# Name of the thesis

Konstantinos Papadimos

## Contents

# 1 Theoritical Overview

## 1.1 Units in particle physics

Particle physics, study the properites of subatomic particles and their interactions. To describe such microscopic phenomena and quantities, an appropriate system of units must be adopted, for if the SI system of units was to be used, one would have to deal with large exponents. More over, it is more practical to utilize a system that is based on the typical length and time scales found in particle physics[modern particle phyics].

The foundamental constants of the special theory of relativity, as well as quantum mechanics are $\hbar$, c, and GeV and they can be used as a basis to form a system of units, the Natural Units. Natural units can be further simplified by chosing:

$$c = \hbar = 1$$

The table below sumarizes the relationship between natural units and S.I

| Relationship between natural units and S.I | | |
|---|---|---|
| Quanity | Natural units($\hbar = c = 1$) | S.I |
| Energy | GeV | $Kgm^2s^{-2}$ |
| Momentum | GeV | $Kgm^2s^{-2}$ |
| Mass | GeV | Kg |
| Time | $GeV^{-1}$ | s |
| Length | $GeV^{-1}$ | m |

Table 1: Some basic quantites in S.I and in Natural units .

## 1.2 Special Relativity

### 1.2.1 Four-Vectors - Lorentz transformations

Let S and S' be two inertial reference systems, with S' moving at (a relativistic) 2velocity $u$ relative to S. The coordinates are chosen such that the motion is along the x-axis in both reference systems. The clocks of both S and S' are synchronized so that when $x = x' = 0$, $t = t' = 0$. For an event with coordinates (x, y, z, t) in S, the coordinates in S' are given by the Lorentz transformations:

$$\begin{aligned} (x')^0 &= \gamma(x^0 - \beta x^1) \\ (x')^1 &= \gamma(x^1 - \beta x^0) \\ (x')^2 &= x^2 \\ (x')^3 &= x^3 \end{aligned} \tag{1}$$

where $\beta \equiv \frac{u}{c}$, $x^0 \equiv ct$, $x^1 \equiv x$, $x^2 \equiv y$, $x^3 \equiv z$

The elements $x^i$, i = 0, 1, 2, 3 define the position four vector. Mathematicaly four vectors are 4 dimentional, rank 1 tensors that transform according to lorentz transformations

With the introduction of four vectors, using Einstein's summation convention, lorentz transformations can be written as:

$$(x')^i = \Lambda^i_j x^j \tag{2}$$

Where $\Lambda$, is the Lorentz transformation matrix, a rank 2 tensor:

$$\Lambda = \begin{pmatrix} \gamma & -\gamma\beta & 0 & 0 \\ -\gamma\beta & \gamma & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{3}$$

The following quantity does not change under lorentz transformation:

$$I^2 = -(x^0)^2 + (x^1)^2 + (x^2)^2 + (x^3)^2 = -(x'^0)^2 + (x'^1)^2 + (x'^2)^2 + (x'^3)^2 \tag{4}$$

or written in a more compact form:

$$I = g_{\mu\nu}x^\mu x^\nu = x^\mu x_\mu \tag{5}$$

where $g_{\mu\nu}$ is the Minkowski (metric) tensor:

$$g_{\mu\nu} = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

Such a quantity as I is called **invariant**

The introduction of four vectors and the metric tensor, yield the Minkowski Space Time where points need 4 coordinates to be fully described(1, time like and 3 space like ) and the distance between them is beeing defined by the Minkowski tensor. The necessity of four vectors, in order to describe non scalar quantities(such as velocity and momentum) in minkowski space time, is therefore evident.

### 1.2.2 Energy and Momentum

According to the principle of relativity, the laws of physics must be the same in all inertial reference systems. Hence, if momentum is conserved in one inertial frame of reference, it must also be conserved in all others. It is evident that the momentum of a moving particle must be defined in an appropriate manner, to satisfy the principle of relativity [3]. The four momentum is therefore defined as:

$$p^\mu = m\eta^\mu \tag{7}$$

Where $\eta^\mu = \frac{dx^\mu}{dt'}$, the four velocity of the particle.

The timelike component of four momentum, expresssed in natural units is $p^0 = \gamma m$. The 3 space like components, constitue the vector momentum, $\vec{p} = \gamma m\vec{\beta}$. The relativistic energy is defined as:

$$E = \gamma m = p^0 \tag{8}$$

Thus, the components of four momentum are:

$$p^\mu = (E, \vec{p}) \tag{9}$$

At this point we are able to calculate the invariant "interval" $p^\mu p_\mu$ :

$$p^\mu p_\mu = E^2 - |\vec{p}|^2 = m^2 \tag{10}$$

What is invariant in the case of momentum four vector, is the particle's mass. This quantity is called **invariant mass** of a particle as all observers, in different frames of references, agree upon its value. It is the invariant mass of particles that we can(or we try to) measure, in CERN as well.

## 1.3 The standard model

The Standard Model (SM)of particle physics is a theoretical framework that describes the fundamental particles and their interactions through the strong, electromagnetic, and weak forces. Each force is described by a corresponding quantum field theory(QFT). Namely, electromagnetic and weak interactions are described by the electroweak theory and the strong interactions by quantum chromodynamics(QCD). The interactions between particles in each QFT, are described in terms of the exchange of a spin-1 gauge boson. The photon, is the gauge boson of QED, while the gluon, which like the photon has

no mass, is the force-carrying particle in the strong interaction. The charged $W^+$and $W^-$ bosons mediate the weak charged current interaction, which is responsible for $\beta$ decay and fusion, while the weak neutral current interaction, is mediated by the electrically neutral Z boson. These interactions are also governed by the principles of symmetry and conservation, which dictate that certain properties, such as charge and energy, are conserved during particle interactions

The fundamental particles that comprise all matter according to SM are the already mentioned gauge bosons, quarks, and leptons. The quarks and leptons are organized into three generations, with each generation containing two types of leptons and two types of quarks. The leptons are either negatively charged, with a charge of -1, or electrically neutral. The quarks, on the other hand, have fractional charges of either -1/3 or +2/3, and are characterized by their color, which can be blue, green, or red. Additionally, for each elementary fermion, there is a corresponding antifermion with the same mass and spin but with an opposite electric charge.

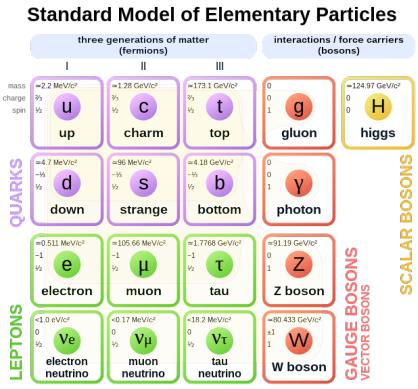



Figure 1: Summary of the elementary particles. All matter around us is made up by 12 fermions!

Completing the picture of fundamental particles is the scalar boson, the Higgs boson, responsible for giving mass to the other particles. A brief summary of the fundamental particles is presented in Figure 1.

The SM has undergone extensive testing through high-energy experiments at CERN, with its predictions confirmed with a high degree of precision. However, the model has limitations, such as its inability to account for dark matter or the observed imbalance between matter and antimatter in the universe.

Despite its limitations, the Standard Model remains a cornerstone of modern physics, and its continued study and refinement is essential to advancing our understanding of the universe at its most fundamental level.

# 2 Accelerators and Detectors: the LHC and the CMS

## 2.1 The Large Hadron Collider

Our knowledge in the field of high energy physics has been largely obtained through fixed target experiments that used proton and electron accelerators. However, over the last decades, the significance of colliding beam experiments has been rising. Such experiments involve two particle beams that rotate in opposite directions and collide at multiple points around the ring. The key advantage of colliding beam machines is their ability to produce new particles due to the high center of mass energy created during the collision. This energy increases linearly as E, rather than as $E^{1/2}$, in fixed target experiments, and almost all of it is utilized in generating new particles[modern particle physics].

One great example of a colliding beam machine is The Large Hadron Collider (LHC). The LHC has been instrumental in many groundbreaking discoveries, with the most famous one beeing the Higgs boson, and has helped scientists to further our understanding of the fundamental nature of the universe.

The LHC encompasses a 27-kilometre ring consisting of superconducting magnets with numerous accelerating structures along its length.

Within the accelerator, a strong magnetic field is accelerating the two counterrotating proton beams to velocities near that of the speed of light uppon collision. The thousands of superconducting magnets, responsible of the generation of the magnetic field, are of varying sizes and types. Dipole magnets, 1232 in total and 15 meters in length, are utilized to bend the beams and quadrupole magnets, 392 in total and 5-7 meters long, focus the beams. Prior to collision, another type of magnet is used to compress the particles, increasing the likelihood of collisions source.

## 2.2 The Compact Muon Solenoid

The main goal of the Compact Muon Solenoid (CMS), as a general purpose particle detector, is to to reconstruct the Feynman diagram associated with any interaction that might happen inside the LHC. The first and foremost interactions that happen are the collisions between the beams, which generate individual interactions known as *events*. Even though most of the particles associated with an event are unstable, their final decay products, are stable enough to reach the detector and be measured. In the rest of the chapter I will give a brief overview of the CMS detector and discus how does it detect particles.

