Univerità degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Tesi

Identificazione e inferenza per clustering multivariato con applicazione a dati di pazienti ischemici

Relatore: Prof. Livio Finos
Dipartimento di Scienze Statistiche
Correlatori: Prof. Federico Ferraccioli
M.Sc Andrea Zanola

Laureanda: Elisa Colorio
Matricola N. 2058302

Anno Accademico 2023/24

Indice

Introduzione

Il clustering è stato oggetto di studio e ricerca per diversi decenni ed è diventato una delle tecniche più rilevanti nell'ambito dell'analisi dei dati, poiché consente di individuare strutture nascoste e pattern significativi all'interno di dataset complessi ed eterogenei. Questa tecnica ha trovato applicazione in una vasta gamma di settori, tra cui medicina, biologia, marketing e molti altri, dimostrandosi utile per il *data exploration*.

Nonostante siano stati fatti notevoli progressi nel corso degli anni, il clustering rimane un problema complesso e in continuo sviluppo, soprattutto nel contesto non parametrico. Inoltre, la definizione stessa di *cluster* può variare a seconda del contesto e degli obiettivi dell'analisi, rendendo ancora più difficile stabilire criteri oggettivi per determinare se un insieme di dati costituisce effettivamente un cluster.

Nel contesto della tesi in questione, ci si concentra su un particolare dataset relativo a pazienti colpiti da ictus, in cui sono stati registrati diversi deficit utilizzando la scala NIHSS (National Institutes of Health Stroke Scale).

L'obiettivo principale è esplorare questo dataset al fine di identificare possibili raggruppamenti di pazienti significativi, qualora fossero presenti, che condividono caratteristiche simili. Per raggiungere l'obiettivo, data la complessità dei dati e la necessità di gestire cluster di forme e dimensioni variabili, si è deciso di adottare un approccio di clustering basato sulla densità.

Per fare ciò, verrà seguito un processo che prevede diverse fasi di trasformazione dei dati. In particolare, verrà adottata una metodologia che include una prima analisi dei dati attraverso una breve analisi descrittiva e una successiva creazione di una matrice delle distanze apposita per i dati ordinali in questione che rappresenta le relazioni tra le osservazioni nel dataset. Successivamente verranno applicate tecniche di embedding spettrale per ridurre la dimensionalità dei dati e proiettarli in uno spazio di dimensioni inferiori in cui la rappresentazione risulta chiara ed efficace consentendo un'interpretazione più semplice.

A questo punto verrà utilizzato un metodo che utilizza il clustering basato sulla densità per identificare cluster significativi sulla trasformazione dei dati. Questo tipo di clustering si basa sulla ricerca di regioni dense nello spazio dei dati, separandole da regioni più sparse, e può rivelarsi particolarmente efficace nel rilevare cluster di forma arbitraria e dimensioni variabili. In particolare verrà analizzato il metodo che si occupa di identificare le mode

della distribuzione dei dati attraverso il metodo *Mean Shift* che consente di individuare le regioni di massima densità nello spazio dei dati senza la necessità di specificare a priori il numero di cluster.

Nel primo capitolo verrà presentata una panoramica generale sul dataset dei pazienti ischemici, descrivendo la natura dei dati e i principali obiettivi della ricerca. Si discuteranno inoltre i concetti fondamentali del clustering e le sfide che comporta.
Nel secondo capitolo viene effettuata un'esplorazione del processo di trasformazione delle informazioni, partendo dai dati originali passando per la creazione della matrice delle distanze e l'embedding spettrale, al fine di preparare i dati per l'analisi dei cluster. Nel terzo capitolo verrà approfondito il concetto di clustering basato sulla densità e come questo sia legato alla ricerca delle mode nei dati. Nel quarto capitolo verranno esposti i metodi utilizzati per l'identificazione delle mode o cluster tra i pazienti ischemici. Infine, nel quinto capitolo verranno presentati i risultati ottenuti dal clustering applicato al dataset in analisi.

Capitolo 1

Introduzione al problema

1.1 Descrizione dataset

L'ictus è una condizione medica che si verifica quando il flusso di sangue verso una parte del cervello viene interrotto o ridotto, provocando danni alle cellule cerebrali a causa della mancanza di ossigeno. Si distingue fra ictus di tipo ischemico, quando il flusso sanguigno viene interrotto, e ictus emorragico, se avviene a causa della rottura di un vaso sanguigno e può comportare gravi disabilità e in alcuni casi la morte. Risulta quindi utile ottenere maggior informazioni sui pazienti al fine di avere una visione più ampia sul come trattarli. Un'analisi mirata all'identificazione di sottogruppi di pazienti ischemici con caratteristiche simili è vantaggiosa per numerosi motivi.

Un primo motivo è che l'appartenenza di un individuo a un gruppo di pazienti può rivelare informazioni utili circa la personalizzazione del trattamento in base alle caratteristiche ed esigenze dei gruppi di pazienti, consentendo l'applicazione di terapie più mirate ed efficaci per i pazienti ischemici e la riduzione degli effetti collaterali. Inoltre l'identificazione di cluster di pazienti ischemici può migliorare l'accuratezza della prognosi, riuscendo a predire in modo più accurato l'andamento della malattia e a pianificare le cure di conseguenza. Un'altra utilità che può portare l'analisi dei gruppi è la scoperta di nuove associazioni tra variabili cliniche, fornendo una migliore comprensione dei fattori che influenzano l'ictus e le sue conseguenze.

I deficit dei pazienti colpiti da ictus vengono misurati attraverso una nota scala clinica comunemente adottata per valutare la gravità dell' ictus, definita come NIHSS (National Institutes of Health Stroke Scale) in Kwah & Diong (2014). Questa viene utilizzata diffusamente per misurare il deterioramento iniziale e la risposta alle terapie acute. Da numerosi studi condotti con le PCA è emerso che i diversi deficit misurati dalla scala NHISS sono correlati tra loro, ossia è possibile che alcuni deficit possano verificarsi insieme più frequentemente di altri, suggerendo appunto una certa correlazione tra i deficit (Tshimanga et al. 2023).

I dati analizzati in questo elaborato provengono dall'Ospedale Universitario di Padova

e dall'Università di Washington, concentrandosi su un gruppo selezionato di 172 pazienti che hanno subito un ictus ischemico. La selezione dei pazienti è stata attentamente eseguita per concentrarsi esclusivamente sull'ictus ischemico, escludendo altri casi al fine di approfondire l'analisi di questa specifica tipologia di ictus. Il dataset comprende dettagliate informazioni su esami clinici dei pazienti tra cui test comportamentali, motori, sensoriali, visivi e linguistici con i quali si vuole misurare la gravità dei deficit neurologici. Precisamente il dataset contiene pazienti la cui età è compresa tra i 21 e i 94 anni con una mediana e media di 64 ± 15. Le variabili misurate secondo la NIHSS sono 14, descrivono i deficit neurologici dei pazienti e includono: livello di coscienza delle domande risposte correttamente, livello di coscienza dei comandi, disturbi visivi, paralisi degli occhi, paralisi facciale, forza del braccio destro, forza del braccio sinistro, forza della gamba destra, forza della gamba sinistra, atassia, sensibilità, linguaggio, disartria e negligenza alle istruzioni. Il punteggio finale del NIHSS è la somma dei punteggi relativi a questi deficit, di conseguenza un valore molto alto indica un maggior grado di danni neurologici, mentre un valore più basso suggerisce meno danni. Sono stati esclusi dall'analisi i pazienti che riportavano un punteggio totale NIHSS pari a 0, poichè indicava l'assenza di danni neurologici. In questo modo il punteggio NIHSS varia da un minimo di 1 a un massimo di 17 tra i pazienti, con una media di 4. Per la maggior parte dei deficit, la distribuzione dei punteggi NIHSS mostra che la maggioranza dei pazienti ha un punteggio pari a 0, indicando l'assenza di quella specifica sintomatologia. Solo una piccola parte dei pazienti ha un punteggio pari a 1, mentre pochi hanno punteggi superiori a 1. Questo schema si osserva per tutte le variabili, tranne per deficit dell'articolazione della parola, della paralisi facciale e della sensibilità, per le quali si nota una distribuzione diversa: molti pazienti presentano punteggi pari a 1 oltre al valore 0. Per osservare le relazioni tra i deficit dei pazienti è stato calcolato il coeficiente di Spearman, una misura di correlazione non parametrica che valuta il grado di associazione tra due variabili ordinali. Il grafico riportato in figura 1.1 fornisce informazioni sulla forza e sulla direzione della relazione tra le variabili. Ad esempio si può notare che all'aumentare del deficit LOC-Questions aumenterà anche il valore della variabile Best Language. Analogamente avviene per i deficit Motor Arm R e Motor Leg R, così come per Motor Arm L e Motor Leg L. Inoltre anche la correlazione tra l'attenzione e i disturbi motori agli arti inferiori e superiori sinistri non è da trascurare.

Per individuare eventuali raggruppamenti all'interno del dataset, verrà esplorato il concetto di clustering come metodo chiave per raggruppare efficacemente i dati, offrendo così una prospettiva più approfondita e strutturata sulla complessità dei dati.

1.2 Introduzione al clustering

Le tecniche di clustering sono numerose e si applicano ad una vasta gamma di contesti e condizioni, risultando particolarmente utili nei casi in cui non è necessario predire classi

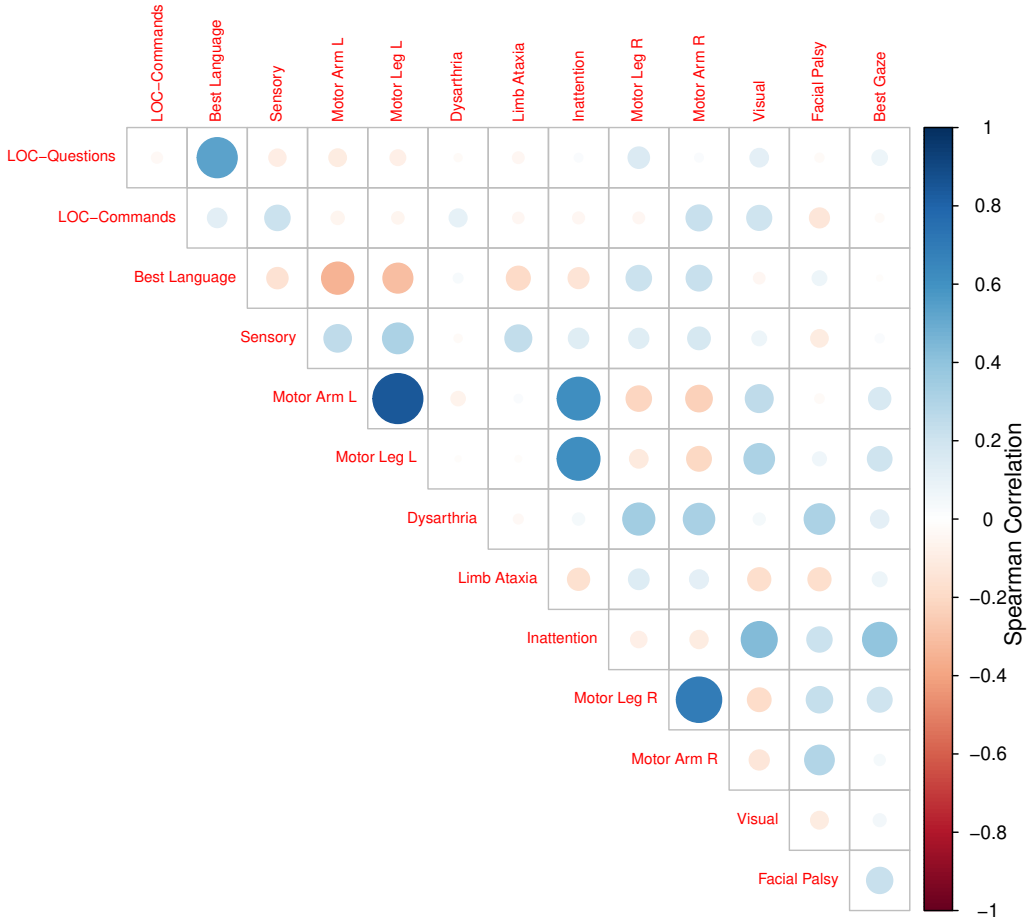


Figura 1.1: Correlazione di Spearman tra i deficit dei pazienti ischemici

specifiche, ma piuttosto si desidera esplorare la naturale suddivisione dei dati. L'obiettivo principale infatti, è la suddivisione degli oggetti in gruppi che riflettano in modo significativo la struttura sottostante dei dati.

I cluster possono assumere diverse forme, ad esempio possono essere tra loro disgiunti, sovrapposti, deterministici, probabilistici, piatti e gerarchici, offrendo una varietà di prospettive per interpretare la distribuzione dei dati. Gli algoritmi utilizzati per formare questi cluster possono anch'essi presentare diverse nature, classificandosi come divisivi o agglomerativi, a seconda del loro approccio nell'assegnare gli oggetti ai gruppi.

Il clustering di dati è stato ampiamente studiato nella letteratura di data mining e di machine learning e si tratta di un metodo di apprendimento non supervisionato il che significa che non richiede etichette o classi predefinite per gli oggetti nel dataset.

Il clustering consiste nel raggruppare un insieme di oggetti in sottoinsiemi definiti *cluster* in modo che gli oggetti all'interno di ciascun gruppo siano più simili tra loro rispetto agli oggetti in altri cluster. Questi gruppi dovrebbero soddisfare due criteri fondamentali: la similarità intra-cluster, ossia gli oggetti all'interno dello stesso gruppo devono essere simili e omogenei, e la dissimilarità inter-cluster, ovvero i gruppi diversi devono contenere oggetti eterogenei.

Esistono principalmente due macro categorie di clustering: l'approccio parametrico e non parametrico. Nell'approccio parametrico, l'analisi è guidata da ipotesi rigide sulla struttura dei dati, come la forma convessa dei cluster e la distribuzione nota a meno di parametri come media e deviazione standard. Questo approccio richiede anche la specificazione del numero di cluster come parametro di input e la stima dei parametri delle distribuzioni di densità basata sui dati stessi. Tuttavia, l'assunzione di una struttura specifica e la necessità di specificare il numero di cluster possono limitare la flessibilità e l'adattabilità del modello ai dati reali. Al contrario, l'approccio non parametrico basato sulla densità, non effettua ipotesi rigide sulla forma dei cluster o sulla distribuzione dei dati. Viene fornita un' ampia descrizione di diverse tipologie e tecniche di clustering in Wasserman (2017) e Shah et al. (2012).

1.3 Procedura generale dell'analisi del clustering

Il clustering può essere applicato a una vasta gamma di contesti, dai dati testuali ai social media, dalle reti sociali ai dati biologici e neurologici. La sua complessità deriva dalla sua dipendenza dal dominio dei dati e dal contesto del problema. Esplorare il clustering implica l'utilizzo della conoscenza del dominio dei dati e l'esplorazione di varie soluzioni possibili (Reddy 2018).

In generale, la procedura di clustering segue una serie di passaggi ben definiti. Il primo passo consiste nell'analisi dei dati identificando i pattern e nell' eventuale selezione delle caratteristiche degli oggetti più rilevanti per l'analisi. Inoltre si può ricorrere ad una o più trasformazioni dei dati per creare nuove caratteristiche basate su quelle precedenti ma maggiormente informative.

Il secondo passo è definire la misura di distanza o similarità più opportuna ai dati e applicarla per ottenere una nuova matrice di distanze o di similarità. Questa misura è utilizzata per valutare quanto due osservazioni siano simili o vicine l'una all'altra e ne esistono di diversi tipi per potersi adattare a molte tipologie di dato.

Successivamente, si procede con l'effettivo clustering, utilizzando uno dei numerosi algoritmi e metodi esistenti per raggruppare i dati in base alla loro prossimità o similarità. L'obiettivo è creare gruppi omogenei o cluster di dati che condividono determinate caratteristiche o proprietà.

Infine, si possono compiere passaggi facoltativi come l'astrazione dei dati, ad esempio un'ulteriore riduzione della dimensionalità o l'aggregazione dei risultati, e l'analisi dei risultati per valutare qualità ed efficacia del processo.
Per affrontare questi passaggi è fondamentale che chi opera abbia una conoscenza dettagliata del dominio specifico e dell'insieme dei dati, poichè queste informazioni risultano molto utili per migliorare la qualità dell'estrazione delle caratteristiche e quindi della trasformazione dei dati, del calcolo della similarità, del raggruppamento e della rappresentazione dei

cluster.

1.4 Problemi nel clustering

L'analisi dei cluster è un campo ricco di complessità dovute alla diversità dei dati e degli obiettivi. I gruppi individuati attraverso l'analisi dei cluster possono derivare da una vasta gamma di forme di dati, infatti possono essere costruiti a partire da valori di variabili, matrici di similarità e dissimilarità, pesi sugli archi in un grafo o altre rappresentazioni. Inoltre, i gruppi possono essere organizzati in una partizione dell'insieme di oggetti, ma possono anche essere sovrapposti o non esaustivi, di conseguenza l'appartenenza ai gruppi può variare da chiaramente definita a meno evidente.

In letteratura non esiste una definizione universalmente accettata di "veri" cluster perché l'analisi di questi viene utilizzata con differenti obiettivi e i ricercatori hanno idee diverse su cosa dovrebbe far appartenere gli oggetti allo stesso gruppo. Ad esempio, alcuni potrebbero privilegiare l'omogeneità all'interno dei cluster, mentre altri potrebbero dare più importanza alla separazione tra i gruppi (Hennig 2015). Alcune definizioni sono state fornite ad esempio da Wishart (1969) che sottolineava l'importanza dei metodi di clustering nell'identificare modalità distinte dei dati, indipendentemente dalla loro forma e varianza e da Hartigan & Lada (1985) descrivendo i cluster come regioni di alta densità separate da altre regioni di densità bassa.

