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

Statistica, Economia e Finanza

Clustering di dati three-way basato su misture di matrici normali: alcuni aspetti teorici e uno studio di simulazione

Relatore: dott.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Jacopo Diquigiovanni
Matricola n. 1052949

Anno Accademico 2014/2015

Indice

Capitolo 1

Introduzione

La rivoluzione a cui si sta assitendo in campo statistico da qualche decennio a questa parte è qualcosa di assolutamente singolare nel corso della storia: i dati crescono, e lo fanno ad un ritmo impressionante. Secondo l'analista americano Doug Laney i *Big Data* tendono, nel corso del tempo, ad aumentare in riferimento a tre caratteristiche interdipendenti: volume, velocità di generazione e varietà. Estendendo tale concetto ai *dataset* con cui siamo soliti confrontarci, la dinamica di fondo rimane la medesima: sono disponibili più unità statistiche, e con queste cresce l'informazione teoricamente estraibile dalle stesse. Per fare ciò è necessario innanzitutto creare strutture capaci di gestire tale mole di dati in maniera adeguata, in modo da minimizzare la perdita di informazione e permetterne la lettura in maniera più semplice possibile. In questa prospettiva, questa tesi si concentra su dati *three-way* che rappresentano, nel senso sopra menzionato, un efficace compromesso tra capacità di sintesi e ricchezza concettuale: questi sono caratterizzati da 3 diverse modalità, ossia *osservazioni, variabili* e *occasioni*, in modo da poter gestire l'evoluzione (nel corso del tempo, dello spazio ecc..) delle differenti variabili rilevate. Ogni osservazione viene pertanto arricchita, rispetto ai tradizionali *dataset two-way*, con una modalità in più, determinando la creazione di una struttura che annovera, oltre a righe e colonne, degli *strati*.

L'aumento dei dati a disposizione e la crescita dell'informazione *reale* spesso però non vanno di pari passo a causa dell'aumento del rumore, ossia di contenuto non rilevante. Risulta pertanto necessaria la specificazione di procedure mirate all'estrapolazione di informazione veramente utile: una tra queste è rappresentata dalla *cluster analysis*. Sotto tale definizione viene racchiuso un vasto insieme di tecniche di analisi multidimensionale aventi come scopo il partizionamento dello spazio campio-

nario in gruppi in accordo con un qualche criterio guida. Fare *clustering* non solo permette di suddividere le osservazioni di cui si dispone, ma anche di generare una regola di classificazione più ampia applicabile a unità future provenienti dalla stessa popolazione. Le applicazione di tale metodologia però non si esauriscono qua: si pensi per esempio alla possibilità di condurre un'analisi di raggruppamento e, in seguito, selezionare un determinato numero di osservazioni per ciascun gruppo. Così facendo si può ottenere una riduzione della numerosità del *dataset*, con benefici in termini computazionali in presenza di *Big Data*, senza andare ad intaccare in maniera significativa l'eterogeneità presente dei dati.

L'obiettivo del presente lavoro è quello di applicare l'azione di *clustering* ai dati *three-way*: a fronte di una procedura più complessa, l'analisi di gruppo per questi dati sfrutta l'informazione portata in dote dalla modalità *in più* per formare *cluster* di individui simili non solo per quanto riguarda le caratteristiche osservate, ma anche la loro evoluzione nel tempo o nello spazio. Tra i tanti approcci possibili quello discusso si fonda sulla specificazione di un modello probabilistico alla base del processo generatore dei dati, inquadrando dunque la metodologia in un contesto statisticamente rigoroso: si parla in tal caso di *clustering parametrico*. A differenza di altri *modus operandi*, quello proposto gode di alcuni dei benefici dovuti all'individuazione di un modello statistico, permettendo ad esempio di applicare procedure inferenziali per determinare il numero di gruppi presenti nei dati o una valutazione della probabilità che ogni unità appartenga ai diversi gruppi. Chiaramente esistono anche dei limiti: la presenza di ipotesi distributive riduce il campo di applicabilità del metodo, portando a risultati qualitativamente peggiori qualora gli assunti di base non siano rispettati. Inoltre, l'utilizzo di un gran numero di parametri (necessari per la stima dei gruppi) appesantisce il modello in maniera evidente, rendendo necessarie, come vedremo in seguito, ulteriori restrizioni.

La trattazione si svolge come segue.

Nel capitolo 2 vengono introdotti i dati *three-way*, definendone struttura e utilizzo. Al fine di inserirli in un contesto statistico, vengono analizzate le distribuzioni matrici-variate, ossia distribuzioni di probabilità o densità specifiche per matrici aleatorie. Tra queste, la più importante è quella gaussiana, che nel corso della presentazione si dimostrerà essere un caso specifico della distribuzione multivariata normale.

Nel capitolo 3 viene presentato il concetto di *cluster analysis* e la sua utilità nei più disparati ambiti. Nell'ottica di combinare il modello distributivo presentato nel capitolo precedente e l'analisi di raggruppa-

mento viene studiato l'approccio parametrico al *clustering*. In tale ottica si ipotizza che i dati provengano da una *mistura* di distribuzioni normali, in cui ogni componente definisce uno specifico gruppo, caratterizzato quindi da un certo numero di parametri: la stima viene condotta con il metodo della massima verosimiglianza grazie all'alternanza di due differenti step all'interno dell'algoritmo iterativo Expectation-Maximization (EM). Per ottenere il numero ottimale di *cluster* presenti nei dati viene utilizzato un approccio Bayesiano basato sul confronto delle probabilità a posteriori di differenti modelli. Dopo aver considerato alcune limitazioni presenti in tale formalizzazione, vengono esplorate alcune tecniche per il *clustering three-way* diverse da quella presentata: aspetto comune in tutte è il tentativo di comprimere i dati in uno schema a due vie, proposito che comporta un'evidente distorsione.

Nel capitolo 4 viene analizzato uno studio di simulazione: principale obiettivo dell'esperimento è quello di valutare la bontà delle stime proposte al variare di alcune condizioni circa la dimensionalità delle osservazioni, il soddisfacimento delle ipotesi e la divisione tra i gruppi. Infine vengono proposte alcune considerazioni generali sul lavoro svolto e messi in luce possibili ampliamenti.

Capitolo 2

Modellazione di matrici casuali

2.1 I dati *three-way*

La definitiva consacrazione del *computer* e degli altri strumenti di raccolta e trasmissione dell'informazione avvenuta negli ultimi anni ha radicalmente modificato il rapporto tradizionalmente esistente tra l'individuo e la statistica: a fronte di un ovvio aumento delle problematiche relative alla gestione di tale mole di dati, l'aumento esponenziale di informazione disponibile ha comportato la possibilità di condurre analisi statistiche sempre più accurate e precise. Ma il gravoso *trade-off* tra pesantezza computazionale e precisione delle stime non è l'unico aspetto da considerare in tale frangente. Difatti, l'esplosione di dati aventi spesso strutture molto articolate porta di pari passo la necessità di gestire, innanzitutto da un punto di vista logico-formale, tale complessità: una possibile formalizzazione, tra le miriadi esistenti, è rappresentata dai *dati three-way*.

I dati *three-way*, o dati a tre vie, sono dati caratterizzati da una struttura a tre modalità, ossia righe, colonne e strati. È proprio quest'ultima modalità a differenziare i dati *three-way* dai tradizionali *dataset two-way* con cui siamo soliti confrontarci, caratterizzati da vettori p-dimensionali di osservazioni. *Dataset* di questo tipo solitamente provengono dall'osservazione di una pluralità di variabili al variare del tempo o dello spazio (nel primo caso si parla di dati longitudinali, nel secondo di dati spaziali), o dall'osservazione simultanea delle due componenti, formando i cosiddetti dati spazio-temporali. Ulteriori esempi sono forniti da studi in cui sono rilevati una serie di giudizi su più oggetti o una serie di at-

tributi da un campione di individui, o dalle serie storiche multivariate. Nella tabella 2.1 sono presentate alcune molto comuni strutture di dati *three-way*.

Formalmente, sia $\mathbf{X} = \{x_{ijh} : i \in \mathbf{I}, j \in \mathbf{J}, h \in \mathbf{H}\}$ la struttura parallelepipeda di dati *three-way*, e siano $\mathbf{I}$=$\{1,\ldots,\text{n}\}$, $\mathbf{J}$=$\{1,\ldots,\text{p}\}$, $\mathbf{H}$=$\{1,\ldots,\text{r}\}$ rispettivamente l'insieme degli indici della modalità i (riga), j (colonna), h (strato). Il generico elemento $x_{ijk} \in \mathbb{R}$ rappresenta, dunque, la i-esima osservazione della j-esima variabile, in accordo con la h-esima situazione. Una formalizzazione di questo tipo può risultare di difficile interpretazione a causa della già citata presenza di una modalitá *in più* rispetto ai tradizionali *dataset* a disposizione; per ovviare a tale ostacolo, è possibile fornire un'interpretazione matriciale di $\mathbf{X}$, ossia rappresentare $\mathbf{X}$ come l'interazione di 3 differenti insiemi di matrici. Tali insiemi risultano essere:

- Un primo insieme di n matrici variabile-occasione, chiamate *matrici orizzontali*, il cui generico elemento $\mathbf{X}_{i..} \equiv \{x_{i..} : j \in \mathbf{J}, h \in \mathbf{H}\}$ $(i \in \mathbf{I})$ rappresenta la matrice risultante dall'osservazione di p variabili in r occasioni su un generico individuo i. Attraverso tale rappresentazione è possibile valutare la relazione tra le p variabili, tenendo in considerazione la loro evoluzione nelle r occasioni, per ogni fissata osservazione i.

- Un secondo insieme di p matrici unità-occasione, chiamate *matrici laterali*, il cui generico elemento $\mathbf{X}_{.j.} \equiv \{x_{.j.} : i \in \mathbf{I}, h \in \mathbf{H}\}$ $(j \in \mathbf{J})$ rappresenta la matrice risultante dall'osservazione di una generica variabile j in r occasioni su un campione di n osservazioni. Attraverso tale rappresentazione è possibile valutare per ogni fissato $j \in \mathbf{J}$ la dinamica delle variabili nelle diverse occasioni sullo stesso campione di osservazioni.

	Righe	**Colonne**	**Strati**
Dati temporali multivariati	Unità	Variabili	Tempi
Dati spaziali multivariati	Unità	Variabili	Luoghi
Dati spazio-temporali multivariati	Luoghi	Variabili	Tempi
Giudizi multi-attributo	Unità	Attributi	Giudizi
Serie storiche multivariate	Unità	Tempi	Variabili

Tabella 2.1: Tabella di alcune strutture di dati *three-way*.

- Un terzo insieme di r matrici unità-variabile, chiamate *matrici frontali*, il cui generico elemento $\mathbf{X}_{..h} \equiv \{x_{..h} : i \in \mathbf{I}, j \in \mathbf{J}\}$ $(h \in \mathbf{H})$ rappresenta la matrice risultante dall'osservazione di p variabili su un campione di n osservazioni data una generica occasione h. Attraverso tale rappresentazione è possibile valutare per ogni fissato $h \in \mathbf{H}$ la relazione tra le variabili rilevate sullo stesso campione di osservazioni.

Ulteriori approfondimenti legati alle rappresentazioni matriciali di $\mathbf{X}$ sono presenti in Rizzi e Vichi (1995).

Nell'ottica di estrapolare dai dati informazione utile per generare conoscenza, obiettivo principale dell'analisi statistica, viene assunto il singolo *dato* come una *variabile casuale*, ossia come esito numerico di un esperimento aleatorio rappresentato da una ben precisa distribuzione probabilistica. Si può facilmente comprendere, dunque, come la scelta di un'opportuna distribuzione di probabilità f sia indissolubilmente legata al buon esito della ricerca statistica, qualunque sia il formato dei dati in oggetto. In presenza di dati *three-way*, oggetto di studio del presente lavoro, la distribuzione di probabilità f è chiamata a modellare non singoli vettori, come accade nell'analisi multidimensionale classica, bensì matrici: in tal caso si parla di *distribuzioni matrici-variate*.

Con un abuso di notazione, poniamo $X = X_{i..} = (X_{hj})$, i=1,. . . ,n; h=1,. . . ,r; j=1,. . . ,p una generica *matrice orizzontale* appartenente allo spazio $\mathbb{R}^{(r\times p)}$, e assumiamo che $f : \mathbb{R}^{(r\times p)} \longrightarrow \mathbb{R}_0^+$ sia la sua distribuzione di probabilità o densità:

$$X \sim f(\theta) \qquad\qquad i = 1, \ldots, n$$

con θ insieme dei parametri caratterizzante la distribuzione di probabilità o densità f. Distribuzioni di questo tipo sono caratterizzate da un numero di parametri generalmente superiore rispetto alle distribuzioni multivariate prima citate, fenomeno dovuto alla maggiore complessità della struttura dati che si è chiamati a gestire.

Una volta definito il modello probabilistico da cui i dati provengono, può risultare utile definire il *momento primo* e il *momento secondo centrale*, ossia il valore atteso e la varianza della matrice casuale.