### 2.2.1 Overview

The CMS detector consists of 5 compartments, each with unique functionality, that are organised in several coaxial layers. The Silicon Tracker, located in the innermost part of CMS, includes silicon pixel vertex detectors and silicon strip detectors, which trace the position and momentum of charged particles. The Electromagnetic Calorimeter (ECAL), the second layer, is composed of PbWO4 crystals and intended to detect photons and electrons. The Hadronic Calorimeter (HCAL), the third layer, is designed to identify hadrons. The Superconducting Solenoid Magnet, the fourth layer, is an solenoid coil that generates a constant magnetic field of 4 T along the direction of the beam. Due to the deflection of the trajectories of charged particles by the magnet, it becomes possible to measure their momentum. The final, outer most layer, is responsible for the measurement of muon the tracks. Figure 2 provides a sectional view of the CMS detector[`https://cms.cern/news/cms-detector-design`].

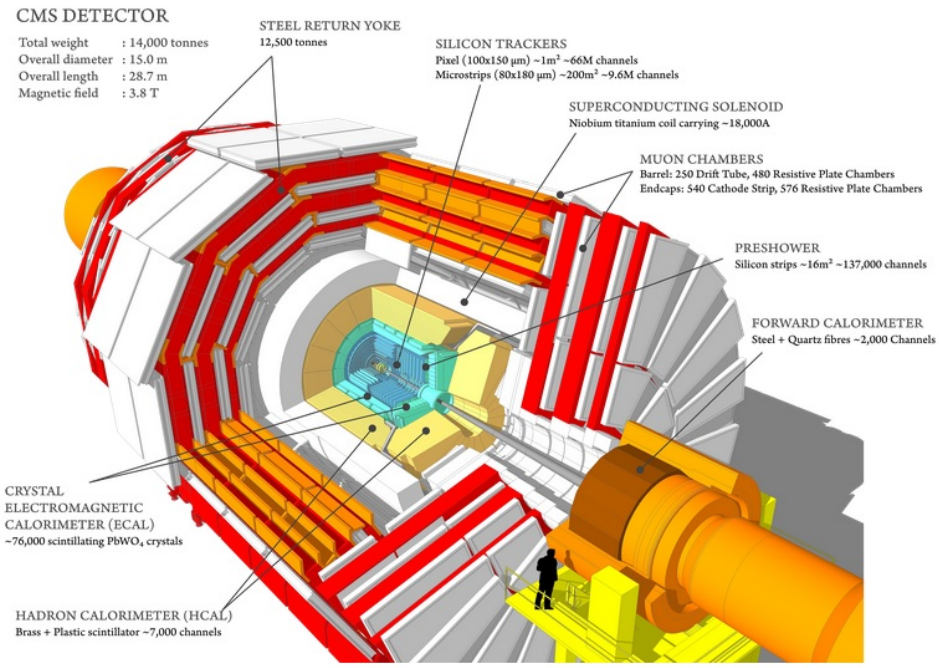



Figure 2: A cross-sectional perspective of the CMS detector

### 2.2.2 Coordinate convention at the CMS

Given the solenoid geometry of the CMS detector, it is more convenient to use a spherical type of coordinates$(r, \phi, \theta)$. The origin is located at the collision point and the z axis is parallel to the beam as shown in figure 3. In this system, the momentum of a particle(or any other vector) can be analyzed in a component parallel to the z axis and one component perpendicular to the z axis(Transverse mometnum). Transverse mometnum is defined as follows:

$$|\vec{P_T}| = \sqrt{P_x^2 + P_y^2} = |\vec{P}| \sin\phi \tag{11}$$

Where $|\vec{P}| = \sqrt{P_x^2 + P_y^2 + P_z^2}$. The CMS detector, measures the transverse energy(source ) of particles, and thus it is useful to work with the transverse momentum $P_T$. The azimuth angle $\phi \in [0, 2\pi)$ coordinate is the angle between $P_t$ and x axis and the polar angle $\theta \in [0, \pi]$ is the angle between the momentum vector and the z axis.

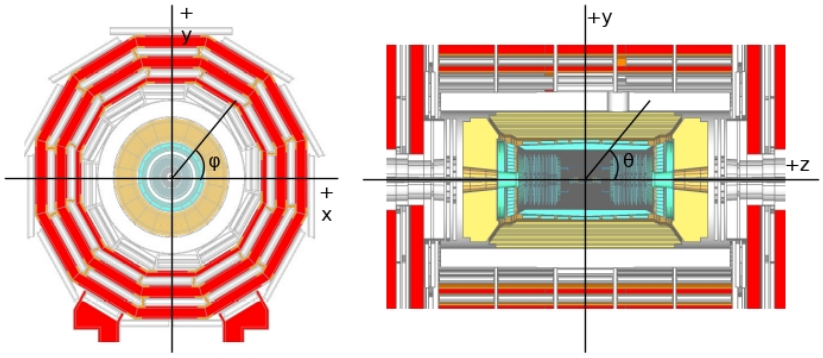



Figure 3: CMS coordinates

Due to the relativistic nature of the phenomena taking place inside LHC, it is more usefull to work with lorentz invariant quantities [V. Chiochia (2010) Accelerators and Particle Detectors from University of Zurich]. Thus, instead of working with the polar angle it is more convenient to introduce the lorentz invariant *pseudorapidity* $\eta \in [-\infty, +\infty]$. Pseudorapidity is defined as:

$$\eta \equiv -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{12}$$

The cartesian$p_x$, $p_y$, $p_z$ momentum components are related to the $P_T$, $\eta$, $\phi$ components by the following transformation relations:

$$\begin{aligned} p_x &= P_T \cos\phi \\ p_y &= P_T \sin\phi \\ p_z &= P_T \sinh\eta \\ |\vec{P}| &= P_T \cosh\eta \end{aligned} \tag{13}$$

### 2.2.3 Position tracking and momentum measurements: Silicon Tracker

The Silicon tracker measures the positions of charged particles at a number of points, thus it is able to record their trajectory. Given the radius of curvature of the particle's track(due to the 4T magnetic field of the super conducting solenoid), the tracker provides sufficient information, to reconstruct the momentum of the particle. More over, the geometrical location of the trajectory gives direct information regarding the position of the particle. Therefore, the silicon tracker measurements provide information regarding the $P_T$, $\eta$ and $\phi$ of the particles that it detects.

A schematic representation of the Sillicon Tracker's corss section, can be viewed on figure 4. The tracker consists of a silicon pixel detector and a silicon strip detector. The silicon pixel detector is

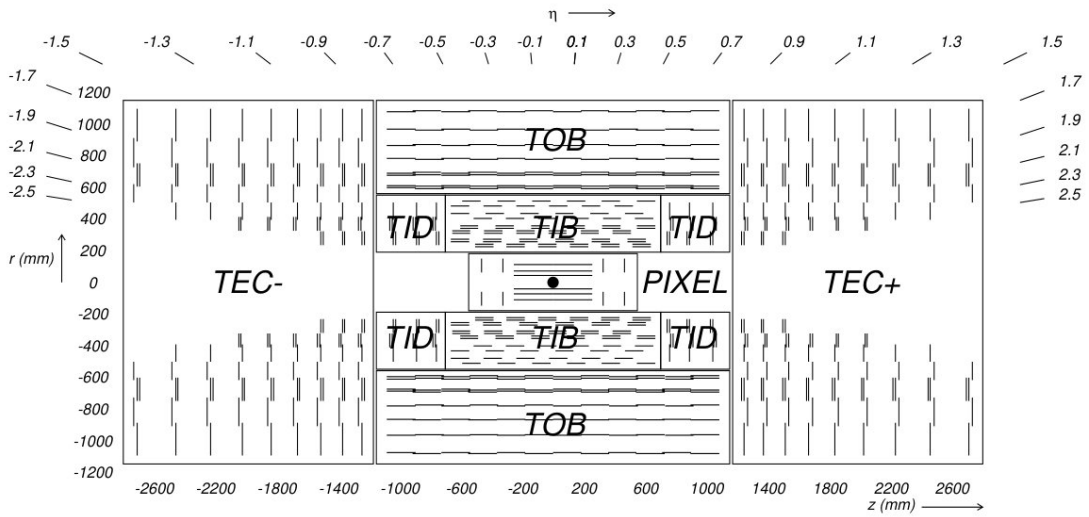



Figure 4: Schemtic illustration of a crossection of the CMS Tracker

composed of two sub-detectors. Namely, the barrel which consists of three layers covering the region $|\eta| < 2.2$ and at $r = 4.4, 7.3$ and $10, 2$cm. The end caps, are two discs of pixel modues, located one on each side, that complete the design of the silicon pixel detector. The pixel detector improrves the trajectory and position measurements, by providing two-dimensional measurements of the charged particles' hit positions. (source).

The silicon strip detector, covers the radial region $r \in [20, 116]$ cm. and is comprised of four inner barrel (TIB) layers and two inner endcaps (TID). The TIBs are assembled in shells and each TID consists of three small discs. The outer barrel (TOB) encompasses both TIB and TID and contains six concentric layers. The tracker is closed off on either end by two endcaps (TEC). Measurements at the silicon strip detector give information regarding the path of each particle allows the distinction of separate particle trajectories.

