Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

Un metodo di stima puntuale per il clustering bayesiano basato su metriche sullo spazio delle partizioni

Relatore Prof. Antonio Canale
Dipartimento di Scienze Statistiche

Laureando Nicola Castelletti
Matricola 2002689

Anno Accademico 2022/2023

Indice

Introduzione

Il clustering è una procedura che consiste nel dividere un insieme eterogeneo di osservazioni in diversi gruppi distinti, detti cluster, ognuno dei quali contenente unità simili per qualche caratteristica. I metodi di clustering sono importanti strumenti nell'investigazione scientifica in molti differenti domini. Esistono diversi metodi per eseguire clustering su un insieme di dati.

Alcuni tra gli algoritmi più popolari sono quelli di tipo gerarchico agglomerativo e quelli di tipo non gerarchico, tra cui k-medie. Questi algoritmi, nonostante si dimostrino efficaci in numerose applicazioni, esplorano solo un ristretto sottoinsieme dello spazio delle partizioni o necessitano della specificazione del numero di cluster a priori, restituiscono una singola soluzione di clustering e sono largamente euristici, non essendo basati su modelli formali.

Nei metodi di clustering basati su modelli mistura finiti, ogni componente del modello corrisponde ad un potenziale cluster. Scegliendo di stimare i parametri del modello tramite stima di massima verosimiglianza, ciascuna osservazione viene assegnata al cluster per il quale è massima la relativa probabilità a posteriori, valore che indica anche il grado di incertezza nell'assegnazione. In questo tipo di modelli non si tiene conto dell'incertezza nella stima dei parametri.

I metodi di clustering basati su modelli mistura bayesiani incorporano informazione a priori sui parametri del modello, e permettono di valutare l'incertezza nella struttura di clustering incondizionatamente alle stime dei parametri, analizzando la distribuzione a posteriori del clustering. I modelli mistura bayesiani non parametrici, inoltre, assumono un numero infinito di componenti. Questa caratteristica permette sia di evitare la specificazione a priori del numero di componenti, sia la possibilità di crescita del numero di cluster presenti nei dati, a mano a mano che nuovi dati vengono raccolti.

In questo elaborato ci si focalizza su questi ultimi modelli e, in particolare, sui risultati di Wade and Ghahramani [2018], nella ricerca di un'appropriata stima puntuale della soluzione di clustering, con relativo insieme di credibilità al 95%. Questo con lo scopo di sintetizzare l'informazione contenuta nella distribuzione a posteriori sullo spazio, molto spesso enorme, delle partizioni.

Capitolo 1

Metodi di clustering

Il clustering è una procedura non supervisionata molto studiata in statistica, che consiste nel partizionamento di un insieme eterogeneo di osservazioni, in un insieme di sottogruppi, detti cluster, contenenti ognuno unità simili per qualche loro caratteristica. Una possibile distinzione dei metodi di clustering prevede la loro divisione in hard clustering e soft clustering. I metodi di hard clustering forniscono una soluzione per la quale ogni unità può appartenere ad uno ed un solo cluster; mentre, nei metodi di soft clustering, si ottiene una soluzione che permette l'attribuzione di ogni unità a più di un cluster, accompagnando ognuna di queste assegnazioni con un relativo grado di confidenza. Si descrivono brevemente i metodi di hard e soft clustering.

1.1 Hard clustering

I metodi di hard clustering prevedono di fornire una soluzione di clustering nella quale ogni unità viene assegnata ad uno ed un solo cluster. Alcuni tra gli algoritmi più popolari di hard clustering sono quelli di tipo gerarchico agglomerativo e quelli di tipo non gerarchico, tra cui k-medie [Hartigan and Wong, 1979]. Questi algoritmi, nonostante si dimostrino efficaci in numerose applicazioni, presentano alcune limitazioni: esplorano solo un ristretto sottoinsieme dello spazio delle partizioni o necessitano della specificazione del numero di cluster a priori, restituiscono una singola soluzione di clustering e sono largamente euristici, non essendo basati su modelli formali, proibendo l'uso di strumenti statistici, per esempio, per determinare il numero di cluster.

Per l'utilizzo di un algoritmo di hard clustering è necessario specificare quali definizioni di distanza si intende utilizzare. In particolare, sono necessarie

due definizioni di distanza: distanza tra due unità e distanza tra un cluster e un cluster/unità. La definizione della distanza tra due unità avvenire generalmente tramite la definizione di distanza di Minkowski, la quale comprende, come casi particolari, la distanza di Manhattan, Euclidea e di Lagrange. Per la definizione della distanza tra un cluster e un cluster/unità, usualmente si ricorre a uno dei seguenti metodi: legame singolo, legame completo, legame medio, distanza dei centroidi o metodo di Ward. Naturalmente, la scelta di diverse combinazioni delle due distanze porta presumibilmente ad ottenere differenti partizioni delle unità; la scelta delle tipologie di distanze da utilizzare sarà dunque guidata dal tipo di informazione che si desidera ottenere dai dati.

I metodi di hard clustering possono essere suddivisi in metodi gerarchici e metodi non gerarchici. Nei metodi gerarchici avviene una sequenziale formazione di nuovi cluster, con la particolarità che ad ogni iterazione, un'unità già precedentemente assegnata ad un cluster non può essere riassegnata ad un nuovo cluster. Nei metodi non gerarchici, invece, si procede per adattamenti successivi, permettendo dunque la riassegnazione ad un nuovo cluster di un'unità già precedentemente assegnata.

1.1.1 Metodi gerarchici

Nei metodi gerarchici di clustering vengono formati sequenzialmente nuovi cluster aggregando o dividendo unità in base alla loro distanza.
I metodi gerarchici si dividono in agglomerativi e divisi: i primi partono dalla situazione in cui ogni unità è considerata un cluster a sè stante, agglomerando ad ogni iterazione le unità più vicine tra loro; i secondi, invece, considerano in partenza tutte le unità come appartenenti ad un unico cluster, separando ad ogni iterazione le unità più lontane tra loro.
Per costruzione, i metodi gerarchici non permettono, ad un'unità già precedentemente assegnata ad un cluster, la riassegnazione ad un nuovo cluster: nel caso di elemento collocato in maniera errata all'inizio della procedura, questa caratteristica può inficiare la bontà della soluzione di clustering.

1.1.2 Metodi non gerarchici

Nei metodi non gerarchici di clustering si procede per adattamenti successivi, permettendo la riassegnazione ad un nuovo cluster di un'unità già precedentemente assegnata. Per questo tipo di metodi è necessario specificare a priori

il numero di cluster desiderati; si possono in seguito confrontare i risultati per diverse numerosità di cluster.

L'algoritmo più conosciuto per eseguire questo tipo di clustering è k-medie. Questo algoritmo considera una suddivisione iniziale di tutte le unità in k gruppi casuali; in seguito, iterativamente calcola il centroide di ogni gruppo e riassegna ciascuna unità al gruppo che ha il centroide più vicino all'unità, fino a quanto, ad una nuova iterazione, la composizione dei cluster non viene ulteriormente modificata. Questo algoritmo, nonostante si dimostri efficace in numerose applicazioni, presenta anche alcune limitazioni: l'algoritmo fornisce una sola soluzione di clustering, non tenendo dunque conto della naturale variabilità ad essa associata, soluzione che tra l'altro rimane vincolata al numero di cluster imposto a priori. In più, rimane una soluzione largamente euristica, non essendo basata su modelli formali, proibendo l'uso di strumenti statistici, per esempio, per determinare il numero di cluster.

1.2 Soft clustering

Con i metodi di soft clustering si ottengono soluzioni di clustering che prevedono la possibilità di attribuire ogni unità a più di un cluster creato, accompagnando ognuna di queste assegnazioni con un relativo grado di confidenza. In questo modo, unità sui bordi di un cluster si possono considerare appartenenti a quel cluster con un grado di confidenza minore rispetto a quello di unità situate al centro del cluster. I metodi di soft clustering basati su modelli utilizzano modelli mistura finiti, dove ogni componente del modello corrisponde potenzialmente ad un cluster [Fraley and Raftery, 2002]. I problemi di determinazione del numero di componenti e di stima dei parametri delle distribuzioni di probabilità possono essere affrontati tramite la selezione statistica del modello, per esempio, tramite vari criteri di informazione.

L'algoritmo di aspettazione-massimizzazione (EM) viene tipicamente utilizzato per la stima di massima verosimiglianza dei parametri del modello mistura. Date le stime di massima verosimiglianza dei parametri, la probabilità a posteriori che un'unità appartenga ad un cluster può essere ottenuta tramite il teorema di Bayes. Ciascuna unità viene assegnata al cluster al quale corrisponde la massima probabilità a posteriori, valore che indica anche il grado di incertezza nell'assegnazione. Questa misura di incertezza, però, ignora l'incertezza nella stima dei parametri del modello, in quanto la soluzione di clustering è fornita al netto della stima puntuale dei parametri in gioco, non tenendo dunque conto della naturale variabilità associata alla stima.

In opposizione alla stima di massima verosimiglianza dei parametri del modello mistura, i modelli mistura bayesiani incorporano informazione a priori sui parametri, che, osservati i dati, diventa informazione a posteriori: la presenza di una distribuzione a posteriori dei parametri del modello permette, nel momento in cui si presenta l'incertezza associata alla soluzione di clustering trovata, di tenere conto anche dell'incertezza nella stima dei parametri [Bouveyron et al., 2019].

I modelli mistura bayesiani non parametrici assumono che il numero di componenti del modello sia infinito. In opposizione ai modelli mistura finiti, questo non solo evita la necessità di dover specificare il numero di cluster a priori, ma permette anche l'eventuale crescita del numero di cluster presenti nei dati, a mano a mano che nuovi dati vengono raccolti.

