Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Individuazione di anomalie collettive mediante metodi non parametrici semi-supervisionati: un'applicazione alla fisica delle particelle

Relatore: dott.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Correlatore: dott. Tommaso Dorigo
Dipartimento di Fisica

Laureando: Alessandro Casa
Matricola n.: 1100339

Anno Accademico 2015/2016

"Good, better, best.
Never let it rest.
Until your good is better
and your better is best."

Indice

Capitolo 1

Introduzione

L'espressione *individuazione delle anomalie* (*anomaly detection*) si riferisce a una serie di tecniche, statistiche e non, che mirano ad identificare comportamenti nei dati che si scostano da quello che è considerato essere il comportamento "normale", al quale nel seguito si farà riferimento con il termine *background*.

Generalmente, per risolvere problemi legati all'*individuazione di anomalie*, si fa riferimento a tecniche di analisi non supervisionata e, in particolare, a metodi di raggruppamento. Infatti è spesso complicato ottenere un campione rappresentativo del comportamento anomalo; qualora ad esempio ci si trovi in una situazione in cui non si ha la certezza che nei dati appaiano comportamenti di questo tipo risulta impossibile disporre di tale campione. Da un punto di vista statistico questo introduce quindi un problema rilevante in quanto preclude la possibilità di utilizzare tecniche di analisi supervisionata che richiederebbero un adeguato numero di osservazioni per entrambe le classi (normale e anomala).
È possibile tuttavia strutturare in maniera differente il problema qualora si supponga di disporre di un campione del quale è nota la provenienza dal *background*. Tale campione è solitamente più facile da ottenere e permette di utilizzare metodologie statistiche di analisi semi-supervisionata.

La maggior parte delle tecniche di analisi semi-supervisionata utilizza i dati provenienti dal *background* per stimare un modello generatore che descriva adeguatamente, in termini probabilistici, il comportamento normale del fenomeno che si sta studiando. Queste tecniche considerano quindi come anomale quelle osservazioni per le quali la probabilità di essere state generate da tale modello è minore di una determinata soglia. Un approccio di questo tipo risulta adeguato solamente nel caso in cui si sia nella situazione in cui le anomalie si trovano in zone a bassa densità dello spazio campionario del *background*. La maggior parte delle tecniche di *individuazione delle anomalie* incontrano maggiori difficoltà nel caso in cui queste anomalie giacciano nel supporto dei dati di *background* e di conseguenza tale problema è stato

raramente affrontato in letteratura.

Questo lavoro trova la sua collocazione nel contesto appena descritto e, in particolare, prende spunto da un problema ben noto nell'ambito della fisica delle particelle. Si tratta di una branca della fisica che studia le componenti della materia mediante l'analisi del comportamento di particelle ad alta energia fatte collidere all'interno di appositi acceleratori. La teoria fisica che al momento attuale rappresenta lo stato dell'arte per spiegare e classificare le componenti della materia è nota con il nome di *Modello Standard* (Beringer et al., 2012). Sebbene tale modello abbia trovato l'ultima conferma con l'osservazione, da decenni teorizzata, del Bosone di Higgs (Chatrchyan et al., 2012, Aad et al., 2012), si ritiene sia ancora incompleto. Per questo motivo sono state condotte analisi dipendenti da determinati modelli teorici in grado di descrivere eventuali fenomeni non ancora individuati ma attesi in quanto previsti da tali modelli. D'altro canto però, qualora le anomalie fossero legate a fenomeni fisici non ancora adeguatamente descritti da relative teorie, le tecniche supervisionate non permetterebbero di individuarle. Invece, questa situazione si presta particolarmente bene all'utilizzo di tecniche semi-supervisionate di individuazione delle anomalie, in quanto il comportamento assunto dal *background* è noto e adeguatamente descritto dal *Modello Standard*.

Da un punto di vista statistico, si assumerà che la distribuzione del *background* sia nota (o più realisticamente, che sia possibile stimarla arbitrariamente bene) e si cercherà la manifestazione di un eventuale segnale fisico di interesse in termini di allontanamento dalla distribuzione del *background*.

Lo scopo che questo lavoro si prefigge consiste nel proporre delle tecniche statistiche semi-supervisionate in grado di affrontare il problema dell'*individuazione di anomalie collettive*. Con tale espressione, proposta da Chandola, Banerjee e Kumar (2009), si fa riferimento alla situazione in cui *"i dati, presi individualmente, non sono considerabili anomali ma è la loro occorrenza come gruppo ad essere anomala"*. In una situazione del genere risulta evidente come l'individuazione di un eventuale picco o *cluster* nel dominio del *background* e usualmente non presente nei dati, fornisca un'indicazione riguardo la presenza di un comportamento anomalo ovvero un segnale. Questo lavoro generalizza ad un contesto non parametrico quanto fatto da Vatanen et al. (2012), i quali hanno affrontato i problemi menzionati in precedenza utilizzando un approccio parametrico e in particolar modo basato su modelli a mistura finita di componenti gaussiane. Si ritiene che, essendo lo scopo delle tecniche studiate quello di individuare eventuali comportamenti anomali ed ignoti nei dati a disposizione, un approccio non parametrico possa portare ad un sostanziale vantaggio in quanto, non assumendo e non vincolandosi particolari forme distributive, permette una maggiore libertà nell'individuazione di tali comportamenti. Inoltre,

i metodi non supervisionati non parametrici, a differenza di quelli parametrici, si basano sull'assunto che i *cluster* si manifestino come regioni ad alta densità, ovvero picchi nel dominio di osservazione.

In seguito si riportano i principali contributi di questa tesi:

- Vengono proposte due diverse tecniche di riduzione della dimensionalità appositamente introdotte per il particolare contesto in cui si sta operando. Tali tecniche cercano infatti di individuare le variabili, tra quelle a disposizione, che presentano un comportamento mutato rispetto al comportamento di *background*. Assumendo che il *background* sia stazionario tale differenza di comportamento è quindi interpretabile in termini di presenza di un comportamento anomalo che potenzialmente fornisce indicazioni su un particolare fenomeno non noto. Questi metodi di riduzione risultano fondamentali dal momento in cui ci si muove in un contesto non parametrico nel quale quindi la *maledizione della dimensionalità* risulta essere un problema particolarmente rilevante. Le due tecniche non solo quindi permettono di ricondursi a sottospazi di dimensione inferiore ma selezionano quelle variabili contenenti segnale; questo comporta un enorme vantaggio in termini di interpretabilità dei risultati;

- Viene proposto un metodo per selezionare la matrice di lisciamento in maniera ottimale, subordinatamente al problema in questione, per stimare non parametricamente una densità multivariata. La scelta del parametro di lisciamento risulta essere un tema particolarmente importante nel caso in cui si faccia riferimento a stime non parametriche della densità; tale problema assume dimensioni ancora maggiori nel caso in cui si operi in spazi multidimensionali e in cui si debba quindi selezionare non un unico parametro ma una matrice di lisciamento. In questo lavoro viene proposto un metodo per individuare tale matrice, che tenga conto del fatto che si opera in un contesto semi-supervisionato e che sia quindi vincolato a tenere in considerazione delle informazioni che si hanno a disposizione riguardo il *background*. Una volta applicato questo metodo di selezione, la matrice di lisciamento ottenuta può essere utilizzata in diversi modi permettendo in particolare di adattare metodi tipicamente non supervisionati al contesto semi-supervisionato;

- Viene proposta una modifica del metodo *MEM* (Li, Ray e Lindsay, 2007), un algoritmo di *clustering*, basato sul più noto algoritmo *Expectation Maximization* (EM), che mira a trovare le mode di una distribuzione di probabilità stimata utilizzando metodi non parametrici. I gruppi vengono definiti come le regioni di attrazione delle mode individuate. In questo lavoro il metodo è stato

modificato per tenere in considerazione il fatto che ci si trova in un contesto semi-supervisionato e per permettere di utilizzare le informazioni provenienti dal *background*. Così facendo è possibile affrontare il tema dell'*individuazione di anomalie collettive* utilizzando fondamentalmente un algoritmo di *clustering* in grado però di vincolarsi ad alcune informazioni a disposizione. Rimanendo nell'ambito non parametrico non si vincola l'eventuale segnale ad assumere particolari forme distributive: si ritiene questo possa costituire un evidente vantaggio soprattutto alla luce del fatto che non è noto se tale segnale sia presente e quindi tantomeno si pensa di conoscere il modo in cui dovrebbe presentarsi;

- A completamento dei contributi sopra menzionati, è stato predisposto il software per l'implementazione degli stessi, nonchè della procedura parametrica di individuazione delle anomalie collettive proposte da Vatanen et al. (2012).

La trattazione si sviluppa come segue.

Nel secondo capitolo vengono introdotti i metodi di analisi semi-supervisionata e quelli di individuazione delle anomalie in un contesto generale che prescinde dalle applicazioni fisiche, fornendo una breve rassegna di quanto presente in letteratura. Viene inoltre presentato nel dettaglio l'approccio adottato da Vatanen et al. (2012), dal quale questo lavoro prende spunto.

Nel terzo capitolo si procede con una maggiore contestualizzazione del problema e delle metodologie statistiche che vengono utilizzate per affrontarlo. Si riportano quindi dei cenni riguardo la stima non parametrica della densità basata sul metodo del nucleo. Si introducono le metodologie di *clustering* non parametrico che sono state successivamente utilizzate ed infine si parla dei problemi di cui tali metodologie risentono nel caso in cui ci si trovi in spazi ad elevata dimensionalità.

Nel quarto capitolo vengono presentati nel dettaglio i contributi di questo lavoro e nel quinto capitolo si riportano alcune analisi numeriche. Queste fanno riferimento ad uno studio di simulazione condotto con il principale obiettivo di studiare il comportamento dei metodi di selezione delle variabili introdotti. Successivamente si prosegue con l'applicazione delle tecniche presentate a dati relativi ad un problema di fisica delle particelle. I risultati saranno inoltre confrontati con quanto ottenuto facendo ricorso ai metodi proposti da Vatanen et al. (2012).

Si conclude infine con considerazioni di ordine generale volte a fornire un'analisi critica dei risultati e dei contributi introdotti, cercando di evidenziare pregi e difetti degli stessi e cercando di fornire eventuali ulteriori spunti.

Capitolo 2

Individuazione di anomalie

2.1 Individuazione di anomalie e apprendimento semi-supervisionato

2.1.1 Individuazione di anomalie

Con l'espressione *individuazione di anomalie* si fa riferimento ad una serie di metodologie applicate per rilevare eventuali osservazioni, o gruppi di queste, che abbiano un comportamento difforme rispetto a quanto ci si aspetti osservando un determinato fenomeno. Facendo riferimento al significato comune del termine, e non a quello statistico, nel seguito si utilizzerà il termine "normale" per indicare il comportamento atteso dei dati, in contrapposizione a quello anomalo che si intende rilevare. Alternativamente, si definirà *background* il processo generatore dei dati aventi comportamento normale e *segnale*, l'eventuale anomalia.

Negli ultimi anni si è assistito ad un crescente interesse nei confronti delle tecniche di individuazione delle anomalie, in quanto applicabili a diversi contesti: esempi ne sono le tecniche per l'individuazione di frodi bancarie, di patologie non note, o come in questo lavoro, di un possibile segnale fisico di interesse.

Tali metodologie sono estremamente varie ed eterogenee, non solo per la pluralità dei contesti in cui trovano applicazione, ma anche per le difficoltà che il problema in esame comporta. A livello definitorio, è complesso riuscire a fornire una definizione precisa e al contempo universale di anomalie. A livello operativo, non è sempre ovvio definire il comportamento normale del fenomeno di interesse, che talvolta presenta caratteristiche di non stazionarietà. Infine, non è scontato riuscire ad ottenere una quantità adeguata di dati utilizzabili per la fase di stima dei modelli, soprattutto per quanto riguarda la classe delle anomalie, che tipicamente è molto rara rispetto alla classe di osservazioni avente un comportamento ritenuto normale.

Di seguito viene fornita una breve rassegna delle principali classi di metodi per l'individuazione di anomalie con lo scopo di fornire una descrizione generale del contesto in cui ci si muoverà nei capitoli successivi. Per una rassegna completa si vedano Chandola, Banerjee e Kumar (2009) e Markou e Singh (2003).

In base al tipo di anomalie che si pensa siano presenti, Chandola, Banerjee e Kumar (2009) distinguono tra metodi per l'individuazione di anomalie puntuali, anomalie contestuali ed anomalie collettive. Con anomalie puntuali si intendono quelle singole osservazioni non conformi alla distribuzione dei dati ritenuta "normale" mentre con anomalie contestuali si intendono le osservazioni che risultano anomale solamente se osservate in uno specifico contesto (ad esempio in un determinato istante temporale). Le anomalie collettive, invece, vengono definite sottolineando come, "*i dati, presi individualmente, non sono considerabili anomali ma è la loro occorrenza come gruppo ad essere anomala*", (Chandola, Banerjee e Kumar, 2009). Le tecniche utilizzabili per individuare anomalie di questo tipo sono sostanzialmente differenti rispetto a quelle a cui si fa ricorso per le altre due tipologie. Gli autori puntualizzano come la rilevazione di tali gruppi anomali sia in generale più complessa, e generalmente affrontata meno frequentemente in letteratura, e richieda spesso informazioni aggiuntive sui dati o la conoscenza di una particolare relazione tra gli stessi: solitamente è richiesta la presenza di una struttura nei dati che sia esprimibile in termini di rete, relazioni sequenziali o spaziali.

In base all'approccio operativo utilizzato per l'individuazione delle anomalie è possibile distinguere tra le seguenti classi di metodi:

- Metodi basati sulla distanza (vicini più vicini): si cercano di individuare le anomalie utilizzando concetti legati alla distanza di un punto rispetto agli altri. L'assunzione alla base richiede che le anomalie si trovino in regioni a bassa densità dello spazio campionario;

- Metodi di classificazione: la maggior parte dei metodi di individuazione delle anomalie è riconducibile a un contesto di classificazione, dove le classi sono rappresentate dal *background* e dal segnale. In base alla quantità di informazione a disposizione, tali metodi si differenziano in supervisionati, non supervisionati e semi-supervisionati. Nel primo caso si assume che sia possibile disporre di una quantità adeguata di dati per la stima del modello statistico, anche appartenenti alla classe relativa alle anomalie; in questo caso i dati normali possono avere una struttura multiclasse. Più flessibili e maggiormente utilizzate sono le metodologie di tipo non supervisionato, ovvero tecniche di *clustering* che non necessitano di una classificazione dei dati a disposizione. Un'assunzione frequente in questo caso prevede che le anomalie non appartengano a nessun

cluster; tale ipotesi incontra dei problemi nel caso in cui si utilizzi un metodo che raggruppa in maniera esaustiva tutte le osservazioni e può quindi essere riformulata in termini di distanza delle stesse dal centroide del cluster più vicino. Un'osservazione viene quindi considerata anomala qualora si trovi ad una distanza maggiore di una determinata soglia fissata rispetto a tale centroide. Le similitudini con i metodi basati sui vicini più vicini sono evidenti, in particolare qualora si faccia ricorso ad un algoritmo di *clustering* che utilizzi principi geometrici per individuare un raggruppamento. Un approccio alternativo, che sarà in parte esplorato in questo lavoro, consiste nell'assumere che le anomalie rappresentino un *cluster* a sè (assunzione verosimile qualora si sia in presenza di anomalie collettive). Nel caso in cui ci si trovi in una delle molte situazioni intermedie in cui nè siano note entrambe le classi, nè siano entrambe non note, l'approccio usato è di tipo semi-supervisionato. Questo approccio verrà descritto nel paragrafo successivo per la particolare rilevanza che assume in questo lavoro;

- Verifica d'ipotesi: il problema in esame può essere affrontato stimando la densità dalla quale provengono i dati aventi un comportamento normale e, successivamente, definendo come anomale quelle osservazioni che non risultano compatibili con tale densità, mediante un test statistico. Usualmente queste tecniche fanno riferimento a metodi riconducibili a verifiche d'ipotesi e richiedono che l'eventuale classe di osservazioni anomale si trovi in una regione dello spazio campionario differente rispetto alle classi note ed aventi comportamento considerabile normale. Essendo necessaria una stima della densità, tali metodi risentono dei ben noti problemi che si incontrino qualora si operi in contesti ad elevata dimensionalità; in queste situazioni è quindi necessario di disporre di un gran numero di osservazioni per garantire delle stime affidabili. Le tecniche facenti riferimento a tale classe possono essere classificate in base all'approccio di stima adottato, parametrico o non parametrico (Markou e Singh, 2003);

- Tecniche di riduzione della dimensionalità: si assume di poter ridurre la dimensionalità mantenendo allo stesso tempo intatta la struttura di interesse dei dati ed evidenziando la presenza di anomalie. Il vantaggio di tali tecniche consiste nel minor onere computazionale mentre il maggior svantaggio riguarda il fatto che è richiesta separabilità tra anomalie e dati normali nei sottospazi in cui si opera.

Alcuni autori hanno differenziato il problema di individuazione delle anomalie da quello dell'individuazione delle novità: l'unica differenza sostanziale tra questi

approcci consiste nel fatto che, in quest'ultimo, una volta rilevati eventuali comportamenti anomali, questi vengono incorporati nel modello in modo tale da fornire una descrizione più completa del fenomeno che si sta studiando. Le anomalie vengono quindi considerate come qualcosa di non ancora osservato e non come dei comportamenti non compatibili con il fenomeno che si sta studiando.

2.1.2 Apprendimento semi-supervisionato

Il problema affrontato in questa tesi riguarda la conoscenza parziale del fenomeno di interesse in termini di appartenenza alla classe del *background* e quella del segnale. Tale conoscenza può concretizzarsi nel fatto che siano noti dei vincoli riguardo la necessità che alcune osservazioni appartengano ad una stessa classe, oppure, come nel caso in esame, nell'osservazione della sola classe di osservazioni normali. In quest'ultimo caso il problema si formula come segue: si dispone di due insiemi di dati $\mathcal{X}_b$ e $\mathcal{X}_{bs}$, del primo è nota l'appartenenza di tutte le osservarzioni alla classe del *background*, mentre del secondo non sono note le etichette di classe.

In tale contesto, la volontà di inserire l'informazione a disposizione nella formulazione e stima del modello statistico si traduce nell'uso di un approccio semi-supervisionato. Allo scopo di contestualizzare quanto sarà presentato, si fornisce di seguito una panoramica generale delle metodologie studiate ed applicate in questo ambito. Per una rassegna si vedano Zhu (2005) e Chapelle, Scholkopf e Zien (2009).

Innanzitutto è importante notare come l'interesse nei confronti di metodologie semi-supervisionate stia crescendo negli ultimi anni. Tali metodologie risultano infatti particolarmente utili in tutti quegli ambiti applicativi nei quali vengono raccolte grandi moli di dati con facilità ma dove, allo stesso tempo, risulta più complicato ottenere una classificazione dei dati stessi: basti pensare a come aumentino i possibili errori e i costi nei casi in cui tale classificazione richieda infatti l'intervento dell'uomo. In queste situazioni quindi risulta utile avere a disposizione alcune tecniche che siano in grado di estrarre informazione dai dati avendone una conoscenza parziale. In letteratura ci si è spesso chiesto se il ricorso a queste tecniche porti a dei reali vantaggi; non è scontato intatti che i dati non etichettati migliorino effettivamente le prestazioni dei metodi di classificazione portando quindi informazione realmente utilizzabile ai fini di descrivere un determinato fenomeno adeguatamente. In qualche caso è stato dimostrato che l'utilizzo di dati non classificati ha portato ad un peggioramento dell'accuratezza previsiva. Sebbene tali considerazioni dipendano dai dati che si hanno a disposizione e dal contesto in cui le metodologie vengono applicate viene comunemente sottolineato che l'utilizzo congiunto di osservazioni etichettate e non, porti dei vantaggi qualora il modello che si sta considerando sia corretto mentre

nel caso contrario porta ad una degradazione delle stime. Da questa considerazione, essendo spesso complicata la formulazione di un modello che fornisca una buona approssimazione del processo generatore dei dati, si evince come tale problema sia particolarmente delicato e come vadano fatte delle valutazioni specifiche nelle diverse situazioni.

I metodi di apprendimento semi-supervisionato fanno usualmente ricorso alle seguenti assunzioni:

- *Smoothness*: se due punti si trovano in una regione ad elevata densità e sono vicini, allora lo saranno anche le corrispondenti etichette di classe;
- *Cluster*: se due punti appartengono ad uno stesso *cluster* è probabile che appartengano alla stessa classe; tale assunzione può esser riformulata in termini di confini di separazione tra le classi che devono trovarsi in regioni a bassa densità;
- *Manifold*: i dati giacciono in uno spazio di dimensionalità inferiore rispetto a quella originale.

I diversi metodi di apprendimento semi-supervisionato che sono stati studiati e proposti in letteratura possono essere suddivisi in modelli generativi, modelli con separazione in regioni a bassa densità e modelli grafici.

I modelli generativi assumono che la distribuzione dei dati, condizionata all'appartenenza ad una classe, sia una mistura le cui componenti risultano identificabili, a livello ideale, anche nel caso in cui si abbia una sola osservazione etichettata per classe accompagnata da un elevato numero di osservazioni non etichettate.

Nei modelli grafici si presuppone che la struttura dei dati possa esser descritta attraverso una rete nella quale i nodi sono le osservazioni sia etichettate che non mentre gli archi riflettono la similarità tra osservazioni. Tali modelli sono probabilmente i più studiati in letteratura nel contesto dell'analisi semi-supervisionata ed esistono delle sostanziali differenze tra i diversi approcci proposti: la caratteristica comune riguarda il fatto che, usualmente, fanno ricorso all'assunzione di *smoothness* e cercano quindi di propagare le etichette di classe, partendo dalle osservazioni per le quali queste sono disponibili, lungo gli archi del grafo in base alla similarità.
Si fa notare come, generalmente, tali metodi abbiano la necessità di avere un certo numero di osservazioni etichettate per ciascuna delle classi presenti.

I modelli con separazione in regioni a bassa densità si focalizzano maggiormente sulla distribuzione di probabilità dei dati, a prescindere dall'etichetta di classe: utilizzando questi dati si cerca quindi di individuare delle classi tali che le superfici di separazione tra le stesse si trovino in regioni dello spazio campionario aventi

bassa densità. Questa classe di modelli presenta evidenti punti di contatto con le procedure di *clustering*.

