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

Statistica, Economia e Finanza

La riduzione della dimensionalità nel clustering basato su modello: un'analisi mediante simulazione e un'applicazione a dati genetici

Relatore: dott.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Alessandro Casa
Matricola n.: 1032854

Anno Accademico 2013/2014

Indice

Capitolo 1

Introduzione

Negli ultimi decenni, con l'avvento e il successivo sviluppo dei calcolatori, abbiamo assistito ad una vera e propria rivoluzione tecnologica che ha radicalmente cambiato non solamente il modo di operare in campo scientifico ma anche il modo di pensare e di lavorare in moltissimi ambiti della vita quotidiana.

L'avvento dei computer, l'abbattimento dei costi derivanti dal loro utilizzo e lo sviluppo di tecnologie a questi collegate hanno fatto in modo che sia diventato via via sempre più semplice raccogliere dati riguardanti i più svariati fenomeni. Si è di conseguenza assistito negli ultimi tempi ad una vera e propria esplosione della quantità di dati disponibili. La disponibilità di una tale mole di dati porta notevoli vantaggi in termini informativi ma è altrettanto facile capire come, di pari passo, aumentino anche il rumore e quindi l'informazione irrilevante contenuta in questi dati. Da ciò si può quindi notare come questa rivoluzione introduca delle nuove problematiche e delle nuove sfide riguardanti la necessità di introdurre dei metodi che permettano di discriminare tra ciò che è informazione utile e ciò che non lo è.

Dal momento in cui lo scopo principale della statistica è quello di estrarre conoscenza da un insieme di dati, risulta evidente come alcune di queste sfide riguardino direttamente questa disciplina. Di pari passo con questa esplosione di dati si è infatti sviluppato il concetto di *data mining*, una branca della statistica che si occupa, attraverso l'utilizzo di tencniche multivariate, di *scavare* nei dati alla ricerca di informazione utile. Una delle tecniche più utilizzate in questo ambito è il *clustering*, o analisi dei gruppi.

Con il termine *clustering* si fa riferimento ad un ampia classe di metodi aventi lo scopo di individuare delle strutture intrinseche nei dati e, in particolare, di trovare gruppi di osservazioni che presentino caratteristiche simili al proprio interno e differenti tra i diversi gruppi. Tipicamente questo obiettivo viene perseguito uti-

lizzando delle tecniche basate su princìpi geometrici ed euristici che differenziano e raggruppano le osservazioni basandosi su concetti di dissimilarità e di distanza. Utilizzando un approccio di questo tipo si riscontrano però alcune criticità importanti principalmente riguardanti il fatto che queste tecniche, non basandosi formalmente su procedure statistiche, non presentano proprietà matematiche ben definite e non permettono di avvalersi di procedure inferenziali per risolvere alcune questioni rilevanti che emergono solitamente dall'analisi dei gruppi. Per superare questo problema sono stati introdotti degli altri metodi, tra cui il *clustering* basato su modelli statistici parametrici e, in particolare, su modelli a mistura finita. Questi metodi collocano l'analisi di raggruppamento in un contesto statisticamente più rigoroso permettendo così una più agevole risposta a domande riguardanti, ad esempio, il numero di gruppi presenti nei dati ed il grado di incertezza di una data partizione.

Nel momento in cui ci si trova ad operare in spazi ad elevata dimensionalità il *clustering* basato su modello presenta dei limiti evidenti. Come sarà messo in evidenza infatti, l'approccio all'analisi dei gruppi basato sui modelli a mistura finita risulta soffrire di un'evidente problema riguardante la rapidità con cui cresce il numero di parametri da stimare all'aumentare della dimensionalità dello spazio in cui si opera, non del tutto risolvibile attraverso l'utilizzo di modelli parsimoniosi. Questo fa sì che, in queste situazioni, prima di poter applicare l'algoritmo di *clustering* siano necessarie delle operazioni di riduzione della dimensione che riescano a discriminare tra variabili rilevanti e variabili irrilevanti e, in generale, a selezionare dei sottospazi di dimensione inferiore sui quali poi operare. Qualora queste operazioni riescano effettivamente a raggiungere il loro obiettivo, il *clustering* basato su modello riesce a fornire delle partizioni qualitativamente migliori e quindi ad agevolare la comprensione della struttura interna ai dati e del fenomeno studiato. Alcune di queste tecniche di riduzione della dimensione inoltre, andando a selezionare le variabili, rendono i risultati più facilmente interpretabili permettendo anche di comprendere quali caratteristiche differenzino in maniera più marcata gli eventuali gruppi presenti.

Lo scopo che questo lavoro si prefigge è quello di studiare il comportamento del *clustering* basato su modelli a mistura finita nel momento in cui si cerchi un raggruppamento di dati aventi un'elevata dimensionalità e di valutare la bontà di metodi differenti di riduzione della dimensione nell'affrontare i problemi che si riscontrano in queste situazioni.

La trattazione si sviluppa come segue.

Nel secondo capitolo si pone l'attenzione sui limiti e sulle criticità presentate dagli approcci classici all'analisi di raggruppamento mostrando come, andando ad inserire questo tipo di analisi in un contesto più statisticamente rigoroso di cui siano note le proprietà matematiche, si possano superare alcuni di questi limiti. In particolare, vengono presentati i modelli a mistura finita e l'approccio al *clustering* basato su modello statistico parametrico.

Nel terzo capitolo si illustrano i problemi e le sfide che si incontrano nell'applicazione di metodi statistici a dati ad elevata dimensionalità - concetti che rientrano sotto il cappello generale di *maledizione della dimensionalità*. Viene mostrato in particolare come l'approccio al *clustering* basato su modello presenti dei limiti in queste situazioni in quanto il numero di parametri da stimare cresce rapidamente all'aumentare della dimensione dello spazio in cui i dati sono definiti. Vengono quindi presentati alcuni metodi di riduzione della dimensionalità che, cercando di estrarre l'informazione utile ai fini della classificazione dei dati in gruppi, hanno in questo frangente lo scopo di superare questi limiti. In questo lavoro ci si è concentrati in particolare sull'analisi delle componenti principali, l'analisi delle componenti principali sparse e il metodo di selezione delle variabili proposto da Raftery e Dean (2006).

Nel quarto capitolo si è infine cercato di comprendere ed osservare, attraverso delle analisi numeriche, il comportamento del *clustering* basato su modello e dei metodi di riduzione della dimensione nel momento in cui si affrontano i problemi legati all'elevata dimensionalità dei dati. Viene presentato uno studio di simulazione che è stato condotto con il principale obiettivo di confrontare il comportamento di diversi metodi di riduzione e di selezione delle variabili, utilizzati preliminarmente al *clustering* basato su modello. Inoltre il *clustering* basato su modello e i metodi di riduzione della dimensione sono stati applicati a dati provenienti da un'analisi di *microarray*.

Si conclude infine con alcune considerazioni di ordine generale volte a fornire un'analisi critica dei risultati ottenuti, cercando di evidenziare i pregi e i difetti dei metodi sui quali questo lavoro è incentrato e mettendo in luce eventuali ulteriori spunti.

Capitolo 2

L'approccio parametrico all'analisi di raggruppamento

2.1 Limiti dei metodi di raggruppamento basati su distanza

Con il termine *clustering* si intende un insieme di tecniche di analisi multivariata che permettono di trovare delle strutture nei dati e, in particolare, di fornirne un partizionamento in gruppi. Lo scopo principale è quello di trovare dei gruppi all'interno dei quali le unità statistiche presentino caratteristiche omogenee e che, allo stesso tempo, si differenzino in maniera rilevante dalle unità statistiche appartenenti agli altri gruppi (si cerca la coesione interna e l'isolamento esterno, Cormack 1971).

L'approccio tradizionale all'analisi dei gruppi si basa principalmente su procedure e concetti di tipo geometrico ed euristico e, in particolare, richiede la definzione formale del concetto di dissimilarità, usualmente mutuato dal concetto di distanza tra osservazioni. L'idea alla base consiste nel cercare di formare dei gruppi andando a minimizzare la distanza tra le osservazioni appartenenti allo stesso gruppo e massimizzando la distanza tra le osservazioni apppartenenti a gruppi differenti.

Nei metodi di raggruppamento appartenenti a questo tipo di approccio si può distinguere tra *metodi di partizione* e *metodi gerarchici*.

Nei *metodi di partizione* si vanno a creare delle suddivisioni dello spazio campionario in un numero di costituenti definito a priori e che devono essere valutate secondo una qualche funzione obiettivo. Probabilmente la procedura di partizione più popolare è il metodo delle *k-medie* (MacQueen *e altri*, 1967). Questo metodo parte da una suddivisione delle unità in k *cluster* iniziali e calcola poi i centroidi (medie) di questi *cluster* riallocando successivamente le unità al *cluster* dal cui cen-

troide presentano distanza minima e prosegue iterativamente fino al momento in cui non si hanno ulteriori cambiamenti nella partizione dei dati. Il metodo delle *k-medie* presenta delle limitazioni in quanto non garantisce la convergenza all'ottimo assoluto e non garantisce che, partendo da una partizione iniziale differente, converga alla stessa partizione finale dei dati. Un metodo alternativo, proposto per ovviare al problema della non robustezza della media, è il *metodo dei medoidi* proposto da Kaufman e Rousseeuw (1987), che procede in maniera analoga al metodo delle *k-medie* utilizzando però i medoidi in sostituzione ai centroidi.

I *metodi gerarchici* creano un insieme di partizioni nidificate valutate successivamente sulla base di qualche criterio che, basandosi comunque sulle distanze o dissimilarità tra osservazioni, regola il modo in cui i gruppi vengono agglomerati o divisi. Tra i criteri più utilizzati possiamo citare: il *metodo del legame singolo*, il *metodo del legame completo* e il *metodo del legame medio*. In questi casi si aggregano due gruppi quando è minima la distanza tra le osservazioni più vicine, più lontane, o rispettivamente quando è minima la distanza media tra le osservazioni appartenenti a gruppi diversi. Si noti che, utilizzando criteri differenti, si perviene a partizioni nidificate differenti e che non è possibile definire un metodo migliore in assoluto.

Per ulteriori approfondimenti riguardo i metodi gerarchici e i metodi di partizione, la cui spiegazione esaustiva non è lo scopo di questo lavoro, si rimanda per esempio a Mardia *e altri* (1980), Azzalini e Scarpa (2004).

A fronte di una semplicità concettuale, i metodi menzionati presentano delle criticità di un certo rilievo. Innanzitutto i metodi di partizione e i metodi gerarchici non affrontano in nessun modo il problema riguardante il numero di gruppi nei quali suddividere i dati. Se nei metodi di partizione il numero di gruppi va espressamente specificato a priori, creando quindi dei problemi qualora non si avessero informazioni a riguardo, nei metodi gerarchici non viene posto il problema, dal momento in cui si creano delle partizioni nidificate, ma è successivamente compito di colui che analizza i risultati decidere, basandosi su criteri spesso euristici e grafici e quindi non formalmente statistici, il numero di *cluster*. Si può far notare inoltre il fatto che questi metodi presentano delle severe limitazioni di tipo computazionale al crescere della numerosità campionaria sia dal punto di vista puramente di calcolo che di archiviazione della memoria in quanto, basandosi sulla distanza tra le coppie di osservazioni, richiedono il calcolo della matrice di dissimilarità che ha un numero di elementi che cresce con il quadrato della dimensione del campione. Un altro

problema che si può riscontrare utilizzando i metodi presentati risiede nel fatto che, ognuno di questi metodi, tende a suddividere i dati in gruppi aventi forme distorte sulla base del metodo al quale si fa riferimento: ad esempio con il *metodo del legame singolo* spesso si ha l'effetto catena, si accorpano cioè osservazioni anche molto lontane nello spazio purchè tra esse esista una successione di punti che li lega. Con il *metodo del legame completo* e con le *k-medie* si ottengono tendenzialmente gruppi di forma in generale ipersferica rischiando così di non cogliere gruppi di forma irregolare. Inoltre si può far notare che i metodi basati su distanza producono una suddivisione in gruppi anche nelle situazioni in cui i dati non presentino una struttura che fornisce particolari indicazioni in tal senso.

Il problema fondamentale dei metodi basati su distanza è l'assenza di principi statistici alla base di questi. Questo fa sì che le proprietà statistiche di questi approcci siano generalmente sconosciute precludendo così la possibilità di risolvere alcune importanti questioni che emergono dall'analisi dei gruppi quali scegliere il numero di gruppi e scegliere quale metodo sia preferibile. Da quanto detto risulta quindi ovvia l'impossibilità di utilizzare procedure inferenziali associate a questi metodi.

2.2 Una formalizzazione del problema di raggruppamento: i modelli a mistura finita

Finora si è visto come i metodi basati su distanza presentino un certo numero di limitazioni nell'affrontare problemi riguardanti l'analisi dei gruppi. L'impossibilità di risolvere alcune importanti questioni in maniera statisticamente rigorosa e con tecniche di natura inferenziale, e di conoscere le proprietà statistiche di questi metodi, costituiscono i problemi probabilmente più rilevanti che portano a ricercare altri metodi di raggruppamento. Si avverte quindi la necessità di procedure basate su criteri statistici più formali e che permettano di affrontare il *clustering* inserendolo in un contesto più rigoroso. A sostegno di questa esigenza si riporta il pensiero di Aitkin *e altri* (1981): "*when clustering samples from a population, no cluster analysis is a priori believable without a statistical model*".

L'approccio illustrato in questo lavoro assume che le osservazioni a disposizione, $\{x_1, \ldots, x_n\}$, $x_i = (x_{i1}, \ldots, x_{ip})$, per $i = 1, \ldots, n$, definite in uno spazio p-dimensionale, siano un campione di realizzazioni indipendenti e identicamente di-

stribuite da una distribuzione di probabilità f, appartenente ad una famiglia parametrica e dunque, caratterizzata da un certo numero di parametri da stimare. In particolare, f rappresenta una opportuna combinazione di un certo numero di distribuzioni omogenee, ciascuna delle quali definisce un gruppo. La scelta più comune per f ricade sulla famiglia di modelli a mistura finita.

La formulazione generale di un modello a mistura finita con K sottopopolazioni è

$$f(x) = \sum_{k=1}^{K} \pi_k f_k(x) \tag{2.1}$$

dove π_k sono detti *proporzioni della mistura* e devono garantire le seguenti proprietà

$$0 \leq \pi_k \leq 1 \qquad\qquad (k = 1, \ldots, K) \tag{2.2}$$

e

$$\sum_{k=1}^{K} \pi_k = 1 \tag{2.3}$$

e dove $f_k(\cdot)$ è la distribuzione di probabilità per la k-esima componente. Spesso le sottopopolazioni sono modellate come provenienti dalla stessa famiglia di distribuzioni, in questo caso la (2.1) può essere riscritta come

$$f(x) = \sum_{k=1}^{K} \pi_k f(x|\theta_k) \tag{2.4}$$

dove θ_k è il vettore dei parametri per la k-esima componente. Conseguentemente, date delle unità statistiche $x = \{x_1, \ldots, x_n\}$, la log-verosimiglianza del modello-mistura è

$$l(\theta; x) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k f(x_i; \theta_k) \right). \tag{2.5}$$

I modelli a mistura finita forniscono un approccio matematico rigoroso, particolarmente utile per modellare un'ampia gamma di fenomeni differenti e di distribuzioni di forma sconosciuta. Il loro grande sviluppo è infatti dovuto alla loro estrema flessibilità: utilizzando questi modelli si riescono a modellare per esempio distribuzioni multimodali o caratterizzate da asimmetria. Il prezzo da pagare, nel momento in cui si utilizza questa classe di modelli, è legato alla loro maggiore complessità rispetto ad altri metodi.

2.2.1 Stima dei parametri del modello

Le prime analisi basate su modelli a mistura finita risalgono alla fine del XIX secolo (Pearson, 1894) ma ebbero poco seguito, fondamentalmente a causa della complessità dei modelli e degli sforzi computazionali necessari per stimarli. Si noti infatti che la forma della funzione di verosimiglianza per un campione generato dalla funzione di densità 2.4 è solitamente complicata e multi-modale e si presta raramente a soluzioni analitiche in forma chiusa. Il loro successivo sviluppo è quindi dovuto all'avvento dei computer e all'introduzione dell'algoritmo EM. Pur non essendo specificatamente dedicato a questi modelli l'algoritmo EM ha infatti permesso una semplificazione del processo di stima e una maggiore comprensione delle proprietà e risulta essere senza dubbio la tecnica di stima più popolare in questo ambito.

