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

Statistica per le Tecnologie e le Scienze

Procedure robuste nel clustering modale: alcuni approfondimenti

Relatore: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Correlatore: prof Luca Greco
Università Giustino Fortunato di Benevento

Laureando: Marco Rudelli
Matricola n. 1225550

Anno Accademico 2021/2022

Indice

Introduzione

In ambito statistico è noto come ogni modello, che miri a descrivere un qualsiasi tipo di fenomeno, sia sempre e solo un'approssimazione, o una versione semplificata, della realtà. È quindi sostanzialmente certa, nella pratica, la presenza di alcune deviazioni dalle assunzioni modellistiche, che mettono possibilmente a rischio la validità dell'inferenza. Le deviazioni dalle ipotesi del modello si materializzano spesso sotto forma di contaminazione, ossia con la presenza, nel campione, di una frazione di osservazioni anomale (*outlier*), lontane dalla massa dei dati, che non provengono dal processo generatore del fenomeno di interesse. Ignorare tali anomalie può condurre ad un'inferenza largamente fuorviante. Si parla di statistica robusta con riferimento all'insieme dei metodi che, anche in queste situazioni, mantengano l'efficacia e l'affidabilità dei risultati inferenziali.

La *cluster analysis*, ossia la ricerca di gruppi latenti in una data popolazione, è un ambito della statistica che, come gli altri, soffre delle problematiche appena citate. In questo caso è pero necessaria una particolare attenzione, in quanto il concetto di *outlier* diventa relativo ai *cluster*, e non relativo alla popolazione. Questo aggiunge un livello di difficoltà al problema dello sviluppo di metodi di *clustering* robusti, che dovranno procedere in parallelo con l'individuazione degli *outlier* e quella dei gruppi. Numerose opzioni di tecniche robuste esistono nel campo del *clustering* parametrico, ma il problema è stato per lo più trascurato nell'ambito del *clustering* modale o non parametrico.

Nel *clustering* modale i gruppi sono associati alle mode della densità della popolazione, stimata non parametricamente; questa formulazione al problema di *clustering* ha il vantaggio di consentire l'individuazione di gruppi di forma arbitraria, il cui numero è una proprietà intrinseca del processo generatore dei dati, ed è quindi parte del procedimento di stima. D'altro canto, essendo nota la vulnerabilità dei metodi di stima non parametrici alla presenza di osservazioni anomale, è chiara la necessità di sviluppare appropriate tecniche per irrobustire il *clustering*.

In questo caso, l'obiettivo dell'identificazione degli *outlier* si intreccia quindi con l'individuazione delle mode spurie, che possono nascere nelle regioni affette da sparsità a causa delle osservazioni anomale. È infatti possibile che un gruppo di *outlier* abbastanza vicini sia scambiato per un *cluster*, ed è, al contrario, possibile che le osservazioni appartenenti ad un gruppo di piccole dimensioni vengano scambiate per *outlier*.

Il contributo di questo lavoro consisterà appunto nell'esplorazione dei problemi di robustezza legati al clustering modale, e nella proposta di una metodologia di *trimming* che permetta di identificare, ed eliminare, le osservazioni valutate come più anomale.

La trattazione si sviluppa come segue.

Nel Capitolo 1, dopo una panoramica generale sui metodi di *clustering*, si espone in dettaglio l'approccio modale del *clustering* modale, e le sue proprietà.

Nel Capitolo 2 si introduce il concetto di robustezza e il problema della contaminazione, sotto forma di valori anomali. Prima i concetti saranno esposti in generale, e poi a nello specifico per il *clustering*. Si descriveranno in breve alcune tecniche di *clustering* robusto esistenti in letteratura.

Nel Capitolo 4 si introdurrà il problema della contaminazione per quanto riguarda i metodi non parametrici, e si procederà con una descrizione del metodo di *trimming* proposto.

Nel Capitolo 5, si esporranno i risultati di uno studio di simulazione per valutare il comportamento del metodo proposto riguardo l'identificazione degli *outlier* e in merito all'irrobustimento del *clustering*. Si confronterà inoltre il metodo con una tecnica proveniente dal mondo parametrico.

Capitolo 1

Clustering non parametrico

1.1 Introduzione

Le tecniche che permettono di ripartire in gruppi un set di unità, secondo le loro caratteristiche, fanno parte dell'ambito noto come *cluster analysis* o *clustering*. La specificazione di cosa effettivamente costituisca nei dati un gruppo (*cluster*) assume un'importanza fondamentale in questo campo, e le differenze in questa definizione (o l'assenza di una definizione formale) hanno portato allo sviluppo di svariati metodi di raggruppamento diversi. Quelli ad oggi esistenti possono essere sostanzialmente categorizzati in due tipologie.

La prima comprende le tecniche che usano come criterio per il raggruppamento una misura di distanza (o dissimilarità) tra le osservazioni. Un cluster è qui intuitivamente definito come insieme di unità relativamente vicine (o simili) tra loro e distanti (o diverse) dalle unità degli altri gruppi. Alcuni esempi che ricadono in questa categoria sono i *metodi gerarchici*, tra i quali i metodi del *legame completo*, *singolo* o *medio* e la tecnica delle *K-medie*. Dato un campione $\mathcal{X} = \{x_1, \ldots, x_n\}, x_i \in \mathbb{R}^d$, quest'ultima costruisce una partizione $S = \{S_1, \ldots, S_K\}$ in K gruppi sulla base del criterio di ottimizzazione espresso da:

$$\underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x - \bar{x}_i\|^2 \tag{1.1}$$

dove $\bar{x}_k = \frac{1}{n} \sum_{x_i \in S_k} x_i$.

Il metodo cerca quindi, tra tutte le possibili partizioni dei dati, quella che minimizzi la distanza euclidea delle osservazioni dai centri dei

cluster, associati alle medie campionarie dei punti a loro appartenenti. Ciò porterà intrinsecamente a prediligere gruppi di simile dimensione e di forma pressochè sferica.

L'ottimizzazione della suddetta funzione avviene in maniera iterativa. Dopo un'inizializzazione casuale dei punti nei gruppi, vengono calcolate le rispettive medie campionarie; queste determinano a loro volta l'aggiornamento dei gruppi tramite l'associazione di ogni punto alla più vicina media. È quindi possibile aggiornare nuovamente le medie e iterare fino a convergenza l'algoritmo.

Pur essendo molto utilizzati a fronte di una semplicità concettuale e computazionale, i metodi di *clustering* basati sulla distanza si fondano su una specificazione euristica del concetto di gruppo, e soffrono appunto della mancanza di una definizione statistica formale di cosa costituisca un *cluster* in una data popolazione. Ciò si traduce nell'assenza di una vera struttura di fondo che il processo di stima cerchi di approssimare utilizzando il campione, provoca l'impossibilità di utilizzare procedure inferenziali per valutare o comparare risultati e, soprattutto, l'impossibilità di definire quale sia il vero numero di gruppi della popolazione in esame. Per un approfondimento si vedano Kaufman e Rousseeuw (2005).

La seconda tipologia di metodi di *clustering* comprende i metodi basati sulla densità, nei quali il concetto di *cluster* è associato ad alcune caratteristiche specifiche della distribuzione di probabilità che contraddistingue il processo generatore dei dati. A differenza dei metodi basati sulla distanza, la presenza di una precisa definizione di gruppo è la base per la specificazione di un modello statistico più rigoroso, e permette di fare inferenza sul numero di cluster o sulla bontà dei risultati ottenuti. Questo *framework* è stato sviluppato in due direzioni: mediante l'utilizzo di modelli parametrici per descrivere la densità, dando origine all'approccio noto come *model-based clustering*, o mediante l'utilizzo di metodi non parametrici, dando origine a ciò che viene definito *clustering modale*. Nel seguito di questo paragrafo si offrirà una panoramica dei primi, mentre ai secondi sarà dedicato un approfondimento nei paragrafi successivi, in quanto oggetto di questo lavoro.

La formulazione parametrica al problema di *clustering* assume che ogni gruppo sia descritto da un modello probabilistico $f_k(x) = f(x; \theta_k)$, $k = 1, \ldots, K$, dipendente da un vettore di parametri θ_k, e che il cam-

pione $\mathcal{X} = \{x_1, \ldots, x_n\}$ sia generato da un modello a mistura finita

$$f(x) = \sum_{k=1}^{K} \pi_k f(x; \theta_k)$$

I parametri del modello $\theta = (\theta_1, \ldots, \theta_K)$ e $\pi = (\pi_1, \ldots, \pi_K)$, $\pi_k > 0, \sum_{k=1}^{K} \pi_k = 1$ possono essere stimati mediante la massima verosimiglianza le cui equazioni, tuttavia, tipicamente non ammettono soluzione esplicita e devono essere risolte numericamente.

Un algoritmo utile a risolvere elegantemente il problema dell'ottimizzazione numerica è noto con il nome di Expectation-Maximization (EM, Dempster *et al.*, 1977). Originariamente sviluppato per massimizzare una verosimiglianza in presenza di dati mancanti, l'algoritmo EM distingue, con riferimento ad un parametro θ di interesse, da stimare mediante i dati x e in mancanza dei dati z la cosiddetta log-verosimiglianza completa

$$\ell(\theta|x, z) = \log p(x, z|\theta)$$

dalla verosimiglianza incompleta

$$\ell(\theta|x) = \log p(x|\theta).$$

Poiché Z non è osservato, anziché massimizzare $\ell(\theta|x, z)$, l'idea dell'algoritmo EM è di massimizzare il suo valore atteso, condizionatamente a x. A questo scopo si ripetono, iterativamente ($t = 1, \ldots, T$), due passi:

1. E-step (Expectation) fissato $\theta = \theta^{(t)}$, si determina il valore atteso

$$E_z \left(\log p(x, z|\theta^{(t)})|x, \theta^{(t)} \right)$$

2. M-step (Maximization) al variare di θ, si massimizza il valore atteso calcolato al passo E e si determina

$$\theta^{(t+1)} = argmax_\theta E_z \left(\log p(x, z|\theta)|x \right)$$

fino a convergenza dell'algoritmo. Mediante questo procedimento, poiché si massimizza ad ogni passo un minorante della log-verosimiglianza, è garantita anche la crescita di quest'ultima, e la convergenza ad un massimo locale.

Questo ragionamento è stato applicato per la massimizzazione della verosimiglianza di un modello mistura, assumendo che il dato mancante z_i esprima per ogni individuo l'appartenenza al gruppo, e sia

descritto da un vettore multinomiale latente $Z_i = (Z_i^{(1)}, \ldots, Z_i^{(K)}) \sim Bin_K((\pi_1, \ldots, \pi_K); 1)$ che assume valore 1 in posizione k quando l'osservazione i appartiene al cluster k, $\pi_j = P(Z^{(k)} = 1)$. Semplici calcoli algebrici mostrano che l'equazione della log-verosimiglianza completa di un modello mistura diventa

$$\ell(\theta, \pi | x_1, \ldots, x_n, z_1, \ldots, z_n) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \left(\pi_k f(x_i | \theta_k) \right). \qquad (1.2)$$

Ad ogni iterazione, il passo E dell'algoritmo si riduce al calcolo delle quantità

$$\tau_{ij}^{(t)} = \frac{\pi_j f(x_i | \theta_j^{(t)})}{\sum_j \pi_j^{(t)} f(x_i | \theta_j^{(t)})}.$$

La stima del parametro θ viene successivamente aggiornata al passo M:

$$\begin{aligned} \hat{\theta}^{(t+1)}, \hat{\pi}^{(t+1)} &= argmax_{\theta,\pi} E_z \left(\log f(x, z | \theta) | x \right) \\ &= argmax_{\theta,\pi} \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ij}^{(t)} \log \left(\pi_j f(x_i | \theta_j) \right) \end{aligned}$$

Il procedimento viene ripetuto fino all'iterazione T che garantisce la convergenza.

Una volta stimati i parametri della mistura, l'osservazione i verrà assegnata alla componente della mistura $\tilde{k}$ per cui risulta massima la probabilità a posteriori stimata

$$\tilde{k} : argmax_k \tau_{ik}^{(T)}.$$

L'utilizzo di un modello parametrico ai fini del *clustering* porta vantaggi e svantaggi. I primi sono soprattutto legati agli strumenti statistici utilizzabili, come la possibilità di valutare l'incertezza legata al raggruppamento attraverso le probabilità a posteriori o l'opportunità di determinare il numero di gruppi attraverso l'uso di criteri di scelta del modello quali il *Bayesian Information Criterion*. Gli svantaggi sono invece soprattutto legati alla rigidità della modellazione della densità: è evidente che i gruppi debbano avere una forma compatibile con il modello parametrico definito per le componenti e che eventuali allontanamenti delle distribuzioni di ogni cluster da quelle specificate parametricamente mettono a rischio la qualità dei risultati. Per un approfondimento recente su questi metodi si veda il testo di rassegna di Bouveyron *et al.* (2019).

1.2 I domini di attrazione delle mode

In risposta agli svantaggi del *clustering* parametrico, l'idea della sua controparte non parametrica, nota anche come *clustering modale*, è proprio quella di non dare una forma predeterminata ai gruppi, ma di specificarli concettualmente come *"regioni dello spazio continue e relativamente densamente popolate, circondate da regioni continue e relativamente vuote"* (Carmichael e Julius, 1968). Tale principio porta quindi ad associare un cluster ad ogni diversa regione ad elevata concentrazione del supporto. Si parla in questo senso di *regione modale* o di *dominio di attrazione* di una moda.

Ciò permette, pur fornendo una grande libertà alla modellazione della forma dei singoli cluster, di mantenere un solido contesto statistico a supporto dei risultati. Il numero di gruppi è inoltre una proprietà intrinseca della densità di interesse ed è automaticamente associato ad ogni stima.

Nei paragrafi che seguono si fornirà una panoramica dell'approccio non parametrico al *clustering* da un punto di vista formale e operativo.

Seppur già informalmente introdotta nella seconda metà del secolo scorso, l'idea di associare i gruppi alle regioni modali della funzione di densità sottostante i dati ha trovato una collocazione formale solo recentemente (Chacón, 2015), grazie al ricorso alla teoria Morse, una branca della geometria differenziale che descrive il comportamento di oggetti matematici attraverso l'analisi dei punti critici di una funzione.

Si assume che la funzione di densità $f : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$, che ha generato il campione a disposizione, appartenga all'insieme di funzioni Morse, un ampio insieme di funzioni lisce e prive di punti critici degeneri (si veda Matsumoto, 2002, per un approfondimento). A partire da f, è possibile definire, per ogni $x \in \mathcal{D}$, la curva integrale $\nu_x : \mathbb{R} \mapsto \mathbb{R}^d$

$$\nu_x(0) = \mathbf{x} \qquad \nu'_{\mathbf{x}}(t) = -\nabla f(\nu_{\mathbf{x}}(t)), \tag{1.3}$$

che definisce la direzione del gradiente in x. Si dimostra che le destinazioni delle curve integrali

$$\theta = \lim_{t\to\infty} \nu_x(t)$$

rappresentano tutte e sole le mode di f e, poiché le curve integrali hanno unica intersezione nei punti critici di f, l'insieme delle destinazioni delle curve integrali definiscono una partizione $\{\mathcal{D}_\theta\}_{\theta\in\Theta}$ di $\mathcal{D}$ in involucri stabili

$$\mathcal{D}_\theta = \{\mathbf{x} : \lim_{t\to\infty} \nu_{\mathbf{x}}(t) = \theta\}$$

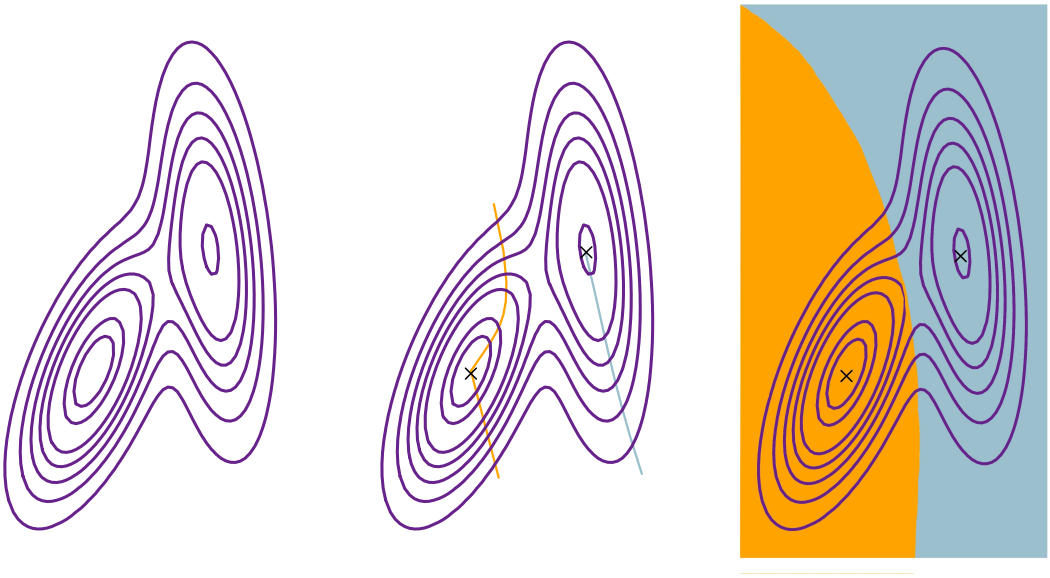

Figura 1.1: Esempio di funzione di densità bimodale (a sinistra), curve integrali associate a tre punti del dominio (al centro) e partizione del dominio in domini di attrazione delle mode.

formati dall'insieme dei punti la cui curva integrale ha la medesima destinazione θ. Ciascuna di tali regioni definisce il dominio di attrazione della moda θ corrispondente.

A dispetto della complessità matematica, il concetto di dominio di attrazione di una moda può essere intuitivamente colto immaginando di associare la densità f ad una catena montuosa i cui picchi rappresentano le mode di f: il dominio di attrazione di una moda è identificato dalla regione che risulterebbe bagnata se, dal picco associato, fuoriuscisse un flusso d'acqua, lasciato scorrere in assenza di attrito. Si veda la Figura 1.1 per una illustrazione in due dimensioni.

