Università degli Studi di Padova

Dipartimento di Fisica e Astronomia "Galileo Galilei"

Tesi di Laurea in Fisica

Studio del decadimento di coppie di bosoni di Higgs con tecniche multivariate

Relatore:
Prof. Tommaso Dorigo

Laureando: Nicola Riolfi
Matricola: 599153

Anno Accademico 2014/2015

Abstract

Questa tesi affronta il problema della corretta ricostruzione dei decadimenti $H \to b\bar{b}$ di coppie di bosoni di Higgs in stati finali comprendenti 4 jets da b quark.

Utilizzando una simulazione di MC del processo di segnale di questo decadimento, si considererà dapprima la sola identificazione di b quarks nei jets adronici, derivando una gerarchia di possibili algoritmi di selezione basati su b-tags e, in un secondo momento, anche sulle coordinate angolari tra i jets.

Lo studio delle variabili discriminanti porterà poi alla costruzione e ottimizzazione di algoritmi multivariati in grado di incorporare informazioni cinematiche nella scelta dei 4 jets originati dal decadimento e del loro giusto accoppiamento agli originari bosoni di Higgs. Sarà dimostrato un metodo di applicazione di algoritmi MVA (ed in particolare BDT) solitamente non praticato per la ricostruzione del segnale studiato.

Indice

Capitolo 1

Introduzione

1.1 Quadro teorico

1.1.1 Il Modello Standard

Il *Modello Standard* (MS) è, al momento, la teoria che meglio descrive tre delle quattro forze fondamentali e tutte le particelle elementari note. Sviluppato nel corso seconda metà del XX secolo, ha raggiunto la formulazione corrente nei primi anni '70 ad opera di S. Glashow, S. Weinberg e A. Salam nella forma di una teoria quantistica di campo che descrive le interazioni forte, elettromagnetica e debole, le ultime due delle quali sono ulteriormente unificate nella forza *elettrodebole*. Tutte le interazioni, ad eccezione di quella gravitazionale, sono quindi descritte da una teoria quantistica.

Il MS raggruppa le particelle elementari in due categorie principali: i *fermioni*, particelle di spin semintero che obbediscono alla statistica di *Fermi-Dirac*, e i *bosoni*, particelle con spin intero descritte dalla statistica di *Bose-Einstein*. Tra questi ultimi rientrano i bosoni di gauge, che sono le particelle mediatrici delle interazioni fondamentali.

I fermioni costituiscono la materia ordinaria del mondo macroscopico e, in base alle interazioni a cui sono soggetti, sono classificati in *quark* e *leptoni*. Ogni fermione ha una corrispondente antiparticella.

Un ruolo unico nel MS è svolto dal *bosone di Higgs*, un bosone scalare la cui esistenza è stata teorizzata nel 1964 e dovrebbe spiegare il meccanismo per il quale le particelle hanno massa. Il 4 luglio 2012 al CERN gli esperimenti ATLAS e CMS hanno annunciato la scoperta di un bosone con massa 125 GeV/c^2, che è stato riconosciuto come il bosone di Higgs [1][2].

1.1.2 Il decadimento $H \to b\bar{b}$

Il lavoro di questa tesi è stato svolto all'interno di un gruppo il cui obiettivo principale è lo studio della sezione d'urto del decadimento $H \to b\bar{b}$ di coppie di bosoni di Higgs in stati finali comprendenti 4 jets da b quark. La produzione di coppie di bosoni di Higgs, la cui osservazione richiederà l'analisi di una mole di dati non ancora raccolta da LHC, costituisce l'unico strumento a nostra disposizione per poter misurare l'auto-accoppiamento del bosone di Higgs. La sezione d'urto di questo processo, molto piccola secondo il MS

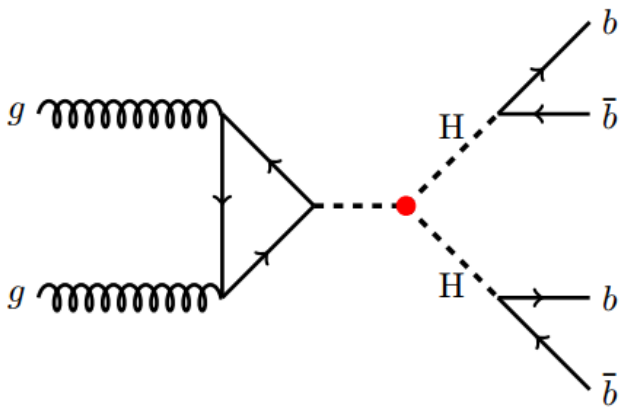

Figura 1.1: Diagramma di Feynman della produzione di doppio Higgs e relativi decadimenti $H \to b\bar{b}$

(σ = 10 fb), sarebbe notevolmente superiore nel quadro del *Minimal Supersymmetric Standard Model* (MSSM).

L'obiettivo di questa tesi consiste nello studio di algoritmi in grado di isolare il segnale di questo decadimento in campioni di segnale ad 8 TeV generati tramite Monte Carlo (per ulteriori dettagli cfr. cap. 1.3). Questo lavoro ha fatto parte di un più generale studio dei dati sperimentali raccolti dall'esperimento CMS nel Run 1 di LHC, che ha delineato le strategie di analisi che verranno applicate ai dati raccolti nel Run 2, che con la più alta energia nel centro di massa (13 TeV) e la prevista maggiore luminosità integrata che verrà raccolta nei prossimi anni offre la speranza di poter isolare un segnale di produzione di coppie di bosoni di Higgs.

1.2 L'apparato sperimentale

1.2.1 Il Large Hadron Collider

Il *Large Hadron Collider* (LHC) è il più grande acceleratore di particelle al mondo. Avviato per la prima volta nel 2008, è l'ultima aggiunta al complesso di acceleratori del CERN, situato al confine tra Francia e Svizzera vicino a Ginevra.

LHC è un acceleratore e collider di protoni che misura 27 km di circonferenza, posizionato ad una profondità media di 100 metri sotto la superficie.

All'interno dell'acceleratore, due fasci di protoni circolano a velocità relativistiche in due tubi distinti e in verso opposto. I fasci sono guidati lungo una traiettoria curvilinea da un forte campo magnetico, mantenuto da elettromagneti superconduttori tra cui 1232 dipoli che mantengono la curvatura dell'orbita dei protoni e 392 quadrupoli che focheggiano i fasci. I magneti sono mantenuti ad una temperatura stabile di circa 1.9 K da un impianto criogenico ad elio liquido.

I fasci sono accelerati da 8 cavità a radiofrequenza lungo l'anello dell'acceleratore. Nel progetto originale ogni fascio di protoni avrebbe dovuto avere già dal 2009 un'energia di 7 TeV, per un totale di 14 TeV nel sistema di riferimento del centro di massa. A causa di un guasto, però, i tempi si sono notevolmente dilatati ed è stata raggiunta un'energia nel centro di massa $\sqrt{s} = 7$ TeV nel 2011 e 8 TeV nel 2012, con le prime collisioni a $\sqrt{s} = 13$ TeV nel maggio 2015.

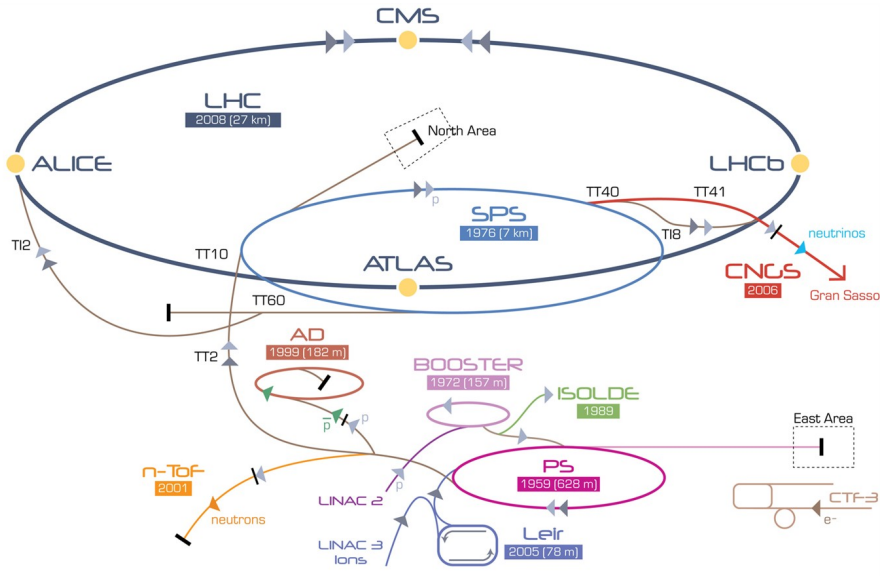

Figura 1.2: Schema del complesso di acceleratori del CERN e dei quattro esperimenti principali: ALICE, ATLAS, CMS e LHCb

Questo ordine di grandezza di energia per le collisioni tra i fasci è necessario per poter ottenere conferme al Modello Standard (MS), quali la scoperta del bosone di Higgs, e per poter osservare eventuali processi non previsti dal MS, verificando così teorie alternative quali la supersimmetria (SUSY). La necessità di mantenere fasci a queste energie in un

collider ha portato alla scelta di passare dalle collisioni elettrone-positrone del LEP a collisioni protone-protone. L'energia emessa da una particella lungo la traiettoria curva nel collider, detta radiazione di sinctrotrone, dipende infatti da m^{-4}, con m la massa della particella.

L'accelerazione dei protoni avviene in diversi stadi, illustrati in Fig. 1.2. I protoni provenienti dalla sorgente vengono iniettati nell'acceleratore lineare *Linac2* e accelerati fino a 50 MeV; passano quindi nel *Proton Synchrotron Booster* (PSB) in cui arrivano all'energia di 1.4 GeV; raggiungono 25 GeV nel *Proton Synchrotron* (PS) e 450 GeV nel *Super Proton Synchrotron* (SPS) prima di essere finalmente iniettati in uno degli anelli di LHC. Prima che LHC acceleri i protoni alla loro energia finale il processo è ripetuto 24 volte, 12 per ognuno dei due anelli.

Una volta raggiunta l'energia finale, i fasci vengono fatti collidere in quattro punti dell'anello, in corrispondenza dei quattro rivelatori principali: *A Toroidal LHC ApparatuS* (ATLAS), *Compact Muon Solenoid* (CMS), *A Large Ion Collider Experiment* (ALICE) e *Large Hadron Collider beauty* (LHCb). ATLAS e CMS sono rivelatori *general purpose* ideati per studiare estensivamente il Modello Standard, compreso il bosone di Higgs. LHCb è specializzato nella ricerca di interazioni del quark bottom (beauty), mentre ALICE è dedicato allo studio di collisioni tra ioni pesanti.