L'algoritmo EM (Expectation-Maximization) (Dempster *e altri*, 1977; McLachlan e Basford, 1988; McLachlan e Krishnan, 2007) è un approccio generale al processo di stima della massima verosimiglianza nei casi in cui si sia in presenza di dati incompleti. Questa formulazione risulta particolarmente conveniente nell'ambito dell'analisi dei gruppi in quanto possiamo considerare come dati "completi" $y_i = (x_i, z_i)$, dove $z_i = (z_{i1}, \ldots, z_{iK})$ con

$$z_{ik} = \begin{cases} 1 & \text{se } x_i \text{ appartiene alla k-esima componente} \\ 0 & \text{altrimenti.} \end{cases} \tag{2.6}$$

Le assunzioni importanti che si fanno sono che la densità di un'osservazione x_i, dato z_i, sia data da $\prod_{k=1}^{K} f(x_i|\theta_k)^{z_{ik}}$ e che le z_i siano indipendenti e identicamente distribuite come delle distribuzioni multinomiali con una singola estrazione e con K categorie con probabilità $\pi_1, \ldots, \pi_K$. La log-verosimiglianza dei dati completi risulta quindi essere:

$$l(\theta_k, \pi_k, z_{ik}|x) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} log[\pi_k f(x_i; \theta_k)]. \tag{2.7}$$

L'algoritmo massimizza iterativamente il valore atteso condizionato della log-verosimiglianza dei dati completi

$$E[l_c(\theta; y; z)|\theta^*] = \sum_{k=1}^{K} \sum_{i=1}^{n} t_{ik} log(\pi_k f(x_i; \theta_k)) \tag{2.8}$$

dove $t_{ik} = E[z = k|y_i, \theta^*]$ e θ^* è un set di valori per i parametri del modello. Partendo da una soluzione iniziale per il valore dei parametri $\theta^{(0)}$ si alternano E-step e M-step. L'*E-step* calcola $E[l_c(\theta; y; z)|\theta^{(q)}]$ condizionatamente al valore assunto dai parametri

al passo q, $\theta^{(q)}$. Si ottiene così facendo:

$$\hat{z}_{ik} \leftarrow \frac{\hat{\pi}_k f(x_i|\hat{\theta}_k)}{\sum_{l=1}^{K} \hat{\pi}_l f(x_i|\hat{\theta}_j).} \tag{2.9}$$

Successivamente *M-step* massimizza $E[l_c(\theta; y; z)|\theta^{(q)}]$ per calcolare un aggiornamento dei valori stimati per l'insieme dei parametri. I due step proseguono iterativamente fino al momento in cui non è soddisfatta una regola d'arresto. Solitamente questa regola d'arresto può essere semplicemente $|l(\theta^{(q)}; y) - l(\theta^{(q-1)}; y)| < \epsilon$ con ϵ scelto positivo e piccolo a piacere.

Facendo in particolar modo riferimento al caso più frequente in cui si ha una mistura di normali multivariate, l'*E-step* procede in maniera analoga andando ad ottenere sempre come soluzione la (2.9) dove si va però a sostituire f con ϕ come definita in (2.13). Cambiamenti più sostanziali avvengono nell'*M-step* che, in questo caso, fornisce come risultati delle espressioni in forma chiusa. Abbiamo infatti:

$$\hat{\pi}_k \leftarrow \frac{n_k}{n}; \qquad \hat{\mu}_k \leftarrow \frac{\sum_{i=1}^{n} \hat{z}_{ik} y_i}{n_k}; \qquad n_k = \sum_{i=1}^{n} \hat{z}_{ik}. \tag{2.10}$$

È stato dimostrato che, sotto certe condizioni, questo metodo converge ad un massimo locale di (2.5). Sebbene queste condizioni non siano sempre verificate, questo algoritmo viene ugualmente utilizzato nel contesto dei modelli a mistura finita con buoni risultati nella pratica (McLachlan e Krishnan, 2007).

L'algoritmo EM presenta comunque anche un certo numero di limitazioni. Innanzitutto non è utilizzabile quando ci si trova in situazioni nelle quali la matrice di covarianza corrispondente ad uno o più componenti presenta problemi di singolarità o di quasi singolarità; più in generale l'algoritmo non può procedere nel caso in cui i gruppi contengano poche osservazioni o nel caso in cui queste osservazioni siano fortemente correlate. Risulta inoltre poco pratico per i modelli con un grande numero di componenti. Infine può convergere molto lentamente, in particolar modo se i valori assegnati in partenza ai parametri (cioè $\theta^{(0)}$) non sono ragionevoli. Questo introduce il problema dell'inizializzazione dell'algoritmo EM. Poichè la funzione di verosimiglianza di un modello-mistura è multimodale e l'algoritmo converge ad un massimo locale, la scelta dei valori di partenza risulta essere fondamentale per trovare il massimo globale della verosimiglianza. L'inizializzazione dell'algoritmo può inoltre modificare la velocità della convergenza stessa. L'approccio che verrà

utilizzato in seguito è quello di scegliere come valori di partenza quelli basati sulle soluzioni del *clustering* gerarchico basato su modello (Fraley, 1998). Si noti che, attualmente, nessuna strategia di inizializzazione si comporta in maniera uniformemente migliore rispetto alle altre in tutte le situazioni. Per ulteriori approfondimenti riguardo questo problema si veda Melnykov e Melnykov (2012).

È importante infine rilevare la presenza di numerosi varianti dell'algoritmo EM che cercano di risolvere alcuni dei problemi sopracitati. Tra le più utilizzate si possono nominare il *classification* EM o CEM, dove le $\hat{z}_{ik}$ sono convertite in una classificazione discreta prima dell'*M-step*, e il *stochastic* EM o SEM (Celeux e Diebolt, 1985), dove le $\hat{z}_{ik}$ sono simulate e non stimate nell'*E-step*.

2.3 Clustering basato su modelli a mistura finita

Da quanto detto finora risulta evidente come basare dei metodi di analisi dei gruppi sui modelli a mistura finita risulti vantaggioso. Infatti, definendo il problema di *clustering* all'interno di un contesto statisticamente rigoroso, è possibile avvalersi di procedure inferenziali per rispondere a molte delle domande alle quali non è possibile rispondere utilizzando i metodi basati sulla distanza tra le osservazioni.

Il *clustering* basato su modello (Fraley e Raftery, 1998) (Fraley e Raftery, 2002) si basa sull'idea che i dati osservati provengano da una popolazione caratterizzata da un certo numero di sotto-popolazioni al suo interno. Tale approccio modella quindi ogni sotto-popolazione separatamente e l'intera popolazione come una mistura di queste, utilizzando i modelli a mistura finita introdotti nel paragrafo precedente. Nel *clustering* basato su modello si assume, quindi, che i dati siano generati da una mistura di distribuzioni nella quale ogni singola componente rappresenta un *cluster* differente. Il modello alla base di questo approccio è quindi:

$$f(x) = \sum_{k=1}^{K} \pi_k f(x|\theta_k) \tag{2.11}$$

dove ora si indica con K il numero totale dei gruppi presenti nei dati, con π_k la proporzione di unità appartenenti al k-esimo gruppo e con $f(x|\theta_k)$ la distribuzione di probabilità per il k-esimo gruppo. Anche l'utilizzo dell'algoritmo EM è quindi giustificato e facilmente comprensibile in quanto, il vedere i dati come "incompleti" ci permette, in questo tipo di analisi, di pensare alle variabili z_{ik} come alle etichette indicanti l'appartenenza dell'osservazione i-esima al k-esimo gruppo. Queste eti-

chette costituiscono la principale differenza tra l'analisi discriminante e l'analisi dei gruppi e, in questa situazione, l'attenzione si concentra sulla loro stima in modo da poter fornire una suddivisione dei dati in *cluster*. La partizione dei dati osservati avviene quindi dopo aver stimato il modello a mistura finita e, in particolar modo, dopo aver ottenuto delle stime per il vettore delle etichette z_{ik}. Nel momento in cui si raggiunge la convergenza dell'algoritmo EM si può infatti ottenere la partizione dei dati $\{\hat{z}_1, \ldots, \hat{z}_K\}$ (si veda il risultato ottenuto in (2.9)) grazie alla probabilità a posteriori $t_{ik} = Pr(Z = k|y_i, \hat{\theta})$ utilizzando la regola MAP (*maximum a posteriori rule*) che assegna le osservazioni y_i al gruppo con la probabilità a posteriori più elevata. La partizione dei dati in gruppi viene quindi fatta classificando l'i-esima osservazione al gruppo j se $\{j|\hat{z}_{ij} = max\ \hat{z}_{ik}\}$.

È molto importante far notare che, utilizzando un approccio di questo tipo, l'algoritmo EM riesce a fornire anche una misura di incertezza nella classificazione rappresentata da $\{1 - max_k \hat{z}_{ik}\}$; questa rappresenta un'informazione che i metodi brevemente spiegati nel paragrafo 2.1 non riescono a darci.

Per quanto detto finora i vantaggi che si riscontrano, nel momento in cui si va ad adottare un approccio al *clustering* basato su un modello a mistura finita, sono molteplici: si possono utilizzare procedure di natura inferenziale e si ha un maggiore flessibilità delle forme assumibili dai *cluster*, in particolar modo se vengono adottate delle distribuzioni flessibili e, di conseguenza, i gruppi non assumo delle forme "tipiche" in base al metodo utilizzato come invece accade con i metodi basati su distanza. Inoltre, come è già stato sottolineato, questo metodo fornisce delle probabilità di appartenenza ai gruppi per ogni unità andando a dare indicazioni quindi sull'incertezza del raggruppamento. In questo contesto poi, come si ha modo di vedere in seguito, il problema della scelta del numero dei gruppi viene affrontato come un problema di selezione del modello. Infine si può far notare che il *clustering* basato su modello non fornisce obbligatoriamente una partizione in gruppi ma può capitare che raggruppi tutte le osservazioni in un singolo *cluster*, cosa che non è invece possibile con i metodi ai quali si è accennato nel paragrafo 2.1.

Nel contesto dell'analisi di raggruppamento spesso si assume che le componenti della mistura appartengano alla famiglia di distribuzione normale; si può fare questa assunzione solo nel caso in cui si abbiano variabili quantitative, continue e con supporto costituito dall'insieme $\mathbb{R}^p$. Nel caso di componenti normali multivariate

abbiamo quindi che

$$f(x_i;\theta_k) = \phi(x_i;\mu_k,\Sigma_k), \qquad\qquad i = 1,\ldots,n \tag{2.12}$$

dove

$$\phi(x_i;\mu_k,\Sigma_k) = (2\pi)^{-\frac{p}{2}}|\Sigma_k|^{-\frac{1}{2}}exp\{-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\} \tag{2.13}$$

denota la densità di una distribuzione normale multivariata con vettore delle medie μ_k e matrice di covarianza Σ_k $(k = 1,\ldots,K)$. Assumendo la normalità otteniamo gruppi ellissoidali e centrati attorno alla media μ_k. Per quel che riguarda le altre caratteristiche geometriche, queste derivano direttamente dal tipo di parametrizzazione che viene imposta alla matrice di covarianza Σ_k.

In questo lavoro viene utilizzata la parametrizzazione proposta da Banfield e Raftery (1989) e ripresa da Celeux e Govaert (1995). Questa parametrizzazione consiste nell'esprimere le matrici di varianza e covarianza dei gruppi Σ_k in termini della loro scomposizione spettrale

$$\Sigma_k = \lambda_k D_k A_k D_k^T \tag{2.14}$$

dove $\lambda_k = |\Sigma_k|^{\frac{1}{d}}$, D_k è la matrice degli autovettori di Σ_k e A_k è una matrice diagonale, tale che $|A_k| = 1$, con gli autovalori di Σ_k sulla diagonale in ordine decrescente. Utilizzando questo tipo di parametrizzazione λ_k determina il volume del k-esimo *cluster*, D_k ne determina l'orientamento e A_k la forma. Lasciando variare alcune di queste quantità tra i vari *cluster* si riescono ad ottenere dei modelli parsimoniosi e facilmente interpretabili che permettono di descrivere molte situazioni differenti (si veda la Tabella 2.1). Ad esempio, se si va a considerare come parametrizzazione $\Sigma_k = \lambda D_k A D_k^T$, si avranno k *cluster* aventi volume uguale tra loro, di forma ellissoidale e uguale per tutti i *cluster* ma con un orientamento che varia da un *cluster* all'altro.

Si sottolinea infine il fatto che in questo contesto si fa spesso ricorso all'assunzione di normalità non necessariamente per un'evidenza empirica ma molto spesso per motivi di semplicità nel trattarla e di flessibilità. Sono stati infatti applicati modelli-mistura con altre distribuzioni quali, per esempio, delle misture di t di Student (utili per la maggiore robustezza) (McLachlan *e altri*, 2002), delle misture di Poisson (Banfield e Raftery, 1989) e misture multinomiali (Melnykov *e altri*, 2010).

Modello	Parametrizzazione	Volume	Forma	Orientamento
λI	Sferica	Uguale	Uguale	NA
$\lambda_k I$	Sferica	Variabile	Uguale	NA
λA	Diagonale	Uguale	Uguale	Asse coordinate
$\lambda_k A$	Diagonale	Variabile	Uguale	Asse coordinate
λA_k	Diagonale	Uguale	Variabile	Asse coordinate
$\lambda_k A_k$	Diagonale	Variabile	Variabile	Asse coordinate
λDAD^T	Ellissoidale	Uguale	Uguale	Uguale
$\lambda D_k A D_k^T$	Ellissoidale	Uguale	Uguale	Variabile
$\lambda_k D_k A D_k^T$	Ellissoidale	Variabile	Uguale	Variabile
$\lambda_k D_k A_k D_k^T$	Ellissoidale	Variabile	Variabile	Variabile

Tabella 2.1: Parametrizzazioni della matrice di covarianza Σ_k con il metodo di Banfield e Raftery (1989)

2.3.1 Selezione del modello

Quando si parla dell'analisi dei gruppi un problema fondamentale rimane quello di determinare il numero dei *cluster* presenti nei dati. Abbiamo già visto nel paragrafo 2.1 che questi problemi risultano non essere risolvibili nel momento in cui si utilizzano metodi di partizione o metodi gerarchici e, anche per questo motivo, è stato introdotto il *clustering* basato su modello che riesce a rispondere a queste domande con tecniche inferenziali. In particolare il problema riguardante il testare il numero di componenti in una mistura, e quindi il numero di *cluster* nei dati, è evidentemente molto importante sia dal punto di vista teorico che pratico ed ha quindi attirato molta attenzione in vari studi negli anni.

Nel *clustering* basato su modello questo problema viene ridotto ad un quesito riguardante la selezione del modello da utilizzare; si usano quindi metodi di selezione del modello. Uno degli approcci più comuni in questo ambito si basa sul *fattore di Bayes* (Kass e Raftery, 1995). L'idea alla base consiste nel confrontare la probabilità a posteriori di un modello condizionatamente ai dati osservati, utilizzando il teorema di Bayes. Siano $M_1, \ldots, M_J$ dei modelli stimati aventi probabilità a priori (spesso considerata uguale per i J modelli) $p(M_j), j = 1, \ldots, J$; allora la probabilità a posteriori del modello dati i dati osservati risulta essere proporzionale alla probabilità dei dati osservati condiziontamente al modello M_j, moltiplicata per la probabilità a priori del modello stesso

$$p(M_j|D) \propto p(D|M_j)p(M_j).$$

Nelle situazioni in cui ci sono parametri non noti, $p(D|M_j)$ è ottenuta integrando rispetto ai parametri

$$p(D|M_j) = \int p(D|\theta_j, M_j)p(\theta_j|M_j)d\theta_j.$$

La quantità $p(D|M_j)$ è nota con il nome di *verosimiglianza integrata* del modello M_j. A questo punto l'approccio bayesiano consiste nell'andare a scegliere il modello più probabile a posteriori; questa scelta, se consideriamo i modelli ugualmente probabili a priori, viene fatta valutando il rapporto tra le verosimiglianze integrate di due modelli differenti

$$\frac{p(D|M_r)}{p(D|M_s)}$$

dove valori maggiori di 1 indicano un'evidenza a favore di M_r e valori minori di 1 indicano un'evidenza a favore di M_s. Questa quantità risulta però particolarmente complicata da calcolare e quindi si fa comunemente uso del *Bayesian Information Criterion* (BIC) (Schwarz *e altri*, 1978) sulla base della seguente approssimazione

$$2logp(D|M_k) \approx 2logp(D|\hat{\theta}_k, M_k) - \nu log(n) = BIC_k \tag{2.15}$$

dove ν è il numero dei parametri da stimare per il modello M_j. Sebbene le assunzioni alla base della (2.15) non siano soddisfatte nel contesto dei modelli a mistura finita, è stato dimostrato che questa approssimazione risulta funzionare bene nell'ambito del *clustering* e che il BIC fornisce un criterio consistente per la scelta del numero dei gruppi (Keribin, 2000). Il confronto tra due modelli differenti viene quindi fatto osservando i valori assunti dal BIC valutato per i modelli in questione. Convenzionalmente una differenza minore di 2 del BIC tra due modelli viene considerata un'evidenza debole, differenze tra il 2 e il 6 vengono considerate come evidenze positive, tra il 6 e il 10 come evidenze forti mentre differenze maggiori di 10 nei valori del BIC valutato in due differenti modelli vengono considerate evidenze molto forti (Kass e Raftery, 1995). Sono stati inoltre proposti altri criteri per scegliere il numero di gruppi nel *clustering* basato su modello; per una comparazione riguardante le performance di questi si rimanda a Biernacki e Govaert (1999).