Una volta definito formalmente l'impianto teorico di riferimento, i principali aspetti di studio legati al *clustering modale* riguardano la stima della densità, effettuata tipicamente mediante metodi non parametrici, e la successiva individuazione delle regioni modali dello spazio campionario, relativamente dense, alle quali associare i cluster. È inoltre di notevole importanza la gestione della dimensionalità dei dati, con le relative criticità che sorgono con la crescita del numero di variabili. Tali aspetti vengono approfonditi nei paragrafi che seguono.

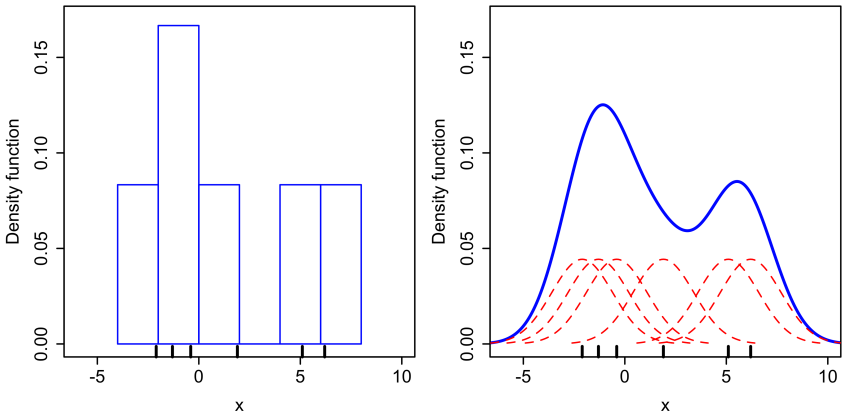

Figura 1.2: Istogramma e stima *kernel* per lo stesso campione

1.3 Stima non parametrica della densità

Il *metodo del nucleo* o *kernel* (Parzen, 1962), è sicuramente la tecnica di stima non parametrica della densità ad oggi più conosciuta e utilizzata. È paragonabile, a livello intuitivo, ad una generalizzazione del più semplice concetto di istogramma, e permette la stima della funzione di densità, legata ad un campione, senza l'appoggio a forme parametriche e alle rigidità a loro collegate (Figura 1.2).

Formalmente, nel caso univariato, dato un campione di n osservazioni $\mathcal{X} = \{x_1, \ldots, x_n\}$ dalla variabile aleatoria X avente densità $f : \mathbb{R} \to \mathbb{R}^+ \cup \{0\}$, lo stimatore *kernel* di f è definito come:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_h (x - X_i) \tag{1.4}$$

$K_h(\cdot)$ è nota come funzione *kernel* scalata, definita come $K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right)$. $K(\cdot)$ è il *kernel* vero e proprio, funzione a valori reali integrabile e non negativa. Nella pratica, per preservare le caratteristiche di $f(\cdot)$ quale stima di una funzione di densità, è richiesto che K rispetti $\int_{-\infty}^{\infty} K(u)\, du = 1$. La scelta di K influenza la forma della stima di f, ma è spesso di poca importanza a livello pratico: purché coerente con le proprietà necessarie, non provoca variazioni di rilievo nei risultati. Il *kernel* normale è la scelta più comune, e pone $K(x) = \phi(x)$, dove $\phi(\cdot)$ è la densità della distribuzione normale standard. Più importante è invece la scelta di h, detto parametro di lisciamento o *bandwidth*. Questo è appunto utile a regolare il grado di liscezza della stima; valori grandi tendono a rendere molto graduali le variazioni del livello di $\hat{f}$,

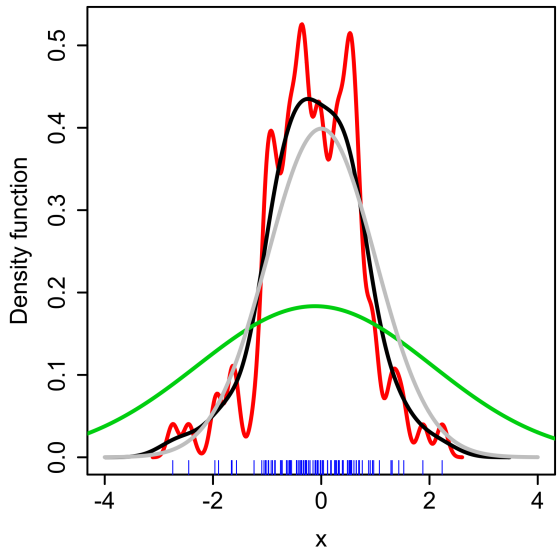

Figura 1.3: Densità stimata al variare di h per lo stesso campione

correndo il rischio di levigare eventuali mode fino a farle scomparire. Valori piccoli di h portano invece ad un risultato frastagliato e possibilmente sovra-adattato al particolare campione. Si veda la Figura 1.3 per un'illustrazione.

La scelta di h può anche essere vista sotto il punto di vista del compromesso distorsione-varianza, con valori piccoli associati a varianze maggiori e bias minori, mentre valori grandi caratterizzati da una maggiore stabilità del risultato ma una distorsione più elevata. Questo suggerisce la possibilità di utilizzare metodi di *cross validation* per la scelta di h; altre opzioni che guidino la selezione del parametro di lisciamento ricorrono invece all'appoggio di particolari famiglie parametriche per ottenere criteri analitici. Per un approfondimento sui metodi di selezione si veda ad esempio Wand e Jones, 1994

La valutazione della qualità dello stimatore si basa su misure di distanza tra la vera funzione di densità e la sua stima. Questa distanza può essere valutata su un punto fissato x appartenente al supporto, utilizzando quindi misure come l'*errore quadratico medio* (MSE), o integrata su tutto il supporto, ottenendo misure come l'*errore quadratico medio integrato* (MISE). Le due quantità sono definite formalmente come:

$$
\begin{aligned}
MSE_x\left(\hat{f}\left(x\right)\right) &= E\left[\left(\hat{f}\left(x\right) - f\left(x\right)\right)^2\right] \\
MISE\left(\hat{f}\right) &= E\left[\int\left(\hat{f}\left(x\right) - f\left(x\right)\right)^2 dx\right]
\end{aligned}
\tag{1.5}
$$

Spesso, per questioni di semplicità matematica, l'analisi non si concentra propriamente sul MISE ma sulla sua versione asintotica, ottenibile tramite sviluppi in serie di Taylor. Tale quantità, detta AMISE, si dimostra essere uguale a:

$$
AMISE\left(\hat{f}\right) = \frac{h^4}{4}\left(\int x^2 K(x)\, dx\right)^2 \int f''\left(x\right)^2\, dx + \frac{1}{nh}\int K\left(z\right)^2\, dz \tag{1.6}
$$

La formula, scomposta nella sommatoria di due termini, evidenzia ancora il ruolo di h come parametro legato al *bias-variance tradeoff*. Si può dimostrare infatti che il primo addendo, che cresce con il valore della bandwidth, sia effettivamente legato alla distorsione dello stimatore, mentre il secondo, inversamente proporzionale ad h, sia legato alla sua varianza.

La generalizzazione multivariata dello stimatore *kernel* non denota particolari criticità dal punto di vista della definizione. Sia $\mathcal{X} = \{x_1, \ldots, x_n\}$ un campione di n osservazioni, ognuna delle quali di dimensione d, $x_i = \{x_{i1}, \ldots, x_{id}\}$, provenienti da una variabile casuale multivariata X con densità $f : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$. Lo stimatore *kernel* è definito come:

$$
\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_H\left(x - X_i\right) \tag{1.7}
$$

Dove $K_H\left(x\right) = \left|H\right|^{-1/2} K\left(H^{-1/2}x\right)$ è la funzione *kernel* scalata. $K(\cdot)$ è la funzione kernel d-variata, avente sostanzialmente le stesse caratteristiche descritte in precedenza per la controparte univariata. Il parametro di lisciamento, H, è ora una matrice quadrata di dimensione d, simmetrica e definita positiva. In questo caso il problema associato alla selezione della bandwidth è potenzialmente legato a $d(d+1)/2$ parametri, portando ad un guadagno di flessibilità ma, nel contempo, introducendo ulteriore complessità alla questione della selezione. È di

conseguenza possibile semplificare la definizione della matrice, adottando ad esempio una struttura diagonale: $H = diag\left(h_1^2, \ldots, h_d^2\right)$. La parametrizzazione in ultimo più semplice prevede di tornare ad un parametro unidimensionale h, ponendo $H = h^2 I$. La semplificazione della struttura di H permette di ricondurre il problema della scelta della bandwidth a situazioni matematicamente più semplici, ed è per questo la strada più comunemente adottata nella pratica.

1.4 Individuazione dei cluster

Il tratto comune alle procedure di clustering non parametrico è quello di associare i cluster alle diverse regioni ad elevata concentrazione, o regioni modali, dello spazio campionario di interesse. Le diverse modalità di ricerca di queste regioni, e una più formale definizione delle stesse, portano poi però a differenze operative nelle modalità di individuare la partizione. Gli approcci più noti sono due: il primo fa uso di curve di livello, mentre il secondo mira alla ricerca delle mode tramite lo studio del gradiente della funzione di densità. Si veda Menardi (2015) per una panoramica sui metodi di *clustering modale* presenti in letteratura.

1.4.1 Metodi basati sulle curve di livello

Questo approccio al problema dell'individuazione delle regioni del supporto ad elevata concentrazione ha origine nelle idee di Hartigan (1975), il quale ha proposto di legare i cluster agli insiemi di livello connessi presenti nella funzione di densità. Formalmente, data una funzione di densità $f : \mathbb{R}^d \to \mathbb{R}^+ \cup \{0\}$ e una soglia $\lambda \in [0, max\,(f)]$, si definisce la regione dello spazio campionario dove la densità è superiore alla soglia

$$L\,(\lambda; f) = L\,(\lambda) = \left\{x \in \mathbb{R}^d : f\,(x) > \lambda\right\}. \tag{1.8}$$

$L(\lambda)$ può essere una regione connessa o meno a seconda del valore fissato di λ. Essendo $L(\lambda)$ ignoto, in quanto nella pratica non si conosce $f(x)$, si ottiene una stima $\hat{L}(\lambda)$ tramite il plug-in nella formula (1.8) di una stima non parametrica $\hat{f}$, spesso ottenuta tramite metodo *kernel*. Si noti come ogni regione connessa contenga al suo interno almeno una moda, e come per ogni moda esista un valore di λ tale che la corrispondente regione connessa contenga solo quella moda. In generale, non esiste un singolo valore di λ che individui tutte le regioni modali; per questo motivo il parametro viene fatto variare nell'intervallo $[0, max\,(f)]$ generando una

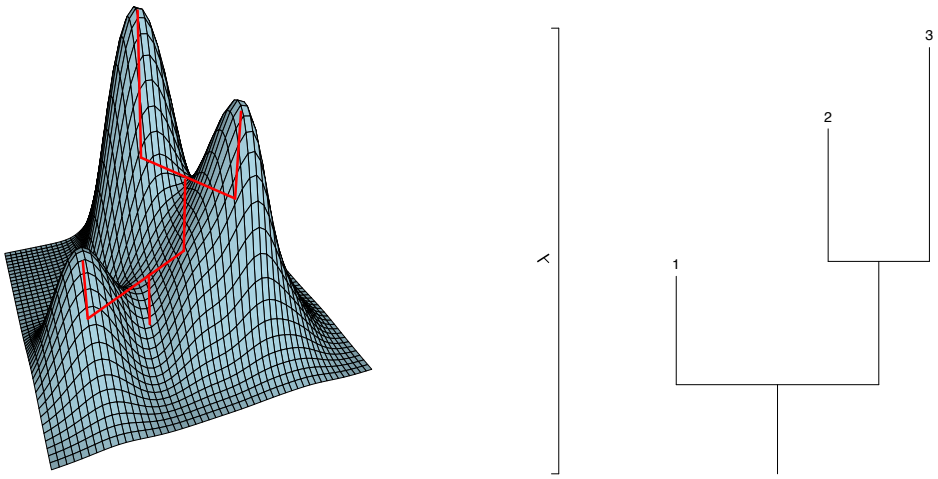

Figura 1.4: Esempio di una distribuzione bivariata trimodale e relativo *albero dei cluster*

costruzione gerarchica nota come *albero dei cluster*. Ad ogni foglia di tale albero corrisponde una moda distinta ed al loro numero corrisponde il numero di gruppi. Si veda, per un'illustrazione nel caso bidimensionale, la Figura 1.4.

Questo approccio operativo, seppur concettualmente intuitivo, trova la sua criticità dal punto di vista computazionale: la ricerca delle componenti connesse, data una stima $\hat{L}(\lambda)$, non è infatti un problema banale quando le dimensioni aumentano rispetto al caso univariato e le componenti diventano più complicate di semplici intervalli. Per risolvere il problema nel caso multidimensionale ci si appoggia usualmente alla teoria dei grafi, e le maggiori differenze tra i vari metodi di clustering basati sulle curve di livello della densità ricadono proprio nelle diverse modalitá di individuazione delle componenti connesse.

Superata questa fase ed ottenute le suddette componenti, le osservazioni ad esse associate andranno a definire i nuclei dei cluster; in generale, nelle code della distribuzione o nelle valli tra due mode, vi saranno osservazioni ancora non associate ad alcun gruppo. Queste andranno trattate ed assegnate ad un cluster tramite metodi di classificazione basati sulle osservazioni già etichettate; il metodo proposto da Azzalini e Torelli (2007) prevede ancora di utilizzare metodi di stima non parametrici: per ogni nucleo di cluster $k = 1, \ldots, K$ verrà calcolata la stima $\hat{f}_k(\cdot)$, basata sui punti già assegnati al dato nucleo. Un'osservazione x_0 sarà poi associata al gruppo per cui $\hat{f}_k(x_0) / max_{j \neq k} \hat{f}_j(x_0)$ è massima.

1.4.2 Metodi basati sulla ricerca delle mode

Questo tipo di approccio comprende metodi di raggruppamento che, attraverso un processo di ricerca degli ottimi locali della densità, assegnano ciascuna osservazione ad una moda di riferimento; ogni *cluster* sarà quindi l'insieme delle unità che fanno riferimento alla medesima moda, e il numero di mode corrisponderà al numero di gruppi nella partizione dei dati. Tra questi il metodo più conosciuto, e origine di gran parte delle evoluzioni in questa direzione, è noto come *mean-shift clustering* (Fukunaga e Hostetler, 1975). L'idea di fondo di questa metodologia consiste nel seguire per ogni osservazione il percorso ascendente dettato dal gradiente della densità, fino a giungere ad un ottimo locale.

In particolare, si consideri il gradiente $\nabla \hat{f}(\cdot)$ dello stimatore *kernel* definito in 1.7; Questo assume la forma:

$$\nabla \hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla K_H (x - X_i) \tag{1.9}$$

dove $\nabla K_H(\cdot)$ è il gradiente della funzione *kernel* scalata.

Partendo da un generico punto del supporto $x^{(0)}$ l'algoritmo aggiorna ricorsivamente la posizione del punto stesso attraverso una media pesata. Al passo s, dato $w_i\left(x^{(s)}\right)$ il vettore dei pesi di ciascuna componente di x_i, l'algoritmo definisce l'aggiornamento come:

$$\begin{aligned} x^{(s+1)} &= \sum_{i=1}^{n} w_i \left(x^{(s)}\right) x_i \\ &= x^{(s)} + \left[\sum_{i=1}^{n} w_i \left(x^{(s)}\right) x_i - x^{(s)}\right] \\ &= x^{(s)} + M(x^{(s)}) \end{aligned} \tag{1.10}$$

dove $M(x^{(s)})$ denota il *mean shift* e i pesi $w_i(x)$ sono, a meno di un fattore di normalizzazione, definiti come i vettori $\nabla K_H(x - x_i)$. Il mean shift si dimostra essere equivalente ad un algoritmo di risalita del gradiente basato su uno stimatore del gradiente normalizzato:

$$x^{(s+1)} = x^{(s)} + a \frac{\nabla \hat{f}\left(x^{(s)}\right)}{\hat{f}\left(x^{(s)}\right)} \tag{1.11}$$

In particolare l'equivalenza si dimostra vera quando $\nabla \hat{f}$ è una stima di ∇f basata su uno *shadow kernel*, che è, almeno in generale, diverso

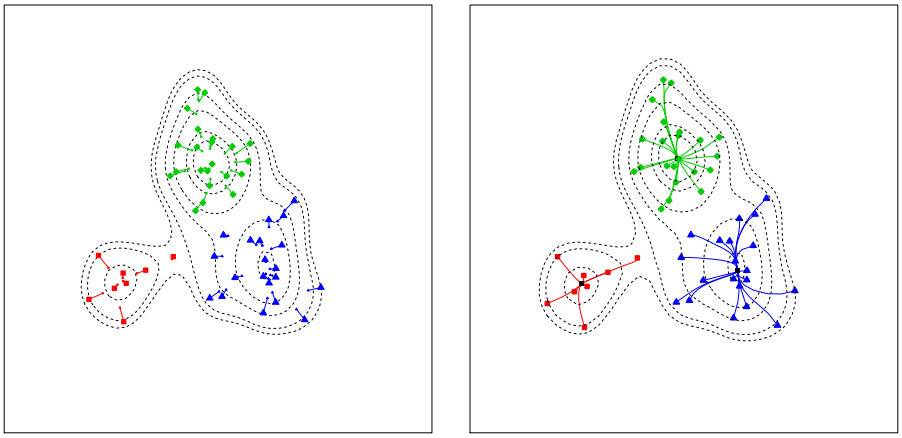

Figura 1.5: Un campione tratto dalla densità trimodale illustrata nella Figura 1.4, con sovraimposte le curve di livello della densità stimata; a sinistra, il *mean-shift* delle osservazioni alla prima iterazione e a sinistra il percorso completo compiuto per raggiungere la moda stimata di pertinenza.

dal *kernel* K usato per la stima $\hat{f}$ e nei pesi $w_i(x)$. L'uso di un gradiente normalizzato, pur non cambiando la direzione della risalita, si dimostra avere proprietà interessanti come una convergenza più rapida ai massimi locali di $\hat{f}$.

Si noti inoltre che l'algoritmo può essere applicato a qualsiasi punto del supporto, ed è quindi potenzialmente in grado di generare una partizione completa di quest'ultimo e non solo delle osservazioni nel dataset. La partizione delinea così i *cluster* come regioni dello spazio campionario che facciano riferimento alla stessa moda, matematicamente definite in precedenza come domini di attrazione della densità stimata.