Il valore atteso è facilmente definibile come:

$$E[X] = E[X_{hj}] \tag{2.1}$$

ossia come una matrice in cui ogni elemento rappresenta il valore atteso dell'elemento ugualmente posizionato all'interno della matrice dei

dati. Se la rappresentazione del momento primo risulta alquanto agevole, più spinosa risulta quella della varianza: difatti esprimere tutte le possibili covarianze con dati di questo tipo comporterebbe l'utilizzo di una struttura dati *a quattro vie*, decisamente poco agevole. Pertanto, al fine di rendere facilmente fruibile l'informazione contenuta nel momento secondo centrale, si applica alla matrice dei dati una particolare trasformazione lineare, ossia la *vettorizzazione*:

$$vec : \mathbb{R}^{(r \times p)} \longrightarrow \mathbb{R}^{rp} \tag{2.2}$$

Tale funzione permette di trasformare una matrice, appartenente ad un generico spazio $\mathbb{R}^{(r \times p)}$, in un vettore, appartenente ad un generico spazio $\mathbb{R}^{rp}$, formato dalla successione dei vettori colonna caratterizzanti la matrice. Grazie a tale rappresentazione, il momento secondo centrale dei dati può essere rappresentato in maniera analoga a quanto noto per il caso multivariato, ossia come una matrice di varianza-covarianza di dimensione $rp \times rp$ definita come:

$$Var[X] = E[vec(X - E[X])vec^T(X - E[X])] \tag{2.3}$$

In maniera analoga è possibile definire il momento terzo centrale, legato all'asimmetria, e il momento quarto centrale, legato alla curtosi della variabile casuale.

Per ulteriori proprietà riguardanti le distribuzioni matrici-variate si rimanda a De Waal (1985).

2.2 La distribuzione matriciale normale

La *distribuzione matriciale normale* rappresenta la naturale estensione della distribuzione normale applicata a variabili casuali aventi la struttura definita nel paragrafo precedente. All'interno del composito universo delle distribuzioni matrici-variate, quella normale risulta essere la più utilizzata e diffusa per i numerosi pregi di cui, similmente al caso univariato e multivariato, essa gode. Difatti, l'elevata trattabilità dal punto di vista matematico, il ruolo fondamentale che essa riveste e le numerose ed utili proprietà che la contraddistinguono rappresentano aspetti che hanno contribuito notevolmente al suo sviluppo anche in riferimento alla modellazione di dati *three-way*. Inoltre, anche in tale ambito, il teorema del limite centrale assicura a tale distribuzione il ruolo di modello di riferimento per un elevato numero di fenomeni *non normali*.

2.2.1 La parametrizzazione

La distribuzione matriciale normale rappresenta una particolare famiglia parametrica: in accordo con la notazione utilizzata nel paragrafo 2.1, sia $X \in \mathbb{R}^{(r \times p)}$ tale che:

$$X \sim \phi^{(r \times p)}(\theta) \tag{2.4}$$

con $\phi^{(r \times p)}(\cdot)$ funzione di densità caratterizzante la distribuzione matriciale normale $r \times p$-variata e θ un vettore di parametri che caratterizzano la distribuzione. Per rendere facilmente interpretabile la funzione appena introdotta, la più diffusa e funzionale parametrizzazione suddivide il vettore dei parametri caratteristici in tre insiemi disgiunti, $\theta = (\mathbf{M}, \Phi, \Omega)$, dove:

- $\mathbf{M} = \{\mu_{hj} : h = 1, \ldots, r \;\; j = 1, \ldots, p\}$ ha dimensione $r \times p$ e il generico elemento μ_{hj} rappresenta il valore atteso della j-esima variabile in accordo con la h-esima situazione. Non deve sorprendere che la dimensione di $\mathbf{M}$ risulti uguale alla dimensione della variabile casuale X: difatti, ad ogni elemento di X viene associato, in $\mathbf{M}$, il corrispondente valore atteso. In altre parole,

$$\mathbf{M} = E[X]$$

- $\Phi = \{\phi_{hl} : h,l = 1, \ldots, r\}$ ha dimensione $r \times r$ e rappresenta la matrice di varianza-covarianza tra le r occasioni. È doveroso far notare fin da subito che la struttura dei dati *three-way* richiederebbe la specificazione di p differenti matrici così definite, in quanto ragionevolmente la matrice di correlazione varia *al variare delle variabili*. L'utilizzo di una sola matrice Φ è reso possibile da una specifica espressione della forma analitica della funzione di densità $\phi^{(r \times p)}(\cdot)$, come sarà illustrato in seguito.

- $\Omega = \{\omega_{jf} : j,f = 1, \ldots, p\}$ ha dimensione $p \times p$ e rappresenta la matrice di varianza-covarianza tra le p variabili. In accordo con il caso precedente, la creazione di un modello che tenga in considerazione il variare della matrice Ω *al variare delle r occasioni* risulta un elemento imprescindibile per la gestione di tale tipologia di dati. L'inconveniente in essere sarà risolto contestualmente al precedente.

Appare chiaro fin da subito come sotto tale parametrizzazione la distribuzione matriciale normale sia un'*estensione* del caso multivariato, con l'aggiunta di una matrice di varianza-covarianza ulteriore (e

la conseguente modifica della dimensione del valore atteso $\mathbf{M}$) dovuta all'introduzione della terza dimensione dei dati.

A fronte di una spiccata chiarezza concettuale, tale modello comporta l'ovvio aumento dei parametri da stimare, inconveniente dovuto proprio alla presenza di un insieme di parametri in più. Nel caso multivariato, in presenza del valore atteso $\mu_{p\times 1}$ ed una matrice di varianza-covarianza simmetrica $\Sigma_{p\times p}$, i parametri da stimare risultano:

$$s_1 = p + \frac{p(p+1)}{2}$$

Nel caso matriciale, in presenza del valore atteso $\mathbf{M}_{r\times p}$ e di due matrici di varianza-covarianza simmetriche $\Phi_{r\times r}$ e $\Omega_{p\times p}$, essi risultano:

$$s_2 = rp + \frac{r(r+1)}{2} + \frac{p(p+1)}{2} \tag{2.5}$$

causando un aumento dei parametri:

$$s = s_2 - s_1 = p(r-1) + \frac{r(r+1)}{2} \tag{2.6}$$

2.2.2 La funzione di densità

La *funzione di densità per una variabile casuale matriciale normale* X *di media* $\boldsymbol{M}$ *e matrici di varianza-covarianza* Φ *e* Ω è definita come:

$$\begin{aligned} f(x|\mathbf{M},\Phi,\Omega) =& (2\pi)^{-\frac{rp}{2}}|\Phi|^{-\frac{p}{2}}|\Omega|^{-\frac{r}{2}} \\ & \times exp\left\{-\frac{1}{2}tr\Phi^{-1}(x-\mathbf{M})\Omega^{-1}(x-\mathbf{M})^T\right\} \end{aligned} \tag{2.7}$$

La (2.7) appartiene alla *famiglia di distribuzioni ellittiche matrici-variate*, generalizzazione dell'omonima famiglia di distribuzioni multivariate: nello specifico, essa appartiene alla classe di distribuzioni *sferiche a sinistra* se $\Phi = I_r$, alla classe di distribuzioni *sferiche a destra* se $\Omega = I_p$, alla classe di distribuzioni *sferiche* se contemporaneamente $\Phi = I_r$ e $\Omega = I_p$.

A fronte di una formulazione certamente non delle più semplici, riveste un ruolo chiave l'interpretazione della distribuzione matriciale normale come *caso specifico* della ben più nota distribuzione normale multivariata realizzato tramite il processo di *vettorizzazione* presentato in precedenza (si veda l'espressione (2.2)). Tale trasformazione lineare permette difatti la creazione di un vettore $vec(X)$ di dimensione $pr \times 1$ contraddistinto dalla successione dei vettori colonna della variabile X: la

distribuzione normale multivariata risulta, dunque, quella appropriata per la modellazione della variabile aleatoria in questione, caratterizzata da pr elementi normali eventualmente correlati tra loro.

Si definisca pertanto:

$$vec(X) \sim N^{(pr)}(\mu, \Sigma)$$

con $\mu_{pr\times 1}$ valore atteso e $\Sigma_{pr\times pr}$ matrice di varianza-covarianza. Se il passaggio del valore atteso da $\mathbf{M}$ a μ deriva dall'intuitiva *vettorizzazione* della matrice $\mathbf{M}$ nel vettore μ ($E[vec(X)] = vec(\mathbf{M}) = \mu$), lo stesso non si può di certo dire per il momento secondo centrale, ancora una volta oggetto di difficoltà non intuitivamente superabili. È difatti necessario condensare l'informazione contenuta nelle due matrici Φ e Ω in una sola matrice Σ, chiamata ad esprimere tutte le possibili associazioni tra le p variabili e le r occasioni fornendo così una soluzione al problema presentato nel paragrafo 2.2.1 .

Al fine di determinare la relazione esistente tra le tre matrici in gioco, si introduce un'operazione matriciale nota come *prodotto di Kronecker*. A differenza del tradizionale prodotto *righe per colonne*, il prodotto di Kronecker può essere calcolato su matrici di qualsiasi dimensione, risultando quindi particolarmente funzionale al nostro caso in cui le due matrici Φ e Ω sono generalmente di dimensioni differenti. Tale operatore, definito come:

$$\otimes : \mathbb{R}^{(p\times q)} \times \mathbb{R}^{(s\times t)} \longrightarrow \mathbb{R}^{(ps\times qt)} \tag{2.8}$$

comporta la moltiplicazione di ogni singolo elemento della prima matrice per *l'intera seconda matrice* (e chiaramente viceversa), ossia nel caso specifico delle due matrici caratterizzanti la distribuzione normale matriciale Φ e Ω:

$$\Phi \otimes \Omega = \begin{bmatrix} \phi_{11}\omega_{11} & \dots & \phi_{11}\omega_{1t} & \dots & \phi_{1q}\omega_{11} & \dots & \phi_{1q}\omega_{1t} \\ \phi_{11}\omega_{21} & \dots & \phi_{11}\omega_{2t} & \dots & \phi_{1q}\omega_{21} & \dots & \phi_{1q}\omega_{2t} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ \phi_{11}\omega_{s1} & \dots & \phi_{11}\omega_{st} & \dots & \phi_{1q}\omega_{s1} & \dots & \phi_{1q}\omega_{st} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ \phi_{p1}\omega_{11} & \dots & \phi_{p1}\omega_{1t} & \dots & \phi_{pq}\omega_{11} & \dots & \phi_{pq}\omega_{1t} \\ \phi_{p1}\omega_{21} & \dots & \phi_{p1}\omega_{2t} & \dots & \phi_{pq}\omega_{21} & \dots & \phi_{pq}\omega_{2t} \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ \phi_{p1}\omega_{s1} & \dots & \phi_{p1}\omega_{st} & \dots & \phi_{pq}\omega_{s1} & \dots & \phi_{pq}\omega_{st} \end{bmatrix}.$$

Attraverso tale rappresentazione quindi vengono espresse tutte le possibili *covarianze incrociate* tra le pr variabili, racchiudendo all'interno di

una sola matrice l'informazione contenuta nelle due distinte matrici Φ e Ω. La relazione cercata risulta pertanto:

$$Var[vec(X)] = \Sigma = \Phi \otimes \Omega \tag{2.9}$$

che come accennato in precedenza permette l'utilizzo di *una sola matrice* Φ e *una sola matrice* Ω senza che ciò comporti l'ipotesi di uguale struttura di covarianza tra le p variabili al variare delle r occasioni, e viceversa.

Grazie al risultato ottenuto è possibile ora notare l'equivalenza tra la (2.7) e la distribuzione normale multivariata della variabile casuale $vec(X)$:

$$\begin{aligned} f(vec(x)|\mu,\Sigma) &= (2\pi)^{-\frac{rp}{2}}|\Sigma|^{-\frac{1}{2}} \\ &\quad \times exp\left\{-\frac{1}{2}(vec(x)-\mu)^T\Sigma^{-1}(vec(x)-\mu)\right\} \\ &= (2\pi)^{-\frac{rp}{2}}|\Phi|^{\frac{-p}{2}}|\Omega|^{-\frac{r}{2}} \\ &\quad \times exp\left\{-\frac{1}{2}(vec(x)-\mu)^T(\Phi\otimes\Omega)^{-1}(vec(x)-\mu)\right\} \end{aligned}$$

resa possibile da tre importanti proprietà riguardanti il prodotto di Kronecker (Gupta e Nagar, 1999):

$$det(A \otimes B) = det(A)^p det(B)^r \tag{2.10}$$

$$(C \otimes D)^{-1} = C^{-1} \otimes D^{-1} \tag{2.11}$$

$$tr(EF^TGFH) = (vec(F))^T(E^TH^T \otimes G)vec(F) \tag{2.12}$$

con A matrice quadrata $r \times r$; B matrice quadrata $p \times p$; C e D matrici *non singolari*, ed E, F, G, H matrici di dimensioni tali per cui esistono tutti i prodotti righe per colonne definiti in 2.12.

Per ulteriori nozioni riguardanti la distribuzione matriciale normale si rimanda a De Waal (1985) e Gupta e Varga (1992); per approfondimenti riguardanti la famiglia di distribuzioni ellittiche nuovamente a Gupta e Nagar (1999).

2.3 Il modello mistura di matrici normali

In molte applicazioni pratiche, l'ipotesi di *identica distribuzione* delle osservazioni di cui si dispone risulta essere un costrutto artificioso poco aderente alla realtà: è frequente notare, in tal caso, *eterogeneità* nei

dati a causa della presenza di più modelli probabilistici condensati in un'unica distribuzione. In tale contesto diviene dunque fondamentale l'individuazione di un metodo capace di gestire la struttura non omogenea appena presentata: una possibile soluzione è fornita dalle *misture finite di distribuzioni*.