Si fa notare che il problema riguardante il determinare il numero di *cluster* è stato affrontato anche utilizzando il test del rapporto di verosimiglianza (si veda per esempio McLachlan e Peel, 2004). Sfortunatamente le assunzioni alla base del test non sono verificate facendo sì che la quantità $-2log\lambda$ non abbia l'usuale distribuzione χ^2 e rendendo quindi non consigliato l'uso di questo tipo di approccio in questo contesto.

Capitolo 3

La riduzione della dimensionalità

3.1 La maledizione della dimensionalità

Negli ultimi decenni, come è già stato fatto notare, le innovazioni tecnologiche hanno generato una vera e propria esplosione di dati disponibili. Spesso questi dati presentano un numero di variabili paragonabile, o talvolta maggiore, al numero di osservazioni a disposizione. È evidente come una quantità maggiore di dati comporti dei vantaggi in termini di informazione disponibile; nella pratica però non tutte le variabili rilevate risultano fornire informazioni rilevanti. Detto ciò è agevole capire come questa esplosione di dati possa aver portato anche una serie di problemi e di nuove sfide riguardanti la necessità di trovare delle tecniche statistiche adeguate che riescano a sintetizzare e a selezionare l'informazione utile nei dati.

L'espressione "maledizione della dimensionalità" (Bellman, 1957) indica il problema derivante dal rapido incremento delle dimensioni dello spazio matematico associato all'aggiunta di variabili; questo incremento porta ad una maggiore dispersione dei dati all'interno dello spazio descritto dalle variabili rilevate, ad una maggiore difficoltà nella stima e, in generale, nel cogliere delle strutture nei dati stessi. Per permettere una comprensione più efficace del problema si riporta un esempio da Azzalini e Scarpa (2004). Nel caso in cui si abbiano a disposizione $n = 500$ osservazioni uni-dimensionali nell'intervallo unitario $(0, 1)$, l'intervallo è descritto in maniera adeguata dalle osservazioni in questione e, nel caso in cui si voglia stimare una funzione $f(x)$, è probabile ottenere una stima attendibile della stessa grazie alla breve distanza che separa gli n punti nell'intervallo. Se le n osservazioni non fossero però uni-dimensionali ma fossero distribuite sul quadrato del piano descritto da $(0, 1)^2$, questi punti risulterebbero essere meno fitti. Proseguendo in questo modo e passando a dimensioni superiori la dispersione dei punti aumenta rapidamente e

conseguentemente peggiora la qualità della stima della funzione $f(x)$ e la capacità di descrivere adeguatamente lo spazio campionario. Nel caso in cui si operi nello spazio $\mathbb{R}^p$, per compensare l'aumento della spaziatura tra i punti sarebbe necessario avere un numero di osservazioni pari a n^p; mentre è comune disporre di un campione di 500 osservazioni, risulta essere pressochè impossibile disporre di un campione, anche solamente nel caso di $p = 10$, di $n = 500^{10}$ osservazioni.

Risulta quindi evidente come, in spazi a dimensionalità elevata, risultino necessarie delle tecniche di riduzione della dimensione che riescano a selezionare e a sintetizzare l'informazione contenuta nelle variabili.

3.1.1 Clustering di dati ad elevata dimensionalità

I metodi di raggruppamento assumono una grande importanza nel contesto della statistica multivariata e quindi la maledizione della dimensionalità va a porre dei problemi che il *clustering* deve cercare di risolvere per poter trovare una struttura e una suddivisione all'interno dei dati.

In linea di principio può sembrare sensato pensare che utilizzare più informazione, e di conseguenza più variabili, possa portare a dei risultati migliori per quanto riguarda l'analisi di raggruppamento. Questo suggerirebbe il fatto che, nel caso in cui si abbia un numero elevato di variabili disponibili, un approccio adeguato sia quello di utilizzarle tutte congiuntamente in modo da non perdere informazione potenzialmente utile. Nella pratica però molte variabili possono essere ridondanti o possono non contenere informazioni utili ai fini del raggruppamento. Oltretutto, qualora si utilizzassero anche queste variabili, non solo non si otterrebbero dei miglioramenti ma, spesso, ci si troverebbe nella situazione in cui queste variabili andrebbero a peggiorare i risultati.

Nel contesto del *clustering* basato su modello a mistura finita, la maledizione della dimensionalità risulta essere un problema molto serio poichè, tipicamente, il numero di parametri da stimare cresce con la dimensionalità del problema. In particolare, nei modelli basati su mistura di componenti gaussiane, se consideriamo la parametrizzazione completa ($\lambda_k D_k A_k D_k^T$) presentata nella Tabella 2.1, il numero di parametri da stimare è uguale a:

$$\nu = (K - 1) + Kp + Kp(p + 1)/2$$

dove $(K-1)$ è il numero di parametri relativi alle proporzioni π_k, Kp è il numero di parametri relativi alle medie dei K *cluster* e $Kp(p+1)/2$ rappresenta il numero di parametri relativi alle matrici di covarianza Σ_k. Si può immediametamente notare come il numero ν dei parametri da stimare sia una funzione quadratica del numero di variabili rilevate; per questo motivo, al crescere di p, per riuscire a stimare questo tipo di modello è necessaria una numerosità campionaria molto elevata. È facile intuire come il problema della dimensionalità, in questo frangente, sia strettamente legato alla stima delle matrici di covarianza Σ_k e vanno letti in questa direzione i tentativi di ottenere modelli più parsimoniosi attraverso delle parametrizzazioni di Σ_k. In effetti quando il numero di osservazioni è piccolo relativamente al numero di parametri da stimare si ottengono delle stime $\hat{\Sigma}_k$ molto variabili che conducono di conseguenza a delle funzioni instabili di classificazione dei dati in gruppi. Nel caso peggiore in cui la matrice Σ_k risulti essere singolare non può essere direttamente applicato l'approccio al *clustering* basato su modello senza ricorrere a delle misure correttive.

3.2 Rimedi alla maledizione della dimensionalità nel clustering

È stato fatto notare (Law *e altri*, 2004) come i metodi studiati per la selezione delle variabili permettano di migliorare di molto i risultati nell'ambito dell'analisi di classificazione e come questi metodi abbiano ricevuto in generale un'attenzione minore nell'ambito dell'analisi di raggruppamento. Gli autori indicano come possibile motivazione il fatto che, nel *clustering*, non è del tutto chiaro secondo quale criterio valutare l'importanza delle variabili a disposizione senza poter fare riferimento alle etichette indicanti l'appartenenza delle osservazioni ad una determinata classe e senza avere, come spesso accade, informazioni a priori riguardanti la struttura dei dati. Questo problema pone una sfida ancora maggiore nel caso in cui non sia noto il numero di *cluster* dal momento che la scelta del numero di *cluster* è interconnessa al problema di selezione delle variabili. Un esempio che aiuti a comprendere meglio questa interconnessione tra i due problemi viene riportato nella figura 3.1. Queste difficoltà sono state sottolineate anche da Gnanadesikan *e altri* (1995) che sostengono che *"uno degli aspetti più spinosi dell'analisi dei cluster continua ad essere quello riguardante il peso da assegnare alle variabili e la loro selezione"*.

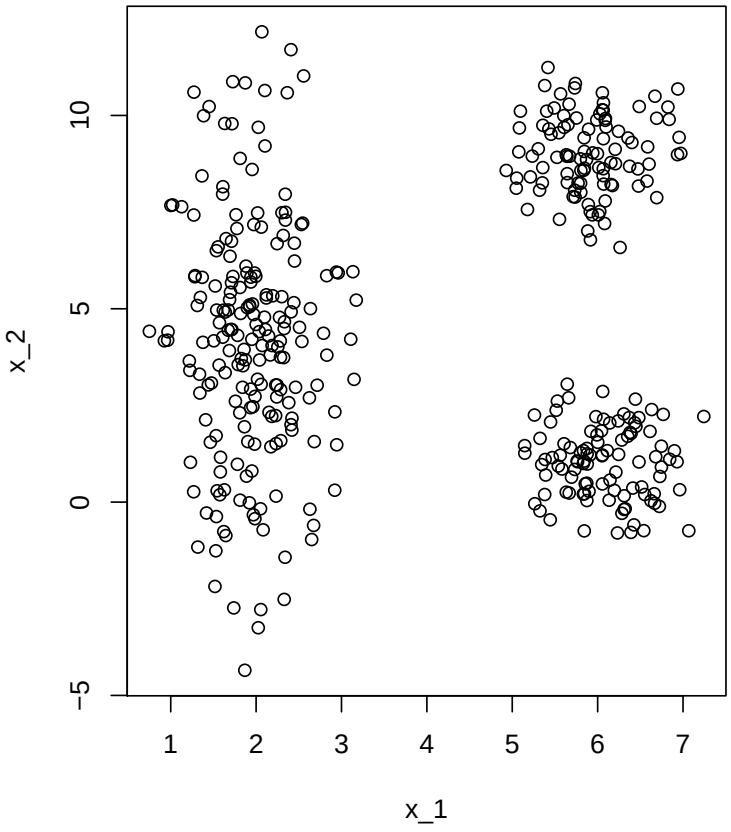

Figura 3.1: Il numero di gruppi è interconnesso al sottoinsieme di variabili utilizzato. Il sottoinsieme ottimale per identificare tre, due o un gruppo è dato rispettivamente da $\{x_1; x_2\}, \{x_1\}$ *e* $\{x_2\}$

.

Tuttavia, per i motivi riportati finora, anche nel *clustering* risulta essere molto importante, prima di cercare di raggruppare i dati, ridurre la dimensionalità degli stessi o selezionare, secondo qualche criterio, le variabili rilevanti ai fini dello scopo finale. Inoltre questo fa sì che, generalmente, si riescano ad ottenere delle soluzioni più facilmente interpretabili. A sostegno della necessità di utilizzare tecniche di riduzione della dimensione o di selezione delle variabili, è stato argomentato (Scott e Thompson, 1983) che gli spazi con dimensionalità elevata risultano essere spesso vuoti e questo ci conferma ulteriormente la possibilità e l'esigenza di trovare dei metodi che modellino i dati in sottospazi di dimensione inferiore rispetto allo spazio originale.

Negli anni sono stati proposti vari approcci per cercare di ovviare al problema della dimensionalità e di ottenenere delle buone partizioni nel contesto del *clustering* basato su modello. Si procede quindi ora andando a fornire alcune possibili classificazioni di questi approcci.

Una prima classificazione dei possibili metodi per trattare il problema della dimensionalità nel contesto del *clustering* è stata introdotta da Bouveyron e Brunet-Saumard (2014). In questo articolo viene innanzitutto evidenziata l'esistenza di alcuni approcci quali i *metodi di selezione delle variabili*, i *metodi di riduzione della dimensione*, i *metodi di regolarizzazione*, i metodi che fanno uso di *modelli vincolati e parsimoniosi*.

La *selezione delle variabili* per il *clustering* si pone come obiettivo quello di ridurre la dimensionalità dei dati andando a selezionare le variabili rilevanti che riescano a cogliere la struttura in gruppi dei dati stessi. L'idea alla base è quella secondo la quale i gruppi si differenziano in maniera sostanziale solo rispetto ad alcune delle variabili misurate. Si possono citare due approcci differenti: uno dei due considera la selezione delle variabili come un problema di selezione del modello mentre l'altro cerca di selezionare le variabili rilevanti utilizzando una penalizzazione nella funzione di verosimiglianza per ottenere la sparsità. Il primo approccio presenta delle limitazioni di tipo computazionale nel momento in cui si abbiano dati a dimensionalità molto elevata mentre il secondo, introducendo un termine di penalizzazione, soffre del problema riguardante la selezione dei valori di penalizzazione che inducono la sparsità, cosa non sempre semplice da fare nell'ambito della classificazione non supervisionata.

I *metodi di riduzione della dimensione* implicitamente assumono che i dati appartengano ad uno spazio di dimensione inferiore e cercano quindi di proiettare i dati su un opportuno sottospazio in modo che sia mantenuta la maggiore informazione possibile, misurata secondo qualche criterio. Una volta fatto questo si applicano le procedure di cui si è parlato in precedenza per ottenere un raggruppamento dei dati originali.

I *metodi di regolarizzazione* sono maggiormente utilizzati nell'ambito dell'analisi di classificazione ma possono essere facilmente estesi anche all'analisi di raggruppamento. Questi metodi si basano sull'utilizzo di penalizzazioni che introducono la sparsità nella stima dei parametri del modello. Così facendo alcuni parametri vengono posti uguali a zero riducendo così la dimensione del vettore dei parametri da

stimare.
I metodi basati su *modelli parsimoniosi* considerano la *maledizione della dimensionalità* come un problema legato al fatto che l'approccio all'analisi di raggruppamento utilizzato in questo lavoro si basa su modelli sovra-parametrizzati, in particolar modo per quanto riguarda le matrici di covarianza Σ_k. Questa classe di metodi cerca quindi di riparametrizzare la matrice di covarianza in modo da ridurre il numero di parametri da stimare. Si noti che la parametrizzazione introdotta da Banfield e Raftery (1989) e riportata nella tabella 2.1 va ad inserirsi esattamente nel contesto dei metodi parsimoniosi.
Per quel che riguarda i *metodi di riduzione della dimensione* si può notare come questi possano causare una perdita di informazione utile per ottenere una buona partizione dei dati. I *metodi di regolarizzazione* non presentano questo tipo di limitazione ma talvolta, come è già stato fatto notare, sono difficili da applicare in quanto richiedono di fissare i parametri di penalizzazione. I metodi basati su *modelli parsimoniosi* sembrano essere una soluzione sotto certi aspetti migliore in quanto vanno a proporre un compromesso tra il modello più adatto e quello che si riesce effettivamente a stimare nella pratica.

In varie pubblicazioni (si veda ad esempio Dy e Brodley, 2004) è stata introdotta un'ulteriore distinzione tra i diversi approcci utilizzabili nel momento in cui si affronta il problema della selezione delle variabili e della riduzione della dimensionalità; questa distinzione individua un metodo *wrapper* ed un metodo *filter*. Il primo approccio cerca di risolvere il problema della selezione di un sottoinsieme di variabili tenendo conto della successiva analisi di raggruppamento dei dati. Il secondo approccio, invece, affronta la selezione delle variabili senza considerare lo scopo successivo che consiste nel fornire una partizione; cerca cioè di selezionare le variabili rilevanti andando a guardare le proprietà intrinsiche dei dati senza tener conto del metodo di raggruppamento da applicare in seguito al sottoinsieme selezionato. A riguardo di questi due differenti approcci viene frequentemente sottolineato (Bouveyron e Brunet-Saumard, 2014) come sia conveniente utilizzare un metodo *wrapper* per ridurre la dimensionalità dei dati in quanto *"la non connessione tra la riduzione della dimensione e l'algoritmo di clustering può portare ad una perdita di informazione che potrebbe essere discriminante per il raggruppamento"*.

3.3 Un approfondimento su alcuni metodi di selezione e riduzione della dimensionalità

Vengono ora presentati più dettagliatamente tre diversi metodi, di seguito esplorati nelle analisi numeriche, per affrontare il problema della dimensionalità nell'ambito del *clustering* basato su modello: l'analisi delle componenti principali, l'analisi delle componenti principali sparse e il metodo di selezione delle variabili proposto da Raftery e Dean (2006).

3.3.1 Analisi delle componenti principali

L'analisi delle componenti principali (Jolliffe, 2005) è, probabilmente, una delle tecniche di analisi multivariata più famose ed utilizzate. Introdotta da Pearson (1901), fu sviluppata poi, indipendentemente, da Hotelling (1933) che propose l'analisi delle componenti principali come un metodo che mira a ridurre la dimensione dei dati cercando di spiegare la maggior quota possibile di variabilità degli stessi.

L'analisi delle componenti principali è un metodo statistico che serve a riassumere l'informazione contenuta in p variabili mediante un numero r (con $r < p$) di nuove variabili (le componenti principali) incorrelate tra loro e ottenute come combinazione lineare delle p variabili di partenza. Questa sintesi che si cerca di ottenere mira a minimizzare la perdita di informazione contenuta nelle variabili originali misurata in termini di riduzione delle variabilità totale delle nuove variabili rispetto alle variabili di partenza. L'enorme successo di questa tecnica è legato al fatto che minimizza la perdita di informazione e che tenta di risolvere allo stesso tempo il problema della dimensionalità e della multicollinearità.