### 2.2.4 Energy Measurements: Calorimeters

Apart from measuring position and momentum, determining the energy of particles produced in LHC collisions is crucial. In the Compact Muon Solenoid (CMS) experiment, this information is obtained from particle interactions with matter in the calorimeters. Particles that are stable enough to reach the detector without decaying are either leptons, photons, or hadrons. The interactions between electrons, photons, and matter are of electromagnetic nature, while those between hadrons (charged or neutral) and matter are strong interactions. Therefore, the CMS experiment employs two types of calorimeters: the Electromagnetic Calorimeter (ECAL), located at the innermost layer, which measures the energy of photons and electrons, and the Hadron Calorimeter (HCAL), situated at the outer shells of the calorimeter section.

- Electromagnetic Calorimeter (ECAL)

  Figure 5(source) provides a view of the Electromagnetic calorimeters inside the CMS. The ECAL is composed of lead tungstate (PbWO4) crystals and is designed with a central barrel section (EB) and two endcaps (EE) that cover a range of pseudorapidities up to $1.48 \leq |\eta| \leq 3.0$(source). The crystals are highly dense and scintillate when high-energy photons or electrons interact with them. When a particle passes through the ECAL, it deposits its energy in the form of electromagnetic showers, which cause the crystals to emit light. The emitted light is then captured and amplified in order to estimate the energy of the incoming particle. The high-density crystals of the ECAL make it possible to accurately measure the energy of photons and electrons with high precision and resolution.

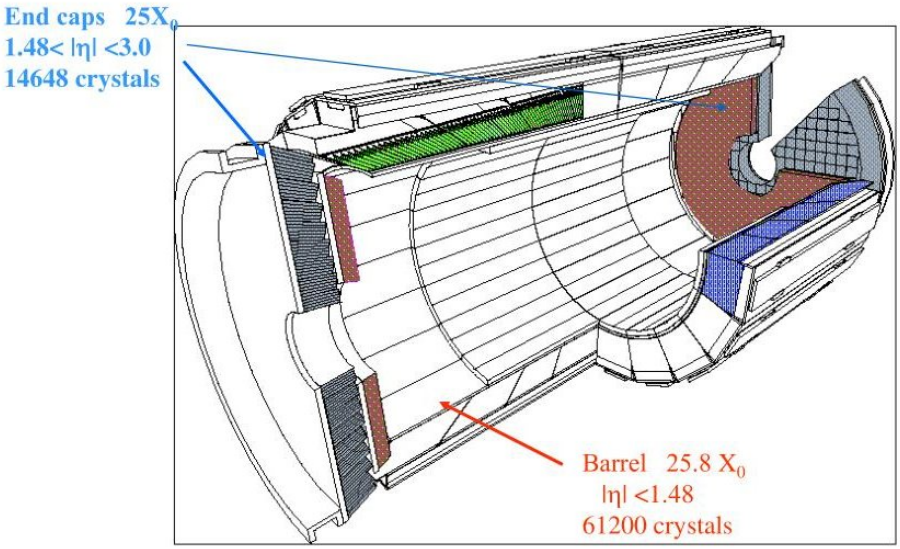



Figure 5: Schemtic illustration of the Ecal parts inside CMS

- Hadron Clorimeter (HCAL)

  The hadrons that manage to reach the detector, fly off the ECAl and interact with the Hadron Calorimeter. The HCAL consists of alternating layers of absorber material and plastic scintillator tiles that detect particles generated by the hadrons as they interact with the absorber. When particles pass through the HCAL, they interact with the absorber material, producing showers of particles that create signals in the scintillator tiles. These signals are then read out and processed to measure the energy of the incoming hadrons. The HCAL has both a barrel section(HB), with pseudorapidity coverage at $|\eta| < 1.3$ and endcap(HE), covering a range of pseudorapidities $1.3 \leq |\eta| \leq 3.0$. The HCAL is highly effective at measuring the energy of hadrons due to its high-density absorber material and precise arrangement of scintillator tiles (source)

#### 2.2.5 Detecting Muons

In the outer regions of the CMS detectors, are located the muon chambers. They are the final part of the detector and are designed soley for the detection of muons, which due to their large mass(207 times greater than the electron mass) muons travel a longer distance in matter than electrons. Thus, their energy cannot be measured in ECAL.

The muon chambers consist of 250 drift tubes (DTs) and 540 cathode strip chambers (CSCs), which track the positions of the particles. Additionally, there are 610 resistive plate chambers (RPCs) and 72 gas electron multiplier chambers (GEMs), making a total of 1400 chamber units . The use of multiple layers of detectors and different types of chambers makes the system robust and able to filter out background noise.

Figure 6 illustrates the arrangement of the four different kinds of chambers. In the "barrel region," which surrounds the beam line, the DTs and square-shaped RPCs are grouped in coaxial cylinders. The CSCs, trapezoidal RPCs, and GEMs are located at the end cap region of the barrel . This arrangement allows for accurate measurements of the muons' trajectories and momenta in different regions of the detector.([[`https://cms.cern/index.php/detector/detecting-muons`][

### 2.3 Event Reconstruction & Detector Calibration

#### 2.3.1 Event Reconstruction

The infrastructure described in the previous sections provides (almost) all the necessary information regarding the particle collisions taking place inside the LHC. The next step is to combine this informa-

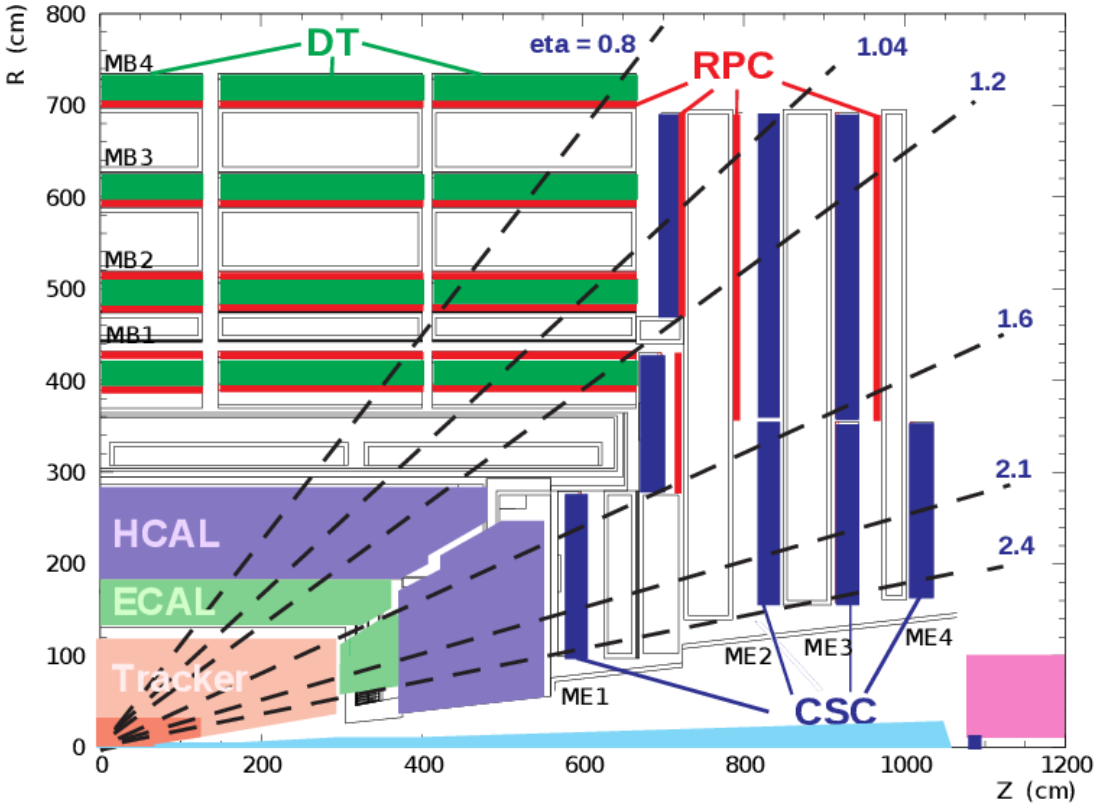



Figure 6: A quarter sectional view of the CMS muon chambers. The beamline is perpendicular to the plane of the page

tion to reconstruct the physical objects (particles, jets, etc.) that are being produced in each collision (event).

As already discussed, the trajectories and momenta of the charged particles are reconstructed using information coming from the pixel tracker, while signals coming from the calorimeters provide energy measurements of the particles. The Particle Flow algorithm is then used to combine information from all subdetectors to create a consistent set of physical objects, including electrons, photons, muons, and neutral and charged hadrons. These objects are further processed by dedicated algorithms to reconstruct composite objects such as taus and jets, and to estimate missing energy. To determine the vertices of the collisions, information from the reconstructed tracks of the particles is used. The primary vertex is defined as the vertex with the largest sum of the transverse momentum of all contributing physical objects.

### 2.3.2 Callibration and energy scale uncertainties

The reconstructed objects, are usually calibrated using well-known resonances, such as the Z boson or J/psi meson, whose masses and decay properties are well-measured. The calibration process involves adjusting the energy scale and resolution of the reconstructed objects such that the resonances in data and simulation appear at the correct mass values with the correct amount of smearing.

However, it is not possible to achieve a perfect agreement between data and simulation due to the complexities of the subdetectors and reconstruction algorithms used in the experiment, as well as non-linear effects such as detector aging or radiation damage. As a result, one defines "energy scale and resolution uncertainties" that reflect the level of disagreement between data and simulation.

Such deviations in the energy scale (energy scale uncertainties) have an effect on the measured momenta and spatial coordinates of the particles, which can lead to inconsistency between the width of the resonant mass distribution, in simulations and measurement.