Capitolo 2

Metodi di clustering bayesiani non parametrici

I modelli mistura sono uno degli strumenti più popolari in statistica bayesiana non parametrica. Le osservazioni sono assunte i.i.d. con densità

$$f(y|P) = \int K(y|\theta)dP(\theta),$$

dove $K(y|\theta)$ è una specificata densità parametrica sullo spazio delle osservazioni con parametro di mistura $\theta \in \Theta$ e P è una misura di probabilità su Θ. In un contesto bayesiano, il modello è completato con la specificazione di una distribuzione a priori del parametro ignoto, che in questo caso è l'ignota misura P. Nel contesto più generale possibile, questo parametro P può essere una qualsiasi misura di probabilità su Θ, richiedendo una distribuzione a priori non parametrica. Tipicamente la distribuzione a priori non parametrica ha realizzazioni P discrete quasi certamente, dove

$$P = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j} \quad \text{q.c.},$$

dove viene spesso assunto che i pesi (ω_j) e gli atomi (θ_j) sono indipendenti e i θ_j sono i.i.d. da una misura di base P_0. Quindi, la densità è modellata con un modello mistura infinito numerabile

$$f(y|P) = \sum_{j=1}^{\infty} \omega_j K(y|\theta_j).$$

Dato che P è discreta q.c., questo modello induce una partizione latente $\mathbf{c}$ dei dati dove due osservazioni appartengono allo stesso cluster se sono generate dalla stessa componente del modello mistura. La partizione può essere rappresentata da $\mathbf{c} = (C_1, \ldots, C_{k_N})$, dove C_j contiene gli indici delle osservazioni nel j-simo cluster, k_N è il numero di cluster nel campione e N è la numerosità campionaria. Alternativamente, la partizione può essere rappresentata da $\mathbf{c} = (c_1, \ldots, c_N)$, dove $c_n = j$ se la n-sima osservazione è nel cluster j-simo. Nel seguito si utilizza la prima rappresentazione.

Una differenza chiave rispetto ai modelli mistura finiti è che il numero di componenti mistura è infinito; questo permette alle osservazioni di determinare il numero di cluster, o componenti occupate, k_N presenti nei dati, il quale può crescere a mano a mano che nuove osservazioni vengono raccolte. Sia $\mathbf{y}_j = \{y_i\}_{i \in C_j}$, la funzione di verosimiglianza marginale per le osservazioni $y_{1:N}$ data la partizione è

$$f(y_{1:N}|\mathbf{c}) = \prod_{j=1}^{k_N} m(\mathbf{y}_j) = \prod_{j=1}^{k_N} \int \prod_{i \in C_j} K(y_i|\theta) dP_0(\theta).$$

La distribuzione a posteriori della partizione, che riflette l'incertezza nella soluzione di clustering date le osservazioni, è proporzionale al prodotto tra la distribuzione a priori della partizione e la funzione di verosimiglianza marginale

$$p(\mathbf{c}|f(y_{1:N})) \propto p(\mathbf{c}) \prod_{j=1}^{k_N} m(\mathbf{y}_j), \tag{2.1}$$

dove la distribuzione a priori della partizione è ottenuta dalla distribuzione a priori della misura P selezionata. Per esempio, un processo a priori di Dirichlet [Ferguson, 1973] per P con parametro di massa totale α corrisponde a

$$p(\mathbf{c}) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \alpha^{k_N} \prod_{j=1}^{k_N} \Gamma(n_j),$$

dove $n_j = |C_j|$ è il numero di osservazioni nel cluster j. Numerose altre distribuzioni a priori, sviluppate nella letteratura non parametrica bayesiana, possono essere considerate per la misura P, come il processo di Pitman-Yor [Pitman and Yor, 1997], conosciuto anche come il processo Poisson-Dirichlet a due parametri, o il processo normalizzato generalizzato Gamma o, più in

generale, una distribuzione a priori nella classe delle misure normalizzate completamente casuali, modelli Poisson-Kingman [Pitman, 2003], o le distribuzioni a priori stick-breaking [Ishwaran and James, 2001]. Consultare Lijoi and Prünster [2010] per una panoramica.

In generale, la funzione di verosimiglianza marginale delle osservazioni data la partizione o la distribuzione a priori della partizione usata per calcolare la distribuzione a posteriori in (2.1) possono non essere disponibili in forma chiusa.

In aggiunta, per un insieme di osservazioni di cardinalità N, il numero di possibili k-partizioni, partizioni costituite da k insiemi, è definito come il numero di Stirling di seconda specie $S_{N,k}$. Anche per N piccolo, questo numero è molto grande, il che rende il calcolo della distribuzione a posteriori in (2.1) intrattabile, anche per scelte semplici per la distribuzione a priori della partizione e la funzione di verosimiglianza. Per queste ragioni, tipicamente vengono utilizzate tecniche MCMC, algoritmi che producono campioni approssimati $(\mathbf{c}^m)_{m=1}^M$ dalla distribuzione a posteriori della partizione, (2.1). Chiaramente, descrivere tutti i singoli campioni approssimati dalla distribuzione a posteriori della partizione, campioni che a volte differiscono tra loro solo per alcune, poche, osservazioni, è impraticabile. Nelle prossime sezioni si presentano i risultati di Wade and Ghahramani [2018], nella ricerca di appropriati strumenti di sintesi per caratterizzare la distribuzione a posteriori della partizione.

2.1 Stime puntuali per la struttura di clustering

In una tipica analisi bayesiana, la distribuzione a posteriori di un parametro di interesse univariato è spesso riassunta riportando una stima puntuale come la media, mediana o moda a posteriori, assieme ad un intervallo di fiducia al 95% per caratterizzare l'incertezza. Nel problema in analisi, il parametro di interesse è la partizione. Si presentano strumenti di sintesi della relativa distribuzione a posteriori, in quel che segue.

Una prima semplice soluzione è quella di utilizzare la moda a posteriori, ovvero la partizione più visitata dall'algoritmo MCMC. Questo approccio può essere problematico: per la presenza di un enorme numero di partizioni, molte delle quali che differiscono l'un l'altra solo di alcune, poche, osservazioni, l'algoritmo visita solitamente la quasi totalità delle partizioni una sola volta,

producendo di conseguenza una stima puntuale della moda della distribuzione a posteriori della partizione altamente instabile e soggetta a molta variabilità. In più, è risaputo che la moda può non essere rappresentativa del centro della distribuzione.

Una seconda soluzione consiste in metodi basati sulla matrice di similarità a posteriori. Per una numerosità campionaria pari a N, gli elementi di questa matrice N per N rappresentano la probabilità che due osservazioni appartengano allo stesso cluster, e possono essere stimati dalla proporzione dei campioni MCMC che raggruppa le due osservazioni assieme. In seguito, basandosi sulla matrice di similarità, vengono applicati alcuni classici algoritmi gerarchici o di partizionamento. Questi metodi hanno lo svantaggio di essere ad-hoc [Wade and Ghahramani, 2018].

Una soluzione più elegante è quella basata sulla teoria della decisione. La teoria della decisione è lo studio matematico-statistico del modo di scegliere tra varie alternative possibili, individuando quali possano essere le decisioni ottimali, determinate considerando un decisore ideale pienamente razionale. Seguendo questo approccio, è necessario definire una funzione di perdita $L(\mathbf{c}, \widehat{\mathbf{c}})$ sullo spazio delle partizioni, che misura la perdita che si ottiene nello stimare il vero clustering $\mathbf{c}$ con $\widehat{\mathbf{c}}$. Siccome il vero clustering è sconosciuto, la funzione di perdita viene mediata su tutti i possibili veri clustering, dove la funzione di perdita associata a ciascun potenziale vero clustering è pesata per la probabilità a posteriori associata a quel possibile vero clustering. La stima puntuale $\mathbf{c}^*$ corrisponde alla stima che minimizza la funzione di perdita a posteriori attesa,

$$\mathbf{c}^* = \arg\min_{\widehat{\mathbf{c}}} \mathbb{E}[L(\mathbf{c}, \widehat{\mathbf{c}})|y_{1:N}] = \arg\min_{\widehat{\mathbf{c}}} \sum_{\mathbf{c}} L(\mathbf{c}, \widehat{\mathbf{c}}) p(\mathbf{c}|y_{1:N}).$$

Per esempio, per un parametro reale θ, la stima puntuale ottimale è la media a posteriori sotto la funzione di perdita ad errore quadratico $L_2(\theta, \widehat{\theta}) = (\theta - \widehat{\theta})^2$, ed è la mediana a posteriori sotto la funzione di perdita ad errore assoluto $L_1(\theta, \widehat{\theta}) = |\theta - \widehat{\theta}|$.

L'attenzione si sposta dunque sul determinare quale possa essere un'appropriata funzione di perdita sullo spazio delle partizioni.

Una scelta semplice per la funzione di perdita è la perdita 0-1, $L_{0-1}(\mathbf{c}, \widehat{\mathbf{c}}) = \mathbf{1}(\mathbf{c} \neq \widehat{\mathbf{c}})$, che assume una perdita di 0 se la stima $\widehat{\mathbf{c}}$ è uguale alla verità $\mathbf{c}$, e una perdita di 1 altrimenti. Sotto la perdita 0-1, la stima puntuale è la moda a posteriori. Questa funzione di perdita non è tuttavia soddisfacente perché non tiene conto della similarità tra due clustering: una partizione che differisce dalla verità nell'allocazione di solo un'osservazione è penalizzata allo stesso modo di

una partizione che differisce dalla verità nell'allocazione di molte osservazioni. In più, è risaputo che la moda può non essere rappresentativa del centro di una distribuzione. Perciò, sono necessarie delle funzioni di perdita più generali.