La derivazione delle componenti principali avviene tramite un procedimento sequenziale: si cerca inizialmente una combinazione lineare delle variabili di partenza massimizzandone la varianza, in seguito si prosegue cercando una seconda combinazione lineare che massimizzi la varianza e che sia incorrelata con la precedente e si prosegue in questo modo. Così facendo risulta facilmente intuibile come si vengano ad ottenere delle nuove variabili con importanza, misurabile come percentuale di varianza spiegata, decrescente e incorrelate tra loro. Si noti che la derivazione delle componenti principali può avvenire utilizzando la matrice di varianza e covarianza dei dati o, in alternativa, la matrice di correlazione (usando quindi le variabili standardizzate). La scelta di quale matrice utilizzare per il calcolo delle componenti

principali è molto importante in quanto questa analisi non è invariante rispetto a trasformazioni di scala.

Formalmente, siano $x^{(1)}, \ldots, x^{(p)}$ le variabili di partenza e $\hat{\Sigma}$ la stima della matrice di varianza e covarianza di x. La prima componente principale z_i è, per definizione, una combinazione lineare di tutte le variabili $x^{(i)}$ con $i = 1, \ldots, p$ tale che abbia varianza massima. La prima componente principale può dunque essere scritta come:

$$z_1 = a_{11}x^{(1)} + a_{12}x^{(2)} + \cdots + a_{1p}x^{(p)} = a_1^T x$$

dove a_1 viene calcolato risolvendo il problema di massimizzazione vincolata:

$$\begin{gathered} \max_{a_1} Var(a_1 x) = \max_{a_1} a_1^T \hat{\Sigma} a_1 \\ s.v.\ a_1^T a_1 = 1 \end{gathered} \tag{3.1}$$

Si prosegue poi sequenzialmente cercando la successiva combinazione lineare che massimizzi la varianza e aggiungendo il vincolo che le nuove variabili siano incorrelate tra loro:

$$a_j \hat{\Sigma} a_k = 0 \ \forall j = 1, \ldots, k-1. \tag{3.2}$$

Si può dimostrare come le soluzioni a questi problemi di massimizzazione siano date da $a_i = \gamma_i$ e $var(z_i) = \lambda_i$ dove λ_i è il massimo autovalore della matrice $\hat{\Sigma}$ e γ_i è il rispettivo autovettore. In questo modo si può vedere come le componenti principali risultino essere ordinate in maniera decrescente in funzione della loro varianza, dove la i-esima componente ha varianza pari all'i-esimo autovalore della matrice di covarianza dei dati originali e dove la matrice dei coefficienti è composta dai corrispondenti autovettori.

Nell'ambito del *clustering* basato su modello le componenti principali sono una tecnica molto utilizzata per cercare di risolvere il problema dell'elevata dimensionalità. Risulta evidente come l'analisi delle componenti principali possa essere vista come un approccio di tipo *filter* e come sia iscrivibile a quel tipo di metodi definiti da Bouveyron e Brunet-Saumard (2014) come *metodi di riduzione della dimensione*.

Precedentemente è stato evidenziato come utilizzare un approccio *filter*, che non tiene quindi conto del fatto che si ricerca una riduzione della dimensione con lo scopo finale di ottenere un buon raggruppamento dei dati, per risolvere il problema della dimensionalità possa produrre soluzioni non ottimali. L'idea che giustifica l'utilizzo dell'analisi delle componenti principali in questo contesto si basa sulla speranza che, utilizzando un numero di componenti principali minore rispetto al numero delle variabili originali ma cercando di mantenere una certa quota di variabilità spiegata,

queste componenti mantengano al loro interno la struttura di gruppo dei dati risolvendo così il problema dell'elevata dimensionalità senza perdere però le informazioni rilevanti allo scopo dell'analisi di raggruppamento. Questo però spesso non avviene; Chang (1983) ha evidenziato come utilizzare l'analisi delle componenti principali in queste situazioni sia la prassi. Viene però successivamente messo in evidenza come le componenti associate agli autovalori maggiori non contengano necessariamente una quantità di informazione utile maggiore per quanto riguarda l'analisi di raggruppamento. Questo fa sì quindi che in alcuni casi l'utilizzo di questa tecnica, in questo contesto, possa oscurare la struttura di gruppo dei dati. Anche Yeung e Ruzzo (2001) hanno sottolineato come le prime m componenti principali spesso non diano una partizione dei dati efficiente mostrando l'esistenza di un sottoinsieme di altre m componenti che permette di raggiungere risultati migliori. Viene sottolineato inoltre come non sembri esserci un trend da seguire nel momento in cui si affronta la scelta delle componenti principali. Si noti inoltre come, essendo le componenti principali un metodo basato su delle combinazioni lineari delle variabili originali, questa tecnica non riesca a tener conto di dipendenze non lineari tra le variabili. Un altro problema legato alla costruzione stessa delle componenti principali, evidenziato da Zou *e altri* (2006), riguarda il fatto che essendo una combinazione lineare di tutte le variabili di partenza, spesso si riscontrano difficoltà nell'interpretazione dei risultati riguardanti eventuali partizioni dei dati e ci si trova nell'impossibilità di comprendere quali variabili giochino un ruolo di maggior rilievo nel differenziare le unità appartenenti a *cluster* diversi.

In questo lavoro si è deciso di utilizzare l'analisi delle componenti principali nonostante i limiti in quanto questo tipo di analisi rimane un metodo molto comune per risolvere i problemi legati alla dimensionalità, anche nell'ambito del *clustering*.

3.3.2 Analisi delle componenti principali sparse

L'analisi delle componenti principali sparse (Zou *e altri*, 2006) è un metodo che cerca di trovare un modo di superare alcuni limiti che si riscontrano quando si utilizzano le componenti principali; uno di questi limiti risiede nel fatto che le componenti principali sono, per costruzione, una combinazione lineare di tutte le variabili rilevate sui dati. Questo rende difficile l'interpretazione stessa delle componenti principali e quindi, nell'ambito del *clustering*, anche l'intepretazione dei gruppi.

Si noti come, essendo il successo e l'utilità delle componenti principali spesso legati alla loro effettiva interpretabilità, questo problema sia stato spesso oggetto di studio e come si siano cercati metodi per risolverlo. Tra questi metodi possiamo citare le tecniche di rotazione ed il metodo SCoTLASS (Jolliffe *e altri*, 2003) in cui si introduce una prima idea embrionale di *componenti principali sparse*.

Il metodo SCoTLASS modifica le componenti principali rendendo possibile che alcuni coefficienti siano pari a zero. Questo approccio infatti va a risolvere l'usuale problema di massimizzazione vincolata alla base dell'analisi delle componenti principali (si veda l'equazione 3.2) aggiungendo però un ulteriore vincolo:

$$\sum_{v=1}^{p} |a_{kv}| \leq t \tag{3.3}$$

per un qualche valore di t. Questo vincolo viene imposto direttamente sulla norma L_1 delle componenti principali. I limiti maggiori di questo tipo di approccio riguardano la scelta del valore di t e il fatto che, nelle situazioni in cui si richiede una elevata percentuale di varianza spiegata, le soluzioni ottenute non risultino essere sufficientemente sparse (Jolliffe *e altri*, 2003).

Zou *e altri* (2006) cercano di superare questi limiti individuando un numero di nuove variabili, combinazione lineare delle variabili osservate, e cercando al contempo di introdurre sparsità guadagnandone così in interpretabilità. In altre parole lo scopo che questa analisi cerca di raggiungere non è solamente quello di ridurre la dimensionalità dei dati ma anche quello di selezionare le variabili più rilevanti.

Per ottenere sparsità nell'analisi delle componenti principali il problema viene riformulato in termini di regressione penalizzata. Infatti, rilevando ogni singola componente principale come una combinazione lineare delle variabili originali, i coefficienti di questa combinazione lineare possono essere visti come i coefficienti di un modello di regressione. In questo modo si riesce ad inserire il calcolo delle componenti principali all'interno del contesto di penalizzazione ottenendo la sparsità nei coefficienti delle componenti stesse.

Per fare questo, nella pratica, viene utilizzata una tecnica di penalizzazione e di selezione delle variabili, proposta da Zou e Hastie (2005), chiamata *elastic net*. Si fa notare come sia stata sottolineata la possibilità di utilizzare altre tecniche di penalizzazione, quali la *regressione ridge* o il *Least Absolute Shrinkage and Selection Operator* (Lasso) (Tibshirani, 1996), che risultano però avere delle limitazioni che l'*elastic net* riesce a risolvere.

L'*elastic net* (Zou e Hastie, 2005) fondamentalmente risulta essere un metodo di regolarizzazione che combina linearmente gli altri due metodi sopraccitati. Nell'ambito del modello di regressione

$$y_i = \sum_{j=1}^{p} x_i^{(j)} \beta_j + \epsilon_i, i = 1, \ldots, n; \tag{3.4}$$

la stima dei parametri si ottiene andando a minimazzare la seguente quantità:

$$\hat{\beta}_j = (1 + \lambda_2) \left\{ \arg\min_{\beta} ||y - \sum_{j=1}^{p} x^{(j)} \beta_j||^2 + \lambda_2 \sum_{j=1}^{p} |\beta_j|^2 + \lambda_1 \sum_{j=1}^{p} |\beta_j| \right\}, \tag{3.5}$$

dove λ_1 e λ_2 sono i parametri che permettono di introdurre la sparsità. Questo approccio assume solamente che λ_1 sia non negativo mentre l'assunzione sulla positività di λ_2 viene fatta solamente nelle situazioni in cui il numero di variabili risulta essere maggiore del numero di unità statistiche. L'estensione al caso delle componenti principali sparse si ottiene sostituendo al primo addendo della (3.5) la seguente:

$$(a_j - \beta)^T X^T X (a_j - \beta) \tag{3.6}$$

dove a_j è il vettore dei coefficienti associati alla prima componente principale ordinaria e X è la matrice $n \times p$ delle osservazioni.
La j-esima componente principale sparsa si otterrà, pertanto, come:

$$z_j = \beta_{j1} x^{(1)} + \cdots + \beta_{jp} x^{(p)}. \tag{3.7}$$

Si è già puntualizzato in precedenza come un limite degli approcci che cercano di ottenere una situazione di sparsità introducendo delle soglie sia dovuto al fatto che richiedono di dare dei valori a tali soglie; anche in questo caso un possibile limite può derivare dalla necessità di fornire dei valori a priori per λ_1 e a λ_2. Dai valori di questi parametri dipende il grado di sparsità ed è una soluzione comune quella di provare diversi scenari in modo da trovare quei valori che permettano di ottenere la sparsità desiderata.

Per quanto detto finora, risulta evidente come sia più complicato classificare questo tipo di metodo andandolo ad inserirlo in una classe di quelle presentate nel paragrafo 3.2: le componenti principali sparse possono essere infatte viste al tempo stesso sia come un *metodo di riduzione della dimensione* sia come un *metodo di*

penalizzazione ed infine anche come un *metodo di selezione delle variabili*, come puntualizzato dagli stessi autori.

3.3.3 Metodo Raftery-Dean

Nello specifico ambito del *clustering* basato su modello statistico, Raftery e Dean (2006) hanno proposto un metodo che ottiene una riduzione della dimensionalità attraverso una selezione delle variabili ritenute rilevanti ai fini di offrire una buona partizione in gruppi dei dati. È evidente come questo approccio sia classificabile come un approccio di tipo *wrapper* in quanto pone il problema della selezione delle variabili come un problema di selezione del modello da utilizzare per l'analisi di raggruppamento.

Questo metodo procede iterativamente e prevede di risolvere il problema della selezione delle variabili ponendolo in termini di comparazione tra diversi modelli.

Si considera innanzitutto di avere un set di variabili X e, ad ogni passo, una loro suddivisione:

- $X^{(1)}$, l'insieme delle variabili considerate rilevanti ai fini del raggruppamento;
- $X^{(2)}$, la/le variabile/i proposte per essere incluse o escluse da $X^{(1)}$;
- $X^{(3)}$, le variabili rimanenti.

Ad ogni iterazione del metodo si prende una decisione in merito all'eventuale inclusione o esclusione di $X^{(2)}$ basandosi su un qualche criterio che riesca a valutare se $X^{(2)}$ sia utile o meno. In analogia con quanto detto nel paragrafo 2.3.1, il criterio che viene utilizzato in questo contesto è il *fattore di Bayes*; per via dei già citati problemi computazionali legati al suo utilizzo, anche in questo caso, si ricorre ad una sua approssimazione data dal BIC. Risulta evidente come questo metodo cerchi di selezionare le variabili ponendo questo problema come fosse una scelta tra modelli competitivi. Siano infatti:

$$\begin{aligned} M1: \quad & p(X|z) \\ & = p(X^{(1)}, X^{(2)}, X^{(3)}|z) \\ & = p(X^{(3)}|X^{(2)}, X^{(1)})p(X^{(2)}|X^{(1)})p(X^{(1)}|z) \qquad (3.8) \\ M2: \quad & p(X|z) \\ & = p(X^{(1)}, X^{(2)}, X^{(3)}|z) \\ & = p(X^{(3)}|X^{(2)}, X^{(1)})p(X^{(2)}, X^{(1)}|z) \qquad (3.9) \end{aligned}$$

dove z è il vettore (non osservato) delle etichette indicanti l'appartenenza delle unità ai gruppi. Si noti come in M1, $X^{(2)}$ dipenda dal vettore z solamente tramite $X^{(1)}$; questo significa che M1 specifica un modello dove $X^{(2)}$ non fornisce nessuna informazione aggiuntiva per quel che riguarda la struttura di gruppo dei dati. In M2, invece, possiamo vedere come $X^{(2)}$ dipenda direttamente dal vettore di etichette formalizzando cosí l'idea secondo la quale, in questo caso, la variabile proposta per essere inclusa (o esclusa) è rilevante ai fini della sucessiva ripartizione. Si noti inoltre come, in M1, non venga assunto che le variabili $X^{(2)}$ siano indipendenti dalle variabili rilevanti: questo permette di evitare di includere variabili non realmente rilevanti solo in quanto correlate con le variabili contententi informazione. La scelta riguardante l'inclusione di $X^{(2)}$ viene dunque fatta andando a comparare, tramite il BIC, i modelli M1 e M2.

Questo metodo procede attraverso un *algoritmo greedy* che, ad ogni passo, cerca la variabile che maggiormente migliora il raggruppamento, valutando questo miglioramento tramite il BIC nel modo appena descritto, e cerca se e quali tra le variabili possono essere tolte dal modello. L'algoritmo si ferma nel momento in cui non sono possibili ulteriori miglioramenti. Nello specifico questo metodo procede come segue:

- Si sceglie K_{max}, il numero massimo di gruppi da considerare nell'analisi;
- *Primo passo*: Si sceglie la prima variabile andando a prendere quella che presenta una differenza maggiore tra il BIC calcolato utilizzando questa variabile per l'analisi dei gruppi (considerando $k \in 2, \ldots, K_{max}$) e il BIC calcolato non utilizzandola;
- *Secondo passo*: Si sceglie la seconda variabile da includere in $X^{(1)}$ andando a considerare la differenza tra il BIC calcolato per il modello contenente questa variabile congiuntamente a quella scelta al primo passo e il BIC calcolato per il modello contenente solamente la variabile scelta al primo passo al quale si somma il BIC calcolato per un modello di regressione dove $X^{(1)}$ é la variabile indipendente e la generica $x^{(j)} \in X^{(3)}$ é la variabile dipendente. Si evidenzia il fatto che il BIC calcolato per il modello di regressione viene inserito nel confronto per formalizzare la possibile correlazione tra le variabili rilevanti e quelle non rilevanti. Si noti infine che non è stata posta nessuna condizione riguardo la positivitá di questa differenza, nè nel primo passo nè in questo, in

quanto si richiede semplicemente che queste siano le migliori variabili possibili tra quelle a disposizione.

- *Generico passo (Inclusione)* La variabile proposta per essere inclusa in $X^{(1)}$ è quella che presenta la maggior differenza tra i BIC, come definiti al punto precedente con l'aggiunta della variabile inclusa al secondo passo e al variare del numero di gruppi k considerati. Successivamente si va a valutare se questa differenza sia positiva o meno: qualora fosse positiva la variabile viene inserita nel set di variabile selezionate per il *clustering*, qualora fosse negativa questo set rimane lo stesso.

- *Generico passo (Esclusione)* La variabile proposta per essere escluda da $X^{(1)}$ è quella che, tra le variabili appartenenti al set di variabili selezionate per il *clustering*, presenta la minore differenza tra il BIC calcolato per il modello contentente tutte le variabili selezionate fino a questo passo, e la somma tra i BIC calcolato per il modello contenente tutte le variabili eccetto quella proposta per l'esclusione e il BIC calcolato per la regressione di questa variabile su tutte le altre variabili individuate come rilevanti. Se questa differenza risulta essere negativa allora la variabile proposta viene rimossa dal set $X^{(1)}$, se la differenza risulta essere positiva viene invece mantenuta e considerata come variabile utile ai fini del *clustering*.

- Dopo il primo e il secondo passo, il passo generico viene iterato e l'algoritmo si ferma nel momento in cui la differenza tra i BIC nel passo di inclusione è negativa e la differenza tra i BIC nel passo di esclusione è positiva. Questo perchè, ovviamente, non avvengono più cambiamenti nel set di variabili considerate rilevanti ai fini dell'analisi di raggruppamento.

