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

Statistica per l'Economia e l'Impresa

Modelli a mistura finita per il clustering: un confronto tra l'approccio parametrico e non parametrico

Relatore: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureanda: Giorgia Caicchiolo
Matricola n. 2045146

Anno Accademico 2023/2024

Indice

Introduzione

L'insieme delle tecniche per suddividere un insieme di dati in gruppi prende il nome di *cluster analysis* o *clustering*.

L'obiettivo del *clustering* è trovare una suddivisione dei dati, una *partizione*, tale per cui i dati siano il più possibile omogenei all'interno del loro gruppo, e quanto più possibile diversi tra gruppi distinti.

Suddividere un *dataset* in gruppi omogenei offre diversi vantaggi in moltissime applicazioni. Nell'analisi di mercato, ad esempio, viene utilizzato per segmentare la clientela in base ai comportamenti di acquisto, permettendo di personalizzare offerte e campagne pubblicitarie. In biologia, nello studio delle relazioni evolutive, è fondamentale per individuare sequenze genetiche simili. Nel *data mining*, il *clustering* è una tecnica essenziale per scoprire *pattern* e relazioni nascoste all'interno di grandi quantità di dati.

I più classici metodi di *clustering* sono stati introdotti basandosi sul concetto di distanza tra le osservazioni: due osservazioni si troveranno nello stesso *cluster* se risulteranno sufficientemente vicine rispetto a una qualche misura di distanza. Tra questi vi sono i *metodi di partizionamento* e i *metodi gerarchici*.

I metodi di partizionamento classificano i dati in gruppi a partire da una suddivisione iniziale e applicano una tecnica di riallocazione iterativa, che migliora la partizione spostando gli oggetti da un gruppo ad un altro. Un classico esempio è il *metodo delle k-medie*, il quale sposta le osservazioni da un gruppo ad un altro sulla base della loro distanza dai centroidi (medie) dei gruppi: ciascuna osservazione si troverà nel gruppo con centroide ad essa più vicino.

I metodi gerarchici, invece, creano una serie di scomposizioni nidificate degli oggetti, iniziando con una partizione dei dati che contiene tutte le osservazioni in uno stesso gruppo (approccio divisivo) o ciascuna osservazione in un gruppo distinto (approccio agglomerativo) e procedono con suddivisioni o fusioni successive dei gruppi fino a che queste sono possibili, o fino al verificarsi di una qualche condizione di terminazione. Ad ogni iterazione la suddivisione o fusione da effettuare sarà scelta minimizzando un particolare criterio di distanza ed una volta che un'iterazione è stata fatta, non può più essere disfatta. Ciascun criterio definisce uno specifico metodo gerarchico, tra ci possiamo citare il *metodo del legame singolo*, il *metodo del legame completo* e il *metodo del legame medio*.

Questi metodi presentano una certa rigidità nell'impostazione che, se da un lato porta ad una maggiore facilità nella loro comprensione e implementazione, comporta allo stesso tempo anche diversi limiti.

Innanzitutto è necessaria una scelta a priori del numero di *cluster* da usare, spesso lasciata a colui che analizza i risultati, sulla base di considerazioni sulla natura dei

dati o di tipo euristico e grafico. Sono inoltre fondamentali la scelta della partizione iniziale dei dati, nei metodi di partizionamento, e la questione su come gestire i valori anomali (o *outlier*), entrambi aspetti che possono modificare notevolmente i risultati della procedura. Un'altra limitazione importante è la tendenza a raggruppare i dati in *cluster* di una specifica forma, dipendente dal metodo utilizzato. I metodi di partizionamento, ad esempio, cercheranno sempre di interpolare i dati in gruppi di forma sferica, richiedendo la necessità di estensioni per individuare gruppi di altre forme. L'ultima questione, possibilmente la più rilevante, è data dal fatto che questi metodi non siano espressi in termini statistici rigorosi, e di conseguenza non sia possibile nessun processo inferenziale. La trattazione esaustiva di questi metodi non è lo scopo di questo lavoro, per questo si rimanda, ad esempio, a Hennig *et al.* (2016),

In tempi più recenti si è sviluppato un approccio alternativo all'analisi dei gruppi, basato su un approccio statistico più rigoroso. La formulazione del problema di *clustering* segue in questo caso la logica classica dell'inferenza statistica, e si basa sull'idea che i dati a disposizione siano realizzazioni di una variabile casuale descritta da una certa funzione di probabilità o densità. I gruppi sono associati a specifiche caratteristiche di tale funzione, e possono essere individuati mediante una stima della distribuzione. La strategia più naturale per inquadrare il problema in questi termini è quella di assumere quale processo generatore dei dati un modello a mistura finita in cui ciascuna componente descrive uno specifico gruppo.

Un approccio di questo tipo ha diversi vantaggi. Inquadra il problema in un contesto probabilistico, in cui il concetto di *cluster* è definito inequivocabilmente dalla sua distribuzione specifica, sostituisce il problema di trovare il numero di *cluster* con un problema di selezione del modello e fornisce un approccio sistematico per trattare gli *outlier*, consentendo di tenere conto dei valori anomali espandendo il modello.

In questa relazione si approfondirà lo studio di modelli a mistura finita per il raggruppamento di dati continui, discutendone le varianti parametrica e non parametrica.

La trattazione si sviluppa come segue.

Nel primo capitolo viene introdotto in modo formale l'approccio al *clustering* basato su modello. Successivamente, vengono descritti i modelli a mistura finita e come questi possano essere utilizzati in un problema di *clustering*. Viene poi illustrato un algoritmo di stima di massima verosimiglianza di modelli mistura parametrici e spiegato come questo possa essere utilizzato per fare *clustering*.

Nel secondo capitolo viene introdotta la stima non parametrica della densità, concentrandosi in particolare sullo stimatore *Kernel*. Vengono poi descritti i modelli mistura non parametrici e la stima di densità non parametrica ed è infine presentato un algoritmo che utilizza lo stimatore *kernel* per la stima di modelli mistura non parametrici.

Nel terzo capitolo vengono messi a confronto i due approcci descritti in precedenza attraverso uno studio di simulazione, volto a valutare le prestazioni dei due algoritmi in relazione a diversi possibili scenari presentati dai dati.

Si conclude infine con alcune considerazioni di carattere generale, con l'obiettivo di evidenziare vantaggi e limiti di ciascuno degli approcci considerati e mettendo in luce eventuali possibili approfondimenti.

Capitolo 1

Misure parametriche per il *clustering*

1.1 Modelli mistura parametrici

Finora è stato evidenziato come i metodi di raggruppamento basati sulla distanza lascino diverse questioni aperte, molte delle quali possano essere affrontate adottando una formulazione più rigorosa del problema, e come l'approccio del *clustering* basato su modelli offra la possibilità di farlo.

Associare un modello distributivo a ciascun gruppo conferisce al problema di *clustering* una formalizzazione che lo trasformi in un comune e ben definito problema di inferenza statistica, cosa non possibile con gli approcci basati sulla distanza, in quanto mancanti di proprietà statistiche: i gruppi possono essere identificati tramite un processo di stima, e la qualità del risultato può essere valutata attraverso metodi di confronto tra modelli, utilizzando criteri di bontà di adattamento e di selezione del modello.

Formalmente, il problema di *clustering* basato su modello può essere definito come segue. Sia $\mathbf{Y} \in S_Y \subset \mathbf{R}^d$ una variabile casuale d-dimensionale, osservata su un campione di soggetti $\mathbf{y}_1, \ldots, \mathbf{y}_n$ che rappresentano i dati a disposizione. Si definisca inoltre una variabile casuale $\mathbf{Z} = (Z_1, \ldots, Z_K)$ non osservabile, che descrive l'appartenenza di un soggetto a uno tra i K possibili gruppi ed assume, pertanto, distribuzione multinomiale

$$\mathbf{Z} \sim Bin_K(1, \pi)$$

dove $\pi = (\pi_1, \ldots, \pi_K)$, $\pi_k > 0$, $\sum_{k=1}^{K} \pi_k = 1$ è un vettore che definisce le probabilità *a priori* di appartenenza ai diversi gruppi.

Si assume che, condizionatamente al gruppo di appartenenza, la distribuzione di $\mathbf{Y}$ segua un modello statistico caratterizzato da uno specifico vettore di parametri

$$\mathbf{Y}|Z_k = 1 \sim f_k(\mathbf{y}; \theta_k) \ , \ k = 1, \ldots, K \tag{1.1}$$

È possibile allora ricavare la funzione di densità o probabilità marginale

$$f(\mathbf{y}) = \sum_{k=1}^{K} P(Z_k = 1) f_k(\mathbf{y}; \theta_k) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}; \theta_k) \tag{1.2}$$

che rappresenta l'equazione di un modello a mistura finita.

I modelli a mistura finita sono estremamente flessibili, poiché in grado di descrivere un'ampia gamma di fenomeni e tipologie di dati differenti a seconda del modello parametrico specificato per un singolo gruppo.

Va comunque notato come questa maggiore versatilità comporti una complessità maggiore rispetto altri metodi, costituendo un possibile limite di questo approccio.