1.2.2 Sistema di coordinate

Il sistema di coordinate adottato dal CMS (Fig.1.3) ha l'origine degli assi centrata al punto nominale di collisione all'interno dell'esperimento, l'asse y che punta in verticale verso l'alto e l'asse x che punta radialmente verso il centro di LHC. L'asse z punta quindi lungo la direzione del fascio. L'angolo azimutale ϕ è misurato a partire dall'asse x nel piano xy. L'angolo polare θ è misurato a partire dall'asse z. La *pseudorapidità* η è definita come

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

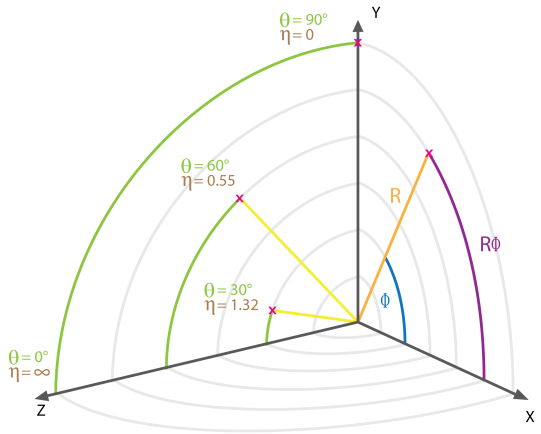

Figura 1.3: Sistema di coordinate usato per descrivere un jet

L'impulso e l'energia trasverse alla direzione del fascio, dette p_T e E_T rispettivamente, sono calcolate dalle componenti x e y [3].

1.2.3 Il Compact Muon Solenoid

Il *Compact Muon Solenoid* (CMS) è uno dei due rivelatori *general purpose* di LHC. Ha un ambito di ricerca molto ampio, che comprende verifiche del Modello Standard ad alte energie (bosone di Higgs incluso) e lo studio di modelli alternativi, ma si propone di esplorare tutta la fenomenologia delle collisioni protone-protone.

La sezione complessiva del rivelatore CMS è mostrata in forma schematica in Fig. 1.4. Il rivelatore complessivo misura 21.6 m di lunghezza e 14.6 m di diametro, per un peso totale di 12500 tonnellate, ed è diviso in due zone principali: il *barrel*, la regione centrale cilindrica, e gli *endcaps*, le due estremità laterali che chiudono il rivelatore [3].

Il rivelatore contiene un solenoide superconduttore lungo 13 m e con 5.9 m di diametro, che è in grado di generare un campo magnetico di 4 T. Un campo magnetico di questa intensità è necessario per dare alla traiettoria delle particelle una curvatura sufficiente da garantire una buona risoluzione in impulso nel tracker nonostante le dimensioni limitate.

All'interno del solenoide si trovano tracker e calorimetri. Nella regione più vicina al vertice di interazione, dove il flusso di particelle è estremamente elevato, sono utilizzati 3 strati di rivelatori a pixel di silicio per gli elevati requisiti di granularità e precisione necessari al fine di individuare i vertici secondari. Il resto del tracker consiste in 10 strati di rivelatori a microstrip di silicio, di dimensione crescente al crescere della distanza dal punto di interazione vista la diminuzione del flusso delle particelle e dei requisiti di granularità correlati.

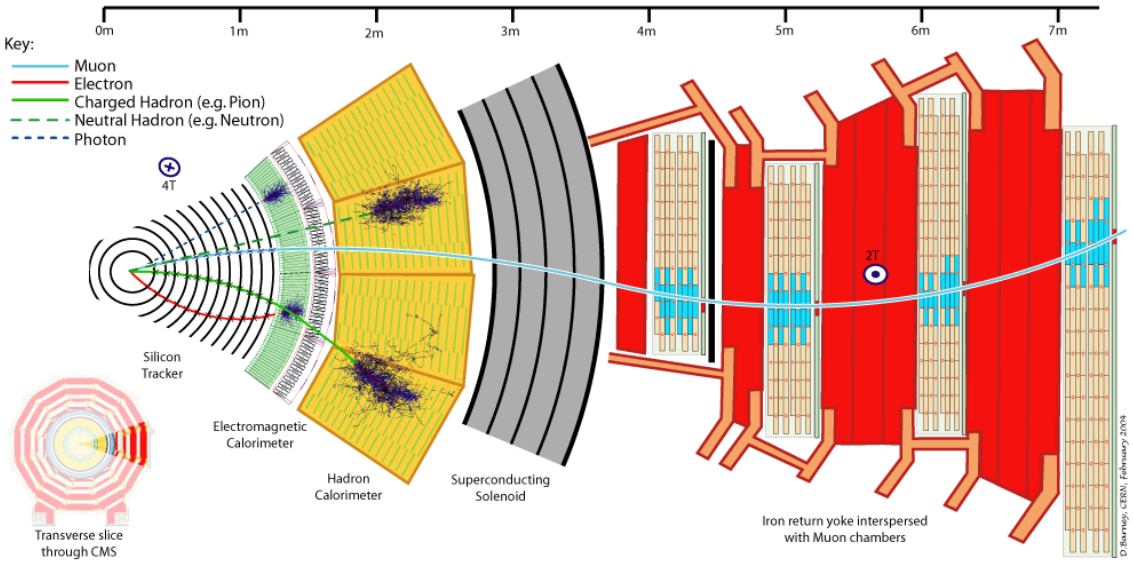

Figura 1.4: Rappresentazione schematica della sezione trasversale del rivelatore CMS, in cui è mostrato il percorso di vari tipi di particelle.

Il calorimetro elettromagnetico (ECAL) è un calorimetro ermetico e omogeneo composto da cristalli scintillanti di tungstato di piombo ($PbWO_4$). Questi cristalli hanno lunghezza di radiazione e raggio di Moliere corti ($X_0 = 0.89$ cm e $R_M = 2.2$ cm), sono veloci (emettono l'80% della luce entro 25 ns) e resistenti alla radiazione. Purtroppo hanno basso light yield, richiedendo l'uso di rivelatori di fotoni con alto guadagno interno che possano operare in campi magnetici molto forti. Per questo scopo sono stati usati fotodiodi a valanga in silicio (APD) nel barrel e fototriodi (VPT) negli endcaps. Lo spessore di ECAL in lunghezze di radiazione è di circa $25X_0$ e la copertura in pseudorapidità arriva a $|\eta| < 3.0$.

Il calorimetro adronico (HCAL) circonda ECAL ed è un calorimetro sampling realizzato con scintillatori plastici intervallati da strati di assorbitore in ottone. La necessità di usare un calorimetro sampling piuttosto che omogeneo viene dalla dimensione degli sciami adronici, molto maggiore di quelli elettromagnetici. La luce di scintillazione è convertita da fibre di *wavelength-shifter* (WLS) inserite nello scintillatore e trasportata a fotodiodi ibridi (HPD), che possono fornire un ottimo guadagno e operare anche in campi magnetici particolarmente intensi. Lo spessore di HCAL in lunghezze di interazione varia da $7\lambda_I$ a $11\lambda_I$, a seconda di η, e la copertura in pseudorapidità nel calorimetro scintillante arriva a $|\eta| < 3.0$. Copertura fino a $|\eta| < 5.0$ è fornita da un altro calorimetro, *Hadron Forward* (HF), che usa acciaio come assorbitore e fibra di quarzo come radiatore Cherenkov. La luce emessa dalle fibre di quarzo è raccolta e rilevata da fotomoltiplicatori.

All'esterno del solenoide si trova il sistema di rivelazione dei muoni, che consiste in rivelatori a gas alternati a strati di ferro di 1.5 m di spessore, saturati dal campo magnetico di ritorno. Dovendo coprire una superficie molto ampia sotto densità di radiazione variabili, i rivelatori sono di tre diverse tipologie: *Drift Tubes* (DT), *Cathode Strip Chambers* (CSC) e *Resistive Plate Chambers* (RPC).

1.3 Panoramica sui dati a disposizione

I dati utilizzati per questa tesi costituiscono puro segnale della produzione di coppie di bosoni di Higgs con decadimento $H \rightarrow b\bar{b}$ di entrambi i bosoni. Il segnale è simulato con metodo Monte Carlo (MC) utilizzando il generatore MadGraph5_aMC [4]; la fase di adronizzazione è gestita dal MC Pythia 6. La catena di generazione fornisce l'informazione al livello di generazione degli eventi. Questi possono quindi essere utilizzati per calibrare algoritmi che siano in grado di rilevare il segnale del decadimento nell'ambito di dati sperimentali. Specificamente, i dati sono raccolti in *ntuple* di ROOT raccolte in *trees* da circa 300k eventi. Un tree è stato generato in base alla fisica del Modello Standard, mentre gli altri riguardano una varietà di modelli alternativi, che includono accoppiamenti anomali del bosone di Higgs a sé stesso o a gluoni. Tutta l'analisi che si trova in questa tesi è basata sui dati relativi al Modello Standard.

Il framework di analisi dati ROOT è stato usato per tutta l'analisi contenuta in questa tesi. Una ntupla di ROOT è una sequenza di variabili che, in questo contesto, descrivono un evento. Tra le variabili a disposizione, le seguenti sono state utilizzate:

- Jets che costituiscono l'evento simulati alla MC, ognuno descritto dalle variabili: CSV, p_T, η, ϕ, energia E. Tra questi, gli algoritmi dovranno trovare i *b-jets* corretti relativi ai prodotti di decadimento di bosoni di Higgs.

- Partoni del livello di generazione dell'evento, specificamente i 4 partoni relativi ai b-jets e i 2 bosoni di Higgs che danno origine ai partoni, inclusa l'informazione di accoppiamento dei primi ai secondi. Ognuno di questi è descritto dalle variabili p_T, η, ϕ, E.

Gli algoritmi di b-tagging (e in particolare CSV) sono descritti nel capitolo 2.

Capitolo 2

Analisi di algoritmi semplici

2.1 Algoritmo di matching al gen-level

Perché sia possibile un confronto tra algoritmi di selezione dei jets è necessario sapere quali sono i veri b-jets e quali di essi provengono dal decadimento dello stesso bosone di Higgs. Infatti, anche gli eventi di produzione di coppie di bosoni di Higgs possono presentare jets addizionali dovuti all'adronizzazione di altri partoni emessi dallo stato iniziale o finale del processo. Lavorando sul Monte Carlo, è possibile ottenere informazione sull'origine dei jets direttamente dal livello di generazione dell'evento.