La costruzione di una funzione di perdita più generale, però, non è semplice poiché, come osservato da Binder [1978], la funzione di perdita dovrebbe soddisfare principi di base come l'invarianza alla permutazione degli indici delle osservazioni e l'invarianza alla permutazione delle etichette dei cluster, sia per il vero clustering $\mathbf{c}$ che per quello stimato $\widehat{\mathbf{c}}$. Binder nota che la prima condizione implica che la perdita debba essere una funzione dei conteggi $n_{ij} = |C_i \cap \widehat{C}_j|$, conteggi che sono uguali alla cardinalità dell'intersezione tra C_i, l'insieme degli indici delle osservazioni nel cluster i sotto $\mathbf{c}$, e $\widehat{C}_j$, l'insieme degli indici delle osservazioni nel cluster j sotto $\widehat{\mathbf{c}}$, per $i = 1, \ldots, k_N$ e $j = 1, \ldots, \widehat{k}_N$, e con k_N e $\widehat{k}_N$ che rappresentano il numero di cluster in $\mathbf{c}$ e in $\widehat{\mathbf{c}}$, rispettivamente. Binder esplora funzioni di perdita che soddisfano questi principi, partendo da semplici funzioni dei conteggi n_{ij}. La cosiddetta funzione di perdita di Binder è una funzione quadratica dei conteggi, che, per tutte le possibili coppie di osservazioni, penalizza i due errori di allocare due osservazioni in cluster diversi quando dovrebbero essere nello stesso cluster (penalità l_1) o di allocarle nello stesso cluster quando dovrebbero essere in due cluster differenti (penalità l_2):

$$B(\mathbf{c}, \widehat{\mathbf{c}}) = \sum_{n<n'} l_1 \mathbf{1}(c_n = c_{n'}) \mathbf{1}(\widehat{c}_n \neq \widehat{c}_{n'}) + l_2 \mathbf{1}(c_n \neq c_{n'}) \mathbf{1}(\widehat{c}_n = \widehat{c}_{n'}).$$

La funzione di perdita di Binder, per una coppia di partizioni composta da una vera partizione $\mathbf{c}$ ed una stimata $\widehat{\mathbf{c}}$, restituisce la somma delle penalità associate a ciascuna coppia di osservazioni che presenta errore di allocazione, prendendo in considerazione tutte le $\binom{N}{2}$ possibili coppie di osservazioni.

2.2 Confronto tra due funzioni di perdita: perdita di Binder e variazione dell'informazione

Meilă [2007] introduce la variazione dell'informazione (VI) per il confronto tra clustering, che è costruita dalla teoria dell'informazione e confronta l'informazione contenuta in due clustering con l'informazione condivisa tra i due clustering. Le due principali quantità a cui la teoria dell'informazione fa riferimento sono l'entropia e la mutua informazione. L'entropia quantifica l'ammontare di incertezza presente nella realizzazione di una variabile casuale, mentre la

mutua informazione quantifica l'informazione riguardo a una variabile casuale ottenuta osservando un'altra variabile casuale, mutualmente dipendente dalla prima.

Più formalmente, la VI è definita come

$$\begin{aligned}\mathrm{VI}(\mathbf{c},\widehat{\mathbf{c}}) &= \mathrm{H}(\mathbf{c}) + \mathrm{H}(\widehat{\mathbf{c}}) - 2\mathrm{I}(\mathbf{c},\widehat{\mathbf{c}}) \\ &= -\sum_{i=1}^{k_N} \frac{n_{i+}}{N} \log\left(\frac{n_{i+}}{N}\right) - \sum_{j=1}^{\widehat{k}_N} \frac{n_{+j}}{N} \log\left(\frac{n_{+j}}{N}\right) - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \frac{n_{ij}}{N} \log\left(\frac{n_{ij}N}{n_{i+}n_{+j}}\right),\end{aligned}$$

dove log denota log in base 2. I primi due termini rappresentano l'entropia dei due clustering, e misurano l'incertezza dell'allocazione in un cluster di una sconosciuta osservazione casualmente selezionata, dato un particolare clustering delle osservazioni. L'ultimo termine è la mutua informazione tra i due clustering, e misura la riduzione dell'incertezza dell'allocazione in un cluster di un osservazione in $\mathbf{c}$, quando si viene a conoscenza del cluster in cui la stessa osservazione è stata allocata in $\widehat{\mathbf{c}}$. La VI varia tra 0 e $\log(N)$.

Viene proposto di utilizzare la VI come funzione di perdita, e se ne fornisce un dettagliato confronto con una versione N-invariante della perdita di Binder, definita come

$$\tilde{B}(\mathbf{c},\widehat{\mathbf{c}}) = \frac{2}{N^2} B(\mathbf{c},\widehat{\mathbf{c}}) = \sum_{i=1}^{k_N} \left(\frac{n_{i+}}{N}\right)^2 + \sum_{j=1}^{\widehat{k}_N} \left(\frac{n_{+j}}{N}\right)^2 - 2\sum_{i=1}^{k_N}\sum_{j=1}^{\widehat{k}_N} \left(\frac{n_{ij}}{N}\right)^2.$$

Entrambe le funzioni di perdita sono considerate N-invarianti dato che dipendono da N solo tramite le proporzioni n_{ij}/N.

Ci si focalizza su queste due funzioni di perdita perché soddisfano diverse proprietà desiderabili.

Proprietà 1. *Sia VI che $\tilde{B}$ sono metriche sullo spazio delle partizioni.*

Dunque, sullo spazio delle partizioni, sia VI che $\tilde{B}$ sono degli insiemi equipaggiati con la nozione di distanza tra i loro elementi, ovvero tra le diverse partizioni. Per una dimostrazione di questa proprietà confrontare Meilă [2007] per VI e Wade and Ghahramani [2018] per $\tilde{B}$.

Per le successive proprietà è necessario vedere lo spazio delle partizioni $\mathbf{C}$ come un insieme parzialmente ordinato, ovvero un insieme equipaggiato

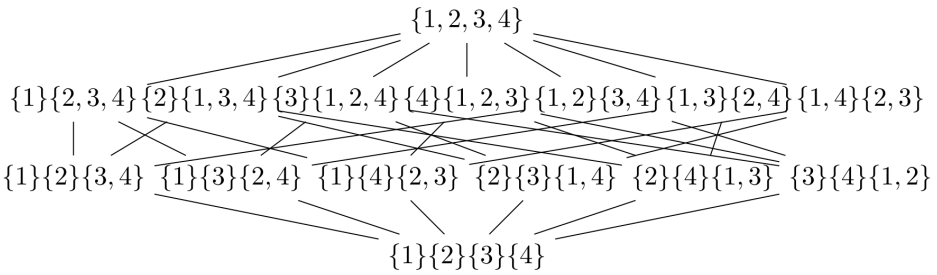


Figura 2.1: Diagramma di Hasse per lo spazio delle partizioni, con dimensione campionaria $N = 4$. Una linea è tracciata da $\mathbf{c}$ a $\widehat{\mathbf{c}}$ quando $\mathbf{c}$ è coperto da $\widehat{\mathbf{c}}$.

con $\leq$, definito dall'inclusione di insiemi: per $\mathbf{c}, \widehat{\mathbf{c}} \in \mathbf{C}, \mathbf{c} \leq \widehat{\mathbf{c}}$ se, per ogni $i = \{1, \ldots, k_N\}$, $C_i \subseteq \widehat{C_j}$ per qualche $j \in \{1, \ldots, \widehat{k_N}\}$.

Per ogni $\mathbf{c}, \widehat{\mathbf{c}} \in \mathbf{C}$, $\mathbf{c}$ è definito coperto da $\widehat{\mathbf{c}}$ se $\mathbf{c} < \widehat{\mathbf{c}}$ e non c'è alcun $\widehat{\widehat{\mathbf{c}}} \in \mathbf{C}$ tale per cui $\mathbf{c} < \widehat{\widehat{\mathbf{c}}} < \widehat{\mathbf{c}}$. Questa relazione di copertura è utilizzata per definire il diagramma di Hasse, nel quale gli elementi di $\mathbf{C}$ sono rappresentati come nodi di un grafo, e una linea è tracciata da $\mathbf{c}$ a $\widehat{\mathbf{c}}$ quando $\mathbf{c}$ è coperto da $\widehat{\mathbf{c}}$. Un esempio di diagramma di Hasse per $N = 4$ è rappresentato in Figura 2.1.

Lo spazio delle partizioni possiede una struttura ancora più ricca; esso forma un reticolo, ovvero un insieme parzialmente ordinato nel quale ogni coppia di elementi ha sia un estremo inferiore che un estremo superiore. Questo deriva dal fatto che ogni coppia di partizioni ha un estremo inferiore, inf, ed un estremo superiore, sup; per un sottoinsieme $\mathbf{S} \subseteq \mathbf{C}$, un elemento $\mathbf{c} \in \mathbf{C}$ è un maggiorante di $\mathbf{S}$ se $\mathbf{s} \leq \mathbf{c}$ per ogni $\mathbf{s} \in \mathbf{S}$, e $\mathbf{c} \in \mathbf{C}$ è l'estremo superiore di $\mathbf{S}$, indicato $\mathbf{c} = \sup(\mathbf{S})$, se $\mathbf{c}$ è un maggiorante di $\mathbf{S}$ e $\mathbf{c} \leq \mathbf{c}'$ per ogni maggiorante $\mathbf{c}'$ di $\mathbf{S}$. Un minorante e il minimo di un sottoinsieme $\mathbf{S} \subseteq \mathbf{C}$ sono analogamente definiti, l'ultimo indicato con $\inf(\mathbf{S})$. Si definiscono gli operatori $\wedge$, chiamato intersezione, e $\vee$, chiamato unione, come $\mathbf{c} \wedge \widehat{\mathbf{c}} = \inf(\mathbf{c}, \widehat{\mathbf{c}})$ e $\mathbf{c} \vee \widehat{\mathbf{c}} = \sup(\mathbf{c}, \widehat{\mathbf{c}})$. Seguendo la convenzione della teoria del reticolo, si usa $\mathbf{1}$ per indicare il più grande elemento del reticolo delle partizioni, ovvero la partizione con ogni osservazione in un unico cluster, $\mathbf{c} = (\{1, \ldots, N\})$, e $\mathbf{0}$ per indicare il più piccolo elemento del reticolo delle partizioni, ovvero la partizione con ogni osservazione nel proprio cluster, $\mathbf{c} = (\{1\}, \ldots, \{N\})$. Consultare Nation [1998] per maggiori dettagli sulla teoria del reticolo e il Supplementary Material del lavoro di Wade e Ghahramani [2017] per dettagli specifici sul reticolo delle partizioni.

Una proprietà desiderabile è che sia VI che $\tilde{B}$ siano allineate con il reticolo delle partizioni. Nello specifico, entrambe le metriche sono allineate verticalmente nel diagramma di Hasse; se $\widehat{\widehat{\mathbf{c}}}$ è connesso con $\widehat{\mathbf{c}}$ e $\widehat{\mathbf{c}}$ è connesso con $\mathbf{c}$, allora la distanza tra $\widehat{\widehat{\mathbf{c}}}$ e $\mathbf{c}$ è la somma verticale delle distanze tra $\widehat{\widehat{\mathbf{c}}}$ e $\widehat{\mathbf{c}}$ e tra $\widehat{\mathbf{c}}$ e $\mathbf{c}$ (confrontare Proprietà 2). Entrambe le metriche sono anche allineate orizzontalmente nel diagramma di Hasse; la distanza tra due partizioni qualunque è la somma orizzontale delle distanze tra ogni partizione e l'intersezione delle due partizioni (confrontare Proprietà 3).