A partire dalla loro introduzione (si veda per esempio Basford e McLachlan, 1985), tali modelli hanno suscitato sempre maggiore interesse grazie all'estrema flessibilità ed utilità, fornendo di fatto un approccio matematico all'analisi di un gran numero di fenomeni aleatori. Le applicazioni possibili sono numerose e legate principalmente alla presenza di sottopopolazioni omogenee all'interno di una stessa popolazione, rivestendo per questo motivo un ruolo di primaria importanza nella *cluster analysis* che verrà ampiamente trattata nel capitolo 3. Per il momento, sarà presentato un punto di vista teorico contestualmente ad alcune proprietà.

Sia $X = \{X_1, \ldots, X_n\}$ un campione di osservazioni con $X_i = (X_{hj}) \in \mathbb{R}^{r \times p}$ per $i = 1, \ldots, n; h = 1, \ldots, r; j = 1, \ldots, p$. Il modello a mistura finita ipotizza che le n osservazioni provengano da k sottopopolazioni, ciascuna caratterizzata dunque da una propria distribuzione f_l e da propri parametri θ_l (in linea di massima diversi per ogni sottopopolazione) per $l = 1, \ldots, k$. L'utilizzo di tale modello dunque mira a racchiudere all'interno di una *sola distribuzione probabilistica* le diverse distribuzioni da cui i dati provengono, attraverso una combinazione lineare convessa delle singole funzioni di densità:

$$\sum_{l=1}^{k} \pi_l f_l(x, \theta_l)$$

Come parziale semplificazione, introduciamo l'ipotesi di *appartenenza alla medesima famiglia di distribuzioni* da parte delle matrici casuali $r \times p$-variate, ossia:

$$\sum_{l=1}^{k} \pi_l f(x, \theta_l) \tag{2.13}$$

All'interno di questo composito scenario, riveste un ruolo rilevante quello in cui f appartiene alla famiglia normale per la sua già citata flessibilità.

Tale scelta, oltre che da numerose prove empiriche, è supportata anche da un importante risultato teorico presentato in McLachlan e Peel (2004):

...any continuous distribution can be approximated arbitrarily well by a finite mixture of normal densities with common variance (or covariance matrix in the multivariate case).

È facile dunque intuire come l'utilizzo di un modello basato su una mistura di distribuzioni normali sia completamente giustificato e rappresenti, nella realtà, *la norma più che l'eccezione.*

Siano $\phi_l^{(r\times p)}(\cdot; M_l, \Phi_l, \Omega_l)$ con $l = 1, \ldots, k$ le funzioni di densità delle k sottopopolazioni in esame: la distribuzione della generica matrice casuale osservata risulta essere:

$$f(x_i|\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k) = \sum_{l=1}^{k} \pi_l \phi_l^{(r\times p)}(x_i; M_l, \Phi_l, \Omega_l) \tag{2.14}$$

sotto i vincoli:

$$0 \leq \pi_l \leq 1 \qquad\qquad l = 1, \ldots, k \tag{2.15}$$

$$\sum_{l=1}^{k} \pi_l = 1 \tag{2.16}$$

I *pesi* π_l rappresentano delle *probabilità a priori* che un'osservazione appartenga al gruppo l, ossia alla l-esima sottopopolazione del modello mistura.

Si dimostra inoltre che il valore atteso della *variabile mistura* X_i è definito come:

$$E[X_i] = \sum_{l=1}^{k} \pi_l M_l \tag{2.17}$$

e la varianza, riprendendo la notazione $vec(X_i)$ per la già menzionata esigenza concettuale, risulterà:

$$\begin{aligned} Var[vec(X_i)] = &\sum_{l=1}^{k} \pi_l vec(M_l) vec(M_l)^T + \sum_{l=1}^{k} \pi_l (\Phi_l \otimes \Omega_l) \\ &- \left(\sum_{l=1}^{k} \pi_l vec(M_l)\right)\left(\sum_{l=1}^{k} \pi_l vec(M_l)\right)^T \end{aligned} \tag{2.18}$$

Per la dimostrazione si rimanda a Viroli (2011).

Capitolo 3

Clustering parametrico per dati *three-way*

3.1 Introduzione alla *cluster analysis*

All'interno del complesso insieme delle attività umane, la classificazione di un insieme di elementi in differenti categorie rappresenta una pratica molto diffusa ed estremamente utile per i più disparati scopi. Si definisce *cluster analysis* un eterogeneo insieme di metodi volti a suddividere i dati in una serie di gruppi attraverso il soddisfacimento di un qualche criterio guida. Nell'accezione più generica, tale proposito è perseguito senza alcuna conoscenza preliminare circa il numero di gruppi o qualsiasi altra informazione utile riguardo le *possibili* strutture presenti nei dati. In tale contesto dunque la *cluster analysis* si differenzia notevolmente dai metodi di classificazione, procedure che sfruttando il raggruppamento noto di alcune unità mirano alla categorizzazione di un altro *set* di osservazioni. Si può facilmente comprendere come tale attività sia fondamentale non solo in statistica, ma in tutta una serie di settori anche molto differenti tra loro: per esempio, nel *marketing* può risultare profittevole partizionare i clienti in base alle scelte d'acquisto così da proporre per determinate fasce d'utenza una campagna pubblicitaria *ad hoc*, oppure in biologia è di fondamentale importanza per la derivazione della tassonomia delle specie viventi. Anche lo studio del comportamento dei cosiddetti "singoletti", ossia di quegli elementi che costituiscono un gruppo a sè stante poiché assumono valori fortemente diversi dal resto della popolazione, come gli *outlier* o valori anomali, trova nella pratica numerose applicazioni: si pensi in tal senso al fenomeno della duplicazione delle carte di credito, in cui pagamenti di denaro in-

genti e concentrati in poco tempo (anomali pertanto rispetto al resto della popolazione) possono suggerire un'attività fraudolenta.

Con riferimento alla trattazione di dati *three-way*, nel corso del tempo tale disciplina è stata capace di ritagliarsi sempre maggiore rilevenza e centralità nel dibattito statistico: una struttura complessa come quella dei dati a tre vie, difatti, comporta una ricchezza informativa tale da rendere ancora più interessante, utile e ricca di significato l'analisi dei gruppi, permettendo quindi di condurre indagini sempre più accurate e precise.

L'obiettivo del presente lavoro è quello di applicare l'azione di *clustering* ai dati *three-way*: a fronte di una procedura più complessa, l'analisi di gruppo per questi dati sfrutta l'informazione portata in dote dalla modalità *in più* per formare *cluster* di individui simili non solo per quanto riguarda le caratteristiche osservate, ma anche la loro evoluzione nel tempo o nello spazio.

3.2 La formalizzazione del problema di raggruppamento

L'obiettivo è quello di individuare, se esistente, la struttura a gruppi presente nei dati e assegnare ogni unità al *cluster* di riferimento.

Sia $X = \{X_1, \ldots, X_n\}$ un campione di unità statistiche con $X_i = (X_{hj}) \in \mathbb{R}^{r \times p}$ una matrice contenente l'osservazione di p variabili rilevate in r occasioni, per $i = 1, \ldots, n; h = 1, \ldots, r; j = 1, \ldots, p$. Si assuma che tali osservazioni provengano dal modello mistura descritto nel paragrafo 2.3 in cui le citate sottopopolazioni risultano descrivere una struttura di gruppo all'interno dello spazio $\mathbb{R}^{r \times p}$: l'approccio basato su modello per dati *three-way* prevede, similmente al caso *two-way*, che i dati siano generati da una mistura di distribuzioni sottostanti di probabilità o densità in cui ogni componente rappresenta un *cluster* specifico. Come precedentemente espresso, una scelta comune per la distribuzione di ciascuna componente è quella della normale matrice-variata con k differenti sottopopolazioni, con tale parametro ignoto e, pertanto, da stimare.

Nell'accezione più generica, sia

$$f(x|\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k) = \sum_{l=1}^{k} \pi_l \phi_l^{r \times p}(x; \theta_l) \tag{3.1}$$

la funzione di densità della variabile d'interesse con $\theta_l = \{M_l, \Phi_l, \Omega_l)$ l'insieme dei parametri definito nel paragrafo 2.2.1 per l' $l - esimo$ gruppo. Si noti come sotto tale formalizzazione il modello presenti la massima flessibilità possibile: essendo la distribuzione normale completamente descritta dal proprio valore atteso e dalla struttura di covarianza, l'assenza di qualsiasi vincolo sugli insiemi M_l, Φ_l, Ω_l comporta la possibilità che i gruppi siano differenti per posizione nello spazio $\mathbb{R}^{r \times p}$ e per variabilità con riferimento ad entrambe le modalità.

Una volta definita la struttura sottostante, il processo di allocazione di un'osservazione in un determinato gruppo necessita di due elementi fondamentali: la stima dei parametri del modello e quella della probabilità di appartenenza ai diversi *cluster*. Non conoscendo a priori il numero di gruppi esistenti nei dati, appare neccessaria la stima di più modelli e, successivamente, la selezione del modello migliore sulla base di un criterio che permetta il confronto per valori differenti di k.

3.2.1 La divisione delle osservazioni in gruppi tramite l'algoritmo EM

Con riferimento al primo obiettivo, ossia la stima dei parametri, assai funzionale risulta la specificazione della *funzione di verosimiglianza*: l'ipotesi distributiva precedentemente formultata rende tale approccio di gran lunga preferibile ad altri metodi di stima, anche grazie alla possibiltà di applicare tecniche inferenziale ai risultati ottenuti.

Assumiamo di disporre di un campione di n osservazioni indipendenti $x = \{x_1, \ldots, x_n\}$ la cui distribuzione è quella presentata nell'equazione (3.1); la funzione di *log-verosimiglianza* associata risulta:

$$l(\pi, \theta; x_1, \ldots, x_n) = \sum_{i=1}^{n} log\Big(\sum_{l=1}^{k} \pi_l \phi_l^{(r \times p)}(x_i; M_l, \Phi_l, \Omega_l)\Big) \qquad (3.2)$$

con $\pi = \{\pi_1, \ldots, \pi_k\}$ e $\theta = \{\theta_1, \ldots, \theta_k\}$. A causa della complessità del modello mistura, tale funzione risulta spesso multimodale e generalmente difficile da massimizzare per via analitica: se il primo inconveniente verrà in seguito affrontato e risolto, il secondo suggerisce l'esigenza di utilizzare un criterio alternativo per giungere alle stime da noi desiderate: una soluzione possibile è data dall'algoritmo Expectation-Maximization (EM) (McLachlan *et al.*, 2004). Lo scopo di tale metodo, presentato per la prima volta nel 1977 (Dempster *et al.*, 1977), è quello di aumentare, e possibilmente massimizzare, la log-verosimiglianza in presenza di dati

incompleti: nel caso specifico della *cluster analysis*, si considera *dato completo* $y_i = (x_i, z_i)$ con x_i *dato osservato* e z_i *dato mancante*.

L'elemento $z_i = (z_{i1}, \ldots, z_{ik})$ è un vettore che rappresenta l'etichetta di gruppo della specifica osservazione mediante la codifica di k differenti variabili:

$$z_{il} = \begin{cases} 1 & \text{se } x_i \in \text{al gruppo } l \\ & \qquad\qquad\qquad\qquad l = 1, \ldots, k \\ 0 & \text{altrimenti} \end{cases}$$

I dati mancanti così definiti sono indipendenti ed identicamente distribuiti secondo una distribuzione multinomiale con k categorie di probabilità rispettivamente $\pi_1, \ldots, \pi_k$:

$$f(z_i|\pi, \theta) = \prod_{l=1}^{k} \pi_l^{z_{il}}$$

e nello specifico:

$$f(z_{il} = 1|\pi, \theta) = \pi_l \tag{3.3}$$

$$f(x_i|z_{il} = 1; \pi, \theta) = \phi_l^{(r\times p)}(x_i; M_l, \Phi_l, \Omega_l). \tag{3.4}$$

Dalla (3.3) e (3.4), posto $z = \{z_1, \ldots, z_k\}$ si ottiene la log-verosimiglianza dei dati completi:

$$l(\pi, \theta; x, z) = \sum_{i=1}^{n}\sum_{l=1}^{k} z_{il}\Big[log\Big(\pi_l \phi_l^{(r\times p)}(x_i; M_l, \Phi_l, \Omega_l)\Big)\Big] \tag{3.5}$$

L'algoritmo EM ha come obiettivo la massimizzazione del valore atteso della log-verosimiglianza dei dati completi condizionatamente ai dati osservati, procedura che necessita la specificazione della *probabilità a posteriori* τ_{il}: tale quantità rappresenta la probabilità che l' i-esima osservazione appartenga all'l-esimo gruppo *condizionatamente* al valore assunto da x_i, caratterizzandosi dunque come una stima del valore atteso di z_{il}.