Come già evidenziato, nel tree dei dati è presente l'informazione sui partoni da cui vengono generati i jets. Questa sarà riferita d'ora in avanti come "informazione al gen-level". Questa informazione può essere utilizzata per creare un algoritmo che indichi, per ogni evento, quali sono i 4 jets relativi ai due b e ai due $\bar{b}$. Ad ogni partone viene associato il jet più vicino in angolo, utilizzando come discriminante la quantità

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2}$$

Questo metodo di selezione richiede però degli accorgimenti per essere applicato a tutti gli eventi. Un b quark, adronizzando, non sempre produce un jet collimato ad esso e con energia al di sopra della soglia di selezione. Questo comportamento può portare a notevoli divergenze tra un partone e il jet generato relativo. Per questa ragione, per tagliare il rumore ed assicurare un buon compromesso tra accurarezza ed efficienza dell'algoritmo è quindi necessario fare un taglio ad un certo valore di ΔR. Per decidere il taglio è stato fatto un istogramma (Fig. 2.1) coi valori di ΔR relativi ai migliori match partoni-jets trovati per evento (quindi 4 valori di ΔR per evento). Si è deciso quindi un taglio a $\Delta R < 0.5$ per isolare un sottocampione di eventi ben ricostruiti.

Una volta scelto l'algoritmo di selezione dei jet "veri", è stato quindi possibile procedere con l'analisi di semplici algoritmi.

2.2 Confronto tra algoritmi basati su b-tag

I b-tag sono tarati per essere discriminanti particolarmente forti per discernere i jets che hanno origine dall'adronizzazione di un quark b. L'algoritmo utilizzato in questa tesi è

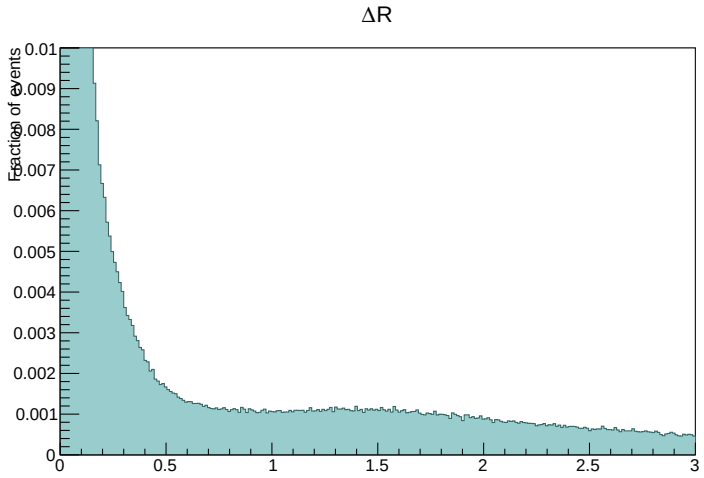

Figura 2.1: Valori di ΔR relativi ai migliori match partoni-jets per evento

denominato CSV (*Combined Secondary Vertex*), e combina l'informazione sulla presenza all'interno del jet di un vertice secondario da cui hanno origine due o più tracce di particelle cariche con altre informazioni utili a discriminare i b-jets da jets generici. Insieme a considerazioni sulla massa invariante dei dijet, possono essere utilizzati per trovare e accoppiare in modo efficace le combinazioni corrette di jets in un evento.

Come analisi preliminare si è scelto di fare un confronto tra vari algoritmi rudimentali basati su b-tag, senza utilizzare l'informazione della massa dei dijet per cercare di non inserire nel background un bias piccato intorno a $m_H = 125$ GeV/c^2. Si definiscono per comodità di utilizzo tre tagli sui valori di b-tag: *tight* per CSV> 0.898, *medium* per $0.244 <$CSV< 0.898 e *loose* per CSV< 0.244. Gli algoritmi testati si basano su questi tagli e selezionano per ogni evento 4 jets che soddisfino, rispettivamente:

- 4 b-tag tight;
- 4 b-tag medium;
- 4 b-tag loose;
- 3 b-tag tight + 1 b-tag medium;
- 3 b-tag tight + 1 b-tag loose.

Nei casi in cui più di 4 jets soddisfassero i requisiti di btag, gli algoritmi hanno scelto i jets con più alto p_T. Il comportamento di ogni algoritmo è stato analizzato con diversi tagli in p_T (20, 25, 30 GeV/c).

Va notato che questi algoritmi non si occupano dell'accoppiamento dei jets, ma solo della loro selezione. Senza usare massa invariante (o altre variabili discriminanti per i dijets) mancano le informazioni per accoppiare i jets in modo ragionevole.

Efficienza e accuratezza

Questo studio richiede in ultimo di identificare i due Higgs in ogni evento, quindi eventi in cui gli algoritmi hanno selezionato meno di 4 jets sono inutilizzabili. Per questa ragione, nell'analisi sono stati considerati due criteri di valutazione paralleli:

- *Efficienza*: la frequenza con la quale un algoritmo seleziona 4 jets, a priori dalla correttezza delle selezione. E' mostrata in percentuale, dove 100% rappresenterebbe la stessa efficienza dell'algoritmo di matching al gen-level nel trovare 4 jets.
- *Accuratezza*: la frequenza con la quale {0,1,2,3,4} dei jets selezionati corrispondono a quelli corretti, considerando solo gli eventi in cui l'algoritmo è riuscito a selezionare 4 jets. Nelle tabelle è mostrata la percentuale dei casi in cui ogni algoritmo sceglie correttamente uno specifico numero di jets.

Il diverso numero di eventi sui quali ogni tabella di risultati è calcolata è dovuto al diverso taglio in p_T, che influenza l'efficienza complessiva dell'algoritmo di matching al gen-level rispetto alla totalità degli eventi a disposizione (minore il taglio, maggiore l'efficienza).

Algoritmo	Efficienza	0	1	2	3	4
4Tight	0.36%	0.00%	0.00%	0.91%	8.60%	90.50%
4Medium	41.98%	0.00%	0.00%	0.41%	11.44%	88.15%
4Loose	90.91%	0.00%	0.04%	3.12%	35.62%	61.22%
3T1M	3.72%	0.00%	0.00%	0.44%	7.75%	91.81%
3T1L	4.86%	0.00%	0.00%	0.84%	24.17%	74.99%

Tabella 2.1: taglio in p_T = 30 GeV/c - 61037 eventi

Algoritmo	Efficienza	0	1	2	3	4
4Tight	0.34%	0.00%	0.00%	0.85%	8.48%	90.68%
4Medium	42.04%	0.00%	0.01%	0.52%	13.35%	86.11%
4Loose	92.74%	0.00%	0.05%	3.86%	39.38%	56.71%
3T1M	3.63%	0.00%	0.00%	0.55%	9.36%	90.08%
3T1L	4.76%	0.00%	0.00%	1.11%	26.44%	72.45%

Tabella 2.2: taglio in p_T = 25 GeV/c - 69661 eventi

Algoritmo	Efficienza	0	1	2	3	4
4Tight	0.33%	0.00%	0.00%	0.79%	10.24%	88.98%
4Medium	42.32%	0.00%	0.02%	0.75%	15.68%	83.55%
4Loose	95.13%	0.00%	0.08%	4.77%	43.23%	51.92%
3T1M	3.53%	0.00%	0.00%	0.84%	10.88%	88.28%
3T1L	4.65%	0.00%	0.00%	1.34%	29.10%	69.57%

Tabella 2.3: taglio in p_T = 20 GeV/c - 77257 eventi

2.3 Gerarchia di algoritmi

E' evidente dai risultati, e d'altronde prevedibile, che gli algoritmi più accurati (4 tight b-tags e 3 tight + 1 medium b-tags) sono anche i meno efficienti. Si può quindi creare una gerarchia di algoritmi, ordinati dal più accurato al meno accurato, e utilizzarla per cercare di ottenere la migliore accuratezza possibile tramite questi algoritmi di base basati su btags, mantenendo anche una buona efficienza complessiva.

Come gerarchia di accuratezza si è presa:

- 4 b-tags tight
- 3 b-tags tight + 1 medium
- 4 b-tags medium
- 3 b-tags tight + 1 loose
- 4 b-tags loose

I risultati nelle tabelle sono stati quindi ricalcolati considerando gli algoritmi come categorie esclusive: per ogni evento è stato provato ogni algoritmo in successione, fermandosi al primo che riuscisse a selezionare 4 jets. Per evitare ripetizioni, è riportata solo la tabella con taglio in p_T = 30 GeV/c.

Algoritmo	Efficienza	0	1	2	3	4
4Tight	0.36%	0.00%	0.00%	0.91%	8.60%	90.50%
3T1M	3.36%	0.00%	0.00%	0.39%	7.56%	92.05%
4Medium	38.26%	0.00%	0.00%	0.39%	11.73%	87.87%
3T1L	1.14%	0.00%	0.00%	0.43%	27.87%	71.70%
4Loose	47.79%	0.00%	0.05%	3.04%	38.11%	58.81%
Totale	90.91%	0.00%	0.03%	1.79%	25.63%	72.56%

Tabella 2.4: taglio in p_T = 30 GeV/c - 61037 eventi

In conclusione a questa analisi preliminare si osserva quindi che la richiesta di 4 b-tags tight, pur avendo accuratezza molto alta, risulta avere un'efficienza troppo bassa per essere realisticamente utilizzabile; il miglior compromesso tra efficienza e accuratezza degli eventi che vengono ricostruiti resta l'algoritmo a 4 b-tags medium.

Capitolo 3

Studio delle variabili discriminanti

3.1 Impostazione dell'analisi

Per proseguire nell'analisi è necessario identificare le variabili cinematiche discriminanti, ovvero le più sensibili alla scelta corretta dei jets. La conoscenza di queste variabili sarà fondamentale per la calibrazione degli algoritmi MVA. Per ogni evento, sono state testate le seguenti variabili:

- CSV, p_T, η dei jets singoli;
- p_T delle coppie di jets: $p_{T,H} = |\vec{p}_{T,b} + \vec{p}_{T,\bar{b}}|$;
- p_T totale dei 4 jets: $p_{T,tot} = |\vec{p}_{T,1} + \vec{p}_{T,2} + \vec{p}_{T,3} + \vec{p}_{T,4}|$;
- angoli η e ϕ tra una coppia di jets: $\Delta\eta_{b,\bar{b}}$ e $\Delta\phi_{b,\bar{b}}$;
- angoli η e ϕ tra le due coppie di jets in un evento: $\Delta\eta_{H_1,H_2}$ e $\Delta\phi_{H_1,H_2}$;
- massimo e minimo tra i 6 possibili angoli $\Delta\phi_{b,\bar{b}}$ e $\Delta\eta_{b,\bar{b}}$ per ogni scelta di 4 jets in un evento, comparato con la combinazione giusta.