Tra i molti contributi ai metodi di ricerca delle mode presenti in letteratura particolarmente interessante è il lavoro di Li *et al.*, 2007. Gli autori propongono uno speciale tipo di *mean shift*, chiamato *modal expectation maximization* (MEM), basato sull'analogia tra la densità kernel e i modelli a mistura. Viene inoltre approfondita l'idea di associare ad ogni *cluster* una distinta stima non parametrica della densità, che permette quindi di ricavare per ciascuna osservazione una probabilità a posteriori di appartenenza ai *cluster*. Si acquisiscono quindi alcuni strumenti interessanti legati alle tecniche di raggruppamento basate su modelli a mistura finita, come la possibilità di fare *soft clustering*, pur mantenendo le caratteristiche della stima ottenuta con metodi non parametrici.

1.5 Il problema della dimensionalità

L'utilizzo di metodi di *clustering* non parametrici va incontro a notevoli difficoltà quando l'obiettivo è la partizione di un dataset con un elevato numero di variabili. Queste difficoltà sono legate al fenomeno noto in letteratura come *maledizione della dimensionalità*.

La *maledizione della dimensionalità* si riferisce appunto alle problematiche che sorgono quando si cerca di estrarre informazioni da dati che risiedono in un supporto ad elevato numero di dimensioni. L'origine comune delle criticità è l'incremento del volume dello spazio matematico conseguente all'aumento della dimensionalità dello stesso; le osservazioni a disposizione diventano progressivamente più distanti tra loro, o sparse.

La sparsità è, anche intuitivamente, un problema per un qualsiasi processo stocastico di stima che sia affetto da variabilità, ma i metodi non parametrici, caratterizzati da una maggiore flessibilità, soffrono questa situazione in maniera ancora più accentuata.

Non fanno eccezione le tecniche di stima non parametrica della densità, come il metodo *kernel*; si può infatti intuire come la bandwidth debba crescere per poter includere un numero sufficiente di osservazioni in una stima $\hat{f}(x_0)$ al fine di contrastarne la variabilità. Ciò può però impedire di cogliere strutture locali nella densità, causando over-smoothing e aumentando la distorsione. Tale *bias-variance tradeoff* è tipicamente noto in questo ambito come *paradosso della vicinanza*: per cogliere strutture locali della densità sarebbe necessario considerare regioni che risulterebbero pressoché vuote, ma per ottenere regioni che non siano vuote non si prenderebbero in considerazione strutture locali.

Matematicamente il problema si può mostrare (Wand e Jones, 1994) considerando lo stimatore *kernel* 1.7 con la sua più semplice parametrizzazione $H = h^2 I$. Il minimo AMISE fissato K, al variare di h, è:

$$\inf_{h>0} AMISE\left\{\hat{f}\left(\cdot;h\right)\right\} =$$
$$= \frac{d+4}{4d}\left(\mu_2(K)^{2d}\{dR(K)\}^4\left[\int\{\nabla^2 f(x)\}^2\,dx\right]^d n^{-4}\right)^{\frac{1}{d+4}} \qquad (1.12)$$

Si può notare che il tasso di convergenza è dell'ordine di $n^{-\frac{4}{d+4}}$, diventa cioè più lento all'aumentare delle dimensioni, peggiorando radicalmente la bontà dello stimatore quando d è elevato. Ciò significa che

serviranno numerosità maggiori per ottenere stime della stessa qualità: per garantire una precisione costante in termini di MISE, la dimensione campionaria deve crescere esponenzialmente all'aumentare del numero di dimensioni. Solitamente lo stimatore basato sul metodo del nucleo viene ritenuto affidabile fino a circa sei dimensioni (Scott e Sain, 2005).

Le criticità della stima *kernel* si trasferiscono ai metodi di *clustering* che ne facciano uso. Il fenomeno della sparsità è collegato ad uno spostamento della massa di probabilità verso le code della distribuzione, dove, a causa dell'elevata variabilità, è possibile o addirittura comune la formazione di mode spurie. Per contro, il forzato aumento della bandwidth rischia di levigare troppo le caratteristiche locali della densità, cancellando possibilmente cluster di piccola dimensione. Nella pratica, nonostante sia chiaro che il *clustering* modale non sia lo strumento più adatto per lavorare con un grande numero di dimensioni, i metodi non parametrici si sono dimostrati in grado di cogliere strutture nella densità anche in spazi ad elevata dimensionalità. Alcuni risultati accettabili sono stati riportati nell'applicazione del *clustering* modale a spazi con decine o anche centinaia di dimensioni (si veda ad esempio Stuetzle e Nugent, 2010). È inoltre da considerare la possibilità, tipica in situazioni di elevata dimensionalità, che i dati giacciano in un sottospazio di dimensione inferiore di quello considerato; in questi casi metodi di riduzione del numero di variabili sono la strada da percorrere.

Capitolo 2

Inferenza robusta

2.1 Valori anomali e contaminazione

Sebbene ogni modello statistico, per sua natura, rappresenti una descrizione semplificata, e dunque incompleta, della realtà, è fondamentale che deviazioni dal modello assunto non siano eccessive, perché rischierebbero di compromettere l'esito delle procedure inferenziali. A dispetto di ciò, è frequente, qualunque sia lo scopo dell'analisi statistica, che i dati presentino una qualche forma di contaminazione, da intendersi come presenza di una quota valori anomali (*outlier*) rispetto al modello assunto come generatore dei dati. E'fondamentale notare come il concetto di *outlier* sia da definirsi solo nei confronti di un determinato modello, secondo il quale l'osservazione anomala è molto inverosimile o addirittura impossibile (si veda, ad esempio, Gather *et al.*, 2002).

Questa definizione può essere a sua volta specializzata in diverse categorie di punti anomali (Farcomeni e Greco, 2016), le più interessanti note come valori estremi, punti influenti e *bridge points*. I valori estremi sono *outlier* caratterizzati da valori anomali rispetto ad una o più dimensioni, ma che seguono comunque il trend generale del campione e non sono associati a evidenze contrastanti. È fondamentale la differenza tra questi e i punti influenti, che sono invece valori anomali che non seguono il trend generale del campione; sono quindi in grado di influenzare fortemente, se ignorati, i risultati delle analisi. Infine, i bridge points sono osservazioni appartenenti alla regione a bassa densità fra due regioni modali. Questi sono particolarmente importanti dal punto di vista del *clustering*, in quanto le regioni ad alta densità sono solitamente associate ai *cluster*, e i *bridge points* sono di dubbia attribuzione a ciascun gruppo.

Un'ulteriore possibile distinzione relativa al concetto di *outlier* è

quella tra *outlier* strutturali e *componentwise*. I primi sono osservazioni anomale provenienti da una distribuzione contaminante completamente separata da quella che genera i dati puliti, mentre i secondi appaiono quando le singole dimensioni delle osservazioni pulite possono essere corrotte separatamente, come spesso accade in caso di errori di misurazione. Nella pratica, la maggior parte delle procedure robuste tratta comunque le osservazioni anomale senza distinzione tra *outlier* strutturali o *componentwise*.

È infine necessaria una distinzione tra il concetto di valore anomalo e quello di valore corrotto. Il primo è definito dal punto di vista statistico: lo stato di *outlier* è assegnato ad un'osservazione, relativamente al modello, in base alla sua posizione nello spazio campionario. Il secondo concetto identifica invece uno stato, inosservabile, che esula da definizioni di tipo statistico, in quanto un osservazione è intuitivamente definita come corrotta quando proveniente da errori di vario tipo che ne modifichino il valore originale. Tale osservazione non necessariamente sarà associata ad un valore anomalo; potrà infatti essere un *inlier*, cioè appartenere a regioni dove risiedono la maggioranza dei dati.

In presenza di contaminazione, è naturale ipotizzare una situazione in cui convivono due meccanismi di generazione casuale diversi, uno responsabile della generazione della maggioranza dei dati, mentre l'altro causa della creazione di osservazioni anomale. Più formalmente, indicato con $\mathcal{X} = \{x_1, \ldots, x_n\}$ il campione osservato e con f il modello probabilistico assunto a generazione di $\mathcal{X}$, il contesto che caratterizza la contaminazione del campione può essere associato ad una mistura, dove la vera funzione di probabilità o densità ϕ sottostante i dati giace in realtà in un intorno di f:

$$\phi(x) = (1 - \lambda)f(x) + \lambda\delta(x)$$

con $\delta(\cdot)$ è una arbitraria funzione, e $0 < \lambda < 1$. Anche se può essere ragionevolmente possibile specificare il modello relativo alla maggioranza dei dati, non è altrettanto semplice, infatti, fare assunzioni riguardo la struttura del meccanismo di generazione degli *outlier*.

2.2 Procedure robuste

Ignorare la presenza di contaminazione dei dati può comportare un deterioramento della qualità dell'inferenza: la variabilità delle stime aumenta, si introduce instabilità a piccoli perturbamenti del campione e distorsione, i test perdono potenza.

Con statistica robusta ci si riferisce all'insieme di procedure statistiche che garantiscono la stabilità dell'inferenza e l'affidabilità dei risultati anche a fronte di allontanamenti del campione dalle assunzioni modellistiche considerate. Al contempo, nel caso le assunzioni modellistiche siano rispettate, una procedura robusta deve garantire che rimanga contenuta l'inevitabile perdita di efficienza rispetto procedure classiche.

L'idea chiave alla base delle procedure robuste è tipicamente quella di produrre una versione aggiustata delle procedure classiche, ignorando o riducendo il peso delle osservazioni anomale. A questo scopo, è opportuno che l'inferenza e l'individuazione degli *outlier* proceda congiuntamente: poiché una procedura non robusta è affetta dalla presenza di valori anomali, non può essere utilizzata per l'identificazione dei valori anomali, conducendo a fenomeni noti come *swamping* e *masking*. Nel primo caso, un'osservazione genuina è erroneamente etichettata come *outlier*; nel secondo caso un *outlier* è erroneamente etichettato come osservazione genuina.

In letteratura sono state proposte procedure robuste per far fronte a innumerevoli problemi statistici: dal semplice uso della mediana campionaria come stimatore di un parametro di posizione, al più sofisticato uso di stimatori M; dagli stimatori S nei modelli di regressione, a versioni robuste di tecniche statistiche multivariate quali l'analisi delle componenti principali. Fare una rassegna di tali metodi va al di là degli scopi del presente lavoro, e si rimanda, ad esempio, ai testi di Ronchetti e Huber (2009) per una trattazione approfondita delle procedure inferenziali e a Farcomeni e Greco (2016) per una rassegna sui metodi robusti di analisi statistica multivariata e riduzione della dimensionalità.

Nel seguito, si offrirà una panoramica dei metodi robusti proposti in letteratura nell'ambito della *cluster analysis*, in quanto oggetto del presente lavoro.

2.3 Clustering Robusto

Il concetto di valore anomalo, descritto e definito all'inizio del capitolo, va ulteriormente specializzato se l'ambito di applicazione è la *cluster analysis*. Tipicamente, quando i dati sono assunti come provenienti da un *set* di gruppi inosservati, gli *outlier* sono da considerarsi tali se sono eccezionali rispetto al *cluster* di appartenenza, e non più rispetto alla popolazione.

Si pensi, ad esempio, ad un *bridge point*, che potrebbe essere assolutamente verosimile dal punto di vista della densità generale, ma che è un punto appartenente alla coda, e possibilmente *outlier*, se considerato rispetto alla densità del *cluster* del quale fa parte. Al contrario, un punto appartenente ad un gruppo di piccola dimensione, anche se verosimile rispetto al suo cluster, potrebbe essere erroneamente considerato come outlier se valutato rispetto alla densità generale della popolazione.

Queste considerazioni portano alla necessità di sviluppare metodi robusti che procedano congiuntamente nel raggruppamento e nell'individuazione dei valori anomali, in quanto la selezione di questi dipende, come detto, dalla configurazione dei gruppi e, viceversa, la configurazione dei gruppi sarà influenzata dalle osservazioni anomale. Non solo sarà quindi necessario individuare correttamente gli *outlier* rispetto ai *cluster*, ma sarà necessario individuare i *cluster* in maniera robusta, dato che un'errata partizione dei dati porterebbe inevitabilmente ad un'errata identificazione dei valori anomali.

In particolare, si noti come la determinazione di un *cluster* aggiuntivo potrebbe rendere verosimili delle osservazioni che prima erano considerate come anomale, e ciò rende poco chiaro, in molti casi, quale sia la natura di un gruppo isolati di piccole dimensioni. Questo, come si vedrà nel capitolo successivo, sarà un problema ancora più grave per i metodi di *clustering* non parametrici, per i quali il numero di gruppi non è fissato a priori ma è un risultato automatico del processo di stima.

Una tecnica semplice ed intuitiva per l'irrobustimento di metodi statistici è nota come *trimming*, e prevede di scartare la proporzione (spesso fissata a priori) del campione corrispondente alle osservazioni che siano considerate più anomale. I metodi di *trimming* si concentreranno quindi sulla selezione e l'eliminazione dei valori più anomali, permettendo, plausibilmente, di cancellare in maggioranza valori corrotti, se questi sono presenti. Si noti però che l'ottimizzazione della probabilità di eliminare osservazioni contaminate non è il criterio alla base della maggioranza delle tecniche di *trimming* degli *outlier*. Una procedura che, invece, volesse massimizzare la suddetta probabilità dovrebbe, in primis, fare delle assunzioni sulla distribuzione delle osservazioni corrotte, rendendo più rigido il modello e più specifici i casi di applicazione. Inoltre, la possibilità di tralasciare valori contaminati che non siano effettivamente *outlier* non è particolarmente grave per quanto riguarda la corretta individuazione dei *cluster*, in quanto un *inlier*, anche corrotto, non può essere distinto a livello matematico da un dato pulito. Per coniugare la ricerca degli *outlier* a quella delle osservazioni corrotte, senza adottare

uno specifico modello per descrivere la generazione per queste ultime, è usuale fare delle assunzioni non restrittive che sostanzialmente assicurino che la contaminazione generi osservazioni distanti dalle masse dei dati puliti, cioè che generi effettivamente *outlier*.

La prima applicazione dell'approccio del *trimming* alla *cluster analysis* è stato il metodo delle *trimmed K-means* introdotto da Cuesta-Albertos *et al.*, 1997. Il metodo si basa sull'ottimizzazione del criterio

$$\underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \sum_{x_i \in S_k} \|x_i - \bar{x}_k\|^2 \tag{2.1}$$

dove K è il numero di *cluster*, $S = \{S_0, \ldots, S_K\}$ è una partizione del campione $\mathcal{X} = \{x_1, \ldots, x_n\}$ e $\bar{x}_k = \frac{1}{n} \sum_{x_i \in S_k} x_i$.

A differenza della versione classica del metodo descritta nel Capitolo 1, l'ottimizzazione ignora il primo elemento S_0 della partizione S, che sarà vincolato ad includere una fissata proporzione del campione e conterrà i valori censurati dal *trimming*; le osservazioni appartenenti a questo insieme potranno pertanto essere arbitrariamente distanti dalle altre senza condizionare la qualità dei risultati.

A livello operativo l'ottimizzazione avviene, come per il metodo non robusto, in maniera iterativa. In questo caso si alterneranno però, dopo un inizializzazione casuale, due tipologie di passi: il *concentration step*, ed il classico aggiornamento basato sui minimi quadrati. Il *concentration step* si occuperà del trimming vero e proprio, assegnando le $n\alpha$ unità più anomale al *set* S_0. L'assegnazione avverrà ordinando le osservazioni secondo il loro grado di *outlyingness* relativa al cluster di appartenenza, corrispondente alla distanza euclidea della data osservazione dal centro ad essa più vicino. A questo punto si procederà con il normale aggiornamento delle $K-$medie, partizionando la restante parte del campione ed aggiornando di conseguenza i centri. I due passi qui descritti saranno ripetuti fino a convergenza.

La maggior parte dei contributi per irrobustire tecniche di clustering siè tuttavia sviluppata nel contesto *model-based*, dove l'assunzione in merito ad una struttura distributiva di ciascun gruppo pone le basi naturali per l'applicazione di tecniche robuste di inferenza sviluppate, ad esempio, quando si assume che i dati abbiano distribuzione gaussiana. In García-Escudero *et al.* (2010), ad esempio, si considerano misture di popolazioni normali e, similmente a quanto visto per le *trimmed K-means*, la funzione di verosimiglianza da ottimizzare (1.2) viene sostituita da una versione nella quale è concesso che una frazione delle osservazioni

sia censurata:

$$\sum_{i=1}^{n} z(x_i) log \left(\sum_{k=1}^{K} \pi_k \phi_d \left(x_i; \mu_k, \Sigma_k \right) \right) \tag{2.2}$$

dove $z(x_i)$ è una funzione indicatrice che restituisce lo stato di ogni osservazione x_i come facente parte del campione non censurato ($z(x_i) = 1$) o meno ($z(x_i) = 0$). La stima sarà sottoposta al vincolo $\sum_{i=1}^{n} z(x_i) = n(1 - \alpha)$, cioè sarà una fissata frazione α delle osservazioni ad essere censurata. La *trimmed maximum likelihood* è stata considerata anche da Neykov *et al.* (2007) e Gallegos e Ritter, 2009. La massimizzazione della (2.2) rappresenta un problema illimitato se non si impongono dei vincoli alle matrici Σ_k, in quanto la presenza di alcune osservazioni quasi collineari porterebbe alla formazione di gruppi spuri con distribuzioni praticamente degeneri. In García-Escudero *et al.* (2010) si sono considerati dei vincoli, mirati all'evitare soluzioni degeneri e spurie, che impongono un limite massimo al rapporto degli autovalori delle matrici Σ_k. L'utilizzo combinato del *trimming* e dei vincoli fornisce sicuramente robustezza e stabilità al *clustering* basato su mistura.

Si noti che, data una stima del modello, le osservazioni che subiranno il *trimming* sono intrinsecamente quelle legate ad una densità più bassa. Questo è coerente con la generica definizione di *outlier*, ma meno con la specificazione degli stessi come valori relativamente anomali rispetto al gruppo di appartenenza. Ciò, in casi estremi, tipicamente collegati ad una sovrastima della quantità di *outlier*, potrebbe portare l'algoritmo a censurare completamente gruppi di piccola dimensione, in quanto la loro densità sarebbe troppo bassa rispetto a quella della popolazione generale. Diventa inoltre più complicata, almeno a livello teorico, l'identificazione dei *bridge points*, in quanto questi gioveranno dell'incremento alla loro densità portato dalle code della distribuzione di più gruppi.

