Higgs at LHC and future colliders therefore may reveal physics beyond the SM.

Flavour and CP Violation The SM is not able to explain *why* there are exactly three flavours of fermions. Furthermore, the mixing of the quark eigenstates via the CKM matrix and the mixing of the neutrino eigenstates in the PMNS matrix are the only places where a violation of the conservation of charge and parity (CP) can be accommodated. The occurence of CP violations might explain the observed asymmetry of matter and antimatter and although the CP in the neutrino sector has to be studied yet in further detail, the CP violation in the SM might be insufficient to explain the asymmetry. A solution might be physics beyond the SM.

Nature of neutrinos An explanation for the large mass difference of the neutrinos to the other fermions is the so called *seesaw mechanism*. Here it is assumed that neutrinos are Majorana particles, which means that neutrinos would be their own anti-particle. Since the observable consequences of this would be very small the nature of neutrinos is not clear yet. One possibility to resolve this is the search for the neutrinoless double β-decay, which could only occur if neutrinos are Majorana particles.

3 Experimental setup

In this thesis data recorded with the Compact Muon Solenoid (CMS) detector located at the Large Hadron Collider in Geneva is analyzed. This chapter briefly describes the main properties of the collider and the structure and sub-systems of the CMS experiment.

3.1 Large Hadron Collider

The Large Hadron Collider (LHC) [11] is the worlds largest and most powerful particle collider ever built. It is located in a ring tunnel with a circumference of 26.7 km approximately 50 to 175 m below the ground.

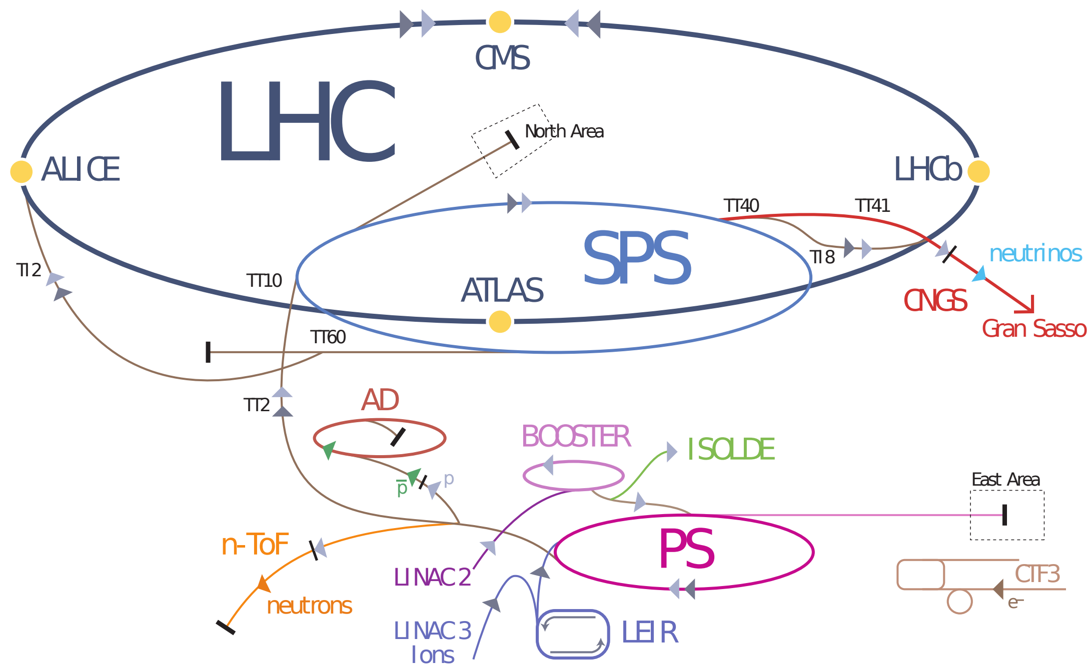


Figure 3.1: Sketch of the LHC ring with the four big experiments and the pre-accelerator components [12].

The collider contains two beam pipes to accelerate two beams in opposite directions and is designed to operate at a maximum energy per proton of 7 TeV. In the first runs in the years 2010 and 2011 the LHC operated at an energy per proton of 3.5 TeV, in 2012 this was increased to 4 TeV. After this, the LHC took a long shutdown to upgrade several components and has been restarted in 2015 with an energy per proton of 6.5 TeV corresponding to a center of mass energy per collision of $\sqrt{s} = 13$ TeV.

Another operation mode is the collision of heavy ions such as lead nuclei and the collision of lead nuclei with protons.

The beam pipes cross at four collision points and at each point an experiment is located. These are CMS [13], ATLAS [14], ALICE [15] and LHCb [16]. CMS and ATLAS are multi-purpose experiments with a symmetric architecture with respect to the collision point. Their main goal beside high precision measurements of SM properties is Higgs physics and the search for physics beyond the Standard Model. ALICE and LHCb are more specialized experiments and will not be discussed here in further detail.

In this work only proton-proton data will be considered.

The protons are pre-accelerated by several components before they get injected in the LHC, a schematic sketch of the whole facility is shown in Fig. 3.1. They get accelerated in bunches and a quantity to measure the collider perfomance is called the *luminosity* $\mathcal{L}$ [17]. It is defined as the number of collisions per area per time and therefore a quantity for the total amount of data collected in a specific timerange is the *integrated luminosity*

$$L_{\text{int}} = \int \mathcal{L}\, dt \; . \tag{3.1.1}$$

In the 2015 data at LHC an integrated luminosity of 4.22 fb^{-1} has been achieved in proton proton collisions at the CMS experiment [18]. From the integrated luminosity and the cross section σ of a specific process one can directly calculate the expected number of events N in which the process occured by

$$N = \sigma \cdot L_{\text{int}} \; . \tag{3.1.2}$$

3.2 Compact Muon Solenoid

The CMS experiment [13] is a symmetric multi-purpose detector and its general structure is sketched in Fig. 3.2. It is centered around the reference collision point of the proton bunches and contains several detector systems such as a silicon tracker system, an electromagnetic and a hadronic calorimeter and the outer muon chambers. The geometrical dimensions of the complete CMS detector are a length of 21.6 m and a diameter of 14.6 m with a total weight of about 12600 tons.

Keystone of the detector is the superconducting solenoid which provides a strong magnetic field of 3.8 T. Due to the diferent subdetector systems the CMS detector is able to reconstruct different particle types from their signature as sketched in Fig. 3.3.

In this section the coordinate system of CMS and the different subsystems are briefly described with a more detailed look on the tracker system which is the key system for this thesis.

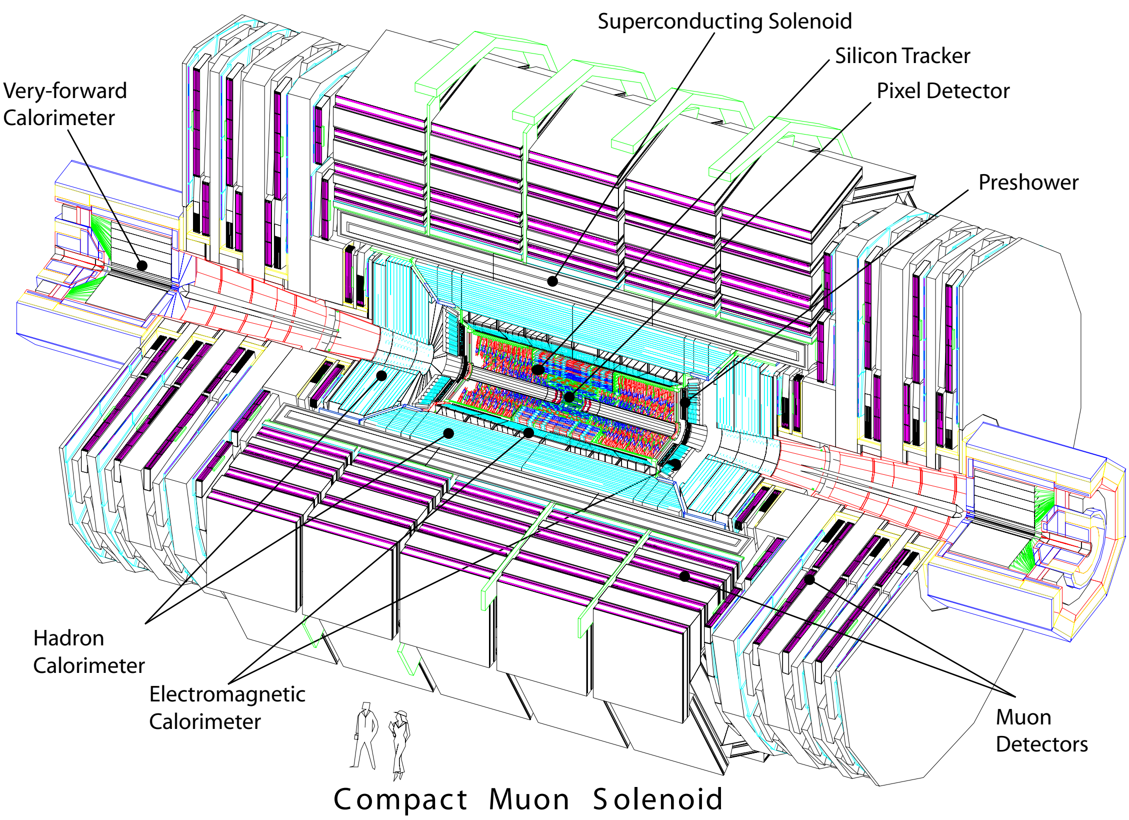


Figure 3.2: Sketch of the components of the CMS detector [13].

3.2.1 Coordinate system

The origin of the coordinate system of the detector is the reference collision point. Its x-axis points to the center of the collider, the y-axis is pointing upwards and the z-axis is parallel to the beam. Since the detector is built as a shell structure around its center, a polar coordinate system is used. The azimuthal angle ϕ is defined in the x-y-plane with $-\pi < \phi \leq \pi$ and $\phi = 0$ is defined to be the direction of the x-axis. The polar angle θ with

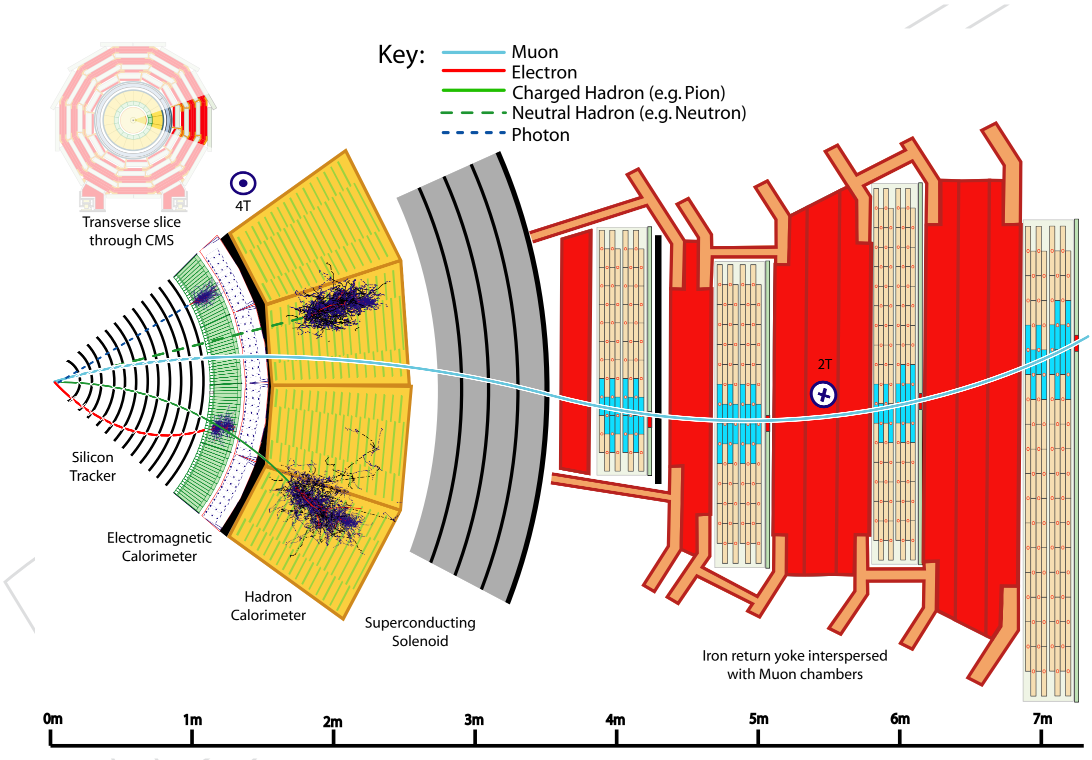


Figure 3.3: Idealized view of the specific particle interactions in a transverse slice of the CMS detector [19].

$0 \leq \theta \leq \pi$ starts in the positive z-axis direction and is commonly expressed in terms of the *pseudorapidity* η which is defined as

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] . \tag{3.2.3}$$

The pseudorapidity is a high energy approximation for the *rapidity* which is a quantity difficult to measure but only changes by a constant under Lorentz transformations in the z-direction. Therefore differences in η are Lorentz invariant with respect to the z-axis. Distances in the η-ϕ-plane are expressed by

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} . \tag{3.2.4}$$

In the collision of hadrons the longitudinal momenta of the interacting quarks are unknown and thus the boost with respect to the z-axis differs for each collision. However, the transversal momenta are expected to be small and therefore a useful quantity is the projection of the measured final state momenta on the transversal plane

$$p_T = p \cdot \sin\theta . \tag{3.2.5}$$

3.2.2 Tracker system

The silicon tracker system is the closest detector part with respect to the collision point and contains two different kinds of detector technology. A schematic sketch of the whole tracker system is shown in Fig. 3.4. Directly next to the collision point comes the inner tracker with silicon pixel technology while the outer shells contain silicon strips. In Fig. 3.5 the layers of the inner tracker are schematically shown and the efficiency to detect a charged particle in two or more layers with respect to the particles pseudorapidity is plotted. The whole tracker system has a geometrical length of 5.8 m and a diameter of 2.5 m.

The inner pixel tracker has a pixel cell size of 100×150 μm^2 and a similar track resolution in $r - \phi$ and z direction. It covers a range of $-2.5 < \eta < 2.5$ which is also the acceptance of the outer strip detector and consists of three barrel layers up to a pseudorapidity of $\eta = 1.4442$ and two endcap disks. The pixel layers are located at radii between 4.4 cm and 10.2 cm with a length of 53 cm in the barrel region while the endcap disks are located at $z = \pm 34.5$ cm and $z = \pm 46.5$ cm with radii between 6 cm and 15 cm.

The outer barrel consists of six cylindrical silicon strip layer modules and surrounds the inner tracker while the outer endcaps consists of nine discs.

The tracker system covers a momentum resolution from 0.7% at 1 GeV to 5% at 1 TeV, the resolution of the primary vertex reconstruction depends on the number and transversal momentum of tracks in a range from 50 μm for vertices with many tracks to 500 μm for few tracks with low transversal momentum [20].

The tracker is operated at temperatures below -10 °C.

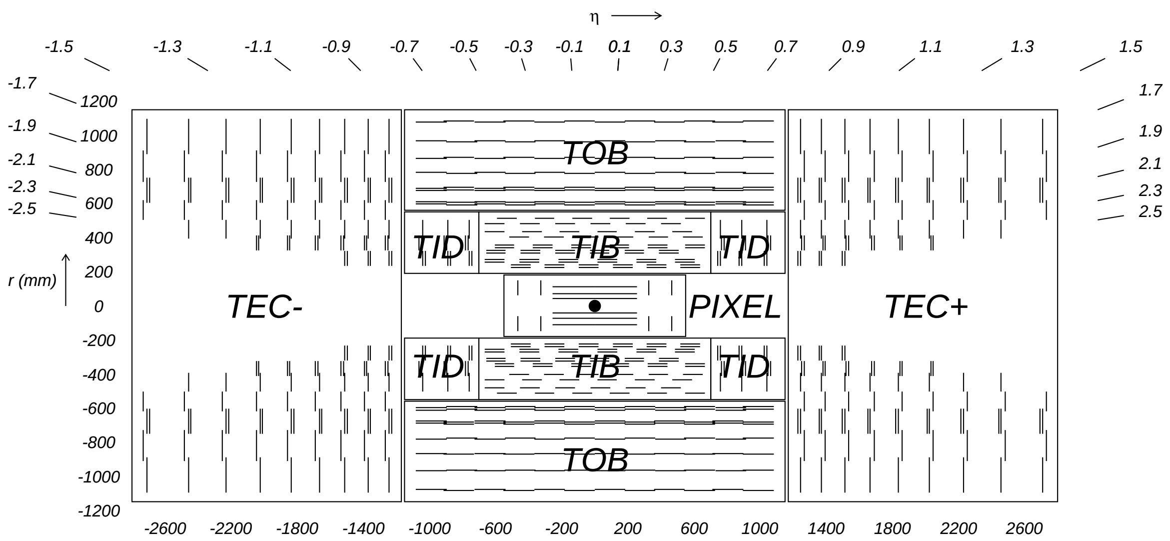


Figure 3.4: Sketch of the CMS tracker system [13]. Each line represents a detector module. Meaning of the shortcuts: Tracker Inner Barrel (TIB), Tracker Outer Barrel (TOB), Tracker Inner Discs (TID), Tracker Endcaps (TEC).

3.2.3 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) consists of lead tungstate crystals. Like the tracker it is divided in barrel (61200 crystals) and endcap (7324 crystals per endcap) parts.

The barrel crystals have a front cross section of 2.2×2.2 cm^2 and are 23 cm long. In the endcaps the crystals are ordered in a rectangular grid and have a front cross section of 2.86×2.86 cm^2 and a length of 22 cm.

Avalanche photodiodes in the barrel, which are able to operate in the strong magnetic field, and vacuum phototriodes in the end cap, which are less sensitive to the higher radiation, detect the light emitted by the scintillators.

The energy resolution of the ECAL can be parametrized as

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E\,[\mathrm{GeV}]}}\right)^2 + \left(\frac{N}{E\,[\mathrm{GeV}]}\right)^2 + C^2 \qquad (3.2.6)$$

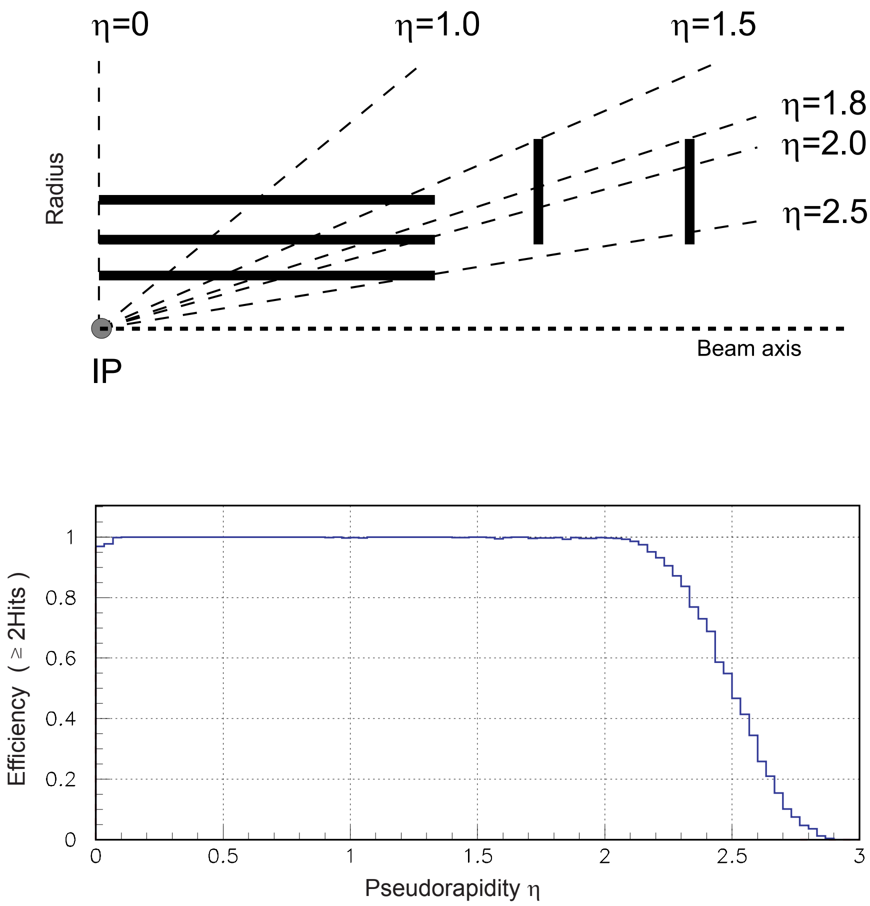


Figure 3.5: Geometrical layout of the inner pixel tracker and its hit coverage as a function of the pseudorapidity [13].

with $S = 2.8\%$ being the stochastic term, $N = 0.12$ the noise term and $C = 0.3\%$ the constant term.

The CMS ECAL reaches a mass resolution for $Z \to ee$ of 1.6% with barrel electrons and 2.4% if both are measured in the endcaps [21].

Preshower detectors consisting of lead absorber and silicon detector layers are installed in front of the endcaps in a pseudorapidity range of $1.6 < |\eta| < 2.6$ to help identifying neutral pions and improve the positional measurement of electrons and photons.

3.2.4 Hadronic calorimeter

The hadronic calorimeter (HCAL) is composed of several subdetectors made of brass, stainless steel absorbers and plastic scintillators. The barrel region covers an absolute pseudorapidity up to 1.3 within the superconducting solenoid and is completed by the outer HCAL which provides sufficient stopping power for the containment of hadronic showers. The endcap covers a pseudorapidity range of $1.3 < |\eta| < 3.0$.

For high pseudorapidity regions to reconstruct very forwaded jets the HCAL is completed by the forward calorimeters covering a range up to $|\eta| = 5.0$.

3.2.5 Superconducting solenoid

Keystone of the particle identification and the measurement of high energy momenta of produced particles is the strong magnetic field in CMS provided by the superconducting solenoid. It is a 4-layer winding composed of a reinforced NbTi conductor and yields a magnetic field of 3.8 T. The geometrical dimensions of the solenoid are a length of 12.5 m and a diameter of 6 m. It is operated at about 4.5 K.

3.2.6 Muon chambers

Muons are the only interacting particles that travel far enough to reach the region outside of the solenoid and therefore the muon detector system is the outermost part of CMS. For the identification and reconstruction three different types of gaseous detectors are used which overlap with respect to the pseudorapidity. In the barrel region up to $|\eta| < 1.2$ four layers of drift tubes are installed between the iron yokes of the magnet. The endcap region consists of cathode strip detectors covering $0.9 < |\eta| < 2.4$. As a third part the muon system contains resistive plate chambers used as a trigger system up to $|\eta| < 1.6$.

The achieved momentum resolution in the muon reconstruction, which is done by the combintion of the muon detectors and the tracker system, is about $1.3\% - 2.0\%$ for

$20 < p_T/\text{GeV} < 100$ in the barrel and below 6% in the endcaps. Muon momenta above 1 TeV can be resolved better than 10% [22].

3.2.7 Trigger system

The proton-proton bunch collision frequency is designed to be 40 MHz with a design luminosity of 10^{34} cm^{-2}s^{-1}. Roughly 20 interactions per collision are expected. Since the amount of data per time can not be streamed by any known technical concept, a two step trigger system is installed to filter the amount of data.