Il *clustering* semi-supervisionato, nella sua accezione più generale, appartiene ai metodi che cercano di fornire superfici di separazione in regioni a bassa densità. L'analisi di raggruppamento è uno di quei casi in cui l'informazione aggiuntiva, che permette di passare da un ambito non supervisionato ad un ambito semi-supervisionato, può essere fornita sia in termini di conoscenza parziale di appartenenza a delle classi di alcune osservazioni sia in termini di vincoli riguardanti la necessità o meno che queste appartengano ad uno stesso *cluster* (si veda ad esempio Basu, Bilenko e Mooney, 2004). Nel caso in cui ci si trovi nella situazione più "usuale" in cui è disponibile una classificazione per una parte dei dati, tale informazione può essere utilizzata per aggiustare un metodo di clustering esistente. Ad esempio Basu, Banerjee e Mooney (2002) modificano il metodo delle *k-medie* inizializzando l'algoritmo con una procedura che fa ricorso alle osservazioni per le quali sia nota la classe di appartenenza e alle informazioni in queste contenute.

2.2 Metodo parametrico di rilevamento di anomalie collettive

Uno dei lavori da cui questa tesi trae spunto è quello di Vatanen et al. (2012) i quali hanno affrontato il problema semi-supervisionato di conoscenza parziale del fenomeno di interesse mediante un approccio parametrico.

Nel lavoro appena menzionato lo studio dei metodi proposti è collegato all'ambito della fisica delle particelle nel quale l'individuazione di comportamenti anomali rispetto a quanto previsto dal *modello standard* potrebbe portare alla formulazione di nuove teorie. Ci si trova formalmente in un contesto di *individuazione delle novità* in quanto, qualora venisse individuato del segnale non conforme alle descrizioni fornite dal *Modello Standard*, si cercherebbe di capire se è possibile inglobare queste nuove informazioni in una teoria più completa.

Gli autori, confermando quanto detto nei paragrafi precedenti, fanno notare come l'approccio più usuale al problema di rilevamento di anomalie consista nello stimare un modello, detto modello di *background*, che descriva adeguatamente il comportamento usuale dei dati per poi cercar di testare se un'eventuale osservazione proveniente da un campione non etichettato possa provenire da tale modello. Ovviamente un'osservazione viene classificata come anomala qualora risulti poco probabile la sua provenienza dal modello di *background*. Un approccio di questo tipo si pone naturalmente in un contesto di verifica d'ipotesi ma permette di con-

siderare come anomale solo quelle osservazioni non compatibili con il modello di *background*. Un limite evidente di tale approccio consiste quindi nella incapacità di tener conto della presenza di eventuali *anomalie collettive*.

Una strada percorribile per l'individuazione di questo tipo di anomalie consiste nel collegare l'individuazione ad un'analisi di raggruppamento; così facendo, infatti, un eventuale *cluster* presente nei dati non etichettati e non rilevato nel campione di *background* risulterebbe essere anomalo anche qualora dovesse presentarsi in una regione dello spazio campionario compatibile con quella del modello che descrive il comportamento usuale dei dati. Per questa ragione, la procedura di Vatanen et al. (2012) si configura come un adattamento al contesto semi-supervisionato di una classe di procedure di *clustering*, in particolare basate sull'uso di modelli parametrici.

Il metodo che verrà in seguito illustrato più in dettaglio si basa sulle seguenti assunzioni:

- disporre di due campioni, $\mathcal{X}_b$ e $\mathcal{X}_{bs}$, il primo dei quali proveniente dal processo di *background* mentre il secondo potenzialmente contenente anche segnale ma per il quale non si dispone delle etichette di classe;
- le anomalie si presentano assieme come un eccesso di massa rispetto alla distribuzione di *background*;
- è presente un numero di osservazioni anomale tale da permettere approcci inferenziali;
- la dimensionalità dei dati è o può essere ridotta ad una grandezza tale da permettere la stima della densità utilizzando modelli di mistura;
- il *background* ha una distribuzione stazionaria.

2.2.1 Modello a background fisso

Per risolvere il problema del rilevamento di anomalie collettive gli autori hanno proposto un procedimento a due passi. Per prima cosa viene stimata, in maniera parametrica, la densità di *background* $p_B(x)$ utilizzando i dati $\mathcal{X}_b$ dei quali è nota la classificazione. In secondo luogo si modellano i dati non etichettati $\mathcal{X}_{bs}$ utilizzando quello che viene chiamato *modello a background fisso* ed indicato come $p_{FB}(x)$ che è definito come:

$$p_{FB}(x) = (1 - \lambda)p_B(x) + \lambda p_S(x), \tag{2.1}$$

dove $p_S(x)$ ha il compito di cogliere eventuali anomalie. Il *modello a background fisso* viene stimato mantenendo costante $p_B(x)$ e stima quindi $p_S(x)$ per catturare comportamenti considerabili non usuali rispetto alla distribuzione di *background*.

Si assume che $p_B(x)$ sia una mistura finita di componenti gaussiane:

$$p_B(x|\theta) = \sum_{j=1}^{J} \pi_j N(x|\mu_j, \Sigma_j), \tag{2.2}$$

dove J è il numero di componenti della mistura, π_j sono le proporzioni della mistura, $N(x|\mu_j, \Sigma_j)$ indica la funzione di densità di una variabile casuale gaussiana con vettore delle medie μ_j e matrice di varianza e covarianza Σ_j (per ulteriori approfondimenti riguardo i modelli mistura si veda McLachlan e Peel, 2004).
Questo modello viene stimato utilizzato l'algoritmo *Expectation-Maximization* (EM) (Dempster, Laird e Rubin, 1977) che permette in questa situazione di ottenere delle stime in forma chiusa per i parametri del modelllo (2.2).

Gli autori hanno successivamente indicato come riuscire a stimare il *modello a background fisso* utilizzando nuovamente l'algoritmo EM e i dati di cui non si dispone di una classificazione; una complicazione deriva dal fatto che $p_B(x)$ è da considerare fissato mentre vanno stimati sia il parametro λ che i parametri relativi a $p_A(x)$ in (2.1). In questo contesto $p_A(x)$ può essere un'ulteriore mistura con Q componenti gaussiane. Il modello in (2.1) viene espresso quindi come:

$$\begin{aligned} p_{FB}(x) &= (1-\lambda)p_B(x) + \lambda \sum_{q=J+1}^{J+Q} \tilde{\pi}_q N(x|\mu_q, \Sigma_q) \\ &= \pi_B p_B(x) + \sum_{q=J+1}^{J+Q} \pi_q N(x|\mu_q, \Sigma_q), \end{aligned} \tag{2.3}$$

con $\pi_B = 1 - \lambda$ e $\pi_q = \lambda\tilde{\pi}_q, q = J+1, \ldots, J+Q$.

L'algoritmo EM in tale situazione opera in maniera analoga al caso standard alternando iterativamente i due passi con la differenza che, nel passo di massimizzazione, i parametri relativi alla media e alla struttura di covarianza delle componenti relative a $p_B(x)$ vengono considerati costanti.
Per poter avere un'indicazione riguardo il numero di componenti nei modelli mistura gli autori hanno utilizzato il *criterio di informazione basato su convalida incrociata* (Smyth, 2000); in generale si può fare riferimento a diversi criteri per prendere tale decisione e, ad esempio, Fraley e Raftery (2002), nel loro approccio all'analisi di raggruppamento basato sui modelli a mistura di componenti gaussiane, utilizzano il *criterio di informazione di Bayes* (BIC).

Per quanto riguarda la valutazione di bontà di adattamento di questi modelli, gli autori fanno ricorso ad un semplice confronto in termini di verosimiglianza: si

prende a riferimento la verosimiglianza relativa al modello per il *background* e vengono combinate con tale modello, sequenzialmente, le componenti della mistura del modello per le anomalie. Qualora tali componenti colgano effettivamente qualche comportamento anomalo nei dati dei quali non si dispone di classificazione, la verosimiglianza tenderà ad aumentare. Nel caso in cui invece non mostri degli incrementi significativi, le stime dei parametri di tale componente vengono poste uguali a dei valori casuali. Questo procedimento viene sfruttato iterativamente per decidere quali componenti rimuovere dal modello $p_A(x)$: una componente viene esclusa dal modello qualora venga re-inizializzata a valori casuali un numero di volte superiore ad una determinata soglia. Utilizzando tale procedura si vuole quindi ottenere $p_{FB}(x)$; tale modello sarà analogo a $p_B(x)$ qualora, nella procedura iterativa descritta in precedenza, tutte le componenti della mistura relativa al modello per le anomalie vengano tolte. Un comportamento del genere, riscontrato durante la procedura di stima, fornirebbe quindi un segnale riguardo il fatto che non siano presenti delle osservazioni anomale nei dati non etichettati che quindi assumono un comportamento assimilabile a quello descritto dal modello di *background*.

In relazione all'assunzione precedentemente riportata riguardo la possibilità di ridurre la dimensionalità dei dati si noti che tale assunzione è necessaria perchè i modelli a mistura finita hanno un numero di parametri da stimare che cresce velocemente all'aumentare della dimensione dello spazio in cui si opera. Spesso, nell'utilizzare questi modelli, è quindi necessario ridurre la dimensione del problema prima di procedere con la stima degli stessi. Nel lavoro di cui si sta parlando gli autori hanno utilizzato come metodo di riduzione l'*analisi delle componenti principali*.

Si fa notare infine come il modello a *background fisso* proposto da Vatanen et al. (2012) permetta di fornire indicazioni per una varietà di diversi obiettivi:

- *Classificazione*: le osservazioni possono essere classificate come anomalie utilizzando come criterio la probabilità a posteriori di appartenere a $p_A(x)$;
- *Proporzione di anomalie*: la stima del parametro λ in (2.1) permette di stimare la proporzione di osservazioni considerate anomale nel campione non etichettato;
- *Significatività delle anomalie*: può essere utilizzato un test (solitamente il test del rapporto di verosimiglianza) per verificare l'ipotesi $\lambda = 0$ e per verificare quindi se le anomalie rilevate siano considerabili significanti.

Capitolo 3

L'approccio non parametrico al clustering

3.1 Introduzione al clustering basato sulla densità

Nel capitolo precedente si è sottolineato come in questo lavoro si prenda spunto da Vatanen et al. (2012) adattando quanto fatto dagli autori ad un contesto non parametrico, poichè si ritiene che tale generalizzazione possa portare dei vantaggi in termini di flessibilità delle procedure utilizzate e dei risultati ottenibili. L'obiettivo è quello di estendere alcune ben note procedure non parametriche di *clustering* ad un contesto semi-supervisionato di individuazione di un eventuale comportamento anomalo, che in questo caso rappresenta il segnale fisico di interesse, presumendo noto il comportamento usuale del *background*.

L'utilizzo di procedure di analisi di raggruppamento è giustificato dal fatto che si fa riferimento ad un problema di *individuazione delle anomalie collettive*, dove cioè un eventuale segnale si presenta come un picco nella distribuzione dei dati, non presente nella distribuzione di *background*. È evidente come si possa identificare tale picco con un nuovo gruppo e come ci si possa quindi ricondurre ad utilizzare metodologie di *clustering* per individuarlo.

Con il termine *clustering* si intendono tutte quelle tecniche che cercano di trovare dei gruppi nei dati. Tali metodologie sono applicate nei più svariati ambiti, sia come analisi preliminare che permette di cogliere in maniera più immediata alcune strutture presenti nei dati, sia come obiettivo conclusivo delle analisi.

Le tecniche esistenti di analisi di raggruppamento possono essere suddivise in due diverse tipologie. La prima di queste, alla quale appartengono i *metodi gerarchici* quali il metodo *del legame singolo* e *del legame completo*, e i *metodi di partizione* quali il metodo delle *k-medie*, utilizza come criterio alla base del raggruppamento

una misura di distanza tra le osservazioni.
Le metodologie appartenenti a questa categoria presentano, a fronte di una semplicità concettuale, alcune criticità strettamente connesse alla mancanza di fondamenti statistici alla loro base. Sebbene vengano frequentemente utilizzate, queste tecniche non permettono ad esempio di ricorrere a procedure inferenziali e non affrontano il problema riguardante il numero dei gruppi presenti nei dati, che risulta essere uno dei temi più delicati nell'ambito dell'analisi di raggruppamento.

Il secondo approccio al *clustering* è basato sulla densità: i gruppi vengono associati a delle specifiche caratteristiche della distribuzione di probabilità che si assume possa descrivere adeguatamente i dati a disposizione. Le tecniche appartenenti a questa categoria hanno il vantaggio di essere inserite in un contesto statisticamente più rigoroso che permette di fare inferenza sul numero dei gruppi o sulla bontà della partizione fornita. Tale approccio è stato sviluppato in due direzioni distinte: da un lato il *clustering* basato su modelli parametrici e dall'altro il *clustering* modale o non parametrico.

L'analisi di raggruppamento parametrica si basa su un modello statistico parametrico, per descrivere la densità dei dati, selezionato tra i modelli a mistura finita con componenti appartenenti ad una determinata famiglia. La procedura di analisi prevede la stima del modello con il metodo della massima verosimiglianza, solitamente utilizzando l'algoritmo *Expectation-Maximization* (EM), che permette di calcolare la probabilità a posteriori di ogni componente della mistura data una determinata osservazione; tale osservazione verrà infine assegnata alla componente con associata la probabilità a posteriori più elevata. Si noti che questo approccio prevede che ogni componente della mistura rappresenti uno specifico gruppo; questo, se da un lato risulta conveniente a livello matematico, dall'altro può essere una limitazione nel caso in cui l'assunzione parametrica alla base sia violata o nel caso in cui i gruppi non siano nettamente separati. Risulta comunque evidente come i vantaggi ottenibili ricorrendo ad un approccio simile siano legati ai concetti statistici alla base, che permettono ad esempio di considerare misure di incertezza dell'analisi di raggruppamento (in termini di probabilità a posteriori) e di scegliere automaticamente il numero di gruppi presenti nei dati con criteri automatici quali il *criterio di informazione di Bayes* (*BIC*). Per ulteriori approfondimenti riguardo questo approccio si veda ad esempio Fraley e Raftery (2002).

Vale la pena precisare che il metodo per l'individuazione del segnale implementato da Vatanen et al. (2012) rappresenta un adattamento semi-supervisionato all'analisi di raggruppamento parametrica. In entrambi i casi ci si basa su un modello a mistura con un numero finito di componenti che vengono stimate utilizzando l'algoritmo EM. Qualora si faccia riferimento al segnale come ad una nuova com-

ponente della mistura non rilevata nei dati di *background* è immediato considerare tale componente come un nuovo gruppo presente nei dati; la presenza di un tale comportamento fa sì che ci si trovi ad affrontare un problema di individuazione di anomalie collettive.

L'idea alla base dei metodi di *clustering* non parametrico è stata introdotta da Carmichael e Julius (1968) dove un *cluster* viene definito come *"una regione dello spazio continua e relativamente densamente popolata, circondata da regioni continue e relativamente vuote"*. Tale idea è stata successivamente ripresa da vari autori, che hanno proposto delle procedure coerenti con tale nozione dei gruppi. Solo negli ultimi anni l'analisi di raggruppamento modale è stata diffusamente studiata, principalmente grazie agli avanzamenti compiuti in ambito computazionale.

La formalizzazione della nozione di gruppo come regione densamente popolata avviene associando i gruppi al dominio di attrazione delle mode della funzione di densità sottostante i dati, stimata non parametricamente. Tale intuizione, se da un lato permette già di inserire l'analisi di raggruppamento in un contesto statistico più rigoroso rispetto ai metodi basati su distanza, dall'altro non vincola i *cluster* ad avere una determinata forma come fatto dai metodi basati su modelli a mistura finita. Inoltre il numero di *cluster*, corrispondente al numero di mode, è una proprietà intrinseca delle funzione di densità e viene automaticamente stimato dalle procedure.

A fronte di tali vantaggi, nel momento in cui si faccia uso di tecniche di analisi di raggruppamento non parametriche vi sono alcune problematicità da tenere particolarmente in considerazione. Questi problemi riguardano:

- Stima della densità, necessaria alla successiva identificazione delle regioni modali;
- Individuazione operativa dei gruppi intesi come regioni ad elevata densità dello spazio campionario;
- Gestione della dimensionalità dei dati.

Nei prossimi paragrafi l'approccio non parametrico al *clustering* viene discusso dettagliatamente, con particolare riferimento agli aspetti ora menzionati.

3.2 Stima non parametrica della densità

Tra i vari stimatori non parametrici di una funzione di densità (si veda ad esempio Wand e Jones, 1994) il metodo del nucleo (Parzen, 1962) è certamente il più noto e diffuso. Esso rappresenta una generalizzazione del concetto di istogramma e consente

di definire uno stimatore per la funzione di densità senza dover ricorrere a delle assunzioni parametriche con i rischi che queste comportano in termini di robustezza dei modelli.

Sia $\mathcal{X} = \{x_1, \ldots, x_n\}$ un campione di n osservazioni da una variabile casuale X avente densità $\mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, lo stimatore *kernel* risulta essere:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i), \tag{3.1}$$

dove h è il parametro di lisciamento e $K_h(\cdot)$ è il nucleo che deve soddisfare alcune proprietà in modo da garantire la validità delle stesse anche per quanto riguarda la stima $\hat{f}(x)$.
In generale la scelta di $K(\cdot)$ dovrebbe determinare la forma della funzione mentre la scelta di h dovrebbe determinare la varianza dello stimatore. Nella pratica si è provato che differenti specificazioni di $K(\cdot)$ non portano a cambiamenti sostanziali nella stima della densità mentre è di fondamentale importanza una scelta adeguata del parametro di lisciamento. La criticità della scelta deriva dal fatto che al tendere verso zero del valore del parametro di lisciamento si avrà infatti una stima della densità molto frastagliata mentre al crescere di tale valore la stima risulterà eccessivamente liscia e tenderà a non rilevare alcune caratteristiche, quali l'eventuale multimodalità, della funzione di densità.

In letteratura sono stati analizzati diversi metodi per scegliere il valore del parametro di lisciamento basati ad esempio sul riferimento ad una determinata famiglia parametrica di distribuzioni o su convalida incrociata; per una rassegna più completa di tali metodi si vedano ad esempio Wand e Jones (1994), Bowman e Azzalini (1997).

Per valutare la bontà della stima ottenuta si considera un criterio che valuti la discrepanza tra lo stimatore utilizzato, $\hat{f}$, e la vera densità f; a tal scopo sono state studiate diverse misure. Le principali sono l'*errore quadratico medio* (*MSE*) e l'*errore quadratico medio integrato* (*MISE*) definiti rispettivamente come:

$$\begin{aligned} MSE_x(\hat{f}(x)) &= E\{(\hat{f}(x) - f(x))^2\} \\ MISE(\hat{f}) &= E \int \{\hat{f}(x) - f(x)\}^2. \end{aligned} \tag{3.2}$$

Mentre l'*MSE* cerca di stabilire la bontà dello stimatore in un singolo punto, il *MISE* mira a valutarne le prestazioni sull'intero spazio campionario. Di più semplice trattazione da un punto di vista matematico, è la versione asintotica di quest'ultimo (*AMISE*). Mediante uno sviluppo in serie di Taylor del secondo ordine tale misura

si dimostra essere pari a:

$$AMISE(\hat{f}) = \frac{h^4}{4}\sigma_k^4 \int f''(x)^2 dx + \frac{1}{nh} \int K(z)^2 dz.$$

L'*AMISE* permette inoltre di vedere, tramite i rapporti di proporzionalità di questa misura rispetto al valore del parametro di lisciamento h, come la scelta di tale parametro sia importante per poter ottenere una buona stima della funzione di densità. Si nota infatti come, se per minimizzare il primo addendo sarebbe sufficiente scegliere $h \to 0$, tale scelta andrebbe a determinare un incremento del secondo addendo: poichè si dimostra che tali quantità sono legati a distorsione e varianza dello stimatore, risulta chiaro come ci si trovi a dover affrontare quindi il ben noto *trade-off* tra varianza e distorsione.

Il metodo del nucleo è facilmente generalizzabile al caso multivariato, situazione in cui sarà utilizzato in questo lavoro.
Dato un campione di n osservazioni $X = \{x_1, \ldots, x_n\}$ con la generica osservazione $x_i = (x_{i1}, \ldots, x_{id}) \in \mathbb{R}^d$, per $i = 1, \ldots, n$ lo stimatore basato sul metodo del nucleo, in ambito multivariato, viene definito come:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_H(x - X_i), \tag{3.3}$$

dove $K(\cdot)$ è in questo caso una funzione *kernel* d-variata mentre H è una matrice quadrata di dimensione d, simmetrica e definita positiva detta matrice di lisciamento. Anche nel caso multivariato l'attenzione va maggiormente posta sulla scelta dei parametri di lisciamento di tale matrice. L'introduzione infatti di una matrice di lisciamento, se da un lato porta ad un guadagno in termini di flessibilità, dall'altro introduce una maggiore complessità vista la presenza potenziale di $d(d+1)/2$ valori differenti da determinare. Per superare tale complessità spesso vengono adottate parametrizzazioni diagonali di tale matrice tali che $H = Diag(h_1^2, \ldots, h_d^2)$ o $H = h^2 I$ con I matrice identica di ordine d; queste semplificazioni saranno anche adottate nel seguito in questa tesi.

Alcuni metodi di selezione automatica del parametro di lisciamento validi nel caso univariato, possono essere utilizzati, con le dovute modifiche, anche nel caso multivariato. In generale tali metodi risultano comunque più complessi e questo è il motivo per cui, nel caso in cui si operi in spazi a più dimensioni, siano stati meno studiati e per cui si adottano frequentemente parametrizzazioni che mirano a ricondursi a situazioni più semplici.

3.3 Individuazione operativa dei gruppi

A fronte di un comune punto di partenza nella definizione di gruppo quale regione ad alta densità, da un punto di vista operativo il *clustering* modale si è sviluppato seguendo due differenti direzioni le quali presentano alcune ovvie analogie legate all'utilizzo di tecniche di stima non parametriche ma con alcune sostanziali differenze nel modo in cui si cerca di giungere ad una partizione dei dati. Il primo approccio si basa sulle curve di livello della densità mentre il secondo si basa sulla ricerca delle mode della distribuzione. In questo e nel prossimo paragrafo si parlerà nello specifico di due differenti metodi appartenenti a queste differenti categorie. Per una rassegna ed una discussione più dettagliata riguardo i metodi di *clustering* modale presenti in letteratura si veda Menardi (2015).

