UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Unsupervised learning searches for anomalies in proton-proton collisions

Relatore

Prof. Tommaso Dorigo

Laureanda

Marija Mojsovska

Anno Accademico 2020/2021

Contents

Chapter 1

Introduction

1.1 The standard model

The standard model is a theory that is capable of describing the fundamental structure of matter to high accuracy. The model classifies the elementary particles and describes in detail the behaviour of three of the four fundamental forces (the electromagnetic force, the strong, and weak interactions; the theory does not include the gravitational force). Although this model has successfully explained nearly all experimental outcomes and predicted a wide variety of phenomena, it never gave an explanation to some important questions, such as the different abundance of matter and antimatter or the nature of dark matter. There is also the problem of naturalness with the mass of the Higgs boson, which is much smaller than what could be expected given the large value of many quantum corrections, which appear to cancel to a high level accuracy through a yet to be explained mechanism. These can be considered as indicators that the standard model is incomplete and some new information is necessary to expand our understanding of the universe.

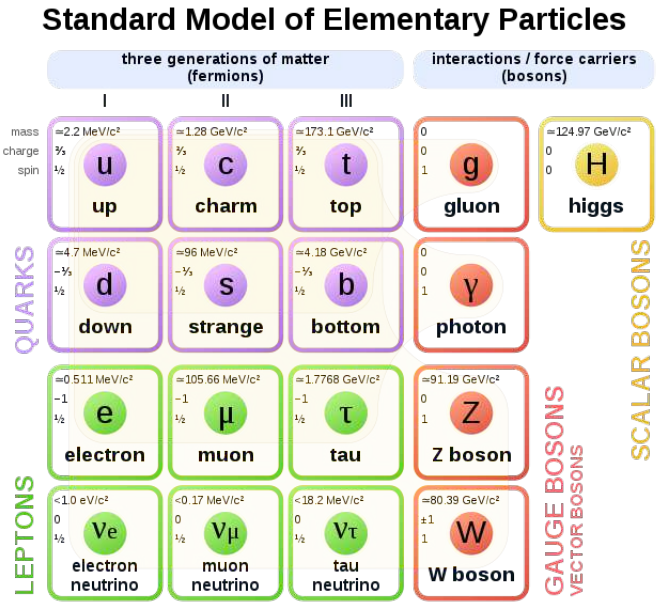

Figure 1.1: The Standard model

1.2 The Large Hadron Collider

The search for new physics beyond the Standard Model is led by the experiments at the world's biggest and most powerful particle accelerator, the Large Hadron Collider (LHC). In the collider two proton

beams are accelerated at velocities close to the speed of light, they travel in separate pipes in opposite direction guided by the strong magnetic field inside the 27 kilometer ring. The beams are made to collide in four locations corresponding to the four particle detectors CMS, ATLAS, ALICE and LHCb.

1.2.1 ATLAS and CMS

ATLAS and CMS are the two detectors with a general purpose that investigate a variety of physics phenomena. They have the same scientific goal but different technical solutions and design.

ATLAS consists of many layers of detectors around the point of collision, each with a specific task. The function of the inner detector is to measure the particle position, momentum and charge. It contains of a Pixel Detector, Semiconductor Tracker, and Transition Radiation Tracker. The inner detector is surrounded by a solenoid magnet. Calorimeters are situated outside the solenoid magnet. They are designed to absorb all the energy of the particles. There is an inner electromagnetic calorimeter and a hadronic calorimeter after that. The particles that leave the calorimeter undetected, apart from the neutrinos, are the muons. The Muon Spectrometer task is to track these muons. It incorporates a Thin Gap Chambers, Resistive Plate Chambers, Monitored Drift Tubes, and Cathode Strip Chambers.

CMS stands for Compact Muon Solenoid. Starting from the point of collision and moving outward it is composed of silicon trackers followed by calorimeters: electromagnetic calorimeter and hadronic calorimeter. These parts of the detector are surrounded by a very powerful solenoid magnet that generates a magnetic field of around 4T. The outside layer is occupied by the muon chambers which have the role of observing the passage of the muons.

1.2.2 Search for new physics at LHC

Most of the research at LHC is done by comparing the results from the experiments with the predictions of the Standard Model. The approach used is supervised classification: a simulation of phenomena consistent with the Standard Model and new hypothetical processes predicted by theories that go beyond the Standard Model. Therefore, the search for new physics is focused on a model specific search.

Currently there is no evidence of new phenomena. There may be many reasons for this: it might be that the searched signals are very rare or that the detectors are not sensitive and/or powerful enough to reveal them. There is also the possibility that the new physics is something that theoreticians have not yet considered, and maybe it is already present in the data but the model specific searches do not allow for it to be discovered.

1.3 Unsupervised research

The difference between a supervised and unsupervised approach to a problem is that in the supervised approach there is a training set of data in which every event is labeled, for example as signal or background. These data allow us to build a model, which may later be used on a new sets of data. In an unsupervised approach the data are not labeled, instead all the data are being confronted in order to find interesting structures in that data set.

In particle physics, the background can be described by models with high precision, but uncertainties in these models can be found in extreme regions of the considered space, like the tails of a distribution. These could be the regions where we expect to find signals of new physics. In such cases, unsupervised searches may provide a useful and complementary, model-independent approach.

The following chapters describe an algorithm that uses an unsupervised approach to search for anomalies in collision data.

Chapter 2

The data

Collisions at LHC generate a large number of particles that interact and decay following complex processes. Around one billion collisions are recorded by the detectors every second. The data are afterwards reduced by selecting interesting events, but still up to one million gigabytes of data are being processed every day at CERN Data Centre.

2.1 Simulated data

To study the performance of the algorithm in controlled and understandable conditions it is useful to generate data with known characteristics that are easy to describe, instead of using the complex data deriving from the detectors.

A simulated dataset is created using a uniform distribution in the interval [0,1] for the background, and a Gaussian distribution with mean μ and variance σ for the signal. The variance is randomly chosen within the interval [0.01, 0.1], while the value of μ is chosen in the interval [3σ, 1-3σ]. While these values appear arbitrary, they allow for the definition of a controlled situation in which it is possible to verify the functioning of the algorithm.

The data are generated fixing the following parameters of the simulation algorithm:

- number of active dimensions
- number of dimensions in which the signal has distinctive behaviour
- number of simulated events
- signal fraction

To each active dimension of all the events, a value for the background is assigned. A signal element is added to this in the given number of dimensions for the signal in order to achieve the desired signal fraction.

2.2 Data preprocessing

The events observed from the proton-proton collisions are characterized by a large number of parameters that form a complex multidimensional space. A dis-uniform distribution with a peak at low values followed by an exponential decrease is common. It is the behavior of the transverse momentum (*Figure 2.1*) or the invariant mass. These wide density distributions are largely the result of the sharply falling density of the proton as a function of parton momentum, and can affect the algorithm. To prevent this possible influence, a series of actions are performed on the original data.

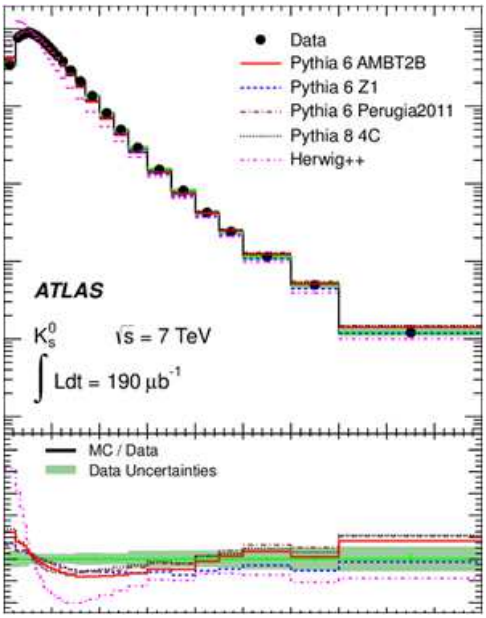

Figure 2.1: Distribution of the transverse moment of Λ barion [8]

2.2.1 Probability integral transformation and copula

A marginal distribution is the probability distribution of each parameter, while the distribution of the vector whose elements are all the parameters is called joint distribution.

To avoid the problem caused by the natural probability distribution of the variables an integral transformation can be applied.