Un modello mistura particolarmente diffuso per dati continui, assume che le singole componenti abbiano densità Normale multivariata e

$$\phi(\mathbf{y}; \mu_k, \Sigma_k) = (2\pi)^{-\frac{p}{2}} |\Sigma_k|^{-1/2} e^{-\frac{1}{2}(\mathbf{y}-\mu_k)^T \Sigma_i^{-1} (\mathbf{y}-\mu_k)}.$$

Di conseguenza la (1.2) diventa

$$f(\mathbf{y}) = \sum_{k=1}^{K} \pi_k \phi(\mathbf{y}_i | \mu_k, \Sigma_k).$$

I gruppi trovati sono centrati nella media μ_k ed hanno forma ellissoidale, caratterizzata dalla matrice di varianza e covarianza Σ_k.

Una mistura di distribuzioni normali è pertanto caratterizzata dal parametri $\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) \in \Theta \subset \mathbf{R}^p$ con $p = k + dk + \frac{d(d+1)k}{2}$. A dispetto della sue semplicità concettuale, quindi, il numero di parametri da stimare può essere anche elevato. Per questo, è comune porre delle restrizioni sulle matrici Σ:

1. $\Sigma_1 = \ldots = \Sigma_K = \sigma^2 I$
2. $\Sigma_1 = \ldots = \Sigma_K = Diag(\sigma_1^2, \ldots, \sigma_K^2)$
3. $\Sigma_1 = \ldots = \Sigma_K = \Sigma$

In alternativa le matrici Σ_k possono espresse attraverso una parametrizzazione più parsimoniosa che sfrutta la loro scomposizione spettrale:

$$\Sigma_k = \lambda_k D_k A_k D_k'.$$

In questa parametrizzazione, proposta da Banfield e Raftery (1993) e successivamente ripresa da Celeux e Govaert (1993), λ_k determina il volume del gruppo, D_k il suo orientamento e A_k la sua forma. In questo modo otteniamo modelli parsimoniosi e facilmente interpretabili, adatti a descrivere svariate situazioni.

1.2 Stima di un modello mistura parametrico

1.2.1 Verosimiglianza completa e osservata

Definire il problema di *clustering* mediante la specificazione di un modello statistico parametrico, consente l'identificazione dei gruppi mediante la stima dei parametri dello stesso.

Sia allora $\mathbf{y}_1, \ldots, \mathbf{y}_n$ un campione di osservazioni da $\mathbf{Y}$ e $\mathbf{z}_1, \ldots, \mathbf{z}_n$ l'n-pla di vettori non osservati, e quindi latenti, che definisce l'appartenenza di ciascun soggetto ad un gruppo.

La verosimiglianza, cosiddetta *osservata*, associata al modello (1.2) è

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\theta)$$

e la log-verosimiglianza

$$\begin{aligned} l(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K; \mathbf{y}_1, \ldots, \mathbf{y}_n) &= \sum_{i=1}^{n} \log f(\mathbf{y}_i; \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\ &= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \theta_k) \end{aligned}$$

da cui si possono ricavare le equazioni di verosimiglianza

$$\begin{aligned} \frac{\partial l(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K; \mathbf{y}_1, \ldots, \mathbf{y}_n)}{\partial \theta_k} &= \sum_{i=1}^{n} \frac{\partial \log \sum_{k=1}^{K} \pi_k f(\mathbf{y}; \theta_k)}{\partial \theta_k} \\ &= \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \theta_k)} \pi_k \frac{\partial f(\mathbf{y}_i; \theta_k)}{\partial \theta_k} \\ &= \sum_{i=1}^{n} \frac{\pi_k f(\mathbf{y}_i; \theta_k)}{\sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \theta_k)} \frac{1}{f(\mathbf{y}_i; \theta_k)} \frac{\partial f(\mathbf{y}_i; \theta_k)}{\partial \theta_k} \\ &= \sum_{i=1}^{n} \frac{\pi_k f(\mathbf{y}_i; \theta_k)}{\sum_{k_1}^{K} \pi_k f(\mathbf{y}_i; \theta_k)} \frac{\partial \log f(\mathbf{y}_i; \theta_k)}{\partial \theta_k} \\ &= \sum_{i=1}^{n} \pi_{i,k} \frac{\partial \log f(\mathbf{y}_i; \theta_k)}{\partial \theta_k} \end{aligned}$$

La forma delle equazioni di verosimiglianza suggerisce un'interpretazione interessante: massimizzare la log-verosimiglianza rispetto ai parametri θ_k corrisponde a massimizzare la log-verosimiglianza dei contributi individuali pesati rispetto ai $\tau_{i,k}$, dove

$$\begin{aligned} \tau_{i,k} &= \frac{\pi_k f(\mathbf{y}_i; \theta_k)}{\sum_{k=1}^{K} \pi_k f(\mathbf{y}_i; \theta_k)} = \frac{P(Z_{i,k} = 1) f(\mathbf{y}_i | Z_{i,k} = 1)}{f(\mathbf{y}_i)} \\ &= \frac{P(Z_{i,k} = 1, \mathbf{y}_i)}{f(\mathbf{y}_i)} = P(Z_{i,k} = 1 | \mathbf{y}_i) \end{aligned}$$

è la probabilità *a posteriori* che l'individuo i appartenga al gruppo k.

Il problema è che, a dispetto della agevole interpretazione, i $\tau_{i,k}$ dipendono essi stessi dai parametri θ_k. Questo, insieme alla presenza del logaritmo all'interno della sommatoria, rende impossibile risolvere le equazioni di verosimiglianza analiticamente, richiedendo l'utilizzo di un metodo numerico.

1.2.2 Algoritmo EM

Un algoritmo utile a risolvere elegantemente il problema dell'ottimizzazione numerica è noto con il nome di EM (*Expectation-maximization*).

L'algoritmo EM (Dempster *et al.*, 1977) è stato sviluppato per massimizzare una verosimiglianza in caso di dati mancanti.

Siano $\mathbf{y}$ i dati osservati e $\mathbf{z}$ i dati mancanti, sia θ il parametro di interesse e sia $f(\mathbf{y}, \mathbf{z}|\theta) = f(\mathbf{z}|\mathbf{y}, \theta) f(\mathbf{y}|\theta)$.

Si definiscano

$$l(\theta|\mathbf{y}, \mathbf{z}) = \log f(\mathbf{y}, \mathbf{z}|\theta) \text{ verosimiglianza completa} \tag{1.3}$$

$$l(\theta|\mathbf{y}) = \log f(\mathbf{y}|\theta) \text{ verosimiglianza osservata}$$

L'idea è di massimizzare, anziché la (1.3), il suo valore atteso condizionatamente a $\mathbf{y}$.

A questo scopo si ripetono, iterativamente ($t = 1, \ldots, T$), 2 passi:

- **E-step** (*expectation step*): fissato $\theta = \theta^{(t)}$ si determina

$$E_Z[\log f(\mathbf{y}, \mathbf{z}|\theta^{(t)})|\mathbf{y}, \theta^{(t)}]$$

- **M-step** (*maximization step*): al variare di θ si massimizza il valore atteso calcolato al passo E e si determina:

$$\theta^{(t+1)} = argmax_\theta \; E_Z[\log f(\mathbf{y}, \mathbf{z}|\theta^{(t)})|\mathbf{y}, \theta^{(t)}]$$

Fino a convergenza dell'algoritmo. Poiché si massimizza ad ogni iterazione un un minorante della log-verosimiglianza, è garantita anche la crescita di quest'ultima, e la convergenza ad un massimo locale.

Nel caso di un modello mistura i dati mancanti sono i vettori $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,K})$, $i = 1, \ldots, n$, la log-verosimiglianza osservata è la log-verosimiglianza di un modello mistura $l(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K|\mathbf{y}_1, \ldots, \mathbf{y}_n) = \sum_{i=1}^n \log \sum_{k=1}^K \pi_i f(\mathbf{y}_i; \theta_k)$ e la log-verosimiglianza completa si costruisce assumendo osservato il gruppo di appartenenza

$$\begin{aligned}
l(\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K|\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n) &= \log f(\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{z}_1, \ldots, \mathbf{z}_n|\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\
&= \log \prod_{i=1}^n f(\mathbf{y}_i, \mathbf{z}_i|\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\
&= \sum_{i=1}^n \log f(\mathbf{z}_i|\pi_1, \ldots, \pi_K) f(\mathbf{y}_i|\mathbf{z}_i, \theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K) \\
&= \sum_{i=1}^n \prod_{k=1}^K \pi_k^{z_{i,k}} \prod_{k=1}^K f(\mathbf{y}_i|\theta_k)^{z_{i,k}} \\
&= \sum_{i=1}^n \log \prod_{k=1}^K (\pi_k f(\mathbf{y}_i|\theta_k))^{z_{i,j}} \\
&= \sum_{i=1}^n \sum_{k=1}^K \log(\pi_k f(\mathbf{y}_i|\theta_k))^{z_{i,k}} \\
&= \sum_{i=1}^n \sum_{k=1}^K z_{i,k} \log(\pi_k f(\mathbf{y}_i|\theta_k)).
\end{aligned}$$

L'algoritmo determina

$$\hat{\theta} = argmax_\theta E_z(\log f(\mathbf{y}, \mathbf{z}|\theta)|\mathbf{y})$$

alternando per $t = 1, \ldots, T$:

- **E-step**: Fissati i parametri, determina

$$\begin{aligned} E_{\mathbf{z}}(logf(\mathbf{y},\mathbf{z}|\theta)|\mathbf{y}) &= E\left(\sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k}\log(\pi_k f(\mathbf{y}_i|\theta_k))|\mathbf{y}\right) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k f(\mathbf{y}_i|\theta_k))E(z_{i,k}|\mathbf{y}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k f(\mathbf{y}_i|\theta_k))f(z_{i,k}|\mathbf{y}) \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k f(\mathbf{y}_i|\theta_k))\tau_{i,k} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k^{(t)} f(\mathbf{y}_i|\theta_k^{(t)}))\tau_{i,k}^{(t)} \end{aligned}$$

  e

$$\tau_{i,k}^{(t)} = \frac{\pi_k f(\mathbf{y}_i|\theta_k^{(t)})}{\sum_{k=1}^{K}\pi_k^{(t)} f(\mathbf{y}_i|\theta_k^{(t)})} \tag{1.4}$$

- **M-step**: Al variare dei parametri (fissati i $\tau_{i,k}^{(t)}$) determina

$$\begin{aligned} \hat{\theta}^{(t+1)}, \hat{\pi}^{(t+1)} &= argmax_{\theta,\pi} E_z(\log f(\mathbf{y},\mathbf{z}|\theta)|\mathbf{y}) \\ &= argmax_{\theta,\pi}\sum_{i=1}^{n}\sum_{k=1}^{K}\log(\pi_k f(\mathbf{y}_i|\theta_k))\tau_{i,k}^{(t)} \end{aligned}$$

  Ipotizzando, ad esempio, che ciascuna componente della mistura segua una distribuzione normale, è possibile ricavare in forma esplicita l'espressione di stima delle medie:

$$\mu_k^{(t)} = \frac{\sum_{i=1}^{n}\tau_{i,k}^{(t)}\mathbf{y}_i}{\sum_{i=1}^{n}\tau_{i,k}^{(t)}}$$

  mentre la stima della matrice di varianza e covarianza dipende dalla sua parametrizzazione. Per il modello più generale (VVV) è data da:

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{n}\tau_{i,k}^{(t)}(\mathbf{y}_i-\mu_k^{(t)})(\mathbf{y}_i-\mu_k^{(t)})^T}{\sum_{i=1}^{n}\tau_{i,k}^{(t)}}$$

Una volta arrivato a convergenza, l'algoritmo ha prodotto una stima del vettore delle probabilità τ e si può associare ad ogni osservazione il gruppo con probabilità massima:

$$\hat{z}_{i,k} = \begin{cases} 1\ se\ i = argmax_k\ \hat{\tau}_{i,k} \\ 0\ altrimenti \end{cases}$$

Va notato come l'algoritmo produca le stime delle probabilità di appartenenza di un'osservazione a ciascun gruppo e non una partizione del dataset come gli algoritmi basati sul concetto di distanza.

Avere una stima della probabilità di appartenenza a ciascun gruppo è un ulteriore vantaggio dell'utilizzo di modelli mistura, in quanto permette di dare una stima dell'incertezza del partizionamento, utile soprattutto per le osservazioni in cui non è evidente il gruppo di appartenenza dalle analisi esplorative.

Inoltre, come è già stato detto in precedenza, un altro vantaggio dell'utilizzo di modelli a mistura finita per il *clustering* è la loro versatilità.

Nel momento in cui si utilizza un algoritmo per affrontare un problema di raggruppamento tramite questo approccio, è necessario che la stessa flessibilità si trovi anche nell'algoritmo in questione.

Per questo motivo sono state proposte diverse modifiche alla struttura originale dell'algoritmo EM, che prevedono, ad esempio, la gestione degli *outlier* o l'utilizzo dell'algoritmo in un contesto supervisionato o semi-supervisionato.

Per tali approfondimenti si veda Bouveyron *et al.* (2019).

1.3 Discussione

Si distinguono di seguito alcuni aspetti di approfondimento legati all'uso di modelli misture nel *clustering*.

- *Scelta dei valori iniziali*: La funzione di verosimiglianza per modelli mistura solitamente non è convessa, ha più massimi, e di conseguenza le equazioni di verosimiglianza hanno più soluzioni. Per questo motivo è necessario far partire l'algoritmo da più partizioni iniziali dei dati. Coleman *et al.* (1999) hanno proposto alcuni metodi di divisione del *dataset* quando non è data una misura di distanza delle osservazioni a priori, sfruttando procedure di *clustering* basate sulla distanza, come il metodo *k-means*. Un'altra possibilità è data da *clustering* gerarchici basati su modello (Banfield e Raftery, 1993).

- *Scelta del numero di gruppi*: La scelta del numero di gruppi può essere interpretata come una scelta del modello. La stima dell'ordine di un modello mistura è spesso fatta utilizzando criteri di informazione che penalizzano la funzione di verosimiglianza per un fattore proporzionale al numero di termini. Un criterio spesso utilizzato è il BIC (*Bayesian Information Criterion*), dato da

  $$-2 \log L(\hat{\theta}) + d \log n$$

  con d numero di parametri del modello. Sotto alcune condizioni, è dimostrato che il BIC si comporta in maniera consistete nel scegliere il giusto numero di componenti della mistura (Keribin, 2000), ma tende a sovrastimare il numero di gruppi per dare una buona approssimazione del modello (Biernacki *et al.*, 2000).

  Un'alternativa è l'ICL (*Integrated Classification Criterion*):

  $$-2 \log L(\hat{\theta}) + d \log n + EN(\hat{\tau})$$

  dove $EN(\hat{\tau})$ è l'entropia calcolata su (1.4).

  Ulteriori strategie possono fondarsi su un *approccio di ricampionamento*, come mostrato da McLachlan (1987), su diversi criteri di informazione, o sulla

convalida incrociata. Tuttavia, a seguito di un confronto tra tali metodi, il BIC sembra essere il criterio che offre i risultati più soddisfacenti (Biernacki e Govaert, 1999).

- *Scelta della distribuzione*: Va infine notato come la maggiore versatilità del modello mistura comporti una complessità maggiore rispetto ad altri metodi. Talvolta può non essere ovvio come scegliere il modello per i gruppi, oppure quest'ultimi potrebbero avere forme diverse. Inoltre, la sua efficacia dipende dalla validità delle ipotesi fatte sulla distribuzione scelta, e può portare a forti distorsioni dei risultati nel caso le ipotesi fatte sui dati non fosse in realtà rispettate.

  Nel capitolo che segue, si presenterà l'approccio *non parametrico* che, non ponendo alcun vincolo sulla funzione di densità, è particolarmente utile per dati che durante l'analisi esplorativa non presentino caratteristiche riconducibili a distribuzioni note.

Capitolo 2

Misure non parametriche per il *clustering*

2.1 Stima non parametrica di una funzione di densità

2.1.1 Introduzione ai metodi non parametrici

L'impiego di modelli mistura parametrici per il *clustering* offre diversi vantaggi, tra cui una buona efficienza computazionale e una descrizione sintetica della distribuzione attraverso pochi parametri, rendendola più semplice da interpretare e comunicare. Inoltre, è particolarmente utile con campioni di piccole dimensioni, poiché le ipotesi forti sulla forma della distribuzione permettono di ottenere stime accurate anche con meno dati. Tuttavia, l'efficacia di tale approccio dipende dalla validità delle ipotesi fatte sulla distribuzione dei dati, e può portare a forti distorsioni nei risultati nel caso tali ipotesi non fossero in realtà rispettate.

Per far fronte a tali limiti, un approccio possibile consiste nel rilassare le ipotesi distributive sul modello statistico volto a descrivere il gruppo, ricorrendo ad un approccio non parametrico.

Nel seguito si presenta un'introduzione a tali metodi, concentrando l'attenzione al metodo *Kernel*. Si rimanda a Silverman (1986) per approfondimenti.

Il primo metodo introdotto e tuttora più largamente utilizzato è l'istogramma.

Sia $y_1, \ldots, y_n$ un campione di osservazioni da $Y \in S_Y \subset \mathbf{R}$ di cui non si hanno informazioni sulla distribuzione. Per costruire l'istogramma divideremo S_Y in intervalli

$$[y_0 + mh; y_0 + (m+1)h)$$

con y_0 l'origine, h la larghezza di ciascun intervallo e m interi positivi e negativi. Successivamente definiamo l'istogramma come

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} I_{m_y}(y_i)$$

con I_{m_y} funzione indicatrice dell'intervallo $[y_0 + m_y h; y_0 + (m_y + 1)h)$ contenente y. L'istogramma può fornire una prima stima della densità arbitrariamente precisa in

base al posizionamento e alla larghezza degli intervalli. Risulta quindi un metodo estremamente utile in ambito univariato per l'analisi esplorativa e la presentazione grafica dei dati, in quanto offre una rappresentazione dei dati facilmente intuibile anche per chi non opera in ambito statistico.

Passando all'ambito multivariato, cominciano a presentarsi una serie di difficoltà legate, ad esempio, all'origine degli assi, alla loro direzione e all'interpretazione stessa del grafico, che può risultare difficoltosa. Anche la sua discontinuità comporta grandi difficoltà nel caso siano necessarie delle derivate della funzione stimata. Può infine essere significativamente migliorato dal punto di vista della descrizione matematica dell'accuratezza, un limite che si traduce in un uso inefficiente dei dati quando gli istogrammi vengono impiegati come stime di densità in procedure come l'analisi dei *cluster* e l'analisi discriminante non parametrica.