Tale algoritmo trova solo un ottimo locale nello spazio dei possibili modelli e, pertanto, questo metodo potrebbe essere migliorato utilizzando un algoritmo di ottimizzazione differente.

Gli stessi autori hanno poi fatto notare come, quando ci si trova a trattare dei dati sui quali è stato rilevato un numero molto elevato di variabili, questo metodo sia troppo lento e computazionalmente oneroso per essere realmente utile nella pratica. Raftery e Dean (2006) suggeriscono di combinare il loro approccio ad un metodo di ricerca del modello alternativo e ad una preselezione delle variabili basandosi su

qualche altro metodo. Si evidenzia come sia stato ipotizzato anche l'utilizzo di questo metodo come via per superare le criticità evidenziate da Chang (1983) riguardo l'utilizzo delle componenti principali in modo da selezionare quest'ultime non sulla base della quota di varianza spiegata ma sulla base del loro effettivo contenuto informativo per quel che riguarda la divisione dei dati in gruppi.

Si fa infine notare come il metodo appena spiegato sia stato studiato e ripreso da Maugis *e altri* (2009) i quali ne hanno proposto una variante nella quale le variabili irrilevanti ai fini del raggruppamento sono spiegate solamente da un sottinsieme di variabili rilevanti. Questo approccio, secondo gli autori, dovrebbe risultare più realistico e versatile.

Capitolo 4

Un'esplorazione numerica

4.1 Alcune considerazioni a partire da uno studio di simulazione

4.1.1 Obiettivi dello studio

In questo paragrafo si cerca di comprendere, mediante uno studio di simulazione, il comportamento del *clustering* basato su modello a mistura finita nel momento in cui si operi in un contesto ad elevata dimensionalità.

Si fa notare che si son dovute operare delle scelte in modo tale da circoscrivere il problema ponendosi in una situazione specifica nella quale si ha solamente un sottoinsieme di variabili osservate rilevanti ai fini dell'analisi di raggruppamento e si osserva un certo numero di variabili irrilevanti, indipendenti tra loro e indipendenti dalle variabili utili ai fini del *clustering*. In questo modo si opera quindi in un contesto non generale ma tuttavia molto frequente anche nelle situazioni reali.

Le domande alle quali si è cercato di rispondere e gli obiettivi che questo studio si prefigge sono di seguito elencati:

- Valutare la capacità del *clustering* basato su modello di fornire una buona partizione dei dati a seconda della forma dei gruppi considerando, in particolare, *cluster* aventi sia forma ellittica che non ellittica;
- Valutare la capacità del *clustering* basato su modello di fornire una buona partizione dei dati al variare del grado di separazione tra i gruppi (situazione esplorata qualora questi abbiano forma ellittica);
- Valutare come variano i risultati ottenuti al variare del numero di variabili irrilevanti ai fini dell'analisi di raggruppamento;

- Valutare e confrontare la bontà, al variare del numero di variabili irrilevanti ai fini del *clustering*, dell'analisi delle componenti principali, dell'analisi delle componenti principali sparse e del metodo proposto da Raftery e Dean (2006) come metodi per risolvere il problema della dimensionalità nell'ambito del *clustering* basato su modello;

- Valutare come variano i risultati ottenuti ai punti precedenti al variare della numerosità campionaria;

- Valutare come variano i risultati ottenuti ai punti precedenti al variare del numero di variabili considerate rilevanti ai fini dell'analisi di raggruppamento.

4.1.2 Descrizione degli scenari di simulazione e dei metodi utilizzati

In questo studio sono stati simulati alcuni scenari differenti, facendo variare la numerosità campionaria n, il numero di variabili rilevanti d e il numero di variabili irrilevanti w; si è invece considerato fissato e uguale a 2 il numero dei gruppi k presenti nei dati.

In particolar modo è possibile delineare la presenza di tre "macroscenari" che si differenziano tra loro per il tipo di distribuzione $1/2f_1(\cdot\,;\theta_1) + 1/2f_2(\cdot\,;\theta_2)$ da cui sono stati generati i dati. In particolare si hanno:

- gruppi ben separati di forma sferica: f_k è una distribuzione $N_d(\mu_k, I_d)$, $k = 1, 2$, dove I_d è la matrice identità di ordine d e μ_1, μ_2 sono due vettori $d-$dimensionali aventi distanza pari a 6;

- gruppi non ben separati di forma sferica: f_k è una distribuzione $N_d(\mu_k, I_d)$, $k = 1, 2$, dove I_d è la matrice identità di ordine d e μ_1, μ_2 sono due vettori $d-$dimensionali aventi distanza pari a 4;

- gruppi di forma non sferica, caratterizzati da una forte asimmetria: f_k è la distribuzione di probabilità di y_k, dove $y_1 = e^{x_1}$ con $x_1 \sim N(\mu_1, I_d)$ e $y_2 = -e^{-x_2}$ con $x_2 \sim N(\mu_2, I_d)$, dove μ_1, μ_2 sono due vettori d-dimensionali aventi distanza pari a 6.

Per un'illustrazione bidimensionale delle tre distribuzioni, si veda la figura 4.1.

Per ogni "macroscenario" si sono fatti variare d, n e w. In particolare sono stati simulati scenari differenti per ogni possibile combinazione di d e n, con $d =$

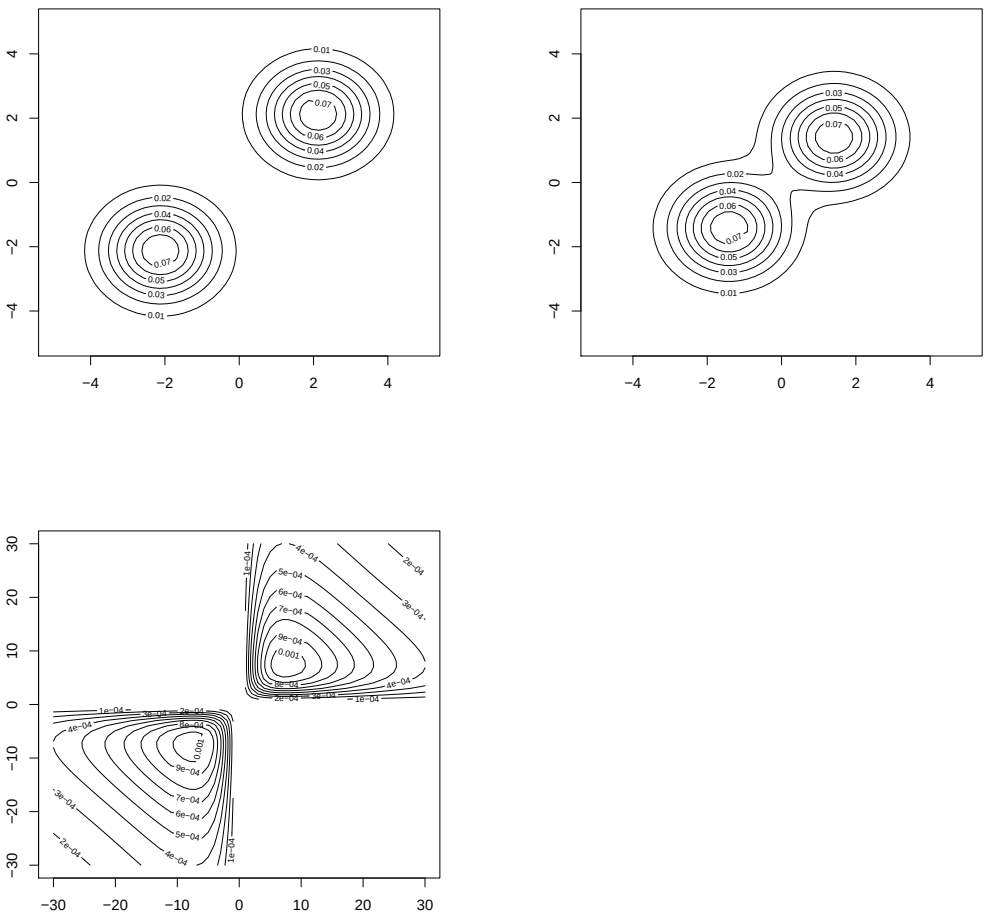

Figura 4.1: Distribuzioni bi-variate nei tre macroscenari considerati: gruppi ben separati e generati da una normale (in alto a sinistra), gruppi non ben separati e generati da una normale (in alto a destra) e gruppi generati da una distribuzione log-normale (in basso a sinistra).

$2, 5, 10$ e $n = 50, 100, 250$. In ogni scenario è stata successivamente aggiunta una quantità di variabili irrilevanti ai fini del *clustering* pari a 10,25,35 e, per ognuno di questi scenari, è stato generato un numero di campioni uguale a 500. Le variabili irrilevanti sono state generate da distribuzioni normali standard indipendenti tra loro e indipendenti dalle d variabili utili ai fini della classificazione.

Per valutare la bontà delle partizioni ottenute applicando l'algoritmo di *clustering* basato su modello a mistura finita, si è fatto ricorso all'*Adjusted Rand Index* [ARI, Hubert e Arabie (1985)]. Questo è un indice derivato dal *Rand Index* ma modificato in modo tale da avere media nulla. Il *Rand Index* rappresenta una misura della similarità tra due diversi raggruppamenti dei dati. In questo studio, avendo a disposizione le etichette indicanti la reale appartenenza dei dati ai due gruppi, questo indice è stato utilizzato per confrontare il raggruppamento fornito dall'algoritmo di *clustering* con la reale classificazione dei dati corrispondente alla componente della mistura da cui i dati sono stati generati. Il *Rand Index* è definito come segue.
Sia dato un insieme di n elementi $X = \{x_1, \ldots, x_n\}$ e due diverse partizioni di X da comparare, $P = \{p_1, \ldots, p_r\}$ e $T = \{t_1, \ldots, t_s\}$ rispettivamente una partizione di X in r gruppi e una partizione di X in s gruppi. Indicati con:

- a = il numero di coppie di elementi appartenenti allo stesso gruppo sia in P che in T;
- b = il numero di coppie di elementi appartenenti ad un gruppo differente sia in P che in T;
- c = il numero di coppie di elementi appartenenti allo stesso gruppo in P ed ad un gruppo differente in T;
- d = il numero di coppie di elementi appartenenti ad un gruppo differente in P ed allo stesso gruppo in T.

il *Rand Index* è definito come:

$$R = \frac{a + b}{a + b + c + d} \tag{4.1}$$

Il numeratore di (4.1) rappresenta intuitivamente il grado di concordanza tra le due partizioni. Risulta $R \in [0, 1]$ con $R = 1$ quando le due partizioni forniscono esattamente lo stesso raggruppamento. L'ARI, di seguito utilizzato, rappresenta un opportuno aggiustamento del *Rand Index* volto ad imporre una media nulla nel

caso in cui le unità siano allocate casualmente entro i gruppi nelle due partizioni. Analogamente al *Rand Index*, l'ARI assume valore 1 in caso di accordo perfetto tra le partizioni ma può assumere anche valori negativi.

Per ogni scenario di simulazione precedentemente delineato è stata individuata inizialmente una partizione dei dati, utilizzando il *clustering* basato su mistura di distribuzioni gaussiane, della quale si è calcolato l'*Adjusted Rand Index*. Successivamente si sono aggiunte le w variabili irrilevanti e, al variare di w, sono state calcolate le partizioni dei dati e l'ARI sia per i dati "originali" (ARI_R), sia per i dati dopo l'applicazione dell'analisi delle componenti principali (ARI_{PC}), sia per i dati dopo l'applicazione dell'analisi delle componenti principali sparse (ARI_{SPC}). Il metodo proposto da Raftery e Dean (2006) (ARI_{RD}), a causa di problemi di natura computazionale già evidenziati in precedenza, è stato utilizzato come possibile soluzione al problema della dimensionalità solo nel caso in cui $w = 10$. Gli autori stessi hanno infatti fatto notare come, nel caso in cui ci si trovi in una situazione nella quale sia stato rilevato un numero elevato di variabili, questa procedura risulti essere troppo lenta; per questo motivo si è stati costretti a ridurre il numero di valutazioni riguardanti questo metodo.

Per quanto riguarda l'*analisi delle componenti principali* si è operato in modo da prendere in considerazione, per ogni differente campione generato, un numero variabile di componenti principali tale che queste spiegassero il 70% della variabilità dei dati originali.
Per quanto riguarda invece l'*analisi delle componenti principali sparse* è stato utilizzato un numero di componenti pari al numero di componenti principali considerate per lo stesso campione. La sparsità è stata introdotta considerando un numero di coefficienti diversi da zero pari al numero di coefficienti maggiori, in valore assoluto, di 0.2 nella matrice dei coefficienti calcolata nell'analisi delle componenti principali. Si noti che in entrambe le analisi si è operato sulla matrice di correlazione.

Dal punto di vista strettamente operativo gli scenari e i metodi appena delineati sono stati implementati con il linguaggio di programmazione *R* (R Development Core Team, 2011). Nello specifico, utilizzando il pacchetto *mvtnorm* (Genz *e altri*, 2014), è stata sviluppata una funzione per la generazione dei dati da mistura (si veda la figura 4.2). Per quanto riguarda il *clustering* basato su modello è stato utilizzato il pacchetto *mclust* (Fraley *e altri*, 2012) che propone un algoritmo basato

```
mistura<-function(n,medie,varianze,pi) {
ngruppi<-length(pi)
quanti<-sample(1:ngruppi,n,rep=T,pi)
quanti<-sort(quanti)
dimgruppi<-table(quanti)
x<-matrix(NA,1,ncol(medie))

for (i in 1:ngruppi) {
xtemp<-rmvnorm(dimgruppi[i],medie[i,],varianze[,,i])
x<-rbind(x,xtemp)
}

x_a<-x[-1,]
x<-cbind(x_a,quanti)
x
}
```

Figura 4.2: Funzione 'mistura' sviluppata per la generazione dei dati nello studio di simulazione.

sulla mistura di componenti gaussiane e utilizza una parametrizzazione delle matrici di covarianza ripresa da quella proposta da Banfield e Raftery (1989) e riportata in tabella 2.1. Per quel che riguarda il calcolo dell'*Adjusted Rand Index* si è usato il pacchetto *pdfCluster* (Azzalini e Menardi, 2014). Infine, per il metodo proposto da Raftery e Dean e l'analisi delle componenti principali sparse, si sono utilizzati rispettivamente i pacchetti *clustvarsel* (Scrucca *e altri*, 2013) ed *elasticnet* (Zou e Hastie, 2012).

4.1.3 Risultati

I risultati dello studio di simulazione sono riportati nelle tabelle 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8 e 4.9.

Si evidenzia innanzitutto come la forma dei gruppi, e quindi le distribuzioni dalle quali sono stati generati i dati, abbia una forte influenza sui risultati del *clustering* basato su modello. Questo si può notare andando a confrontare i valori assunti dall'ARI nei casi in cui i gruppi abbiano forma sferica con lo stesso indice nelle situazioni in cui i dati siano invece generati da distribuzioni non gaussiane e fortemente asimmetriche; in questo caso infatti l'ARI fornisce risultati sistematicamente inferiori, per qualsiasi d, rispetto ai casi in cui i gruppi siano gaussiani. I peggiori risultati ottenuti utilizzando il *clustering* basato su modello in questo frangente so-

no imputabili alla non sfericità e all'asimmetria dei gruppi (si veda la figura 4.1); si ricordi infatti che questo tipo di approccio all'analisi di raggruppamento cerca di descrivere i dati facendo ricorso ai modelli a mistura finita con componenti gaussiane. Questo comportamento permette quindi di capire come questo metodo risulti essere poco robusto nelle situazioni in cui i gruppi abbiano forme irregolari.

Osservando le differenze tra i valori assunti dall'ARI nel caso in cui i gruppi siano ben separati rispetto al caso in cui non lo siano, è facile intuire inoltre come anche la separazione dei *cluster* assuma una forte rilevanza in questo tipo di analisi. Infatti, nelle situazioni in cui le medie dei gruppi abbiano distanza pari a 6, l'ARI assume valori costantemente più elevati rispetto alle situazioni nelle quali la distanza tra le medie sia uguale a 4. Il motivo di questo comportamento risulta essere di facile comprensione, anche osservando la figura 4.1. Si fa notare come l'algoritmo di *clustering* fornisca comunque delle buoni partizioni, e di conseguenza l'ARI assuma dei valori elevati, anche nel caso in cui i gruppi non siano ben separati: questo è dovuto al fatto che si è deciso di operare in uno scenario in cui, pur diminuendo la distanza tra le medie dei *cluster*, il grado di separazione dei gruppi è limitato e rimane quindi ben delineata la struttura di gruppo (si veda la figura 4.1). È ovvio aspettarsi che, diminuendo ulteriormente il grado di separazione tra i gruppi, il *clustering* basato su modello fornisca delle partizioni qualitativamente inferiori. Si evidenzia infine come, nel caso in cui la distanza tra le medie di gruppo sia pari a 4, l'ARI presenti uno *standard error* sistematicamente più elevato rispetto alla situazione in cui la distanza sia maggiore; questo comportamento fornisce un'indicazione in linea con quanto detto finora in quanto, andando a diminuire la separazione tra i gruppi, l'aumento di variabilità può essere sintomo di una maggiore difficoltà riscontrata dall'algoritmo di raggruppamento nel momento in cui cerchi di partizionare i dati in gruppi.