3.3.1 Metodi basati sulle curve di livello della densità

L'approccio basato sulle curve di livello della densità segue quanto proposto da Hartigan (1975) il quale ha introdotto un concetto di *cluster* legato agli insiemi di livello connessi della funzione di densità. Tale approccio quindi non collega i gruppi direttamente alle mode della distribuzione, ma ne fornisce una definizione in termini di regioni ad alta densità dello spazio campionario, definite per l'appunto dalle curve di livello.

Formalmente, fissata una soglia $\lambda > 0$, si definisce:

$$L(\lambda; f) = L(\lambda) = \{x \in \mathbb{R}^d : f(x) > \lambda\} \qquad (0 \leq \lambda \leq max f) \tag{3.4}$$

la regione dello spazio campionario la cui densità è superiore al livello λ fissato. Al variare di tale parametro, $L(\lambda)$ può essere una regione connessa o sconnessa e ogni sua componente connessa includerà almeno una moda della funzione di densità. Essendo $L(\lambda)$ ignoto, una sua stima $\hat{L}(\lambda)$ si ottiene sostituendo ad $f(x)$ in (3.4) una sua stima non parametrica $\hat{f}(x)$; usualmente si fa ricorso ad una stima *kernel* della densità sebbene tale scelta non sia vincolante.

Poichè in generale non è garantito che esista un valore λ capace di identificare tutte le regioni modali, λ viene fatto variare nell'intervallo dei suoi possibili valori: si viene quindi a creare una gerarchia nota come *albero dei cluster*.
In conclusione l'*albero dei cluster* conta il numero di componenti connesse di $L(\lambda)$ al variare di λ. Le foglie di tale albero rappresentano le mode della distribuzione e identificano il numero di gruppi. Si veda figura 3.1 per un'illustrazione semplice nel caso univariato.

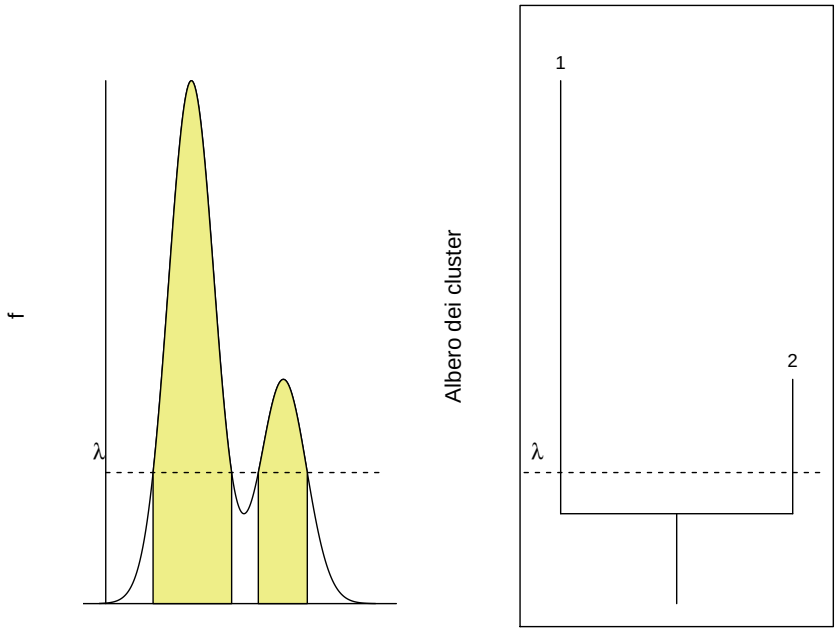

Figura 3.1: Esempio di distribuzione univariata bimodale con rispettivo *albero dei cluster*

Uno dei maggiori problemi che hanno limitato l'utilizzo di questo tipo di approccio all'analisi di raggruppamento riguarda il fatto che trovare le componenti connesse corrispondenti ad una data $\hat{L}(\lambda')$ risulta essere computazionalmente e concettualmente semplice solamente nel caso unidimensionale dove queste componenti sono degli intervalli. Nel caso multidimensionale, per individuare tali componenti in letteratura si è usualmente fatto ricorso alla teoria dei grafi.

Un grafo G è un modo per rappresentare le relazioni tra coppie di osservazioni utilizzando un insieme di vertici collegati tra loro da un insieme di archi. Un grafo viene detto connesso qualora si riesca sempre ad individuare un percorso che colleghi due differenti vertici; qualora il grafo non sia connesso, ogni sottografo connesso identifica una componente connessa. È quindi diretto il riferimento alla teoria dei grafi per l'individuazione delle componenti connesse di $L(\lambda)$. Si indichi infatti con $G(\lambda)$ il sottografo G associato ad una soglia λ, tale sottografo sarà quindi composto da quei vertici aventi densità superiore a λ collegati qualora rispettino una determinata condizione. In questo modo le componenti connesse di $G(\lambda)$ forniscono un'approssimazione di quelle di $L(\lambda; f)$ e, al variare di λ, si riesce a ottenere l'*albero dei cluster*.

Le maggiori differenze tra i metodi esistenti in letteratura che propongono approcci al *clustering* non parametrico basato sulle curve di livello della densità riguardano i differenti modi proposti per individuare le componenti connesse.

In questo lavoro si fa riferimento, e verrà quindi utilizzato nelle analisi successive, quanto proposto da Azzalini e Torelli (2007). Gli autori, per superare il problema di cui si è parlato, ricorrono alla *triangolazione di Delaunay*, duale della *tassellatura di Voronoi*. Dato un insieme $S = \{x_1, \ldots, x_n\}$, con $x_i \in \mathbb{R}^d$, la *tassellatura di Voronoi* definisce una partizione dello spazio $\mathbb{R}^d$ formata da n poliedri $V(x_1), \ldots, V(x_n)$ tali che un generico punto x apparterrà all'insieme $V(x_i)$ se x_i è l'elemento di S più vicino ad x. A partire dalla *tassellatura di Voronoi* due punti x_i e x_j vengono collegati con un arco qualora i corrispettivi elementi $V(x_i)$ e $V(x_j)$ condividano una porzione di una faccia del poliedro. In questo modo si viene a formare un grafo, detto *triangolazione di Delaunay*.

Operativamente, data una stima $\hat{f}(x)$ della funzione di densità $f(x)$, ottenuta dal campione a disposizione S, si considera la quantità $\hat{L}(\lambda)$ definita in precedenza e in particolare se ne considera una restrizione

$$S(\lambda; \hat{f}) = S(\lambda) = \{x_i \in S : \hat{f}(x_i) > \lambda\} \qquad (0 \leq \lambda \leq max\hat{f}) \tag{3.5}$$

Si opera su tale restrizione in quanto lo scopo non è quello di partizionare l'intero spazio $\mathbb{R}^d$ ma è quello di fornire un raggruppamento delle osservazioni appartenenti al campione in esame. Dopo aver costruito la *triangolazione di Delaunay* per individuare le componenti connesse di $S(\lambda)$, per ogni λ, vengono rimossi dal grafo tutte quei vertici $x_i \notin S(\lambda)$ e tutti gli archi che condividono almeno uno di questi vertici. In questo modo si otterranno dei gruppi di osservazioni connesse dagli archi rimasti; tali gruppi costituiscono quindi le componenti connesse di $S(\lambda)$. Queste operazioni andrebbero idealmente ripetute per tutti quei livelli λ tale che $0 < \lambda < max\hat{f}$; a livello pratico si considera solamente una griglia finita di valori per λ. Così si ottiene quindi l'*albero dei cluster*.

Si fa notare infine come tale metodo fornisca M insiemi di punti che vengono detti *nuclei dei cluster* mentre allo stesso tempo vi sono delle osservazioni, sulle code della distribuzione o in corrispondenza della valle tra due mode, per le quali non viene fornita un'etichetta di appartenenza ad un gruppo. È quindi necessario, dopo aver trovato tali M insiemi, utilizzare un metodo di classificazione per allocare le osservazioni non etichettate. Azzalini e Torelli (2007) propongono, come idea generale, di calcolare la densità stimata $\hat{f}_j(x_0)$, con x_0 generica osservazione non allocata, basata sui punti già assegnati ad un determinato gruppo j, per $j = 1, \ldots, M$ e successivamente di assegnare x_0 al gruppo per cui è massima la quantità $r_j(x_0) = \hat{f}_j(x_0)/max_{k \neq j}\hat{f}_k(x_0)$.

3.3.2 Metodi basati sulla ricerca delle mode

Le tecniche appartenenti a questo tipo di approccio all'analisi di raggruppamento mirano a trovare direttamente le mode della densità dei dati a disposizione per poi associare ogni singola osservazione alla moda di pertinenza.

Le principali metodologie di questo tipo sfruttano procedure per la ricerca dei punti di ottimo della funzione di densità quali, ad esempio, il *mean-shift clustering* (Fukunaga e Hostetler, 1975). L'idea di fondo della procedura consiste nel muovere ciascuna osservazione lungo il percorso ascendente del gradiente della densità, fino a convergere ad un punto di massimo, ovvero una moda.

In questo lavoro si concentra l'attenzione su quanto proposto da Li, Ray e Lindsay (2007), un algoritmo chiamato *modal EM* (*MEM*) che riprende in parte l'algoritmo EM adattandolo ad un contesto di ricerca delle mode in ambito non parametrico. L'idea alla base della procedura è che uno stimatore *kernel* è un modello mistura a n componenti. Selezionando ad esempio un *kernel* gaussiano la stima *kernel* viene così definita

$$\hat{f}(x) = \sum_{i=1}^{n} \frac{1}{n} N(x|x_i, \Sigma)$$

dove $N(x|x_i, \Sigma)$ è la funzione di densità di una variabile casuale normale con media uguale a x_i e struttura di covarianza descritta dalla matrice Σ che in questo caso contiene i parametri di lisciamento utilizzati per la stima della densità. Si è fatto finora riferimento ad un *kernel* gaussiano non per motivi sostantivi ma perchè, oltre ad esser la scelta più frequentemente adottata, in questo caso porta a dei vantaggi sotto l'aspetto computazionale; ciò non toglie che l'algoritmo possa essere utilizzato anche nel caso in cui si considerino *kernel* appartenenti a differenti famiglie distributive.

A differenza dell'algoritmo EM, capace di stimare i parametri di una mistura massimizzando la sua verosimiglianza, lo scopo dell'algoritmo *MEM* è quello di trovare massimi locali, e quindi le mode, della distribuzione.

Formalmente, sia $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$, con $x \in \mathbb{R}^d$ e dove π_k è la probabilità a priori della k-esima componente della mistura e $f_k(x)$ è la densità di tale componente. Dato un punto iniziale $x^{(0)}$, l'algoritmo trova i massimi locali della funzione di densità $f(x)$ espressa in termini di mistura, alternando i seguenti due passi fintanto che non viene raggiunto un determinato criterio di convergenza. Partendo con $r = 0$:

- Sia $p_k = \frac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}$, k=1,...,K.
- Aggiorna $x^{(r+1)} = \underset{x}{\text{argmax}} \sum_{k=1}^{K} p_k log f_k(x)$.

Nel primo dei due passi riportati viene calcolata, dato il punto $x^{(r)}$ in cui ci si trova, la probabilità a posteriori di ogni componente della mistura. Il secondo passo è invece il passo di massimizzazione, in analogia con quanto previsto dall'algoritmo EM. Viene assunto generalmente che $\sum_{k=1}^{K} p_k log f_k(x)$ abbia un unico massimo; tale assunzione viene rispettata nel caso in cui le componenti della mistura $f_k(x)$ abbiano una distribuzione Gaussiana. Gli autori dimostrano come questo algoritmo, al raggiungimento di un determinato criterio di convergenza, fornisca un valore $x^{(r')}$, con r' ultima iterazione dell'algoritmo, che risulta essere un massimo locale della distribuzione. Si noti l'algoritmo sfrutti una proprietà valida solamente nel caso in cui si vogliano trovare i punti di massimo locale di una distribuzione espressa in termini di un modello mistura.

Il raggruppamento dei dati viene ottenuto mediante la seguente procedura:

1. Viene stimata la funzione di densità, in seguito indicata come $\hat{f}(x|S, \sigma^2)$ usando il metodo del nucleo e una matrice di lisciamento $H = \sigma^2 I_d$;

2. $\hat{f}(x|S, \sigma^2)$ viene utilizzata come densità per i due passi dell'algoritmo *MEM* riportato in precedenza. Ogni singola osservazione $x_i, i = 1, \ldots, n$ viene usata come valore iniziale dell'algoritmo con lo scopo di trovare le mode di $\hat{f}(x|S, \sigma^2)$. La moda identificata partendo da x_i viene indicata con $M_\sigma(x_i)$;

3. Viene formato un insieme G contenente i valori distinti delle mode trovate al passo precedente;

4. Se $M_\sigma(x_i)$ è uguale al k-esimo elemento in G, si considera x_i come appartenente al k-esimo *cluster*.

L'algoritmo quindi "muove" ciascun valore x_i verso una moda e i punti iniziali che portano a una stessa moda vengono considerati come appartenenti allo stesso gruppo. L'algoritmo di *clustering* così definito viene detto *Mode Association Clustering* (*MAC*).

Gli autori sottolineano come tale algoritmo presenti alcuni vantaggi rilevanti rispetto all'approccio parametrico all'analisi di raggruppamento: i principali sono legati alla possibilità di ottenere una funzione di densità specifica per ogni singolo *cluster* e alla maggiore robustezza del metodo qualora vengano violate le assunzioni parametriche.

Anche questo algoritmo, come quelli basati sulle curve di livello della densità, fornisce una struttura di raggruppamento gerarchica seppur di diverso tipo. È noto infatti che, all'aumentare del parametro di lisciamento, la densità tende a diventare via via più liscia fino a coprire eventuali strutture multimodali e facendo sì che,

da qualsiasi punto iniziale x_i si parta, si raggiunga la medesima moda. È possibile quindi definire una gerarchia considerando una sequenza crescente di parametri di lisciamento. Nel caso estremo in cui tale parametro tenda a 0 si avrà che l'insieme di mode corrispondente sarà uguale all'insieme $S = \{x_1, \ldots, x_n\}$.

3.4 Gestione della dimensionalità

Una delle criticità alla quale si va incontro utilizzando metodi di *clustering* non parametrico e alla quale si è cercato di fornire una risposta nel capitolo successivo è legata ai problemi che vanno sotto il nome di *maledizione della dimensionalità*.

Il concetto di *maledizione della dimensionalità* si riferisce a quella serie di criticità che, paradossalmente, si incontrano nel momento in cui si cerca di estrarre informazione rilevante da un numero elevato di variabili. L'aggiunta di variabili implica un incremento delle dimensioni dello spazio matematico associato e tale incremento comporta una maggiore dispersione dei dati all'interno di questo spazio con conseguenti difficoltà riguardanti il processo di stima o, più in generale, il cogliere una struttura all'interno dei dati stessi (per maggiori approfondimenti si veda, ad esempio, Hastie, Tibshirani e Friedman, 2009, paragrafo 2.5). Questo fenomeno assume aspetti particolarmente critici nel momento in cui, come in questo lavoro, si utilizzino metodi statistici non parametrici. La maggiore flessibilità che si riesce ad ottenere con questi metodi è dovuta al fatto che, non imponendo una determinata struttura al problema, si lascia che siano i dati a parlare; ovviamente questo, nel caso in cui si abbia un numero elevato di variabili, è particolarmente complicato a causa della sparsità dello spazio in cui si lavora.

Riportandosi alla particolare situazione in cui si opera in questa tesi risulta ovvio come la stima non parametrica della densità risenta di questo problema per via del fatto che la massa di probabilità tende a presentarsi nelle code della distribuzione e che all'aumentare del numero di dimensioni aumenta anche il numero di parametri di lisciamento richiesti.
In modo più formale si può mostrare (Wand e Jones, 1994) che, quando si utilizza ad esempio lo stimatore del nucleo (3.3), anche ammettendo una parametrizzazione della matrice di lisciamento del tipo $H = h^2 I$ che riduce al minimo il numero di

parametri da selezionare, il minimo $AMISE$ ottenibile al variare di h, sia:

$$\inf_{h>0} AMISE\{\hat{f}(\cdot;h)\} = \\ = \frac{d+4}{4d}\left(\mu_2(K)^{2d}\{dR(K)\}^4\left[\int\{\bigtriangledown^2 f(x)\}^2 dx\right]^d n^{-4}\right)^{\frac{1}{(d+4)}}. \tag{3.6}$$

Si noti come il tasso di convergenza della (3.6) sia dell'ordine di $n^{-\frac{4}{(d+4)}}$ e come quindi diventi più lento all'aumentare delle dimensioni. Questo rallentamento del tasso di convergenza è una chiara manifestazione della maledizione della dimensionalità e può peggiorare radicalmente la bontà dello stimatore basato sul metodo del nucleo in dimensioni elevate.
A causa della sparsità, per poter includere nel procedimento di stima una quantità sufficiente di osservazioni, i parametri di lisciamento devono assumere valori elevati che renderebbero impossibile il cogliere comportamenti locali della densità. Fondamentalmente ci si trova in una situazione in cui si risente del *paradosso della vicinanza*; qualora si considerino regioni dello spazio locali queste risultano esser vuote e, per contro, qualora queste non siano vuote non si stanno prendendo in considerazioni regioni locali. Tale paradosso altro non è che un modo di riportare a questo contesto il noto compromesso tra distorsione e varianza. Scott (2015) mostra alcuni esempi nei quali vengono evidenziate nella pratica le difficoltà legate a questo paradosso e si vede quindi come, congiuntamente all'aumentare delle dimensioni, la numerosità campionaria debba crescere esponenzialmente per garantire precisione costante dello stimatore in termini di $MISE$.

Per quanto detto finora usualmente si ritiene che lo stimatore basato sul metodo del nucleo risulti di scarsa utilità qualora si operi in più di cinque dimensioni. Tale considerazione deve però esser contestualizzata all'obiettivo specifico dell'analisi. Alcune applicazioni hanno mostrato (Li, Ray e Lindsay, 2007) come nella pratica lo stimatore riesca a cogliere la struttura nei dati anche in spazi di dimensione superiore a condizione di accettare che vi sia un'accuratezza globale generalmente inferiore.

Un ulteriore aspetto rilevante e che va affrontato in contesti multidimensionali riguarda il fatto che l'informazione di interesse giaccia, di fatto, in un sottospazio di dimensione inferiore rispetto a quello dei dati. Scott (2015) sostiene che "la presenza di una diminuzione di rango nei dati multivariati, più che l'elevata dimensionalità di per sé, è la componente più importante della maledizione della dimensionalità".

Per risolvere questo problema è quindi importante disporre di tecniche che selezionino alcune variabili o che permettano di ottenere nuove variabili a partire dalle

originali; lo scopo di tali tecniche deve essere dunque quello di permettere idealmente di operare in quel sottospazio in cui giacciono le strutture di interesse presenti all'interno dei dati. In tale direzione quindi va quanto sarà proposto nel capitolo successivo dove si è cercato di studiare il comportamento di alcune procedure per selezionare le variabili contenenti del segnale e che permettono quindi di lavorare in sottospazi che non solo sono esplorabili con maggiore facilità ma che dovrebbero mantenere al proprio interno le strutture di interesse.

Capitolo 4

Un approccio non parametrico globale al problema di individuazione di anomalie collettive

4.1 Formalizzazione del problema

In questo capitolo vengono introdotte alcune metodologie proposte per superare le criticità, che si incontrano nell'ambito dell'analisi di raggruppamento non parametrica, alle quali si è accennato nel capitolo precedente. Lo scopo è quindi quello di contestualizzare tali problematiche e di proporre delle soluzioni per superarle rifacendosi all'ambito semi-supervisionato in cui si opera ed utilizzando le informazioni aggiuntive che si hanno a disposizione.

Le ipotesi alla base dei metodi che saranno proposti, riprendendo in parte quanto assunto da Vatanen et al. (2012) nel loro lavoro, sono le seguenti:

- La distribuzione di *background* è nota o quantomeno stimabile arbitrariamente bene;
- La distribuzione di *background* è unimodale;
- La distribuzione di *background* è stazionaria;
- Le anomalie si presentano in maniera collettiva come un picco non presente nella distribuzione di *background*;
- Il segnale da individuare, seppur raro, si presenta in una proporzione tale da permetterne l'individuazione.

Si noti come tali assunzioni siano tutte contestualizzabili all'ambito fisico, dal quale questo lavoro prende spunto, nel quale il segnale è usualmente raro e si presenta

come un picco non noto nella distribuzione di *background* la quale è assunta nota e descrivibile tramite il *Modello Standard*.

Nei paragrafi successivi si presentano inizialmente due differenti tecniche di riduzione della dimensionalità che permettono, selezionando le variabili contenenti segnale e quindi rilevanti ai fini dell'analisi, di operare in spazi ridotti agevolando così l'interpretazione e l'utilizzo di tecniche non parametriche. Successivamente si propone un metodo *ad hoc* per selezionare la matrice di lisciamento in maniera ottimale condizionatamente al contesto semi-supervisionato in cui si lavora. Infine viene illustrata una modifica dell'algoritmo *MEM* che ha la finalità di tener conto, nell'analisi di raggruppamento, delle informazioni che si hanno a disposizione relative al *background*.

Si introduce ora la notazione che si è adottata nel resto di questo capitolo e nel successivo. Si denotano con $\mathcal{X}_b = \{x_1, \ldots, x_{n_b}\}$ e $\mathcal{X}_{bs} = \{x_1, \ldots, x_{n_{bs}}\}$ i campioni a disposizione che si suppongono generati rispettivamente dalla distribuzione del *background* f_b e da quella che potenzialmente include un segnale f_{bs}, con f_b e f_{bs} : $\mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$. Specificatamente ciascuna delle x_i ($i = 1, \ldots, n_b$ o $i = 1, \ldots, n_{bs}$) sono assunte essere realizzazioni i.i.d. delle variabili casuali d-dimensionali $X_b = (x^{(1)}, \ldots, x^{(j)}, \ldots, x^{(d)})$ e $X_{bs} = (x^{(1)}, \ldots, x^{(j)}, \ldots, x^{(d)})$. Qualora si faccia riferimento a delle stime non parametriche della densità ($\hat{f}_b$ e $\hat{f}_{bs}$) ottenute con il metodo del nucleo con H_b e H_{bs} si indicano le rispettive matrici di lisciamento utilizzate.