Inoltre è importante anche considerare la complessità dei dati e l'incertezza associata ad essi. Spesso i dati reali contengono rumore o outlier che incidono sui risultati del clustering identificando cluster significativi anche se non sono tali.

Di conseguenza i risultati dell'analisi dei cluster non sono mai inequivocabilmente definiti, ma piuttosto soggetti a interpretazione e soggettività. Non esiste una regola generale che garantisca la validità dei cluster, poiché la definizione stessa di cosa costituisca un cluster e la sua interpretazione possono variare. Si può concludere affermando che la soggettività e l'interpretazione sono componenti intrinseche dell'analisi dei cluster.

Un altro aspetto importante da considerare è la valutazione della presenza o meno dei cluster all'interno dei dati considerati. Infatti, in linea generale tutti gli algoritmi di clustering tenderanno a produrre cluster anche se i dati possono non contenerne, perciò è importante valutare se i cluster identificati sono effettivamente significativi e riflettono la struttura reale nei dati stessi. A tal fine è utile esaminare i dati per determinare se esiste una tendenza al clustering. Questo significa verificare se i dati mostrano evidenze di raggruppamenti o pattern distinti che potrebbero essere identificati tramite clustering.

Inoltre molti algoritmi di clustering richiedono che vengano specificati alcuni parametri come ad esempio il numero di gruppi o la distanza critica per determinare se due punti sono sufficientemente simili da essere raggruppati o meno. La scelta di questi parametri può influenzare significativamente i risultati del clustering e la struttura dei raggruppa-

menti ottenuti. Inoltre la scelta di valori inappropriati per questi parametri può portare a raggruppamenti non ottimali o addirittura errati rendendo la scelta dei parametri un aspetto fondamentale e un'ulteriore sfida in ambito di clustering.

Tuttavia, la selezione dei parametri non è l'unica decisione da prendere nell'ambito del clustering. Come illustrato in precedenza durante i passaggi dell'analisi generale del clustering, è essenziale scegliere una specifica metrica di similarità poichè anche questa selezione può influenzare notevolmente i risultati del clustering, richiedendo un'attenta considerazione della tipologia e del dominio dei dati.

Ai problemi precedentemente elencati se ne aggiungono altri quando il numero di dimensioni supera la bidimensionalità, una situazione comune nella realtà poichè la maggior parte dei problemi coinvolge dati di dimensioni superiori.

Mentre in una rappresentazione unidimensionale o bidimensionale è possibile fare affidamento ad un interpretazione intuitiva dei dati attraverso una visualizzazione grafica, questo non avviene quando si superano le tre dimensioni, rendendo difficile anche l'interpretazione visiva dei risultati del clustering.

Inoltre rispetto al contesto univariato c'è una disponibilità inferiore di tecniche applicative.

Nei casi in cui si utilizzano metodi non parametrici subentra la maledizione della dimensionalità ossia, all'aumentare della dimensione, i dati si disperdono nello spazio e diventano sempre più sparsi e quindi distanti tra loro, rendendo più difficile identificare relazioni significative e strutture nei dati. Questo fenomeno comporta che alcune misure di distanza tra gli oggetti possono diventare meno informative a causa della maldizione della dimensionalità.

Inoltre, un problema comune ad esempio nel contesto del clustering di dati multimediali è la loro inefficacia nel gestire vettori di caratteristiche ad alta dimensionalità. La maggior parte degli algoritmi non è progettata per lavorare con dati di elevata dimensionalità, e la loro performance diminuisce rapidamente con l'aumento delle dimensioni o addirittura non funzionano affatto. Inoltre, pochi algoritmi possono gestire database contenenti grandi quantità di rumore Hinneburg & Keim (1998).

Infine, la complessità computazionale degli algoritmi di clustering aumenta significativamente con l'aumento delle dimensioni del dataset, poiché il numero di calcoli e confronti cresce con il numero di variabili.

Nei seguenti capitoli verranno fornite alcune metodologie per la risoluzione di alcune di queste problematiche, permettendo l'analisi dei dati in questione.

Capitolo 2

Trasformazione dei dati

Il tipo di dati analizzati influisce significativamente sulla scelta della procedura di clustering. Spesso gli algoritmi si basano sull'assunzione implicita che i dati siano numerici continui, tuttavia nella realtà questo non sempre si verifica. I dati infatti potrebbero appartenere a diverse tipologie come ad esempio dati discreti, qualitativi, temporali o spaziali. Nel caso specifico dei punteggi NIHSS, essi sono considerati dati quantitativi, in quanto rappresentano misurazioni numeriche delle condizioni neurologiche del paziente, tuttavia, è importante notare non sono valori continui, ma sono spesso trattati come variabili ordinali. Questo è dovuto al fatto che i valori più alti indicano un maggiore grado di disabilità o compromissione neurologica, ma non necessariamente sono equidistanti o distribuiti uniformemente.

È essenziale sviluppare funzioni di similarità adeguate per comparare i punteggi NIHSS tra i pazienti, dato che questo aspetto riveste un ruolo cruciale nella successiva fase di clustering. Durante tale fase, l'obiettivo primario è individuare pattern o gruppi tra i pazienti, basandosi sui loro punteggi NIHSS.

Nel caso di variabili ordinali, le uniche relazioni possibili sono "uguale a", "maggiore di", "minore di", pertanto la costruzione della misura di distanza per questo tipo di dati dovrebbe considerare tali relazioni e basarsi sulle relazioni tra i due oggetti analizzati e gli altri oggetti. Nella letteratura sull'analisi dei dati statistici una delle misure che tiene conto di questi principi è la misura di distanza generalizzata Walesiak & Dudek (2010).

2.1 Misura di distanza generalizzata

Esistono numerose definizioni di misura di distanza generalizzata, in questa sezione verrà esposta una distanza generalizzata basata sul seguente coefficiente di correlazione generalizzato.

Considerando i vettori di osservazioni $(x_{1j}, ..., x_{nj})$, $(x_{1h}, ..., x_{nh})$ e le due variabili j e h

viene presentato il coeffiente di correlazione generalizzato:

$$\Gamma_{jh} = \frac{\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ikj}b_{ikh}}{[\sum_{i=2}^{n}\sum_{k=1}^{i-1} a_{ikj}^2 \sum_{i=2}^{n}\sum_{k=1}^{i-1} b_{ikh}^2]^{1/2}}, \tag{2.1}$$

dove $i, k = 1, ..., n$ con n numero di oggetti mentre $j, h = 1, ..., m$ con m numero di variabili, mentre $a_{ikj} = (x_{ij} - x_{kj}), \quad b_{ikh} = (x_{ih} - x_{kh})$.
La funzione di distanza $d : A \times A \to \mathbb{R}$, dove A è l'insieme degli oggetti e $\mathbb{R}$ è l'insieme dei numeri reali, deve rispettare alcuni vincoli imposti per poter essere definita una misura di distanza:

- non negativa: $d_{ik} \geq 0$;
- riflessiva: $d_{ik} = 0$ quando $i = k$;
- simmetrica: $d_{ik} = d_{ki}$;

per $i, k = 1, ..., n$. Tuttavia il coefficiente di correlazione generalizzato non soddisfa i vincoli di non negatività e riflessività, pertanto è opportuno effettuare alcune trasformazioni. Per quanto riguarda la non negatività sarà sufficiente porre $d_{ik} = (1-\Gamma_{ik})/2$ in modo da rendere i valori di d_{ik} all'interno dell'intervallo $[0, 1]$. Invece, non è possibile soddisfare l'assunzione di riflessività per questo viene presa in considerazione la distanza generalizzata che riesce ad unire tutte e tre le assunzioni citate in precedenza. Questo tipo di distanza è basata sul coefficiente di correlazione generalizzato e viene formulata nella seguente espressione:

$$d_{ik} = \frac{1 - s_{ik}}{2} = \frac{1}{2} - \frac{\sum_{j=1}^{m} a_{ikj}b_{kij} + \sum_{j=1}^{m}\sum_{l=1;l\neq i,k}^{n} a_{ilj}b_{klj}}{\left[\sum_{j=1}^{m}\sum_{l=1}^{n} a_{ilj}^2 \sum_{j=1}^{m}\sum_{l=1}^{n} b_{klj}^2\right]^{1/2}}, \tag{2.2}$$

considerando $x_{ij}(x_{kj}, x_{lj})$ i-esima(k-esima, l-esima) osservazione della j-esima variabile con $i, k, l = 1, ..., n$, dove n è il numero degli oggetti, $j = 1, ..., m$ con m numero delle variabili.
Per le variabili misurate su scale continue o in intervalli vengono considerati i termini a_{ipj} e b_{kij} come i seguenti:

$$\begin{aligned} a_{ipj} &= x_{ij} - x_{pj} \text{ per } p = k, l, \\ b_{krj} &= x_{kj} - x_{rj} \text{ per } r = i, l. \end{aligned} \tag{2.3}$$

Considerando invece la scala ordinale è possibile solamente confrontare tra loro i valori che assumono le variabili per ogni osservazione. I termini a_{ipj} e b_{krj} avranno la seguente forma:

$$a_{ipj}(b_{kij}) = \begin{cases} 1 \text{ se } x_{ij} > x_{pj}(x_{kj} > x_{rj}) \\ 0 \text{ se } x_{ij} = x_{pj}(x_{kj} = x_{rj}) \text{ per } \ p = k, l; \ r = i, l. \\ -1 \text{ se } x_{ij} < x_{pj}(x_{kj} < x_{rj}) \end{cases} \tag{2.4}$$

Quindi nel denominatore della formula (2.2) il primo fattore indica il numero dei rapporti "maggiore di" e "minore di" per l'oggetto i e il secondo fattore è il numero dei rapporti "maggiore di " e "minore di " per oggetto k. In conclusione emerge che è utilizzata solamente l'idea del coefficiente di correlazione generalizzato poichè le indicazioni per la costruzione della misura (2.2) con l'utilizzo di (2.3) e (2.4) sono rispettivamente il coefficiente di correlazione di Pearson e il coefficiente tau di Kendall.

2.1.1 Proprietà della misura della distanza generalizzata

La distanza proposta nella sezione 2.1 presenta alcune proprietà:

- può essere applicata quando le variabili vengono misurate su scala ordinale, intervallare o proporzionale;
- assume valori nell'intervallo [0,1] dove 0 indica che per gli oggetti confrontati i e k la relazione che li lega è solamente "uguale a"; mentre 1 indica che gli oggetti i, k comparati si verificano relazioni di tipo "maggiore di ";
- soddisfa le condizioni $d_{ik} \geq 0, d_{ii} = 0, d_{ik} = d_{ki}$ per tutti gli i,k=1,...n;
- l'analisi empirica dimostra che la distanza a volte non soddisfa la disuguaglianza triangolare;
- necessita di almeno una coppia di oggetti non identici per evitare lo zero al denominatore;
- la trasformazione dei dati con una qualsiasi funzione strettamente crescente (formula (2.4)) o con qualsiasi funzione lineare (formula (2.3) non modifica il valore della distanza d_{ik}.

La distanza (2.2) tiene conto delle variabili ponderate equamente, ma in caso i pesi non siano uguali la definizione di distanza viene definita come:

$$d_{ik} = 1/2 - \frac{\sum_{j=1}^{m} w_j a_{ikj} b_{kij} + \sum_{j=1}^{m} \sum_{l=1}^{n} (w_j a_{ilj} b_{klj})}{[\sum_{j=1}^{m} \sum_{l=1}^{n} w_j a_{ilj}^2 \sum_{j=1}^{m} \sum_{l=1}^{n} w_j b_{klj}^2]^{1/2}}, \tag{2.5}$$

dove i pesi w_j con $j = 1, ..., m$ $w_j \in (0; m), \sum_{j=1}^{m} w_j = m$.

Uno studio della misura di distanza generalizzata è approfondito in Jajuga et al. (2003).

Una volta calcolata la distanza generalizzata è difficile immaginare la forma dello spazio, per questo i dati verranno trattati attraverso i grafi usufruendo dello *spectral embedding* per la trasformazione dei dati.

2.2 Embedding

Come già discusso nella sezione 1.4 quando la dimensione dei dati aumenta, il clustering può comportare diverse complicazioni.

Un approccio efficace per affrontare tali problemi è l'embedding dei dati, che mira a ridurre la dimensionalità rappresentando i dati in uno spazio a dimensioni inferiori, pur mantenendo la struttura e le relazioni intrinseche tra gli oggetti. Uno dei vantaggi principali è che rende la visualizzazione dei dati molto più chiara ed efficace consentendo un'interpretazione più semplice. Un'altro vantaggio dell'embedding è la riduzione degli effetti della maledizione della dimensionalità, poiché con la riduzione delle dimensioni dei dati si ottengono spazi di dati più densamente popolati e più significativi per l'analisi. Inoltre, l'embedding dei dati può contribuire a ridurre la complessità computazionale degli algoritmi di clustering, rendendo il processo di clustering più efficiente e meno oneroso computazionalmente.

2.2.1 *Spectral embedding*

In aggiunta alla riduzione della dimensionalità dei dati, i metodi spettrali mirano a mappare i dati originali in uno spazio euclideo a dimensionalità inferiore.

Questo significa che non solo cercano di ridurre la complessità dei dati, ma anche di proiettarli in uno spazio in cui le relazioni tra gli oggetti sono più facilmente comprensibili ed analizzabili. In altre parole, i metodi spettrali cercano di trovare una rappresentazione degli elementi in cui le distanze euclidee tra i punti corrispondono alle similarità o alle distanze tra i dati originali.

Piuttosto di lavorare direttamente con i punti dati originali e le loro dimensioni, i metodi spettrali operano su una matrice che rappresenta la similarità o la distanza tra i dati. Questo consente di lavorare con le relazioni tra i dati anziché con i dati stessi.

Uno dei principali vantaggi è la capacità di lavorare con oggetti arbitrari per la riduzione della dimensionalità, consentendo di applicare la riduzione della dimensionalità a una vasta gamma di dati. Tuttavia, vi sono anche degli svantaggi, come la complessità computazionale, poichè il tempo richiesto per creare la matrice di similarità aumenta rapidamente con il numero di punti dati, e la complessità del processo di determinazione degli autovettori di questa matrice.

Lo *spectral embedding* risulta particolarmente utile per eseguire il clustering su oggetti arbitrari, come insiemi di nodi in un grafo. Questo significa che può essere utilizzato per identificare gruppi o pattern in una vasta gamma di dati.

Si consideri un database contenente n osservazioni, il primo passaggio è quello di creare una matrice $n \times n$ dei pesi che rappresenti la similarità tra i diversi punti. Dato un insieme di punti $x_1, x_2, ..., x_n$ e una nozione di similarità $s_{ij} \geq 0$ tra tutte le coppie di punti x_i e x_j, un modo efficace di rappresentare i dati è sotto forma di grafo di similarità $G = (V, E)$ dove V è l'insieme dei nodi o vertici e E l'insieme degli archi. Ogni vertice v_i in questo

grafo rappresenta un punto dato x_i. Due vertici sono connessi se la similarità s_{ij} tra i punti dati corrispondenti x_i e x_j è positiva o maggiore di una certa soglia, e l'arco è pesato da s_{ij}. All'interno delle analisi riguardanti i dati di pazienti ischemici verrà considerata la matrice di similarità:

$$s_{ij} = 1 - d_{ij}, \tag{2.6}$$

dove d_{ij} è la distanza generalizzata definita in precedenza in 2.2. Si vuole trovare una partizione del grafo tale che i bordi tra gruppi diversi abbiano pesi molto bassi, ossia i punti in diversi cluster siano meno simili tra loro e i bordi all'interno di un gruppo abbiano pesi elevati ovvero i punti all'interno dello stesso cluster siano simili tra loro.

Innanzitutto viene definito $G = (V, E)$ un grafo non diretto con insieme di vertici $V = \{v_1, v_2, ...v_n\}$ e pesato, cioè ogni arco tra due nodi v_i e v_j porta un peso non negativo $w_{ij} \geq 0$. La matrice di adiacenza non negativa e pesata del grafo è la matrice $W = (w_{ij})_{i,j=1,...,n}$ e se $w_{ij} = 0$, significa che i vertici v_i e v_j non sono connessi da un arco. Inoltre, poiché G è non diretto, si richiede che $w_{ij} = w_{ji}$. Il grado di un vertice $v_i \in V$ è definito come:

$$d_i = \sum_{j=1}^{n} w_{ij}. \tag{2.7}$$

La matrice dei gradi D è definita come la matrice diagonale con i gradi $d_1, \ldots, d_n$ sulla diagonale.

Per creare la matrice W del grafo è possibile utilizzare diversi metodi a partire dalla matrice delle distanze tra i dati. Quello utilizzato nell'analisi dei dati dei pazienti ischemici è l'ε *neighborhood graph* che consiste nel connettere i punti che sono vicini tra loro, ovvero quelli le cui distanze reciproche sono più piccole di un certo valore ε, chiamato raggio. Quest'ultimo definisce la distanza massima a cui due punti possono essere considerati vicini. Quando si considera questo tipo di grafo, le distanze tra i punti connessi sono generalmente tutte entro lo stesso ordine di grandezza, poiché sono limitate da ε. Di conseguenza, assegnare un peso agli archi, ovvero dare maggiore importanza alle connessioni tra certi punti rispetto ad altre, non aggiungerebbe molte informazioni al grafo, poiché tutte le connessioni sono già abbastanza vicine tra loro. Di solito si considera l'ε *neighborhood graph* come un grafo non pesato, dove ogni connessione ha lo stesso peso, indicando semplicemente la presenza o l'assenza di un collegamento tra i punti, tuttavia è comunque possibile considerare la sua versione pesata. É importante specificare che se ε non viene scelta con attenzione, si rischia di creare grafi con componenti sconnesse. Una variante dell' ε *neighborhood graph* è la sua versione che considera il kernel, ossia:

$$W_{ij} = \exp\left(-\frac{||d_{ij}||^2}{h}\right), \tag{2.8}$$

dove h è il parametro deciso dall'utente. Inoltre viene mantenuta l'idea che se la distanza tra due osservazione d_{ij} risulta maggiore rispetto ad una soglia, il valore della distanza verrà posto pari a 0.

