Study of Alternative Statistical Methods for a Model Unspecific Search in CMS

von
Mark Olschewski

Diplomarbeit in Physik

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

im Juli 2011

angefertigt im
III. Physikalischen Institut A
Prof. Dr. Thomas Hebbeker

Zusammenfassung

Seit dem Jahr 2010 kollidieren am Teilchenbeschleuniger LHC erstmals Protonen bei einer Schwerpunktsenergie von 7 TeV miteinander. Der Compact Muon Solenoid (CMS) ist eines der Experimente, die die daraus neu entstehenden Teilchen messen. Das Ziel ist es, ein genaues Bild der Physik der Elementarteilchen in diesem Regime zu erhalten. Dazu gehört, die Parameter des heutigen Standardmodells genau zu vermessen sowie sich auf die Suche nach neuen, vom Standardmodell nicht erklärten, physikalischen Gesetzmäßigkeiten zu machen.

Es existiert eine Reihe von Modellen, die Physik jenseits des Standardmodells beschreiben und die bisher nicht widerlegt sind. Viele Analysen der CMS-Daten beschäftigen sich mit einem solchen Modell: Sie überprüfen, ob das Modell ausgeschlossen werden kann oder ob es die gesammelten Daten besser erklärt als das Standardmodell.

Die modellunabhängige Suche in CMS (MUSiC) analysiert die CMS-Daten ohne das Zugrundelegen eines Modells neuer Physik, sondern vergleicht die Daten ausschließlich mit den Vorhersagen von Standardmodell-Monte-Carlo-Generatoren. Im Gegensatz zu vielen anderen Analysen wird dabei nicht nur eine kleine Anzahl von Ereignissen in einem ausgewählten Bereich des Phasenraums untersucht, sondern ein großer Teil der aufgezeichneten Daten.

Diese Arbeit erweitert die MUSiC-Analyse um zwei Komponenten:

- Komplementär zum MUSiC-Algorithmus wird ein sogenannter *"Bump Hunter"* vorgeschlagen. Es handelt sich um eine Algorithmus, der ausschließlich Daten analysiert. Er sucht nach Abweichungen im glatten Verlauf bestimmter Verteilungen und kommt somit ohne Modellvorhersage aus. Diese Analyse verwendet die CMS-Daten des Jahres 2010.

- Um die Signifikanz für eine Abweichung zwischen den Daten und dem Modell zu bestimmen, wird bei MUSiC ein *p-Wert* bestimmt. Dieser zeigt jedoch in bestimmten Bereichen Anomalien, die korrigiert werden müssen. Deshalb wird ein neuer p-Wert erarbeitet und analysiert.

Abstract

Since 2010, protons are collided with a center of mass energy of 7 TeV for the first time ever using the particle accelerator LHC. The Compact Muon Solenoid (CMS) is one of the experiments designed to measure the produced particles. Its purpose is to get a detailed picture of elementary particles in that regime. This includes the measurement of the parameters of the Standard Model and the search for new physics, not explained by the Standard Model.

There are a number of models suggesting physics beyond the Standard Model. Many analyses of CMS data test whether a specific model can either be excluded by the data or if it explains the data more convincingly than the Standard Model.

The Model Unspecific Search in CMS (MUSiC) analyses CMS data without considering such a model of new physics, but instead by comparing the data solely with predictions made by Standard Model Monte Carlo generators. In contrast to many conventional analyses MUSiC not only examines a small selected fraction of the data but the larger part of the recorded events.

This thesis extends the MUSiC analysis by two components:

- Complementary to the standard MUSiC algorithm, a different algorithm is proposed, called *"Bump Hunter"*. It is a purely data driven analysis, searching for deviations in the smooth distribution of certain observables. As a result, it doesn't rely on Standard Model Monte Carlo predictions. It is applied to 2010 CMS data.

- In order to determine the significance of the deviation between data and Monte Carlo prediction, MUSiC calculates a *p-value*. However, the current implementation shows anomalies for certain cases, which have to be corrected for. As a consequence, an alternative p-value will be derived and studied.

Contents

1. Introduction

The CMS experiment at the Large Hadron Collider is a huge experiment producing a massive amount of data. Nearly all its analyses make use of sophisticated statistical methods in order to test the data against a certain model. The Model Unspecific Search in CMS (MUSiC) evaluates a large portion of the available CMS data by comparing it to Standard Model Monte Carlo simulations.

The development of MUSiC has been going on for several years now, but it wasn't until 2010 when first data from collisions at a center of mass energy of 7 TeV were available. Therefore MUSiC could be validated with first data. It did not show major differences between experimental data and the prediction [1]. Nevertheless, such "real life" conditions always give rise to new issues and require the reevaluation of applied methods.

A major task is the handling of a low Monte Carlo statistics. This issue is not unique to the MUSiC analysis, usually this is solved by determining some Standard Model contribution from the data instead of simulating them. But as MUSiC covers a great range of different event topologies, these conventional methods are challenging.

In this thesis, the methods used by MUSiC are reevaluated under the aspect of low Monte Carlo statistics complementary approaches to a model independent search are shown.

As it compares CMS data to Standard Model Monte Carlo simulations, this thesis provides an introduction to both, the theory in chapter 2 and the experiment in chapter 3.

The software used for data analysis is presented in chapter 4. Chapter 5 outlines the reconstruction of particles from the detector signatures and the particle selection.

The standard MUSiC analysis is explained in chapter 6.

Subsequently in chapter 7, we introduce the concept of the Bump Hunter. This is a data driven analysis looking for resonances in invariant mass spectra. We develop an algorithm and apply it to CMS data.

A slight change to the statistical method used by MUSiC currently is propsed in chapter 8. MUSiC uses a p-value to determine the deviation between Monte Carlo and data. An alternative p-value is developed and its performance is compared to the one currently used.

1.1. Formal Remarks

As there are different conventions in the world of physics, some should be mentioned to avoid any misunderstandings. If not stated otherwise, we use natural units. This means that the speed of light and the Planck constant are set to 1:

$$\hbar = c = 1. \tag{1.1}$$

Which leads to one remaining unit. We use electron volts (eV) as the unit of energy.

A completely different issue is the coordinate system used inside the detector. Beside the usual x, y, z coordinates another system is frequently used. Here ϕ gives the angle measured perpendicularly to the beam axis. η is the pseudorapidity

$$\eta = -\ln \tan\left(\frac{\theta}{2}\right) \tag{1.2}$$

with θ being the angle in the x-y-plane coplanar to the beam, measured from the beam axis. Therefore η, ϕ and z are used as a basis.

2. Theoretical Background

2.1. The Standard Model of Particle Physics

The Standard Model of particle physics (SM) is a description of matter and its interactions. The theory is well tested, has not been disproved, and is widely acknowledged by the scientific community[1]. Therefore, it is the model of choice to test against CMS data in a model unspecific analysis like MUSiC. It does not explain gravity, but effects resulting from gravity can be neglected in collider experiments.

The following discussion is based on [3] if not explicitly stated otherwise.

2.1.1. Particles and Interactions

The foundation of this theory are pointlike structureless particles, which can be transformed into each other under certain circumstances. Each particle is of a certain type, defined by observables like spin, mass, and charge. (figure 2.1)

[1] Several Nobel prices have been awarded for both the theory and experimental tests, e.g. 2008 "for the discovery of the origin of the broken symmetry which predicts the existence of at least three families of quarks in nature" to Kobayashi and Maskawa [2].

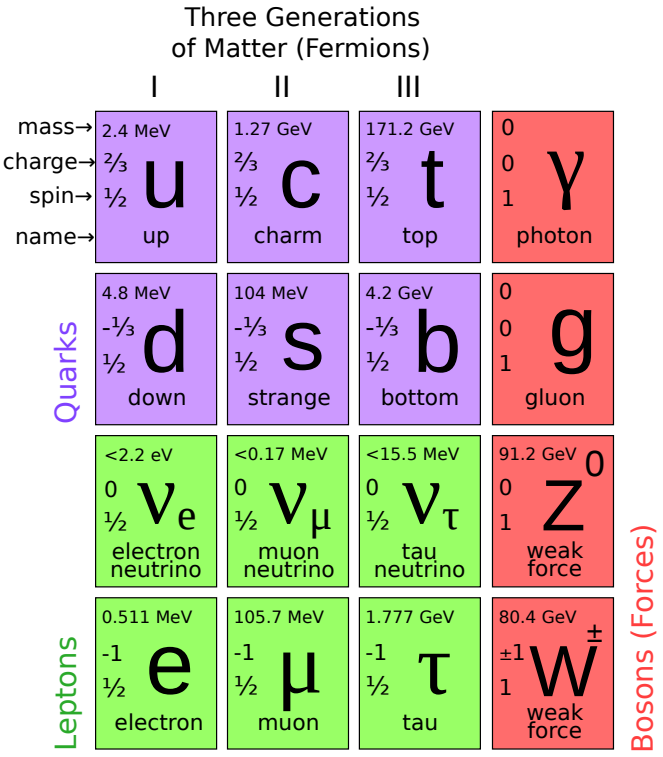


Figure 2.1.: Known elementary particles of the Standard Model [4].

Interaction	Mediator	Relative Strength	Range / m
Strong	8 gluons	1	10^{-15}
Electromagnetic	photon	10^{-2}	∞
Weak	$Z^0, W^{\pm}$	10^{-2}	10^{-16}

Table 2.1.: The interactions considered in the Standard Model [5].

What is called matter (including antimatter) is assembled from spin-1/2 particles, called fermions. They obey the Fermi-Dirac statistics and therefore follow Pauli's exclusion principle of fermions with the same quantum numbers. They can be subdivided into leptons and quarks. Six types of leptons, called flavours, have been found: the electron, muon, and tau as well as the electron neutrino, the muon neutrino, and the tau neutrino. The other part of matter is formed by quarks: the up and down, the charm and strange, as well as the top and bottom quark are the observed flavours. In contrast to leptons, quarks do not exist as free particles but form bound states called hadrons.

The interactions (sometimes called forces) are mediated via bosons (i.e. particles with an integer spin). They are subject to the Bose-Einstein statistics, i.e. they can occupy the same quantum state. There are gluons for the strong interaction, photons for the electromagnetic interaction, and Z-Bosons as well as $W^{\pm}$-Bosons for the weak interaction. Only charged particles couple to photons. Analogously, a colour charge is necessary for strong interaction. All particles interact via weak interaction.

As it has been measured that neutrinos change their family by oscillating, at least two of them need to carry a mass. Therefore, right handed neutrinos have to exist, though they haven't been observed yet. From these couplings, only the charged weak interaction ($W^{\pm}$) can change the flavour of the particles, e.g. $b \to W^- + u$.

2.1.2. Quantum Electro Dynamics (QED)

To describe the effects of the electromagnetic interaction, a quantum field theory can be used in which particles can be created and annihilated.

The Schrödinger equation describes the development of a quantum mechanical wave function of a particle. One possible relativistic extension to it was introduced by Dirac:

$$\left(i\gamma^\mu \partial_\mu - m\right)\psi(x) = 0 \tag{2.1}$$

with $\gamma^{\mu=0,1,2,3}$ being the Dirac Matrices. This is a linear differential equation requiring ψ to be a 4-dimensional spinor. It describes spin-1/2 particles of one type as well as their antiparticles.

This can also be expressed by the Lagrangian:

$$\mathscr{L} = \bar{\psi}(x)\left(i\gamma^\mu \partial_\mu - m\right)\psi(x) \tag{2.2}$$

Using the Euler-Lagrange equation, one obtains equation 2.1.

We now consider unitary local gauge transformations

$$U^\dagger = U^{-1} \tag{2.3}$$

of the form

$$\psi(x) \to \psi(x)' = U(x)\psi(x) \tag{2.4}$$

with

$$U(x) = e^{i \cdot \alpha(x)} \tag{2.5}$$

where α is real.

The Dirac equation (equation 2.1) is not invariant under such transformation. A covariant modification can be constructed by changing the partial derivative ∂_μ to

$$D_\mu = \partial_\mu - iQA_\mu. \tag{2.6}$$

The charge Q describes the strength of the coupling. This introduces a new field A which transforms

$$A_\mu \to A'_\mu = A_\mu + \frac{1}{Q} \cdot \partial_\mu \alpha(x). \tag{2.7}$$

A represents the field of spin 1 photons, which are the mediators of the electromagnetic interaction. Equation 2.1 transforms accordingly to

$$\left(i\gamma^\mu(\partial_\mu - iQA_\mu) - m\right)\psi(x) = 0. \tag{2.8}$$

This is the kinetic term of the electron and the term describing the interaction with the photon field A_μ. For photon propagation, a kinetic term

$$-\frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.9}$$

with

$$F_{\mu\nu}(x) = \partial_\mu A_\nu(x) - \partial_\nu A_\mu(x) \tag{2.10}$$

has to be added.

2.1.3. Quantum Chromo Dynamics (QCD)

Since the first half of the 20th century it was obvious that there must be a force beside the electromagnetic and weak force in order to bind protons and neutrons in the nucleus. It wasn't until 1964 when the introduction of a colour charge solved the spin-statistic problem: The Δ^{++}, a uuu-baryon, possesses a spin of $3/2$, i.e. the quarks have the same quantum numbers. This fact contradicts the Pauli theorem. Therefore a further quantum number has to exist: the colour charge. The term colour charge has been chosen as there are three different states, called red, green and blue, and their antistates. By demanding that all real particles have an effective colour charge of white, one can elegantly accommodate the fact that quarks only exist in clusters of two ($q\bar{q}$, e.g. with colours $r\bar{r}$), three (e.g. qqq, with colours rgb) or more.

The QCD is similarly constructed compared to the QED with the important difference that it acts on a different symmetry group and that the gauge bosons can interact with each another. The corresponding symmetry group is the SU(3). The Lagrangian reads:

$$\mathscr{L} = \bar{q}^a(i\gamma^\mu\partial_\mu - m)q^a \tag{2.11}$$

where q is a 12-dimensional spinor with three colour components times the Dirac components as described in QED. The gauge transformations

$$U(x) = e^{i\theta_s(x)t^s} \tag{2.12}$$

have eight degrees of freedom which correspond to eight gauge bosons, in the case of QCD they are called gluons.

Because gluons change the colour of an interacting quark, they also carry a colour charge. That also means that gluons can interact with each other. This leads to a "running" of the coupling constant α_s: The coupling decreases with higher energy (smaller distances). This provides an idea of why quarks do not exist as free particles: When their distance increases, the stored binding energy increases as well, which leads to confinement.

On the other hand, at high energies quarks behave like free particles, a phenomenon called asymptotic freedom. Therefore at collider experiments, an interaction between quarks rather than between hadrons is observed.

2.1.4. Quantum Flavour Dynamics (QFD)

At the beginning of the 20th century, Hahn and Meitner measured the spectrum of β-decays. This lead to Pauli's postulation of a very weakly interacting neutral fermion, today known as neutrino. Another 30-40 years later, Glashow, Salam and Weinberg developed a theory of the electroweak interaction, postulating the Z and W bosons.

The change in flavour is a remarkable feature of the weak interaction. To describe it, at least a SU(2) symmetry is necessary. It has three generators, therefore it postulates three gauge bosons, W_1, W_2 and W_0. These should be massless according to QFD, which is not realized in nature. The short range of the weak interaction, its low cross section and the direct observation of the mass resonances make it evident that they are massive. How this can be resolved is described in section 2.1.5. The gauge bosons only couple to left handed fermions and right handed anti fermions. A quantum number, the weak isospin is introduced to describe that difference (table 2.2). A linear combination of W_1 and W_2 gives the $W^{\pm}$ bosons:

$$W^+ = \frac{1}{\sqrt{2}}(W_1 - iW_2) \tag{2.13}$$

$$W^- = \frac{1}{\sqrt{2}}(W_1 + iW_2). \tag{2.14}$$

The W_0 boson does not correspond to the Z boson, as the Z also couples to right handed fermions. By unifying the weak and the electromagnetic interaction, one can resolve this problem. A combined symmetry group $SU(2) \times U(1)$ generates a fourth boson, called B. This is not the photon field from QED, but from B and W_0 the photon and the Z boson can be built:

$$A = \cos\theta_W \cdot B + \sin\theta_W \cdot W_0, \tag{2.15}$$

$$Z = -\sin\theta_W \cdot B + \cos\theta_W \cdot W_0. \tag{2.16}$$

The parameter describing the mixing, θ_W is called the Weinberg angle. It has been measured to be $\sin^2\theta_W \approx 0.23$. The quantum number corresponding to the U(1) symmetry is called weak hypercharge Y (table 2.2).

The Quantum Flavour Dynamics is successful in describing and unifying the electromagnetic and the weak interaction. A typical application is the beta decay via a W-boson.

2.1.5. Higgs Mechanism

Although the masses of fermions can be inserted into the QED-Lagrangian (similar in QCD) without violating its invariance under gauge transformation, this is not the case for the boson masses and the QFD-Lagrangian. To account for them, a potential V is added to the Lagrangian:

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2 \tag{2.17}$$

fermion	Q	I	I_3	Y
ν_L	0	1/2	1/2	-1
e_L	-1	1/2	-1/2	-1
e_R	-1	0	0	-2
u_L	2/3	1/2	1/2	1/3
d_L	-1/3	1/2	-1/2	1/3
u_R	2/3	0	0	4/3
d_R	-1/3	0	0	-2/3

Table 2.2.: The quantum numbers describing the electroweak couplings. Q is the charge, I is the weak isospin, Y is the electroweak hypercharge. For anti fermions, the signs invert [3].

with the SU(2) doublet

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_3 + i\phi_4 \\ \phi_1 + i\phi_2 \end{pmatrix}. \tag{2.18}$$

This potential introduces new interactions to the Lagrangian, eventually resulting in three mass terms for the three gauge bosons. The model predicts the masses to be correlated:

$$\frac{m_W}{m_Z} = cos\theta_W, \tag{2.19}$$

which turns out to be in good agreement with electroweak θ_W measurements.

As a consequence of the Higgs model, an additional boson, the Higgs boson is postulated. Experiments at LEP set a limit of $m_H > 114.4\,\text{GeV}$ [6] for a Standard Model Higgs boson. From electroweak measurements one obtains an upper limit of 158 GeV. A combined analysis from direct Higgs searches at CDF and DØ excludes the mass range of 158 GeV to 175 GeV [7].

2.2. Beyond the Standard Model

2.2.1. Deficits of the Standard Model

Indisputable, the Standard Model is successful in describing a wide range of today's observations, not only in collider physics. But on the other hands, there are theoretical limitations, and some astrophysical observations are not described by the Standard Model. A lot of questions remain open [8, 9]:

- The model has 19 free parameters, can they be deduced from first principles?
- Can all theories be unified in a single gauge group? Can gravity be included?
- Why are there three generations of quarks and of leptons? Is there a hidden substructure?
- Why is there an asymmetry between matter and antimatter?

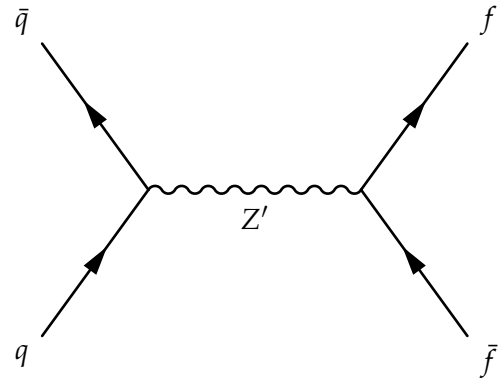


Figure 2.2.: Leading order Feynman graph of a Z'.

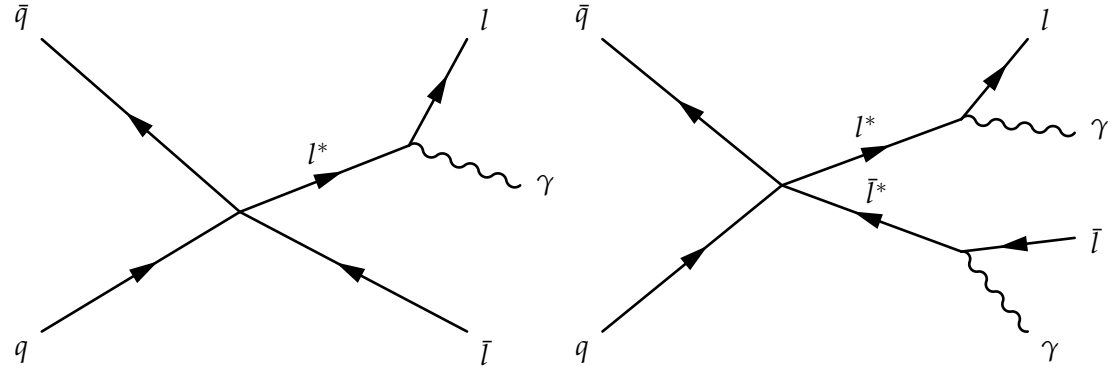


Figure 2.3.: Leading order Feynman graphs of excited lepton production and decay. On the left single excited muon production and on the right pair production.

- As stated above, the Higgs mechanism plays an important role in the Standard Model. The "smoking gun", a Higgs boson resonance, has not been found yet. Additionally, loop diagram corrections suggest a high Higgs mass in the order of the Planck scale [8]. If there is no new physics in the TeV regime, corrections must be extremely fine tuned to produce a Higgs mass in the order of a few hundred GeV as predicted by experiments like L3 and the Standard Model. How can this hierarchy problem be resolved?

- How can the neutrino oscillation (and as a result their masses) observed by other experiments be included into the model?

2.2.2. Heavy Gauge Bosons

A rather simple extension of the Standard Model would be the introduction of another U(1) gauge group. This leads to a new neutral gauge boson, commonly called Z'. There are different models for such a boson, e.g. superstring inspired theories [10] but they mainly differ by the couplings and the breaking scale. We concentrate on the sequential Standard Model, in which the Z' has the same fermion coupling as the Standard Model Z (figure 2.2). It could occur as an excited Z boson in the case of extra dimensions at the weak scale or in models with exotic fermions. It is commonly used as a reference model to compare exclusion results of different experiments [11].

2.2.3. Excited Leptons

Excited fermion states would be an evidence for a fermion substructure. Commonly, the subparticles are called "preons". Standard model fermions can be seen as the ground state of a preon system. Excited leptons carry the same lepton number as their Standard Model relatives. Therefore, they can be produced singly via $q\bar{q} \rightarrow l\bar{l}^*,\ l^*\bar{l}$ as well as pairwise through $q\bar{q} \rightarrow l^*\bar{l}^*$ (figure 2.3). The coupling between excited fermions and Standard Model fermions can be described by a contact interaction [12]:

$$\mathscr{L}_{\text{contact}} = \frac{g_*^2}{\Lambda^2}\frac{1}{2}j^\mu j_\mu \tag{2.20}$$

with

$$j_\mu = \eta_L \bar{f}_L \gamma_\mu f_L + \eta'_L \bar{f}^*_L \gamma_\mu f^*_L + \eta''_L \bar{f}^*_L \gamma_\mu f_L + \text{h. c.} + \text{ right handed.} \tag{2.21}$$

Λ is called the substructure scale below which the interaction occurs. It is estimated to be not much smaller than 1 TeV and the mass of the excited state m^* shouldn't be much lighter than Λ [12]. For excited muons, which are used as a benchmark scenario in section 7.3.2, mass limits of $m > 103.2\,\text{GeV}$ for $\mu^*\mu^*$ and $m > 221\,\text{GeV}$ for $\mu\mu^*$ have been determined at LEP [13].

The cross sections of the single excited quark process can be calculated to [12]

$$\sigma(q\bar{q} \rightarrow l\bar{l}^*,\ l^*\bar{l}) = \frac{\pi}{6s}\left(\frac{s}{\Lambda^2}\right)^2 \cdot \left(1 + \frac{v}{3}\right) \cdot \left(1 - \frac{m^{*2}}{s}\right)^2 \left(1 + \frac{m^{*2}}{s}\right) \tag{2.22}$$

with the velocity of the excited lepton

$$v = \frac{s - m^{*2}}{s + m^{*2}}. \tag{2.23}$$

2.2.4. Leptoquarks

The symmetry between quarks and leptons has no underlying common principle in the Standard Model. It can be explained by other models, such as SU(5) grand unification, Pati-Salam SU(4), composite models, technicolour, and super-string inspired models. All of them introduce a new type of boson, called leptoquark (LQ) [15].

A leptoquark decays into a quark and a lepton. Therefore, it carries a charge of $\pm 1/2$. Baryon or lepton number violating couplings are excluded for light leptoquarks as this would lead to proton decay [16]. Instead, leptoquarks carry both, a lepton number as well as a quark number. Experimental constrains favour the decay into same generation particles [15].

At proton proton colliders, leptoquarks are predominantly pair produced via gluon gluon fusion [15], but also by quark antiquark annihilation (figure 2.4).

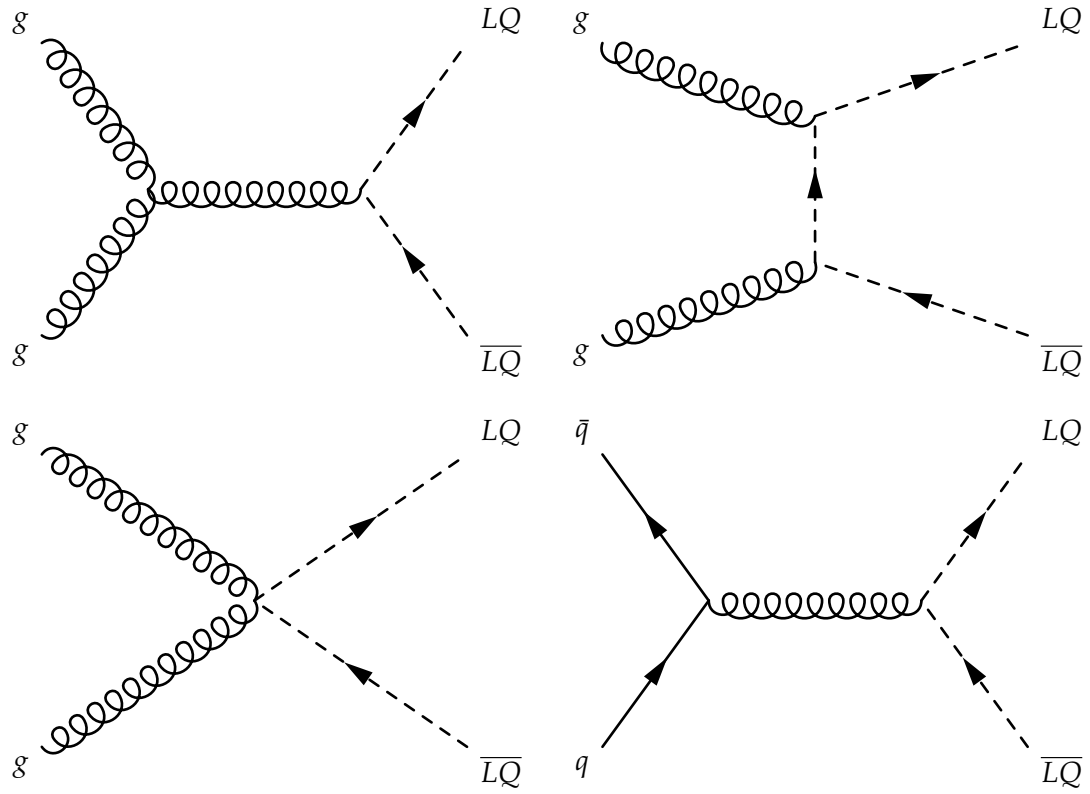


Figure 2.4.: Leading order Feynman graphs of leptoquark production at the LHC [14].

3. Experimental Setup

The data used in this work were acquired by the CMS experiment. This experiment is located around an interaction point of the proton-proton collider LHC. Due to the complexity of the experiment, this thesis can only give a short overview of the experimental setup. For detailed information please refer to [17, 18].

3.1. Collider Experiment

Usually, high energy physics deals with the interaction of two particle systems. Two different types of experiments exist: Fixed target experiments evaluate the interaction of accelerated particles with a static block of matter, whereas in collider experiments, both interacting particles are accelerated towards each other.

Two important quantities are the center of mass energy $\sqrt{s}$ and the luminosity $\mathscr{L}$. For most experiments a high center of mass energy is favorable. For the production of yet unknown heavy particles, a center of mass energy of at least their mass is essential, as energy and mass are equivalent according to the special relativity:

$$E^2 = m^2 + p^2. \tag{3.1}$$

Also small substructures can only be resolved by high energies. This is a consequence of Heisenberg's uncertainty principle:

$$\Delta x \cdot \Delta p \leq \hbar. \tag{3.2}$$

It suggests an antiproportional correlation, so that one requires an energy transfer of roughly 200 MeV to resolve length scales of order 1 fm.

Luminosity is a measure for the expected number of events of a certain event type. The event rate of a certain type of process can be determined by the luminosity and the process cross section:

$$\dot{N} = \sigma \cdot \mathscr{L}. \tag{3.3}$$

The cross sections depend on the type of interacting particles and their center of mass energy. Some important cross sections are depicted in figure 3.1. They span several orders of magnitude, making it a challenge to extract processes with a small cross section.

The instantaneous luminosity at a collider experiment with a Gaussian beam profile can be calculated from the number of bunches n, the number of particles inside a bunch N, the revolution frequency f and the beam cross sections σ_x, σ_y to

$$\mathscr{L} = \frac{nN^2 f}{4\pi\sigma_x\sigma_y}. \tag{3.4}$$

The idea of collider experiments emerged in the 1950s [20]. They have a lower luminosity than fixed target experiments, where ideally nearly all particles of a single beam interact with a block of material. The advantage of collider experiments is the high center of mass energy

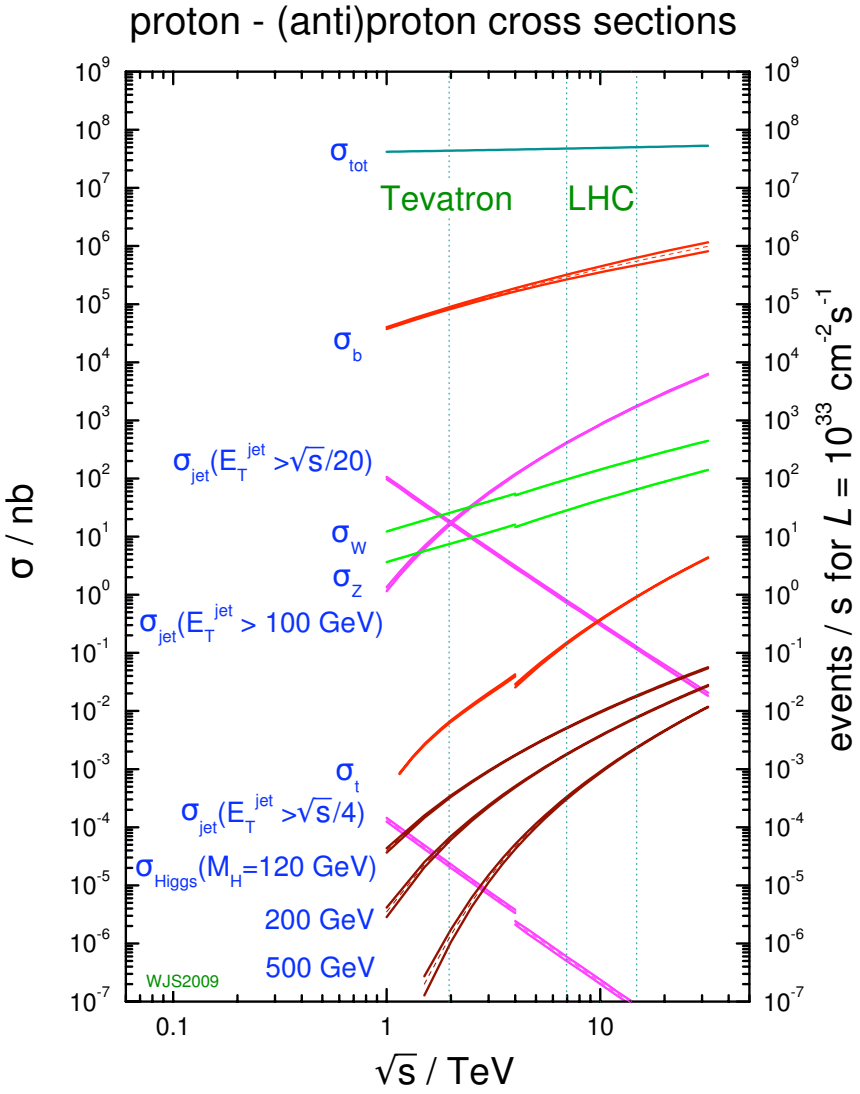


Figure 3.1.: Cross sections at hadron colliders. Marked are the Tevatron at $\sqrt{s} = 1.96\,\text{TeV}$ and the two LHC configurations at $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$. The gap at around 4 TeV indicates the difference between $p\bar{p}$ (left) and pp (right) interactions [19](modified).

$\sqrt{s}$. In fixed target experiments $\sqrt{s} \propto \sqrt{E_{\text{beam}}}$ applies whereas in collider experiments with particles of the same mass and energy

$$\sqrt{s} = 2 \cdot E_{\text{beam}} \tag{3.5}$$

applies.

Ring colliders have the advantage of being able to recycle the beam for a lot of collisions and therefore it is easier to achieve a higher luminosity than at linear colliders. In addition, accelerating paths can be reused each time the particles orbit, resulting in less infrastructure and a more compact scaling. On the other hand, linear colliders are less limited by synchrotron radiation or by the necessity of a high B field to keep the particles on the curved track.

Different kinds of particles can be used for collision experiments. Typical choices are proton-proton/antiproton, electron-electron/positron and electron-proton. Not suitable are unstable particles (e.g. tau) and neutral particles. As electrons are rather light ($m \approx 511\,\text{keV}$), their kinetic energy in rings is limited by synchrotron radiation. The disadvantage of proton experiments manifests in the fact that protons are not elementary particles (section 2.1.1). In a hard interaction, to first order only one quark or gluon interacts with one constituent of the other proton. Neither the elementary particles of the interaction nor their center of mass energy is usually known. Additionally, this can lead to proton remnants in the detector, making it a challenge to identify the hard interaction properly.

3.2. Large Hadron Collider (LHC)

The LHC [17] is a particle accelerator and collider located near Geneva, Switzerland. It is capable of accelerating protons and heavier nuclei in two rings, though this work takes into account only proton-proton collisions. The design center of mass energy of the colliding protons is 14 TeV. In 2010, the accelerator has run with a center of mass energy of $\sqrt{s} = 7\,\text{TeV}$, which has never been reached before at a collider experiment. A maximal luminosity of $10\,\text{nb}^{-1}\text{s}^{-1}$ is envisaged[1], using 2808 bunches with a bunch distance of 25 ns. With this configuration 19 interactions per beam crossing are expected on average [21]. This pile up of events is a challenge for the analyses as the superposition of different processes has to be considered.

