UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Corso di Laurea in Fisica

Tesi di Laurea

Anomaly detection nei dati dell'esperimento CMS

Relatore

Prof. Tommaso Dorigo

Laureanda

Chiara Maccani

Anno Accademico 2019/2020

Indice

Introduzione

Ogni volta che ci si approccia ad un problema di analisi dati si parte dall'implicita assunzione che il processo in osservazione segua delle determinate regole che vengono rispecchiate dai dati stessi. A partire da questi si può, dunque, formulare delle ipotesi sul modello che descrive tale fenomeno, che possono essere verificate mediante la raccolta di ulteriori campioni. Queste ipotesi, se i dati usati per formularle sono abbastanza rappresentativi, dovrebbero descrivere il comportamento *normale* del sistema.

Viene chiamata *anomaly detection* una tecnica che permette di rilevare se in un determinato campione sono presenti dei dati che differiscono in modo sostanziale da ciò che viene considerato come norma [1].

Il problema principale da affrontare consiste nel fatto che non è possibile dare una definizione univoca di cosa significhi "differire dalla norma". Nei casi più semplici in cui viene considerata una sola variabile che si distribuisce in modo normale, una misura del grado di anomalia di un dato potrebbe essere espressa dal numero di deviazioni standard con cui esso si discosta dal valor medio. Quando si considerano, invece, distribuzioni non normali e multidimensionali, tale definizione non si può più applicare e sorge, quindi, la necessità di definire una metrica che dia un'indicazione di quanto un dato sia anomalo rispetto agli altri.

Esistono vari tipi di approccio a un problema di questo tipo:

- approccio *supervisionato*: si suppone di avere a disposizione un campione di dati, il così detto training set, con ignota distribuzione di provenienza ma di cui si ha a disposizione l'etichetta di classificazione. Ad esempio, è noto se un determinato evento rappresenta un segnale o è rumore di fondo. Dato un nuovo insieme di dati è possibile allora stimare le etichette con il modello costruito precedentemente sui dati di training e calcolare le distanze a partire dalle previsioni fatte con il modello.
- approccio *non supervisionato*: in questo caso le etichette, dette *labels*, non sono note per nessun dato. Il calcolo delle distanze si basa, quindi, sul confronto tra tutti i dati dell'intero campione.

Un algoritmo non supervisionato, per essere considerato efficace, deve, quindi, riuscire a definire dinamicamente i comportamenti considerati normali, senza la necessità di utilizzare un campione di training data, adattandosi alle diverse caratteristiche di dominio, senza richiedere una sua conoscenza approfondita. Inoltre, i valori anomali devono poter essere rilevati in modo efficace anche se la distribuzione dei dati è sconosciuta.

Quando viene applicato un algoritmo di anomaly detection è necessario formulare un'ipotesi che si vuole verificare e, in riferimento a questa, i risultati che si ottengono possono essere classificati in quattro categorie:

- rivelazione corretta di un'anomalia: l'insieme di dati anomali rivelati nel campione dall'algoritmo corrispondono effettivamente a processi anomali
- rivelazione corretta dell'assenza di anomalie: l'algoritmo non rivela alcun insieme di dati anomali se non ve ne sono
- falso positivo: l'algoritmo classifica come anomalo un'insieme di dati che non lo è, e che, invece, è espressione, ad esempio, del rumore intrinseco del sistema.

- falso negativo: l'algoritmo non registra un'insieme di dati anomalo effettivamente presente. Ciò può essere dovuto al fatto che l'intensità del suo segnale non è abbastanza forte rispetto al rumore di fondo del sistema.

Risulta, quindi, necessario tentare di minimizzare sia il numero di falsi positivi che quello di falsi negativi.

Le tecniche di anomaly detection trovano applicazioni in numerosi ambiti, come ad esempio in quello della cybersecurity, in cui vengono utilizzate per individuare comportamenti inconsueti e quindi proteggere da attacchi informatici, ma sono anche ampiamente usate in campo scientifico, dato che, spesso, è stata proprio l'osservazione di dati anomali a incentivare la formulazione di nuovi modelli.

L'idea alla base dell'algoritmo *RanBoxIter* è quella di analizzare i dati provenienti dall'esperimento CMS cercando di rivelare anomalie rispetto a ciò che è previsto dal Modello Standard. L'individuazione di questi eventi anomali, che rappresentano segnali di nuova fisica, potrebbe risultare utile per cercare di costruire una teoria che estenda tale modello. L'approccio utilizzato è quello non supervisionato: non viene formulata alcuna ipotesi sul modello che descrive l'ipotetico segnale [2]. Ciò risulta essere vantaggioso poichè esistono alcune regioni estreme dello spazio delle fasi che descrive un evento osservato dai rilevatori di LHC che non sono rappresentate in modo accurato e un approccio supervisionato potrebbe non mostrare alcuna evidenza di nuova fisica [3]. L'algoritmo si basa sulla ricerca di addensamenti in uno spazio standardizzato tramite una scatola che analizza i suoi sottospazi aumentando in modo via via incrementale le dimensioni. Esso lavora in ambiente ROOT e risulta essere l'estensione di un precedente algoritmo *RanBox* in cui l'analisi veniva effettuata tramite una scatola che ispezionava sottospazi di dimensione fissata [4] .

L'algoritmo RanBoxIter

2.1 Considerazioni Preliminari

Ciascun evento osservato all'interno dei rivelatori dell'acceleratore LHC è caratterizzato da una grande quantità di parametri misurati che formano uno spazio multidimensionale complesso da analizzare. Tali parametri, come ad esempio il momento trasverso associato a una particella prodotta o la massa invariante di un particolare decadimento, dipendono sia dai processi fisici che originano le osservazioni che dalle proprietà dell'apparato di rivelazione, e spesso presentano una distribuzione di probabilità che è caratterizzata da un picco seguito da una decrescita esponenziale (a titolo esplicativo si osservi la Figura 2.1).

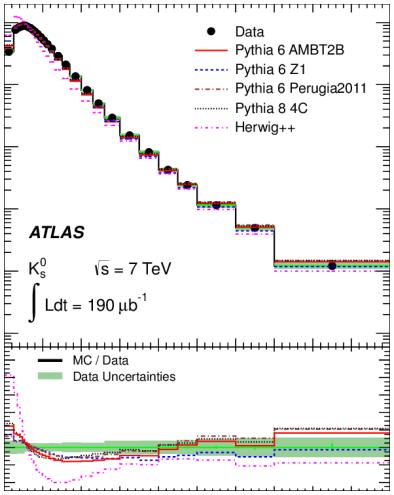

Figura 2.1: Distribuzione del momento trasverso p_T del barione Λ [5]

Queste distribuzioni associate a ciascuna variabile si dicono *distribuzioni marginali* in riferimento alla *distribuzione congiunta*, che rappresenta la distribuzione di probabilità associata al vettore che contiene tutti i parametri in considerazione.

I segnali di nuova fisica tipicamente contribuiscono ai dati osservati con un surplus di eventi in zone ristrette dello spazio dei parametri e, quindi, se lo scopo dell'algoritmo è di individuare tali addensamenti, è naturale che esso convergerà nella regione in cui ciascuna variabile presenta il suo picco di probabilità. È necessario quindi applicare una serie di operazioni preliminari di standardizzazione in modo tale che sia più facile evidenziare segnali di anomalie nelle distribuzioni associate ai parametri di interesse.

2.1.1 Principal Component Analysis

L'analisi delle componenti principali è una tecnica che permette di ridurre la dimensione di un vettore di variabili casuali correlate $X = (X_1, ..., X_p)$ a $dim(X) = k < p$ limitando la perdita di informazioni.

Siano

$$\begin{aligned} Y_1 =& a_{11}X_1 + ... + a_{1p}X_p \\ Y_2 =& a_{21}X_1 + ... + a_{2p}X_p \\ Y_p =& a_{p1}X_1 + ... + a_{pp}X_p \end{aligned}$$

delle combinazioni lineari delle variabili nel vettore X.

Le componenti principali sono quelle variabili $Y_1, ..., Y_p$ che hanno varianza massima e che sono tra loro incorrelate. Vale, infatti, il seguente teorema.

Teorema (Teorema fondamentale della PCA): Sia Σ la matrice di dimensione $p \times p$ che contiene le covarianze tra le p componenti del vettore X e siano $(\lambda_1, e_1), (\lambda_2, e_2), ...(\lambda_p, e_p)$ le coppie autovalore-autovettore ad essa associate, con $\lambda_1 \geq \lambda_2, ... \geq \lambda_3 \geq 0$. La i-esima componente principale è data da

$$Y_i = e_{i1}X_1 + e_{i2}X_2 + ... + e_{1p}X_p$$

a cui è associata una varianza pari a $var(Y_i) = \lambda_i$ e covarianza $cov(Y_i, Y_j) = 0$ per $i \neq j$.

Tra queste p nuove variabili è auspicabile che solo $k < p$ siano, in realtà, rilevanti; in tal modo la dimensionalità del problema diminuisce, semplificandone enormemente il trattamento. Per scegliere il numero di componenti principali da tenere in considerazione è necessario valutare diversi fattori.

Definizione (Varianza spiegata e varianza residua) [6]: Sia $\mathcal{U}$ una popolazione e sia $\{\mathcal{U}_m\}_{m=1,...,M}$ una partizione di $\mathcal{U}$ in m sottopopolazioni. Si considerino le variabili condizionate $Y|\mathcal{U}_m$. Vale allora la seguente relazione:

$$var(Y) = \sum_{m=1}^{M} var(Y|\mathcal{U}_m)\frac{N_m}{N} + \sum_{m=1}^{M}(\mathbf{E}(Y|\mathcal{U}_m) - \mathbf{E}(Y))^2\frac{N_m}{N}$$

con N_m numerosità campionaria della sottopopolazione e N numerosità campionaria totale. Il secondo termine rappresenta la varianza delle medie condizionate ed è indicatore di quanto le sottopopolazioni siano diverse tra loro. Esso viene definito *varianza spiegata*, in quanto è spiegata dalla disomogeneità delle sottopopolazioni. Il primo termine rappresenta, invece, la media delle varianze condizionate e viene detto *varianza residua*.

La prima dimensione Y_1 è quella che spiega la maggior parte della varianza. Tipicamente si fissa una soglia di varianza spiegata che si vuole descrivere e si sceglie il numero di componenti principali in modo tale da raggiungere tale soglia. Per prendere questa decisione si possono valutare, anche, l'entità degli autovalori e l'interpretazione delle variabili.