La derivazione della sua espressione analitica è resa possibile dal noto teorema di Bayes; sia S uno spazio campionario partizionato dagli eventi $A_1, \ldots, A_k$ a due a due disgiunti, di probabilità positiva tali che $\bigcup_{l=1}^{k} A_l = S$. Allora:

$$P(A_l|E) = \frac{P(E|A_l)P(A_l)}{\sum_{l=1}^{k} P(E|A_l)P(A_l)}. \tag{3.6}$$

Pertanto, se A_l è l'evento "x è stato generato dalla componente l" ed E è l'evento "x=x_0", la probabilità che l'i-esima osservazione appartenga alla l-esima componente della mistura, noto il suo valore x=x_0, è:

$$\begin{aligned}\tau_{il}(x|\pi_1,\ldots,\pi_k,\theta_1,\ldots,\theta_k) &= \frac{\pi_l \phi_l^{(r\times p)}(x; M_l, \Phi_l, \Omega_l)}{\sum_{t=1}^{k} \pi_t \phi_t^{(r\times p)}(x; M_t, \Phi_t, \Omega_t)} \\ &= \frac{\pi_l \phi_{il}}{\sum_{t=1}^{k} \pi_t \phi_{it}}.\end{aligned} \tag{3.7}$$

Per inizializzare l'algoritmo, è necessario disporre di un valore di partenza per z_i $\forall i = 1,\ldots,n$: questo può essere scelto arbitrariamente o, per velocizzare la procedura successiva, provenire da un altro processo di raggruppamento.

In seguito si alternano due *step*: in primis l'M-step che, a partire dal valore noto z_i, massimizza la verosimiglianza fornendo così le stime di $M_l, \Phi_l, \Omega_l, \pi_l$ $\forall l = 1,\ldots,k$. Nello specifico, la funzione in questione risulta:

$$E[l(\pi,\theta;x,z)] = E[l(\pi,\theta;x|z) + l(\pi,\theta;z)] \tag{3.8}$$

Si osservi che i termini del secondo membro possono essere massimizzati separatamente poichè dipendenti da insiemi di parametri differenti, rendendo dunque più agevole il calcolo. Con riferimento al primo termine, dal calcolo delle derivate prime rispetto ai parametri della distribuzione consegue:

$$\hat{M}_l = \frac{\sum_{i=1}^{n} \tau_{il} x_i}{\sum_{i=1}^{n} \tau_{il}} \tag{3.9}$$

$$\hat{\Phi}_l = \frac{\sum_{i=1}^{n} \tau_{il} (x_i - \hat{M}_l) \hat{\Omega}_l^{-1} (x_i - \hat{M}_l)^T}{p \sum_{i=1}^{n} \tau_{il}} \tag{3.10}$$

$$\hat{\Omega}_l = \frac{\sum_{i=1}^{n} \tau_{il} (x_i - \hat{M}_l) \hat{\Phi}_l^{-1} (x_i - \hat{M}_l)^T}{r \sum_{i=1}^{n} \tau_{il}} \tag{3.11}$$

Dal secondo termine, invece, si possono ottenere le stime delle probabilità a priori sotto i vincoli di positività e somma unitaria (si veda l'equazione 2.15 e 2.16):

$$\hat{\pi}_l = \frac{\sum_{i=1}^{n} \tau_{il}}{n} \tag{3.12}$$

L'E-step, invece, prevede il calcolo di $Pr(z_{il} = 1|x_i)$ $\forall i = 1,\ldots,n$ $\forall l = 1,\ldots,k$ ossia, in altre parole, la stima della probabilità a posteriori

τ_{il}; questa, in accordo con la notazione in (3.7), risulta pertanto:

$$\hat{\tau_{il}}(x_i|\hat{\pi},\hat{\theta}) = \frac{\hat{\pi}_l \phi_l^{(r\times p)}(x_i;\hat{M}_l,\hat{\Phi}_l,\hat{\Omega}_l)}{\sum_{t=1}^{k}\hat{\pi}_t \phi_t^{(r\times p)}(x_i;\hat{M}_t,\hat{\Phi}_t,\hat{\Omega}_t)} \tag{3.13}$$

Le due procedure proseguono iterativamente fino al soddisfacimento di un prefissato criterio di convergenza, solitamente $|l(\pi^{(q)},\theta^{(q)};x,z) - l(\pi^{(q-1)},\theta^{(q-1)};x,z)| < \epsilon$ (con ϵ quantità sufficientemente piccola), ossia fintanto che le stime dei gruppi non risultano significativamente migliorabili.

L'allocazione delle osservazioni nei diversi gruppi, vero e proprio obiettivo della *cluster analysis*, viene condotta sulla base della stima del vettore τ_i: la classificazione dell'osservazione i-esima sarà allocata nel gruppo l_0 se:

$$l_0|\hat{\tau_{il_0}} = arg\ max_l(\hat{\tau_{il}}) \qquad l = 1,\ldots,k \tag{3.14}$$

ossia verrà assegnata al gruppo che a posteriori sembra il più probabile. Da tale formulazione è possibile ricavare $(1 - \max_l \hat{\tau_{il}})$, un indicatore dell'incertezza collegata al raggruppamento effettuato; valori elevati potrebbero suggerire un partizionamento incerto, sebbene sia doveroso tenere in considerazione l'influenza del numero di gruppi presenti sul valore in questione.

Sotto determinate condizioni (McLachlan e Krishnan, 2007) è stato dimostrato che il metodo converge ad un massimo locale della funzione di massima verosimiglianza del modello mistura. La convergenza ad un massimo *locale* e non globale può essere affrontata in due modi: considerando diversi valori di partenza, procedura che però appesantisce in maniera consistente le tempistiche del processo, oppure fornendo etichette attraverso un raggruppamento iniziale ragionevole, per esempio, *agglomerativo gerarchico*.

3.2.2 Il criterio di informazione Bayesiano per il confronto tra modelli

Rimane da risolvere, dopo aver presentato la procedura per stimare i parametri del modello ed effettuare il raggruppamento corrispondente, una questione pendente di fondamentale importanza: il numero di *cluster* presenti nei dati. Tale problema risulta essere di difficile soluzione per dati *two-way* e - a maggior ragione - per quelli *three-way*, la cui

analisi grafica è esclusa a priori per l'elevata dimensionalità delle osservazioni. La conoscenza del numero di gruppi, difatti, è difficilmente nota a priori ed è uno dei principali temi di ricerca in ambito di *cluster analysis*.

La metodologia descritta sinora rende possibile il superamento di tale problema grazie all'applicabilità di procedure inferenziali con il duplice proposito di selezionare il modello ottimale e di individuare il numero di componenti. L'idea alla base è quella di utilizzare un criterio di selezione del modello al variare del numero di gruppi.

La procedura più comunemente utilizzata si basa sul *fattore di Bayes* (Kass e Raftery, 1995): l'intuizione di fondo è di utilizzare (ancora una volta) la probabilità a posteriori condizionatamente ai dati osservati dei possibili modelli e di scegliere il più verosimile sulla base di ciò. Selezionato K come numero massimo di gruppi possibile, siano $M_1, \ldots, M_K$ i modelli da confrontare con rispettivamente $1, \ldots, K$ componenti al loro interno, e siano $P(M_j)$ con $j = 1, \ldots, K$ le probabilità a priori degli stessi; un'ipotesi frequente è $P(M_r) = P(M_s) = \frac{1}{K} \forall r, s = 1, \ldots, K$ poichè, senza alcuna informazione preventiva, tutte le possibilità risultano equiprobabili.

Dal noto teorema di Bayes, la probabilità a posteriori del modello M_j condizionata ai dati osservati (per semplicità, D) è proporzionale alla probabilità inversa, ossia alla probabilità di osservare i dati D in presenza del modello M_j, moltiplicata per la probabilità a priori $P(M_j)$:

$$P(M_j|D) \propto P(D|M_j)P(M_j) \tag{3.15}$$

Da tale espressione, e in presenza di parametri ignoti, $P(D|M_j)$ è ottenibile attraverso il calcolo dell'integrale:

$$P(D|M_j) = \int P(D|\theta_j, M_j)P(\theta_j|M_j)d\theta_j \tag{3.16}$$

con $P(\theta_j|M_j)$ distribuzione a priori del vettore parametrico θ_j.

La quantità $P(D|M_j)$ viene chiamata *verosimiglianza integrata* del modello M_j e riveste un ruolo di primaria importanza all'interno dell'approccio Bayesiano. Difatti, se le probabilità a priori sono equivalenti come ipotizzato, la selezione del modello basata sulla probabilità a posteriori si riduce ad un semplice confronto tra le verosimiglianze integrate dei differenti modelli. Nell'ottica di confrontare due tra i K possibili modelli, il fattore Bayesiano viene definito come il rapporto:

$$B_{rs} = \frac{P(D|M_r)}{P(D|M_s)} \qquad r, s \in 1, \ldots, k$$

da cui consegue la preferenza per il modello M_r se $B_{rs} > 1$, per il modello M_s se $B_{rs} < 1$. Svolgendo tutti i $\binom{K}{2}$ possibili confronti, è possibile derivare un preciso ordinamento tra gli M_j: ci limiteremo a selezionare il modello g il cui fattore Bayesiano B_{gs} risulta maggiore di $1\ \forall s = 1, \ldots, K, s \neq g$.

Tuttavia nella pratica tale metodologia non risulta direttamente applicabile a causa delle difficoltà nel calcolo dell'integrale dell'equazione 3.16; si utilizza dunque l'approssimazione nota come *Bayesian information criterion* (BIC):

$$2logP(D|M_j) \approx 2logP(D|\hat{\theta}_j, M_j) - \nu_j log(n) = BIC_j \qquad (3.17)$$

con ν_j numero di parametri indipendenti del modello M_j. Tale formulazione però neccessita di alcune condizioni di regolarità che non sono soddisfatte dai modelli a mistura finita da noi utilizzati: fortunatamente molti risultati pratici ne permettono l'applicabilità nella *cluster analysis* così condotta. Per esempio, Leroux *et al.* (1992) ha dimostrato che il confronto tra i BIC non sottostima asintoticamente il numero di componenti presenti nei dati, e sei anni dopo Keribiin (1998) che tale indicatore risulta consistente in riferimento al numero di gruppi. Nel confronto tra due modelli, una differenza in termini di BIC minore di 2 viene considerata un'evidenza debole, tra il 2 e il 6 evidenza media, tra il 6 e il 10 evidenza forte e superiore al 10 evidenza molto forte.

Esiste purtroppo un gravoso *trade-off* tra la scelta del modello e le stime ottenute: un modello semplice probabilmente comporta la specificazione di *più cluster* per cogliere appieno le strutture presenti nei dati, mentre per un modello a maggiore complessità, con le connesse problematiche computazionali elencate nel prossimo paragrafo, ne sono sufficienti meno.

Alcune varianti al BIC sono state proposte nel corso del tempo: tra queste risulta, per esempio, quella suggerita da Banfield e Raftery (1993) in cui il criterio di informazione Bayesiano viene sostituito dall' *Approximate Weight of Evidence* (AWE), che però in diversi esperimenti si è dimostrato consistentemente meno efficace del BIC, o l'*Integrated Completed Likelihood* (ICL) (Biernacki *et al.*, 2000).

3.3 Alcune considerazioni sull'approccio parametrico

La metodologia proposta presenta alcune limitazioni: innanzitutto può richiedere tempistiche assai elevate, fenomeno dovuto al gran numero di parametri da stimare e al numero di iterazioni necessarie per la convergenza al valore ottimale; come precedentemente specificato, ragionevoli valori iniziali e l'esistenza di gruppi ben separati possono ridurre, più o meno significativamente, il secondo inconveniente. Inoltre, esso non può essere utilizzato se in presenza di matrici singolari o, più generalmente, se uno o più gruppi sono caratterizzati da poche osservazioni al loro interno.

Per superare tali difficoltà, nel corso del tempo sono state presentate alcune varianti all'algoritmo EM, come il *classification* EM o CEM (Celeux e Govaert, 1992) in cui i valori τ_{il} vengono discretizzati prima dell'M-step o lo *stochastic* EM o SEM (Celeux *et al.*, 1996) in cui invece tali scalari vengono simulati invece che stimati.

Con riferimento all'elevato numero di parametri, tale approccio presenta ovvi limiti computazionali a fronte di una struttura estremamente elastica: dall'equazione (2.5) deduciamo che un aumento unitario p comporta un incremento di parametri pari a:

$$\begin{aligned} n_{\theta_1} - n_{\theta_2} &= r(p+1) + \frac{r(r+1)}{2} + \frac{(p+1)(p+2)}{2} - \\ &\quad - rp + \frac{r(r+1)}{2} + \frac{p(p+1)}{2} = \\ &= r + p + 1 \end{aligned}$$

Allo stesso risultato si giunge se si considera un aumento unitario di r. In particolare, per valori elevati di p e r la complessità computazionale tende ad aumentare velocemente creando non pochi problemi in fase di stima: per tale motivo, è ragionevole considerare la possibilità di imporre alcune restrizioni al fine di ottenere modelli più parsimoniosi. Per raggiungere tale scopo, è necessario tenere in primaria considerazione l'imposizione di vincoli che risultino appropriati con i dati in questione, in modo tale da non creare distorsioni eccessive che, di fatto, renderebbero poco efficace l'azione di *clustering*.