Proprietà 2. *Sia per VI che per $\tilde{B}$, se $\boldsymbol{c} \geq \widehat{\mathbf{c}} \geq \widehat{\widehat{\mathbf{c}}}$, allora*

$$d(\mathbf{c}, \widehat{\widehat{\mathbf{c}}}) = d(\mathbf{c}, \widehat{\mathbf{c}}) + d(\widehat{\mathbf{c}}, \widehat{\widehat{\mathbf{c}}}).$$

Proprietà 3. *Sia per VI che per $\tilde{B}$,*

$$d(\mathbf{c}, \widehat{\mathbf{c}}) = d(\mathbf{c}, \widehat{\mathbf{c}} \wedge \mathbf{c}) + d(\widehat{\mathbf{c}}, \widehat{\mathbf{c}} \wedge \mathbf{c}).$$

Dimostrazioni possono essere trovate nel Supplementary Material del lavoro di Wade e Ghahramani [2017]. Queste due proprietà implicano che, se il diagramma di Hasse è allungato per riflettere la distanza tra ogni partizione e $\mathbf{1}$, la distanza tra due partizioni qualunque può essere facilmente determinata dal diagramma di Hasse allungato. Le Figure 2.2 e 2.3 rappresentano il diagramma di Hasse per $N = 4$ di Figura 2.1 allungato secondo $\tilde{B}$ e VI rispettivamente.

Dai diagrammi di Hasse allungati si possono ottenere diverse informazioni riguardo le similarità e le differenze tra le due metriche. Una differenza evidente è la scala dei due diagrammi.

Proprietà 4. *Una distanza sullo spazio delle partizioni che soddisfa le Proprietà 2 e 3 ha la proprietà che, per due partizioni qualunque $\boldsymbol{c}$ e $\widehat{\boldsymbol{c}}$,*

$$d(\mathbf{c}, \widehat{\mathbf{c}}) \leq d(\mathbf{1}, \mathbf{0}).$$

Quindi,

$$\mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}}) \leq \log(N) \quad e \quad \tilde{B}(\mathbf{c}, \widehat{\mathbf{c}}) \leq 1 - \frac{1}{N}.$$

Una dimostrazione può essere trovata nel Summplementary Material del lavoro di Wade e Ghahramani [2017]. In entrambi i casi, il limite sulle distanze

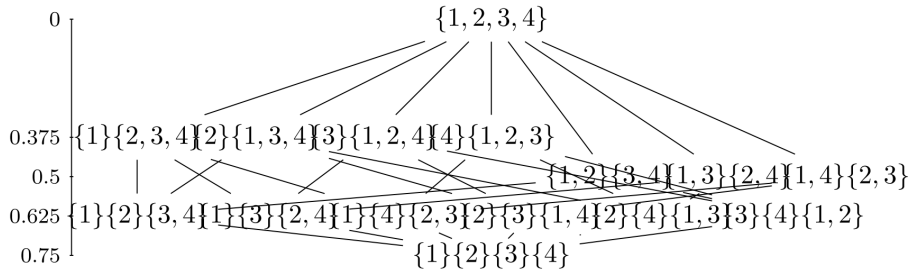


Figura 2.2: Diagramma di Hasse allungato secondo $\tilde{B}$, con una dimensione campionaria pari a $N = 4$. Dal diagramma di Hasse allungato secondo $\tilde{B}$ si può determinare la distanza tra due partizioni qualunque. Esempio: se $\mathbf{c} = (\{1, 2\}, \{3, 4\})$ e $\widehat{\mathbf{c}} = (\{1\}, \{3\}, \{2, 4\})$, allora $\mathbf{c} \wedge \widehat{\mathbf{c}} = (\{1\}, \{2\}, \{3\}, \{4\})$ e $d(\mathbf{c}, \widehat{\mathbf{c}}) = d(\mathbf{c} \wedge \widehat{\mathbf{c}}, \mathbf{1}) - d(\mathbf{c}, \mathbf{1}) + d(\mathbf{c} \wedge \widehat{\mathbf{c}}, \mathbf{1}) - d(\widehat{\mathbf{c}}, \mathbf{1}) = 0.75 - 0.5 + 0.75 - 0.625 = 0.375$.

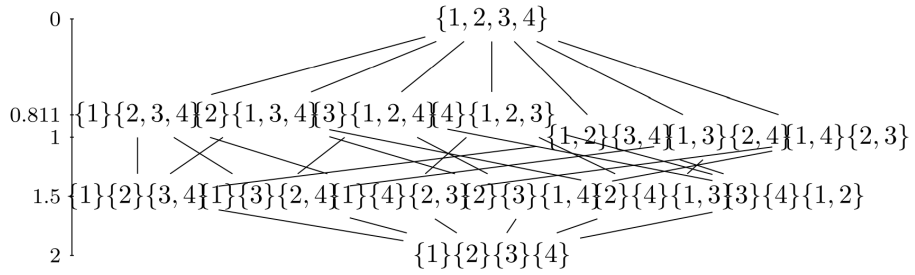


Figura 2.3: Diagramma di Hasse allungato secondo VI, con una dimensione campionaria pari a $N = 4$. Dal diagramma di Hasse allungato secondo VI si può determinare la distanza tra due partizioni qualunque. Esempio: se $\mathbf{c} = (\{1, 2\}, \{3, 4\})$ e $\widehat{\mathbf{c}} = (\{1\}, \{3\}, \{2, 4\})$, allora $\mathbf{c} \wedge \widehat{\mathbf{c}} = (\{1\}, \{2\}, \{3\}, \{4\})$ e $d(\mathbf{c}, \widehat{\mathbf{c}}) = d(\mathbf{c} \wedge \widehat{\mathbf{c}}, \mathbf{1}) - d(\mathbf{c}, \mathbf{1}) + d(\mathbf{c} \wedge \widehat{\mathbf{c}}, \mathbf{1}) - d(\widehat{\mathbf{c}}, \mathbf{1}) = 2 - 1 + 2 - 1.5 = 1.5$.

tra due clustering dipende dalla numerosità campionaria N. Però, il comportamento di questi due limiti è molto differente; per VI il limite tende ad infinito per $N \to \infty$, mentre per $\tilde{B}$ il limite tende a uno per $N \to \infty$. Se N cresce, il numero totale B_N di partizioni aumenta drasticamente. Dunque è naturale pensare che il limite per la distanza tra due clustering sullo spazio delle partizioni debba crescere all'aumentare della dimensione dello spazio. In particolare, $\mathbf{1}$ e $\mathbf{0}$ diventano più distanti per $N \to \infty$, dato che c'è un crescente numero, $B_N - 2$, di partizioni tra questi due estremi; per $\tilde{B}$, la perdita associata alla stima di uno di questi due estremi con l'altro estremo tende al numero fissato uno, mentre per VI, tende ad infinito.

Per i diagrammi di Hasse allungati in Figure 2.2 e 2.3 si possono determinare le partizioni più vicine a $\mathbf{c}$. Per esempio, le partizioni più vicine a $\mathbf{1}$ sono quelle partizioni che dividono $\mathbf{1}$ in due cluster, un singoletto e uno contenente tutte le altre osservazioni; e le partizione più vicine a $(\{1\}, \{2\}, \{3, 4\})$ sono quelle che uniscono le due partizioni più piccole $(\{1, 2\}, \{3, 4\})$ e quelle che dividono il cluster di dimensione due $(\{1\}, \{2\}, \{3\}, \{4\})$.

Proprietà 5. *Per entrambe le metriche VI e $\tilde{B}$, le partizioni più vicine alla partizione $\boldsymbol{c}$ sono:*

- *se $\boldsymbol{c}$ contiene almeno due cluster di dimensione uno e almeno un cluster di dimensione due, le partizioni che uniscono due cluster qualunque di dimensione uno e le partizioni che dividono un cluster qualunque di dimensione due.*
- *se $\boldsymbol{c}$ contiene almeno due cluster di dimensione uno e nessun cluster di dimensione due, le partizioni che uniscono due cluster qualunque di dimensione uno.*
- *se $\boldsymbol{c}$ contiene al massimo un cluster di dimensione uno, le partizioni che dividono il più piccolo cluster di dimensione maggiore di uno in un singoletto e un cluster con tutte le rimanenti osservazioni del cluster originale.*

Una dimostrazione può essere trovata nel Supplementary Material del lavoro di Wade a Ghahramani [2017]. Si nota, dunque, che la distanza tra due partizioni viene definita sulla base della similarità di allocazione delle osservazioni nelle due partizioni, e non esclusivamente sul numero di cluster presenti all'interno delle due partizioni.

La Proprietà 5 caratterizza l'insieme delle partizioni stimate alle quali è associata la più piccola perdita. Sotto entrambe le funzioni di perdita, la più piccola perdita di zero avviene quando la partizione stimata è uguale alla verità. Altrimenti, la più piccola perdita avviene quando il clustering stimato differisce dalla verità unendo due cluster singoletti o dividendo un cluster di dimensione due, o, se nessuno dei due casi è possibile, dividendo il più piccolo cluster di dimensione n in un singoletto e un cluster di dimensione $n-1$. Si nota inoltre che la perdita di stimare il vero clustering con un clustering che unisce due singoletti o divide un cluster di dimensione due è $\frac{2}{N}$ e $\frac{2}{N^2}$ per le metriche VI e $\tilde{B}$ rispettivamente, che convergono a 0 quando $N \to \infty$ per entrambe le metriche, ma ad un ritmo più rapido per $\tilde{B}$.

A seguire, si nota che il diagramma di Hasse allungato da $\tilde{B}$ in Figura 2.2 appare asimmetrico, nel senso che $\mathbf{1}$ è più separato dagli altri quando confrontato con il diagramma di Hasse allungato da VI nella Figura 2.3. La Proprietà 6 riflette l'apparente asimmetria della Figura 2.2.