L'algoritmo per l'ottimizzazione della 2.2 procede, dopo un'inizializzazione casuale, alternando passi di *concentration*, dove si rilevano e censurano le osservazioni più anomale, a passi di ottimizzazione vincolata basati sull'EM. In particolare, la configurazione di partenza prevede il campionamento casuale di un sottoinsieme di $K(d + 1)$ osservazioni dal campione originale, attraverso le quali calcolare un valore iniziale per medie μ_k e covarianze Σ_k (sottoposte ai vincoli) di ogni gruppo. I valori di π_k saranno inizializzati casualmente. A questo punto, si alterneranno i passi sopra citati; quello di *concentration* ordinerà le osservazioni x_i secondo l'ultimo aggiornamento della stima della densità, e scarterà

quelle associate alle stime $\sum_{k=1}^{K} \hat{\pi}_k \phi_d \left(x_i; \hat{\mu}_k, \hat{\Sigma}_k \right)$ più basse. Si procede poi con il passo EM, applicato considerando solo le osservazioni non censurate. Eventuali vincoli sulle matrici Σ_k saranno considerati in questo passo. Il valore ottimo locale al quale l'algoritmo converge sarà dipendente dal dato punto di partenza; la procedura sarà quindi ripetuta più volte, variando il campione di inizializzazione, in modo da cercare il massimo globale.

Una più famosa alternativa, nota come *tclust*(García-Escudero *et al.*, 2008), pur mantenendo la stessa struttura dell'algoritmo appena descritto, sostituisce l'ottimizzazione tramite EM con quella tramite Classification EM (Celeux e Govaert, 1992). La differenza principale tra i due metodi sarà quindi la funzione obiettivo, che diventa:

$$\sum_{k=1}^{K} \sum_{x_i \in S_k} log \left(\pi_k \phi_d \left(x_i; \mu_k, \Sigma_k \right) \right) \tag{2.3}$$

e il criterio utilizzato per ordinare le osservazioni nel passo di *concentration*; per ogni x_i, non si userà più la densità stimata $\hat{f}(x_i)$, ma solo il contributo del cluster $\tilde{k}$ alla quale x_i è assegnata $\hat{\pi}_{\tilde{k}} \phi_d \left(x_i; \hat{\mu}_{\tilde{k}}, \hat{\Sigma}_{\tilde{k}} \right)$. Si noti che nella (2.3), la partizione del campione sarà $S = \{S_0, \ldots, S_K\}$ e S_0 conterrà gli $n\alpha$ punti censurati dalla procedura.

Capitolo 3

Clustering non parametrico robusto

3.1 Introduzione al problema

La caratterizzazione della struttura probabilistica di una popolazione tramite la specificazione di un modello completamente non parametrico ha, tra le sue proprietà, quella di evitare l'imposizione di vincoli di forma alla funzione di densità, lasciando che siano solamente i dati a definirne l'aspetto tramite il processo di stima. Questo guadagno di flessibilità rende però meno intuitivo ed immediato lo studio, a partire dalla loro definizione, dei valori anomali, quali osservazioni appartenenti, secondo un dato modello, a regioni dello spazio campionario a bassa probabilità.

La rigidità di un modello parametrico permette infatti, a fronte di una determinata proporzione di osservazioni non corrotte, e grazie all'uso di stimatori robusti dei parametri, di ottenere in maniera ragionevolmente corretta la stima della densità su tutto lo spazio parametrico, anche in presenza di contaminazione. La naturale conseguenza è allora l'individuazione di *outlier* come punti residenti in aree associate ad una densità particolarmente bassa, proprio perché questi non ne influenzano la stima in maniera notevole.

Il discorso non è invece così immediato in campo non parametrico; uno stimatore non parametrico della densità (come lo stimatore *kernel*) è, per sua natura, locale, in quanto la stima è influenzata dalle osservazioni limitatamente ad un loro intorno. Concettualmente questo è dovuto all'assenza della rigidità data dalla forma funzionale del modello; sono i dati stessi a definire, in maniera completamente libera ed appunto locale, dove ricade la massa della funzione di densità. Si cade

quindi in una sorta di circolo vizioso, in quanto si vuole cercare osservazioni associate ad una bassa densità, ma sono le osservazioni stesse, possibilmente anomale, a spostare la densità in ogni dato punto dello spazio campionario; data questa mancanza di robustezza locale tipica degli strumenti non parametrici, sarà difficile definire un'osservazione come inverosimile secondo il dato modello. Si può intuire come, in questo *framework*, il concetto di osservazione anomala diventi strettamente legato a quello di sparsità, in quanto un *outlier* può essere definito come tale solamente quando è relativamente distante dalle altre osservazioni, o analogamente, appartenente ad una regione soggetta a sparsità.

Come discusso nel Capitolo 1, i metodi di stima non parametrica della densità, e i metodi di raggruppamento che ne fanno uso, soffrono della problematica denominata *maledizione della dimensionalità*. La sparsità dei dati conseguente all'aumento del numero di dimensioni diminuisce la qualità delle stime, portando alla possibile creazione di mode spurie. D'altra parte, la necessità di sovralisciare le stime per contrastare tale problema comporta l'aumento del *bias*. A causa del legame tra *outlier* e sparsità dei dati, pertanto, i medesimi problemi emergono in presenza valori anomali, che aggravano i problemi legati alla dimensionalità; viceversa, i problemi legati ai valori anomali sono più gravi al crescere delle dimensioni dello spazio campionario.

Nell'ambito del *clustering* il problema è, concettualmente e praticamente, ancora più complesso, in quanto, come esposto nel capitolo precedente, un valore è definito come anomalo solamente rispetto al gruppo al quale appartiene e non nei confronti del totale delle unità. In questo ambito la ricerca degli *outlier* è quindi un processo concomitante con la partizione delle osservazioni, e dipendente da esso dal punto di vista della qualità dei risultati. In presenza di contaminazione, però, l'instabilità dello stimatore non parametrico della densità, dovuta alla sparsità che contraddistingue gli *outlier*, rende difficile ottenere una suddivisione in gruppi robusta. È qui importante evidenziare una differenza tra i modelli di *clustering* parametrici ed il *clustering* modale; nei primi il numero di gruppi, essendo un'assunzione legata alla specificazione del modello, è fissato a priori. Il processo di stima che ottimizza la funzione di verosimiglianza è subordinato a questa scelta, che sarà poi, solo in seguito, guidata da vari strumenti come la valutazione di criteri automatici. Per ogni data stima del modello il numero di *cluster* è quindi fissato ed assunto come corretto, rendendo perciò più facile identificare gli *outlier* rispetto ai gruppi. Nel *clustering* modale, invece, il numero di gruppi è una quantità automaticamente ottenuta attraverso la sti-

ma non parametrica della funzione di densità, in quanto corrisponde al numero di mode della suddetta funzione; l'identificazione dei valori anomali rispetto ai *cluster* non può più quindi fare affidamento sulla fissata numerosità dei gruppi, perché la reale esistenza di ogni moda stimata può essere messa in discussione. L'apparizione di mode spurie nelle regioni sparse connesse alla contaminazione del campione diventa quindi deleteria per l'identificazione degli *outlier*. Ogni *cluster* spurio renderà verosimili, quando considerate rispetto allo stesso, le osservazioni anomale al suo interno, impedendone la corretta identificazione, che dovrebbe avvenire rispetto ai gruppi veramente presenti nella popolazione. Si veda, per un'illustrazione, la Figura 3.1.

Il problema della robustezza è stato largamente trascurato nella letteratura relativa al clustering modale. I limitati contributi in questa direzione si riferiscono prevalentemente all'uso di uno stimatore della densità basato sui vicini più prossimi e sfruttano la distanza tra le osservazioni. Si vedano, ad esempio, Yang *et al.* (2021). Alcuni contributi estendono l'idea del *mean-shift* a misure di posizione più robuste dando luogo al *median-shift* o *medoid-shift* (Shapira *et al.*, 2009; Sheikh *et al.*, 2007).

Sebbene i problemi sopra descritti affliggono entrambe le metodologie di ricerca dei *cluster* non parametriche descritte nel Capitolo 1, a livello pratico, il metodo basato sulle curve di livello risulta operativamente più robusto. Gli insiemi di livello connessi sono infatti ricercati per un *set* discreto di valori di λ, e non in maniera continua. Questo porterà a cogliere con certezza le mode più forti e, spesso, a trascurare quelle deboli, tipicamente generate dagli *outlier*. L'albero dei *cluster* fornisce inoltre uno strumento grafico per valutare, ed eventualmente aggregare, i gruppi associati alle mode mode più deboli.

Per questa ragione, nel seguito di questo capitolo, si esplorerà e proporrà un metodo per rilevare gli *outlier* e robustificare il *clustering* non parametrico, seguendo l'approccio basato sulla ricerca diretta delle mode tramite lo studio del gradiente della stima della densità. In questo caso infatti ogni moda, anche se generata ipoteticamente da una sola osservazione, sarà sicuramente associata ad un *cluster*, rendendo quindi più complicato lo studio della struttura dei veri gruppi associati alla popolazione di riferimento, nonché la stima del numero degli stessi.

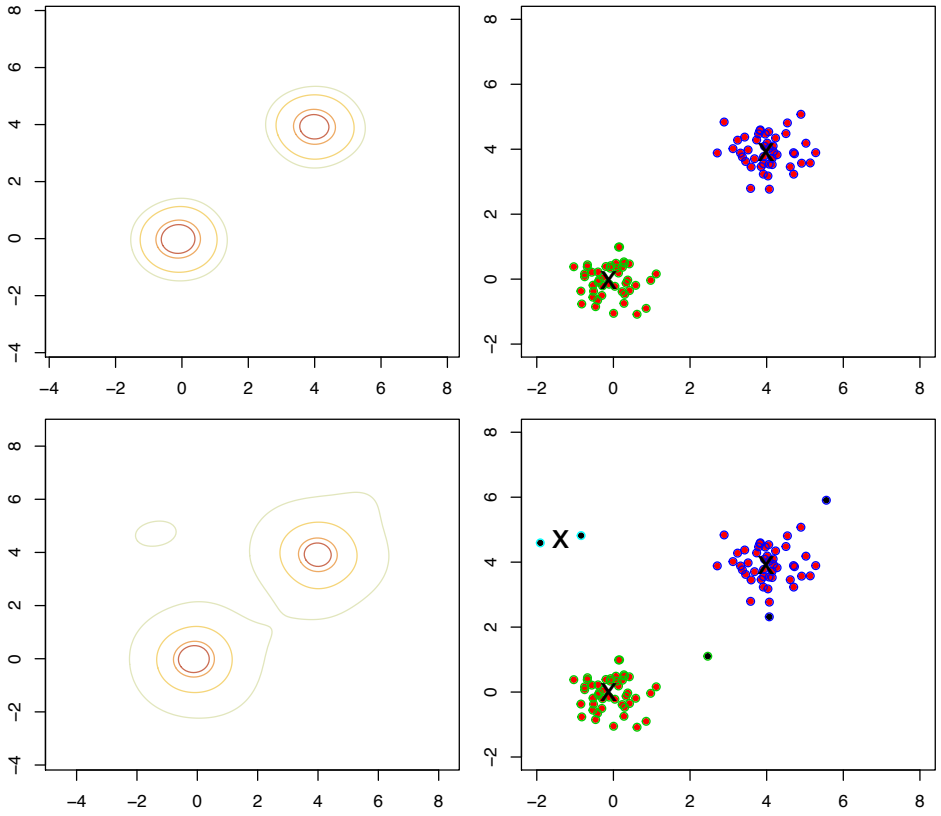

Figura 3.1: Esempio bidimensionale di campione generato da un modello centrale a due gruppi: in alto, stima della densità e *clustering*, ottenuto mediante ricerca delle mode, (indicate con una `x`). In basso, effetto sulla stima della densità e sul *clustering* della contaminazione del campione con una piccola quota di *outlier*.

3.2 Un metodo di individuazione degli *outlier* nel clustering non parametrico

3.2.1 Il contesto teorico

L'idea chiave per l'identificazione dei valori anomali in un contesto di clustering modale, è quella di associare, come suggerito dalla definizione, e coerentemente con la *ratio* dei metodi robusti descritti nel capitolo precedente, il concetto di *outlier* con quello di punto a bassa densità. Tuttavia, poiché nella *cluster analysis*, il grado di *outlyingness* fa riferimento al gruppo di appartenenza, e non alla popolazione generale, si pone il problema di definire cosa identifica in questo contesto la densità di un singolo *cluster*.

Seguendo la notazione indicata nel Capitolo 1, sia $f : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$,

una funzione di densità caratterizzata da K mode, e siano $\{\mathcal{D}_k\}_{k\in\{1,\dots,K\}}$ i loro domini di attrazione $\cup_{k=1}^{K}\mathcal{D}_k = \mathcal{D}$. Questi definiscono a tutti gli effetti, nel *clustering* modale, i *cluster* teorici associati alla data densità, e sono effettivamente gli obiettivi asintotici della stima associata a metodi come il *mean shift*.

Appare naturale definire la funzione di densità densità $g_k(\cdot)$ del singolo gruppo k come segue:

$$g_k(x) := \begin{cases} f(x)/\pi_k & \text{se } \; x \in \mathcal{D}_k \\ 0 & \text{se } \; x \in \mathcal{D} \setminus \mathcal{D}_k, \end{cases} \tag{3.1}$$

dove $1/\pi_k$ è una costante di normalizzazione definita da

$$\pi_k = \int_{\mathcal{D}_k} f(x)\, dx. \tag{3.2}$$

Un valore anomalo sarà allora definito tale se sarà caratterizzato da una bassa densità g_k relativamente al gruppo k di pertinenza.

Rispetto ai metodi parametrici, nel contesto non parametrico, dove l'identificazione dei gruppi non è guidata dall'ottimizzazione di una funzione obiettivo quale la verosimiglianza, la scelta di considerare una densità normalizzata, in luogo della densità globale, comporta il vantaggio principale di evitare che gruppi di dimensione contenuta vengano eccessivamente valutati come anomali, in quanto associati ad una densità globale mediamente più bassa rispetto a gruppi di dimensione maggiore.

Seppur completamente fedele alla definizione di *cluster* come dominio di attrazione di una moda, questa specificazione delle funzioni $g_i(\cdot)$ descrive delle densità che difficilmente risultano realistiche o naturali dal punto di vista pratico. È infatti difficile pensare che il reale meccanismo che ha generato i dati non preveda regioni di intersezione tra le osservazioni di gruppi diversi, cioè che le code relative al punto di incontro tra due gruppi siano troncate. Ciò porterà, nella fase di definizione degli stimatori, ad una più naturale implementazione delle densità di *cluster*, basata sul concetto di mistura e sulle proprietà dello stimatore *kernel*.

3.2.2 La controparte empirica

Poiché le quantità $f, \mathcal{D}_k, \pi_k, K$ coinvolte nella definizione della (3.1) non sono note, è necessario procedere ad una loro stima.

Dato il campione $\mathcal{X} = \{x_1, \dots, x_n\}$, è agevole ottenere una stima $\hat{f}$ di f mediante l'impiego dello stimatore kernel (1.7). L'applicazione di un

metodo di risalita del gradiente quale il *mean-shift* consente la successiva stima delle mode e delle regioni modali di $\hat{f}$, nonché la partizione delle osservazioni campionarie nei cluster $S = \{S_1, \ldots, S_{\hat{K}}\}$.

A partire dalla (3.2), è possibile notare che

$$\begin{aligned} \pi_k &= \int_{\mathcal{D}_k} f(x)\, dx \\ &= \int_{\mathcal{D}} \mathbf{1}_{\mathcal{D}_i}(x) f(x)\, dx \\ &= \mathbb{E}\left[\mathbf{1}_{\mathcal{D}_i}(x)\right], \end{aligned} \tag{3.3}$$

dove $\mathbf{1}_A(\cdot)$ è la funzione indicatrice dell'insieme A.

La naturale controparte empirica delle π_k è dunque rappresentata dalle medie campionarie

$$\hat{\pi}_k = \frac{\sum_{i=1}^{n} \mathbf{1}_{\hat{\mathcal{D}}_k}(x_i)}{n} = \frac{\sum_{x \in S_k} 1}{n} = \frac{n_k}{n} \tag{3.4}$$

ossia, tramite la frazione delle osservazioni del campione assegnate ad ogni *cluster*.

Nel passaggio precedente si stanno usando i domini di attrazione $\hat{\mathcal{D}}_k$ basati su $\hat{f}$ in sostituzione dei veri domini di attrazione delle mode di f. Si noti che non necessariamente questi saranno in numero uguale, in quanto $\hat{K}$ è solo una stima di K. Si può tuttavia mostrare che (Chacón, 2015, Teorema 4.1), se le derivate prima e seconda dello stimatore di densità convergono quasi certamente alle derivate prima e seconda della vera f, la stima del numero di regioni modali è consistente. Poiché, sotto opportune condizioni di regolarità di f, e quando il parametro di lisciamento è selezionato in maniera ottimale, lo stimatore *kernel* e le sue derivate sono consistenti, abbiamo garanzia della fondatezza teorica del modo di procedere.

Una stima delle densità relative ai *cluster*, per ogni osservazione, sarà quindi ottenibile tramite *plug-in*:

$$\hat{g}_k(x) := \begin{cases} \hat{f}(x)/\hat{\pi}_k & \text{se } \; x \in S_k \\ 0 & \text{se } \; x \in S \setminus S_k \end{cases} \tag{3.5}$$

Come anticipato nel paragrafo 3.2.1, questa rappresentazione delle densità relative ai gruppi, pur essendo la naturale derivazione della definizione di gruppo nell'ambito del *clustering* modale, è poco realistica e naturale a livello pratico. È quindi possibile definire un'alternativa, basata su un'approssimazione di f e sulle proprietà dello stimatore *kernel*.