All'interno dell'algoritmo *RanBoxIter* tale analisi è stata implementata tramite l'apposita classe `TPrincipal` disponibile nelle librerie di ROOT. Detto `NAD`("Number Of Active Dimensions") il numero di dimensioni attive iniziali, è così possibile ridurre la quantità di variabili interessanti fissando il numero `NPCA` di componenti principali da considerare.

2.1.2 Trasformazione integrale di probabilità e copule

Per ovviare il problema dovuto al fatto che le variabili di interesse presentano intrinsecamente delle zone in cui la probabilità di essere rivelate è più alta rispetto ad altre, si può applicare la cosiddetta trasformazione integrale di probabilità.

Data una qualsiasi distribuzione di probabilità continua $f(t)$ di una variabile casuale, se ad essa viene applicata la funzione cumulativa

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

la variabile casuale y che si ottiene da questa trasformazione risulta essere distribuita uniformemente in $[0, 1]$. Infatti,

$$\begin{aligned} F_y(y) =& P(Y \leq y) = P(F_X(X) \leq y) \\ =& P(X \leq F_X^{-1}(y)) = F_X(F_X^{-1}(y)) \\ =& y \end{aligned} \tag{2.1}$$

Tale procedura viene applicata a ciascuna distribuzione marginale.

Si definisce copula d-dimensionale la mappa $C : [0, 1]^d \to [0, 1]$ che rappresenta la funzione di ripartizione congiunta delle variabili marginali dopo che sono state rese uniformi attraverso la trasformazione integrale di probabilità. Essa è caratterizzata dal fatto che il suo grafico risulta stare sempre all'interno del cubo d-dimensionale di lato 1. Il teorema di Sklar afferma come sia proprio questa funzione a contenere i vari tipi di dipendenze tra le variabili aleatorie in gioco.

Teorema (Teorema di Sklar): Ogni funzione di distribuzione multivariata $H(x_1...x_d) = Pr[X_1 \leq x_1, ..., X_d \leq x_d]$ di un vettore casuale $(X_1, X_2, ..., X_d)$ può essere espressa tramite le sue distribuzioni marginali $F_i(x_i) = Pr[X_i \leq x_i]$ ed una copula C. Infatti $H(x_1, ..., x_d) = C(F_1(x_1), ..., F_d(x_d))$.

A questo punto ciò che consideriamo come background risulta essere rappresentato da un cubo d-dimensionale di lato 1, con d pari al numero di parametri analizzati. Un ipotetico segnale, che supponiamo abbia distribuzione Gaussiana, sarà quindi ora più facile da rivelare, sottoforma di addensamento di punti sopra un fondo che ci aspettiamo essere più uniformemente distribuito. L'incremento di densità locale causato dal segnale non è visibile nelle distribuzioni marginali a causa della trasformazione integrale, ma è ancora presente nello spazio multidimensionale.

2.2 Simulazione di dati

Per validare l'algoritmo prima che esso venga eseguito di sui dati reali, è possibile verificare le sue prestazioni su datsets simulati in modo tale che il segnale e il background presentino direttamente le caratteristiche descritte nella conclusione del paragrafo precedente (modalità `mock`). In questo caso, risulta essere, quindi, noto quali tra i punti generati appartengono al fondo e quali al segnale, ossia le etichette (o labels) che in situazioni reali sarebbero ignote sono così note perchè simulate.

L'algoritmo di simulazione richiede i seguenti parametri:

- `NAD`: numero totale di dimensioni attive
- `Nmock`: numero di eventi simulati (che consideriamo essere statisticamente indipendenti)
- `FlatFrac`: frazione degli Nmock dati che appartengono al background
- `Gaussian_dims`: numero di dimensioni in cui viene iniettato un segnale Gaussiano ($\leq$ `NAD`)

Il numero di punti `NAD`-dimensionali che appartengono al background risultano quindi essere $N_b =$ `Nmock`$\cdot$ `FlatFrac` mentre quelli che appartengono al segnale $N_s =$ `Nmock` $- N_b$.

Sfruttando la classe `TRandom3` di ROOT, a ciascuna dimensione di tutti gli `Nmock` dati viene assegnato un valore scelto uniformemente nell'intervallo $[0, 1]$, dopodichè per N_s di questi eventi viene sommato un contributo scelto a partire da una distribuzione normale lungo un numero pari a `Gaussian_dims` delle sue dimensioni. Tali distribuzioni normali univariate sono a loro volta generate casualmente scegliendo il valore della loro varianza σ in modo uniforme nell'intervallo $[0.01, 0.1]$ e quello della loro media μ nell'intervallo $[3\sigma, 1 - 3\sigma]$.

2.3 Inizializzazione della scatola

L'algoritmo *RanBoxIter* (abbreviazione di Random Box Iterative) ha come scopo la rilevazione di addensamenti di punti e lo fa investigando lo spazio tramite una scatola di dimensione massima

$n =$ `Nvar` dai lati mobili. Per partire con la ricerca è necessario, quindi, inizializzare la posizione e le dimensioni di tale scatola. Tale procedura viene chiamata, anche, *seeding*. Detto v il suo volume ci aspettiamo che il numero di punti interni ad essa, se questi ultimi si distribuiscono uniformemente in tutte le loro dimensioni, sia pari a

$$N_{exp} = N_{tot} \cdot v$$

L'algoritmo implementato in questa fase consiste nella creazione di una scatola n - dimensionale caratterizzata dal fatto di avere k lati i cui estremi sono inizializzati ad un valore casuale in $[0, 1]$ ed $n-k$ lati la cui lunghezza è pari all'intero intervallo $[0, 1]$, dove k è un numero intero calcolato in modo tale che la box iniziale contenga 10 punti. Questa scelta è totalmente arbitraria ma funziona bene per le applicazioni che andremo a considerare in questo lavoro; una scelta che porti a prevedere un numero molto inferiore di punti rischia di essere troppo soggetta a fluttuazioni Poissoniane, una che consideri numerosità molto superiori tende a non essere sensibile a segnali anomali di piccola entità, quali quelli di nostro interesse.

La lunghezza media dei k lati casuali si ottiene calcolando il valore atteso della distanza tra gli estremi. Siano X e Y le variabili casuali indipendenti che rappresentano tali estremi e siano $f_X(x)$ ed $f_Y(y)$ le rispettive distribuzioni uniformi in $[0, 1]$, si ottiene:

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

Si ottiene, quindi, che in media il volume della scatola vale

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

Sotto ipotesi di uniformità il volume risulta essere, anche, pari a

$$v = \frac{N_{exp}}{N_{tot}}$$

dove imponiamo $N_{exp} = 10$. Allora

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

Ad ogni nuovo ciclo l'algoritmo di inizializzazione casuale dei lati della scatola viene ripetuto finché il suo volume è minore di una determinata soglia fissata `MaxBoxVolume`.

2.4 La statistica test Z_{PL}

2.4.1 Test di ipotesi

La verifica di un test di ipotesi è una procedura che permette di decidere se e con quale probabilità un determinato fenomeno rappresentato da un campione di dati è descritto da una particolare ipotesi.

Chiamando X la variabile casuale che descrive tale fenomeno sia $f(x;\Theta)$ la distribuzione di probabilità ad esso associata, con Θ sottoinsieme di parametri ignoti che appartengono a un insieme Ω, detto *insieme parametrico*.

Quando si effettua un test statistico si parte dalla definizione delle ipotesi. Nel nostro caso, sotto l'ipotesi nulla H_0 si assume che il campione ottenuto sia il risultato di fenomeni casuali, sotto l'ipotesi alternativa H_1 si afferma invece che i dati provengano da un segnale differente dal fondo casuale. Tali ipotesi si esprimono tramite determinati parametri

$$H_0 : \theta \in \Theta_0 \qquad\qquad H_1 : \theta \in \Theta_1$$

con $\Theta_0 \cap \Theta_1 = \emptyset$ e $\Theta_0 \cup \Theta_1 = \Omega$.

Data la numerosità campionaria n, sia $X_n = (X_1, ..X_n)$ la variabile casuale campionaria che descrive un insieme C detto *spazio campionario*. Lo scopo del test è di individuare una regione C_1 in C, detta *regione critica*, tale che, se il campione di dati ottenuto dagli esperimenti $x_n = (x_1, ..., x_n) \in C_1$ c'è evidenza per rifiutare l'ipotesi nulla H_0, mentre se $x_n \in C_0 = C - C_1$, con C_0 detta *regione di accettazione*, non c'è evidenza per rifiutarla.

A seconda della decisione presa, si possono verificare 4 diversi scenari:

- si accetta H_0 quando è vera
- si rifiuta H_0 quando è falsa
- si rifiuta H_0 quando in realtà è vera $\Rightarrow$ *errore del primo tipo*
- si accetta H_0 quando in realtà è falsa $\Rightarrow$ *errore del secondo tipo*

All'errore del primo tipo è associata la probabilità $\alpha(C_1, \theta)$, mentre a quello del secondo tipo la probabilità $\beta(C_1, \theta)$ (Figura 2.2). Viene chiamata *potenza* del test la quantità $P = 1 - \beta(C_1, \theta)$.

È necessario, quindi, individuare la regione critica in modo tale che le due probabilità di errore siano piccole, qualsiasi sia il valore di θ. Poiché non è possibile minimizzare contemporaneamente i due errori $\alpha(C_1, \theta)$ e $\beta(C_1, \theta)$, ciò che si può fare è fissare $\alpha(C_1, \theta)$ ad una soglia desiderata e contemporaneamente minimizzare $\beta(C_1, \theta)$,cosicché la potenza del test sia massima.