The cumulative distribution function of a random variable with a probability density function $f(t)$ is given by

$$y = F(x) = \int_{\infty}^{x} f(t)dt \tag{2.1}$$

The transformed variable thus obtained, $y = F(x)$ is uniformly distributed in [0,1].

$$F_y(y) = P(Y \leq y) = P\left(F_X(X) \leq y\right) = P\left(X \leq F_x^{-1}(y)\right) = F_X\left(F_X^{-1}(y)\right) = y \tag{2.2}$$

After this procedure is applied on the marginal distribution of each variable, the data structure and dependencies of the variables are contained in a copula C. $C : [0,1]^d \to [0,1]$ is a joint cumulative distribution function of a d-dimensional random vector on the unit cube $[0,1]^d$ with uniform marginals.

Sklar's theorem: *Every multivariate cumulative distribution function* $H\left(x_1, \ldots, x_d\right) = \Pr\left[X_1 \leq x_1, \ldots, X_d \leq x_d\right]$ *of a random vector* $\left(X_1, X_2, \ldots, X_{d_2}\right)$ *can be expressed in terms of its marginals* $F_i\left(x_i\right) = \Pr\left[X_i \leq x_i\right]$ *and a copula* C. *Indeed:* $H\left(x_1, \ldots, x_d\right) = C\left(F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right)$.

In case that the multivariate distribution has a density h, *and if this is available, it holds further that* $h\left(x_1, \ldots, x_d\right) = c\left(F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right) \cdot f_1\left(x_1\right) \cdot \cdots \cdot f_d\left(x_d\right)$ *where* c *is the density of the copula.*

The theorem also states that, given H, *the copula is unique on* $\operatorname{Ran}\left(F_1\right) \times \cdots \times \operatorname{Ran}\left(F_d\right)$ *which is the cartesian product of the ranges of the marginal cumulative distribution functions. This implies that the copula is unique if the marginals* F_i *are continuous.*

The converse is also true: given a copula $C : [0,1]^d \to [0,1]$ *and marginals* $F_i(x)$ *then* $C\left(F_1\left(x_1\right), \ldots, F_d\left(x_d\right)\right)$ *defines a d-dimensional cumulative distribution function with marginal distributions* $F_i(x)$.

After the application of these transformations our data are now neatly packed inside a d-dimensional unit cube where d is equal to the number of analysed parameters. As a consequence it is now easier to detect localized overdensities caused by a hypotetical signal in the multidimensional space.

2.2.2 Principal Component Analysis (PCA)

The technique called "Principal Component Analysis" (PCA) can be used to reduce the dimensionality of a dataset with minimal loss of information. It returns a set of uncorrelated variables (principal components) by applying an orthogonal transformation on the original D dimensional space. This can be seen as a hyperellipsoid fit to the data, where the principal components correspond to the major axes of the ellipse. The first principal component, once an ordering is performed by the components variance, accounts for a large part of the variance of the data, and each next principal component accounts for a smaller residual variance compared to the previous. As a consequence a subspace with $D' < D$ dimensions can be selected by considering the first D' principal components that explain a sufficient amount of the total variance.

We want to keep the algorithm as generic as possible therefore the implementation of the PCA is optional. Its application can be useful when working with a dataset that has an elevated number of characteristic features. However in the cases discussed in the following chapters the PCA is omitted because it was seen that its implementation did not lead to improvement in the results. The reason why PCA can be nocuous in the search for a small signal in a large background is that the variables which show a different behaviour in the signal component may be ones with small variance in the total dataset, when the latter is dominated by a majority of background events.

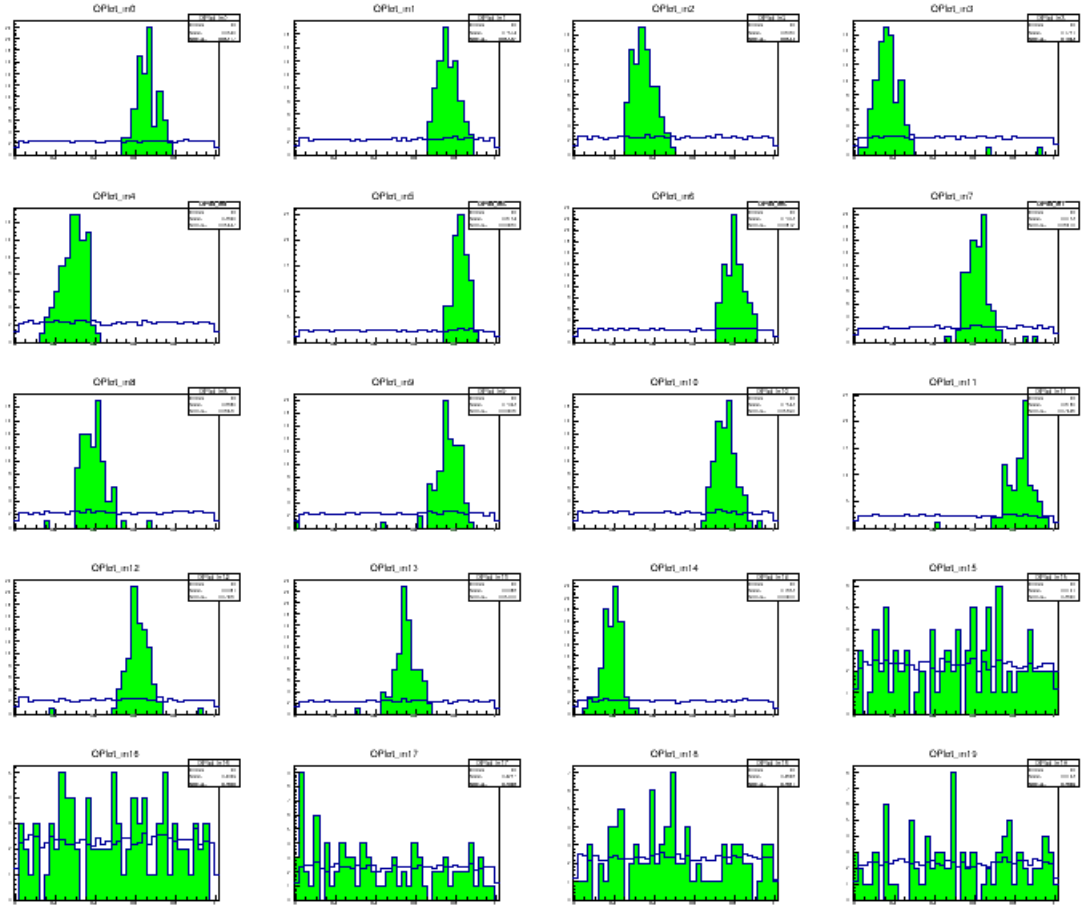

Figure 2.2: Distribution of the simulated data using 1% signal fraction in 10000 events, where 15 of the 20 variables show a Gaussian distribution in the signal component. The blue histogram shows the totality of the data before the application of the RanBox search. The algorithm identifies as the most overdense region a narrow box which exploits the distinctive features of the signal: the green histogram shows the data in the selected region, normalized to the same area of the blue histogram. It is evident how the selected region is enriched in the signal component.

Chapter 3

The RanBox algorithm

The purpose of the RanBox algorithm is to search for overdensities using a box with sides of variable length that is moved around in the multidimensional space trying to maximize a certain statistic. The algorithm is executed multiple times, each time with a different initialization in a different subspace, with smaller dimension compared to the original. The reason for this is that the signal is not expected to present distinctive characteristics in all the analyzed variables.

3.1 Initialization of the box

The algorithm allows for three different choices of how to initialize the box before a gradient descent search is performed in the selected subspace. It can be initialized at random, by finding a dense region with a cluster search, or by maximizing the kernel density. The first method has an advantage of being very fast, while the others give better results at a cost of longer execution time.

3.1.1 Random initialization

The algorithm creates an n-dimensional box that has k sides with bounds chosen randomly in $[0, 1]$, and $n - k$ fixed sides with upper bound equal to 1 and lower bound equal to 0. The value of k is chosen in such a way that the expected value of the number of events contained in the box is equal to 10, so that fluctuations are unlikely to make it too small or zero.