Lo *stimatore naive* parte dalla definizione della funzione di densità

$$f(y) = \lim_{h \to 0} \frac{1}{2h} P(y - h < Y < y + h)$$

e stima $P(y-h < Y < y+h)$ attraverso la proporzione di osservazioni nell'intervallo $(y - h; y + h)$. Lo stimatore può essere scritto come

$$\hat{f}(y) = \frac{1}{2hn} \sum_{i=1}^{n} I_y(y_i) \tag{2.1}$$

con I_y funzione indicatrice dell'intervallo $(y - h; y + h)$.

Definendo dei pesi w come

$$w(y) = \begin{cases} \frac{1}{2} \text{ se } |y| < 1 \\ 0 \text{ altrimenti} \end{cases}$$

è possibile riscrivere la (2.1) come

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} w\left(\frac{y - y_i}{h}\right).$$

Lo stimatore *naive* può essere interpretato come un istogramma in cui ciascun intervallo, di larghezza $2h$, è centrato in un'osservazione del campione, slegando la stima della densità dal posizionamento degli intervalli, ma mantenendola dipendente dalla loro larghezza.

È importante notare che anche questo stimatore, così come l'istogramma, è discontinuo, comportando gli stessi limiti discussi in precedenza.

Una possibile soluzione a questo problema è sostituire i pesi w con delle funzioni di densità. A partire da questa idea viene definito lo *stimatore Kernel*, che verrà trattato nel paragrafo successivo.

Altri metodi che seguono invece diverse elaborazioni possono essere il *metodo del vicino più vicino* e i metodi che si basano sulle *serie ortogonali*. Per approfondimenti su questi metodi si veda Silverman (1986).

2.1.2 Lo stimatore *Kernel* di densità

L'idea da cui si parte per definire lo stimatore *kernel* è di costruire una sorta di istogramma lisciato dei dati osservati, "sommando" tra loro più curve, ciascuna centrata in un'osservazione.

Formalmente, lo stimatore *kernel* K con parametro di lisciamento h è definito da

$$\hat{f}(y) = \frac{1}{n}\sum_{i=1}^{n} K_h(y - y_i)$$

con $K_h(y) = \frac{1}{h}K\left(\frac{y}{h}\right)$, K funzione positiva che integra a 1, nota come *nucleo* o *kernel* e $h > 0$ un parametro di lisciamento che governa la forma di $\hat{f}$. Il ruolo di h è quello di determinare quanto "liscia" si presenta la funzione stimata: per valori piccoli di h le curve poste su ciascuna osservazione saranno piccole, e quindi la loro somma poco omogenea (Figura 2.1 a sinistra). Aumentando il valore di h l'ampiezza delle curve aumenta e la funzione risultante sarà sempre più omogenea e meno dettagliata (Figura 2.1 a destra).

Per valutare la bontà della stima è necessario definire una misura di divergenza tra la vera densità da cui sono generati i dati f e la sua stima $\hat{f}$.

Nell'ambito della stima *puntuale*, una naturale misura della discrepanza tra vero valore e valore stimato è data dall'*errore quadratico medio* (o *mean square error*, MSE), definito da

$$MSE_y(\hat{f}) = E[\hat{f}(y) - f(y)]^2 \tag{2.2}$$

che, sfruttando proprietà elementari di media e varianza, può essere riscritto come

$$MSE_y(\hat{f}) = \{E[\hat{f}(y)] - f(y)\}^2 + Var[\hat{f}(y)].$$

Minimizzare l'MSE, come si vorrebbe fare per ottenere un buon stimatore, equivale quindi a trovare un buon compromesso tra varianza e distorsione dello stimatore. La questione del compromesso tra varianza e distorsione nel momento in cui si sceglie uno stimatore è un tema ricorrente in ambito statistico.

Quando si stima una densità, una misura della bontà della stima può essere trovata estendendo la (2.2) ad una stima di misura *globale* attraverso l'*errore quadratico integrato medio* (o *mean integrated square error*, MISE), definito da

$$MISE(\hat{f}) = E\int \{\hat{f}(y) - f(y)\}^2 dy. \tag{2.3}$$

Anche il MISE può essere scomposto in

$$MISE(\hat{f}) = \int \{E[\hat{f}(y)] - f(y)\}^2 dy + \int Var[\hat{f}(y)]dy.$$

Andando a calcolare il valore atteso dello stimatore Kernel

$$E[\hat{f}(y)] = \int \frac{1}{h}k\left(\frac{y - y_i}{h}\right) f(y)dy$$

e la sua varianza

$$Var[\hat{f}(y)] = \frac{1}{h}\int \frac{1}{h^2}k\left(\frac{y - y_i}{h}\right)^2 f(y)dy - \left\{\frac{1}{h}\int k\left(\frac{y - y_i}{h}\right) f(y)dy\right\}^2$$

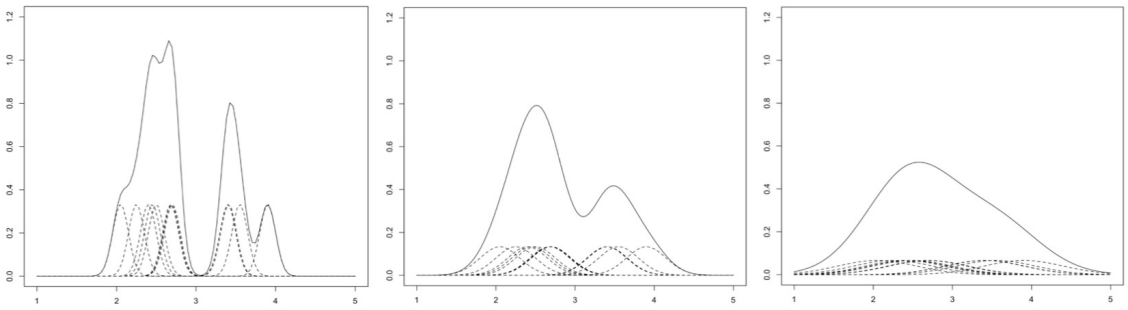

Figura 2.1: Stima di densità *Kernel* con evidenziati i singoli *kernel* centrati nelle osservazioni campionarie e h rispettivamente pari a 0.1, 0.25 e 0.5

possiamo notare come sia l'espressione della distorsione che quella della varianza dipendano direttamente dalla scelta del parametro di lisciamento h. In altre parole, l'accuratezza della stima dipende in modo critico dalla scelta del parametro h, per il ruolo che quest'ultimo ha nel controllare il *trade-off* tra distorsione e varianza della stima.

La scelta di h è innanzitutto influenzata dagli obiettivi per cui si utilizza la stima della densità. Se l'obiettivo dell'analisi, ad esempio, è di tipo esplorativo, come punto di partenza per la costruzione di modelli o la verifica di ipotesi, la scelta del parametro può essere fatta soggettivamente, così come se la stima della densità viene utilizzata per presentare delle conclusioni fatte da analisi precedenti, può essere utile una stima leggermente sotto lisciata e lasciare allo spettatore il compito, più semplice, di lisciare ulteriormente la funzione.

Per procedimenti di stima che vengono eseguiti frequentemente e su grandi *dataset*, invece, è necessario un procedimento automatico di scelta del parametro.

Un primo approccio è dato dalla minimizzazione dell'AMISE (*Approximate mean integrated square error*), una versione approssimata e più facilmente trattabile della (2.3), definito da

$$\frac{1}{4}h^4k_2^2\int f''(y)^2dy + \frac{1}{nh}\int K(y)^2dy$$

con $k_2 = \int y^2K(y)dy$. Il valore di h così trovato, detto *valore ideale* è dato da

$$h_{opt} = k_2^{-2/5}\left\{\int K(y)^2dy\right\}^{1/5}\left\{n\int f''(y)^2dy\right\}^{-1/5}. \tag{2.4}$$

È immediato notare come la (2.4) dipenda dalla densità f ignota. Per ovviare a questo problema, una soluzione è basarsi su una *distribuzione standard di riferimento*, ed andare a sostituzione la derivata seconda della sua funzione di densità a $\int f''(y)^2dy$ nell'espressione (2.4). Tra le distribuzioni più usate vi è la distribuzione normale. È chiaro che la selezione di h sarà appropriata solamente se la densità di riferimento scelta si avvicina alle caratteristiche della vera distribuzione f.

Altri criteri per la scelta di h possono basarsi sul bootstrap, sulla convalida incrociata, sulla funzione di log-verosimiglianza penalizzata o su test grafici. Per una trattazione più esaustiva dei criteri di selezione del parametro di lisciamento si rimanda a Silverman (1986).

È infine necessaria un'ultima osservazione sull'aspetto della complessità computazionale del metodo *kernel*. Così come tutti i metodi di stima non parametrici della densità, lo stimatore *kernel* effettua un numero ingente di stime, che può diventare oneroso anche con calcolatori ad elevate prestazioni, se il campione è di grandi dimensioni. Sono quindi necessari alcuni accorgimenti durante il suo utilizzo, come la scelta di una funzione *kernel* che possa essere calcolata velocemente o lo spostamento di termini moltiplicativi costanti al di fuori della sommatoria nella quale la funzione *kernel* viene calcolata.