Con riferimento al lavoro in ambito multivariato di Banfield e Raftery (1993), una possibile soluzione è fornita dalla restrizione delle matrici di varianza e covarianza Φ_l e Ω_l attraverso la scomposizione spettrale derivante dall'omonimo teorema. Applicando tale procedura per esempio

alla matrice Φ_l si ottiene :

$$\Phi_l = \lambda_l D_l A_l D_l^T \tag{3.18}$$

con D_l matrice degli autovettori di Φ_l, A_l matrice diagonale i cui elementi risultano proporzionali agli autovalori corrispondenti e λ_l la costante di proporzionalità associata; ognuna di queste quantità governa specifiche caratteristiche geometriche dell'$l-$esima componente gaussiana, ossia rispettivamente l'orientamento, la forma ed il volume. Si giunge alle stesse conclusioni considerando Ω_l. Si osservi come l'utilizzo del teorema di scomposizione spettrale sia particolarmente funzionale anche in termini interpretativi: ad esempio, se l'autovalore maggiore risulta essere abbondantemente più grande degli altri, l'$l-$esima componente della mistura (nel caso multivariato) sarà concentrato grosso modo lungo una retta nello spazio $r-$variato, o viceversa se gli autovalori risultano pressochè costanti la componente avrà forma sferica. L'dea di fondo, in ambito *three-way*, risulta essere la medesima: far variare alcune di queste quantità tra i *cluster* e mantenerne altre invece fissate così da diminuire il numero di parametri non vincolati. Le proprietà geometriche dei gruppi, in tal caso, deriveranno dalla duplice specificazione delle matrici Φ_l e Ω_l.

Così definito, il modello proposto è in grado di gestire un gran numero di scenari differenti, tra cui i principali sono quelli a componenti omoschedastiche (EEE), a componenti omoschedastiche diagonali (EEI), a componenti eteroschedastiche (VVV), a componenti eteroschedastiche diagonali (VVI), a componenti sferiche con medesimo volume (EII) e a componenti sferiche con volume differente (VII).

Nella tabella 3.1 è riassunto il numero di parametri da stimare per ciascuna combinazione elencata (Viroli, 2011).

Meno frequente, ma talvolta molto utile, risulta l'imposizione di vincoli sull'altro insieme, ossia $\mathbf{M}_l$: ad esempio, se le occasioni risultano essere misurazioni ripetute per una determinata variabile, è sensato assumere l'uguaglianza in media dei dati al variare delle occasioni, portando così il numero di parametri della matrice delle medie da rp a p.

$\Phi_l \backslash \Omega_l$	**EEE**	**EEI**	**VVV**
EEE	$\alpha + \frac{r(r+1)}{2} + \frac{p(p+1)}{2}$	$\alpha + \frac{r(r+1)}{2} + \mathrm{p}$	$\alpha + \frac{r(r+1)}{2} + \frac{kp(p+1)}{2}$
EEI	$\alpha + \mathrm{r} + \frac{p(p+1)}{2}$	$\alpha + \mathrm{r} + \mathrm{p}$	$\alpha + \mathrm{r} + \frac{kp(p+1)}{2}$
VVV	$\alpha + \frac{kr(r+1)}{2} + \frac{p(p+1)}{2}$	$\alpha + \frac{kr(r+1)}{2} + \mathrm{p}$	$\alpha + \frac{kr(r+1)}{2} + \frac{kp(p+1)}{2}$
VVI	$\alpha + \mathrm{kr} + \frac{p(p+1)}{2}$	$\alpha + \mathrm{kr} + \mathrm{p}$	$\alpha + \mathrm{kr} + \frac{kp(p+1)}{2}$
EII	$\alpha + 1 + \frac{p(p+1)}{2}$	$\alpha + 1 + \mathrm{p}$	$\alpha + 1 + \frac{kp(p+1)}{2}$
VII	$\alpha + \mathrm{k} + \frac{p(p+1)}{2}$	$\alpha + \mathrm{k} + \mathrm{p}$	$\alpha + \mathrm{k} + \frac{kp(p+1)}{2}$
$\Phi_l \backslash \Omega_l$	**VVI**	**EII**	**VII**
EEE	$\alpha + \frac{r(r+1)}{2} + \mathrm{kp}$	$\alpha + \frac{r(r+1)}{2} + 1$	$\alpha + \frac{r(r+1)}{2} + \mathrm{k}$
EEI	$\alpha + \mathrm{r} + \mathrm{kp}$	$\alpha + \mathrm{r} + 1$	$\alpha + \mathrm{r} + \mathrm{k}$
VVV	$\alpha + \frac{kr(r+1)}{2} + \mathrm{kp}$	$\alpha + \frac{kr(r+1)}{2} + 1$	$\alpha + \frac{kr(r+1)}{2} + \mathrm{k}$
VVI	$\alpha + \mathrm{kr} + \mathrm{kp}$	$\alpha + \mathrm{kr} + 1$	$\alpha + \mathrm{kr} + \mathrm{k}$
EII	$\alpha + 1 + \mathrm{kp}$	$\alpha + 1 + 1$	$\alpha + 1 + \mathrm{k}$
VII	$\alpha + \mathrm{k} + \mathrm{kp}$	$\alpha + \mathrm{k} + 1$	$\alpha + \mathrm{k} + \mathrm{k}$

Tabella 3.1: Numero di parametri per ciascuna combinazione con $\alpha = krp$

3.4 Metodi alternativi per il *clustering three-way*

Esistono chiaramente differenti approcci all'analisi di gruppo, ognuno contraddistinto da determinati *pro* e *contro*. Questi, sebbene molto diffusi per i tradizionali *dataset two-way*, sono invece meno noti e studiati in presenza di strutture *three-way*, chiaro specchio della disparità di approfondimento e attenzione rivolta alle due tipologie di dati. Per tale motivo, le principali metodologie di *clustering three-way* consistono nella compressione del *dataset* in un usuale schema a due vie e la successiva ricerca dei gruppi secondo i tradizionali metodi.

Una soluzione tanto semplice quanto intuitiva in tal senso è rappresentata dall'utilizzo delle componenti principali (si veda per esempio Mardia *et al.*, 1979). Secondo la notazione introdotta nel capitolo precedente, indicata con $\mathbf{X}_{.j.} \equiv \{x_{.j.} : i \in \mathbf{I}, h \in \mathbf{H}\}$ $(j \in \mathbf{J})$ la matrice $n \times r$ contenente l'osservazione delle r occasioni per una fissata variabile p sul campione di interesse, la prima componente principale risulta essere:

$$\mathbf{X}_{.j.}\gamma_1 \qquad\qquad j = 1, \ldots, p \tag{3.19}$$

con γ_1 autovettore corrispondente all'autovalore massimo ottenuto dalla scomposizione spettrale della matrice di varianza-covarianza del gene-

rico vettore $\mathbf{x}_{ij.}$ formante la matrice (si ricordi l'ipotesi di identica distribuzione delle osservazioni che porta all'uguaglianza della matrice di varianza-covarianza al variare di n). Tale trasformazione permette di condensare all'interno di un unico vettore buona parte dell'informazione contenuta nella $p-esima$ matrice laterale: per costruzione la prima componente principale mantiene la massima variabilità possibile per una trasformazione che, di fatto, riduce i dati dalla struttura $r-variata$ a quella univariata. Applicando tale procedimento a tutte le p matrici così definite, si ottiene un nuovo *dataset* $n \times p$ in cui in ognuna delle p variabili è riassunta l'informazione contenuta al variare delle r occasioni: si è pertanto ottenuta la prima citata riduzione della dimensionalità dei dati a disposizione attraverso la *sintesi* di una modalità in un singolo valore. Chiaramente il metodo può essere equivalentemente sviluppato con riferimento alla matrice frontale $\mathbf{X}_{..h} \equiv \{x_{..h} : i \in \mathbf{I}, j \in \mathbf{J}\}$ $(h \in \mathbf{H})$ contenente l'osservazione delle p variabili per una fissata occasione r.

I limiti di tale approccio sono evidenti: l'inevitabile perdita di variabilità che l'analisi delle componenti principali comporta rischia di intaccare, in maniera più o meno grave, la possibile struttura a gruppi presente, fuorviando di fatto la *cluster analysis*. Si veda a tal proposito il lavoro in ambito multivariato proposto da Chang (1983).

Un altro possibile *modus operandi*, che può essere considerato una sorta di generalizzazione del metodo appena presentato, si basa sulla contemporanea e iterativa riduzione dei dati e azione di *clustering*, processo mirante al soddisfacimento di un qualche criterio precedentemente imposto. Così facendo, ad ogni successiva riduzione della dimensionalità ottenuta grazie all'applicazione di una qualche trasformazione lineare segue la corrispondente divisione in gruppi delle nuove unità, finchè non si giunge ad una partizione ottimale. L'ottimalità, chiaramente, non è assoluta ma deve essere calcolata sulla base di un approccio fissato a priori, come ad esempio quello dei minimi quadrati (Vichi *et al.*, 2007).

Le procedure sinora analizzate sono comunque contraddistinte da un orientamento al *clustering* alquanto tradizionale basato su criteri largamente euristici e geometrici che non prevedono dunque alcun esplicito modello probabilistico per i gruppi: in questa nuova ottica, una prima formalizzazione è la seguente. Riprendendo il modello a mistura finita di distribuzioni utilizzato nel paragrafo 2.3, sia, con riferimento ad una determinata occasione $h = 1, \ldots, r$:

$$X_{ih} \sim \phi_l^p(\mu_{lh}, \Omega_l) \qquad l = 1, \ldots, k$$

il vettore p-dimensionale relativo all'osservazione i, la cui appartenenza

al gruppo l risulta nota. Si noti come, per costruzione, venga imposta la medesima struttura di covarianza al variare delle r occasioni,ossia:

$$\Omega_{lh} = \Omega_l \qquad \forall h = 1, \ldots, r \tag{3.20}$$

Imponendo l'ulteriore ipotesi:

$$X_{is} \perp X_{it} \qquad \forall s, t = 1, \ldots, r \tag{3.21}$$

il modello mistura assume la notazione:

$$f(x_i) = \sum_{l=1}^{k} \pi_l \prod_{h=1}^{r} \phi_l^p(x_{ih}; \mu_{lh}, \Omega_l)$$

Il passaggio dalla struttura a tre vie a quella a due vie e la contemporanea specificazione di un modello statistico sottostante permettono di condurre un'analisi contraddistinta da un maggiore rigore formale rispetto ai metodi basati su distanza: difatti la presenza di criteri per la scelta del modello ottimale, e di conseguenza del numero di gruppi, insieme a quella di principi statistici alla base (con la conseguente possibilità di effetturare procedure inferenziali) rendono per molti versi l'approccio *model-based* (Fraley e Raftery, 1998, 2002) preferibile ed applicabile nei più disparati contesti.

I limiti di tale formalizzazione risultano però essere tanto più gravi quanto maggiore è l'allontanamento dalle ipotesi introdotte. Nello specifico, l'ipotesi 3.20 di identica struttura di covarianza al variare delle occasioni rappresenta a ben vedere una forzatura rilevante, così come l'ipotesi 3.21 che ignora completamente la possibile dipendenza delle variabili al variare delle occasioni. I dati *three-way* difatti prevedono un'ovvia dinamica *al variare di r* (si pensi, per esempio, ai citati dati temporali multivariati) che non può essere assolutamente tralasciata: non considerarla significherebbe di fatto ridurre i dati così definiti ad un insieme di $n \times r$ osservazioni $p-$variate.

Capitolo 4

Studio di simulazione

4.1 Obiettivi

Con il presente studio di simulazione si vuole valutare l'efficacia dell'approccio di *clustering* basato su misture di matrici normali per la suddivisione in gruppi di dati *three-way*: si noti pertanto come l'ipotesi di normalità sia un presupposto alla base di tale metodologia che, per forza di cose, andrà a condizionarne i risultati. I possibili scenari esplorabili sono chiaramente infiniti, in questo paragrafo ne saranno analizzati alcuni nel tentativo di comprendere le dinamiche alla base del *clustering* così condotto.

I principali obiettivi perseguiti sono:

- Valutare la bontà del raggruppamento proposto al variare della distribuzione probabilistica delle osservazioni. In particolare sono stati considerati dati con distribuzione normale matriciale, dunque in linea con le ipotesi del modello, e dati che violano tale assunto: in tal caso la scelta è ricaduta su distribuzioni asimmetriche.

- Valutare il numero di gruppi trovati dall'algoritmo di *clustering* con riferimento ai casi esplorati nel punto precedente.

- Valutare la bontà del raggruppamento proposto per gruppi differenti in media ma con medesima variabilità, e gruppi a differente variabilità nelle occasioni ma identica media.

- Valutare il numero di gruppi trovati dall'algoritmo di *clustering* con riferimento ai casi esplorati nel punto precedente.

- Valutare come variano i due indicatori di nostro interesse (bontà di raggruppamento e numero di gruppi) al variare della numerosità campionaria.

- Valutare come variano i due indicatori di nostro interesse (bontà di raggruppamento e numero di gruppi) al variare della dimensionalità delle matrici frontali, ossia al variare di p.

- Valutare come variano i due indicatori di nostro interesse (bontà di raggruppamento e numero di gruppi) al variare della dimensionalità delle matrici laterali, ossia al variare di r.

4.2 Cosa, come, perchè: presentazione degli scenari

Nell'ottica di presentare una situazione verosimile, è stato scelto di simulare un fenomeno ampiamente studiato in presenza di osservazioni *three-way*, ossia quello delle *serie storiche multivariate* (si veda la Tabella 2.1): oltre ad essere utilizzate assai frequentemente, queste sono caratterizzate anche da una struttura ben precisa, altamente funzionale e relativamente semplice in rapporto ai nostri scopi.