In 2010, a maximal instantaneous luminosity of $\mathscr{L} = 205\,\mu\text{b}^{-1}\text{s}^{-1}$ has been achieved. Integrated over time a total luminosity of $43.17\,\text{pb}^{-1}$ has been recorded [22].

A radio frequency (RF) acceleration system as used in the LHC cannot handle low energy (non relativistic) particles. Therefore an injection chain of different accelerators is used (figure 3.2): Ionized hydrogen is successively inserted into Linac2, Proton Synchrotron Booster (PSB), Proton Synchrotron (PS) and Super Proton Synchrotron (SPS), which have been upgraded to fulfill the requirements of LHC. Protons with an energy of 450 GeV are then injected into the LHC.

The RF system consists of eight cavities per beam. They are fed by 300 kW klystrons and accelerate the proton bunches with electromagnetic waves of a frequency of 400 MHz. The system not only accelerates the particles from 450 GeV to 7 TeV but also collimates the bunches longitudinally and compensates energy losses from synchrotron radiation.

In order to bend the beam trajectory to a ring, superconducting dipole magnets are put into place. For proton accelerators, the strength of the dipoles is the limiting factor for the maximum beam energy. Therefore, 1232 14 m-magnets with a maximum field strength of 8.33 T are used. The superconductors are cooled with superfluid helium in order to keep a temperature of 1.9 K.

In addition, higher order magnets (quadrupoles etc.) are used to influence the beam structure.

[1] $1\,\text{b} = 10^{-28}\,\text{m}^2$

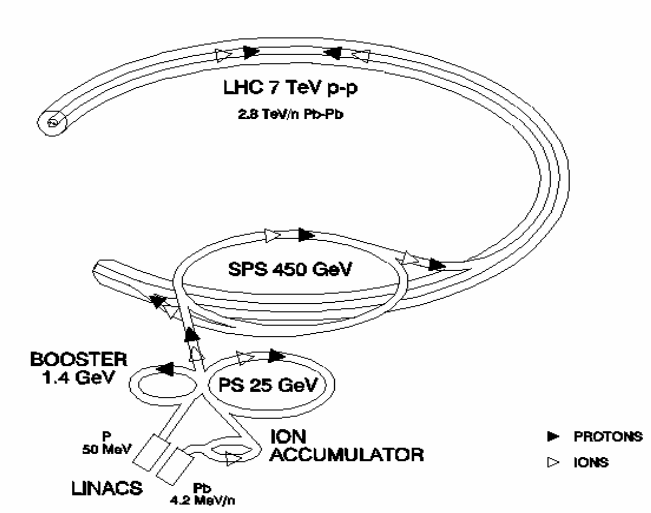


Figure 3.2.: The injection chain of the Large Hadron Collider [21].

The LHC, sometimes referred to as the world's largest machine, possesses a spectrum of very different sophisticated technologies in a variety of different fields: cryogenic systems, superconducting magnets, RF systems, vacuum technology, and beam instrumentation to mention only some of them. Each is vital for the LHC and was constructed with a tremendous effort, but a detailed description is beyond the scope of this thesis.

3.3. Compact Muon Solenoid (CMS)

The Compact Muon Solenoid (CMS) is one of two multi purpose experiments at the Large Hadron Collider. It is located in an underground cavern near Cessy, France. The detector is assembled from five wheels and two endcaps (figure 3.3). It weighs 14 000 t but is 21 m long and has a diameter of 15 m. Therefore the detector is called "compact". It consists of several concentric layers. Namely the tracker to determine interaction vertices and the trajectory of charged particles, the calorimeter to measure the energy of particles, and the muon chambers for a precise identification and measurement of muons. Beside that, CMS also incorporates a 3.8 T solenoid magnet and an iron return joke. The strong magnetic field is necessary to determine the momentum of relativistic particles from their curvatures.

3.3.1. Tracker

The tracker system is the inner most detector part. It includes the pixel and the strip tracker.

Silicon Pixel Tracker

On the very inside the pixel detector is located, which consists of three layers in the barrel and two in the endcaps. This allows it to measure three trajectory points, with three coordinates each, over nearly the whole range up to $|\eta| < 2.5$. The pixel detector incorporates 66 million pixels with a size of 100 µm × 150 µm. Usually more than one p-n junction of the detector is affected by a particle crossing. Using this so called "charge sharing" effect a resolution of down to 15 µm × 15 µm can be achieved. The axis of the detector elements are rotated by 20° from the expected particle trajectory to further enhance this effect. It is of special importance, as a pixel detector near the interaction point is the only possibility to determine the multiple

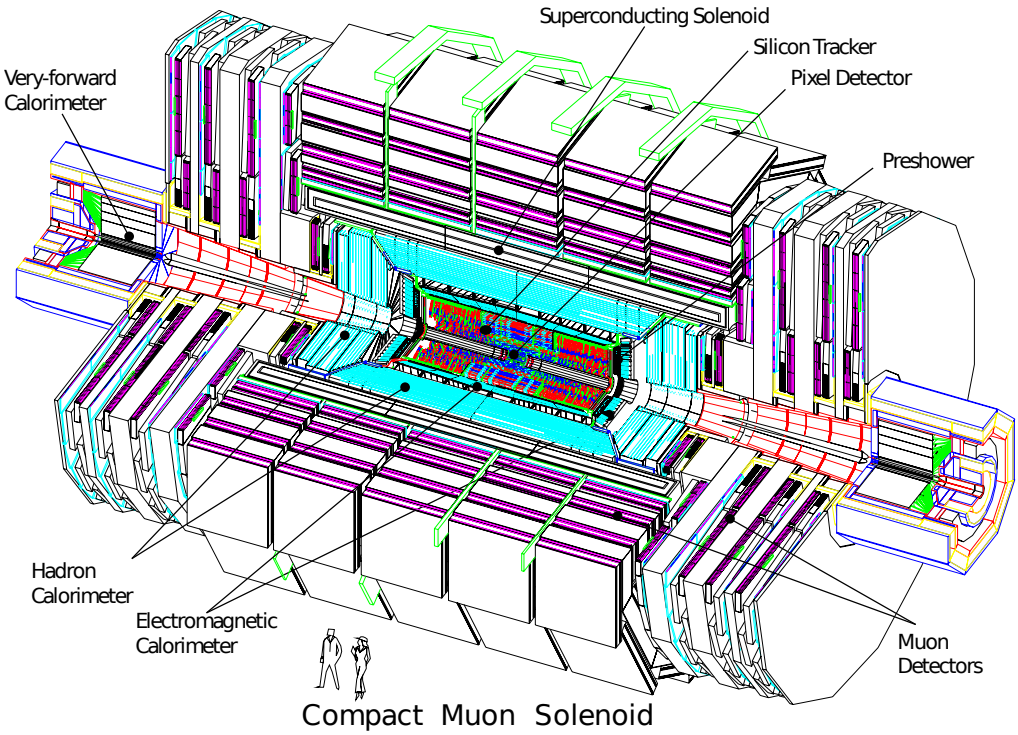


Figure 3.3.: The Compact Muon Solenoid [18].

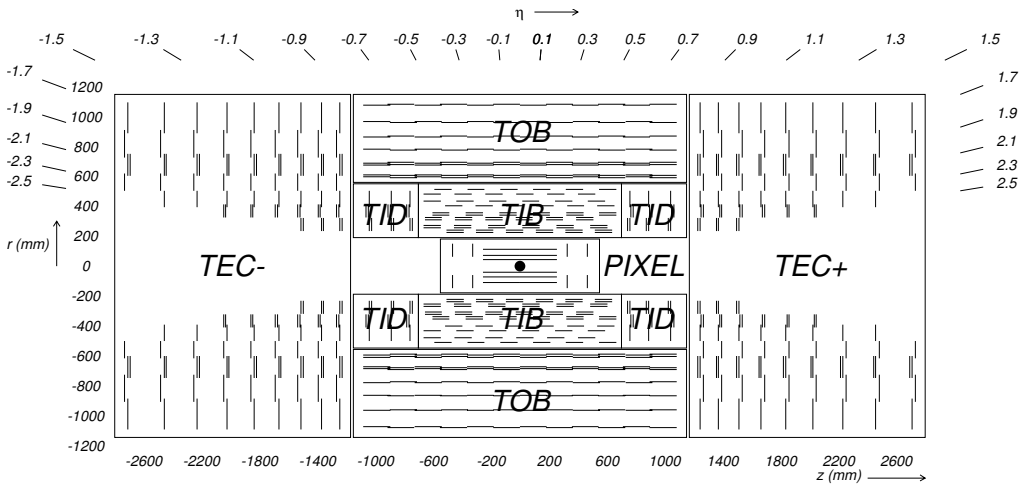


Figure 3.4.: The Layout of the CMS tracker as a cross section in the *r*-*z* plane [18].

primary vertices originating from piled up events and secondary vertices resulting from decays of long-lived particles like b-quarks and τ-leptons.

Silicon Strip Tracker

The SIlicon Strip Tracker is subdivided into four parts: The tracker inner barrel (TIB) and tracker inner disks (TID) on the inside are surrounded by the tracker outer barrel (TOB) and the tracker endcaps (TEC).

The main task of the strip tracker is to measure the curvature of the trajectories. As the solenoid produces a homogeneous 3.8 T magnetic field in z-direction inside the tracker (see section 3.3.5), the measurement of the r-ϕ coordinates has a high priority.

TIB and TOB strips can measure the r-ϕ co-ordinates of a trajectory point as they are aligned along the beam axis. TID and TEC strips are aligned perpendicularly to the beam axis. Some strip layers have a stereo layer attached. This is tilted by 0.1 rad. By that, the third coordinate of a trajectory point can be determined with a resolution of 230 µm - 530 µm. The layout of the tracker ensures that at least nine layers are passed by each trajectory in an η-range up to 2.4. The parts of the tracker are constructed as follows:

- The TIB consists of four layers of silicon strips and has a resolution of 23 µm-35 µm.
- The six TOB layers are more separated and thus have a resolution of 35 µm-53 µm.
- Each TEC consists of nine discs and provides up to nine ϕ measurements.
- Each TID consists of three discs and can measure coordinates with a resolution of 23 µm-35 µm.

Typical transverse momentum resolutions of the total tracker system can be found in figure 3.5. For a 100 GeV muon, the resolution is better than $< 2\%$ in the barrel.

3.3.2. Electromagnetic Calorimeter

To measure the energy of electrons and photons, generated during the collisions, an electromagnetic calorimeter is used. The CMS electromagnetic calorimeter (ECAL) is a hermetic and homogeneous calorimeter assembled from 75848 lead tungstate crystals. The dominant interactions of the electrons are ionisation and bremsstrahlung. Photons lose energy via the photoelectric effect, Compton scattering, and pair production. Electrons and photons deposit most of their energy in the ECAL, whereas hadrons and muons only deposit a portion of the energy here.The particles excite electrons of the scintillation material and eventually eV-photons are generated. The number of photons is proportional to the energy deposited.

A typical energy resolution for $E < 500\,\text{GeV}$ was determined to [18]:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E/\text{GeV}}}\right)^2 + \left(\frac{0.12}{E/\text{GeV}}\right)^2 + (0.30\%)^2. \tag{3.6}$$

Crystals

The crystals used in the CMS ECAL are made of $PbWO_4$. This is an anorganic scintillator, which is reasonably radiation hard. Though it is not a plastic scintillator, it is fast enough to cope with the LHC bunch distances: After 25 ns, which is the minimal time between two bunch crossings, 80% of the photons are emitted.

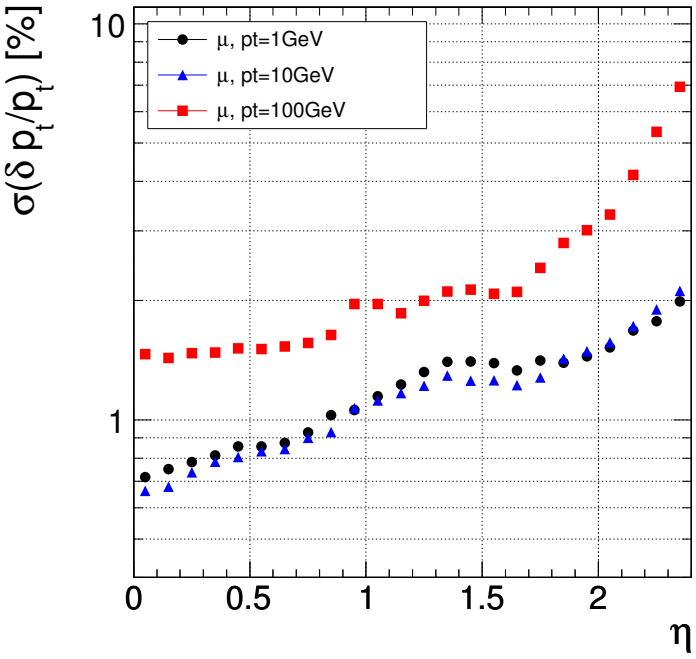


Figure 3.5.: Momentum resolution of muons as measured by the tracker [18]. Depicted is the width σ of the normal distribution descibing the relative deviations of p_t.

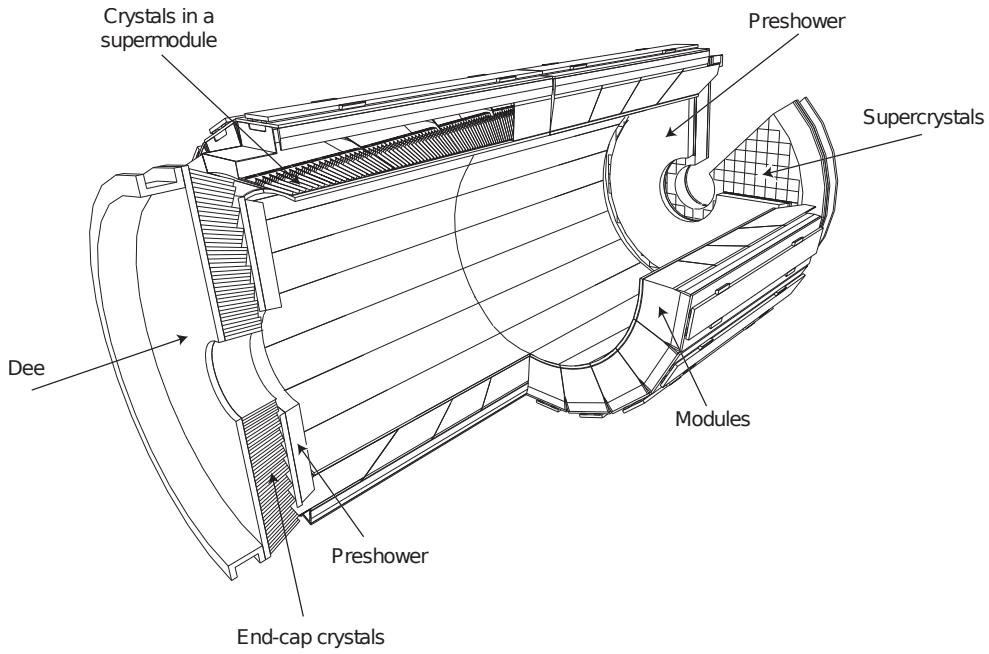


Figure 3.6.: The Electromagnetic Calorimeter of CMS [18].

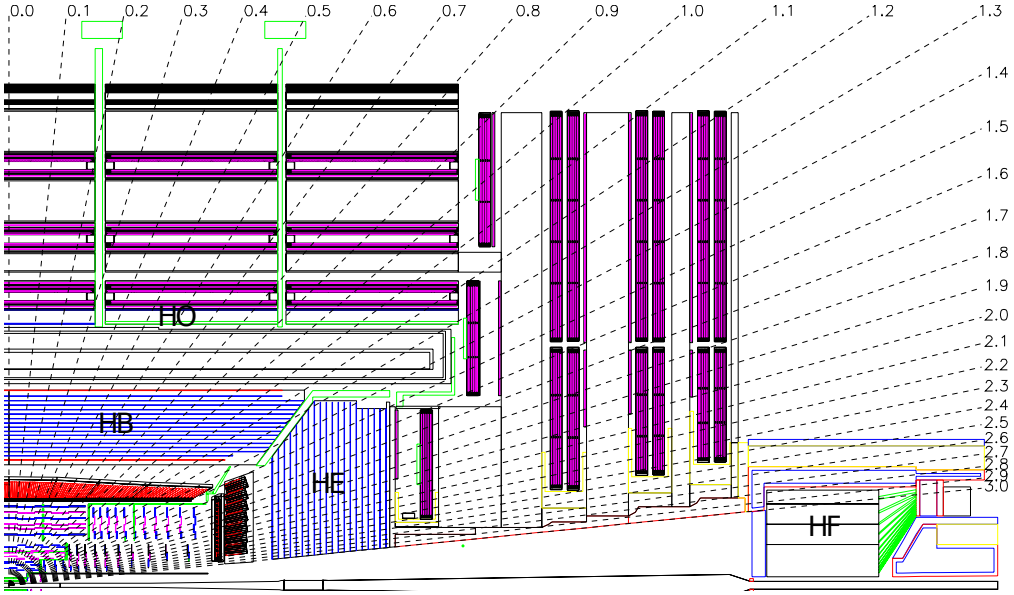


Figure 3.7.: The Hadronic Calorimeter of CMS [18].

The material has a radiation length of $X_0 \approx 0.89\,\text{cm}$ and a Molière radius[2] of $r_M \approx 2.2\,\text{cm}$. These are the scales of an electromagnetic shower. The CMS crystals are approximately 25 radiation lengths long and have a width of about 1 Molière radius. Their dimensions in $\Delta\phi$-$\Delta\eta$ are approximately 0.0174×0.0174, leading to a high spatial resolution. This is important, for instance to detect collimated photons from a Higgs decay.

The shape of the crystals is that of truncated pyramids. Their axes are tilted by 3° to the beamspot, to prevent cracks in the crystals from being aligned along possible trajectories. Therefore, crystals differ in shape depending on their η-position.

Layout and Photodetectors

Following the overall cylindrical layout of CMS, the ECAL is assembled from a barrel part ($|\eta| < 1.479$) and two endcaps ($1.479 < |\eta| < 3.0$).

In the barrel, avalanche photodiodes are used as photodetectors. Because of differences in the radiation and the magnetic field in the region, the endcap crystals are read out by vacuum phototriodes.

Another notable difference between the two regions is the 20 cm thick preshower detector in front of the actual endcap ECAL. It is designed to identify neutral pions, in order to better distinguish between electrons and minimal ionizing particles, and to improve the spatial resolution of the ECAL.

3.3.3. Hadronic Calorimeter

For heavy charged particles the dominant effect of losing energy is via strong interactions with nuclei of the calorimeter material. Their interaction length λ_I is usually longer than the radiation length for electrons and photons.

The hadronic calorimeter (HCAL) includes four different parts, as can be seen in figure 3.7. The hadron calorimeter barrel and endcap surround the electromagnetic calorimeter. They are

[2]The radiation length of a material is the distance by which an electron has reduced its energy by a factor of e. 90% of the energy is deposited in a cylinder with a radius of $r \approx r_M$ [23].

supplemented by the hadron outer calorimeter, which is located outside the solenoid, and the hadron forward calorimeter.

- The *Hadron Barrel Calorimeter* (HB) is a sampling calorimeter, consisting of alternating layers of absorber and scintillator material. Brass is used as absorber, only the most inner and most outer plate is made of stainless steel. Between them, plastic scintillators produce the photons which are guided via wavelength shifting fibers to the photo detectors. The detector covers a range of $|\eta| < 1.3$ and the scintillator tiles have a granularity of $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$.
- The *Hadron Outer Calorimeter* (HO) is the continuation of the HB beyond the superconducting magnet and covers the same η-range. It is used to measure the shower energy deployed beyond the HB and is necessary as it extends the material thickness at $\eta = 0$ from $5.82 \cdot \lambda_I$ to $11.8\lambda_I$. The HO consists of one scintillation layer located in front of the first muon system layer. In the central detector wheel, it is supplemented by a second scintillation layer between the solenoid and the return yoke.
- The *Hadron Endcap Calorimeter* (HE) covers the range from $|\eta| = 1.3$ to $|\eta| = 3.0$. It has a similar structure as the HB. The granularity varies from $(\Delta\eta, \Delta\phi) = (0.087, 0.087)$ for $|\eta| < 1.6$ to $(\Delta\eta, \Delta\phi) = (0.17, 0.17)$ for $|\eta| > 1.6$.
- In the forward region from $|\eta| = 2.9$ up to $|\eta| = 5.2$, the *Hadron Forward Calorimeter* (HF) is located. On average, the energy deposited in this part is more than seven times higher than the energy deposited in the rest of the detector. For this reason a different, more radiation hard detector type is used. The HF is a Cherenkov detector made out of steel with quartz fibers inserted. The readout resolution is $(\Delta\eta, \Delta\phi) = (0.175, 0.175)$

Test beam studies determine the typical combined energy resolution of the two calorimeters (ECAL+HCAL) for pions to [24]:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{101\%}{\sqrt{E/\mathrm{GeV}}}\right)^2 + (4.0\%)^2. \tag{3.7}$$

3.3.4. Muon System

As muons traverse the whole detector without depositing much energy, a special detector can be located in the outer part of the detector (figure 3.8). Beside the solenoid magnet, the muon detection system is the name-giving feature of CMS. It is designed to measure the trajectory and the momentum of muons to a high precision. Usually it is assumed that all particles reaching the muon chambers are muons, even though some high energetic hadrons might pass the hadron calorimeter (This effect is called ”punch-through“.).

As muons are charged particles, they are also detected by the silicon tracker system. For cost reasons, gas detectors are used for the muon system. A comparison of the muon p_t resolution between tracker and muon system can be found in figure 3.9. The synergy of both leads to an optimal resolution over the entire p_t region of interest up to the TeV regime.

There are three different types of muon chambers built into CMS:

- In the barrel, *drift tubes* (DT) are located. Their choice is motivated by the comparably low muon rate, the low neutron-induced background and the fact that the magnetic field is mostly contained in the return yoke. They cover the range $|\eta| < 1.2$ and are arranged in four stations around and inside the iron return yoke structure. Each of them contains eight DT layers to measure the r-ϕ-coordinates. In the first three stations another four layers of DTs, rotated by 90°, determine the z-coordinate as well.

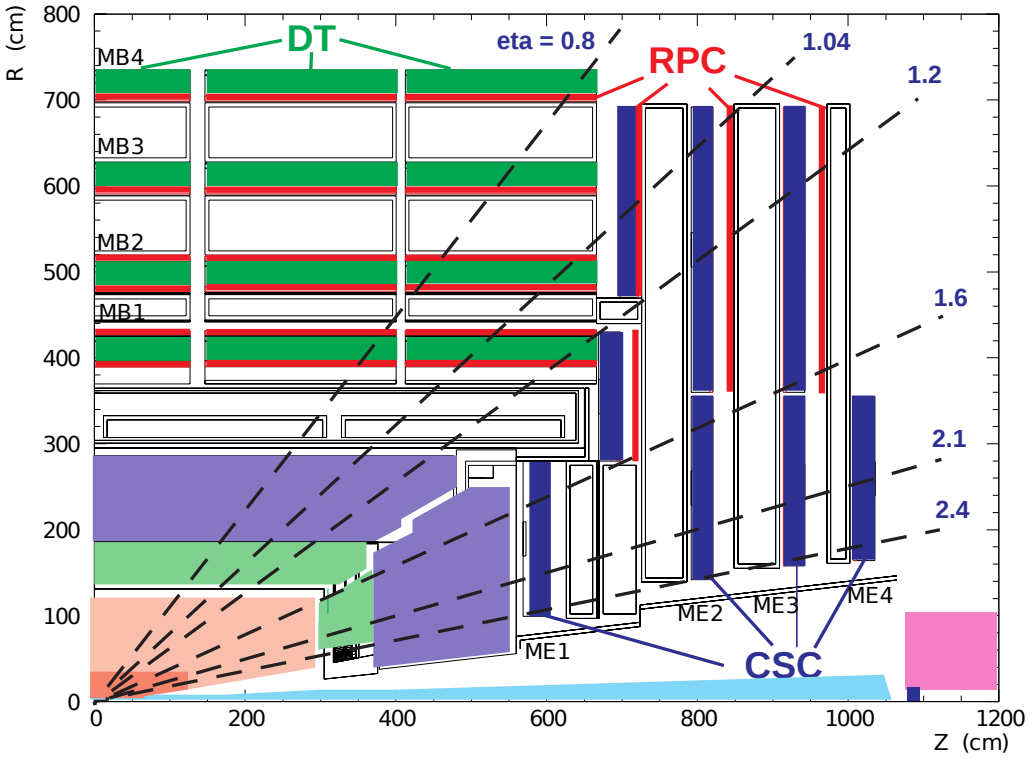


Figure 3.8.: The Muon System of CMS [25].

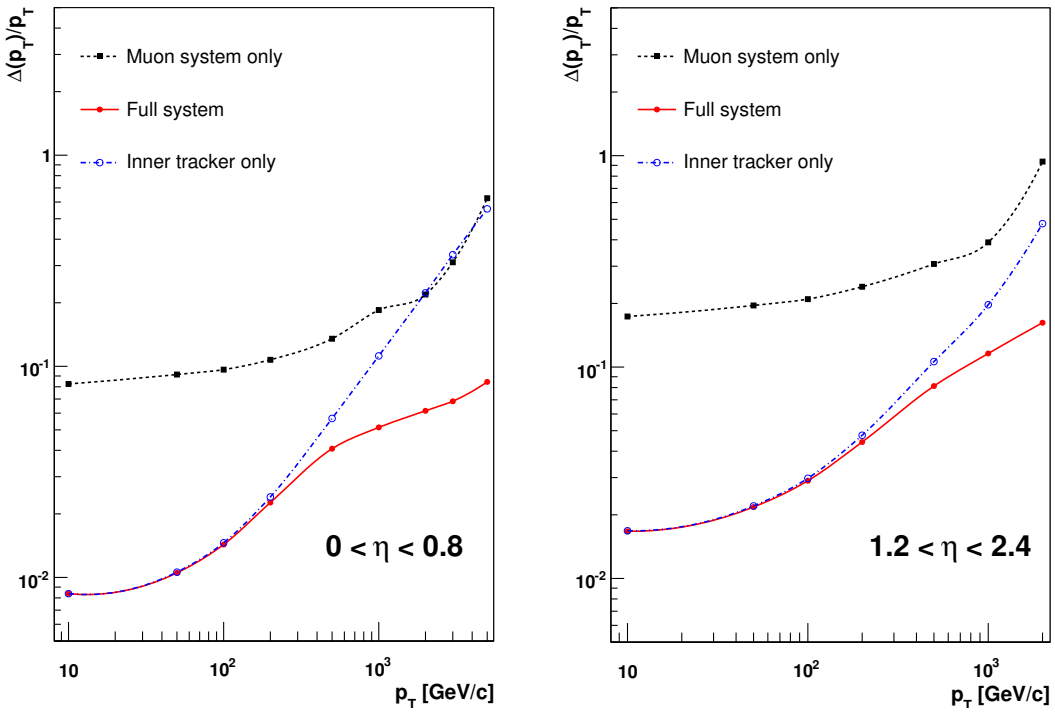


Figure 3.9.: Comparison between the transverse momentum resolution of a muon in terms of p_t and the used detector elements [18].

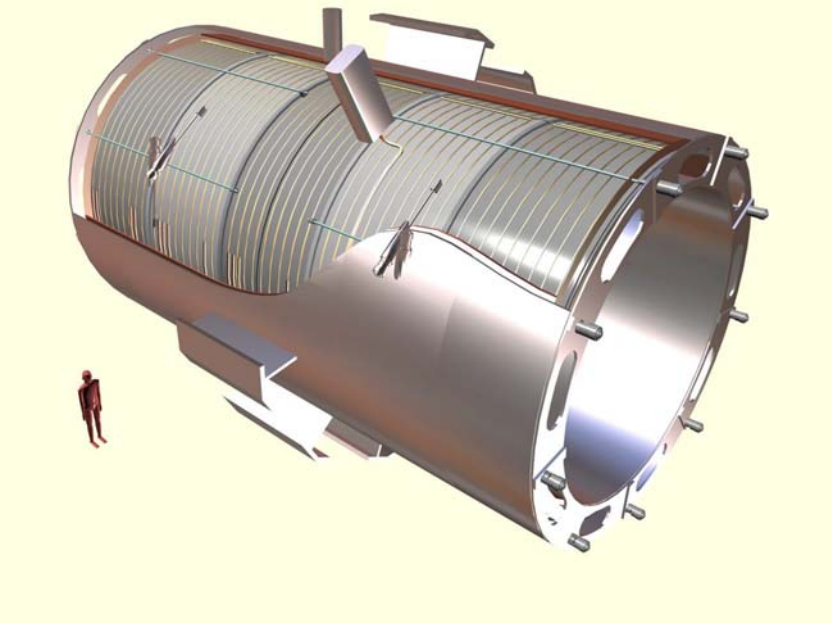


Figure 3.10.: Artwork of the CMS superconducting solenoid [18].

- In the endcaps, the magnetic field is non-uniform and high event rates are present. In this region, which ranges from $|\eta| = 0.9$ to $|\eta| = 2.4$, *cathode strip chambers* (CSC) are used instead of DTs. There are four stations with 6 layers each. The CSC is made of multiwire proportional chambers, where both the cathode strips and anode wires are read out. This way, all three coordinates and timing information are available.

- *Resistive plate chambers* are gaseous detectors with two parallel plates. They have an excellent time resolution, better than the bunch distance of 25 ns. Due to this fact, they are used for triggering, complementing the other two types of muon chambers. The chambers are built into the barrel as well as the endcap region up to $|\eta| = 1.6$.

3.3.5. Solenoid

The momentum of a particle is determined from the curvature of its trajectory, measured in the tracker and the muon system. Comparing centrifugal force and Lorentz force results in

$$r = \frac{p_t}{q \cdot B}. \tag{3.8}$$

Where r is the radius of the curvature, p_t is the transversal component of the momentum, q is the charge and B the magnetic field. The greater r becomes, the greater is its uncertainty. Therefore, a high magnetic field B is necessary to determine large particle momenta.

The CMS magnet is operating at a magnetic field of $B = 3.8\,\mathrm{T}$. With a maximum stored energy of 2.6 GJ it sets a record for collider experiments.

The magnet is a superconducting solenoid, i.e. it has a cylindrical coil which causes the magnetic field lines to be mainly parallel to the beam axis. It is located between the HCAL Barrel and the HCAL Outer. The superconductor is made of four layers of NbTi which are cooled down by liquid helium to 4.6 K. The coil is embedded inside an aluminium alloy support structure.

The flux is returned by a 10 000 t iron yoke on the outside which accomodates part of the muon system. At this point, the magnetic field is antiparallel compared to the inside and has about half the strength.

3.3.6. Trigger System

As mentioned earlier, the design of CMS includes a bunch crossing distance of 25 ns, i.e. an event frequency of 40 MHz. Obviously it is not possible to store the arising amount of data completely. The CMS trigger system filters the events, so that the most interesting ones can be kept and the others are dismissed. This is done in two steps:

- The *Level 1 trigger* (L1) is implemented via programmable electronics and reduces the event rate to around 50 kHz. Low-resolution information from the calorimeters and the muon chambers are processed, while the complete data is buffered. The triggers are multi-staged and the allowed run time for them is 3.2 µs in order not to lose any events stored in the pipeline.
- The *High Level Triggers* (HLT) run on the data reduced by L1 triggers. They also use the muon system and calorimeter information, and perform particle reconstruction. Computer farms are used to process the data. The event rate is further reduced to about 150 Hz. Eventually, these events are stored and can be analysed.

As the experience with the data and the luminosity change over time, the HLT change accordingly. A description of the HLT used in this analysis can be found in section 5.2.

4. Computing and Software Framework

To analyse the large amount of CMS data a number of different software applications is used. Some important tools are presented in this chapter. They range from standard analysis tools (ROOT) to high energy physics and CMS specific software. As most analyses are similar at some steps synergy effects can be exploited. Additionally, the usage of the LHC computing grid is explained.

4.1. WLCG

The huge amount of data produced by the LHC experiments evoked the concept of the *World-wide LHC Computing Grid* (WLCG). CMS data is not only stored at CERN, but distributed to more than 100 data centers worldwide. Each dataset is transfered on more than one site. The analyses are then run at these sites and the results are sent back. Currently, some 100 MB per second of CMS data is transferred between the sites for this purpose altogether [26].

4.2. CMSSW

CMSSW is a collection of software used for CMS event processing. A large amount of modules is available, intended for particle reconstruction and analysis. They are executed sequentially and are able to evaluate or add something to the event content. Interfaces to Monte Carlo event generators exist to produce appropriate samples.

4.3. Monte Carlo Generation and Processing

This analysis uses officially produced data and Monte Carlo samples as noted in table A.1 in the appendix. The standard chain of processing for these differ:

First, Monte Carlo samples are generated, including the hard interaction, possible initial and final state radiation and further processes like hadronisation of quarks or decay of unstable particles (GEN). These objects are then fed into a GEANT 4 detector simulation (SIM) and the virtual response of the detector is determined (DIGI). From this, the event content is reconstructed (RECO). This last step is the same for the actual data. Sometimes, an updated reconstruction is run later (RERECO) [27].

4.4. Miscellaneous Software

Besides the above mentioned software, a number of other libraries and application were used for this analysis, some playing an important role. The PXL [28] library and its file format PXLIO is extensively used by MUSiC. It offers a processing framework and the PXLIO file format is a lightweight protocol which can handle the large amount of events evaluated by MUSiC.

The ROOT [29] package incorporates scientific analysis and visualization tools, which are especially used to work with histograms.

For numerical integration, the GNU Scientific Library [30] is used.

5. Data Selection

With the help of software as stated in section 4, the CMS data acquired from the various detector elements (section 3.3) is used to reconstruct the original particle content of the events. In this chapter, the particle reconstruction, selection criteria for events and particles as well as the evaluated data samples are described. The reconstruction criteria are common CMS criteria used by various groups. The selection criteria are chosen especially for the MUSiC analysis.

5.1. Data and Monte Carlo Samples

To describe the Standard Model, QCD, γ, W, Drell-Yan, $t\bar{t}$, Di-Boson and Upsilon samples are used. A detailed list of the samples can be found in table A.1 in the appendix. Events from all samples can include jets from initial and final state radiation. In the case of PYTHIA they are simulated using a phenomenological parton shower approach. ALPGEN and MADGRAPH can simulate a limited number of jets directly.