We consider the bounds of an interval in $[0, 1]$ as two independent random variables X, Y with uniform distribution $f_X(x)$ and $f_Y(y)$ in $[0, 1]$. The average length of an interval with bounds chosen randomly is:

$$\begin{aligned}
\mathbf{E}(|X-Y|) &= \int_0^1\int_0^1 |x-y| f_X(x) f_Y(y) dxdy = \int_0^1\int_0^1 |x-y| dxdy \\
&= \int_0^1\int_0^1 (x-y)\cdot \mathbf{I}_{x>y} + \int_0^1\int_0^1 (y-x)\cdot \mathbf{I}_{y>x} \\
&= 2\cdot\int_0^1\int_y^1 (x-y) dxdy = 2\cdot\int_0^1 \left[\frac{x^2}{2} - xy\right]_y^1 dy \\
&= 2\cdot\int_0^1 \left[\frac{1}{2} - y - \frac{y}{2} + y^2\right] dy = \int_0^1 \left[1 + y^2 - 2y\right] dy \\
&= \left[y + \frac{y^3}{3} - y^2\right]_0^1 = \frac{1}{3}
\end{aligned} \tag{3.1}$$

Now the volume of the box is equal to

$$v = 1^{(n-k)} \cdot \left(\frac{1}{3}\right)^k \tag{3.2}$$

Inside a box of volume v we expect a $N_{exp} = N_{tot} \cdot v$ number of points under the hypothesis of uniformity. Fixing $N_{exp} = 10$ we obtain

$$\frac{10}{N_{tot}} = 1^{(n-p)} \cdot \left(\frac{1}{3}\right)^k \quad \Longrightarrow \quad k = \frac{\log\left(\frac{10}{N_{tot}}\right)}{\log\left(\frac{1}{3}\right)} \tag{3.3}$$

For each cycle of the algorithm k out of the n variables are randomly selected, with k calculated using the above formula, and their bound are chosen randomly in $[0, 1]$.

3.1.2 Clustering

This initialization method is based on the calculation of the Euclidean distance between points in the data. The first step is to randomly choose k variables, with k given by the *Equation 3.3*, and search for the nearest neighbour j of every data point i in the k-dimensional space. Then for every point i we calculate the number of neighbouring points which have i as their closest neighbour. The box is now centered around the point which has the shortest distance for the largest number of points compared to the others.

In the next iterations $k/2$ variables that in the previous iteration presented the largest number of data points in an interval of their domain are maintained while the rest are chosen randomly.

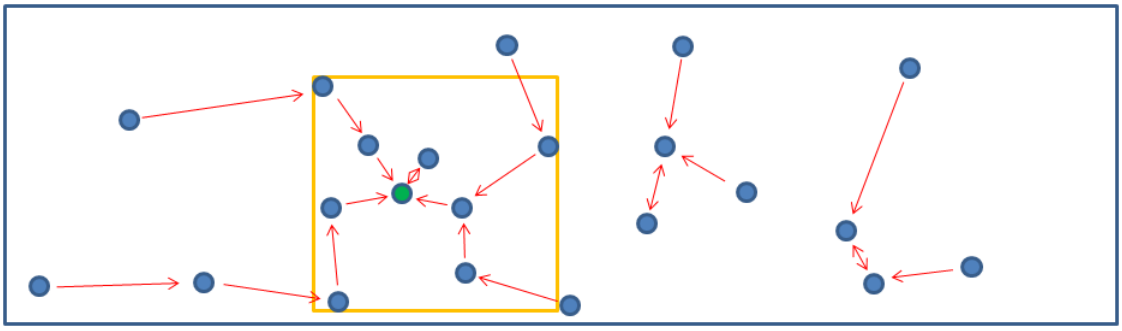

Figure 3.1: Graphical representations of the clustering algorithm

3.1.3 Initialization using kernel density

This algorithm searches for the initial box boundaries using a kernel estimation of the density. All data points are substituted with a d-dimensional Gaussian distribution. The density is then calculated at the position of each of the N events as sum of the N d-dimensional Gaussian distributions. Now the initial box boundaries are $[\max(x_{hd} - k, 0.), \min(x_{hd} + k, 1.)]$ where x_{hd} is the point of highest density and k is fixed at 0.2.

3.2 Test statistic

To verify the efficiency of a proposed model in describing an observed phenomena a statistical hypotheses testing is performed on the data sample.

We consider that the phenomena are governed by the random variable X with a probability distribution $f(x; \Theta)$ where Θ is a set of unknown parameters, and define the hypotheses. In our case the null hypothesis H_0 states that the data are a result of random background phenomena, whereas the alternative hypothesis H_1 assumes that the data are a product of a signal that differs from the background. These hypotheses are described by a specific parameters

$$H_0 : \theta \in \Theta_0 \tag{3.4}$$

$$H_1 : \theta \in \Theta_1 \tag{3.5}$$

with $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Omega$.

At this point we need to choose a test statistic $t(X)$ and fix a critical region. If t is located inside the critical region the null hypothesis H_0 is rejected. The probability of rejecting the null hypotheses when it is true is called error of the first kind α, correspondingly error of the second kind β is the probability of accepting H_0 when false. Power of a test $1 - \beta$ is the probability that the test correctly rejects the null hypotheses. We can fix the value of α and try to maximize the power of the test. According to the lemma of Neyman and Pearson, if both hypotheses are simple the most powerful test statistic is the likelihood ratio. The hypotheses are simple if the unknown parameters are specified as single values, $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$.

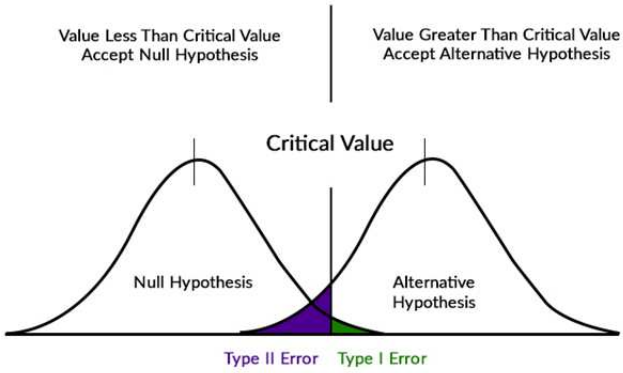

Figure 3.2: Distribution of the H_0 and H_1 hypotheses with the errors of first and second type

The likelihood L of a hypothesis H to which corresponds a probability density $f(x, \theta)$ is equal to the product of the probability density function of the n independent observations x_i

$$L\left(\mathbf{X} \mid \theta\right) = \prod_{i=1}^{n} f\left(x_i, \theta\right) \tag{3.6}$$

It is possible to find a parameter $\hat{\theta}$ that maximizes the likelihood, this is the maximum likelihood estimate.

$$L\left(\mathbf{X} \mid \hat{\theta}\right) = \max_{\theta \in \Omega} L\left(\mathbf{X}, \theta\right) \tag{3.7}$$

Wilks's theorem: *Assume $X = (x_1, x_2, ..., x_n)$ is the observed data and $\Theta = (E, T) = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_r, \tau_1, \tau_2, \ldots, \tau_S)$ are the unknown parameters. The null hypothesis is $H_0 : E_0 = (\epsilon_{10}, \epsilon_{20}, \ldots \epsilon_{r0})$ and $H_1 : E_1 \neq E_0$ is the alternative hypothesis. The likelihood ratio is defined as*

$$\lambda(\mathbf{x_n}) = \frac{L\left(X \mid E_0, \hat{T}_c\right)}{L(X \mid \hat{E}, \hat{T})} = \frac{P_r\left(X \mid E_0, \hat{T}_c\right)}{P_r(X \mid \hat{E}, \hat{T})} \tag{3.8}$$

$\hat{E}, \hat{T}$ are the maximum likelihood estimate of the parameters E and T, $\hat{T}_c$ is the maximum likelihood estimate when $E = E_0$. Under the null hypothesis the variable $-2\log\left(\lambda\left(\mathbf{x_n}\right)\right)$ asymptotically approaches the χ^2 distribution with r degrees of freedom.

$$-2\log\left(\lambda\left(\mathbf{x_n}\right)\right) \longrightarrow \chi^2_r \tag{3.9}$$

The test statistic $Z = \sqrt{-2\log\left(\lambda\left(\mathbf{x_n}\right)\right)}$ will be used in the present case. According to Wilks's theorem, under the null hypothesis Z depends on a single parameter and follows a normal distribution since it is a square root of a variable with a χ^2 distribution.