Altri grafi di similarità sono il grafo *k-nearest neighbor* e il grafo *fully connected*. Il primo è costruito connettendo i vertici v_i con i vertici v_j se v_j è tra i k vicini di v_i oppure per la sua versione mutuale si connettono i vertici v_i e v_j solo se v_i è tra i k vicini di v_j e v_j è tra i k vicini di v_i. Diversamente il secondo collega tutti i punti con similarità positiva tra loro e assegna un peso a tutti gli archi in base alla similarità s_{ij}.

La matrice di similarità può essere considerata come la matrice di adiacenza di un grafo, in cui ogni nodo corrisponde a un elemento dati, e il peso di un arco corrisponde alla similarità tra questi elementi dati. Il problema quindi si riduce a trovare un "taglio ottimale" nel grafo cioè suddividere il grafo in parti in modo che i nodi all'interno di ciascuna parte siano strettamente connessi tra loro, mentre i nodi tra parti diverse siano debolmente connessi. Questo concetto di "taglio ottimale" può essere raggiunto analizzando gli autovettori della matrice di adiacenza oppure della matrice Laplaciana del grafo. Di conseguenza i metodi spettrali possono essere considerati una tecnica basata su grafi per il clustering di qualsiasi tipo di dati, mediante la conversione della matrice di similarità in una struttura di rete. Come discusso nel paragrafo 1.4 uno dei problemi del clustering era legato al rumore nei dati che incide sui risultati del clustering identificando cluster significativi anche se non sono tali. La matrice di similarità è uno strumento potente anche in presenza di distanze rumorose, poiché codifica una vasta quantità di informazioni sui dati. L'analisi spettrale si rivela particolarmente utile in questo contesto, in quanto consente di estrarre il rumore dalla rappresentazione della similarità attraverso l'uso degli autovettori della matrice.

Una volta ottenuto il grafo di similarità G e la rispettiva matrice di adiacenza W il passo successivo consiste nel calcolare le matrici Laplaciane del grafo. La costruzione effettiva dell'embedding non è definita in modo univoco nella letteratura poichè ogni autore varia il metodo per creare la matrice Laplaciana del grafo. Si esploreranno brevemente due tipologie: la Laplaciana non normalizzata e quella normalizzata, quest'ultima con due possibili varianti. Il grafo Laplaciano non normalizzato è definito come

$$L = D - W, \tag{2.9}$$

esso non dipende dagli elementi diagonali della matrice di adiacenza W. Infatti ogni matrice di adiacenza che coincide con W su tutte le posizioni al di fuori della diagonale porta alla stessa Laplaciana non normalizzata del grafo L. In particolare, i *self-edge* in un grafo non cambiano la corrispondente Laplaciana del grafo. Per quanto riguarda il grafo Laplaciano normalizzato ci sono due diverse matrici chiamate entrambe così. Sono strettamente correlate tra loro e sono definite come:

$$L_{\text{sym}} := D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}, \tag{2.10}$$

$$L_{\text{rw}} := D^{-1} L = I - D^{-1} W, \tag{2.11}$$

dove D è la matrice diagonale dei gradi dei nodi e L è la Laplaciana non normalizzata. La prima è una matrice simmetrica mentre la seconda è strettamente correlata a un random

walk sul grafo, poiché riflette la probabilità di spostarsi da un nodo all'altro. Con random walk su un grafo si intende un processo casuale che si muove attraverso i nodi del grafo seguendo connessioni casuali tra i nodi.

La scelta di quale di questi tre grafi Laplaciani usare per calcolare gli autovettori è influenzata dalla distribuzione dei gradi nel grafo di similarità. Se il grafo è uniforme e i gradi dei nodi sono simili, tutte le Laplaciane produrranno risultati simili e funzioneranno in modo efficace per il clustering. Diversamente, se i gradi dei nodi sono ampiamente distribuiti ossia alcuni nodi hanno un numero molto elevato di connessioni, mentre altri ne hanno poche i risultati dei grafi Laplaciani possono differire notevolmente poiché ciascuna Laplaciana tiene conto della distribuzione dei gradi in modo diverso. Secondo Von Luxburg (2007) ci sono diversi argomenti a favore dell'uso del clustering spettrale normalizzato piuttosto che non normalizzato, e nel caso normalizzato è meglio utilizzare gli autovettori di L_{rw} piuttosto che quelli di L_{sym}. Seguendo i suggerimenti, nell'analisi di pazienti ischemici è stata utilizzata la matrice L_{rw}. Verrà quindi approfondita la procedura per ottenere gli autovettori della matrice L_{rw}.

L'algoritmo utilizzato per ottenere L_{rw} effettua i seguenti passaggi:

1. Viene costruito un grafo di similarità utilizzando uno dei metodi elencati in precedenza (nel caso di pazienti ischemici il grafo ε *neighborhood*). Viene definita W come la matrice di adiacenza del grafo e D la matrice dei gradi;

2. Si calcola il Laplaciano non normalizzato L;

3. Vengono considerati i primi k autovettori $u_1, ..., u_k$ di $Lu = \lambda Du$;

4. Si considera $U \in \mathbb{R}^{n \times k}$ la matrice che contiene gli autovettori $u_1, ..., u_k$ come colonne.

Questo algoritmo utilizza gli autovettori generalizzati di L, che corrispondono agli autovettori della matrice L_{rw}. Questo cambiamento di rappresentazione migliora le proprietà dei cluster nei dati, in modo che i cluster possano essere rilevati in modo semplice nella nuova rappresentazione.

Un'ampia descrizione dell'embedding e dello *spectral embedding* è fornita in Von Luxburg (2007), Reddy (2018) e Ng et al. (2001).

2.3 Applicazione ai dati in analisi

Una volta ottenuta la distanza generalizzata ai dati relativi ai pazienti ischemici, si ottiene una matrice $n \times n$ ossia 172×172, contenente tutte le distanze o connessioni tra i pazienti. Utilizzando il grafo dell' ε *neighborhood*, vengono eliminate, ovvero poste a 0 tutte le connessioni tra i pazienti che presentano una distanza maggiore di ε. La scelta di ε è fondamentale per il problema e di conseguenza cambieranno anche i risultati ottenuti al variare di tale parametro soglia. Questa scelta è stata valutata prendendo in considerazione

la distribuzione dei valori delle distanze tra i pazienti. Le distanze variano nell'intervallo $[0, 1]$, tuttavia il massimo valore raggiunto in termini di distanza tra pazienti è 0.6, con una media di 0.25 circa come mostrato in figura 2.1.

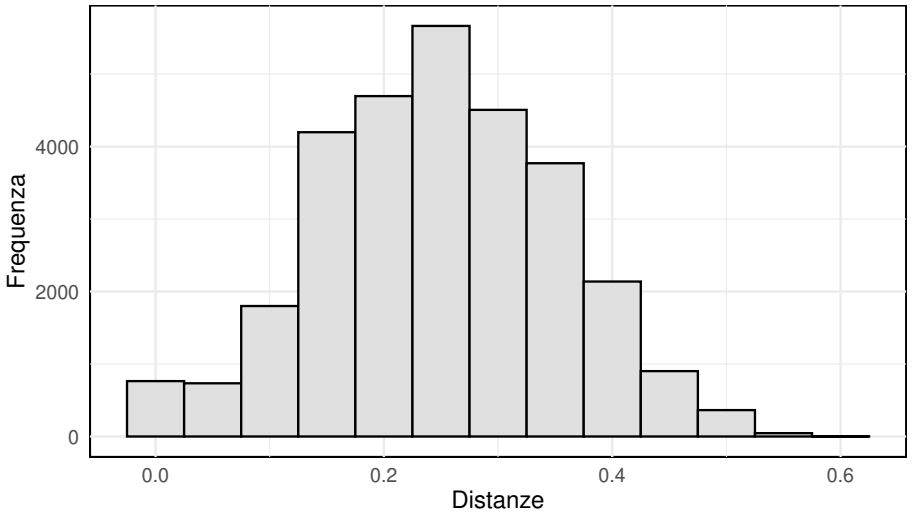


Figura 2.1: Istogramma delle distanze ottenute dalla distanza generalizzata sui dati di pazienti ischemici.

É importante considerare però, come è stato citato precedentemente, che alcune scelte del valore di ε possono creare grafi con componenti non connesse, ossia pazienti non connessi a nessun'altro paziente. Questa situazione nel caso dei dati relativi ai pazienti colpiti da ictus si verifica con una soglia inferiore a 0.18 e un esempio di cosa accade se dovessimo ulteriormente abbassare la soglia è proposto in figura 2.2. In questo caso infatti, al diminuire della soglia ci saranno sempre più pazienti che non presentano connessioni. Per l'analisi di clustering verranno quindi utilizzati $\varepsilon \geq 0.18$, considerando che all'aumentare della soglia verranno considerate sempre più connessioni. Una situazione analoga si è verificata quando è stata considerata la variante del grafo dell'ε *neighborhood* che includeva il kernel, i risultati al variare di ε rimangono gli stessi, quindi si è preferito considerare la prima versione per semplicità.

A questo punto si è proceduto nella costruzione dell'embbedding ottenendo la matrice L_{rw}, la quale dipenderà anch'essa dal valore di ε scelto in precedenza e dalla quale sono stati estratti gli autovettori. Vengono considerati i primi 4 autovettori eliminando il primo poichè costante per costruzione. Quindi verrà considerato lo spazio tridimensionale creato dal secondo, terzo e quarto autovettore della matrice L_{rw} per efettuare l'operazione di clustering. La rappresentazione tridimensionale è una scelta legata alla miglior visualizzazione e interpretabilità.

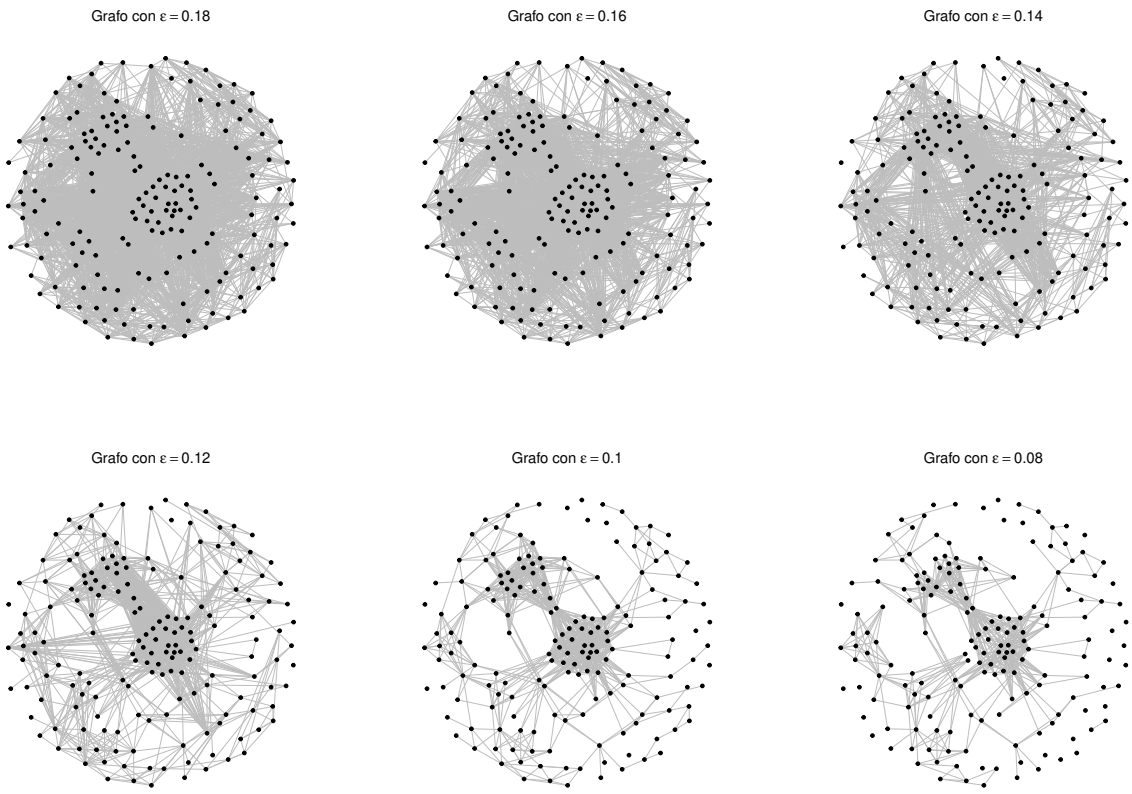


Figura 2.2: Grafo ε-neighborhood al variare di ε.

Capitolo 3

Clustering basato sulle densità

Nel corso degli anni, c'è stata un'evoluzione dei metodi di clustering nella letteratura con una vasta gamma di approcci e tecniche adattabili a diverse tipologie di dati, contesti applicativi e obiettivi di clustering. Questo studio si focalizzerà su uno dei metodi più significativi scoperti dagli anni '90 in poi: il clustering basato sulla densità. Questo approccio ha dimostrato di essere particolarmente efficace nella gestione di dati complessi.

Esso è un metodo che si concentra sulla scoperta di regioni di alta densità nei dati e sulla loro suddivisione in cluster. Uno dei principali vantaggi di questo approccio è la sua natura non parametrica che elimina la necessità di specificare il numero di cluster a priori. Questo rende l'approccio estremamente flessibile, consentendo ai cluster di assumere forme diverse nello spazio dei dati e di adattarsi meglio a strutture complesse.

Uno dei più diffusi algoritmi per l'identificazione dei cluster in questo contesto è DBSCAN (Density-Based Spatial Clustering of Applications with Noise) che identifica regioni dense nello spazio dei dati e considera i punti in queste regioni come facenti parte dello stesso cluster. Inoltre è particolarmente efficace nella gestione di cluster di forma arbitraria e dimensioni diverse e nella presenza di rumore nei dati, cioè i punti che non appartengono a nessun cluster specifico. Un altro algoritmo è DENCLUE (Density-Based Clustering using Clue Discovery), un'estensione di DBSCAN che introduce il concetto di funzione di influenza e utilizza le informazioni sulla densità locale per identificare cluster nello spazio dei dati e si basa sull'analisi delle proprietà di densità dei punti. Infine un altro algoritmo di clustering basato sulla densità è il *mean shift* che trova i massimi locali nella densità dei dati. Questo algoritmo non richiede la specificazione del numero di cluster a priori ed è efficace nel trovare cluster con forme eterogenee e nella gestione di dati non lineari. Nelle ricerche precedentemente condotte sul tema, sono stati proposti numerosi ulteriori algoritmi, tuttavia in seguito si farà riferimento solamente a quelli di interesse per i metodi di clustering che verranno presentati.

Nella letteratura sul clustering basato sulla densità, emergono concetti chiave essenziali

per la definizione e l'implementazione di vari metodi. Tra questi emerge la suddivisione dei dati mediante i metodi della divisione orizzontale, un metodo che consente di identificare e suddividere regioni dei dati, contribuendo così alla formazione dei cluster. Esso prevede la separazione orizzontale della distribuzione dei dati utilizzando una soglia di densità e si traccia una linea orizzontale che divide la distribuzione dei dati come mostrato in figura 3.1.

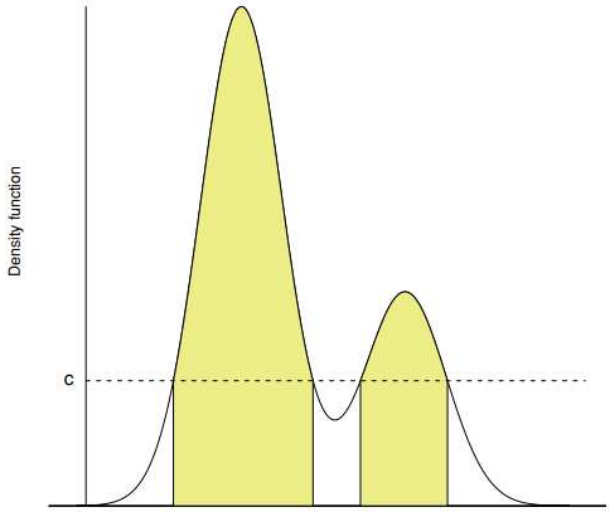


Figura 3.1: Esempio di suddivisione orizzontale della densità nel caso $d = 1$ secondo una soglia c, in questo modo è possibile individuare gli insiemi di punti per cui la densità è maggiore di c, Azzalini & Torelli (2007).

Tutti i punti con densità di probabilità superiore a c formano un cluster. Abbassando la soglia, alcuni punti che erano inizialmente considerati singoli cluster potrebbero unirsi ad altri punti vicini con densità di probabilità più alta. In Figura 3.1 viene illustrato il caso di una densità unidimensionale come presentato in Azzalini & Torelli (2007), tuttavia la situazione diventa più complessa quando si considerano più dimensioni.

Nel contesto dell'analisi della densità, è fondamentale definire alcune notazioni chiave che saranno utilizzate nel corso di questo studio. Si consideri $X_1, ...X_n$ un campione casuale d-variato da una funzione di densità p, il simbolo $g(x)$ viene utilizzato per rappresentare il gradiente di p in x mentre $\mathscr{H}(x)$ indica l'Hessiana di p in x. Si assume che la funzione di densità $p(x)$ sia differenziabile due volte rispetto al vettore delle variabili casuali x, questo implica che le derivate parziali della funzione esistono e sono continue per due volte. Il gradiente di p fornisce informazioni sulla direzione di massimo incremento della densità, mentre l'Hessiana fornisce dettagli sulla curvatura della densità in quel punto. Analogamente la stima della media di una densità kernel con *bandwidth* h verrà definita

con $\hat{p}_h$, di conseguenza la corrispondente Hessiana è indicata con $\mathscr{H}_h$, mentre viene definita come $\mathscr{H}_{Z,h}$ quando è necessario rendere chiaro il dataset Z. In maniera analoga l'hessiana di uno stimatore $\hat{p}_h$ viene definita $\hat{\mathscr{H}}_h$.