Per riuscire a comprendere se il *clustering* basato su modelli a mistura finita riesca a fornire delle partizioni sensate dei dati nel caso in cui siano state rilevate anche variabili irrilevanti ai fini dell'analisi di raggruppamento, bisogna fare riferimento ai valori assunti da ARI_R nei vari scenari. Risulta evidente il fatto che il numero di variabili w rilevate abbia una forte influenza sulla qualità delle partizioni ottenute: si nota infatti come all'aumentare di w diminuiscano i valori assunti da ARI_R. È possibile riscontrare un comportamento differente nel caso in cui le medie dei gruppi abbiano distanza uguale a 4 rispetto al caso in cui queste abbiano distanza mag-

giore: l'andamento dell'indice sembra quindi dipendere dalla struttura dei gruppi. Com'è lecito aspettarsi, infatti, nel caso in cui i gruppi presentino una separazione più netta, l'algoritmo di *clustering* è in grado di fornire delle buone partizioni, in particolar modo in corrispondenza di numerosità campionarie elevate, anche quando il numero di variabili irrilevanti ai fini dell'analisi di raggruppamento è elevato. Nel caso invece in cui i gruppi non siano ben separati i risultati ottenuti dall'analisi di raggruppamento degradano più velocemente all'aumentare di w e, nel caso in cui d sia uguale a 5 o a 10, in corrispondenza di un numero elevato di variabili irrilevanti ai fini della classificazione, il *clustering* basato su modello fornisce partizioni dei dati pressochè casuali. Da questa considerazione si può trarre spunto per evidenziare come, a parità di altre condizioni, l'ARI_R tenda a peggiorare all'aumentare di d: questo può dare un'indicazione riguardo il fatto che, senza un metodo di riduzione della dimensionalità che riesca a discriminare tra variabili rilevanti e rumore, l'algoritmo di *clustering* non riesca a fare questa distinzione percependo quindi l'aumento di d come un problema legato all'aumentare della dimensionalità dei dati e non come un aumento della quantità di informazione rilevante. Si evidenzia infine come la struttura di gruppo sembri influenzare anche la variabilità dei risultati ottenuti: si vede infatti come, tendenzialmente, lo *standard error* di ARI_R tenda ad aumentare nel caso in cui i gruppi non siano ben separati rispetto al caso in cui lo siano. Questo, riprendendo quanto detto prima, ci fornisce un'indicazione riguardante la maggior difficoltà riscontrata dall'algoritmo nel fornire un raggruppamento dei dati e la minore stabilità dei risultati.

Si è detto finora di come aggiungendo variabili irrilevanti ai fini della classificazione, in particolar modo negli scenari in cui i gruppi non sono ben separati, il *clustering* basato su modello non riesca a fornire delle buone partizioni e come sia quindi necessario l'utilizzo di qualche metodo di riduzione della dimensionalità che permetta di estrarre l'informazione realmente rilevante dai dati. Per riuscire a valutare la bontà dei metodi presi in esame in questo lavoro si deve fare riferimento agli indici ARI_{PC}, ARI_{SPC} e, quando disponibile, ARI_{RD}.

Per quel che riguarda l'analisi delle componenti principali si nota come questo metodo risulti essere fortemente influenzato sia dalla numerosità campionaria che dal numero di variabili irrilevanti ai fini dell'analisi di raggruppamento. Si può riscontrare, anche in questo frangente, il fatto che questo metodo fornisce delle partizioni qualitativamente migliori nel momento in cui i gruppi sono ben separati rispetto a

quando questi hanno una separazione minore. Inoltre, in ogni situazione, l'ARI_{PC} assume valori maggiori all'aumentare di d: questo andamento dell'indice può essere dovuto al fatto che, all'aumentare del rapporto tra variabili rilevanti ed irrilevanti, l'analisi delle componenti principali riesce effettivamente ad estrarre una quantità di informazione maggiore dai dati ottenendo così risultati migliori. Si fa notare come questo non accada valutando i valori assunti da ARI_R e come quindi ci sia un'indicazione riguardo il fatto che le componenti principali riescano a discriminare le variabili in maniera più efficiente.

L'analisi delle componenti principali sparse sembra essere un metodo tendenzialmente più stabile rispetto agli altri presi in esame in questo lavoro. Queste infatti, pur fornendo delle partizioni peggiori all'aumentare di w, riescono generalmente a cogliere la struttura di gruppo nei dati. In particolare, nel momento in cui si ha una numerosità campionaria elevata, ARI_{SPC} assume valori elevati sia nel caso di gruppi ben separati che di gruppi non ben separati e sembra inoltre non avere una forte influenza il numero di variabili irrilevanti ai fini del raggruppamento rilevate sui dati. La relazione tra l'andamento di ARI_{SPC} e la numerosità campionaria risulta evidente anche dal fatto che i risultati presentano una maggiore variabilità nelle situazioni in cui $n = 50$ mentre tendono a stabilizzarsi su valori più elevati e meno variabili con $n = 100, 250$. Per quel che riguarda il rapporto tra l'analisi delle componenti principali sparse e la quantità di variabili rilevanti ai fini della classificazione presenti nel sistema, sembra esserci una tendenza simile a quella riscontrata con l'utilizzo delle componenti principali sebbene questa risulti in questo caso meno evidente, presumibilmente a causa della maggiore stabilità di questo tipo di analisi. Un confronto tra questo metodo di riduzione della dimensionalità e gli altri presi in considerazione permette di concludere a favore dell'utilizzo delle componenti principali sparse che sembrano essere l'unico metodo, tra quelli presi in esame, ad essere in grado di mantenere una soddisfacente struttura di gruppo nella riduzione della dimensionalità e a fornire delle buone partizioni nella maggioranza degli scenari simulati.

Le valutazioni riguardanti la bontà del metodo proposto da Raftery e Dean (2006) sono meno generali in quanto si è riusciti a valutare questo approccio solamente nel caso in cui $w = 10$; risulta ovvio come potrebbe essere d'interesse riuscire ad ottenere i risultati anche nelle situazioni in cui si aggiunga una quantità di variabili irrilevanti maggiore, estendendo così l'analisi. I valori assunti da ARI_{RD} presentano

una relazione chiara con la numerosità campionaria, come già riscontrato con gli altri metodi. Si nota inoltre come, anche in questo caso, i risultati degradino nel passaggio dalle situazioni in cui i dati sono generati da una mistura di componenti gaussiane ben separate alle situazioni in cui queste presentino un grado di separazione minore. Questo metodo, a differenza degli altri due presi in considerazione in questo lavoro, non sembra portare a risultati migliori all'aumentare di d; sembra infatti seguire un andamento più simile a quello di ARI_R andando a fornire partizioni peggiori all'aumentare del numero di variabili rilevanti. Pur sottolineando nuovamente il fatto che sarebbe necessario riuscire a valutare la bontà di questo metodo in più situazioni rispetto a quanto fatto in questo studio, questo tipo di andamento sembra costituire un limite serio in quanto fornisce un'indicazione riguardo l'incapacità del metodo di Raftery e Dean (2006) di riuscire a selezionare le variabili contenenti una struttura di gruppo. Concludendo sembra che questo metodo, sebbene sia l'unico tra quelli esaminati ad esser stato pensato appositamente per essere utilizzato in un contesto di analisi di raggruppamento basata su modello, non fornisca un tipo di approccio valido per risolvere il problema dell'elevata dimensionalità.

Vengono infine fatte alcune valutazioni riguardanti i risultati ottenuti nel caso in cui i dati siano generati da distribuzioni non gaussiane e caratterizzate dalla presenza di una forte asimmetria. Si è deciso di non valutare congiuntamente questa situazione alle altre prese in esame in quanto i risultati riguardanti questo scenario presentano degli andamenti anomali, non del tutto in linea con le relazioni evidenziate fino ad ora e difficilmente interpretabili. Si è già notato come la forma dei gruppi sembri avere una forte influenza sui risultati forniti dal *clustering* basato su modello. È invece più complicato riuscire a spiegare il motivo per il quale l'indice ARI_R presenti valori migliori rispetto a quelli assunti da ARI; sembra infatti che l'aggiunta di variabili irrilevanti ai fini della partizione migliori la qualità della partizione stessa. In questo scenario si può poi vedere come, a causa della forma stessa dei gruppi, tendenzialmente si inverta la relazione tra i valori assunti dagli indici e la numerosità campionaria.

Una considerazione di ordine generale che si può fare riguarda il fatto che i risultati ottenuti sono molto variabili sia in media che in varianza; questo può dare un segnale riguardante la difficoltà riscontrata dall'algoritmo di *clustering* basato su modello nel fornire una partizione in gruppi dei dati. Una possibile motivazione di questa maggiore variabilità può essere riscontrata nella natura stessa dei gruppi

che, essendo di forma fortemente asimmetrica e quindi non ellittica, fa in modo che questo tipo di approccio all'analisi di raggruppamento, in quanto basato su modelli a mistura finita di componenti gaussiane, non sia adatto in queste situazioni. Un'ulteriore analisi riguardante il numero di gruppi forniti dal metodo di *clustering* in questi scenari, mette chiaramente in evidenza come la qualità inferiore delle partizioni fornite sia legata al fatto che spesso i valori anomali non vengono riconosciuti come appartenenti ad uno dei due gruppi presenti nei dati ma vengono raggruppati in dei *cluster* a sè stanti: questo fa sì che spesso l'algoritmo di *clustering* non riesca a cogliere il corretto numero dei gruppi andando di conseguenza a fornire dei valori inferiori dell'ARI.

		n=50	n=100	n=250
	ARI	0.988 [sd=0.047]	0.994 [sd=0.017]	0.994 [sd=0.009]
w=10	ARI_R	0.993 [sd=0.037]	0.994 [sd=0.015]	0.994 [sd=0.009]
	ARI_{PC}	0.068 [sd=0.245]	0.327 [sd=0.451]	0.973 [sd=0.021]
	ARI_{SPC}	0.991 [sd=0.026]	0.994 [sd=0.015]	0.994 [sd=0.010]
	ARI_{RD}	0.847 [sd=0.301]	0.992 [sd=0.021]	0.994 [sd=0.011]
w=25	ARI_R	0.988 [sd=0.067]	0.994 [sd=0.016]	0.994 [sd=0.010]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.317 [sd=0.435]
	ARI_{SPC}	0.988 [sd=0.032]	0.993 [sd=0.016]	0.993 [sd=0.010]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.115 [sd=0.317]	0.598 [sd=0.485]	0.975 [sd=0.133]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.046 [sd=0.203]
	ARI_{SPC}	0.989 [sd=0.029]	0.993 [sd=0.017]	0.994 [sd=0.010]
	ARI_{RD}	-	-	-

Tabella 4.1: Risultati delle simulazioni nel caso con $d = 2$, gruppi normali e ben separati. Vengono riportati il valore medio e lo *standard error* dell'ARI per ogni metodo utilizzato e al variare della numerosità campionaria n e del numero di variabili irrilevanti w.

		n=50	n=100	n=250
	ARI	0.878 [sd=0.139]	0.904 [sd=0.064]	0.908 [sd=0.037]
w=10	ARI_R	0.890 [sd=0.100]	0.901 [sd=0.063]	0.907 [sd=0.038]
	ARI_{PC}	0.000 [sd=0.004]	0.002 [sd=0.038]	0.714 [sd=0.311]
	ARI_{SPC}	0.838 [sd=0.198]	0.899 [sd=0.064]	0.907 [sd=0.038]
	ARI_{RD}	0.537 [sd=0.394]	0.543 [sd=0.446]	0.906 [sd=0.039]
w=25	ARI_R	0.224 [sd=0.385]	0.666 [sd=0.389]	0.901 [sd=0.056]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{SPC}	0.611 [sd=0.389]	0.879 [sd=0.097]	0.905 [sd=0.037]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.002 [sd=0.038]	0.004 [sd=0.059]	0.046 [sd=0.199]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{SPC}	0.485 [sd=0.430]	0.877 [sd=0.095]	0.901 [sd=0.039]
	ARI_{RD}	-	-	-

Tabella 4.2: Risultati delle simulazioni nel caso con $d = 2$, gruppi normali e non ben separati. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.235 [sd=0.266]	0.196 [sd=0.226]	0.196 [sd=0.164]
w=10	ARI_R	0.011 [sd=0.085]	0.014 [sd=0.083]	0.483 [sd=0.220]
	ARI_{PC}	0.000 [sd=0.005]	0.000 [sd=0.003]	0.000 [sd=0.003]
	ARI_{SPC}	0.117 [sd=0.250]	0.096 [sd=0.229]	0.086 [sd=0.196]
	ARI_{RD}	0.213 [sd=0.258]	0.197 [sd=0.226]	0.189 [sd=0.168]
w=25	ARI_R	0.000 [sd=0.002]	0.000 [sd=0.004]	0.001 [sd=0.004]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{SPC}	0.318 [sd=0.358]	0.203 [sd=0.309]	0.238 [sd=0.299]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.000 [sd=0.001]	0.000 [sd=0.000]	0.000 [sd=0.003]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{SPC}	0.144 [sd=0.277]	0.138 [sd=0.270]	0.100 [sd=0.232]
	ARI_{RD}	-	-	-

Tabella 4.3: Risultati delle simulazioni nel caso con $d = 2$ e gruppi non normali. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.989 [sd=0.034]	0.994 [sd=0.014]	0.995 [sd=0.009]
w=10	ARI_R	0.993 [sd=0.024]	0.994 [sd=0.015]	0.995 [sd=0.009]
	ARI_{PC}	0.985 [sd=0.050]	0.990 [sd=0.019]	0.993 [sd=0.010]
	ARI_{SPC}	0.985 [sd=0.077]	0.994 [sd=0.014]	0.995 [sd=0.009]
	ARI_{RD}	0.687 [sd=0.378]	0.984 [sd=0.100]	0.995 [sd=0.009]
w=25	ARI_R	0.923 [sd=0.247]	0.993 [sd=0.016]	0.995 [sd=0.011]
	ARI_{PC}	0.075 [sd=0.261]	0.858 [sd=0.324]	0.991 [sd=0.011]
	ARI_{SPC}	0.988 [sd=0.052]	0.993 [sd=0.016]	0.995 [sd=0.009]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.117 [sd=0.317]	0.304 [sd=0.452]	0.704 [sd=0.450]
	ARI_{PC}	0.000 [sd=0.000]	0.102 [sd=0.299]	0.990 [sd=0.012]
	ARI_{SPC}	0.989 [sd=0.300]	0.994 [sd=0.015]	0.994 [sd=0.009]
	ARI_{RD}	-	-	-

Tabella 4.4: Risultati delle simulazioni nel caso con $d = 5$, gruppi normali e ben separati. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.889 [sd=0.094]	0.903 [sd=0.094]	0.908 [sd=0.034]
w=10	ARI_R	0.888 [sd=0.091]	0.897 [sd=0.057]	0.907 [sd=0.034]
	ARI_{PC}	0.134 [sd=0.323]	0.388 [sd=0.442]	0.899 [sd=0.035]
	ARI_{SPC}	0.853 [sd=0.165]	0.897 [sd=0.056]	0.906 [sd=0.034]
	ARI_{RD}	0.391 [sd=0.379]	0.562 [sd=0.434]	0.886 [sd=0.143]
w=25	ARI_R	0.071 [sd=0.246]	0.163 [sd=0.346]	0.719 [sd=0.364]
	ARI_{PC}	0.000 [sd=0.007]	0.000 [sd=0.000]	0.137 [sd=0.324]
	ARI_{SPC}	0.743 [sd=0.304]	0.880 [sd=0.075]	0.901 [sd=0.036]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.000 [sd=0.000]	0.002 [sd=0.040]	0.004 [sd=0.058]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.004 [sd=0.057]
	ARI_{SPC}	0.519 [sd=0.428]	0.882 [sd=0.074]	0.901 [sd=0.036]
	ARI_{RD}	-	-	-

Tabella 4.5: Risultati delle simulazioni nel caso con $d = 5$, gruppi normali e non ben separati. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.518 [sd=0.142]	0.421 [sd=0.093]	0.314 [sd=0.044]
w=10	ARI_R	0.858 [sd=0.197]	0.710 [sd=0.162]	0.552 [sd=0.093]
	ARI_{PC}	0.092 [sd=0.282]	0.067 [sd=0.247]	0.313 [sd=0.460]
	ARI_{SPC}	0.697 [sd=0.257]	0.698 [sd=0.242]	0.734 [sd=0.169]
	ARI_{RD}	0.417 [sd=0.195]	0.421 [sd=0.095]	0.312 [sd=0.045]
w=25	ARI_R	0.445 [sd=0.493]	0.830 [sd=0.310]	0.663 [sd=0.125]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.002 [sd=0.043]
	ARI_{SPC}	0.819 [sd=0.238]	0.729 [sd=0.348]	0.802 [sd=0.220]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.022 [sd=0.147]	0.296 [sd=0.452]	0.734 [sd=0.173]
	ARI_{PC}	0.000 [sd=0.001]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{SPC}	0.454 [sd=0.450]	0.645 [sd=0.418]	0.671 [sd=0.407]
	ARI_{RD}	-	-	-