The Level-1 trigger is a programmable electronic trigger and designed to reduce the output to 30 kHz. It forms trigger objects like lepton candidates in limited spatial regions using local components like hit patterns in the calorimeters and the muon chambers. The global components combine all information and form more complex observables. The L1 trigger uses information that is promptly available to perform the calculations in each bunch crossing.

The high level trigger system is a software system in a filter farm and uses all the data collected by the L1 trigger. It reduces the data rate to 150 Hz [23].

4 Motivation and analysis strategy

4.1 Overview

Since one of the main goals of the LHC experiments is the search for physics beyond the Standard Model, it is nescessary to achieve a very good understanding and prediction of the Standard Model processes to substract them as backgrounds. One of the aspects that has to be considered is the misreconstruction of objects. In this thesis a closer look to the misidentification of electrons as photons is taken. Even with the high track reconstruction efficiency of the CMS pixel detector described in Section 3.2.2 some fraction of electrons do not register a track seed and therefore appear as photons in the reconstructed events.

This is of particular interest in analyses which are sensitive to final states with photons. The purpose of this thesis is therefore to estimate the fraction of true electrons that are misidentified as photons using a data driven method. In this chapter the definition of the studied quantity, which is referred to as the *fake rate*, is given and the general method for its determination is introduced.

4.2 Definition of the fake rate

The goal is to estimate the number of *true* electrons[1] originating from the decay of some particle X but which have been misreconstructed as photons. This quantity is denoted as $N^X_{e\to\gamma}$. The total number of events containing a particle X is denoted as N_X. Here the specific nature of X does not matter for the following general considerations as long as it can decay into an electron, one example for X might be the W boson.

The probability, that a real electron is misreconstructed (and thus measured) as a photon, is denoted as $p(e \to \gamma)$, while the probability, that a real electron is reconstruted correctly as an electron is denoted as $p(e \to e)$. The number of predicted real electrons faking photons can then be estimated by

$$N^X_{e\to\gamma} = N_X \cdot p(e \to \gamma)\,, \tag{4.2.1}$$

[1]From now on, particles include their anti-particles as well.

while the number of real electrons that are reconstructed correctly as electrons is

$$N^X_{e\to e} = N_X \cdot p(e \to e)\,. \tag{4.2.2}$$

To achieve the number of true electrons misreconstructed as photons, the ratio of the probability that a real electron is misreconstructed as a photon and the probability that a real electron is reconstructed as an electron has to be estimated:

$$N^X_{e\to\gamma} = N^X_{e\to e} \cdot \frac{p(e\to\gamma)}{p(e\to e)} \equiv N^X_{e\to e} \cdot R \tag{4.2.3}$$

Here, the *transfer factor* R is introduced.

In general, the sum of the probabilities $p(e \to \gamma)$ and $p(e \to e)$ should *not* be equal unity since it has to be taken into account that a real electron could be reconstructed as something else or not reconstructed at all. If this probability is denoted as $p(e \to X)$, this can be written as

$$1 = p(e\to\gamma) + p(e\to e) + p(e\to X)\;. \tag{4.2.4}$$

It will be motivated in Section 4.3 that due to a reasonable selection of considered events, $p(e \to X)$ can be neglected. Then the convenient way to express R is to define a more intuitive quantity with a clear definition which is called the *fake rate* and denoted $f_{e\to\gamma}$. The idea is that $f_{e\to\gamma}$ directly represents the fraction of times an electron gets misidentified as a photon.

It is introduced by noting that the searched fraction corresponds to the probability $p(e \to \gamma)$ and thus $p(e \to \gamma) \equiv f_{e\to\gamma}$. Due to Eq. 4.2.4 it is clear that $1 - f_{e\to\gamma}$ represents the fraction of times an electron gets reconstructed correctly and therefore corresponds to $p(e \to e)$.

From this the transfer factor can be expressed by the fake rate as

$$R = \frac{f_{e\to\gamma}}{1 - f_{e\to\gamma}}\;. \tag{4.2.5}$$

4.3 Tag and probe on the Z-boson resonance

This section explains the basic idea for the estimation of $f_{e\to\gamma}$. The details of the implementation and the definition of the objects named here will be given in the next chapters.

Since the fake rate represents the inefficiency of the inner pixel tracker to register electron track seeds and therefore might be written as $f_{e\to\gamma} = 1 - \varepsilon_{\text{pixel}}$, its determination is basically an efficiency measurement. For this it is a convenient and elegant way to use a tag and probe approach [24]. Its concept is to use a well known physics process that features two objects of the same type. An appropriate choice in the case of electrons is the decay of a Z boson. The origination of two measured objects from a Z-boson decay can be ensured by reconstructing the invariant mass distribution of both in the region of the Z-boson resonance, as briefly described in Section 2.7. It is assumed here that the inner pixel tracker efficiency is independent of the physical process that produces the electrons.

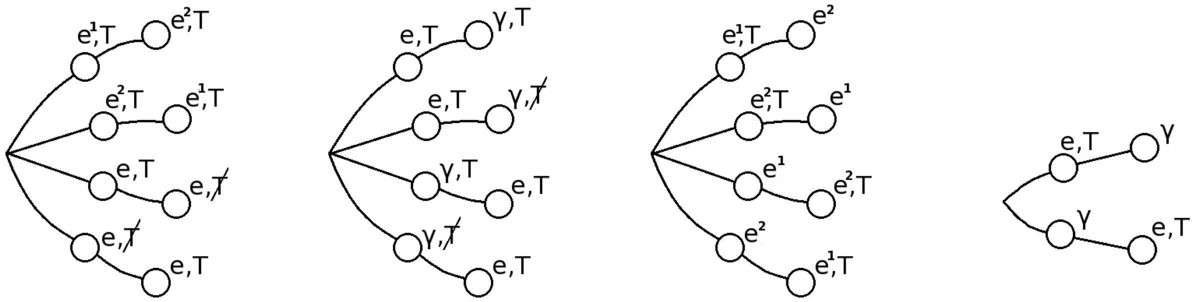
To measure such an efficiency, one of the reconstructed electrons originating from the Z boson decay is required to fulfill the corresponding requirements of the efficiency quantity to be measured, this is called the tag. The second electron is called the *probe* and leads to the efficiency estimation by dividing the number of probe electrons fulfilling the efficiency requirements by the total number of probe electrons.

To filter events that are of further interest, in this thesis a combination of the pixel seed requirement and a high level trigger that filters only events containing at least one electron is used as a preselection requirement for the tag object. This makes sure that only events are considered where a Z boson might have appeared and subsequently decayed into two electrons.

What can be measured by this kind of selection is the number of events where both electrons originating from a Z-boson decay are reconstructed correctly, denoted by $N^{Z}_{ee\to ee}$ and the number of events where one of the electrons has been misreconstructed as a photon, denoted by $N^{Z}_{ee\to e\gamma}$.

For clarification of what the fake rate means if two objects are measured per selected event, the reconstruction cases of two produced electrons and how the fake rate is defined then is explained in the following.

Considering an event with a Z decay, this process leads to two final state electrons and it is assumed that they are reconstructed each as an electromagnetic object that corresponds either to an electron or a photon. Figure 4.1 a) and b) shows the possibility tree for the assignment of the two objects to the electron-electron and electron-photon sample, respectively. Each "blob" represents an electron-photon candidate which is labeled with its reconstruction case, thus each "branch" of the shown trees represent a reconstruction possibility of both objects. Here, e indicates the reconstruction as an electron, γ the misreconstruction as a photon, T that the object fired the trigger and $\not{T}$ that the object did not fire the trigger, comma seperated labels mean a logical and-combination of both. In case that both objects fulfill the criteria in an indistinguishable manner, the objects are subscripted by numbers 1 and 2.



(a) $N^Z_{ee\to ee}$ (considering probe trigger)
(b) $N^Z_{ee\to e\gamma}$ (considering probe trigger)
(c) $N^Z_{ee\to ee}$ (only tag trigger)
(d) $N^Z_{ee\to e\gamma}$ (only tag trigger)

Figure 4.1: Cases to assign objects to an event class. See the text for the definition of the symbols.

In general, the probabilities of the electron to be reconstructed as any object and that it fires the electron trigger are correlated. If a true electron fulfills all the trigger criteria, one can assume that it is more lileky a correctly reconstructed electron and not misreconstructed as something else. For the same reason, an object that has not fired the electron trigger is more unlikely to be reconstructed as an electron.

The probability, that a true electron is reconstructed as an electron and fulfills the trigger criteria is denoted by $p(e \to e, \mathrm{T})$, and the other propabilities by $p(e \to e, \not{\mathrm{T}})$, $p(e \to \gamma, \mathrm{T})$ and $p(e \to \gamma, \not{\mathrm{T}})$ respectiveley. For the number of counted objects $O\left(N^Z_s\right)$, assigned to each event reconstruction case $s = \{ee \to ee, ee \to e\gamma\}$, meaning that both electrons are reconstructed correctly or one of the electrons is misreconstructed as a photon, this leads to the following general expressions:

$$\begin{aligned}\frac{O\left(N^Z_{ee\to ee}\right)}{N_Z} &= [p(e \to e, \mathrm{T}) \cdot p(e \to e, \mathrm{T})] + [p(e \to e, \mathrm{T}) \cdot p(e \to e, \mathrm{T})] + \\ &\quad \left[p(e \to e, \mathrm{T}) \cdot p(e \to e, \not{\mathrm{T}})\right] + \left[p(e \to e, \not{\mathrm{T}}) \cdot p(e \to e, \mathrm{T})\right] \\ &= 2 \cdot [p(e \to e, \mathrm{T}) \cdot p(e \to e, \mathrm{T})] + \\ &\quad 2 \cdot \left[p(e \to e, \mathrm{T}) \cdot p(e \to e, \not{\mathrm{T}})\right] \end{aligned} \tag{4.3.6}$$

$$\begin{aligned}\frac{O\left(N^Z_{ee\to e\gamma}\right)}{N_Z} &= 2 \cdot [p(e \to e, \mathrm{T}) \cdot p(e \to \gamma, \mathrm{T})] + \\ &\quad 2 \cdot \left[p(e \to e, \mathrm{T}) \cdot p(e \to \gamma, \not{\mathrm{T}})\right] \end{aligned} \tag{4.3.7}$$

The searched electron-electron and electron-gamma sample sizes N_s^Z correspond to the object number per sample divided by the object-per-event-number, thus $N_s^Z = O\left(N_s^Z\right)/2$.

Since the correlation between trigger and reconstruction is unknown, this can not be used for the estimation. Instead, one can neglect the probes trigger requirement and only look for the tag object to fulfill the requirements. Then the possible cases reduce to Figure 4.1 c) and d), where again in c) for the $N^Z_{ee\to ee}$ set the two possible combinations to assign the tag and probe attributes to the electromagnetic objects have to be taken into account.

Let $k_{t\leftrightarrow p}$ be the fraction of events which would be assigned to the $N^Z_{ee\to ee}$ set and where the probe object also fulfills the tag criteria and so the event has to be counted twice here. Then $(1 - k_{t\leftrightarrow p})$ is the fraction where this is not the case. Denoting the probabilities without trigger requirements as $p(e \to e)$ and $p(e \to \gamma)$ respectiveley leads to

$$\begin{aligned}
\frac{O\left(N^Z_{ee\to ee}\right)}{N_Z} &= k_{t\leftrightarrow p} \cdot [2 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to e) + 2 \cdot p(e \to e) \cdot p(e \to e, \mathrm{T})] \\
&\quad + (1 - k_{t\leftrightarrow p}) \cdot [2 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to e)] \\
&= k_{t\leftrightarrow p} \cdot [4 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to e)] \\
&\quad + (1 - k_{t\leftrightarrow p}) \cdot [2 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to e)] \\
&= (1 + k_{t\leftrightarrow p}) \cdot 2 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to e) \qquad \text{and} \qquad (4.3.8) \\
\frac{O\left(N^Z_{ee\to e\gamma}\right)}{N_Z} &= 2 \cdot p(e \to e, \mathrm{T}) \cdot p(e \to \gamma)\ . \qquad (4.3.9)
\end{aligned}$$

The transfer factor is then given by

$$R = \frac{p(e \to \gamma)}{p(e \to e)} = \frac{O\left(N^Z_{ee\to e\gamma}\right)}{(1 + k_{t\leftrightarrow p}) \cdot O\left(N^Z_{ee\to ee}\right)} = \frac{N^Z_{ee\to e\gamma}}{(1 + k_{t\leftrightarrow p}) \cdot N^Z_{ee\to ee}}\ . \qquad (4.3.10)$$

The connection to the fake rate $R = f_{e\to\gamma}/(1 - f_{e\to\gamma})$ then requires

$$f_{e\to\gamma} = \frac{N^Z_{ee\to e\gamma}}{(1 + k_{t\leftrightarrow p}) \cdot N^Z_{ee\to ee} + N^Z_{ee\to e\gamma}}\ . \qquad (4.3.11)$$

If the probe object fulfills the tag criteria in every electron-electron event, then $k_{t\leftrightarrow p} = 1$ and one would end up with a factor of 2 in the denominator.

A common notation for this term is

$$f_{e\to\gamma} = \frac{N_{e\to\gamma}}{N_{e\to e} + N_{e\to\gamma}} \qquad (4.3.12)$$

where $N_{e\to e} \equiv (1 + k_{t\leftrightarrow p}) \cdot N^Z_{ee\to ee}$ and $N_{e\to\gamma} \equiv N^Z_{ee\to e\gamma}$.

4.4 Analysis strategy and the template method

In this section the basic analysis strategy of this thesis is briefly described in a general way. It is assumed that two sets of pairwise objects assigned to $N_{ee\to ee}$ and $N_{ee\to e\gamma}$ have been selected from data and the kinematic properties of each stored object are accessible. The technical details of the selecton of events where two electrons originating from a Z-boson decay appear will be described in Chapter 6.

Following the arguments in Section 2.7 the invariant mass spectrum of the decay products of the Z boson will follow a specific distribution. In general, the invariant mass of n particles can be obtained by summing up the four vectors $p = (E, \vec{p})$ of the particles and calculating the sum square:

$$M^2_{\text{inv}} = \left(\sum_{i=1}^{n} p^i \right)^2 . \tag{4.4.13}$$

A usefull relation is the invariant mass of two particles with known transversal momenta, which is a high-energy approximation neglecting the masses of the particles themselves:

$$M^2_{\text{inv}} = 2 p^1_T p^2_T \left(\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2) \right) \tag{4.4.14}$$

If the selection was ideal, meaning that each selected $N_{ee\to ee}$ event would really originate from a Z-boson decay, the task would be simple. One would have to count the events in each set and could directly calculate the fake rate using Eq. 4.3.11.

In a statistical data analysis, the selection will never be ideal and therefore also object pairs are selected where the invariant mass appears in the considered range by chance. The question then is how to estimate the fraction of unwanted selected events, appearing as a background in the considered distribution.

A powerful method that is applied in this thesis is the *template method*. Here the known processes that occur in the considered region are extracted from simulation or by data driven methods. The formed distributions build the templates. Now the differences of the shapes between the selected data and the sum of the templates is exploited. The fraction of each template is a free parameter that is estimated by optimising the agreement between the data and the sum of the templates. The template method can be motivated in a Bayesian statistical framework and the details can be found e.g. in Ref. [24]. A schematic visualization of optimizing the fractions of two templates to a data distribution is given in Fig. 4.2 for the simple case of gaussian shapes.

The goal of this analysis is therefore the selection of events with two electrons originating from a Z-boson decay and the appropriate modeling of a background and a signal template to estimate the electron fake rate.

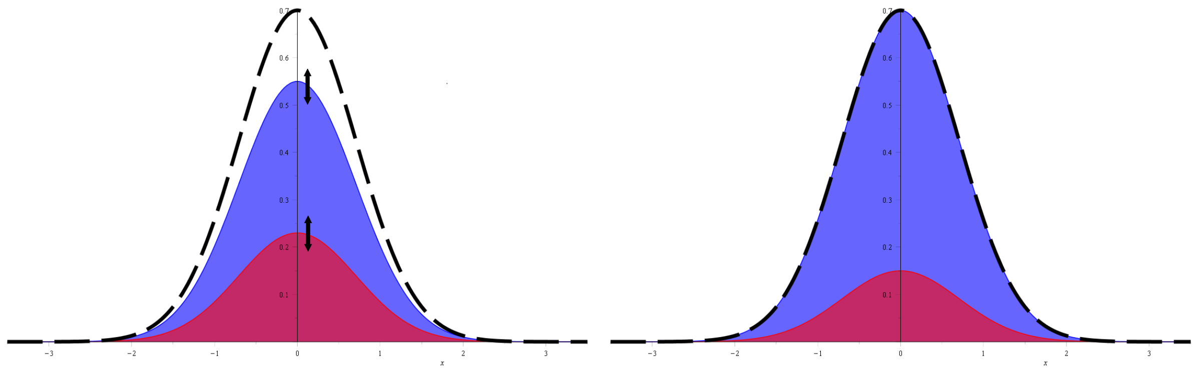


Figure 4.2: Schematic example of the idea of fitting the fraction of different template shapes. The black dashed line represents some data distribution and in the left figure two initial template shapes are shown. In the right figure the fractions of the template shapes normalizations are optimized to reproduce the data distribution. The optimization is not drawn to scale.

5 Object definition

In the proton-proton collisions in the center of the CMS detector many different physical processes occur leading to several final-state particles. These travel through the different subdetectors of CMS and most of the known SM particles interact in a measurable way with the detectors described in Chapter 3. The task is then to reconstruct the physical objects from the detected signatures. In this chapter a brief description of the Particle Flow algorithm will be given and the physical objects used in this thesis, which are electrons, photons and muons will be defined.

5.1 Particle-Flow Algorithm

The most commonly used algorithm to identify and reconstruct all the particles from a collision in CMS is the Particle Flow algorithm (PF) [19,25,26]. It is based on an efficient track reconstruction, a clustering algorithm to achieve a good assignment and abscission of overlapping showers and an efficient procedure to link all the different subdetector energy deposits of each particle.

The first step is the reconstruction of the energy deposits in the tracker system and their assignment to charged particle tracks. Each track is then extrapolated through the calorimeter entries, which are combines to clusters, in the sense that if it falls within the boundary of at least one cluster these clusters are associated to the track. In the case of muons the energy deposits in the outer muon chambers are combined with the tracker information. The associated subdetector entries are then assigned to a particle category where the algorithm identifies muon, electrons and photons as well es charged and uncharged hadrons. Electrons, photons and muons will be described in the next two sections in further detail.

Since hadronic final states occur due to the process of hadronisation mentioned in Sec. 2.5 which leads to spatial cone-like bounded tracks originating from the same vertex called jets, tracks from the primary vertex and all neutral particles are then clustered to jets by common algorithms like the anti-k_T-algorithm [27].

For the search for new physics beyond the SM it is assumed that stable unknown particles would most likely not (or very weakly) interact with the known detector technologies. The CMS detector due to its hermetic architecture is able to identify *missing energy* of an event. Assuming that the colliding protons have negligible transversal momenta, the transversal momentum vectors of all final state particles should sum up to zero due to momentum conservation. The calculated difference of this sum to zero should therefore correspond to the total momentum of undetected particles, which from the SM particle content are the neutrinos. This is called the *missing transverse energy* and is defined as

$$\not{E}_T = \left|\vec{\not{E}}_T\right| = \left|-\sum_{i=1}^{N} \vec{p}_T^{\,i}\right| \tag{5.1.1}$$

with N being the number of reconstructed particles.

The reconstructed objects are then stored in the `RECO` data format which contains the complete information of the reconstruction, and then reprocessed to the analysis object data format (`AOD`) which is a subset of `RECO` and is designed to be directly used in data analysis [28]. In this thesis the compressed `AOD` format `MiniAOD` is used where a lot of detailed track information which is not of further interest to a particular analysis is packed in a compact way. Details about the content of `MiniAOD` can be found in Ref. [29, 30].

5.2 Electrons and photons

Electons and photons are the key objects of this analysis. Both have a similar signature in the ECAL and therefore the key attribute for the distinction is their interaction with the tracker system. While electrons are electrically charged and therefore ionize the silicon pixels in the tracker, photons are electrically neutral and do not leave a track seed. Here the *pixel seed* is defined as a boolean which is true if the reconstructed electromagnetic (egamma) object, defined in the following, registered a pixel hit in two or more layers of the inner pixel tracker as shown in the upper sketch of Fig. 3.5.

In general photon objects are reconstructed by deposits in the ECAL superclusters. Electrons are reconstructed by matching ECAL superclusters to track seeds found in, or close to, the pixel detectors [31]. On the tracker seeds the tracks are reconstructed using the GSF tracking algorithm [32].

To discriminate true electron-gamma objects from background, several requirements to further object variables are made. The *loose* identification (ID) is defined in Tab. 5.1.

This identification is recommended to be used when backgrounds are rather low as it is expected on the Z-boson resonance and therefore this identification is used in this thesis.

Table 5.1: Photon loose-ID parameter criteria for barrel photon objects [33].

Parameter	Criterium
H/E	< 0.05
$\sigma_{i\eta i\eta}$	< 0.0102
$\mathrm{Iso_{CH}}$	< 3.32 GeV
$\mathrm{Iso_{NH}}$	$< 1.92 + 0.014 \cdot p_{T,\gamma} + 0.000019 \cdot p_{T,\gamma}^2$
Iso_γ	$< 0.81 + 0.0053 \cdot p_{T,\gamma}$

The identification parameters are given by:

H/E: The ratio of the energy deposit in the HCAL tower which is closest to the seed of the ECAL supercluster assigned to the track and the ECAL deposit. The usage of only the single HCAL towr closest to the superclaster makes this variable less sensitive to pile-up. The H/E requirement suppresses jets thats are misidentified as photons.

$\sigma_{i\eta i\eta}$: A quantity for the width of the shower in the ECAL. It is calculated in 5×5 crystal array in the $\eta - \phi-$plane around the supercluster seed and is technically defined by

$$\sigma^2_{i\eta i\eta} = \frac{\sum_i^{5\times5} w_i(\eta_i - \eta_{seed})^2}{\sum_i^{5\times5} w_i}, \qquad w_i = \max\left[0, 4.7 + \ln\left(\frac{E_i}{E^{5\times5}}\right)\right] . \tag{5.2.2}$$

$\mathrm{Iso_{CH}}$: This is the isolation with respect to charged hadrons. It is calculated by the energy deposit in a cone $\Delta R < 0.3$ around the reconstructed egamma object that can be assigned to charged hadrons and the isolation energy is corrected for pile-up. The assignment of energy deposits to charged and neutral hadrons as well as photons is done by the PF algorithm.

$\mathrm{Iso_{NH}}$: The isolation energy for neutral hadrons, analogous to the charged hadron isolation. The p_T-dependency corrects for the correlation between the photon's p_T and the neutral hadron isolation energy.

Iso_γ: Similar to the charged and neutral hadron isolation this is the photon isolation which is also p_T dependent. The requirement suppresses photons originating from jets.

5.3 Muons

Muons are reconstructed by consistent hits in the inner and outer tracker, the energy deposit in the calorimeters hits and the muon detector system where muon candidates are reconstructed using the Kalman-Filter algorithm [34]. In this thesis muons with a cut based loose ID are used which are Particle Flow muons either reconstructed as global muons or arbitrated tracker muons [35, 36]. This ID is highly efficient for prompt muons as expected from the decay of heavy vector bosons such as the Z boson.

6 Event selection

In this chapter the technical details of the event and object selection are described. The used datasets and triggers are stated and motivated and the implementation of the tag and probe method is explained.