The arising question then is: how do the various analysis techniques that scientists have in their disposal respond to energy scale uncertainties? In other words, what is the distinctive ability of the various analysis techniques? Our work will focus on the effects that energy scale uncertainties have, in a traditional fit-based analysis and a more modern Boosted Decision Tree-based analysis, using the generic diobject production process as the working example.

# 3 Multi variate and single variate classification techniques

## 3.1 Multivariate techniqies: Decision trees and Supervised Learning

The distinction between different particles, can be regarded as a classification problem where the target, is the prediction of a categorical output variable(i.e. lepton, boson), based on one or more input variables(i.e momenta components). Classification problems in machine learning can be solved with supervised learning. In such procedure, a training data set is being used for the development(training) of a model that is able to perform the classification task. The output of the model is then beeing tested and evaluated on previously unseen data.

Before presenting any specific method of solving classification problems It is important to present an overview of the key elements in supervised learning.

### 3.1.1 Supervised learning

Let us pose the following problem: Given a data set $D = (\vec{X}, \vec{y})$, where $\vec{X}$ is a matrix of the independet variables and $\vec{y}$ is a vector of dependent variables, we want to find a model $f(\vec{x}; \vec{\theta})$, that can predict an output from a set of input variables. Moreover, we want to be able to judge the performance of the model on a given data set. To do that we need to define a cost function $C(\vec{y}, f(\vec{X}; \vec{\theta}))$, such that the model will have to find the parameters $\theta$ that minimize the cost function.[5]

This the mathimatical pustulation of a supervised learning problem. I will now, in brief, discuss the role and interpretation of each of the 'ingredients' stated above.

- Model

  The model, is a mathematical function $f : \vec{x} \to y$ of the parameters $\theta$. Given a set of parameters, the output of the function, the prediction $y_i$, is derived from the input variables $\vec{x}$. The parameters are undefined. The task of the training is to estimate the set of parameters from the training data set. In a classification problem(something is of type a or it is not), it is possible to use the logistic transformation of the function output, to obtain the probability of the positive class.

- Cost function

  The cost function, also known as an objective function, is represented by mathematical function and it measures how well a model fits the training data. The cost function is used to train the model by finding the best set of parameters $\theta$ that minimize the function. In machine learning, the objective function, usually consists of two parts: a training loss function (L) and a regularization term ($\Omega$).

  $$obj(\theta) = L(\theta) + \Omega(\theta) \tag{14}$$

  The training loss function measures how predictive the model is with respect to the training data. A common choice of training loss function is the logistic loss, which is used for logistic regression(classification) and is given by

  $$L(\theta) = \sum_i [y_i \ln(1 + e^{-\hat{y}_i}) + (1 - y_i \ln(1 + e^{\hat{y}_i}))] \tag{15}$$

  where $y_i$ is the true label and $\hat{y}_i$ is the predicted label.

The regularization term, $\Omega(\theta)$, controls the complexity of the model, which helps to avoid overfitting. Overfitting occurs when a model is too complex and starts to extract local features from the training data. The model thus, looses its generalization power to new unseen data. Regularization helps to prevent overfitting by adding a penalty term to the cost function, which discourages the model from having too many parameters or too complex a structure.

The following figure gives an example of overfitting due to a very complex and very simple model.

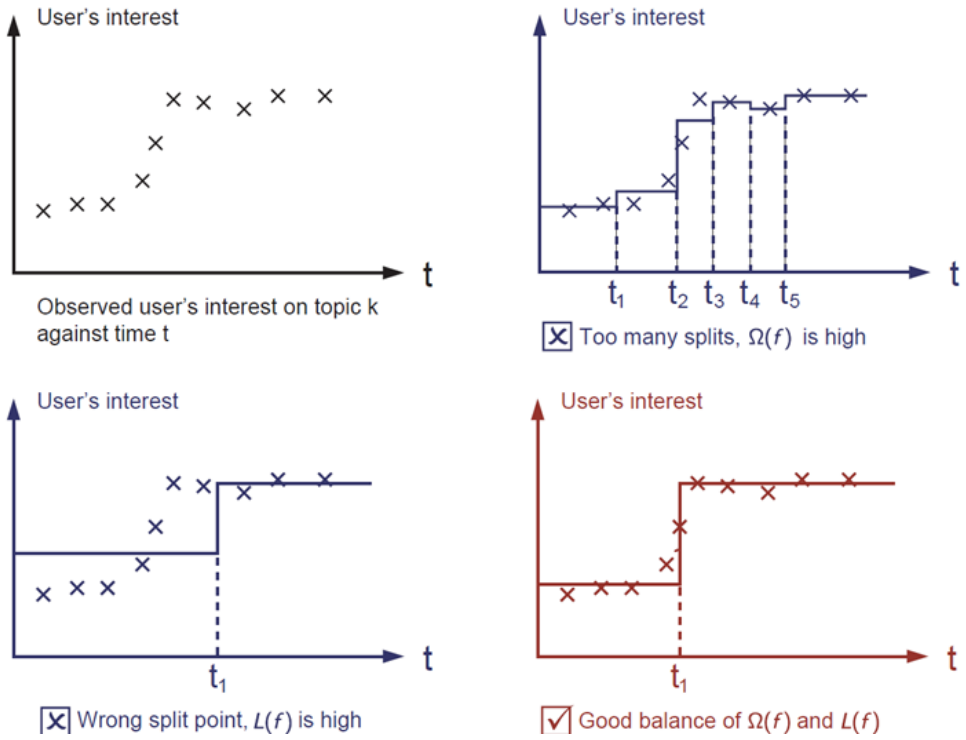



Figure 7: Examples of over fitting and under fitting. The top right model, places too many cuts. Even though it succesfully describes the trend, the splits seem to correspond only on the specific data set, therefore it is overfitted. The bottom left model places too few and impresice cuts. The bottom right model seems to succesfully describe the trend while its simplicity infers that It has not sacrificed its generalization power.

### 3.1.2 Decision Trees

A decision tree is a flowchart-like tree structure, where each internal node represents a feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome.

Formally, a decision tree can be represented as a set of rules or conditions in the form of:

```
f(X)={condition1,condition2,..condition_n}
```

where each condition is a tuple of the form (feature, threshold, comparison operator) and the final outcome is represented by the leaf node. For example, consider the decision tree of figure 8 that classify fruits based on color, shape, size, and taste. Let `X` be the input `X = {"red", "smal", "sour"}`. Then `f(X) = "grape"` [2]

- Decision Tree Ensembles

  The tree ensemble model consists of a set of classification and regression trees (CART). Let $\mathcal{F}$ be the set of all possible CART's and $f_k \in \mathcal{F}$, a function that represents a CART. The model in

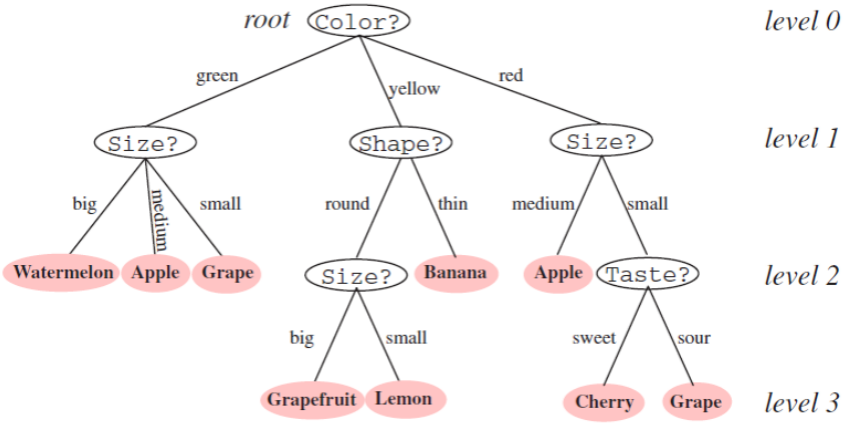



Figure 8: Example of a a decision tree that clasifies fruits

discussion then, can be written as:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in \mathcal{F} \tag{16}$$

If $\hat{y}_i$ represents the prediction of the tree, given an input variable $x_i$, the real label of $x_i$ will be denoted as $y_i$ . The objective function will be of the form:

$$obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{i=1}^{t} \omega(f_i) \tag{17}$$

where $\omega(f_i)$ is the complexity of a given tree and l is the loss function.

- Tree boosting

  As stated earlier, the model is beeing trained, to learn those trees $f_k$ that minimize the objective. The resulting model then, will be an ensemble of those functions $f_k$. The optimization of the objective, is a problem that cannot be solved with the traditional methods. Instead, the model is being iteratively trained in an additive manner.[1] let the prediction value at the t-th iteration be $\hat{y}_i^{(t)}$. In the next iteration(t+1), the chosen function $f_{t+1}$, is such that if added to the model, the resulting prediction $\hat{y}_i^{(t+1)}$ will minimize the cost function:

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= \hat{y}_i^{(1)} + f_2(x_i) \\ &\cdots \\ \hat{y}_i^{(t)} &= \hat{y}_i^{(t-1)} + f_t(x_i) = \textstyle\sum_{k=1}^{K} f_k(x_i) \end{aligned} \tag{18}$$

  The objective at step t is:

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \omega(f_i) = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \omega(f_i(t)) \tag{19}$$

  Taylor expanding the loss function $l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i))$, around $f_t$, up to the second order and neglecting terms, referring to previous rounds, the specific objective becomes:

$$\sum_{i=1}^{n}\left[g_i f_t(x_i)+\frac{1}{2}h_i f_t^2(x_i)\right]+\omega(f_t) \tag{20}$$

Where

$$\begin{aligned} g_i &= \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \\ h_i &= \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \end{aligned} \tag{21}$$

This is the minimization goal for $f_t$ . [4]

### 3.1.3 Signal from background separation using BDT

When the trained BDT model is applied to a given dataset, it returns the probability (BDT score) for an event to be signal or background. If the returned probability of an event is less than $50\%$, then the event in question is more "background-like." The predicted results can then be visualized as a histogram, allowing us to determine the number of signal and background events for each BDT score. This information can be used to define a value of the BDT score to place a "cut" and keep all the signal and background events from that value onwards.