The data was acquired in the 2010 7 TeV run. This analysis uses an integrated luminosity of $36.1\,\mathrm{pb}^{-1}$.

To achieve this, a high instantaneous luminosity is necessary. As a consequence several proton-proton interaction happen at one bunch crossing. This effect is called pileup. To account for this, simulated minimum bias events are mixed into the Monte Carlo events. The only requirement for minimum bias events is to observe any signature in the detector. The probability distribution used for the pileup events added to the Monte Carlo events is depicted in figure 5.1.

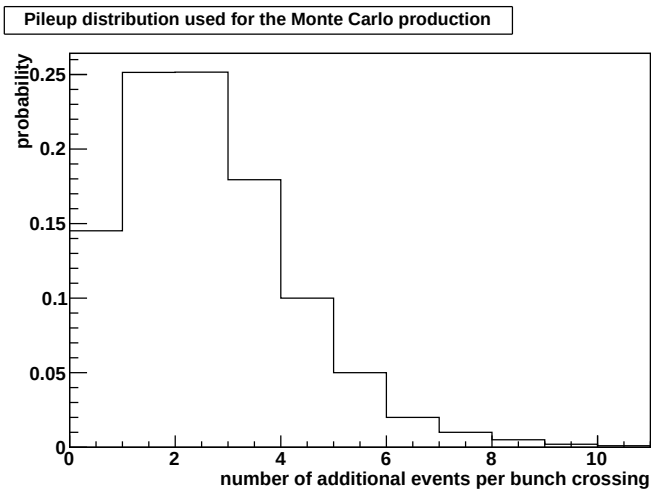


Figure 5.1.: Probability distribution used to add pileup events to the Monte Carlo.

CMS Trigger Name	CMS Run Range	Integrated Luminosity
Mu9	135808 to 147116	$8.3\,\mathrm{pb}^{-1}$
Mu11	147117 to 148818	$9.5\,\mathrm{pb}^{-1}$
Mu15_v1	148819 to 149063	$18.5\,\mathrm{pb}^{-1}$
Ele20_LW_L1R	135808 to 141949	$0.3\,\mathrm{pb}^{-1}$
Ele20_SW_L1R	141950 to 144114	$2.9\,\mathrm{pb}^{-1}$
Ele17_SW_EleId_L1R	144115 to 147116	$5.1\,\mathrm{pb}^{-1}$
Ele17_SW_TightEleId_L1R	147117 to 148818	$9.5\,\mathrm{pb}^{-1}$
Ele22_SW_TighterEleId_L1R_v2	148819 to 149063	$10.3\,\mathrm{pb}^{-1}$
Ele22_SW_TighterEleId_L1R_v3	149064 to 149442	$8.1\,\mathrm{pb}^{-1}$

Table 5.1.: List of used muon and electron triggers. No trigger is prescaled, i.e. all triggered events were recorded. In the first column, the number after the lepton abbreviation gives the p_t threshold in GeV of the trigger [31].

5.2. Event Selection

Only events that pass the high level triggers (section 3.3.6) as summarised in table 5.1 are taken into account.

Only events with at least one muon or one electron are selected. For them, robust triggers exist. γ and jet-events without leptons are dominated by QCD. This is a large background which is hard to handle and for which not enough Monte Carlo events are available resulting in bad statistics. This restriction should be reevaluated for future analyses.

Additionally, at least one muon must fulfil $p_t > 25\,\mathrm{GeV}$ or one electron must fulfil $p_t > 30\,\mathrm{GeV}$. This is necessary to have the same selection for all data samples as the data is acquired using triggers with different p_t thresholds. It is slightly higher than the highest criterion of the triggers to exclude the phase space region in which the triggers have a bad efficiency. Further leptons in the event have loser restrictions as described in the following sections.

Apart from that, events with more than 10 tracks are rejected if more than 25% of the tracks are badly reconstructed. Additionally, the following criteria for a good primary vertex are applied:

- A minimum of four tracks, each having a goodness of the fit of $\frac{\chi^2}{\mathrm{ndf}} < 10$, were used for its reconstruction.
- The vertex is less than 2 cm away from the interaction point in the r-ϕ plane.
- The vertex is less than 24 cm away from the interaction point in z-direction.

5.3. Particle Reconstruction and Selection

The reconstruction of the main physics objects used by MUSiC, electrons, muons, photons, jets and missing transverse energy, is described below. Figure 5.2 illustrates how different particles interact with the detector.

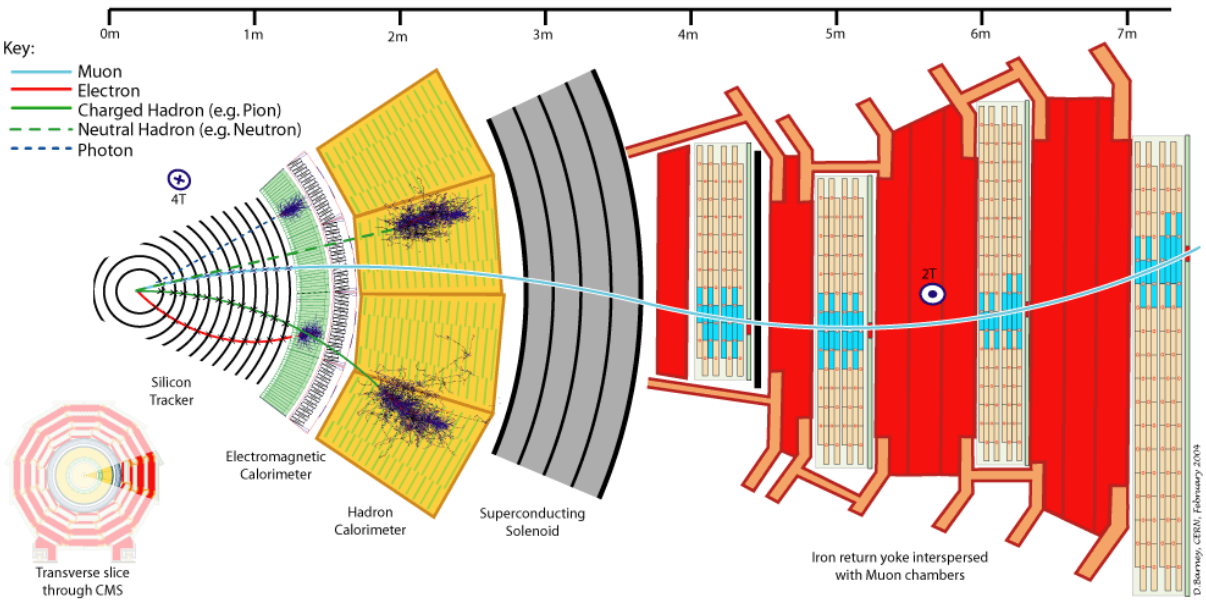


Figure 5.2.: Scheme how particles interact with the CMS detector. The muon trajectory are identified in the tracker and in the muon chambers. Photons and electrons deposit their energy in the ECAL, whereas electrons can be seen in the tracker as well. Hadrons deposit most of their energy in the HCAL. Charged hadrons leave hits in the tracker as well [32].

One has to keep in mind that the reconstruction is subject to uncertainties. Therefore it is crucial to have a reliable detector simulation so that Monte Carlo and data behave in a similar way. Also a realistic estimation of the uncertainties is essential. This is described in section 6.6.

5.3.1. Muon

MUSiC uses global muons, i.e. muons that are reconstructed in both the muon system and the silicon tracker. Hits from one DT or CSC chamber are first combined to segments, which correspond to trajectories inside one chamber. Each segment can be used as a seed to start propagating a track through the muon system. At each step, the properties of the muon candidate are updated to be less dependent on the seed properties. If the combined track fulfills quality criteria such as a good χ^2, it will give rise to a so called standalone muon.

Reconstructed trajectories in the silicon tracker are matched with the standalone muon by propagating both to the same surface. Only if they fulfill certain momentum and spatial criteria, a refit is performed using hits from the standalone muon and the corresponding track. If there is more than one tracker track matching the standalone muon, the global fit with the smallest χ^2 is chosen to define a global muon [33].

For this analysis, the following selection criteria are applied to reconstructed muon candidates:

- muon must be reconstructed globally and in in the tracker
- $p_t > 18\,\text{GeV}$. p_t criteria are introduced as particles need a certain momentum in order to be well reconstructed. This *particle selection* criterion is looser than the *event selection* criterion for muons as described in section 5.2.
- $|\eta| < 2.1$ is the muon trigger threshold corresponding to the detector coverage.

- $|d_{xy}| < 0.2\,\text{cm}$ distance from the primary vertex. This rejects cosmic muons and muons originating from pile up as they have a different primary vertex.
- global track fit $\frac{\chi^2}{ndf} < 10$. This and the following criteria are applied for a well reconstructed trajectory.This *particle selection* criterion is looser than the *event selection* criterion for muons as described in section 5.2.
- number of hits in the tracker > 10.
- number of hits in the pixel tracker > 0. This criterion rejects muons originating from the decay in flight of another particle.
- two or more matched segments in the muon chambers with at least one hit compatible with the reconstruction refit as described above.
- In a cone of $\Delta R < 0.3$ the particle energies excluding the muon track must not exceed 3 GeV. This tracker isolation criterion rejects non prompt muons. For instance these can be produced from hadrons decaying in flight.

5.3.2. Electron

The electron reconstruction uses the tracker and ECAL information. Electrons shower inside the ECAL and deploy energy in a number of crystals, e.g. a typical 120 GeV electron deposits 97% of its energy in a range of 5×5 crystals. These crystals are combined to a cluster. For this, an algorithm considering fixed size crystal arrays is used as it shows the best performance.

While traversing the tracker material, electrons are subject to bremsstrahlung. The radiated photons are distributed in ϕ in the ECAL due to the magnetic field. They have to be taken into account in order to determine the correct electron energy. This is not a small effect: 35% of the electrons lose more than 70% of their energy due to bremsstrahlung [34]. To accumulate all the energy, the clusters generated from an electron and its photons are combined to a supercluster. Starting from a local maximum, a seed cluster, other clusters nearby are selected, which fulfil certain quality criteria [35]. The energy is corrected depending on the number of hit crystals.

In a second step, the electron trajectory inside the tracker is reconstructed. Starting from the ECAL supercluster, the algorithm searches a track seed made of one pixel detector hits. For this, the possible trajectory is calculated, starting from the ECAL supercluster using the determined energy and considering the cases of e^+ and e^-. From the seed in the first pixel layers, a track is reconstructed to the outside. A Gaussian Sum Filter is used to model the nonlinear energy loss due to bremsstrahlung in the tracker. As a criterion for the goodness of the fit, a χ^2 test is performed and at least five tracker hits are required [34].

For this analysis the following selection criteria are applied:

- $p_t > 25\,\text{GeV}$ as soft electrons are more likely to be misidentified. This *particle selection* criterion is looser than the *event selection* criterion for electrons as described in section 5.2.
- $|\eta| < 2.5$ which is the range covered by the tracker.
- $\Delta\phi < 0.09$ between the extrapolated tracker trajectory and the ECAL supercluster at the interaction vertex to assure a good matching between these two parts of the reconstruction.
- $H/E < 0.05$, being the ratio of hadron calorimeter energy and the electromagnetic calorimeter energy in a cone of radius 0.15 around the electron position in the calorimeter. Too much hadronic activity is evidence for a jet.

For electrons in the barrel region in addition the following applies [36]:

- $\Delta\eta < 0.05$ between the tracker trajectory and the ECAL supercluster at the interaction vertex.
- $E_{2\times5}/E_{5\times5} > 0.94$ or $E_{1\times5}/E_{5\times5} > 0.83$. The fraction of the energy deposited in 2×5 crystals divided by the energy in the total cluster is a measure for the spreading of the energy. As the bremsstrahlung photons are distributed in ϕ due to the magnetic field, the deposited energy should be localised in η but spread in ϕ.
- $E + H1$ isolation $< 2\,\text{GeV} + 0.03 \cdot E_T$. This is the isolation in the ECAL and in the first HCAL layer. E is the electromagnetic energy in a cone of radius 0.3 around the position of the electron excluding an inner cone of radius 3 crystals (to exclude the electron) and an eta strip of total width of 3 crystals (to exclude electron bremsstrahlung). $H1$ is the HCAL energy at transverse depth 1 in a cone of radius 0.3 around the position of the electron, excluding a radius of 0.15. This criterion rejects electrons that are part of a jet.
- Tracker isolation $< 7.5\,\text{GeV}$ also to reject non-prompt electrons.
- $E_2/E_9 < 0.9$ to exclude spikes which are due to an interaction in the read out instead of the active material. E_2 is the energy of the crystal hit by the extrapolated electron track plus the energy of second nearest crystal.

And in the endcaps, we have

- $\Delta\eta < 0.07$. This criterion is loosened as the alignment is not understood as well as in the barrel.
- $\sigma_{i\eta i\eta} < 0.03$ the energy spreading in η measured in crystals in the 5×5 block around the seed. This is similar to the $E_{2\times5}/E_{5\times5}$ criterion.
- $E + H1$ isolation $< 2.5\,\text{GeV} + 0.03 \cdot (E_T - 50\,\text{GeV})$
- H2 isolation $< 0.5\,\text{GeV}$. This criterion on the isolation in the second layer of the HCAL rejects high energetic jets.
- Tracker isolation $< 15\,\text{GeV}$. The isolation selection is loosened as there is more deposited energy in the forward region.

5.3.3. Photon

Like electrons, photons deposit their energy in the ECAL. The same cluster algorithm as for electrons (section 5.3.2), using a fixed array of 5×5 crystals, is used to reconstruct photons. The energy is slightly corrected, e.g. for crystal gaps.

Because of the amount of material in the tracker, there is a fair chance ($\mathcal{O}(10\%)$) for a photon to convert into an electron-antielectron pair inside the detector. The pair will be boosted in direction of the original photon but will split up in ϕ due to the magnetic field. Therefore the supercluster algorithms used for electron reconstruction are also used to reconstruct such photons. The energy scale for converted photons is corrected. The energy resolution is slightly worse than for unconverted ones [37].

Only photons with the following properties are selected:

- $p_t > 25\,\text{GeV}$ to select well reconstructable photons.
- $|\eta| < 1.442$ this is the barrel region of the ECAL.

- $H/E < 0.05$ The fraction of the energy deposited in the HCAL must not be greater than 5% of the ECAL energy.
- No pixel seed in the tracker as this would be evidence for a charged particle.
- Track isolation $< 3.5\,\text{GeV} + 0.001 \cdot p_t$. This and the following isolations reject non-prompt photons.
- ECAL isolation $< 4.2\,\text{GeV} + 0.006 \cdot p_t$.
- HCAL isolation $< 2.2\,\text{GeV} + 0.0025 \cdot p_t$.
- $E_2/E_9 < 0.9$ to reject spikes.

5.3.4. Jet

For jet reconstruction, the *Particle Flow anti-k_T algorithm* is used, which is *collinear safe* and *infrared safe* [38, 39]. That means, it is stable in the case that the energy splits up onto two nearby collinear entities instead of one, and stable in the case of adding a low energetic particle (e.g. from a radiated gluon).

The anti-k_T algorithm looks for the smallest distance d. The distance between two entities (i.e. a particle or a pseudojet) i and j is defined by

$$d_{ij} = min(k_{T,i}^{-2}, k_{T,j}^{-2}) \cdot \frac{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}{R^2}, \tag{5.1}$$

and the distance between one entity and the beam is given by

$$d_{iB} = k_{T,i}^{-2}. \tag{5.2}$$

Here, η is the pseudorapidity, ϕ is the azimuth angle, and $k_{T,i}$ is the transverse momentum of the entity i. For the cone radius parameter R, MUSiC uses the value $R = 0.5$. The two entities having the smallest distance d_{ij} are combined to a single pseudojet by adding up the four-momenta. When d_{iB} is the minimal distance inside the system, the corresponding pseudojet i is promoted to a jet and removed from the set. This procedure is repeated until no particle is left.

The idea of Particle Flow is to identify all particles from an event by combining all detector elements [40]. Particle Flow has an inherent cleaning procedure, therefore the particle identification of a certain type is not independent of the identification of other types. The concept of Particle Flow requires both a high efficiency and a low fake rate. For the track finder algorithm, this is achieved by starting with relatively tight constraints, removing the corresponding tracker hits and loosening the constraints for the next iteration.

The Particle Flow jets are not built from simple calorimeter entries, but from applying the anti-k_T algorithm to reconstructed particle candidates, each of which can be a hadron, electron, muon or photon [41]. A jet energy scale factor is applied to correct the energy depending on η and p_t. It should be mentioned that this correction is in the order of $< 10\%$, whereas calorimeter jets have to be corrected for about 20% to 50% [42].

The used, so called "loose", Particle Flow jet identification requires:

- $p_t > 50\,\text{GeV}$. In this regime the jet energy scale is most reliable.
- $|\eta| < 2.5$ is the regime of the tracker and HCAL.
- Neutral hadron fraction < 0.99. Removes jets from HCAL noise. This and the next selection criterion requires the jet to have a charged constituent in both, ECAL and HCAL.

- Neutral electro-magnetic fraction < 0.99. Removes photons and and ECAL noise.
- Number of constituents > 1 to exclude single particles.

Additionally, in the forward region $|\eta| > 2.4$ it asks for:

- Charged hadron fraction > 0 to have charged hadrons in the jet. Removes HCAL and ECAL noise as well as cosmic rays.
- Charged multiplicity > 0 to have charged constituents in the jet.
- Charged electro-magnetic fraction < 0.99.

The charged hadron energy fraction is determined by matching tracks to the HCAL energy depositions. The fraction of the jet's matched depositions divided by the jet's total energy in ECAL and HCAL is the charged hadron fraction. To determine the neutral energy fraction the remaining energy inside the jet is used. The electro-magnetic fractions are determined analogously.

5.3.5. Missing Transverse Energy

The Particle Flow E_t^{miss} is determined by adding up the negative vectorial sum of the transverse energies of all particle candidates. This corresponds to the missing transverse energy [42].

The selection criteria are:

- $E_t^{\text{miss}} > 30\,\text{GeV}$. Small values of E_t^{miss} are neglected as they are due to the insufficient detector resolution.
- The ϕ-distance between the direction of the missing transverse energy and each electron $\Delta\phi > 0.1$ to reject E_t^{miss} caused by an electron which interacts with an uninstrumented part of the calorimeter.

5.4. Event Cleaning

Sometimes a misidentification of a particle leads to the reconstruction of two particles. For instance, an electron could be reconstructed as a photon as well. A cleanup procedure is used, if a particle is not isolated, i.e. a second particle is found inside a cone with $\Delta R < 0.2$. Particles are removed in the following order:

- Jets are removed near photons and electrons.
- Photons are removed near photons and electrons.
- Electrons are removed near electrons.
- Muons are removed near muons.

5.5. Results

By requiring the above described criteria, 827000 data events are selected. They are sorted into 88 exclusive and 100 inclusive classes (as described in section 6.4). The highest jet multiplicity can be found in the $1\mu + 8$jets class. The largest number of leptons is observed in the 4μ class.

6. The MUSiC Analysis

The search for new physics is most often performed by model specific analyses. Each of these compares the outcome of the experiment with the prediction of a theoretical model. The model should not yet be excluded, should be well motivated and testable on the data. Typical examples, which are also examined by the CMS collaboration, are supersymmetric models (SUSY), heavy gauge bosons as described in section 2.2.2, unparticles and large extra dimensions.

An alternative to these specific analyses is the model unspecific or model independent approach. One should be aware of the fact that a total model independence can never be achieved and can therefore not be the goal of an analysis. Instead, this approach considers the currently best motivated and tested model, the Standard Model of particle physics (see chapter 2) and tests the data against it. It is therefore minimally biased towards any model of new physics and should be sensitive to a variety of different signatures that might appear due to new physics.

The model independent analysis in collider physics has some tradition already. An analysis at the L3 experiment has been performed [43], as well as at H1 [44], CDF [45] and DØ [46]. At CMS, it's the first time that a model independent analysis monitors the experiment right from the beginning. A lot of effort has been made to establish, study and perform the MUSiC analysis [47, 19, 48, 49].

6.1. Concept

The standard way of comparing theory and experiment at collider experiments is by comparing certain properties of data events and Monte Carlo simulated events. We follow this approach with the Model Unspecific Search in CMS [50].

The idea of a model independent search opens a great field of possibilities. The following guidelines lead the way of the analysis [19].

- The analysis should be minimally biased towards certain models of new physics. This is achieved by minimal selection criteria and the consideration of all possible final states.
- MUSiC considers well understood physics objects. A central η range provides for a well understood detector and concentrates on hard interaction particles, which is well described by the Standard Model Monte Carlo. Event classes are built from well reconstructable physics objects only.
- Keeping the analysis and the statistical methods straightforward is important, as a positive analysis result has to be understandable.
- The test should be powerful in terms of being able to discover possible differences between the Standard Model and the experimental data.

6.2. Work Flow

The MUSiC software mainly consists of three applications, a flow chart is depicted in figure 6.1.

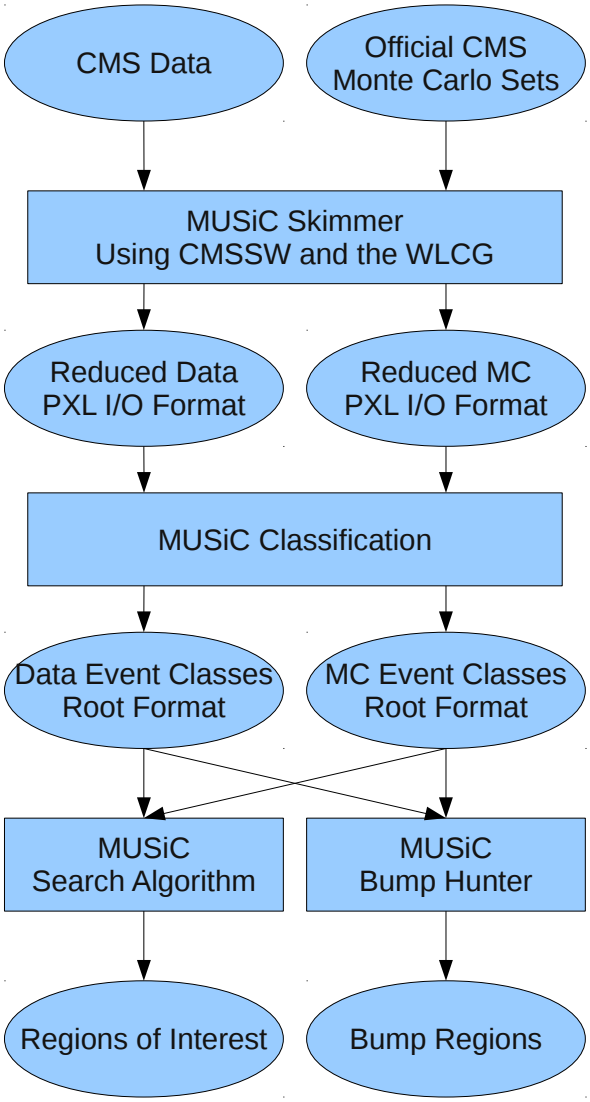


Figure 6.1.: Work flow of the MUSiC software package.

The Skimmer is a CMSSW module which can be run on the WLCG. It preselects the data and Monte Carlo by applying some lose cuts to it (chapter 5) and stores the necessary information in PXLIO files (section 4.4).

The second step of the processing is the classification. MUSiC sorts the events into classes depending on their physical content. This process can also be parallelised and usually runs on the local condor cluster. In this step, also a number of control plots is generated to validate the process.

In the last step, the distributions of the event classes are evaluated. Originally, this happens by applying the MUSiC search algorithm, called Region of Interest algorithm. This is also the point, where the Bump Hunter developed in the present thesis fits in.

The following three sections describe each step in detail.

6.3. Data Selection and Information Reduction

As a first step, we select those events and particles[1] which fulfill certain criteria as described in chapter 5. Also more detailed detector information is dropped from the data file if it is not needed for this analysis.

The physics objects considered by MUSiC are:

- electrons (e),
- muons (μ),
- photons (γ),
- jets (jet) and
- missing transverse energy (E_t^{miss}).

MUSiC is capable of distinguishing between jets from b-hadrons and non-b jets [49]. It can also separate leptons by their charge in two different categories. These two features are not used in this analysis.

6.4. Classification

The data events are sorted into event classes depending on their physics content. There are exclusive as well as inclusive classes. Exclusive classes only include those events with a distinct physics content after the selection. Inclusive classes include all events which have at least a certain physics content. An example is sketched in figure 6.2: The event is part of the class $1\mu, 1e, 2$jets and several inclusive classes, denoted with the suffix "+X". The number of event classes is not predetermined but results from all possible classes which can be built from Monte Carlo or data events.

The Monte Carlo events are scaled using the measured luminosity $\mathscr{L}$ and cross section of the physical process σ to:

$$b_{\text{total}} = \sigma \int \mathscr{L}\, dt. \tag{6.1}$$

Often this cross section is determined using higher order corrections than are used in the actual simulation process.

We introduce a scale factor α describing the ratio of the actual number of produced events N_{SM}^{total} and the expected number of events b_{total} as described by equation 6.1:

$$\alpha = \frac{N_{SM}^{\text{total}}}{b_{\text{total}}}. \tag{6.2}$$

This means, the number of Monte Carlo events must be divided by α to scale it.

To get a better estimation in the case of small expected event numbers, as a rule of thumb Monte Carlo simulation should produce at least a factor 10 times the number of expected events. Unfortunately this is not always possible and as a result, some samples of certain physics processes are scaled up instead of down, due to a lack of Monte Carlo statistics.

For each event class, up to three distributions are investigated further. To calculate these in the case of inclusive classes, only those particles included in the name of the event class are used. If there are more particles of a certain kind, only the ones with the highest momenta are taken into account.

[1] More precise would be the word "physics content", as a jet and the missing transverse energy are also treated as if they were particles.

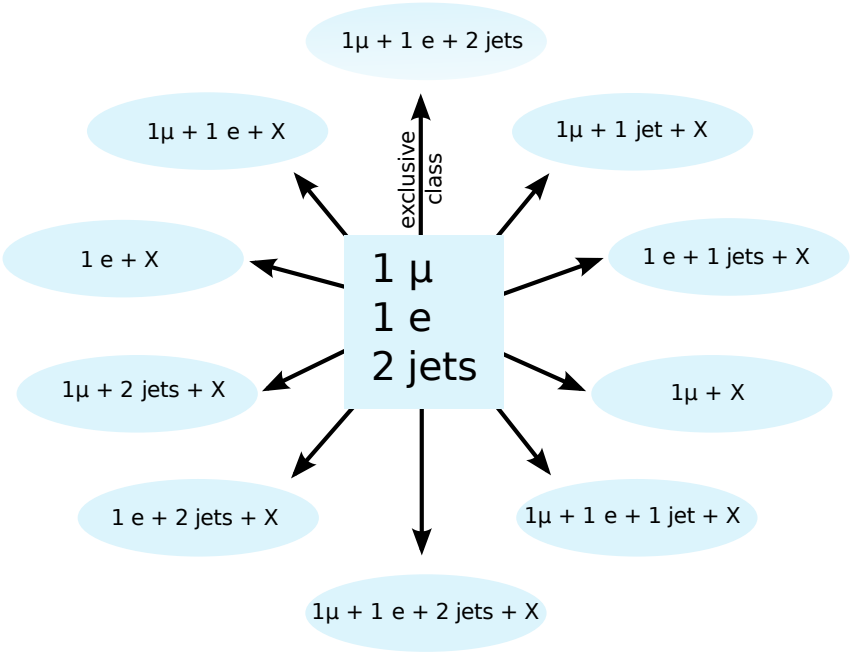


Figure 6.2.: The event $1\mu, 1e, 2$jets is sorted into one exclusive and several inclusive event classes.

- The sum of all scalar momenta (Σp_t). This is the most general distribution which can be determined for all classes.
- The invariant mass (M_inv). This distribution is sensitive to unstable heavy particles appearing as excess at their respective mass. It can only be calculated for classes described by at least two particles.

Event classes that include missing transverse energy in their physics content are treated slightly different. In that case, the investigated distributions are:

- The sum of all scalar momenta (Σp_t), also including the missing transverse energy as one momentum.
- The transverse invariant mass (M^T_inv), as a measure for the total invariant mass.
- The missing transverse energy (E_t^miss). Stable minimum interacting particles could be found as an excess in this distribution.

To limit the maximum number of bins, these distributions are binned with a bin width of 10 GeV. In regions, where the detector resolution is worse than 10 GeV, bins are merged to reflect the resolution. This can only be an estimate, as properties of different particle types, measured with different precisions, determine the observables. For Σp_t we assume that the momenta are equally distributed among the event content entities. In the following we therefore assume that each constituent has a transverse momentum of $\frac{\sum p_t}{N}$.

Different CMS studies have determined the p_t resolution for the considered particles: For muons, we fit a function of the form $A \cdot e^{B \cdot x}$ to the muon endcap resolution [33]. This yields:

$$\sigma_{p_t}(\mu)/\mathrm{GeV} = 0.016 \cdot \frac{\sum p_t/N}{\mathrm{GeV}} + 1.5 \cdot 10^{-4} \cdot \left(\frac{\sum p_t/N}{\mathrm{GeV}}\right)^2 . \tag{6.3}$$

For electrons and photons we use a fit for the golden electron resolution [34].:

$$\sigma_{p_t}(e)/\text{GeV} = 3.88 \cdot 10^{-2} + 6.25 \cdot 10^{-4} \cdot \frac{\sum p_t/N}{\text{GeV}} + 2.704 \cdot 10^{-5} \cdot \left(\frac{\sum p_t/N}{\text{GeV}}\right)^2. \tag{6.4}$$

$$\sigma_{p_t}(\gamma)/\text{GeV} = \sigma_{p_t}(e)/\text{GeV} \tag{6.5}$$

High energy jets can be resolved with [51]:

$$\sigma_{p_t}(\text{jet})/\text{GeV} = \sqrt{\frac{\sum p_t/N}{\text{GeV}} \cdot \sqrt{1.05 \cdot \left(\frac{\sum p_t/N}{1.05\,\text{GeV}}\right)^{-2} + 0.218 \cdot \left(\frac{\sum p_t/N}{\text{GeV}}\right)^{-0.81}}}. \tag{6.6}$$

this formula fails for small energies. So if it gives a resolution of smaller than 10 GeV, a standard 10 GeV resolution is used instead. The missing transverse energy resolution is determined by a linear fit to data from [52], which yields:

$$\sigma_{p_t}(E_t^{\text{miss}})/\text{GeV} = 0.567 \cdot \sqrt{\Sigma p_\text{t}/\text{GeV}}. \tag{6.7}$$

This can only be an estimate as it does not yet consider the types of particles reconstructed in the event. The total p_t-resolution can then be determined by

$$\sigma_{p_t} = \sqrt{\sum_{\text{constituents}} (\sigma_{p_t}(\text{constituent}))^2}. \tag{6.8}$$

For the $1\mu, 1e, 2$jets class this yields

$$\sigma_{p_t} = \sqrt{(\sigma_{p_t}(\mu))^2 + (\sigma_{p_t}(e))^2 + 2 \cdot (\sigma_{p_t}(\text{jet}))^2}. \tag{6.9}$$

For the resolution of M_inv, we use the same as for Σp_t, assuming that $\sum p_t \approx M_\text{inv}$. This is true for the case of maximal M_inv, i.e. all particles are distributed in the $\eta = 0$ plane and are not boosted. So for muons, the resolution can be determined by

$$\sigma_{M_\text{inv}}(\mu)/\text{GeV} = 0.016 \cdot \frac{M_\text{inv}/N}{\text{GeV}} + 1.5 \cdot 10^{-4} \cdot \left(\frac{M_\text{inv}/N}{\text{GeV}}\right)^2. \tag{6.10}$$

and for the other constituents equations 6.4 to 6.7 can be applied analogously. The total resolution can be determined by:

$$\sigma_{M_\text{inv}} = \sqrt{\sum_{\text{constituents}} (\sigma_{M_\text{inv}}(\text{constituent}))^2}. \tag{6.11}$$

The resolution of the missing transverse energy distribution depends on the resolution of the individual particles. A single bin in the E_t^{miss}-distribution can be composed of events with different event topologies corresponding to different resolutions, i.e. the resolution can only be estimated. Because of $\Sigma p_\text{t} \geq E_t^{\text{miss}}$, it is conservative (i.e. it results in a larger number of bins) to use

$$\sigma_{p_t}(E_t^{\text{miss}})/\text{GeV} = 0.567 \cdot \sqrt{E_t^{\text{miss}}/\text{GeV}}. \tag{6.12}$$

6.5. Region of Interest Algorithm

The Region of Interest Algorithm is an algorithm used to locate the most significant deviations and give a measure on how significant they are. A different approach, the Bump Hunter, can be found in chapter 7.

6.5.1. Step 1: Determining the Minimal P-Value

The resulting distributions from the classification process are scanned for deviations between data and Monte Carlo. Therefore, a region as depicted in figure 6.4 is compiled by adding up all bins in this region. The following procedure is applied to all possible regions of adjacent bins of all distributions. For each region the deviation between data and Monte Carlo is determined. One of the challenges of a model independent analysis is the measure of discrepancy. P-values provide such a measure, depending on the number of events, its expectation and the modelled uncertainties.

A p-value can be defined as:

> "The p-value is the probability, calculated assuming H_0 is true, of obtaining a test statistic value at least as contradictory to H_0 as the value that actually resulted.
>
> The smaller the P-value, the more contradictory is the data to H_0" [53]

In our case, the hypothesis H_0 is the agreement of the data with the Monte Carlo simulation. As test statistic we use the number of events in a certain region of the phase space.

Both, the experiment and the Monte Carlo production are Poisson processes. When generating events, the number of events in two time intervals are independent, the number of events only depends on the measuring duration, and two events do not happen simultaneously. Therefore, the probability P to observe a certain number of events n in a bin region is Poisson distributed with an unknown expectation value λ for a large sample size:

$$P = \frac{e^{-\lambda}\lambda^n}{n!}. \tag{6.13}$$

And the p-value for the region can be determined by:

$$p = \sum_{i=N_{\text{data}}}^{\infty} \frac{e^{-\lambda}\lambda^i}{i!} \qquad \text{if } \lambda \leq N_{\text{data}} \tag{6.14}$$

$$p = \sum_{i=0}^{N_{\text{data}}} \frac{e^{-\lambda}\lambda^i}{i!} \qquad \text{if } \lambda > N_{\text{data}} \tag{6.15}$$

with N_{data} being the number of observed data events.