Tabella 4.6: Risultati delle simulazioni nel caso con $d = 5$ e gruppi non normali. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.988 [sd=0.060]	0.995 [sd=0.014]	0.994 [sd=0.010]
w=10	ARI_R	0.994 [sd=0.023]	0.994 [sd=0.015]	0.994 [sd=0.010]
	ARI_{PC}	0.978 [sd=0.108]	0.992 [sd=0.017]	0.993 [sd=0.010]
	ARI_{SPC}	0.980 [sd=0.104]	0.994 [sd=0.015]	0.994 [sd=0.010]
	ARI_{RD}	0.499 [sd=0.414]	0.741 [sd=0.412]	0.994 [sd=0.010]
w=25	ARI_R	0.531 [sd=0.480]	0.786 [sd=0.397]	0.993 [sd=0.010]
	ARI_{PC}	0.862 [sd=0.328]	0.984 [sd=0.079]	0.992 [sd=0.011]
	ARI_{SPC}	0.987 [sd=0.042]	0.992 [sd=0.017]	0.994 [sd=0.010]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.055 [sd=0.222]	0.152 [sd=0.351]	0.201 [sd=0.393]
	ARI_{PC}	0.083 [sd=0.274]	0.547 [sd=0.494]	0.992 [sd=0.011]
	ARI_{SPC}	0.988 [sd=0.031]	0.993 [sd=0.016]	0.993 [sd=0.011]
	ARI_{RD}	-	-	-

Tabella 4.7: Risultati delle simulazioni nel caso con $d = 10$, gruppi normali e ben separati. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.890 [sd=0.092]	0.900 [sd=0.060]	0.906 [sd=0.037]
w=10	ARI_R	0.887 [sd=0.108]	0.897 [sd=0.059]	0.904 [sd=0.038]
	ARI_{PC}	0.210 [sd=0.383]	0.519 [sd=0.448]	0.902 [sd=0.038]
	ARI_{SPC}	0.855 [sd=0.159]	0.891 [sd=0.063]	0.902 [sd=0.040]
	ARI_{RD}	0.237 [sd=0.294]	0.312 [sd=0.407]	0.624 [sd=0.418]
w=25	ARI_R	0.014 [sd=0.112]	0.009 [sd=0.091]	0.038 [sd=0.180]
	ARI_{PC}	0.001 [sd=0.276]	0.002 [sd=0.045]	0.099 [sd=0.286]
	ARI_{SPC}	0.692 [sd=0.341]	0.865 [sd=0.069]	0.895 [sd=0.039]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.000 [sd=0.000]	0.000 [sd=0.000]	0.000 [sd=0.000]
	ARI_{PC}	0.000 [sd=0.000]	0.000 [sd=0.000]	0.002 [sd=0.043]
	ARI_{SPC}	0.227 [sd=0.374]	0.854 [sd=0.090]	0.893 [sd=0.040]
	ARI_{RD}	-	-	-

Tabella 4.8: Risultati delle simulazioni nel caso con $d = 10$, gruppi normali e non ben separati. Cfr. Tabella 4.1.

		n=50	n=100	n=250
	ARI	0.522 [sd=0.127]	0.393 [sd=0.090]	0.291 [sd=0.038]
w=10	ARI_R	0.861 [sd=0.167]	0.668 [sd=0.151]	0.445 [sd=0.097]
	ARI_{PC}	0.959 [sd=0.167]	0.985 [sd=0.047]	0.970 [sd=0.059]
	ARI_{SPC}	0.752 [sd=0.163]	0.743 [sd=0.142]	0.718 [sd=0.114]
	ARI_{RD}	0.385 [sd=0.153]	0.375 [sd=0.072]	0.286 [sd=0.033]
w=25	ARI_R	0.984 [sd=0.064]	0.915 [sd=0.144]	0.641 [sd=0.106]
	ARI_{PC}	0.590 [sd=0.490]	0.644 [sd=0.479]	0.987 [sd=0.100]
	ARI_{SPC}	0.873 [sd=0.133]	0.888 [sd=0.105]	0.849 [sd=0.105]
	ARI_{RD}	-	-	-
w=35	ARI_R	0.987 [sd=0.095]	0.979 [sd=0.075]	0.724 [sd=0.127]
	ARI_{PC}	0.012 [sd=0.109]	0.056 [sd=0.230]	0.828 [sd=0.378]
	ARI_{SPC}	0.909 [sd=0.155]	0.926 [sd=0.106]	0.925 [sd=0.075]
	ARI_{RD}	-	-	-

Tabella 4.9: Risultati delle simulazioni nel caso con $d = 10$ e gruppi non normali. Cfr. Tabella 4.1.

4.2 Un'applicazione a dati genetici

4.2.1 Presentazione del problema e dei dati

In questo paragrafo i metodi presentati ed utilizzati fino ad ora vengono applicati ad un insieme di dati di espressione genica rilevati mediante tecnologia *microarray*. I *microarray* sono delle collezioni di sonde di DNA attaccate ad un supporto solido; in questo modo si forma una matrice (*array*) che permette di analizzare in maniera simultanea un insieme di geni molto grande all'interno di un campione di DNA. L'utilizzo forse più tipico dei *microarray* riguarda la possibilità di confrontare il profilo genetico di un individuo affetto da una qualche patologia con il profilo di un individuo sano; in questo modo si cerca di comprendere quali geni codifichino per quella determinata patologia permettendo poi in seguito la sintesi e l'implementazione di terapie geniche mirate. In letteratura infatti si è notato come alcuni virus modifichino l'espressione di alcuni geni per riuscire a sopravvivere nella cellula ospite. Ad esempio, in ambito oncologico, l'analisi dei *microarray* può, mostrando valori significativamente alterati tra soggetti sani e soggetti malati, essere utile sia per cogliere la presenza di un eventuale tumore, sia per riuscire a scoprirne nuove classi.

In questo frangente, la statistica assume una certa importanza poichè ha il compito di trovare dei metodi che riescano ad estrarre informazione rilevante dall'analisi dei *microarray* e dalle diverse espressioni assunte dai geni studiati. È facile capire come studi di questo tipo pongano problemi risolvibili solamente con l'utilizzo di tecniche statistiche multivariate in quanto si hanno matrici di *microarray* tipicamente con un numero di variabili rilevate p molto maggiore rispetto al numero di unità campionarie n.

In questo contesto il *clustering* ha assunto una particolare rilevanza differenziandosi in due percorsi aventi scopi differenti: fornire una partizione dei geni in gruppi sulla base dei tessuti (unità campionarie) studiati oppure una partizione dei tessuti sulla base dei geni rilevati (variabili). In questo lavoro si considera una situazione nella quale si vuole ottenere una partizione dei tessuti, e quindi dei soggetti, utilizzando l'analisi di raggruppamento basata su modello. In particolare, l'utilizzo di metodi di *clustering* per ottenere una partizione dei soggetti si pone tipicamente come obiettivo quello di individuare particolari patologie, o di distringuere, in campioni di individui affetti dalla medesima patologia, delle manifestazioni alternative della stessa.

Si è già visto in precedenza come, nel caso in cui si abbia una numerosità campionaria bassa se confrontata con il numero di variabili rilevate, l'approccio parametrico all'analisi di raggruppamento fornisca risultati non buoni e molto variabili a causa della sovraparametrizzazione della matrice di varianze e covarianze. Si cerca quindi di applicare i metodi spiegati ed utilizzati in precedenza per risolvere il problema della dimensionalità in una situazione in cui, essendo $n < p$, una riduzione o una selezione delle variabili risulta essere necessaria per riuscire a cogliere l'informazione realmente rilevante e, in questo caso, a capire quali geni assumano espressioni differenziate tra pazienti sani e pazienti malati o entro diverse manifestazioni della stessa malattia.

Nello specifico in questo lavoro sono stati utilizzati dei dati raccolti da Alon *e altri* (1999) riguardanti l'espressione di 2000 geni rilevati su 62 individui. Questi ultimi possono essere divisi in due classi: per 40 di essi è nota la presenza di una forma tumorale al colon, mentre i restanti 22 sono pazienti sani. L'etichetta di appartenenza al gruppo di pazienti sani o malati potrà essere utilmente utilizzata per valutare la qualità del partizionamento ottenuto e a fini interpretativi.

4.2.2 Metodi utilizzati e risultati ottenuti

In questo studio, per cercare di ottenere delle buone partizioni dei dati che riescano a differenziare tra individui sani e individui affetti da patologia tumorale, si è utilizzato un approccio ripreso da De Bin e Risso (2011). Questo approccio prevede tre *step* successivi:

- Normalizzazione e filtraggio dei geni;
- Riduzione della dimensionalità;
- *Clustering* nello spazio ridotto.

Le giustificazioni che portano ad operare in questo modo sono riconducibili al fatto che, filtrando i geni e riducendo la dimensionalità dei dati prima di proseguire con l'analisi di raggruppamento, si cerca di fare in modo di utilizzare le variabili realmente rilevanti e di operare in sottospazi di dimensione inferiore dove quindi l'approccio basato su modello presenta risultati più soddisfacenti.

Seguendo Dudoit *e altri* (2002) si è applicata innanzitutto una trasformazione logaritmica alle colonne della matrice sulla quale si opera; questo perchè è noto che

le variabili relative alle espressioni genetiche presentano spesso una distribuzione asimmetrica e quindi, grazie a questa trasformazione, si riesce generalmente a ricondursi ad una forma maggiormente compatibile con l'ipotesi di normalità e ad alleggerire il problema riguardante l'eventuale presenza di valori anomali. Questa trasformazione è a maggior ragione utile nell'ottica di una successiva applicazione di modelli a mistura finita con componenti gaussiane. Inoltre si vanno a standardizzare le righe (unità statistiche) della matrice contenente i dati in modo tale che abbiano media nulla e varianza unitaria: viene fatto ciò per evitare che un eventuale livello elevato assunto dall'espressione di un gene influenzi pesantemente il livello medio tra le osservazioni. Questo tipo di standardizzazione è coerente con il fatto che, solitamente, in analisi di questo tipo per misurare la similarità tra due tessuti si utilizza la correlazione tra le espressioni dei geni studiati.

In accordo con Dudoit *e altri* (2002) si è inoltre eseguita un'operazione di *thresholding* in cui le espressioni minori di una certa soglia minima assunte dai geni vengono poste uguali al valore della soglia stessa e, analogamente, le espressioni maggiori di una soglia massima sono vincolate ad essere uguali a questa.

Inoltre è stata applicata un'operazione di *filtering* che va per prima ad eseguire un filtraggio e una selezione delle variabili da utilizzare; questa infatti, dati due valori, esclude dal data set quelle variabili che hanno il rapporto tra il valore massimo e il valore minimo minore o uguale al primo dei due valori e la differenza tra il massimo e il minimo valore assunto minore uguale al secondo. È facile comprendere come anche questa operazione miri a limitare l'impatto dei valori anomali sulle successive analisi.

Si fa notare come queste operazioni di preprocessamento siano state applicate ai dati in questione utilizzando la funzione *preprocess* presente in R nel paccetto *plsgenomics* (Boulesteix *e altri*, 2012), che riprende le operazioni proposte da Dudoit *e altri* (2002) ed è stata implementata appositamente per lo studio di dati da *microarray*.

A questa prima operazione di filtraggio e preprocessamento dei dati si è deciso, seguendo quanto proposto da De Bin e Risso (2011), di aggiungere una fase di preselezione andando a considerare solamente le variabili la cui distribuzione univariata presentasse una qualche struttura di gruppo ed escludendo le altre. Si son quindi mantenuti nel data set solamente quei geni i quali, andando ad applicare un algoritmo di *clustering*, forniscono un numero di gruppi maggiore o uguale a

Metodo	Num. Variabili	Num. Gruppi	ARI
No preprocess, no preselezione	2000	5	0.0033
Preprocess	327	3	0.3227
Preprocess + preselezione	87	4	0.4220
Preprocess + preselezione + componenti principali	8	4	0.5257
Preprocess + preselezione + componenti principali sparse	8	3	0.5518
Preprocess + preselezione + Raftery e Dean	5	1	0.0000

Tabella 4.10: Risultati ottenuti sui dati riguardanti la presenza o assenza di una patologia tumorale al colon con diversi *setting*.

2. Un'operazione di questo tipo trova giustificazione nel fatto che, così facendo, si considerano come irrilevanti le variabili che al loro interno non presentano una differenziazione in gruppi.

Per quanto riguarda invece il secondo passo, e cioè la riduzione della dimensionalità, si è fatto ricorso ai tre metodi spiegati nel paragrafo 3.2 e utilizzati nello studio di simulazione. Per introdurre la sparsità dei coefficienti nell'analisi delle componenti principali sparse, si è operato nello stesso modo descritto nel capitolo 4.

I risultati ottenuti sono riportati nella tabella 4.10. Si sottolinea il fatto che con *preprocess* si intendono le operazioni di filtraggio e normalizzazione proposte da Dudoit *e altri* (2002) mentre con *preselezione* si intende l'operazione di selezione delle variabili proposta da De Bin e Risso (2011) a cui si è accennato in precedenza. Per quanto riguarda le componenti principali e le componenti principali sparse si è deciso di mantenere un numero di componenti tale che queste spieghino il 70% della variabilità originale del sistema, mantenendo così il parallelismo con quanto fatto nello studio di simulazione.

È immediato notare come non sia possibile ottenere una partizione sensata dei dati nel caso in cui non si operi nessuna operazione preliminare; si vede infatti come l'ARI fornisca un valore molto basso, e indicante una partizione in completo disaccordo con la vera struttura di gruppo, nel caso in cui non si vada ad applicare nessun tipo di preprocessamento ai dati. Si può vedere invece come, anche solamente applicando le operazioni proposte da Dudoit *e altri* (2002), si riduca di molto il numero di variabili e migliori notevolmente il valore dell'ARI. Questi valori migliorano ulteriormente dopo l'applicazione dell'operazione di preselezione e dell'analisi

	Gruppo 1	Gruppo 2	Gruppo 3	Gruppo 4
Soggetti sani	11	1	2	8
Soggetti malati	0	3	34	3

Tabella 4.11: Tabella di contingenza riguardante la partizione ottenuta dopo aver applicato il preprocessamento, la preselezione e l'analisi delle componenti principali.

	Gruppo 1	Gruppo 2	Gruppo 3
Soggetti sani	1	14	7
Soggetti malati	34	3	3

Tabella 4.12: Tabella di contingenza riguardante la partizione ottenuta dopo aver applicato il preprocessamento, la preselezione e l'analisi delle componenti principali sparse.

delle componenti principali e delle componenti principali sparse. Si sottolinea come questi metodi ci permettano di ridurre il numero di variabili a tal punto da avere $n > p$, pur mantenendo una quota elevata di variabilità spiegata.

Applicando i metodi di riduzione a questi dati, risulta più complicato fornire un ordinamento, in termini di qualità degli stessi, rispetto al caso in cui sono stati applicati ai dati simulati. Una considerazione generale e valida per qualsiasi metodo si sia applicato riguarda l'evidente difficoltà che il *clustering* basato su modello presenta nel cogliere il corretto numero dei gruppi. Questa limitazione, in una situazione in cui i due gruppi indicano la presenza o l'assenza di una patologia tumorale, potrebbe in realtà fornire un'indicazione riguardo il fatto che esistano delle sottocategorie tumorali. In questo caso, come si può osservare dalle tabelle 4.11 e 4.12, non sembra delinearsi una situazione di questo tipo. Si noti, tuttavia, che il gruppo dei soggetti malati viene rilevato con una buona precisione sia applicando l'algoritmo di *clustering* dopo aver utilizzato le componenti principali che le componenti principali sparse. Invece le partizioni individuate suggeriscono una scarsa omogeneità nel gruppo di pazienti sani. Non è noto se tali disomogeneità siano dovute alla presenza di qualche altra patologia e tali risultati dovrebbero essere interpretati da un esperto in campo genetico.