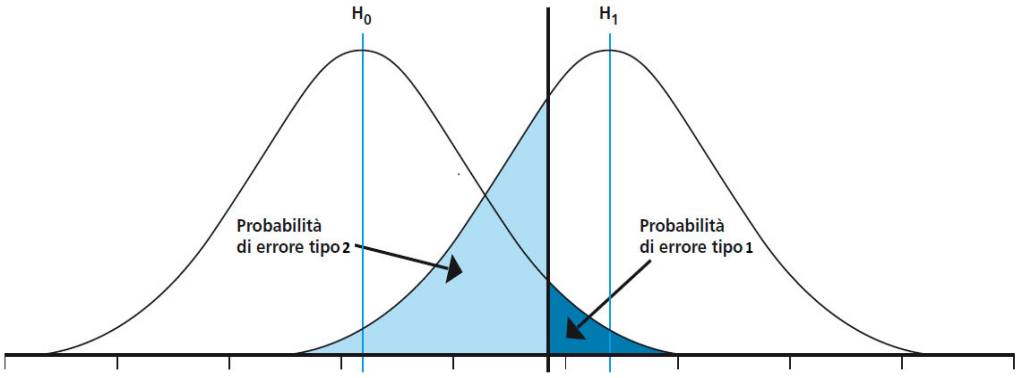

Figura 2.2: Distribuzione di probabilità associate alle ipotesi H_0 e H_1

A questo punto si cerca di individuare una funzione dei dati, chiamata statistica test, caratterizzata dal fatto che la sua distribuzione è completamente determinata sotto ipotesi nulla. Dato un campione $\mathbf{x_n} = (x_1, ..., x_n)$ la cui variabile casuale associata si distribuisce come $f(x, \theta)$, si costruisce la funzione di verosimiglianza

$$L(\mathbf{x_n}|\theta) = \prod_{i=1}^{n} f(x_i, \theta) \tag{2.2}$$

che sotto l'ipotesi nulla H_0 sarà:

$$L(\mathbf{x_n}|\theta_0) = \prod_{i=1}^{n} f(x_i, \theta_0).$$

Al variare del parametro θ si chiama stima di massima verosimiglianza quel $\hat{\theta}$ tale che

$$L(\mathbf{x_n}|\hat{\theta}) = \max_{\theta\in\Omega} L(\mathbf{x_n}, \hat{\theta}).$$

Vale il seguente teorema:

Teorema (Teorema di Wilks). Siano $\mathbf{x_n} = (x_1, x_2, ..., x_n)$ i dati osservati e $\Theta = (E, T) = (\epsilon_1, \epsilon_2, ..., \epsilon_r, \tau_1, ...\tau_S)$ i parametri ignoti. Sia H_0: $E_0 = (\epsilon_{10}, \epsilon_{20}, ...\epsilon_{r0})$ l'ipotesi nulla e H_1:$E_1 \neq E_0$ l'ipotesi alternativa. Si definisce il rapporto di verosimiglianza come

$$\lambda(\mathbf{x_n}) = \frac{L(X|E_0, \hat{T}_c)}{L(X|\hat{E}, \hat{T})} = \frac{\Pr(X|E_0, \hat{T}_c)}{Pr(X|\hat{E}, \hat{T})}$$

con $\hat{E}, \hat{T}$ stime di massima verosimiglianza dei parametri E, T e $\hat{T}_c$ stima di massima verosimiglianza condizionata quando vale l'ipotesi nulla $E = E_0$. Sotto l'ipotesi nulla la variabile $-2log(\lambda(\mathbf{x_n}))$ si distribuisce asintoticamente come una distribuzione χ^2 a con r gradi di libertà

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

Nel caso in esame viene utilizzata come statistica test la funzione $Z = \sqrt{-2log(\lambda(\mathbf{x_n}))}$ che, come conseguenza del teorema di Wilks, sotto ipotesi nulla, dipendendo da un solo parametro, si distribuisce come una normale standard, essendo la radice di una variabile che si distribuisce come un chi-quardo.

2.4.2 *On-off problem*

A questo punto è necessario calcolare la funzione Z in modo che essa descriva il fenomeno che si sta analizzando. In questo contesto si ha a che fare con un problema di tipo on-off, già studiato dagli astrofisici Li e Ma nel 1983 [7]. Ciò che si cerca di fare è di individuare la variazione associata ad un segnale rispetto ad un tasso di fondo non noto con precisione; la statistica Z dovrebbe quindi rappresentare una misura che esprime quanto un determinato eccesso di eventi in una regione sia, effettivamente, dovuto alla presenza di segnale.

Per quanto riguarda il caso in esame, ciò si traduce nel definire una misura che esprima quanto sia significativa l'osservazione di N_{on} punti all'interno della scatola su un totale di $N = N_{on} + N_{off}$ punti, con N_{off} conteggi esterni. Sia $\hat{N}_s$ il numero di eventi di segnale dentro a scatola e $\hat{N}_b$ il numero di punti interni associati al background, per cui $N_{on} = \hat{N}_s + \hat{N}_b$. Si introduce, quindi, l'ipotesi nulla H_0: "non siamo in presenza di segnale", che corrisponde all'affermare che il valore di aspettazione degli eventi associati al segnale è pari a zero $\mathbf{E}(\hat{N}_s) = 0$. L'ipotesi alternativa è rappresentata dallo statement "siamo in presenza di segnale" e si può esprimere come $\mathbf{E}(\hat{N}_s) \neq 0$. Sotto ipotesi di uniformità del background ci si aspetta che il numero di punti associati al fondo all'interno di una scatola di volume v in un volume totale pari a 1 sia

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

e che, quindi, il numero di eventi di segnale sia pari a

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

Se vale l'ipotesi nulla, non ci si aspetta segnale e il valore atteso dei punti associati al background nella scatola sarà una frazione degli N totali, tanto più grande quanto lo è il volume considerato

$$\mathbf{E}(\hat{N}_s) = 0 \qquad \Rightarrow \qquad \mathbf{E}(\hat{N}_b) = v \cdot (N_{on} + N_{off})$$

Si applica, quindi, il Teorema di Wilks ponendo $\mathbf{x_n} = (N_{on}, N_{off})$ e $\Theta = (\mathbf{E}(\hat{N_s}), \mathbf{E}(\hat{N_b}))$. Si assume che N_{on} e N_{off} si distribuiscano come due variabili casuali Poisson. Allora:

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

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

Si può quindi calcolare il rapporto di verosimiglianza come

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

Di conseguenza, se N_{off} e N_{on} sono abbastanza grandi ($\sim\geq$ 10), la variabile $\sqrt{-2log(\lambda(\mathbf{x_n}))}$ si distribuisce come una normale standard, poichè l'ipotesi nulla è caratterizzata solo dal parametro $E(\hat{N_s}) = 0$. Chiamo questa quantità Z_{PL}, in quanto essa corrisponde al valore della "*profile likelihood*".

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

2.5 Massimizzare Z_{PL}

Il criterio con cui la scatola si muove è, dunque, la massimizzazione della statistica Z_{PL}. Viene allora eseguito, per un numero di volte pari a `NGDloop` ("Number of Gradient Descent loops"), un algoritmo che modifica i lati della scatola di una quantità additiva $\lambda_{(k,i)}$, dove $k \in \{0, ..., \texttt{Nvar-1}\}$ indica la dimensione considerata ed $i \in \{1, ..., 4\}$ uno dei 4 tipi di movimenti possibili che vengono saggiati. In ciascuna dimensione, infatti, la scatola può muovere sia l'estremo inferiore che quello superiore avanti oppure indietro [1]. In base ai risultati che l'algoritmo ottiene, il passo subisce una modifica ad ogni iterazione. Esso viene inizialmente impostato a $\lambda_{(k,i)} = 0.2$.

Viene, poi, eseguito un loop su tutti i dati per contare quanti punti sono interni alla nuova scatola.

Viene calcolata, allora, il valore di Z_{PL} per ciascun tipo di movimento e se uno di questi valori risulta maggiore di quello migliore ottenuto dalle precedenti iterazioni, esso viene aggiornato. Vengono salvati anche la dimensione k e il tipo di movimento i che sto considerando e vengono aggiornati gli estremi. Infine, viene modificato il valore del passo $\lambda_{(k,i)}$ in questo modo:

- se il movimento migliore è quello per cui viene ampliato l'intervallo riducendo l'estremo inferiore allora $\lambda_{(k,i)}$ viene moltiplicata per un fattore f. Se, però, il valore di tale estremo risulta essere minore di zero, significa che è stato spostato più là rispetto ai limiti originali dello spazio standardizzato. Esso viene posto, allora, pari a zero e $\lambda_{(k,i)}$ pari a un valore ϵ molto piccolo.

[1] in caso uno degli estremi coincida con 0 o 1 il numero di movimenti viene ridotto di conseguenza

- se il movimento migliore è quello che riduce l'intervallo, allora $\lambda_{(k,i)}$ viene moltiplicata per un fattore f.
- se il movimento migliore è quello per cui viene ampliato l'intervallo aumentando l'estremo superiore allora $\lambda_{(k,i)}$ viene moltiplicata per un fattore f. Se, però, il valore di tale estremo risulta essere maggiore di 1, esso viene posto pari a 1 e $\lambda_{(k,i)}$ pari a ϵ.

Si pone $f = 1.5$. Tale scelta è arbitraria, ma empiricamente si osserva che permette di accelerare la convergenza senza compromettere le performances dell'algoritmo.

Quando, invece, l'i-esimo movimento non porta a nessun miglioramento del valore di Z_{PL} l'estremo resta quello iniziale e il passo viene ridotto di ϵ. Se $\lambda_{(k,i)} < \epsilon$ per tutti i tipi di movimento, significa che l'ampiezza dell'intervallo è stata ottimizzata in quella determinata direzione k . Quando ciò si verifica per tutte le k dimensioni significa che è stata individuata la scatola migliore.

2.6 Implementazione di RanBoxIter

L'idea alla base dell'algoritmo *RanBoxIter* è quella di andare a sondare lo spazio tramite un numero di scatole pari a `Nbest` di dimensione progressiva `Nvar`= 2, 3, 4 ... `NvarMax`.È necessario tenere in considerazione più di una scatola poiché, in base a come esse sono state casualmente inizializzate, risultano convergere in regioni dello spazio diverse che possono presentare diversi addensamenti di segnale.

Per `Nvar`=2 vengono testate tutte le possibili scatole bidimensionali verificando quali tra tutte le possibili combinazioni di due tra le `NPCA` variabili in gioco portano alla scatola a cui è associata il massimo valore di Z_{PL}. Per ottenere ciò viene, quindi, eseguito per ogni combinazione di variabili l'algoritmo di inizializzazione descritto nel paragrafo 2.3 e quello per massimizzare Z_{PL} descritto nel paragrafo 2.5. Viene quindi stilata una lista delle `Nbest` migliori coppie che viene copiata nel ciclo successivo. Viene, quindi, aggiunta una dimensione alla scatola e, a partire dalla lista proveniente dal ciclo precedente, vengono saggiate tutte le combinazioni di 3 variabili, avendo cura che esse non vengano ripetute. Ogni qualvolta viene individuata una box che presenta un valore di Z_{PL} maggiore rispetto a quelle in lista, la classifica viene aggiornata. Tale procedura viene ripetuta aumentando progressivamente le dimensioni. Il numero massimo di combinazioni che vengono analizzate è

$$N_{trials} = (\texttt{NPCA} - 1) \cdot \texttt{NPCA} + \texttt{Nbest} \cdot \sum_{i=1}^{\texttt{NPCA}-2} i = (\texttt{NPCA} - 1) \cdot \texttt{NPCA} + \texttt{Nbest} \cdot \frac{(\texttt{PCA -1 })(\texttt{ PCA -2 })}{2}$$

dove il primo termine della somma rappresenta il numero di possibili modi con cui posso scegliere due variabili e il secondo la quantità che si deve aggiungere a questo conteggio ogni volta viene aumentata la dimensione. Talvolta accade che una scatola $(n-1)$ - dimensionale presenti un valore di Z_{PL} maggiore di tutte quelle n - dimensionali, per cui non è detto che le migliori box abbiano dimensione `Nvar`.