Per stimare la densità della distribuzione, viene utilizzato uno stimatore di densità kernel definito come:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h^d}K\left(\frac{x - X_i}{h}\right), \tag{3.1}$$

dove n è la numerosità campionaria, K è una funzione kernel simmetrica e liscia, $h > 0$ è la larghezza di banda detta *bandwidth*.

La media dello stimatore di densità è data da:

$$\hat{p}_h(x) = E[\hat{p}_h(x)] = \int K(t)p(x + th)\,dt, \tag{3.2}$$

che rappresenta la stima della densità media attorno al punto x, considerando la distribuzione dei dati nella vicinanza di x.

In generale, possiamo utilizzare una matrice di larghezza di banda H nello stimatore di densità:

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

dove K è una funzione kernel di densità di probabilità simmetrica, mentre H è una matrice di larghezza di banda simmetrica e definita positiva e K_H è la funzione di kernel scalata definita come $K_H(x) = |H|^{-\frac{1}{2}}K(H^{-\frac{1}{2}}x)$ dove $|H|$ è il determinante della matrice H.

All'interno dello studio in questione saranno di fondamentale importanza la derivata prima e seconda. La stima della derivata prima del kernel è definita come:

$$\hat{g}(x) = \widehat{\nabla^{(1)}p(x)} = \frac{1}{n}\sum_{i=1}^{n}\nabla K_H(x - X_i), \tag{3.4}$$

dove ∇ è il vettore colonna delle d derivate parziali di primo ordine e

$$\nabla K_H(x) = |H|^{-\frac{1}{2}}H^{-\frac{1}{2}}\nabla K\left(H^{-\frac{1}{2}}x\right). \tag{3.5}$$

Per convenzione si considerano le derivate dopo il ridimensionamento con la matrice della larghezza di banda H.

Analogamente lo stimatore della curvatura del kernel verrà definito come:

$$\hat{\mathscr{H}} = \widehat{\nabla^{(2)}p(x)} = \frac{1}{n}\sum_{i=1}^{n}\nabla^{(2)}K_H(x - X_i), \tag{3.6}$$

dove $\nabla^{(2)}$ denota la matrice di tutte le derivate parziali di secondo ordine, e

$$\nabla^{(2)}K_H(x) = |H|^{-\frac{1}{2}}H^{-\frac{1}{2}}\nabla^{(2)}K\left(H^{-\frac{1}{2}}x\right)H^{-\frac{1}{2}}. \tag{3.7}$$

3.1 Problematiche nel testare le mode

Nel complesso panorama dell'analisi delle mode presenti nei dati, che richiede la valutazione di quantità, tipologie e posizioni delle stesse, emergono una serie di problematiche. Innanzitutto un primo problema sorge nel considerare l'ipotesi nulla che un punto x non sia una moda, ossia che non sia un punto di massimo locale della densità, questo comporta delle difficoltà simili a quelle riscontrate nel testare l'ipotesi nulla che la media di una distribuzione sia zero.

Quando si testa l'ipotesi nulla che il punto x non sia una moda, l'alternativa, ovvero l'ipotesi che ci sia almeno una moda, forma un insieme di misura zero nello spazio dei parametri. Questo significa che la probabilità di ottenere esattamente un dato valore per la funzione, ad esempio il gradiente della densità o gli autovalori dell'Hessiana, è teoricamente zero.

In altre parole il problema è che, sebbene l'ipotesi alternativa possa essere vera, la sua probabilità di verificarsi è così piccola che non è possibile costruire un test significativo per rilevarla (Genovese et al. 2016).

Inoltre, un secondo problema che sorge nell'identificazione delle mode all'interno della distribuzione di dati riguarda il fatto che possono potenzialmente esistere un numero infinito di posizioni in cui le mode possono manifestarsi. Questa caratteristica può generare una problematica di rilievo, in quanto richiede la valutazione e il confronto di molteplici posizioni possibili delle mode, aumentando la complessità del processo di identificazione e dell'interpretazione dei risultati. Tale molteplicità di posizioni potenziali per le mode introduce una situazione complessa nei test multipli, rendendo necessaria un'attenta considerazione delle implicazioni nel processo di identificazione delle mode.

Infine, l'identificazione e la verifica dell'esistenza delle mode all'interno della distribuzione dei dati comportano la necessità di fare inferenze sugli autovalori dell'Hessiana, una matrice che contiene le seconde derivate della funzione di densità. Tuttavia, un ulteriore problema si presenta dal fatto che gli autovalori dell'Hessiana non variano continuamente con la matrice stessa ovvero piccoli cambiamenti nella matrice possono causare cambiamenti discontinui negli autovalori. Questa non continuità rende inadatti metodi statistici come il bootstrap e il metodo delta per l'inferenza sugli autovalori dell'Hessiana, poiché questi metodi si basano sull'assunzione di continuità. Di conseguenza, risulta difficile confermare l'esistenza e determinare le proprietà delle mode nella distribuzione dei dati mediante queste tecniche statistiche tradizionali (Genovese et al. 2016).

3.2 Mode e cluster

Mode e cluster sono due concetti strettamente correlati nell'ambito dell'analisi dei dati poichè le mode di una distribuzione di probabilità multivariata sono utilizzate per definire i cluster dei dati. In particolare le mode possono essere utilizzate per raggruppare i dati in modo significativo, evidenziando le loro strutture intrinseche e le relazioni tra di esse. Affinché questa identificazione sia possibile, è necessario che la funzione di densità sia di tipo *Morse*, ovvero che l'Hessiana della distribuzione di probabilità sia non degenere in ogni punto stazionario (Hessiana non singolare). Questo assicura che la direzione di massimo aumento e diminuzione sia ben definita intorno ai punti critici.

Inoltre è importante non siano presenti curve di biforcazione le quali connetterebbero punti critici di diverso indice, compromettendo la struttura ordinata delle linee di flusso intorno ai punti critici. Questo garantisce la struttura "a cascata" delle linee di flusso intorno ai punti critici ovvero il movimento in modo organizzato e ordinato attorno ai punti critici, un'idea di linee di flusso è presentata in figura 3.2 (b). Questo implica che vi sia una sola curva integrale di massimo incremento per ogni punto nello spazio delle caratteristiche, ovvero che da qualsiasi punto nello spazio delle caratteristiche, esiste un'unica direzione in cui la densità di probabilità aumenta più rapidamente. Seguendo questa direzione, si arriva infine a una delle mode della distribuzione, che rappresenta i punti in cui la densità di probabilità raggiunge valori relativamente alti rispetto ai punti circostanti. Quando si parla di curva integrale attraverso un punto si fa riferimento ad una traiettoria che inizia in quel punto e segue la direzione di massimo aumento della funzione di densità di probabilità (Genovese et al. 2016).
Le curve integrali, eccetto nei punti stazionari, non si intersecano mai, e questo significa che non si sovrappongono mai ad altre curve nel loro percorso. Questo assicura che ogni punto nello spazio sia assegnato a una sola curva integrale e che le curve dividano lo spazio in regioni distinte. Ogni curva integrale attraverso un punto non modale ha una destinazione finale. Questa destinazione è definita come il punto verso cui la curva si avvicina quando il tempo tende all'infinito. Nella maggior parte dei casi, questa destinazione è una delle mode (o antimode) della distribuzione, indicando che le curve integrali tendono a convergere verso le mode. Definendo la curva integrale attraverso un punto x:

$$\pi_x : \mathbb{R} \to \mathbb{R},\ \pi_x(t) = x,\ \pi_x^{'}(t) = \nabla p\{\pi_x(t)\}, \tag{3.8}$$

la destinazione della curva integrale attraverso un generico punto x può essere definita come:

$$dest(x) = lim_{t\to\infty}\pi_x(t), \tag{3.9}$$

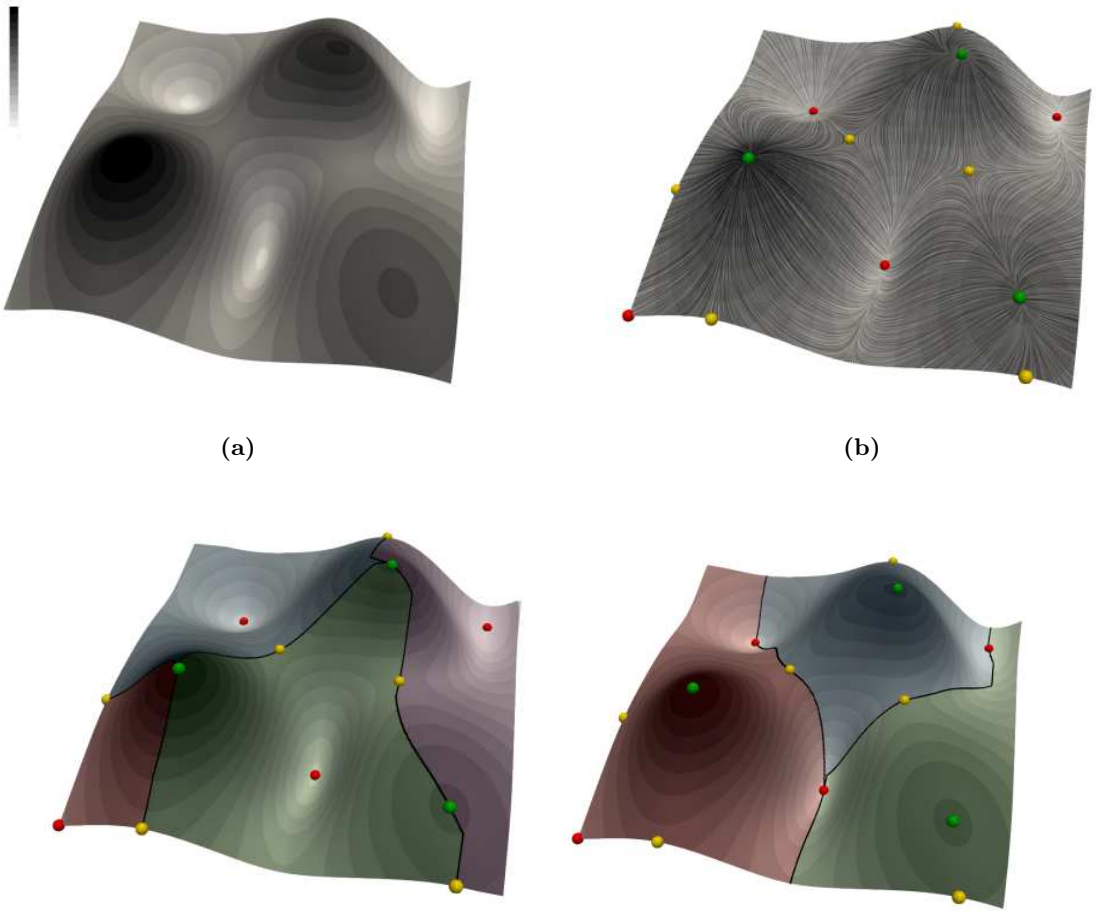
dunque la moda m è definita $dest(m) = m$. É possibile dimostrare che per tutti i punti x, ad eccezione di un insieme di misura zero che convergerà a punti di sella della distribuzione, la destinazione della curva integrale è una delle mode della distribuzione di probabilità m_j

con $dest(x) = m_j$. Si conclude affermando che quasi la totalità delle curve integrali portano a mode della distribuzione (Genovese et al. 2016).

Vengono inoltre definiti i cluster come i bacini di attrazione delle mode o *ascending manifolds*:

$$\mathscr{A}_j = \{x : dest(x) = m_j\}, \tag{3.10}$$

corrispondono agli insiemi di punti il cui percorso di ascensione del gradiente porta alla stessa moda. Queste regioni infatti, rappresentano le aree nello spazio il cui le curve integrali convergono verso una particolare moda, formando così i noti cluster associati a quella moda. In figura 3.2 è riportato un esempio di bacini di attrazione (Nilsson et al. 2023). I cluster campionari $C_1, C_2, ...C_k$ vengono definiti come insiemi di punti che appartengono ai rispettivi ascending manifolds $C_j = \{X_i : X_i \in \hat{\mathscr{A}_j}\}$ dove $\hat{\mathscr{A}_1}, ...\hat{\mathscr{A}_k}$ sono gli ascending manifolds ottenuti dalla stima della distribuzione di probabilità $\hat{p}$ come presentato in Genovese et al. (2016).



(a) **(b)**

(c) **(d)**

Figura 3.2: Funzione di densità (a), (b) mostra il campo vettoriale gradiente con in rosso i punti di minimo in verde i punti di massimo e in giallo i punti di sella, invece (c) mostra i bacini di attrazione dei minimi e (d) i bacini di attrazione dei massimi (Nilsson et al. 2023).

Capitolo 4

Metodi di inferenza di clustering

Nel presente capitolo verranno esplorati alcuni test che, combinati tra loro, consentono l'identificazione dei cluster nei dati. I test verranno prima presentati dal punto di vista teorico e successivamente applicati ai dati provenienti da pazienti ischemici.

Il primo test proposto è il *dip test*, finalizzato a determinare se i dati seguono una distribuzione unimodale o meno. Successivamente, ci si concentrerà sul test sulle *caratteristiche significative* della distribuzione dei dati, questo metodo fornisce una visualizzazione grafica delle regioni di interesse, come ad esempio le regioni del gradiente o della curvatura. Tali informazioni saranno fondamentali per l'identificazione dei cluster. Infine, verrà adottato un approccio non parametrico basato sulla densità per ottenere i valori delle mode, che costituiranno i cluster effettivi dei dati.

4.1 *Dip test*

Il test *"dip"* è un test non parametrico ed è generalmente utilizzato per valutare se un insieme di dati ha una distribuzione unimodale o multimodale. Il primo passo consiste nel calcolare la statistica *dip* ottenuta come la differenza tra il valore massimo di una funzione di distribuzione cumulativa unimodale e la funzione di distribuzione cumulativa empirica. Tuttavia questo test è applicabile solamente nel caso univariato, pertanto verrà esaminata l'estensione al multivariato fornita da Guo & Shah (2023).

L'approccio proposto riduce il problema multivariato a una serie di test univariati utilizzando una tecnica di proiezione univariata. In sostanza, verifica l'unimodalità multivariata attraverso la valutazione dell'unimodalità univariata per diverse direzioni nell'embedding dei dati.

Il test *dip* univariato viene reso applicabile al multivariato attraverso una procedura *hunt and test* per verificare l'unimodalità della distribuzione multivariata. Tale procedura prevede una suddivisione dei dati nella quale la prima parte dei dati viene utilizzata per selezionare il valore più opportuno dei parametri coinvolti nelle ipotesi mentre la seconda parte viene impiegata per testare l'ipotesi nulla. Diversamente dagli schemi già esistenti il

metodo proposto gode del controllo dell'errore di primo tipo che si avvicina asintoticamente al livello nominale (Guo & Shah 2023).

Nel contesto dell'identificazione dei cluster di una distribuzione è di interesse l'applicazione della procedura *hunt and test* all'interno di un contesto applicativo di inferenza di clustering.
Spesso utilizzare algoritmi come k-means e clustering gerarchico, può risultare problematico poichè indicano sempre la presenza di cluster anche quando i dati provengono da una popolazione omogenea. É possibile vedere questo problema come la verifica che il vero numero di cluster sia uno. Nella presentazione di questo metodo vengono considerati dati euclidei e si assume che esista un solo cluster se la distribuzione della popolazione è unimodale. É possibile formalizzare l'ipotesi nulla come un test per l'unimodalità.

Mentre in ambito univariato la definizione di unimodalità è unica, esistono molte nozioni di unimodalità multivariata. Nella formulazione di questo test verrà considerata l'unimodalità lineare: un vettore casuale $(X_1, ..., X_p)$ è unimodale se $\sum_i a_i X_i$ è unimodale per ogni coefficiente $a = (a_1, ..., a_p)$ diverso da 0. $X \in \mathbb{R}^p$ è unimodale se $a^T X$ è unimodale per ogni a diversa da 0, pertanto è possibile ottenere

$$\begin{cases} H_0 : \bigcap_{a \neq 0} \{\sum_{i=1}^p a_i X_i \text{ è unimodale}\} \\ H_1 : \exists a \neq 0 \text{ tale che } \sum_{i=1}^p a_i X_i \text{ non è unimodale} \end{cases} \tag{4.1}$$

Seguendo la procedura deve essere effettuata una suddivisione casuale dei dati in una parte A e in una parte B. Successivamente sarà individuata una direzione $\hat{a}$ nel primo set di dati e verrà effettuato il test con i dati rimanenti.

- Per identificare un'opportuna direzione $\hat{a}$, sulla parte A viene eseguito un algoritmo a 2 medie e viene scelto $\hat{a}$ come vettore normalizzato che collega i due centri dei cluster.

- Quindi, per verificare l'unimodalità, viene utilizzato un test basato sulla statistica *dip*. Si consideri F_n la distribuzione empirica della parte B dei dati e U l'insieme delle distribuzioni univariate unimodali. La statistica *dip* viene definita come:

$$\rho_n := \inf_{Q \in U} \|F_n - Q\|_\infty. \tag{4.2}$$

  Può essere calcolata mediante pacchetto `diptest` (Maechler 2013) in R.

É importante effettuare una precisazione: Hartigan (1985) suggerisce di confrontare la statistica *dip* ρ_n con la statistica dip di un campione estratto da un uniforme (0,1) che funge da distribuzione nulla meno favorevole. Uno dei limiti di tale metodologia è che produce risultati conservativi.