4.2 Stima non parametrica della densità

Nel capitolo precedente, nell'introdurre lo stimatore non parametrico basato sul metodo del nucleo, si è parlato di come il problema principale nella procedura di stima sia legato alla scelta del parametro di lisciamento o, più in generale nel caso in cui si operi in ambito multivariato, della matrice di lisciamento. Si è sottolineato come la scelta di tale matrice risulti particolarmente critica in quanto influenza pesantemente i risultati che si ottengono con lo stimatore *kernel*: una scelta errata di tali parametri può infatti coprire completamente alcune strutture di interesse nei dati. Per questi motivi in letteratura sono stati proposti diversi criteri di selezione automatica di questi parametri ai quali si è brevemente accennato.

In questo lavoro si è ritenuto che, muovendosi in un contesto semi-supervisionato dove quindi si hanno a disposizione alcune informazioni aggiuntive sul comportamento assunto dai dati, fosse possibile sfruttare tali informazioni per ottenere un criterio *ad hoc* di scelta dei parametri di lisciamento. Come evidenziato in precedenza, la distribuzione sottostante i dati non è nota a causa della possibile presenza di un segnale che si manifesta mediante un picco di densità che emerge dalla distribuzione

del *background*. D'altra parte, essendo quest'ultima nota con un margine di errore arbitrariamente piccolo, ed essendo i dati osservati provenienti da essa almeno in larga misura, è ragionevole assumere che la distribuzione di interesse sia, almeno sotto certi aspetti, riconducibile a quella del *background*.

Su questa idea generale si basa la procedura proposta di selezione del parametro di lisciamento. In particolare, coerentemente con l'assunto tipico del *clustering* non parametrico secondo il quale ogni gruppo è associato ad una moda della distribuzione sottostante i dati, la selezione del parametro di lisciamento avverrà in modo tale da garantire che rimanga invariata la moda che rappresenta il *background* che è nota.

A questo scopo, si sfrutta la proprietà secondo la quale il gradiente di una qualsiasi funzione risulta essere pari a zero qualora venga calcolato in un punto di massimo locale. Tale considerazione può essere ovviamente riportata anche nel caso in cui si consideri una funzione di densità nella quale quindi i massimi locali rappresentano le mode di tale densità.

Si consideri l'usuale espressione dello stimatore *kernel* in un generico caso multidimensionale (3.3) dove, per ora, si assume che H sia genericamente una matrice $d \times d$ simmetrica e definita positiva e dove si considera un kernel $K(\cdot)$ qualunque; nelle analisi si è usualmente utilizzato un kernel gaussiano ma tale scelta non risulta in nessun modo vincolante. È facile verificare che il gradiente di $\hat{f}(x)$ calcolato nel generico punto x sia un vettore $\hat{f}(x)'$ di componenti:

$$\frac{\partial \hat{f}(x)}{\partial x^{(j)}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial K_H(x - X_i)}{\partial x^{(j)}}, \quad j = 1, \ldots, d. \tag{4.1}$$

Sia ora M la moda individuata nella distribuzione f_b del *background*. Si individua $H = H_{bs}$ in modo che

$$\hat{f}_{bs}(M)' = \frac{1}{n} \sum_{i=1}^{n} K'_{H_{bs}}(M - X_{i_{bs}}) = 0. \tag{4.2}$$

Così facendo è possibile ottenere quindi un criterio per identificare i parametri di lisciamento in modo tale che gli stessi garantiscano la nullità del gradiente nella moda della funzione di densità f_b stimata con il metodo del nucleo facendo sì che M sia un ottimo locale di $\hat{f}_{bs}$.

Nel particolare ambito in cui ci si sta muovendo si suppone innanzitutto di avere due distinti insiemi di dati: uno proveniente dalla distribuzione di *background* mentre per il secondo non si hanno informazioni in tal senso ed è quindi possibile che contenga del segnale. Operativamente quindi si procede determinando le coordinate della moda del *background*: nelle analisi successive si è fatto ricorso al *MEM* per coerenza con quanto fatto nel resto del lavoro, ma tale scelta non è vincolante e per tale

scopo potrebbe essere utilizzato un qualsiasi algoritmo di ricerca delle mode. Una volta ottenuta l'informazione riguardo il massimo locale di f_b si procede ponendo uguale a zero il gradiente dello stimatore *kernel* della densità ottenuto sul campione $\mathcal{X}_{bs}$, calcolato in tale massimo locale. Si ottiene così un sistema di equazioni che permette di calcolare i parametri di lisciamento per stimare non parametricamente la densità f_{bs} condizionatamente al fatto che tale stima deve aver gradiente nullo nella moda del *background*.
In questo modo si giunge ad avere una stima della densità sull'insieme di dati potenzialmente contenente del segnale vincolata ad alcune caratteristiche note della distribuzione di *background*. La stima non parametrica così ottenuta è poi il punto di partenza per utilizzare uno degli algoritmi di *clustering* di cui si è parlato nel capitolo precedente.

Si ritiene infine opportuno fare alcune precisazioni riguardo al metodo di selezione dei parametri di lisciamento appena presentato. Innanzitutto si è fatto riferimento ad una situazione multidimensionale ma ovviamente quanto detto vale anche nel caso in cui si operi in una singola dimensione. D'altro canto, lavorare in uno spazio a più dimensioni, introduce alcune problematicità. Si noti infatti come quanto riportato nella (4.2) altro non sia che un sistema in d equazioni. Per poter essere in una situazione determinata è quindi necessario che in tale sistema siano presenti d incognite; questo fa sì che, in caso di unimodalità del *background*, la matrice di lisciamento H non possa esser considerata come una matrice piena ma vada considerata la parametrizzazione $H = (h_1^2, \ldots, h_d^2)I_d$. Si ritiene comunque che questo non sia un problema particolarmente rilevante in quanto questa parametrizzazione della matrice H è una scelta comune qualora si voglia utilizzare il metodo *kernel* per la stima non parametrica della densità in un contesto multidimensionale.

Più rilevante potrebbe risultare il problema legato alla presenza di un numero di mode, nella distribuzione di *background*, pari a m con $m > 1$. In questo caso infatti si otterrebbero $d \times m$ equazioni, e non vi sono dunque garanzie che il problema abbia soluzione o che questa sia unica. In questo lavoro si assume una distribuzione di *background* unimodale, coerentemente con le informazioni note sul *background* nel contesto fisico di riferimento.
Si noti infine come sarebbe opportuno porre un ulteriore vincolo sull'hessiano della funzione per avere la certezza di lavorare con dei punti di massimo e non di minimo; in questo caso non è però necessario in quanto si considera f_b unimodale.

4.3 Individuazione operativa dei gruppi

Nel capitolo precedente sono stati introdotti alcuni metodi di *clustering* non parametrico basati sulla ricerca delle mode della distribuzione di densità. Ci si è particolarmente sofformati sull'algoritmo *MEM* (Li, Ray e Lindsay, 2007), una variante dell'algoritmo EM che anzichè individuare i punti di massimo della funzione di verosimiglianza, come accade tipicamente nel *clustering* parametrico, individua i punti di massimo di una funzione di densità e associa a tali punti la rispettiva regione di attrazione che corrisponde a uno specifico *cluster*. L'algoritmo alterna iterativamente due passi con lo scopo, dato un punto iniziale $x^{(0)}$, di trovare il percorso che da tale punto porta ad una moda della distribuzione. Una condizione necessaria al fine di poter adeguatamente utilizzare questo algoritmo riguarda il fatto che la distribuzione della quale si ricercano i punti di massimo deve essere espressa in termini di un modello a mistura con un numero di componenti finito. Gli autori hanno sfruttato tale condizione per adattare questo algoritmo di ricerca delle mode all'ambito del *clustering* modale, notando come lo stimatore *kernel* della densità sia di fatto un modello mistura con n componenti.

Si è quindi avvertita la necessità di trovare un modo di modificare l'algoritmo *MEM* in modo tale da permettere di sfruttare in maniera più efficiente le informazioni a disposizione. L'idea esplorata in questo lavoro consiste nel cercare le mode non della stima non parametrica della densità ottenuta attraverso il metodo del nucleo ma di una sua modifica che tenga conto della distribuzione del *background*.
La densità della quale si ricercano i punti di massimo locale viene allora espressa in termini di una mistura di misture:

$$\begin{aligned} \hat{f}_A(x) &= \pi_1 f_b(x) + (1-\pi)\hat{f}_{bs}(x) \\ &= \pi_1 f_b(x) + (1-\pi)\left(\frac{1}{n_{bs}}\sum_{i=1}^{n_{bs}} K_{H_{bs}}(x - X_{i_{bs}})\right). \end{aligned} \tag{4.3}$$

Dall'espressione (4.3) è immediato notare che, esprimendo la densità in questo modo, si continua ad utilizzare un modello a mistura finita e questo rende possibile l'utilizzo dell'algoritmo *MEM*. A questo punto quindi si ricercano i punti di massimo locali della densità in (4.3) e successivamente viene fornito un raggruppamento delle sole osservazioni appartenenti al campione $\mathcal{X}_{bs}$. Operando in questo modo si ritiene che si stia utilizzando in maniera completa l'informazione a disposizione e che quindi ci si muova in un contesto semi-supervisionato imponendo un vincolo all'algoritmo in modo tale che tenga conto, nella ricerca dei massimi, anche delle caratteristiche della distribuzione di probabilità del *background*.

Si noti come, nel caso in cui anche f_b non sia nota ma venga di fatto stimata mediante il metodo del nucleo (seppure con arbitraria accuratezza) la (4.3) sia, per specifiche scelte di π, una stima *kernel* della densità ottenuta a partire da tutti i dati disponibili (sia $\mathcal{X}_{bs}$ che $\mathcal{X}_b$).

Alcune prove hanno mostrato come questa modifica dell'algortimo *MEM* porti a dei risultati che risentono effettivamente di un vincolo determinato dalle coordinate della moda del *background*. La procedura quindi in generale vincola l'algoritmo a determinare una moda pressochè uguale a quella individuata nella distribuzione di *background* anche qualora venga applicata al campione del quale non si hanno informazioni ed etichette di classe.

4.4 Gestione della dimensionalità

4.4.1 Una procedura semisupervisionata per la selezione di variabili

Nel capitolo precedente ci si è soffermati sull'evidenziare come l'elevata dimensionalità dello spazio in cui si opera introduca una serie di problematicità che peggiorano le prestazioni delle metodologie di *clustering* e come sia un problema particolarmente rilevante nel caso in cui si faccia riferimento a tecniche non parametriche. La soluzione più comune per risolvere questo tipo di problemi è quella di ridurre la dimensione dello spazio, prima di applicare procedure di *clustering*.
Sebbene in questo lavoro ci si muova in un contesto di apprendimento semi supervisionato, sembra opportuno notare, viste le affinità a livelli di problematiche, alcune difficoltà che si incontrano nei problemi non supervisionati nel ridurre la dimensionalità dello spazio. Non avendo informazioni su un'eventuale relazione asimmetrica tra variabili in un contesto non supervisionato risulta infatti più problematico sviluppare dei criteri per valutare l'importanza di una variabile per il raggiungimento di un determinato obiettivo. Un esempio immediato riguarda l'utilizzo dell'*analisi delle componenti principali* (per una trattazione completa e recente si veda ad esempio Jolliffe, 2002). È stato infatti evidenziato come, sebbene siano spesso utilizzate anche in un contesto non supervisionato, non vi sia nessuna garanzia che le componenti associate agli autovalori maggiori contengano informazione utile ai fini di un'eventuale analisi di raggruppamento (Chang, 1983). È quindi importante tenere in considerazione, nel momento in cui si riduce la dimensionalità, delle specifiche caratteristiche del contesto in cui si opera cercando di tenerne conto nelle metodologie che si utilizzano.

Per quanto detto, si riscontra dunque la necessità di operare una riduzione della dimensionalità che sia funzionale agli obiettivi che si vogliono raggiungere, ovvero in modo da mantenere, o se possibile far risaltare ulteriormente, l'eventuale segnale.

Usualmente le due strade principali per ottenere una riduzione della dimensionalità consistono nel selezionare delle variabili considerate rilevanti secondo qualche criterio, o nel crearne delle nuove a partire da quelle originali cercando di concentrarne l'informazione. In questo lavoro si è preferito adottare il primo tipo di approccio in quanto permette una maggiore interpretabilità sia dello spazio in cui si opera sia dei risultati che si possono in seguito ottenere. Facendo brevemente riferimento al contesto pratico dal quale prende spunto questo lavoro risulta ovvio come la possibilità di selezionare esclusivamente quelle variabili che contengono del segnale porti ad un grande vantaggio in termini di interpretazione del fenomeno fisico sottostante.

Nel seguito quindi verrà presentato un approccio generale al raggiungimento di tale obiettivo, e verranno discusse alcune implementazioni alternative finalizzate alla selezione di variabili aventi un comportamento potenzialmente differente tra f_b e f_{bs}.

Sebbene con motivazioni alla base differenti la procedura proposta presenta alcune analogie con le *foreste casuali* (Breiman, 2001) in quanto entrambi i metodi sono iterativi e selezionano un sottoinsieme di variabili differenti ad ogni passo.

L'idea di fondo della procedura è quella di confrontare la distribuzione del *background*, supposta nota, con la distribuzione stimata sulla totalità dei dati, che potenzialmente includono la componente del segnale, limitatamente a diversi sottoinsiemi di variabili.

Più in dettaglio, ad ogni iterazione, la procedura si sviluppa nei seguenti passi:

1. Si selezionano $k < d$ variabili casualmente;

2. Si confronta la distribuzione delle k variabili su $\mathcal{X}_b$ con quella delle stesse su $\mathcal{X}_{bs}$, secondo uno specifico criterio (descritto nei paragrafi 4.4.2 e 4.4.3). Avendo assunto che il *background* sia stazionario, tale differenza nel comportamento, qualora vi fosse, sarebbe un'indicazione di presenza di segnale all'interno delle variabili esaminate;

3. Se il confronto rivela difformità tra le due distribuzioni, per ciascuna delle k variabili viene aggiornato un contatore.

Alla fine del ciclo di iterazioni si disporrà di un conteggio che, per ogni singola variabile, indica quante volte tale variabile è stata considerata assumere un comportamento anomalo. L'assunzione che si fa è che, qualora i criteri utilizzati riescano a cogliere adeguatamente l'eventuale presenza di segnale, tale conteggio tenderà ad evidenziare quali variabili siano realmente rilevanti.

Una variante di questo modo di procedere consiste nel selezionare le k variabili, al passo 1, in accordo a un vettore p di probabilità che ad ogni iterazione viene aggiornato a seconda che le variabili selezionate siano risultate discriminanti o meno al passo precedente. L'aggiornamento avviene assumendo che il contatore provenga da una distribuzione multinomiale e facendo ricorso alla stima di massima verosimiglianza.
Si è fatto questo ritenendo che, qualora i criteri utilizzati riescano a discriminare bene tra i due diversi insiemi di dati, tale procedura permetterà, all'aumentare del numero di iterazioni, di focalizzarsi principalmente sulle variabili realmente contenenti segnale in quanto tenderà a selezionarle con maggiore frequenza nei sottoinsiemi di lunghezza k.

I criteri esplorati per il confronto tra f_b e f_{bs} verranno presentati in seguito; si fa comunque notare come tutti questi criteri facciano ricorso a tecniche non parametriche di stima.
La selezione delle k variabili ad ogni passo ha la finalità di agevolare l'applicazione di tali metodi riportandosi ad operare in sottospazi di dimensione inferiore. Infine si noti come tutti i criteri utilizzati presuppongano un certo grado di conoscenza del *background*. Tale considerazione si inserisce nel contesto di analisi semi-supervisionata in cui si sta lavorando e in cui quindi si è cercato di tratte vantaggio dalle informazioni a disposizione nell'utilizzo delle procedure.

4.4.2 Metodo basato su rilevazione di multimodalità

Una delle assunzioni fatte da Vatanen et al. (2012), e ripresa in questo lavoro, è che le anomalie si presentino all'interno del dominio del *background* in zone ad elevata densità e che quindi non siano considerabili valori anomali nel senso comune del termine ma che diventino anomali qualora se ne consideri il comportamento a livello di gruppo. Nella situazione in cui si opera le eventuali anomalie costituiscono il segnale che si cerca di individuare e si assume quindi che si presenti come una nuova moda non rilevata nella distribuzione di *background* nota.

L'idea alla base del primo criterio proposto per valutare la rilevanza delle variabili è quello di confrontare il numero delle mode della distribuzione di *background* con quello delle distribuzione di cui non conosciamo la classificazione. Nella figura 4.1 viene illustrato il comportamento che dovrebbe assumere una variabile contenente segnale in un semplice esempio.

Esistono molti approcci differenti allo studio di eventuale multimodalità nell'ambito delle densità stimate non parametricamente. Silverman (1981), ad esempio, fa riferimento al teorema per il quale il numero di mode di una stima della densità

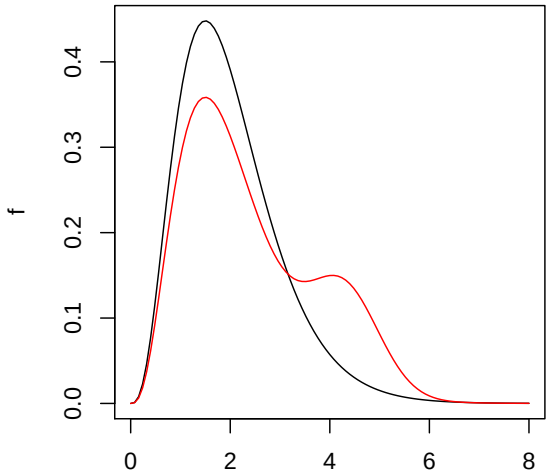

Figura 4.1: Esempio di funzioni di densità, f_b e f_{bs}, di una variabile rilevante per l'individuazione del segnale

non parametrica è una funzione decrescente del parametro di lisciamento h. Viene definita quindi *l'ampiezza di banda critica k-esima*, h_{crit}, che viene utilizzata come misura per testare la presenza di multimodalità; è infatti sufficiente stimare la densità dei dati non etichettati utilizzando un parametro di lisciamento scelto seguendo qualche criterio di ottimalità e confrontare tale parametro con h_{crit} con $k = 1$.
In Hartigan e Hartigan (1985) viene proposto un test che utilizza la *statistica dip*, la massima distanza tra la funzione di ripartizione empirica F_n e la più vicina funzione di ripartizione F appartenente alla classe di tutte le distribuzioni unimodali U. Si è in presenza di multimodalità nel caso in cui *dip* assuma un valore grande.
Hartigan e Mohanty (1992) propongono il cosiddetto *RUNT test*. Questo test inserisce lo studio della multimodalità nel contesto dell'analisi di raggruppamento. Tale approccio fornisce una struttura gerarchica dove ogni *cluster*, se composto da due o più osservazioni, si divide in un certo numero di sotto-gruppi. Si associa quindi ad ogni *cluster* C il numero di punti $n(C)$ nel suo più piccolo sotto-gruppo (o *"runt"*). La statistica test viene quindi definita come $\max_C n(C)$.

Altri metodi studiano se, qualora si valuti la densità stimata nel segmento che collega due diverse mode, si riscontri la presenza di un'anti-moda (si veda ad esempio Burman e Polonik, 2009).

Poichè è usuale pensare alle mode di una distribuzione come a dei *cluster* all'interno della stessa, risulta conveniente riformulare il problema in questione in termini

di analisi di raggruppamento seppur con opportuni accorgimenti finalizzati a tener conto che il contesto di studio è in realtà semi-supervisionato. Nel nostro caso si conosce la vera distribuzione dei dati nel caso in cui ci si trovasse in una situazione di assenza di segnale. Questo criterio assume quindi che la distribuzione di *background* sia unimodale e quindi presenti un unico *cluster*; alla luce di questo, una variabile si ritiene contenga del segnale qualora, valutatane la distribuzione sui dati per i quali non si dispone di una classificazione, venisse rilevata una situazione di multimodalità. In termini pratici, riportandosi alla struttura riportata nel paragrafo precedente, una variabile viene considerata rilevante qualora sia stata selezionata in un sottoinsieme di k variabili in cui il numero di *cluster* risulta essere maggiore di uno.

4.4.3 Metodo basato su test di verifica d'ipotesi

Il secondo criterio esplorato per la selezione di variabili consiste nell'applicazione di un test statistico per il confronto tra distribuzioni. In generale il sistema di ipotesi, tenendo in considerazione di quanto detto, può essere espresso come:

$$\begin{cases} H_0 : f_b = f_{bs} \\ H_1 : \overline{H_0} \end{cases} \tag{4.4}$$

Un test, applicato in questo caso come criterio discriminante nella procedura descritta nel paragrafo 4.4.1, dovrà quindi essere in grado di cogliere eventuali differenze tra le due densità marginalmente rispetto alle variabili selezionate ad ogni iterazione, e giudicare come rilevanti quelle variabili che hanno condotto al rifiuto dell'ipotesi nulla.

Vale la pena ricordare che parlando dei diversi approcci al problema della rilevazione delle anomalie, si è visto come sia immediato porre tale problema in termini di test di verifica d'ipotesi. Nelle situazioni in letteratura in cui la rilevazione di anomalie è stata affrontata come una verifica d'ipotesi tale obiettivo è semplificato dal fatto che spesso vengono considerate anomalie di tipo puntuale le quali si presentano in regioni del supporto campionario a bassa densità.

Nel contesto in cui si colloca questo lavoro l'eventuale presenza di anomalie collettive, che indicherebbe una presenza di segnale, costringe invece a considerare il test di verifica d'ipotesi in termini di un confronto tra due diverse densità.