Se l'algoritmo viene eseguito in modalità `mock` è possibile, anche, conteggiare quanti eventi che sappiamo appartenere al segnale sono presenti nella scatola e quindi calcolare la percentuale di segnale iniettato che viene catturato.

L'output finale consiste nell'elenco delle variabili e degli intervalli che caratterizzano la box migliore e la lista delle `Nbest` scatole ordinate per Z_{PL}. Vengono riportate, anche, il loro volume `Vol`, il numero totale di punti `Nin`,il numero di punti che ci aspetterebbe se ci fosse solo background `Nexp`, il numero di eventi di segnale `Ns` e la frazione di esso che la scatola contiene `BoxSF`.

Prestazioni di RanBoxIter

Vengono analizzate le prestazioni dell'algoritmo testando la sua sensibilità di analisi sui dati simulati in modalità `mock`.

3.1 Distribuzione di H_0

Se l'algoritmo viene eseguito su un campione di dati in cui non sono presenti eventi associati al segnale (`FlatFrac` = 1) esso convergerà comunque in una zona dello spazio dei parametri in cui vi è un addensamento di punti, causato dalle fluttuazione casuali del background, tale per cui il valore della statistica Z_{PL} è massimo. È necessario, quindi, costruire inizialmente le distribuzioni di Z_{PL} sotto ipotesi nulla H_0: "non è presente segnale", così da poter poi individuare il valore di soglia che corrisponde a una determinata probabilità di errore del primo tipo α. Tali distribuzioni assumono forma diversa al variare dei parametri `Nbox`, numero di scatole con cui l'algoritmo indaga lo spazio degli eventi, e `NvarMax`, numero delle loro dimensioni. Vengono effettuate simulazioni facendo variare tali parametri, in particolare vengono analizzati i casi in cui `Nbox` vale $5, 8, 10$ e `NvarMax` vale $8, 10, 12, 15$, e per ciascuna delle loro combinazioni l'algoritmo viene eseguito 500 volte. In tutte le simulazioni si fissa `NAD` pari a 30 e `NPCA` pari a 20 e il numero di dati analizzati `Nmock` pari a 10000. Viene riportato in un istogramma il valore della statistica Z_{PL} corrispondente alla scatola in cui essa è massima (a titolo esplicativo, l'istogramma relativo alla distribuzione di H_0 con parametri `Nbox` = 5 e `NvarMax` = 12 è visibile in Figura 3.1).

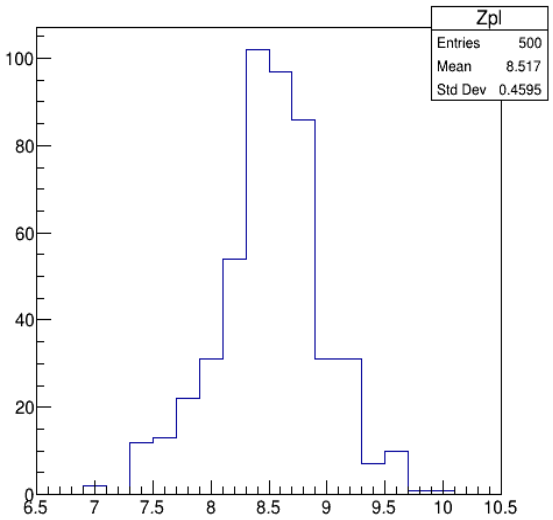

Figura 3.1: Distribuzione di Z_{PL} sotto ipotesi nulla con Nbox=5 e NvarMax=12

Tali istogrammi vengono poi normalizzati in modo tale da ottenere una distribuzione di probabilità. Si sceglie di analizzare due possibili scenari fissando la probabilità di errore del primo tipo al 5%, $\alpha_5 = 0.05$, e all'1%, $\alpha_1 = 0.01$. Essendo la numerosità del campione delle Z_{PL} pari a 500, la soglia per cui vale α_1, chiamata $s_{1\%}$, deve essere scelta in modo tale solo 5 elementi del campione risultino

essere maggiori di essa, mentre quella corrispondente a α_5, chiamata $s_{5\%}$, deve essere scelta in modo che 25 elementi siano maggiori di essa. Viene calcolata anche la media e la deviazione standard σ di ciascun campione.

I risultati delle simulazioni sono riportati in Tabella 3.1. In Figura 3.2 sono visibili i grafici che rappresentano l'andamento del valor medio del campione e di $s_{5\%}$ in funzione di `Nbox` e di `NavrMax`.

Nbox	**NvarMax**	**Media**	σ	$\mathbf{s_{1\%}}$	$\mathbf{s_{5\%}}$
5	8	7.67	0.3	8.5	8.2
5	10	8.13	0.4	9.0	8.7
5	12	8.52	0.5	9.6	9.2
5	15	8.91	0.6	10.2	9.8
8	8	7.82	0.3	8.8	8.3
8	10	8.34	0.3	9.2	8.9
8	12	8.75	0.3	9.6	9.3
8	15	9.28	0.5	10.5	10.0
10	8	7.89	0.3	8.8	8.5
10	10	8.37	0.3	10.0	9.3
10	12	8.83	0.3	9.8	9.4
10	15	9.33	0.4	10.2	10.1

Tabella 3.1

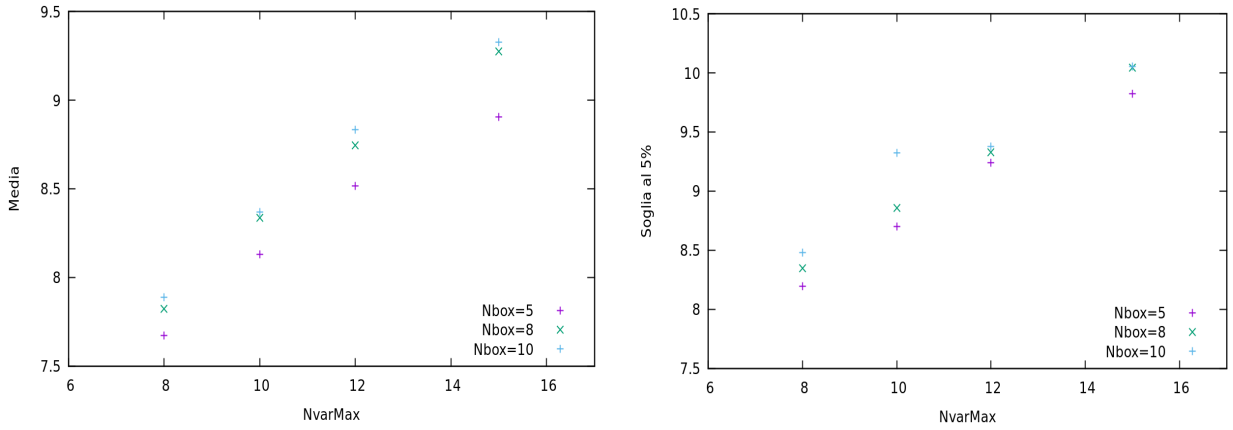

Figura 3.2: Grafici della media e di $s_{5\%}$ in funzione di NvarMax e di Nbox

Si nota come il valor medio della statistica Z_{PL} e il valore della soglia al 5% aumentino al crescere delle dimensioni della scatola, in particolare per `Nbox` pari a 5 e 8 i loro andamenti risultano essere lineare con coefficiente di correlazione lineare maggiore di 0.99. Ciò è dovuto al fatto che all'aumentare dei sottospazi considerati vengono analizzati più parametri che caratterizzano i dati e, quindi, vi sono più modi in cui essi possono distribuirsi dentro o fuori la scatola e di conseguenza può succedere che il valore di Z_{PL} sia maggiore. Anche la dispersione attorno al valor medio σ aumenta con `NvarMax`. Infine si nota che anche al crescere del numero di scatole che ispezionano lo spazio aumenta leggermente il valore medio della statistica test.

3.2 Distribuzione di H_1 e power del test

A partire da questi risultati è possibile andare a studiare la sensibilità dell'algoritmo nella rivelazione di anomalie quando viene iniettato un segnale. Vengono fissati `Nbox`= 5 e `NvarMax`= 12; sotto ipotesi nulla a questa combinazione di parametri erano state associate le soglie $s_{5\%} = 9.2$ e $s_{1\%} = 9.6$. Viene eseguito l'algoritmo prima fissando il numero delle dimensioni del segnale gaussiano `Gaussian_dims` a 15 e facendo variare la frazione di eventi di background e poi ponendo `FlatFrac` a 0.95 e facendo variare `Gaussian_dims`. Si definisce la frazione di eventi di segnale come $F_s = 1-$`FlatFrac`. Per

$\mathbf{F_s}$	**Media**	σ	$\mathbf{P_{1\%}}$	$\mathbf{P_{5\%}}$
0.005	8.7	0.4	0	0.18
0.01	8.5	0.5	0	0.10
0.015	8.9	0.8	0.16	0.26
0.02	8.9	0.9	0.16	0.26
0.025	9.6	1.4	0.36	0.54
0.03	9.8	1.4	0.40	0.49
0.035	10.6	1.9	0.66	0.76
0.04	17.7	3.8	0.98	1.00
0.045	15.6	3.3	0.98	0.98
0.05	14.2	3.4	0.94	0.98
0.055	18.8	3.8	1.00	1.00

Tabella 3.2: Gaussian_dims = 15

Gaussian_dims	**Media**	σ	$\mathbf{P_{1\%}}$	$\mathbf{P_{5\%}}$
8	8.6	0.4	0.02	0.04
10	8.8	0.8	0.16	0.34
11	9.5	1.4	0.34	0.58
13	11.1	2.8	0.72	0.72
12	10.3	1.6	0.68	0.84
15	14.2	3.4	0.94	0.98
18	28.3	3.6	1.00	1.00

Tabella 3.3: $F_s = 0.05$

ciascuna di queste condizioni vengono raccolti dei campioni con numerosità pari a 50 della statistica Z_{PL} corrispondente alla migliore scatola. Poiché l'algoritmo lavora in modalità `mock` ai dati generati dalla simulazione è associata una label ed è quindi possibile conoscere quanti tra i punti racchiusi nella scatola appartengono al segnale ($\hat{N}_s$). Viene quindi costruita la distribuzione dei valori di Z_{PL}, che, una volta normalizzata, rappresenta la distribuzione di probabilità dell'ipotesi alternativa H_1:"è presente del segnale". Viene calcolato il power $P_{5\%}$ e $P_{1\%}$ associato a questo test di ipotesi come l'integrale della distribuzione di probabilità di H_1 tra i valori di soglia $s_{5\%}$ e $s_{1\%}$ ed ∞. I risultati sono riportati nelle Tabelle 3.3 e 3.2. Le curve che rappresentano il power sono visibili in in Figura 3.3.