Il progetto di simulazione rientra nello schema generale che segue: siano stati generati campioni X di n osservazioni, p variabili, r occasioni e composti da k gruppi. Tralasciando l'appartenenza ad uno specifico gruppo e di conseguenza il pedice l, la generica osservazione $x_{r\times p} = (x_{.1}, \ldots, x_{.p}) = (x_{1.}^T, \ldots, x_{r.}^T)^T$ è stata generata da una matrice normale $\phi^{(r\times p)}(\cdot; M, \Phi, \Omega)$. Nello specifico:

- $x_{t.} \sim N^p(\mu_t, \Omega) \forall t = 1, \ldots, r$ con Ω matrice diagonale a componenti eteroschedastiche. Le variabili pertanto risultano indipendenti ma a variabilità differente.

- $x_{.h} \forall h = 1, \ldots, p$ il vettore composto dall'evoluzione temporale della data variabile h. E' stato scelto di ipotizzare un processo autoregressivo di ordine 1 $AR(1)$ senza costante con origine nell'osservazione univariata x_{1h}; a partire pertanto dallo schema generale $X_{th} = \eta_1 X_{(t-1)h} + \epsilon_t$ con ϵ_t processo white noise a media nulla e varianza σ_ϵ^2 e per $t = 2, \ldots, r$, risulta (Di Fonzo e Lisi, 2005):

$$E[X_{th}] = \eta_1 E[X_{(t-1)h}] \tag{4.1}$$

con$E[X_{1h}] = [\mu_1]_h$, e:

$$\Phi = Var[X_{th}] = \begin{cases} \frac{\sigma_\epsilon^2}{1-\eta_1^2} & \text{se } t = 1 \\ \eta_1^t Var[X_{1h}] & \text{se } t = 2, \ldots, r \end{cases} \tag{4.2}$$

Il vettore $x_{.h}$ sarà quindi distribuito normalmente con media (4.1) e varianza (4.2).

In tutte le simulazioni proposte, i dati provengono da 2 distinti gruppi: la scelta è ricaduta su tale valore (minimo per un'azione di *clustering* significativo) per questioni di semplicità, in quanto la complessità del modello aumenta vertiginosamente all'aumentare del numero di gruppi (si pensi, per esempio, all'incremento dei parametri da stimare per $k_1 = k_0 + 1$). Per ogni campione analizzato, le due componenti risultano equipresenti nei dati, ossia $n_1 = n_2 = \frac{n}{2}$ con n_1, n_2 numerosità del primo e secondo gruppo rispettivamente.

Lo studio di simulazione è stato condotto come segue: sono stati analizzati due principali scenari in cui i gruppi differivano per il valore atteso e per la struttura di varianza-covarianza rispettivamente, e per ognuno di questi sono stati considerati gruppi provenienti da una distribuzione normale o da una distribuzione asimmetrica. Questi quattro macro-casi sono di seguito presentati nel dettaglio:

1. gruppi differenti in media provenienti da distribuzioni normali matriciali $\phi_l^{(r\times p)}(\cdot; M_l, \Phi, \Omega)$ per $l = 1, 2$ con distanza[1] $d(M_1, M_2) \simeq 6$. I gruppi pertanto presentano media differente ma stesse matrici di covarianza, con il parametro autoregressivo $\eta_1 = 0.8$.

2. gruppi differenti in media provenienti da distribuzioni asimmetriche. Le osservazioni x_1 del primo gruppo provengono da $x_1 = y^2$, con $Y \sim \phi^{(r\times p)}(\cdot; M, \Phi, \Omega)$, mentre le osservazioni x_2 del secondo gruppo

[1] Il calcolo della distanza tra le matrici delle medie citato nei primi due scenari simulativi è stato condotto come segue: vista l'ipotesi di eteroschedasticità della matrice Ω, innanzitutto è necessario *riscalare* i valori corrispondenti alle medie delle p variabili dividendole per la propria deviazione standard,ossia:

$$m_{l_{1h}} = \frac{M_{l_{1h}}}{\sqrt{[\Omega]_{hh}}} \qquad \forall h = 1, \ldots, p, l = 1, 2 \tag{4.3}$$

Ora è possibile calcolare la distanza cercata:

$$d(M_1, M_2) = \sqrt{\sum_{h=1}^{p} \sum_{t=1}^{r} |m_{1_{th}} - m_{2_{th}}|} \tag{4.4}$$

con $m_{l_{th}} = \eta_1 m_{l_{t-1,h}}$ $\forall t = 2, \ldots, r$.

da $x_2 = -y^2$. Come nel caso precedente, Φ e Ω sono le stesse per i gruppi con parametro autoregressivo $\eta_1 = 0.8$, mentre la distanza tra le medie dei due gruppi è stata impostata pari circa a 12. La distanza è stata appositamente aumentata in quanto in tale scenario le ipotesi del modello non sono soddisfatte, e dunque per valutare la bontà del partizionamento sono necessari gruppi maggiormente separati nello spazio.

3. gruppi differenti in variabilità provenienti da distribuzioni normali matriciali $\phi_l^{(r\times p)}(\cdot; M, \Phi_l, \Omega)$ per $l = 1, 2$. I gruppi presentano stessa media $M = 0_{(r\times p)}$ ma componente autoregressiva differente: nel primo gruppo $\eta_1 = 0.8$, nel secondo $\eta_1 = -0.2$. La matrice Ω è costante tra i gruppi.

4. gruppi differenti in variabilità provenienti da distribuzioni asimmetriche. Le osservazioni sono ottenute come nel macro-caso 2, con $Y \sim \phi_l^{(r\times p)}(\cdot; M, \Phi_l, \Omega)$ e componente autoregressiva differente in accordo con il macro-caso 3.

Di seguito vengono presentati alcuni grafici esplorativi per ciascuna situazione analizzata. Per ogni scenario descritto viene presentato per un fissato tempo $r = 1$ la disposizione delle osservazioni nello spazio $\mathbb{R}^2$ con riferimento alle variabili 1 e 2 e per una fissata variabile $p = 1$ la sua evoluzione nel tempo per $r = 1$ e $r = 2$. I due gruppi vengono rappresentati con simboli differenti. Per quanto chiaramente non esaustivi, tali grafici possono dare un'idea di fondo sull'andamento nello spazio $\mathbb{R}^{(r\times p)}$ dei due differenti gruppi (tanto più che le p variabili sono indipendenti e la componente autoregressiva - all'interno di un fissato gruppo e scenario - è la medesima e dunque, sebbene sia presentata una sola delle possibili combinazioni tra le p variabili o tra le r occasioni, questa può essere considerata una buona sintesi della dinamica generale).

Per ognuno degli scenari presentati sono state valutate differenti numerosità campionarie $n = 250, 1000$, diverso numero di variabili $p = 2, 6$, diverso numero di occasioni $r = 3, 5$. Il numero di campioni generati per ciascuna delle 32 combinazioni è pari a 250.

Al fine di interpretare i risultati dell'azione di *clustering* è necessario confrontare il raggruppamento ottenuto con la reale appartenenza delle osservazioni ai diversi gruppi: un indicatore comunemente utilizzato per questo scopo è l'*Adjusted Rand Index* (Hubert e Arabie, 1985), o ARI. Tale indicatore deriva dal *Rand Index*, da cui varia per il fatto di avere valore atteso nullo quando l'accordo tra le due partizioni è solo affetto

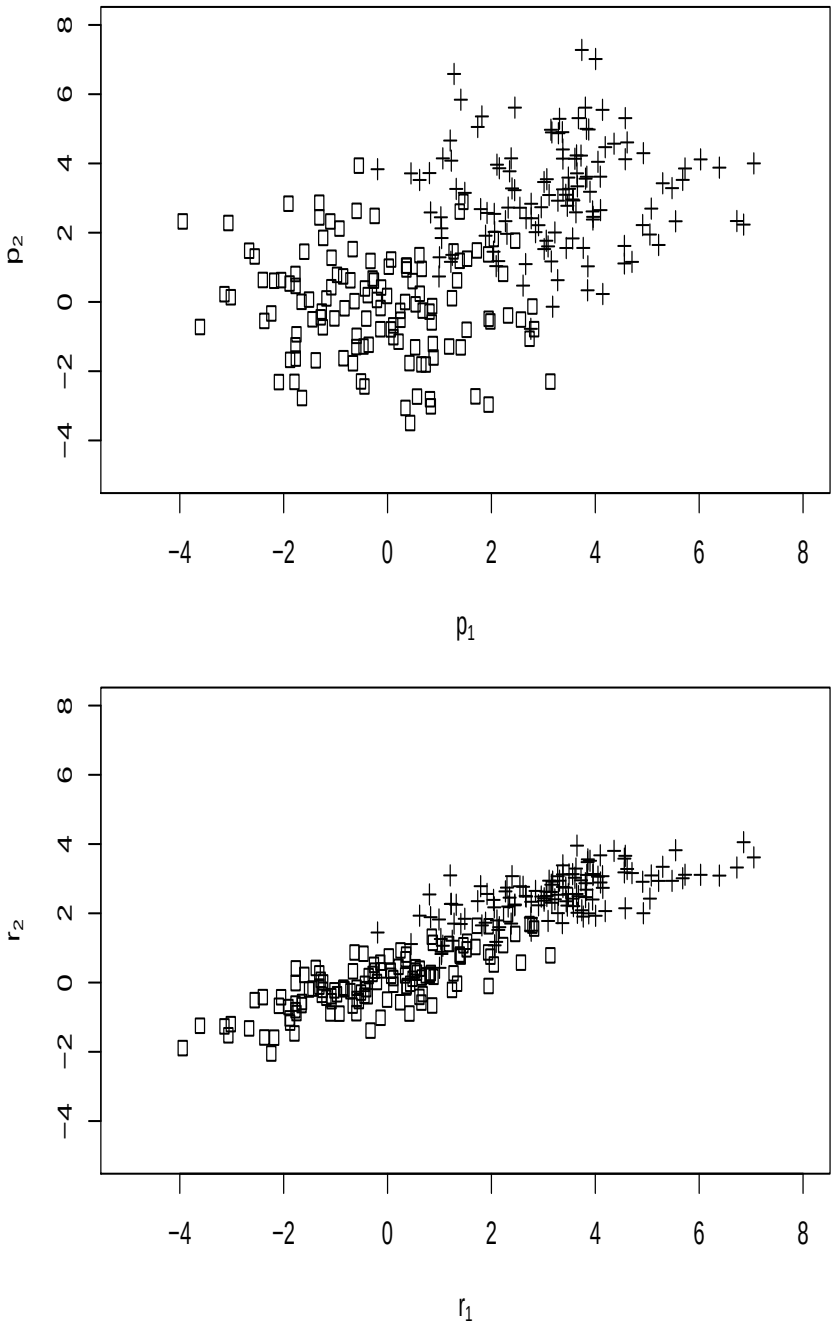

Figura 4.1: Un esempio di campione generato dal macro-scenario 1, $n = 250, p = 2, r = 3$. In alto: confronto tra la variabile 1 e la variabile 2 per un fissato tempo $r = 1$. In basso: confronto tra l'occasione 1 e l'occasione 2 per una fissata variabile $p = 1$.

dal caso. Sia $X = (x_1, \ldots, x_n)$ un insieme di n elementi da cui derivano i raggruppamenti $U = (u_1, \ldots, u_w)$ e $B = (b_1, \ldots, b_q)$ con w e q numero di gruppi rispettivo. Siano inoltre:

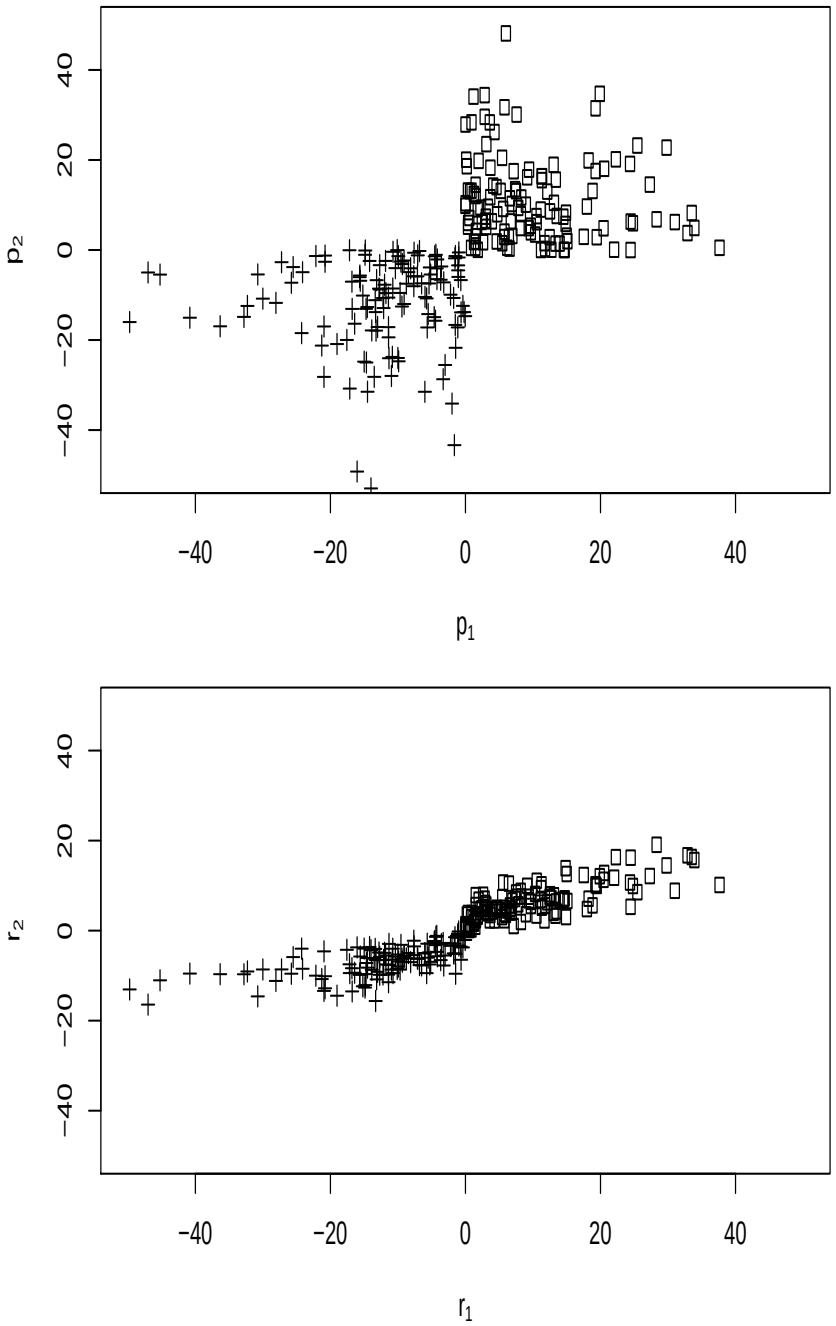

Figura 4.2: Un esempio di campione generato dal macro-scenario 2, $n = 250, p = 2, r = 3$. Cfr Figura 4.1