The on-off problem studied by the astrophysicists Li and Ma in the context of searches for gamma-ray sources in the sky, which produce a signal only when the telescope points "on source", can be applied here. The objective is to see how significant is the observation of N_{on} points inside of the box out of $N = N_{on} + N_{off}$ total data points. The null hypothesis in this case states that there is no signal in the box $H_0 : \mathbf{E}(N_S) = 0$, on the contrary the alternative hypothesis is $H_1 : \mathbf{E}(N_s) \neq 0$. Under the assumption of a uniform background distribution the expected values for the background and signal in a box of volume v are:

$$\mathbf{E}\left(\hat{N}_b\right) = \frac{v}{1-v} N_{off} \tag{3.10}$$

$$\mathbf{E}\left(\hat{N}_s\right) = N_{on} - \mathbf{E}\left(\hat{N}_b\right) = N_{on} - \frac{v}{1-v} N_{off} \tag{3.11}$$

If the null hypothesis is true, we do not expect to find signal in the box

$$\mathbf{E}\left(\hat{N}_s\right) = 0 \quad \Rightarrow \quad \mathbf{E}\left(\hat{N}_b\right) = v \cdot (N_{on} + N_{off}) \tag{3.12}$$

Now we can calculate the likelihood functions and apply Wilks's theorem.

$$\begin{aligned} L\left(X \mid E_0, \hat{T}_c\right) &= \Pr\left[N_{on}, N_{off} \mid \mathbf{E}\left(N_s\right) = 0, \mathbf{E}\left(N_b\right) = v \cdot (N_{on} + N_{off})\right] \\ &= \Pr\left[N_{on} \mid \mathbf{E}\left(N_{on}\right) = v\left(N_{on} + N_{off}\right)\right] \cdot \Pr\left[N_{off} \mid \mathbf{E}\left(N_{off}\right) = \frac{1}{1-v} \cdot (N_{on} + N_{off})\right] \\ &= \left[\frac{\left[v \cdot (N_{on} + N_{off})\right]^{N_{on}}}{N_{on}!}\right] \cdot \exp\left[-v \cdot (N_{on} + N_{off})\right] \cdot \\ &\quad \cdot \left[\frac{\left[\frac{1}{1-v} \cdot (N_{on} + N_{off})\right]^{N_{off}}}{N_{off}!}\right] \cdot \exp\left[\frac{1}{1-v} \cdot (N_{on} + N_{off})\right] \end{aligned} \tag{3.13}$$

$$\begin{aligned} L(X \mid E, T) &= \Pr\left[N_{on}, N_{off} \mid \mathbf{E}\left(N_s\right) = N_{on} - \frac{v}{1-v} N_{off}, \mathbf{E}\left(N_b\right) = \frac{v}{1-v} N_{off}\right] \\ &= \Pr\left[N_{on} \mid \mathbf{E}\left(N_{on}\right) = N_{on}\right] \cdot \Pr\left[N_{off} \mid \mathbf{E}\left(N_{off}\right) = N_{off}\right] \\ &= \left[\frac{N_{on}^{N_{on}}}{N_{on}!}\right] \cdot \exp\left[-N_{on}\right] \cdot \left[\frac{N_{off}^{N_{off}}}{N_{off}!}\right] \cdot \exp\left[-N_{off}\right] \end{aligned} \tag{3.14}$$

The likelihood ratio is given by

$$\lambda\left(\mathbf{x_n}\right) = \frac{L\left(X \mid E_0, \hat{T}_c\right)}{L(X \mid \hat{E}, \hat{T})} = \left[v \cdot \left(\frac{N_{on} + N_{off}}{N_{on}}\right)\right]^{N_{on}} \left[\frac{1}{1-v} \cdot \left(\frac{N_{on} + N_{off}}{N_{off}}\right)\right]^{N_{off}} \tag{3.15}$$

If N_{on} and N_{off} are not too small ($N_{on}, N_{off} \geq 10$), $\sqrt{-2\log\left(\lambda\left(\mathrm{x_n}\right)\right)}$ follows a normal distribution as a consequence of the single parameter dependence of the null hypothesis ($\mathbf{E}(\hat{N_S}) = 0$).

$$Z_{PL} = \sqrt{-2\log\left(\lambda\left(\mathrm{x_n}\right)\right)} = \sqrt{2} \cdot \left[N_{on} \log\left[v \cdot \left(\frac{N_{on} + N_{off}}{N_{on}}\right)\right] + N_{off} \log\left[\frac{1}{1-v}\left(\frac{N_{on} + N_{off}}{N_{off}}\right)\right]\right] \tag{3.16}$$

We arrived at this result assuming that the background is distributed uniformly, which can be considered true for synthetic datasets. When there is a departure from this distribution, which is what we expect in complex real datasets from collider data taking, we can define a *sidebands* region that surrounds the box and use the data in immediate vicinity of the box to estimate the expected number

of data points in the box. By confining the region where we collect the data density we reduce the statistical power of our estimate of the density in the box, but we reduce the bias due to the variations in the overall density of background events in the vicinity of the box, which may be significantly different from the average.

If $[x^i_{min}, x^i_{max}], i = 1, ...D$ are the boundaries of the search box in the D-dimensional space, the sidebands region is defined by

$$\begin{aligned} \delta_i &= 0.5\left(x^i_{\max} - x^i_{\min}\right)\left(2^{1/D} - 1\right) \\ x^i_{\min,SB} &= \max\left(0, x^i_{\min} - \delta_i\right) \\ x^i_{\max,SB} &= \min\left(1, x^i_{\max} + \delta_i\right) \end{aligned} \tag{3.17}$$

Equations 3.10 and 3.11 can be rewritten assuming that N_{off} is the number of data points in the sidebands region

$$\mathbf{E}\left(\hat{N}_b\right) = \tau N_{off} \tag{3.18}$$

$$\mathbf{E}\left(\hat{N}_s\right) = N_{on} - \mathbf{E}\left(\hat{N}_b\right) = N_{on} - \tau N_{off} \tag{3.19}$$

where $\tau = \frac{v_{box}}{v_{SB}}$.

With these adjustments the Z_{PL} function becomes:

$$Z_{PL} = \sqrt{2} \cdot \left[N_{on} \log\left[(1+\tau) \cdot \left(\frac{N_{on}}{N_{on} + N_{off}} \right) \right] + N_{off} \log\left[\frac{1+\tau}{\tau} \left(\frac{N_{off}}{N_{on} + N_{off}} \right) \right] \right] \tag{3.20}$$

The Z_{PL} test statistic is effective when searching for anomalies that occupy large volumes, the expected data points in the box go up to hundreds. Small anomalies well confined in the search volume can be identified more effectively with the R test statistic, this is the case when we expect only a few data points inside the box.

$$R_{reg} = \frac{N_{in}}{S + N_{exp}} \tag{3.21}$$

S is a normalization constant used to prevent the convergence of the algorithm to arbitrary small boxes, its value is usually set equal to 1.

The R test statistic is advantageous when considering real data with complicated distributions of the parameters. Z_{PL} is more sensitive to disuniformities on a large scale that in this case can be result of the collective behavior of the background.

3.3 Maximization of the test statistic

After the initialization, the box is modified and moved around in order to maximize the test statistic. The sides of the box can be increased or decreased by a quantity $\lambda_{(k,i)}$ where $k = 1, ...D$ indicates the variable and i is the iteration of the algorithm, the algorithm performs a maximum of N_{GD} iterations with N_{GD} set at 100. The modifications of the box are performed independently for each variable k. Both upper and lower bound of a side of the box can be modified with caution not to go out of the interval $[0, 1]$. At the beginning the value of the step $\lambda_{(k,i)}$ is set equal to 0.2. For each iteration the value of the test statistic t is calculated with two possible outcomes:

1. $t_{new} \leq t$: The movement of the box leads to a region with smaller value of the test statistic, in this case the box boundaries remain unchanged and the step $\lambda_{(k,i)}$ is reduced by a $\epsilon = 0.01$.
2. $t_{new} > t$: The new box position increases the value of the test statistic, the box boundaries are updated and the value of $\lambda_{(k,i)}$ is modified using the following approach:

- if the movement of the box expands the length of the side by reducing the lower bound but this one remains > 0, $\lambda_{(k,i)} \to 1.5\lambda_{(k,i)}$, if instead it is < 0, $\lambda_{(k,i)}$ is set equal to ϵ and the lower bound is set equal to 0.
- if the movement of the box reduces its size, $\lambda_{(k,i)} \to 1.5\lambda_{(k,i)}$.
- if the movement of the box expands the length of the side by increasing the upper bound, $\lambda_{(k,i)} \to 1.5\lambda_{(k,i)}$, but if its value exceeds 1 $\lambda_{(k,i)}$ is set equal to ϵ and the upper bound is set equal to 1.