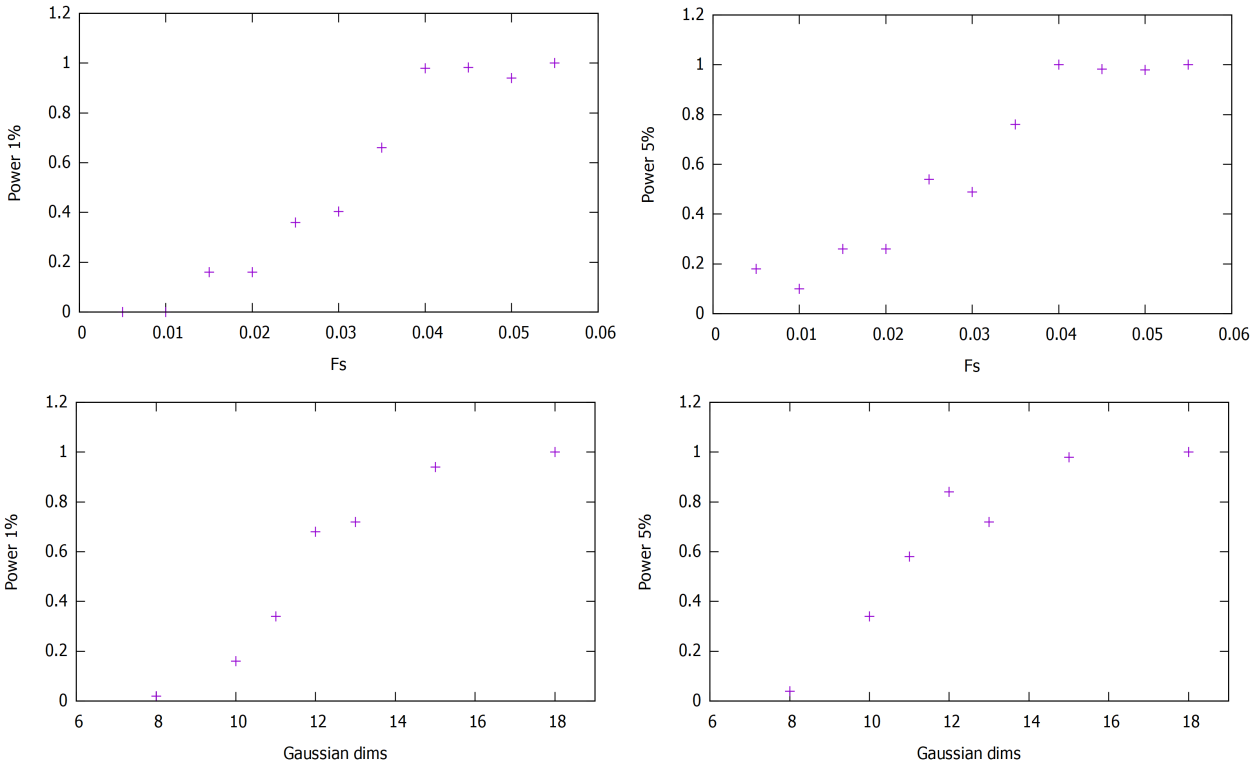

Figura 3.3: Grafici del power in funzione di Gaussian_dims (con $F_s = 0.05$) e di F_s (con Gaussian_dims = 15) , con Nbox =5 e NvarMax =12

Osservando le tabelle si evince che quando nel set di dati sono presenti dei segnali deboli, ossia costituiti da un basso numero di punti oppure presenti solo in un numero ridotto di dimensioni, l'algoritmo non riesce ad individuarli correttamente poiché i valori di Z_{PL} che si ottengono in queste condizioni sono compatibili con quelli risultanti dall'analisi delle fluttuazioni del background. In questi casi la distribuzione associata a H_1 si sovrappone totalmente o quasi totalmente con quella associata a H_0 e, di conseguenza, per tutte le volte che si ottiene un valore di Z_{PL} minore del valore di soglia

(molto spesso) non si può affermare che ci siano evidenze per rifiutare l'ipotesi nulla. Al crescere di F_s e di `Gaussian_dims` anche il valor medio della statistica cresce e la distribuzione H_1 si trasla in avanti aumentando anche la sua dispersione. Conseguentemente,il power, che rappresenta la capacità di rivelare le anomalie correttamente, aumenta.

È possibile analizzare le performance dell'algoritmo andando ad analizzare anche il numero di eventi di segnale che sono presenti nella scatola migliore. In Figura 3.4 sono mostrati gli scatter plots in cui in asse x sono riportati i valori di Z_{PL} e in asse y i valori di $\hat{N}_s$. In tutti i casi rappresentati `Nbox` vale 5, `NvarMax` è pari a 12 e `Gaussian_dims` vale 15. La prima coppia di grafici rappresenta i risultati delle simulazioni con frazione di segnale $F_S = 0.005$, la seconda coppia quelli relativi ai run con $F_S = 0.025$, mentre la terza quelli con $F_S = 0.050$.

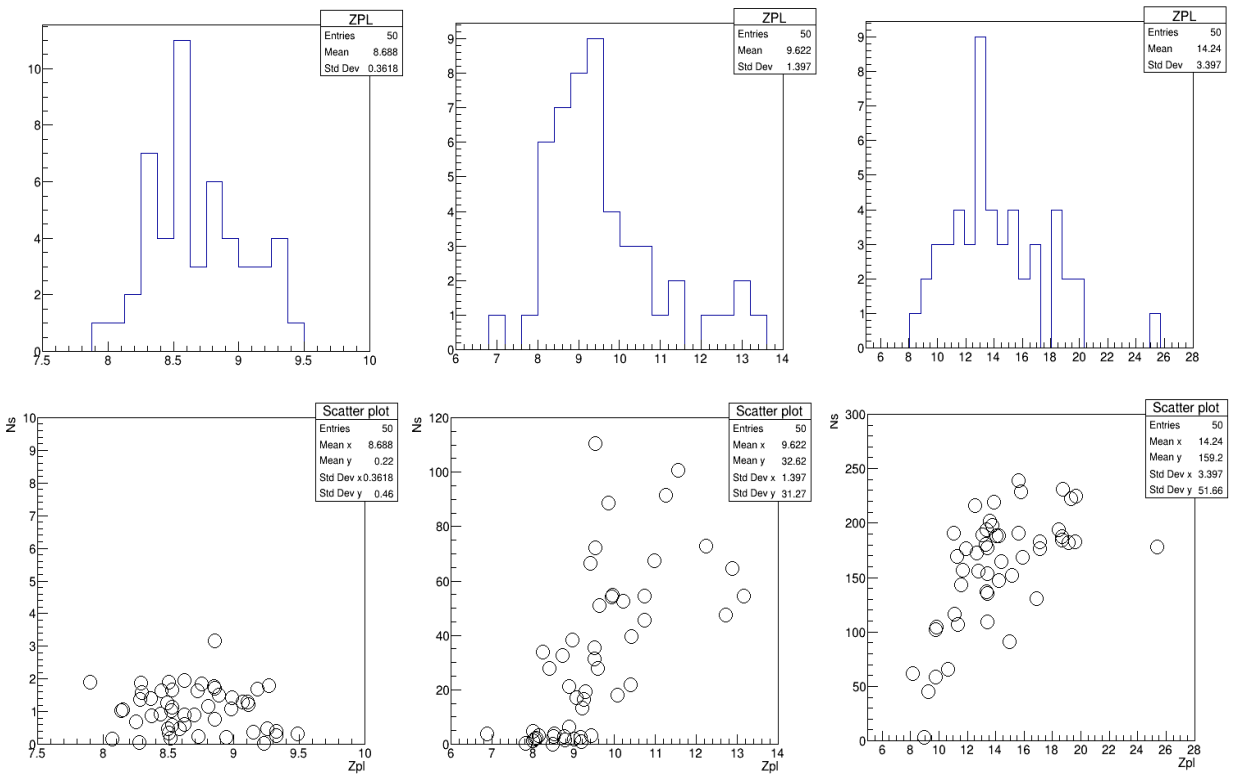

Figura 3.4: Istogrammi relativi alla distribuzione di Z_{PL} e scatter plot di Z_{PL} e $\hat{N}_s$ con $Nbox = 5$, $NvarMax = 12$, $Gaussian\_dims = 12$ e $F_s = 0.005, 0.025, 0.050$

Si osserva che quando il segnale è debole ($F_s = 0.005$) la scatola riesce a raccogliere al massimo 3 eventi sui 50 che erano stati iniettati. In questo caso si nota come la distribuzione dei valori di Z_{PL} sia quasi totalmente sovrapposta con quella relativa all'ipotesi nulla riportata in Figura 3.1. L'algoritmo non risulta, quindi, essere sensibile in queste condizioni: i power risultano essere, infatti, $P_{1\%} = 0$ e $P_{5\%} = 0.18$. Per $F_s = 0.025$ si nota che quando il valore di Z_{PL} è inferiore al valore di soglia la scatola cattura un basso numero di eventi associati al segnale, mentre ad alti valori della statistica test corrispondono alti valori di $\hat{N}_s$. Nei pressi della soglia si verificano entrambi i comportamenti. Infine, quando F_s vale 0.05 si nota come la miglior scatola riesca a catturare una grande quantità di eventi di segnale.

Applicazione ai dati di CMS

Si vuole testare la capacità dell'algoritmo di rivelare anomalie anche su un campione di dati reali. Per questo studio vengono scelti i dati raccolti dall'esperimento CMS nel 2016 con lo scopo di studiare il processo di produzione di coppie di bosoni di Higgs.