The parameter λ is still unknown. When trying to calculate the p-value for the hypothesis "Monte Carlo equals data" N_{SM}/α is a first estimation for λ. In practice, the Monte Carlo simulation is subject to a number of uncertainties. Their influence can be modelled by a so called prior-predictive p-value. The distribution of the uncertainties has to be taken into account, here we assume a Gaussian shape. The probability then reads

$$\mathrm{P} = A \cdot \int_0^{\infty} d\lambda\, e^{-\frac{(\lambda-b)^2}{2\sigma^2}} \cdot \frac{e^{-\lambda}\lambda^N}{N!}, \tag{6.16}$$

where

- $A = \sigma \int_b^{\infty} dt\, e^{-t^2/(2\sigma^2)}$ is a normalisation factor. This is due to the fact that the Gaussian distribution is > 0 for $\lambda < 0$. But as the integration starts at zero this distribution has a lower cut off value of 0.
- b is the estimator for the expectation value of the event count. It is determined by equation 6.2 and

$$b = \sum_{\text{MC processes (i)}} \frac{N_{\text{SM},i}}{\alpha_i}. \tag{6.17}$$

- σ combines the systematic uncertainties. It includes the statistical uncertainties of the Monte Carlo samples:

$$\sigma = \sqrt{\sum \sigma^2_{\text{syst.}} + \sum \sigma^2_{\text{stat,MC}}}. \tag{6.18}$$

The p-value can be calculated as the sum of the probabilities (6.16) over all possible event numbers N in the region which are more extreme than the measured data, i.e.:

$$p = \sum_{i=N_{\text{data}}}^{\infty} \int_0^{\infty} d\lambda\, A \cdot e^{-\frac{(\lambda-b)^2}{2\sigma^2}} \cdot \frac{e^{-\lambda}\lambda^i}{i!} \qquad \text{if b} \leq N_{\text{data}} \tag{6.19}$$

$$p = \sum_{i=0}^{N_{\text{data}}} \int_0^{\infty} d\lambda\, A \cdot e^{-\frac{(\lambda-b)^2}{2\sigma^2}} \cdot \frac{e^{-\lambda}\lambda^i}{i!} \qquad \text{if b} > N_{\text{data}}. \tag{6.20}$$

Naturally, one can only *estimate* the systematic uncertainties. To take a Gaussian as the underlying distribution is an assumption which is often made, but it also has its drawbacks: It implies that the uncertainties are known absolutely and not relatively in respect to the Monte Carlo event count. This is a valid assumption for some uncertainties but might be difficult for others, e.g. the luminosity usually has a relative uncertainty whereas the uncertainty of the jet energy scale is primarily known absolutely. The problem can be seen intuitively when the number of Monte Carlo events is small compared to the total systematic uncertainty: As the Gaussian has a cut off at zero, it is then assumed to be more likely that the count is underestimated than that it is overestimated. A different approach to that problem is to assume fixed relative errors which results in a log-normal distribution. A detailed study can be found in [48].

The p-value calculation is repeated for all regions of the distribution. The region with the smallest p-value is declared the Region of Interest, an example is displayed in figure 6.4.

When there is no Monte Carlo event in one bin for a specific process, the systematic uncertainties are null as well[2]. Often, this is an underestimation of the uncertainty. As an example one can consider 10 bins, each having an expectation value of 1/10 events and a Monte Carlo with a scaling factor $\alpha = 1$ correctly describing the data. Commonly, one event will be found in Monte Carlo and data each, but not necessarily in the same bin. In that case, the filled data bin is not described by a filled Monte Carlo bin. This leads to a p-value of 0.

This behaviour is an effect of insufficient Monte Carlo statistics and the uncertainties should account for it. Therefore MUSiC uses an *uncertainty fill-up* method. Starting from each filled bin at energy[3] x, the number of empty bins up to the bin at position $2 \cdot x$ is counted. The uncertainty for all n empty bins is set to

$$\sigma = \frac{1}{\sqrt{n}\alpha}. \tag{6.21}$$

This corresponds to the Monte Carlo statistics uncertainty if the bin would be filled with $\frac{1}{n\alpha}$ events. The factor α is the scaling factor described in equation 6.2.

An example of the fill-up procedure is depicted in figure 6.3. The blue sample and the red sample each have one bin filled with one unscaled event. They have different scaling factors α. The shaded regions represent the uncertainty fill-up only, other uncertainties are not depicted.

The red sample has one entry at 50 GeV. Therefore the uncertainties fill-up is applied up to 100 GeV. As the fill-up procedure is performed for each sample individually, the 80 GeV bin is filled up although it has an entry of the blue sample.

[2] This is obvious for relative uncertainties. As later described in section 6.6 all considered uncertainties are relative except for the jet energy scale uncertainty and the parton distribution function uncertainty.

[3] This means, Σp_{t}, M_{inv} or E_t^{miss}.

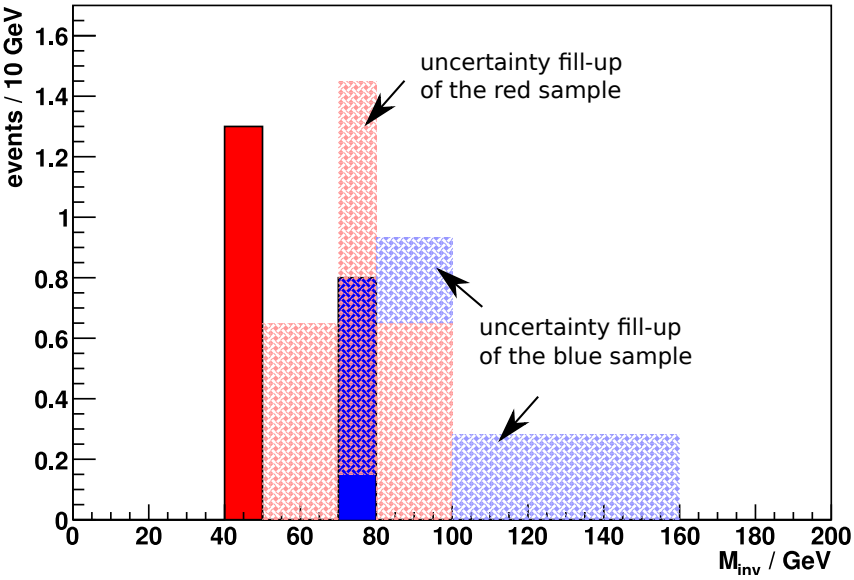


Figure 6.3.: Scheme of an uncertainty fill-up scenario. A blue sample and a red sample with one event each but different weights are shown. The depicted uncertainties (shaded) are only the fill-up uncertainties. For reasons of simplicity, all other uncertainties are not shown in this very figure. The red sample at 50 GeV affects all bins up to $2 \cdot 50\,\text{GeV} = 100\,\text{GeV}$. Analogously, the blue sample affects all bin up to 160 GeV.

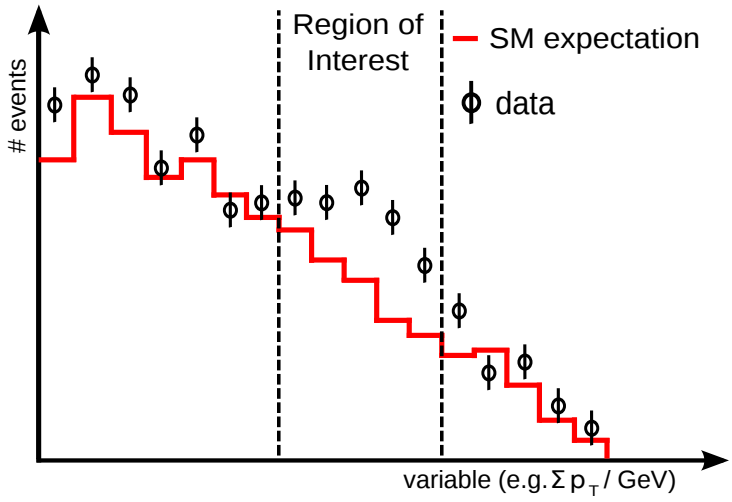


Figure 6.4.: Scheme of a Region of Interest as found by the MUSiC algorithm.

Analogously, the fill-up using the blue sample is carried out. In the 90 GeV and 100 GeV bin the fill-up is performed for both samples, therefore the uncertainties are added up[4]. The fill-up of the blue samples continues up to 160 GeV.

A study about how well the fill-up method performs can be found in chapter 8.

[4]Their squares are added up and the square root is applied to the sum.

6.5.2. Step 2: The Look Elsewhere Effect

Assume an experiment with a chance of 1% to reject the null hypothesis even though it's true. If one carries out such an experiments and it has a positive result (i.e. the hypothesis is in agreement with the data) one can claim with 99% confidence level that the hypothesis is true. But if one conducts 100 of such experiments, chances that at least one of these experiments has a positive result just by chance, rise to

$$\sum_{k=1}^{100} \binom{100}{k} \cdot 0.01^k \cdot (1-0.01)^{100-k} \approx 99.5\%. \tag{6.22}$$

Of course this is a statistical effect which has to be accounted for and which we refer to as the "look elsewhere effect".

When scanning one distribution of one class, a lot of different regions are examined. If N_{bins} is the number of bins, $N_{\text{regions}} = \frac{N_{\text{bins}} \cdot (N_{\text{bins}}+1)}{2}$ is the number of regions. As stated above, this has to be accounted for. Obviously, those regions are highly correlated as there is a lot of overlap in the bins. This fact makes an analytical calculation of the impact of the look elsewhere effect difficult.

Instead, to find out how big the probability to get a certain deviation by pure statistical fluctuations is, one can conduct pseudo experiments. Based on the Standard Model expectation obtained from Monte Carlo simulations we can use the assumed probability (6.16) to dice hypothetical experiment outcomes. For each of these, a minimal p-value p_{min} can be determined (as done in the first step for the data). This gives an estimate of how probable it is to have a higher significance (i.e. an even smaller p-value than p_{min}) than seen in the data, if our hypothesis is true. This leads us to a modified p-value:

$$\tilde{p} = \frac{\text{number of pseudo experiments with } p_{\text{min}}^{\text{pseudo}} < p_{\text{min}}^{\text{data}}}{\text{total number of pseudo experiments}}. \tag{6.23}$$

A typical distribution of $p_{\text{min}}^{\text{pseudo}}$ is pictured in figure 6.5. The fraction of the shaded part of the histogram gives $\tilde{p}$.

6.5.3. Step 3: $\tilde{p}$-Distribution

As MUSiC does not only examine one distribution but up to three distributions in a lot of different classes, one has to keep the look elsewhere effect in mind. Unfortunately it is difficult to determine an overall significance which takes this into account. One solution is to compare the distribution of data $\tilde{p}$-values with the $\tilde{p}$-distribution derived from the Monte Carlo expectation. Obvious differences can be easily spotted.

6.6. Uncertainties

The determination of the systematic uncertainties (σ in equation 6.16) is one of the challenges of the MUSiC analysis. As visible in the equation, all uncertainties are assumed to be normally distributed and the resulting distribution is cut off at 0. The contributions to σ are combined as described in equation 6.18. For the 2010 data, the following uncertainties have been taken into account [1]:

- The *integrated luminosity* is measured with an accuracy of 4% [54]. It is fully correlated between all bins, i.e. when dicing a pseudo experiment, all bins are scaled with the same diced integrated luminosity value.

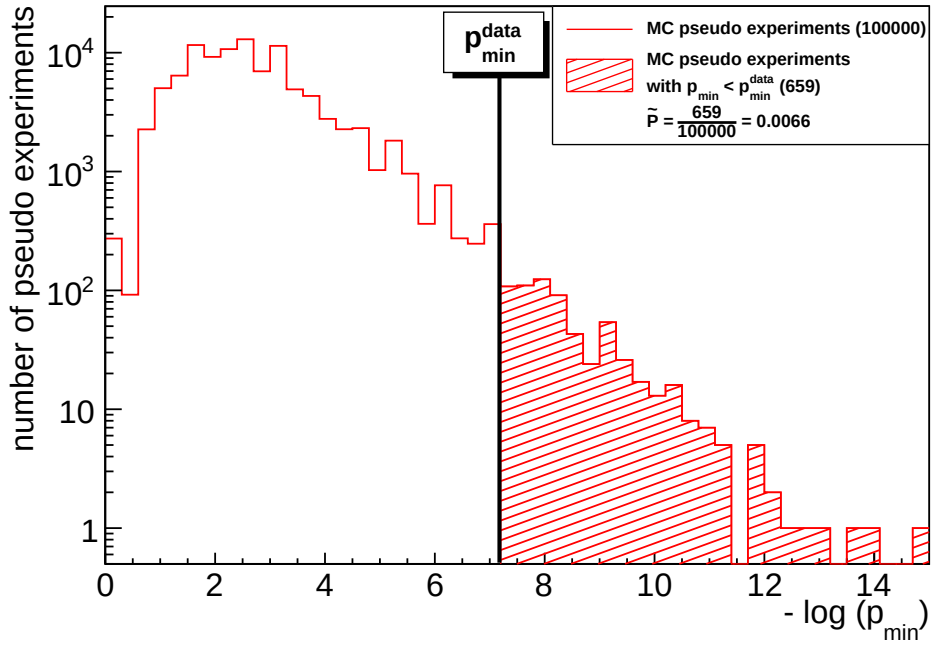


Figure 6.5.: A typical distribution of $p_{\text{min}}^{\text{pseudo}}$. The value of $p_{\text{min}}^{\text{data}}$ yields a $\tilde{p}$ of 0.0066.

Process	Cross Section Uncertainty
QCD	50%
$t\bar{t}$	10%
Drell-Yan $m_{l\bar{l}} > 50\,\text{GeV}$	5%
Drell-Yan $10\,\text{GeV} < m_{l\bar{l}} < 50\,\text{GeV}$	10%
Drell-Yan $m_{l\bar{l}} < 10\,\text{GeV}$	20%
W	5%
γ	50%
Di-Vector-Boson (WW, WZ, ZZ)	10%
Y(1S), Y(2S), Y(3S)	30%

Table 6.1.: Assumed uncertainties on cross sections.

- Total *Cross section* uncertainties depend on how well they are calculated or measured. Details can be found in table 6.1. Again, cross section uncertainties are fully correlated between the bins. The uncertainties of the parton distribution functions are handled separately.

- The *jet energy scale* (JES) describes the scaling between measured energy and the actual jet energy depending on p_t and η. Its effect on the number of events in a bin is not intuitive: As the energy of the jet is changed, events migrate between bins and sometimes also betweens classes, the latter being due to the selection of "good" particles. To account

Object	Efficiency	Fake Probability
Muons	4%	50%
Electrons	3%	100%
Photons	1%	30%
Jets	1%	0

Table 6.2.: Assumed uncertainties on reconstruction efficiency and fake probability [1].

for this, the JES is changed up and down once by about 3%-5%, depending on p_t and η, and the classification is rerun. The differences between the up/down scaled distributions and the standard distribution in each bin are symmetrised and the result is used as the Gaussian σ parameter.

- Under certain circumstances, a particle is not correctly reconstructed. The fraction of the correctly reconstructed particles is called efficiency. In other cases a particle is wrongly identified. E.g. a jet is identified as an electron. The number of wrongly identified particles to the total number of that particle type is the fake probability. Reconstruction *efficiency* and *fake probability* are derived from Monte Carlo and we assume them to :-only depend on the type of the particle. Their uncertainties, listed in table 6.2, are taken from earlier CMS studies.

- The limited *Monte Carlo statistics* are taken into account. The Poisson standard deviation $\sqrt{N_{SM}}/\alpha$ is used as a contribution to the uncertainty as in equation 6.18.

- *Parton distribution functions* (PDF) describe the probability of a certain parton with flavour f at the factorization scale Q to carry the fraction x of the longitudinal momentum at interaction. Therefore, the cross section depend on the PDF of the proton. The straight forward method to determine uncertainty of the number of events would be to generate Monte Carlo samples with varying PDFs. For a model independent analysis this is not feasible due to the large number of Monte Carlo samples. This especially applies as the used CTEQ 6.1 PDF fit [55, 56] includes uncertainties with 20 degrees of freedom, which results in 40 PDF sets (PDF^j) plus the best fit PDF set (PDF^0).

  Instead, we use the *reweighting method*. For each event and PDF set a weight w is calculated considering the two interacting protons:

$$w^j = \frac{\text{PDF}^j(x_1, f_1, Q)}{\text{PDF}^0(x_1, f_1, Q)} \cdot \frac{\text{PDF}^j(x_2, f_2, Q)}{\text{PDF}^0(x_2, f_2, Q)} \tag{6.24}$$

  For each j, the bin content X_j is determined by weighting each event with w^j. The uncertainty for that bin is then determined by:

$$\Delta X^+ = \sqrt{\sum_{i=1}^{40} [\max(X_i - X_0, 0)]^2} \tag{6.25}$$

$$\Delta X^- = \sqrt{\sum_{i=1}^{40} [\max(X_0 - X_i, 0)]^2} \tag{6.26}$$

  The greater of ΔX^+ and ΔX^- is taken as the PDF uncertainty for that bin.

7. Bump Hunter

MUSiC is a model independent analysis but it still relies on the Standard Model of particle physics. Another model independent approach to look for anomalies in data does not rely on Monte Carlo simulations but is completely data driven: By identifying the shape of a given distribution, bumps can be detected, which might be interesting in terms of new physics.

In this chapter, the concept of a Bump Hunter algorithm is developed. After discussing the algorithm, the Bump Hunter is tested on a number of benchmark scenarios. Finally first results using the 2010 CMS data are shown.

7.1. Concept

Figure 7.1 shows the invariant mass spectrum of $\mu^+\mu^-$ as seen by the CMS experiment. A smooth shape interrupted by several peaks is depicted. Each peak energy corresponds to the invariant mass of one type of particle which decays into two muons. Physics beyond the Standard Model is often assumed to have similar signatures at higher energies. A number of models predict new unstable particles. Some of them are discussed in section 2.2: new heavy gauge bosons, leptoquarks, and excited leptons are all expected to have similar signatures.

The idea of a Bump Hunter is to find such peaks by taking only the data distribution into account. The only assumed hypothesis is that of a smooth invariant mass distribution where a new physical phenomenon would show as a bump.

Some design goals of a bump hunting algorithm include:

- The Bump Hunter ought to find a similar set of bumps as a human observerer does.

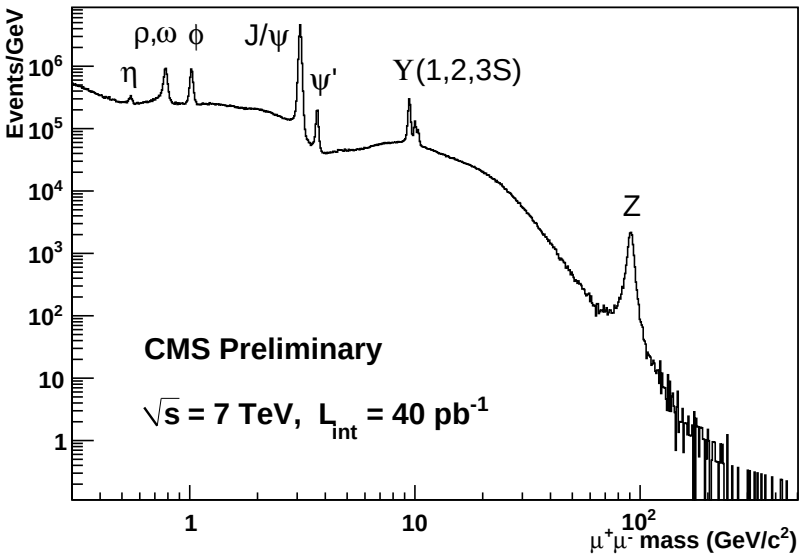


Figure 7.1.: The opposite sign dimuon mass spectrum from the first $40\,\mathrm{pb}^{-1}$ at $\sqrt{s} = 7\,\mathrm{TeV}$ [57].

- It should be reasonably stable, i.e. a slight change of the parameters of the algorithm should not affect the set of found bumps by much. Otherwise the found set would be too arbitrary.

- A peak finder is not considered to be sufficiently general, as a bump superpositioned on top of a steeply falling distribution might not show as a peak. Instead, the form of the underlying distribution must be interpolated.

The reader should be aware that the term "bump hunter" is used for a variety of very different algorithms. Sometimes the term bump refers to an accumulation of events in a background free distribution [58]. The model independent search at CDF also uses a bump hunting algorithm. It utilises sidebands as well, but it is not a data driven method, instead it compares the data event count with the Monte Carlo expectation [45]. A more generic idea of a bump hunter is described in [59].

7.2. Algorithm

The Bump Hunter can be applied to any one dimensional distribution. Here, the invariant mass and transverse invariant mass distributions generated by MUSiC, as described in section 6.4, are used.

Like the MUSiC Region of Interest algorithm, the Bump Hunter uses sliding windows which are examined independently. The Bump Hunter uses a window of three adjacent regions. Each region width is increased by one bin for each iteration individually. An example is shown in figure 7.2. The inner region (here: 440 GeV-520 GeV) is the signal region where the algorithm looks for a bump. The two outer regions are the sideband regions which are used to obtain the shape of the background distribution. The shape is interpolated by fitting a function to those regions. In a second step the goodness of the fit is determined for the sideband regions and the signal region in between. The fit should be reasonably good in the sidebands, otherwise the window won't be considered. For the inner region, the opposite holds true: A potential bump shows up as a significant deviation from the fit.

As a first step, the three regions are chosen. Three regions have four bounds and therefore the number of windows rises with $\mathcal{O}([\text{number of bins}]^4)$. A compromise is made in order to reduce the necessary computing time and cope with the huge amount of data. As the distributions are already rebinned to resolution, most constraints can be made in terms of bins. For this analysis, the following constraints are applied:

- The range from the first filled bin up to the last filled bin plus an additional 5 empty bins at the end of the spectrum is considered.

- The minimal inner region width is 1 bin.

- The maximal inner region width is 20 bins.

- The maximal combined sideband region width is 60 bins.

- The minimal single sideband region width is 2 bins.

- Each sideband must have at least the width of the inner region.

- Each sideband must include at least 2 data events.

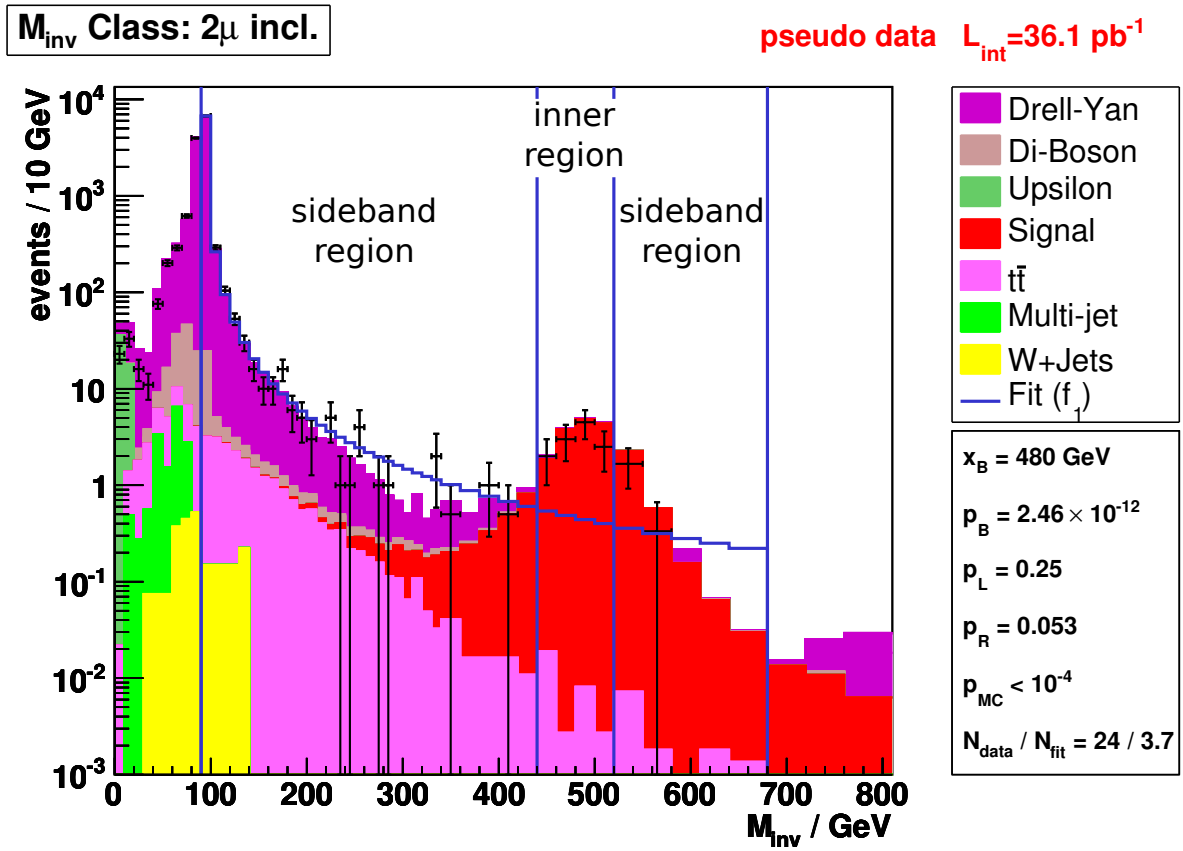


Figure 7.2.: Z' bump in the inclusive $2\mu + X$ class. The bump at around 480 GeV has been found by the Bump Hunter, using the f_2 fit function. The bin 520 GeV-550 GeV has not been assigned to the bump region but to the right sideband as more than 2 events are required in each sideband. The Monte Carlo is not used for the standard Bump Hunter algorithm but only to create test scenarios and after the scan to compare the compatibility of the bump with Monte Carlo.
The filled areas are the stacked Monte Carlo expectations of the corresponding samples. The black markers are given by one pseudo experiment from all Monte Carlo samples. The fit function is drawn in blue. The vertical blue lines denote the borders of the sideband regions and the bump region.
The values in the lower box of the legend refer to the bump region. The median of the bump region (x_B), the p-value of the inner test region (p_B), the p-value of the left sideband region (p_L), the p-value of the right sideband region (p_R), the p-value derived from the Monte Carlo test (p_{MC}), the number of events in the inner region (N_{data}), and the expected number of events in that region (N_{fit}), given from the fit, are denoted.

The motivation for this choice is to have sidebands which are big enough to follow the background distribution. The cut off for the inner region of 20 bins should be big enough to identify localised resonances. The values might have to be changed for future data when more integrated luminosity is available in order to cope with the increasing number of filled bins at high energies.

For the fit the minimization of the negative log likelihood is used. A χ^2 fit would be a bad choice as it fails for low statistics. As the bins are of variable widths, the standard fitting procedure, which uses the fit function value at the bin centre, cannot be applied. Instead, the function is integrated over each bin and compared to the bin content.

Invariant mass distributions are usually rapidly falling, the reason for this is the decrease of the parton distribution function and the reduction of the available phase space in the case of large momentum transfers. Like the invariant mass distributions, the fit function should be a steeply decreasing function. An exponential describes the typical invariant mass distribution well with a small number of degrees of freedom:

$$f_1(x) = P_0 \cdot e^{P_1 \cdot x}. \tag{7.1}$$

A function of the form

$$f_2(x) = \frac{P_0}{(P_1 + x)^{P_3}} \tag{7.2}$$

has been used for fitting similar distributions before [60] and has only three free parameters. It seems well motivated for the purpose of a Bump Hunter. The function falls less rapidly than f_1 for sufficiently high values of x. An example for the f_1 function can be seen in figure 7.3, for the f_2 sample in figure 7.2.

To determine which sidebands provide a good fit and to determine the significance of a possible bump, we need a global measure for the goodness of the fit. The commonly used χ^2 goodness of fit can be translated directly into a p-value for the fit hypothesis. Unfortunately, the χ^2 test assumes a Gaussian shaped distribution of the test statistic, therefore it performs badly in the case of low statistics. It completely fails in the case of an observed bin value of zero [61]. The log likelihood ratio test is also not valid for small expectation values [62].

Instead we use pseudo experiments to determine a p-value as a measure of the goodness of the fit. The integral of the fit function over one bin is used as the expectation value of a Poisson distribution. From this distribution, one pseudo experiments is diced. The Poisson likelihoods of all bins are than multiplied to determine the total likelihood. To be better able to handle the small numbers calculated by this methods, the logarithm of the likelihood is used. By dicing pseudo experiments in each bin of each sideband region (left / right), we obtain a distribution of log likelihood values ($\log L^{\mathrm{L}}_{\mathrm{pseudo}}$ / $\log L^{\mathrm{R}}_{\mathrm{pseudo}}$), each determined by

$$\log L_{\mathrm{pseudo}} = \sum_{\mathrm{bins}} \Big(-b_{\mathrm{fit}} + N_{\mathrm{pseudo}} \cdot \log b_{\mathrm{fit}} - \ln \Gamma(N_{\mathrm{pseudo}} + 1) \Big) \tag{7.3}$$

where N_{pseudo} is the number of events in the bin as determined by one pseudo experiment and b_{fit} is the integral of the fit function in that bin. The likelihood of the data is determined analogously for each sideband region:

$$\log L_{\mathrm{data}} = \sum_{\mathrm{bins}} \left(-b_{\mathrm{fit}} + N_{\mathrm{data}} \cdot \log b_{\mathrm{fit}} - \ln \Gamma(N_{\mathrm{data}} + 1) \right). \tag{7.4}$$

We treat the fraction of pseudo experiments having a smaller likelihood than the data as the p-value for our goodness of fit test. The p-value is calculated for each sideband separately:

$$p_{\mathrm{L,\,R}} = \frac{\text{number of pseudo experiments with } \log L^{\mathrm{L,R}}_{\mathrm{pseudo}} < \log L^{L,R}_{\mathrm{data}}}{\text{total number of pseudo experiments}}. \tag{7.5}$$

For the inner region this procedure is not feasible. Because the p-value should be low in the case of a bump, a large number of pseudo experiments would have to be conducted. Instead, we treat the inner region as a single bin and calculate the p-value by summing over the Poisson distribution:

$$p_{\mathrm{B}} = \sum_{i=N_{\mathrm{data}}}^{\infty} \frac{e^{-b_{\mathrm{fit}}} \cdot b_{\mathrm{fit}}^{i}}{i!} \quad \text{if } N_{\mathrm{data}} \geq b_{\mathrm{fit}} \tag{7.6}$$

$$p_{\mathrm{B}} = \sum_{i=0}^{N_{\mathrm{data}}} \frac{e^{-b_{\mathrm{fit}}} \cdot b_{\mathrm{fit}}^{i}}{i!} \quad \text{if } N_{\mathrm{data}} < b_{\mathrm{fit}} \tag{7.7}$$

with b being the result of the fit in that region. An advantage of this method is that it automatically levels an up-down fluctuation in the data, e.g. two bins in the inner region where one has a deficit and the other has an excess. Such a signature is not looked for.

For a region to qualify as a bump we define the following criteria:

- To ensure a good compliance of the fit function with the sideband data we require $p_{\mathrm{L}} > 0.05$ and $p_{\mathrm{R}} > 0.05$, which approximately corresponds to a Gaussian 2σ deviation limit.
- A bump must be sufficiently significant, therefore we require $p_{\mathrm{B}} < 10^{-7}$, corresponding to about 5 Gaussian standard deviations.
- The bump must be an excess in data relative to the fit. Although models are conceivable that produce downward variations in the invariant mass distribution due to interference terms, they are unlikely to result in localised dips.

When a region is identified as a bump, often other nearby regions, e.g. shifted by one bin, also qualify as a bump region. They essentially describe the same bump. Therefore a cleanup procedure is applied: The bump with the smallest p_B is selected and all overlapping bumps are dismissed. That means that more than one bump can only be found in each distribution if they are not overlapping.

It should be mentioned that the Bump Hunter does not account for the look elsewhere effect. One should be aware of this fact when interpreting a bump.

Comparison with Standard Model Monte Carlo

As an addition to the Bump Hunter, the identified bumps are compared to Monte Carlo in order to identify those bumps that are least compatible with the Standard Model. This is done by dicing pseudo data from the Standard Model Monte Carlo. p_{MC} is then determined by the fraction of pseudo experiments for that $p_{\mathrm{B,pseudo}} \leq p_{\mathrm{B,data}}$ is true. It describes whether in the inner region the fit is more compatible with the Monte Carlo or with the data. A small p_{MC} means that in the inner region, the Monte Carlo complies better with the fit than the data does. And if the fit describes the expectation but not the observed data, this is an indication for a very interesting bump.

We require $p_{\mathrm{MC}} < 0.01$ to pass this test. Nevertheless, this is only an additional information and we also look at those bumps that do not pass this test.

	Mass m / GeV	Decay Channel	Cross Section × BR / pb
ZprimeSSMToMuMu_M500	500	2μ	2.00
ZprimeSSMToEE_M500	500	$2e$	2.00

Table 7.1.: Z' cross sections [63].

7.3. Sensitivity Tests

After introducing this new method for bump hunting, its performance has to be tested. We test whether generic bumps based on simple benchmark scenarios can be found. One should be aware of the fact that the Bump Hunter can never claim the exclusion of a new theory or proof its existence. A possible bump has to be evaluated further.

In this section we evaluate resonances in models of new physics. The Standard Model resonances depicted in figure 7.1 cannot be found by the MUSiC Bump Hunter. The Standard Model Z resonance is not found because it lacks a reasonable left sideband. Also the low mass resonances cannot be detected, as they are mostly cut off by the selection and the minimal bin resolution used by MUSiC is too large to resolve them.

7.3.1. Heavy Neutral Gauge Boson

A Z' is an intuitively appealing test scenario for the Bump Hunter. The Z' is a new neutral gauge boson, which is predicted in many models of new physics, for example superstring inspired theories (section 2.2.2). The Monte Carlo samples of a sequential Standard Model Z' as described in table 7.1 are used. For the background the Standard Model Monte Carlo sample compilation as described in table A.1 in the appendix is used. We produce pseudo data using an integrated luminosity of $36.1\,\mathrm{pb}^{-1}$, which corresponds to the luminosity acquired in the CMS 2010 run.

For the exponential fit function (f_1, equation 7.1) the Bump Hunter finds significant bumps in the classes 2μ, $2e$, $2\mu + X$, and $2e + X$. The $2e$ class is depicted in figure 7.3. Although the Monte Carlo test is not passed, the value of $P_{\mathrm{MC}} = 0.09$ indicates that the fit describes the Monte Carlo better than the data.

The second fit function (f_2, equation 7.2) shows similar results. Again, the bumps are found in all four distributions. Figure 7.2 shows the $2\mu + X$ class. For this scenario both fit functions are appropriate, though f_1 fits the underlying distribution slightly better.