6.1 Data samples

6.1.1 Data

While the LHC delivered an integrated luminosity of 4.22 fb^{-1} in the year 2015, the amount of data recorded by the CMS detector due to the detector run time only amounts to 3.81 fb^{-1}. In this thesis only the CMS data where the magnet operated at $B = 3.8$ T and with a proton-proton bunch-crossing interval of $\Delta T = 25$ ns is used, corresponding to an integrated luminosity of $L_{\mathrm{int}} = 2.3\ \mathrm{fb}^{-1}$.

By applying combinations of high level trigger paths as described in Sec. 3.2.7, different data sets are created for further usage in data analyses.

The searched event topology includes at least one real electron and therefore the `SingleElectron` dataset is used where one or more electrons appear in each event. To form the background template, which will be described in Section 6.4, a dataset filtered for at least one muon and one electromagnetic object per event is used. As a control dataset also a subset containing at least one muon has been applied where its usage is described in Section 6.2. All datasets are shown in Tab. 6.1 with their technical path names in the CMS data aggregation system (DAS) which is described in Ref. [37].

6.1.2 Simulation

Also multiple Monte Carlo simulation sets are used for comparison with the data, to construct the probability distributions and for cross checking the results. These are listed in Tab. 6.2 with the technical DAS path names.

Table 6.1: Datasets used for event selection and background estimation.

/SingleElectron/Run2015C_25ns-16Dec2015-v1/MINIAOD
/SingleElectron/Run2015D-16Dec2015-v1/MINIAOD
/MuonEG/Run2015C_25ns-16Dec2015-v1/MINIAOD
/MuonEG/Run2015D-16Dec2015-v1/MINIAOD
/SingleMuon/Run2015C_25ns-16Dec2015-v1/MINIAOD
/SingleMuon/Run2015D-16Dec2015-v1/MINIAOD

Table 6.2: Simulated datasets.

Process	Technical path name
$Z \to ee$	/DYToEE_NNPDF30_13TeV-powheg-pythia8/*/MINIAODSIM
$t\bar{t}$ + jets	/TTJets_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/*/MINIAODSIM
$W \to l\nu$	/WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/*/MINIAODSIM

For the calculation of the hard processes the *POWHEG* [38] and *MadGraph* [39] event generators are used and the parton shower and fragmentations are generated by *PYTHIA8* [40]. The reconstruction of the simulated particle tracks in the CMS detector is simulated by the detector simulation package *GEANT* [41]. All samples are generated with a 25 ns bunch crossing.

Pile-up reweighting

Besides the hard interaction which is defined as the highest energetic collision in an event, several low energy interactions occur in the crossing of two proton bunches. These particle tracks originating from such interaction vertices that are not the primary vertex are called pile-up. In simulation, the expected distribution of the number of collisions per bunch-crossing differs from the measured distribution in data. Thus for comparisons between data and simulation, simulated events have to be scaled with a weight that depends on the number of estimated proton-proton vertices. The estimation of the weight factors is done by the ratio of the number of events in data and the expected number of events for specific numbers of vertices.

Luminosity reweighting

In addition to the pile-up reweighting, Monte Carlo samples are typically generated with different integrated luminosities compared to the data. Assuming that the cross section σ

of a certain physical process used in the generated sample is the same as the cross section for that process in the measured events, the simulated events can be reweighted by

$$w = \frac{N_{\text{exp}}}{N_{\text{gen}}} = \frac{\sigma \cdot L_{\text{int}}}{N_{\text{gen}}} \tag{6.1.1}$$

where Eq. 3.1.2 has been used and N_{gen} is the number of generated Monte Carlo events. The cross sections and generated events for the used MC samples are listed in Tab. 6.3.

Table 6.3: Cross sections [42] and number of generated events for the used MC samples.

Process	σ [pb]	N_{gen}
$Z \to ee$	1921.8	$4.96535 \cdot 10^7$
$W \to l\nu$	61526.7	$4.71613 \cdot 10^7$
$t\bar{t}$ + jets	831.76	$9.89924 \cdot 10^6$

Both the pile-up and the luminosity reweighting are applied.

6.2 Trigger

For the selections the signal triggers listed in Tab. 6.4 are applied and their usage will be explained further in the following.

Table 6.4: Signal and baseline trigger with technical path names.

Single Electron	Signal	`HLT_Ele27_eta2p1_WPLoose_Gsf`
	Baseline	`HLT_Ele22_eta2p1_WPLoose_Gsf`
Single Muon	Signal	`HLT_Mu38NoFiltersNoVtx_Photon38_CaloIdL`
	Baseline	`HLT_Mu24_eta2p1`

The main selection should filter events with at least one electron which naturally leads to the choice of a single electron trigger. In particular a single electron trigger with the requirements $p_T > 27$ GeV and $|\eta| < 2.1$ is used. The low cuts on the transversal momenta and the pseudorapidity range allow more strict requirements in the further analysis.

For the construction of a proper background template a trigger filtering for events with at least one muon with $p_T > 38$ GeV and one photon with $p_T > 38$ GeV has been used. It will be described in Section 6.4 that the selection only uses the muon part of the trigger which therefore is of further interesst. The choice of a muon-photon trigger is motivated by the studies on the electron fake rate that have been performed on the 8 TeV data [43–45] and the chosen trigger is the 13 TeV equivalent to the one used before [46].

From each selection that is done in this analysis, a ratio will be calculated such as the fake rate definition from Eq. 4.3.12. Due to this, the efficiencies of the triggers will cancel out and thus are not of further interest and therefore not studied here.

To validate the p_T turn-on of each trigger a simple application of a common efficiency measurement method is done. The efficiency is measured using only events passing a baseline trigger and the efficiency of the signal trigger results to

$$\varepsilon_{\text{signal}} = \frac{\#\,(\text{signal} \wedge \text{baseline})}{\#\text{baseline}} \,. \tag{6.2.2}$$

To measure the total efficiency and not only the turn-on, the baseline trigger should have requirements which are independent of those of the signal trigger such that the baseline requirements cancel out. Since here only the turn-on of each trigger is validated, the requirement of signal and baseline trigger to be independent is not nescessary. For the single muon trigger the same trigger but with a lower p_T requirement is used while for the muon-photon trigger a single muon trigger is an appropriate choise to check the turn-on of the muon part of the signal trigger. The measurement for the electron trigger is done in the `SingleMuon` dataset to be not biased from the event topology and the muon-photon trigger is measured in the `SingleElectron` dataset.

It should be noted that the plateaus in the efficiency plots shown in Fig. 6.1 thus do *not* correspond to the efficiencies and the shown distrubutions are only proportional to the actual efficiency. Both turn-ons can be seen to be at the expected transversal momentum requirements. The muon part of the muon-photon-trigger reaches its plateau at $p_T = 40$ GeV and thus this is used as a minimum transversal momentum threshold for all the object selections described in this chapter.

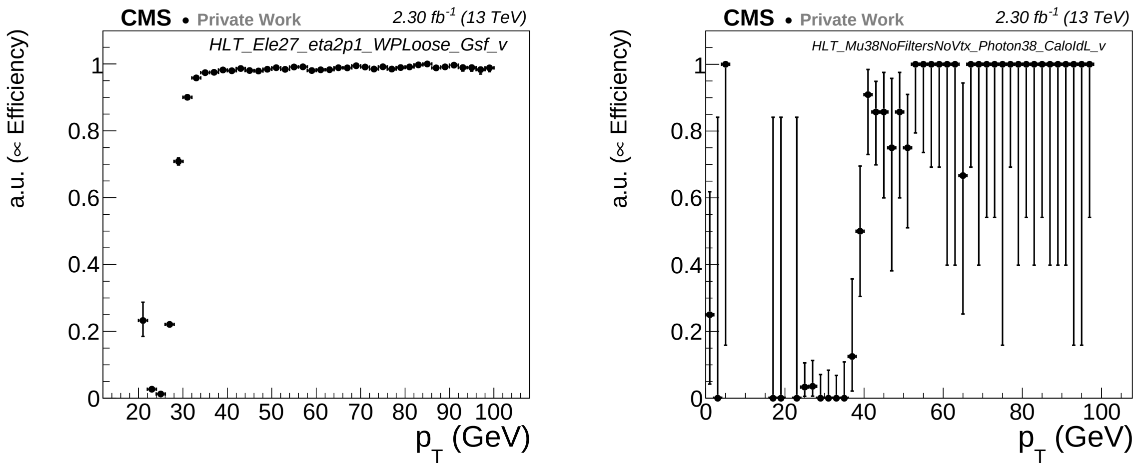


Figure 6.1: Turn-ons of the used triggers. The plateaus do not correspond to the trigger efficiencies, see the text.

6.3 Tag and probe

For the selection of tag and probe the `SingleElectron` dataset is used. After applying the single electron trigger on the data, electromagnetic objects taken from the photon collection are used. Due to the turn-on of the muon part of the used muon trigger an offline cut on the transversal momentum on the objects is applied. Also only objects with an absolute pseudorapidity corresponding to the barrel part of the inner pixel tracker are used.

The applied offline base selection is therefore

- $p_T > 40$ GeV,
- $|\eta| < 1.4442$.

The selection for the tag object is the following.

First, only the preselected electromagnetic objects that possess a pixel seed are used. Second, to make sure that the chosen tag object really has fired the trigger, it has to be matched to an electron reconstructed by the HLT and which is used to trigger this event.

For this the spatial separation ΔR of the reconstructed tracks of the electromagnetic objects (emObj) and the triggered eletron objects (trObj) as well as their transversal momentum ratio $p_T^{\mathrm{emObj}}/p_T^{\mathrm{trObj}}$ is considered. The two dimensional distribution of these two quantities is shown in Fig. 6.2.

Electromagnetic objects that match within

- $\Delta R(\mathrm{emObj}, \mathrm{trObj}) < 0.15$,

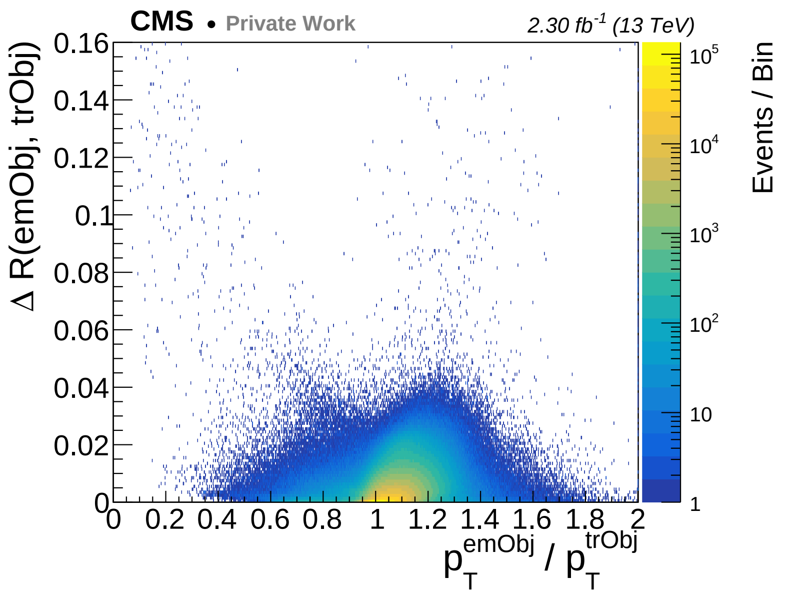


Figure 6.2: Distribution of electromagnetic objects and trigger objects.

- $0.8 < \frac{p_T^{\text{emObj}}}{p_T^{\text{trObj}}} < 1.4$

to a trigger object are tag candidates.

If more than one candidate is available the tag object is chosen randomly from the candidates to not bias the selection with respect to further requirements.

For the probe object, one of the electromagnetic objects that has not been chosen yet is taken. Also in case that there is more than one available the choice is done randomly. If more than one appropriate tag candidate has been available the chosen probe might also fulfill the tag criteria.

The so prepared probe object is then checked if it possesses a pixel seed. Due to this, two samples are prepared which represent the numerator and the denominator of Eq. 4.3.11 and where the numerator is a full subset of the denominator.

If the probe doesn't possess a pixel seed, the event represents the $N_{ee\to e\gamma}$ subset and is assigned to both the numerator and the denominator sample.

If the probe posess a pixel seed, the event belongs to the $N_{ee\to ee}$ subset and is assigned to the denominator once. Also the probe is checked if it fulfulls the tag criteria mentioned above and thus tag and probe are indistinguishable, in that case the event is assigned to the denominator twice. This makes sure that the factor $k_{t\leftrightarrow p}$ is included in the denominator sample.

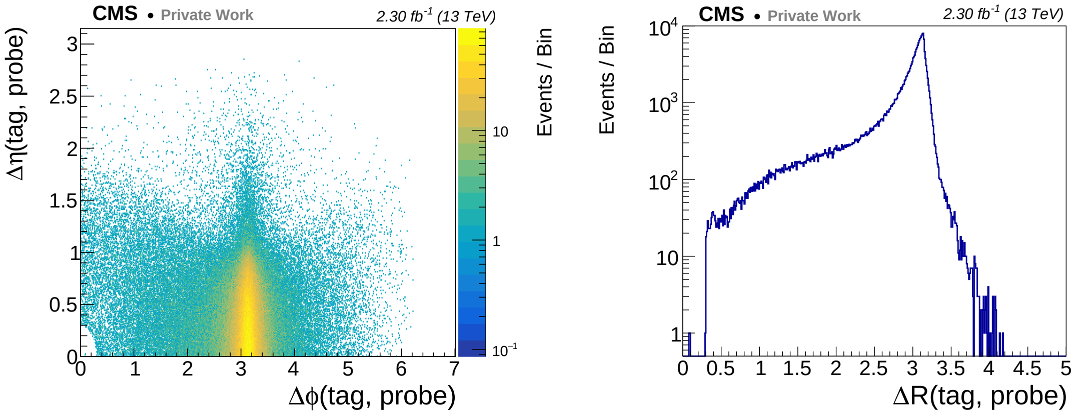


Figure 6.3: Seperation of tag and probe object selection in the η-ϕ-plane and in terms of ΔR. For a more clear visualization, the $\Delta\phi$ calculation has been done in the positive sense of rotation only, leading to a range of $\Delta\phi \in [0 \ldots 2\pi)$.

The spatial separation of the selected tag and probe objects is shown in Fig. 6.3. As it can be seen the objects are seperated by $\Delta R \gtrsim 0.3$ due to the selection and therefore no further ΔR requirement between tag and probe is applied.

In Fig. 6.4 the transversal momentum distributions of the selected tag and probe samples is shown. The shapes look as expected with most of the objects being low energetic and a steep slope towards higher energy.

In addition the same selection as described for data has been done for the three simulated samples.

Since the reconstruction of electrons and photons in the CMS detector has been studied very well, correction factors on the objects momenta due to the detector resolution and scaling in the reconstruction has been developed [47]. These are accessible for data and simulation and have been applied in this analysis. The achieved improvement will be shown below when it comes to a comparison of data and simulation.

For the prepared numerator and denominator samples now for each tag and probe pair the invariant mass is calculated and considered in the range $(60, 120)$ GeV. In Fig. 6.5 the tag and probe invariant mass spectra for numerator and denominator sample in data and simulation is shown where the energy smearing and scaling corrections has *not* been applied. Both are normalized to unity for a better shape comparison. As it can be seen the shapes of the data distributions are slightly shifted with respect to the simulated distributions.

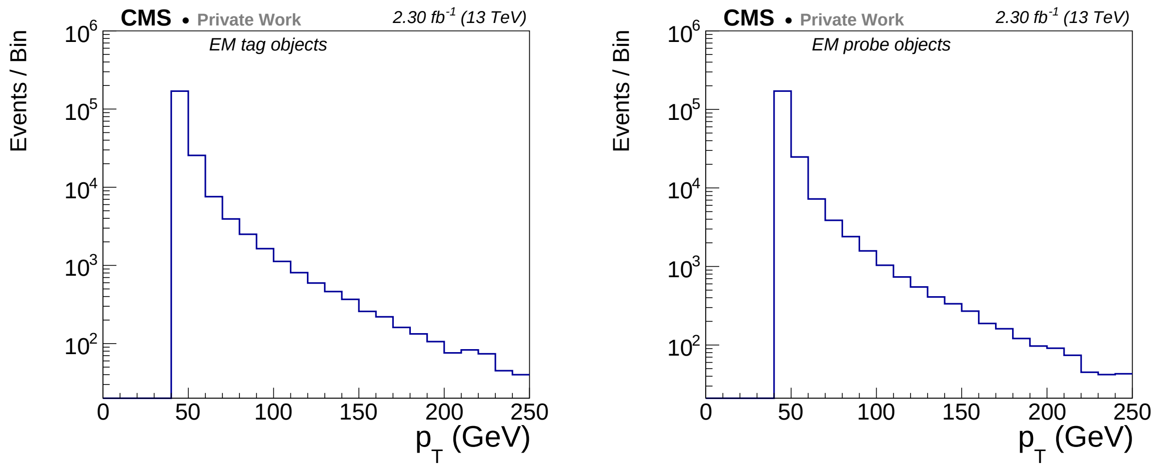


Figure 6.4: Transversal momentum distributions of the selected tag and probe samples.

After the energy smearing and scaling corrections have been applied, the shapes of the data and simulation fits much better as shown in Fig. 6.6. It can be seen that the corrections resolve the shape shifting quite well.

The unnormalized distributions will be used to estimate the fake rate as it will be described in Chapter 7.

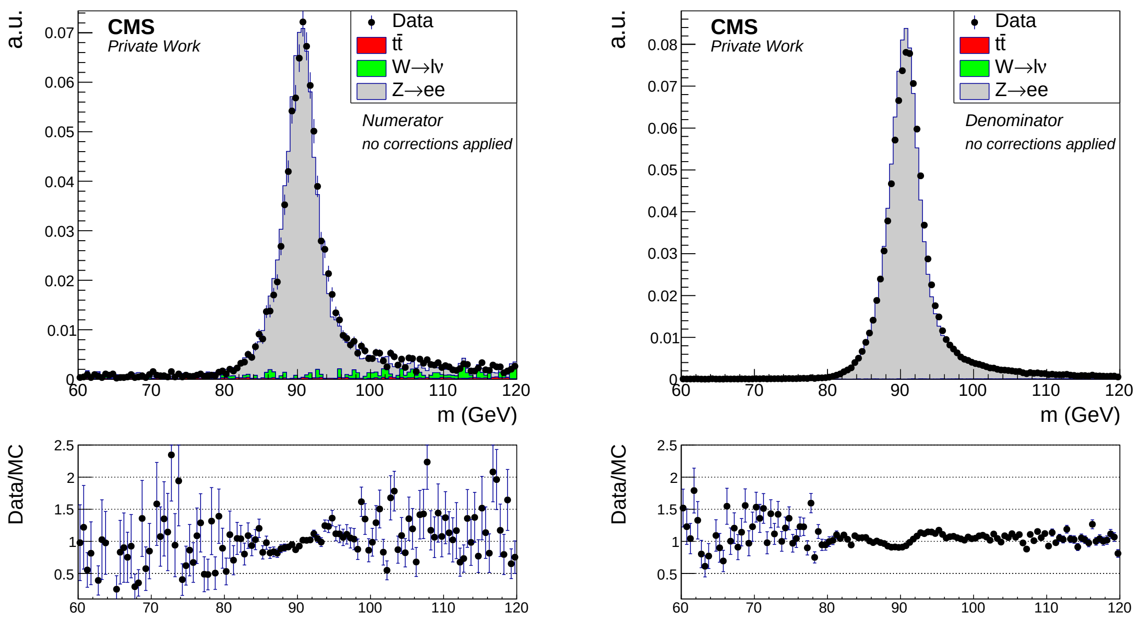


Figure 6.5: Comparison of data and simulation with no smearing corrections applied. Both are normalized to unity for a better shape comparison.

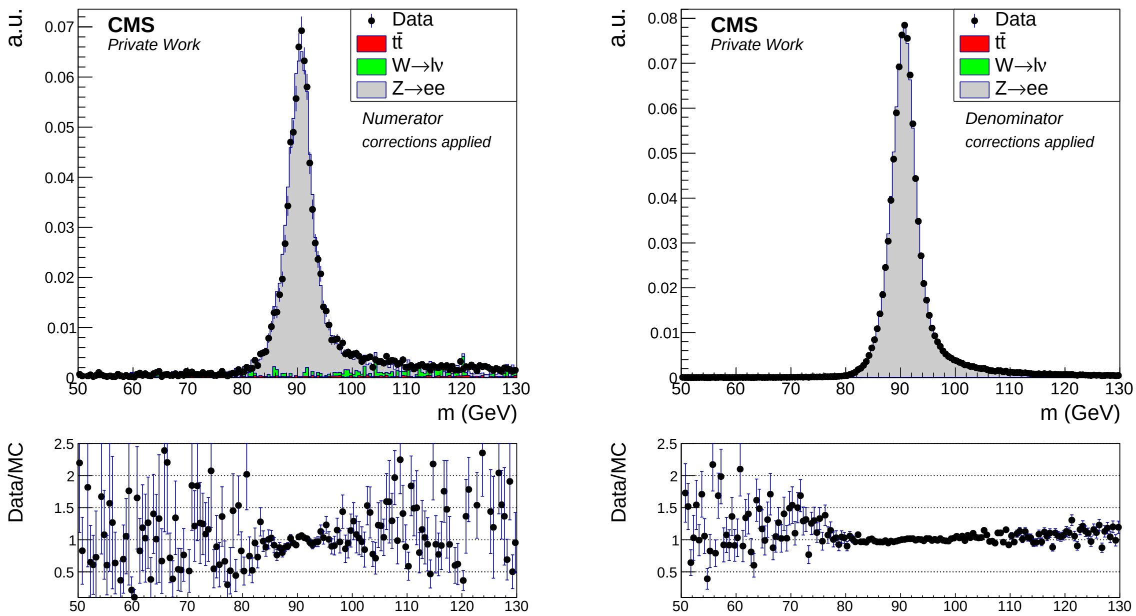


Figure 6.6: Comparison of data and simulation with smearing corrections applied. Both are normalized to unity for a better shape comparison.

6.4 Construction of the templates

To estimate the fraction of selected object pairs in the considered invariant mass distributions that truly originate from a Z-boson decay, referred to as the *signal* subsample, two template are created.

For all selected objects the same offline cuts as in Section 6.3 are applied:

- $p_T > 40$ GeV
- $|\eta| < 1.4442$

6.4.1 Signal

The signal sample should represent the ideal invariant mass distribution if only real electrons originating from a Z-boson decay would be selected. For this, the $Z \to ee$ simulation sample is used and tag and probe objects are selected by the same criteria as described in Section 6.3.

Instead of preparing a nominator and a denominator sample, only the $(1 + k_{t\leftrightarrow p}) \cdot N^{Z}_{ee\to ee}$ part of the denominator is used, meaning only tag and probe pairs where both are commutable. This makes sure to reproduce the Z resonance as good as possible. As an additional requirement it is demanded that the tag and probe objects both match to a generated electron originating from an intermediate Z boson. Figure 6.7 shows the two dimensional distribution in ΔR and the transversal momentum ratio between all selected electromagnetic objects (emObj) and generated electrons (genEl). From this the requirements

- $\Delta R(emObj, genEl) < 0.15$,
- $0.8 < p_T^{\mathrm{emObj}}/p_T^{\mathrm{genEl}} < 1.2$

have been chosen.

The signal sample thus prepared is shown in Fig. 7.3 when it comes to the construction of the corresponding density distribution. It is assumed that the signal shape should be present in both the numerator and denominator and thus only one signal template is constructed. This assumption is validated in Chapter 7 and the density functions based on the templates are able to reproduce the data distributions after the fit is applied.