In contesto multivariato, la necessità di stime di densità non parametriche per recuperare la struttura nei dati multivariati è, forse, ancora maggiore, poiché la modellazione parametrica è più difficile rispetto al caso univariato.

Nella sua forma più generale, lo stimatore *kernel* d-dimensionale è definito da

$$\hat{f}(\mathbf{y}; H) = \frac{1}{n} \sum_{i=1}^{n} K_H(\mathbf{y} - \mathbf{y}_i)$$

con H matrice di parametri di lisciamento simmetrica definita positiva di dimensione $d \times d$ e $K_H(\mathbf{y}) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}\mathbf{y})$ e K funzione positiva d-dimensionale che integra a 1.

Senza porre restrizioni, la matrice H ha $\frac{1}{2}d(d+1)$ parametri di lisciamento, che possono diventare un numero molto elevato di parametri da scegliere anche con valori di d moderati. Può essere quindi utile sostituire H con sue parametrizzazioni più parsimoniose, come una matrice diagonale $H = diag(h_1^2, \ldots, h_d^2)$. Un'ulteriore semplificazione è porre $H = hI$, con I matrice identità, implicando che le singole superfici curve d-dimensionali poste su ciascuna osservazione abbiano la stessa scala in tutte le dimensioni considerate.

Anche per quanto riguarda la funzione K, ci sono diverse definizioni per estendere la funzione *kernel* univariata al caso multivariato, uno di questi è definire la funzione *kernel* multivariata K come il prodotto di d funzioni *kernel* K univariate

$$K^P(\mathbf{y}) = \prod_{j=1}^{d} k(y_j)$$

che conduce allo stimatore *Kernel* prodotto:

$$\hat{f}(\mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{d} K_h(y_j - y_{i,j}). \tag{2.5}$$

Per ulteriori approfondimenti sull'estensione dello stimatore *kernel* al caso multivariato si veda Wand e Jones (1994).

2.2 Modelli mistura non parametrici

Per far fronte ai limiti discussi in merito all'uso di modelli mistura parametrici per il *clustering*, è ragionevole rilassare le ipotesi parametriche in merito al modello distributivo di ciascun gruppo, e consentire una modellazione più flessibile.

Mantenendo le notazioni introdotte nel capitolo precedente, le equazioni (1.1) e (1.2) diventeranno allora

$$\mathbf{Y}|Z_k = 1 \sim f_k(\mathbf{y}) \ , \ k = 1, \dots, K$$

$$f(\mathbf{y}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y}) \tag{2.6}$$

con $f_k(\mathbf{y})$ funzione di densità del gruppo k selezionata all'interno di una più ampia famiglia *non parametrica* di funzioni, come verrà descritto nel paragrafo che segue.

È necessario, come avviene sempre per modelli non parametrici, introdurre alcune ipotesi sulle densità f_k per garantire l'identificabilità del modello. L'assunzione più comune, nota come *indipendenza condizionale*, è stata proposta da Hall e Zhou (2003a) e assume che, condizionatamente ad ogni gruppo, le componenti di $Y = (y_1, \dots, y_d)$ siano indipendenti tra loro:

$$f_k(\mathbf{y}) = \prod_{j=1}^{d} f_{k,j}(y_j). \tag{2.7}$$

Questa assunzione assicura che il modello sia univocamente determinato da un solo valore del vettore $(\pi_1, \dots, \pi_K)$ e da ciascuna $f_{j,k}$, a meno della permutazione di etichette tra i gruppi.

Sebbene inizialmente possa sembrare un'ipotesi fittizia, è stato dimostrato da Hall e Zhou (2003b) come questa ipotesi regga in sostanzialmente qualsiasi caso in cui $d \geq 3$, a patto che si escludano situazioni in cui vi sia una relazione lineare tra le densità dei singoli gruppi. Inoltre risulta essere un'ipotesi comune, e forse più naturale, in contesti affini ai modelli a mistura finita, come i modelli ad effetti misti, dove si suppone che le osservazioni siano indipendenti condizionatamente al gruppo da cui sono tratte. Per una raccolta di risultati significativi sul tema di identificabilità del modello si rimanda a Hunter (2024).

Il modello che consideriamo diventa quindi

$$f(\mathbf{y}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{d} f_{k,j}(y_j).$$

In letteratura sono molteplici gli studi che affrontano l'utilizzo di modelli mistura non parametrici nel *clustering*.

In quanto segue, adottando l'approccio di Levine *et al.* (2011), le funzioni $f_k(\cdot)$ saranno descritte da uno stimatore *kernel*, che nella formulazione prodotto di componenti, descritta nella (2.5) risulta compatibile con l'ipotesi di indipendenza condizionale (2.7).

2.3 Stima di un modello mistura non parametrico

2.3.1 Limiti dell'algoritmo EM e sue estensioni

In ambito non parametrico, la mancanza di un'espressione esplicita della densità, rende necessaria l'implementazione di un nuovo algoritmo di stima del modello che includa anche un procedimento di stima non parametrico della densità.

L'idea più naturale sarebbe partire dall'algoritmo EM ed estenderlo includendo uno step di stima della densità non parametrico. In letteratura sono diversi i tentativi di seguire questa strada, come l'algoritmo proposto da Benaglia *et al.* (2009a). Gli algoritmi così trovati hanno portato a buoni risultati empirici, ma difficilmente riescono a fornire dei risultati teorici che garantiscano l'incremento del valore della funzione di verosimiglianza ad ogni iterazione e di conseguenza la convergenza della funzione ad un massimo locale, come è invece stato fatto da Dempster *et al.* (1977) con l'algoritmo EM.

Un diverso processo di implementazione di un algoritmo che supera il problema della mancanza di una dimostrazione teorica della convergenza si basa sull'utilizzo di un algoritmo MM (*Majorization-Minimization*).

Gli algoritmi MM sono una famiglia di algoritmi, di cui l'algoritmo EM può essere visto come un caso particolare, che sfruttano la concavità o convessità per massimizzare o minimizzare una funzione obiettivo. Uno dei vantaggi degli algoritmi MM è la loro duplice funzione. Possono essere utilizzati sia per affrontare problemi di minimizzazione di una funzione sia per problemi di massimizzazione. Nel primo caso la strategia sarà di trovare una funzione maggiorante e più facilmente manipolabile della funzione obiettivo e successivamente minimizzarla, mentre nel secondo caso, si troverà una funzione minorante e la si massimizzerà. Un buon algoritmo MM in sostanza sostituisce un problema di ottimizzazione complesso con un altro problema di ottimizzazione più semplice. La maggiore semplicità può essere ottenuta in diversi modi: evitando inversioni di matrici molto grandi, linearizzando un problema di ottimizzazione, separando i parametri da ottimizzare, gestendo con eleganza vincoli di uguaglianza e disuguaglianza o, come per l'algoritmo proposto in seguito, trasformando un problema non differenziabile in un problema liscio. Per approfondimenti sull'intuizione alla base degli algoritmi MM e alcune tecniche di costruzione, si veda Hunter e Lange (2004).

L'algoritmo riportato in questo elaborato fa parte di questa classe di modelli ed è stato proposto da Levine *et al.* (2011). L'idea è di formulare un algoritmo di minimizzazione (*Majorization-Minimization algorithm*) per ottimizzare una funzione obiettivo che, per la sua forma che ricorda una funzione di log-verosimiglianza, viene chiamata *funzione di log-verosimiglianza lisciata*. Questa funzione, trovata attraverso la definizione e successiva applicazione di un operatore di lisciamento non lineare alla funzione di densità del modello mistura non parametrico, può essere interpretata come una misura di distanza tra la vera funzione di densità del modello e una sua versione lisciata non linearmente. Per questo motivo, la sua minimizzazione equivarrà a trovare una stima della densità il più vicina possibile alla vera funzione di densità del modello. Della funzione di log-verosimiglianza lisciata così costruita Levine *et al.* (2011) ne ha provato la monotonicità, garantendo la convergenza delle stime verso un minimo locale.

2.3.2 Algoritmo MM

Per stimare il modello definito dalla (2.6) utilizzando un algoritmo MM dobbiamo innanzitutto definire una funzione da ottimizzare.

Sia f la densità del campione, che ipotizziamo essere definita dalla (2.6), e $K(\cdot)$ uno stimatore *kernel* prodotto definito dalla (2.5), con H matrice diagonale i cui

elementi sulla diagonale sono tutti costanti pari ad h. Utilizziamo K per definire un *operatore di lisciamento* S, da applicare a ciascuna componente f_k di f

$$Sf_k(\mathbf{y}) = \int_{S_Y} K(\mathbf{y} - \mathbf{x}) f_k(\mathbf{x}) d\mathbf{x}$$

ed una sua versione non lineare N

$$Nf_k(\mathbf{y}) = exp\{(S \log f_k)(\mathbf{y})\} = exp \int_{S_Y} K(\mathbf{y} - \mathbf{x}) \log f_k(\mathbf{x}) d\mathbf{x}. \tag{2.8}$$

Sostituendo alle f_k le loro versioni lisciate trovate applicando la (2.8), possiamo trovare una versione lisciata della (2.6)

$$Nf(\mathbf{y}) = \sum_{k=1}^{K} \pi_k Nf_k(\mathbf{y}).$$

Definendo infine la funzione

$$l(\theta) = \int_{S_Y} f(\mathbf{y}) \log \frac{f(\mathbf{y})}{Nf(\mathbf{y})} d\mathbf{y} \tag{2.9}$$

abbiamo trovato la nostra funzione obiettivo, da ottimizzare rispetto a $\theta = (f, \pi)$.