Si noti che $b, \bar{b}$ sono sempre intesi relativi allo stesso H.

Per ognuna di queste variabili si sono poste a confronto le distribuzioni normalizzate di segnale e rumore, ottenute applicando un taglio in $p_T = 30$ GeV/c ai jets selezionati. Per le variabili relative a singoli jet il segnale è semplicemente rappresentato dai 4 jets selezionati dal gen-level matching e il rumore dagli altri jets. Per quanto riguarda le variabili legate a combinazioni di jets, il segnale è rappresentato dall'insieme delle combinazioni corrette, ovvero i 4 b-jets accoppiati correttamente 2 a 2 ai prodotti di decadimento dei bosoni di Higgs. Il rumore in questo caso viene da tutte le altre possibili combinazioni. Infine, per variabili relative ad una sola coppia di jets sono stati considerati i valori di entrambe le coppie nelle distribuzioni per ogni combinazione di 4 jets.

Il confronto tra le distribuzioni per le variabili analizzate è presentato in Fig. 3.1-3.3, dove il segnale è indicato in blu e il rumore in rosso.

3.2 Variabili discriminanti

Dal confronto tra le distribuzioni in Fig.3.1, 3.2 e 3.3 è evidente la differenza di potere discriminante tra le variabili scelte.

L'istogramma del p_T evidenzia la rapida crescita della distribuzione del background al diminuire dell'impulso trasverso, convalidando la scelta di $p_T = 30$ GeV/c come valore di taglio. Sia per jet singoli sia per dijet si trova una sostanziale traslazione delle distribuzioni di p_T tra segnale e background, rendendo questa variabile potenzialmente utile per complementare altre valutazioni anche se il suo valore discriminante isolato è in generale basso.

Il CSV, come ci si aspettava, ha un ottimo potere discriminante e vede una netta prevalenza del segnale a valori alti e background a valori bassi, anche se rimane un certo grado di contaminazione del background per CSV alti in accordo con l'accuratezza dell'algoritmo precedentemente provato a 4 b-tag tight, che non ha mai superato il 91% di accuratezza. Allo stesso modo si nota il comportamento inverso, ovvero la possibilità di trovare segnale anche sotto la soglia del b-tag loose.

La massa invariante delle coppie di jets associate al decadimento dei bosoni di Higgs è fortemente piccata intorno al valore previsto di $m_H = 125$ GeV/c^2, con una dispersione molto stretta rispetto alle combinazioni di background. Si nota tuttavia un picco intorno ad m_H anche per il background, che può introdurre un bias per certi tipi di richieste sulla massa delle coppie di jets.

Alcune variabili angolari rivelano comportamenti interessanti: $\Delta\phi$ tra H_1 e H_2 vede il segnale fortemente piccato intorno a π, che suggerisce un centro di massa in moto lento rispetto al rivelatore, in accordo con la massa elevata dello stato precedente al decadimento. $\Delta\eta$ tra b e $\bar{b}$, seppura non offra un buon grado di separazione tra segnale e rumore, è sostanzialmente contenuta entro $\Delta\eta < 1.5$ per il segnale e può costituire un utile taglio in un algoritmo semplice. $\Delta\phi$ tra b e $\bar{b}$ segue un simile comportamento, anche se meno definito, e può rivelarsi utile in vista della calibrazione di algoritmi multivariati.

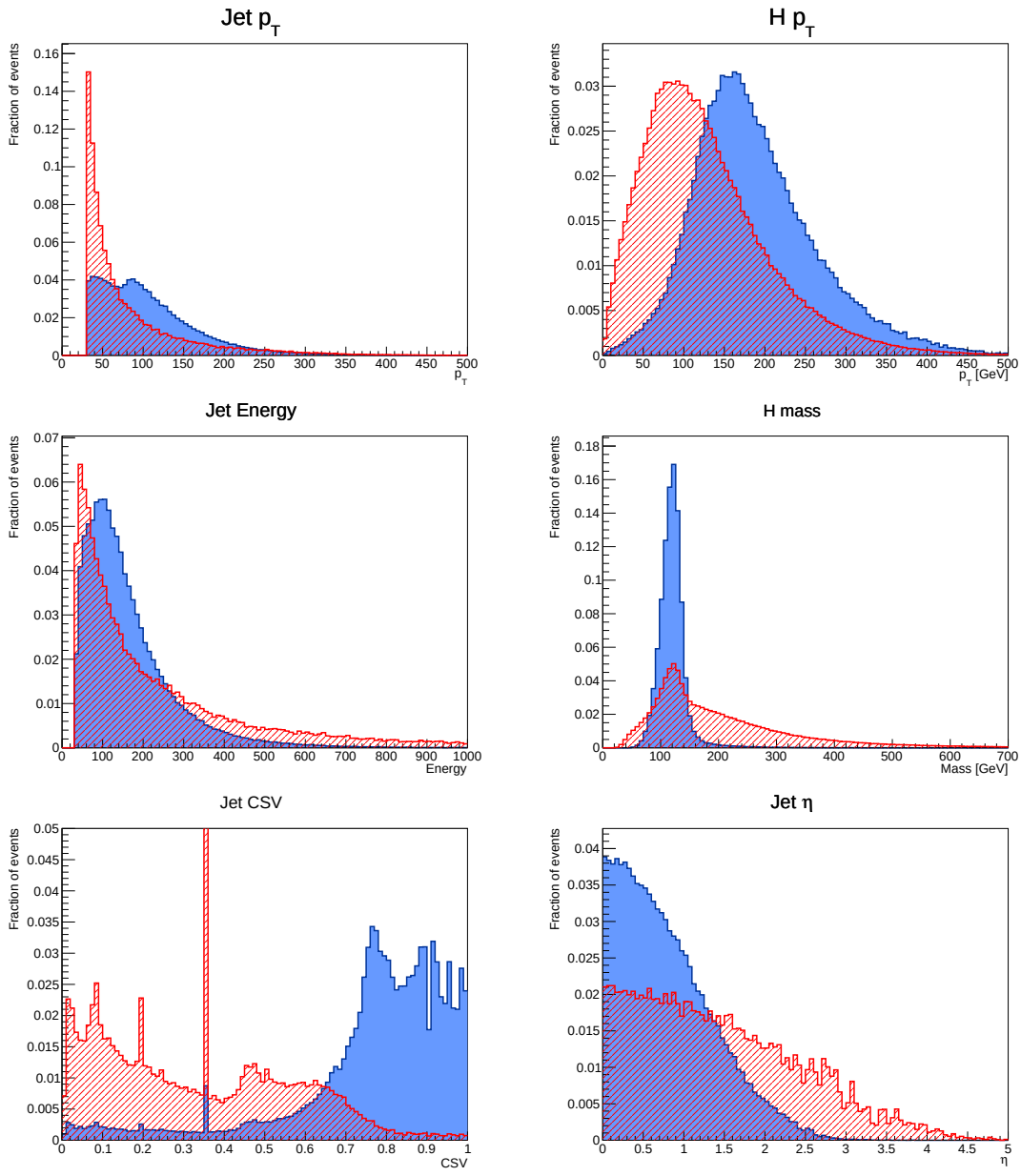

Figura 3.1: Confronto di variabili cinematiche potenzialmente discriminanti tra segnale (in blu) e background (in rosso).

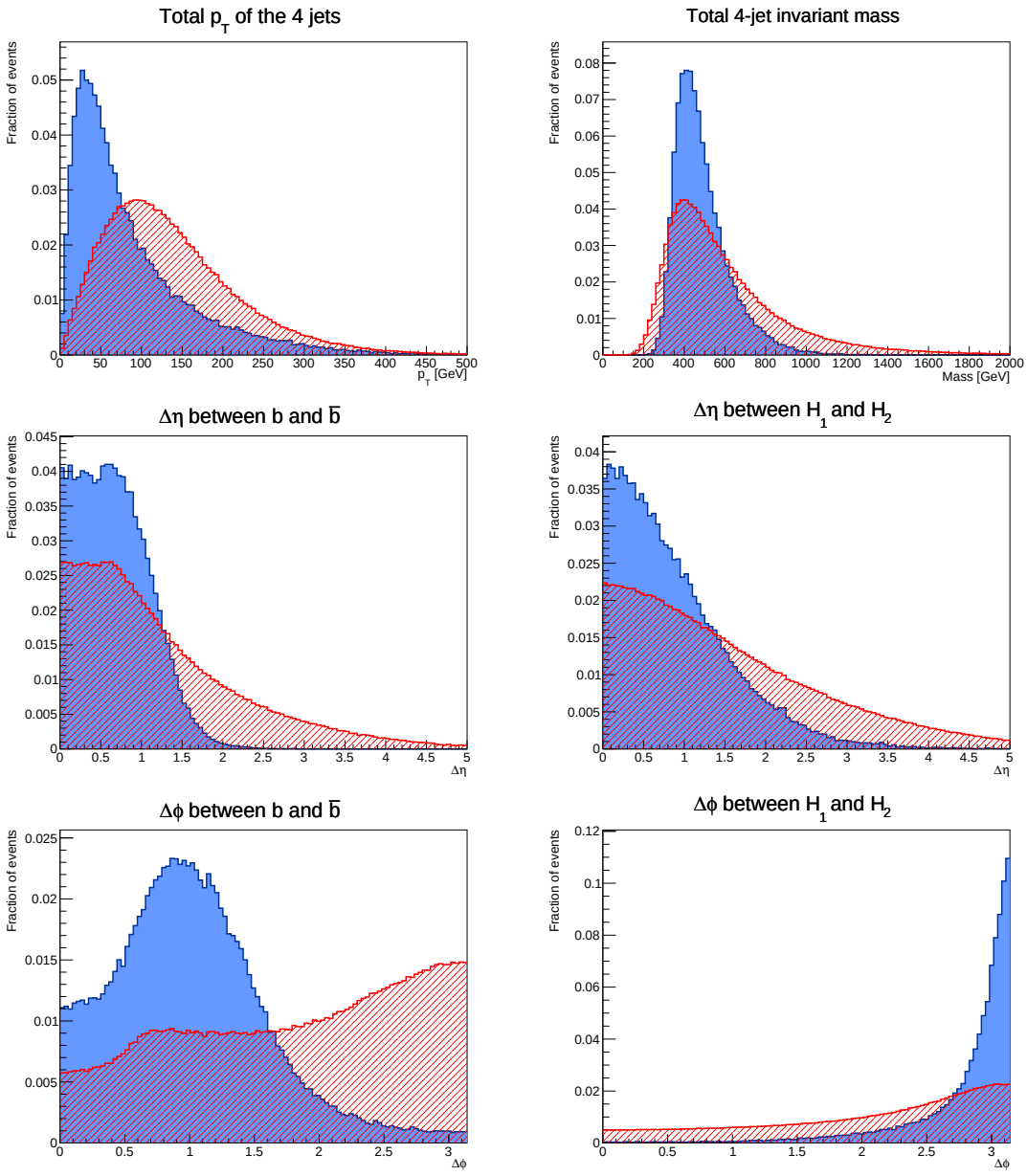

Figura 3.2: Confronto di variabili cinematiche potenzialmente discriminanti tra segnale (in blu) e background (in rosso).