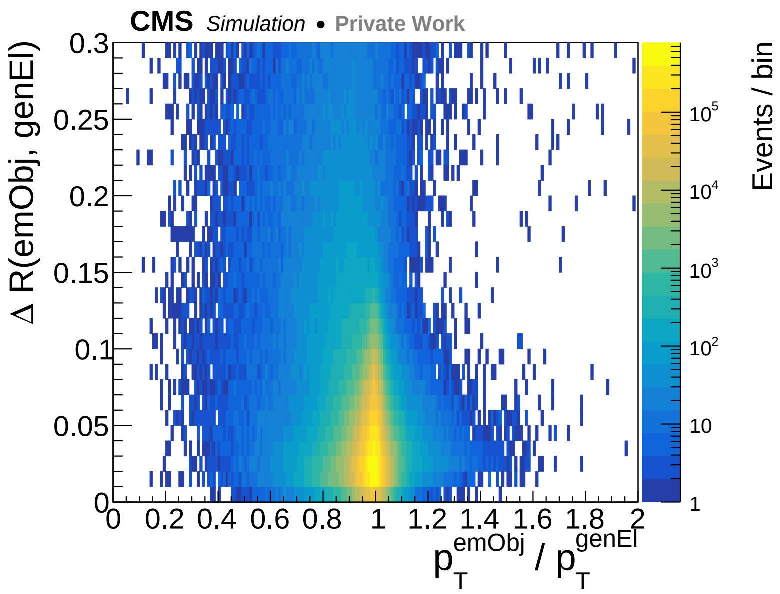


Figure 6.7: Distribution of selected electromagnetic objects and generated electrons in the $Z \to ee$ simulation sample.

6.4.2 Background

Now the fraction of selected events that do *not* correspond to a Z-boson decay will be estimated.

The main backgrounds on the Z peak will arise from processes containing a real electron such as $W \to l\nu + \text{jets}$ and $t\bar{t} + \text{jets}$. The idea is to exploit the lepton flavour symmetry in QED processes. Assuming that the fraction of muons faking photons is negligible, the background can be approximated by the invariant mass distribution using a muon as tag. For this the `MuonEG` dataset is used where the tag object (muon) has to match the muon leg of the single muon trigger (trObj) mentioned in Section 6.2. The matching criteria are the same as for the trigger-object matching in the signal template with

- $\Delta R(muon, trObj) < 0.15$,
- $0.8 < p_T^{\text{muon}} / p_T^{\text{trObj}} < 1.2$

and are set from the distribution shown in Fig. 6.8.

One background template is constructed for numerator and denominator separately and both are shown in Fig. 7.4.

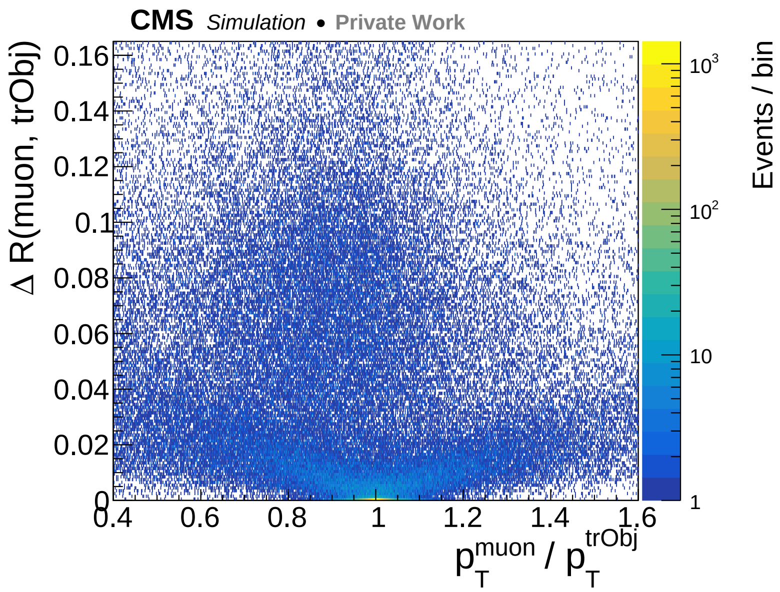


Figure 6.8: Distribution of trigger objects corresponding to the muon leg of the trigger and reconstructed muons.

7 Fake rate estimation

In this chapter a brief description of constructing density functions from the templates and how to fit them to the data is given. The fit results are shown and a parametrized model is presented as an alternative approach.

7.1 Estimation of probability density functions from templates

As described in Section 4.4 the constructed templates will be used to estimate the fractions of signal and background on the Z-boson resonance respectively. The selected unbinned invariant mass datapoints from the tag and probe method are assumed to follow an underlying probability density function (pdf), denoted as f, with respect to the invariant mass m. In this section the key concepts of density estimation, which deals with the problem of estimating a density function based on some data sampled by this pdf, will be described.

7.1.1 Histogram based pdf

A commonly used method to achieve an approximation of the sample pdf of a dataset is to create a histogram. For any practical application this is already done in an implicit meaning, e.g. if one wants to create a visualization of a selected dataset. It simply means to define ranges with respect to the variable of interest, called *bins*, then to assign each data point to the appropriate bin and finally sum up all the assigned points per bin.

In a more formal notation this means that one defines an *indicator function* j [48] such that

$$j(x, w) = \begin{cases} 1 & -w/2 \leq x < w/2 \\ 0 & else \end{cases} \tag{7.1.1}$$

where x denotes the variable of interest and w is the width of the bin. The histogram distribution may then be written as

$$h(x) = \sum_{n=1}^{N} j\left(x - \tilde{x}_n, w\right) \tag{7.1.2}$$

with a total number of datapoints N and $\tilde{x}_n$ meaning the center of the bin where the n^{th} datapoint is assigned to. In general the bin width w might also depend on the corresponding bin center such that $w \to w_n$, but is assumed to be equal for each bin here for simplicity.

The *estimator* $\hat{f}(x)$ for the underlying sample distribution is the nomalized histogram distribution

$$\hat{f}(x) = \frac{1}{Nw} h(x) \ . \tag{7.1.3}$$

This simple form of a histogram might also be referred to as a histogram based pdf estimation with a zeroth order interpolation, meaning that the pdf is described as a piecewise defined function and for each bin a constant density is assumed. Assuming that the true sample distribution $f(x)$ is continuous it is now possible to increase the *smoothness* of the estimator by increasing the interpolation order.

Each bin center $\tilde{x}_k$ corresponds to a value $\tilde{y}_k$ representing the total amount of datapoints assigned to the bin where $k \in [1, K-1]$ and K being the total number of bins. Increasing the interpolation order means to replace the indicator function by a polynomial $g_{k\to k+1}(x)$ of the order M that connects the bin center point $(\hat{x}_k, \hat{y}_k)$ with the point $(\hat{x}_{k+1}, \hat{y}_{k+1})$ and is zero outside the corresponding x-range:

$$g_{k\to k+1}(x) = \begin{cases} \sum_{l=0}^{M} \hat{g}_{k,l} x^l & x \in [\hat{x}_k, \hat{x}_{k+1}] \\ 0 & else \end{cases} \tag{7.1.4}$$

The histogram distribution h is then defined by the sum of the polynomials and the density estimator $\hat{f}$ is again the normalized h. Since this only leads to $K-1$ polynomials the density function is only defined from the first bin center to the last bin center.

Now in general $(K-1) \cdot (M+1)$ unknown coefficients $\hat{g}_{k,l}$ have to be estimated in a way that the resulting density distribution is kind of smooth. For an interpolation order of $M = 1$ only one solution exists due to the requirement that each linear polynomial goes through the two bin center points that have to be connected.

For $M \geq 2$, meaning that each polynomial has ≥ 3 free parameters, additional constraints are needed. As an estimator for the *global smoothness* it is reasonable to calculate the sum of the squared derivative differences of the polnomials at each bin enter $\hat{k}$:

$$\hat{S} = \sum_{\hat{k}=2}^{K-2} \left(\left| \frac{\partial g_{\hat{k}\to\hat{k}+1}(x)}{\partial x} \right|_{x=\hat{k}} - \left| \frac{\partial g_{\hat{k}-1\to\hat{k}}(x)}{\partial x} \right|_{x=\hat{k}} \right)^2 \tag{7.1.5}$$

Now $\hat{S}$ has to be minimized in the $(K-1)\cdot(M-1)$ dimensional parameter space where $M-1$ means that 2 coefficients per polynomial are fixed due to the point connection requirement mentioned above. Since this poblem in general has no unique solution it is useful to add appropriate constraints to stabilize numerical iterative solvers. A natural requirement it that the differences of the derivates at each point become as small as possible. This might be formulated as

$$\left| \frac{\partial g_{\hat{k}\to\hat{k}+1}(x)}{\partial x} - \frac{\partial g_{\hat{k}-1\to\hat{k}}(x)}{\partial x} \right|_{x=\hat{k}} \overset{!}{\leq} \epsilon \qquad \text{for} \qquad \hat{k} \in [2, K-2] \tag{7.1.6}$$

with a global derivative changing tolerance ϵ.

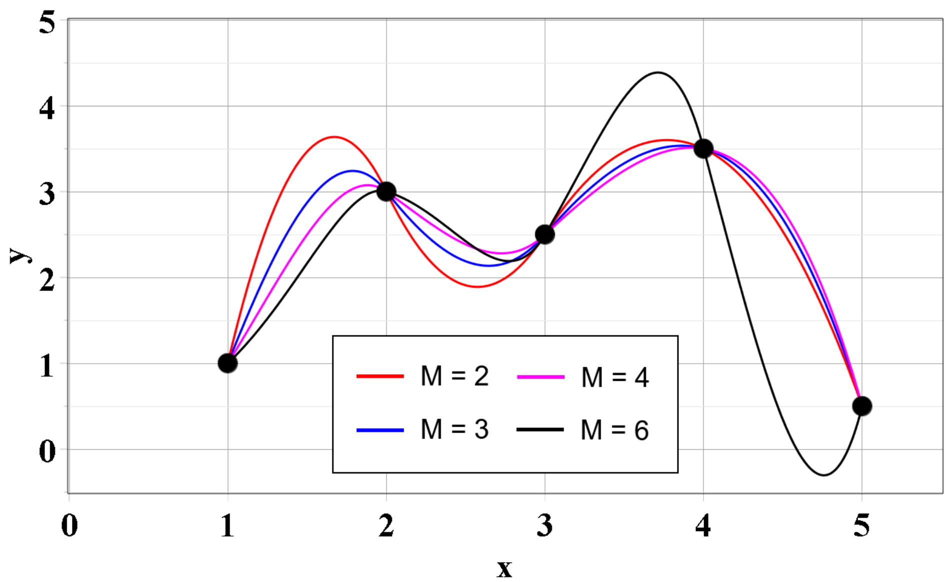


Figure 7.1: Examples for a histogram based pdf estimation in different interpolation orders as described in the text with $K = 5$ datapoints and $\epsilon = 0.01$. The black dots represent the bin centers.

The minimization of $\hat{S}$ defined by Eq. 7.1.5 under the $(K-1)$ constraints from Eq. 7.1.6 with an appropriate choice of ϵ leads to a smooth histogram based density function as shown in Fig. 7.1. In common algorithms ϵ is increased in a way that the data can be resolved by a converging minimization.

As it can be seen, the choice of the interpolation order has to be done in a way that the smoothness is valid in comparison to the data. A good rule of thumb is as much as needed, as little as possible.

The histogram based pdf estimation with interpolation order M is implemented in the *KEYS* package as desribed in Ref. [49] and this package is reimplemented in the *RooFit* framework [50] based on the analsis software *ROOT* [51].

7.1.2 Gaussian kernel estimation

A more sophisticated method to obtain the underlying probability density function of a distribution are Kernel Estimators [49, 52].

Here the indicator function is replaced with something else:

$$\hat{f}(x) = \frac{1}{N} \sum_{n=1}^{N} k(x - x_n, w) \tag{7.1.7}$$

The *kernel function* $k(x, w)$ is normalized to unity. This constraint allows to directly obtain a continuous estimation. The usual form of the kernel function is

$$k(x - x_n, w) = \frac{1}{w} K\left(\frac{x - x_x}{w}\right) \tag{7.1.8}$$

and thus for fixed w the fixed kernel estimator is

$$\hat{f}_0(x) = \frac{1}{Nw} \sum_{n=1}^{N} K\left(\frac{x - x_x}{w}\right) \ . \tag{7.1.9}$$

Now the parameter w is a *smoothing* parameter in a sense that each data point is assumed to be smeared in x direction over regions of order w. It is obvious that the smoothing parameter should be able to reproduce regions of higher density by using a narrower smoothing under the assumption, that regions with more data can be estimated better, and therefore should be variable. This is known as *adaptive kernel estimation*. If the fluctuation of data in a region x_n with N entries is assumed to follow a poissonian, the relative uncertainty scales as

$$\frac{1}{\sqrt{N}} \propto \frac{1}{\sqrt{f(x_n)}} \tag{7.1.10}$$

where $f(x)$ is the sample distribution. Hence it is reasonable to scale the smoothing parameter by

$$w(x_n) = \frac{\tilde{w}}{\sqrt{f(x_n)}} \ . \tag{7.1.11}$$

Now $\tilde{w}$ and the sampling distribution f are unknown. For f one can substitute the fixed kernel estimator $\hat{f}_0(x)$. Due to Eq. 7.1.8 the dimension of $w(x_n)$ should be the same as the dimension of x and the sample distribution $f(x)$ has to have the inverse dimension of x. Thus from Eq. 7.1.11 it follows that $\tilde{w}$ has the same dimension as $\sqrt{x}$ or $\sqrt{w(x_n)}$.

For a gaussian kernel function

$$K(x - x_n, w_n) = \frac{1}{\sqrt{2\pi w_n^2}} \exp\left[-\frac{(x - x_n)^2}{2w_n^2}\right] \tag{7.1.12}$$

it can be shown that the *mean integrated squared error* is minimized when $\tilde{w}$ is choosed as

$$\tilde{w}^* = \left(\frac{4}{3}\right)^{1/5} \sigma N^{-1/5} \tag{7.1.13}$$

where σ denotes the standard deviation of the data [49]. Since $\tilde{w}$ has the dimension of $\sqrt{\sigma}$ or $\sqrt{w(x)}$ as shown above, this leads to

$$w(x_n) = w_n = \rho \left(\frac{4}{3}\right)^{1/5} \sqrt{\frac{\sigma}{\hat{f}_0(x_n)}} N^{-1/5} \tag{7.1.14}$$

where a factor ρ for further tuning has been introduced which is typically set to unity. The kernel estimated density function thus reads

$$\hat{f}_1(x) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{w_n} K\left(\frac{x - x_n}{w_n}\right) \ . \tag{7.1.15}$$

An optimization of ρ might be done if local variances σ_{local} in the data seem to be of about two order of magnitude smaller than the global variance σ and in that case

$$\rho = \sqrt{\frac{\sigma_{\mathrm{local}}}{\sigma}} \ . \tag{7.1.16}$$

Roughly spoken, choosing $\rho > 1$ *promotes smoothness over detail preservation* and vice versa for $\rho < 1$.

In this thesis a gaussian adaptive kernel density estimator is used as embedded in the *RooFit* framework. A qualitative understanding of the obtained density function from a dataset using gaussian kernels can be obtained from Fig. 7.2.

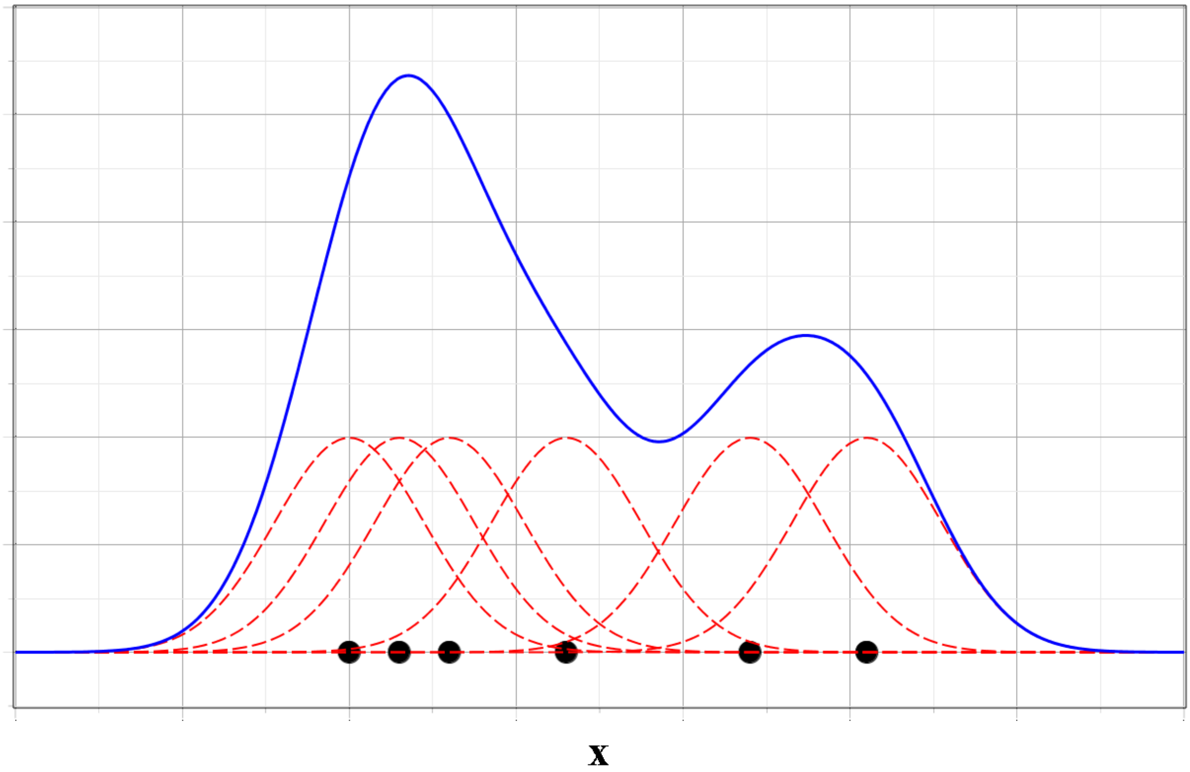


Figure 7.2: Example of a simple gaussian kernel density estimation. The black dots represent some data with respect to some variable x and are approximated with gaussian kernels, the blue line is the density estimator.

7.1.3 Signal and background shape

The templates for signal and background on the Z-boson resonance are now used for the estimation of density functions.

For the signal a histogram based pdf with a third order interpolation as described in Section 7.1.1 is built using an invariant mass binning with a bin width of 1 GeV. In addition, to consider resolution effects and differences between simulation and data the template pdf is convolved with a gaussian in which the mean and width are floating parameters.

The convolution is performed using the Fast Fourier Transformation algorithm [53]. Thus the signal pdf can be written as

$$\hat{S}(x, \mu_{\mathrm{gaus}}, \sigma_{\mathrm{gaus}}) = \hat{h}_{\mathrm{histo}}(x) * g\left(x, \mu_{\mathrm{gaus}}, \sigma_{\mathrm{gaus}}\right) \ . \tag{7.1.17}$$

Since for numerator and denominator the same signal template is used only one signal pdf is estimated. The result is shown in Fig. 7.3.

The background pdf is estimated using a gaussian adaptive kernel estimator as described in Section 7.1.2 where the ρ parameter was set to unity. Thus the background pdf only depends on the invariant mass

$$\hat{B}(x) = \hat{f}_{\mathrm{kde}}(x) \ . \tag{7.1.18}$$

The kernel estimation is done with the unbinned templates. The constructed pdf shapes for the background are shown in Fig. 7.4.

To obtain the total model, signal and background are summed up and weighted with a total event factor each:

$$\hat{P}(x, N_{\mathrm{sig}}, N_{\mathrm{bkg}}, \mu_{\mathrm{gaus}}, \sigma_{\mathrm{gaus}}) = N_{\mathrm{sig}} \cdot \hat{S}(x, \mu_{\mathrm{gaus}}, \sigma_{\mathrm{gaus}}) + N_{\mathrm{bkg}} \cdot \hat{B}(x) \tag{7.1.19}$$

The numbers of signal and background events are constrained to add up to the total number of events $N_{\mathrm{sig}} + N_{\mathrm{bkg}} = N_{\mathrm{tot}}$ in the considered invariant mass range. In Fig. 7.5 the complete model for numerator and denominator is shown with initial chosen values for the parameters. The optimization of the parameters will be discussed in the next section.

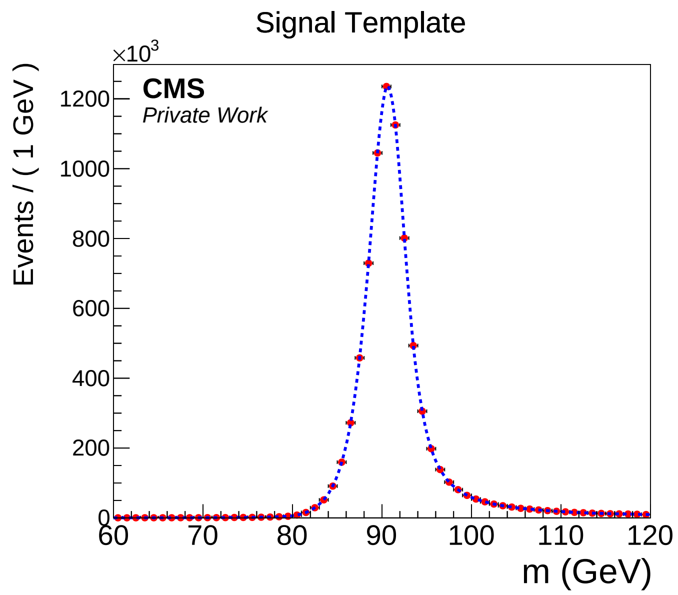


Figure 7.3: Selected signal template where the red dashed line is the corresponding histogram based density estimator. The plotted binning corresponds to the binning used in the density estimation.

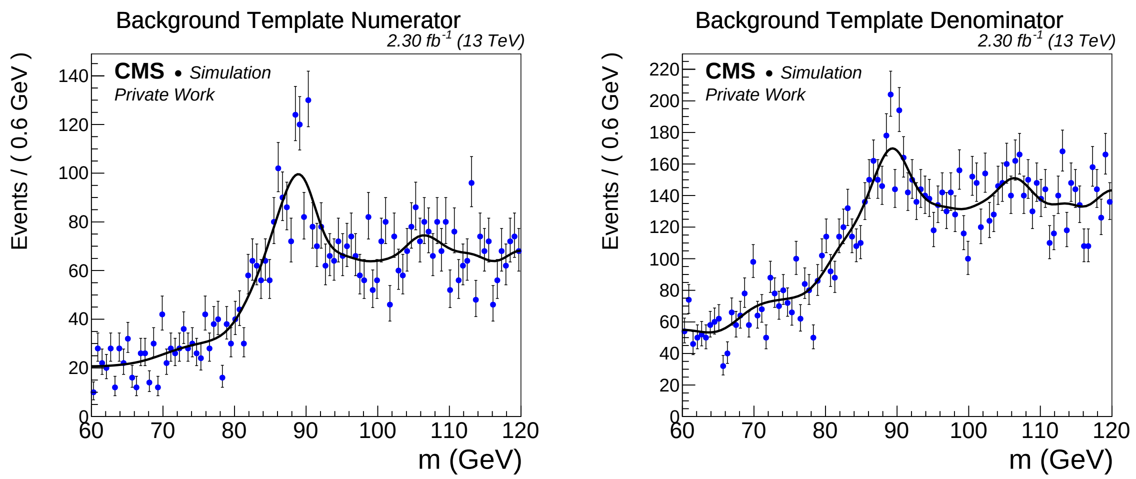


Figure 7.4: Background templates with the black lines being the estimated density functions. The chosen binning is for visualization only, the kernel estimation is done unbinned.