## 3.2 Signlevariate Techniques: Fit based classification

A fit-based analysis can be considered as single-feature classification where, signal events are separated from background events, by fitting the mass histogram (mass spectrum) of the two components.

After fitting the signal and background invariant mass, the observed data can be modeled as

$$observation(x) = sig(x) + bkg(x), x \in \mathcal{M} \tag{22}$$

Where $sig$ and $bkg$, are the fitted signal and background and $\mathcal{M}$ is the mass range of the dataset in question.

Let $I \subseteq \mathcal{M}$ be a region of interest in the invariant mass spectrum. The number of observed events, background events, and signal events in $I$ can be estimated as follows:

$$O = \int_I observation(x)dx \tag{23}$$

$$B = \int_I bkg(x)dx \tag{24}$$

$$S = O - B \tag{25}$$

## 3.3 Statistical interpretation of the results

Usually, once the signal is separated from the background, a specific region is defined (either a BDT score range in case of a BDT-based analysis or a mass range in case of a fit-based analysis), and only the signal and background events falling within the defined region are selected. The question that arises is whether the selected signal is simply a statistical fluctuation of the background or not. In other words, how statistically significant is the signal? Because measurements at CMS are Poissonian in nature, the measured signal is compared to the Poissonian deviation of the background. Therefore, the significance we are interested in is defined as follows:

$$\text{Significance} = \frac{Signal}{\sqrt{Background}} \tag{26}$$

Were signal and background signify the number of signal and background events present in the selected region.

# 4 Data set and analysis methods

## 4.1 The $Y \to XX$ channel

The $Y \to XX$ dataset is composed of simulated events that represent the generic process of a resonance Y decaying into an XX pair, in the mass range between 100 and 300 GeV. The background consists of 50,839 events, while the signal comprises 5,946 events. The particular set is made up of miscellaneous pre-existing Monte Carlo (MC) samples, and the selected events contain only leptonic final states (one lepton and one antilepton of the same flavor). For the purpose of this study, only generator-level events were used, and given that we are not interested in any particular process, but rather in the most general case, no kinematic constraints were placed in the event selection. It should be noted that despite the fact that the parent MC samples contain lepton pairs in the final states, the resulting set can represent any diobject production in the selected mass range. The following table summarizes the features of the dataset in question:

| Feature | Description |
|---|---|
| $Pt_1$ | The transverse momentum of the first particle in the XX pair |
| $\eta_1$ | The psudorapidity of the first particle in the XX pair |
| $\phi_1$ | azimuth angle of the first particle in the XX pair |
| $Pt_2$ | The transverse momentum of the second particle in the XX pair |
| $\eta_2$ | The psudorapidity of the second particle in the XX pair |
| $\phi_2$ | azimuth angle of the second particle in the XX pair |

Table 2: Summary of the data set features

The mass spectrum for the $Y \to XX$ decay, can be seen in figure 9

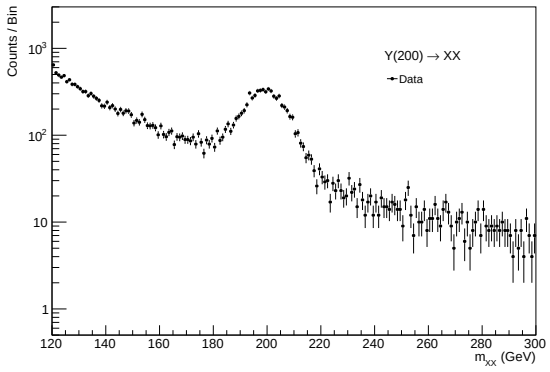



Figure 9: The $Y \to XX$ invariant mass spectrum

## 4.2 Energy scale uncertainties

Energy scale uncertainties have an effect on the transverse momenta $P_t$ of the produced $XX$ pair. Such uncertainties are modeled as random noise (Gaussian smearing) to the signal component. To smear the data by $x\%$, we iterate over every signal event and multiply the transverse momenta by numbers sampled from a Gaussian distribution of $\mu = 1$ and $\sigma = x/100$, where $x$ is the percentage of smearing (each $P_t$ is multiplied by a different number).

In the present study, we compare the performance of a Boosted Decision Tree (BDT)-based analysis and a fit-based analysis on a given dataset, for various cases of smearing (various values of $\sigma$). Table 3 summarizes the smearing percentages used for this analysis. The values are chosen such that we examine both cases of mild noise ($0 - 5\%$) up to extreme noise ($50\%$). Nevertheless, it should be noted

that what is considered as an extreme case or a mild case of smearing is not defined in an absolute manner but rather is related to the mass range of the data upon which the blur is applied.

| Percentage of smearing |
|---|
| 0% |
| 5% |
| 10% |
| 15% |
| 20% |
| 30% |
| 40% |
| 50% |

Table 3: Summary of the smearing cases that will be studied in this work

## 4.3 Analysis Method I: Training a BDT Classifier

### 4.3.1 The Train/Test/Application data sets

For the BDT training (and due to the lack of an infinite number of events), the original dataset had to be split into three parts. The training set is used to train the classifier. As the reader may have noticed, the signal events in the original dataset are much less than the background events. This class imbalance makes the training of the model much harder, and for that reason, the training set has been enriched with more (unseen) signal events, so that the two classes have the same number of events. The testing set is used to evaluate the training. For that purpose, the two signal and background classes also have the same number of events. Finally, the application set is used for the analysis. Through trial and error, we noticed that working with smaller statistics enhances the magnitude of statistical fluctuations in the analysis. To avoid such confusion, the application set contains a part of the testing set to ensure large statistics. This doesn't interfere with the training, since the BDT has never seen the testing events during training. Finally, it should be mentioned that in order to have an "apples-to-apples" comparison between the BDT-based analysis and the Fit-based analysis, the application set will be analyzed in both cases. For that reason, the smearing cases that will be discussed for the rest of this chapter will concern only the application set.

Table 4 summarizes the number of events used in each dataset.

| Data Set | No.Signal Events | No. Background Events |
|---|---|---|
| Training | 3882 | 3882 |
| Testing | 3881 | 3881 |
| Application | 2973 | 20827 |

Table 4: Sumarry of the Train Test Application number of events

### 4.3.2 Training

One of the key aspects of a successful training is the feature space that is being used. Although the Train/Test sets consist of the features described in Table 2, those features are not optimal for the particular classification problem. The feature space that is found to be optimal for the problem in question consists of five features and is summarized in Table 5 and Figure 10.

To assess the performance of the trained model, the Area Under the Receiver Operating Curve (ROC-AUC) is used as a metric. However, it is important to note that in addition to evaluating the model's performance, we also consider the possibility of overfitting. To quantify overfitting, we examine the ratio of the number of training events to the number of testing events at a particular BDT score. If the

| Feature | Description |
|---|---|
| $Pt_1$ | the transverse momentum of the first particle in the XX pair. |
| $Pt_2$ | the transverse momentum of the second particle in the XX pair. |
| $\Delta\phi = \phi_2 - \phi_1$ | the difference in the azimuthal angles between the two particles in the XX pair. |
| $\Delta\eta = \eta_2 - \eta_1$ | the difference in the pseudorapidity values between the two particles in the XX pair. |
| $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ | the separation in the eta-phi plane between the two particles in the XX pair. |