Proprietà 6. *Supponiamo che N sia divisibile per k, e sia $\boldsymbol{c}_k$ una partizione con k cluster di dimensione uguale N/k.*

$$\tilde{B}(\mathbf{1}, \mathbf{c}_k) = 1 - \frac{1}{k} > \frac{1}{k} - \frac{1}{N} = \tilde{B}(\mathbf{0}, \mathbf{c}_k).$$

$$\mathrm{VI}(\mathbf{1}, \mathbf{c}_k) = \log(k) \leq \log(N) - \log(k) = \mathrm{VI}(\mathbf{0}, \mathbf{c}_k), \quad \textit{per } k \leq \sqrt{N},$$

e

$$\mathrm{VI}(\mathbf{1}, \mathbf{c}_k) = \log(k) \geq \log(N) - \log(k) = \mathrm{VI}(\mathbf{0}, \mathbf{c}_k), \quad \textit{per } k \geq \sqrt{N}.$$

Inoltre, dalla Figura 2.2 si osserva che le partizioni con due cluster di dimensioni uno e tre sono equidistanti tra i due estremi sotto $\tilde{B}$. La seguente proprietà generalizza questa osservazione.

Proprietà 7. *Supponiamo che N sia un numero intero pari e quadrato. Allora, le partizioni con due cluster di dimensioni $n = \frac{1}{2}(N - \sqrt{N})$ e $N - n$ sono equidistanti da $\boldsymbol{1}$ e $\boldsymbol{0}$ sotto $\tilde{B}$.*

Questa proprietà non è auspicabile per una funzione di perdita, poiché afferma che la perdita di stimare una partizione composta da due cluster di dimensioni $\frac{1}{2}(N-\sqrt{N})$ e $\frac{1}{2}(N+\sqrt{N})$ con la partizione di un solo cluster o con la partizione di tutti i singoletti è la stessa. Tuttavia, intuitivamente, $\mathbf{1}$ è una stima migliore.

Il comportamento di VI è molto più ragionevole, poiché le partizioni con due cluster saranno sempre meglio stimate con $\mathbf{1}$ rispetto che con $\mathbf{0}$ per $N > 4$ e le partizioni con $\sqrt{N}$ cluster di dimensioni uguali sono equidistanti da $\mathbf{0}$ e da $\mathbf{1}$.

Infine, si nota che, dato che sia VI che $\tilde{B}$ sono metriche sullo spazio delle partizioni, si può costruire una sfera intorno a $\mathbf{c}$ di dimensione ϵ, definita come:

$$B_\epsilon(\mathbf{c}) = \{\widehat{\mathbf{c}} \in \mathbf{C} : d(\mathbf{c}, \widehat{\mathbf{c}}) \leq \epsilon\}.$$

Dalla Proprietà 5, la sfera non-triviale più piccola, ovvero la più piccola sfera attorno a $\mathbf{c}$ con almeno due partizioni, sarà la stessa per entrambe le metriche. Quando si considera la successiva sfera più piccola, emergono delle differenze; un esempio dettagliato è fornito nel Supplementary Material del lavoro di Wade e Ghahramani [2017]. Secondo l'opinione degli autori, la sfera VI riflette più da vicino l'intuizione del più vicino insieme di partizioni a $\mathbf{c}$.

2.3 Stima puntuale e insiemi di credibilità per la variazione dell'informazione

Come dettagliato nella sezione precedente, sia VI che $\tilde{B}$ condividono diverse proprietà desiderabili, tra cui l'allineamento con il reticolo delle partizioni e la coincidenza nella più piccola sfera non-triviale intorno a qualsiasi partizione. Tuttavia sono emerse anche delle differenze. In particolare, si è riscontrato che $\tilde{B}$ presenta alcune asimmetrie peculiari, preferendo dividere i cluster anziché unirli, mentre si è constatato che la sfera VI riflette più precisamente l'intuizione del vicinato di una partizione. Alla luce di ciò, si propone di utilizzare VI come funzione di perdita nell'analisi bayesiana dei cluster.
Sotto VI, la partizione ottimale $\mathbf{c}^*$ è:

$$\begin{aligned}\mathbf{c}^* &= \arg\min_{\widehat{\mathbf{c}}} \mathbb{E}[\mathrm{VI}(\mathbf{c}, \widehat{\mathbf{c}})|\mathcal{D})] \\ &= \arg\min_{\widehat{\mathbf{c}}} \sum_{n=1}^{N} \log(\sum_{n'=1}^{N} \mathbf{1}(\widehat{c}_{n'} = \widehat{c}_n)) - 2\sum_{n=1}^{N} \mathbb{E}[\log(\sum_{n'=1}^{N} \mathbf{1}(c_{n'} = c_n, \widehat{c}_{n'} = \widehat{c}_n))|\mathcal{D}],\end{aligned}$$

dove $\mathcal{D}$ indica i dati.

Per caratterizzare l'incertezza nella stima puntuale $\mathbf{c}^*$, si propone la costruzione di una sfera di credibilità di un dato livello di credibilità $1 - \alpha$, con

$\alpha \in [0, 1]$, definita come

$$B_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} : d(\mathbf{c}^*, \mathbf{c}) \leq \epsilon^*\},$$

dove ϵ^* è il più piccolo valore $\epsilon \geq 0$ tale che $P(B_\epsilon(\mathbf{c}^*)|\mathcal{D}) \geq 1 - \alpha$. La sfera di credibilità è la più piccola sfera attorno a $\mathbf{c}^*$ con una probabilità a posteriori almeno pari a $1 - \alpha$. La sfera riflette l'incertezza a posteriori nella stima puntuale $\mathbf{c}^*$; con probabilità $1 - \alpha$, si ritiene che il vero clustering sia entro una distanza di ϵ^* dalla stima puntuale $\mathbf{c}^*$, dati i dati. La sfera può essere definita in base a qualsiasi metrica nello spazio delle partizioni, come VI e $\tilde{B}$. Se la più piccola sfera non-triviale secondo VI o $\tilde{B}$ ha una probabilità a posteriori di almeno $1-\alpha$, le sfere di credibilità costruite secondo le due metriche coincideranno (confrontare Proprietà 5). Tuttavia, solitamente le due sfere saranno diverse.

La sfera di credibilità è riassunta dai relativi limiti verticali superiori, verticali inferiori ed orizzontali, definiti, rispettivamente, come le partizioni nella sfera di credibilità con il minor numero di cluster che sono più distanti da $\mathbf{c}^*$, con il maggior numero di cluster che sono più distanti da $\mathbf{c}^*$, e con la maggior distanza da $\mathbf{c}^*$. I limiti sono definiti in modo più formale di seguito, dove la notazione $k(\mathbf{c})$ viene utilizzata per indicare il numero di cluster in $\mathbf{c}$.

Definizione 1 (Limiti verticali superiori). *I limiti verticali superiori della sfera di credibilità $B_{\epsilon^*}(\boldsymbol{c}^*)$, indicati con $v^u_{\epsilon^*}(\boldsymbol{c}^*)$, sono definiti come*

$$v^u_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : k(\mathbf{c}) \leq k(\mathbf{c}') \; \forall \; \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ e } d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}'', \mathbf{c}^*) \; \forall \; \mathbf{c}'' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ con } k(\mathbf{c}) = k(\mathbf{c}'')\}.$$

Definizione 2 (Limiti verticali inferiori). *I limiti verticali inferiori della sfera di credibilità $B_{\epsilon^*}(\boldsymbol{c}^*)$, indicati con $v^l_{\epsilon^*}(\boldsymbol{c}^*)$, sono definiti come*

$$v^l_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : k(\mathbf{c}) \geq k(\mathbf{c}') \; \forall \; \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ e } d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}'', \mathbf{c}^*) \; \forall \; \mathbf{c}'' \in B_{\epsilon^*}(\mathbf{c}^*) \text{ con } k(\mathbf{c}) = k(\mathbf{c}'')\}.$$

Definizione 3 (Limiti orizzontali). *I limiti orizzontali della sfera di credibilità $B_{\epsilon^*}(\boldsymbol{c}^*)$, indicati con $h_{\epsilon^*}(\boldsymbol{c}^*)$, sono definiti come*

$$h_{\epsilon^*}(\mathbf{c}^*) = \{\mathbf{c} \in B_{\epsilon^*}(\mathbf{c}^*) : d(\mathbf{c}, \mathbf{c}^*) \geq d(\mathbf{c}', \mathbf{c}^*) \; \forall \; \mathbf{c}' \in B_{\epsilon^*}(\mathbf{c}^*)\}.$$

Questi limiti descrivono gli estremi della sfera di credibilità e, con una probabilità a posteriori di $1 - \alpha$, quanto si ritiene che la partizione possa differire da $\mathbf{c}^*$. Un esempio è fornito nel Supplementary Material del lavoro di Wade e Ghahramani [2017]. Nella pratica, si definiscono i limiti verticali e orizzontali basandosi sulle partizioni nella sfera di credibilità con una probabilità a posteriori stimata positiva.

Nella letteratura esistente, la quantificazione dell'incertezza nella struttura di clustering viene tipicamente descritta attraverso una mappa di calore della matrice di similarità a posteriori stimata. Tuttavia, a differenza della sfera di credibilità con un livello di confidenza bayesiano $1 - \alpha$, non vi è una quantificazione precisa di quanta incertezza sia rappresentata dalla matrice di similarità a posteriori. Inoltre, le sfere di credibilità hanno l'ulteriore vantaggio di caratterizzare l'incertezza attorno alla stima puntuale $\mathbf{c}^*$.

Capitolo 3

Applicazione dei metodi di clustering

Nella presente sezione si mettono a confronto diversi metodi di clustering su due insiemi di dati, galaxies unidimensionale ed iris con più di due dimensioni. In particolare, si utilizzano metodi di soft clustering e hard clustering: k-medie, clustering gerarchico, clustering bayesiano non parametrico con funzione di perdita di Binder e con la variazione dell'informazione.

3.1 Dataset galaxies

Il dataset galaxies, disponibile nel pacchetto MASS di **R**, contiene un vettore unidimensionale di misurazioni delle velocità in km/sec di 82 galassie della regione della Corona Boreale. La presenza di cluster fornisce evidenza di vuoti e super-ammassi nell'universo lontano.
Un'analisi preliminare del dataset, Figura 3.1, mostra una distribuzione empirica della velocità multimodale: si osservano tre picchi nella densità stimata, con un possibile quarto picco se si considera quello centrale come diviso in due. Quest'analisi fa pensare alla presenza di tre o quattro cluster all'interno dei dati.