Per quanto concerne lo specifico test, l'attenzione è stata concentrata, per coerenza, sull'uso di test non parametrici. Ad esempio Hall (1984) ha sottolineato come, da un punto di vista pratico, una statistica test naturale per verificare se una certa

densità f possa esser realmente la densità dalla quale provengono i dati è fornita dall'errore quadratico integrato (ISE)

$$I = \int (\hat{f} - f)^2, \tag{4.5}$$

dove $\hat{f}$ è una stima non parametrica di f. Nella situazione in cui si opera, tale test prevederebbe di confrontare la vera funzione di densità del *background* supposta nota con la stima non parametrica ottenuta utilizzando i dati non etichettati. Un approccio di questo tipo risulta essere particolarmente adatto al contesto in esame nel quale si vuole testare se i dati per i quali non si dispone di una classificazione possano provenire dalla densità di *background* o se presentino un comportamento sostanzialmente differente. Tale differenza, viste le assunzioni fatte, darebbe un'indicazione riguardo la presenza del segnale che si sta cercando.
Tale approccio è stato inizialmente preso in considerazione ma successivamente scartato in quanto si ritiene che possa essere limitante considerare che l'informazione sui dati di *background* riguardi la perfetta conoscenza della funzione di densità dalla quale provengono gli stessi. Questa situazione potrebbe infatti non esser del tutto verosimile nel caso in cui si operi su dati reali; l'informazione riguardo la densità di *background* potrebbe essere infatti differente e riguardare il fatto di conoscere alcune caratteristiche di tale densità o il fatto di poter simulare a piacere dalla stessa.

Si è deciso quindi di analizzare in seguito alcuni metodi che presuppongano una conoscenza meno completa riguardo la distribuzione di *background* in modo tale che il metodo proposto possa avere una maggiore flessibilità ed applicabilità in diverse situazioni.

Vanno in tale direzione le considerazioni fatte da Anderson, Hall e Titterington (1994) i quali propongono la statistica test

$$T_{h_1 h_2} = \int (\hat{f}_1 - \hat{f}_2)^2, \tag{4.6}$$

dove, per $j = 1, 2$, $\hat{f}_j$ è una stima non parametrica della densità basata sul j-esimo campione utilizzando il parametro di lisciamento h_j. Gli autori propongono di sfruttare la distribuzione *bootstrap* di tale quantità per poter ottenere una conclusione riguardo questa verifica d'ipotesi.

Duong, Goud e Schauer (2012) prendono spunto da quanto proposto da Anderson, Hall e Titterington (1994) cercando allo stesso tempo di costruire una statistica test che abbia una distribuzione nota sotto l'ipotesi nulla e che permetta quindi di non dover ricorrere a delle tecniche di verifica d'ipotesi basate su ricampionamento. Si suppone quindi di avere due insiemi di dati d-dimensionali $Z_1, \ldots, Z_{n_1}$ e $Y_1, \ldots, Y_{n_2}$

provenienti dalle rispettive funzioni di densità f_1 e f_2. Utilizzando questi due differenti dataset si ottengono, utilizzando il metodo del nucleo, le stime $\hat{f}_1(z; H_1)$ e $\hat{f}_2(y; H_2)$ dove H_l, con $l = 1, 2$, è una matrice di lisciamento. Per testare l'ipotesi nulla $H_0 : f_1 = f_2$ gli autori riprendono quindi la misura di discrepanza proposta da Anderson, Hall e Titterington (1994) in (4.6).
Questa misura di discrepanza viene poi riscritta come $T = \psi_1 + \psi_2 - (\psi_{1,2} + \psi_{2,1})$ dove $\psi_l = \int f_l(z)^2 dz$ e $\psi_{l_1,l_2} = \int f_{l_1}(z) f_{l_2}(z) dz$ ovvero

$$\begin{aligned}
\psi_1 &= \frac{1}{n_1^2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_1} K_{H_1}(Z_{i_1} - Z_{i_2}) \\
\psi_2 &= \frac{1}{n_2^2} \sum_{j_1=1}^{n_2} \sum_{j_2=1}^{n_2} K_{H_2}(Y_{j_1} - Y_{j_2}) \\
\psi_{1,2} &= \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{H_1}(Z_i - Y_j) \\
\psi_{2,1} &= \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} K_{H_2}(Z_i - Y_j).
\end{aligned} \tag{4.7}$$

La statistica test T può essere interpretata come una comparazione tra una differenza a coppie dentro al campione $Z_{i_1} - Z_{i_2}$ e $Y_{i_1} - Y_{i_2}$ e una differenza a coppie tra campioni $Z_i - Y_j$: nel caso in cui quest'ultima sia maggiore della precedente questo fornisce un indicazione sul fatto che i due campioni sono differenti.

Sotto l'ipotesi nulla $H_0 : f_1 = f_2$ la quantità $\frac{T - \mu_T}{\sigma_T \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \xrightarrow{d} N(0, 1)$, dove $\mu_T = [n_1^{-1}|H_1|^{-1/2} + n_2^{-1}|H_2|^{-1/2}]K(0)$ e $\sigma_T^2 = 3[\int f(z)^3 dx - (\int f(z)^2 dz)^2]$.
Per poter ricorrere alla distribuzione nulla asintotica è necessario quindi avere una stima dei parametri μ_T e σ_T^2. Chacón e Duong (2010) hanno presentato un algoritmo per ottenere stimatori consistenti delle matrici di lisciamento H_1 e H_2 in modo tale che minimizzino l'errore quadratico medio asintotico e sia possibile una stima di μ_T. Pongono inoltre $\hat{\sigma_T}^2 = (n_1 \hat{\sigma_1}^2 + n_2 \hat{\sigma_2}^2)/(n_1 + n_2)$.

È stato quindi utilizzato questo test come criterio per valutare l'eventuale difformità tra *background* e campione non etichettato. Per ricondursi ad una situazione di analisi semi-supervisionata, e sfruttare quindi la conoscenza che si possiede sul comportamento assunto dai dati appartenenti al campione $\mathcal{X}_b$, si è deciso di stimare in maniera accurata la matrice di lisciamento relativa a questo determinato campione utilizzando un numero di dati molto elevato. Si ritiene così di riuscire a tenere in considerazione della conoscenza che si ha a disposizione senza però vincolarsi alla conoscenza della specifica funzione di densità. La decisione di ricondursi in questo modo ad una situazione semi-supervisionata è stata fatta nella consapevolezza di

utilizzare in maniera meno stringente l'informazione a disposizione ma si è ritenuto che così facendo si renda maggiormente applicabile il metodo a diverse situazioni in cui potrebbe non essere scontato conoscere la esatta distribuzione di densità del *background*.

4.4.4 Discussione critica

In questo paragrafo si procede quindi sottolineando e giustificando alcune scelte compiute per quel che riguarda i metodi proposti nei paragrafi precedenti.

Innanzitutto si noti che la procedura di selezione delle variabili ad ogni iterazione opera su un numero di variabili k, con $k < d$ dove d è il numero totale di variabili a disposizione. Questa scelta è stata fatta al fine di tenere in considerazione le indicazioni, di cui si è parlato anche in precedenza e riportate da diversi autori (si veda ad esempio Wand e Jones, 1994), riguardanti le difficoltà incontrate dalla stima della densità basata sul metodo del nucleo nel caso in cui si operi in uno spazio di dimensione superiore a cinque. Tale decisione va giustificata inoltre facendo riferimento alla volontà di rimanere in ambito multidimensionale in modo tale che, sebbene lo spazio venga ridotto, rimanga l'opportunità di cogliere l'eventuale struttura presente a livello di relazioni tra diverse variabili.

Da un altro punto di vista si è consapevoli che il fatto di selezionare ad ogni iterazione un sottoinsieme di k variabili introduce alcune criticità. È ovvio infatti come, qualora nel sottoinsieme selezionato siano presenti congiuntamente delle variabili realmente rilevanti e delle variabili non rilevanti, i criteri utilizzati tenderanno erroneamente a considerare come contenenti del segnale tutte le k variabili prese in considerazione in quella determinata iterazione. Tale situazione sarà presumibilmente molto frequente in quanto si assume che il numero di variabili rilevanti sia inferiore rispetto al numero di variabili che non presentano traccia di segnale. Si ritiene perciò che tale errore sia inevitabile ma allo stesso tempo è lecito assumere che l'aleatorietà dello stesso lo renda ininfluente, dopo un numero adeguato di iterazioni della procedura di selezione.

Per quanto concerne il metodo di selezione basato sulla rilevazione di multimodalità un'eventuale critica che potrebbe essere avanzata riguarda il fatto che, basandosi su di una stima di densità non parametrica, tale stima potrebbe presentare delle mode spurie in particolar modo sulle code della distribuzione. A tal proposito si son fatte delle prove, in diversi contesti per quanto riguarda distribuzione dei dati e dimensionalità, per testare la capacità del metodo di individuare una singola moda qualora la distribuzione dalla quale son stati generati i dati sia effettivamente unimodale; si è concluso che il metodo risulta cogliere in maniera sufficientemente

adeguata tale struttura. Inoltre, in questa fase dell'analisi, si ritiene questo possa essere un problema marginale in quanto si è preferito essere anti-conservativi dando maggior peso ad un'eventuale incapacità nel cogliere l'importanza di una variabile contenente del segnale rispetto alla situazione opposta.

Infine, per quanto riguarda il metodo di selezione delle variabili basato sul test d'ipotesi, si è consapevoli del fatto che, poichè si esegue un test ad ogni iterazione della procedura di selezione delle variabili, ci si trova in una situazione di test multipli e quindi risulterebbe necessario correggere i valori dei p-value ottenuti dall'applicazione del test sopra presentato. Gli autori hanno però sottolineato che il test risulta essere conservativo; tende cioè ad accettare l'ipotesi nulla più spesso rispetto al livello di confidenza nominale. Tale indicazione è stata poi ulteriormente confermata da alcuni studi simulativi che hanno mostrato come, utilizzando anche differenti tipi di correzione per i p-value, il test tenderebbe ad accettare l'ipotesi nulla anche in situazioni in cui il segnale, e quindi la differenza tra i due campioni, risulti abbastanza evidente.
Si è quindi deciso di utilizzare tale procedura senza correzioni dei p-value ottenuti; tale decisione può inoltre essere giustificata dal fatto che l'obiettivo di questi test non è realmente quello di confrontare le distribuzioni, quanto piuttosto quello di selezionare le variabili più discriminanti. Si preferisce essere anti-conservativi in modo da ottenere innanzitutto delle indicazioni che potranno poi in seguito essere analizzate più dettagliatamente con altre metodologie applicabili nello spazio ridotto delle variabili selezionate.

Capitolo 5

Un'esplorazione numerica

5.1 Alcune considerazioni a partire da uno studio di simulazione

5.1.1 Obiettivi dello studio

Lo scopo di questo paragrafo è quello di analizzare, mediante uno studio di simulazione, il comportamento dei metodi di selezione delle variabili quando ci si trova ad operare in un contesto di analisi semi-supervisionata dove solo alcune variabili contengono delle anomalie, che in questo contesto rappresentano il segnale.

Nello studio si son dovute operare delle scelte per circoscrivere il problema e analizzarlo in alcune situazioni particolari; si è tuttavia cercato di mantenersi in situazioni verosimili tentando inoltre di rifarsi esplicitamente allo specifico problema fisico dal quale questo lavoro prende spunto.

Le domande alle quali si è voluto rispondere e gli obiettivi che ci si prefigge sono di seguito elencati:

- Valutare la capacità dei due approcci presentati nei paragrafi 4.4.2 e 4.4.3 di selezionare le variabili contenenti segnale al variare della quantità dello stessp;
- Valutare la capacità dei due approcci di selezionare le variabili contenenti segnale al variare della separazione tra il picco del segnale e quello nella distribuzione di *background*;
- Valutare analogie ed eventuali differenze, per entrambi gli approcci, tra la variante basata su campionamento di variabili equiprobabile ad ogni passo dell'algoritmo e la variante che prevede l'aggiornamento della probabilità di selezione;

- Valutare analogie e differenze tra l'approccio basato su rilevazione di multi-modalità e quello basato su test di verifica d'ipotesi mettendone in luce pregi e difetti in relazione allo specifico obiettivo per il quale sono stati utilizzati.

5.1.2 Descrizione degli scenari

Visto quanto detto nei capitoli precedenti e dovendo riportarsi ad operare in un contesto di analisi semi-supervisionata, si sono generati, ad ogni passo di simulazione, due differenti campioni. ll primo insieme rappresenta i dati di *background* $\mathcal{X}_b$, per i quali si può assumere di essere a conoscenza del fatto che non sono presenti delle anomalie, mentre il secondo campione $\mathcal{X}_{bs}$ costituisce l'insieme di dati non etichettato nei quali quindi potrebbero essere presenti dei comportamenti anomali rispetto al *background*.

Per riuscire ad operare in una situazione verosimile rispetto a quella che ci si aspetta di trovare nel particolare contesto applicativo reale al quale si fa riferimento, si è deciso per il *background* di simulare da una distribuzione d-variata nella quale sia presente un certo grado di asimmetria ($sk \in [0.5, 1)$) e una moderata correlazione tra le variabili generate ($cor \in [0.10, 0.35]$). Per fare questo è stata considerata la distribuzione *normale asimmetrica canonica fondamentale* ($CFUSN$) presentata in Arellano-Valle e Genton (2005) cui si rimanda per i dettagli.
I dati di *background* $\mathcal{X}_b$ vengono quindi generati da una variabile casuale X_b tale che

$$\begin{aligned} X_b &\sim f_b(x) \\ \text{con } \ f_b &= CFUSN_d(\mu_1, \Sigma_1, \Delta) \end{aligned} \tag{5.1}$$

con Σ_1 tale da far sì che le variabili siano omoschedastiche e correlate con grado di correlazione specificato in precedenza.

I dati dei quali non si dispone di classificazione $\mathcal{X}_{bs}$ sono invece stati generati da una variabile casuale multivariata X_{bs} tale che

$$\begin{aligned} X_{bs} &\sim f_{bs}(x) = (1-\pi) f_b(x) + \pi f_s(x) \\ \text{con } \ f_s &= N_d(\mu_2, \Sigma_2) \end{aligned} \tag{5.2}$$

con $\mu_2 = \mu_1 + dist$ e $dist$ è un vettore non nullo nelle componenti contenenti segnale mentre Σ_2 è una matrice che mantiene la correlazione tra le variabili invariata rispetto ai dati di *background* mentre la varianza, pur rimanendo omogenea tra le diverse dimensioni, risulta essere quattro volte inferiore rispetto a quella della

distribuzione di *background* delineata da Σ_1. Questa scelta è stata fatta per fare in modo che l'eventuale segnale risulti nei dati come una nuova moda, più o meno evidente sulla base di π e di *dist*, solo nelle variabili nelle quali $\mu_1 \neq \mu_2$ mentre per le altre variabili la presenza di una determinata proporzione di dati provenienti da f_s non modifica la struttura fondamentale che rimane quella assimilabile ad una distribuzione asimmetrica multivariata e unimodale.

I *setting* di simulazione sono stati fatti variare con riferimento alla proporzione π di segnale presente nei dati e alla distanza (*dist*) di questo dalla media della distribuzione multivariata di *background*.

Si discute ora dei parametri che variano nei diversi scenari di simulazione. Si sono considerate tre diverse situazioni $\pi = (0.10, 0.05, 0.01)$. Tale decisione è giustificata da quanto fatto da Vatanen et al. (2012), i quali hanno fatto variare la proporzione dallo 0.01 allo 0.20, e inoltre ciò ha permesso di rimanere in una situazione quantomeno verosimile all'ambito applicativo della fisica delle particelle nel quale, qualora fosse presente, la quantità di segnale sarebbe molto sbilanciata rispetto a quella dei dati che riflettono il normale comportamento del *background*.

Per il parametro *dist* relativo alla distanza del segnale dalla media della distribuzione di *background* si sono simulate due differenti situazioni: per le variabili contenenti segnale si ha $dist = 2$ nel primo caso e $dist = 3$ nel secondo, per le altre variabili $dist = 0$. Le due diverse situazioni sono state considerate per cercare di studiare se ed eventualmente come variano le prestazioni dei metodi di selezione utilizzati nel momento in cui vari il grado di separazione tra la principale moda della distribuzione e la moda generata dalla presenza di dati anomali rispetto al comportamento normale degli stessi. Ci si aspetta che i metodi incontrino una maggiore difficoltà nell'individuazione del segnale nel caso in cui la separazione sia inferiore. Si noti inoltre che, pur prendendo in considerazione questi due differenti scenari, le anomalie sono state generate in entrambi i casi in modo tale che possano essere quantomeno compatibili con il dominio del *background*; questa decisione è stata presa per rimanere in un ambito di rilevazione di anomalie collettive, delle quali si è parlato nei capitoli precedenti.

Nei risultati che saranno presentati in seguito sono state analizzate le prestazioni dei metodi implementati anche nel caso in cui non si sia in presenza di segnale ($\pi = 0$); tale analisi è stata condotta come controllo, per cercare di capire se il modo in cui sono stati generati i dati avesse in qualche modo distorto la procedura di selezione facendo sì che venissero selezionate determinate variabili anche qualora non ci fosse segnale da rilevare.

Sono stati invece considerati fissati la numerosità campionaria ($n = 3500$), il numero di variabili generate ($d = 30$), il numero di variabili contenenti del segnale

($s = 5$), la numerosità del sottoinsieme sul quale si opera ad ogni iterazione ($k = 4$) e la forma distributiva.

Per quel che riguarda i due metodi di selezione delle variabili vanno fatte alcune precisazioni:

- Si ricorda innanzitutto che, per quanto concerne il metodo basato sulla rilevazione di multimodalità, ci si è ricondotti ad operare in ambito semi supervisionato assumendo che i dati di *background* abbiano distribuzione unimodale (si son infatti generati in tal senso) e si son quindi considerate come rilevanti quelle variabili che presentano un numero di *cluster* maggiore o uguale a due. Per quel che riguarda invece il metodo basato su verifica d'ipotesi ci si è riportati a lavorare in condizioni semi-supervisionate stimando, prima di applicare il test, la matrice di lisciamento asintoticamente ottimale per la distribuzione di *background*;

- Entrambi i metodi presentano alcune limitazioni dal punto di vista computazionale. Per questo motivo si è dovuto, ad ogni passo dell'algoritmo presentato nei capitoli precedenti, selezionare non solamente $k = 4$ variabili su cui operare ma anche un numero n' di osservazioni con $n' < n$ ($n' = 500$ per il metodo basato su rilevazione di multimodalità e $n' = 1000$ per il metodo basato su test d'ipotesi). Tale scelta ha alla base solamente motivazioni di ordine computazionale ed è stata compiuta nella consapevolezza che operare con un numero ridotto di osservazioni può portare ad un peggioramento delle prestazioni delle procedure di selezione considerate. D'altro canto si ritiene comunque di interesse analizzare tali prestazioni tenendo anche in considerazione che i problemi computazionali emersi nelle simulazioni, nelle quali quindi l'algoritmo di selezione dev'esser testato un numero elevato di volte, potrebbero essere molto meno rilevanti nell'analisi di dati reali;

- Per ogni scenario sopra delineato, qualora si sia utilizzato il metodo basato su rilevazione di multimodalità, è stato generato un numero di campioni pari a 1000, di cui 500 provenienti da f_b e 500 provenienti da f_{bs}, e per ogni coppia di insiemi di dati l'algoritmo ha compiuto $M = 100$ iterazioni. Qualora invece si sia fatto riferimento al metodo basato sul test di verifica d'ipotesi, problemi di ordine computazionale hanno portato alla necessità di ridurre a 600 il numero di campioni generati, 300 da f_b e 300 da f_{bs}. Anche in questo caso il numero di iterazioni è stato posto pari a $M = 100$. La scelta di ridurre il numero di campioni simulati sui quali testare questo metodo è stata fatta tenendo in considerazione, in seguito ad alcune prove, che in questo caso risulta essere preferibile utilizzare un sottoinsieme di osservazioni n' più elevato rispetto a

quanto preso inizialmente in considerazione: la scelta di $n' = 1000$ ha portato quindi alla necessità di simulare un numero ridotto di campioni;

- Nel caso in cui si sia utilizzato l'aggiornamento multinomiale per quel che riguarda il vettore di probabilità di selezione delle k variabili ad ogni passo le prime, 20 iterazioni di entrambi gli algoritmi sono state considerate di *warm-up* e son servite quindi per ottenere delle successive stime più stabili del vettore di probabilità non portando però ad un aggiornamento dello stesso. Questa scelta è stata fatta in quanto alcune prove preliminari hanno infatti evidenziato un problema legato all'aggiornamento del vettore di probabilità; si è visto infatti come avessero una forte rilevanza i risultati ottenuti nelle prime iterazioni sui risultati finali forniti. L'inserimento di un numero di iterazioni di *warm-up* permette di evitare di considerare come contenente segnale una variabile che nella realtà non ne presenta ma che è stata erroneamente selezionata nei primi passi dell'algoritmo;

- Per valutare la bontà della selezione compiuta dalle procedure, nel caso in cui le k variabili vengano selezionate ad ogni iterazione in maniera equiprobabile, si è deciso di considerare il numero medio di volte, per ogni coppia di campioni, nel quale una variabile è stata considerata rilevante. Nel caso in cui si consideri la procedura con aggiornamento del vettore di probabilità, si è deciso di valutare la stima finale di tale probabilità come misura di bontà della selezione: qualora infatti la procedura riesca in maniera adeguata a selezionare le variabili contenenti anomalie, tale vettore dovrebbe presentare valori più elevati per queste variabili rispetto ai valori relativi alle variabili invarianti nei due diversi insiemi di dati.

Dal punto di vista strettamente operativo gli scenari e i metodi appena delineati sono stati implementati con il linguaggio di programmazione *R* (R Core Team, 2016). Nello specifico, per generare i dati sono stati utilizzati i pacchetti *mvtnorm* (Genz et al., 2008) e *sn* (Azzalini, 2015) mentre per i due metodi di selezione delle variabili studiati si sono utilizzati i pacchetti *pdfCluster* (Azzalini e Menardi, 2014) e *ks* (Duong et al., 2007).

5.1.3 Risultati

I risultati dello studio di simulazione sono riportati nelle tabelle da 5.1 a 5.4.

Si ritiene importante in prima battuta sottolineare come entrambi gli algoritmi di selezione, in entrambi i *setting* di campionamento, non selezionino delle variabili

come rilevanti qualora non siano state generate delle anomalie (e quindi $\pi = 0$). Tale scenario è stato analizzato come controllo per garantire che i metodi studiati non tendano in ogni caso a selezionare determinate variabili.