4.1 Descrizione del processo fisico

Il *bosone di Higgs* è una particella neutra con spin 0 la cui esistenza è predetta dal Modello Standard. Esso è il portatore di forza di un campo scalare chiamato *campo di Higgs* e risulta essere di particolare importanza poiché è responsabile del fenomeno chiamato *rottura spontanea di simmetria del campo elettrodebole* che conferisce massa ai bosoni di gauge $W^{\pm}$ e Z^0 [8]. Il bosone di Higgs è stato osservato per la prima volta nel 2012 dagli esperimenti CMS e ATLAS e la sua notevole massa, non direttamente prevista dallo SM, è stata misurata essere pari a $m_h \simeq 125\text{GeV}$ [9].

Questo bosone è tale da presentare la possibilità di accoppiarsi con se stesso e questa proprietà di *self-coupling* viene tipicamente studiata analizzando processi di produzione di coppie non risonanti hh. Secondo il Modello Standard la produzione hh avviene principalmente attraverso un fenomeno chiamato *fusione gluone-gluone* che presenta un loop interno di fermioni (Figura 4.1).

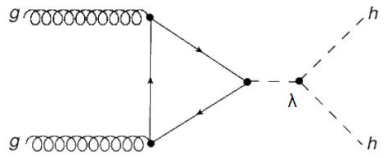

Figura 4.1: Diagramma di Feynman del processo di produzione di una coppia hh

La lagrangiana associata all'Higgs presenta un termine

$$\mathcal{L}_h = \cdots - k_\lambda \lambda_{SM} v H^3 \cdots$$

dove λ_{SM} è la costante di accoppiamento triplo predetta dal Modello Standard, v è una grandezza chiamata valore atteso del campo di Higgs nel vuoto, e k_λ una costante che esprime la deviazione rispetto λ_{SM}. Essa è un parametro libero e viene stimata a partire dai dati. Se dovesse risultare che $k_\lambda \neq 1$ allora ciò indicherebbe la presenza di nuova fisica. L'osservazione di questo fenomeno è, quindi, di grande rilevanza fisica perché permette di verificare se le predizione dello SM sono corrette, tuttavia esso risulta essere un processo estremamente raro (la sezione d'urto ad esso associata è dell'ordine $\sigma_{hh} = \text{o}(10\text{fb})$).

La ricerca di processi di produzione hh viene effettuata dall'esperimento CMS attraverso collisioni protone-protone con energia nel centro di massa $\sqrt{s} = 13\text{TeV}$ [10]. Essendo il bosone di Higgs una particella molto massiva è possibile che essa decada in una coppia di quark bottom e anti-bottom ($m_b \simeq 4.18\text{GeV}$) producendoli on shell. il rapporto di decadimento BR (*Branching Ratio*) legato a $hh \to b\bar{b}b\bar{b}$ è $BR = 33.3\%$ ed esso risulta essere il maggiore tra tutti quelli associati ai possibili modi di decadimento. Ciascuno dei quark bottom produce successivamente un jet adronico la cui presenza viene facilmente registrata dal rilevatore. Questo tipo di processi sono rappresentati in Figura 4.2.

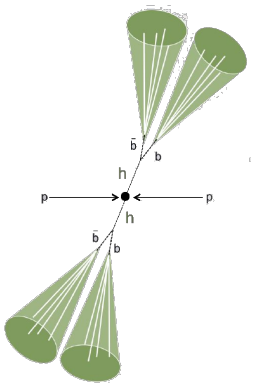

Figura 4.2: Schema dell'evento di collisione e di decadimento $pp \to hh \to b\bar{b}b\bar{b}$

Attraverso un sistema di trigger, vengono, allora, selezionati tutti gli eventi che presentano un numero maggiore o uguale a 4 di jets totali di cui almeno due originati da un quark bottom (una selezione offline richiede poi che vi siano almeno 4 jets con questa proprietà). Questi algoritmi, chiamati algoritmi di *b-tagging*, sfruttano il fatto che i quark b presentano una vita media abbastanza lunga da far sì che il vertice di decadimento che corrisponde alla produzione di adroni si presenti ad una distanza rivelabile ($c\tau \sim 450\mu m$). Inoltre, poiché il quark bottom presenta una massa molto maggiore di quella dei suoi prodotti di decadimento, questi jets risultano essere composti da una grande molteplicità di particelle che possono presentare anche un alto valore del momento trasverso. Un algoritmo che attua questo processo di identificazione è il CSV (*Combined Secondary Vertex*) che associa ad ogni jet un valore compreso tra 0 e 1, tanto più prossimo all'unita quanto il getto presenta caratteristiche tipiche di un b-jet.

Nel campione così costruito sono presenti, secondo le predizioni del Modello Standard, tre tipi di processi:

- fenomeni legati alla QCD (*Quantum Chromo Dynamics*): coppie di quark e antiquark prodotte del processo di *splitting* di gluoni (particelle con spin e massa nulla che sono i mediatori della forza forte) possono subire un fenomeno di frammentazione per cui la grande energia contenuta in $q\bar{q}$ viene convertita in due getti adronici. Si possono, quindi, verificare eventi in cui vengono prodotti, tramite processi legati alla forza forte, due coppie di quarks b, oppure eventi che presentano jets originati da quark più leggeri ma che sono stati incorrettamente taggati come bottom.

- produzione di coppie $t\bar{t}$: il quark top decade debolmente in quark bottom con branching ratio $BR = 99.8\%$ producendo un b-jet. È possibile che vengano prodotti, attraverso il fenomeno della radiazione di gluoni, un'altra coppia di b-jets, oppure, altri jets falsamente b-taggati

- produzione di coppie di bosoni di Higgs (in piccolissima frazione, finora non ancora messa in evidenza sperimentalmente).

4.2 Campione di dati e variabili cinematiche

Si è scelto di testare l'algoritmo RanBoxIter su due campioni diversi. Il primo contiene i dati ottenuti tramite simulazioni Monte Carlo di processi di produzione di coppie di top (che sono fenomeni che risultano essere ben caratterizzati dal punto di vista cinematico) ed è possibile andare a verificare il comportamento dell'algoritmo quando viene iniettato artificialmente un piccolo segnale in alcune variabili cinematiche. Il secondo campione contiene, invece, i dati reali raccolti da CMS. Si può, quindi, studiare le regioni dello spazio dei parametri in cui l'algoritmo trova addensamenti di eventi. È possibile leggere anche solo una frazione `readfrac` di questi dati, evitando problemi legati sia al grande tempo di calcolo (essenzialmente dovuta all'algoritmo di clustering) che alla memoria disponibile.

Ciascun evento contenuto in entrambi i campioni è caratterizzato dal numero di jet rivelati `njets`, il numero di jets b-taggati `ntags` ed altri parametri. Per ogni jet sono, poi, noti il valore del momento trasverso `jpt`, la pseudorapidità `jeta`, l'angolo azimutale `jphi`, il valore assegnato dall'algoritmo CSV `jcsv` e l'energia rilasciata nel calorimetro `je`. La pseudorapidità è una grandezza che è funzione dell'angolo θ compreso tra la direzione del fascio e il momento della particella prodotta ed è definita come

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

Essa tende a ∞ se $\theta \to 0$ e tende a 0 se $\theta \to 90$. Questo insieme di variabili identifica completamente tutte le caratteristiche del jet. È possibile fissare una soglia `cutcsv` sopra la quale si considera un jet b-taggato.

A partire da questi dati di input è possibile costruire un set di variabili cinematiche di alto livello, che meglio caratterizzano il campione e rendono più proficua una sua investigazione nella ricerca di processi di nuova fisica (come, ad esempio, risonanze di coppie di jets). Esse risultano essere 45 in totale (per cui `ND`= 45). Tra queste sono di particolare interesse la massa invariante totale dei 4 jets M_{1234}, la massa invariante e il momento trasverso associato alle 2 migliori combinazioni di jet che presentano il più alto valore CSV M_{12}, M_{34}, $p_{T,12}$, $p_{T,34}$, la somma dei momenti di tutti i jet h_T, il momento del primo jet $p_{T,1}$ e quello del quarto $p_{T,4}$. La massa invariante totale viene calcolata considerando il valore assoluto della differenza tra la massa invariante di tutte le permutazioni possibili di due coppie di jets che presentano i maggiori indici CSV $\Delta M_{klmn} = |M_{kl} - M_{mn}|$. Viene scelta la combinazione che presenta la minor differenza di massa. Le altre variabili cinematiche calcolate sono, ad esempio, il valore assoluto di tutte le combinazioni di differenze dei valori di pseudorapidità e dei valori dell'angolo azimutale, oppure la massa invariante e il momento associato ai jet che si propagano avanti o indietro rispetto a un asse chiamato *thrust axis*.

Tra tutti questi 45 parametri si considerano l20 variabili più interessanti, che si denominano `NAD` = 20. Per mezzo dell'analisi delle componenti principali PCA è possibile, infine, ridurre ulteriormente questo numero a `NPCA`= 15. Questa riduzione della dimensionalità dello spazio dei parametri associati agli eventi è necessaria per via di quella che in machine learning è comunemente chiamata *curse of dimensionality*: per quanto grande sia un campione di dati, gli eventi non sono in grado di popolare in maniera descrittiva uno spazio a dimensionalità troppo elevata. Ad esempio, con 1 milione di eventi e 20 dimensioni dello spazio, la divisione di ciascuna dimensione in 2 bins rende minore di uno il numero di eventi attesi in ciascun bin multidimensionale.

Quando si verifica il comportamento di RanBoxIter sul campione di $t\bar{t}$, si ipotizza per un test dell'algoritmo che se si dovesse presentare un segnale di nuova fisica, esso sarebbe individuabile nelle grandezze M_{1234}, M_{12}, M_{34}, $p_{T,12}$, $p_{T,34}$, h_T, $p_{T,1}$, $p_{T,4}$. È possibile, quindi, iniettare un segnale (la cui frazione rispetto ai dati totali è pari a $N \cdot$ `FakeFrac`) nelle dimensioni associate a questi parametri. Si assume che il segnale associato alle masse invarianti si distribuisca in modo Gaussiano, mentre quello associato ai momenti segua una distribuzione di probabilità Landau.

$$L(x) = \frac{1}{\pi}\int_0^\infty e^{-t\log t - xt}\sin(\pi t)dt$$

In particolare $M_{1234} = G(1000, 100)$, $M_{12} = M_{34} = G(300, 30)$, $p_{T,12} = p_{T,34} = L(400, 60)$, $h_T = L(1500, 200)$, $p_{T,1} = L(400, 60)$, $p_{T,4} = L(100, 20)$.