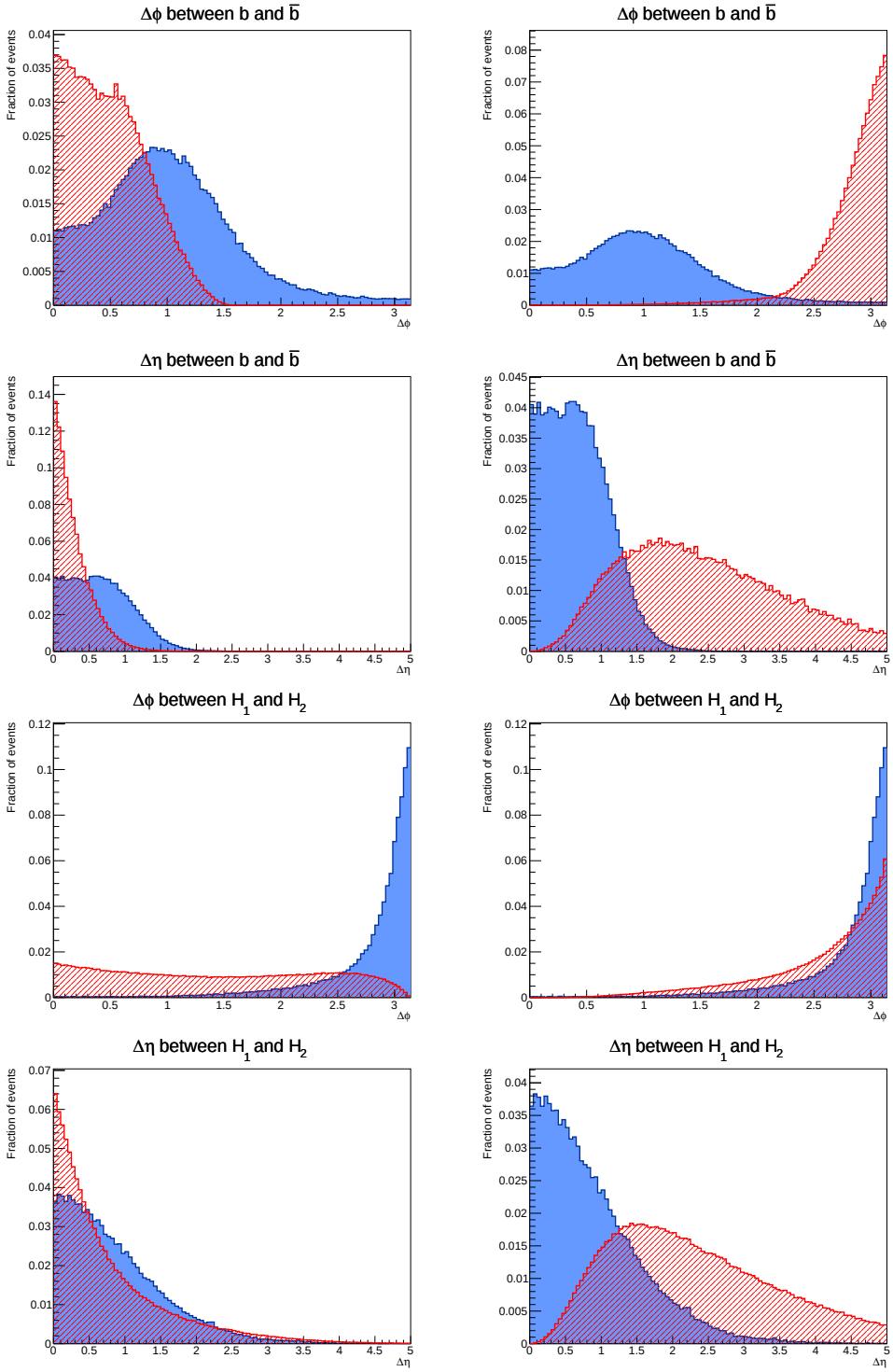

Figure 3.3: Confronto di variabili cinematiche potenzialmente discriminanti tra segnale (in blu) e background (in rosso).

3.3 Pairing tramite coordinate angolari

Vista la possibilità di usare l'angolo tra i jets come variabile discriminante dallo studio precedente, si è deciso di tentare un approccio con un algoritmo di test basato su informazioni angolari per accoppiare due a due i jets. L'accoppiamento è stato valutato sia separatamente (fungendo anche da selezione), sia in combinazione con altri metodi di selezione a b-tag già utilizzati in precedenza.

L'algoritmo provato compie le seguenti operazioni:

1. Taglia in $p_T = 30$ GeV su tutti i jets.
2. Se utilizzato in combinazione con specifiche richieste di b-tag, taglia i jets che non soddisfano tali richieste.
3. Per ogni evento con almeno 4 jets rimanenti dopo i tagli, itera su ognuna delle possibili combinazioni di 4 jets, accoppiati 2 a 2.
4. Elimina le combinazioni per le quali almeno una coppia ha $\Delta\eta_{b\bar{b}} \geq 1.5$.
5. Calcola la variabile $\Delta\phi_{H_1H_2}$ e taglia le combinazioni al di fuori dell'intervallo $|\Delta\phi_{H_1H_2} - \Pi| < 0.4$.
6. Calcola $(\Delta\phi_{H_1H_2} - \Pi)^2$per ogni combinazione. Per ogni iterazione, sceglie la combinazione col valore minimo.

I risultati sono riportati nella Tabella 3.1, in cui questo algoritmo puramente angolare è denominato "base" e viene mostrato in combinazione con algoritmi basati su b-tags.

Algoritmo	Efficienza	0	1	2	3	4
base	95.40%	0.03%	0.60%	9.01%	36.70%	53.67%
base + 4T	0.26%	0.00%	0.00%	0.63%	3.80%	95.57%
base + 4M	30.41%	0.00%	0.01%	0.43%	9.47%	90.08%
base + 4L	74.97%	0.01%	0.23%	4.87%	29.18%	65.70%

Tabella 3.1: Risultati ottenuti con l'algoritmo puramente angolare e la sua combinazione con algoritmi basati su b-tags.

Capitolo 4

Analisi MVA

4.1 Il pacchetto TMVA

Il *Toolkit for MultiVariate Analysis* (TMVA) fornisce un environment di apprendimento automatico (*machine learning*) integrato in ROOT con lo scopo di processare e valutare varie tecniche multivariate di classificazione e regressione. Il pacchetto TMVA è specificamente disegnato per soddisfare le necessità delle applicazioni di fisica alle alte energie. Il pacchetto contiene una varietà di tecniche di analisi multivariate, alcune delle quali saranno in seguito descritte con maggiore dettaglio.

TMVA consiste in una implementazione in C++ per ognuno dei metodi multivariati a disposizione e fornisce algoritmi di training, testing e performance evaluation oltre a script per visualizzare i risultati. Rende inoltre disponibile un'ampia gamma di opzioni per modificare il comportamento di ogni metodo, così da poterne ottimizzare le performance in diverse situazioni.

Una tipica analisi con il pacchetto TMVA, che sia classificazione o regressione, consiste in due fasi principali: la fase di *training*, in cui le tecniche multivariate utilizzate vengono addestrate, provate e valutate e la fase di *applicazione*, in cui le tecniche vengono applicate al problema concreto per il quale è stato fatto il training.

4.2 Fase preliminare

4.2.1 Preparazione dei dati

La selezione dei jets corretti in un evento è un problema di classificazione, sebbene non usuale e in quanto tale non immediatamente supportato. Il pacchetto TMVA si limita infatti a fornire la risposta degli algoritmi selezionati in una scala da *rumore* a *segnale* per ogni evento, mentre nel caso sotto esame serve una selezione più complessa: per ogni evento devono essere selezionati 4 jets, che devono essere accoppiati correttamente due a due. Questo non può essere fatto direttamente dal TMVA nell'implementazione attuale e ha reso necessaria una fase preliminare di preparazione dei dati prima di poterli utilizzare con algoritmi multivariati.

Per la preparazione si è deciso di procedere in questo modo:

- Si è fatto un taglio preliminare in $p_T = 30$ GeV per tutti i jets (coincidente con quello impostato per l'algoritmo di matching al passo successivo) per eliminare una buona frazione del rumore e alleggerire notevolmente la parte combinatoria seguente.

- Per ogni evento con almeno 4 jets al di sopra del taglio in p_T, sono stati trovati i jets corretti tramite l'algoritmo di matching al gen-level. Tutti gli eventi in cui non si siano trovati 4 jets sono stati eliminati, in quanto per questi eventi sarebbe stato impossibile controllare l'accuratezza degli MVA.

- Per ogni evento rimanente si sono trovate tutte le possibili combinazioni di 4 jets abbinati due a due, evitando quelle equivalenti (scambi di jet che non si traducessero in cambiamenti delle variabili cinematiche associate alla combinazione).

- Tutte le combinazioni così trovate sono state scritte nel tree di input insieme a tutte le relative variabili cinematiche rilevanti discusse nel capitolo precedente.