Per quanto riguarda i risultati ottenuti al variare della quantità di anomalie presenti nei dati si può notare in generale come tale quantità influisca molto, come era lecito aspettarsi, sulle prestazioni dei diversi metodi. Questo comportamento non è sorprendente in quanto è naturale supporre che, nel caso in cui si sia in presenza di un segnale molto debole, i metodi implementati incontrino maggiori difficoltà nel coglierlo non riuscendo di conseguenza ad individuare le variabili più rilevanti.
Questa tendenza generale assume però caratteristiche differenti qualora si analizzino separatamente le performance dei due diversi metodi di selezione. Per quanto riguarda il metodo basato sulla rilevazione di multimodalità questo sembra fornire un'indicazione corretta ed evidente anche qualora la proporzione di segnale presente sia pari a $\pi = 0.05$. Nello scenario simulato avente segnale più debole ($\pi = 0.01$) il metodo non riesce invece a cogliere e a selezionare quelle variabili che contengono delle anomalie.
Per ciò che concerne il metodo basato sulla verifica d'ipotesi, se da un lato le sue prestazioni risentono ancora in maniera evidente della proporzione di segnale generato, dall'altro sembra essere meno soddisfacente sotto questo punto di vista rispetto al metodo precedente. Tale approccio sembra fornire indicazioni chiare qualora la proporzione di anomalie presenti sia pari a $\pi = 0.1$ mentre non sembra cogliere la presenza di segnale qualora il segnale si presenti in proporzioni inferiori.

Per quel che riguarda le performance dei metodi al variare della distanza tra la media del *background* e la media delle anomalie generate si possono notare comportamenti meno omogenei tra i due metodi.
Per il primo metodo la distanza sembra influire negativamente sulle prestazioni dell'algoritmo qualora la proporzione di anomalie sia pari a $\pi = 0.1$ o $\pi = 0.05$ mentre, sebbene la differenza sia lieve, quando $\pi = 0.01$ il metodo sembra avere un comportamento più soddisfacente nel caso in cui la distanza sia inferiore.
Per quanto riguarda la seconda procedura analizzata invece, tenendo conto di quanto detto in precedenza riguardo l'incapacità di fornire indicazioni rilevanti in presenza di segnale debole, sembra risentire in maniera differente della distanza. Si nota infatti che le indicazioni riguardo la presenza di variabili contenenti segnale è più evidente nel caso in cui tale segnale sia posto ad una distanza inferiore. Tale comportamento sembra andare in controtendenza rispetto a quanto sarebbe lecito aspettarsi in quanto si reputa che, generalmente, possa essere più agevole individuare il segnale qualora questo sia ben distinto dalla distribuzione di *background* e si trovi quindi in una regione a densità inferiore. Potrebbe essere d'interesse valutare questo metodo

in più scenari differenti tra loro per via di questa caratteristica.

I commenti fatti finora, sebbene riguardino le prestazioni generali dei due diversi approcci di selezione delle variabili, fanno principalmente riferimento al caso in cui si sia utilizzato un campionamento equiprobabile per estrarre le k variabili sulle quali operare ad ogni iterazione. Nel caso in cui si sia utilizzato l'aggiornamento multinomiale le differenze tra le variabili contenenti segnale e le altre variabili, e di conseguenza le indicazioni di cui si è parlato in precedenza risultano essere, seppur presenti, meno marcate. Bisogna infatti tener conto di come, per ogni coppia di campioni simulati, gli algoritmi abbiano compiuto $M = 100$ iterazioni e come, nella caso di aggiornamento del vettore di probabilità, le prime 20 iterazioni siano state considerate di *warm-up*. Se da un lato si ritiene quindi che il numero di iterazioni totali possa non esser sufficientemente elevato per fornire indicazioni più evidenti, dall'altro si pensa che i risultati ottenuti possano essere comunque promettenti ed indicare che, all'aumentare del numero di iterazioni, si potrebbe giungere ad una situazione nella quale lo schema di selezione delle variabili ad ogni iterazione permette di lavorare nella maggior parte dei casi in un sottospazio che fa riferimento alle variabili realmente contenenti del segnale. Si ritiene quindi che potrebbe essere di interesse valutare le prestazioni degli algoritmi al crescere di M.

Si osservi che il metodo basato sulla rilevazione di multimodalità inoltre opera ad ogni iterazione con una numerosità $n' = 500$ a fronte di una numerosità $n' = 1000$ utilizzata per il metodo basato su verifica d'ipotesi; questo sottolinea ulteriormente il miglior comportamento del primo metodo e permette di ipotizzare che risultati migliori sarebbero ottenibili con tale procedura all'aumentare del valore di n'.
Come detto, per motivi computazionali si è dovuto applicare questa procedura utilizzando un sottoinsieme n' di osservazioni; si fa notare quindi come, con prove differenti, il metodo sia sensibile rispetto alla scelta di tale valore e presenti risultati migliori all'aumentare dello stesso. Potrebbe quindi essere interessante verificarne le prestazioni senza selezionare un sottoinsieme di osservazioni ma utilizzando gli interi campioni di dati.

Alcuni commenti aggiuntivi vanno fatti per quanto riguarda la procedura basata sulla verifica d'ipotesi. Nei capitoli precedenti si era già parlato di come gli autori stessi (Duong, Goud e Schauer, 2012) avessero sottolineato il fatto che questo test risulta essere conservativo. Tale osservazione sembra quindi ripercuotersi nei risultati ottenuti in questo studio di simulazione dove si può generalmente notare come una variabile venga considerata come rilevante un numero molto basso di volte. Questo comportamento è anche il motivo pratico per il quale si è preferito non tener conto di correzioni per i test multipli che avrebbero portato ad una completa incapacità del metodo di evidenziare differenze tra la densità stimata di *background* e la densità

stimata sui dati dei quali non si dispone di una classificazione.
Si ritiene potrebbe essere interessante studiare in maniera più approfondita il comportamento del test proposto da Duong, Goud e Schauer (2012) cercando di capire se ci si possa ricondurre ad operare in una situazione meno conservativa che, nel contesto di questo lavoro, potrebbe permettere di ottenere indicazioni più evidenti per quanto riguarda le variabili contenenti segnale.

Quanto fatto potrebbe risentire di scelte arbitrarie compiute riguardo alcuni parametri quali il numero di variabili da utilizzare da ogni iterazione k, la lunghezza del periodo di *warm-up*. Tale scelte sono state compiute in seguito ad alcune prove e nell'esigenza di fissare questi valori in modo da non aumentare i diversi scenari da analizzare; si è comunque consapevoli dell'arbitrarietà di tali scelte e di come potrebbe essere interessante studiare come cambiano i risultati al variare delle stesse.

	$dist = 3$						$dist = 2$					
	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}
π=0.1	8.45	12.92	13.35	13.00	13.28	13.37	8.26	12.64	13.05	12.69	12.87	13.05
π=0.05	8.75	11.99	12.21	12.36	11.90	12.13	8.23	10.88	11.12	11.14	10.82	11.09
π=0.01	7.77	8.07	7.98	8.21	8.03	7.84	7.82	8.19	8.07	8.29	8.15	7.95
π=0	7.85	8.03	7.95	8.00	7.98	7.58	-	-	-	-	-	-

Tabella 5.1: Numero medio di volte in cui le variabili che non contengono segnale (x_0) e quelle che ne contengono (x_s) sono state considerate rilevanti su 100 iterazioni. Criterio di selezione basato sulla multimodalità (par 4.4.2), campionamento equiprobabile delle variabili ad ogni iterazione.

	$dist = 3$						$dist = 2$					
	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}
π=0.1	0.029	0.052	0.054	0.054	0.055	0.056	0.029	0.053	0.054	0.054	0.054	0.056
π=0.05	0.030	0.050	0.049	0.050	0.047	0.050	0.030	0.049	0.047	0.048	0.047	0.049
π=0.01	0.033	0.035	0.034	0.035	0.035	0.033	0.033	0.036	0.035	0.035	0.037	0.033
π=0	0.033	0.033	0.035	0.034	0.033	0.033	-	-	-	-	-	-

Tabella 5.2: Probabilità media di estrazione delle variabili che non contengono segnale (x_0) e di quelle che ne contengono (x_s) dopo 100 iterazioni. Criterio di selezione basato sulla multimodalità (par 4.4.2), campionamento con aggiornamento multinomiale del vettore di probabilità ad ogni iterazione.

	$dist = 3$						$dist = 2$					
	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}
π=0.1	0.468	0.687	0.683	0.737	0.633	0.767	0.566	0.913	0.893	1.040	0.867	0.957
π=0.05	0.006	0.003	0.010	0.007	0.003	0.007	0.005	0.003	0.007	0.007	0.003	0.003
π=0.01	0.001	0.000	0.000	0.000	0.003	0.000	0.001	0.000	0.000	0.000	0.003	0.000
π=0	0.000	0.000	0.000	0.000	0.000	0.000	-	-	-	-	-	-

Tabella 5.3: Numero medio di volte in cui le variabili che non contengono segnale (x_0) e quelle che ne contengono (x_s) sono state considerate rilevanti su 100 iterazioni. Criterio di selezione basato su verifica d'ipotesi (par 4.4.3), campionamento equiprobabile delle variabili ad ogni iterazione.

	$dist = 3$						$dist = 2$					
	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}	x_0	x_{s_1}	x_{s_2}	x_{s_3}	x_{s_4}	x_{s_5}
π=0.1	0.031	0.043	0.041	0.048	0.044	0.045	0.031	0.045	0.046	0.049	0.047	0.047
π=0.05	0.033	0.033	0.033	0.034	0.033	0.033	0.033	0.034	0.033	0.034	0.034	0.034
π=0.01	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033	0.033
π=0	0.033	0.033	0.033	0.033	0.033	0.033	-	-	-	-	-	-

Tabella 5.4: Probabilità media di estrazione delle variabili che non contengono segnale (x_0) e di quelle che ne contengono (x_s) dopo 100 iterazioni. Criterio di selezione basato su verifica d'ipotesi (par 4.4.3), campionamento con aggiornamento multinomiale del vettore di probabilità ad ogni iterazione.

5.2 Un'applicazione alla fisica delle particelle

5.2.1 Descrizione del problema

Il Modello Standard (MS) è una teoria fisica che descrive tre delle quattro forze fondamentali (interazione forte, elettromagnetica e debole) e tutte le particelle elementari note.

Tale modello si concentra sullo studio delle particelle subatomiche e sullo studio delle interazioni tra queste. Le particelle subatomiche vengono classificate in due diverse famiglie: da una parte i fermioni (quark e leptoni) e dall'altra i bosoni, i quali risultano essere le particelle mediatrici delle interazioni fondamentali. Sebbene la teoria sottostante il MS sia stata in gran parte verificata sperimentalmente, ad oggi il modello non risulta ancora essere sufficientemente completo a spiegare la realtà e quindi la ricerca di nuovi fenomeni fisici non ancora osservati sta proseguendo in tale direzione.

La sede dove i più importanti e principali esperimenti hanno luogo è il CERN, ad oggi il più grande laboratorio al mondo di fisica delle particelle, che comprende l'acceleratore di particelle a forma di anello, noto con il nome di Large Hadron Collider (LHC). All'interno del LHC, in condizioni sperimentali simili a quelle immediatamente successive al Big Bang, vengono lanciati fasci di protoni che viaggiano a velocità vicine a quelle della luce. I protoni vengono fatti collidere in quattro stazioni di rilevamento del LHC: ATLAS, CMS, LHCb e ALICE. In questo caso ci si concentra maggiormente sulla stazione CMS (Compact Muon Solenoid), una struttura che permette di "fotografare" gli eventi immediatamente seguenti alla collisione tra protoni e quindi di raccogliere informazioni riguardo questi eventi e i comportamenti assunti dalle particelle generate dalla collisione. In generale lo studio del comportamento delle particelle generate dalle collisioni non può avvenire tramite osservazione diretta, a causa del loro rapido decadimento, ma avviene attraverso i segnali emessi dalle particelle durante il decadimento stesso, registrati da appositi rilevatori.

L'esperimento di riferimento per le analisi che seguono riguarda l'interazione forte fra i costituenti dei protoni, quarks e gluoni, attraverso lo scambio di gluoni che sono i mediatori dell'interazione forte. In una frazione non trascurabile delle collisioni, si possono osservare getti collimati di particelle cariche e neutre, che sono il risultato della frammentazione di quarks e gluoni emessi con alta energia dal punto della collisione. Tale esperimento è noto produrre due processi fisici distinti: il background, *QCD multijet background*, include eventi in cui i getti energetici provengono da quarks leggeri e gluoni e il segnale, *top pair production*, che corrisponde

alla creazione di coppie top-antitop.
Il quark top è il più pesante dei 6 quarks e la sua fenomenologia è del tutto distinta da quella dei processi di QCD, in quanto la sua alta massa rende possibile l'immediato decadimento, che può dar vita a uno o tre getti energetici. Il decadimento del quark top, che avviene per interazione debole, comporta quasi esclusivamente la creazione di un quark bottom, che produce un getto adronico, e l'emissione di un bosone W. Il bosone W a sua volta decade istantaneamente, il più delle volte in coppie di quarks leggeri che a loro volta sono osservabili come getti di particelle.

A livello sperimentale la differenza fra i due processi sopra descritti consiste principalmente nella maggiore energia dei getti prodotti da decadimenti delle coppie di quark top. La particolarità della struttura dei protoni fa sì che la probabilità di collisioni decresca con l'energia liberata dalle collisioni per interazione forte. Una selezione che richieda la presenza di getti energetici riduce fortemente il *background* di QCD, che tuttavia rimane il processo dominante, che si manifesta in circa l'84% degli eventi, a fronte di un segnale che si verifica nel rimanente 16%. Le caratteristiche più discriminanti fra *background* di QCD e segnale di *top pair production* sono legate alla cinematica dei getti emessi, e alle loro particolarità.

I dati utilizzati nell'applicazione, messi a disposizione dal CMS tramite l'Istituto Nazionale di Fisica Nucleare (sede di Padova), non sono reali ma provengono da una simulazione Monte-Carlo e garantiscono pertanto la conoscenza completa del fenomeno oggetto di studio, relativamente a quali eventi danno luogo al segnale e quali al background. Questo permette di valutare le metodologie proposte nel capitolo precedente.

Si dispone, in particolare, di due insiemi differenti di dati, $\mathcal{X}_b = \mathcal{X}_{QCD}$, relativo esclusivamente al processo di QCD e $\mathcal{X}_{bs} = \mathcal{X}_{QCD+TT}$, contenente anche osservazioni provenienti da un processo di *top pair production*, di numerosità rispettivamente n_b = 20000 e n_{bs} = 10000. Nel secondo insieme, pur avendo a disposizione l'informazione relativamente a quali eventi danno luogo al segnale e quali al *background*, tale informazione verrà usata nel seguito solo allo scopo di valutare i risultati. Nei dati utilizzati il segnale, e quindi le osservazioni relative al processo di creazione della coppia top-antitop, è in una percentuale pari al 16.22%.

Prima di passare alla descrizione delle variabili considerate nell'analisi si ritiene sia importante, per poter introdurre alcune quantità di interesse e utili nel seguito, parlare del sistema di coordinate del CMS (figura 5.1). Per ottenere alcune informazioni riguardo le caratteristiche dei getti generati da una collisione è infatti necessario disporre di un sistema di coordinate la cui origine è il punto di collisione. L'asse x è orientata verso il centro dell'anello del LHC, l'asse y verso l'alto mentre l'asse z è orientata lungo la direttrice del fascio attraverso le montagne della Giura.

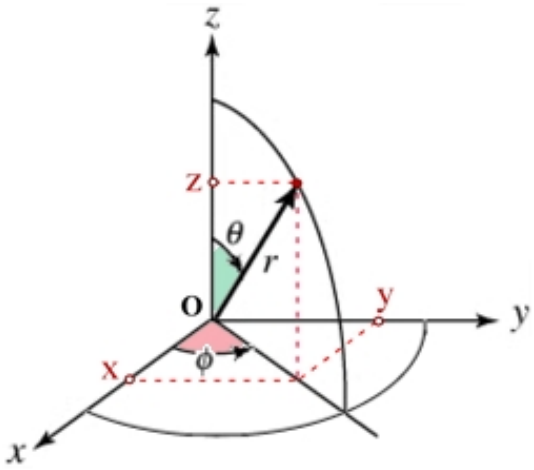

Figura 5.1: Sistema di coordinate del CMS. Il vettore orientato rappresenta un esempio di jet.

L'angolo azimutale ϕ è misurato a partire dall'asse x sul piano xy e r rappresenta la coordinata radiale in questo piano. L'angolo polare θ è definito nel piano rz ma viene usualmente espresso in termini di pseudorapidità $\eta = ln(tan(\theta/2))$. La pseudorapidità assume valore zero nel caso in cui una particella si muova perpendicolarmente alla direzione del fascio mentre assume valore infinito qualora il movimento sia parallelo o antiparallelo rispetto all'asse z. La quantità p_T denota il momento trasverso trasferito nell'urto primario e rappresenta la componente del momento perpendicolare all'asse z.

Avendo definito il sistema di coordinate del CMS si può passare a descrivere le variabili prese in considerazione nelle analisi riportate in seguito. Innanzitutto si elencano le variabili cosiddette di *alto livello* le quali sono comuni tra tutti i getti rilevati dalla collisione o rappresentano una trasformazione di alcune misure rilevate sui getti stessi:

- *njets*: numero di getti adronici, generati dalla collisione, con energia superiore a 20 GeV e pseudorapidità η<2.4;

- *ntags*: numero di getti, tra quelli con le caratteristiche sopra elencate, aventi $csv > 0.8$;

- *metx*: componente x dell'energia trasversa mancante;

- *mety*: componente y dell'energia trasversa mancante;

- *MET*: energia trasversa mancante definita come $MET = \sqrt{metx^2 + mety^2}$, un valore significativamente diverso da zero di tale misura può indicare la

produzione di una o più particelle non interagenti, molto importanti nella ricerca di fenomeni fisici finora non noti;

- *Centralità*: indice che varia tra 0 e 1 definito come $centr = \sum_{jets} p_T/E$. Un valore prossimo ad uno indica una forte dispersione;

- *drmin*: delta R minimo tra coppie di jets. È definito come il quadrato della distanza minima tra jets, operando nel piano definito da pseudorapidità ed azimut;

- *ptfrac*: frazione di p_T dei due jets più energetici rispetto al totale dei primi 6 jets. Si ritiene che nel caso in cui ci si trovi in presenza di segnale, l'energia dovrebbe avere una migliore spartizione tra i jets;

- *dp12*: angolo azimutale tra i primi due jets ordinati rispetto al valore assunto dalla variabile *csv*. Si ritiene che, nella situazione di *background*, l'angolo dovrebbe essere minore;

- *dp23*: angolo azimutale tra il secondo ed il terzo jets ordinati rispetto al valore del loro momento trasverso p_T.

In seguito si procede poi con l'elencare le variabili cosiddette di *basso livello* le quali forniscono informazioni specifiche sui singoli jets. Tali variabili sono quindi disponibili per tutti i jets presi in considerazione.

- p_T: momentro trasverso;

- E_t: energia trasversa, definita come $E_t = E \sin\theta$;

- η: pseudorapidità, come definita in precedenza;

- ϕ: angolo azimutale, come definito in precedenza;

- *csv*: indice che assume valori tra 0 e 1, assume valori più elevati qualora il getto sia generato da un b-quark.

Poichè le coppie top-antitop possono quindi dar luogo a quattro o a sei getti adronici, su suggerimento degli esperti i dati considerati corrispondono a processi di interazione forte in cui si osservano almeno 4 getti energetici. Tra questi, sono stati selezionati per le analisi solamente i quattro jets più energetici, escludendo quindi eventuali getti aggiuntivi rilevati per una determinata collisione; si è compiuto tale scelta nella consapevolezza di ottenere una parziale perdita di informazione rilevante, visto anche quanto detto sui processi fisici che si sono studiati, ma con lo scopo

di garantire una maggiore uniformità tra unità statistiche e una più agevole applicabilità delle metodologie proposte e presentate in questo lavoro. Avendo dunque a disposizione 10 variabili di alto livello e 5 variabili di basso livello, ciascuna delle quali specifica di ognuno dei quattro jets, si dispone in totale di 30 variabili.

Inoltre ci si è concentrati solamente su collisioni con un'energia superiore a 500 GeV: tale scelta è stata compiuta per motivi di disponibilità di un numero adeguato di osservazioni in quanto, pur essendo le più frequenti, le collisioni a bassa energia raramente generano un numero di jets superiore o uguale a 4.

5.2.2 Descrizione dell'applicazione

Poichè il problema in esame si inserisce pienamente in un contesto di individuazione di anomalie collettive, lo scopo delle analisi che seguono è quello di valutare punti di forza e di debolezza dell'approccio semi-supervisionato non parametrico discusso in questo lavoro, mediante la sua applicazione al problema della ricerca di *top pair production* all'interno del processo QCD.

Si riporta di seguito una descrizione delle analisi effettuate, anche in questo caso svolte nell'ambiente di programmazione R (R Core Team, 2016). Si rimanda all'appendice per il codice predisposto per l'implementazione delle principali procedure realizzate.

Selezione delle variabili Si sono applicate le procedure proposte nei paragrafi (4.4.2) e (4.4.3) in entrambi gli scenari di campionamento. Per riuscire ad utilizzare tali procedure si è dovuto apportare alcune modifiche rispetto a quanto detto nel contesto dell'analisi su dati simulati. Innanzitutto non si è potuto far riferimento ad una situazione di unimodalità del *background*: questo ha fatto sì che si sia dovuto, per quanto riguarda la procedura basata su rilevazione di multimodalità, considerare come rilevanti quelle variabili che presentano un numero di mode superiore nel campione $\mathcal{X}_{bs}$ rispetto a $\mathcal{X}_b$. Si noti inoltre che gli algoritmi sono stati iterati 500 volte e si è considerato come *warm-up* un periodo di 75 iterazioni. Son rimaste invariate, rispetto allo studio di simulazione, tutte le scelte riguardanti ulteriori altri parametri.

Procedura non parametrica di individuazione delle anomalie Dopo aver selezionato le variabili si è proseguito utilizzando i metodi proposti nel capitolo precedente per l'individuazione del segnale in ambito semi-supervisionato.
In particolare si è operato in tal modo:

- Si ottiene la matrice di lisciamento per il *background* H_B in modo tale che f_B sia unimodale;
- Si trova la moda della densità stimata, utilizzando tale matrice di lisciamento, di *background*;
- Si ottiene la matrice di lisciamento per stimare adeguatamente la densità f_{BS} relativa cioè sia al processo di *QCD* che di *top-pair production*. Tale matrice viene ottenuta in modo tale che garantisca la nullità del gradiente della densità calcolato nella moda;
- Infine tenendo in considerazione delle due differenti matrici di lisciamento individuate, si utilizza il *MEM* modificato per ottenere un raggruppamento dei dati provenienti da entrambi i processi fisici.