4.2 Caratteristiche significative per la stima della densità kernel

Nella presente sezione verrà esaminato l'approccio proposto da Duong et al. 2008. Tale articolo risulta di importante rilevanza all'interno dell'elaborato in questione. Infatti, la tematica centrale dell'articolo è il concetto di significatività delle caratteristiche della densità di dati multidimensionali. Il metodo rappresenta un'importante strategia per valutare se determinate caratteristiche dei dati, come i massimi e minimi locali, sono statisticamente significative in una struttura di dati in più dimensioni. Per ottenere una base solida per l'analisi della significatività delle caratteristiche nei dati multidimensionali vengono combinati l'utilizzo di stimatori della derivata della densità kernel con test di ipotesi per regioni modali. Questo approccio consente di discernere se le caratteristiche evidenziate nei dati sono il risultato di vere differenze o semplici variazioni casuali, contribuendo così alla robustezza e all'accuratezza dell'analisi statistica condotta.
É fondamentale chiarire il concetto di caratteristiche significative. Nell'analisi univariata o bivariata le caratteristiche che risultano importanti includono non solo i massimi locali ma anche minimi locali, valli, creste, punti di sella e gradienti ripidi; tuttavia nell'analisi di dati tridimensionali o di maggiore dimensione si fa riferimento esclusivamente ai massimi locali. Questo fenomeno si verifica poichè, con l'aumentare delle dimensioni, i massimi locali diventano le caratteristiche più rilevanti, perché sono più evidenti e influenzano in modo significativo la struttura complessiva dei dati, mentre altre caratteristiche come le valli diventano meno significative perché sono meno evidenti e possono essere sopraffatte dalla dispersione dei dati in grandi dimensioni. Infatti, se si desidera indagare la struttura di densità di dati multivariati, sarà di primo interesse identificare le caratteristiche significative piuttosto che la stima dell'intera densità.

All'aumentare della dimensione emergono alcune differenze che non riguardano solamente le caratteristiche significative, ma influenzano anche il contesto e i test d'ipotesi proposti. Ad esempio, per quanto riguarda la stima multivariata della densità del kernel, le matrici della larghezza di banda generale inducono una rotazione nei kernel che non è presente nel caso unidimensionale (Duong et al. 2008). Inoltre, l'analisi di dati multivariati complessi o ad alta dimensionalità richiede l'adozione di approcci diversificati per l'analisi, specialmente quando si affronta la sparsità dei dati in spazi ad alta dimensione. Le stime di densità sono fondamentali per comprendere le caratteristiche presenti nei dati. Poiché le informazioni quantitative sulle caratteristiche sono contenute nella prima e nella seconda derivata della vera densità p, gli stimatori naturali di quest'ultime saranno le derivate dello stimatore di densità kernel $\hat{p}$. Nel contesto della stima della densità kernel o della sua derivata, la scelta della larghezza di banda per le stime, incide notevolmente sulla loro prestazione. Per questo in seguito verrà considerato un range di lunghezza di banda piuttosto che il valore ottimale di quest'ultimo.

4.2.1 Test d'ipotesi per le stime delle derivate

Si presentata una metodologia per valutare la significatività delle caratteristiche attraverso il test delle derivate, mirato ad individuare regioni in cui le derivate assumono valori significativamente differenti da zero. In particolare, vengono proposti test distinti per le derivate del gradiente e della curvatura, differenziandosi da metodologie precedenti. Sebbene i risultati distribuzionali dei due tipi di test siano simili, le regioni in cui ci si aspetta di rigettare l'ipotesi nulla sono molto diverse o addirittura complementari.

Si considera per $x \in \mathbb{R}^d$ un'ipotesi nulla per il test del gradiente:

$$H_0 : ||\nabla p(x)|| = 0, \tag{4.3}$$

dove $||\cdot||$ è la norma euclidea. La distribuzione asintotica sotto l'ipotesi nulla dopo la normalizzazione è :

$$\left\{\Sigma_H^{(1)}(x)\right\}^{-\frac{1}{2}} g(x) \sim \mathcal{N}(0, I), \tag{4.4}$$

dove $\nabla p(x) = g(x)$, $\Sigma_H^{(1)}(x) = n^{-1}|H|^{-\frac{1}{2}}H^{-\frac{1}{2}}R(\nabla K)H^{-\frac{1}{2}}p(x)$, $R(\nabla K) < \infty$ considerando $R(t(x)) = \int_{\mathbb{R}^d} t(x)t(x)^T dx$ e K è la densità kernel stimata. Una statistica test appropriata per il gradiente è:

$$W^{(1)}(x) = \|\{\Sigma_H^{(1)}(x)\}^{-\frac{1}{2}} g(x)\|^2, \tag{4.5}$$

che richiede una stima di $\Sigma_H^{(1)}(x)$ che si può ottenere sostituendo la stima densità kernel $\hat{p}(x)$ alla verà denistà $p(x)$:

$$\hat{\Sigma}_H^{(1)}(x) = n^{-1}|H|^{-\frac{1}{2}}H^{-\frac{1}{2}}R(\nabla K)H^{-\frac{1}{2}}\hat{p}(x). \tag{4.6}$$

Infine, poichè il termine di $\hat{\Sigma}_H^{(1)}(x)$, in particolare $n^{-1}|H|^{-1/2}\hat{p}(x)$, è positivo la sua componente matriciale $H^{-1/2}R(\nabla K)H^{-1/2}$ è definita positiva e invertibile, si pone

$$\hat{W}^{(1)}(x) = ||\{\Sigma_H^{(1)}(x)\}^{-\frac{1}{2}} \hat{g}(x)||^2. \tag{4.7}$$

Seguirà che $\hat{W}^{(1)}(x)$ asintoticamente approssimabile ad un Chi-quadro con d gradi di libertà (Duong et al. 2008):

$$\hat{W}^{(1)}(x) \overset{\text{approx.}}{\sim} \chi_d^2. \tag{4.8}$$

Il prossimo passaggio consiste nel valutare la significatività dei valori $\hat{W}^{(1)}(x)$ in diversi punti x. Tuttavia, data la correlazione tra i dati , è necessario affrontare questa correlazione in modo adeguato durante l'analisi dei test. Per risolvere il problema , è stata selezionata una procedura di test multipli, piuttosto che l'approccio classico di Bonferroni, che potrebbe risultare eccessivamente conservativo in presenza di correlazione tra i test. La procedura scelta offre una soluzione che riesce a equilibrare il controllo del tasso complessivo di errore di primo tipo con la gestione della correlazione tra i test. La procedura è la seguente. Si consideri α pari al livello di significatività del test. I p-value degli m singoli test verranno disposti in ordine crescente $P_{(1)}, ..., P_{(m)}$ con le rispettive ipotesi nulle $H_{0,(1)}, ..., H_{0,(m)}$ e

confrontando ciascun p-value con una soglia corretta per il numero di test effettuati. Se un p-value j con $1 \leq j \leq m$ risulta inferiore alla soglia corretta $P_{(j)} \leq \alpha/(m-j+1)$, si rigetta l'ipotesi nulla corrispondente e tutte le precedenti $H_{0,(1)}, ..., H_{0,(j)}$. Tuttavia, la soglia viene adattata in base al numero di test eseguiti e ai p-value osservati. Questo approccio consente di valutare in modo accurato la significatività dei dati senza cadere in conclusioni erronee dovute alla correlazione tra le osservazioni. Il p-value associato a ciascun punto x è stato calcolato come $P(U > \hat{W}^{(1)}(x))$, dove U segue una distribuzione chi-quadro con d gradi di libertà. Il p-value indica la probabilità di osservare un valore $\hat{W}^{(1)}(x)$ uguale o superiore sotto l'ipotesi nulla che U segua la distribuzione chi-quadro.

Tuttavia, al fine di ridurre il rischio di effetti spuri dovuti a una dimensione del campione insufficiente, il test di significatività è stato limitato alle regioni in cui il numero di punti dati è sufficientemente grande. Per valutare la dimensione del campione efficace, è stato calcolato un indice basato sulla somma ponderata dei kernel gaussiani centrati sui punti dati, diviso per il valore del kernel gaussiano nello stesso punto. Questa metrica fornisce una stima della densità dei dati in ciascuna regione, consentendo di identificare le aree con un numero adeguato di osservazioni per supportare il test di significatività. Di conseguenza, il test è stato condotto solo nelle regioni in cui si è ritenuto di avere abbastanza dati disponibili per ottenere risultati affidabili.

Viene considerata l'ipotesi nulla che la norma euclidea della derivata seconda della distribuzione dei dati, nonchè l'Hessiana, sia uguale a 0:

$$H_0 : ||vech\nabla^{(2)}p(x)|| = 0. \tag{4.9}$$

In altri termini testiamo l'ipotesi nulla che la curvatura della funzione intorno al punto x sia nulla. Per condurre questo test, viene utilizzata una matrice di banda H adatta per gli stimatori di curvatura della densità del kernel, insieme alla funzione

$$\Sigma_H^{(2)}(x) = n^{-1}|H|^{-1/2}R\left(\text{vech}\left(H^{-1/2}\nabla^{(2)}KH^{-1/2}\right)\right)p(x), \tag{4.10}$$

dove n rappresenta il numero di punti dati, R è una funzione di ridimensionamento, *vech* è l'operatore che converte una matrice in un vettore. La distribuzione nulla asintotica di $vech\hat{\mathscr{H}}$ dopo la normalizzazione è:

$$\{\Sigma_H(x)\}^{-1/2}\text{vech}\mathscr{H} \sim \mathcal{N}(0, Id^*), \tag{4.11}$$

dove I_d* è la matrice identità con $d^* = (d+1)d/2$. Analogamente a quanto visto per la statistica test del gradiente, la statistica test della curvatura sarà:

$$W^{(2)}(x) = ||\Sigma_H^{(2)}(x)\}^{-1/2}\text{vech}\hat{\mathscr{H}}||^2. \tag{4.12}$$

Come già affermato in precedenza, per dimensioni maggiori di due le caratteristiche significative si riferiscono a mode significative. L'obiettivo è stabilire se ci sia una curvatura significativa, senza dover necessariamente caratterizzare questa curvatura attraverso la sua

struttura degli autovettori. In sostanza, ci si concentra sulla presenza di curvature rilevanti piuttosto che su come queste curvature siano orientate nello spazio multidimensionale. Un vantaggio di questo approccio è che si evita la necessità di simulare punti critici della distribuzione nulla, che è richiesta in altre procedure statistiche. Invece, la statistica test $W^{(2)}(x)$ ha una distribuzione nulla approssimativa di forma chiusa, il che semplifica il processo di valutazione della significatività della curvatura senza dover ricorrere a complessi calcoli o simulazioni (Duong et al. 2008).

Una stima di $W^{(2)}(x)$ è:

$$\hat{W}^{(2)}(x) = \|\{\hat{\Sigma}_H^{(2)}(x)\}^{-1/2}\text{vech}\mathscr{H}\|^2 \overset{\text{approx.}}{\sim} \chi^2_d*, \tag{4.13}$$

dove

$$\hat{\Sigma}_H^{(2)}(x) = n^{-1}|H|^{-1/2}R(\text{vech}(H^{-1/2}\nabla^{(2)}KH^{-1/2}))\hat{p}(x). \tag{4.14}$$

Diversamente da altri metodi proposti in letteratura, l'approccio in questione stima $\hat{\Sigma}_H^{(2)}(x)$ in una procedura a livello di matrice anzichè stimare individualmente ogni elemento della matrice, garantendo la positività definita della matrice stimata. Inoltre, l'approccio proposto è computazionalmente più efficiente poiché non richiede il calcolo delle varianze e covarianze campionarie di ogni singolo elemento di vech $\hat{\mathscr{H}}$. Un altro vantaggio del nostro approccio è che consente l'utilizzo di matrici di banda generiche anzichè l'utilizzo di una banda più restrittiva.

Per quanto riguarda la significatività della curvatura si procede in maniera analoga a quanto mostrato per il gradiente. Nel caso dei test di curvatura, il p-value associato ad un determinato punto x è calcolato come $P(U > \hat{W}^{(2)}(x))$ dove $U \sim \chi^2_{d^*}$.

La principale differenza rispetto al test del gradiente è l'aumento dei gradi di libertà della distribuzione chi-quadro da d a $d^* = (d+1)d/2$. Questo aumento dei gradi di libertà è dovuto al fatto che la valutazione della curvatura coinvolge più direzioni nello spazio rispetto al test del gradiente che si concentra solamente nella direzione del gradiente in uno spazio d-dimensionale.

4.2.2 Le regioni di rifiuto del gradiente e della curvatura

Attraverso il confronto tra i test basati sul gradiente e quelli basati sulla curvatura per identificare le regioni modali emerge la complementarità delle regioni di rifiuto. Nei test basati sulla curvatura, le regioni di rifiuto sono progettate per includere le regioni modali, poichè si presume che deviazioni significative dall'ipotesi nulla si verifichino nelle vicinanze di modalità effettive. Di conseguenza le regioni di rifiuto dei test basati sulla curvatura si sovrappongono alle regioni modali, sottolineando l'importanza dell'individuare le deviazioni significative nei dati. Al contrario, nei test basati sul gradiente, le regioni di rifiuto escludono molto probabilmente le regioni modali e antimodali, ma possono includere tutte le altre regioni, a condizione che siano disponibili dati sufficienti. Queste considerazioni indicano che i test basati sulla curvatura rappresentano l'aspetto di maggior interesse, in

quanto consentono di individuare le variazioni significiative nella curvatura associate alle mode presenti nei dati.

Tuttavia, i test basati sul gradiente possono ulteriormente arricchire i risultati dei test basati sulla curvatura, fornendo dettagli o conferme aggiuntive sui pattern di variazione nei dati. Inoltre, è importante notare che, sebbene i test basati sulla curvatura siano direttamente correlati all'identificazione delle modalità, i test basati sul gradiente possono offrire un contributo complementare.

In seguito, viene considerato l'effetto della dimensione del campione sui risultati dei test statistici. É noto che all'aumentare della dimensione del campione, la frequenza con cui si rigetta l'ipotesi nulla aumenta. Questo fenomeno è dovuto alla struttura assimmetrica dei test statistici tra la regione di accettazione e la regione di rifiuto. Nei test basati sulla curvatura, la regione di rifiuto aumenta proporzionalmente alla dimensione del campione. Questo implica che all'aumentare del numero di dati raccolti, la regione di rifiuto si estende e include non solo le regioni modali, ma anche le regioni adiacenti a tali modalità. Questo fenomeno è considerato positivo perché l'ampliamento della regione di rifiuto nei test basati sulla curvatura consente di catturare in modo più efficace le deviazioni e di identificare le modalità o le loro vicinanze come aree di interesse durante l'analisi statistica dei dati. Analogamente alle regioni di rifiuto dei test per la curvatura, nei test basati sul gradiente le regioni di rifiuto tendono ad aumentare con l'aumentare della dimensione del campione, tuttavia il problema è che le regioni di rifiuto potrebbero espandersi fino a includere completamente le regioni modali. Questo fenomeno potrebbe comportare una sovrastima delle deviazioni significative dalla distribuzione attesa, il che potrebbe non essere desiderabile in quanto potrebbe portare a una interpretazione erroneamente pessimistica dei dati. Una strategia alternativa per mitigare questo effetto potrebbe essere quella di ridurre il livello di significatività del test basato sul gradiente. In questo modo si potrebbe trovare un equilibrio tra la dimensione del campione e la dimensione della regione di rifiuto, consentendo una maggiore sensibilità nel rilevare deviazioni significative senza estendere eccessivamente la regione di rifiuto e includere erroneamente le regioni modali.

Scelta della larghezza di banda

Un altro aspetto da non sottovalutare nell'implementazione del metodo è la scelta della larghezza di banda. Infatti tale parametro è un aspetto cruciale per il calcolo delle statistiche test. Questo concetto è particolarmente rilevante nell'ambito della stima delle derivate della densità del kernel, dove la corretta specifica della larghezza di banda può influenzare significativamente i risultati dell'analisi. Per decidere l'opzione migliore, è necessario considerare una serie di metodi di *smoothing*.

Una singola larghezza di banda, applicata uniformemente a tutte le dimensioni, può offrire una visione generale dei dati, ma potrebbe non essere ottimale per analizzare dati correlati o con variazioni di scala diverse. Al contrario, l'uso di una matrice generale di

larghezza di banda consente una maggiore flessibilità e può essere più adatto per un'analisi più approfondita.

Inoltre, è importante considerare che la scelta della larghezza di banda può influenzare i risultati del test statistico. Ad esempio, nei test basati sul gradiente, una larghezza di banda eccessivamente ampia potrebbe portare a una sovrastima delle deviazioni significative dalla distribuzione attesa, con conseguenze negative sull'interpretazione dei risultati.

Per adattarsi a diverse applicazioni, nel metodo proposto è possibile utilizzare approcci alternativi per la selezione della larghezza di banda come valori forniti dall'utente o una matrice di larghezza di banda oggettivamente scelta e guidata dai dati.

4.3 Inferenza non parametrica basata sulla densità per le mode

All'interno dell'articolo di Genovese et al. 2016 viene presentato l'approccio oggetto di analisi in questa sezione. Questo metodo costituisce un'importante base teorica e metodologica per l'approfondimento in questione, focalizzandosi su un'analisi approfondita circa la natura e l'importanza dei cluster nei dati esaminati.

Il metodo si occupa di identificare le mode di una distribuzione multivariata e di fornire intervalli di confidenza che presentino informazioni sulle forme delle mode stimate. In particolar modo indaga la significatività delle mode affinchè si possa distinguere una moda dovuta a fluttuazioni casuali da una che è realmente tale.

Questo approccio di tipo non parametrico si basa sulla suddivisione dei dati identificando le potenziali mode nella prima metà dei dati e conducendo inferenze nella seconda metà. Per ottenere intervalli di confidenza validi per gli autovalori, si utilizza un bootstrap basato sulla trasformazione polinomiale simmetrica elementare, argomento che sarà trattato nella sezione 4.3.1. Questo assicura intervalli di confidenza validi indipendentemente dalle molteplicità degli autovalori.