4.3 Test sul campione

Si testa RanBoxIter sul questo campione di eventi simulati $t\bar{t}$, iniettando, ad esempio, un numero di eventi associati al segnale pari al 5% di quelli letti, `FakeFrac` = 0.05. Si nota che l'algoritmo converge verso l'identificazione di una scatola di dimensioni molto ampie, il cui il volume risulta essere pari anche al 30% dello spazio totale. Questo accade perché Z_{PL} tende a identificare delle macrostrutture nei dati piuttosto che piccoli addensamenti di segnale, come quelli che sono stati iniettati. Questa statistica infatti, rappresenta quanto sia significativo un eccesso di punti nell'ipotesi che il numero di eventi attesi

si distribuisca in modo Poissoniano; tuttavia i dati reali sono caratterizzati da parametri che presentano distribuzioni molto complicate con correlazioni così complesse da non essere completamente eliminate dai processi di standardizzazione descritti nel Paragrafo 2.1. Risulta che Z_{PL} è, quindi, più sensibile a queste disuniformità che si manifestano su grande scala e che sono, in realtà, dovute a comportamenti collettivi del background non previsti dall'ipotesi Poissoniana, piuttosto che al segnale.

È necessario, quindi, apportare alcune correzioni all'algoritmo in modo da aumentarne la sensibilità. Si introduce, innanzitutto, un'altra statistica test, R, che rappresenta il rapporto tra il numero di eventi contati nella scatola e il numero di eventi interni attesi sotto ipotesi di uniformità del background più una costante di normalizzazione introdotta per evitare che l'algoritmo converga verso scatole arbitrariamente piccole:

$$R = \frac{N_{in}}{1 + N_{exp}}$$

Tale statistica risulta essere più adatta al problema fisico in esame poiché rende più significative le deviazioni che avvengono su piccola scala. Le fluttuazioni individuate da R risultano, quindi, coinvolgere tipicamente un numero di eventi minore rispetto a Z_{PL}, tuttavia essa riesce a individuare regioni in cui gli addensamenti di eventi sono più localizzati.

Poiché si è visto che non è possibile assumere che il background sia propriamente uniforme, vengono costruite delle regioni chiamate *sidebands* che circondano la scatola e presentano lo stesso volume. Se l'algoritmo riesce a individuare bene la regione che presenta un eccesso di eventi, si può assumere che il numero di eventi attesi nella scatola in assenza di segnale sia pari al numero di punti contenuto nelle sidebands.

Infine, si può migliorare la procedura di clustering introducendo un nuovo algoritmo di inizializzazione che permette di partire con la ricerca da un regione in cui sono già presenti alcuni punti vicini. Per ciascun evento i nello spazio multidimensionale viene calcolata la distanza con tutti gli altri punti e vengono identificati gli eventi per cui l'i-esimo punto risulta essere il loro primo vicino (considerando solo le dimensioni da 0 a `Nvar`/2 in cui la distanza risulta essere minore [1]). Si cerca, allora, l'evento con indice `imax` che rappresenta il punto più vicino al maggior numero di punti possibile, indicati con $j_1, \cdots, j_n$. Si identificano, infine, tutti i punti per i quali gli eventi j risultano essere i loro primi vicini e si costruisce la più piccola scatola che contiene tutti questi eventi (Figura 4.3).

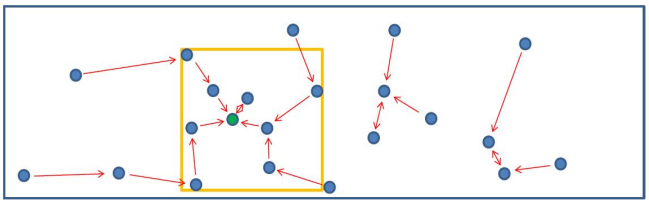

Figura 4.3: Schema che mostra la procedura di inizializzazione della scatola

Viene, quindi, eseguito RanBoxIter sul campione contenete i dati $t\bar{t}$ simulati, impostando `readfrac` in modo tale che vengano analizzati $N = 15759$ eventi. Poiché il segnale viene iniettato in 8 variabili, le dimensioni massime della scatola vengono impostate a `NvarMax` $= 8$. Viene fatta variare la frazione di segnale iniettata `FakeFrac` e il numero di scatole con cui si indaga lo spazio `Nbest` e si osservano il valore della statistica R, il numero di eventi interni `Nin`, il numero di eventi interni attesi `Nexp` e il numero di eventi interni associati al segnale `Ns` relativi alla miglior scatola. Viene riportata anche la percentuale di segnale nella scatola rispetto al totale degli eventi contenuti $\%S_{in}$ e la percentuale di segnale catturato dalla scatola rispetto al numero totale di eventi iniettati $\%S_{tot}$ (Tabella 4.1).

Si osserva che fino a `FakeFrac` $= 0.04$ e `Nbest` $= 8$ l'algoritmo riesce a individuare una scatola in cui la maggioranza dei punti contenuti appartiene al segnale (oltre il 94%). Diminuendo ulteriormente

[1] Questo accorgimento migliora le prestazioni della inizializzazione in quanto la ricerca di anomalie presuppone che siano "anomale" solo alcune delle dimensioni dello spazio. Dato che l'algoritmo non ha nozione di quale esse siano, è efficace concentrarsi su una metrica che ignori le grandi deviazioni di alcune delle direzioni dello spazio, per identificare meglio le zone ad alta densità in sottospazi appropriati.

FakeFrac	N_{fake}	Nbest	R	N_{in}	N_{exp}	N_s	$\%S_{in}$	$\%S_{tot}$
0.05	788	10	98.2	257	1.61	248	96%	32%
0.04	630	10	34.3	87	1.53	86	99%	14%
0.03	473	10	69.5	180	1.58	0	0%	0%
0.04	630	8	146.7	391	1.66	368	94%	59%
0.04	630	5	46.6	116	1.48	0	0%	0%

Tabella 4.1: $N = 15759$, NvarMax = 8

il numero di eventi di segnali iniettati, essi non risultano essere individuabili poiché il valore di R calcolato nelle regioni in cui esso è presente risulta essere simile a quello calcolato nelle regioni caratterizzate dalla presenza di addensamenti dovuti alla fluttuazione del background. Anche diminuendo `Nbest` diminuiscono le prestazioni dell'algoritmo poiché, quando vengono considerate poche scatole, si scartano alcune combinazioni di variabili potenzialmente promettenti.

È possibile infine applicare la trasformazione inversa del processo di PCA e osservare le distribuzioni associate ai parametri iniziali. In Figura 4.4 sono riportati gli istogrammi associati alle 20 variabili originarie relativi ai risultati ottenuti con i parametri `FakeFrac` = 0.05 e `Nbest` = 10. In blu è rappresentata la distribuzione a priori, mentre in verde la distribuzione del segnale individuato dalla scatola. Procedendo in orizzontale il primo istogramma è quello relativo a M_{1234}, il secondo e il terzo a $M_{1,2}$ e $M_{2,3}$, il quinto e il sesto a $p_{T,1}$ e $p_{T,4}$, il diciassettesimo e il diciottesimo a $p_{T,12}$, $p_{T,34}$ e il diciannovesimo a h_T. Si nota che, in corrispondenza di queste variabili, in cui era stato iniettato il segnale, l'algoritmo riesce a individuare degli addensamenti di punti in corrispondenza della coda delle loro distribuzioni a priori.

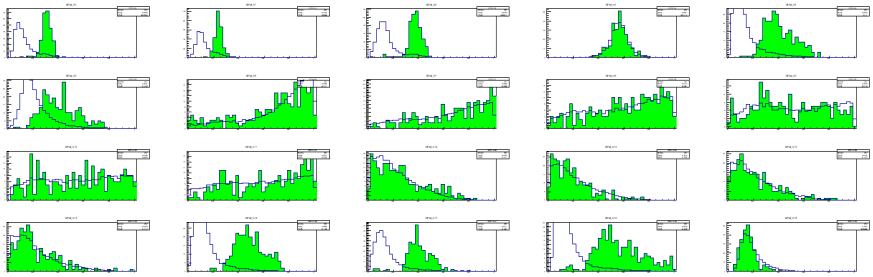

Figura 4.4: Distribuzioni associate al background (blu) e al segnale (verde) delle 20 variabili cinematiche originarie relative al test effettuato sul campione $t\bar{t}$ con parametri FakeFrac = 0.05 e Nbest = 10

In Figura 4.5 sono riportati gli istogrammi associati alle variabili originarie relativi al test con parametri `FakeFrac` = 0.03 e `Nbest` = 10. Si nota che quando l'algoritmo non riesce a rilevare il segnale è perché esso risulta presentare una distribuzione molto simile a quella del background.

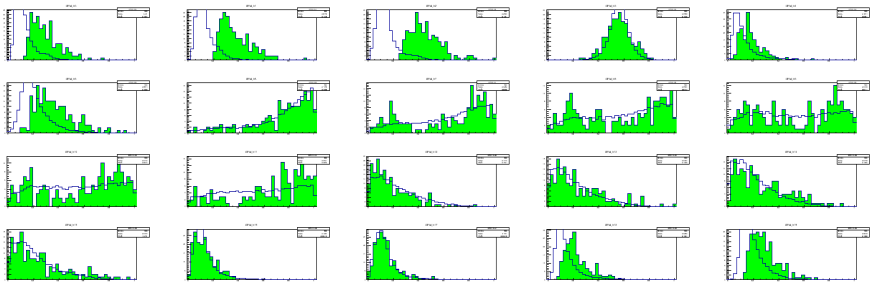

Figura 4.5: Distribuzioni associate al background (blu) e al segnale (verde) delle 20 variabili cinematiche originarie relative al test effettuato sul campione $t\bar{t}$ con parametri FakeFrac = 0.03 e Nbest = 10

Si esegue, infine, l'algoritmo sui dati reali raccolti da CMS due volte, la prima considerando Z_{PL} come statistica test, la seconda utilizzando R. Vengono considerati $N = 20000$ eventi e si fissano `NvarMax`

= 8 e `Nbest` = 10. I risultati sono riportati in Tabella 4.2.

	Test statistic	$\mathbf{N_{in}}$	$\mathbf{N_{exp}}$
Z_{PL}	149.1	14241	2266
R	84.6	103	0.21

Tabella 4.2: $N = 20000$, NvarMax = 8, Nbest = 10

Il test effettuato usando Z_{PL} come statistica test dà come risultato una scatola 5-dimensionale di volume molto grande $V = 0.71$. Osservando gli istogrammi relativi alle variabili originarie (Figura 4.6) si nota come la distribuzione associata al segnale rivelato sia completamente sovrapposta a quella associata background. Questo risulta essere una conferma della inefficienza di tale statistica nella sua applicazione ai dati reali, come discusso precedentemente.