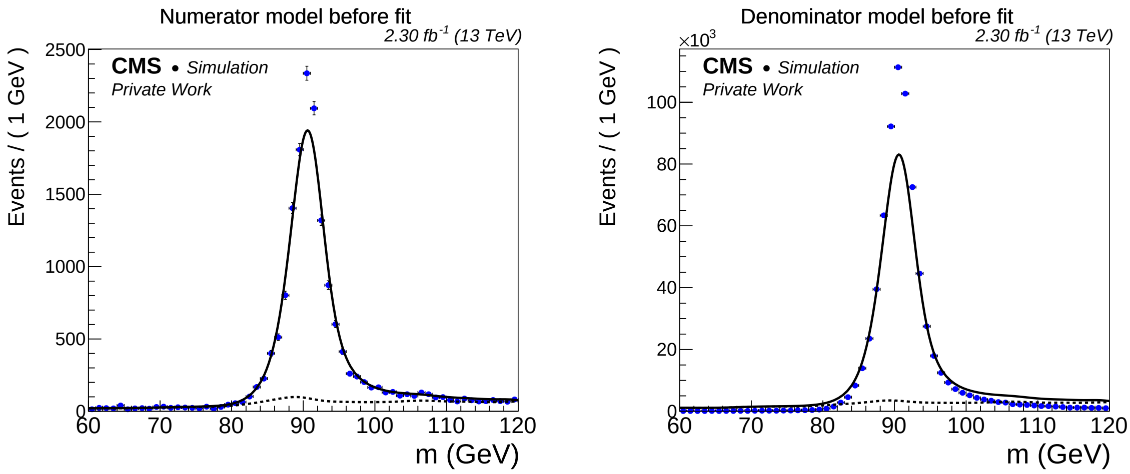


Figure 7.5: Selected numerator and denominator samples with the initial combined pdf model from eq. 7.1.19. The solid black line shows the complete model while the dashed line is the background contribution.

7.2 Extended maximum likelihood fit

In the constructed probability density function the parameters that have to be estimated are N_{sig}, N_{bkg} and the gaussian mean μ_{gaus} and width σ_{gaus}. Combining these in a parameter vetor ξ, the density can be written as $\varrho(x, \xi)$. The question then is how to formulate an estimator $\hat{\xi}$ given an indepedent sample $\{x_1, \ldots, x_N\}$.

In general, the probability that a specific choice of ξ yields a value in the interval $[x, x + dx]$ is just $\varrho(x, \xi)dx$ and the probability for an independent sample $\{x_1, \ldots, x_N\}$ is calculated by [54]

$$\mathcal{L}(x, \xi) = \prod_{i=1}^{N} \varrho(x_i, \xi) \ . \tag{7.2.20}$$

This is called the *likelihood function* of the sample. The maximum-likelihood estimator $\hat{\xi}$ of the parameters ξ are the values that maximize the likelihood [55].

It is common practice to consider the negative logarithm of the likelihood function [24] as

$$-\ln \mathcal{L}(x, \xi) = -\sum_{i=1}^{N} \ln \varrho(x_i, \xi) \tag{7.2.21}$$

and search for the minimum where it is a nescessary condition that

$$-\frac{\partial \ln \mathcal{L}(x,\hat{\xi})}{\partial \xi_j} = 0 \quad \text{for} \quad j = 1 \dots m \tag{7.2.22}$$

and m is the number of parameters combined in ξ.

The parameters for the numbers of signal and background events in the selected data samples are *absolute* rates of physics processes. When repeating an experiment under identical conditions the observed rate N of a process is excpected to fluctuate following a Poisson distribution around the excpected value ν. This poissonian term can be included in the likelihood as a multiplicative factor

$$\mathcal{L}(x,\nu,\xi) = \exp(-\nu)\frac{\nu^N}{N!}\prod_{i=1}^{N}\varrho(x_i,\xi) \ . \tag{7.2.23}$$

This is called the *extended likelihood*. If the expected number of events ν does not depend of the parameters ξ its estimated value corresponds to the observed number $\hat{\nu} = N$.

Using the density function estimator described in Section 7.1.3 an extended unbinned maximum likelihood fit is performed in the invariant mass window of 60 to 120 GeV to determine the floating parameters. The floating parameters are $\xi = (N_{\text{sig}}, N_{\text{bkg}}, \mu_{\text{gaus}}, \sigma_{\text{gaus}})$ and the constraint when performing the fit is that $N_{\text{sig}} + N_{\text{bkg}} = N$.

Figure 7.6 shows the fitted model with respect to the data. The binning shown in the plot is for visualization only since the fit is done unbinned.

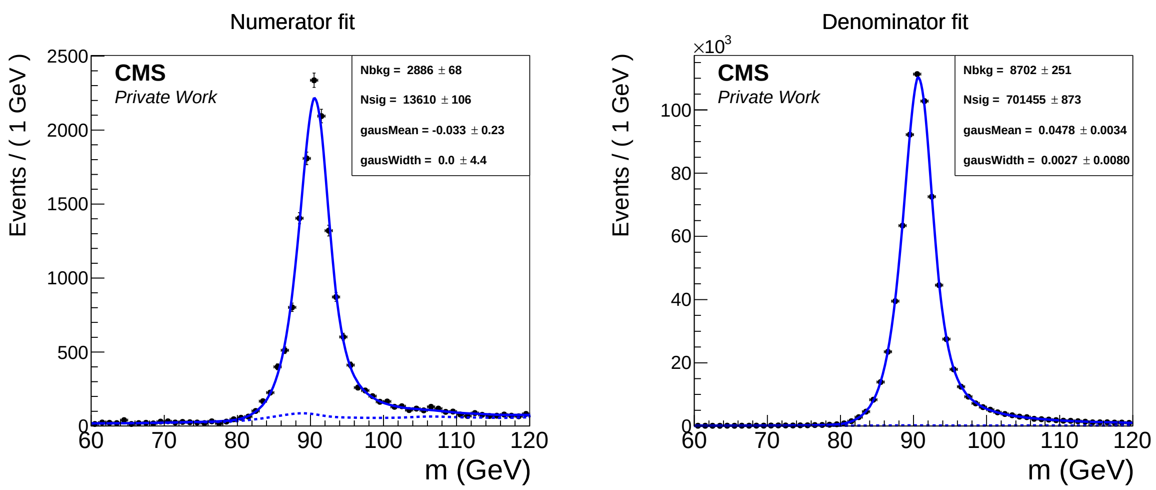


Figure 7.6: Fitted models for numerator and denominator. The solid line shows the complete model while the dashed line is the background contribution.

The resulting fit parameters are summarized in Tab. 7.1.

From the obtained signal yields the fake rate can be estimated as

$$f_{e\to\gamma} = \frac{\hat{N}_{\text{sig}}^{\text{numerator}}}{\hat{N}_{\text{sig}}^{\text{denominator}}} = 1.940 \pm {}^{0.017}_{0.016}(\text{stat})\,\% \,. \tag{7.2.24}$$

The given statistical uncertainties correspond to the 1-σ-Clopper-Pearson interval [56] which is an estimation approach for confidence intervals which can be used for the ratio of poissonian means [57]. Since for the yields the estimated uncertainties from the fit do roughly correspond to a poissonian uncertainty $\sim \sqrt{N}$, this has been assumed to be an accurate estimator.

Table 7.1: Estimated values for the parameters obtained from the fit. The shown uncertainties include only statistics.

	Numerator	Denominator
N_{sig}	$(1.361 \pm 0.011) \cdot 10^4$	$(7.0145 \pm 0.0087) \cdot 10^5$
N_{bkg}	$(2.886 \pm 0.068) \cdot 10^3$	$(8.72 \pm 0.25) \cdot 10^3$
μ_{gaus} [GeV]	$(-3.3 \pm 23) \cdot 10^{-2}$	$(4.78 \pm 0.34) \cdot 10^{-2}$
σ_{gaus} [GeV]	$(0.0 \pm 4.4) \cdot 10^{-2}$	$(2.70 \pm 0.80) \cdot 10^{-3}$

Fake rate on simulation

The same method as described above has been applied to estimate the fake rate on the stacked simulation samples. The results from the fits are shown Fig. 7.7.

The resulting global fake rate value is

$$f_{e\to\gamma}^{\text{MC}} = 0.954 \pm {}^{0.007}_{0.007}(\text{stat})\,\% \tag{7.2.25}$$

which is about half of the fake rate in data. This is due to the better efficiency of hits in the pixel tracker in the simulation that significantly influences the size of the numerator sample.

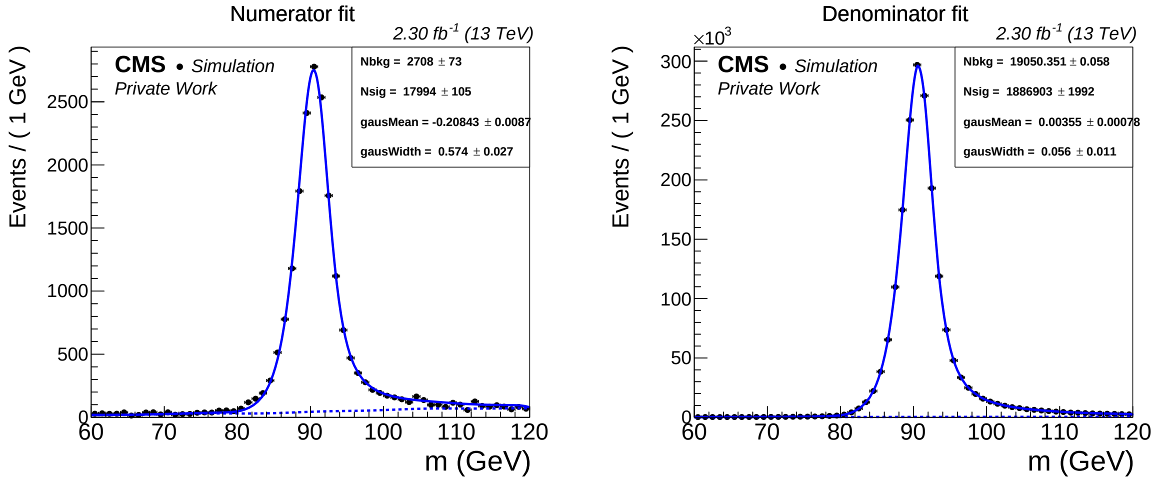


Figure 7.7: Fitted models for numerator and denominator in simulation. The solid line shows the complete model while the dashed line is the background contribution.

7.3 Parametrized Model

As an alternative approach besides the density estimation described in Section 7.1 a parametrized form has been considered.

In Section 2.7 it has been motivated that the Z-boson resonance follows a Breit-Wigner (BW) distribution which therefore should be a reasonable model for the signal peak. To consider resolution effects it is convolved with a double sided crystal ball function (DCB) which is described in Appendix A.1. The parameters for the signal that have to be estimated then are

$$\xi_{\text{sig}} = (\mu^{\text{BW}}, \sigma^{\text{BW}}, \mu^{\text{DCB}}, \sigma^{\text{DCB}}, \alpha_1^{\text{DCB}}, \alpha_2^{\text{DCB}}, n_1^{\text{DCB}}, n_2^{\text{DCB}}) \tag{7.3.26}$$

where the upper indices refer to the variables of the Breit Wigner and the crystal ball function respectively. In this notation the signal function reads

$$\hat{S}_{\text{parametrized}}(x, \xi_{\text{sig}}) = BW(x, \xi_{\text{sig}}^{\text{BW}}) * DCB(x, \xi_{\text{sig}}^{\text{DCB}}) \; . \tag{7.3.27}$$

The background shape $\hat{B}_{\text{parametrized}}$ is the product of an exponential and an error function, also called the *CMS shape* described in Appendix A.2. The open parameters are

$$\xi_{\text{bkg}} = (\alpha^{\text{cms}}, \beta^{\text{cms}}, \gamma^{\text{cms}}, \mu^{\text{cms}}) \; . \tag{7.3.28}$$

As before the complete model is the sum of both signal and background:

$$\hat{P}_{\text{parametrized}}\left(x, \xi_{\text{sig}}, \xi_{\text{bkg}}\right) = N_{\text{sig}}\hat{S}_{\text{parametrized}}\left(x, \xi_{\text{sig}}\right) + N_{\text{bkg}}\hat{B}_{\text{parametrized}}\left(x, \xi_{\text{bkg}}\right) \quad (7.3.29)$$

The parametrized model therefore has 14 open parameters with $N_{\text{sig}} + N_{\text{bkg}} = N_{\text{tot}}$. To decrease the dimension of the open parameter space as a reasonable assumption the mean and width of the Breit Wigner signal shape are fixed to the current experimental values of the mass und decay width of the Z boson [58] with

$$\mu^{\text{BW}} = 91.1876 \text{ GeV},$$
$$\sigma^{\text{BW}} = 2.495 \text{ GeV}.$$

The results after the fit are summarized in Tab. 7.2 and the shapes are shown in Fig. 7.8. In the fit shapes it can be seen that the parametrized model is not able to reproduce the denominators shape for invariant masses below the Z-boson resonance. Also the background contribution does not look like the $t\bar{t} + W \rightarrow l\nu$ expectation from the simulation and is significantly larger in the numerator in comparison to the template method.

The obtained fake rate from the parametrization is

$$f_{e\rightarrow\gamma}^{\text{parametrized}} = 1.835 \pm {}^{0.016}_{0.016}(\text{stat})\,\% \,. \quad (7.3.30)$$

Again instead of the uncertainties from the fit poissonian unertanties for the yields are used to estimate the 1-σ-Clopper-Pearson interval.

The parametrized result for the fake rate differs from the template method result by $\Delta \sim 6\sigma$ which corresponds to a relative deviation of $\sim 5\%$. In the shown fit results the model is not able to reproduce the data distributions in a reasonable way. The main problem of this model is the large number of floating parameters which have to be set to reasonable initial values to achieve stable fit results. One possibility might be to fit the model in several steps and try to fix the signal parameters first using a signal only shape and subsequently adding the background model which could also be pre-fitted to some background shape model. Since this thesis concentrates on the density estimation based on templates, the parametrized model has not been studied in further detail.

Table 7.2: Estimated values for the parameters obtained from the parametrized fit. The shown uncertainties include only statistics.

	Numerator	Denominator
N_{sig}	$(1.2695 \pm 0.0004) \cdot 10^4$	$(6.9171 \pm 0.0084) \cdot 10^5$
N_{bkg}	$(6.4819 \pm 0.0029) \cdot 10^3$	$(1.8125 \pm 0.0028) \cdot 10^3$
μ^{DCB} [GeV]	$(-6.08 \pm 0.27) \cdot 10^{-1}$	$(-5.418 \pm 0.033) \cdot 10^{-1}$
σ^{DCB} [GeV]	1.047 ± 0.035	1.3767 ± 0.0043
α_1^{DCB}	3.223 ± 0.028	5.0 ± 3.7
α_2^{DCB}	4.9 ± 3.3	1.8639 ± 0.0069
n_1^{DCB}	$(1.00 \pm 0.15) \cdot 10^{-1}$	9 ± 20
n_2^{DCB}	5.0 ± 3.7	1.931 ± 0.019
α^{cms}	$(8.630 \pm 0.013) \cdot 10^2$	$(9.803 \pm 0.019) \cdot 10^2$
β^{cms}	$(2.500 \pm 0.012) \cdot 10^{-1}$	$(2.000 \pm 0.057) \cdot 10^{-1}$
γ^{cms} [1/GeV]	$(6.82 \pm 0.19) \cdot 10^{-2}$	$(4.24 \pm 0.20) \cdot 10^{-2}$
μ^{cms} [GeV]	9.9 ± 6.0	3.6 ± 5.6

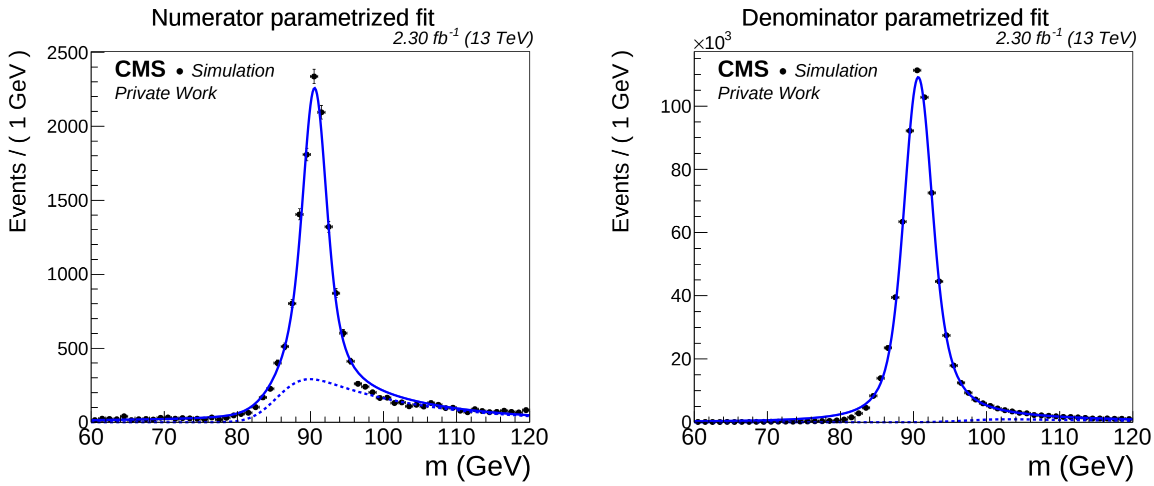


Figure 7.8: Fitted parametrized models for numerator and denominator. The solid line shows the complete model while the dashed line is the background contribution.

8 Systematic uncertainties and results

Up to now only statistical uncertainties on the estimated fake rate values have been stated. In this chapter the systematic uncertainties are discussed and the final results are summarized.

8.1 Systematic uncertainties

Two main systematical effects are considered in the estimation of the fake rate. These are the mismodelling of the utilized templates and the dependency of the fake rate on the event signature and kinematic properties of the considered electromagnetic objects. In this section the contribution and estimation of both effects will be briefly described.

8.1.1 Mismodelling of the templates

The influence of the mismodelling of the templates, meaning that the chosen template histogram and thus the estimated density distribution for signal and background does not represent the respective fraction of the selected event distribution in a proper way, has been studied using different approaches.

For the signal template it is assumed that the shape of the selected tag and probe pairs from the simulated $Z \rightarrow ee$ sample should be mismodelled at the utmost in the order of the simulation differences from data and therefore be represented by the influence of the convolved gaussian.

The estimation then consists of doing the fit in the data without the convolution where the floating parameters are reduced to the total event count of signal and background contribution. Doing this leads to a relative deviation of the resulting fake rate of below 0.1 % which is expected due to the small contribution of the gaussian as can be seen in Table 7.1. Since this corresponds to about 4% of the statistical uncertainty, this effect is neglected.

The mismodelling of the background template has been estimated using a modified backgroud template. For this the stacked simulation samples have been used and the *true* backgrund on the Z-boson resonance was evaluated by only using tag and probe pairs for which both objects do *not* correspond to a generated electron originating from a generated intermediate Z boson. The anti-matching criteria for this selection have been chosen to be the inverted criteria as for the generator matching in Section 6.4.1:

- $\Delta R(emObj, genEl) > 0.15$ and
- $p_T^{\mathrm{emObj}}/p_T^{\mathrm{genEl}} \notin (0.8, 1.2)$.

With the modified template the maximum likelihood fit has been performed on simulation and the estimated fake rate differs by 3% with respect to the original template fit. This is assumed to be the systematic uncertainty due to the background mismodelling for the fake rate in data.

8.1.2 Dependencies on event observables

By far the largest systematical influence comes from the dependency of the fake rate with respect to several variables. To estimate this quantitatively, the preparation of the numerator and denominator samples as described in Section 6.3 is performed in bins with respect to the transversal momentum p_T^{probe} and the absolute pseudorapidity $\left|\eta^{\mathrm{probe}}\right|$ of the probe object as well as the number of reconstructed vertices N_{vertex} and the number of reconstructed charged hadron tracks N_{ch} originating from the primary vertex per consdidered event. The binning for each variable is chosen in a way that the corresponding fits are viable with respect to the available statistics. The binnings are stated in Table 8.1.

In each bin the fit procedure from Section 7 is applied, the resulting fits are given in Appendix B. From the fits in each bin the corresponding fake rate is estimated and the resulting distributions are shown in Fig. 8.1.

Similar to the fake rate studies performed on $\sqrt{s} = 8$ TeV [43] a deacrease of $f_{e\to\gamma}$ with respect to the transversal momentum and the absolute value of the pseudorapidty can be seen. The large increase in the first $|\eta|$ bin comes due to the inefficiency of the inner tracker as can be seen in Fig. 3.5 for $\eta \lesssim 0.1$.

An appropriate approach would be to use the depedencies of the fake rate to estimate an appropriate parametrized function for $f_{e\to\gamma}$ with respect to the observables which could then be used for the application in regions beyond the considered ranges. The available data taken in 2015 which is used in this thesis is too few to consider multi-dimensional binnings as can be seen in the shown plots. The statistical uncertainties increase with the increasing observables and the dependencies of the observables of interest like the transversal momentum can merely be fitted with parametric functions although trends can be seen.

Therefore the systematic influence of the dependencies is rather estimated in a simplified way to achieve an appropriate order of magnitude. This is done by calculating the unweighted mean value of the fake rate in each distribution. The uncertainty of the mean value is then increased in a way that it covers the variation of the fake rate bins within the distribution range for most values. This leads to an uncertainty of 30 % which is assumed to be valid as an indicator for the systematical influence of these dependencies on the global fake rate value given in Eq. 7.2.24.

Both the unweighted mean value and the 30 % uncertainty band for each distribution are shown Fig. 8.1 together with the fitted fake rate values for each bin.

The total systematic uncertainty which is assumed in this thesis is therefore

$$\Delta(\mathrm{sys}) \simeq \sqrt{30^2 + 3^2}\,\% \simeq 30\,\% \,. \tag{8.1.1}$$

Table 8.1: Chosen binning for the considered observables which are explained in the text. The given numbers define the lower bounds of each bin while the last number is the upper bound of the last bin, respectively.