Of course this only gives a qualitative idea of the performance of the Bump Hunter. To measure the sensitivity quantitatively, one can conduct a sufficiently large number of pseudo experiments and measure how often the algorithm claimed a discovery in the expected region. The sensitivity can be determined by:

$$S = \frac{\text{number of pseudo experiments where the resonance was found}}{\text{total number of pseudo experiments}}. \tag{7.8}$$

In that sense, figure 7.4 shows the sensitivity of the algorithm for a Z' with the mass $M = 500\,\mathrm{GeV}$ at an integrated luminosity of $36\,\mathrm{pb}^{-1}$ and varying cross sections in the $2\mu + X$ event class. At the expected cross section in the Sequential Standard Model of 2.00 pb (table 7.1) the bump is found with a likelihood of $S_1 = 0.800^{+0.027}_{-0.029}$ for f_1 and $S_2 = 0.855^{+0.024}_{-0.026}$ for f_2. At this value, also the saturation for this search is reached. When the sidebands do not follow the data distribution, regardless of the quality of the bump, there will not be a positive result. This is why the saturation value is below 100%.

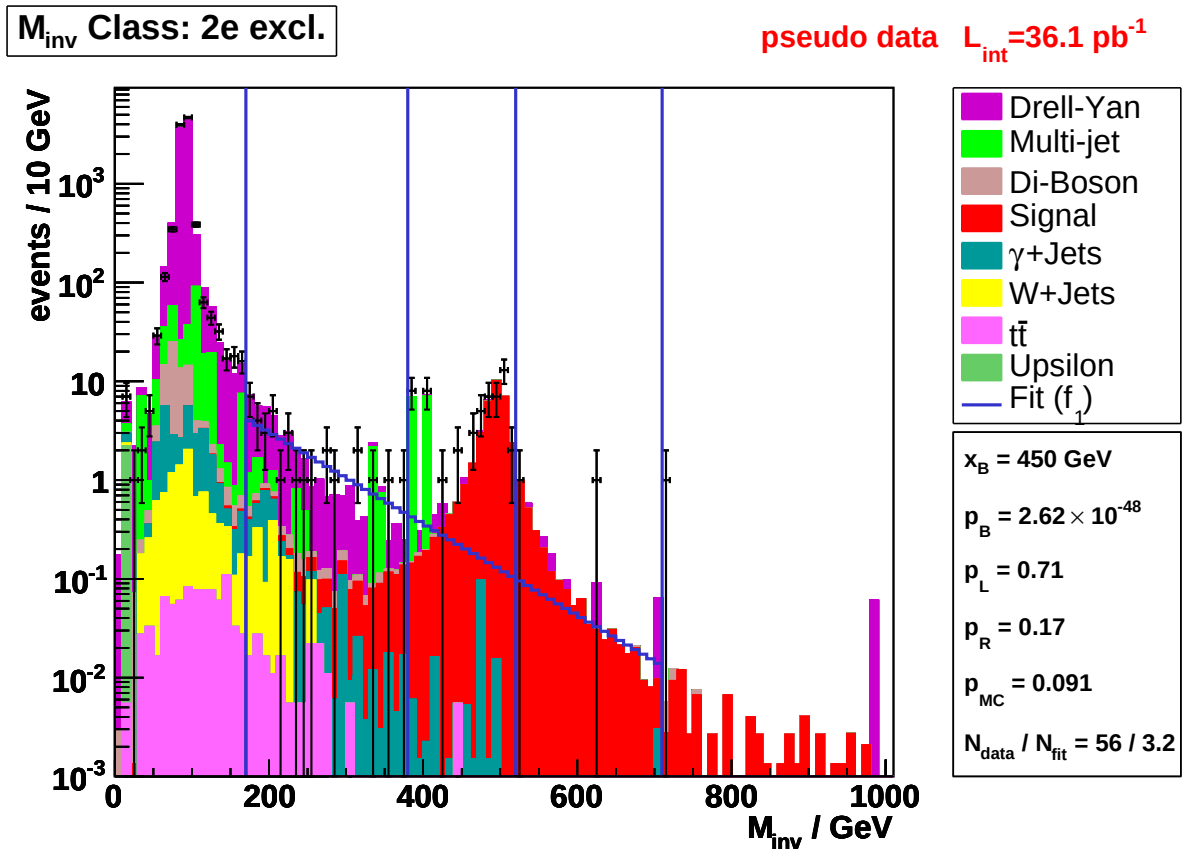


Figure 7.3.: Z' bump in the exclusive $2e$ class. The Bump Hunter has found the peak at around 450 GeV, using the f_1 fit function. The Z peak has not been found due to the lack of an appropriate left sideband. The spikes in the distribution are caused by low Monte Carlo statistics.

The sensitivities using f_1 and f_2 for the Z' peak are similar, f_2 performs only slightly better.

7.3.2. Excited Muon

If leptons are composite particles, a spectrum of excited lepton states might be observed at the LHC. In the model tested here, a μ^* is produced together with a Standard Model μ via a Drell-Yan process and then decays into $\mu + \gamma$.

The cross section of a 400 GeV excited muon at $\Lambda_0 = 10$ TeV is taken from [64] and is rescaled to $\Lambda_1 = 2$ TeV by (section 2.2.3)

$$\sigma_1 = \sigma_0 \cdot \left(\frac{\Lambda_0}{\Lambda_1}\right)^4 . \tag{7.9}$$

The values are listed in table 7.2.

In the class $1\mu, 1\gamma + X$ the signal resonance has been found using the fit function f_2. It is depicted in figure 7.5. Data and Monte Carlo in the left sideband are well described by the fit. Similar to the Z' scenario, the background is vanishing in the right sideband. Instead, part of the bump does not lie in the inner region but in the right sideband.

The bump for the $\Lambda = 2$ TeV is discovered in around 80% of the pseudo experiments. The sensitivities for different cross sections are depicted in figure 7.6. The values for the two fit func-

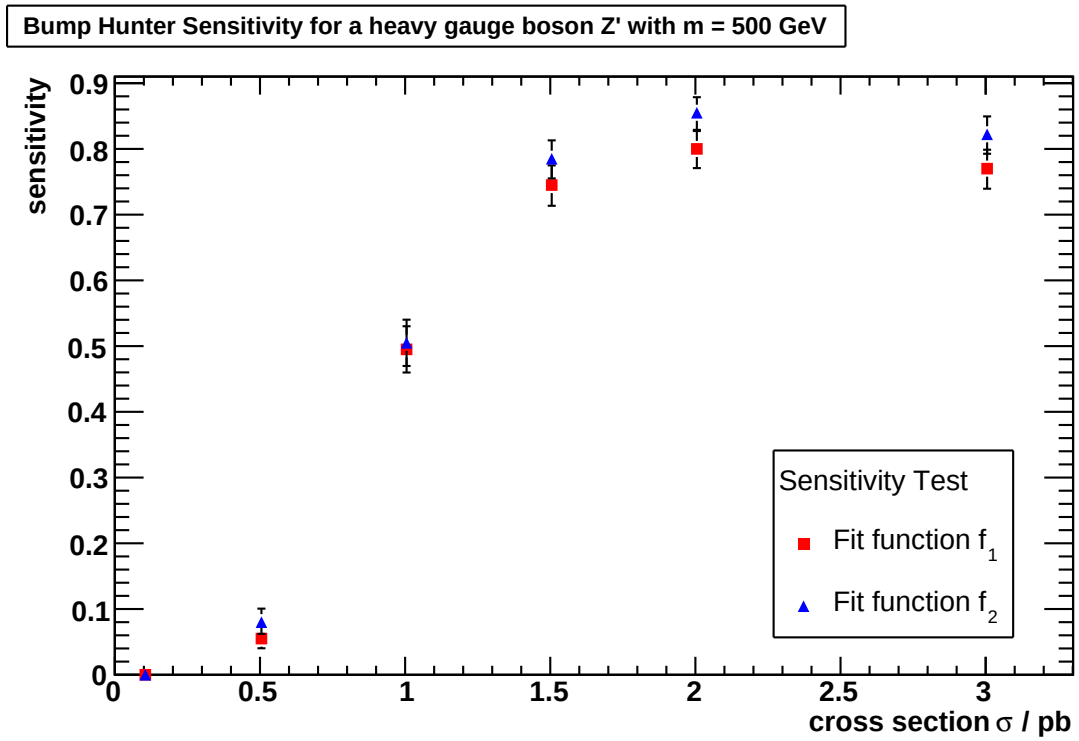


Figure 7.4.: Sensitivity of the Bump Hunter for a 500 GeV Z' depending on the cross section. All criteria from section 7.2 have been applied.

μ^*mass/GeV	Λ/TeV	Cross Section $\times$ BR / pb
400	10	0.0023
400	2	1.44

Table 7.2.: Excited muon cross sections.

tions are compatible within their uncertainties. The saturation value is larger in comparison to the Z' scenario.

Apparently, the Bump Hunter can handle resonances in the tail of a distribution very well, even though it has no well described right sideband.

7.3.3. Leptoquark

Leptoquarks are hypothetical particles carrying both a quark and a lepton quantum number. The existence of leptoquarks is postulated by a number of models. They could for instance explain the number of lepton and quark generations (section 2.2.4).

Second generation leptoquarks are predominantly pair produced and decay into a muon and a c quark. In MUSiC, a second generation leptoquark should be best observable in the $1\mu, 1\text{jet} + X$ class.

For this sensitivity test, a second generation leptoquark with a mass of $m = 350\,\text{GeV}$ is used. Pseudo experiments at an integrated luminosity of $36.1\,\text{pb}^{-1}$ are produced. The cross section $\sigma = 0.251\,\text{pb}$ as in the mBRW model [65] is too small for the resonance to be discovered by the Bump Hunter. Still this scenario with a different cross section is an important benchmark

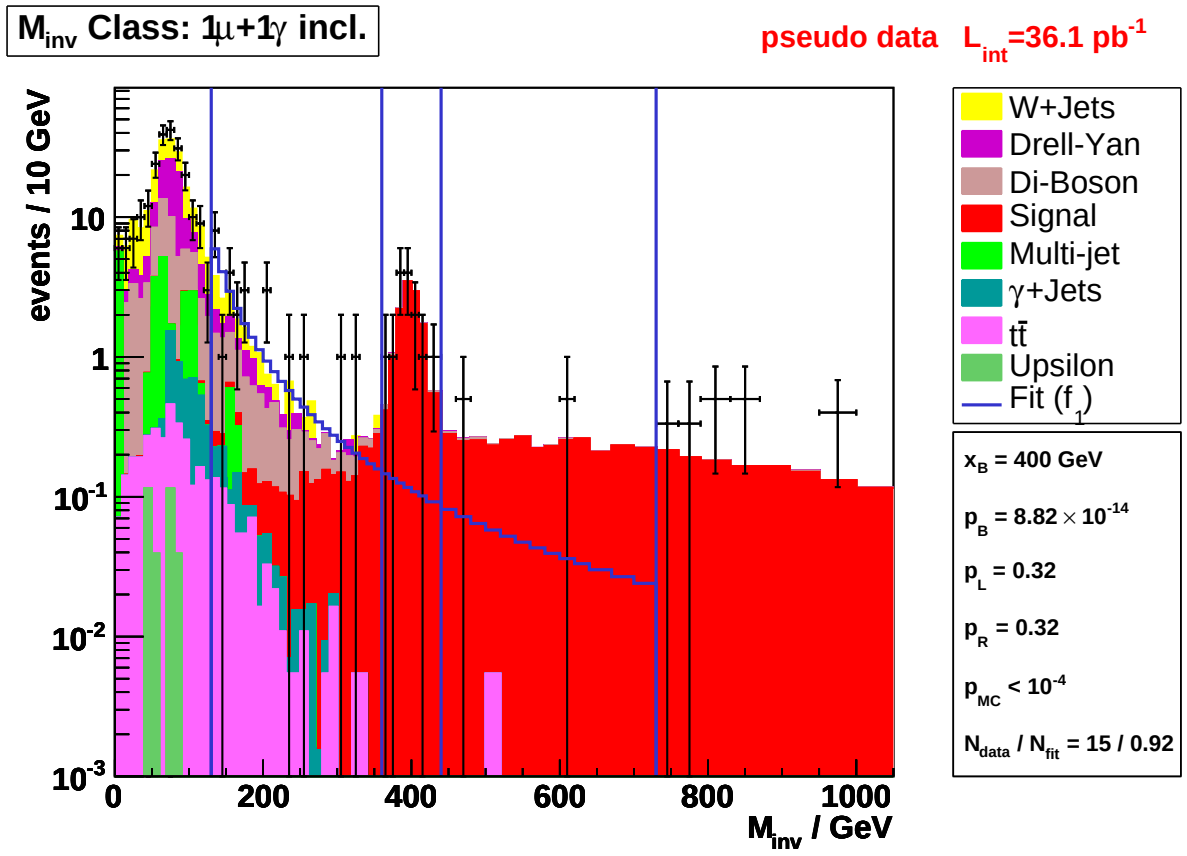


Figure 7.5.: Invariant mass spectrum of the $1\mu, 1\gamma + X$ class including an excited muon with a mass of 400 GeV and $\Lambda = 2$ TeV. The bump has been found, using the f_2 fit function.

scenario for the Bump Hunter. In contrast to the evaluated Z' and the excited muon scenario, there is a large number of background events in the bump region provided by the Standard Model.

In figure 7.7 the Bump Hunter has found the leptoquark scaled to a cross section of 250 pb. The algorithm is capable of finding such resonances even if further distributions are superimposed on them.

The sensitivity for the case of a leptoquark with varying cross section is pictured in figure 7.8. Using the f_1 fit function, the Bump Hunter discovers the signal when it has a sufficiently large cross section. The sensitivity spectrum of f_2 shows a different effect. A bump is "discovered" in around 60% of the pseudo experiments, even if the signal cross section is 0. As there is a smooth distribution in the Standard Model Monte Carlo distribution, this is a misidentification. The sensitivity rises for larger cross sections and has a saturation value of around 80%-90%. This is less than for f_1, which means that f_2 usually describes the sidebands not as well as f_1.

7.4. 2010 Data Results

When running on the 2010 CMS data from proton proton collisions, the Bump Hunter finds a number of bumps in the invariant mass distributions. Table 7.3 shows all 23 interesting regions found in the 2010 data, using the exponential fit f_1. The results using f_2 in table 7.4 comprise 33 bumps. Not every bump can be discussed here in detail. Therefore, the two most signifi-

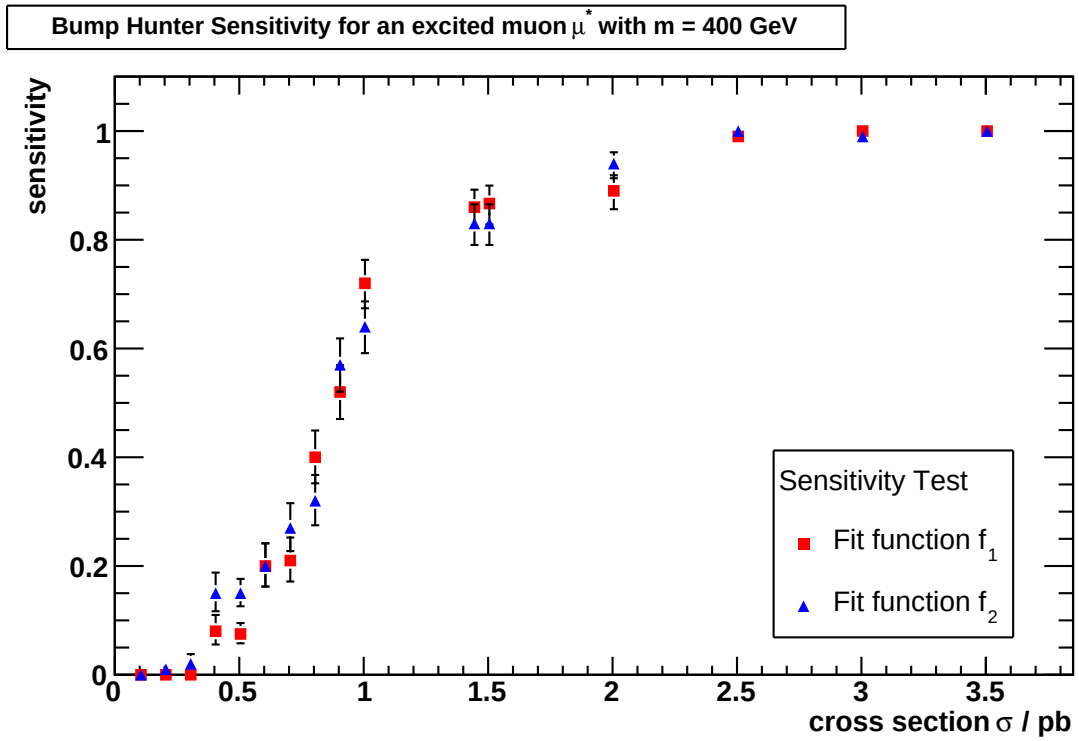


Figure 7.6.: Bump Hunter sensitivity for a μ^* scenario with a mass of 400 GeV at 36.1 pb^{-1}. All criteria from section 7.2 have been applied.

cant ones, and those passing the Monte Carlo test are presented in detail. Additionally a few particularly interesting ones are shown and compared with the results of the standard MUSiC algorithm.

A lot of the bumps in data also appear in Monte Carlo. One example is the most significant bump using the f_1 fit. It is observed in the $2\mu + E_t^{\text{miss}}$ class, pictured in figure 7.9. This bump results from a turn-on effect caused by the selection thresholds for muons and E_t^{miss}.

The majority of the bumps, also the second most significant one found with f_1 (figure 7.10), are of that kind. It is inherent to the Bump Hunter to find such bumps, as they fulfill all presented criteria. So even if they are not very interesting in terms of new physics, it is a good indicator for the performance of the Bump Hunter that they have been found.

The classes which pass the Monte Carlo test using f_1 are $2\mu + E_t^{\text{miss}} + X$ (figure 7.11), 2μ, 1jet $+ E_t^{\text{miss}} + X$ (figure 7.12) and 2μ, 1jet $+ E_t^{\text{miss}}$ (figure 7.13). The identified regions all show a bump structure in data as well as in Monte Carlo, but in each case the data has an excess compared to the simulation and the peak appears to be shifted to higher values.

In the class $1e$, 1jet $+ E_t^{\text{miss}}$ one bump is detected using f_1 (figure 7.14). It is located in the tail of the distribution at around 600 GeV. Five of six bins lie well above the sideband fit. Noticeably, the bump is not formed like a Gaussian or a Breit-Wigner. It has no fast decline to the sides so it does not look like a typical resonance. Also, the Monte Carlo expectation lies between the fit and the data, indicating that the fit function might fall too steeply. Indeed, this region has not qualified as a bump when using the f_2 fit function. Nevertheless, it is one of the more interesting bumps, worth further examination with a higher integrated luminosity.

When evaluating the same distribution with the MUSiC region of interest algorithm a subsection of the region was determined as the most significant region(figure 7.15). MUSiC calculated a significance of $p_{\text{data}} \approx 0.003$, when corrected for the look elsewhere effect a $\tilde{p} \approx 0.18$ resulted.

Class	x_B	p_{B}	p_{L}	p_{R}	p_{MC}	Figure
2 mu, 1 met	90 GeV - 150 GeV	$1.4 \cdot 10^{-31}$	0.29	0.61	0.011	7.9
1 mu, 1 gam	60 GeV - 100 GeV	$2.4 \cdot 10^{-28}$	0.14	0.41	0.81	7.10
1 e, 1 mu, 1 met + X	90 GeV - 170 GeV	$2.2 \cdot 10^{-26}$	0.68	0.27	0.45	
2 e, 1 met + X	100 GeV - 160 GeV	$1.4 \cdot 10^{-19}$	0.14	0.21	0.2	
1 e, 1 gam, 1 met	100 GeV - 150 GeV	$1.4 \cdot 10^{-18}$	0.23	0.54	0.59	
1 mu, 1 gam, 1 met + X	80 GeV - 150 GeV	$6.6 \cdot 10^{-18}$	0.8	0.15	0.97	
1 e, 1 gam, 1 met + X	100 GeV - 150 GeV	$1.1 \cdot 10^{-15}$	0.56	0.66	0.67	
1 e, 4 jet + X	600 GeV - 860 GeV	$3.4 \cdot 10^{-14}$	0.087	0.39	0.49	
1 e, 2 jet	300 GeV - 440 GeV	$3.9 \cdot 10^{-13}$	0.058	0.61	0.93	
1 mu, 2 jet	490 GeV - 1000 GeV	$5.2 \cdot 10^{-13}$	0.88	0.24	0.39	
2 mu, 2 jet + X	320 GeV - 480 GeV	$3 \cdot 10^{-12}$	0.084	0.099	0.011	
1 e, 1 mu, 1 jet, 1 met + X	210 GeV - 300 GeV	$6 \cdot 10^{-12}$	0.68	0.49	0.24	
1 e, 3 jet, 1 met	410 GeV - 550 GeV	$6.1 \cdot 10^{-12}$	0.059	1	0.29	
2 mu, 1 met + X	100 GeV - 140 GeV	$1.7 \cdot 10^{-11}$	0.23	0.39	**0.009**	7.11
2 mu, 1 jet, 1 met + X	220 GeV - 400 GeV	$2.5 \cdot 10^{-10}$	0.15	0.76	**0.0032**	7.12
2 mu, 1 jet, 1 met	210 GeV - 360 GeV	$1.2 \cdot 10^{-09}$	0.32	0.59	**0.0042**	7.13
1 e, 2 jet + X	280 GeV - 380 GeV	$1.7 \cdot 10^{-09}$	0.25	0.17	0.94	
1 e, 1 jet, 1 met	540 GeV - 680 GeV	$2.3 \cdot 10^{-09}$	0.063	0.28	0.04	7.14
1 e, 4 jet	640 GeV - 1010 GeV	$8.7 \cdot 10^{-09}$	0.13	0.8	0.6	
1 e, 3 jet, 1 met + X	430 GeV - 550 GeV	$1.2 \cdot 10^{-08}$	0.062	0.34	0.28	
1 e, 1 jet, 1 met + X	540 GeV - 680 GeV	$4.8 \cdot 10^{-08}$	0.1	0.074	0.19	
1 mu, 4 jet + X	480 GeV - 620 GeV	$5.6 \cdot 10^{-08}$	0.25	0.93	0.14	
2 mu + X	180 GeV - 240 GeV	$8.8 \cdot 10^{-08}$	0.25	0.11	0.13	

Table 7.3.: List of bumps found in 2010 data with the Bump Hunter using an exponential fit function, sorted by ascending p_{B}. p_{B} is the p-value of the bump region, p_{L} and p_{R} are the p-values of the two sideband regions. p_{MC} describes the compatibility with the Monte Carlo, i.e. whether the bump can also be observed in Monte Carlo. The values of p_{MC} where the bumps pass $p_{\mathrm{MC}} < 0.01$ are printed in bold.

Class	x_{Bump}	p_{inner}	p_{left}	p_{right}	p_{MC}	Figure
1 e, 4 jet + X	620 GeV - 920 GeV	$1.3 \cdot 10^{-18}$	0.061	0.46	0.48	7.16
2 mu, 2 jet + X	320 GeV - 480 GeV	$2.8 \cdot 10^{-15}$	0.34	0.073	**0.0099**	7.17
1 e, 2 jet	300 GeV - 400 GeV	$3.1 \cdot 10^{-15}$	0.22	0.091	0.92	
1 mu, 2 jet	250 GeV - 330 GeV	$7.2 \cdot 10^{-14}$	0.059	0.14	0.69	
1 e, 3 jet, 1 met	390 GeV - 490 GeV	$7.3 \cdot 10^{-14}$	0.064	0.083	0.23	
1 mu, 2 jet + X	240 GeV - 310 GeV	$1 \cdot 10^{-12}$	0.057	0.085	0.7	
1 e, 2 jet + X	300 GeV - 380 GeV	$1.1 \cdot 10^{-11}$	0.054	0.063	0.93	
2 mu, 2 jet	300 GeV - 480 GeV	$1.3 \cdot 10^{-11}$	0.051	0.71	0.088	
1 mu, 1 jet, 1 met + X	190 GeV - 220 GeV	$2.7 \cdot 10^{-11}$	0.51	0.22	0.2	
1 e, 2 jet + X	780 GeV - 900 GeV	$3.2 \cdot 10^{-11}$	0.069	0.09	0.79	
1 e, 3 jet, 1 met + X	450 GeV - 570 GeV	$4.3 \cdot 10^{-11}$	0.058	0.3	0.19	
1 e, 3 jet + X	990 GeV - 1230 GeV	$5.2 \cdot 10^{-11}$	0.052	0.98	0.65	
1 mu, 2 jet, 1 met	320 GeV - 360 GeV	$5.7 \cdot 10^{-11}$	0.051	0.73	0.046	7.19
1 e, 1 mu, 1 jet, 1 met + X	210 GeV - 300 GeV	$6.1 \cdot 10^{-11}$	0.85	0.37	0.24	
1 e, 3 jet	630 GeV - 810 GeV	$1.7 \cdot 10^{-10}$	0.068	0.59	0.77	
1 e, 1 jet, 1 met	200 GeV - 240 GeV	$2.2 \cdot 10^{-10}$	0.41	0.073	0.51	
2 mu, 1 jet, 1 met	220 GeV - 380 GeV	$5.7 \cdot 10^{-10}$	0.11	0.78	**0.0061**	
2 mu, 1 jet, 1 met + X	220 GeV - 360 GeV	$6.3 \cdot 10^{-10}$	0.66	0.76	**0.0027**	
1 mu, 4 jet	560 GeV - 810 GeV	$1 \cdot 10^{-09}$	0.3	0.78	0.14	
1 e, 1 jet, 1 met + X	170 GeV - 190 GeV	$2.4 \cdot 10^{-09}$	0.063	0.23	0.56	
1 mu, 1 jet + X	240 GeV - 290 GeV	$8.1 \cdot 10^{-09}$	0.12	0.49	0.3	
2 mu, 1 jet + X	360 GeV - 520 GeV	$9 \cdot 10^{-09}$	0.46	0.8	0.036	
1 e, 3 jet + X	530 GeV - 670 GeV	$9.1 \cdot 10^{-09}$	0.11	0.066	0.81	
1 mu, 2 jet	510 GeV - 630 GeV	$1.9 \cdot 10^{-08}$	0.068	0.54	0.3	
1 mu, 3 jet + X	680 GeV - 920 GeV	$2.4 \cdot 10^{-08}$	0.057	0.23	0.34	7.21
1 e, 4 jet	640 GeV - 980 GeV	$3.1 \cdot 10^{-08}$	0.49	0.73	0.64	
1 mu, 4 jet + X	900 GeV - 1360 GeV	$3.6 \cdot 10^{-08}$	0.59	0.95	0.27	
1 mu, 1 jet + X	330 GeV - 390 GeV	$4.4 \cdot 10^{-08}$	0.23	0.78	0.27	
1 e, 4 jet, 1 met + X	520 GeV - 720 GeV	$4.5 \cdot 10^{-08}$	0.17	0.61	0.21	
2 mu, 2 jet + X	580 GeV - 680 GeV	$4.5 \cdot 10^{-08}$	0.85	0.056	0.012	
2 e, 2 jet + X	500 GeV - 540 GeV	$5.5 \cdot 10^{-08}$	0.88	0.18	**0.0016**	7.18
1 e, 2 jet + X	460 GeV - 560 GeV	$5.6 \cdot 10^{-08}$	0.34	0.1	0.92	
2 e, 1 jet, 1 met + X	220 GeV - 360 GeV	$5.8 \cdot 10^{-08}$	0.13	0.85	0.031	

Table 7.4.: List of bumps found in 2010 data with the Bump Hunter using the fit function f_2.

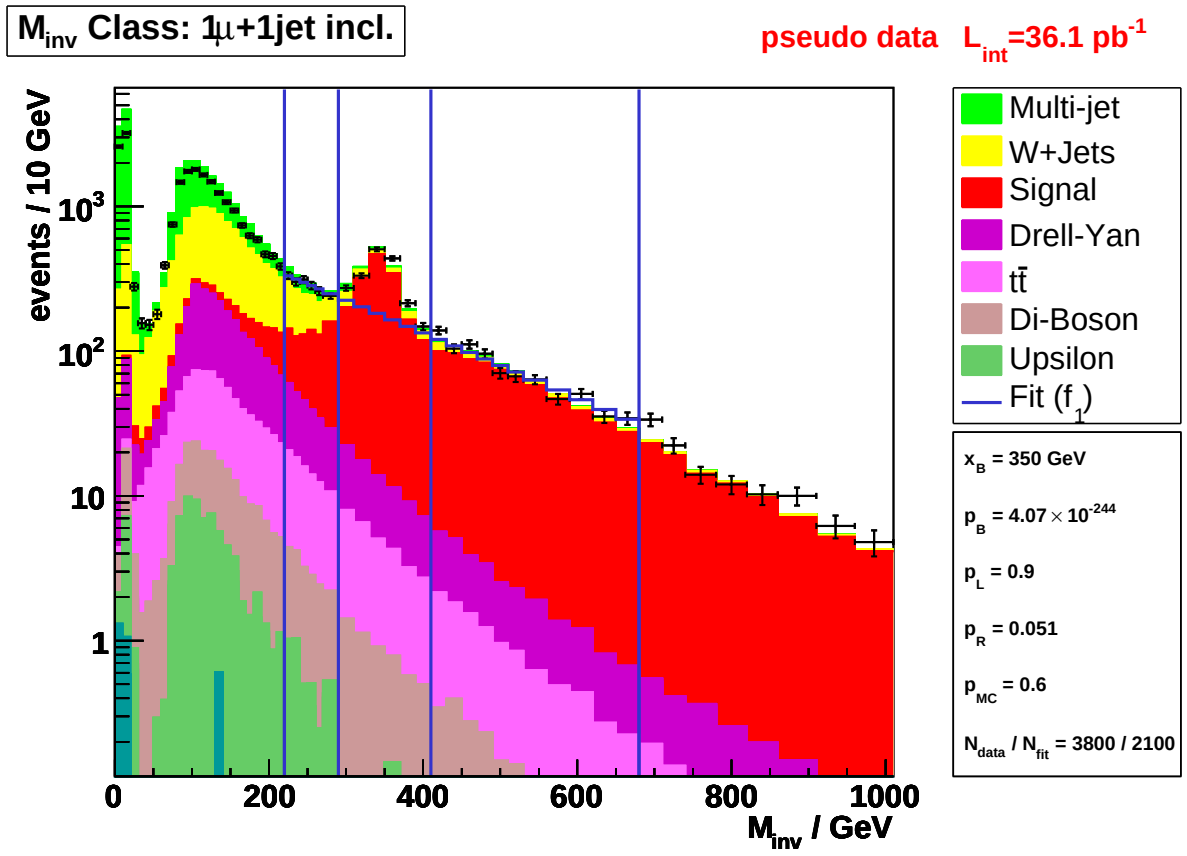


Figure 7.7.: Invariant mass spectrum of the $1\mu, 1\text{jet} + X$ class with a second generation leptoquark as signal. The signal cross section is $\sigma = 250\,\text{pb}$. The bump has been found, using the f_1 fit function.

Using f_2, 33 bumps are detected. The most significant one, in the class $1e, 4\text{jets} + X$, is depicted in figure 7.16. It is compatible with the Monte Carlo prediction.

The second most significant bump in the f_2 scan, in the $2\mu, 2\text{jets} + X$ class (figure 7.17), passes the Monte Carlo test. Without considering the Monte Carlo it looks quite interesting as it appears to be a bump on an underlying falling distribution. It also qualifies as an excess in relation to the Monte Carlo. For clarification, a higher luminosity is needed.

Other bumps that pass the Monte Carlo test are identified in the $2\mu, 1\text{jet} + E_t^{\text{miss}}$ class and the $2\mu, 1\text{jet} + E_t^{\text{miss}} + X$ class. They are in the same regions as when using f_1 and are therefore not depicted again.

Also the bump in the $2e, 2\text{jets} + X$ class (figure 7.18) passes the Monte Carlo test. The fit function does not describe the data well in a more global sense but only in the six fitted bins. When comparing to the Monte Carlo prediction one can see an excess of data in the inner region and a deficit in the right sideband. Because of the small size region, this leads to a very steeply falling fit function.

Also for the f_2 fit, we find a narrow bump in the $1\mu, 2\text{jets} + E_t^{\text{miss}}$ class, as depicted in figure 7.19. Here, the Bump Hunter has found a two bin excess. It demonstrates the capability of the Bump Hunter to find narrow resonances as well. This fact is important, as for long living particles, the decay width is small and the bump width is only determined by the resolution of the detector. As we rebin to resolution, narrow bumps are an expected signature.

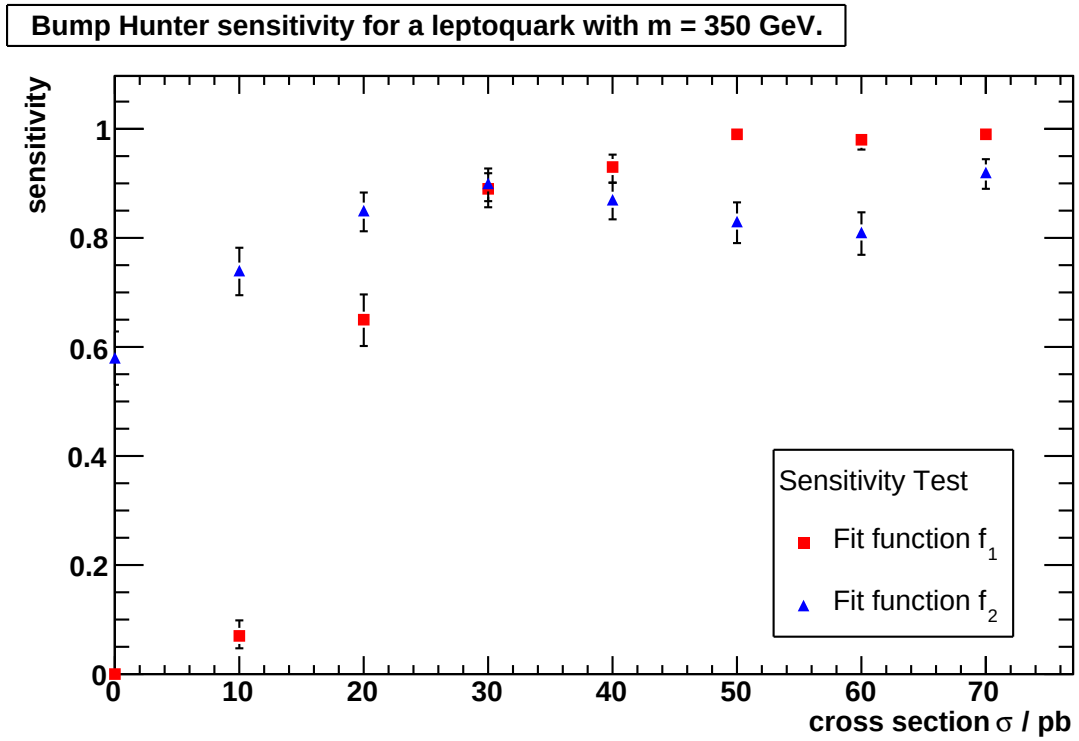


Figure 7.8.: Sensitivity for the LQ scenario at 36 pb^{-1}. Using the fit function f_2, a bump is discovered in the region regardless of a leptoquark signal. All criteria from section 7.2 have been applied.