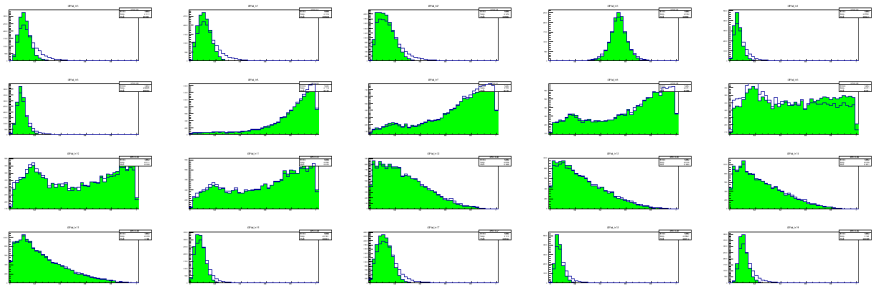

Figura 4.6: Distribuzioni associate al background (blu) e al segnale (verde) delle 20 variabili cinematiche originarie relative al test effettuato sul campione di dati reali usando Z_{PL} come statistica test

Vengono, quindi, analizzati i risultati ottenuti utilizzando R come statistica test. In Figura 4.7 sono riportate le distribuzioni marginali delle 8 variabili che compongono la miglior scatola: in blu è rappresentata la distribuzione a priori (che risulta essere uniforme poiché è rappresentata nello spazio standardizzato), in verde la distribuzione degli eventi selezionati dalla scatola e in rossa quella relativa agli eventi che risulterebbero interni alla scatola in tutte le altre 7 dimensioni, ma non in quella rappresentata. Si nota la presenza di addensamenti in alcune di queste variabili.

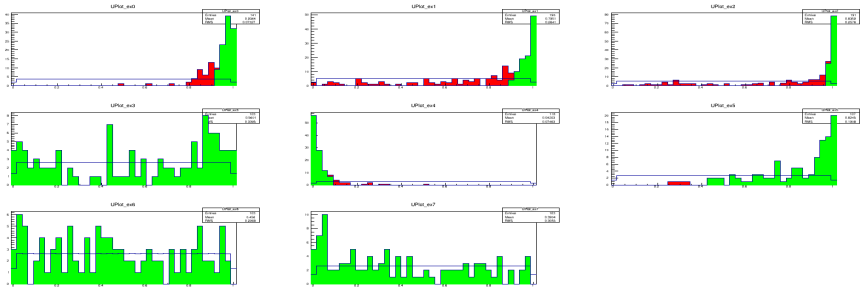

Figura 4.7: Distribuzioni associate al background (blu), al segnale (verde), e agli eventi esclusi solo nella dimensione mostrata (rosso) delle 8 variabili che compongono la miglior scatola relativa al test effettuato sul campione di dati reali usando R come statistica test

In Figura 4.8 sono riportati gli scatterplots in cui vengono rappresentate tutte le possibili combinazioni coppie di variabili tra le 8 che compongono la scatola. In blu sono rappresentati tutti gli eventi, in verde quelli selezionati dalla scatola e in rosso quelli che sono interni in tutte le dimensioni eccetto quelle rappresentate.

Infine, osservando la distribuzione delle variabili originarie (Figura 4.9), si può notare come l'algoritmo abbia effettivamente rivelato delle anomalie nella coda delle distribuzioni relative alle masse invarianti M_{1234}, M_{12} e M_{34} riportate nei primi tre istogrammi.

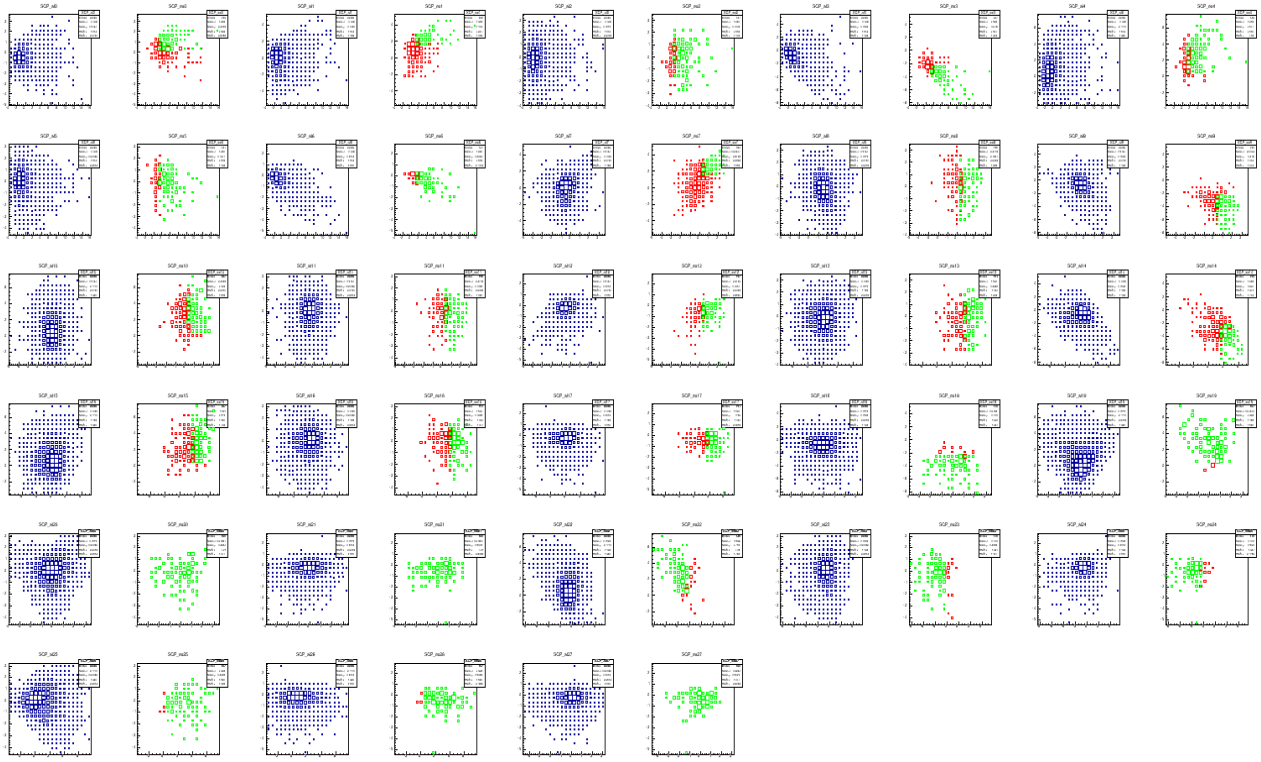

Figura 4.8: Scatterplots che rappresentano tutte le possibili coppie di variabili che compongono la miglior scatola. In blu: tutti gli eventi; in verde: eventi interni; in rosso: eventi esclusi dalla scatola solo nella dimensione mostrata. Test effettuato sul campione di dati reali usando R come statistica test.

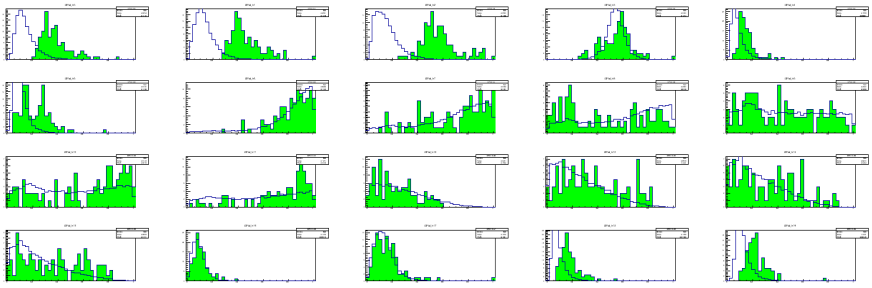

Figura 4.9: Distribuzioni associate al background (blu) e al segnale (verde) delle 20 variabili cinematiche originarie relative al test effettuato sul campione di dati reali usando R come statistica test

4.4 Conclusioni

In questa tesi è stato sviluppato e messo a punto un algoritmo di ricerca di anomalie in spazi multidimensionali, utile per la identificazione di processi di nuova fisica in collisioni di alta energia quali quelle prodotte dall'acceleratore LHC e raccolte dall'esperimento CMS. L'algoritmo si basa su una standardizzazione dei dati multidimensionali che permetta la ricerca di zone ad alta densità locale nello spazio multidimensionale senza venir influenzato dalla disuniformità dei processi di background. La ricerca viene effettuata in intervalli multidimensionali nello spazio standardizzato, massimizzando una statistica di test che metta in evidenza l'accumulo di eventi nell'intervallo. Si sono studiate le proprietà a priori di potenza e type-1 error rate dell'algoritmo in condizioni standardizzate, e si è poi applicato l'algoritmo a un caso di fisica considerando eventi raccolti da CMS con getti adronici da b-quarks.

Bibliografia

[1] Mehrotra, Kishan G., Chilukuri K. Mohan e HuaMing Huang (2017), *Anomaly Detection Principles and Algorithms* 1st. Springer Publishing Company,Incorporated. ISBN: 978-3319675244

[2] De Simone A., Jacques T. (2019), *Guiding new physics searches with unsupervised learning.* In: European Physical Journal C 79.4. ISSN: 14346052, DOI: 10.1140/epjc/s10052-019-6787-3, arXiv: 1807.06038.

[3] Collins J. H., Howe K., Nachman B. (2018), *Anomaly Detection for Resonant New Physics with Machine Learning.* In: Physical Review Letters, ISSN: 10797114, DOI: 10.1103/PhysRevLett.121.241803, arXiv: 1805.02664.

[4] Fumanelli M., *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.

[5] 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.* In: Physical Review D 85

[6] Pace L., Salvan A. (2008), *Introduzione alla Statistica, vol. 1 Statistica Descrittiva*, CEDAM, ISBN: 8813199392

[7] Li, T.-P. e Y.-Q. Ma (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.

[8] MartinB. R., Shaw G. (2017), *Particle Physics*, John Wiley & Sons, Ltd. ISBN: 978-1-118-91190-7

[9] The CMS Collaboration (2012), *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC.* In Physics Letters b, DOI:10.1016/j.physletb.2012.08.021, arXiv:1207.7235

[10] The CMS Collaboration (2019), *Search for nonresonant Higgs boson pair production in the $b\bar{b}b\bar{b}$ final state at $\sqrt{s} = 13\,TeV$* In: Journal of High Energy Physics, DOI:10.1007/JHEP04(2019)112, arXiv:1810.11854