The increase of the step $\lambda_{(k,i)}$ by a factor of 1.5 is an arbitrary choice, but it is a value that leads to good results, in terms of the amount of signal found in the final box, and faster convergence.

If $\lambda_{(k,i)} \leq \epsilon$ for all the k variables the best box is found and the iteration cycle is interrupted.

3.4 Example of algorithm performance using synthetic data

A synthetic dataset is generated using a random number generator as described in *Chapter 2*. Signal is produced with distinctive behavior, Gaussian distribution, in 10 out of the 20 active dimensions with 1% signal fraction in 5000 events. A 6-dimensional search box is used with random initialization and Z_{PL} as the test statistic to be maximized.

With the above choice of parameters, the algorithm converges after 59 iterations to a box that is characterized with $N_{in} = 44$ data points when $N_{exp} = 0.21$ are expected, out of the 44 events in the box $N_s = 43$ are signal events. The volume of the box is 4×10^{-5}.

The 6 variables of the best box are shown in *Figure 3.3*. In these graphs the background distribution is shown in blue, the signal is green, while the red indicates events that are excluded from the box only by virtue of the chosen box interval in the shown dimension.

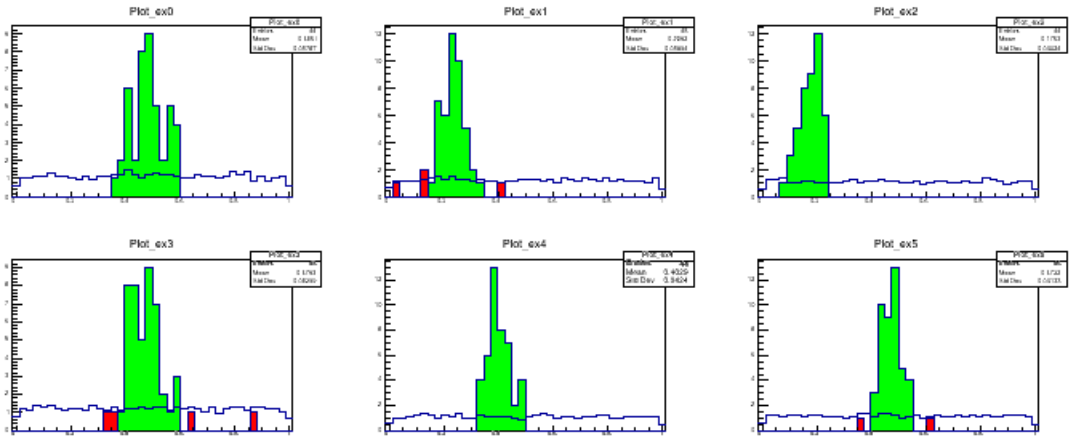

Figure 3.3: Distribution of the background (in blue), signal (in green) and of the events excluded from the box only in the shown dimension (in red), for the six variables that define the best box found by the RanBox algorithm in a run on 5000 synthetic events with 1% signal present in 10 of the 20 variables. The distribution in blue and the sum of the distributions in green and red are normalized to 1.

In *Figure 3.4* are shown scatterplots of all the combinations of the 6 variables of the best box. In blue are all the events, those that are included in the box are shown in green and in red are events that are excluded from the best box only because of their value in the shown dimensions.

Table 3.1 presents the characteristics of the five most significant boxes. We can see that the boxes contain a large amount of signal events and almost all the data points inside them are signal related.

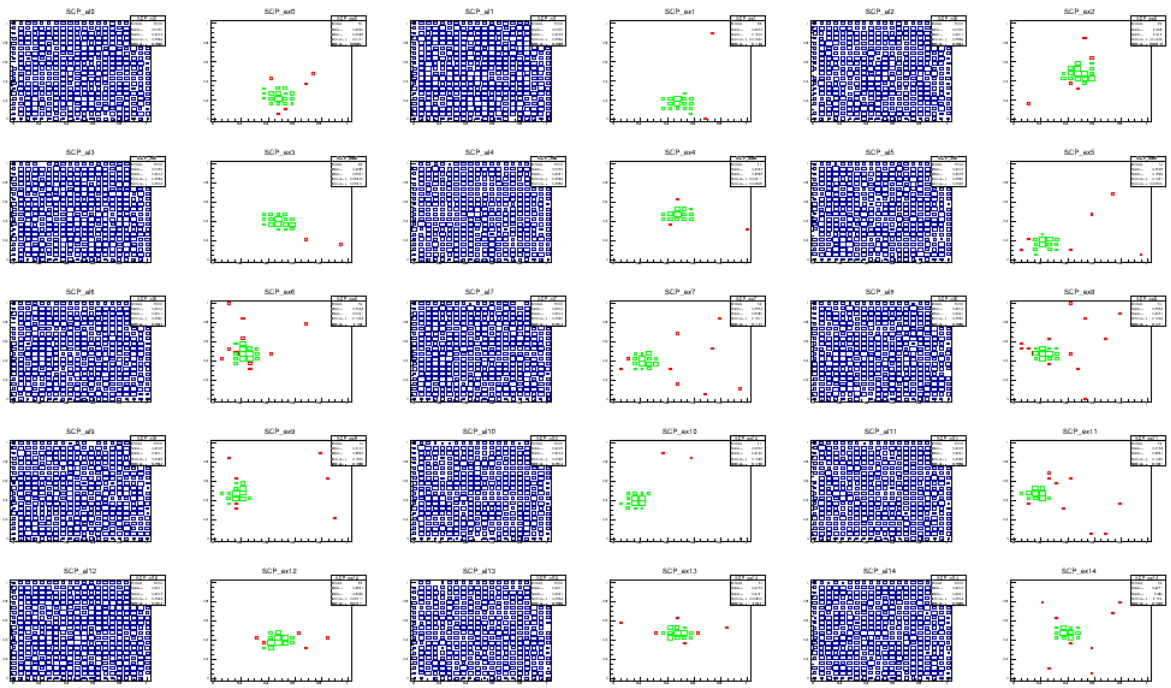

Figure 3.4: Scatterplots of the six variables defining the best box to which the RanBox algorithm converges in a run on 5000 synthetic events with 1% signal present in 10 of the 20 variables. On the left side of each pair of graphs in blue is shown the distribution of the entire data, while on the right side is the distribution of the selected events in green, and in red the distribution of events that would have been included in the box if not for their value in the shown dimensions

Z_{PL}	N_{in}	N_s	signal fraction in box	signal to background gain
19.52	44	43	86	97.7273
19.49	42	42	84	100
18.71	45	45	90	100
17.85	46	45	90	97.8261
16.65	39	37	74	94.8718

Table 3.1: Characteristics of the five most significant boxes

3.5 RanBoxIter

Given the dimensionality of the search box, its dimensions are chosen at random by the RanBox algorithm. The probability that the random choice of b dimensions out of n results in selecting x dimensions in which signal is present when s is the number of dimensions with signal, is given by the formula:

$$P(n, b, x, s) = \frac{C(s, x)\ C(n - s, b - x)}{C(n, b)} \tag{3.22}$$

where $C(n, k) = \frac{n!}{k!(n-k)!}$ represents the number of k-combinations from a set of n elements.