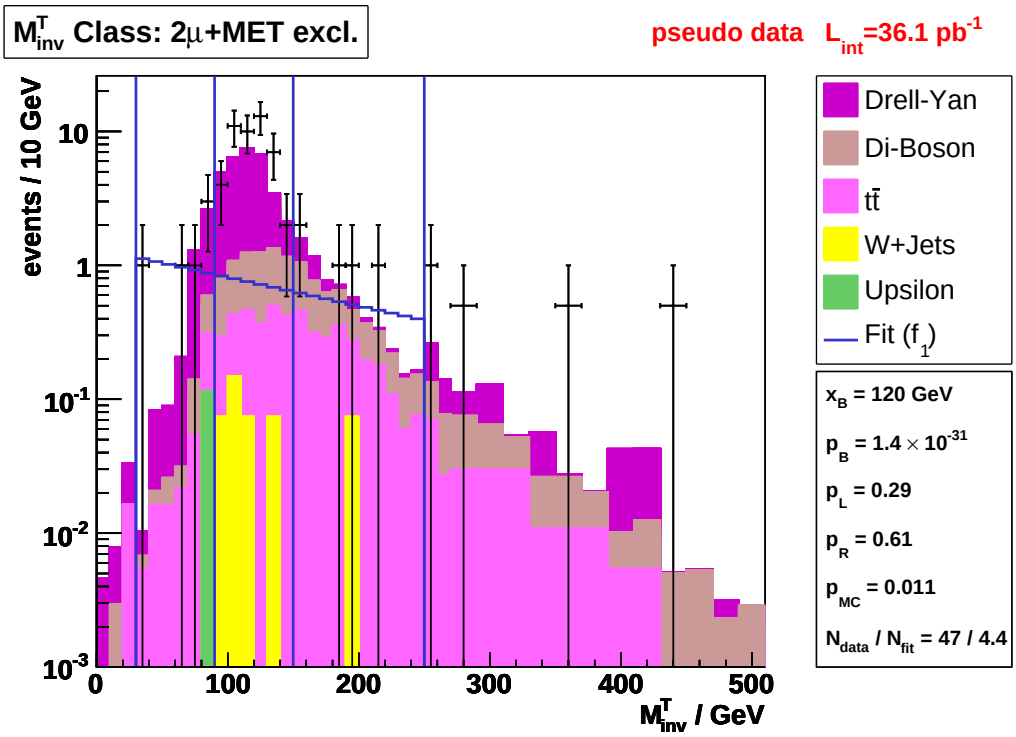


Figure 7.9.: Most significant bump, found in the M_{inv}^T distribution of the $2\mu, +E_t^{\mathrm{miss}}$ class, in 2010 data with the f_1 fit.

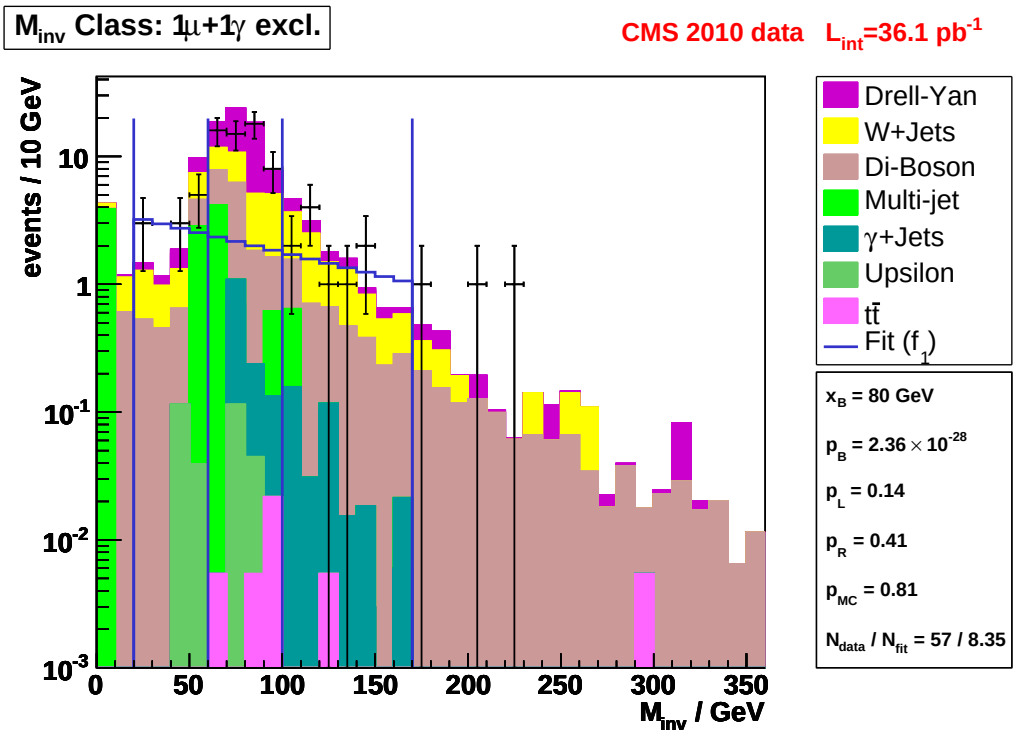


Figure 7.10.: Second most significant bump, found in the M^{T}_{inv} distribution of the $1\mu, 1\gamma$ class, in 2010 data using the f_1 fit.

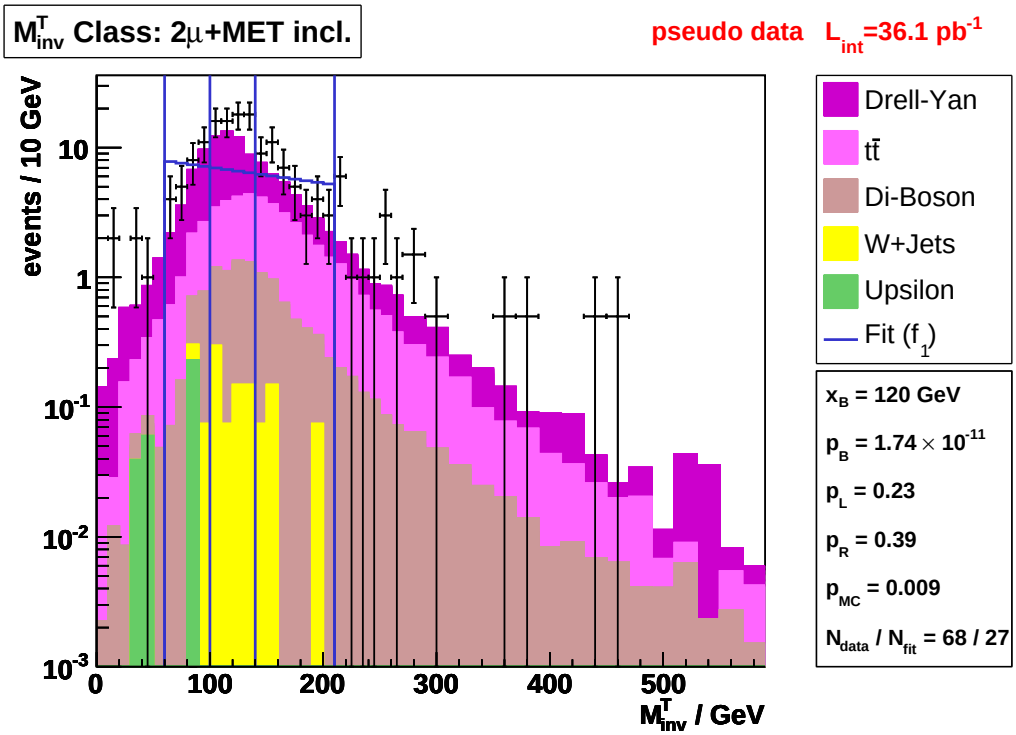


Figure 7.11.: Bump in the $2\mu + E_t^{\text{miss}} + X$ class in 2010 data using the f_1 fit. This bump has passed the Monte Carlo test.

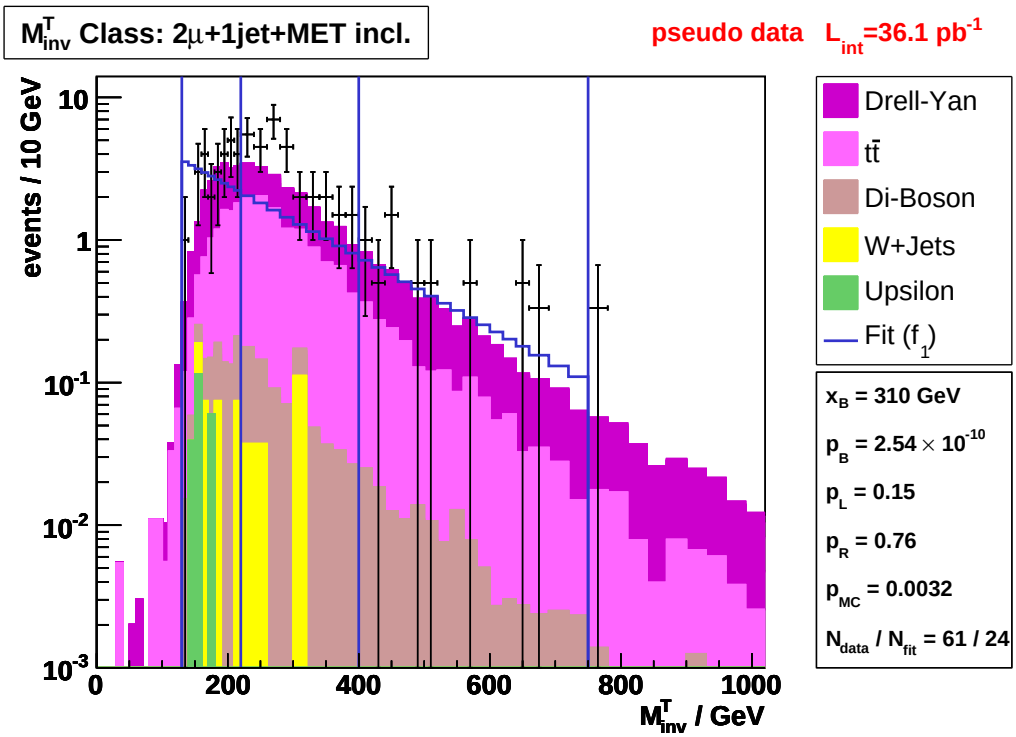


Figure 7.12.: Bump in the 2μ, 1jet + E_t^{miss} + X class in 2010 data using the f_1 fit. This bump has passed the Monte Carlo test.

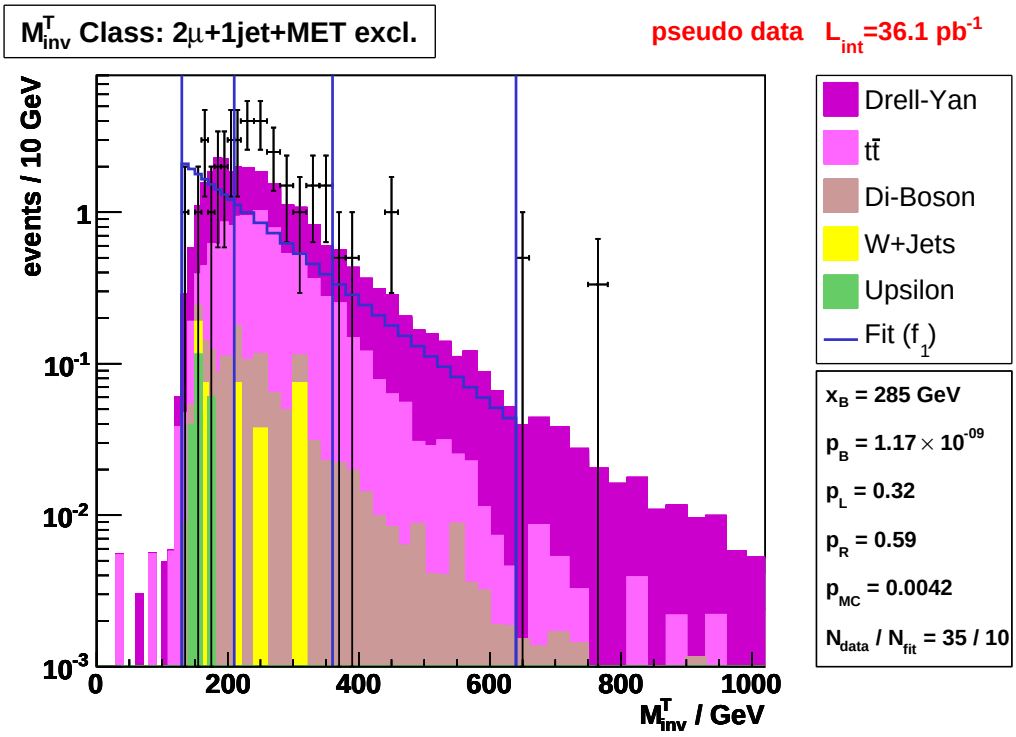


Figure 7.13.: Bump in the 2μ, 1jet + E_t^{miss} class in 2010 data using the f_1 fit. This bump has passed the Monte Carlo test.

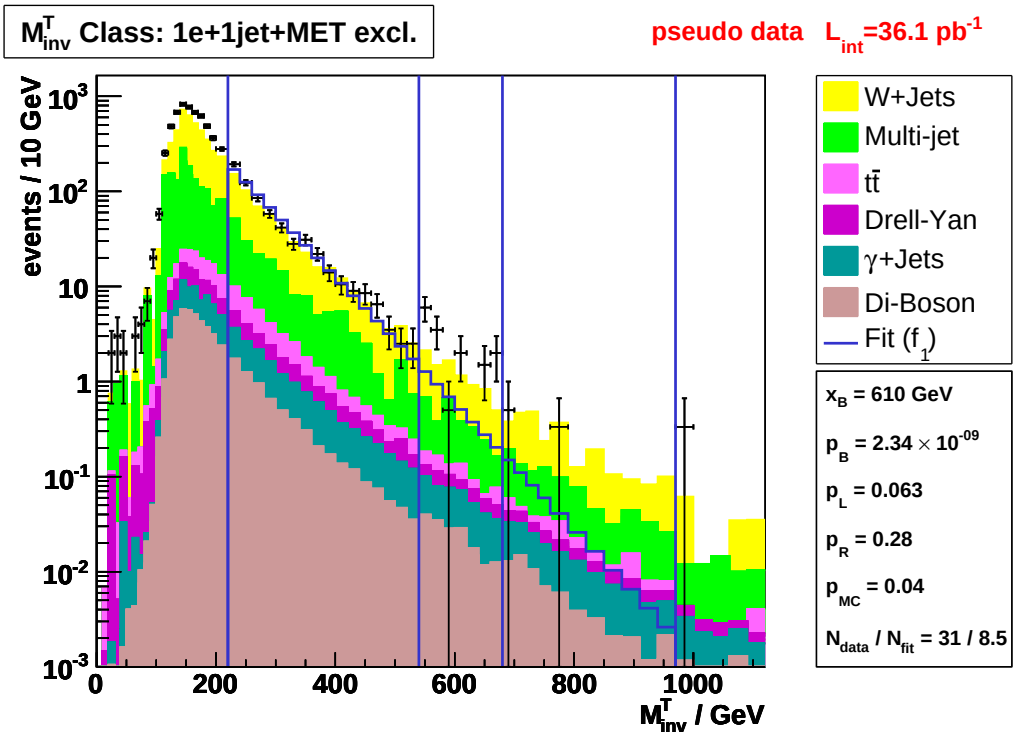


Figure 7.14.: Interesting bump in the $1e$, 1jet + E_t^{miss} event class. The fit function is f_1.

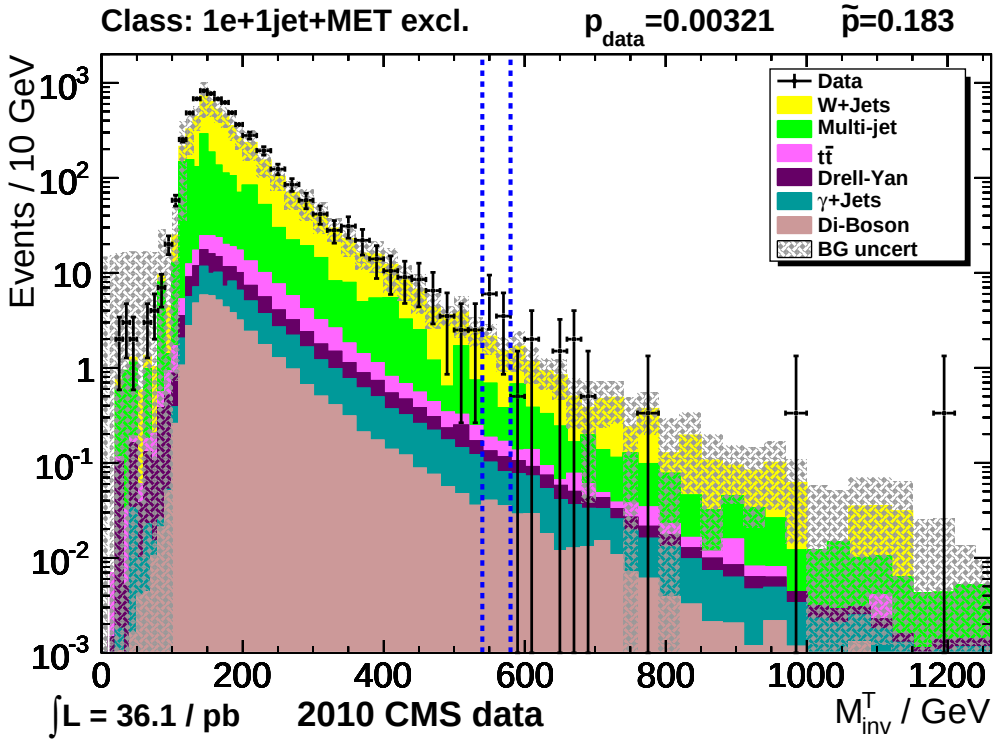


Figure 7.15.: The missing transverse energy of the $1e$, 1jet + E_t^{miss} event class examined by the MUSiC region of interest algorithm.

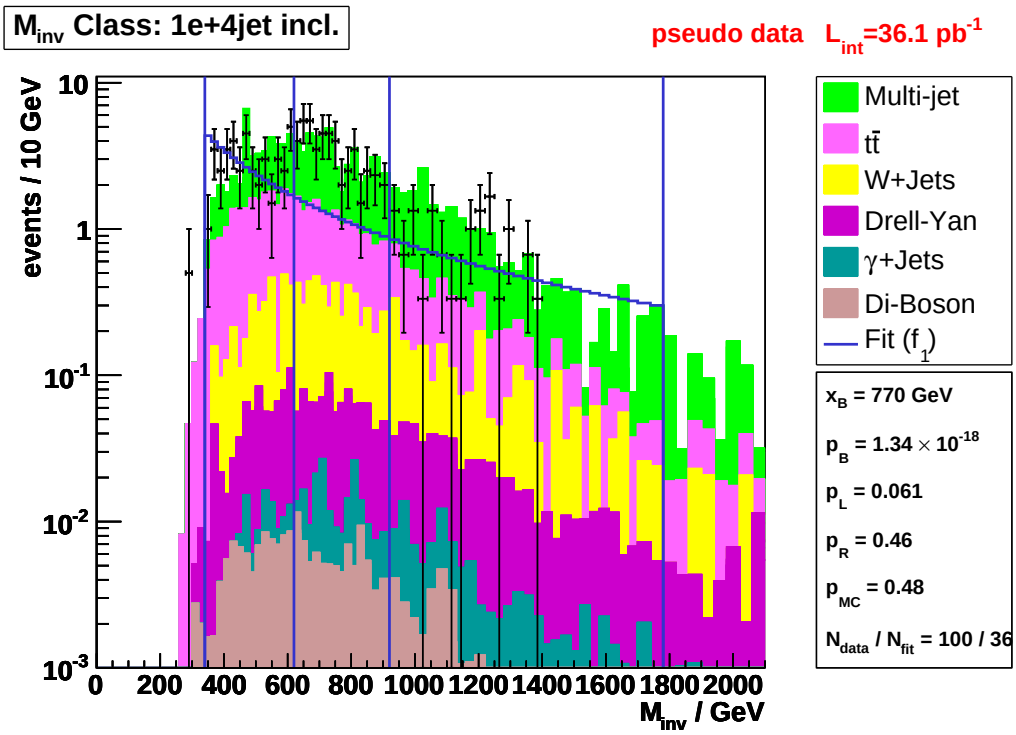


Figure 7.16.: Most significant bump in data using the f_2 fit function.

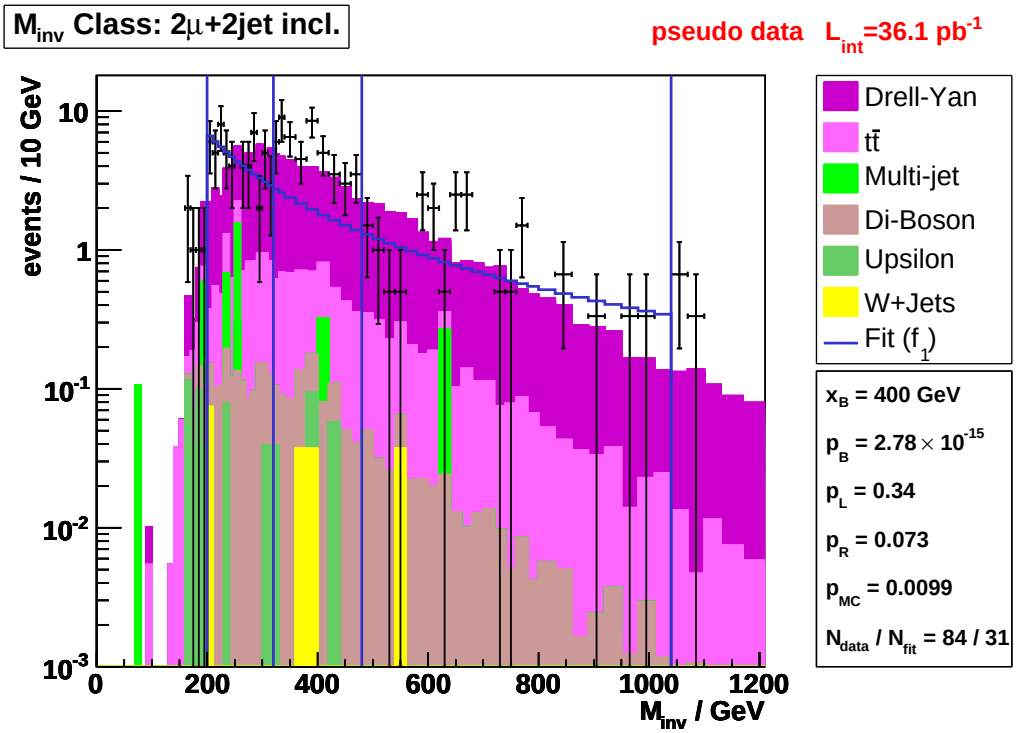


Figure 7.17.: Second most significant bump in data using the f_2 fit function.

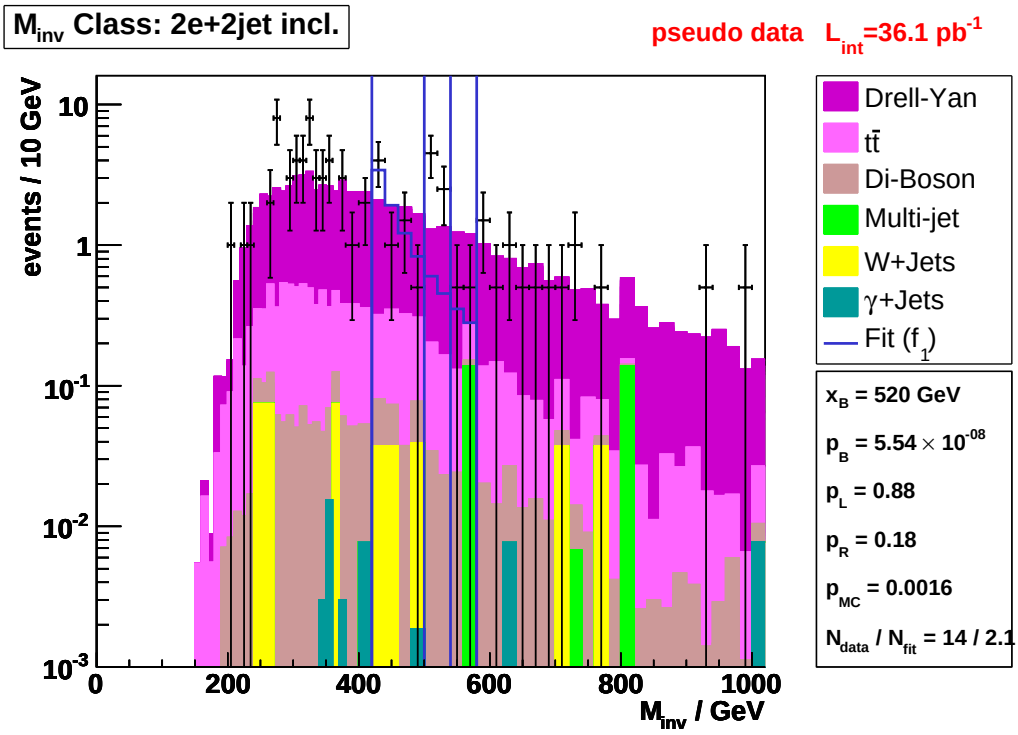


Figure 7.18.: $2e$, 2jets $+ X$ class, scanned using the f_2 fit function.

Figure 7.20 shows the evaluation of the same class by the MUSiC region of interest algorithm. About the same region was selected. A significance of around 10^{-3} was determined but considering the look elsewhere effect it decreases to $\tilde{p} \approx 0.1$.

The rate of detecting fake bumps seems to be quite low for the selected criteria and parameters of the algorithm. Unfortunately, this is difficult to quantify as a definition of what a fake bump is would be arbitrary. Figure 7.21 shows a bump, found by the Bump Hunter using f_2. This bump seems to be an artifact, originating from an inappropriate fit function. The shape of the function does not follow the shape of the data, even though it describes the data in the sidebands well. This bump was not detected using the fit function f_1.

7.5. Conclusion and Outlook

The Bump Hunter is a new algorithm to search for resonances, especially in invariant mass spectra. The concept of a completely *data driven* and *model independent* analysis was established and an algorithm was developed and implemented.

Two different fit functions have been tested and both show a good performance. This has been demonstrated with test scenarios of a Z', a leptoquark, and an excited muon.

The 2010 data has been evaluated using the presented Bump Hunter algorithm. 23 / 33 bumps have been found using the fit function f_1 / f_2. A few of them are interesting and should be reevaluated with a higher integrated luminosity. A bump strongly indicating a phenomenon not explainable by the Standard Model could not be found.

As this is a completely new approach, a number of questions remain open: A study on the look elsewhere effect of the Bump Hunter would be helpful. The set of fit functions could well be extended. The MUSiC event classes cover a great range of the invariant mass spectrum. But not all combinations of particles in an event are combined to evaluate their invariant masses.

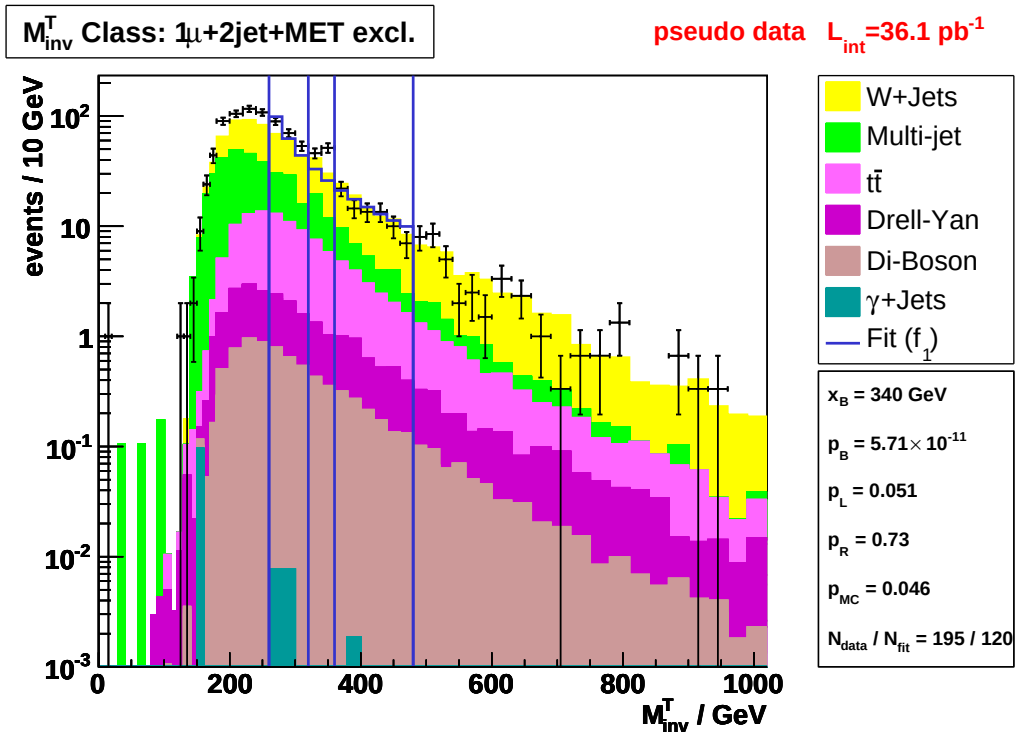


Figure 7.19.: Narrow bump discovered in the 1μ, 2jets + E_t^{miss} class, discovered using f_2.

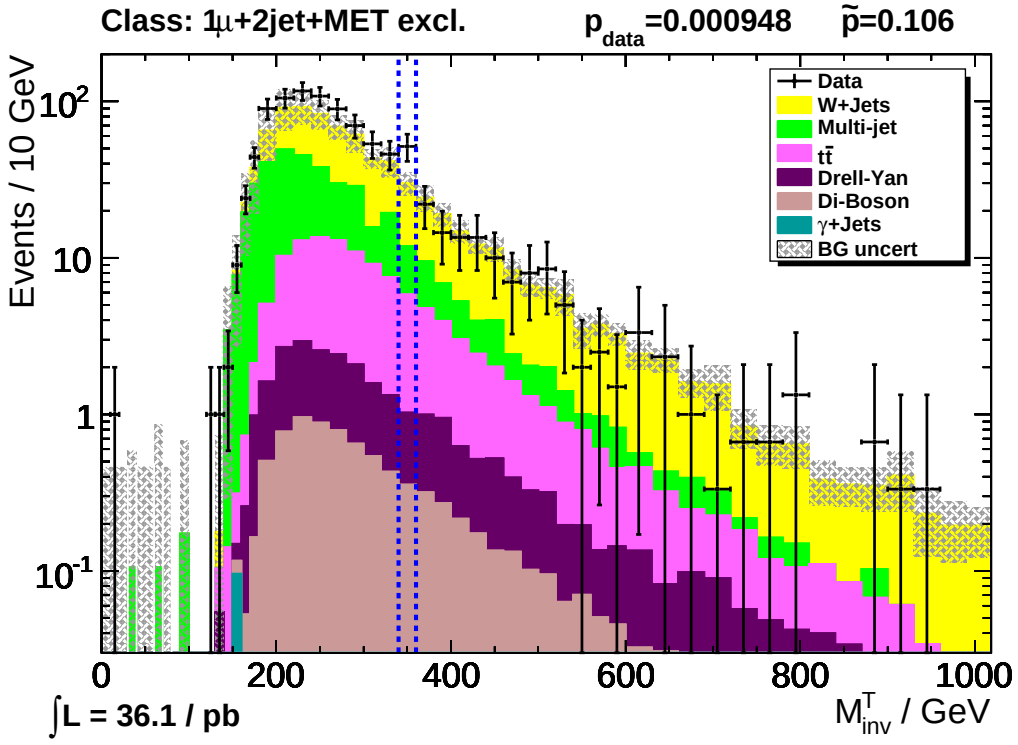


Figure 7.20.: Evaluation of the 1μ, 2jets + E_t^{miss} event class by the MUSiC region of interest algorithm.

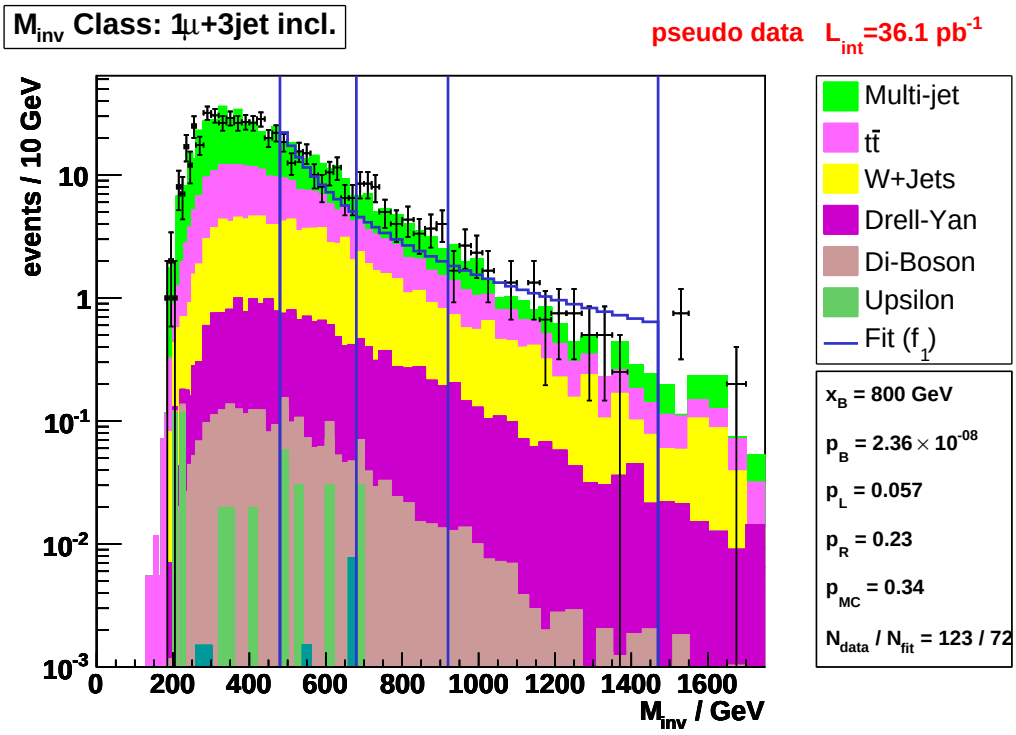


Figure 7.21.: Example for the impact of the choice of fit function. This bump found with the function f_2 has not been found while using f_1. Obviously, this bump is an artifact due to the form of the fit function.