L'applicazione dell'algoritmo k-medie, guardando alla percentuale di varianza dei dati spiegata dalla soluzione di clustering in Figura 3.2, suggerisce la presenza di tre (79.6% di varianza spiegata) o di quattro (93.7% di varianza spiegata) cluster all'interno dei dati. Le due soluzioni di clustering sono rappresentate in Figura 3.3, con le osservazioni colorate per cluster di appar-

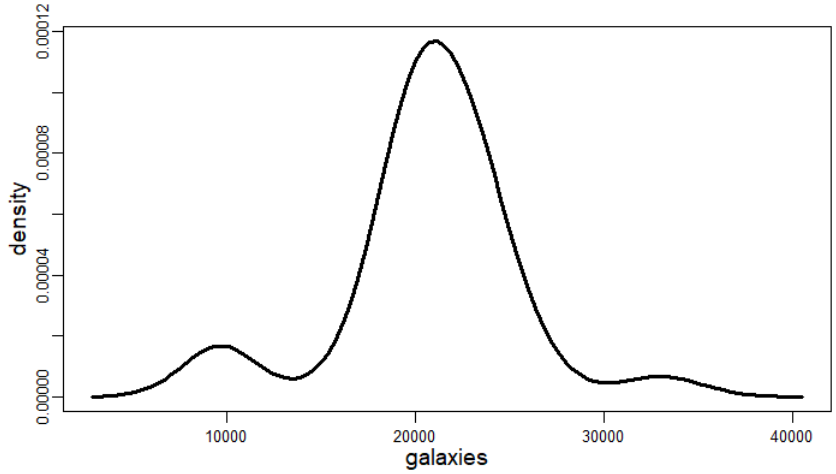


Figura 3.1: [galaxies] stima non parametrica della densità.

tenenza stimato.

Si applica la procedura di clustering gerarchico agglomerativo, considerando tutte le combinazioni delle seguenti scelte di distanze: come distanza tra unità, la distanza Euclidea e la distanza di Manhattan, mentre come distanza tra unità e cluster/unità, legame singolo, completo, medio e metodo di Ward. Si ottengono i medesimi risultati per entrambe la scelte di distanza tra unità: il legame singolo, completo e medio forniscono la medesima partizione delle unità in tre cluster, mentre il metodo di Ward divide le unità in quattro cluster, Figura 3.4.

In un contesto bayesiano non parametrico, si modellano i dati con un processo di Dirichlet. Si utilizza un campionatore a posteriori [Corradin et al., 2021], con parametri di default della funzione `PYdensity()`, per estrarre 10000 campioni, dopo aver eliminato i primi 1000 come rodaggio, dalla distribuzione sullo spazio delle partizioni. La Figura 3.5 rappresenta le stime puntuali trovate dall'algoritmo di ricerca greedy per la funzione di perdita di Binder e per la VI. I punti sono rappresentati contro la stima dei valori della densità del modello mistura di Dirichlet. Si osserva che la perdita di Binder tende a collocare osservazioni che hanno un'allocazione incerta in cluster singoletti, per un totale di cinque cluster, due dei quali sono singoletti; mentre la soluzione VI è costituita da tre cluster, di cui nessun singoletto.

Una sfera di credibilità al 95% contiene tutte quelle partizioni attorno alla

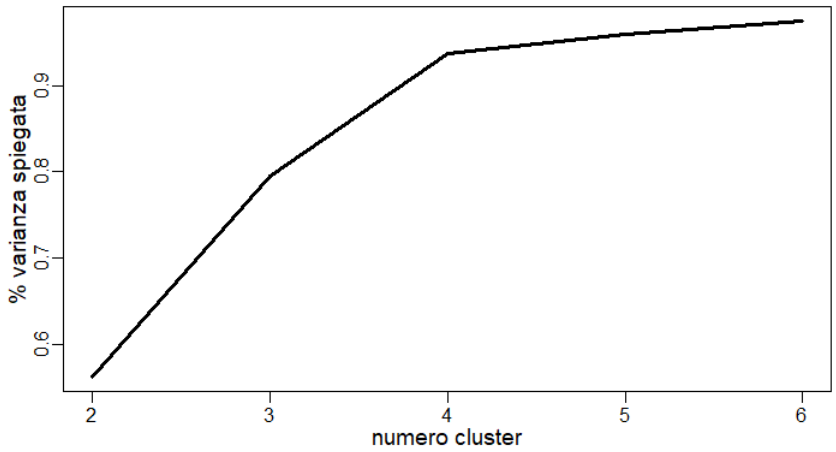
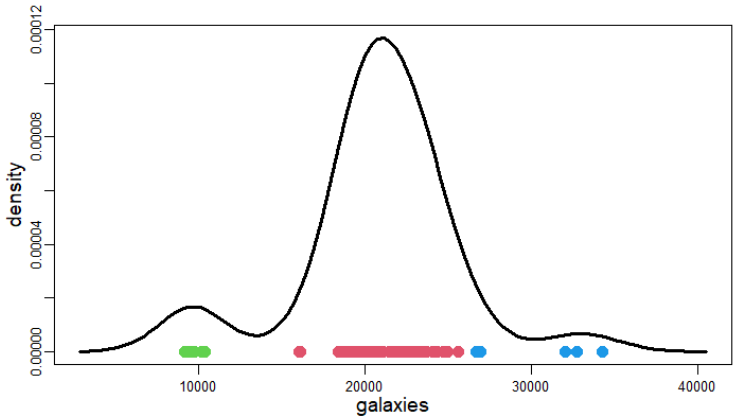
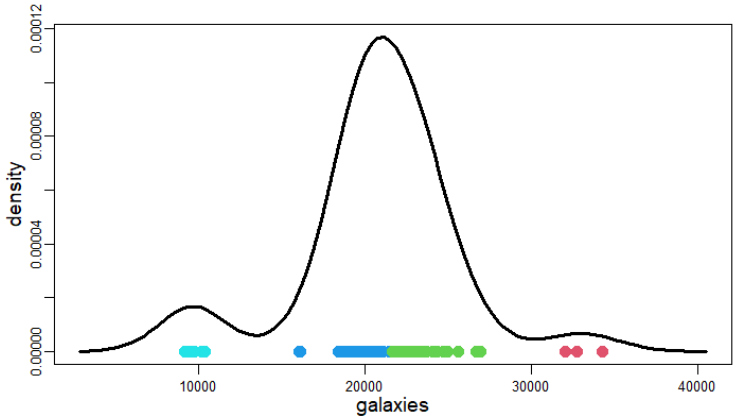


Figura 3.2: [galaxies] k-medie, percentuale di varianza spiegata per numero di cluster.

stima puntuale $\mathbf{c}^*$ che si ritiene, con una credibilità del 95%, possano essere il vero clustering. Si utilizzano le funzioni proposte in Wade and Ghahramani [2018] per calcolare e rappresentare la sfera di credibilità per VI tramite i limiti verticali superiori, verticali inferiori ed orizzontali (Figura 3.6). Si osserva una grande variabilità attorno alla stima puntuale della partizione. Con una probabilità a posteriori del 95%, si crede che, ad un estremo, si possano modellare i dati usando solo due componenti, di cui una con una grande varianza, che possa tenere conto dei valori anomali (il cluster rosso in Figura 3.6a). All'altro estremo, i dati possono essere divisi in un cluster di medie dimensioni (il cluster arancione in Figura 3.6b) e tanti, 12, piccoli cluster. Il limite orizzontale assegna un solo cluster a ciascuna delle due mode laterali della distribuzione, ma divide la moda centrale parzialmente nei due cluster laterali, e per la restante parte in altri tre cluster. La Figura 3.6d enfatizza come la matrice di similarità a posteriori sottorappresenti l'incertezza attorno alla stima puntuale, in confronto alla sfera di credibilità.

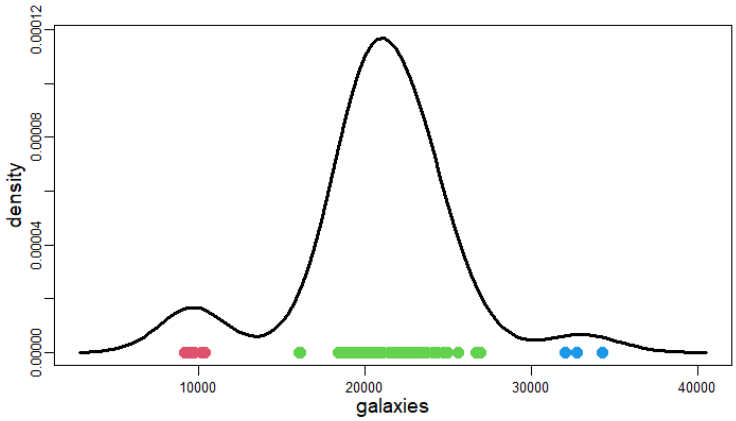


(a) K-medie, 3 cluster

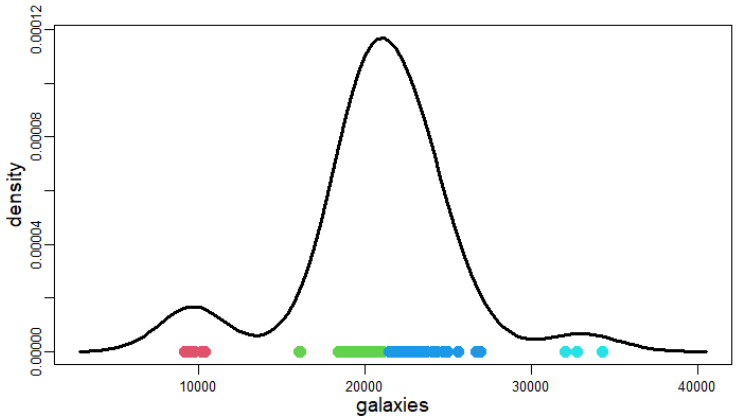


(b) K-medie, 4 cluster

Figura 3.3: [galaxies] partizioni dei dati ottenute con k-medie, con 3 e 4 cluster, rispettivamente.