The following tables show the probability associated with $b = 5, 6...10$ dimensional box in $n = 20$ dimensional space. In *Table 3.2* the signal is considered to have distinguishing features in 8 of the 20 dimensions, in *Table 3.3* in 10 dimensions, and in *Table 3.4* in 12 dimensions. The values of the probability in these tables indicate that to find a significant signal associated with an overdensity, the RanBox algorithm has to try a large number of combinations, especially when the probability to select most of the significant components of the signal is small. The probabilities are therefore a measure of the inverse of the number of trials required in order to have sensibility to the signal in different conditions.

	x=3	4	5	6	7	8	9	10
b=5	0.23839	0.05418	0.00361					
6	0.31785	0.11919	0.01734	0.00072				
7	0.35758	0.19866	0.04768	0.00433	0.00010			
8	0.35208	0.27506	0.09781	0.01467	0.00076	7.9384×10^{-6}		
9	0.30807	0.33008	0.16504	0.03667	0.00314	7.1445×10^{-5}	X	
10	0.24006	0.35008	0.24006	0.07502	0.00953	0.00036	X	X

Table 3.2: Probability calculations when the signal is distinguishable in 8 of the 20 dimension ($s = 8$)

	x=3	4	5	6	7	8	9	10
b=5	0.3483	0.13545	0.01625					
6	0.37152	0.24381	0.06501	0.00542				
7	0.32508	0.32508	0.14628	0.02709	0.00155			
8	0.24006	0.35008	0.24006	0.07502	0.00953	0.00036		
9	0.15004	0.31507	0.31507	0.15004	0.03215	0.00268	5.9538×10^{-5}	
10	0.07794	0.23869	0.34372	0.23869	0.07794	0.01096	0.00054	5.4125×10^{-6}

Table 3.3: Probability calculations when the signal is distinguishable in 10 of the 20 dimensions ($s = 10$)

	x=3	4	5	6	7	8	9	10
b=5	0.39732	0.25542	0.05108					
6	0.31785	0.35758	0.16347	0.02384				
7	0.19866	0.35758	0.28607	0.09536	0.01022			
8	0.0978	0.27506	0.35208	0.20538	0.0503	0.00393		
9	0.03667	0.16504	0.33008	0.30807	0.13203	0.02358	0.00131	
10	0.00953	0.07502	0.24006	0.35008	0.24006	0.07502	0.00953	0.00036

Table 3.4: Probability calculations when the signal is distinguishable in 12 of the 20 dimensions ($s = 12$)

RanBoxIter is a variation of the original RanBox algorithm that uses a different approach for selecting the box variables. An incremental scan is performed using a fixed number of boxes, usually 20. Starting with 2 variables, all the possible boxes are tested and ordered by the value of the test statistic. This means that for each combination of two variables the box is initialized and moved around in search for maximization of the test statistic. Then a third dimension is added to the 20 most significant 2-dimensional boxes, and the cycle is repeated. At each iteration a new dimension is included and the process is repeated until the number of desired dimensions for the final box is reached. At the end the algorithm returns the box that corresponds to the highest value of the test statistic in the last iteration as the best box. However, it is possible that a box with smaller dimensionality gives better results in comparison with the best box and therefore it is useful to also look at the full list of the most significant boxes and their characteristics.

To study the performance of the algorithm, power tests are carried out. First we want to see how the results vary in function of the signal fraction. 5000 events are simulated in a 20 dimensional space with signal present in 15 out of the 20 dimensions. The random initialization of the box is used and Z_{PL} is the applied test statistic.

The distribution for the null hypothesis is obtained running the algorithm 500 times using only flat

background data. A fit with Gamma function is performed in order to find the critical regions corresponding to error of first type $\alpha = 0.05, 0.01$ and 0.001. Now for each value of the signal fraction 10 runs of the algorithm are done from which the value of the power is calculated as the fraction of values in the critical regions for every value of α. The power functions are shown in *Figure 3.6*. We can see that for signal fractions $\geq$0.003 the value of the power converges at 1.

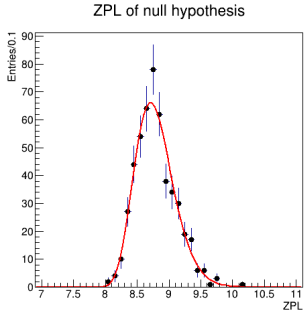

Figure 3.5: Distribution of the null hypotheses

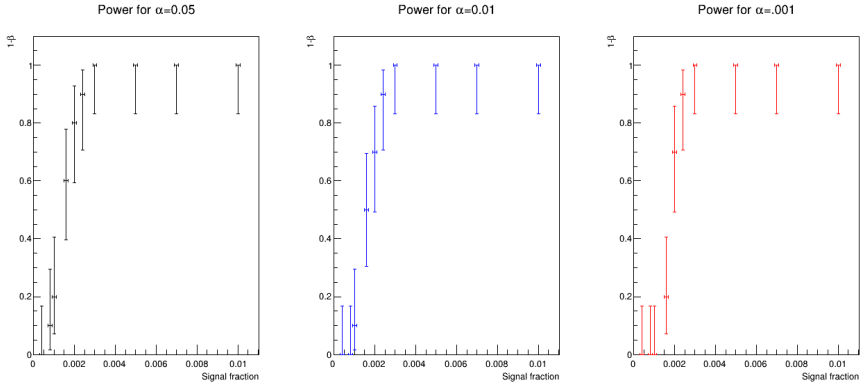

Figure 3.6: Power curves as a function of the signal fraction. Confidence intervals are estimated using the Clopper-Pearson method [7]

Another power test is done varying the number of dimensions in which the signal is present. The signal fraction for this test is fixed at 1%. Observing *Figure 3.7* it is very clear that the performance of the algorithm starts dropping when the number Gaussian dimensions gets to 4 and lower.

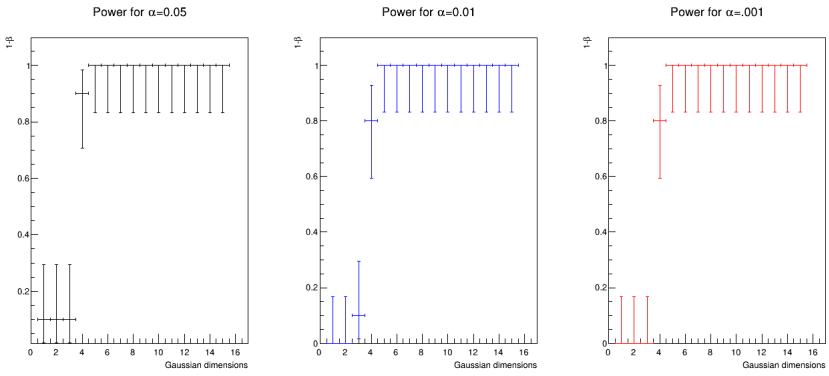

Figure 3.7: Power curves as a function of the number of dimensions in which a signal is present. Confidence intervals are estimated using the Clopper-Pearson method [7]

Chapter 4

Application on data from proton-proton collisions

The algorithm was designed to search for signs of new physics phenomena in data coming from particle collisions. A local increase of density in the multidimensional space of the kinematic variables that describe the particles involved in the collision can indicate a potentially interesting correlation that can be studied further.

4.1 HEPMASS dataset

To test the algorithm we use the "HEPMASS" dataset from the UCI Machine Learning Repository database [5]. The events in this dataset are produced by simulating a decay of an unknown particle X in $t\bar{t}$ and the following decay mode:

$$t\bar{t} \to W^+ b W^- \bar{b} \to qq' b\ell\nu\bar{b} \tag{4.1}$$

For background events the Standard model $t\bar{t}$ production is considered. As a result the final state is identical, but the kinematic features are different due to the presence of intermediate resonance. In *Figure 4.1* are shown Feynman diagrams of the signal and background processes.

The events are generated considering ATLAS configuration for detector simulation and only events that agree with the hypothesis of being a result of a decay of pair of top quarks are selected.

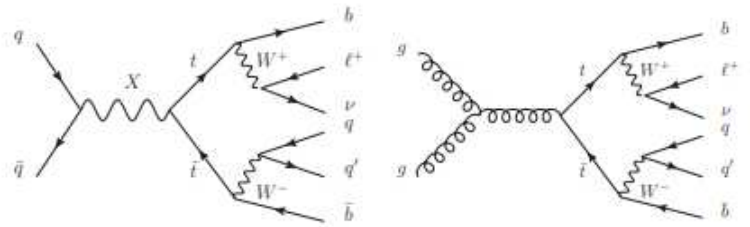

Figure 4.1: Feynman diagrams of the signal process on the left, and the background process on the right [5]

The data are characterized by 27 features that describe each event. A standard reconstruction of a proton-proton collision is performed, identifying jets, leptons and b-jets. From the reconstruction of the event, the low-level kinematic features obtained are the four vectors: the momentum of the leading lepton, the momentum of the four leading jets, b-tagging information for the jets, and the missing transverse momentum magnitude and angle. The transverse momentum cannot be measured directly, it is instead deduced from the imbalance of momentum in the final state, it is the opposite of the sum of the momentum vectors of all observed particles, calculated in the transverse plane of the

beams. The high level features of the set are the values of the invariant masses of the intermediate objects calculated using the low-level kinematic features. The invariant masses are: $m_{\ell\nu}$ from the decay process $W \to \ell\nu$, m_{jj} from the $W \to qq'$ process, m_{jjj} from the $t \to Wb \to bqq'$ process, $m_{j\ell\nu}$ from the $t \to Wb \to \ell\nu b$ process, and the m_{WWbb} mass of the unknown particle $X \to t\bar{t}$.