Sebbene l'obiettivo sia quello di identificare f e i suoi domini di attrazione, mediante uno stimatore *kernel* di parametro H sarà al più possibile fare inferenza sulla versione convoluta $f_H(\cdot)$, che ne rappresenta una approssimazione:

$$\begin{aligned} f_H(\mathbf{x}) &= \int_{\mathcal{D}} K_H(x-u) f(u)\, du \\ &= \mathbb{E}(K_H(x-X)). \end{aligned} \tag{3.6}$$

Fissata la partizione indotta dai domini di attrazione delle sue mode, f_H può essere equivalentemente rappresentata come segue

$$\begin{aligned} f_H(x) &= \int_{\mathcal{D}} K_H(x-u) f(u)\, du \\ &= \sum_{k=1}^{K} \int_{\mathcal{D}_k} K_H(x-u) f(u)\, du \\ &= \sum_{k=1}^{M} \pi_k \int_{\mathcal{D}_k} K_H(x-u) \frac{1}{\pi_k} f(u)\, du \\ &= \sum_{k=1}^{M} \pi_k g_{kH}(x) \end{aligned} \tag{3.7}$$

dove g_{kH} rappresenta la versione convoluta della funzione di densità del gruppo k. Infatti:

$$g_{kH}(x) = \int_{\mathcal{D}} K_H(x-u) g_k(u)\, du = \mathbb{E}(K_H(x-\tilde{X}_k)$$

e $\tilde{X}_k$ è una variabile generata *cluster* k, e descritta dalla densità g_k definita in (3.1).

Volendo approssimare g_{kH} attraverso i dati, è naturale definire la

controparte empirica

$$\begin{aligned}\hat{f}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_H(x - x_i) \\ &= \frac{1}{n}\sum_{k=1}^{K}\sum_{x_i \in S_k} K_H(x - x_i) \\ &= \sum_{k=1}^{M}\frac{n_k}{n}\sum_{x_i \in S_k}\frac{1}{n_k} K_H(x - x_i) \\ &= \sum_{k=1}^{M}\hat{\pi}_k \tilde{g}_k(x)\end{aligned} \tag{3.8}$$

dove

$$\tilde{g}_k(x) = \sum_{x_i \in S_k}\frac{1}{n_k} K_H(x - x_i) \tag{3.9}$$

è la corrispondente campionaria della densità convoluta di gruppo g_{kH}.

Secondo questa rappresentazione, la stima *kernel* della densità può essere descritta secondo una mistura di stime *kernel* delle densità unimodali, ottenute con le sole osservazioni appartenenti a ciascun gruppo. È quindi ragionevole, come alternativa alle $\hat{g}_k(x)$, l'utilizzo di $\tilde{g}_k(x)$ per stimare g_k.

Questo approccio, che associa ad ogni *cluster* una stima di densità non parametrica ottenuta sui soli punti ad esso appartententi, è utilizzato, ad esempio, in Azzalini e Torelli (2007) e Li *et al.* (2007). In particolare, in quest'ultimo lavoro, è approfondito l'aspetto legato al decadimento troppo veloce delle code associato ad una stima della densità basata sui soli punti associati al dato gruppo. Il problema, visibile soprattutto quando i *cluster* non sono ben separati, potrebbe essere risolto adottando la filosofia del *soft clustering*, senza associare ogni osservazione ad un solo gruppo, e permettendo che questa contribuisca, in maniera pesata, alle stime delle densità di più *cluster*.

3.2.3 L'algoritmo

Avendo definito le quantità utili a valutare il grado di *outlyingness* delle osservazioni, e le stime di queste, l'identificazione dei valori anomali verrà eseguita, sulla base di tali quantità, congiuntamente ai processi di stima e *clustering*.

La procedura che si propone, di natura iterativa, è descritta nell'Algoritmo 1 ed è volta all'identificazione ad una quota di *outlier* pari ad α. Al generico passo r, l'algoritmo stima la densità $\hat{f}^{(r)}$ sulla base di un sottocampione $\mathcal{X}^{(r)}$ selezionato da $\mathcal{X}$, e mediante l'applicazione del *mean-shift* risale il gradiente di $\hat{f}^{(r)}$ per identificare le mode e l'insieme $S = \{S_1, \ldots, S_{K_r}\}$ dei gruppi di pertinenza di ciascuna osservazione campionaria. La partizione S consente di determinare, per ogni x_i la densità del gruppo di riferimento $\hat{g}_{k_i}^{(r)}(x_i)$ mediante la (3.5) (o $\tilde{g}_{k_i}^{(r)}(x_i)$ mediante la 3.9). Le $n(1-\alpha)$ osservazioni x_i cui è associata la densità $\hat{g}_{k_i}^{(r)}(x_i)$ (o $\tilde{g}_{k_i}^{(r)}(x_i)$) più alta entrano a questo punto a far parte del sottocampione $\mathcal{X}^{(r+1)}$ utilizzato nell'iterazione successiva.

Il procedimento viene inizializzato mediante un sottocampione $\mathcal{X}^{(1)}$ selezionato casualmente, e replicato fino a quando $\mathcal{X}^{(r+1)} = \mathcal{X}^{(r)}$ o al raggiungimento di un fissato numero massimo di iterazioni.

In linea di principio, la stima della densità ottenuta alla fine del processo iterativo è robusta in quanto basata su un campione da cui sono state troncate le $n\alpha$ osservazioni candidate ad essere anomale. Tuttavia, qualora il sottocampione $\mathcal{X}^{(1)}$, selezionato casualmente, dovesse dar luogo ad una stima di densità caratterizzata da una o più mode spurie, il ricorso alle densità normalizzate rischierebbe di convalidare la presenza di tali mode, che potrebbero quindi permanere anche a conclusione delle iterazioni poiché, a fronte di una densità $\hat{f}^{(r)}$ contenuta, il rapporto $\hat{g}_k^{(r)} = \hat{f}^{(r)}/\hat{\pi}_k$ risulterebbe elevato.

Al fine di svincolare la procedura dalla dipendenza del campione iniziale, conferendole maggior robustezza, si propone di adottare due accorgimenti: in primo luogo il sottocampione $\mathcal{X}^{(1)}$ sarà selezionato in modo da includere una quota $1-\alpha_0$ relativamente contenuta di osservazioni. In secondo luogo, l'Algoritmo 1 sopra descritto sarà esso stesso replicato B volte, dando luogo, per ogni osservazione x_i alle stime $\hat{g}_{k_i}^{(b)}(x_i)$, $b \in \{1, \ldots, B\}$. Al termine delle B replicazioni, si assegnerà a ciascuna osservazione un valore di *outlyingness* pari a

$$\hat{g}(x_i) = mediana(\hat{g}_{k_i}^{(1)}(x_i), \ldots, \hat{g}_{k_i}^{(B)}(x_i))$$

(o l'analoga controparte $\tilde{g}$), che sfrutta la robustezza intrinseca legata all'uso della mediana e si effettuerà il *trimming* del campione scartando le $n\alpha$ osservazioni anomale, associate ad un più basso valore di $\hat{g}(x_i)$.

La logica sottostante tale procedimento, schematizzato nell'Algoritmo 2, è la seguente. Se x_i è un punto appartenente ad uno dei *cluster* nominali, questo sarà associato a dei valori di $g_{k_i}^{(b)}(x_i)$ abbastanza sta-

Algoritmo 1 Algoritmo interno

Input: $\mathcal{X}$, H, α_0, α, R_{max}
$\mathcal{X}^{(0)} \leftarrow \mathcal{X}$
$\mathcal{X}^{(1)} \leftarrow$ *sottocampione casuale di* $\mathcal{X}$ *di dimensione* $n(1-\alpha_0)$
$r \leftarrow 1$
while $\mathcal{X}^{(r)} \neq \mathcal{X}^{(r-1)}$ **and** $r <= R_{max}$ **do**
  determina $\hat{f}^{(r)}(x;H)$ *basata su* $\mathcal{X}^{(r)}$
  applica mean-shift a $\hat{f}^{(r)}(x;H)$ *e determina cluster* $S_{k_i}^{(r)}$ di $x_i \, \forall x_i$
  determina $\hat{g}_{k_i}^{(r)}(x_i) \, \forall x_i$
  $r \leftarrow r+1$
  $\mathcal{X}^{(r)} \leftarrow$ *sottocampione di* $\mathcal{X}$ *formato dalle* $n(1-\alpha)\, x_i$ *con* $\hat{g}_{k_i}^{(r)}(x_i)$ più alta
end while
return $\hat{g} = \{\hat{g}_{k_1}^{(r)}, \ldots, \hat{g}_{k_n}^{(r)}\}$

Algoritmo 2 Algoritmo esterno

for $b \leftarrow 1$ to B **do**
  $\hat{g}^{(b)} \leftarrow$ *Algoritmo Interno* ($\mathcal{X}$, H, α_0, α, R_{max})
end for
for $i \leftarrow 1$ to n **do**
  $\hat{g}(x_i) \leftarrow mediana(\hat{g}_{k_i}^{(1)}(x_i), \ldots, \hat{g}_{k_i}^{(B)}(x_i))$
end for
return G

bili, in quanto la moda relativa al vero gruppo, forte di una numerosità associata che rende molto probabile che almeno alcuni dei suoi punti facciano parte di $\mathcal{X}^{(1)}$, rimarrà presente nella maggior parte delle replicazioni. Per contro se x_i è un *outlier*, quindi appartenente ad una zona soggetta a sparsità, ma comunque legato ad una moda spuria, sarà associato a dei valori di $\hat{g}_{k_i}^{(b)}(x_i)$ alti solamente quando la moda spuria verrà inserita nel campione iniziale $\mathcal{X}^{(1)}$, evento tipicamente poco probabile, dato un valore di α_0 sufficientemente alto, se la moda è composta da pochi punti o troppo vicina ad una moda più forte.

Il *clustering* robusto si otterrà ora applicando il *mean-shift* sulla base della densità stimata solamente con i punti che non sono stati identificati come anomali.

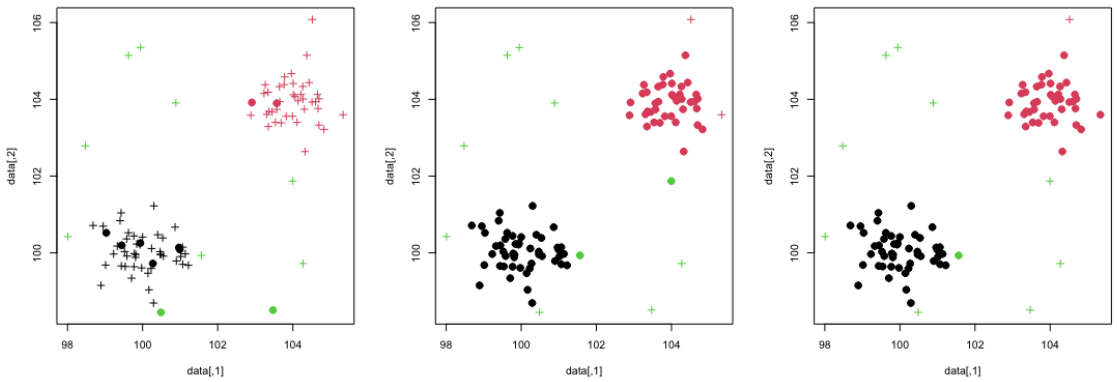

Figura 3.2: Esempio di *run* dell'algoritmo interno; a sinistra il campione iniziale rispetto al campione totale; al centro il campione selezionato dopo un'iterazione; a destra il campione ottenuto a convergenza. Le osservazioni scartate dal processo sono indicate con '+'.

3.2.4 Discussione

In questa sezione si discutono alcune questioni, concettuali o operative, collegate all'applicazione della procedura ora proposta.

- *Scelta di* α *e* α_0
  Sebbene una determinazione automatica della quota α di osservazioni anomale sarebbe indubbiamente desiderabile, sono svariate in letteratura le procedure che, similmente a quella proposta, ne richiedono una specificazione a priori (si vedano, ad esempio, García-Escudero *et al.*, 2008; Cuesta-Albertos *et al.*, 1997). In realtà, come si vedrà Capitolo 4, il metodo si dimostra solitamente efficace nella pratica anche in caso di una moderata sovrastima della frazione di *outlier*, in quanto, al più, cancellerà erroneamente punti appartenenti alle code dei cluster, senza intaccare i risultati del clustering in maniera significativa. Ciò è una diretta conseguenza della scelta di normalizzare le densità di cluster, rendendo più raro che un gruppo di piccola dimensione venga completamente eliminato.

  La metodologia proposta richiede anche la specificazione di un parametro di iniziale *trimming* α_0, che rappresenta la quota di osservazioni da scartare per la selezione del campione iniziale $\mathcal{X}^{(1)}$. Tale parametro dovrà assumere valori sufficientementi elevati in modo da favorire la selezione di un sottocampione iniziale che contenga un numero ristretto, o possibilmente nullo, di *outlier*. Per contro,

è auspicabile che $\mathcal{X}_1$. sia composto da punti provenienti da ciascun vero *cluster* presente nella popolazione. È per questo motivo che α_0 dovrebbe crescere al crescere di n, magari mantentendo pressochè costante il numero assoluto di osservazioni in $\mathcal{X}_1$, ma dovrebbe assumere valori più bassi, a parità di n, se si sospetta che la configurazione sotto studio sia caratterizzata da un numero di gruppi latenti elevato.

- *Scelta di H*
  Un aspetto finora non discusso, ma particolarmente critico, riguarda la scelta del parametro di lisciamento H per la stima della densità. Infatti, poiché H governa il grado di lisciamento della densità, e dunque il numero di mode, la sua selezione è, in generale, un aspetto fondamentale per la qualità dei risultati del *clustering* modale, a prescindere dalla presenza, o meno, di valori anomali. È per questo motivo che in questo lavoro, che si focalizza sulla questione della robustezza, non si darà particolare spazio allo studio della selezione del parametro di lisciatura.

  Nelle analisi empiriche, si è fatto ricorso ad una struttura diagonale della matrice H e a criteri di ottimalità asintotica del MISE sotto ipotesi di normalità dei dati, come suggerito ad esempio da Azzalini e Torelli (2007)

$$H = \frac{4}{n(d+2r+2)}^{\frac{2}{d+2r+4}} S \tag{3.10}$$

  dove d è il numero di dimensioni dello spazio, r è il grado della derivata di cui si vuole ottimizzare la stima ($r = 0$ per la stima ottimale della densità, $r = 1$ per la stima ottimale del gradiente, utile per il *mean-shift*), n è la numerosità del campione usato e S è una matrice diagonale contenente le stime della varianza per ogni dimensione.

  Pur essendo ovvio che in una popolazione avente più gruppi distinti l'ipotesi di normalità su cui si basa l'ottimalità dello stimatore non sarà rispettata, negli scenari presi in esame ciò non porterà a particolari problemi. La stima delle varianze in S avviene inoltre sul totale dei dati, portando ad una *bandwidth* che sicuramente causerà *oversmoothing* se il *focus* sono le singole densità di *cluster*. Inoltre, la presenza di valori anomali potrebbe ulteriormente inflazionare le stime delle varianze. Nella pratica, l'*oversmoothing* risultante, che migliora la stabilità dello stimatore *kernel*, non sembra deleterio per

la qualità dei risultati del *trimming* se i gruppi sono ben separati, ma sarà possibilmente problematico se i gruppi sono più vicini. Si noti, inoltre, che gli scenari studiati nel capitolo successivo disporranno i gruppi in diagonale (la distanza tra i centri sarà la stessa in ogni dimensione); tale disposizione è necessaria per non portare ad una stima di H che abbia evidenti differenze di scala tra le dimensioni, aspetto che sarebbe deleterio per dei gruppi pressochè sferici. Per contro, l'uso di una matrice diagonale stimata sul totale dei dati, unita alla disposizione diagonale dei gruppi e alla tendenza all'*oversmoothing*, porta ad una stima di H che favorisce *cluster* sferici, e porta a problemi, ad esempio, in caso di *cluster* ellittici.

Per mitigare questo problema si è sviluppata una versione del metodo che, per la stima *kernel* delle densità di *cluster*, utilizza delle *bandwidth*, H_k, stimate separatamente per *cluster* (si utilizzerà lo stesso stimatore descritto sopra, ma sui soli punti assegnati al dato gruppo). Si noti che si sarebbe potuta utilizzare una struttura non diagonale per le matrici H_k, sostituendo alla matrice diagonale S l'usuale stimatore per la varianza in campo multivariato. Sarebbe inoltre probabilmente utile vincolare la forma delle stesse, in maniera simile a quella adottata nell'algoritmo *tclust*.

È quindi chiaro come l'aspetto della selezione della *bandwidth* sia di imprescindibile importanza per la qualità dei risultati della procedura, e debba essere necessariamente approfondito per ottenere risultati che siano stabili nella varietà di situazioni legate a dati reali.

- Un terzo aspetto interessante riguarda una questione che parte da una considerazione puramente concettuale; secondo la struttura delle stime *kernel*, ogni osservazione contribuisce ad incrementare la densità nei suoi dintorni. Questo significa, nell'ottica del *trimming*, che un'osservazione può covalidare se stessa come *inlier*. Si potrebbe pensare invece, in un'ottica simile a quella della *leave-one-out cross-validation*, che ogni osservazione vada valutata solamente sui contribuiti alla densità portati dagli altri punti, escludendo dal campione l'osservazione stessa. Quest'idea può essere implementata, a livello pratico, penalizzando le osservazioni in $\mathcal{X}^{(r)}$, che contribuiscono alla stima delle densità, sottraendo effettivamente il contributo alla stima della densità che tali osservazioni portano a se stesse (e scalando opportunamente il risultato, per considerare che, dopo la penalità, si è usato un punto in meno nella stima).

- Un ultimo aspetto, che concerne puramente l'implementazione, è la possibilità di fissare un limite minimo di osservazioni perchè una moda venga considerata *cluster* dal processo di *mean-shift*. Questa possibilità offre una stabilità maggiore ai risultati

Capitolo 4

Studio di simulazione

4.1 Obiettivi

In questo capitolo si illustrano i risultati relativi ad uno studio di simulazione, svolto con l'obiettivo di valutare le capacità del metodo proposto di identificare i valori anomali, e la qualità della partizione a cui conduce.

Il comportamento del metodo sarà valutato rispetto alle seguenti caratteristiche del problema:

- configurazione di gruppo: valutata in termini di forma, separazione, e grado di bilanciamento dei gruppi
- numerosità campionaria e dimensione dello spazio di riferimento
- frazione di *outlier* nel campione.

Un ulteriore aspetto di valutazione saranno le prestazioni del metodo a seconda del valore del parametro α, impostato dall'utente per definire la frazione di punti che il metodo segnala come *outlier*.