Variable	Binning
p_T^{probe} [GeV]	$[40, 45, 65, 90, 120, 160]$
$\left\|\eta^{\mathrm{probe}}\right\|$	$[0.0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4]$
N_{vertex}	$[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]$
N_{ch}	$[0, 20, 40, 60, 80, 100, 120, 140, 160]$

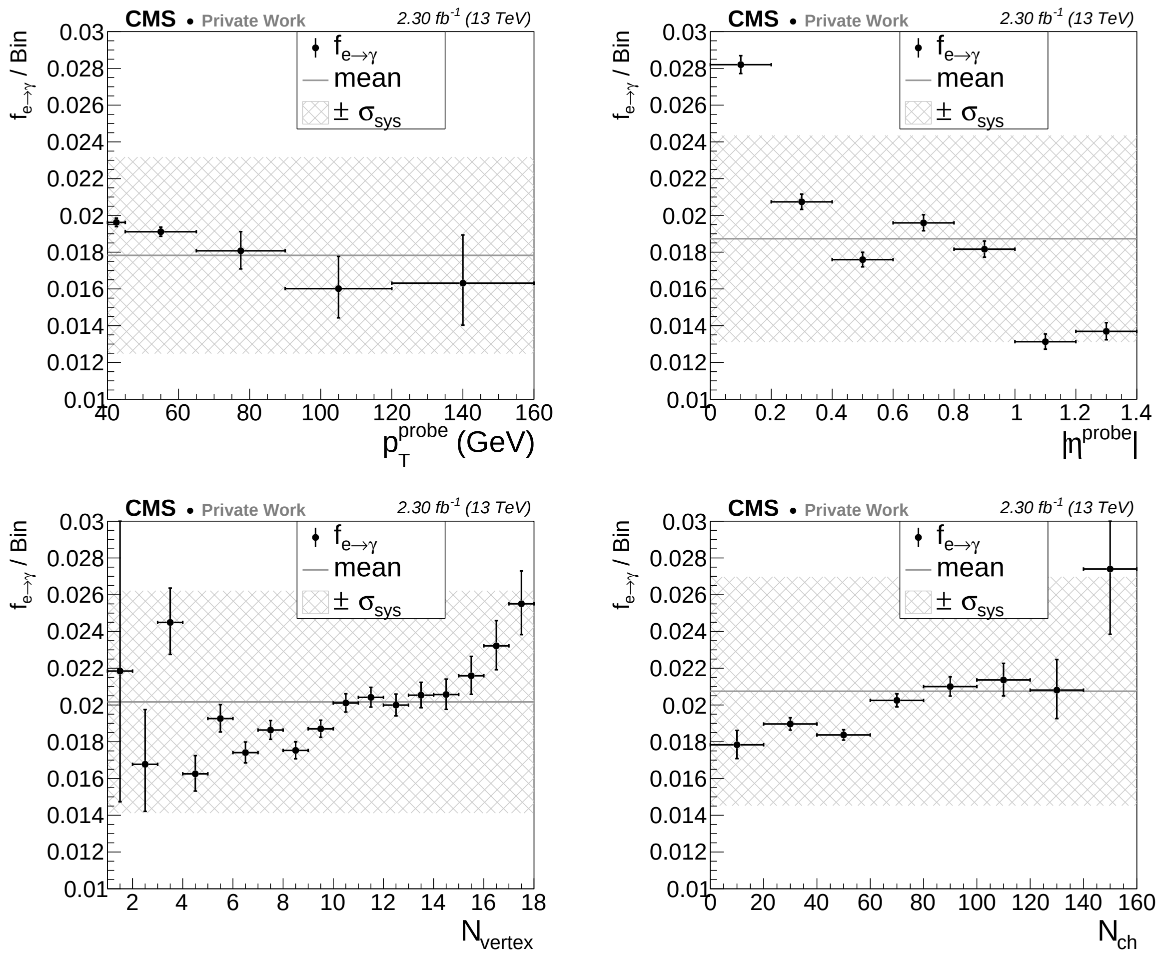


Figure 8.1: Dependency of the fake rate with respect to the probes transversal momentum, the probes absolute value of its pseudorapidity, the number of reconstructed vertices N_{vertex} per event and the number of charged hadrons N_{ch} originating from the primary vertex per event. The systematic uncertainty of the mean value is explained in the text.

8.2 Closure on simulation

With the estimated fake rate on simulation given in Eq. 7.2.25 a cross check of the applied method is performed. For this in each used simulated process a simplified selection is applied.

As done throughout this thesis, electromagnetic objects within the kinematic cuts $p_T >$ 40 GeV and $|\eta| < 1.4442$ are selected.

The $N_{e\to e}$ sample, corresponding to the fake rate notation from Eq. 4.3.12, is then prepared by counting in each event the number of thus selected eletromagnetic objects that possess a pixel seed. The $N_{e\to\gamma}$ sample are therefore all electromagnetic objects without a pixel seed and in addition match to a generated electron within the common matching criteria for generated objects.

The latter sample represents the directly simulated electrons misreconstructed as photons while $N_{e\to e}$ has to be weighted with the transfer factor $f_{e\to\gamma}/(1 - f_{e\to\gamma})$ and should then match to $N_{e\to\gamma}$ if the estimation of $f_{e\to\gamma}$ is valid.

The thus prepared samples are binned with respect to the objects transversal momentum and are shown in Fig. 8.2. As it can be seen, the directly simulated event counts are reproduced from the prediction within the uncertainties. Due to the expected dependency of the fake rate which decreases towards higher transversal momenta, the shown deviations should be corrected if an appropriate parametrization is applied.

The closure plots validate the applied data driven method based on templates to estimate the fake rate.

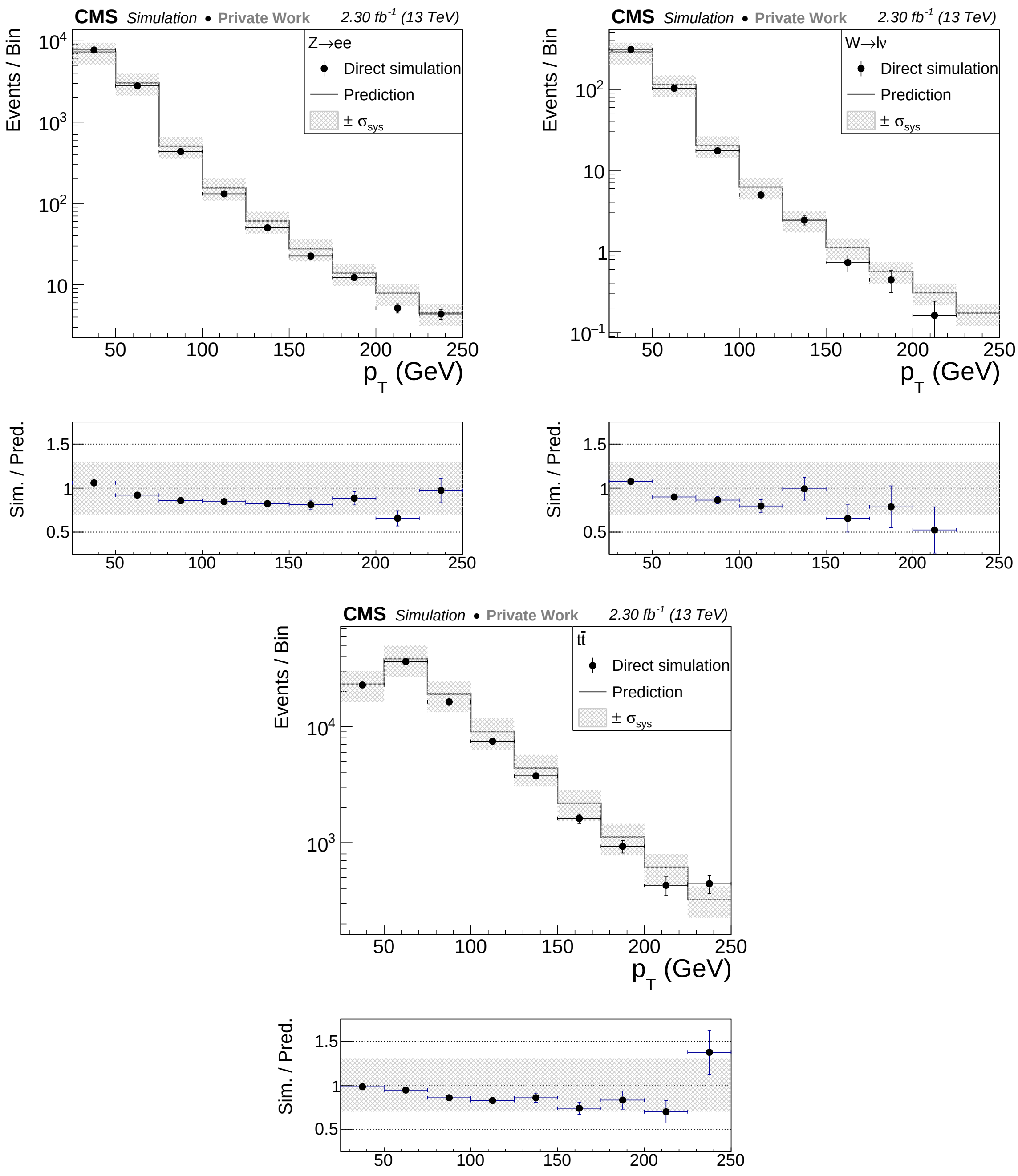


Figure 8.2: Closure plot of the fake rate with respect to the probes transversal momentum for each used simulated sample. The systematic uncertainty is assumed to be 30 % as explained in Section 8.1.2.

8.3 Summary and outlook

Throughout this thesis, a data driven method using tag and probe and template based fits on the Z-boson resonance, for the estimation of the misidentification probability of electrons as photons has been presented. The fake rate has been estimated in data recorded in 2015 by the CMS experiment. From the estimation in simulated processes the method has been checked to be valid and able to achieve a good estimator for the sought quantity. The presented results have been used in Ref. [59].

The obtained results for the fake rate value in data and simulation with the statistical and systematical uncertainties are:

$$f_{e\to\gamma} = 1.940 \pm {}^{0.017}_{0.016}(\mathrm{stat}) \pm 0.582(\mathrm{sys})\,\% \tag{8.3.2}$$

$$f^{\mathrm{MC}}_{e\to\gamma} = 0.954 \pm {}^{0.007}_{0.007}(\mathrm{stat}) \pm 0.286(\mathrm{sys})\,\% \tag{8.3.3}$$

A strong dependency of the fake rate on several kinematic observables has been obtained. Studies of these dependencies in greater depth are nescessary to estimate the fake rate in regions beyond the studied kinematic regions and the used data from 2015 does not provide enough statistical power for such an extrapolation. The data of 2016 with a much higher integrated luminosity should allow to continue with more detailed studies on this.

Appendices

A Parametrized functions

A.1 Double sided crystal ball function

Gaussian curve with polynomials on each side. Besides normalization:

$$f(x, \mu, \sigma, \alpha_1, \alpha_2, n_1, n_2) = \begin{cases} \exp\left[-\frac{\alpha_1^2}{2}\right]\left[1 - \frac{\alpha_1}{n_1}\left(\alpha_1 + \frac{x-\mu}{\sigma}\right)\right]^{-n_1} & \text{if} \quad x < \mu - \alpha_1\sigma \\ \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] & \text{if} \quad \mu - \alpha_1\sigma < x < \mu + \alpha_2\sigma \\ \exp\left[-\frac{\alpha_2^2}{2}\right]\left[1 - \frac{\alpha_2}{n_2}\left(\alpha_2 - \frac{x-\mu}{\sigma}\right)\right]^{-n_2} & \text{if} \quad x > \mu + \alpha_2\sigma \end{cases} \tag{1.1.1}$$

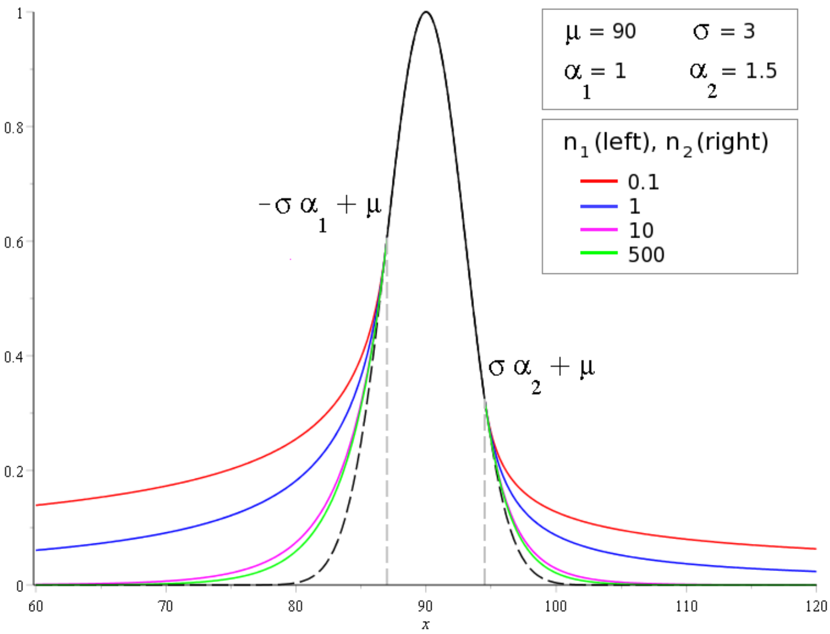


Figure A.1: Example plots of the double sided crystal ball function with different parameter sets. The black dashed line represents the ongoing gaussian.

A.2 CMS Shape

Product of falling exponential and error function. Besides normalization:

$$\mathrm{erf}(y) = \int_0^y \exp(-\tau^2)\, d\tau \tag{1.2.2}$$

$$f(x, \mu, \alpha, \beta, \gamma) = (1 + \mathrm{erf}\,[(x - \alpha)\beta]) \cdot \exp\,[-\gamma(x - \mu)] \tag{1.2.3}$$

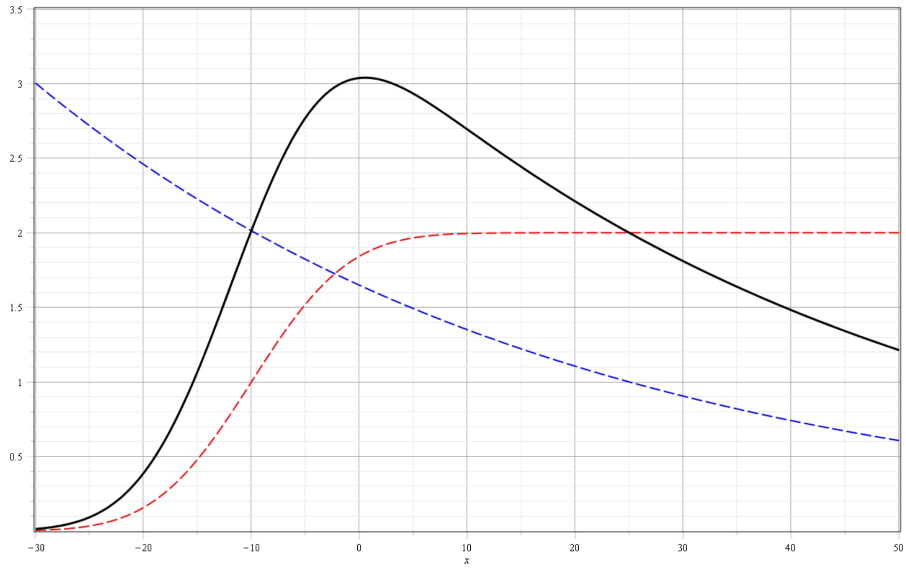


Figure A.2: Example of the *CMS shape*. The blue dashed line represents the falling exponential, the red dashed line is proportional to the error function. Plotted with $\alpha = -10$, $\beta = 0.1$, $\gamma = 0.02$, $\mu = 25$.

B Fit results for the dependecies

Here all fits in the tag and probe bins are presented. The notation in the figures is:

- *pt* means p_T in GeV
- *eta* means $|\eta|$
- *nvtx* means N_{vertex}
- *ntrk* means N_{ch}
- "$< X <$" technically means "$\leq X <$"