4.2 Tests with the HEPMASS dataset

The difficulty of using this dataset in comparison with a synthetic dataset is the fact that overdensities may also be a product of background processes. To get better results, the sidebands region is used for the estimation of the density, and because we are searching for small anomalies produced by the signal, the R test statistic is favoured over the Z_{PL} test.

A number of different tests are performed to see how the behavior of the algorithm changes varying its parameters. In the first test the RanBox algorithm is used to search for 2% signal in 10000 events. We vary the dimensionality of the search box from 5 to 12 and for each value 100, 500, 1000, 5000 and 10000 trials are carried out. More trials means more explored subspaces as every trial considers different dimensions for the search box. For every combination of parameters the algorithm is run 10 times and in *Table 4.1* and *Table 4.2* the average values are shown. As expected we can see how the signal to background gain ($SBG = \frac{N_s}{N_{in}} \frac{1}{signal fraction}$) increases with the increasing of N_{trials}.

N_{trials}	100			500			1000		
N_{var}	R	N_s	SBG	R	N_s	SBG	R	N_s	SBG
5	15.68	6.9	11	22.02	12.3	21.11	23.97	10.9	15.28
6	19.254	7.6	12.171	21.604	8.6	16.578	24.471	10.7	12.77
7	18.561	5	9.687	25.647	6.7	13.355	25.795	6.2	10.488
8	22.045	10.1	20.906	27.808	6.3	10.566	27.857	5.4	7.586
9	21.79	2	4.56	28.036	5.7	8.966	30.283	14.7	15.349
10	21.879	2.3	4.826	32.554	6.5	8.526	37.203	6.4	7.748
11	23.537	1.3	2.837	28.408	8	12.233	35.225	12.5	17.18
12	26.181	5.8	9.379	35.975	7.4	8.786	36.647	14.8	17.154

Table 4.1: Results of the RanBox algorithm when varying the dimensionality of the search box for N_{trials} = 100, 500 and 1000

N_{trials}	5000			10000		
N_{var}	R	N_s	SBG	R	N_s	SBG
5	27.671	18.4	24.894	26.237	22.2	28.545
6	30.568	11.4	17.886	28.164	16.5	23.807
7	34.427	14.3	16.656	32.751	18.6	22.774
8	34.856	14.3	17.428	34.911	17.2	22.015
9	39.234	27.5	26.863	36.564	7.5	8.662
10	40.973	16.2	18.481	36.371	14.7	16.425
11	42.698	14	15.539	40.169	14.5	11.329
12	41.696	14.1	15.876	41.79	9.1	10.628

Table 4.2: Results of the RanBox algorithm when varying the dimensionality of the search box for N_{trials} = 5000 and 10000

For comparison we look at the results that the iterative version of the algorithm RanBoxIter produces in the same situation (2% signal in 10000 events). Looking at *Table 4.3* we can notice how the increment of the dimensionality of the box produces an increase in the signal to background gain.

N_{varmax}	R	N_s	SBG
5	25.47455	9.1	12.71
6	30.80839	9.5	13.95
7	30.71683	15.7	20.62
8	37.43612	16	19.916
9	38.53549	23.6	16.603
10	34.60224	12.7	17.261
11	39.35078	10.3	11.884
12	33.01037	22.5	26.058

Table 4.3: Results of the RanBoxIter algorithm when varying the dimensionality of the search box

For the previous tests the random initialization of the box was used in order to save computation time. However, *Table 4.4* shows how the performance of the RanBox algorithm is improved when using the kernel density initialization instead of the random one.

	R	N_s	SBG
random initialization	27.857	5.4	7.586
kernel density initialization	33.132	12.4	17.293

Table 4.4: Differences in the performance of the RanBox algorithm with 8-dimensional box and $N_{tials} = 1000$ when random initialization is used and when kernel density initialization is used to search for 2% signal in 10000 events.

Now using the kernel density initialization and 12-dimensional search box we make 10 runs of the RanBox algorithm for signal fraction equal to $1.5\%, 2\%, 2.5\%$ and 3%. For each run 1000 trials are performed. In *Table 4.5* we can examine how the SBG changes as a result of the used signal fraction.

signal fraction	1.5%	2%	2.5%	3%
number of boxes with $SBG > 1$	4	6	6	6
number of boxes in the first five with $SBG > 1$	3	3	4	4
average SBG in the first five boxes	4.09	9.77	13.34	24.48

Table 4.5: Changes in SBG when varying the signal fraction. For each value of the signal fraction RanBox algorithm is executed 10 times.

The same test is done using the RanBoxIter algorithm. The results are shown in *Table 4.6*.

signal fraction	1.5%	2%	2.5%	3%
number of boxes with $SBG > 1$	6	8	8	10
number of boxes in the first five with $SBG > 1$	2	4	4	5
average SBG in the first five boxes	7.05	13.74	19.36	24.67

Table 4.6: Changes in SBG when varying the signal fraction. For each value of the signal fraction RanBoxIter algorithm is executed 10 times.

A single run of the RanBox algorithm is performed using the following parameters: kernal density initialization, 12 dimensional search box, 1000 trials, 500 signal events in total of 10000 events. The algorithm converges after 51 iterations to a box that is characterized with $N_{in} = 60$ data points when $N_{exp} = 0.04$ are expected, out of the 60 events in the box $N_s = 25$ are signal events. *Table 4.7* presents the characteristics of the five most significant boxes.

R	N_{in}	N_s	signal fraction in box	signal to background gain
57.66	60	25	5	8.333
52.9	59	53	10.6	17.966
42.81	51	34	6.8	13.333
41.86	48	21	4.2	8.75
39.41	47	31	6.2	13.19

Table 4.7: Characteristics of the five most significant boxes

We can see that the boxes contain a not negligible amount of signal events even though the caught signal fraction is small. This can be a result of the similarities between the signal and the background distributions. Applying the inverse transformation we can observe in *Figure 4.2* the original distributions of signal (in green) and background (in blue) of the 12 variables defining the best box.