For example, one could imagine a model that shows as a signature in a $3e$ class but the two softer electrons are the decay product of a yet unknown particle. For the MUSiC concept this might not be feasible as soft particles might be poorly described by the Monte Carlo. The data driven Bump Hunter could be used to evaluate these distributions as well.

8. An Improved P-Value for MUSiC

The way to calculate the p-value in MUSiC is well motivated and has been extensively tested [50, 1]. Nevertheless, it is only one possible statistical model and also has its drawbacks. In this chapter an alternative approach is derived and the two methods are compared.

The variables used in the derivation are listed in table 8.1.

8.1. Status Quo

The status quo of the statistical methods used in MUSiC is described in Section 6.5.1. The p-value of a bin region, in the case[1] of $N_{data} \geq N_{SM}$, is determined by:

$$p = \sum_{i=N_{data}}^{\infty} P(i|b) \tag{8.1}$$

where $P(i|b)$ is the probability to observer i events if the Monte Carlo expectation is b. For the current Implementation of MUSiC this means:

$$p = \sum_{i=N_{data}}^{\infty} A \cdot \int_0^{\infty} d\lambda \exp\left(\frac{-(\lambda - b)^2}{2\sigma^2}\right) \cdot \frac{e^{-\lambda}\lambda^i}{i!} \tag{8.2}$$

where b is the scaled number of Monte Carlo events

$$b = \sum_{\text{MC processes (i)}} \frac{N_{\text{SM},i}}{\alpha_i}, \tag{8.3}$$

σ is the systematic uncertainty and A is the normalisation. If $b = 0$ and $N_{data} > 0$, this leads to a p-value of 0, i.e. an infinite significance. But as b is always determined by the finite statistics of Monte Carlo experiments, it is mathematically incorrect to claim a p-value of exactly 0. The value b can only be an estimator for the expection value of the number of events. How good this estimation is depends strongly on how many Monte Carlo events contribute to b, i.e. how big N_{SM} is. As they are scaled with luminosity and process cross section, this does not necessarily mean that b must be big if the scaling factor α is sufficiently large. The best solution to the described problem is sufficient Monte Carlo statistics. Table A.1 states the sample cross sections and their statistics. For instance, the "QCD_Pt50to80" sample has 3.2 million events but an expected number of events of

$$N = \sigma \cdot \int \mathscr{L}\, dt = 5.5 \cdot 10^6 \text{pb} \cdot 36.1 \text{pb}^{-1} = 2 \cdot 10^8 \text{ events.} \tag{8.4}$$

Approximately a factor of 60 times more events are expected than are available through Monte Carlo. One can see that this QCD statistics are clearly insufficient. Dedicated analyses therefore use QCD samples derived from data (e.g. [66]). The constrains of a model independent search make it difficult to follow this approach because MUSiC examines a lot of different final states with different physical processes contributing.

[1] For simplicity only this case is considered here. The other case behaves analogously.

$P(a\|b)$	Probability to observe a under the condition of b.
$\pi(a)$	Prior probability to observe a.
n	Number of observed events. This is not N_{data} but rather corresponds to the running variable i in equation 8.2.
N_{SM}	Number of actual Monte Carlo events.
b	Scaled number of Monte Carlo events.
α	Scaling factor for Monte Carlo.
λ	Running variable for the true expectation value.

Table 8.1.: List of variables and functions.

8.2. Derivation of an Improved P-Value

If the Monte Carlo statistics are small, b is no longer a good estimator for the Standard Model expectation value. One approach to take this into account is described in this section.

The goal is to determine a new discrete probability function $P(n|b)$ (as in equation 8.1), which describes the probability to dice n events, in the case of observing a scaled number of Monte Carlo events of b. Under the assumption that the number of events in the considered bin is independent of the number of events in the other bins it yields:

$$P(n|b) = \int d\lambda \, P(n|\lambda) \cdot P(\lambda|b). \tag{8.5}$$

The probability to observe n events if the scaled number of Monte Carlo events is b ($P(n|b)$) is equal to the probability to observe n events in the case the true expectation value is λ ($P(n|\lambda)$) times the probability that the true expectation value is λ in the case that b is the scaled number of Monte Carlo events ($P(\lambda|b)$), integrated over all possible values of λ.

The probability to dice n events if the expectation value λ is known, can be described by a Poisson distribution.

$$P(n|\lambda) = \frac{e^{-\lambda}\lambda^n}{n!}. \tag{8.6}$$

According to Bayes theorem [67] the probability for the hypothesis that a certain assumption for the expectation value is true, given the observation of b, reads:

$$P(\lambda|b) = \frac{P(b|\lambda) \cdot \pi(\lambda)}{\int d\lambda' P(b|\lambda') \cdot \pi(\lambda')}. \tag{8.7}$$

The prior probability π has to be estimated. There is a variety of models describing which prior is the most reasonable. Bayes himself suggested a uniform prior in the case of having no knowledge of the underlying distribution whatsoever. Obviously, this choice is not invariant under non-linear variable transformations. Jeffreys proposed a prior that is invariant under transformations of the form [68]

$$x \to x^n. \tag{8.8}$$

Another often considered approach is a prior function derived from the principle of maximum entropy [68]. The fact that a prior is necessary and the choice of it is the most criticised point

in Bayesian probability theory. In fact, in many cases the prior is unknown and can only be estimated.

For this study, a uniform prior function is considered. As mentioned above, it is not invariant under a non-linear transformation of the observed variable. But for a counting experiment this argument is less valid, as the number of events is a natural choice and is unlikely to be transformed in a non-linear way. One should also keep in mind that only reasonable bins are examined: An event with an energy which is much bigger than the center of mass energy of the colliding particles should never appear in the measurement. A uniform prior would not make sense in such a regime.

Using a uniform prior, equation 8.7 reads:

$$P(\lambda|b) = \frac{P(b|\lambda)}{\int d\lambda' P(b|\lambda')}. \tag{8.9}$$

In the case of large statistics and that the probability to find an event in the evaluated bin region is sufficiently smaller than 1, one can assume a Poisson distribution for $P(b|\lambda)$ (or $P(b|\lambda')$). This is usually true in the studied distributions as the samples contribute to several event classes each comprising a number of bins. The assumption might be inappropriate under some circumstances: If the algorithm examines a large bin region and most of the events of the samples are in that region. In that case, the statistics will be reasonably large. Later it will be shown, that this yields the convergence of the proposed p-value to the one described in section 6.5.1.

Usually the number of generated Monte Carlo events does not correspond to the integrated luminosity of the data and the process cross section. A rule of thumb is to use a factor of 10 times more Monte Carlo events than expected in data. Unfortunately this is not always possible. We introduced the factor α (equation 6.2) which is calculated from the integrated luminosity, cross section and the number of Monte Carlo events:

$$\alpha = \frac{N_{SM}^{\text{total}}}{\int \mathscr{L}\, dt \cdot \sigma_{\text{SM}}}. \tag{8.10}$$

It is used as the factor between the scaled number of Monte Carlo events and the number of Monte Carlo events in the considered region:

$$b = N_{SM}/\alpha. \tag{8.11}$$

With that, one obtains the scaled Poisson distribution in terms of N_{SM} instead of b:

$$P(b|\lambda) = P(N_{SM}|\lambda) = \alpha \frac{e^{-\alpha\cdot\lambda}(\alpha\cdot\lambda)^{N_{SM}}}{N_{SM}!}. \tag{8.12}$$

The integral in 8.9 can be calculated to

$$\int d\lambda'\, P(b|\lambda') = \int d\lambda'\, P(N_{SM}|\lambda') = \int d\lambda' \alpha \frac{e^{-\alpha\cdot\lambda'}(\alpha\cdot\lambda')^{N_{SM}}}{N_{SM}!} = 1. \tag{8.13}$$

So that from equation 8.9 one can conclude

$$P(\lambda|b) = P(b|\lambda) = \alpha \frac{e^{-\alpha\cdot\lambda}(\alpha\cdot\lambda)^{N_{SM}}}{N_{SM}!} \tag{8.14}$$

and therefore equation 8.5 yields:

$$P(n|N_{SM}) = \int_0^\infty d\lambda \frac{e^{-\lambda}\lambda^n}{n!} \cdot \alpha \frac{e^{-\alpha\cdot\lambda}(\alpha\cdot\lambda)^{N_{SM}}}{N_{SM}!}. \tag{8.15}$$

This integral can be solved analytically, one obtains the simple expression

$$P(n|N_{SM}) = \alpha^{N_{SM}+1} \cdot (\alpha+1)^{-n-N_{SM}-1} \cdot \binom{n+N_{SM}}{n}. \tag{8.16}$$

Neglecting a normalisation, this is can be transformed to the equation Przyborowski et al. [69] obtained. They did not explicitly use Bayesian statistics for their derivation which is another indication that the choice of the prior is reasonable.

In the special case of $\alpha = 1$ the expression reads:

$$P(n|N_{SM}) = \left(\frac{1}{2}\right)^{n+N_{SM}+1} \binom{n+N_{SM}}{n}. \tag{8.17}$$

Equation 8.16 does not take systematic errors into account. To describe them, we start from equation 8.15 again and introduce a normal distribution to modulate the Monte Carlo expectation value λ:

$$P = \int_0^\infty d\lambda \int_0^\infty d\mu \, \frac{e^{-\mu}\mu^n}{n!} \cdot A(\mu) \cdot e^{\frac{-(\mu-\lambda)^2}{2\sigma^2}} \alpha \frac{e^{-\alpha\cdot\lambda}(\alpha\cdot\lambda)^{N_{SM}}}{N_{SM}!}. \tag{8.18}$$

Where σ is the total systematic uncertainty and

$$A(\mu) = \sigma \int_\mu^\infty dt \, e^{-t^2/(2\sigma^2)} \tag{8.19}$$

is a normalisation due to the fact that the integrals only run over positive values.

As Bayes' theorem was used to derive this p-value, it is referred to as the Bayesian p-value. The p-value from section 6.5.1 is called the current p-value.

8.3. Implementation Details

MUSiC, as a model unspecific search, relies on a number of different Monte Carlo samples, generated using a variety of different simulation programs and techniques. For the 2010 data analysis 75 different samples were included to describe the data. Instead of equation 8.12, a convolution $C(\lambda, \alpha_1, N_{SM,1}, \ldots, \alpha_m, N_{SM,m}))$ of such distributions has to be determined, with different α_i and $N_{SM,i}$. Analogously to equation 8.18, one obtains:

$$P = \int_0^\infty d\lambda \int_0^\infty d\mu \, \frac{e^{-\mu}\mu^n}{n!} \cdot A(\mu) \cdot e^{\frac{-(\mu-\lambda)^2}{2\sigma^2}} \cdot C(\lambda, \alpha_1, N_{SM,1}, \ldots, \alpha_m, N_{SM,m}). \tag{8.20}$$

For two distributions the convolution C can be calculated by[2]:

$$C = \int d\lambda' \, \alpha_1 \frac{e^{-\alpha_1\cdot\lambda'}(\alpha_1\cdot\lambda')^{N_{SM,1}}}{N_{SM,1}!} \alpha_2 \frac{e^{-\alpha_2\cdot(\lambda-\lambda')}(\alpha_2\cdot(\lambda-\lambda'))^{N_{SM,2}}}{N_{SM,2}!}. \tag{8.21}$$

[2]Here, one integrates over all possible combinations that the sum of the expectation values of the two samples (λ' and $\lambda - \lambda'$) is equal to λ.

Convolutions of this form can be solved by [70]:

$$C \propto \int_0^y dx\, e^{ax} x^m e^{b\cdot(y-x)} (y-x)^n \tag{8.22}$$

$$= \int_0^y dx\, e^{(a-b)x} x^m e^{by} \sum_{i=0}^{n} (-x)^i y^{n-i} \tag{8.23}$$

$$= \sum_{i=0}^{n} (-1)^i e^{by} y^{n-i} \int_0^y dx\, e^{(a-b)x} x^{m+i} \tag{8.24}$$

$$= \sum_{i=0}^{n} (-1)^i e^{by} y^{n-i} \left[e^{(a-b)x} \left(\sum_{k=0}^{m+i} \frac{(-1)^k k! \binom{m+i}{k}}{(a-b)^{k+1}} x^{m+i-k} \right) \right]_{x=0}^{y} . \tag{8.25}$$

Which again is a sum of terms having the same form $A \cdot e^{ax} x^n$. It can therefore be recursively convoluted with other Monte Carlo distributions.

Although in principle the convolutions can be solved analytically, there are severe numerical issues with that. As the Poisson distribution has fast rising values in its nominator as well as in its denominator, it is usually not calculated as written. ROOT for example calculates

$$\text{Poisson}(n, \lambda) = \exp\left(n \cdot \ln(\lambda) - \lambda - \ln\Gamma(n+1)\right),$$

using the ln and $\ln\Gamma$ functions to avoid the division of large numbers. This trick is unfortunately not fully applicable to the convolution. By multiplying and summing different, partially very big and very small numbers, the calculation encounters numerical issues.

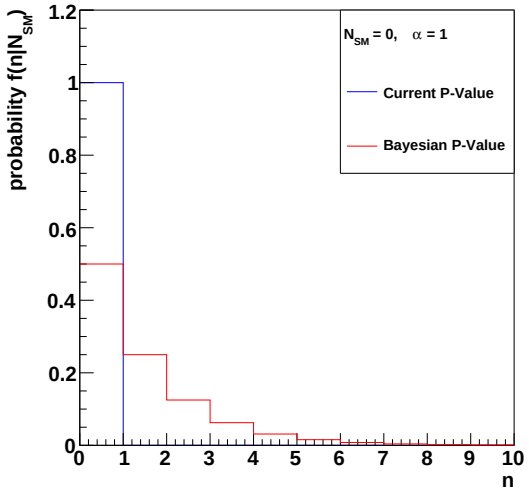
Several methods have been applied to solve this problem. Kahans summation algorithm [71] and arbitrary precision computation are two of them. Kahans summation algorithm takes the finite precision of the numbers into account while summing them up using a computer. Arbitrary precision computation increases the precision by using a lot more memory at the expense of calculation speed. Unfortunately these issues could not yet be completely resolved. Therefore a concrete application of the Bayesian p-value could not be carried out.

Figures 8.1 show the differences and similarities of the likelihood distributions used to calculate the Bayesian p-value and the current p-value. The motivation for the Bayesian p-value was the behaviour of the current p-value in the special case of $N_{SM} = 0$. The Bayesian p-value clearly accounts for this special case as can be seen in figure 8.1(a). Notably in figure 8.1(b), the probability distribution is shifted to higher values, especially for the case of small Monte Carlo statistics. In the case of $\alpha < 1$ the Monte Carlo events are scaled up (figure 8.1(c)). When the Monte Carlo statistics is high ($\alpha > 10$), the two probability distribution, and thus also the p-values, converge (figure 8.1(d)).

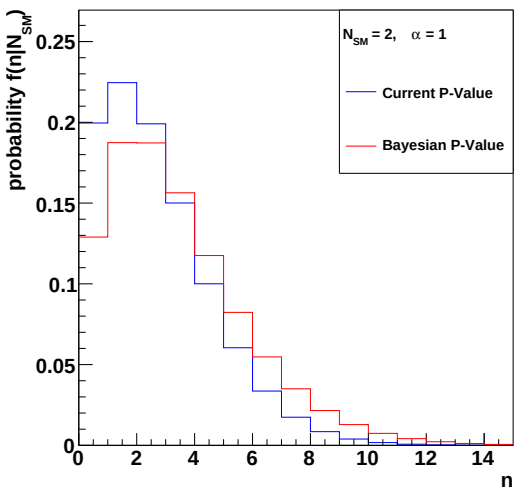
8.4. Coverage Tests

The following coverage tests describe the performance of the derived Bayesian p-value and compare it to the current p-value. Similar tests can be found in [48]. They have been conducted to compare a p-value with log-normally distributed uncertainties to the p-value with normally distributed uncertainties. Using varying uncertainties, these two p-values have been shown to perform comparably while each of them has its own advantages.

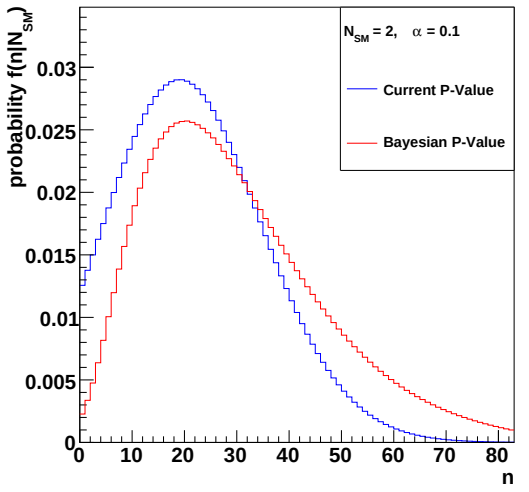
A p-value describes the probability to find a deviation at least as big as the one observed under the assumption that the hypothesis (the Monte Carlo simulation describes reality) is true. If the null hypothesis is in fact true, the p-values of a drawn sample must be uniformly distributed between 0 and 1. An inadequate model of the uncertainties and discretisation effects,



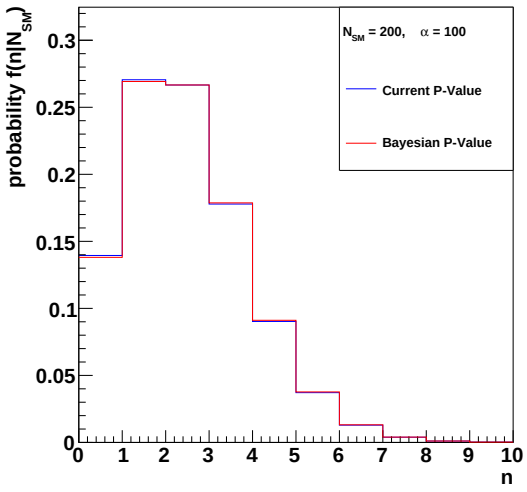
(a) $N_{SM} = 0, \alpha = 1$



(b) $N_{SM} = 2, \alpha = 1$



(c) $N_{SM} = 2, \alpha = 0.1$



(d) $N_{SM} = 200, \alpha = 100$

Figure 8.1.: Discrete probability distribution used to calculate the p-value for different parameter configurations. The systematic uncertainty is always chosen to be $\sigma = 0.1 \cdot N_{SM} \cdot \alpha$. This uncertainty is chosen as most of the uncertainties considered by MUSiC are relative uncertainties. This does not mean, that a distribution with equal relative uncertainties as a log-normal distribution is assumed, Instead a Gaussian distribution is used to model the systematic uncertainties.

being due to the fact that one considers integer event counts, can lead to a different distribution. In that case, the claim of the p-value, describing how big the discrepancy is, might be inadequate.

We can test the p-value by calculating pseudo experiments and measuring how many p-values are smaller than a given test value and how many should be smaller.

The coverage test is performed in term of the parameters $\langle N_{\text{data}} \rangle$, which is the expectation value of the number of events in a certain region, and α, which is the scaling parameter of the Monte Carlo sample and therefore a measure for the Monte Carlo statistics. The following describes the details of the algorithm. Its pseudo code can be found in section A.2 in the appendix.

The value of N_{data} is diced from a Poisson distribution with expectation value $\langle N_{\text{data}} \rangle$. To determine N_{SM} first the systematic uncertainties have to be taken into account. The expectation value $\langle N_{\text{SM}} \rangle$ is diced from a Gaussian distribution with mean $\langle N_{\text{data}} \rangle$ and standard deviation $\langle N_{\text{data}} \rangle \cdot \sigma_{\text{rel}}$. The systematic uncertainty $\langle N_{\text{data}} \rangle \cdot \sigma_{\text{rel}}$ is assumed to be relative to the expectation value. This is a valid estimation as most uncertainties are determined relative to the number of Monte Carlo events[3].

From the Monte Carlo expectation value, the number of Monte Carlo events is diced using a Poisson distribution, taking the scaling parameter α into account:

$$N_{\text{SM}} = \text{Poisson}(\langle N_{\text{data}} \rangle \cdot \alpha)/\alpha \tag{8.26}$$

where $\text{Poisson}(x)$ is a random sample from the Poisson distribution with expectation value x.

To calculate the p-value, the systematic uncertainty is assumed to be:

$$\sigma = N_{\text{SM}} \cdot \sigma_{\text{rel}}. \tag{8.27}$$

In the case of the current p-value, it is combined with Monte Carlo statistical uncertainty. In the case it is zero and the fill-up procedure is applied, it is set to $\frac{1}{\sqrt{N_{\text{fill-up}} \cdot \alpha}}$.

With that, the p-value can be calculated. Using the results of all pseudo experiments the value of

$$p_{\text{diced}} = \frac{\text{number of rounds with p-values with } 2 \cdot p \leq p_{\text{test}}}{\text{total number of rounds}} \tag{8.28}$$

is determined. An ideal p-value should be uniformly distributed. As a result, for a sufficiently large number of dicing rounds, considering any test value p_{test},

$$p_{\text{diced}} = p_{\text{test}} \tag{8.29}$$

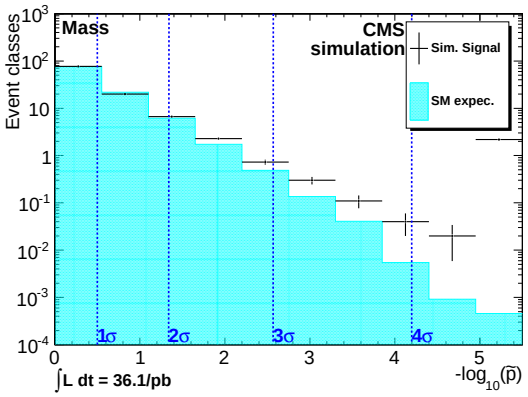
should hold true.

One should be aware that the possible values of p depend on the symmetry of the distribution. For a symmetric distribution, which is the case for a large number of events, the maximum value of p is 0.5. The factor 2 in equation 8.28 is necessary to account for that. The more asymmetric the distribution becomes, the higher p can become, up to 1 for a completely asymmetric distribution. In that case, the factor 2 is only an estimate.

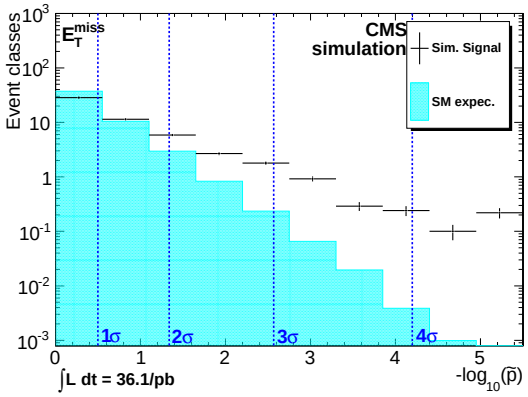
On average, the p-values can then either be too small, which means the p-value determination is too liberal, one says, the p-value has undercoverage. The opposite, overestimated p-values, thus a too conservative approach, is called overcoverage. While both deviations should be avoided, undercoverage is considered less desirable [72].

For a good coverage, the p-value must be uniformly distributed in the complete interval $[0, 1]$. By using the described coverage test, the uniformity is only checked for by comparing

[3] This does not mean, that a distribution with equal relative uncertainties as a log-normal distribution is assumed.



(a) P-value distribution of M_{inv} scans. In blue the expectation of a Standard Model Monte Carlo, the black markers show an added SSM Z' with mass $m = 500\,\text{GeV}$ scenario [1].



(b) P-value distribution of E_t^{miss} scans. In blue the expectation of a Standard Model Monte Carlo, the black markers show an added SUSY LM0 scenario [1].

Figure 8.2.: Two scenarios of new physics as seen by MUSiC. The SUSY scenario shows small excesses in a number of classes whereas the Z' shows a significant excess in just a couple of classes. Only exclusive classes were evaluated. The number of classes are expectation values.

the two regions of $p \leq p_{\text{test}}$ and $p > p_{\text{test}}$, instead of looking at the complete p-distribution. Therefore the choice of p_{test} is important. For the MUSiC analysis, two scenarios play an important role:

- In some scenarios a slight deviation in a large number of classes can be observed. This indication for new physics is unique to the MUSiC analysis as conventional analyses do not usually evaluate such a number of different classes. An example are SUSY scenarios with a lot of different decay channels. The $\tilde{p}$ distribution can be seen in figure 8.2(b). Here, from 2σ onwards, more deviations than expected in the Standard Model occur.

- The Z' scenario shows a different behaviour. The $\tilde{p}$ distribution follows the Standard Model expectation except for two classes showing a very high significance. This scenario is depicted in figure 8.2(a). Usually one speaks of a discovery when observing a difference of at least 5 Gaussian standard deviations, which corresponds to a p-value of $5.7 \cdot 10^{-7}$.

These scenarios already incorporate the trial factor due to the look elsewhere effect. The 2σ deviations from the SUSY scenario approximately correspond to a 3σ deviation in the original p-value (figure 8.3). This requires MUSiC to have a good p-value coverage in the low σ regime as well as in the high σ (low p) regime. The lower p-values are obviously the most crucial for this analysis. Unfortunately, their coverage is hard to estimate because a lot of pseudo experiments would be necessary. Therefore, we conduct the test for $p = 0.05$ and $p = 0.001$.

For small p-values, the discretisation of the p-value spectrum is less dominant. The reason for that is that there are more possibilities to encounter a small p-value as there are more different event counts corresponding to small p-values than to high p-values. The p-value range is denser near 0. So at least the discretisation effect should not worsen the p-value coverage for small values of p_{test}.

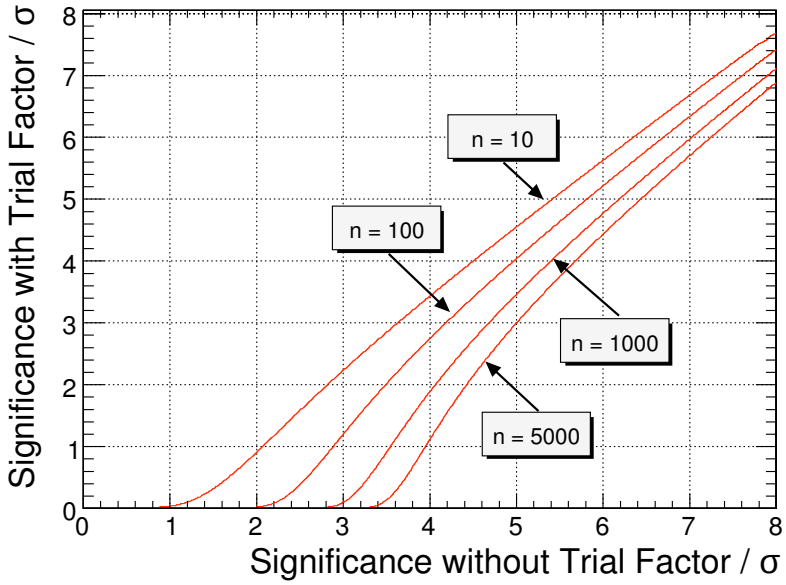


Figure 8.3.: Significances with and without trial factor, thus with and without taking the look elsewhere effect into account [19].

8.5. Results of the Coverage Test

Figures 8.4, 8.5, 8.6 and 8.7 show the coverage of the Bayesian p-value, the current p-value without fill-up (section 6.5.1) and with a typical fill-up scenario. For the fill-up scenario we assume that the uncertainty fill-up is distributed over 100 bins, which is the case when the last filled bin is at 1000 GeV. This corresponds to the highest filled bins in 2010 data. This assumption is conservative as a fill-up with less bins leads to a lower uncertainty.

The coverage is described as $\log_{10} \frac{p_{\text{diced}}}{p_{\text{test}}}$. It can be seen as the number of magnitudes lying between p_{diced} and p_{test}.

The current p-value (figure 8.6) shows a clear undercoverage in the case of low statistics. As expected, the regime of low α and $b = 0$ is too liberal. This is the reason to the introduce the fill-up and the Bayesian p-value. For sufficiently large statistics, the p-value becomes conservative. There is a small region in which the coverage is correct. This is due to the transition from the conservative to the liberal regime.

When looking at the coverage of the Bayesian p-value one can see a different behaviour. Instead of an undercoverage in the the low statistics regime, the p-value is too conservative. Especially noticeable is the an area with an extreme undercoverage in the form of a nose. When looking at a broader range (figure 8.5 one can see that this phenomenon is locally confined. For even lower statistics the coverage becomes significantly better. One can assume that it is due to a discretisation effect. The p-value corresponding to a certain event count approaches the threshold of 0.05 and at the point when it passes and becomes smaller than 0.05 the coverage becomes less conservative. This hypothesis is encouraged by the fact, that the area becomes bigger for smaller values of p_{test} (figure 8.8).

In the case of good statistics (large α) one can see the same coverage for all p-values. The reason for this is the fact that the likelihoods converge in this regime, as can be seen in figure 8.1(d). It can be noted that also in regimes where the statistics are very high, still undercoverage dominates for $b < 3$. This changes for higher values of b and from about $b > 4$ the coverage is very good regardless of α (figure 8.5).

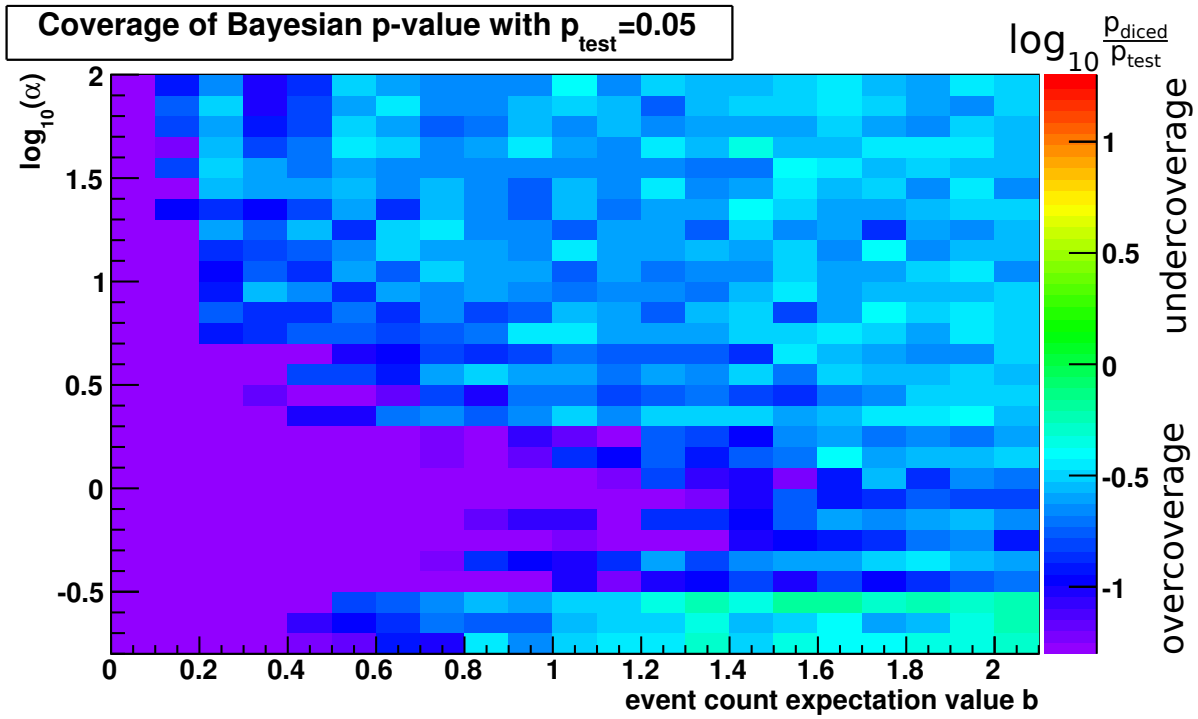


Figure 8.4.: Coverage of the Bayesian p-value for $p_{\text{test}} = 0.05$ (algorithm A.1 in the appendix).

To test the fill-up p-value properly is difficult as the fill-up depends on the observed resolution and not just the single bin tested. As an example, a fill-up scenario with 100 bins is tested (figure 8.7). Two discontinuities can be observed in figure 8.7 at around $\log(\alpha) = 0.6$ and $\log(\alpha) = -0.5$. It can be speculated that these are due to the property of the algorithm that fills the uncertainty only in the case when now events are observed. This is a hardly predictable and discontinously behaviour.

Figure 8.8 shows the coverage plot of the Bayesian p-value for $p_{\text{test}} = 0.001$ with a small resolution. An overcoverage with similar properties as the $p_{\text{test}} = 0.05$ can be seen. But in this example the overcoverage is also extended to higher values of b and smaller values of α. Just at around $b = 5$ the p-values covers well. The coverage is still conservative in the whole regime.

8.6. Conclusion

An alternative p-value has been introduced to MUSiC. It has been derived by using Bayes' theorem to take the finite statistics of the Monte Carlo into account. It assumes a Poisson distribution for the statistical uncertainty of the number of Monte Carlo events and a Gaussian distribution for systematic uncertainties. Some numerical issues could not be resolved yet and prevented the Bayesian p-value from being implemented and used by the MUSiC analysis.

Coverage test have been conducted to evaluate three different p-values: The current p-value without fill-up, the current p-value with a 100 bins fill-up and the Bayesian p-value. The first one is liberal (i.e. it shows undercoverage) for low statistics and therefore not appropriate. The p-value with fill-up is mostly conservative and is well suited to be used by MUSiC. The Bayesian p-value has a good coverage over most of the α-b plane, shows a predictable behaviour and would be a good alternative to the fill-up method if it was feasible.