Il metodo proposto sarà valutato nelle seguenti varianti:

NP1 stima delle densità di gruppo mediante $\hat{g}_k(x)$ definite in (3.5);

NP2 stime delle densità di gruppo mediante $\tilde{g}_k(x)$ definite in (3.9);

NP2H$_k$ stime delle densità di gruppo mediante $\tilde{g}_k(x)$ definite in (3.9) e parametri di lisciamento specifici di gruppo.

La procedura sarà valutata in termini assoluti, e in termini relativi in confronto a:

TCLUST (García-Escudero *et al.*, 2008), metodo parametrico robusto, basato su modelli a mistura finita con componenti Gaussiane.

4.2 Dettagli di implementazione

Lo studio comprende una varietà di scenari di simulazione diversi, raggruppabili in 6 categorie al variare del numero di gruppi, della loro forma, della distanza e del bilanciamento tra essi.

In particolare, indicata con $f(x)$ la densità del modello centrale, si considerano:

1. Due gruppi ben separati, di forma sferica e numerosità bilanciata; $f(x) = 0.5f_1(x) + 0.5f_2(x)$, dove f_m $(m = 1, 2)$ è la densità di una distribuzione $N_d(\mu_m, 0.3I_d)$, e μ_m la distanza euclidea tra le due medie μ_m, disposte sulla bisettrice, è pari a $\sqrt{32}$.

2. Due gruppi ben separati, di forma sferica e numerosità sbilanciata; $f(x) = 0.2f_1(x) + 0.8f_2(x)$, dove f_m $(m = 1, 2)$ è la densità di una distribuzione $N_d(\mu_m, 0.3I_d)$, la distanza euclidea tra le due medie μ_m, disposte sulla bisettrice, è pari a $\sqrt{32}$.

3. Due gruppi non ben separati, di forma sferica e numerosità bilanciata; $f(x) = 0.5f_1(x) + 0.5f_2(x)$, dove f_m $(m = 1, 2)$ è la densità di una distribuzione $N_d(\mu_m, 0.3I_d)$, la distanza euclidea tra le due medie μ_m, disposte sulla bisettrice, è pari a $\sqrt{8}$.

4. Due gruppi non ben separati, di forma sferica e numerosità sbilanciata; $f(x) = 0.2f_1(x) + 0.8f_2(x)$, dove f_m $(m = 1, 2)$ è la densità di una distribuzione $N_d(\mu_m, 0.3I_d)$, la distanza euclidea tra le due medie μ_m, disposte sulla bisettrice, è pari a $\sqrt{8}$.

5. Due gruppi ben separati, di forma ellittica e numerosità bilanciata; $f(x) = 0.5f_1(x) + 0.5f_2(x)$, dove f_m $(m = 1, 2)$ è la densità di una distribuzione $N_d(\mu_m, \Sigma_d)$, la distanza euclidea tra le due medie μ_m, disposte sulla bisettrice, è pari a $\sqrt{32}$ e Σ_d è una matrice diagonale, con valori pari 0.3 nei primi $d - 1$ elementi della diagonale e 2 nell'ultimo.

6. Un gruppo di forma altamente asimmetrica; $f(x)$ è la densità di una distribuzione le cui marginali, indipendenti, sono $\chi^2(3)$.

Per ognuna delle categorie appena definite saranno inoltre fatti variare i seguenti parametri:

1. La numerosità campionaria totale ($n = 100$, $n = 500$), di cui faranno parte anche gli eventuali *outlier*

2. Il numero di dimensioni dello spazio campionario ($d = 2$, $d = 5$)

3. Il numero di *outlier* ($n_{out} = 0$, $n_{out} = n/10$)

In ultimo, ogni scenario sarà analizzato per due valori differenti del parametro α, pari a 0.1 e 0.15.

Gli *outlier* saranno generati a partire da una distribuzione uniforme su un iperrettangolo che contenga i *cluster*; ogni valore sarà poi valutato, e sostituito, se appartiene ad una zona ad alta densità rispettivamente al gruppo più vicino (si sostituiscono quindi gli *inlier*). Nel caso di densità normali, ad esempio, si valuterà la distanza di Mahalanobis dei candidati *outlier* da ogni media μ_m, si sostituiranno quindi le osservazioni che abbiano la distanza dalla media più vicina pià piccola del quantile 0.95 di una chi-quadro con gradi di liberà uguali al numero delle dimensioni del supporto.

La capacità di *detection* dei valori anomali verrà valutata in termini di *true positive rate* (TPR), che corrisponde al numero di veri *outlier* segnalati dal metodo come anomali sul numero totale dei veri *outlier*. Si valuterà inoltre il *false positive rate* (FPR) legati alla segnalazione come anomali di punti appartenenti a ciascuno dei veri *cluster*, ovvero, per ogni vero gruppo, il numero di punti erroneamente segnalati come anomali dal metodo sul totale dei punti del gruppo.

Per quanto concerne, invece, la qualità del *clustering* conseguente al *trimming*, coerentemente alla prassi usata nella letteratura relativa al *clustering*, verrà utilizzato l'*Adjusted Rand Index* (ARI, Hubert e Arabie, 1985). L'indicatore viene costruito a partire dall'indice di Rand, che conta la proporzione di coppie di osservazioni rispetto alla cui allocazione due partizioni sono in accordo, e misurato sui campioni depurati dalle (vere) osservazioni contaminate. Fa eccezione il solo scenario 6, composto da un gruppo, dove si è usato, in luogo dell'ARI, l'indice di Fowlkes e Mallows (FMI, Fowlkes e Mallows, 1983), più appropriato per il confronto tra partizioni quando una di queste è di fatto formata da un solo gruppo.

La densità necessaria per l'implementazione della procedura non parametrica sarà stimata mediante il metodo *kernel* utilizzando come parametro di lisciamento h il valore asintoticamente ottimale sotto l'ipotesi che i dati siano Gaussiani. Nella variante $\mathbf{NP2H}_k$, i parametri di lisciamento di ciascuna densità di gruppo saranno definiti analogamente, come discusso nel paragrafo 3.2.4.

Il parametro α_0 sarà, negli scenari 1, 2, 3, 4, 6, caratterizzati dalla presenza di due *cluster*, posto uguale a 0.9 per i campioni di numerosità

$n = 100$, e a 0.98 per i campioni di numerosità $n = 500$. Nello scenario 5, caratterizzato dalla presenza di un solo gruppo, α_0 sarà invece uguale a 0.95 per $n = 100$ e a 0.99 quando $n = 500$.

Non si sfrutterà la possibilità di fissare un numero minimo di unità perchè una moda nella stima della densità formi effettivamente un *cluster*.

Con riferimento al metodo parametrico, questo sarà applicato fissando a priori il numero di gruppi pari al vero numero di *cluster*. Questo pone il suddetto metodo in vantaggio rispetto a quelli non parametrici proposti, che non assumono da un numero di *cluster* fissato. È inoltre utile ricordare che tutti i test che prevederanno la generazione del campione da una mistura di componenti Gaussiane metteranno il metodo parametrico nella situazione perfetta per il suo funzionamento, in quanto il modello assunto per la generazione dei dati puliti sarà correttamente specificato, cosa che nella realtà è raramente vera.

Lo svolgimento delle analisi sarà svolto nell'ambiente di programmazione R (R Development Core Team, 2011), mediante l'ausilio del pacchetto `ks` (Duong, 2022) che implementa l'algoritmo *mean-shift* e il pacchetto `tclust` (Fritz *et al.*, 2012) che implementa l'omonimo metodo parametrico.

4.3 Risultati

I risultati delle simulazioni sono riportati nelle Tabelle 4.1-4.12 e mostrano una generale appropriatezza della procedura proposta sia nell'identificazione degli *outlier*, quando presenti, che nel partizionamento dei dati in gruppi.

Non sorprendentemente, i risultati in generale migliori, per tutti i metodi, si ottengono quando i *cluster* sono ben separati (scenari 1 e 2, Tab. 4.1, 4.2, 4.3, 4.4), in quanto quest'aspetto rende più facile l'individuazione dei gruppi e, di conseguenza, quella degli *outlier*. In questa situazione, il bilanciamento delle dimensioni dei gruppi è un aspetto rilevante solamente per il metodo parametrico, che tende, in maniera sproporzionata, ad identificare come *outlier* punti appartenenti al gruppo di dimensione più piccola. Questo, associato ad una sovrastima del vero numero di *outlier*, porta a risultati inferiori nel *clustering* (Tab. 4.3, 4.4). In questo senso, appare giustificata la strategia, adottata nella procedura proposta, di valutare il grado di *outlyingness* normalizzando le densità dei *cluster*.

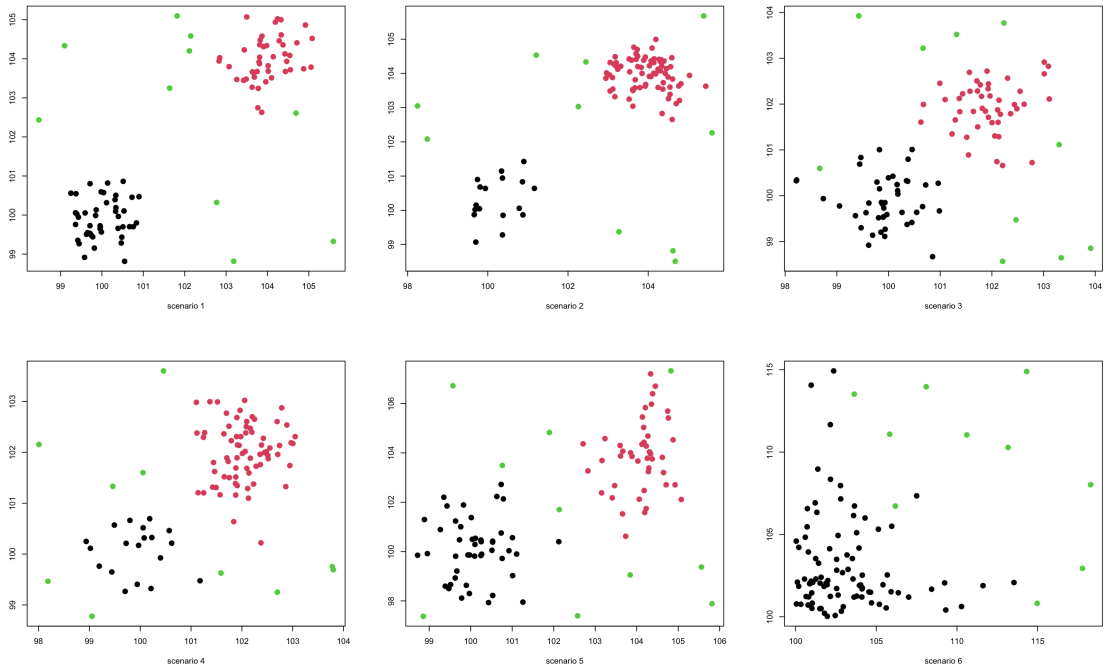

Figura 4.1: Esempi di campioni dai vari scenari ($n = 100$, $n_{out} = n/10$, $d = 2$); in verde gli *outlier* nominali.

Gli scenari (3 e 4) caratterizzati da gruppi più vicini (Tab. 4.5, 4.6, 4.7, 4.8) rappresentano situazioni in generale più complesse, sia dal punto di vista dell'individuazione dei valori anomali, sia dal punto di vista del *clustering* conseguente. I metodi valutati sembrano portare ancora a risultati comparabili, in generale quando i gruppi sono bilanciati, e finchè la frazione di *outlier* non è sovrastimata quando i gruppi sono di dimensione diversa; sono però incrementate le criticità per TCLUST, legate ad una sbilanciata eliminazione di punti dal gruppo più piccolo.

La variazione della forma, da sferica a ellittica, dei gruppi (scenario 5, Tab 4.9, 4.10), è un aspetto apparentemente più critico per i metodi non parametrici NP2 e NP1. Tuttavia, il lieve deterioramento dei risultati è più verosimilmente dovuto all'*oversmoothing* legato all'uso di un unico parametro di lisciamento H. Lo scenario 5, per $d = 2$, dimostra invece il vantaggio di flessibilità del metodo NP2.H$_k$ rispetto alla controparte NP2, e a NP1 (tabella 4.10). La causa sostanziale di questa differenza è da ricercarsi però nell'*oversmoothing* associato alla all'uso di un'unica *bandwidth* H, in quanto, proprio a causa dell'eccessiva lisciatura, la forma del *kernel* influenza direttamente quella della densità di ogni *cluster*, costringendola ad assumere una configurazione più sferica.

L'effetto in sostanza svanisce per $d = 5$, essendo qui l'*oversmoothing* meno presente. La lontananza tra i *cluster* impedisce comunque alla problematica di mostrarsi nei risultati di *clustering*. In effetti, il *clustering* modale è noto per le sue capacità di gestire, a differenza dei metodi parametrici, gruppi di forma arbitraria. Questo è anche testimoniato dai buoni risultati delle procedure non parametriche nello scenario 6 (Tab. 4.12), che presenta in generale estreme difficoltà a causa della presenza di un unico gruppo di forma molto asimmetrica. In questo caso, valori elevati dell'indicatore di qualità del *clustering* utilizzato suggeriscono la capacità di massima degli algoritmi di identificare il gruppo nominale, pur a fronte di una naturale tendenza all'individuazione di gruppi spuri.

Il numero di dimensioni sembra essere, in generale, un aspetto favorevole all'individuazione degli *outlier*. Questo è probabilmente dovuto alla maggiore distinzione, in termini di sparsità, tra le regioni legate ai cluster e quelle legate ai valori anomali. Questi, infatti, dato l'aumento di volume dello spazio campionario conseguente all'aumento delle dimensioni, tenderanno a trovarsi più isolati, sia rispetto alle masse dei dati associate ai *cluster*, sia relativamente agli altri *outlier*. Sarà, per questo motivo, più difficile la formazione di mode spurie di numerosità considerevole, che mettano in difficoltà il processo di *trimming*.

Infine, la numerosità sembra essere un'aspetto particolarmente interessante. Se da un lato all'aumentare della dimensione campionaria la stima dei domini di attrazione dei veri *cluster* diventa più stabile e accurata, dall'altro le regioni associate agli *outlier*, a parità di volume, diventano effettivamente meno sparse. L'effetto è concettualmente inverso a quello appena descritto per il numero di dimensioni. I metodi non parametrici potrebbero avere quindi maggiori difficoltà ad individuare quando una moda spuria composta da un maggior numero di punti è effettivamente tale. Per contrastare questo fenomeno, è fondamentale la scelta di valori più alti di α_0 al crescere di n, magari rendendo costante il numero di osservazioni che compongono il campione iniziale. Nella maggior parte dei casi (soprattutto se il numero di parametri da stimare è più alto, come per il metodo NP2H$_k$) l'incremento nell'accuratezza delle stime legata all'aumento della dimensione campionaria ha un peso maggiore nei risultati rispetto problematica sopra citata; è pero sicuramente necessario un approfondimento da questo punto di vista, collegato a variazioni della numerosità campionaria di almeno un'ordine di grandezza.

Con riferimento al comportamento della procedura in presenza di contaminazione, a parità di altre condizioni, un'analisi poco attenta del

tasso di corretta identificazione degli outlier (TPR, Tab. 4.2, 4.4, 4.6, 4.8, 4.10, 4.12)) potrebbe apparentemente suggerire una tendenza, condivisa dalla totalità dei metodi sotto esame, a non identificare una quota di valori anomali. Tuttavia, è utile notare che le osservazioni generate come *outlier* non saranno necessariamente i valori più anomali in un dato campione, ma una quota di osservazioni, sulle code dei *cluster* risulta tipicamente più estrema rispetto ai valori anomali nominali. Ad esempio, il TPR di circa il 0.9, raggiunto, per $d = 2$ e $\alpha = 0.1$, dai metodi NP1 e NP2 negli scenari 1 e 2, dove i gruppi sono sferici e ben separati (tabelle 4.2, 4.4), è evidenza di un'individuazione sostanzialmente perfetta dei valori più anomali dei campioni. A riprova di questo è possibile osservare come il risultato non migliori all'aumentare della numerosità campionaria, ma mostri un miglioramento solo all'aumentare della frazione di eliminazione α. Lo stesso fenomeno è osservabile nelle tabelle sopra citate anche per $d = 5$, ma in questo caso il TPR si stabilisce su un valore di circa 0.97, per $\alpha = 0.1$, in virtù del maggior volume a disposizione nello spazio campionario all'aumentare del numero delle dimensioni, che permette ai valori generati come anomali di disporsi, con probabilità più alta, in posizioni distanti dalle masse dei *cluster*.

In assenza di contaminazione, si può notare come i metodi non parametrici di *trimming* non sembrino, in generale, minare l'efficienza delle procedure di *clustering*. Se i gruppi sono in origine separati, i risultati sono effettivamente perfetti anche dopo l'eliminazione delle code dei *cluster* (Tabelle 4.1, 4.3); anche quando i gruppi sono vicini, non sembra esserci una differenza nel risultato del *clustering* al variare del valore di α (tabelle 4.5, 4.5). Il discorso è diverso, come detto, per il metodo parametrico, quando i gruppi sono di dimensione sbilanciata, e soprattutto quando sono vicini (tabelle 4.3, 4.7).

Nella tabella 4.8 si può notare però come, nonostante i metodi non parametrici normalizzino le densità di cluster per valurate l'*outlyingness*, ci sia comunque una disparità tra i tassi di errata classificazione dei gruppi; ciò è verosimilmente dovuto alla tecnica del sotto-campionamento iniziale utilizzata per aumentare la robustezza del metodo, che intrinsecamente va a penalizzare maggiormente le zone più sparse. Essendo sostanzialmente la sparsità il metro di paragone che permetta di riconoscere gli *outlier*, è naturale che un gruppo più piccolo, e quindi più sparso, venga intrinsecamente penalizzato. Tale disparità nell'errata selezione dei valori anomali dai vari gruppi è però, per i metodi non parametrici, comunque più contenuta rispetto a TCLUST, e tende a diminuire all'aumentare della dimensione del campione.