Le combinazioni sono così diventate gli eventi per il TMVA. Ogni combinazione può essere corretta o sbagliata, e in quanto tale un valore di risposta può essere assegnato ad ognuna. Dopo aver ottenuto la risposta del TMVA, per ogni evento del Monte Carlo si è quindi scelta la combinazione con miglior risposta per ogni algoritmo, che è stata presa a rappresentare il miglior candidato secondo quell'algoritmo.

Dal sample iniziale di circa 300k eventi l'algoritmo di matching al gen-level ha trovato 4 jets associati col criterio $\Delta R < 0.5$ ai b-jets provenienti dal decadimento di bosoni di Higgs in circa 60k eventi usando un taglio in $p_T = 30$ GeV, come mostrato nelle tabelle dell'analisi preliminare. Di questi, 20k eventi sono stati utilizzati per la fase di training e i rimanenti 40k eventi per la fase di testing. Dopo averne ricavato le combinazioni di jets, questi si sono tradotti in 590k combinazioni usate per il training e 1175k combinazioni per l'applicazione per il TMVA.

4.2.2 Correlazioni

Per fare una scelta consapevole delle variabili da utilizzare nel training degli algoritmi non basta sapere quali sono discriminanti migliori, è necessario anche controllarne la correlazione. Correlazioni significative, infatti, non aggiungono informazione utile ma solo complessità a livello di training. Nell'analisi di un campione in cui gli eventi del segnale sono caratterizzati da un certo numero di variabili parzialmente correlate tra loro, molti algoritmi ottengono risultati migliori quando il training è fatto su un numero ristretto di variabili non correlate piuttosto che sull'intero spazio di variabili discriminanti a disposizione. I risultati per ogni algoritmo possono anche dipendere pesantemente dal tipo di correlazione, lineare o non lineare.

Per questa ragione sul campione del segnale è stato fatto un controllo delle correlazioni relative tra le variabili discriminanti che sono emerse dallo studio nel capitolo precedente. I coefficienti di correlazione lineare tra le variabili sono riportati nella matrice in Fig.4.1. Nella matrice è riportata per completezza anche la massa invariante degli h-dijets, nonostante questa non verrà utilizzata per calibrare gli algoritmi.

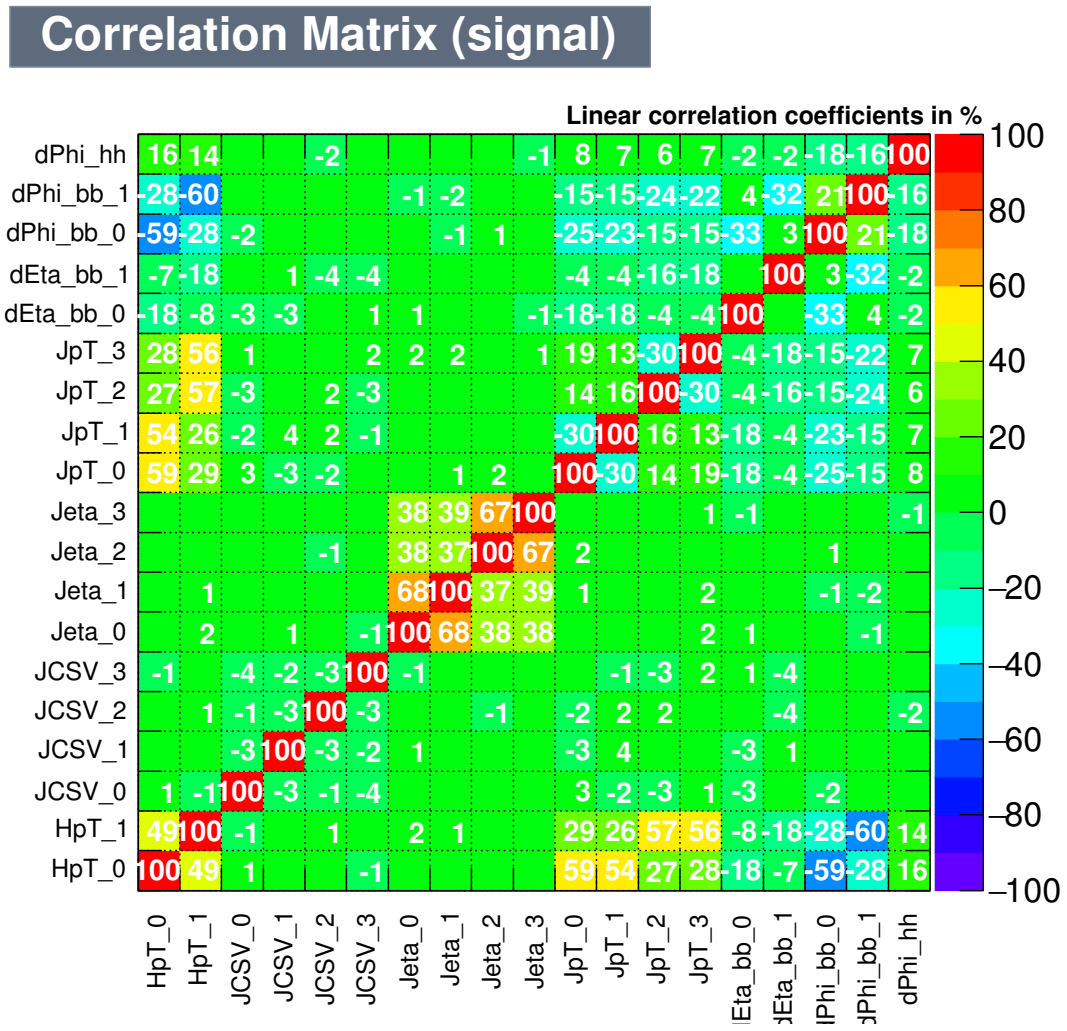

Figura 4.1: Matrice di correlazione tra tutte le variabili precedentemente analizzate con un potere discriminante soddisfacente.

In Fig.4.2 sono mostrati gli scatter plots delle correlazioni più significative. Si nota, in accordo con la matrice, che la correlazione lineare più rilevante è quella positiva tra gli angoli η dei b-jets accoppiati. Esistono poi correlazioni lineari tra il p_T di un b-jet e il p_T totale della coppia di cui fa parte, nonché tra il p_T delle due coppie di jets.

La correlazione negativa tra l'angolo $\Delta\phi$ tra due b-jets accoppiati e il p_T totale della coppia è anch'essa fondamentalmente lineare, soprattutto considerando la distribuzione di $\Delta\phi_{b\bar{b}}$.

Da un'analisi dei rimanenti scatter plots non emergono correlazioni non lineari significative.

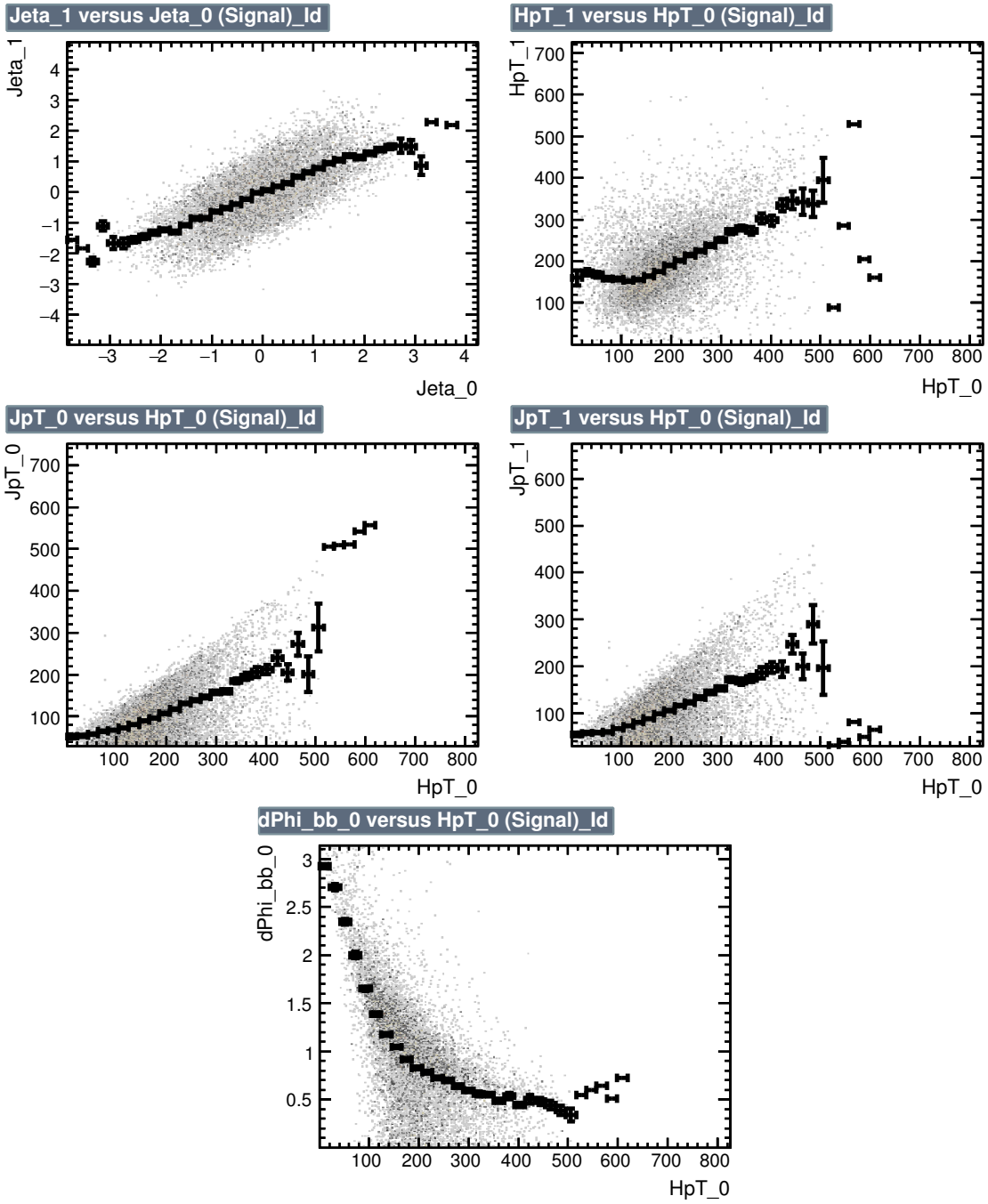

Figura 4.2: Scatter plots di correlazione più significativi tra le variabili analizzate.