Si è inoltre notato (si vedano figure 4.3 e 4.4) come, sia nel caso delle componenti principali sia nel caso delle componenti principali sparse, l'informazione contenente la struttura di gruppo sembri essere contenuta nella prima componente principale e si perda andando a considerare le altre componenti. In questa situazione quindi

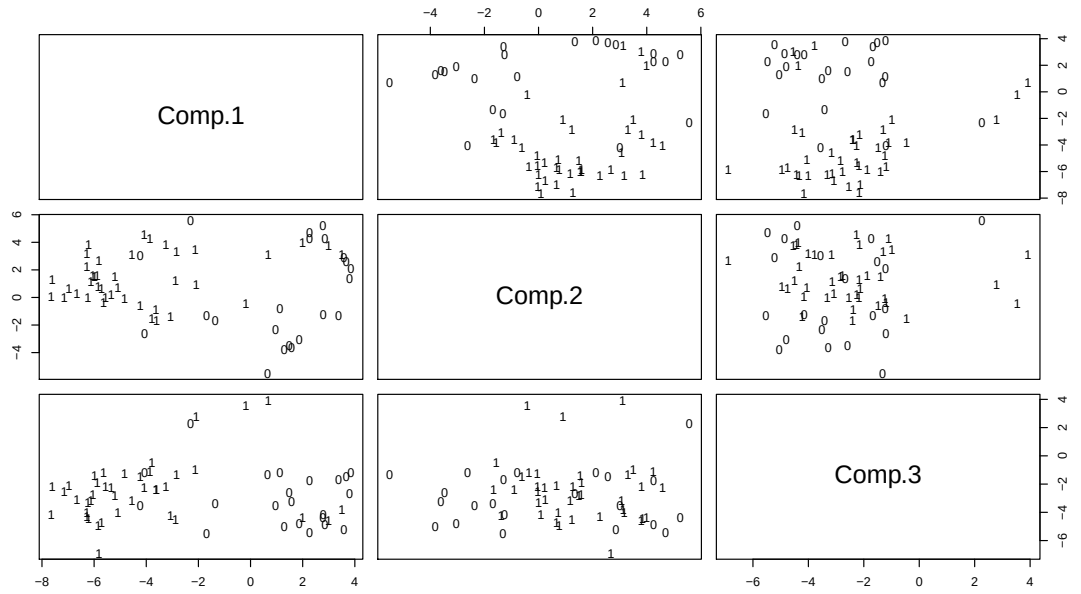

Figura 4.3: Prime tre componenti principali calcolate sui dati dopo preprocess e preselezione. Con 0 vengono indicati i soggetti sani mentre con 1 vengono indicati i soggetti malati.

non sembrano presentarsi le criticità evidenziate nel paragrafo 3.3.1 (Chang, 1983) in quanto la componente che spiega la maggior quota di variabilità mantiene al proprio interno una struttura di gruppo e quindi informazione rilevante ai fini dell'analisi di raggruppamento.

Per quanto riguarda il metodo proposto da Raftery e Dean (2006) quel che si può facilmente vedere è l'incapacità di distinguere una struttura di gruppo. Si evidenzia come questa limitazione sia stata notata anche nel caso in cui si siano provati *setting* e operazioni di preprocessamento differenti; in queste altre situazioni, anche nei casi in cui viene rilevato un numero di *cluster* maggiore di uno, l'ARI mantiene un valore comunque molto basso andando ad indicare una partizione dei dati pressochè casuale e dando un'indicazione forte a riguardo dell'incapacità di questo metodo di fornire una buona selezione ai fini del raggruppamento.

Infine si è condotta un'ulteriore analisi per cercare di comprendere quali possano essere i geni che, secondo i metodi utilizzati, maggiormente si differenziano tra i pazienti sani e i pazienti malati. Per quel che riguarda il metodo di Raftery e Dean (2006) questo tipo di analisi risulta immediata in quanto l'algoritmo stesso seleziona le variabili considerate rilevanti mentre, nel contesto dell'analisi delle componen-

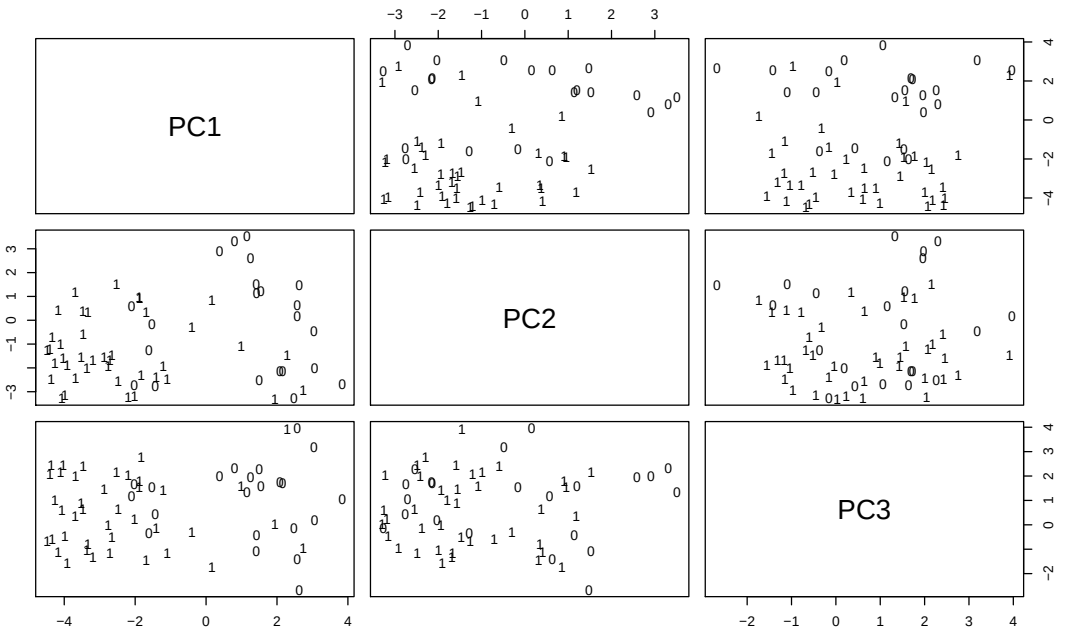

Figura 4.4: Prime tre componenti calcolate applicando le componenti principali sparse sui dati dopo preprocess e preselezione. Con 0 vengono indicati i soggetti sani mentre con 1 vengono indicati i soggetti malati.

ti principali e delle componenti principali sparse, si son considerate come rilevanti le variabili che presentano, in valore assoluto, coefficienti elevati in corrispondenza delle componenti selezionate. Si può far notare come alcune variabili siano considerate rilevanti ai fini del raggruppamento da tutti e tre i metodi. Tra le variabili che sembrano maggiormente differenziare il gruppo dei soggetti sani dal gruppo dei soggetti malati si possono citare: Hsa.8192, Hsa.1039, Hsa.3068, Hsa.1825, Hsa.848, Hsa.3409. Potrebbe quindi risultare d'interesse approfondire il ruolo di questi geni confrontando quanto ottenuto in questo lavoro con risultati di altri studi.

4.3 Conclusioni

L'obiettivo di questo capitolo, e più in generale di questo lavoro, è stato quello di studiare il comportamento dell'approccio al *clustering* basato su modelli a mistura finita nel caso in cui si operi in spazi di dimensione elevata. Inoltre si è cercato di comprendere, nel caso in cui l'analisi di raggruppamento non fornisca delle buone partizioni, quali metodi riescano a ridurre la dimensionalità dei dati permettendo di ottenere dei risultati qualitativamente migliori.

Per rispondere a questi obiettivi, i differenti metodi trattati sono stati utilizzati sia in uno studio di simulazione sia applicati a dei dati reali provenienti da un analisi di *microarray*.

Alla luce dei risultati ottenuti in queste due differenti analisi condotte si possono trarre alcune considerazioni di ordine generale.
Innanzitutto si è visto come il *clustering* basato su modello riscontri effettivamente delle serie difficoltà, dovute alla sparsità dei dati e all'elevato numero di parametri da stimare, nel cogliere la struttura di gruppo, qualora sui dati sia stato rilevato un numero elevato di variabili. La gravità del problema si riduce con l'applicazione di opportune tecniche di riduzione della dimensionalità, che riescano a ricondurre l'analisi su sottospazi di dimensione inferiore selezionando l'informazione maggiormente rilevante. Nell'analizzare i dati provenienti dallo studio di *microarray*, dove si ha un numero di variabili rilevate maggiore rispetto alla numerosità campionaria, questa necessità risulta essere ancora più evidente.

Si è osservato poi come, tra i metodi presi in considerazione in questo lavoro, l'analisi delle componenti principali sparse sembri essere quello meno influenzato dai cambiamenti di *setting* fornendo così un'indicazione forte riguardo la stabilità di questo approccio. Le componenti principali sparse hanno infatti fornito risultati migliori in quasi tutti gli scenari simulati e hanno inoltre mostrato un buon comportamento anche nel momento in cui siano state applicate ai dati di *microarray*.

Un'osservazione importante riguarda il metodo proposto da Raftery e Dean (2006). Questo, tra i metodi approfonditi in questo lavoro, risulta essere l'unico introdotto appositamente per risolvere il problema della dimensionalità nel caso in cui si utilizzi un approccio al *clustering* basato su modelli a mistura finita. I risultati ottenuti in questo studio sono però scoraggianti e, se si considerano inoltre i problemi di natura computazionale del metodo, portano a sconsigliarne l'utilizzo in queste situazioni.

Un'ulteriore considerazione riguarda la scarsa robustezza dell'approccio al *clustering* basato su modelli mistura quando i gruppi non siano compatibili con l'ipotesi di normalità. Questo dipende ovviamente dal fatto che le analisi di raggruppamento, in questo lavoro, sono state condotte basandosi su modelli a mistura finita con componenti gaussiane.

Ovviamente le considerazioni che si possono trarre da questo lavoro risultano non essere generali e, soprattutto per quanto riguarda lo studio di simulazione, sono

circoscritte agli scenari ai quali i metodi studiati sono stati applicati. Eventuali approfondimenti potrebbero quindi riguardare un'estensione dell'analisi valutando la bontà di questo approccio al *clustering* e dei metodi di riduzione delle variabili in contesti più ampi e generali. In particolare, sarebbe opportuno considerare situazioni in cui i gruppi abbiano un maggior grado di sovrapposizione, non siano ugualmente rappresentati nel campione o presentino delle forme differenti. Potrebbe essere d'interesse, in particolar modo negli scenari in cui i gruppi non siano generati da distribuzioni gaussiane, utilizzare componenti della mistura descritte da distribuzioni più flessibili cercando così di alleviare le difficoltà sopracitate. Inoltre potrebbe essere utile considerare situazioni in cui le variabili irrilevanti ai fini del raggruppamento non siano indipendenti tra loro e/o non siano indipendenti dalle variabili considerate rilevanti.

Per quanto riguarda la capacità del *clustering* di cogliere in maniera corretta la struttura di gruppo quando si modellano dati provenienti da un'analisi di *microarray*, si potrebbero fare diverse considerazioni. Pur non essendo lo scopo di questo lavoro quello di valutare specificatamente il funzionamento dell'approccio parametrico all'analisi di raggruppamento nel caso in cui si abbiano dati genetici, si evidenzia come, per ottenere buoni risultati in tali situazioni, sia necessaria una maggiore integrazione tra diversi ambiti scientifici e tra diversi studi. Si può comunque notare come, se abbinato a tecniche di preprocessamento e di riduzione della dimensionalità, il *clustering* basato su modello sembri poter essere un buon approccio da approfondire ulteriormente.

Bibliografia

Aitkin M.; Anderson D.; Hinde J. (1981). Statistical modelling of data on teaching styles. *Journal of the Royal Statistical Society. Series A (General)*, **144**(4), 419–461.

Alon U.; Barkai N.; Notterman D. A.; Gish K.; Ybarra S.; Mack D.; Levine A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, **96**(12), 6745–6750.

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

Azzalini A.; Scarpa B. (2004). *Analisi dei dati e data mining*. Springer.

Banfield J. D.; Raftery A. E. (1989). Model-based gaussian and non-gaussian clustering. Relazione tecnica, DTIC Document.

Bellman R. (1957). *Dynamical programming*. Princeton University Press, Princeton, NJ.

Biernacki C.; Govaert G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, **64**(1), 49–71.

Boulesteix A. L.; Lambert-Lacroix S.; Peyre J.; Strimmer K. (2012). *Plsgenomics: PLS analysis for genomics*. R package version 1.2-6.

Bouveyron C.; Brunet-Saumard C. (2014). Model-based clustering of high-dimensional data: a review. *Computational Statistics & Data Analysis*, **71**, 52–78.

Celeux G.; Diebolt J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem. *Computational Statistics Quarterly*, **2**(1), 73–82.

Celeux G.; Govaert G. (1995). Gaussian parsimonious clustering models. *Pattern recognition*, **28**(5), 781–793.

Chang W. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, **32**(3), 267–275.

Cormack R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (General)*, **134**(3), 321–367.

De Bin R.; Risso D. (2011). A novel approach to the clustering of microarray data via nonparametric density estimation. *BMC Bioinformatics*, **12**(1), 49.

Dempster A. P.; Laird N. M.; Rubin D. B. *e altri* (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal statistical Society*, **39**(1), 1–38.

Dudoit S.; Fridlyand J.; Speed T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**(457), 77–87.

Dy J. G.; Brodley C. E. (2004). Feature selection for unsupervised learning. *The Journal of Machine Learning Research*, **5**, 845–889.

Fraley C. (1998). Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, **20**(1), 270–281.

Fraley C.; Raftery A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, **41**(8), 578–588.

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

Fraley C.; Raftery A. E.; Murphy T. B.; Scrucca L. (2012). *Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation.*

Genz A.; Bretz F.; Miwa T.; Mi X.; Leisch F.; Scheipl F.; Hothorn T. (2014). Mvtnorm: Multivariate normal and t distributions. *R package version 0.9-99992, URL http://CRAN.R-project.org/package=mvtnorm*.

Gnanadesikan R.; Kettenring J.; Tsao S. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, **12**(1), 113–136.

Hotelling H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, **24**(6), 417–441.

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

Jolliffe I. (2005). *Principal component analysis*. Wiley Online Library.

Jolliffe I. T.; Trendafilov N. T.; Uddin M. (2003). A modified principal component technique based on the lasso. *Journal of Computational and Graphical Statistics*, **12**(3), 531–547.

Kass R. E.; Raftery A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**(430), 773–795.

Kaufman L.; Rousseeuw P. (1987). *Clustering by means of medoids*. North-Holland.

Keribin C. (2000). Consistent estimation of the order of mixture models. *Sankhya Ser. A*, **62**(1), 49–66.

Law M. H.; Figueiredo M. A.; Jain A. K. (2004). Simultaneous feature selection and clustering using mixture models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **26**(9), 1154–1166.

MacQueen J. *e altri* (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. California, USA.

Mardia K. V.; Kent J. T.; Bibby J. M. (1980). *Multivariate analysis*. Academic press.

Maugis C.; Celeux G.; Martin-Magniette M. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, **65**(3), 701–709.

McLachlan G. J.; Basford K. E. (1988). *Mixture models. Inference and applications to clustering*. New York: M. Dekker.

McLachlan G. J.; Krishnan T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

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

McLachlan G. J.; Bean R.; Peel D. (2002). A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics*, **18**(3), 413–422.

Melnykov V.; Melnykov I. (2012). Initializing the em algorithm in gaussian mixture models with an unknown number of components. *Computational Statistics & Data Analysis*, **56**(6), 1381–1395.

Melnykov V.; Maitra R. *e altri* (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, **4**, 80–116.

Pearson K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, **185**, 71–110.

Pearson K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **2**(11), 559–572.

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

Raftery A. E.; Dean N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, **101**(473), 168–178.

Schwarz G. *e altri* (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Scott D. W.; Thompson J. R. (1983). Probability density estimation in higher dimensions. In *Computer Science and Statistics: Proceedings of the Fifteenth Symposium on the Interface*. North-Holland, Amsterdam.

Scrucca L.; Raftery A. E.; Dean N. (2013). Clustvarsel: A package implementing variable selection for model-based clustering in R. *(to be submitted) Journal of Statistical Software.*

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Yeung K. Y.; Ruzzo W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**(9), 763–774.

Zou H.; Hastie T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301–320.

Zou H.; Hastie T. (2012). *Elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.1.

Zou H.; Hastie T.; Tibshirani R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**(2), 265–286.

RINGRAZIAMENTI

Alla professoressa Giovanna Menardi va un ringraziamento particolare per aver sopportato di seguire un non-smanettone, per tutto quello che mi ha insegnato, per avermi seguito con una pazienza infinita e per aver saputo spronarmi a fare del mio meglio.

A mia mamma, che per tutta la vita ha dovuto lavorare il doppio per riuscire a darmi più del doppio, per essere stata sempre esigente e per avermi insegnato che l'impegno e la passione in quel che si fa battono tutto il resto.

A mia nonna, che "guai se non ci fosse stata".

Ad Irene, per essere la cugina più fastidiosa e petulante sulla faccia della terra.

A Fabio, per la sua amicizia continua e duratura e per essere, in realtà, non un amico ma parte della famiglia.

A chi mi ha insegnato a volare in alto.

A Sara, per la condivisione dei sogni.

A tutti quelli che hanno creduto in me anche nei momenti più difficili, che mi hanno sopportato e che mi sono stati vicini. Conoscendomi, non deve essere stato facilissimo.

E infine un grazie speciale va a chi invece non ci credeva.