- c_1 il numero di coppie di elementi di X appartenenti al medesimo gruppo in U e B
- c_2 il numero di coppie di elementi di X appartenenti ad un gruppo diverso sia in U che in B

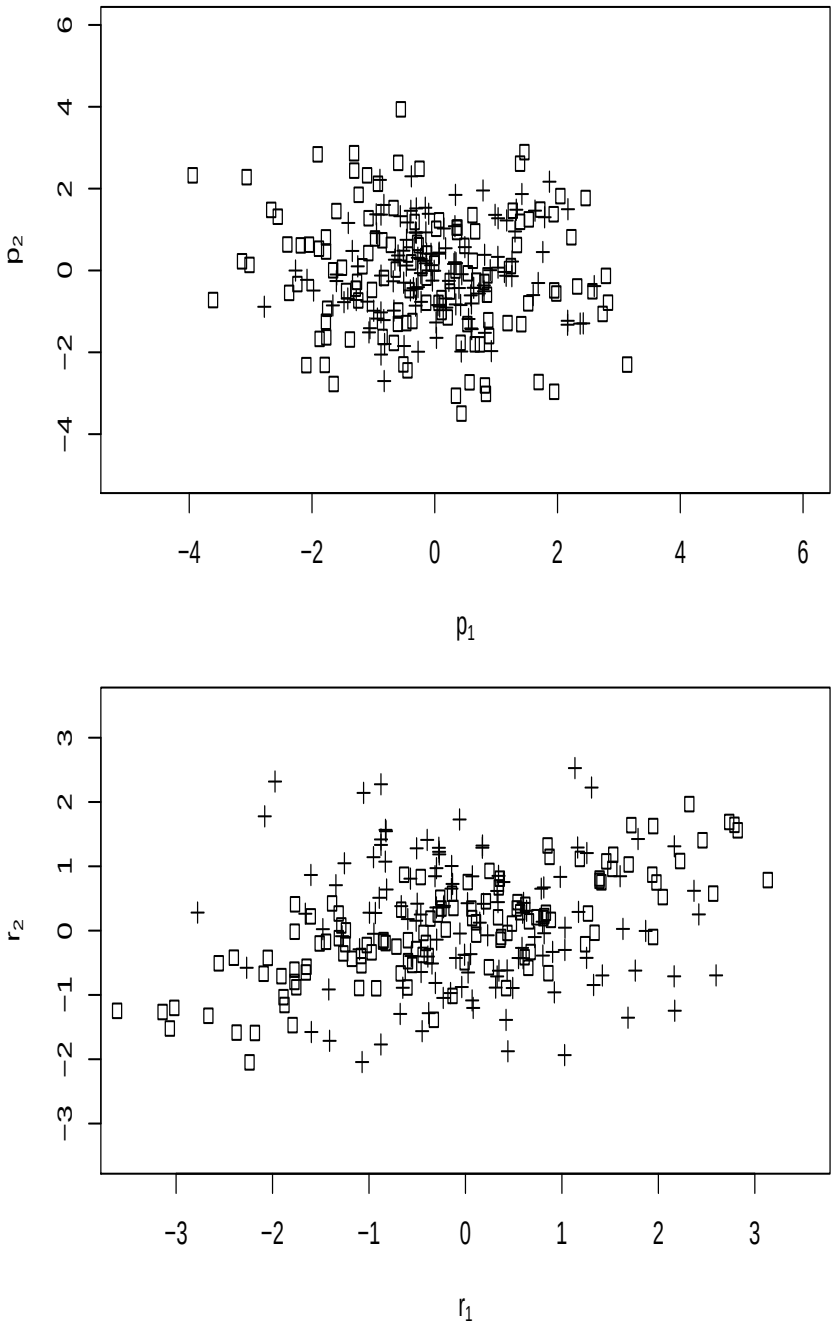

Figura 4.3: Un esempio di campione generato dal macro-scenario 3, $n = 250, p = 2, r = 3$. Cfr Figura 4.1

- d_1 il numero di coppie di elementi di X appartenenti allo stesso gruppo in U, ma gruppo diverso in B
- d_2 il numero di coppie di elementi di X appartenenti ad un gruppo differente in U, ma medesimo gruppo in B

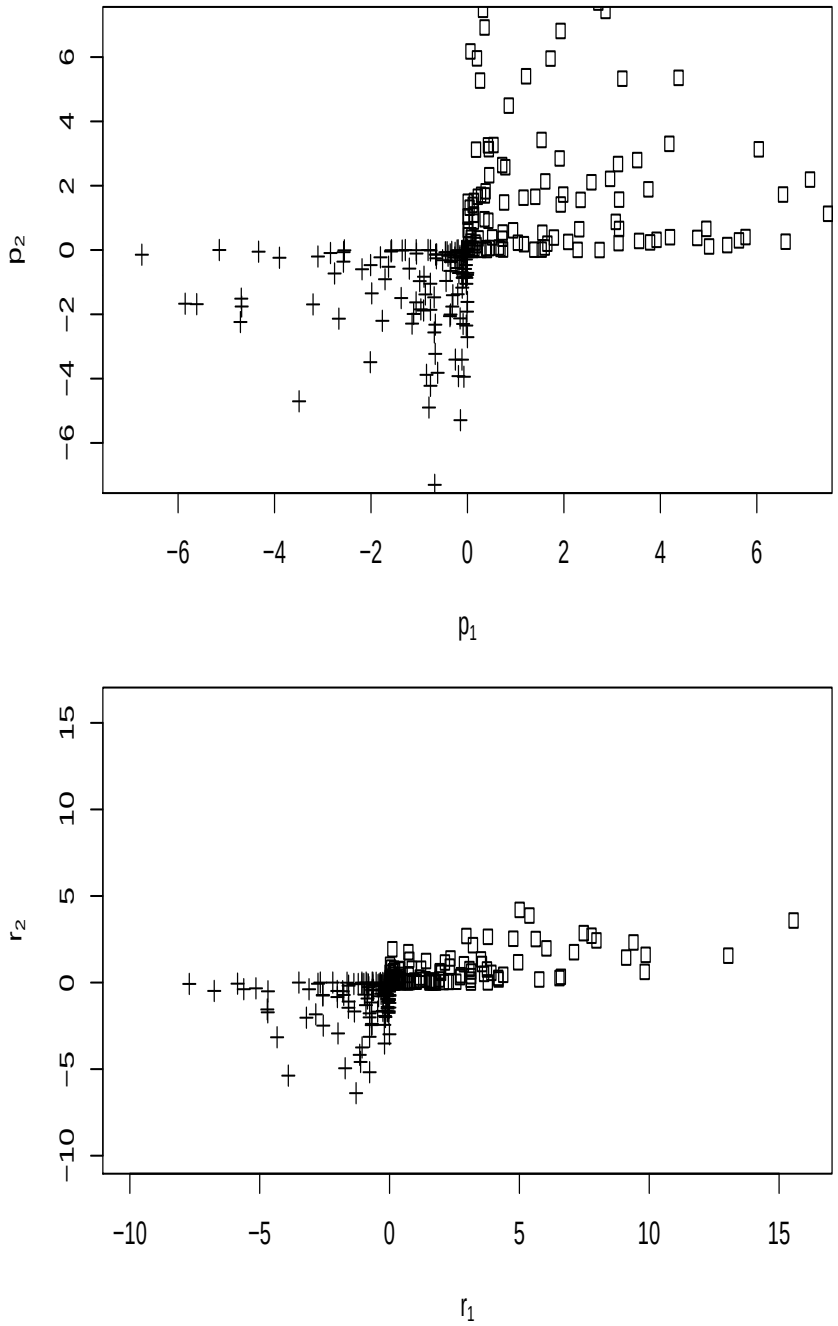

Figura 4.4: Un esempio di campione generato dal macro-scenario 4, $n = 250, p = 2, r = 3$. Cfr Figura 4.1

Allora il *Rand Index* risulta:

$$R = \frac{c_1 + c_2}{c_1 + c_2 + d_1 + d_2} \tag{4.5}$$

L'indice utilizzato nel presente lavoro, l'ARI, può assumere valori negativi e vale 1 in caso di perfetta uguaglianza tra le due partizioni.

Per quanto riguarda la parte computazionale, il lavoro è stato svolto con il linguaggio di programmazione R (R Core Team, 2013): la creazione ed il successivo raggruppamento delle osservazioni (tutte derivanti da distribuzioni matriciali normali o loro trasformazioni) sono stati realizzati con il pacchetto *mmn.em* (Viroli, 2010), mentre il calcolo dell' *Adjusted Rand Index* con il pacchetto *pdfCluster* (Azzalini e Menardi, 2014) .

4.3 Risultati: descrizione e commento

I risultati, divisi per scenario e numerosità campionaria, sono riportati nelle tabelle dalla 4.1 alla 4.8.

Per ogni esperimento sono state riportate due quantità: il valore medio dell'ARI e il numero medio di gruppi trovati dall'algoritmo al variare dei campioni generati. I due indicatori appaiono strettamente interdipendenti in quanto l'ARI è in media tanto più basso quanto più il numero di gruppi individuati si discosta da quello reale; in particolare le partizioni migliori sono state trovate in presenza di 2 gruppi stimati. È possibile quindi concludere che la simulazione da noi proposta presenta generalmente una struttura di gruppo alquanto evidente, fattore che rende dunque più semplice il processo di raggruppamento.

		$p = 2$	$p = 6$
$r = 3$	ARI	0.967 [dev.std=0.023]	0.967 [dev.std=0.023]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]
$r = 5$	ARI	0.823 [dev.std=0.166]	0.963 [dev.std=0.024]
	GRUPPI	1.964 [dev.std=0.187]	2.000 [dev.std=0.000]

Tabella 4.1: Valore medio dell'ARI e numero medio di gruppi individuati al variare delle simulazioni. Tra parentesi deviazione standard associata. I risultati si riferiscono al macroscenario 1 con n=250

		$p = 2$	$p = 6$
$r = 3$	ARI	0.953 [dev.std=0.000]	0.972 [dev.std=0.011]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]
$r = 5$	ARI	0.865 [dev.std=0.023]	0.971 [dev.std=0.011]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]

Tabella 4.2: Risultati riferiti al macroscenario 1 con n=1000. (Cfr. anche Tabella 4.1).

Una prima dinamica di fondo è riscontrabile dall'osservazione dei macroscenari caratterizzati da osservazioni normali rispetto a quelli contraddistinti da osservazioni con distribuzione asimmetrica: i primi tendono ad individuare *cluster* più aderenti alla realtà e, soprattutto, colgono in maniera nettamente migliore il numero di gruppi realmente esistente. La motivazione di tale diversità risiede nelle più volte citate ipotesi del modello, ossia l'esistenza di una mistura di componenti normali alla base della distribuzione probabilistica delle osservazioni: va da sè che l'algoritmo fatichi, con conseguenze evidenti in termini di prestazioni, a riconoscere gruppi che, per costruzione, non risultano gaussiani. In altre parole la procedura mostra scarsa robustezza rispetto alla deviazione dall'ipotesi distributiva sui gruppi.

Altro tratto contraddistintivo emerso dall'analisi è certamente la differenza nella qualità dei risultati tra i primi due macroscenari - con grup-

		$p=2$	$p=6$
$r=3$	ARI	0.872 [dev.std=0.161]	0.822 [dev.std=0.162]
	GRUPPI	2.580 [dev.std=0.752]	2.756 [dev.std=0.695]
$r=5$	ARI	0.544 [dev.std=0.161]	0.554 [dev.std=0.159]
	GRUPPI	3.516 [dev.std=0.718]	3.836 [dev.std=0.807]

Tabella 4.3: Risultati riferiti al macroscenario 2 con n=250. (Cfr. anche Tabella 4.1).

		$p=2$	$p=6$
$r=3$	ARI	0.564 [dev.std=0.120]	0.560 [dev.std=0.096]
	GRUPPI	4.536 [dev.std=0.531]	4.896 [dev.std=0.306]
$r=5$	ARI	0.413 [dev.std=0.035]	0.411 [dev.std=0.054]
	GRUPPI	4.980 [dev.std=0.140]	4.988 [dev.std=0.109]

Tabella 4.4: Risultati riferiti al macroscenario 2 con n=1000. (Cfr. anche Tabella 4.1).