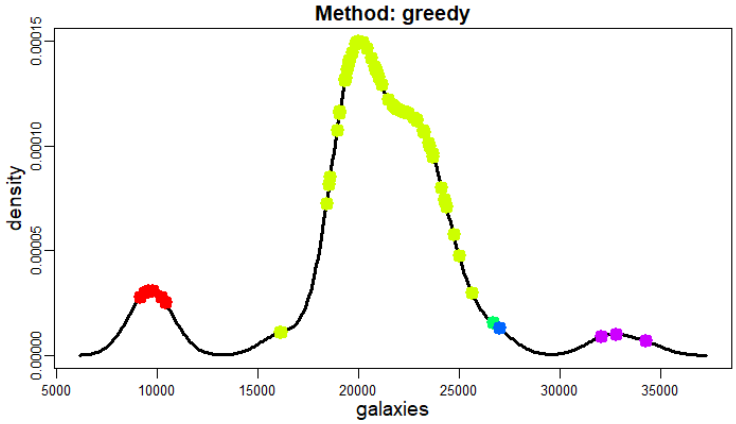


(a) Legame singolo, 3 cluster

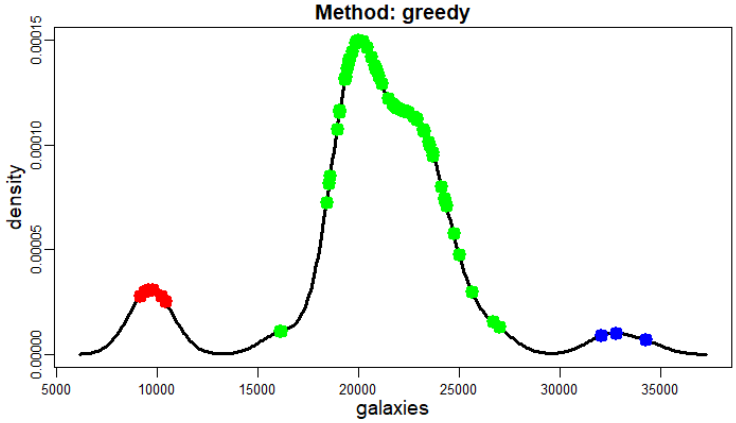


(b) Metodo di Ward, 4 cluster

Figura 3.4: [galaxies] partizioni dei dati ottenute con clustering gerarchico agglomerativo con distanza Euclidea, con legame singolo (a) e con metodo di Ward (b).

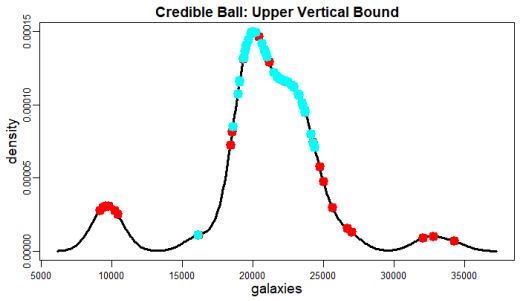


(a) Perdita di Binder, 5 cluster

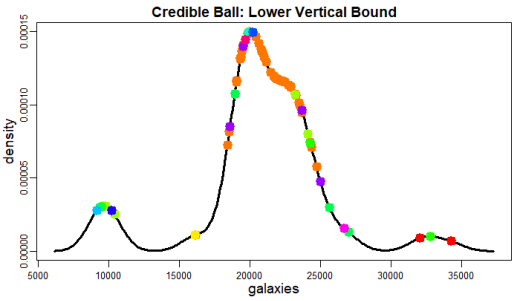


(b) VI, 3 cluster

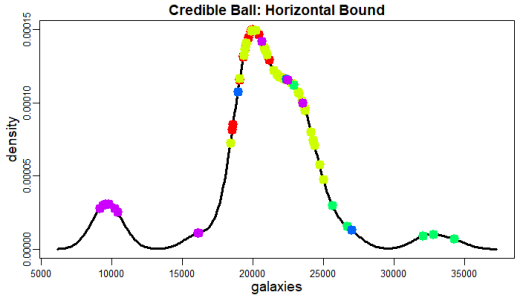
Figura 3.5: [galaxies] stime puntuali per la partizione, per Binder e per VI.



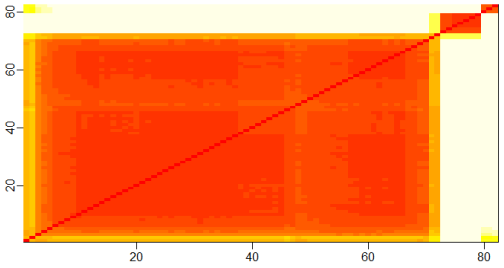
(a) Limite verticale superiore, 2 cluster



(b) Limite verticale inferiore, 13 cluster



(c) Limite orizzontale, 5 cluster



(d) Matrice di similarità a posteriori

Figura 3.6: [galaxies] sfera di credibilità al 95% con VI, rappresentata da (a) il limite verticale superiore, (b) il limite verticale inferiore e (c) il limite orizzontale; e (d), una mappa di calore della matrice di similarità a posteriori.

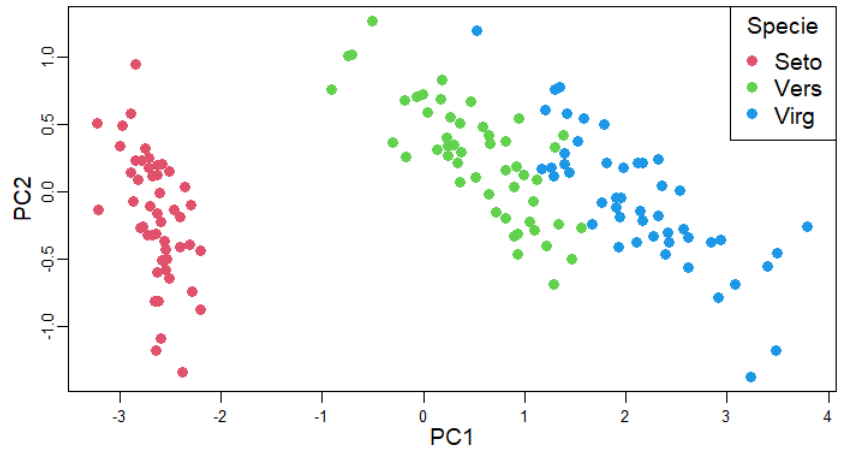


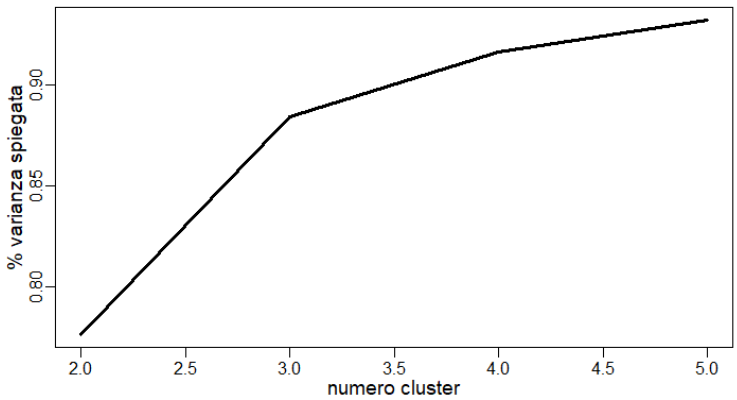
Figura 3.7: [iris] insieme di dati rappresentato sul piano delle prime due componenti principali, con colore corrispondente alla specie di appartenenza.

3.2 Dataset iris

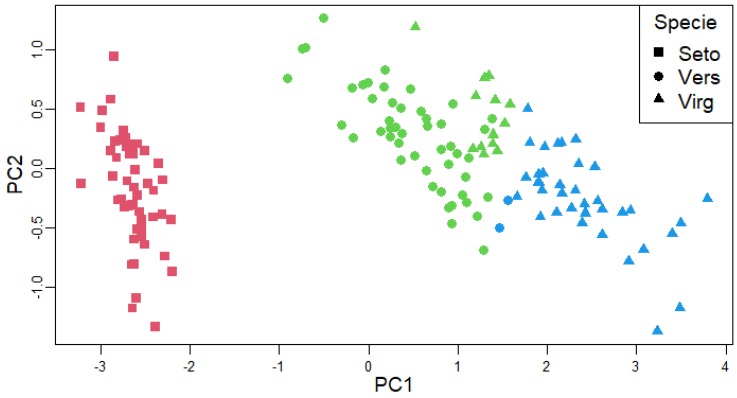
Il dataset iris contiene le misurazioni in centimetri delle variabili lunghezza e larghezza del sepalo e lunghezza e larghezza del petalo, per 150 fiori, 50 per ciascuna delle tre seguenti specie: *Iris setosa*, *versicolor* e *virginica*.
Un'analisi preliminare del dataset sul piano delle prime due componenti principali, Figura 3.7, mostra una chiara separazione della specie *setosa* dalle altre due; *versicolor* e *virginica* sono quasi interamente separate, all'infuori di alcune osservazioni nell'area centrale tra le due specie.

L'algoritmo k-medie riesce a cogliere bene la presenza di tre cluster nei dati, soluzione che porta ad avere una percentuale di varianza dei dati spiegata dalla partizione pari all'88.4% (Figura 3.8). Questa soluzione sbaglia nella classificazione di 16 unità, in particolare: *setosa*, la nuvola di punti separata dal resto delle osservazioni, viene tutta correttamente allocata; della specie *versicolor*, due sole osservazioni vengono classificate come *virginica*; mentre per *virginica*, 14 osservazioni vengono classificate come *versicolor*.

Si applica la procedura di clustering gerarchico agglomerativo, considerando, anche per questo insieme di dati, tutte le combinazioni delle seguenti scelte di distanze: come distanza tra unità, la distanza Euclidea e la distanza di Manhattan, mentre come distanza tra unità e cluster/unità, legame singolo,



(a) Percentuale di varianza spiegata per numero di cluster



(b) Partizione con tre cluster

Figura 3.8: [iris] k-medie, (a) percentuale di varianza spiegata per numero di cluster, (b) partizione con tre cluster, con colore corrispondente alla specie stimata di appartenenza.