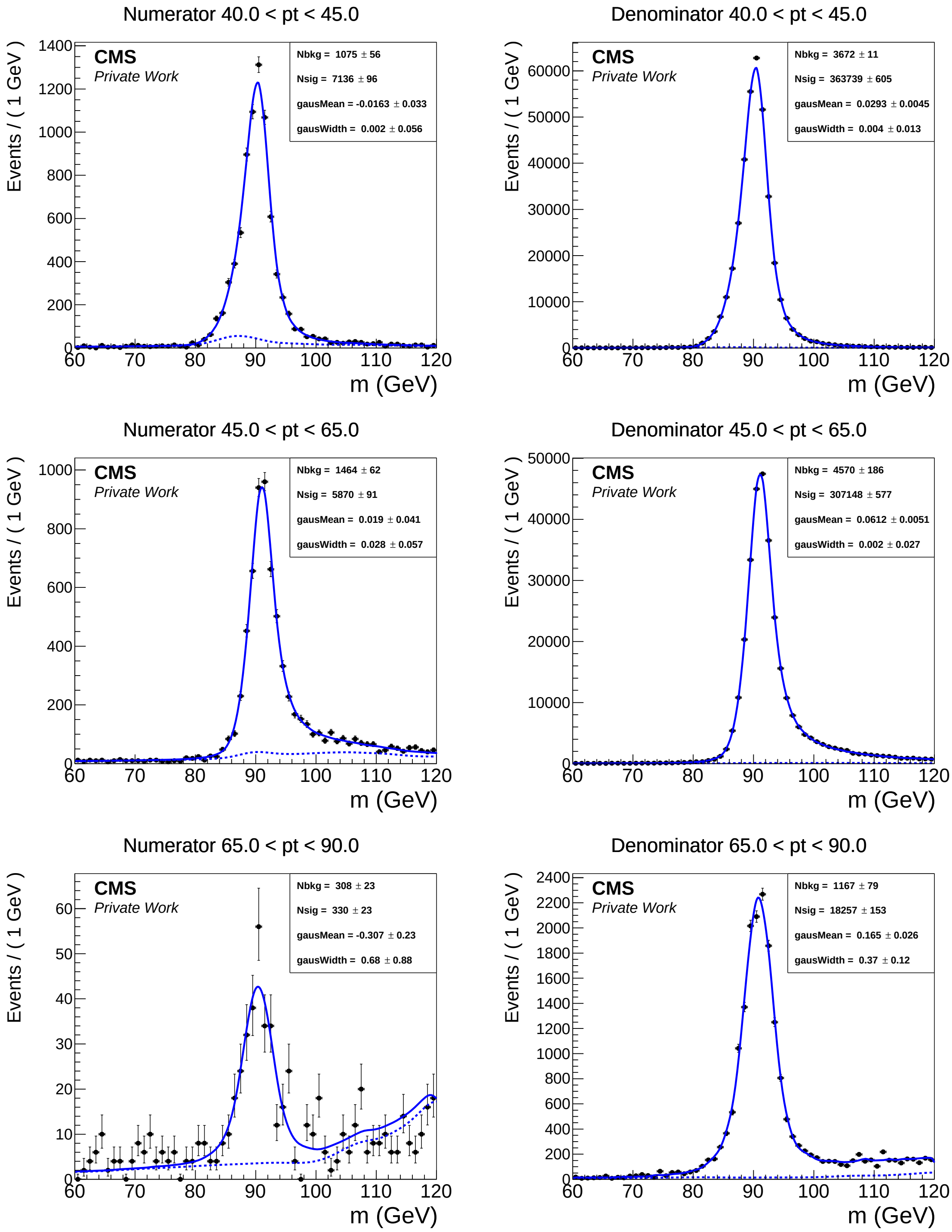


Figure B.1: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

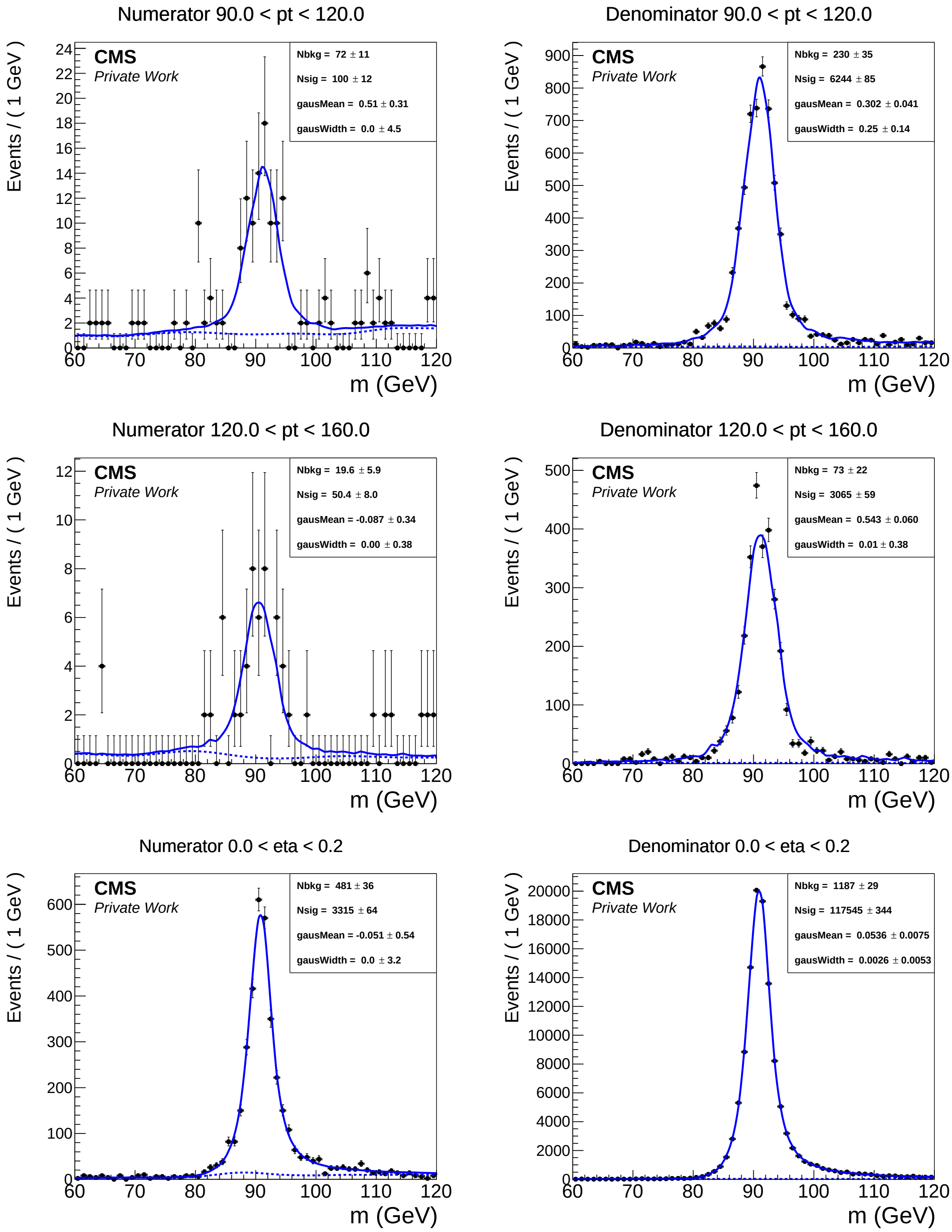


Figure B.2: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

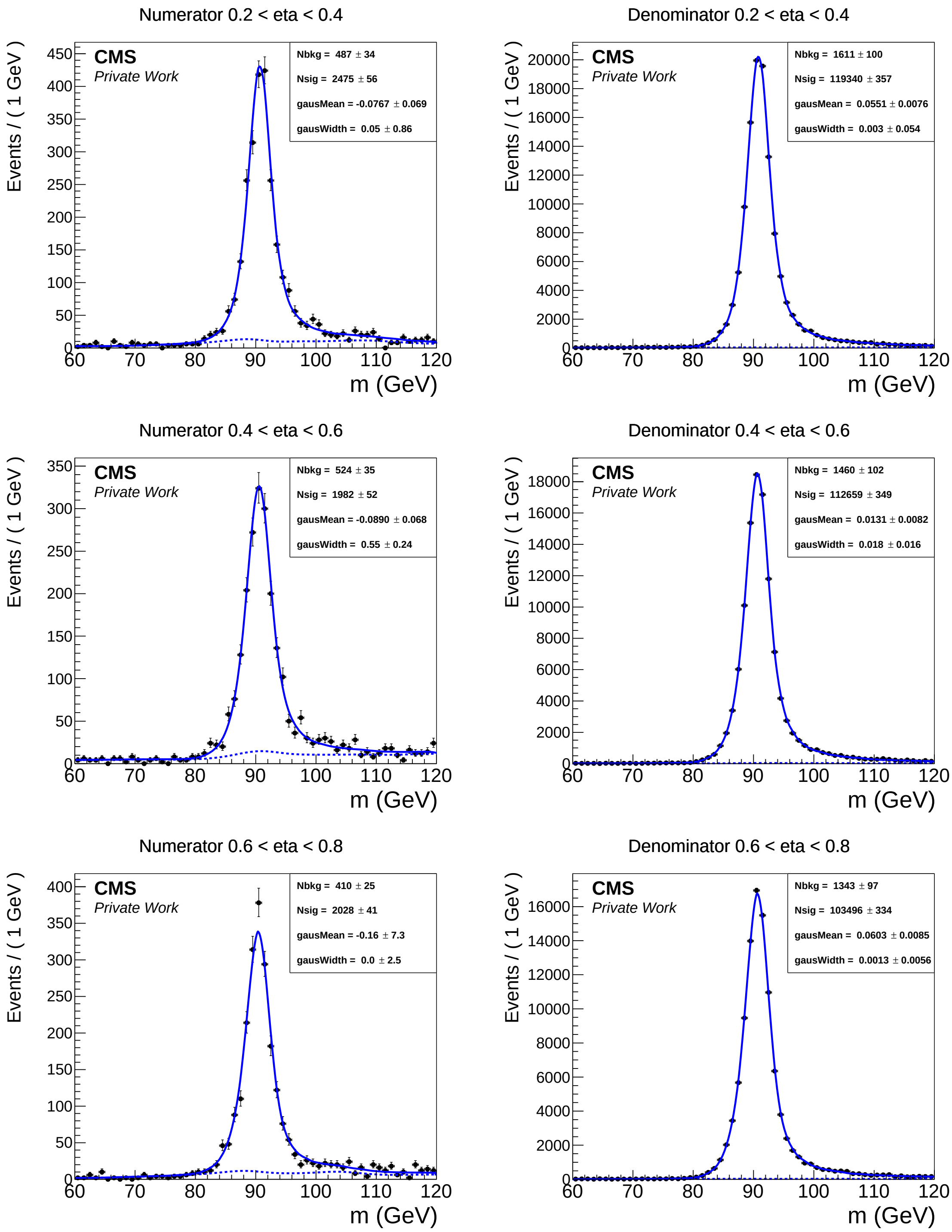


Figure B.3: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

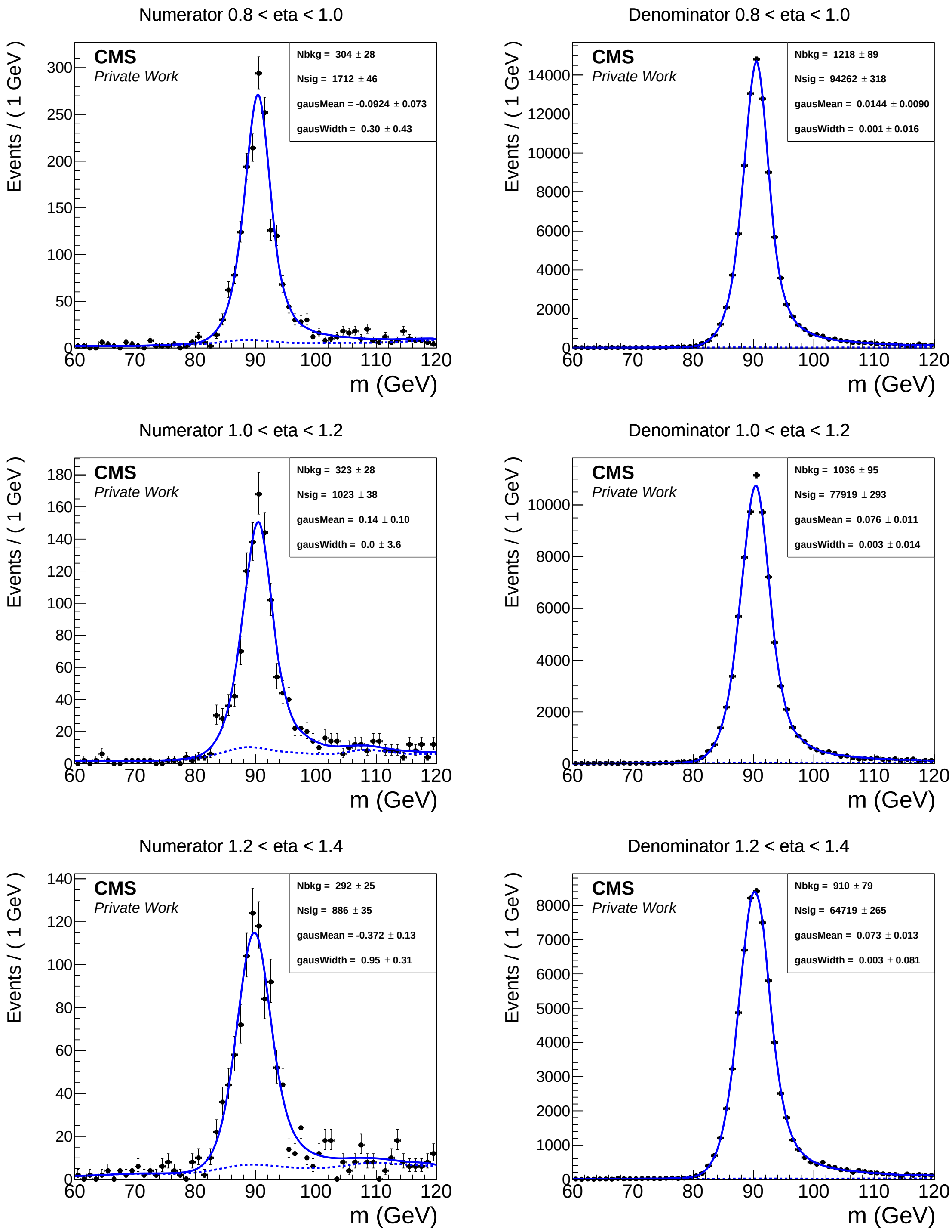


Figure B.4: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

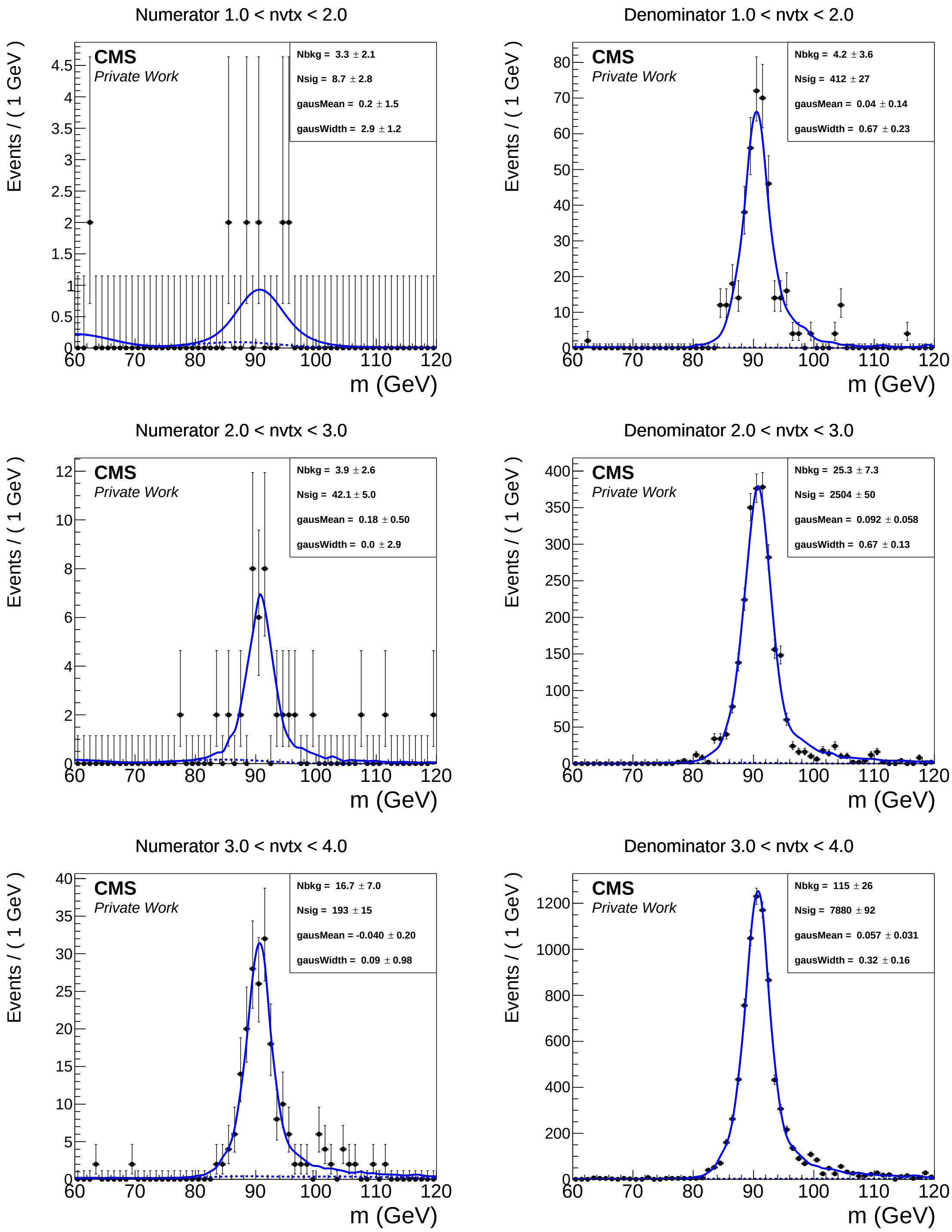


Figure B.5: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

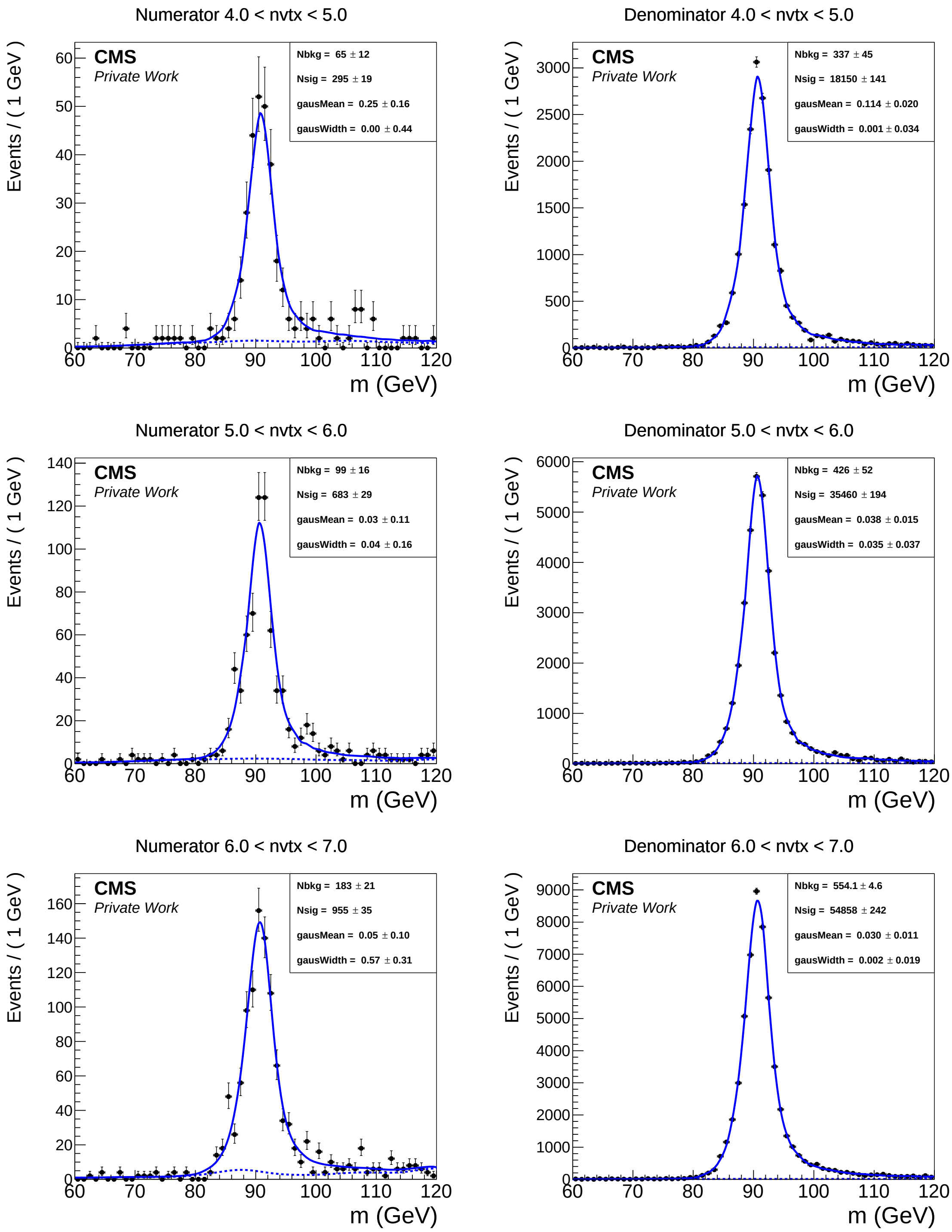


Figure B.6: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

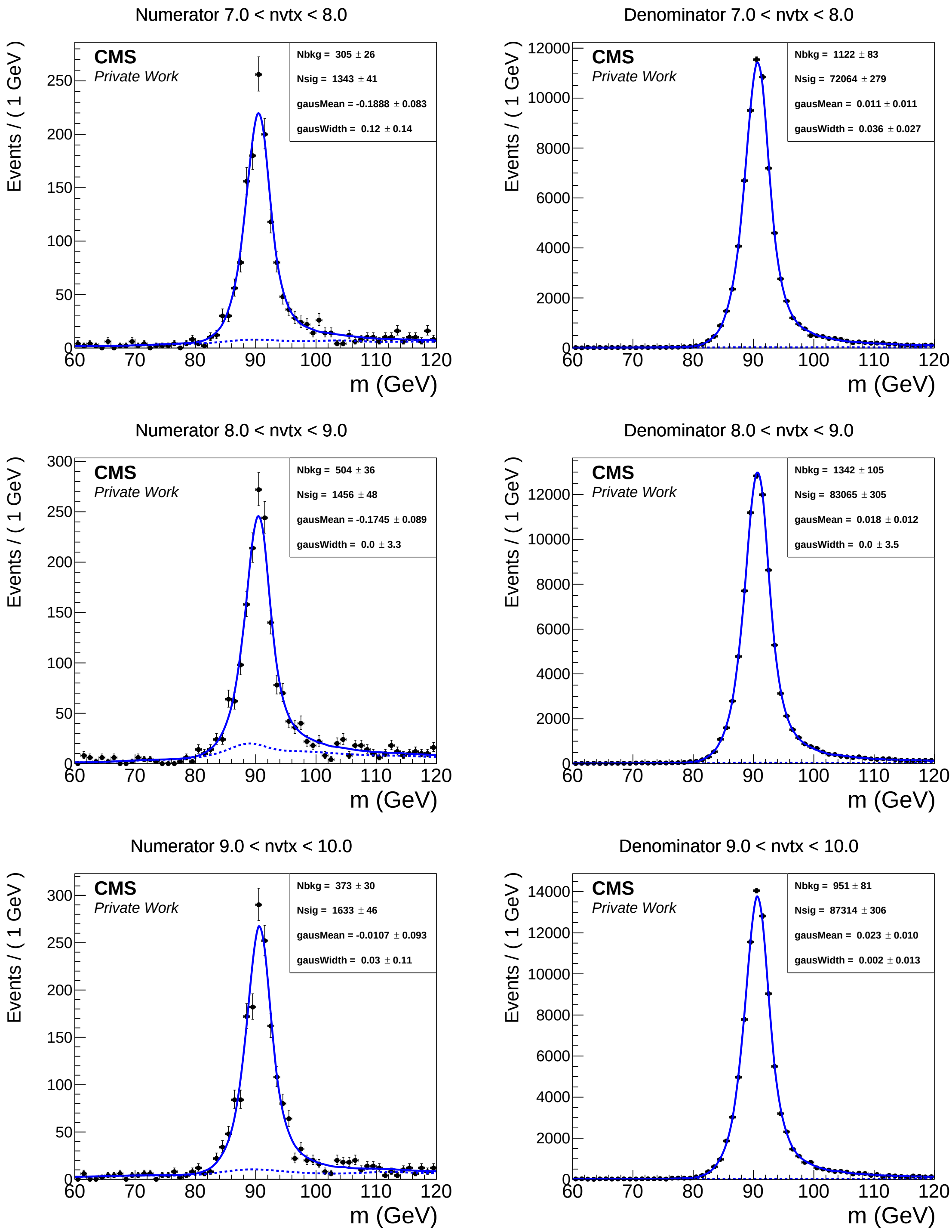


Figure B.7: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

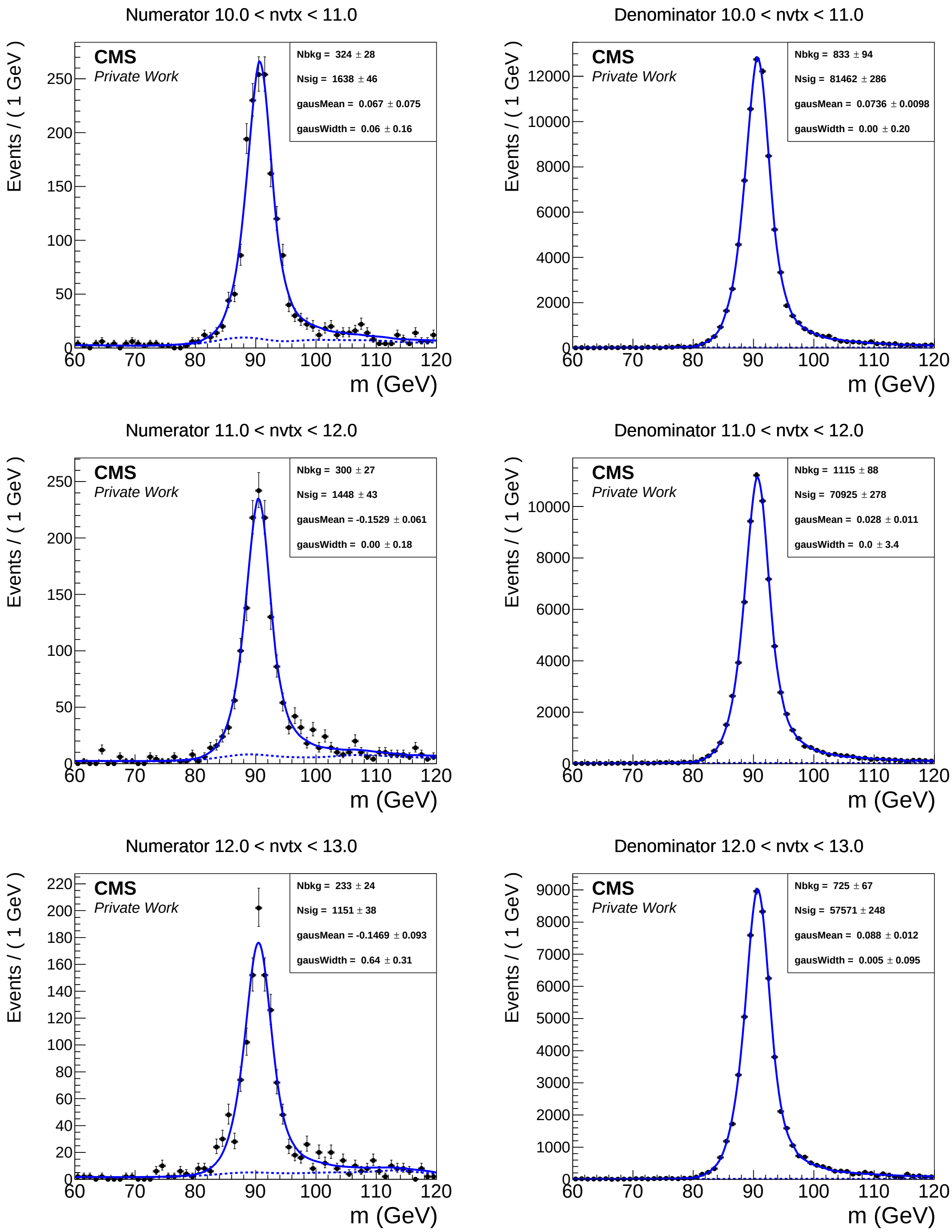


Figure B.8: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

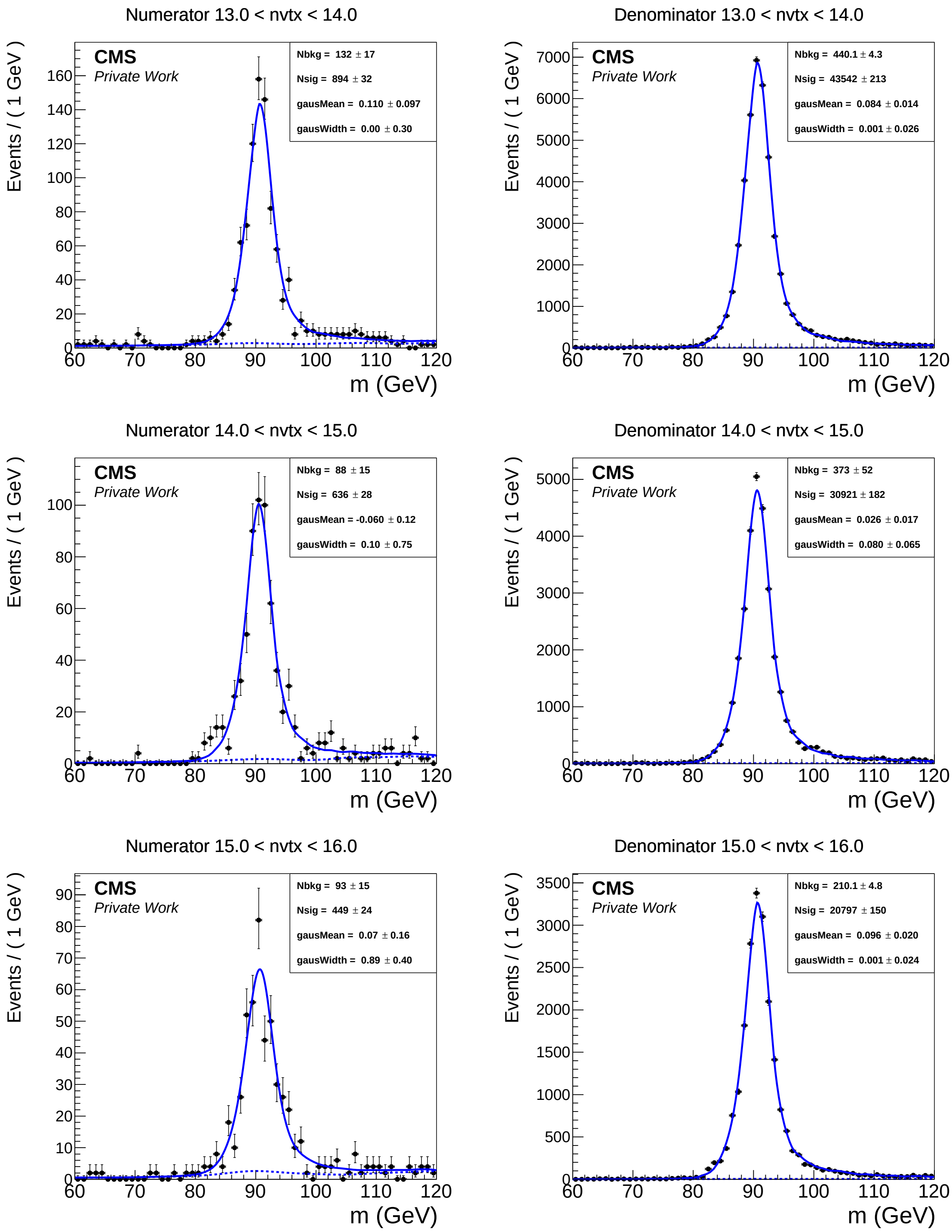


Figure B.9: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

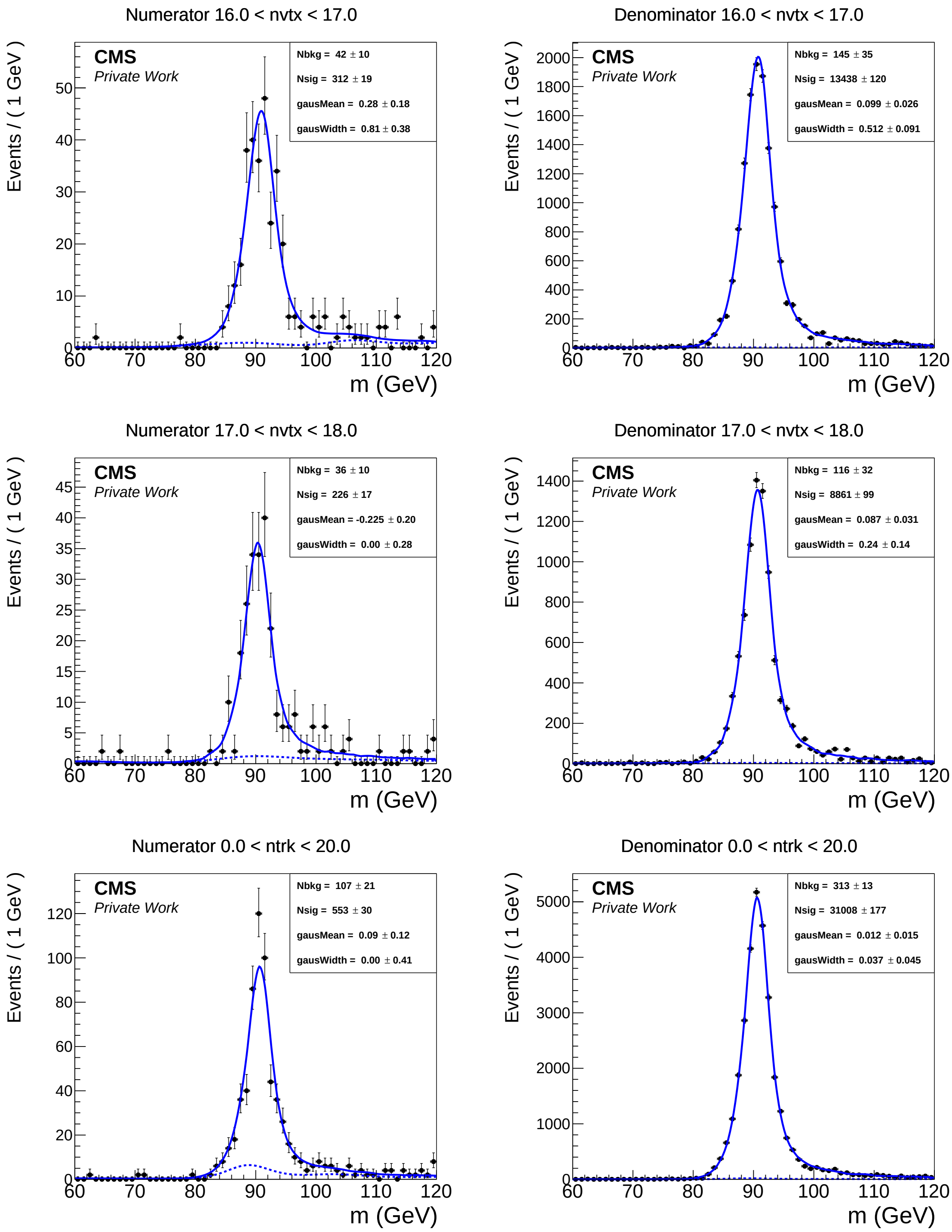


Figure B.10: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

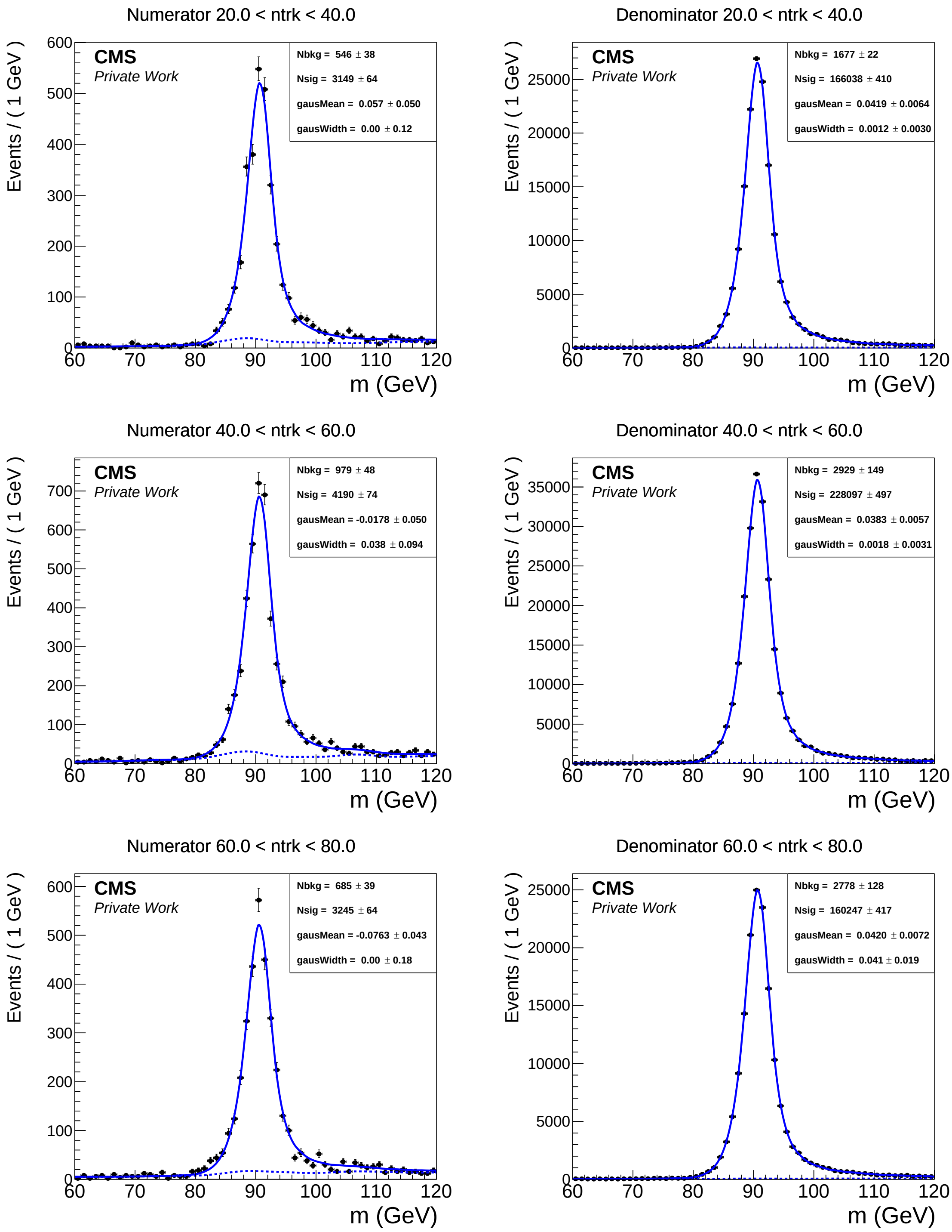


Figure B.11: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

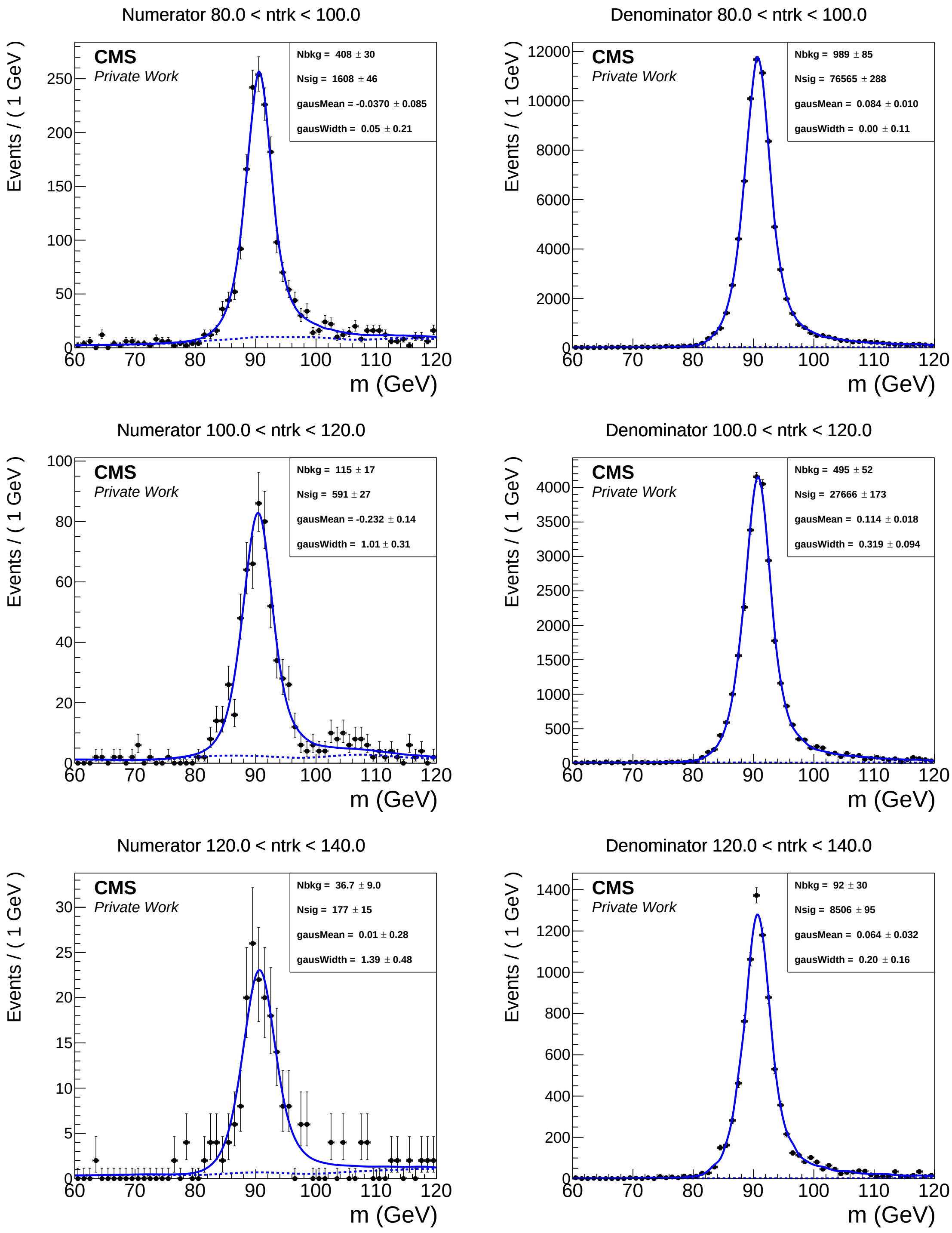


Figure B.12: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

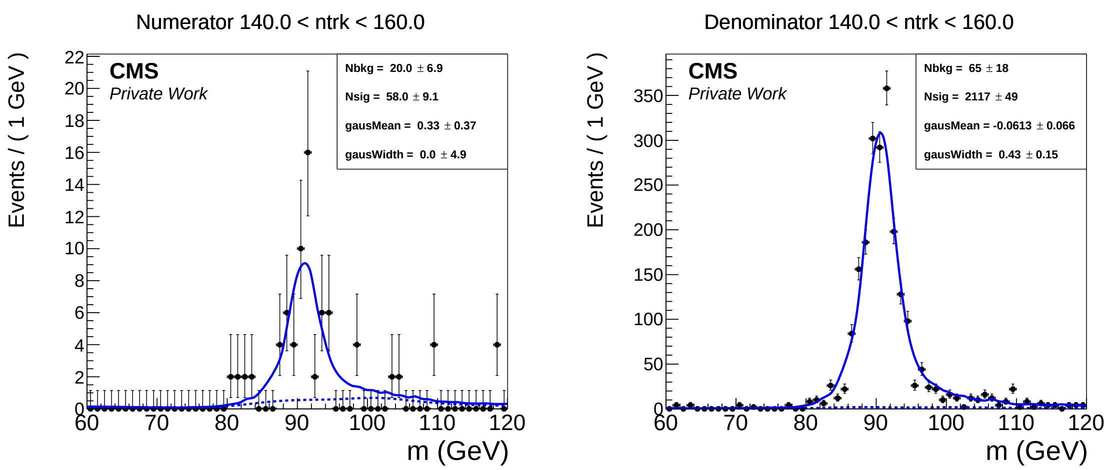


Figure B.13: Fit results for all tag and probe bins corresponding to the plots in Sec. 8.1.2

Bibliography

[1] R. Feynman, “The Feynman Lectures on Physics, Vol. 1”. Addison-Wesley Longman, Amsterdam, 1. edition, 1964.

[2] Y. Nagashima, “Elementary Particle Physics Volume 1: Quantum Field Theory and Particles”. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1. edition, 2010.

[3] Y. Nagashima, “Elementary Particle Physics Volume 2: Foundations of the Standard Model”. Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1. edition, 2013.

[4] M. Thomson, “Modern Particle Physics”. Cambridge University Press, United Kingdom, Cambridge, 1. edition, 2013.

[5] S. L. Glashow, “Partial Symmetries of Weak Interactions”, *Nucl. Phys.* **22** (1961) 579, `doi:10.1016/0029-5582(61)90469-2`.

[6] S. Weinberg, “A Model of Leptons”, *Phys. Rev. Lett.* **19** (1967) 1264, `doi:10.1103/PhysRevLett.19.1264`.

[7] A. Salam and J. Ward, “Electromagnetic and weak interactions”, *Physics Letters* **13** (1964), no. 2, 168, `doi:10.1016/0031-9163(64)90711-5`.

[8] 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, `doi:10.1016/j.physletb.2012.08.020`, `arXiv:1207.7214`.

[9] 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, `doi:10.1016/j.physletb.2012.08.021`, `arXiv:1207.7235`.

[10] C. Autermann, “Experimental status of supersymmetry after the LHC Run-I”, *Prog. Part. Nucl. Phys.* **90** (2016) 125, `doi:10.1016/j.ppnp.2016.06.001`, `arXiv:1609.01686`.

[11] LHC Study Group Collaboration, "The Large Hadron Collider: conceptual design", Technical Report CERN-AC-95-05-LHC, Oct, 1995.

[12] C. Lefevre, "LHC: The Guide". https://cds.cern.ch/record/1165534, Feb, 2009. Last accessed on 15.10.2016.

[13] CMS Collaboration, "The CMS experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008) S08004.

[14] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* **3** (2008), no. 08, S08003.

[15] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008) S08002.

[16] LHCb Collaboration, "The LHCb Detector at the LHC", *Journal of Instrumentation* **3** (2008) S08005.

[17] K. Wille, "The Physics of Particle Accelerators". Oxford University Press, Oxford, 1. edition, 2000.

[18] CMS Collaboration, "CMS Luminosity Public Results". https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults, Oct, 2015. Last accessed on 15.10.2016.

[19] CMS Collaboration, "Particle Flow Reconstruction and Global Event Description with the CMS Detector". CMS Paper PRF-14-001, 2014.

[20] CMS Collaboration, "CMS tracking performance results from early LHC operation", *The European Physical Journal C* **70** (2010), no. 4, 1165, `doi:10.1140/epjc/s10052-010-1491-3`.

[21] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV", *Journal of Instrumentation* **8** (2013) P09009.

[22] CMS collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *Journal of Instrumentation* **7** (2012) P10002.

[23] V. Gori, "The CMS High Level Trigger", *Int. J. Mod. Phys. Conf. Ser.* **31** (2014) 1460297, `doi:10.1142/S201019451460297X`, `arXiv:1403.1500`.

[24] O. Behnke, K. Kröniger, G. Schott, and T. Schörner-Sadenius, "Data Analysis in High Energy Physics". Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1. edition, 2013.

[25] J. Weng, "A Global Event Description using Particle Flow with the CMS Detector", in *Proceedings, 34th International Conference on High Energy Physics (ICHEP 2008): Philadelphia, Pennsylvania, July 30-August 5, 2008*. 2008. `arXiv:0810.3686`.

[26] CMS Collaboration, F. Beaudette, "The CMS Particle Flow Algorithm", in *Proceedings, International Conference on Calorimetry for the High Energy Frontier (CHEF 2013): Paris, France, April 22-25, 2013*, p. 295. 2013. `arXiv:1401.8155`.

[27] M. Cacciari, G. P. Salam, and G. Soyez, "The Anti-k(t) jet clustering algorithm", *JHEP* **04** (2008) 063, `doi:10.1088/1126-6708/2008/04/063`, `arXiv:0802.1189`.