Per la sua forma, che ricorda una funzione di log-verosimiglianza, l viene chiamata *log-verosimiglianza lisciata*. In realtà, si può notare come la (2.9) sia una distanza penalizzata di Kullback-Leibler tra la vera densità del modello mistura e la sua versione lisciata, motivo per cui l'algoritmo MM viene in seguito utilizzato per minimizzarla, e non massimizzarla come è usuale fare con le funzioni di verosimiglianza.

Quando utilizzato per minimizzare una funzione, l'algoritmo MM prevede un primo step di maggiorazione della funzione obiettivo ed un secondo step di minimizzazione della funzione maggiorante.

Per trovare la funzione maggiorante definiamo dei pesi

$$\omega_k = \frac{\pi_k Nf_k(\mathbf{y})}{\sum_{k=1}^{K} \pi_k Nf_k(\mathbf{y})}$$

tali che $\sum_{k=1}^{K} \omega_k = 1$. Si noti che gli ω_k definiscono le probabilità a posteriori di appartenenza ai gruppi nel modello mistura con densità lisciata.

La funzione maggiorante sarà definita come

$$b(\theta) = -\int_{S_Y} f(\mathbf{y}) \sum_{k=1}^{K} \omega_k \log\{\pi_k Nf_k(\mathbf{y})\} d\mathbf{y}.$$

Scelti $\theta^{(0)} = (f^{(0)}, \pi^{(0)})$ i valori iniziali dei parametri, l'algoritmo MM alterna per $t = 1, \ldots, T$:

- ***Majorization step***: Fissati i parametri, calcola per ogni i e k

$$\omega_{i,k}^{(t)} = \frac{\pi_k^{(t)} Nf_k^{(t)}(\mathbf{y}_i)}{\sum_{k=1}^{K} \pi_k^{(t)} Nf_k^{(t)}(\mathbf{y}_i)}$$

- ***Minimization step*** Al variare dei parametri (fissati gli $\omega_{i,k}^{(t)}$) determina

$$\hat{f}_{k,j}^{(t+1)}, \hat{\pi_k}^{(t+1)} = argmax_{f,\pi}\ b(f,\pi)$$

di cui si possono scrivere esplicitamente le espressioni

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\omega_{i,k}^{(t)}$$

$$f_{k,j}^{(t+1)}(y) = \frac{\sum_{i=1}^{n}\omega_{i,k}^{(t)}K(y-y_{i,j})}{\sum_{i=1}^{n}\omega_{i,k}^{(t)}}$$

Una volta arrivato a convergenza, l'algoritmo ha prodotto una stima dei valori della densità calcolata in ogni osservazione del campione, una stima delle proporzioni della mistura π ed una stima delle probabilità a posteriori ω. Per ottenere una partizione dei dati, si associa ogni osservazione al gruppo con probabilità ω_k massima.

2.4 Discussione

Si distinguono di seguito alcuni aspetti legati all'uso di modelli mistura non parametrici nel *clustering*.

- *Scelta della matrice H*: Nell'algoritmo presentato nel paragrafo precedente, lo stimatore *kernel* utilizza una matrice H diagonale, con valori sulla diagonale costanti. Questo equivale ad utilizzare lo stesso parametro di lisciamento h per tutte le dimensioni. Non sempre questa assunzione risulta adatta al campione da analizzare, ma sostituire la matrice H così definita con una parametrizzazione meno restrittiva non permetterebbe di assicurare la convergenza dell'algoritmo presentato ad un minimo locale da un punto di vista teorico. Pertanto, la ricerca di un algoritmo che permetta di scegliere una parametrizzazione della matrice H meno restrittiva da utilizzare nello stimatore *kernel* rimane un argomento di ulteriore indagine. Va inoltre sottolineato che una parametrizzazione della matrice più flessibile comporta un numero maggiore di parametri da stimare, e di conseguenza aggiunge un ulteriore carico computazionale ad un contesto in cui quest'ultimo è già elevato.

- *Maledizione della dimensionalità*: In contesti ad elevata dimensionalità, la sparsità dei dati rende difficile l'utilizzo di stimatori non parametrici, a meno che il campione di osservazioni non abbia una numerosità campionaria molto elevata. Questo fenomeno, noto come *maledizione della dimensionalità*, risulta in una grande difficoltà ad ottenere una buona stima non parametrica in dimensioni elevate (indicativamente superiori a cinque) con campioni di dimensioni ragionevoli, che consentano cioè di mantenere un carico computazionale non troppo elevato. Lo studio di simulazione presentato nel prossimo capitolo, metterà in luce da un punto di vista empirico questo aspetto.

- *Interpretazione del modello stimato*: Entrambi gli algoritmi di stima presentati forniscono una stima delle probabilità a posteriori di appartenenza ai gruppi, utilizzabili per trovare una partizione dei dati. Differiscono invece le informazioni fornite dai due algoritmi sulle distribuzioni dei singoli gruppi. L'algoritmo EM fornisce una stima dei parametri caratterizzanti le componenti della mistura, consentendo una rappresentazione sintetica dei gruppi trovati e di conseguenza la possibilità di interpretarli. L'algoritmo MM, invece, così come tutti gli algoritmi di stima non parametrici, fornisce una stima della densità in ciascuna osservazione, ma nessuna informazione sintetica sulla distribuzione dei gruppi, rendendo molto difficile la loro interpretazione.

In definitiva, l'utilizzo di un modello mistura non parametrico offre una buona alternativa al modello parametrico nel caso in cui le assunzioni necessarie a definire quest'ultimo non sembrino ragionevoli per i dati da analizzare, ma può comportare anche alcuni limiti, e richiede pertanto adeguate considerazioni.

Capitolo 3

Studio di simulazione

3.1 Obiettivi dello studio, descrizione degli scenari e dei metodi

In questo capitolo si riportano i risultati di uno studio di simulazione volto a confrontare empiricamente gli approcci di raggruppamento parametrico e non parametrico esposti nei capitoli precedenti. L'obiettivo è valutare il comportamento di quest'ultimi:

- al variare della numerosità campionaria
- al variare della dimensione delle osservazioni
- al variare della forma dei gruppi

Per svolgere lo studio sono stati generati dei campioni da diversi scenari. In tutti i casi la struttura di gruppo è descritta da una mistura di distribuzioni in cui ciascuna componente rappresenta un gruppo. Il numero di gruppi K è stato tenuto costante e pari a 2 così come sono state tenute costanti le proporzioni di appartenenza ai gruppi π_1 e π_2 pari a $\frac{1}{2}$. Sono state invece fatte variare la numerosità campionaria $n \in \{250, 1000\}$, la dimensione dei dati $d \in \{2, 5\}$ e le distribuzioni dei singoli gruppi f_k, in modo da valutare il comportamento degli algoritmi al variare della forma e della distanza dei singoli gruppi.

In particolare, la distribuzione da cui sono stati generati i dati può essere scritta come $\frac{1}{2}f_1(\cdot;\theta_1) + \frac{1}{2}f_2(\cdot;\theta_2)$ in cui si hanno:

- gruppi ben separati di forma sferica: $f_k \sim N_d(\mu_k, I_d)$, $k = 1, 2$, dove I_d è la matrice di identità di ordine d e μ_1, μ_2 sono due vettori d-dimensionali con distanza pari a 6;
- gruppi non ben separati di forma sferica: $f_k \sim N_d(\mu_k, I_d)$, $k = 1, 2$, dove I_d è la matrice di identità di ordine d e μ_1, μ_2 sono due vettori d-dimensionali con distanza pari a 3;
- gruppi caratterizzati da forte asimmetria: f_1 è la distribuzione di $\mathbf{Y}_1 = e^{\mathbf{X}_1}$, con $\mathbf{X}_1 \sim N_d(\mu_1; I_d)$ e f_2 è la distribuzione di $\mathbf{Y}_2 = -e^{-\mathbf{X}_2}$ con $\mathbf{X}_2 \sim N_d(\mu_2, I_d)$ dove le mode di f_1 ed f_2 hanno distanza pari 4.