Si noti che per entrambe le matrici di lisciamento, H_B e H_{BS}, è stata presa in considerazione una parametrizzazione $H = h_i^2 I$ con I matrice identica. Si è presa questa decisione per gli evidenti vantaggi computazionali che questa comporta.

Confronto con altre procedure Le analisi condotte hanno avuto poi come obiettivo quello di confrontare quanto proposto con altre procedure presenti in letteratura. In particolare, a scopo comparativo, si son considerati:

- Uso delle componenti principali per ottenere una riduzione della dimensionalità;
- Uso di procedure di *clustering* non parametrico quali la procedura proposta da Azzalini e Torelli (2007) e l'algoritmo *MEM*, non supervisionato;
- Uso della procedura proposta da Vatanen et al. (2012) per individuazione di anomalie collettive in ambito parametrico.

5.2.3 Discussione

Prima di discutere i risultati ottenuti, è opportuno sottolineare, come accade nella maggioranza delle applicazioni reali, che i dati analizzati, di cui si è potuto disporre solo in tempi recenti e dopo aver sviluppato la maggior parte di questo lavoro di tesi, risultano particolarmente complessi e non rispettano diverse assunzioni fatte.
In particolare una criticità rilevante deriva dal fatto che la distribuzione di *background* non è unimodale poichè alcune variabili presentano una evidente multimodalità. Inoltre analisi esplorative hanno mostrato come non solo il segnale sia

relativamente raro ma anche come questo si presenti, per la maggior parte delle variabili considerate, in regioni dello spazio campionario ad elevata densità creando quindi una sovrapposizione evidente con le osservazioni provenienti dal processo di *background*. Questo comportamento rende quindi più complessa l'individuazione del segnale presente nel campione $\mathcal{X}_{bs}$ e non presente nel campione $\mathcal{X}_b$.
Le problematiche appena descritte fanno sì che i risultati, per quanto promettenti sotto alcuni punti di vista, non siano quelli auspicati.

Si consideri, per iniziare, il risultato ottenuto dall'applicazione della procedura parametrica di Vatanen et al. (2012) sui dati aventi dimensione ridotta mediante un'analisi delle componenti principali e riportato in tabella 5.5. In questo caso, per motivi computazionali e per coerenza con il resto del lavoro, sono state considerate le prime quattro componenti principali, contenenti una proporzione di variabilità originale dei dati pari al 31.5%; si ritiene che questo comportamento sia un'ulteriore conferma della complessità dei dati in questione, non facilmente riassumibili in un numero limitato di variabili.
La procedura stima una mistura a 10 componenti gaussiane di cui 9 per descrivere il *background* e una per descrivere il segnale. L'elevato numero di componenti a modellazione del *background* è verosimilmente dovuto all'asimmetria della distribuzione. Inoltre, sebbene sia noto che non esista una corrispondenza tra il numero di componenti di una mistura e il numero di mode della stessa, questa rappresenta un'ulteriore indicazione in merito alla multimodalità del *background*. Osservando una percentuale pari a circa il 26% di osservazioni appartenenti al segnale erroneamente classificate dal modello come provenienti dal *background* si può trarre un'ulteriore indicazione di come le due classi non abbiano un confine di separazione netto e di come i picchi che rappresentano il segnale si manifestino in zone ad elevata densità della distribuzione di *background*.
A dispetto di queste criticità, la procedura nel complesso commette un errore di classificazione pari al 4.6%. È tuttavia doveroso riportare come si sia riscontrato nella pratica un problema di cui è noto soffra l'algoritmo EM (e quindi la sua modifica) che consiste nell'individuazione frequente di massimi locali della funzione di verosimiglianza. Tale comportamento si è manifestato fornendo risultati differenti anche quando applicato più volte al medesimo insieme di dati. Inoltre, si tenga presente che individuare 10 componenti si traduce, di fatto, nell'individuazione di 10 cluster, a fronte dei soli due noti.

Relativamente all'approccio non parametrico, la procedura descritta non risulta efficace nel raggiungimento dell'obiettivo ultimo di individuazione del segnale e, considerata nel complesso, individua un solo gruppo. Per questa ragione risulta conveniente analizzare singolarmente i tre passi proposti di riduzione della dimensio-

	$Comp_1$	$Comp_2$	$Comp_3$	$Comp_4$	$Comp_5$	$Comp_6$	$Comp_7$	$Comp_8$	$Comp_9$	$Comp_{10}$
Bkg_t	560	1795	662	728	971	989	918	330	1385	38
Sgn_t	32	25	41	154	23	23	56	34	34	1202

Tabella 5.5: Risultati ottenuti utilizzando la procedura parametrica di individuazione delle anomalie proposta da Vatanen et al. (2012). Con Bkg_t e Sgn_t vengono indicate le vere classi di appartenenza.

nalità, selezione del parametro di lisciamento e individuazione delle regioni modali, evidenziandone i pregi e le criticità.

- Con riferimento alla riduzione della dimensionalità le due procedure di selezione delle variabili presentate nei paragrafi 4.4.2 e 4.4.3 sono state eseguite in entrambe le varianti di campionamento. A fini interpretativi i risultati, illustrati nella figura 5.2, sono estramemente promettenti: l'indicazione generale riguarda il fatto che le variabili selezionate includono quelle che gli esperti di fisica ritengono avere un più elevato potere discriminante per l'individuazione del segnale, ad esempio le variabili di *alto livello* e *jcsv*.

  Le indicazioni ottenute risultano generalmente meno nette rispetto a quanto visto per lo studio di simulazione in quanto le procedure producono degli indicatori di rilevanza che decrescono gradualmente. D'altra parte tale comportamento rappresenta una conferma sia di quanto detto in precedenza riguardo alla complessità dei dati in questione, sia di quanto sostenuto dagli esperti del settore che ritengono che la maggior parte delle variabili possa avere un certo potere discriminante. L'unica eccezione è la procedura basata sulla rilevazione di multimodalità, nella variante con aggiornamento delle probabilità di selezione (in basso a sinistra nella figura 5.2) che seleziona nettamente la sola variabile *je3* ma potrebbe essere il risultato di una scelta troppo limitata del periodo di *warm-up*. Sebbene infatti alcune analisi esplorative abbiano mostrato come la variabile sembri discriminare parzialmente tra *background* e segnale, tale risultato sembra comunque poco plausibile dopo quanto detto sulle caratteristiche dei dati e delle variabili a disposizione.

  A dispetto della generale efficacia delle procedure nell'attribuire maggiore rilevanza alle variabili che anche la comunità fisica ritiene maggiormente discriminanti, le stesse non si prestano ad essere utilizzate nelle fasi successive della procedura. Tra le variabili maggiormente rilevanti, alcune sono intrinsicamente discrete e con supporto limitato. Ovviamente in questa situazione i metodi non parametrici che sono stati applicati non possono essere ottimali in quanto fanno ricorso sovente ad una stima della densità basata su *kernel* gaussiani: potrebbe essere utile modificare i metodi proposti in questo lavoro utilizzando dei *kernel* la cui natura vari adeguatamente sulla base delle va-

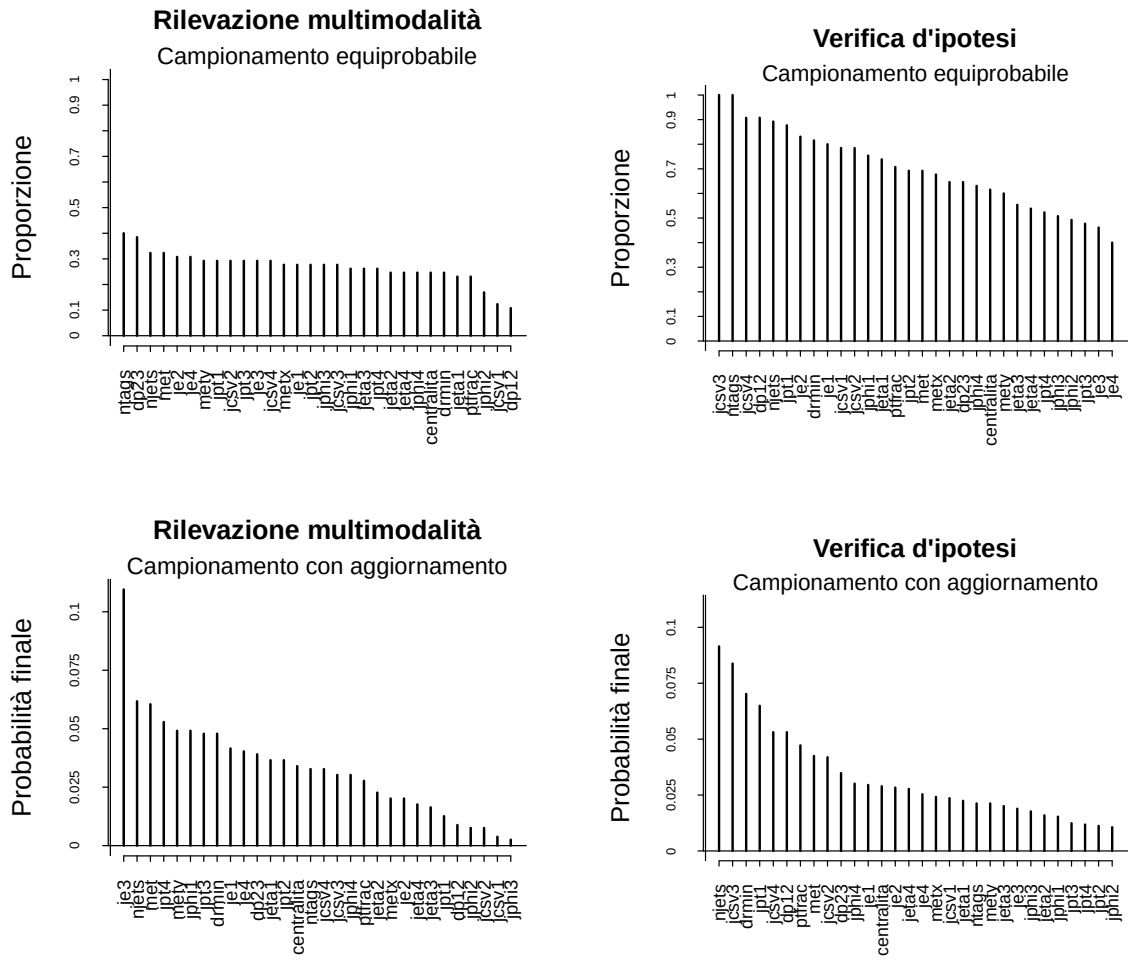

Figura 5.2: Risultati delle procedure di selezione delle variabili

riabili a disposizione (si veda, ad esempio, Li e Racine, 2003). Un ulteriore problema riguarda la multimodalità di alcune variabili quali ad esempio *jcsv*, relativa ai quattro diversi jets. Questo, come già sottolineato, ha richiesto un adattamento dell'algoritmo di selezione basato su rilevazione di multimodalità ma compromette le fasi successive dell'analisi;

- Con riferimento alla stima delle distribuzioni f_b e f_{bs}, necessarie per la successiva individuazione dei gruppi, un problema rilevante riguarda il fatto che, a fronte dell'ipotesi di unimodalità del *background* alla base della procedura proposta, f_b si presenta multimodale, e la multimodalità è in particolare da attribuirsi ad alcune delle variabili individuate dai metodi di selezione.
  Selezionare, pertanto, H_b in modo da rendere unimodale la distribuzione di *background*, comporta un eccessivo lisciamento della stessa che conduce a mascherare il segnale in f_{bs}. Come sarà illustrato a breve, è tuttavia utile sottolineare come il ricorso a variabili che presentano una meno netta multimodalità conduca a risultati interessanti;

- Con riferimento all'algoritmo *MEM* aggiustato tenendo conto dell'informazione disponibile sul *background*, sono emerse due principali criticità: sebbene da analisi preliminari svolte su dati simulati, la stessa sia risultata efficace, in una situazione di sbilanciamento tra la classe del segnale e quella del *background*, l'impiego del campione di *background*, oltre che del campione completo, ai fini della stima della densità, comporta che questo predomini e il segnale venga sopraffatto. Quanto proposto introduce di fatto una sorta di vincolo nel procedimento di raggruppamento che consiste nel forzare l'algoritmo a tener conto della specifica distribuzione di *background* nota o stimata arbitrariamente bene. Si ritiene che questo vincolo risulti adeguato, nel caso in cui si voglia operare nell'ambito semi-supervisionato che si è studiato in questa tesi, ma d'altra parte si è consapevoli che, in situazioni in cui il segnale è raro e si presenta in regioni ad elevata densità della distribuzione di *background*, può introdurre una penalizzazione eccessiva che porta l'algoritmo ad individuare un unico *cluster* e quindi a concludere in favore di una totale assenza di segnale. Potrebbe essere di interesse cercare di introdurre l'informazione legata alla conoscenza della distribuzione di *background* in maniera differente per far in modo di non ottenere una situazione di questo tipo. Inoltre, anche la procedura originale proposta da Li, Ray e Lindsay (2007) non riesce a individuare una corretta struttura di gruppo anche nei casi in cui, come illustrato a breve, un differente approccio non parametrico produce risultati promettenti distinguendo adeguatamente la presenza di segnale. Questo comportamento può

	Cl_1	Cl_2	Cl_3	Cl_4	Cl_5	Cl_6	Cl_7
Bkg_t	2125	2027	2291	1377	42	327	187
Sgn_t	52	69	144	59	287	514	499

Tabella 5.6: Risultati ottenuti utilizzando l'approccio all'analisi di raggruppamento non parametrica proposto da Azzalini e Torelli (2007) sulle variabili *jcsv3*, *jcsv4*, *dp12*, *ntags*. Con Bkg_te Sgn_t vengono indicate le vere classi di appartenenza.

essere dovuto ad una maggiore difficoltà della procedura nel determinare una struttura di gruppo nel caso in cui i gruppi non siano ben separati.

Per questa ragione è stata impiegata una procedura alternativa di individuazione delle regione modali, utilizzando l'algoritmo di *clustering* proposto da Azzalini e Torelli (2007). L'obiettivo è quello di comprendere da una parte le potenzialità dell'approccio non parametrico (a dispetto dell'inefficacia della specifica procedura *MEM*), dall'altra il comportamento delle procedure proposte di selezione delle variabili e di selezione della matrice di lisciamento.

Nella tabella 5.6 si mostrano i risultati ottenuti facendo ricorso al metodo di *clustering* proposto da Azzalini e Torelli (2007) utilizzando le quattro variabili considerate maggiormente rilevanti dalla procedura di selezione basata su verifica d'ipotesi nella variante di campionamento equiprobabile. In questo caso il segnale viene individuato con una percentuale di errore globale relativamente contenuta (di poco inferiore al 9%). Tuttavia, la procedura individua sette regioni modali di cui quattro sostanzialmente associate al *background* e tre associate al segnale. Il risultato, coerente con la precedente osservazione circa la multimodalità della distribuzione, è interessante in quanto permette di cogliere alcuni vantaggi legati alle procedure di selezione delle variabili proposte. Anche in questa analisi si sono utilizzate quattro variabili; tale scelta è stata fatta più per motivi legati a coerenza e di ordine computazionale che non sostantivi. In una situazione come quella in cui si sta operando si ritiene che l'utilizzo di sole quattro variabili possa non essere realmente adeguato in quanto probabilmente comporta una perdita troppo rilevante di informazione. Tuttavia i risultati forniti sono in linea con quanto ottenuto dal metodo parametrico e mostrano indicazioni chiare riguardo la presenza di segnale che, in questo caso, possono essere anche interpretabili in termini fisici.

In seconda battuta, la procedura è stata applicata, analogamente al caso parametrico, alle prime quattro componenti principali, poichè da un'analisi esplorativa delle stesse sul *background* è emersa una multimodalità meno marcata rispetto alle variabili originali. In questa situazione è stato possibile valutare, seppur non in condizioni ottimali, anche la tecnica sviluppata per la selezio-

	Cl_1	Cl_2	Cl_3	Cl_4
Bkg_t	3820	3830	466	260
Sgn_t	143	114	736	631

Tabella 5.7: Risultati ottenuti utilizzando l'approccio all'analisi di raggruppamento non parametrica proposto da Azzalini e Torelli (2007) sulle prime quattro componenti principali utilizzando matrici di lisciamento selezionate tramite la procedura proposta nel paragrafo (4.2). Con Bkg_te Sgn_t vengono indicate le vere classi di appartenenza.

ne del parametro di lisciamento. Dopo aver quindi selezionato la matrice di lisciamento H_b in modo tale da ottenere $\hat{f}_b$ unimodale, la moda M_b di tale distribuzione è stata utilizzata nella procedura per la scelta di H_{bs}, selezionata in modo tale da portare ad una stima che abbia gradiente nullo in M_b. Una volta selezionata la matrice H_{bs} si è proceduto utilizzando l'algoritmo di *clustering* non parametrico proposto da Azzalini e Torelli (2007) che fornisce la partizione riportata in tabella 5.7. La procedura ha fornito due gruppi per il *background* e due gruppi per il segnale. Si noti come quindi, selezionando la matrice H_{bs} nel modo proposto la distribuzione $\hat{f}_{bs}$ risulta comunque essere multimodale, con multimodalità associata, oltre alla presenza del segnale, anche al *background*. Tuttavia è garantita l'individuazione della moda rilevata in $\hat{f}_b$. Tale risultato è più fedele alla situazione reale rispetto a quanto fornito dall'approccio parametrico che, operando sullo stesso sottospazio, individuava nove componenti e quindi nove diversi gruppi. In questo caso l'errore di classificazione è di poco inferiore al 10%.

5.3 Conclusioni

Questo lavoro ha tratto origine da un problema reale che si pone frequentemente in un contesto di fisica delle particelle. Esso nasce dall'esigenza di individuare una possibile fonte di un segnale di interesse, molto spesso debole, che si manifesta come un picco nella distribuzione di *background*, supposta nota o stimabile arbitrariamente bene grazie alla disponibilità potenzialmente illimitata di dati generati dal processo.

In questo contesto, è stato proposto un approccio non parametrico globale di tipo semi-supervisionato che, partendo dalla conoscenza del *background*, ha come obiettivo ultimo quello di individuare eventuali osservazioni anomale. L'approccio, scomponibile nelle diversi fasi di riduzione della dimensionalità dei dati, stima della funzione di densità e individuazione delle regioni modali, intende superare alcune problematicità usuali che si riscontrano quando si utilizzano metodologie statistiche non parametriche. Si è introdotta una procedura generale di selezione delle variabili

avente lo scopo di individuare le variabili rilevanti al fine dell'identificazione del segnale e di permettere quindi di operare in uno spazio di dimensione ridotta, con i vantaggi che questo comporta.
È stato presentato un criterio per la scelta della matrice di lisciamento per la stima della densità basata sul metodo del nucleo, che si avvale delle informazioni aggiuntive a disposizione sul processo di *background*.
Infine è stata approntata una modifica ad un algoritmo di *clustering* non parametrico che, ancora una volta, ha lo scopo di permettere all'algoritmo di tener conto, nella procedura di raggruppamento, delle informazioni che si hanno a disposizione sul comportamento di *background* assunto dai dati.

Per testare quanto proposto si è fatto ricorso sia a uno studio di simulazione sia a un'applicazione a dati relativi all'ambito della fisica delle particelle. I risultati ottenuti, pur evidenziando diverse criticità con riferimento alla specifica applicazione considerata, dove si è riscontrata una grave violazione di alcune delle ipotesi alla base dell'approccio, hanno comunque permesso di trarre alcune indicazioni generali di interesse.
La procedura di selezione delle variabili è risultata efficace nell'evidenziare le variabili più discriminanti, sia nello studio di simulazione che sui dati reali, dove l'interpretazione dei risultati è apparsa in linea con quanto sostenuto dagli esperti nel settore. Si ritiene che una procedura di questo tipo, che permette di selezionare alcune variabili tra quelle a disposizione, possa essere particolarmente utile nel caso in cui si vogliano superare le criticità introdotte dall'elevata dimensionalità dei dati senza però perdere l'interpretabilità dei risultati.
Sebbene la procedura *MEM* sia risultata inadatta a rilevare eventuali anomalie rare e non chiaramente distinte dalla distribuzione del *background*, nel lavoro si è mostrato come un algoritmo di *clustering* basato sulle curve di livello della densità abbia portato a risultati promettenti pur operando in un contesto non supervisionato e non tenendo quindi conto delle informazioni aggiuntive a disposizione. Questo suggerisce che l'approccio non parametrico abbia comunque le potenzialità per raggiungere risultati utili.

Visto quanto detto si ritiene che potrebbe essere interessante proseguire nel perfezionamento delle tecniche proposte. Eventuali generalizzazioni potrebbero essere introdotte in particolare per quanto riguarda il metodo di selezione della matrice di lisciamento rendendo tale procedura utilizzabile anche nei casi in cui la distribuzione di *background* risulti essere multimodale, o includendo vincoli alternativi circa la conoscenza della distribuzione di *background*.
Si ritiene possibile anche la modifica dei metodi proposti per fare in modo che possano tenere in considerazione in maniera più adeguata la natura delle variabili a

disposizione; così facendo si pensa sia possibile migliorare i risultati ottenuti vista la presenza in questo caso di variabili discrete e limitate.
La principale criticità si reputa sia quella legata alla modifica dell'algoritmo *MEM*. Si ritiene perciò possa essere interessante indagare ulteriormente il comportamento del *MEM* in situazioni di gruppi non bilanciati e non adeguatamente separati. Inoltre, appare promettente l'idea di concentrarsi, similmente a quanto fatto con riferimento al *MEM*, sull'aggiustamento di metodi non parametrici alternativi e più efficaci, per tener conto dell'informazione sul *background*.