4.3 Confronto tra algoritmi MVA

E' stato fatto qualche un confronto preliminare tra algoritmi, per identificarne le prestazioni "out of the box" e decidere su quali potesse valere la pena concentrarsi. I più promettenti si sono rivelati essere *Boosted Decision Trees* (BDT), *Likelihood* e *Linear Discriminant* (LD). Nell'appendice si trova una descrizione dettagliata di questi algoritmi. Questo esito preliminare corrisponde in buona parte alle aspettative, visti i dati a disposizione e prendendo come riferimento la *TMVA User Guide* [5].

Nel training degli algoritmi non è stata utilizzata la massa invariante delle coppie di jets, per due ragioni. Innanzitutto per utilizzarla come valore di controllo, ma anche per evitare di introdurre un bias sul background che, a causa del comportamento di questa variabile notato nel cap. 3, potrebbe risultare piccato sul valore di m_H.

Come menzionato nella sezione sulla preparazione dei dati, per ogni evento la combinazione con la risposta più alta da un algoritmo è stata scelta come candidata di segnale per quell'algoritmo. Per questa ragione, l'accuratezza è data dalla percentuale di match perfetti tra i candidati di un algoritmo e il segnale. Non vengono considerati match parziali.

Gli algoritmi scelti sono stati provati con varie combinazioni di variabili e opzioni. I risultati migliori per ogni algoritmo sono riportati in Tabella 4.1. Considerando come base l'insieme di tutte le variabili discriminanti a disposizione, tranne naturalmente la massa invariante delle coppie di jets per quanto menzionato precedentemente, i risultati migliori per BDT e Likelihood sono stati ottenuti rimuovendo dal training anche η e p_T dei b-jets. Per quanto riguarda LD, invece, i migliori risultati si sono ottenuti rimuovendo dal training solo η dei b-jets.

Algoritmo	Accuratezza
BDT	79.73%
LD	71.24%
Likelihood	70.53%

Tabella 4.1: Confronto dell'accuratezza degli algoritmi MVA utilizzati.

Le distribuzioni delle masse invarianti delle coppie di jets per questi algoritmi sono mostrate in Fig.4.5. La distribuzione per il metodo BDT, in particolare, mostra un ottimo accordo con il segnale.

Le distribuzioni della risposta di ogni algoritmo sono mostrate in Fig.4.3 e le relative curve ROC in Fig.4.4. Si noti che queste distribuzioni di segnale e rumore riguardano gli eventi combinatori su cui ha lavorato il TMVA.

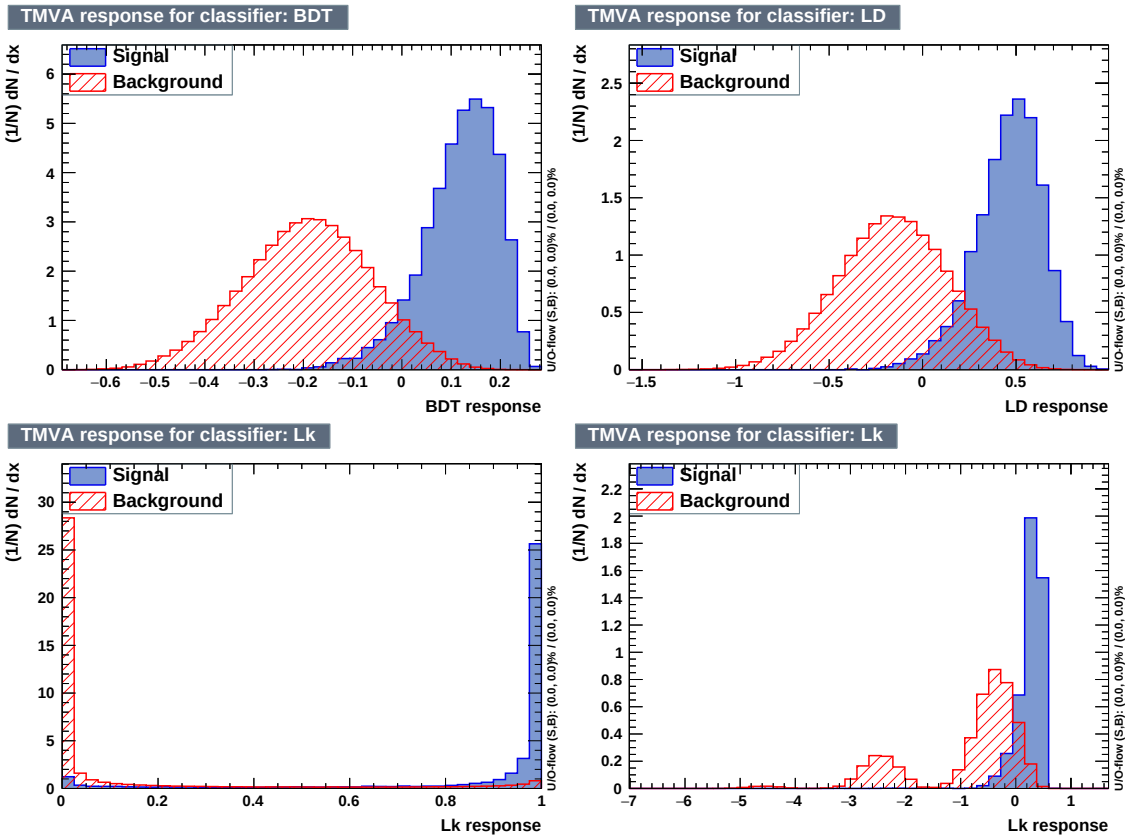

Figura 4.3: Risposta degli algoritmi BDT, LD e Likelihood. La risposta della Likelihood è mostrata due volte, prima e dopo la trasformazione dell'output (cfr. Appendice A.1).

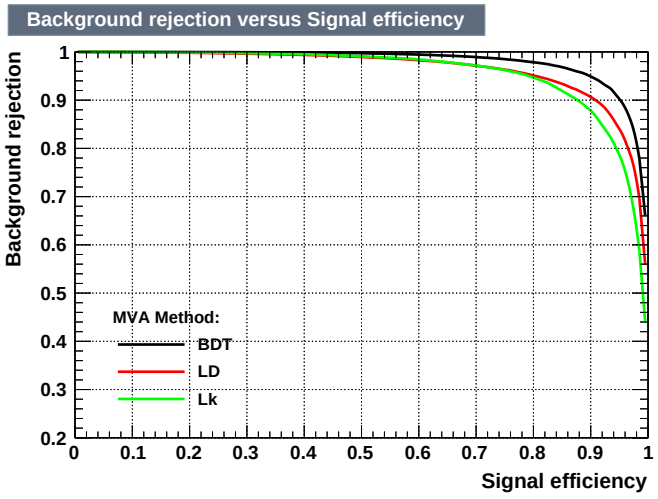

Figura 4.4: Curve ROC per gli algoritmi BDT, LD e Likelihood.

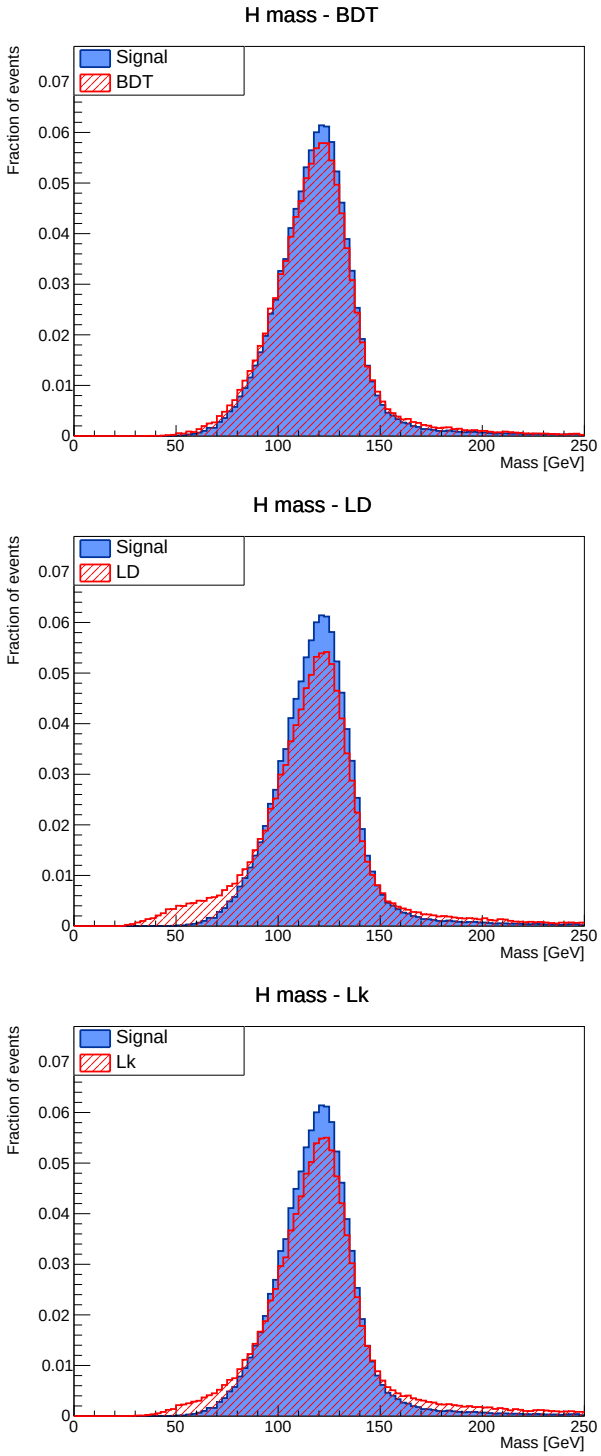

Figura 4.5: Confronto delle distribuzioni delle masse invarianti delle coppie di jets ottenute rispettivamente tramite BDT, LD e Likelihood rispetto al segnale.

Capitolo 5

Conclusioni

In questa tesi si è affrontato il problema della corretta ricostruzione dei decadimenti di coppie di bosoni di Higgs in stati finali comprendenti 4 jets da b quark.

Utilizzando una simulazione di MC del processo di segnale allo studio, si è inizialmente considerato il solo problema di identificazione dei b quarks nei jets adronici, derivandone una gerarchia di possibili strategie di selezione basate su b-tags. Tra gli algoritmi provati il miglior bilanciamento tra efficienza e accuratezza è stato trovato nell'algoritmo a 4 medium b-tags. In seguito allo studio delle variabili discriminanti questi algoritmi sono stati complementati dall'utilizzo dell'informazione angolare tra i jets, ma non sono stati riscontrati notevoli miglioramenti.