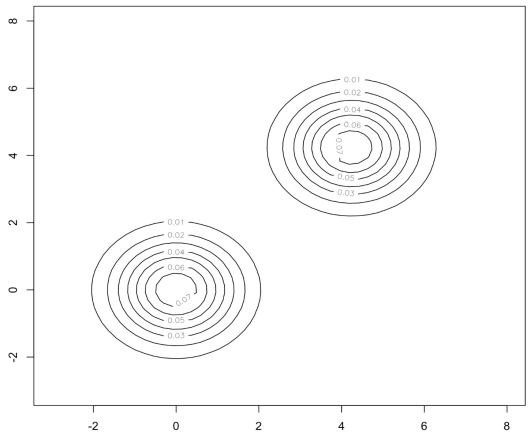

Figura 3.1: Rappresentazione bidimensionale del primo scenario:gruppi ben separati e di forma sferica

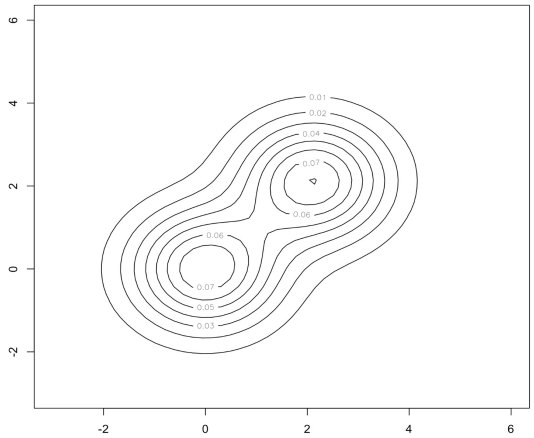

Figura 3.2: Rappresentazione bidimensionale del secondo scenario:gruppi non ben separati e di forma sferica

Le figure 3.1, 3.2 e 3.3 riportano le illustrazioni bidimensionali dei tre scenari considerati.

L'obiettivo di considerare gli scenari descritti è simulare diverse situazioni progressivamente meno adatte all'utilizzo di un algoritmo piuttosto che di un altro.

Lo scenario con gruppi sferici e ben separati rappresenta lo scenario ideale per un problema di raggruppamento, per la facilità con cui i gruppi possono essere distinti.

Il secondo scenario, con gruppi sferici ma non ben separati, vuole rappresentare uno scenario intermedio, in cui la difficoltà dovuta alla vicinanza dei gruppi è mitigata dalla loro forma sferica, quando si utilizzano misture di distribuzioni normali.

L'ultimo scenario rappresenta una situazione particolarmente difficile per la stima di misture di distribuzioni normali, dovuta alla forte asimmetria dei gruppi. Al contrario, per l'algoritmo di stima non parametrico rappresenta il contesto per cui è stato concepito. Per questo motivo ci si aspetta che questo sia lo scenario in cui i risultati varino di più tra i due metodi utilizzati.

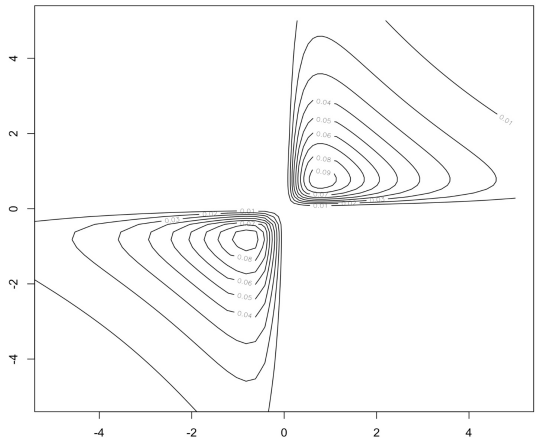

Figura 3.3: Rappresentazione bidimensionale del terzo scenario: gruppi di forma fortemente asimmetrica

Per ogni scenario e valore di n e d sono stati generati 100 campioni. Per ogni campione generato sono state ricavate delle partizioni mediante la stima di una mistura di distribuzioni parametriche, e nell specifico gaussiane, e la mistura di distribuzioni non parametriche, come descritto nel capitolo 2. Il numero di gruppi è stato determinato automaticamente stimando in entrambi i casi modelli mistura da 2 a 9 componenti e selezionando il modello che presentava il valore minimo del BIC. Successivamente è stata valutata la bontà delle partizioni confrontando la partizione trovata con la vera partizione dei dati, attraverso l'*Adjusted Rand Index*.

L'*Adjusted Rand Index* (Hubert e Arabie, 1985) è un indice ricavato dal *Rand Index* e modificato per avere media nulla. Il *Rand Index* è un indice per misurare la similarità tra due diversi raggruppamenti degli stessi dati, e definito come segue.

Sia $Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ un insieme di dati e siamo $P = \{p_1, \ldots, p_r\}$ e $T = \{t_1, \ldots, t_s\}$ due partizioni dei dati Y, rispettivamente di r e s gruppi. Siano

- a: il numero di coppie di elementi di Y appartenenti allo stesso gruppo sia in P che in T
- b: il numero di coppie di elementi di Y appartenenti a gruppi diversi sia in P che in T
- c: il numero di coppie di elementi in Y appartenenti allo stesso gruppo in P ed a gruppi diversi in T
- d: il numero di coppie di elementi in Y appartenenti a gruppi diversi in P ed allo stesso gruppo in T

Allora in *Rand Index* è definito come

$$R = \frac{a+b}{a+b+c+d} \tag{3.1}$$

con $R \in [0, 1]$ e $R = 1$ quando la corrispondenza tra le due partizioni è perfetta.

L'*Adjusted Rand Index* (ARI) è ricavato dal *Rand Index*, sottraendo a quest'ultimo il suo valore atteso, per fare in modo che sia compreso in $[-1; 1]$ ed abbia media nulla nel caso il cui le unità siano allocate casualmente in una delle due partizioni.

Lo studio di simulazione è stato effettuato nell'ambiente di programmazione R (R Core Team, 2024), con l'ausilio dei pacchetti *mclust* (Scrucca *et al.*, 2023) e *mixtools* (Benaglia *et al.*, 2009b) per le stime dei modelli parametrici e, rispettivamente, non parametrici, e il pacchetto *pdfCluster* (Azzalini e Menardi, 2014) per la valutazione delle partizioni.

3.2 Risultati

I risultati dello studio di simulazione sono sintetizzati nelle Tabelle 3.1 e 3.2 e riportano il valore medio e le deviazioni standard dell'ARI al variare dei campioni generati, e il numero medio di gruppi identificati.

Andando ad analizzare i valori medi dell'ARI, i risultati sono coerenti con le aspettative. Entrambi gli algoritmi hanno prodotto partizioni perfette nella maggior parte dei campioni generati dal primo scenario e buone partizioni per il secondo scenario. Riguardo al terzo, i risultati sono molto buoni per l'algoritmo di stima non parametrica mentre l'algoritmo di stima di misture normali non è stato in grado di identificare correttamente i gruppi nella quasi totalità dei campioni simulati.

Anche il numero di gruppi è stato stimato correttamente in tutti i campioni generati dal primo e dal secondo scenario, fatta eccezione per il caso in 5 dimensioni per l'algoritmo MM, che ha sovrastimato il numero di gruppi nella maggior parte dei campioni, sottolineando la difficoltà di questo algoritmo ad operare in grandi dimensioni. Nel terzo scenario, invece, entrambi gli algoritmi hanno sovrastimato il numero di gruppi in modo sistematico, nel tentativo di includere i valori posizionati sulle code della distribuzione all'interno del modello, raggruppandoli in gruppi a sè stanti. Questo comportamento può essere facilmente previsto dall'algoritmo parametrico, mentre può risultare più sorprendente per l'algoritmo non parametrico, dal quale ci si aspetta che riesca a gestire campioni asimmetrici, e necessiterebbe degli approfondimenti su come quest'algoritmo tratti i valori anomali. C'è comunque da precisare che i valori molto superiori per i valori medi dell'ARI relativi al secondo algoritmo facciano pensare che i valori riconosciuti come *outlier*, e posti in gruppi a sé stanti, siano molto inferiori rispetto a quelli classificati come tali dal primo algoritmo, e siano solo le osservazioni molto distanti dalle mode dei gruppi.

Valutando l'influenza della numerosità campionaria sui risultati, si nota come con l'aumentare della difficoltà dello scenario in relazione all'algoritmo utilizzato, il miglioramento dovuto all'aumento delle osservazioni del campione è sempre più evidente. Si osservi, ad esempio, come nel caso in cui i gruppi siano di forma sferica e abbastanza vicini, l'ARI medio relativo alla partizione trovata dall'algoritmo di stima non parametrica in 5 dimensioni sia quasi raddoppiato, così come è più che raddoppiato nelle partizioni calcolate dall'algoritmo di stima di misture di distribuzioni normali nel caso di gruppi fortemente asimmetrici in 2 dimensioni. Al contrario, quando lo scenario presentato risulta facilmente trattabile dall'algoritmo, come il primo scenario o il terzo per l'algoritmo di stima non parametrico, non sembra esserci una particolare differenza tra campioni esigui e più numerosi, conse-

guenza del fatto che già con poche osservazioni i risultati siano molto buoni. Inoltre, nella maggior parte dei casi, l'aumento della numerosità campionaria ha diminuito la deviazione standard, indicando come le stime siano più stabili in campioni più numerosi.

Passando a considerare l'influenza della dimensione dello spazio campionario, anche l'aumento dei quest'ultima comporta forti cambiamenti nei risultati. Questo è dovuto alla cosiddetta *maledizione della dimensionalità* (Bellman, 1961): quando la dimensione dei dati aumenta, aumenta la sparsità dei dati, richiedendo un numero molto più elevato di osservazioni per ottenere buone stime. La conseguenza di questo fenomeno si può notare in modo particolare nelle stime prodotte dall'algoritmo non parametrico in tutti gli scenari in 5 dimensioni. L'incremento della dimensione dello spazio campionario, infatti, sembra avere un'influenza molto più forte sulla stima di misture non parametriche che sulla stima di misture parametriche, dovuta al numero molto maggiore di valori da stimare, peggiorando le partizioni prodotte dall'algoritmo MM in modo molto più rilevante rispetto all'algoritmo EM.