Appendice A

Appendice A: Codice

Algoritmo EM modificato La funzione implementa la procedura di Vatanen et al. (2012) descritta nel paragrafo 2.2. Richiede in input l'insieme di dati $\mathcal{X}_b$ (datib) e $\mathcal{X}_{bs}$ (datibs), il numero massimo di componenti del modello (tot), una soglia di convergenza ϵ(eps), il numero massimo di iterazioni (max.iter) e il numero massimo di volte in cui una componente della mistura può essere re-impostata a valori casuali (max.whititer).
In output fornisce i parametri del modello $p_B(x)$ e del modello $p_{FB}(x)$.

```
EM_mod <- function(datib,datibs,tot,eps=1e-7,max.iter=1000,max.whititer) {
library(mclust)
library(mvtnorm)
library(ks)
bkg <- densityMclust(data=datib)
meanbkg <- bkg$parameters$mean
varbkg <- bkg$parameters$variance$sigma
compbkg <- bkg$G
propbkg <- bkg$parameters$pro
aictot <- matrix(NA,nrow=(tot-compbkg)+1,ncol=max.whititer)
b <- compbkg
while (b<tot) {
b <- b+1
withiter <- 0
aic <- 0
while ((aic==0) & (max.whititer>withiter)) {
iter <- 1
withiter <- withiter+1
abs.e <- 1
```

```
p0 <- matrix(rep(1/b,b),nrow=1)
mean <- array(NA,dim=c(NCOL(datibs),b,max.iter))
m <- matrix(runif((b-compbkg)*NCOL(datibs),-5,5),nrow=NCOL(datibs),
ncol=(b-compbkg))
mean[,,1] <- cbind(meanbkg,m)
sigma <- array(NA,dim=c(NCOL(datibs),NCOL(datibs),b))
sigma[,,1:compbkg] <- varbkg
sigma1<-sapply(1:(b-compbkg),function(x) sigma[,,x] <-
diag(runif(NCOL(datibs),0.5,1.5),NCOL(datibs)),simplify=FALSE)
sigma1 <- matrix(unlist(sigma1),nrow=NCOL(datibs),ncol=NCOL(datibs))
sigma[,,(compbkg+1):b] <- sigma1
sigmasgn <- array(NA,dim=c(NCOL(datibs),NCOL(datibs),b))
while ((abs.e>eps) & (iter<max.iter))
{
dens <- matrix(NA,nrow=NROW(datibs),ncol=length(p0))
pesi <- matrix(NA,nrow=NROW(datibs),ncol=length(p0))
pesi2 <- matrix(NA,nrow=NROW(datibs),ncol=1+(b-compbkg))
gamma <- matrix(NA,nrow=NROW(datibs),ncol=1+(b-compbkg))
for (j in 1:b) {
dens[,j] <- dmvnorm(datibs,mean=mean[,j,iter],sigma=sigma[,,j])
}
for (i in 1:NCOL(pesi)) {
pesi[,i] <- p0[i]*dens[,i]
}
for (i in 1:NCOL(pesi2)) {
if(i==1) {
pesi2[,i] <- apply(pesi[,1:compbkg],1,sum)
} else {
pesi2[,i] <- pesi[,compbkg+(i-1)]
}
}
for (i in 1:NCOL(gamma)) {
if (i==1 & NCOL(gamma)==2) {
gamma[,i] <- pesi2[,i]/(pesi2[,i]+pesi2[,2])
}
if (i==1 & NCOL(gamma)!=2) {
gamma[,i] <- pesi2[,i]/(pesi2[,i]+apply(pesi2[,(i+1):NCOL(pesi2)],1,sum))
}
```

```
if(i!=1) {
if(NCOL(gamma)==2) {
gamma[,i] <- pesi2[,i]/(pesi2[,1]+pesi2[,2])
} else {
gamma[,i] <- pesi2[,i]/(pesi2[,1]+apply(pesi2[,-1],1,sum))
}
}
}
gm <- apply(gamma,2,mean)
p02 <- c(gm[1]*propbkg,gm[2:length(gm)])
if(any(p02==0) | sum(is.na(p02))!=0) {
p02[which(p02==0)] <- rep(0.05,length(which(p02==0)))
p02[which(is.na(p02))] <- rep(0.05,length(which(is.na(p02))))
p02 <- p02/sum(p02)
}
mean1 <- t(gamma[,2:(1+(b-compbkg))])%*%datibs/apply(gamma,2,sum)
[2:(1+(b-compbkg))]
mean2 <- mean[,,iter]
mean2[,(compbkg+1):dim(mean)[2]] <- mean1
if (sum(is.na(mean2))!=0) {
mean2[which(is.na(mean2))]<-runif(length(which(is.na(mean2))),0.2,2)
}
for (k in 1:(length(p0)-compbkg)) {
a <- array(NA,dim=c(NCOL(datibs),NCOL(datibs),NROW(datibs)))
for (v in 1:nrow(datibs)) {
a[,,v] <- gamma[v,1+k]*(as.vector(datibs[v,]-mean[,compbkg+k,iter])%*%
t(as.vector(datibs[v,]-mean[,compbkg+k,iter])))
}
a <- apply(a,c(1,2),sum)
sigmasgn[,,compbkg+k] <- a/apply(gamma,2,sum)[1+k]
}
if(sum(is.na(sigmasgn[,,(compbkg+1):b]))!=0) {sigmasgn[,,(compbkg+1):b] <-
array(diag(runif(1,0,3),NCOL(datibs)),dim=dim(sigmasgn[,,(compbkg+1):b]))
}
abs1 <- sum(abs(mean2-mean[,,iter]))
abs2 <- sum(abs(sigmasgn[,,(compbkg+1):b]-sigma[,,(compbkg+1):b]))
abs3 <- sum(abs(p02-p0))
abs.e <- abs1+abs2+abs3
```

```
mean[,,iter+1] <- mean2
sigma[,,(compbkg+1):b] <- sigmasgn[,,(compbkg+1):b]
p0 <- p02
iter <- iter+1
}
likbkg <- sum(log(dmvnorm.mixt(datibs,mus=t(rbind(mean[,1:(b-1),iter])),
Sigmas=matrix(rbind(varbkg[,,1:compbkg]),
ncol=NCOL(datibs),byrow=T),props=propbkg)))
liktot <- sum(log(dmvnorm.mixt(datibs,mus=t(rbind(mean[,1:b,iter])),
Sigmas=matrix(sigma[,,1:b],ncol=NCOL(datibs),
byrow=T),props=p0)))
parbkg <- sum(varbkg!=0)+(dim(meanbkg)[1]*dim(meanbkg)[2])+length(propbkg)
partot <- sum(sigma!=0)+(dim(mean[,,iter])[1]*dim(mean[,,iter])[2])+
length(p0)
aictot[1,1] <- 2*parbkg-2*likbkg
aictot[(b-compbkg)+1,withiter] <- 2*partot-2*liktot
if ((2*partot-2*liktot)<aictot[b-compbkg,(which(is.na(aictot[b-compbkg,]))[1]
-1)]) {
aic <- 1
out <- list(medie=mean[,,iter],varianza=sigma,prop=p0,componenti=b)
}
}
if (max.whititer==withiter) {break}
}
out <- list(res=out,bkg=list(medie=meanbkg,varianza=varbkg,
componenti=propbkg))
return(out)
}
```

Procedura di selezione della matrice di lisciamento H_{bs} La funzione implementa la procedura descritta nel paragrafo 4.2.Richiede in input l'insieme di dati $\mathcal{X}_b$ (datib) e $\mathcal{X}_{bs}$ (datibs).
In output fornisce la diagonale della matrice H_{bs}.

```
nullgrad <- function(datib,datibs) {
library(mvtnorm);library(Matrix);library(ks);library(pdfCluster)
library(sn);library(Modalclust);library(snowfall)
d <- ncol(datab)
hnorm <- h.norm(datib)
griglia <- seq(from=0.5,to=5,by=0.05)%*%t(hnorm)
cl <- phmac(datib,parallel=TRUE,npart=3)
moda <- cl$mode[unique(cl$n.cluster)==1][[1]]
grad<-function(eval,data,sigma) {
out <- apply(data,1,function(x) -(dmvnorm(eval,mean=x,sigma=sigma))*
(solve(sigma))%*%t(eval-x))
out <- t(out)
out <- apply(out,2,mean)
return(out)
}
sfInit(parallel=TRUE,cpus=3)
sfExportAll()
sfLibrary(mvtnorm)
gradiente <- sfApply(griglia,1,function(x) grad(moda,datibs,sigma=diag(x)))
gradiente <- t(gradiente)
gradiente1 <- apply(abs(gradiente^2),1,sum)
graddiff <- abs(diff(gradiente1))
ot <- which.min(gradiente1)
hopt <- as.numeric(griglia[ot,])
out <- list()
out$Hbs <- hopt
return(out)
}
```

Algoritmo MEM modificato La funzione implementa la procedura descritta al paragrafo 4.3. Richiede in input l'insieme di dati $\mathcal{X}_b$ (datib) e $\mathcal{X}_{bs}$ (datibs) e una matrice contenente i due vettori diagonali delle matrici H_b e H_{bs} (Sigmas).
In output fornisce il numero di gruppi trovati, le mode che identificano tali gruppi e l'appartenenza delle singole osservazioni ai gruppi.

```
HMACmixtmixt <- function(datibs,datib,Sigmas) {
library(mvtnorm);library(zoo)
n <-nrow(datibs)
m <- ncol(datibs)
n.cluster=c();G=datibs;member=seq(n);member.n=c();modes=c();members=c()
g=1;clust=c();f=c();M=c();
for (i in 1:n) {
x0=datibs[i,];x0_old=1;d=1;
while (d>10^(-5))
{
datitot <- rbind(datibs,datib)
f1 <- dmvnorm(datib,x0,diag(Sigmas[1,],ncol(datib)))
f2 <- dmvnorm(datibs,x0,diag(Sigmas[2,],ncol(datibs)))
f <- c(f1,f2)
p=f/sum(f)
x0=p%*%datitot
if (sqrt(sum(x0^2))!=0) {
d=sqrt(sum((x0_old-x0)^2))/sqrt(sum(x0^2))
} else {d <- 10}
x0_old <- x0
}
if (length(dim(M))!=0)
{
temp=c(1:dim(M)[1])[as.logical(apply(abs(M-matrix(1,
dim(M)[1])%*%x0)<0.001*matrix(1,dim(M)[1])%*%s_hat,1,prod))]
if (length(temp)==0) {
clust[i]<-g
g <-g+1
} else {
clust[i]=clust[temp[1]]
}
}
else {clust[i]=g;g=g+1}
```

```
M=rbind(M,x0)
member.n[member==i]=clust[i]
}
if (m==1) {
G=as.matrix(M[match(c(1:max(clust)),clust)])
} else {
G=M[match(c(1:max(clust)),clust),]
}
n.cluster[1]=max(clust);
modes[[1]]=G
members[[1]]=member.n
member=member.n;
}
out=list()
out$n.cluster<-n.cluster
out$mode<-modes[c(1:length(Sigmas))[!duplicated(n.cluster)]]
out$membership<-unique(members)
return(output)
}
```

Procedura di selezione delle variabili La funzione implementa le procedure descritte nei paragrafi 4.4.2 e 4.4.3. Richiede in input l'insieme di dati $\mathcal{X}_b$ (datib) e $\mathcal{X}_{bs}$ (datibs), il numero di iterazioni dell'algoritmo (iter), il numero di variabili da selezionare ad ogni iterazione (nvar), la lunghezza del periodo di *warm-up* (warmup), il metodo di selezione da utilizzare (method, possibili opzioni "pdf" e "kde"), il metodo di campionamento delle variabili ad ogni iterazione da utilizzare (sampling, possibili opzioni "equipr" e "multinom"), la dimensione dei sottocampioni di osservazioni sui quali operare rispettivamente per $\mathcal{X}_b$ (datib) e $\mathcal{X}_{bs}$ (sub1, sub2), la matrice di lisciamento asintoticamente ottimale calcolata in precedenza per il *background* qualora si utilizzi il metodo basato su verifica d'ipotesi (H1).
In output fornisce il contatore che misura l'importanza delle variabili, qualora si sia utilizzato il metodo con campionamento equiprobabile, o il vettore di probabilità di estrazione all'ultima iterazione, qualora si sia utilizzato l'aggiornamento multinomiale.

```
selvar <- function(datib,datibs,iter,nvar,warmup,method,sampling,
sub1,sub2,H1) {
out <- list()
tot <- matrix(0,nrow=iter,ncol=NCOL(datib))
tot[1,] <- rep(1,ncol(datib))
keep <- matrix(NA,nrow=iter,ncol=nvar)
p <- matrix(NA,nrow=iter+1,ncol(datib))
p[c(1,2),] <- rep(1/ncol(datib),ncol(datib))
nlogL <- function(param,data) {
-sum(dmultinom(data,prob=param,log=TRUE))
}
if (method=="pdf") {
noc <- matrix(NA,nrow=iter,ncol=2)
for (i in 2:iter) {
keep[i,] <- sample(1:ncol(datib),size=nvar,prob=p[i,])
s1 <- sample(1:nrow(datib),size=sub1,replace=F,prob=rep(1/
nrow(datib),nrow(datib)))
s2 <- sample(1:nrow(datibs),size=sub2,replace=F,prob=
rep(1/nrow(datibs),nrow(datibs)))
cl1 <- try(pdfCluster(datib[s1,keep[i,]])@noc,silent=TRUE)
if (!is.character(cl1)) {
noc[i,1] <- cl1
cl2 <- try(pdfCluster(datibs[s2,keep[i,]])@noc,silent=TRUE)
if (!is.character(cl2)) {
```

```
noc[i,2] <- cl2
if(noc[i,1]<noc[i,2]) {
tot[i,] <- tot[i-1,]
tot[i,keep[i,]] <- tot[i,keep[i,]]+1
} else {
tot[i,] <- tot[i-1,]
}
} else {
noc[i,2] <- NA
tot[i,] <- tot[i-1,]
}
} else {
noc[i,] <- c(NA,NA)
tot[i,] <- tot[i-1,]
}
out$ris <- tot[iter,]
if (sampling=="multinom") {
if (i>warmup) {
p[i+1,] <- optim(par=c(rep(1/ncol(datib),ncol(datib))),fn = nlogL,
data=tot[i,],lower = 1e-8,upper=1-1e-8)$par
p[i+1,] <- p[i+1,]/(sum(p[i+1,]))
} else {
p[i+1,] <- p[i,]
}
out$ris <- p[iter,]
} else {p[i+1,] <- p[i,]}
}
return(out)
}
if (method=="kde") {
poss <- matrix(NA,nrow=iter,ncol=1)
for (i in 2:iter) {
keep[i,] <- sample(1:ncol(datib),size=nvar,prob=p[i,])
s1 <- sample(1:nrow(datib),size=sub1,replace=F,prob=rep(1/nrow(datib),
nrow(datib)))
s2 <- sample(1:nrow(datibs),size=sub2,replace=F,prob=rep(1/nrow(datibs),
nrow(datibs)))
hpi <- diag(H1[keep[i,]],length(keep[i,]))
```

```
poss[i,] <- kde.test(datib[s1,keep[i,]],datibs[s2,keep[i,]],H1=hpi)$pvalue
if(poss[i,]<=0.05) {
tot[i,] <- tot[i-1,]
tot[i,keep[i,]] <- tot[i,keep[i,]]+1
} else {
tot[i,] <- tot[i-1,]
}
if (sampling=="multinom") {
if (i>warmup) {
p[i+1,] <- optim(par = c(rep(1/ncol(datib),ncol(datib))),
fn = nlogL,data=tot[i,],lower = 1e-8,upper=1-1e-8)$par
p[i+1,] <- p[i+1,]/(sum(p[i+1,]))
} else {
p[i+1,] <- p[i,]
}
} else { p[i+1,] <- p[i,] }
}
if(sampling=="multinom") {out$ris <- p[iter+1,]}
else {out$ris <- tot[iter,]}
return(out)
}
}
```

Bibliografia

Aad, G. et al. (2012). «Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC». In: *Physics Letters B* 716.1, pp. 1–29.

Anderson, N. H., P. Hall e D. M. Titterington (1994). «Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates». In: *Journal of Multivariate Analysis* 50.1, pp. 41–54.

Arellano-Valle, R. B. e M. G. Genton (2005). «On fundamental skew distributions». In: *Journal of Multivariate Analysis* 96.1, pp. 93–116.

Azzalini, A. (2015). *The R package `sn`: The Skew-Normal and Skew-t distributions (version 1.3-0)*. Università di Padova, Italia. URL: `http://azzalini.stat.unipd.it/SN`.

Azzalini, A. e G. Menardi (2014). «Clustering via Nonparametric Density Estimation: The R Package pdfCluster». In: *Journal of Statistical Software* 57.11, pp. 1–26. URL: `http://www.jstatsoft.org/v57/i11/`.

Azzalini, A. e N. Torelli (2007). «Clustering via nonparametric density estimation». In: *Statistics and Computing* 17.1, pp. 71–80.

Basu, S., A. Banerjee e R. Mooney (2002). «Semi-supervised clustering by seeding». In: *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer.

Basu, S., M. Bilenko e R. J. Mooney (2004). «A probabilistic framework for semi-supervised clustering». In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 59–68.

Beringer, J. et al. (2012). «Review of particle physics particle data group». In: *Physical Review D (Particles, Fields, Gravitation and Cosmology)* 86.1, p. 010001.

Bowman, A. W. e A. Azzalini (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*. Vol. 18. OUP Oxford.

Breiman, L. (2001). «Random forests». In: *Machine learning* 45.1, pp. 5–32.

Burman, P. e W. Polonik (2009). «Multivariate mode hunting: Data analytic tools with measures of significance». In: *Journal of Multivariate Analysis* 100.6, pp. 1198–1218.

Carmichael, J. e R. Julius (1968). «Finding natural clusters». In: *Systematic Biology* 17.2, pp. 144–150.

Chacón, J. E. e T Duong (2010). «Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices». In: *Test* 19.2, pp. 375–398.

Chandola, V., A. Banerjee e V. Kumar (2009). «Anomaly detection: A survey». In: *ACM computing surveys (CSUR)* 41.3, p. 15.

Chang, W.-C. (1983). «On using principal components before separating a mixture of two multivariate normal distributions». In: *Applied Statistics*, pp. 267–275.

Chapelle, O., B. Scholkopf e A. Zien (2009). «Semi-Supervised Learning». In: *IEEE Transactions on Neural Networks* 20.3, pp. 542–542.

Chatrchyan, S. et al. (2012). «Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC». In: *Physics Letters B* 716.1, pp. 30–61.

Dempster, A. P., N. M. Laird e D. B. Rubin (1977). «Maximum likelihood from incomplete data via the EM algorithm». In: *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Duong, T., B. Goud e K. Schauer (2012). «Closed-form density-based framework for automatic detection of cellular morphology changes». In: *Proceedings of the National Academy of Sciences* 109.22, pp. 8382–8387.

Duong, T. et al. (2007). «ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R». In: *Journal of Statistical Software* 21.7, pp. 1–16.

Fraley, C. e A. E. Raftery (2002). «Model-based clustering, discriminant analysis, and density estimation». In: *Journal of the American statistical Association* 97.458, pp. 611–631.

Fukunaga, K. e L. Hostetler (1975). «The estimation of the gradient of a density function, with applications in pattern recognition». In: *IEEE Transactions on information theory* 21.1, pp. 32–40.

Genz, A. et al. (2008). «mvtnorm: Multivariate Normal and t Distributions». In: *R package version 0.9-2, URL http://CRAN. R-project. org/package= mvtnorm*.

Hall, P. (1984). «Central limit theorem for integrated square error of multivariate nonparametric density estimators». In: *Journal of multivariate analysis* 14.1, pp. 1–16.

Hartigan, J. (1975). *Clustering Algorithms*. New York: John Wiley & Sons Inc.

Hartigan, J. e S. Mohanty (1992). «The runt test for multimodality». In: *Journal of Classification* 9.1, pp. 63–70.

Hartigan, J. A. e P. Hartigan (1985). «The dip test of unimodality». In: *The Annals of Statistics*, pp. 70–84.

Hastie, T, R Tibshirani e J Friedman (2009). *The elements of statistical learning 2nd edition*.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Li, J., S. Ray e B. G. Lindsay (2007). «A nonparametric statistical approach to clustering via mode identification». In: *Journal of Machine Learning Research* 8.Aug, pp. 1687–1723.

Li, Q. e J. Racine (2003). «Nonparametric estimation of distributions with categorical and continuous data». In: *journal of multivariate analysis* 86.2, pp. 266–292.

Markou, M. e S. Singh (2003). «Novelty detection: a review—part 1: statistical approaches». In: *Signal processing* 83.12, pp. 2481–2497.

McLachlan, G. e D. Peel (2004). *Finite mixture models*. John Wiley & Sons.

Menardi, G. (2015). «A Review on Modal Clustering». In: *International Statistical Review*. 10.1111/insr.12109, n/a–n/a. ISSN: 1751-5823. DOI: 10.1111/insr.12109. URL: http://dx.doi.org/10.1111/insr.12109.

Parzen, E. (1962). «On estimation of a probability density function and mode». In: *The annals of mathematical statistics* 33.3, pp. 1065–1076.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Scott, D. W. (2015). *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Silverman, B. W. (1981). «Using kernel density estimates to investigate multimodality». In: *Journal of the Royal Statistical Society. Series B*, pp. 97–99.

Smyth, P. (2000). «Model selection for probabilistic clustering using cross-validated likelihood». In: *Statistics and computing* 10.1, pp. 63–72.

Vatanen, T. et al. (2012). «Semi-supervised detection of collective anomalies with an application in high energy particle physics». In: *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, pp. 1–8.

Wand, M. P. e M. C. Jones (1994). *Kernel smoothing*. Crc Press.

Zhu, X. (2005). *Semi-Supervised Learning Literature Survey*. Rapp. tecn. 1530. Computer Sciences, University of Wisconsin-Madison.

RINGRAZIAMENTI

Alla professoressa Giovanna Menardi per la passione e la dedizione che mette nel suo lavoro e per il modo in cui è riuscita a trasmettermele.

Al professor Tommaso Dorigo per la pazienza, la disponibilità e l'entusiasmo che mette nel far avvicinare le persone alla sua materia.

A mia mamma, per gli insegnamenti e gli esempi che mi hanno aiutato a crescere.

A mia nonna, per tutto l'affetto.

A Irene, per le discussioni.

A tutta la mia famiglia che, in un modo o nell'altro, mi sopporta da 25 anni.

A Piera, per i commenti caustici e lo stress da biblioteca.
A Valentina, per essere la mia bvdc da una vita ormai.
A Santiago, Bargiu, Silvia e Marco. Fondamentalmente per il risiko. Ah e le cene pugghiesi.
E a tutti gli altri amici fuori e dentro l'università per ogni momento passato assieme.

E infine, a Sara. Per la realizzazione dei sogni.