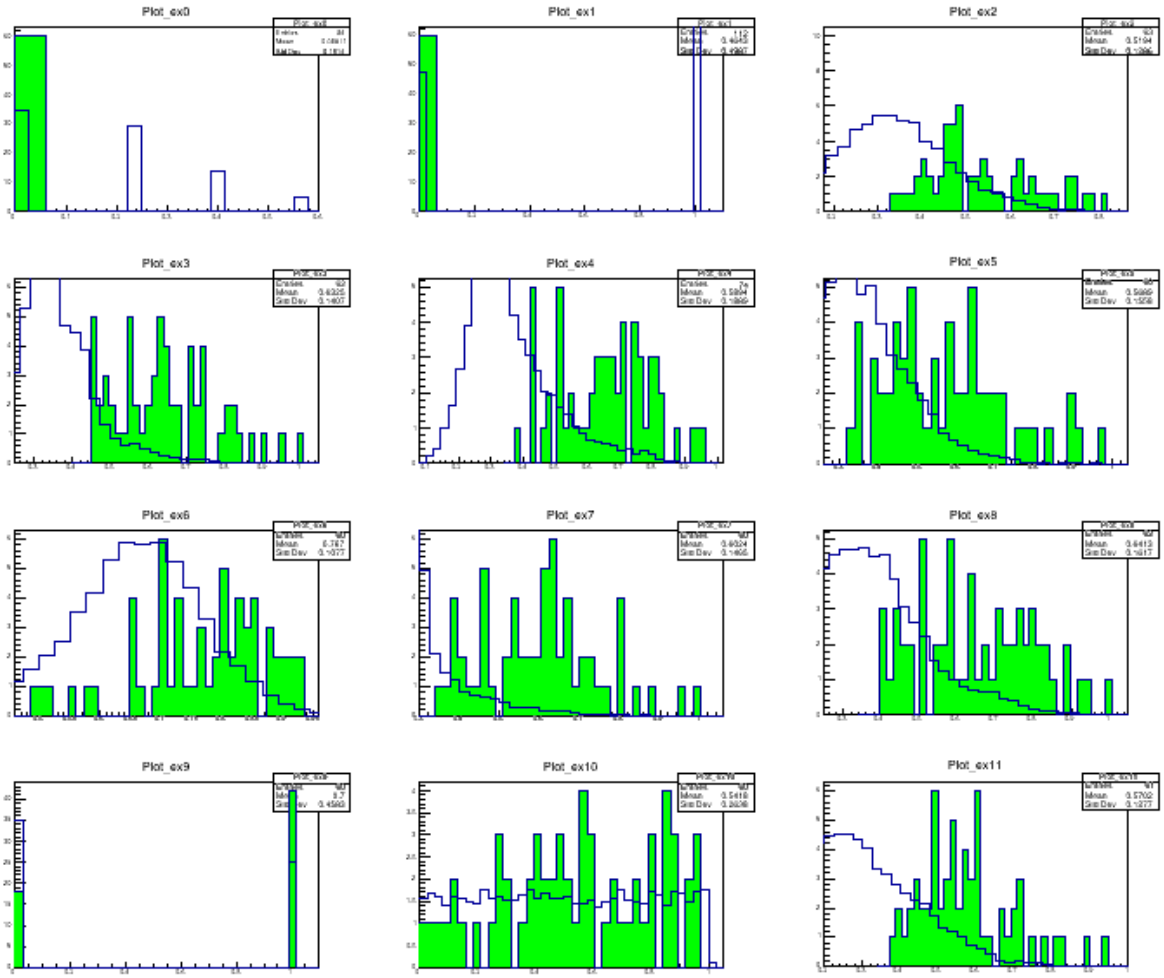

Figure 4.2: The original distribution of the background (in blue) and signal (in green), for the 12 variables that define the best box found by the RanBox algorithm in a run on 10000 events from the HEPMASS dataset with 5% signal fraction. The identified box captures 25 signal events out of a total of 60, and shows distributions in marked disagreement with those of the full data sample.

Lastly a study of power is done with the RanBoxIter algorithm considering 10000 events and a 12 dimensional search box. In this case the clustering method is used for the initialization of the box. The null hypothesis distribution (*Figure 4.3*) is obtained by using only background events and the critical regions corresponding to type I error $\alpha = 0.01, 0.05$ and 0.1 are found. The algorithm is run 10 times for each value of the signal fraction in order to calculate the power. The results are shown in *Figure 4.4*

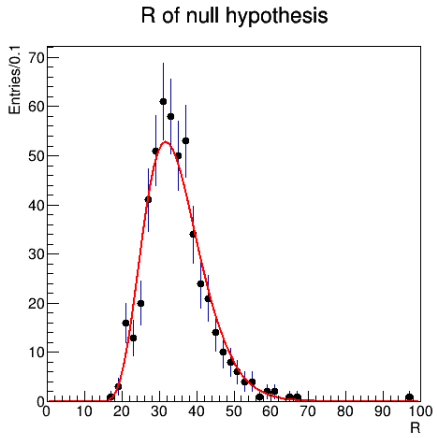

Figure 4.3: Distribution of the null hypotheses

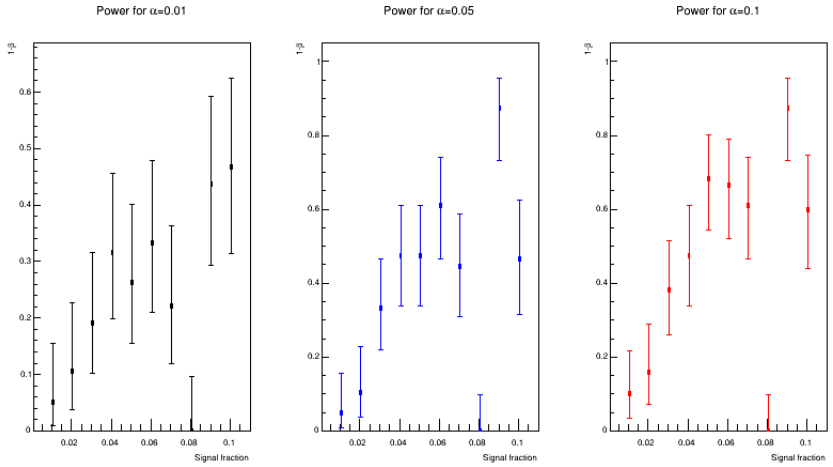

Figure 4.4: Power curves in function of the number of the signal fraction. Confidence intervals are estimated using the Clopper-Pearson method [7]

As expected the performance of the algorithm with the HEPMASS dataset can not be compared to its results when using a synthetic dataset, however we can still expect to find regions and combinations of features that are interesting to study and may lead to new discoveries.

Chapter 5

Conclusion

In this thesis, an algorithm for anomaly detection was presented. This unsupervised method can be very useful in high energy physics research as a fresh perspective in the search for new phenomena without having to focus on a predefined model.

The performed studies demonstrate good sensitivity of the algorithm to very small signal fractions or to signals present in only a few variables when using a synthetic dataset. However when applied to data produced with simulations of proton-proton collisions, where the background presents an irregular structure, we observed that the algorithm was not always able to distinguish between overdensites created by a signal and overdensities that are result of the background distribution. In any case, the algorithm remains capable of identifying even small signal contaminations effectively, and should thus be used as an exploratory tool.

An interesting possibility for improvement is to use a semi-supervised approach. The background can be modeled using known processes of the Standard model and in this way we can avoid confusing background trends for signal. Having the background observations classified, the algorithm can focus on events that remain without a label.

Bibliography

[1] B. R. Martin, G. Shaw (2017), *Particle Physics*, John Wiley & Sons Ltd., ISBN 978-0-470-03293-0

[2] Gerhard Bohm, Günter Zech (2010), *Introduction to Statistics and Data Analysis for Physicists*, Verlag Deutsches Elektronen-Synchrotron, ISBN: 978-3-935702-41-6, DOI: 10.3204/DESY-BOOK/statistics

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman (2008), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer

[4] P. Baldi, P. Sadowski, D. Whiteson (2014), *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, Nature Communications 5:4308, DOI: 10.1038/ncomms5308, arXiv:1402.4735

[5] Pierre Baldi, Kyle Cranmer, Taylor Faucett, Peter Sadowski, Daniel Whiteson (2016), *Parameterized Machine Learning for High-Energy Physics*, The European Physical Journal C, Volume 76, Issue 5, article id.235, 7 pp., DOI: 10.1140/epjc/s10052-016-4099-4, arXiv:1601.07913, URL: https://archive.ics.uci.edu/ml/datasets/HEPMASS

[6] Li T.-P., Ma Y.-Q. (1983), *Analysis methods for results in gamma-ray astronomy.* In: The Astrophysical Journal 272.March 2015, p. 317. ISSN:0004-637X, DOI: 10.1086/161295

[7] C.Clopper, E. S. Pearson (1934), *The use of confidence or fiducial limits illustrated in the case of the binomial*, Biometrika, DOI: 10.1093/biomet/26.4.404.

[8] The Atlas Collaboration (2012), *K_0 and Λ production in pp interactions at $\sqrt{s} = 0.9$ and 7 TeV measured with the ATLAS detector at the LHC*, Physical Review D 85

[9] CERN, *LHC Experiments*, URL: https://home.cern/science/experiments

[10] M. Fumanelli, *Un nuovo metodo per la rilevazione di anomalie in fisica delle particelle*, Tesi di Laurea Magistrale in Scienze Statistiche, Università degli studi di Padova, Anno Accademico 2019/2020, Relatore: Dorigo T.

[11] C. Maccani, *Anomaly detection nei dati dell'esperimento CMS*, Tesi di Laurea Triennale in Fisica, Università degli studi di Padova, Anno Accademico 2019/2020, Relatore: Dorigo T.