Questo metodo affronta le difficoltà nel definire test per le mode, in particolare il problema di testare l'ipotesi nulla di "nessuna moda" e la presenza di un numero potenzialmente infinito di posizioni delle mode. Inoltre, si affronta il problema di fare inferenza sugli autovalori dell'Hessiana $\hat{\mathscr{H}}$, indicati da $\lambda(x) = (\lambda_1(x), ..., \lambda_d(x))$

Nel contesto dell'analisi statistica della densità dei dati, è fondamentale stabilire un insieme di assunzioni che garantisca la validità delle procedure utilizzate. Di seguito i presupposti del metodo:

1. Si suppone che la densità p dei dati sia limitata e continua, definita su un insieme compatto $X \in \mathbb{R}^d$.

2. È richiesto inoltre che il gradiente e l'Hessiana della densità siano limitati e continui.

3. Si assume che l'Hessiana sia non degenere in tutti i punti stazionari, garantendo così la stabilità delle derivate della densità.

4. Si presume che la densità abbia un numero finito di mode, o massimi locali, all'interno dell'insieme X.

5. Si richiede che esista una distanza positiva minima tra le mode e che il massimo degli autovalori dell'Hessiana in ciascuna moda sia negativo.

6. Infine si suppone che il kernel utilizzato nella stima di densità sia una densità di probabilità simmetrica con derivate continue del primo e del secondo ordine, e con momento del secondo ordine limitato.

Queste condizioni assicurano una separazione adeguata tra le mode, una curvatura concava della densità intorno a ciascuna moda e che il kernel sia adatto per la stima della densità.

4.3.1 Il metodo

La media dello stimatore di densità è data da:

$$\hat{p}_h(x) = E[\hat{p}_h(x)] = \int K(t)p(x+th)\,dt, \tag{4.15}$$

che rappresenta la stima della densità media attorno al punto x, considerando la distribuzione dei dati nella vicinanza di x.

Come spiegato già in precedenza, si può utilizzare una matrice di larghezza di banda H nello stimatore di densità:

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i), \tag{4.16}$$

definita come $K_H(x) = |H|^{-\frac{1}{2}}K(H^{-\frac{1}{2}}x)$. Per semplicità verrà utilizzata una larghezza di banda scalare h, corrispondente a $H = h^2 I$.

Per individuare nello spazio le mode dello stimatore della densità $\hat{p}_h$ viene utilizzato l'algoritmo *Mean shift* che individua le mode approssimando i percorsi di massima pendenza nei dati. In questo modo è possibile ottenere un insieme di mode candidate $\hat{\mathscr{M}} = \{\hat{m}_1, ..., \hat{m}_k\}$. Il metodo *Mean Shift* è un algoritmo di clustering non parametrico ed è utilizzato per identificare i cluster nei dati senza la necessità di specificare a priori il numero di cluster. L'approccio principale che utilizza è quello di spostare iterativamente ciascun punto dei dati in direzione della sua moda locale più vicina, cioè verso la regione di maggiore densità dei punti, mostrando verso quale moda è attratto ogni punto. Tale algoritmo convergerà verso i cluster, ne è riportato un esempio in figura 4.1

Algoritmo Mean Shift

Il processo inizia dando in input la funzione di densità di probabilità stimata e una griglia di punti $A = \{a_1, ..., a_N\}$ che generalmente sono i punti del dataset.
Per ogni punto a_j, ponendo $a_j^0 = a_j$ si itera la seguente equazione fino a convergenza,

$$a_j^{(s+1)} \leftarrow \frac{\sum_{i=1}^{n} X_i K\left(\frac{||a_j^{(s)} - X_i||}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{||a_j^{(s)} - X_i||}{h}\right)}, \tag{4.17}$$

dove X_i sono i punti dati, K è la funzione kernel, h è la larghezza di banda e infine s è l'iterazione corrente dell'algoritmo.

In questo passaggio per ogni punto, si calcola il *mean shift* o spostamento medio, che indica la direzione verso la quale il punto deve essere spostato per massimizzare la densità

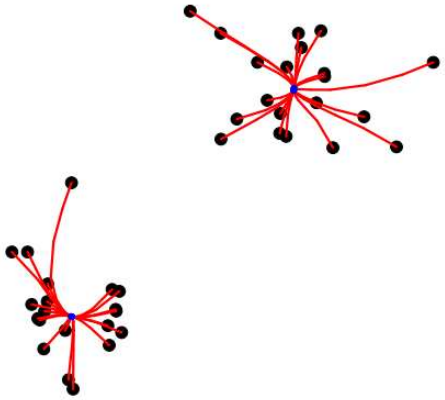


Figura 4.1: Un esempio dell'algoritmo *Mean Shift*: i punti in blu rappresentano le mode mentre i punti che convergono verso di esse sono assegnati al cluster corrispondente (Wasserman 2017).

locale. Questo spostamento viene calcolato utilizzando la funzione Kernel, che assegna un peso ai punti circostanti in base alla loro distanza. A questo punto, ogni punto della griglia viene spostato verso la direzione del gradiente della funzione di densità di probabilità stimata. L'algoritmo restituirà come output $\hat{\mathscr{M}}$, definito come un insieme di valori unici dell'insieme $\{a_1^{(\infty)}, ..., a_N^{(\infty)}\}$. Questo processo viene ripetuto iterativamente fino a quando non si raggiunge la convergenza, ossia quando i punti non si spostano significativamente tra un'iterazione e l'altra. Infatti il simbolo ∞ indica che si considerano gli elementi a_j dopo un numero non definito di iterazioni o fino alla convergenza.

Alla fine del processo, i punti tenderanno a raggrupparsi attorno ai massimi locali della densità dei dati, che rappresentano le mode nei dati.

Algoritmo per il test delle mode locali

La procedura proposta mira a risolvere i problemi elencati nel paragrafo 3.1. Il metodo consiste nelle seguenti fasi:

1. Viene effettuato un data splitting sui dati ossia l'insieme dei dati viene suddiviso in due sottoinsiemi distinti $X = (X_1, ..., X_n)$ e $Y = (Y_1, ..., Y_n)$ dove il primo è utilizzato per individuare le mode candidate e l'altro per condurre test d'ipotesi. Questo approccio consente di concentrarsi esclusivamente sui test d'ipotesi, riducendo il numero complessivo di test effettuati durante la fase di individuazione delle mode. Tale semplificazione limita il rischio di falsi positivi e assicura risultati più robusti e affidabili durante l'analisi dei dati.

   Inoltre, lo splitting dei dati è importante per garantire la validità degli intervalli di confidenza. Senza questa divisione, con un approccio alternativo si potrebbe ottenere un test valido trattando l'Hessiana stimata come un processo casuale sull'intero spazio

dei dati e quindi stimando le massime fluttuazioni di questo processo. Tuttavia, lo splitting dei dati e il focus su un numero finito di punti è molto più semplice e pratico.

2. Viene utilizzato il primo insieme X per costruire la densità stimata $\hat{p}_{X,h}$ e usando l'algoritmo *mean shift* trovare un insieme finito di mode candidate $\hat{\mathscr{M}}$ della densità stimata.

3. Utilizzando il secondo insieme Y, si costruisce $\hat{p}_{Y,h}$ e si calcola il suo gradiente $\hat{g}_{Y,h}$ e l'Hessiana $\hat{\mathscr{H}}_{Y,h}$.

4. Per ogni moda candidata $m = m_j \in \hat{\mathscr{M}}$:

   (a) viene stimata l'Hessiana nel punto m e vengono calcolati gli autovalori $\hat{\lambda}_i(m)$ della matrice Hessiana per $i = 1, ..., d$ dove d è la dimensione dello spazio

   (b) Successivamente, utilizzando gli autovalori ottenuti si calcolano i polinomi simmetrici elementari $s_i(m)$ per $i = 1, ..., d$. Questi polinomi forniscono informazioni importanti sulla forma delle mode candidate.

   (c) Per valutare la variabilità delle statistiche calcolate e ottenere delle stime affidabili dei parametri viene eseguita una procedura bootstrap. Vengono generati B campioni di Hessiane nel punto m tramite il campionamento bootstrap consentendo così di ottenere intervalli di confidenza e condurre test statistici sulla base della variabilità osservata. Infine si determinano anche i B corrispondenti polinomi simmetrici elementari.

   (d) Si calcola il rettangolo di confidenza G_j e l'intervallo di confidenza $\hat{\mathscr{C}}_j$ associato alla moda candidata. Vengono creati in maniera tale da coprire un'area specifica della distribuzione di probabilità intorno alla moda candidata, tenendo conto della variabilità stimata. Il livello di confidenza è $(1 - \alpha)/k$ dove k è il numero di mode candidate.

   (e) A questo punto se l'intervallo di confidenza per la moda candidata giace interamente sopra lo zero allora la moda candidata m verrà dichiarata come moda effettiva. Questo criterio viene utilizzato per determinare in modo significativo la presenza di una moda, considerado la variabilità stimata e garantendo un certo livello di confidenza.

Per ogni moda candidata verrà ottenuto un rettangolo di confidenza, l'insieme di questi rettangoli ci porterà informazioni sulle mode ed è noto come *eigenportrait*.

In precedenza si è discusso di alcuni problemi legati all'identificazione delle mode tra cui quello di testare che una moda m_j non sia tale, la soluzione proposta da questo metodo è l'idea di considerare un ipotesi alternativa più flessibile ovvero che m_j sia un approssimazione di una moda piuttosto che una moda esatta. Grazie allo splitting dei dati testare queste ipotesi modificate è asintoticamente equivalente a testare quelle originali.

Polinomi simmetrici elementari

Per ottenere una regione di confidenza per gli autovalori, si devono considerare i valori stimati degli autovalori $\hat{\lambda}_1(x) \geq \hat{\lambda}_2(x) \geq ... \geq \hat{\lambda}_d(x)$ della matrice $\hat{\mathscr{H}}_{Y,h}(x)$. Questi valori saranno ottenuti attraverso il bootstrap. Tuttavia, effettuare quest'operazione presenta alcuni problemi. In generale, $\lambda(x) = (\lambda_1(x), ..., \lambda_d(x))$ non è una funzione continuamente differenziabile di $\mathscr{H}_h(x)$ pertanto, l'applicazione del bootstrap non produrrà insiemi di confidenza validi per gli autovalori. Tuttavia è stato osservato che, se gli autovalori vengono trasformati utilizzando polinomi simmetrici elementari, l'insieme di confidenza ottenuto è valido. Vengono definiti i polinomi elementari simmetrici $s_1(x), ..., , s_d(x)$:

$$
\begin{aligned}
s_1(x) &= \sum_{i=1}^{d} \lambda_i(x), \\
s_2(x) &= \sum_{i_1=1}^{d} \sum_{i_2=i_1+1}^{d} \lambda_{i_1}(x)\lambda_{i_2}(x), \\
&\vdots \\
s_k(x) &= \sum_{i_1=1}^{d} \sum_{i_2=i_1+1}^{d} \cdots \sum_{i_k=i_{k-1}+1}^{d} \lambda_{i_1}(x)\lambda_{i_2}(x)\cdots\lambda_{i_k}(x), \\
&\vdots \\
s_d(x) &= \lambda_1(x)\lambda_2(x)\cdots\lambda_d(x).
\end{aligned}
\tag{4.18}
$$

I polinomi simmetrici sono utilizzati in questo caso per esprimere i coefficienti del polinomio caratteristico di una matrice in termini dei suoi autovalori. É importante specificare che $s_1(x), ..., , s_d(x)$ sono polinomi di $\lambda(x) = (\lambda_1(x), ..., \lambda_d(x))$ e non di x. Sarebbe corretto usare la notazione $s_j(\lambda_1(x), ..., \lambda_d(x))$ tuttavia per semplicità verrà utilizzato $s_j(x)$. Si può osservare che tutti gli autovalori sono negativi se e solo se $(-1)^k s_k > 0$ per tutti i k. Questa condizione riveste particolare importanza nelle analisi di stabilità e comportamento del sistema. Grazie a questo risultato infatti il polinomio $s(x)$ è una funzione continuamente differenziabile dell'Hessiana e la mappatura da $\lambda(x)$ a $s(x)$ è biunivoca. Ciò implica che può essere espressa come $s(x) = w\{\lambda(x)\}$ e $\lambda(x) = w^{-1}\{s(x)\}$.

4.3.2 Scelta della larghezza di banda

La scelta della larghezza di banda per l'identificazione delle mode attraverso un kernel è un problema complesso affrontato da numerosi approcci in letteratura. É stato osservato che con una ridotta larghezza di banda si tende ad individuare numerose mode, molte delle quali però vengono classificate come fluttuazioni casuali della densità stimata (Genovese et al. 2016). D'altra parte, quando la larghezza di banda è estremamente ridotta, emergono mode spurie che vengono eliminate dal test d'ipotesi. Questo comportamento è dovuto al

fatto che le dimensioni del rettangolo di confidenza variano al variare di h, infatti quando h è eccessivamente ridotto il numero di mode significative diventa 0.

Diversamente con una larghezza di banda ampia, il numero di mode identificate diminuisce fino a raggiungere il caso limite di una moda. É opportuno quindi trovare un equilibrio tra l'individuazione di vere mode di una distribuzione e il rischio di rilevare mode spurie dovute a rumore o fluttuazione nei dati.

L'obiettivo è scegliere la larghezza di banda h in modo tale da massimizzare il numero di mode significative. Considerando $N(h)$ il numero delle mode significanti trovate dal test come funzione di h viene posto $m = max\{N(h) = m\}$ e si definisce $\hat{h} = inf\{h : N(h) = m\}$. Una possibile scelta dell'intervallo della larghezza di banda è $\sigma\{log(n)/n\} \leq h^d \leq \sigma$ dove σ è la standard deviation (si assume che tutte le variabili abbiano standard deviation comune). É stato dimostrato che scegliere la larghezza di banda massimizzando il numero di mode significative porta ad individuare il numero corretto di mode (Genovese et al. 2016). La scelta della larghezza di banda nel caso in cui la densità sia singolare ossia nel caso ci siano picchi stretti o punti di forte accumulazione di probabilità diventa un problema complesso.

Infatti, le tecniche usuali per la scelta della larghezza di banda sono progettate per distribuzioni con andamenti regolari o forme simmetriche e quindi nel caso di densità singolare queste tecniche potrebbero non essere in grado di adattarsi adeguatamente alla struttura della distribuzione, inoltre potrebbero risultare computazionalmente complesse. Di conseguenza le tecniche convenzionali si rivelano inefficaci o addirittura inappropriate, quindi si propone un approccio che mira a massimizzare il numero di mode significative identificate, anziché ottimizzare la precisione della stima nella norma L_2. I risultati delle simulazioni supportano l'efficacia e la validità dell'approccio. Mentre i risultati empirici possono essere promettenti e suggerire l'efficacia del metodo proposto, la sua validità teorica non è stata completamente stabilita, in particolare quando si considerano larghezze di banda molto piccole.

4.3.3 Bias nella stima della densità

Come già affermato in precedenza, per stimare la densità di probabilità di una distribuzione di dati verrà utilizzato uno stimatore kernel della densità $\hat{p}_h$ che dipende dalla larghezza di banda $h > 0$. In particolare saranno le mode di $\hat{p}_h$ ad essere considerate stime delle mode di p_h. Chiaramente è importante sottolineare che, nonostante l'efficacia di $\hat{p}_h$ nell'approssimare p_h, è presente un bias, genericamente di ordine $O(h^2)$, che separa p_h da p. In particolare per quanto riguarda le mode

$$\max_j ||m_j - m_{hj}|| = \mathcal{O}(h^2), \tag{4.19}$$

dove m_j sono le mode di p e m_{hj} le mode di p_h. Con probabilità che tende a 1, $\hat{p}_h$ ha lo stesso numero di mode di p_h definite come $\hat{m}_{h1}, ..., \hat{m}_{hk}$:

$$\max_j ||\hat{m}_{hj} - m_{hj}|| = \mathcal{O}_P\left(\sqrt{\frac{1}{nh^{d+2}}}\right), \tag{4.20}$$

e

$$\max_j \|\hat{m}_{jh} - m_j\| = \mathcal{O}(h^2) + \mathcal{O}_P\left(\sqrt{\frac{1}{nh^{d+2}}}\right). \tag{4.21}$$

La formula 4.20 indica che l'errore nella stima del numero di mode diminuisce all'aumentare delle osservazioni n, della larghezza di banda h e dal numero di dimensioni che riflette la complessità del problema. La formula 4.19 afferma che il bias introdotto dall'utilizzo del KDE non è di fondamentale importanza quando si studiano le mode di una distribuzione, in particolar modo quando l'obiettivo finale è il clustering basato sulle mode. Infatti, nonostante l'influenza del bias è importante precisare che l'appartenenza dei punti ai cluster rimane piuttosto robusta. Anche se le posizioni esatte dei cluster saranno spostate a causa del bias i punti che condividono caratteristiche simili o che sono vicini nella distribuzione originale e che quindi inizialmente appartengono allo stesso cluster tenderanno comunque ad essere assegnati allo stesso cluster. Questo risultato evidenzia la stabilità e la coerenza dei cluster ottenuti dal metodo.

Il metodo proposto non fornisce una potenza del test , tuttavia se vengono fatte assunzioni sulla dimensione e sulla separazione delle mode potrebbe essere possibile calcolare la potenza del test. I risultati proposti da questo metodo sono promettenti in quanto è possibile dimostrare che i cluster basati sulle mode significative sono una buona approssimazione dei cluster della popolazione $\mathscr{A}_1, ..., \mathscr{A}_{k_0}$ definiti nell'equazione 3.10 .

4.4 Confronto tra metodi

Mettendo a confronto i due metodi descritti in Duong et al. (2008) e Genovese et al. (2016) si possono notare differenze e soprattutto complementarietà. Innanzitutto è importante precisare che i due metodi si distinguono per i diversi obiettivi. Genovese et al. (2016) si concentrano sull'identificazione di un insieme definito e finito di potenziali mode, utilizzando un test di significatività per ciascuna di esse, mentre Duong et al. (2008) adottano un approccio più visivo e descrittivo, che fornisce una visualizzazione grafica. Inoltre il metodo di Genovese et al. (2016) mira a fornire un insieme di intervalli di confidenza per gli autovalori dell'Hessiana stimati presso le modalità identificate, il che risulta importante per comprendere meglio la struttura della densità nei punti di interesse.