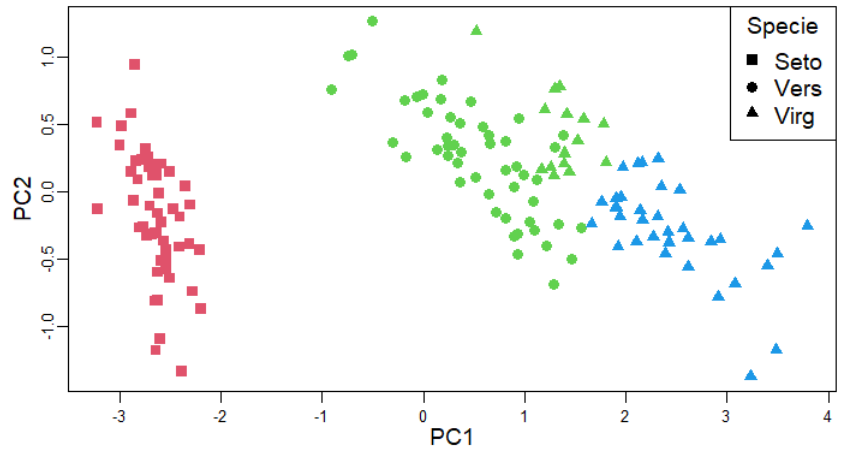
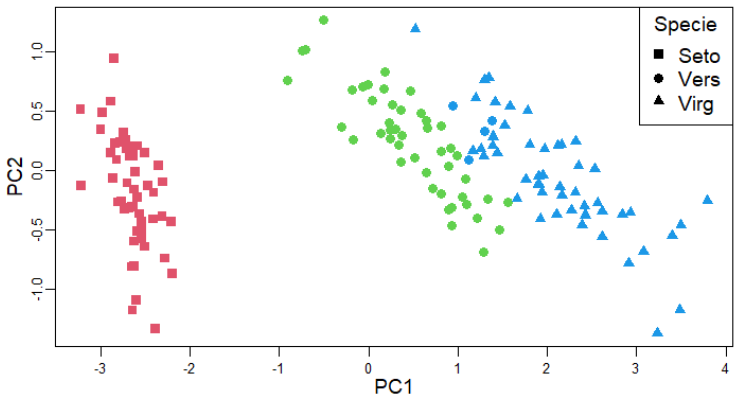


Figura 3.9: [iris] clustering gerarchico con distanza di Manhattan e legame completo, con colore corrispondente alla specie di appartenenza stimata.

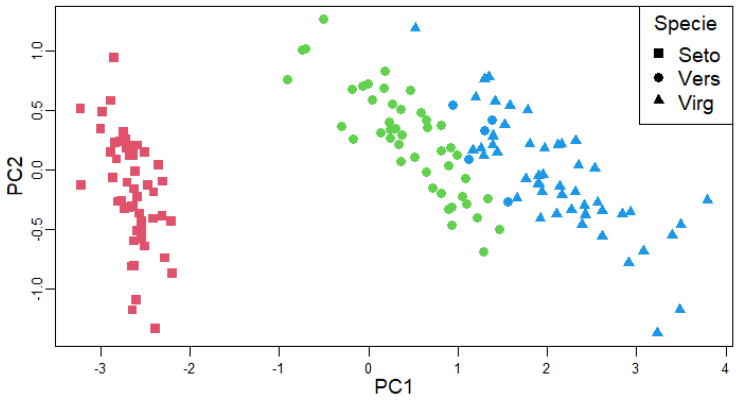
completo, medio e metodo di Ward.
Il dendrogramma più appropriato per un taglio a tre cluster, con il minor numero di unità non correttamente allocate, sembra essere quello ottenuto con distanza di Manhattan e legame completo (non riportato): l'errata allocazione avviene per 16 unità della specie *virginica*, che vengono invece collocate in *versicolor* (Figura 3.9).

In un contesto bayesiano non parametrico, si modellano i dati con un processo di Dirichlet. Si utilizza un campionatore a posteriori [Corradin et al., 2021], con parametri di default della funzione `PYdensity()`, per estrarre 10000 campioni, dopo aver eliminato i primi 1000 come rodaggio, dalla distribuzione sullo spazio delle partizioni. La Figura 3.10 rappresenta le stime puntuali trovate dall'algoritmo di ricerca greedy per la funzione di perdita di Binder e per la VI. Entrambe le funzioni di perdita commettono degli errori di allocazione delle unità: Binder assegna quattro unità di *versicolor* in *virginica*, mentre VI commette lo stesso errore, ma per cinque unità.

Una sfera di credibilità al 95% contiene tutte quelle partizioni attorno alla stima puntuale $\mathbf{c}^*$ che si ritiene, con una credibilità del 95%, possano essere il vero clustering. Si utilizzano le funzioni proposte in Wade and Ghahramani [2018] per calcolare e rappresentare la sfera di credibilità per VI tramite i limiti verticali superiori, verticali inferiori ed orizzontali. Si sono trovati quattro limiti verticali superiori con tre cluster ognuno, un limite verticale inferiore



(a) Perdita di Binder, 3 cluster



(b) VI, 3 cluster

Figura 3.10: [iris] stima puntuale per la partizione, per Binder e per VI, con colore corrispondente alla specie di appartenenza stimata.

Metodo	Indice di Rand
k-medie	0.88
clustering gerarchico	0.88
perdita di Binder	0.97
VI	0.96

Tabella 3.1: Indice di Rand per ogni metodo utilizzato.

con cinque cluster ed 89 limiti orizzontali con quattro cluster ciascuno. Con una probabilità a posteriori del 95%, dunque, si crede che i dati si possano modellare usando tra i tre e i cinque cluster.

In questo esempio non si è riscontrata un'evidente differenza tra l'uso della funzione di perdita di Binder e VI, in quanto entrambe le metriche hanno portato ad una stima puntuale di una partizione di tre cluster. Nonostante ciò, per questa impostazione del processo di Dirichlet sottostante i dati, si è riscontrata una notevole riduzione del numero di osservazioni non correttamente allocate, rispetto ai metodi di hard clustering: su un totale di 150 fiori, i metodi di hard clustering ne hanno allocati 16 non correttamente, mentre i metodi di clustering bayesiani non parametrici hanno ridotto questo numero a 4 o 5. Si riporta nella Tabella 3.1 l'indice di Rand, [Rand, 1971], calcolato per ciascun metodo di clustering utilizzato.

Conclusione

Diversi metodi di clustering sono largamente diffusi ed utilizzati, tra i quali clustering gerarchico e non gerarchico, k-medie, e clustering basato su modelli mistura finiti. Questi metodi, nonostante si dimostrino efficaci in numerose applicazioni, presentano delle limitazioni: esplorano solo un ristretto sottoinsieme dello spazio delle partizioni o necessitano della specificazione del numero di cluster a priori, restituiscono una sola soluzione di clustering o non tengono conto dell'incertezza nella stima dei parametri del modello.

L'analisi dei cluster bayesiana offre un vantaggio rispetto all'analisi classica dei cluster, in quanto la procedura bayesiana restituisce una distribuzione a posteriori sull'intero spazio delle partizioni, riflettendo l'incertezza nella struttura di clustering dati i dati, invece di restituire una singola soluzione o di condizionarsi alle stime dei parametri e al numero di cluster. Ciò consente di valutare le proprietà statistiche del clustering dati i dati. Tuttavia, a causa dell'ampia dimensione dello spazio delle partizioni, un punto importante nell'analisi dei cluster bayesiana è come riassumere adeguatamente la distribuzione a posteriori. Per affrontare questo problema, in questo elaborato si sono ripercorsi i risultati di Wade and Ghahramani [2018] nello sviluppo di strumenti utili ad ottenere una stima puntuale del clustering basata sulla distribuzione a posteriori, e per descrivere l'incertezza intorno a questa stima tramite una sfera di credibilità al 95%.

Ottenere una stima puntuale attraverso un quadro teorico formale di teoria della decisione richiede la specificazione di una funzione di perdita. Nella letteratura precedente, l'attenzione era posta sulla funzione di perdita di Binder. Si propone di utilizzare una misura informativa, la variazione dell'informazione, come funzione di perdita. Si è riscontrato che la funzione di perdita di Binder presenta asimmetrie peculiari, preferendo la divisione rispetto alla fusione dei cluster, mentre la variazione dell'informazione è più simmetrica in questo senso. Questo comportamento della funzione di perdita di Binder fa sì che la partizione ottimale sovrastimi il numero di cluster, assegnando osservazioni

incerte a piccoli cluster aggiuntivi.

Per rappresentare l'incertezza intorno alla stima puntuale, si sono costruite sfere di credibilità al 95% intorno alla stima puntuale, rappresentate tramite i limiti verticali superiori, verticali inferiori ed orizzontali. In aggiunta rispetto ad una mappa di calore della matrice di similarità a posteriori, che è spesso riportata in letteratura, una sfera di credibilità al 95% arricchisce la comprensione dell'incertezza presente. Infatti, la sfera fornisce una quantificazione precisa dell'incertezza presente intorno alla stima puntuale e, negli esempi, si è riscontrato che un'analisi basata esclusivamente sulla matrice di similarità a posteriori può portare ad essere troppo certi rispetto alla struttura di clustering trovata.

Bibliografia

David A Binder. Bayesian cluster analysis. *Biometrika*, 65(1):31–38, 1978.

Charles Bouveyron, Gilles Celeux, T Brendan Murphy, and Adrian E Raftery. *Model-based clustering and classification for data science: with applications in R*, volume 50. Cambridge University Press, 2019.

Riccardo Corradin, Antonio Canale, and Bernardo Nipoti. BNPmix: An R package for Bayesian nonparametric modeling via Pitman-Yor mixtures. *Journal of Statistical Software*, 100:1–33, 2021.

Thomas S. Ferguson. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):209–230, 1973.

Chris Fraley and Adrian E Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108, 1979.

Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453): 161–173, 2001.

Antonio Lijoi and Igor Prünster. Models beyond the Dirichlet process. *Bayesian nonparametrics*, 28(80):342, 2010.

Marina Meilă. Comparing clusterings—an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

James B Nation. Notes on lattice theory, 1998.

Jim Pitman. Poisson-kingman partitions. *Lecture Notes-Monograph Series*, pages 1–34, 2003.

Jim Pitman and Marc Yor. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*, 25(2):855–900, 1997.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Sara Wade and Zoubin Ghahramani. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2), 2018.