La forte influenza della maledizione della dimensionalità sul *clustering* basato su modelli è un fatto noto in letteratura, e solitamente viene affrontato con metodi di riduzione della dimensionalità. Un'esposizione esaustiva di questo aspetto richiederebbe una trattazione a parte, ed esula dagli obiettivi di questo elaborato, per questo si rimanda a Bouveyron *et al.* (2019) per approfondimenti.

Un'ultima valutazione viene fatta sui tempi computazionali necessari per l'esecuzione degli algoritmi, i cui valori medi sono riportati nella tabella 3.3. È immediato notare che, come ci si potrebbe aspettare, l'aumento della numerosità campionaria e del numero di dimensioni allunghi i tempi computazionali ma, mentre per l'algoritmo EM questo aumento non sembra essere particolarmente rilevante, lo stesso non si può dire per l'algoritmo MM. Il tempo di esecuzione dell'algoritmo di stima non parametrica, infatti, cresce in maniera più che esponenziale quando si passa ad analizzare campioni con numerosità o dimensioni più elevate e anche solo considerando un caso non particolarmente critico, con $n = 1000$ e $d = 5$, i tempi di esecuzione sono molto elevati. Sembra quindi poco fattibile l'utilizzo di questo algoritmo con campioni particolarmente numerosi e dimensioni dello spazio campionario anche non troppo elevate.

scenario	d	n	Mistura parametrica	Mistura non parametrica
1°	2	250	0.994 (0.010)	0.993 (0.009)
		1000	0.995 (0.004)	0.994 (0.005)
	5	250	0.994 (0.009)	0.766 (0.148)
		1000	0.995 (0.004)	0.743 (0.124)
2°	2	250	0.750 (0.093)	0.745 (0.056)
		1000	0.746 (0.028)	0.743 (0.027)
	5	250	0.746 (0.058)	0.278 (0.033)
		1000	0.746 (0.031)	0.437 (0.076)
3°	2	250	0.115 (0.156)	0.823 (0.058)
		1000	0.269 (0.046)	0.872 (0.034)
	5	250	0.279 (0.043)	0.601 (0.088)
		1000	0.225 (0.023)	0.687 (0.076)

Tabella 3.1: Risultati delle simulazioni in termini di valore medio e deviazione standard dell'ARI al variare di scenario, d e n

scenario	d	n	Mistura parametrica	Mistura non parametrica
1°	2	250	2.000	2.000
		1000	2.000	2.000
	5	250	2.000	5.150
		1000	2.000	8.120
2°	2	250	1.990	2.010
		1000	2.000	2.000
	5	250	2.000	8.570
		1000	2.000	8.840
3°	2	250	6.510	6.620
		1000	8.930	8.420
	5	250	8.060	8.450
		1000	8.990	8.790

Tabella 3.2: Valore medio del numero di gruppi stimato dai due algoritmo al variare dello scenario, di d e n

scenario	d	n	Mistura parametrica	Mistura non parametrica
1°	2	250	0.390	0.032
		1000	1.197	0.153
	5	250	0.633	4.463
		1000	2.506	33.093
2°	2	250	0.451	0.228
		1000	1.357	0.855
	5	250	0.755	8.201
		1000	2.815	36.303
3°	2	250	0.226	2.677
		1000	0.976	12.856
	5	250	0.492	5.409
		1000	1.854	28.850

Tabella 3.3: Tempi medi in secondi di esecuzione degli algoritmi per campione al variare dello scenario, di d e n

Conclusioni

L'obiettivo di questo elaborato è stato quello di descrivere come i modelli a mistura finita possano essere utilizzati per affrontare problemi di *clustering*, offrendo un'alternativa ai classici metodi basati su distanza.

In particolare, si è approfondito il confronto tra modelli a mistura finita di distribuzioni parametriche e non parametriche, descrivendone i relativi algoritmi di stima e valutando il loro comportamento quando applicati a diversi scenari attraverso uno studio di simulazione.

I risultati dello studio sono riusciti a delineare quali siano i vantaggi ed i limiti di ciascuno approccio, indicando in modo abbastanza chiaro gli scenari in cui ciascuno dei due approcci risulta più adatto.

Cercando di trarre delle considerazioni di ordine generale, i modelli parametrici, come le misture di distribuzioni gaussiane, hanno dimostrato la loro efficacia in contesti in cui le assunzioni relative alla forma delle componenti sono ragionevolmente note. Questi approcci, inoltre, garantiscono una rappresentazione chiara della struttura dei gruppi, consentendo di riassumerne le caratteristiche attraverso pochi parametri facilmente interpretabili.

D'altro canto, i modelli non parametrici si sono rivelati particolarmente utili per scenari complessi, dove non è possibile fare assunzioni rigide sulla distribuzione. Questi metodi offrono maggiore flessibilità, consentendo di modellare le caratteristiche dei dati in maniera più naturale e dinamica. Tuttavia, tale flessibilità comporta una maggiore complessità computazionale.

L'analisi comparativa ha mostrato che, in scenari dove la struttura dei dati è fortemente variabile o sconosciuta, i modelli non parametrici superano le misture parametriche in termini di accuratezza e adattabilità. Tuttavia, in contesti in cui i dati seguono ipotesi ben definite, le misture parametriche si confermano come soluzioni rapide ed efficienti. Di conseguenza, la scelta tra approcci parametrici e non parametrici dovrebbe essere guidata dalle caratteristiche specifiche del problema da affrontare, dalla disponibilità di risorse computazionali e dalle esigenze di interpretabilità dei risultati.

Ovviamente, le conclusioni che si possono trarre da questo lavoro non risultano generali, ma circoscritte agli scenari a cui i metodi sono stati applicati. Eventuali approfondimenti potrebbero riguardare la varietà di scenari considerati, analizzando situazioni in cui i gruppi non hanno forma sferica, ad esempio se sono generati da distribuzioni note diverse dalla distribuzione normale, o casi in cui presentino forme differenti. Si potrebbe inoltre valutare la possibilità di utilizzare metodi di riduzione della dimensionalità con l'algoritmo di stima non parametrico, per cercare di migliorarne l'efficienza computazionale in contesti con dimensioni elevate.

Bibliografia

Azzalini A.; Menardi G. (2014). Clustering via nonparametric density estimation: The R package pdfCluster. *Journal of Statistical Software*, **57**(11), 1–26.

Banfield J.; Raftery A. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**.

Bellman R. E. (1961). *Adaptive Control Processes: A Guided Tour*. Princeton, New Jersey, USA.

Benaglia T.; Chauveau D.; Hunter D. (2009a). An em-like algorithm for semi- and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, **18**(2).

Benaglia T.; Chauveau D.; Hunter D. R.; Young D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, **32**(6), 1–29.

Biernacki C.; Govaert G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, **64**, 49–71.

Biernacki C.; Celeux G.; Govaert G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**, 719 – 725.

Bouveyron C.; Celeux G.; Murphy T. B.; Raftery A. E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Celeux G.; Govaert G. (1993). Gaussian parsimonious clustering models. *Pattern Recognit*, **28**.

Coleman D.; Dong X.; Hardin J.; Rocke D.; Woodruff D. (1999). Some computational issues in cluster analysis with no a priori metric. *Computational Statistics & Data Analysis*, **31**, 1–11.

Dempster A.; Laird N.; Rubin D. (1977). Maximum likelihood from incomplete data via em algorithm. *J. Royal Statistical Soc., Series B*, **39**, 1 – 38.

Hall P.; Zhou X.-H. (2003a). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, **31**.

Hall P.; Zhou X.-H. (2003b). Nonparametric estimation of component distributions in a multivariate mixture. *The annals of statistics*, **31**(1), 201–224.

Hennig C.; Meila M.; Murtagh F.; Rocci R. (2016). *Handbook of Cluster Analysis*. Chapman & Hall.

Hubert L.; Arabie P. (1985). Comparing partitions. *Journal of Classification*, **2**(1), 193–219.

Hunter D. R. (2024). Unsupervised clustering using nonparametric finite mixture models. *Wiley Interdisciplinary Reviews: Computational Statistics*, **16**(1), e1632.

Hunter D. R.; Lange K. (2004). A tutorial on mm algorithms. *The American Statistician*, **58**(1), 30–37.

Keribin C. (2000). Consistent estimate of the order of mixture models. *Sankhy=a, Series A*, **62**, 49–66.

Levine M.; Hunter D. R.; Chauveau D. (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika*, **98**(2), 403–416.

McLachlan G. J. (1987). On bootstrapping the likelihood ratio test stastistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **36**(3), 318–324.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scrucca L.; Fraley C.; Murphy T. B.; Raftery A. E. (2023). *Model-Based Clustering, Classification, and Density Estimation Using mclust in R*. Chapman and Hall/CRC.

Silverman B. W. (1986). *Density estimation for statistics and data analysis: Monographs onstatistics and applied probability*. Chapman and Hall.

Wand M. P.; Jones M. C. (1994). *Kernel smoothing*. CRC press.