Un'altra differenza tra i due approcci consiste nella procedura applicata. Mentre Duong et al. (2008) utilizzanno due test statistici, uno relativo al gradiente pari a 0 e uno alla norma dell'Hessiana pari a 0, diversamente Genovese et al. (2016) si occupano di identificare mode tramite test basati sugli autovalori dell'Hessiana.

Inoltre Duong et al. (2008) effettua i test per numerosi punti ed utilizza una correzione per i test multipli. Le regioni in cui l'ipotesi nulla riguardante il gradiente non viene rigettata e, contemporaneamente, l'ipotesi nulla riguardante l'Hessiana viene rigettata sono considerate interessanti. Queste aree della densità rappresentano casi in cui non ci sono evidenze statistiche di un gradiente significativamente diverso da zero, ma vi sono indicazioni che l'Hessiana è significativamente diversa da zero. Queste regioni possono essere particolarmente interessanti poiché suggeriscono la presenza di mode o strutture complesse nella densità dei dati. Al contrario, l'approccio di Genovese et al. individua le mode utilizzando l'algoritmo *mean shift* e fornisce valori numerici specifici per rappresentarle. In conclusione si può affermare che il metodo di Genovese et al. (2016) fornisce informazioni aggiuntive rispetto all'approccio di Duong et al. (2008) riguardo alle mode da stimare.

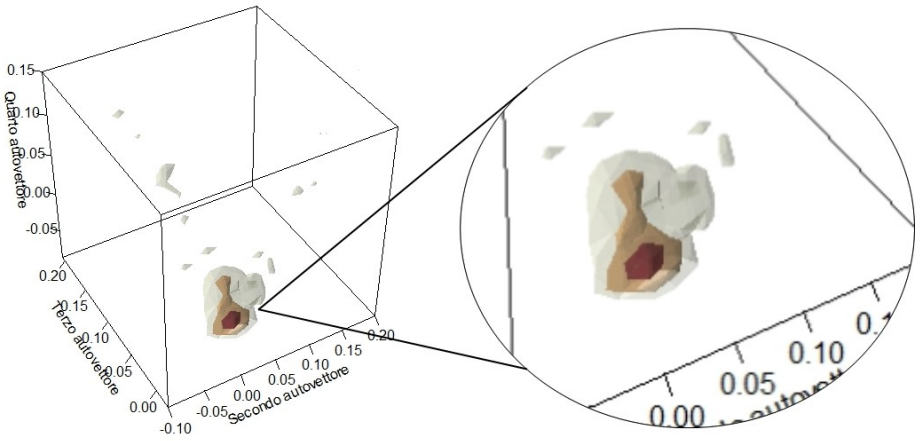
Capitolo 5

Risultati

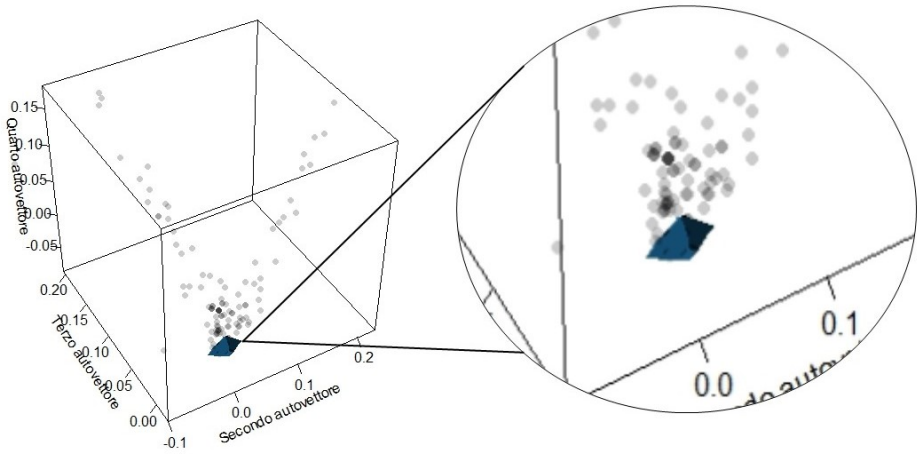
Nel presente capitolo verranno illustrati i risultati delle analisi condotte sul dataset dei pazienti ischemici. Il test *dip* rifiuta l'ipotesi nulla di unimodalità e considera la presenza di più di una moda nella distribuzione.

5.1 Caratteristiche significative

Per quanto invece riguarda il metodo per identificare le caratteristiche significative dell'embedding dei dati, esso fornisce la visualizzazione dei due grafici in figura 5.1. Il primo (a) illustra la stima della densità kernel e le regioni più scure corrispondono a zone in cui i dati sono più concentrati, mentre le regioni più chiare indicano una minore densità di dati, il secondo (b) evidenzia le regioni della curvatura significative con i punti dell'embedding sottostanti. Si può notare che la regione di curvatura e la regione con altà concentrazione della stima kernel coincidono. La presenza di una regione di curvatura significativa suggerisce la presenza di mode o antimode o variazioni rilevanti nella direzione e nell'intensità dei cambiamenti nei dati. Può anche essere il caso di fenomeni non lineari o comportamenti complessi della stima di densità kernel. Va però notato che questo metodo non fornisce direttamente un numero di cluster; piuttosto, indica la possibile presenza di gruppi in base a tali variazioni di direzione nei dati. Non sono presenti regioni del gradiente significative, tuttavia questo risultato può essere dovutoal fatto che è possibile che in alcune regioni dello spazio dell'embedding non siano stati effettuati i test in quanto la procedurra necessita di un certo numero di punti per essere applicata.



(a) Stima della densità Kernel multivariata dell'embedding dove l'intensificazione del colore corrisponde ad una maggior concentrazione dei punti.



(b) Regioni di curvatura significative in blu e i punti dell'embedding in grigio.

Figura 5.1: Rappresentazione delle regioni significative dell'embedding.

5.2 Identificazione dei cluster

Il numero di mode identificate dal metodo, considerate significative, varia in base a diversi parametri come la soglia ε e il *bandwidth* e alla specifica suddivisione del dataset associata al seme scelto per la generazione casuale.

ε	0.18	0.20	0.25	0.30	0.35	0.40	0.45	0.5
Numero di mode	2	2	2	2	1	2	1	2

Tabella 5.1: Numero di mode al variare del parametro soglia ε.

Nella tabella 5.1 viene riportato il numero di mode al variare del parametro ε, considerando numerose prove effettuate su diversi semi per la generazione casuale, mentre per quanto riguarda il *bandwidth*, è stato selezionato quello che, come suggerito dalla letteratura del metodo, permette la massimizzazione delle mode significative. Si nota che la scelta della soglia del valore di ε può influire molto sui risultati.

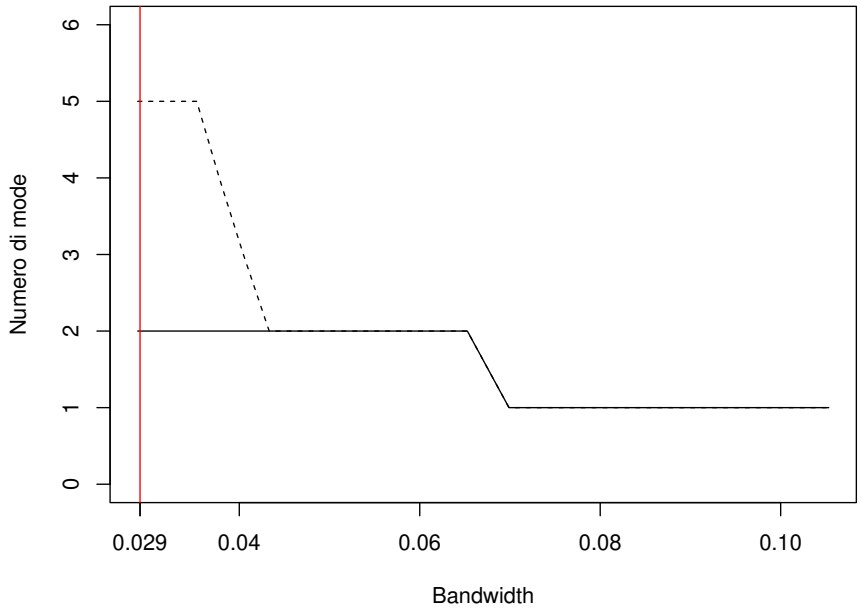


Figura 5.2: Numero di mode identificate dal metodo (significative in linea continua e non significative in linea tratteggiata) al variare del *bandwidth*, in rosso viene delineato il valore scelto per il *bandwidth*.

In seguito verranno presentati i risultati per ε fissato a 0.18 in quanto questo valore ha dimostrato di fornire una rappresentazione più accurata e interpretabile dei dati rispetto ad altri valori di ε. È importante sottolineare che, sebbene i risultati ottenuti con altri valori di ε siano stati esaminati, essi sono generalmente analoghi e coerenti con quelli riportati per $\varepsilon = 0.18$, senza presentare evidenti differenze.

Quando si fissano il seme e la soglia ε è necessario effettuare la scelta della larghezza di banda con cui stimare il kernel. In figura 5.2 è riportato l'andamento del numero di mode al variare del *bandwidth* selezionato. Si nota che, all'aumentare della larghezza di banda, il numero di mode diminuirà, fino a raggiungere il plateau di 1. In questo caso

verrà selezionato il *bandwidth* pari a 0.029, come evidenziato in figura, poichè massimizza il numero di mode significative, tuttavia è possible notare che ad un numero di mode significative corrisponde un numero di mode non significative come indicato dalla linea tratteggiata. Queste saranno le mode identificate come "mode candidate" attraverso il metodo *Mean Shift*, tra queste vengono incluse anche quelle rilevate significative attraverso la procedura di test di significatività delle mode riportata alla sezione 4.3.1. Una volta fissati tutti i parametri variabili vengono individuate due mode, come è possibile osservare in figura 5.3.

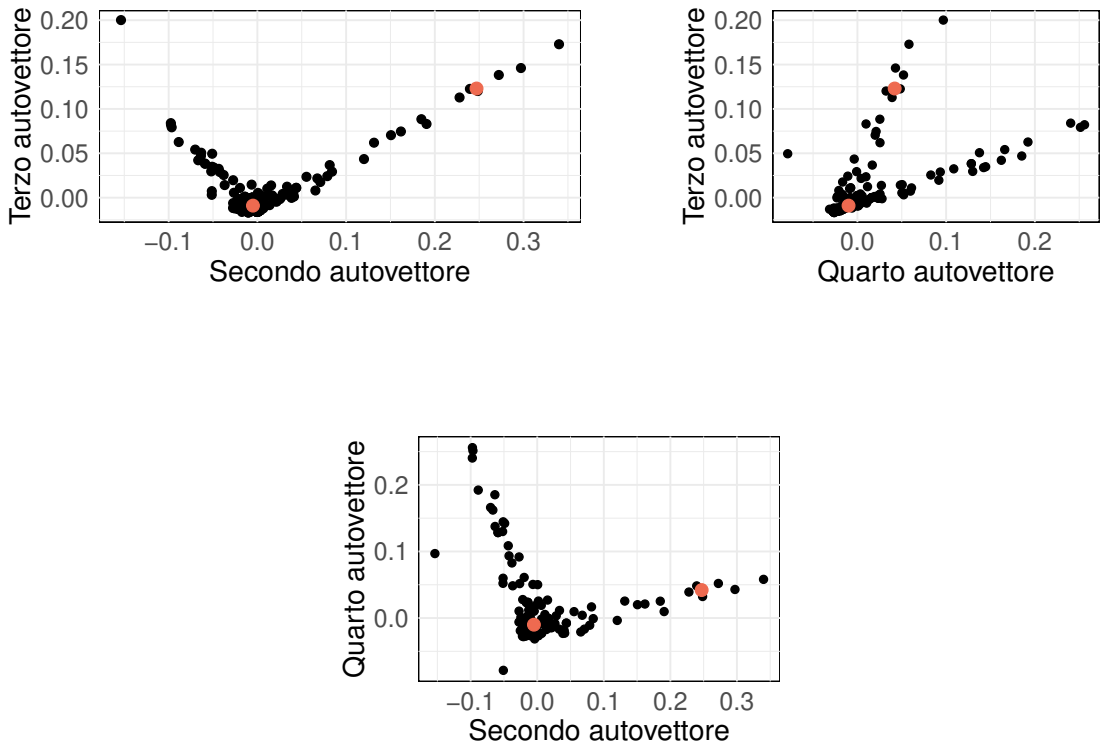


Figura 5.3: Scatter plot della distribuzione dell'embedding con le due mode significative rappresentate dai punti colorati.

Nella figura 5.4 sono riportati tutti e 5 i cluster, inclusi quelli non significativi, al fine di osservare i diversi gruppi dell'embedding dei dati. Dei 5 gruppi, solo il primo e il quinto sono significativi, con 68 e 6 componenti rispettivamente. Gli altri 3 non risultano significativi, con una numerosità pari a 2, 7 e 3. È interessante notare che il quinto gruppo, pur essendo significativo, contiene meno individui rispetto al primo, il quale è invece molto più numeroso

In figura 5.5 invece sono riportati esclusivamente i cluster significativi: il primo è rappresentato in rosa, mentre il quinto in verde, i punti neri appartengono a cluster non

significativi. Successivamente, risulta vantaggioso condurre un'analisi post-clustering per esaminare e commentare le caratteristiche dei gruppi identificati.

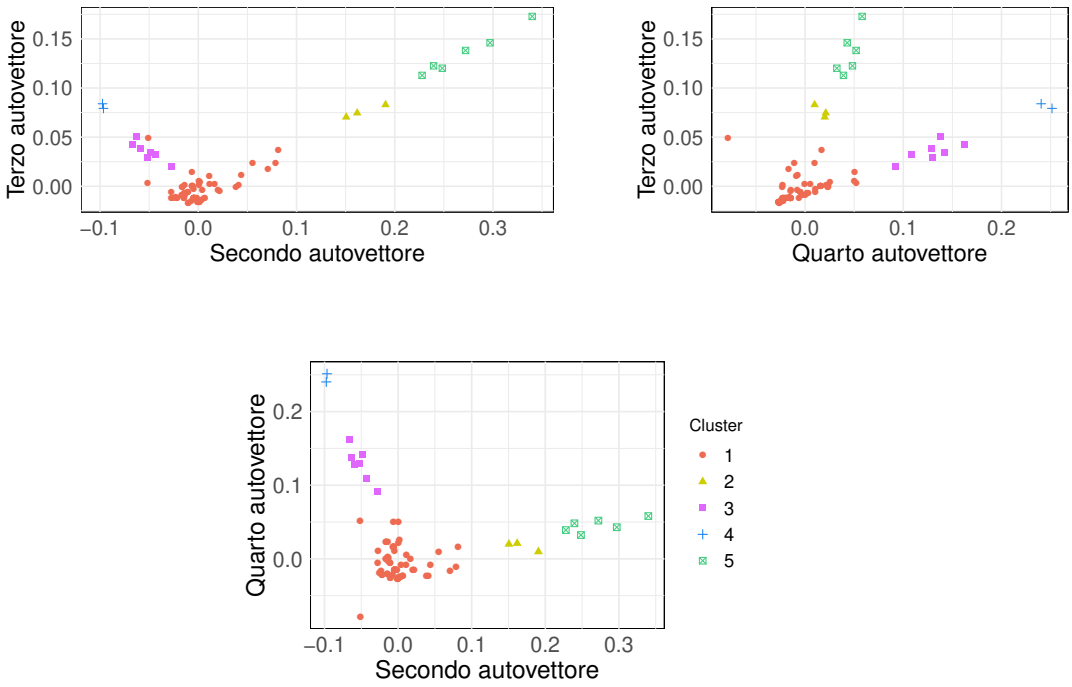


Figura 5.4: Scatter plot della distribuzione dell'embedding con i cluster significativi e non significativi evidenziati.

Nella figura 5.6 è rappresentata la distribuzione dei punteggi totali NIHSS per ciascun cluster. Ogni boxplot illustra la distribuzione dei punteggi totali NIHSS all'interno di un cluster specifico. Un'osservazione interessante emersa dall'analisi è che, all'aumentare del numero del cluster, ci sia un aumento tendenziale del valore della mediana dei punteggi totali NIHSS. Questo suggerisce che, i pazienti assegnati a cluster con un numero più alto, possono avere una maggiore gravità dei sintomi, come indicato dai punteggi NIHSS. In particolare osservando i due cluster significativi 1 e 5 si può notare la loro forte differenza in termini di punteggio totale NIHSS, i primi avranno una gravità inferiore dei sintomi mentre i pazienti appartenenti al quinto gruppo presentano una gravità maggiore di danni neurologici.

In figura 5.7 si possono notare delle differenze, seppur non molto marcate, in termini di età mediana. In particolare, il primo cluster presenta un'età mediana più elevata rispetto agli altri, ma è anche caratterizzato da una certa variabilità e presenza di osservazioni distanti dalla mediana. Al contrario, il quinto cluster mostra un'età mediana inferiore rispetto al primo.