One can think of other problems, outside of MUSiC, where the Bayesian p-value can be

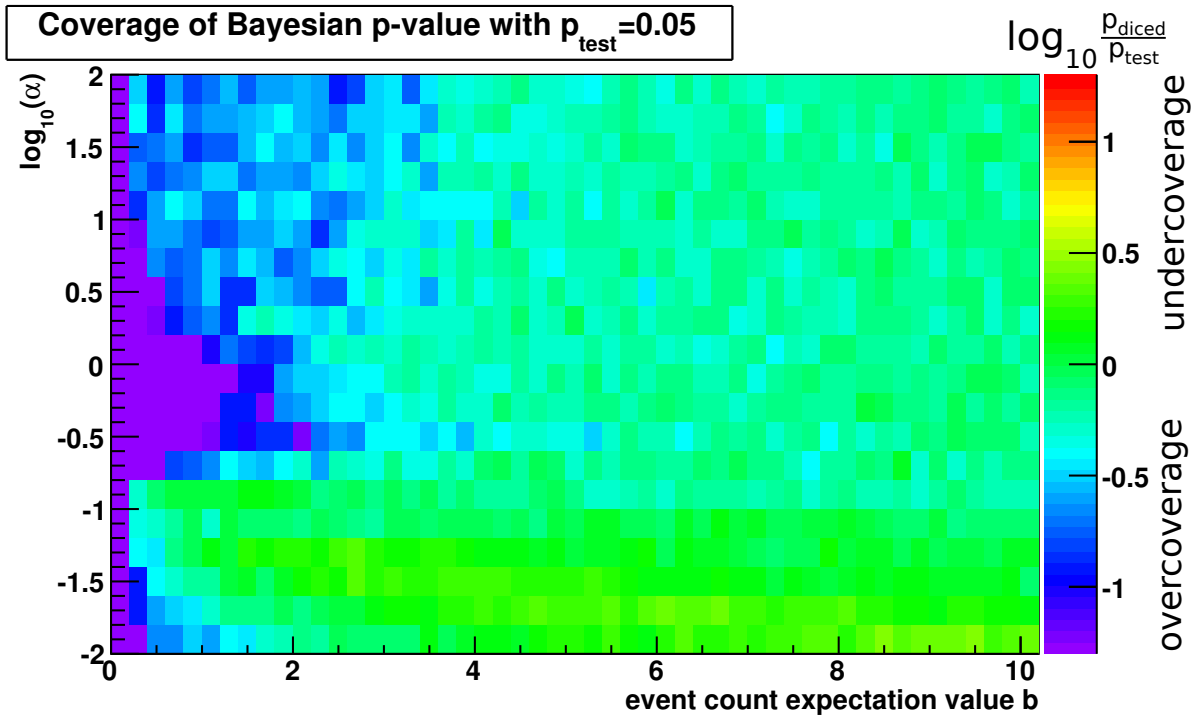


Figure 8.5.: Coverage of the Bayesian p-value for $p_{\text{test}} = 0.05$ covering a large range of α and b (algorithm A.1 in the appendix).

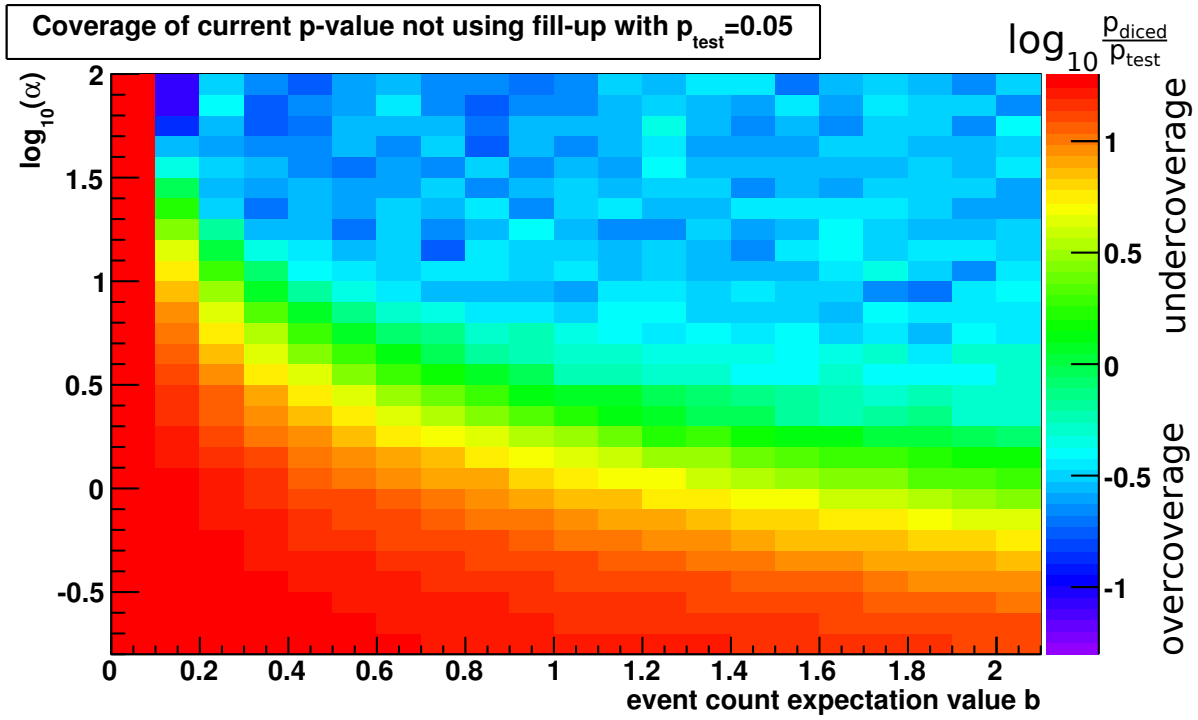


Figure 8.6.: Coverage of the current p-value without fill-up for $p_{\text{test}} = 0.05$ (algorithm A.2 in the appendix).

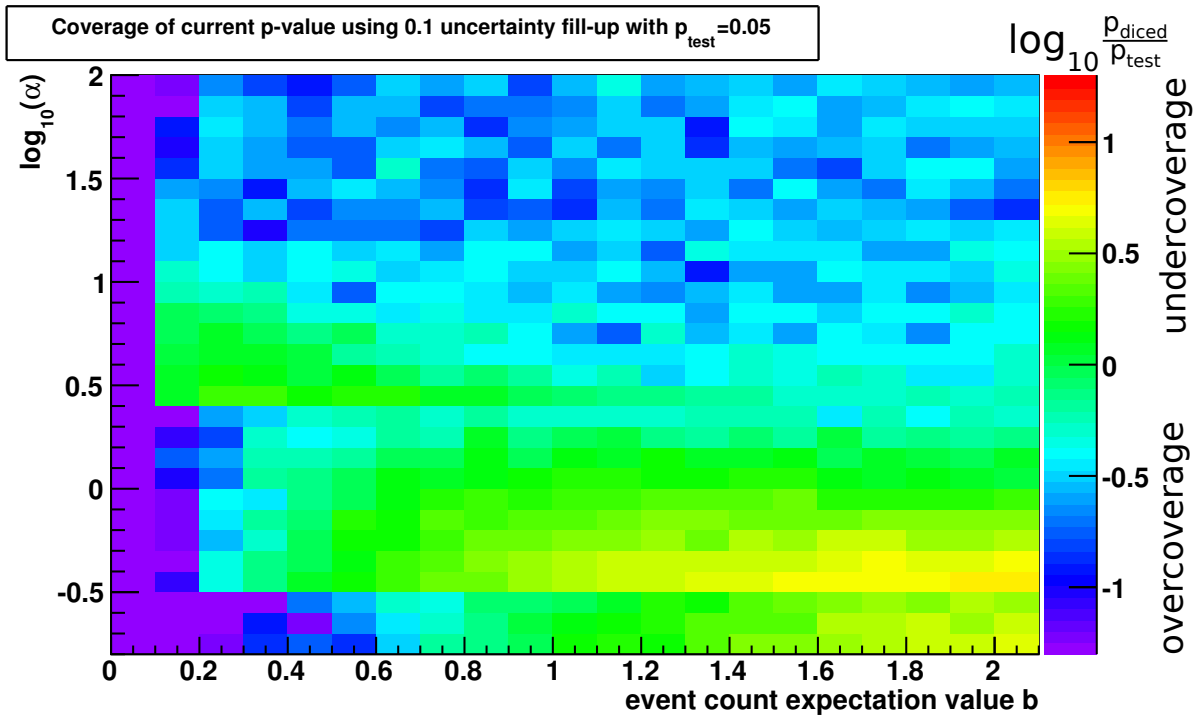


Figure 8.7.: Coverage of the current p-value with fill-up for $p_{\text{test}} = 0.05$ (algorithm A.2 in the appendix).

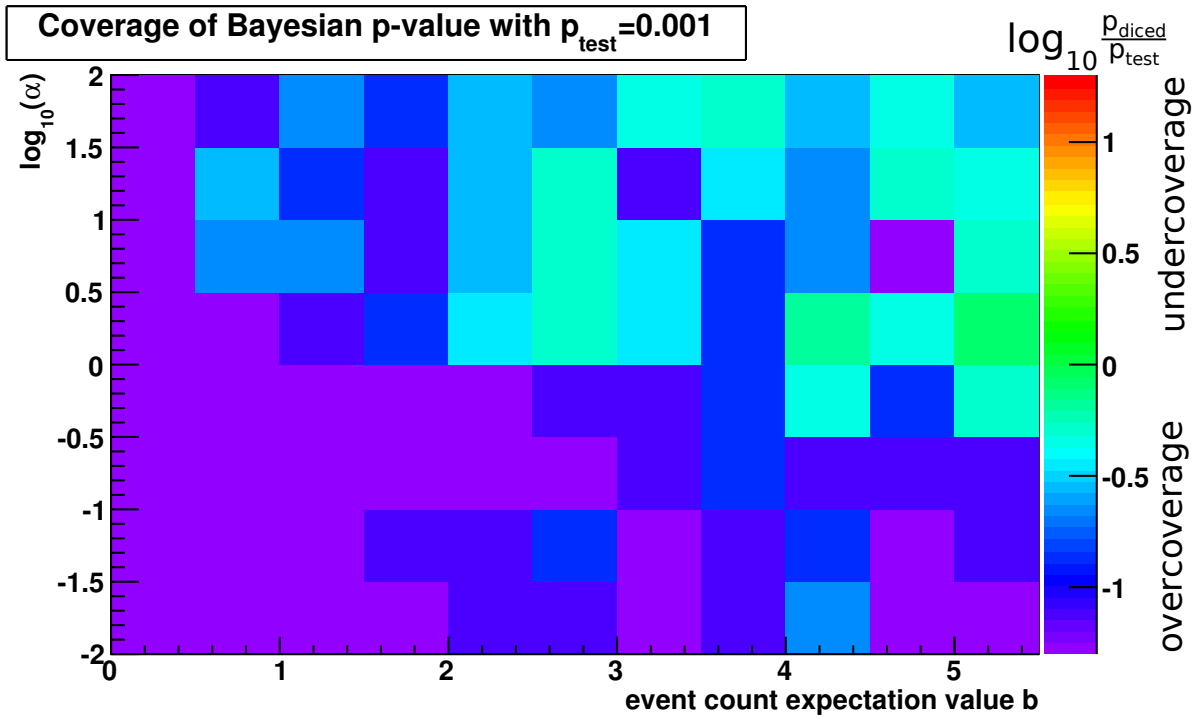


Figure 8.8.: Coverage of the Bayesian p-value for $p_{\text{test}} = 0.001$ (algorithm A.1 in the appendix).

used: When having a number of test values with different statistics but which are not built from different samples, the Bayesian p-value is applicable.

For MUSiC, the current implementation – the p-value with fill-up – is the most suitable and applicable method to today's knowledge.

9. Conclusions and Outlook

In this work, studies of alternative statistical methods for a model independent analysis of CMS data have been conducted. Two innovative ideas have been presented:

- The Bump Hunter is a complementary approach to find abnormalities, especially resonances in the m_t spectrum. An algorithm has been developed and commissioned. Several test scenarios have shown that it is capable of finding bumps. The CMS data of 2010 has been scanned. Some interesting bumps have been found.
- The Bayesian p-value is a different ansatz to calculate a p-value. It is more complex than the one currently used by MUSiC. Its coverage performance has been examined and compared to the current p-value with and without uncertainty fill-up of empty bins. The Bayesian p-value shows a good coverage for a large range of expected events and Monte Carlo statistics without the aid of a separate fill-up procedure for underestimated uncertainties. It could not yet be completely implemented into MUSiC and therefore could not be tested in a practical scenario.

Not having enough Monte Carlo statistics is already an issue with 2010 data and will be a major challenge for the Model Unspecific Search in CMS in the near future. Improved statistical methods will not be able to compensate for that, unlike the apocryphal student lab report states: "We didn't have enough data so we had to use statistics." The MUSiC team puts a great effort into finding a solution for that issue. Possible ideas are the development of a "QCD from data" algorithm or a massive production of QCD Monte Carlo samples.

A. Appendix

A.1. Cross Sections

Table A.1.: Standard model processes used for comparison with the 2010 data in MUSiC [31]. The cross section given is the one used in the analysis. This includes a possible k-factor to consider a higher order cross section than used during the production (see comments).

Data set	σ/pb	N_{events}	Generator	Comments
Upsilon1SToEE	12.2 k	230 k	PYTHIA	LO, cross section measured [73]
Upsilon2SToEE	8.8 k	220 k	PYTHIA	LO, cross section measured [73]
Upsilon3SToEE	2.3 k	110 k	PYTHIA	LO, cross section measured [73]
Upsilon1SToMuMu	14.3 k	2.3 M	PYTHIA	LO, cross section measured [73]
Upsilon2SToMuMu	6.0 k	1.1 M	PYTHIA	LO, cross section measured [73]
Upsilon3SToMuMu	1.6 k	930 k	PYTHIA	LO, cross section measured [73]
QCD_Pt0to5	47.6 G	550 k	PYTHIA	LO
QCD_Pt5to15	36.7 G	1.6 M	PYTHIA	LO
QCD_Pt15to30	810.6 M	5.4 M	PYTHIA	LO, overlap removed
QCD_Pt30to50	49.7 M	3.3 M	PYTHIA	LO, overlap removed
QCD_Pt50to80	5.5 M	3.2 M	PYTHIA	LO, overlap removed
QCD_Pt80to120	640.1 k	3.2 M	PYTHIA	LO, overlap removed
QCD_Pt120to170	92.3 k	3.0 M	PYTHIA	LO, overlap removed
QCD_Pt170to300	23.5 k	3.2 M	PYTHIA	LO, overlap removed
QCD_Pt300to470	1.1 k	3.2 M	PYTHIA	LO, overlap removed
QCD_Pt470to600	67.3	2.0 M	PYTHIA	LO, overlap removed
QCD_Pt600to800	14.9	2.0 M	PYTHIA	LO, overlap removed
QCD_Pt800to1000	1.8	2.1 M	PYTHIA	LO, overlap removed
QCD_Pt1000to1400	319.0 m	1.1 M	PYTHIA	LO, overlap removed
QCD_Pt1400to1800	10.5 m	1.0 M	PYTHIA	LO, overlap removed
QCD_Pt1800	345.6 μ	530 k	PYTHIA	LO, overlap removed
QCD_BC_Pt20to30	129.7 k	2.2 M	PYTHIA	LO, overlap removed
QCD_BC_Pt30to80	131.4 k	2.0 M	PYTHIA	LO, overlap removed
QCD_BC_Pt80to170	8.6 k	1.0 M	PYTHIA	LO, overlap removed
QCD_EM_Pt20to30	2.4 M	37.2 M	PYTHIA	LO, overlap removed
QCD_EM_Pt30to80	3.8 M	71.8 M	PYTHIA	LO, overlap removed

QCD_EM_Pt80to170	137.2 k	8 M	PYTHIA	LO, overlap removed
QCD_Mu_Pt15to20	1.5 M	2.8 M	PYTHIA	LO
QCD_Mu_Pt20to30	1.2 M	11.4 M	PYTHIA	LO
QCD_Mu_Pt30to50	578.5 k	11.4 M	PYTHIA	LO
QCD_Mu_Pt50to80	144.4 k	10.7 M	PYTHIA	LO
QCD_Mu_Pt80to120	29.0 k	3.1 M	PYTHIA	LO
QCD_Mu_Pt120to150	4.4 k	1.0 M	PYTHIA	LO
QCD_Mu_Pt150	2.8	1.0 M	PYTHIA	LO ($150 < \hat{p}_T/\mathrm{GeV} < \infty$)
G_Pt0to15	84.2 M	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt15to30	171.7 k	1.0 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt30to50	16.7 k	1.0 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt50to80	2722	1.0 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt80to120	447.2	1.0 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt120to170	84.2	1.0 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt170to300	22.6	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt300to470	1.5	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt470to800	0.1	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt800to1400	3.5 m	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt1400to1800	12.7 μ	1.1 M	PYTHIA	LO ($\gamma +$ jets)
G_Pt1800	293.6 n	1.1 M	PYTHIA	LO ($\gamma +$ jets) ($1800 < \hat{p}_T/\mathrm{GeV} < \infty$)
W+ Jets	24.4 k	15 M	MADGRAPH	k-Factor NNLO/LO = 1.28
DYToMuMu_M2To10	93.6 k	2.1 M	PYTHIA	k-Factor NNLO/LO = 1.24
DYJetsToLL_M10to50	413	188 k	MADGRAPH	k-Factor NNLO/LO = 1.24
Z+0 Jets	2.4 k	1.4 M	ALPGEN	k-Factor NNLO/LO = 1.24
Z+1 Jets ($p_T^Z 0-100$)	470	320 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+1 Jets ($p_T^Z 100-300$)	11	270 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+1 Jets ($p_T^Z 300-800$)	90 m	110 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+1 Jets ($p_T^Z 800-1600$)	170 μ	33 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+2 Jets ($p_T^Z 0-100$)	130	120 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+2 Jets ($p_T^Z 100-300$)	11	130 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+2 Jets ($p_T^Z 300-800$)	140 m	110 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+2 Jets ($p_T^Z 800-1600$)	370 μ	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+3 Jets ($p_T^Z 0-100$)	28	55 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+3 Jets ($p_T^Z 100-300$)	4.9	55 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+3 Jets ($p_T^Z 300-800$)	100 m	54 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+3 Jets ($p_T^Z 800-1600$)	310 μ	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+4 Jets ($p_T^Z 0-100$)	5.7	44 k	ALPGEN	k-Factor NNLO/LO = 1.24

Z+4 Jets ($p_T^Z 100-300$)	1.6	44 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+4 Jets ($p_T^Z 300-800$)	49 m	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+4 Jets ($p_T^Z 800-1600$)	170 μ	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+5 Jets ($p_T^Z 0-100$)	1.4	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+5 Jets ($p_T^Z 100-300$)	590 m	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+5 Jets ($p_T^Z 300-800$)	24 m	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
Z+5 Jets ($p_T^Z 800-1600$)	89 μ	11 k	ALPGEN	k-Factor NNLO/LO = 1.24
TT	165	1.1 M	PYTHIA	NNLL
WW	43.0	2.1 M	PYTHIA	NLO
WZ	18.2	2.2 M	PYTHIA	NLO
ZZ	5.9	2.1 M	PYTHIA	NLO
PhotonVJets	173	1.1 M	MADGRAPH	LO

A.2. Coverage Algorithms

Algorithm A.1 Calculation of the Bayesian p-value coverage.

set the test value e.g. to $p_{\text{test}} = 0.05$
set the relative systematic uncertainty $\sigma_{\text{sys}}^{\text{rel}} = 10\%$
for all data expectation values $\langle N_{\text{data}} \rangle$ and all scale factors α **do**
  for a sufficent number of times **do**
    dice the expectation value $\langle N_{\text{SM}} \rangle$ from a Gaussian distribution with mean $\langle N_{\text{data}} \rangle$ and standard deviation $\langle N_{\text{data}} \rangle \cdot \sigma_{\text{sys}}^{\text{rel}}$
    dice N_{SM} from a scaled Poisson distribution $N_{\text{SM}} = \text{Poisson}(\langle N_{\text{SM}} \rangle \cdot \alpha)/\alpha$
    dice N_{data} from a Poisson distribution with mean $\langle N_{\text{data}} \rangle$
    set the systematic uncertainty to $\sigma_{\text{sys}} = \sigma_{\text{sys}}^{\text{rel}} \cdot N_{\text{SM}}$
    calculate the Bayesian p-value p from N_{data}, N_{SM}, α, and σ_{sys}
    count how often $2 \cdot p \leq p_{\text{test}}$
  end for
  calculate $p_{\text{diced}} = \frac{\text{number of rounds with p-values with } 2 \cdot p \leq p_{\text{test}}}{\text{total number of rounds}}$
  plot $\log_{10} \frac{p_{\text{diced}}}{p_{\text{test}}}$
end for

Algorithm A.2 Calculation of the current p-value coverage with and without fill-up.

set the test value e.g. to $p_{\text{test}} = 0.05$
set the relative systematic uncertainty $\sigma_{\text{sys}}^{\text{rel}} = 10\%$
set the number of filled up bins to $N_{\text{fill-up}}$
for all data expectation values $\langle N_{\text{data}} \rangle$ and all scale factors α **do**
  for a sufficent number of times **do**
    dice the expectation value $\langle N_{\text{SM}} \rangle$ from a Gaussian distribution with mean $\langle N_{\text{data}} \rangle$ and standard deviation $\langle N_{\text{data}} \rangle \cdot \sigma_{\text{sys}}^{\text{rel}}$
    dice N_{SM} from a scaled Poisson distribution $N_{\text{SM}} = \text{Poisson}(\langle N_{\text{SM}} \rangle \cdot \alpha)/\alpha$
    dice N_{data} from a Poisson distribution with mean $\langle N_{\text{data}} \rangle$
    set the systematic uncertainty to $\sigma_{\text{sys}} = \sqrt{(\sigma_{\text{sys}}^{\text{rel}} \cdot N_{\text{SM}})^2 + (N_{\text{SM}}/\alpha)}$
    if fill-up **and** $\sigma_{\text{sys}} = 0$ **then**
      set the systematic uncertainty to $\sigma_{\text{sys}} = \frac{1}{\sqrt{N_{\text{fill-up}}} \cdot \alpha}$
    end if
    calculate the current p-value p from N_{data}, N_{SM}, and σ_{sys}
    count how often $2 \cdot p \leq p_{\text{test}}$
  end for
  calculate $p_{\text{diced}} = \frac{\text{number of rounds with } 2 \cdot p \leq p_{\text{test}}}{\text{total number of rounds}}$
  plot $\log_{10} \frac{p_{\text{diced}}}{p_{\text{test}}}$
end for

Bibliography

[1] CMS Collaboration, "Model Unspecific Search for New Physics in pp Collisions at sqrt(s) = 7 TeV", *CMS PAS* **EXO-10-021** (2011). http://cdsweb.cern.ch/record/1360173/files/EXO-10-021-pas.pdf.

[2] Nobel Media Collaboration, "The Nobel Prize in Physics 2008", May, 2011. http://nobelprize.org/nobel_prizes/physics/laureates/2008/.

[3] T. Hebbeker, "Skriptum zur Vorlesung Elementarteilchenphysik I,". RWTH SS 2007.

[4] "Standard Model of Elementary Particles", June, 2006. http://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[5] D. Griffiths, "Introduction to Elementary Particles". Wiley-VCH, 2008.

[6] K. Nakamura et al. (Particle Data Group), "Particle Physics Booklet". J. Phys. G 37, 075021, 2010.

[7] LEP Collaboration, "The LEP Electroweak Working Group", July, 2010. http://lepewwg.web.cern.ch/LEPEWWG/.

[8] J. Ellis, "Beyond the Standard Model for Hillwalkers", arXiv:hep-ph/9812235.

[9] G. Altarelli, "The Standard Model and Beyond", arXiv:hep-ph/9809532.

[10] G. Costa, J. Ellis, G. L. Fogli et al., "Neutral currents within beyond the standard model", *Nuclear Physics B* **297** (1988), no. 2, 244 – 286. doi:DOI: 10.1016/0550-3213(88)90020-X.

[11] P. Langacker, "The Physics of Heavy Z' Gauge Bosons", *Rev.Mod.Phys.* **81:1199-1228,2008** (January, 2008) arXiv:0801.1345.

[12] U. Baur, M. Spira, and P. M. Zerwas, "Excited-quark and -lepton production at hadron colliders", *Phys. Rev. D* **42** (Aug, 1990) 815–824. doi:10.1103/PhysRevD.42.815.

[13] K. Nakamura and P. D. Group, "Review of Particle Physics", *Journal of Physics G: Nuclear and Particle Physics* **37** (2010), no. 7A, 075021.

[14] M. Krämer, T. Plehn, M. Spira et al., "Pair production of scalar leptoquarks at the LHC", *Phys.Rev.D* **71:057503,2005** (2005) arXiv:hep-ph/0411038.

[15] CMS Collaboration, "Search for First Generation Scalar Leptoquarks in the evjj channel in pp collisions at sqrt(s) = 7 TeV", arXiv:1105.5237.

[16] W. Buchmüller, R. Rückl, and D. Wyler, "Leptoquarks in lepton-quark collisions", *Physics Letters B* **191** (1987), no. 4, 442 – 448. doi:DOI: 10.1016/0370-2693(87)90637-X.

[17] (ed.) L. Evans and (ed.) P. Bryant, "LHC Machine", *JINST* **3** (2008) S08001. doi:10.1088/1748-0221/3/08/S08001.

[18] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004. doi:10.1088/1748-0221/3/08/S08004.

[19] C. Hof, "Implementation of a Model-Independent Search for New Physics with the CMS Detector exploiting the World-Wide LHC Computing Grid". PhD thesis, RWTH Aachen, October, 2009. http://web.physik.rwth-aachen.de/~hebbeker/theses/hof_phd.pdf.

[20] D. W. Kerst, F. T. Cole, H. R. Crane et al., "Attainment of Very High Energy by Means of Intersecting Beams of Particles", *Phys. Rev.* **102** (Apr, 1956) 590–591. doi:10.1103/PhysRev.102.590.

[21] LHC Collaboration, "LHC Design report", 05, 2011. http://lhc.web.cern.ch/lhc/LHC-DesignReport.html.

[22] CMS Collaboration, "CMS Luminosity Collision Data 2010", June, 2011. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults2010.

[23] P. G. R. Claude Leroy, "Principles of radiation interaction in matter and detection". World Scientific, 2004.

[24] CMS-HCAL Collaboration, "Studies of the response of the prototype CMS hadron calorimeter, including magnetic field effects, to pion, electron, and muon beams", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **457** (2001), no. 1-2, 75 – 100. doi:DOI: 10.1016/S0168-9002(00)00711-7.

[25] CMS Collaboration, "The CMS Physics Technical Design Report, Volume 1", *CERN/LHCC* **2006-001** (2006).

[26] WLCG Collaboration, "Worldwide LHC Computing Grid Technical Site", June, 2011. http://lcg.web.cern.ch/lcg/.

[27] CMS Collaboration, "How to Configure and Run Detector Simulation and Digitization", June, 2011. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookSimDigi.

[28] O. Actis, M. Erdmann, A. Hinzmann et al., "PXL: Physics eXtension Library", June, 2011. http://pxl.sourceforge.net/doxygen3a/.

[29] Root Collaboration, "User's Guide v5.26", June, 2011. http://root.cern.ch/download/doc/Users_Guide_5_26.pdf.

[30] GSL Collaboration, "GNU Scientific Library – Reference Manual", June, 2011. http://www.gnu.org/software/gsl/manual/html_node/.

[31] T. Hebbeker, S. Malhotra, A. Meyer et al., "MUSiC - A model independent search with 2010 data", *CMS AN* **2011/042** (2011).

[32] CMS Collaboration, "Slice of the CMS detector", June, 2011. http://cms.web.cern.ch/cms/Resources/Website/Media/Videos/Animations/files/CMS_Slice.gif.

[33] G. Abbiendi, N. Adam, J. Alcaraz et al., "Muon Reconstruction in the CMS Detector", Technical Report CMS-NOTE-2008-097, CERN, July, 2009.

[34] S. Baffioni, C. Charlot, F. Ferri et al., "Electron reconstruction in CMS", Technical Report CMS-NOTE-2006-040, Feb, 2006.

[35] E. Meschi, T. Monteiro, C. Seez et al., "Electron Reconstruction in the CMS Electromagnetic Calorimeter", Technical Report CMS-NOTE-2001-034, CERN, Jun, 2001.

[36] CMS Collaboration, "HEEP Electron ID and isolation", June, 2011. `https://twiki.cern.ch/twiki/bin/viewauth/CMS/HEEPElectronID`.

[37] J. Nysten, "Photon Reconstruction in CMS", *CMS CR* **2004/004** (2004).

[38] G. P. Salam and M. Cacciari, "Jet clustering in particle physics, via a dynamic nearest neighbour graph implemented with CGAL", *unpublished* (March, 2006). `http://www.lpthe.jussieu.fr/~salam/repository/docs/kt-cgta-v2.pdf`.

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

[40] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET", *CMS PAS* **PFT-09-001** (Apr, 2009).

[41] F. Pandolfi and D. del Re, "Particle Flow Jet Composition", *CMS AN* **2010/005** (January, 2010).

[42] The CMS Particle Flow Physics Object Group Collaboration, "Particle Flow Reconstruction of Jets, Taus, and MET", *CMS AN* **2009/039** (2009).

[43] T. Hebbeker, "A Global Comparison between L3 Data and Standard Model Monte Carlo - a first attempt", *L3 Note* **2305** (1998). `http://web.physik.rwth-aachen.de/~hebbeker/l3note_2305.pdf`.

[44] H1 Collaboration, "A general search for new phenomena in ep scattering at HERA", *Physics Letters B* **602** (2004), no. 1-2, 14 – 30. `doi:DOI: 10.1016/j.physletb.2004.09.057`.

[45] G. Choudalakis, "Model Independent Search For New Physics At The Tevatron", `arXiv:0805.3954`.

[46] D0 Collaboration, "Quasi-Model-Independent Search for New High pT Physics at D0", *Phys. Rev. Lett.* **86** (Apr, 2001) 3712–3717. `doi:10.1103/PhysRevLett.86.3712`.

[47] P. A. Biallass, "Commissioning of the CMS Muon Detector and Development of Generic Search Strategies for New Physics". PhD thesis, RWTH Aachen, March, 2009. `http://web.physik.rwth-aachen.de/~hebbeker/theses/biallass_phd.pdf`.

[48] S. A. Schmitz, "Model Unspecific Search for New Physics with High p_t Photons in CMS", Diplomarbeit, RWTH Aachen, October, 2009. `http://web.physik.rwth-aachen.de/~hebbeker/theses/schmitz_diploma.pdf`.

[49] E. Dietz-Laursonn, "Model Unspecific Search for New Physics with b-Hadrons in CMS", Diplomarbeit, RWTH Aachen, October, 2010. `http://web.physik.rwth-aachen.de/~hebbeker/theses/dietz-laursonn_diploma.pdf`.

[50] CMS Collaboration, "MUSIC – An Automated Scan for Deviations between Data and Monte Carlo Simulation", *CMS PAS* **EXO-08/005** (Oct, 2008).

[51] CMS Collaboration, "Determination of the Jet Energy Resolutions and Jet Reconstruction Efficiency at CMS", *CMS PAS* **JME-09-007** (Jul, 2009).

[52] F. Rebassoo, "Monte Carlo study of 7 TeV MinBias: MET resolution vs. sumEt", October, 2010. https://indico.cern.ch/getFile.py/access?contribId=7&resId=0&materialId=slides&confId=95048.

[53] J. L. Devore, "Probability and statistics for engineering and the sciences". Thomson Brokks/Cole, 2008.

[54] CMS Collaboration, "Absolute luminosity normalization", *CMS DP* **2011-002** (March, 2011).

[55] D. Stump, J. Huston, J. Pumplin et al., "Inclusive Jet Production, Parton Distributions, and the Search for New Physics", *JHEP* **0310** (2003) 046, arXiv:hep-ph/0303013.

[56] J. Pumplin, D. R. Stump, J. Huston et al., "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis", *JHEP* **0207:012,2002** (2002).

[57] CMS Collaboration, "CMS Muon Results Twiki Page", May, 2011. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsMUO.

[58] J. H. Friedman and N. I. Fisher, "Bump hunting in high-dimensional data", *Statistics and Computing* **9** (April, 1999) 123–143.

[59] G. Choudalakis, "On hypothesis testing, trials factor, hypertests and the BumpHunter", arXiv:1101.0390.

[60] A. Ferapontov, G. Landsberg, and P. Tsang, "Searches for Black Holes Production in pp Collisions at sqrt(s) = 7 TeV with CMS Detector", *CMS AN* **2010/313** (November, 2010).

[61] G. Cowan, "Statistical Data Analysis". Oxford Science Publications, 1998.

[62] J. G. Heinrich, "The Log Likelihood Ratio of the Poisson Distribution for Small mu", *CDF Note* **5718** (2001).

[63] CMS Collaboration, "Fall 2010 CMS MonteCarlo Production (7 TeV)", July, 2011. https://twiki.cern.ch/twiki/bin/view/CMS/ProductionFall2010.

[64] M. Gataullin, M. S. Soares, and Y. Yang, "Search for excited muon resonance with the CMS detector", *CMS AN* **2010/319** (2011).

[65] O. V. B. K. K. Emanuela Barberis, Darin Carl Baumgartel, "Search for Pair Production of Second Generation Scalar Leptoquarks with the CMS detector", *CMS AN* **2010/255** (2010).

[66] CMS Collaboration, "Performance of Methods for Data-Driven Background Estimation in SUSY Searches", *CMS PAS* **SUS-10-001** (2010).

[67] F. James, "Statistical Methods in Experimental Physics". World Scientific Publishing, 2010.

[68] E. Jaynes, "Prior Probabilities", *Systems Science and Cybernetics, IEEE Transactions on* **4** (sept., 1968) 227 –241. doi:10.1109/TSSC.1968.300117.

[69] P. Przyborowski and H. Wilenski, "Homogenity of Results in Testing Samples from Poisson Series: With an Application to Testing Clover Seed for Dodder", *Biometrika* **31** (March, 1940) 313–323.

[70] I. S. Gradshteyn and I. M. Ryzhik, "Table of integrals, series, and products". Elsevier/Academic Press, Amsterdam, seventh edition, 2007.

[71] W. Kahan, "Pracniques: further remarks on reducing truncation errors", *Commun. ACM* **8** (January, 1965) 40–. doi:10.1145/363707.363723.

[72] R. D. Cousins, J. T. Linnemann, and J. Tucker, "Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process", *Nuclear Instruments and Methods in Physics Research A* (2008) 480–501, arXiv:physics/0702156.

[73] CMS Collaboration, "Measurement of the Inclusive Upsilon production cross section in pp collisions at sqrt(s)=7 TeV", arXiv:1012.5545.

Selbstständigkeitserklärung

Ich, Mark Olschewski, habe diese Arbeit ohne fremde Hilfe und ohne Verwendung anderer als der hier angegebenen Quellen und Hilfsmittel angefertigt. Ausführungen, die sinngemäß oder wörtlich übernommen wurden, sind als solche gekennzeichnet.

Aachen, im Juli 2011