[28] CMS Collaboration, "Public Twiki Workbook Data Formats". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDataFormats, Oct, 2016. Last accessed on 15.10.2016.

[29] CMS Collaboration, "Public Twiki MiniAOD Information". https://twiki.cern.ch/twiki/bin/view/CMS/MiniAOD, Oct, 2016. Last accessed on 15.10.2016.

[30] CMS Collaboration, "Public Twiki Workbook MiniAOD". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMiniAOD, Oct, 2016. Last accessed on 15.10.2016.

[31] CMS Collaboration, "Electron and Photon Physics Object Offline Guide ". https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideEgamma, 2013. Last accessed on 15.10.2016.

[32] W. Adam, R. Frühwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC", *Journal of Physics G: Nuclear and Particle Physics* **31** (2005), no. 9, N9.

[33] S. Kyriacou, "Photon Cut Based ID". https://indico.cern.ch/event/370513, 2014. Last accessed on 15.10.2016.

[34] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. Meth. A* **262** (1987) 444, `doi:10.1016/0168-9002(87)90887-4`.

[35] CMS Collaboration, "Public Workbook Muon Analysis". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMuonAnalysis, Oct, 2016. Last accessed on 15.10.2016.

[36] CMS Collaboration, "Baseline muon selections for Run-II". https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideMuonIdRun2, Oct, 2016. Last accessed on 15.10.2016.

[37] CMS Collaboration, "Data Aggregation System". https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookLocatingDataSamples, Oct, 2016. Last accessed on 15.10.2016.

[38] C. Oleari, "The POWHEG-BOX", *Nucl. Phys. Proc. Suppl.* **205-206** (2010) 36, `doi:10.1016/j.nuclphysbps.2010.08.016`, `arXiv:1007.3893`.

[39] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, `doi:10.1007/JHEP07(2014)079`, `arXiv:1405.0301`.

[40] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *JHEP* **05** (2006) 026, `doi:10.1088/1126-6708/2006/05/026`, `arXiv:hep-ph/0603175`.

[41] GEANT4 Collaboration, "GEANT4: A Simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250, `doi:10.1016/S0168-9002(03)01368-8`.

[42] CMS Collaboration, "Summary table of samples produced for the 1 Billion campaign with 25ns bunch-crossing". https://twiki.cern.ch/twiki/bin/viewauth/CMS/SummaryTable1G25ns, Oct, 2016. Last accessed on 15.10.2016.

[43] CMS Collaboration, "Search for supersymmetry with photons in *pp* collisions at $\sqrt{s} = 8$ TeV", *Phys. Rev. D* **92** (2015) 072006, `doi:10.1103/PhysRevD.92.072006`.

[44] CMS Collaboration, "Search for supersymmetry in events with a photon, a lepton, and missing transverse momentum in pp collisions at $\sqrt{s} = 8$ TeV", *Physics Letters B* **757** (2016) 6, `doi:10.1016/j.physletb.2016.03.039`.

[45] CMS Collaboration, "Search for supersymmetry in electroweak production with photons and large missing transverse energy in pp collisions at $\sqrt{s} = 8$ TeV", *Physics Letters B* **759** (2016) 479, `doi:10.1016/j.physletb.2016.05.088`.

[46] L. Antonelli et al., "Trigger Proposal from Displaced Supersymmetry group". https://indico.cern.ch/event/336052, 2014. Last accessed on 15.10.2016.

[47] CMS Collaboration, "Energy smearing and scale correction". https://twiki.cern.ch/twiki/bin/viewauth/CMS/EGMSmearer, 2016. Last accessed on 15.10.2016.

[48] I. Narsky and F. C. Porter, "Statistical Analysis Techniques in Particle Physics". Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, 1. edition, 2014.

[49] K. S. Cranmer, "Kernel estimation in high-energy physics", *Comput. Phys. Commun.* **136** (2001) 198, `doi:10.1016/S0010-4655(00)00243-5`, `arXiv:hep-ex/0011057`.

[50] CERN, "RooFit Toolkit for Data Modeling with ROOT". https://root.cern.ch/roofit, 2014-2016.

[51] CERN, "ROOT Data Analysis Framework". https://root.cern.ch, 2014-2016.

[52] E. Parzen, "On Estimation of a Probability Density Function and Mode", *Ann. Math. Statist.* **33** (1962), no. 3, 1065, `doi:10.1214/aoms/1177704472`.

[53] J. W. Cooley and J. W. Tukey, "An Algorithm for the Machine Calculation of Complex Fourier Series", *Math. Comput.* **19** (1965) 297, `doi:10.1090/S0025-5718-1965-0178586-1`.

[54] J. Honerkamp, "Statistical Physics: An Advanced Approach with Applications". Springer, Heidelberg, 3. edition, 2012.

[55] J. Aldrich, "R.A. Fisher and the making of maximum likelihood 1912-1922", *Statist. Sci.* **12** (1997), no. 3, 162, `doi:10.1214/ss/1030037906`.

[56] C. J. Clopper and E. S. Pearson, "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial", *Biometrika* **26** (1934), no. 4, 404.

[57] R. D. Cousins, K. E. Hymes, and J. Tucker, "Frequentist evaluation of intervals estimated for a binomial parameter and for the ratio of Poisson means", *Nucl. Instrum. Meth. A* **612** (2010) 388, `doi:10.1016/j.nima.2009.10.156`.

[58] Particle Data Group, C. Patrignani et al., "Review of particle physics", *Chin. Phys. C* **40** (2016), no. 10, 100001, `doi:10.1088/1674-1137/40/10/100001`.

[59] CMS Collaboration, "Search for supersymmetry in final states with at least one photon and $E_{\mathrm{T}}^{\mathrm{miss}}$ in pp collisions at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-SUS-16-023, 2016.