La figura 5.8 offre una panoramica delle caratteristiche dei due cluster significativi e le confronta tra loro. In particolare i punti rappresentano le medie di ogni variabile all'interno

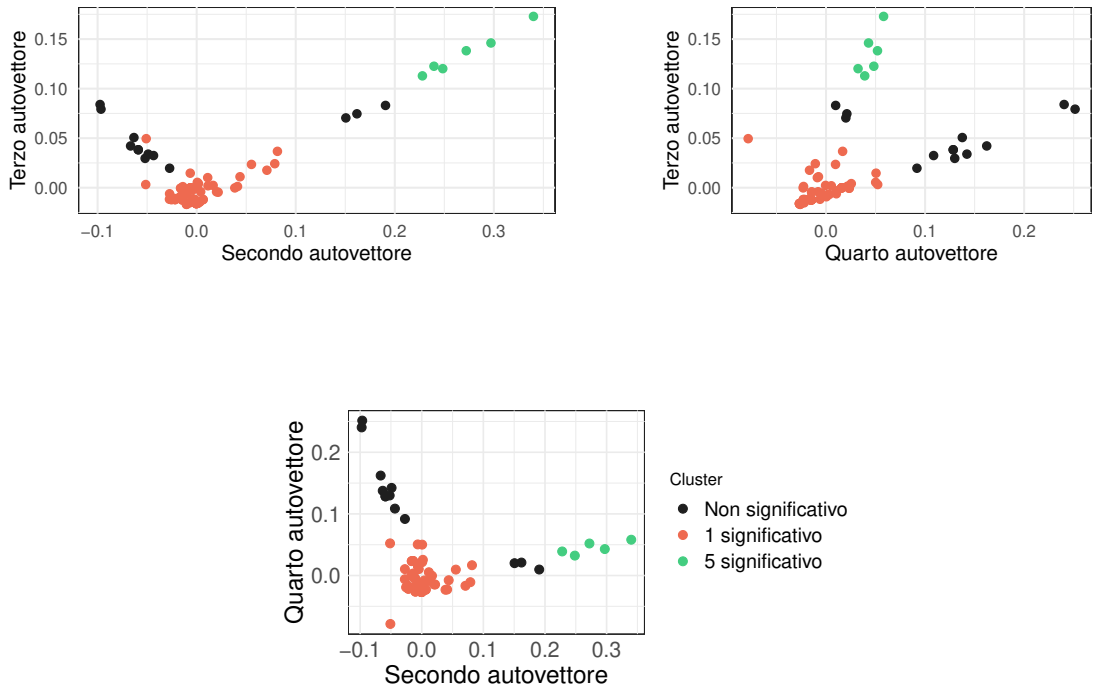


Figura 5.5: Scatter plot della distribuzione dell'embedding con i due cluster significativi (punti colorati) messi in evidenza.

dei cluster, mentre il raggio indica il massimo rilevato per ciascuna variabile, con il deficit minimo posizionato sul cerchio interno, che rappresenta lo zero.

Si può notare che il primo cluster (indicato in rosa) presenta valori medi molto più bassi rispetto al quinto (indicato in blu) e non evidenzia nessun picco significativo, suggerendo che mediamente i pazienti appartenenti a questo cluster assumono valori di deficit inferiori o nulli. Questi risultati sono in linea con quelli emersi dalla figura 5.6. Diversamente il quinto cluster presenta valori medi molto diversi in base alla variabile, ad esempio si nota che per alcuni deficit nessun paziente appartenenti al quinto cluster ha riportato valori superiori a 0 indicando l'assenza di sintomatologia specifica. Questo è il caso dei deficit relativi alla capacità del paziente di seguire comandi, esprimersi verbalmente, coordinare i movimenti degli arti e il movimento dell' arto superiore e inferiore destro. Al contrario i rimanenti deficit riportano valori più alti come evidenziato dai valori prossimi al massimo per il movimento dell' arto superiore e inferiore sinistro e all'attenzione del paziente. Inoltre anche la paralisi facciale e la vista del paziente riportano valori abbastanza alti.

In conclusione, l'analisi dei dati evidenzia differenze significative tra i due cluster considerati. Si può confermare ulteriormente la gravità dei sintomi riportati dai pazienti del quinto cluster che mostrano una maggiore variabilità nei valori medi dei deficit riportati,

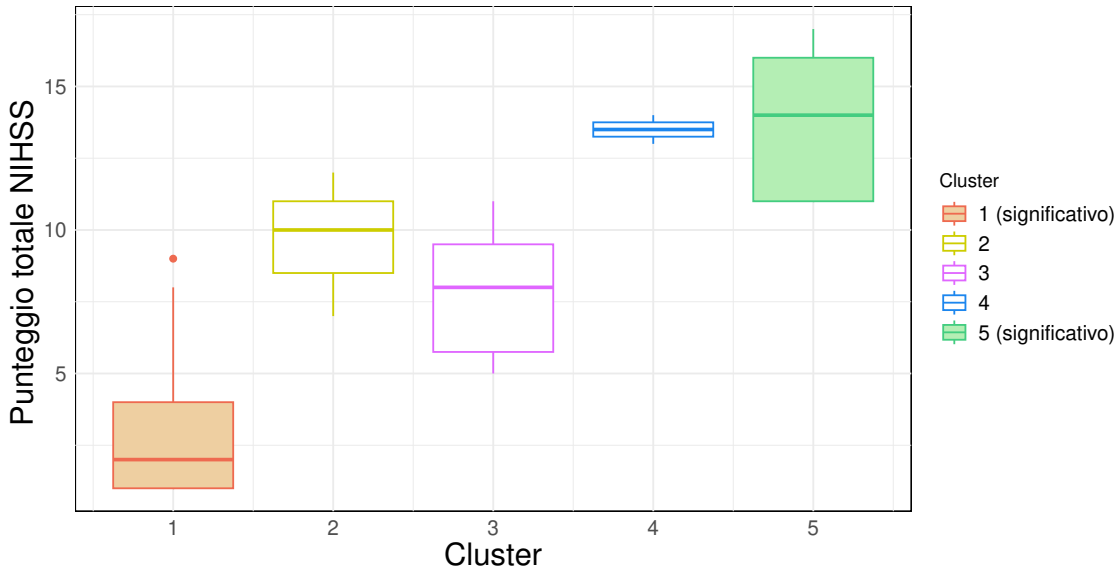


Figura 5.6: Distribuzione dei punteggi totali NIHSS per ogni cluster identificato (1 e 5 significativi in arancione e in verde).

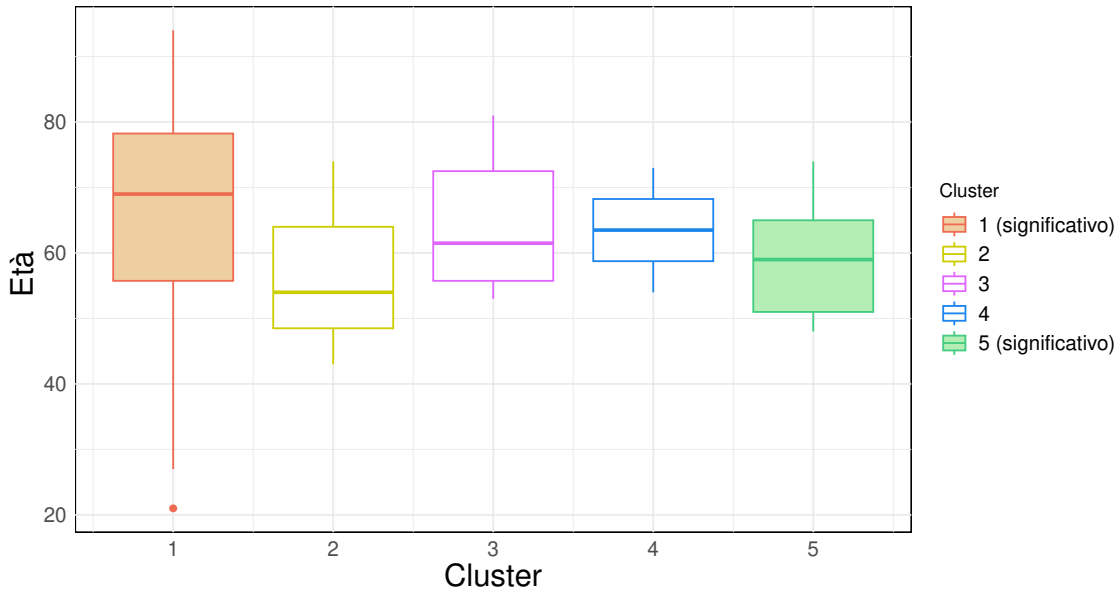


Figura 5.7: Distribuzione dell' età dei pazienti per ogni cluster identificato (1 e 5 significativi in arancione e in verde).

con alcune aree specifiche di sintomi particolarmente gravi. Questi risultati suggeriscono che i pazienti del quinto cluster potrebbero richiedere un approccio terapeutico più specifico e mirato, tuttavia sono necessarie delle competenze mediche per comprendere appieno

le implicazioni cliniche di queste differenze e guidare le decisioni terapeutiche appropriate.

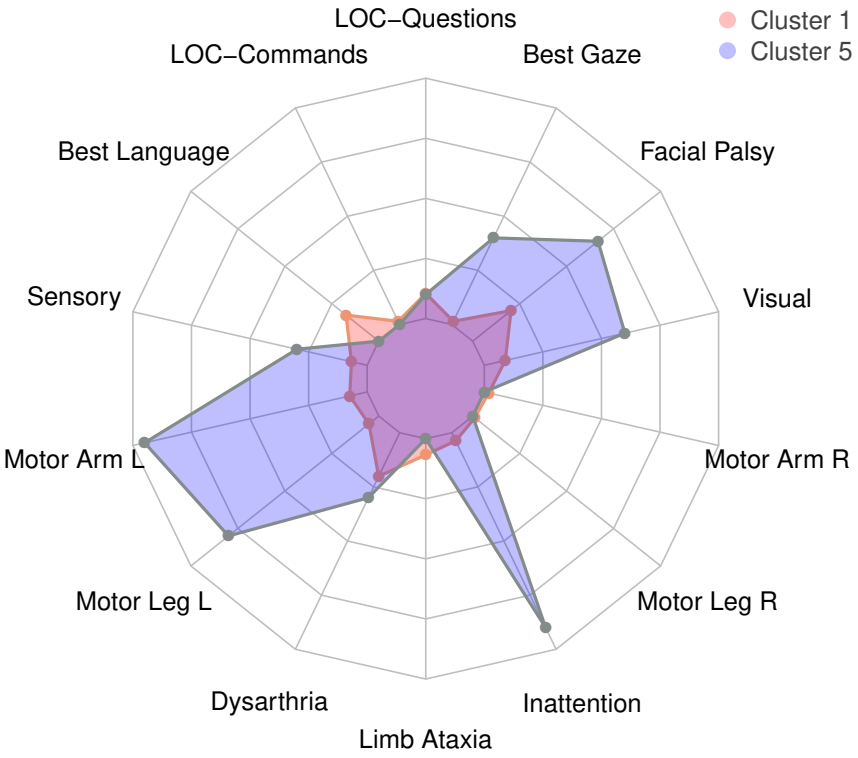


Figura 5.8: Radar plot dei valori medi dei due cluster significativi (gruppo 1 in rosa e 5 in blu) per ogni deficit calcolato con NIHSS. I punteggi medi sono misurati su una scala tra il minimo e il massimo della variabile in questione.

Conclusione

In questo lavoro, è stata esaminata una serie di metodi non parametrici per dentificare cluster nei dati relativi ai pazienti ischemici.

L'obiettivo principale della tesi era quello di individuare pattern significativi tra i pazienti ed estrarre informazioni sui soggetti attraverso il raggruppamento dei dati al fine di migliorare la comprensione della condizione dei pazienti.

L'elaborato si focalizza nell'affrontare i problemi intrinsechi al clustering presentando il clustering basato sulla densità per superarli. Quest'ultimo si concentra sulla creazione di gruppi identificando le regioni di alta densità nei dati.

Al fine di trattare i dati ordinali in modo appropriato e interpretabile, è stata adottata una trasformazione dei dati che come prima cosa ha coinvolto la creazione di una matrice di distanza generalizzata, consentendo una rappresentazione più accurata delle relazioni tra le osservazioni. Successivamente, è stato applicato l'embedding spettrale per risolvere diversi problemi legati al clustering sfruttando anche la riduzione della dimensionalità. Questa combinazione di trasformazione dei dati mediante la distanza generalizzata e il successivo embedding spettrale ha permesso di proiettare i dati originali su uno spazio euclideo, migliorando la comprensione e la visualizzazione delle relazioni tra i punti.

Per affrontare la complessità dei dati e individuare pattern significativi, è stato adottato un approccio basato su tre fasi. Inizialmente è stato eseguito un test sull'unimodalità, noto come *dip* test, con una modifica per poterlo effettuare anche in contesti multidimensionali. Esso ha rifiutato l'ipotesi nulla, suggerendo una possibile distribuzione multimodale dell'embedding dei dati. Successivamente sono state combinate due metodologie complementari. La prima consiste in un'analisi preliminare finalizzata a ottenere una panoramica iniziale delle caratteristiche significative, identificando regioni che forniscono informazioni sulla struttura dell'embedding. Questo metodo, ha permesso di individuare una regione significativa della curvatura dove si presume possano esserci variazioni significative o possibili mode. Successivamente, è stata approfondita l'analisi attraverso l'applicazione del metodo non parametrico basato sulla densità, il quale ha confermato la presenza di una moda significativa nella regione presa in analisi. Questo approccio utilizza il *Mean shift* per identificare le mode nella distribuzione dei dati, per poi testare la loro significatività attraverso la creazione di intervalli di confidenza per le stesse. L'integrazione di queste due metodologie offre una visione completa e approfondita dei dati, consentendo di indivi-

duarne la possibile suddivisione in gruppi e permette di discernere tra cluster significativi e non, caratteristica non comune a tutti i metodi di clustering.

Attraverso l'applicazione di questa procedura sono stati identificati due gruppi significativi che si distinguono principalmente in base al loro punteggio totale NIHSS, il quale riflette l'entità dei deficit cerebrali dei pazienti. Questi due gruppi rappresentano una suddivisione efficace dei pazienti in base alla gravità della loro condizione, con il gruppo più numeroso caratterizzato da punteggi NIHSS più bassi o medi e il gruppo meno numeroso comprendente pazienti con una condizione clinica particolarmente compromessa da gravi danni neurologici. Inoltre è emerso che il gruppo con punteggi più alti risulta avere deficit specifici come ad esempio problemi motori agli arti inferiore e superiore sinistri, paralisi facciale e problemi alla vista.

Infine è importante sottolineare i vantaggi dell'utilizzo del clustering non parametrico basato sulla densità, che è stato il centro dell'analisi presentata. In primo luogo, grazie alla stima non parametrica della densità, il metodo è estremamente flessibile e non richiede vincoli rigidi sulla forma dei cluster. Questa flessibilità consente al clustering modale di adattarsi meglio alle molteplici strutture dei dati che possono presentarsi in diversi contesti.

Inoltre, il numero di cluster non è fissato a priori, ma è una proprietà intrinseca del metodo. Questo significa che il clustering basato sulla densità segue la struttura naturale dei dati anziché imporre una suddivisione prestabilita, distinguendosi così da molti altri metodi di clustering, come il K-means, che richiedono la specificazione del numero di cluster a priori. Uno svantaggio del metodo è che dipende da alcuni parametri come la soglia ε e la scelta del *bandwidth* e anche dal seme fissato per la suddivisione dei dati. In conclusione si può affermare che questa tesi si è occupata di creare una metodologia precisa e particolare per i dati ordinali in questione, al fine di individuare dei cluster di pazienti ischemici, sfruttando il metodo non parametrico basato sulla densità, vantaggioso per i motivi appena elencati e per l'assenza di forti assunzioni.

Bibliografia

Azzalini, A. & Torelli, N. (2007). Clustering via nonparametric density estimation. *Statistics and Computing* **17**, 71–80.

Duong, T., Cowling, A., Koch, I. & Wand, M. P. (2008). Feature significance for multivariate kernel density estimation. *Computational Statistics & Data Analysis* **52**, 4225–4242.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I. & Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **78**, 99–126.

Guo, F. R. & Shah, R. D. (2023). Rank-transformed subsampling: inference for multiple data splitting and exchangeable p-values. *arXiv preprint arXiv:2301.02739* .

Hartigan, P. & Lada, C. J. (1985). Ccd images of suspected herbig-haro objects. *Astrophysical Journal Supplement Series (ISSN 0067-0049), vol. 59, Nov. 1985, p. 383-396.* **59**, 383–396.

Hennig, C. (2015). What are the true clusters? *Pattern Recognition Letters* **64**, 53–62.

Hinneburg, A. & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. In *Knowledge Discovery and Datamining (KDD'98)*.

Jajuga, K., Walesiak, M. & Bak, A. (2003). On the general distance measure. In *Exploratory Data Analysis in Empirical Research: Proceedings of the 25 th Annual Conference of the Gesellschaft für Klassifikation eV, University of Munich, March 14–16, 2001*. Springer.

Kwah, L. K. & Diong, J. (2014). National institutes of health stroke scale (nihss). *Journal of physiotherapy* .

Maechler, M. (2013). Package 'diptest'. *R Package Version 0.75–5. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing* .

Ng, A., Jordan, M. & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems* **14**.

Nilsson, E., Lukasczyk, J., Masood, T. B., Garth, C. & Hotz, I. (2023). Probabilistic gradient-based extrema tracking. In *2023 Topological Data Analysis and Visualization (TopoInVis)*. IEEE.

Reddy, C. K. (2018). *Data clustering: algorithms and applications*. Chapman and Hall/CRC.

Shah, G. H., Bhensdadia, C. & Ganatra, A. P. (2012). An empirical evaluation of density-based clustering techniques. *International Journal of Soft Computing and Engineering (IJSCE) ISSN* **22312307**, 216–223.

Tshimanga, L. F., Zanola, A., Facchini, S., Bisogno, A. L., Pini, L., Atzori, M. & Corbetta, M. (2023). Behavioral clusters in ischemic stroke based on nihss similarity. *medRxiv* , 2023–11.

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing* **17**, 395–416.

Walesiak, M. & Dudek, A. (2010). Finding groups in ordinal data: an examination of some clustering procedures. In *Classification as a Tool for Research: Proceedings of the 11th IFCS Biennial Conference and 33rd Annual Conference of the Gesellschaft für Klassifikation eV, Dresden, March 13-18, 2009*. Springer.

Wasserman (2017). Clustering.

Wishart, D. (1969). 256. note: An algorithm for hierarchical classifications. *Biometrics* , 165–170.