Table 5: Sumarry of the features used for training

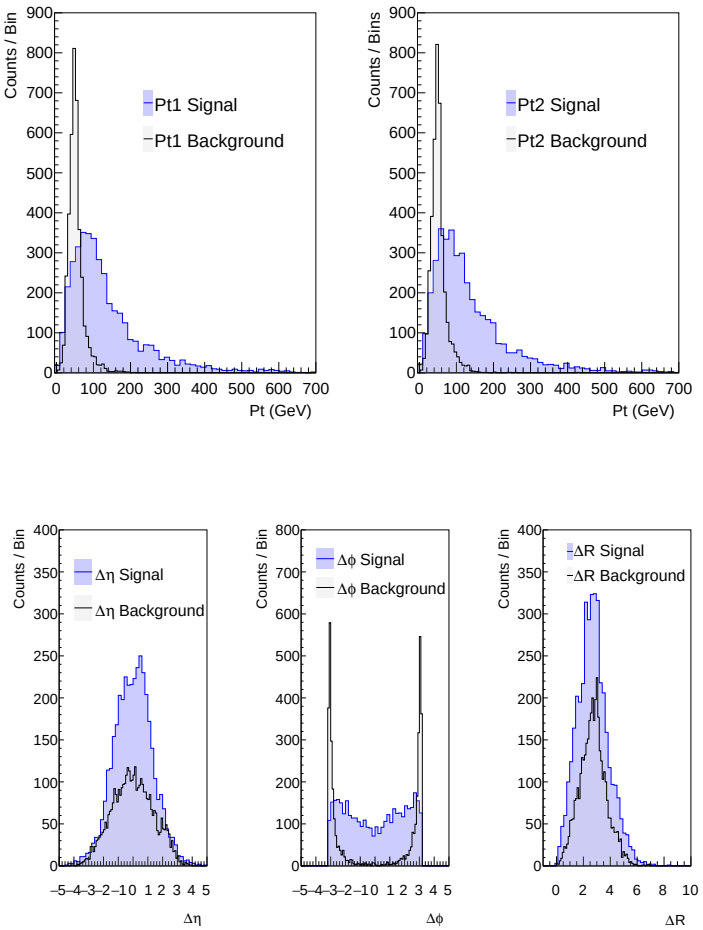
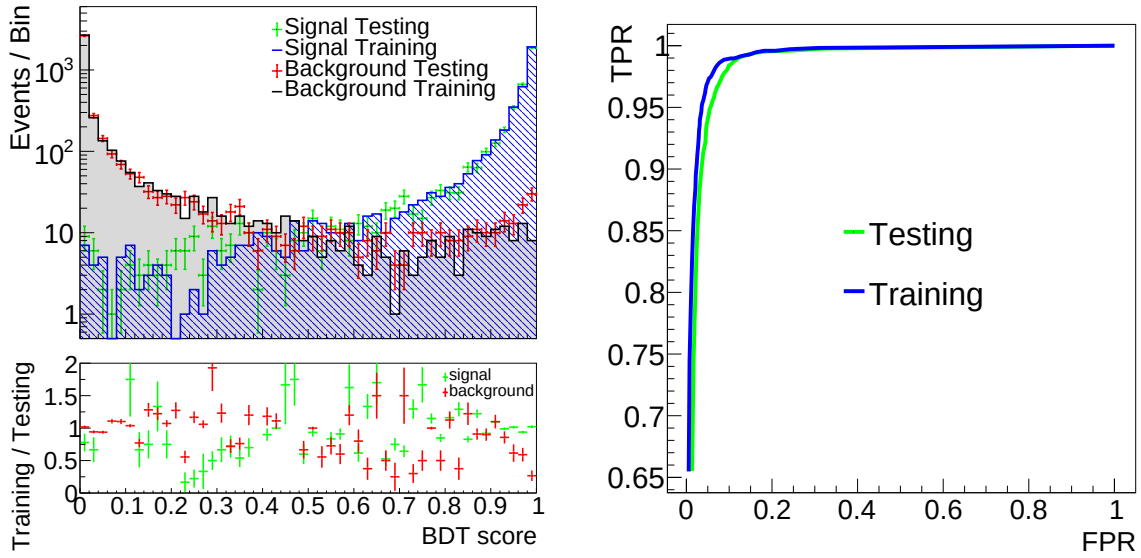



Figure 10: Sumarry of the features used for training

model is not overtrained, the performance on the training and testing set will be approximately the same, and therefore, the ratio will fluctuate around 1 across the BDT score range.

Figure 11 displays the BDT score plot of the training and testing set, as well as the corresponding ROC curves. The performance of the model on the testing set yields an AUC score of 0.98. Looking at the Training/Testing ratios, one can notice that they fluctuate around one. The absence of a profound trend in both signal and background ratios implies that the model is not severely overfitted.

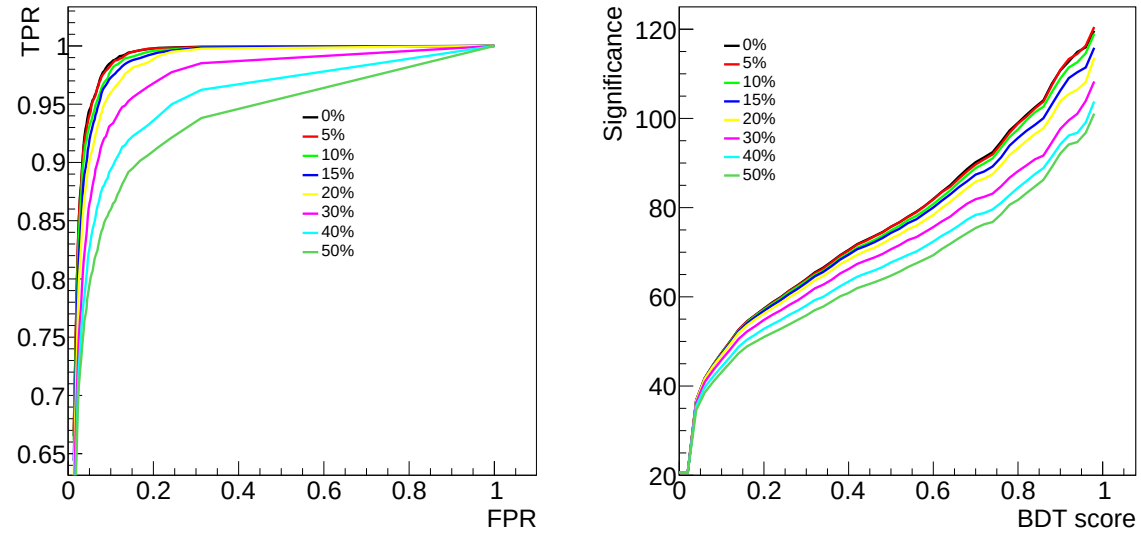



(a) (b)

Figure 11: A: The BDT score of the Testing and Training sets. B: The roc curves for the training and testing sets

### 4.3.3 Application

The application of the model to the application set is a rather straightforward process. The data are simply "fed" to the model, and the latter performs the classification. The ROC curves for the performance of the trained model on the data for each of the smearing cases of table 3 are illustrated in figure 12a.




(a) (b)

Figure 12: a: Summary of the ROC curves for the performance of the model on the data for each smearing case. b: Significances calculated across the BDT score range for the smearing cases of Table 3. The way that these curves are made is analogous to the calculation of the ROC curve.

To find the optimal BDT score to place the cut, the whole BDT score range is scanned, and the significance is evaluated in the range [c,1], $\forall c \in [0, 0.98]$ with step size 0.02, same as the bin width of the BDT score histogram(it is pointless to calculate the significance at a single point since it will be 0). The scan of significance for every case of smearing is illustrated in figure 12b.

Looking at the plot, the cut that gives the best significance is c = 0.98 (the region will be [0.98,1]). Moreover, it is of interest to consider a wider region as well, for the reason that more signal is accepted despite the rejection of less background. To make this argument a bit clearer, let us consider figure 11a. At BDT score ~ 7 and onwards, the amount of background events in each bin remains somewhat constant (within statistical fluctuations) while the signal events increase rapidly. It is therefore interesting for our analysis to see how this behavior reflects on the evolution of significance for the various cases of smearing. Looking at figure 12b again, we conclude that a cut at c = 0.86 is good enough for our purpose.

The results can be seen in Figure 13 which compares the evolution, in terms of smearing percentage, of the significance for the cuts c = 0.98 and c = 0.86. Table 6 presents the amount of signal and background events present for the two cuts.

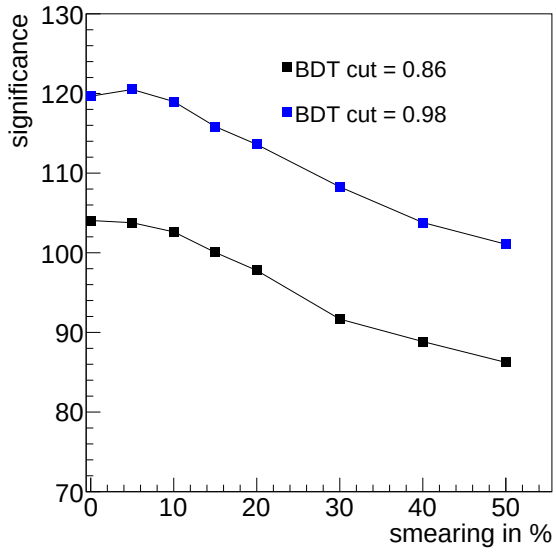



Figure 13: Evolution of significance for the smearing cases of table 3.

| Smearing % | No. Sig. Events at BDT cut = 0.86 | No. Bkg.Events at BDT cut = 0.86 | No. Sig. Events at BDT cut = 0.98 | No. Bkg.Events at BDT cut = 0.98 |
|---|---|---|---|---|
| 0 | 2622.0 | 635.0 | 1977.0 | 273.0 |
| 5 | 2615.0 | 635.0 | 1991.0 | 273.0 |
| 10 | 2586.0 | 635.0 | 1966.0 | 273.0 |
| 15 | 2521.0 | 635.0 | 1914.0 | 273.0 |
| 20 | 2464.0 | 635.0 | 1877.0 | 273.0 |
| 30 | 2310.0 | 635.0 | 1789.0 | 273.0 |
| 40 | 2239.0 | 635.0 | 1715.0 | 273.0 |
| 50 | 2173.0 | 635.0 | 1670.0 | 273.0 |

Table 6: Signal and background events at BDT cut 0.86 and 0.98 for different smearing percentages.

## 4.4 Analysis Method II: Fit based analysis

### 4.4.1 Invariant mass reconstruction

The invariant mass of the XX pair is calculated using the features in Table 2. The resulting spectrum is shown in Figure 14.