		$p=2$	$p=6$
$r=3$	ARI	0.540 [dev.std=0.063]	0.918 [dev.std=0.035]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]
$r=5$	ARI	0.794 [dev.std=0.054]	0.992 [dev.std=0.011]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]

Tabella 4.5: Risultati riferiti al macroscenario 3 con n=250. (Cfr. anche Tabella 4.1).

pi di media diversa e stessa struttura di covarianza - e i secondi due - con diversa matrice di varianza Φ e medesima media -, con raggruppamenti qualitativamente migliori nel primo caso: è necessario però fare una distinzione. Se nel caso di componenti normali la differenza risulta tutto sommato contenuta grazie alla capacità da parte dell'algoritmo di gestire strutture di covarianza a maggiore complessità, essa è decisamente più evidente in presenza di osservazioni non gaussiane che di fatto rendono il *clustering* poco efficace nel raggruppare osservazioni *sparse* e asimmetriche.

Di primario interesse risulta certamente il ruolo di p ed r all'interno dei diversi scenari: nei casi con gruppi separati in media l'aumento di r sembra decisamente predominante in termini di diminuzione dell'ARI rispetto alla variazione del numero di variabili, e viceversa in presenza di gruppi separati per struttura di covarianza. Negli scenari 1 e 2, di-

		$p = 2$	$p = 6$
$r = 3$	ARI	0.557 [dev.std=0.030]	0.926 [dev.std=0.017]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]
$r = 5$	ARI	0.809 [dev.std=0.025]	0.993 [dev.std=0.004]
	GRUPPI	2.000 [dev.std=0.000]	2.000 [dev.std=0.000]

Tabella 4.6: Risultati riferiti al macroscenario 3 con n=1000. (Cfr. anche Tabella 4.1).

		$p = 2$	$p = 6$
$r = 3$	ARI	0.253 [dev.std=0.071]	0.485 [dev.std=0.094]
	GRUPPI	4.772 [dev.std=0.430]	4.480 [dev.std=0.609]
$r = 5$	ARI	0.322 [dev.std=0.071]	0.624 [dev.std=0.100]
	GRUPPI	4.504 [dev.std=0.569]	4.036 [dev.std=0.719]

Tabella 4.7: Risultati riferiti al macroscenario 4 con n=250. (Cfr. anche Tabella 4.1).

		$p = 2$	$p = 6$
$r = 3$	ARI	0.230 [dev.std=0.045]	0.426 [dev.std=0.037]
	GRUPPI	4.996 [dev.std=0.063]	4.992 [dev.std=0.089]
$r = 5$	ARI	0.305 [dev.std=0.043]	0.531 [dev.std=0.083]
	GRUPPI	5.000 [dev.std=5.000]	4.996 [dev.std=0.063]

Tabella 4.8: Risultati riferiti al macroscenario 4 con n=1000. (Cfr. anche Tabella 4.1).

fatti, la presenza per ipotesi di una dinamica temporale comune per i due gruppi porta, all'aumentare del parametro r, al peggioramento delle stime prodotte dall'algoritmo, così come negli scenari 3 e 4 la medesima media tra i due *cluster* tende a smorzare la struttura di gruppo condensando le osservazioni intorno all'origine (si ricordi che $M_1 = M_2 = 0$). L'effetto descritto risulta molto più marcato in presenza di osservazioni provenienti dalla distribuzione non normale.

Inoltre, il valore medio dell'ARI in presenza di gruppi separati in media tende a peggiorare all'aumentare della dimensionalità, in maniera limitata nel caso normale (addirittura migliora per $n = 1000$) e decisamente più evidente nel caso quadratico. L'insoddisfacente partizionamento proposto, in quest'ultimo caso, risulta causato dall'elevato numero di gruppi che l'algoritmo coglie nei dati, chiaro specchio della notevole dispersione delle osservazioni nello spazio: probabilmente conoscendo a priori il numero di componenti e dunque imponendo al processo $k = 2$ sarebbe stato possibile trovare partizioni decisamente migliori. Si noti come tutti i risultati, ma nello specifico quest'ultimo, siano condizionati dal vincolo imposto per motivi tempistici e computazionali circa il numero massimo di gruppi presenti nei dati, ossia 5: con ogni probabilità, se si fosse lasciato libero il modello di valutare un numero maggiore di componenti, avremmo trovato partizioni più inadeguate.

Al contrario, nei macroscenari 3 e 4 si assiste (tabelle 4.5, 4.7) ad una tendenza opposta e, *ceteris paribus*, l'ARI medio tende ad aumentare all'aumentare di r e p. La motivazione di ciò risiede proprio nella struttura stessa dei gruppi in questi scenari: la presenza di una differente dinamica in varianza richiede spazi ampi, ossia di valori elevati di p ed r, per divenire evidente in quanto essendo la media dei due gruppi la medesima in uno spazio a bassa dimensionalità i punti tendono a confondersi e sovrapporsi. Ad ulteriore riprova di ciò, si osservi il caso anomalo (rispetto agli altri presentati nel presente studio) in riferimento al rapporto ARI/ numero di gruppi sintetizzato in tabella 4.5: con $p = 2$ i risultati non sono particolarmente soddisfacenti nonostante il numero di gruppi trovati sia esattamente uguale a 2 a causa della scarsa capacità dell'algoritmo di discriminare i punti del primo gruppo rispetto a quelli del secondo.

Le considerazioni fin qui proposte non possono assolutamente prescindere dalla numerosità campionaria: il fenomeno da noi analizzato necessita di molte osservazioni a causa dell'elevato numero di parametri da stimare, e pertanto la scelta di n è ricaduta su due valori abbastanza elevati come 250 e 1000. L'aumento della numerosità campionaria

comporta due effetti *apparentemente* contrastanti a seconda del caso considerato: in presenza di osservazioni normali, il passaggio di n da 250 a 1000 provoca un aumento medio dell'*Adjusted Rand Index*, con osservazioni asimmetriche una diminuzione. È presumibile che tale diminuzione derivi dal fatto che l'incremento delle osservazioni a disposizione rende il campione maggiormente identificativo del fenomeno reale sottostante: così facendo i dati distribuiti normalmente tenderanno a dividersi in gruppi sempre più definiti, mentre i dati distribuiti asimmetricamente si allontaneranno sempre di più dalla struttura normale ricercata dall'algoritmo.

Infine, lo standard error, come preventivabile, tende a calare all'aumentare della numerosità campionaria, risultando costantemente maggiore in presenza di osservazioni distribuite asimmetricamente a causa della poca robustezza del metodo in tali occasioni.

4.4 Conclusioni

Lo scopo del presente lavoro è stato quello di studiare, da un punto di vista teorico prima e da uno empirico in seguito, le peculiarità del *clustering* per dati *three-way* basato su un approccio parametrico, evidenziandone limiti e punti di forza.

La specificazione di un modello statistico sottostante inquadra l'analisi così condotta in un contesto matematico assolutamente rigoroso, estremamente funzionale in riferimento all'applicazione di procedure inferenziali. Lo studio di simulazione ha evidenziato ottimi risultati in presenza di gruppi distribuiti normalmente, cogliendo praticamente nella totalità dei casi la reale classe di appartenenza delle osservazioni. Al contrario, tale metodologia ha evidenziato delle carenze quando l'assunzione di normalità dei gruppi non è soddisfatta: la motivazione risiede nel criterio utilizzato, che come precedentemente messo in luce ricerca gruppi normali all'interno dei dati. Tale risultato, per quanto preventivabile, rappresenta un limite evidente del metodo in quanto, nei casi reali, l'ipotesi di normalità è difficilmente rispettata.

In una struttura complessa come quella dei dati *three-way*, un ruolo di primaria importanza è rivestito dal numero di variabili p e dal numero di occasioni r: non è possibile, almeno nell'esperimento da noi condotto, tracciare una dinamica unica per i due parametri, ma molto dipende dallo specifico scenario considerato. In linea generale, si è potuto assistere ad un miglioramento dei raggruppamenti all'aumentare di p ed

r in presenza di gruppi normali, così come in presenza di distribuzioni asimmetriche e gruppi separati in varianza, mentre un peggioramento qualora i gruppi fossero asimmetrici e separati in media. In quest'ultimo caso, difatti, le osservazioni tendono a disperdersi eccessivamente nello spazio $\mathbb{R}^{(r \times p)}$ e l'algoritmo fatica a cogliere la reale situazione.

Un primo gravoso limite che ha segnato l'intera simulazione è rappresentato dall'instabilità del metodo al variare del raggruppamento di partenza dei dati: come illustrato, l'algoritmo EM difatti necessita di un valore di *start* per l'iniziale stima dei parametri della mistura. In più occasioni è parso lampante di come al variare di tale valore il processo generasse, sullo stesso campione, partizioni completamente differenti per ARI, numero di gruppi rilevato e caratteristiche geometriche dei *cluster*. Evidentemente l'algoritmo in presenza di dati a tre vie tende a convergere sistematicamente a massimi locali della funzione di massima verosimiglianza e non, come sperato, di massimo globale risultando così estremamente instabile.

I tempi di stima dei diversi scenari sono risultati assai lunghi a causa della pesantezza computazionale del metodo: il gran numero di parametri da stimare iterativamente richiede per forza di cose uno sforzo notevole che si traduce in tempistiche assai elevate. Tale aspetto ha condizionato il lavoro svolto, per esempio facendo propendere la scelta del numero massimo di variabili ed occasioni verso valori piuttosto contenuti. Tra le variabili in input all'algoritmo utilizzato, ossia n,p,r,k, quella che ha maggiormente inciso sule tempistiche è stata la numerosità campionaria n, ma per trarre conclusioni più precise sarebbe necessario uno studio sistematico circa l'incidenza delle variabili menzionate sul tempo t condotto tramite un modello lineare. Nell'ottica sempre di ridurre i tempi del processo, è possibile imporre alcune limitazioni sulle componenti della mistura come presentato nel paragrafo 3.3, pervenendo però così a risultati subottimali.

Tutte le considerazioni finora proposte sono comunque da inquadrare negli specifici scenari affrontati e non sono da ritenere esaustive per l'analisi dell'approccio parametrico al *clustering three-way*.

I possibili sviluppi del presente lavoro sono numerosi: un'idea potrebbe essere quella di aumentare la dimensionalità dei dati sia rispetto al numero di variabili che al numero di occasioni per definire in maniera più chiara la dipendenza tra i risultati forniti e tali valori; un ulteriore utile approfondimento richiederebbe di considerare dati non provenienti da un processo autoregressivo AR (1) ma qualche dinamica evolutiva alternativa e inoltre dati rilevati in spazi geografici diversi. Alla lu-

ce anche delle considerazioni precedentemente esposte, si può valutare l'imposizione di un valore massimo maggiore del numero di gruppi così da permettere il confronto attraverso il fattore Bayesiano tra un numero maggiore di modelli.

Bibliografia

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

Banfield J. D.; Raftery A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, pp. 803–821.

Basford K. E.; McLachlan G. J. (1985). The mixture method of clustering applied to three-way data. *Journal of Classification*, **2**(1), 109–125.

Biernacki C.; Celeux G.; Govaert G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **22**(7), 719–725.

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

Celeux G.; Chauveau D.; Diebolt J. (1996). Stochastic versions of the em algorithm: an experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, **55**(4), 287–314.

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

De Waal D. (1985). Matrix-valued distributions. *Encyclopedia of Statistical Sciences*.

Dempster A. P.; Laird N. M.; Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Di Fonzo T.; Lisi F. (2005). *Serie storiche economiche: analisi statistiche e applicazioni*. Carocci.

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.

Gupta A.; Varga T. (1992). Characterization of matrix variate normal distributions. *Journal of Multivariate Analysis*, **41**(1), 80–88.

Gupta A. K.; Nagar D. K. (1999). *Matrix variate distributions*, volume 104. CRC Press.

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

Kass R. E.; Raftery A. E. (1995). Bayes factors. *Journal of the american statistical association*, **90**(430), 773–795.

Keribiin C. (1998). Estimation consistante de l'ordre de modèles de mélange. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, **326**(2), 243–248.

Leroux B. G. *et al.* (1992). Consistent estimation of a mixing distribution. *The Annals of Statistics*, **20**(3), 1350–1360.

Mardia K. V.; Kent J. T.; Bibby J. M. (1979). *Multivariate analysis*. Academic press.

McLachlan G.; Krishnan T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

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

McLachlan G. J.; Krishnan T.; Ng S. K. (2004). The em algorithm. Relazione tecnica, Papers/Humboldt-Universität Berlin, Center for Applied Statistics and Economics (CASE).

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rizzi A.; Vichi M. (1995). Representation, synthesis, variability and data preprocessing of a three-way data set. *Computational statistics & data analysis*, **19**(2), 203–222.

Vichi M.; Rocci R.; Kiers H. A. (2007). Simultaneous component and clustering models for three-way data: within and between approaches. *Journal of Classification*, **24**(1), 71–98.

Viroli C. (2010). *mmn.em: Mixture of Matrix-Normal distributions*. R package version 2.0.

Viroli C. (2011). Finite mixtures of matrix normal distributions for classifying three-way data. *Statistics and Computing*, **21**(4), 511–522.