Confrontando in generale le procedure proposte e TCLUST, è utile ricordare che quest'ultimo è messo nella condizione perfetta per il suo funzionamento in tutti gli scenari definiti come mistura di componenti normali. L'apparente vantaggio, negli scenari da 1 a 4, in termini di TPR nell'individuazione degli *oultier*, dei metodi NP2 e NP1 rispetto a TCLUST e NP2H$_k$ è da giustificarsi nella congiunzione dell'*oversmoothing* con forma sferica dei gruppi; è quindi un risultato altamente dipendente dalle condizioni di simulazione, e non generalizzabile, come si vede quando la forma dei cluster non è più sferica (scenario 5). Il metodo NP2H$_k$ è comunque, in generale, comparabile con TCLUST, e abbastanza flessibile da essere applicato in situazioni più varie. È necessario però notare che il metodo parametrico non deve affrontare il problema forse più complesso, legato alla valutazione di eventuali mode spurie, in quanto il numero di gruppi è fissato a priori.

Dal punto di vista del *clustering* è poi utile ricordare che le differenze nei risultati tra i metodi non parametrici e TCLUST sono dovute, non solo alla diversa identificazione degli *outlier*, ma effettivamente anche al diverso metodo di *clustering* utilizzato.

Per quanto concerne i risultati riguardanti lo scenario 6 (tabelle 4.11, 4.12), è interessante notare come, in presenza di contaminazione, il metodo TCLUST, anche se applicato ad una distribuzione non normale, riesce ad individuare gli *outlier* in maniera soddisfacente e apparentemente migliore rispetto ai metodi non parametrici. Questo è dovuto però alla maggiore difficoltà legata al problema del *trimming* dal punto di vista non parametrico, in quanto, se il metodo parametrico deve semplicemente individuare gli *outlier* rispetto all'unico gruppo fissato, i metodi non parametrici devono prima riconoscere ed eliminare eventuali gruppi spuri, per poi identificare i valori anomali rispetto ai *cluster* rimanenti. In questo particolare scenario infatti, soprattutto per $n = 500$, i metodi proposti faticano a eliminare tutte le mode spurie (composte a volte da un discreto numero di *outlier*); tale criticità si è in parte risolta aumentando il valore del parametro α_0 rispetto agli altri scenari, da 0.98 a 0.99 per $n = 500$. Sono sicuramente necessarie analisi approfondite in questa direzione, per comprendere l'effetto della numerosità campionaria sulla qualità dell'individuazione delle mode *spurie*, e su come α_0 si relazioni a tale problema. Si noti inoltre che, per $d = 5$, il *clustering* modale, anche a fronte di un'eliminazione quasi totale degli *outlier*, o all'assenza di contaminazione, soffre delle problematiche legate alla maledizione della dimensionalità, ma offre risultati di raggruppamento comunque accettabili. È utile comunque tenere a mente che il legge-

ro calo della qualità dei risultati di *clustering*, all'aumentare di α, è verosimilmente riconducibile anche alla diminuzione delle varianze stimate per ogni dimensione, conseguente all'eliminazione dei valori più estremi; questo infatti, portando ad una più piccola stima della *bandwidth*, aumenta la possibilità dell'apparizione di mode spurie anche tra le osservazioni non anomale.

Per un'analisi più approfondita sarebbe sicuramente utile sottoporre i metodi a scenari ancora più complessi, in modo da far emergere in maniera più netta le differenze tra le varie tecniche.

4.4 Considerazioni finali

Sebbene nei contesti non parametrici, ancor più che in quelli parametrici, le procedure di stima sono note per la loro vulnerabilità alla presenza di *outlier*, il tema della robustezza è stato ampiamente trascurato dalla letteratura relativa al *clustering* modale. In questo lavoro sono stati approfonditi alcuni aspetti legati a questo tema, concentrando l'attenzione soprattutto sul versante concettuale e computazionale e con particolare riferimento ai metodi di clustering basati sulla ricerca delle mode. Si è inoltre sviluppata una procedura iterativa di *trimming* per identificare i valori anomali di gruppi modali e irrobustire l'individuazione di gruppi.

Le analisi empiriche hanno mostrato un comportamento soddisfacente della procedura proposta al variare delle condizioni sperimentali e anche comparativamente rispetto ad un metodo robusto parametrico, applicato in condizioni più vantaggiose.

A completamento di questo lavoro, sono auspicabili alcuni approfondimenti volti ad un più formale studio delle proprietà teoriche della tecnica proposta, nonché della sua efficienza; in particolare è di interesse lo studio, sia teorico che empirico, dell'influenza di α_0 sulle proprietà del metodo al variare del numero di *cluster* e, soprattutto, della numerosità campionaria. Inoltre, sebbene il problema della scelta del parametro di lisciamento è noto per essere particolarmente critico, considerazioni utili potrebbero essere tratte da uno studio di sensibilità della procedura proposta, e in generale delle tecniche di clustering non parametrico, a variazioni di H. Ancora, è auspicabile sviluppare una procedura guidata dai dati per la determinazione della più appropriata quota α di osservazioni da scartare. Infine, variazioni del metodo proposto che facciano uso, in sostituzione della mediana, di un quantile $q < 0.5$, sono degne di essere prese in considerazione.

Tabella 4.1: Risultati relativi allo scenario 1 in assenza di contaminazione dei campioni. False Positive Rate medio (FPR) associato ai gruppi 1 e 2, gr_1, gr_2 di probabilità pari a $\pi_1 = 0.5, \pi_2 = 0.5$, Adjusted Rand Index medio (ARI) relativo al confronto con la vera partizione dei dati, e sua deviazione standard (SD).

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0996	0.0988	0.0996	0.1004
		FPR gr_2	0.1004	0.1012	0.1004	0.0996
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP1	FPR gr_1	0.0994	0.0982	0.0998	0.1004
		FPR gr_2	0.1006	0.1018	0.1002	0.0996
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.102	0.1002	0.1012	0.1010
		FPR gr_2	0.098	0.0998	0.0988	0.0989
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	TCLUST	FPR gr_1	0.101	0.0963	0.0999	0.0980
		FPR gr_2	0.099	0.1038	0.1001	0.1021
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.1475	0.1499	0.1499	0.1484
		FPR gr_2	0.1526	0.1501	0.1501	0.1516
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP1	FPR gr_1	0.1471	0.1507	0.1499	0.1484
		FPR gr_2	0.1530	0.1493	0.1501	0.1516
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.1469	0.1480	0.1505	0.1528
		FPR gr_2	0.1532	0.1521	0.1495	0.1471
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	TCLUST	FPR gr_1	0.1418	0.1489	0.1521	0.1479
		FPR gr_2	0.1584	0.1511	0.1479	0.1522
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]

Tabella 4.2: Risultati relativi allo scenario 1 in presenza di contaminazione dei campioni. Cfr. Tab 4.1. Si riporta anche il True Positive rate medio (TPR) di identificazione degli *outlier*.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0101	0.0020	0.0113	0.0033
		FPR gr_2	0.0137	0.0037	0.0113	0.0042
		TPR	0.8930	0.9740	0.8984	0.9664
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP1	FPR gr_1	0.0103	0.0020	0.0111	0.0033
		FPR gr_2	0.0140	0.0037	0.0114	0.0043
		TPR	0.8910	0.9740	0.8988	0.9660
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.0138	0.0047	0.0123	0.0051
		FPR gr_2	0.0187	0.0042	0.0129	0.0049
		TPR	0.8540	0.9600	0.8868	0.9548
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	TCLUST	FPR gr_1	0.0116	0.005	0.0124	0.0034
		FPR gr_2	0.0178	0.004	0.0109	0.0043
		TPR	0.8680	0.960	0.8952	0.9656
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.0557	0.0531	0.0569	0.0568
		FPR gr_2	0.0586	0.0582	0.0550	0.0544
		TPR	0.9860	0.9990	0.9964	1.0000
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP1	FPR gr_1	0.0544	0.0536	0.0569	0.0565
		FPR gr_2	0.0599	0.0577	0.0550	0.0546
		TPR	0.9860	0.9990	0.9964	1.0000
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.0566	0.0576	0.0597	0.0568
		FPR gr_2	0.0624	0.0558	0.0545	0.0555
		TPR	0.9650	0.9900	0.9864	0.9944
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
	TCLUST	FPR gr_1	0.0546	0.0549	0.0584	0.0568
		FPR gr_2	0.0647	0.0573	0.0543	0.0546
		TPR	0.9640	0.9950	0.9928	0.9984
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]

Tabella 4.3: Risultati relativi allo scenario 2 in assenza di contaminazione dei campioni. False Positive Rate medio (FPR) associato ai gruppi 1 e 2, (gr1, gr2) di probabilità pari a $\pi_1 = 0.2, \pi_2 = 0.8$. Cfr. Tab 4.1.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.1024	0.1220	0.1067	0.1046
		FPR gr_2	0.0994	0.0945	0.0983	0.0988
		$\overline{ARI}[SD]$	1[0]	0.9908[0.0859]	1[0]	1[0]
	NP1	FPR gr_1	0.1075	0.1230	0.1075	0.1052
		FPR gr_2	0.0982	0.0943	0.0981	0.0987
		$\overline{ARI}[SD]$	1[0]	0.9910[0.0859]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.1199	0.1412	0.1105	0.1179
		FPR gr_2	0.0952	0.0898	0.0973	0.0955
		$\overline{ARI}[SD]$	1[0]	0.9909[0.0859]	1[0]	1[0]
	TCLUST	FPR gr_1	0.3061	0.2607	0.3077	0.2388
		FPR gr_2	0.0503	0.0602	0.0473	0.0649
		$\overline{ARI}[SD]$	0.9917[0.0312]	0.9513[0.1389]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.1502	0.1715	0.1522	0.1556
		FPR gr_2	0.1500	0.1447	0.1495	0.1486
		$\overline{ARI}[SD]$	1[0]	0.9906[0.0866]	1[0]	1[0]
	NP1	FPR gr_1	0.1569	0.1780	0.1534	0.1566
		FPR gr_2	0.1483	0.1431	0.1492	0.1483
		$\overline{ARI}[SD]$	1[0]	0.9906[0.0866]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.1744	0.1967	0.1495	0.1815
		FPR gr_2	0.1441	0.1385	0.1501	0.1420
		$\overline{ARI}[SD]$	1[0]	0.9912[0.0860]	1[0]	1[0]
	TCLUST	FPR gr_1	0.4496	0.3938	0.5099	0.3563
		FPR gr_2	0.0777	0.0897	0.0613	0.0978
		$\overline{ARI}[SD]$	0.9441[0.1614]	0.8693[0.02322]	0.9973[0.0120]	0.9996[0.0019]

Tabella 4.4: Risultati relativi allo scenario 2 in presenza di contaminazione dei campioni. False Positive Rate medio (FPR) associato ai gruppi 1 e 2, (gr1, gr2) di probabilità pari a $\pi_1 = 0.2, \pi_2 = 0.8$. Cfr. Tab 4.2.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0159	0.0066	0.0157	0.0040
		FPR gr_2	0.0106	0.0024	0.0116	0.0043
		TPR	0.8950	0.9710	0.8880	0.9616
		$\overline{ARI}[SD]$	1[0]	0.9998[0.0013]	1[0]	1[0]
	NP1	FPR gr_1	0.0159	0.0066	0.0159	0.0040
		FPR gr_2	0.0104	0.0025	0.0115	0.0043
		TPR	0.8970	0.9700	0.8884	0.9616
		$\overline{ARI}[SD]$	1[0]	0.9998[0.0013]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.0153	0.0093	0.0144	0.0054
		FPR gr_2	0.0146	0.0042	0.0150	0.0057
		TPR	0.8670	0.9530	0.8664	0.9492
		$\overline{ARI}[SD]$	1[0]	0.9998[0.0013]	1[0]	1[0]
	TCLUST	FPR gr_1	0.0505	0.0153	0.0395	0.0096
		FPR gr_2	0.0065	0.0029	0.0055	0.0027
		TPR	0.8640	0.9510	0.8900	0.9636
		$\overline{ARI}[SD]$	1[0]	1[0]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.0744	0.0701	0.0651	0.0655
		FPR gr_2	0.0525	0.0518	0.0540	0.0533
		TPR	0.9890	1.0000	0.9940	0.9988
		$\overline{ARI}[SD]$	1[0]	0.9992[0.0054]	1[0]	1[0]
	NP1	FPR gr_1	0.0772	0.0729	0.0665	0.0655
		FPR gr_2	0.0519	0.0513	0.0537	0.0533
		TPR	0.9880	0.9990	0.9940	0.9988
		$\overline{ARI}[SD]$	1[0]	0.9992[0.0054]	1[0]	1[0]
	NP2H$_k$	FPR gr_1	0.0812	0.0915	0.0633	0.0764
		FPR gr_2	0.0537	0.0478	0.0565	0.0514
		TPR	0.9680	0.9900	0.9792	0.9928
		$\overline{ARI}[SD]$	1[0]	0.9995[0.0054]	1[0]	1[0]
	TCLUST	FPR gr_1	0.2044	0.1458	0.2006	0.1426
		FPR gr_2	0.0243	0.0336	0.0236	0.0345
		TPR	0.9640	0.9930	0.9724	0.9968
		$\overline{ARI}[SD]$	0.9941[0.0389]	0.9766[0.0608]	1[0]	1[0]

Tabella 4.5: Risultati relativi allo scenario 3 in assenza di contaminazione dei campioni. Cfr Tab. 4.1.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.1003	0.0979	0.0986	0.0982
		FPR gr_2	0.0997	0.1021	0.1014	0.1018
		$\overline{ARI}[SD]$	0.9790[0.0337]	0.9617[0.0670]	0.9801[0]	0.9773[0]
	NP1	FPR gr_1	0.1020	0.0963	0.0992	0.0982
		FPR gr_2	0.0979	0.1037	0.1008	0.1018
		$\overline{ARI}[SD]$	0.9790[0.0337]	0.9598[0.0686]	0.9801[0.0154]	0.9771[0.0142]
	NP2H$_k$	FPR gr_1	0.0961	0.0989	0.0977	0.0986
		FPR gr_2	0.1040	0.1011	0.1023	0.1014
		$\overline{ARI}[SD]$	0.9786[0.0337]	0.9621[0.0654]	0.9807[0.0147]	0.9768[0.0150]
	TCLUST	FPR gr_1	0.1005	0.0967	0.1008	0.0970
		FPR gr_2	0.0995	0.1033	0.0992	0.1031
		$\overline{ARI}[SD]$	0.9650[0.0487]	0.9572[0.0540]	0.9792[0.0150]	0.9754[0.0158]
$\alpha = 0.15$	NP2	FPR gr_1	0.1450	0.1465	0.1478	0.1482
		FPR gr_2	0.1551	0.1536	0.1522	0.1518
		$\overline{ARI}[SD]$	0.9790[0.0341]	0.9622[0.0687]	0.9800[0.0156]	0.9766[0.0143]
	NP1	FPR gr_1	0.147	0.1445	0.1482	0.1498
		FPR gr_2	0.153	0.1556	0.1518	0.1502
		$\overline{ARI}[SD]$	0.9790[0.0337]	0.9619[0.0670]	0.9803[0.0154]	0.9721[0.0342]
	NP2H$_k$	FPR gr_1	0.1496	0.1469	0.1469	0.148
		FPR gr_2	0.1504	0.1532	0.1531	0.152
		$\overline{ARI}[SD]$	0.9794[0.0324]	0.9583[0.0696]	0.9796[0.0152]	0.9766[0.0143]
	TCLUST	FPR gr_1	0.1468	0.1481	0.1522	0.1471
		FPR gr_2	0.1532	0.1520	0.1478	0.1529
		$\overline{ARI}[SD]$	0.9630[0.0484]	0.9452[0.0710]	0.9792[0.0151]	0.9747[0.0151]

Tabella 4.6: Risultati relativi allo scenario 3 in presenza di contaminazione dei campioni. Cfr. Tab. 4.2.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0193	0.0098	0.0153	0.0091
		FPR gr_2	0.0165	0.0089	0.0148	0.0093
		TPR	0.8390	0.9160	0.8648	0.9176
		$\overline{ARI}[SD]$	0.9797[0.0303]	0.9598[0.0481]	0.9798[0.0134]	0.9781[0.0145]
	NP1	FPR gr_1	0.0188	0.0103	0.0149	0.0096
		FPR gr_2	0.0156	0.0109	0.0154	0.0093
		TPR	0.8450	0.9050	0.8636	0.9148
		$\overline{ARI}[SD]$	0.9797[0.0303]	0.9601[0.0476]	0.9800[0.0134]	0.9781[0.0145]
	NP2H$_k$	FPR gr_1	0.0232	0.0123	0.0163	0.0114
		FPR gr_2	0.0235	0.0111	0.0199	0.0105
		TPR	0.7900	0.8950	0.8368	0.9012
		$\overline{ARI}[SD]$	0.9801[0.0303]	0.9602[0.0481]	0.9800[0.0135]	0.9778[0.0147]
	TCLUST	FPR gr_1	0.0223	0.0111	0.0150	0.0082
		FPR gr_2	0.0201	0.0106	0.0157	0.0081
		TPR	0.8090	0.9020	0.8616	0.9268
		$\overline{ARI}[SD]$	0.9692[0.0418]	0.9610[0.0464]	0.9769[0.0168]	0.9763[0.0161]
$\alpha = 0.15$	NP2	FPR gr_1	0.0604	0.0553	0.0553	0.0571
		FPR gr_2	0.0587	0.0587	0.0580	0.0555
		TPR	0.9640	0.9870	0.9904	0.9932
		$\overline{ARI}[SD]$	0.9805[0.0290]	0.9621[0.0531]	0.9793[0.0133]	0.9783[0.0145]
	NP1	FPR gr_1	0.0600	0.0546	0.0566	0.0572
		FPR gr_2	0.0598	0.0589	0.0577	0.0555
		TPR	0.9610	0.9890	0.9856	0.9932
		$\overline{ARI}[SD]$	0.9805[0.0290]	0.9620[0.0527]	0.9802[0.0136]	0.9781[0.0143]
	NP2H$_k$	FPR gr_1	0.0604	0.0598	0.0598	0.0581
		FPR gr_2	0.0641	0.0580	0.0571	0.0568
		TPR	0.9400	0.9700	0.9736	0.9828
		$\overline{ARI}[SD]$	0.9806[0.0302]	0.9658[0.0452]	0.9798[0.0131]	0.9777[0.0143]
	TCLUST	FPR gr_1	0.0619	0.0595	0.0552	0.0563
		FPR gr_2	0.0632	0.0580	0.0592	0.0559
		TPR	0.9370	0.9710	0.9852	0.9952
		$\overline{ARI}[SD]$	0.9684[0.0425]	0.9631[0.0455]	0.9774[0.0137]	0.9753[0.0172]