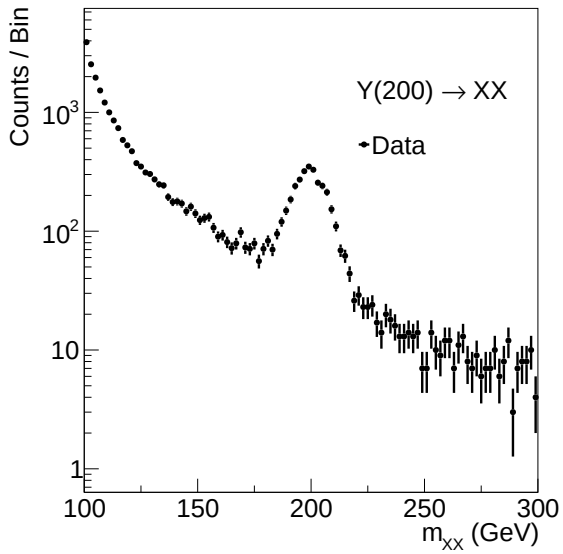



Figure 14: The invariant mass spectrum of the application set

Events with invariant mass $m_{XX} < 120$GeV make the background fit significantly harder without contributing significantly to the analysis. Therefore, such events are excluded from this study, and the working mass spectrum is limited to the range $[120, 300]$GeV.

### 4.4.2 Background Fitting

As discussed in previous sections, the applied smearing only affects the signal component of the application set. For this reason, and to simplify the analysis, the background shape is fitted separately and kept constant throughout the signal fits.

Despite this simplification, determining the shape of the background was not a trivial process. Through trial and error, the function shown in Equation 27 was found to be the best fit.

$$bkg(x) = \alpha + \beta x^{-1/2} + \gamma x^{-1} + \delta x^{3/2} \tag{27}$$

The parameters $\alpha$, $\beta$, $\gamma$, and $\delta$ are free parameters of the fit. The modeled background is illustrated in Figure 15.

### 4.4.3 Signal Fitting

The signal is fitted using a Gaussian function with $\sigma$ and magnitude as free parameters, and $\mu = 200$GeV (the mass of the resonance). Figure 16 shows the fitted invariant mass spectra for smearing percentages of $0\%$, $5\%$, $10\%$, $15\%$, and $20\%$. As illustrated in Figure 17, the signal mass in the extreme cases of $30\%$, $40\%$ and $50\%$ smearing is indistinguishable from the background. Therefore, attempting to fit those spectra would be a pointless exercise.

### 4.4.4 Signal from background separation

As with the BDT method, we want the region of interest that yields the best significance. To do so, we scan various mass windows around the center of the signal. We scanned six different regions (in the

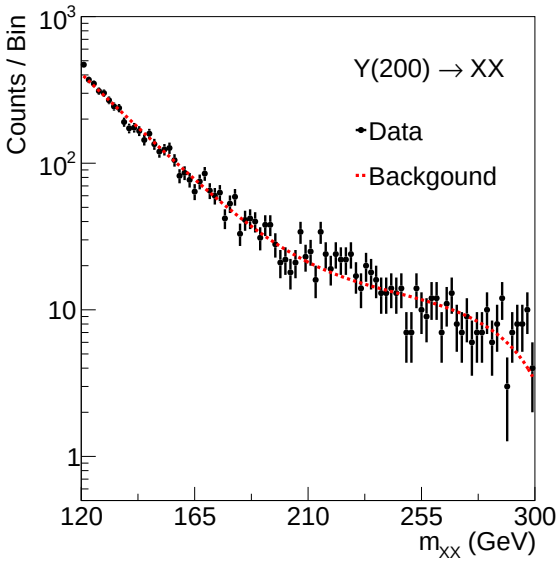



Figure 15: The fitted background

$0\%$ case), beginning from $\pm 0.5\sigma$ up to $\pm 3\sigma$ with a step of $0.5\sigma$. The results can be seen in Figure 18. It is evident that the region $\pm 1.5\sigma$ provides the best performance in terms of significance.

We can then study how the significance changes in the selected region for the various smearing cases in two ways, based on the interpretation of the $\pm 1.5\sigma$ region. One can interpret $\sigma$ as the Gaussian spread of the $0\%$ case and calculate every significance value in the same mass window, resulting in a fixed window study. On the other hand, one can interpret $\sigma$ as the Gaussian spread of each smearing case. That is, the significance will still be calculated at a $\pm 1.5\sigma$, but the range will be different based on the different values of $\sigma$ for every fit, resulting in an adaptive window study. For completeness, we did both studies, and the results are presented in Figure 19. Table 7, summarizes the the values of $\sigma$ (resulting from the fits), and the corresponding mass window for the adaptive widnow search, while table 8, summarizes the amount of signal and background events present in the region of interest of both studies(fixed and adaptive window).

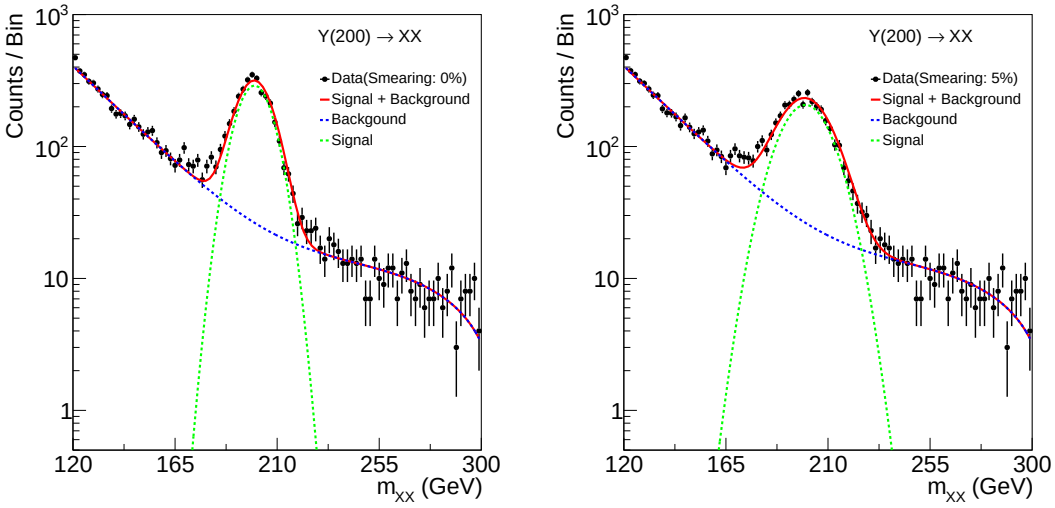
| Smearing % | $\sigma$ in GeV | Invarian Mass $\pm 1.5\sigma$ window in GeV |
|---|---|---|
| 0 | 7.62 | 23.01 |
| 5 | 11.15 | 33.47 |
| 10 | 16.33 | 48.98 |
| 15 | 22.90 | 68.70 |
| 20 | 28.87 | 86.60 |

Table 7: Summary of the invariant mass windows used used in adapitve window study. Note that the resulting window of $0\%$ smearing corresponds to the fixed window case as well.

## 5 Results

So far, in sections 4.3 and 4.4, we have studied how does a multivariate and a singlevariate calssifcation technique responds to energy scale uncertainties, in a signal from background sepparation task. In this chapter, we are going to summarize the results and provide commentary regarding each method, in terms of performance and robustness. Moreover will will draw a comparizon between the two methods.

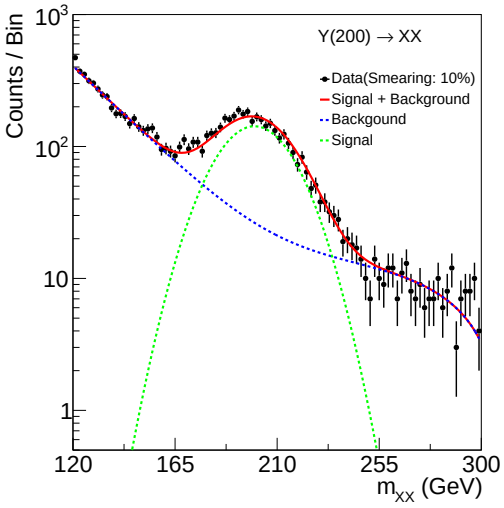
Figure 20, compares the significance yielded by each method, as a function of smearing. Even though, at $0\%$ of smearing, the fit based method, yields a better significance than the BDT method,



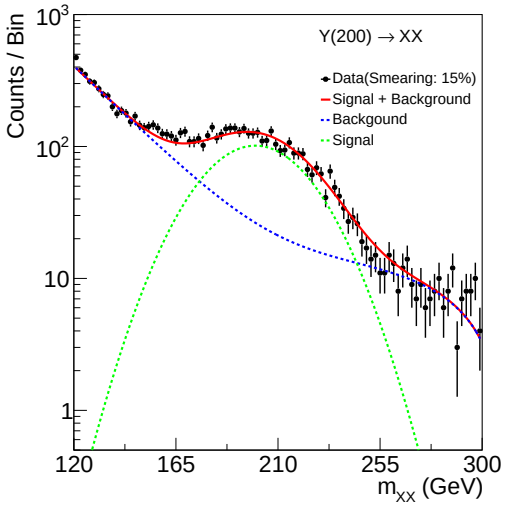

(a)