Sono stati quindi costruiti algoritmi multivariati al fine di incorporare al meglio le informazioni cinematiche nella scelta dei 4 jets originati dal decadimento e del loro giusto accoppiamento agli originari bosoni di Higgs. A causa della peculiarità delle richieste di combinazione dei jets, lo studio ha dimostrato un modo di applicazione di algoritmi MVA solitamente non praticato per la ricostruzione del segnale studiato. I risultati ottenuti, in particolare col metodo BDT, si considerano soddisfacenti.

Appendice A

Approfondimento sui metodi TMVA

A.1 Likelihood

Il metodo della *maximum likelihood* (massima verosimiglianza) consiste nel costruire un modello di funzioni di densità di probabilità che riproducano le variabili di input per segnale e background. Questo specifico approccio assume le variabili come indipendenti, ed è anche detto "naive Bayes estimator" per differenziarlo da altri approcci di likelihood multidimensionali.

Per l'evento i la probabilità di essere segnale è data dal rapporto di likelihood $y_{\mathcal{L}}(i)$, definito come

$$y_{\mathcal{L}} = \frac{\mathcal{L}_S(i)}{\mathcal{L}_S(i) + \mathcal{L}_B(i)} \tag{A.1}$$

dove

$$\mathcal{L}_{S(B)}(i) = \prod_{k=1}^{n_{var}} p_{S(B),k}(x_k(i)) \tag{A.2}$$

e $p_{S(B),k}\,(i)$ indica la funzione densità di probabilità per la k-esima variabile di input x_k. Le $p_{S(B),k}\,(i)$ sono normalizzate

$$\int_{-\infty}^{+\infty} p_{S(B),k}(x_k)dx_k = 1 \tag{A.3}$$

Si può dimostrare che, in assenza di imprecisioni nel modello (quali correlazioni tra variabili di input non rimosse dalla procedura di decorrelazione, o modelli inaccurati per la densità di probabilità), il rapporto mostrato nell'equazione A.1 fornisce la separazione ottimale del segnale dal rumore per un dato set di variabili di input.

Dal momento che la forma parametrica delle funzioni di densità di probabilità è generalmente sconosciuta, la loro forma è approssimata dai dati del campione di training da funzioni non parametriche che possono essere scelte individualmente per ogni variabile.

Se un problema offre un gran numero di variabili di input, o variabili di input con ottima discriminazione, la risposta $y_{\mathcal{L}}$ è spesso piccata a 0 (background) e 1 (segnale).

Questo tipo di risposta è svantaggioso per i passi successivi dell'analisi ed è possibile trasformarlo tramite una funzione sigmoide inversa che allarga i picchi

$$y_{\mathcal{L}}(i) \longrightarrow y'_{\mathcal{L}}(i) = -\tau^{-1} \ln(y_{\mathcal{L}}^{-1} - 1) \tag{A.4}$$

con $\tau = 15$.

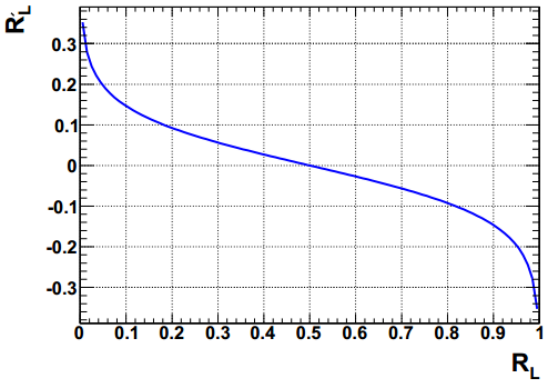

Figura A.1: Trasformazione dell'output della likelihood tramite funzione sigmoide inversa

A.2 Boosted Decision Trees

Un *boosted decision tree* (BDT) è un classificatore binario strutturato ad albero. Il tree (di cui un esempio è mostrato in Fig. A.2) consiste in un albero di nodi, ognuno dei quali contiene una decisione binaria presa su una specifica variabile discriminante x_i alla volta. Ogni decisione utilizza la variabile e il taglio che, al livello di quel nodo, forniscono la miglior separazione tra segnale e background. La stessa variabile può quindi essere usata in diversi nodi, mentre altre potrebbero non comparire mai. I nodi alla fine del tree sono contrassegnati con "S" per il segnale e "B" per il background, a seconda della maggioranza degli eventi che finiscono nel nodo relativo.

Il processo di scelte binarie divide lo spazio delle fasi in molte regioni, ognuna delle quali viene classificata come segnale o background durante il processo di training a seconda del tipo di eventi che finisce in maggioranza nel nodo finale. Essendo poi visualizzabile in forma di albero, un decision tree mantiene un'ottima trasparenza ed è facile da interpretare.

Una debolezza dei decision trees è la loro instabilità rispetto alle fluttuazioni statistiche nel campione di training da cui deriva la struttura del tree. Il *boosting* stabilizza la risposta dell'algoritmo estendendo il concetto da un singolo albero ad un gran numero di alberi che derivano dallo stesso insieme di training e sono infine combinati in un singolo classificatore, la cui risposta viene dalla media pesata dei singoli decision trees. In questo caso si perde però la facilità di interpretazione del singolo tree, anche se la struttura complessiva della selezione può emergere dal confronto tra un numero limitato di trees.

Il boosting in generale fornisce le migliori prestazioni quando è applicato a trees limitati (con una profondità di 3 nodi), che da soli hanno poco potere di classificazione. La profondità limitata offre anche il vantaggio di eliminare quasi completamente la pericolosa tendenza di overtraining dei singoli decision trees, che tipicamente crescono a grandi profondità prima di essere potati.

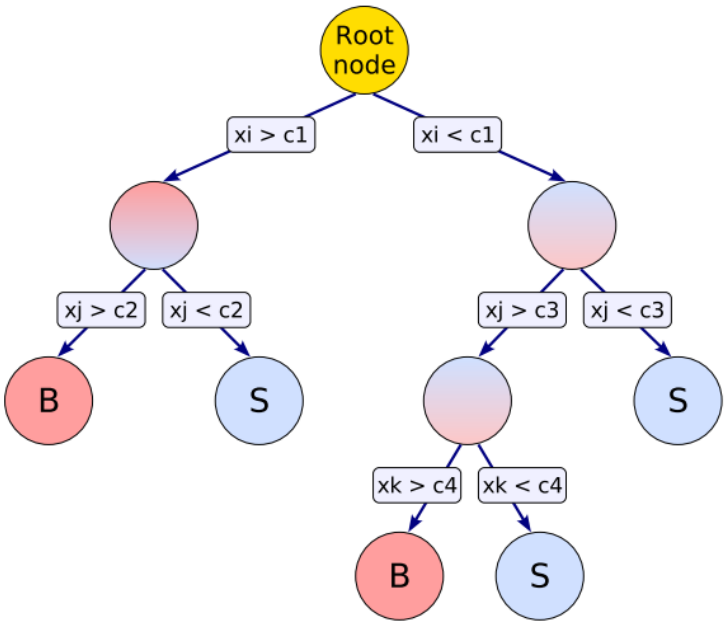

Figura A.2: Rappresentazione schematica di un decision tree

A.3 Linear Discriminant

L'analisi con *Linear Discriminant* (LD) consiste nell'individuazione di una combinazione lineare di variabili che sia in grado di separare il segnale dal background. Questo metodo classifica i dati utilizzando un modello lineare, riferito al fatto che la funzione discriminante

$$y(x) = x^\top \beta + \beta_0 \tag{A.5}$$

è lineare nel parametro vettoriale β. In questa funzione, il parametro β_0 (detto il *bias*) è calibrato in modo che sia $y(x) \geq 0$ per il segnale e $y(x) < 0$ per il background.

Assumendo che ci siano $m + 1$ parametri $\beta_0, \ldots, \beta_m$ da stimare utilizzando un set di training di n eventi, l'equazione che definisce il vettore di parametri β si scrive in forma matriciale come

$$Y = X\beta \tag{A.6}$$

dove β_0 è stato assorbito nel vettore β e sono state introdotte le matrici

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{pmatrix} \tag{A.7}$$

dove la colonna costante in X rappresenta il bias β_0 e Y è composto dai valori bersaglio, con $y_i = 1$ o $y_i = 0$ a seconda che l'evento i-esimo sia rispettivamente segnale o background. Applicando il metodo dei minimi quadrati si possono quindi ottenere le *equazioni normali* per il problema di classificazione, date da

$$X^T X\beta = X^T Y \Longleftrightarrow \beta = (X^T X)^{-1} X^T Y \tag{A.8}$$

La matrice $X^+ = (X^T X)^{-1} X^T$ è detta *pseudo inversa* di X e può essere considerata una generalizzazione della matrice inversa al caso di matrici non quadrate, nel caso in cui X abbia rango massimo.

In un problema di classificazione a due classi (segnale e background in questo contesto) è detta *decision boundary* la ipersuperficie che separa lo spazio vettoriale delle coordinate in due parti, una relativa ad ogni classe. Applicando questa idea al contesto attuale, l'algoritmo classificherebbe tutti gli eventi da una parte del decision boundary come segnale e tutti quelli dall'altra come background. Considerando due eventi x_1e x_2 sul decision boundary, si ha $y(x_1) = y(x_2) = 0$ e quindi $(x_1 - x_2)^T \beta = 0$. Il LD può quindi essere geometricamente interpretato come la ricerca del decision boundary attraverso il vettore β ad esso ortogonale.

Il LD è ottimale per variabili distribuite in modo gaussiano con correlazioni lineari e può essere competitivo con Likelihood. Quando una variabile ha la stessa sample mean per segnale e background non si ottiene discriminazione, ma il LD può spesso beneficiare da trasformazioni adeguate delle variabili di input. E' possibile dimostrare che questo metodo è equivalente al *Fisher Discriminant*, che cerca di massimizzare il rapporto tra varianza tra classi e varianza all'interno delle classi proiettando i dati su un sottospazio lineare.

Bibliografia

[1] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. 2012.

[2] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. 2012.

[3] CMS Collaboration. CMS Physics - Technical Design Report Volume I: Detector Performance and Software. 2006.

[4] Benoit Hespel, David Lopez-Val, and Eleni Vryonidou. Higgs pair production via gluon fusion in the Two-Higgs-Doublet Model. *JHEP*, 1409:124, 2014.

[5] Andreas Hoecker, Peter Speckmayer, Joerg Stelzer, Jan Therhaag, Eckhard von Toerne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.