Tabella 4.7: Risultati relativi allo scenario 4 in assenza di contaminazione dei campioni. Cfr. Tab 4.3.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.1263	0.1635	0.1105	0.1247
		FPR gr_2	0.0936	0.0841	0.0974	0.0938
		$\overline{ARI}[SD]$	0.9664[0.0385]	0.9011[0.1182]	0.9765[0.0132]	0.9662[0.0233]
	NP1	FPR gr_1	0.1248	0.108	0.1097	0.1181
		FPR gr_2	0.0940	0.098	0.0976	0.0955
		$\overline{ARI}[SD]$	0.9669[0.0382]	0.9175[0.0778]	0.9770[0.0130]	0.9664[0.0226]
	NP2H$_k$	FPR gr_1	0.1463	0.2105	0.1077	0.1285
		FPR gr_2	0.0888	0.0724	0.0981	0.0929
		$\overline{ARI}[SD]$	0.0.9651[0.0429]	0.8879[0.1295]	0.9764[0.0130]	0.9650[0.0232]
	TCLUST	FPR gr_1	0.3166	0.2505	0.3173	0.2521
		FPR gr_2	0.0474	0.0624	0.0454	0.0618
		$\overline{ARI}[SD]$	0.9017[0.0951]	0.8208[0.9677]	0.9677[0.0188]	0.9721[0.0183]
$\alpha = 0.15$	NP2	FPR gr_1	0.1765	0.2145	0.1630	0.1747
		FPR gr_2	0.1436	0.1339	0.1467	0.1438
		$\overline{ARI}[SD]$	0.9627[0.0965]	0.8939[0.1374]	0.9774[0.0130]	0.9666[0.0214]
	NP1	FPR gr_1	0.1729	0.1685	0.1617	0.1712
		FPR gr_2	0.1444	0.1454	0.1471	0.1447
		$\overline{ARI}[SD]$	0.9711[0.0336]	0.9056[0.1269]	0.9772[0.0131]	0.9565[0.0747]
	NP2H$_k$	FPR gr_1	0.2153	0.2800	0.1630	0.1855
		FPR gr_2	0.1341	0.1175	0.1467	0.1411
		$\overline{ARI}[SD]$	0.9680[0.0351]	0.8546[0.1870]	0.9765[0.0138]	0.9551[0.0602]
	TCLUST	FPR gr_1	0.4517	0.370	0.4921	0.3654
		FPR gr_2	0.0767	0.095	0.0640	0.0960
		$\overline{ARI}[SD]$	0.8134[0.1792]	0.7205[0.2359]	0.9365[0.0545]	0.9638[0.0237]

Tabella 4.8: Risultati relativi allo scenario 4 in presenza di contaminazione dei campioni. Cfr. Tab 4.4.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0546	0.0278	0.0282	0.0151
		FPR gr_2	0.0119	0.0063	0.0174	0.0068
		TPR	0.8170	0.9040	0.8240	0.9236
		$\overline{ARI}[SD]$	0.9480[0.1062]	0.8947[0.0858]	0.9784[0.0152]	0.9524[0.0567]
	NP1	FPR gr_1	0.0394	0.0229	0.0196	0.0155
		FPR gr_2	0.0125	0.0119	0.0177	0.0085
		TPR	0.8400	0.8730	0.8376	0.9108
		$\overline{ARI}[SD]$	0.9480[0.1062]	0.8970[0.0945]	0.9770[0.0173]	0.9523[0.0525]
	NP2H$_k$	FPR gr_1	0.0580	0.037	0.0295	0.0137
		FPR gr_2	0.0145	0.007	0.0197	0.0089
		TPR	0.7920	0.882	0.8048	0.9112
		$\overline{ARI}[SD]$	0.9467[0.1056]	0.8862[0.1023]	0.9784[0.0152]	0.9594[0.0216]
	TCLUST	FPR gr_1	0.0687	0.037	0.0526	0.0192
		FPR gr_2	0.0072	0.007	0.0085	0.0048
		TPR	0.8260	0.882	0.8432	0.9304
		$\overline{ARI}[SD]$	0.9550[0.0580]	0.9124[0.1057]	0.9785[0.0154]	0.9810[0.0127]
$\alpha = 0.15$	NP2	FPR gr_1	0.1132	0.0936	0.0642	0.0766
		FPR gr_2	0.0487	0.0491	0.0564	0.0510
		TPR	0.9470	0.9760	0.9780	0.9940
		$\overline{ARI}[SD]$	0.9529[0.1050]	0.9011[0.1133]	0.9784[0.0147]	0.9633[0.0209]
	NP1	FPR gr_1	0.1042	0.0789	0.0647	0.0709
		FPR gr_2	0.0496	0.0535	0.0563	0.0525
		TPR	0.9570	0.9720	0.9784	0.9940
		$\overline{ARI}[SD]$	0.9622[0.0421]	0.8985[0.1178]	0.9788[0.0145]	0.9634[0.0214]
	NP2H$_k$	FPR gr_1	0.1217	0.1089	0.0666	0.0755
		FPR gr_2	0.0490	0.0465	0.0589	0.0530
		TPR	0.9300	0.9670	0.9560	0.9820
		$\overline{ARI}[SD]$	0.9529[0.1055]	0.8917[0.1274]	0.9786[0.0148]	0.9627[0.0209]
	TCLUST	FPR gr_1	0.2034	0.1415	0.1885	0.1403
		FPR gr_2	0.0288	0.0377	0.0290	0.0349
		TPR	0.9310	0.9700	0.9488	0.9928
		$\overline{ARI}[SD]$	0.9302[0.0834]	0.8830[0.1184]	0.9762[0.0164]	0.9788[0.0147]

Tabella 4.9: Risultati relativi allo scenario 5 in assenza di contaminazione dei campioni. Cfr. Tab 4.1.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.1040	0.0927	0.1007	0.1021
		FPR gr_2	0.0961	0.1074	0.0993	0.0979
		$\overline{ARI}[SD]$	0.9964[0.0128]	1[0]	0.9981[0.0038]	0.9998[0.0011]
	NP1	FPR gr_1	0.1034	0.0945	0.1006	0.102
		FPR gr_2	0.0967	0.1056	0.0994	0.098
		$\overline{ARI}[SD]$	0.9964[0.0128]	1[0]	0.9981[0.0038]	0.9998[0.0011]
	NP2H$_k$	FPR gr_1	0.1094	0.0943	0.1006	0.1004
		FPR gr_2	0.0907	0.1058	0.0994	0.0996
		$\overline{ARI}[SD]$	0.9972[0.0117]	0.9985[0.0152]	0.9987[0.0030]	0.9998[0.0011]
	TCLUST	FPR gr_1	0.1030	0.0962	0.0991	0.1011
		FPR gr_2	0.0971	0.1038	0.1009	0.0989
		$\overline{ARI}[SD]$	0.9992[0.0056]	1[0]	0.9997[0.0016]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.1544	0.1418	0.1499	0.1515
		FPR gr_2	0.1456	0.1583	0.1501	0.1485
		$\overline{ARI}[SD]$	0.9956[0.0138]	0.9985[0.0152]	0.9976[0.0043]	0.9998[0.0011]
	NP1	FPR gr_1	0.1542	0.1410	0.1499	0.1515
		FPR gr_2	0.1458	0.1591	0.1501	0.1485
		$\overline{ARI}[SD]$	0.9952[0.0153]	0.9985[0.0152]	0.9976[0.0043]	0.9998[0.0011]
	NP2H$_k$	FPR gr_1	0.1542	0.1473	0.1485	0.1509
		FPR gr_2	0.1458	0.1527	0.1515	0.1491
		$\overline{ARI}[SD]$	0.9964[0.0139]	0.9985[0.0152]	0.9984[0.0032]	1[0]
	TCLUST	FPR gr_1	0.1585	0.1324	0.1456	0.1506
		FPR gr_2	0.1416	0.1678	0.1544	0.1494
		$\overline{ARI}[SD]$	0.9988[0.0069]	1[0]	0.9997[0.0016]	1[0]

Tabella 4.10: Risultati relativi allo scenario 5 in presenza di contaminazione dei campioni. Cfr. Tab 4.2.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0225	0.0070	0.0170	0.0061
		FPR gr_2	0.0239	0.0036	0.0194	0.0057
		TPR	0.7910	0.9520	0.8364	0.9472
		$\overline{ARI}[SD]$	0.9973[0.0106]	0.9958[0.0296]	0.9993[0.0024]	1[0]
	NP1	FPR gr_1	0.0265	0.0070	0.0179	0.0061
		FPR gr_2	0.0248	0.0036	0.0196	0.0057
		TPR	0.7690	0.9520	0.8312	0.9468
		$\overline{ARI}[SD]$	0.9969[0.0114]	0.9958[0.0296]	0.9993[0.0024]	1[0]
	NP2H$_k$	FPR gr_1	0.0218	0.0094	0.0167	0.0068
		FPR gr_2	0.0197	0.0045	0.0177	0.0054
		TPR	0.8130	0.9370	0.8452	0.9452
		$\overline{ARI}[SD]$	0.9973[0.0106]	0.9958[0.0296]	0.9993[0.0024]	1[0]
	TCLUST	FPR gr_1	0.0196	0.0081	0.0145	0.0054
		FPR gr_2	0.0188	0.0052	0.0160	0.0049
		TPR	0.8270	0.9400	0.8628	0.9532
		$\overline{ARI}[SD]$	0.9987[0.0098]	1[0]	1[0]	0.9998[0.0013]
$\alpha = 0.15$	NP2	FPR gr_1	0.0635	0.0590	0.0612	0.0556
		FPR gr_2	0.0654	0.0536	0.0608	0.0566
		TPR	0.9200	0.9930	0.9508	0.9952
		$\overline{ARI}[SD]$	0.9960[0.0128]	0.9937[0.0362]	0.9989[0.0029]	1[0]
	NP1	FPR gr_1	0.0644	0.0585	0.0608	0.0555
		FPR gr_2	0.0665	0.0541	0.0613	0.0567
		TPR	0.9110	0.9930	0.9508	0.9952
		$\overline{ARI}[SD]$	0.9960[0.0128]	0.9937[0.0362]	0.9989[0.0029]	1[0]
	NP2H$_k$	FPR gr_1	0.0606	0.0612	0.0573	0.0557
		FPR gr_2	0.0614	0.0538	0.0577	0.0572
		TPR	0.9510	0.9820	0.9824	0.9920
		$\overline{ARI}[SD]$	0.9964[0.0121]	0.9957[0.0303]	0.9993[0.0024]	1[0]
	TCLUST	FPR gr_1	0.0592	0.0583	0.0555	0.0551
		FPR gr_2	0.0608	0.0554	0.0580	0.0566
		TPR	0.9600	0.9880	0.9896	0.9976
		$\overline{ARI}[SD]$	0.9996[0.0044]	1[0]	1[0]	0.9998[0.0013]

Tabella 4.11: Risultati relativi allo scenario 6 in assenza di contaminazione dei campioni. Cfr. Tab 4.1.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.1	0.1	0.1	0.1
		$\overline{FMI}[SD]$	0.9721[0.0861]	0.6720[0.1062]	0.9921[0.0461]	0.8701[0.0365]
	NP1	FPR gr_1	0.1	0.1	0.1	0.1
		$\overline{FMI}[SD]$	0.9670[0.0914]	0.6903[0.1201]	0.9926[0.0426]	0.8750[0.0375]
	NP2H$_k$	FPR gr_1	0.1	0.1	0.1	0.1
		$\overline{FMI}[SD]$	0.9673[0.0943]	0.6814[0.1077]	0.9957[0.0217]	0.8649[0.0382]
	TCLUST	FPR gr_1	0.1	0.1	0.1	0.1
		$\overline{FMI}[SD]$	1[0]	1[0]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.15	0.15	0.15	0.15
		$\overline{FMI}[SD]$	0.9581[0.1075]	0.6124[0.1111]	0.9852[0.0767]	0.8385[0.0518]
	NP1	FPR gr_1	0.15	0.15	0.15	0.15
		$\overline{FMI}[SD]$	0.9598[0.1075]	0.6441[0.1089]	0.9852[0.0767]	0.8521[0.0415]
	NP2H$_k$	FPR gr_1	0.15	0.15	0.15	0.15
		$\overline{FMI}[SD]$	0.9558[0.1072]	0.6153[0.1073]	0.9835[0.0808]	0.8312[0.0772]
	TCLUST	FPR gr_1	0.15	0.15	0.15	0.15
		$\overline{FMI}[SD]$	1[0]	1[0]	1[0]	1[0]

Tabella 4.12: Risultati relativi allo scenario 6 in presenza di contaminazione dei campioni. Cfr. Tab 4.2.

			n=100		n=500	
			d=2	d=5	d=2	d=5
$\alpha = 0.1$	NP2	FPR gr_1	0.0236	0.0129	0.0226	0.0118
		TPR	0.7880	0.8840	0.7968	0.8936
		$\overline{FMI}[SD]$	0.9754[0.0299]	0.7629[0.0771]	0.9944[0.0106]	0.8890[0.0176]
	NP1	FPR gr_1	0.022	0.0127	0.0209	0.0158
		TPR	0.802	0.8860	0.8116	0.8580
		$\overline{FMI}[SD]$	0.9760[0.0294]	0.7585[0.0774]	0.9944[0.0104]	0.8914[0.0168]
	NP2H$_k$	FPR gr_1	0.0238	0.016	0.0209	0.0180
		TPR	0.7860	0.856	0.8116	0.8384
		$\overline{FMI}[SD]$	0.9710[0.0364]	0.7567[0.0756]	0.9924[0.0112]	0.8861[0.0192]
	TCLUST	FPR gr_1	0.0216	0.0113	0.0145	0.0111
		TPR	0.8060	0.8980	0.8628	0.9004
		$\overline{FMI}[SD]$	1[0]	1[0]	1[0]	1[0]
$\alpha = 0.15$	NP2	FPR gr_1	0.0613	0.0576	0.0619	0.0579
		TPR	0.9480	0.9820	0.9432	0.9792
		$\overline{FMI}[SD]$	0.9654[0.0683]	0.6773[0.0961]	0.9965[0.0123]	0.8609[0.0234]
	NP1	FPR gr_1	0.0613	0.0587	0.0620	0.0578
		TPR	0.9480	0.9720	0.9424	0.9796
		$\overline{FMI}[SD]$	0.9656[0.0681]	0.6801[0.0977]	0.9965[0.0123]	0.8600[0.0230]
	NP2H$_k$	FPR gr_1	0.0631	0.0576	0.0615	0.0582
		TPR	0.9320	0.9820	0.9464	0.9764
		$\overline{FMI}[SD]$	0.9747[0.0467]	0.6748[0.0986]	0.9954[0.0144]	0.8699[0.0243]
	TCLUST	FPR gr_1	0.062	0.0578	0.0626	0.0581
		TPR	0.942	0.9800	0.9368	0.9768
		$\overline{FMI}[SD]$	1[0]	1[0]	1[0]	1[0]

Bibliografia

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

Bouveyron C.; Celeux G.; Murphy T.; Raftery A. E. (2019). *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press.

Carmichael J. W.; Julius R. S. (1968). Finding natual clusters. *Systematic Biology*, **17**, 144–150.

Celeux G.; Govaert G. (1992). A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, **14**(3), 315–332.

Chacón J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science*, **30**(4), 518–532.

Cuesta-Albertos J. A.; Gordaliza A.; Matrán C. (1997). Trimmed k-means: An attempt to robustify quantizers. *The Annals of Statistics*, **25**(2), 553–576.

Dempster A. P.; Laird N. M.; Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Duong T. (2022). *ks: Kernel Smoothing*. R package version 1.13.5.

Farcomeni A.; Greco L. (2016). *Robust methods for data reduction*. CRC press.

Fowlkes E. B.; Mallows C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American statistical association*, **78**(383), 553–569.

Fritz H.; Garcia-Escudero L.; Mayo-Iscar A. (2012). tclust: An R package for a trimming approach to cluster analysis. *Journal of Statistical Software*, **47**(12), 1–26.

Fukunaga K.; Hostetler L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, **21**, 32–40.

Gallegos M.; Ritter G. (2009). Trimming algorithms for clustering contaminated grouped data and their robustness. *Advances in Data Analysis and Classification*, **3**(2), 135–167.

García-Escudero, L.and Gordaliza A.; Matrán C.; Mayo-Iscar A. (2010). A review of robust clustering methods. *Advances in Data Analysis and Classification*, **4**(2), 89–109.

García-Escudero L. A.; Gordaliza A.; Matrán C.; Mayo-Iscar A. (2008). A general trimming approach to robust cluster analysis. *The Annals of Statistics*, **36**(3), 1324–1345.

Gather U.; Kuhnt S.; Pawlitschko J. (2002). Concepts of outlyingness for various data structures. *Industrial Mathematics and Statistics*, pp. 545–585.

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

Hubert L.; Arabie P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Kaufman L.; Rousseeuw P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Li J.; Ray S.; Lindsay G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, **8**, 1687–1723.

Matsumoto Y. (2002). *An introduction to Morse theory*, volume 208. American Mathematical Soc.

Menardi G. (2015). A review on modal clustering. *International Statistical Review*, **84**, 413–433.

Neykov N.; Filzmoser P.; Dimova R.; Neytchev P. (2007). Robust fitting of mixtures using the trimmed likelihood estimator. *Computational Statistics & Data Analysis*, **52**(1), 299–308.

Parzen E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, **33**, 1065–1076.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ronchetti E.; Huber P. J. (2009). *Robust statistics*. John Wiley & Sons.

Scott D.; Sain S. (2005). Multidimensional density estimation. *Handbook of Statistics*, **24**, 229–261.

Shapira L.; Avidan S.; Shamir A. (2009). Mode-detection via median-shift. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 1909–1916. IEEE.

Sheikh Y. A.; Khan E. A.; Kanade T. (2007). Mode-seeking by medoidshifts. In *2007 IEEE 11th international conference on computer vision*, pp. 1–8. IEEE.

Stuetzle W.; Nugent R. (2010). A generalized single linkage method for estimating the cluster tree of a density. *Journal of Computational and Graphical Statistics*, **19**, 397–418.

Wand M. P.; Jones M. C. (1994). *Kernel smoothing*. Chapman and Hall/CRC.

Yang J.; Rahardja S.; Fränti P. (2021). Mean-shift outlier detection and filtering. *Pattern Recognition*, **115**, 107874.