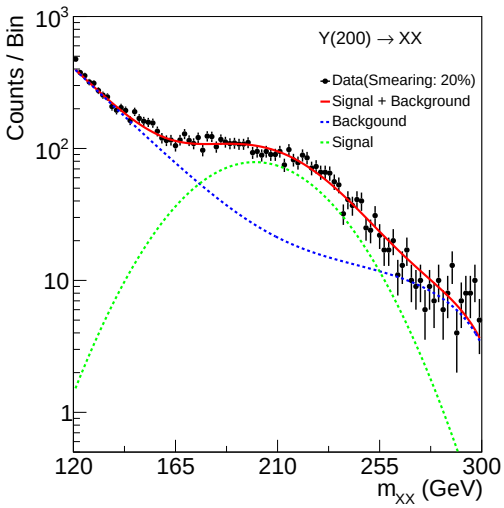

(b)




(c)

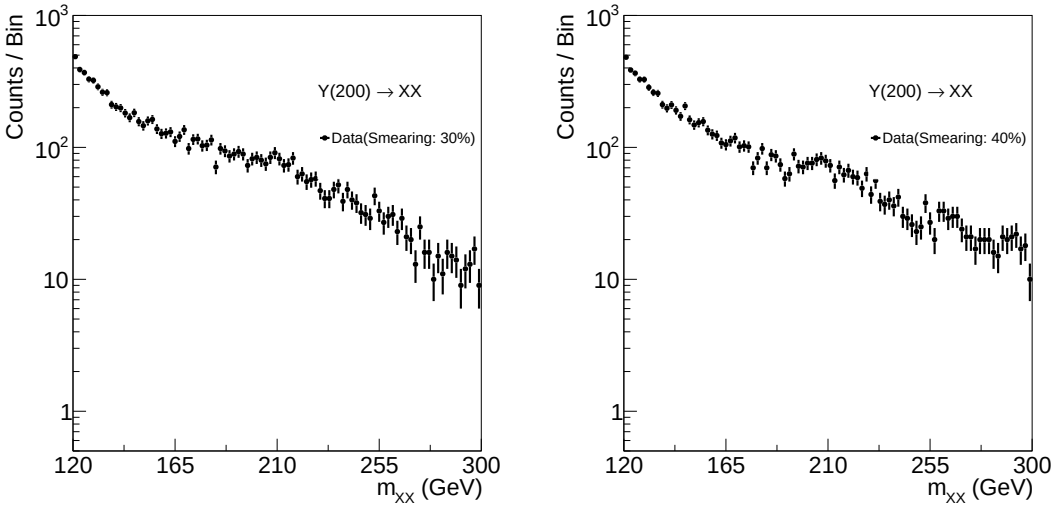



(d)




(e)

Figure 16: Fits for the following smearing cases a: 0%, b:5%, c:10%, d:15%, e:20%



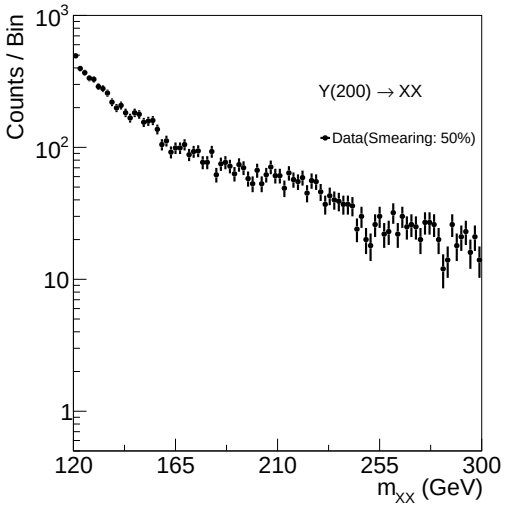




Figure 17: Invariant mass spectra for the extreme smearing cases of : a:30%, b:40% and c:50%. The signal seems indistinguishable from the background in these cases, and therefore a fit based saparation cannot work.

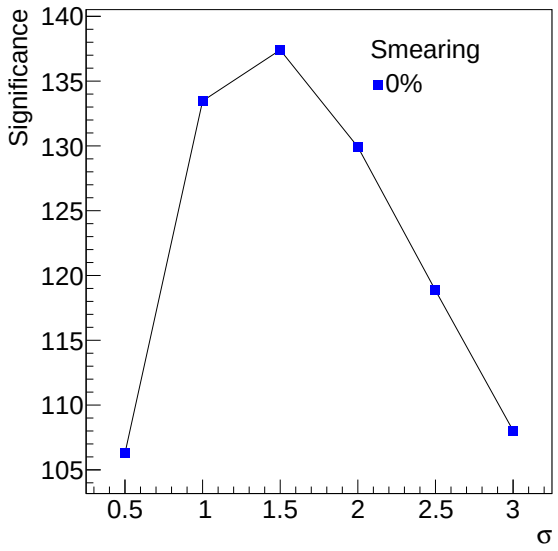



Figure 18: Scan of significance for various values of $\sigma$, in the $0\%$ smearing case. We see that the regrion $\pm 1.5\sigma$ around $\mu = 200GeV$, gives the best significance.

| Smearing % | No. Sig. Events (fixed window) | No. Bkg.Events (fixed window) | No. Sig. Events (adaptive window) | No. Bkg.Events (adaptive window) |
|---|---|---|---|---|
| 0 | 2426 | 311 | 2426 | 311 |
| 5 | 2012 | 311 | 2506 | 476 |
| 10 | 1511 | 311 | 2539 | 752 |
| 15 | 1118 | 311 | 2524 | 1202 |
| 20 | 884 | 311 | 2465 | 1662 |

Table 8: Signal and background events in the 23Gev fixed window region and in the $\pm 1.5\sigma$ adaptive window region, for different smearing percentages.

the bdt is more robust in general. To explain this, on must pay close attention to the feature space used for the training. The classifier learns not only the energy related Pts of the X particles, but also the geometrical features, $\Delta\phi$, $\Delta R$ and $\Delta\eta$. As discribed in section 4.2, smearing has an effect only on the Pt variables, while the spatial features remain invariant under such process. That is, the BDT model, learns to classify the signal, using features that do not change through out the smearing process, and is therefore able to deliver a better performance, when compared to the fit based analysis, which only makes use of the invariant mass, a feature that gets heavilly altered by uncertainties on the energy scale as figures 16 and 17 indicate.

## References

[1] Tianqi Chen and Carlos Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

[2] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. Wiley, 2006. ISBN: 9788126511167. URL: https://books.google.com.cy/books?id=NR-SzW2t7WYC.

[3] D. Griffiths. *Introduction to Elementary Particles*. Physics textbook. Wiley, 2008. ISBN: 9783527618477. URL: https://books.google.com.cy/books?id=Wb9DYrjcoKAC.

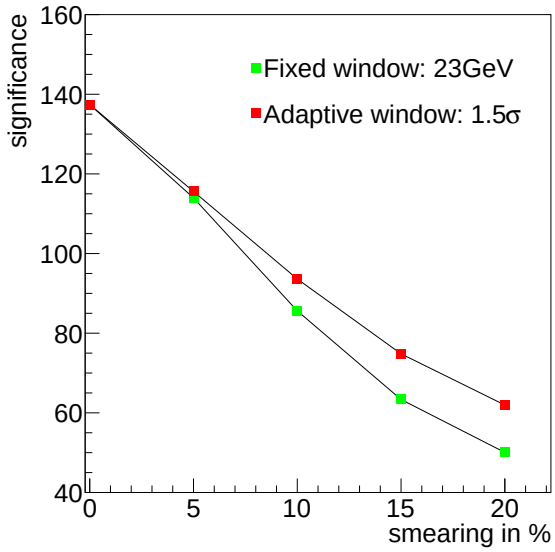



Figure 19: Copmarison of the significance evolution as caclulated in the fixed widow and adaptive window case.

[4] *Introdution to Boosted Trees*. URL: https://xgboost.readthedocs.io/en/stable/tutorials/model.html (visited on 01/28/2022).

[5] Pankaj Mehta et al. "A high-bias, low-variance introduction to Machine Learning for physicists". In: *Physics Reports* 810 (May 2019), pp. 1–124. DOI: 10.1016/j.physrep.2019.03.001. URL: https://doi.org/10.1016%2Fj.physrep.2019.03.001.

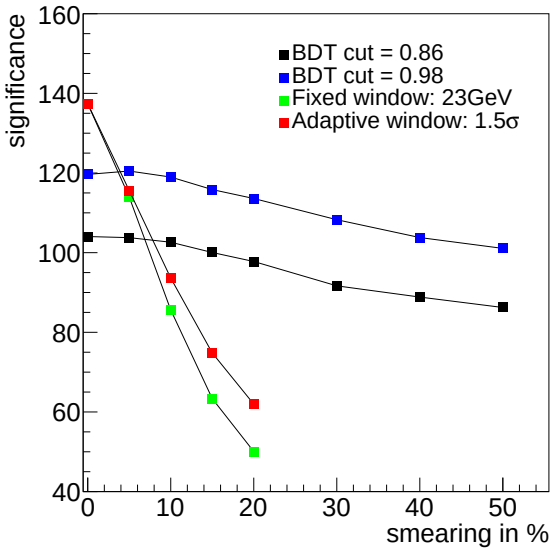



Figure 20: Comparison of the perfomance of the BDT and Fit based analysis, in terms of sifnificance, as a function of the smearing cases. We can see that BDT based analysis, is more robust.