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

Statistica per l'Economia e l'Impresa

Modellazione dell'eterogeneità non osservata mediante variabili latenti continue e categoriali

Relatore: prof.ssa Giovanna Menardi
Dipartimento di Scienze Statistiche

Laureando: Enrico Guglielmino
Matricola n. 2104124

Anno Accademico 2025/2026

Indice

Introduzione

Nell'ambito della statistica multivariata e della moderna analisi dei dati, molti fenomeni osservati sono il risultato di strutture non direttamente misurabili. In contesti applicativi quali la psicometria, le scienze sociali, l'economia o la biologia, le variabili osservate rappresentano spesso manifestazioni parziali e rumorose di dimensioni sottostanti più generali. Di conseguenza, diventa necessario disporre di strumenti in grado di ridurre la complessità dei dati e, al tempo stesso, di fornire una rappresentazione interpretabile della loro struttura interna.

L'Analisi Fattoriale costituisce uno degli strumenti più noti e storicamente più rilevanti per affrontare questa criticità. Essa consente di spiegare la covariazione tra molteplici variabili osservate attraverso un numero ridotto di fattori latenti continui, permettendo così di sintetizzare l'informazione contenuta nei dati e di interpretarla in termini di costrutti non osservabili. Tuttavia, il modello fattoriale classico si fonda sull'ipotesi che la popolazione analizzata sia omogenea, cioè che un'unica struttura latente sia valida per tutti gli individui. Tale assunzione può risultare restrittiva quando i dati provengono da sottopopolazioni differenti, non osservabili direttamente, ma caratterizzate da comportamenti o strutture distributive diverse.

I modelli a mistura finita affrontano questo secondo aspetto introducendo una variabile latente categoriale, che consente di rappresentare la popolazione come composta da un numero finito di gruppi non osservati. Anche questo approccio presenta però limiti importanti: quando il numero di variabili osservate cresce, il numero di parametri da stimare cresce più che linearmente, con conseguenti problemi di instabilità, sovradattamento e perdita di interpretabilità.

Il presente lavoro si colloca all'intersezione tra questi due approcci. L'obiettivo centrale è lo studio di modelli statistici in grado di descrivere l'eterogeneità non osservata attraverso l'integrazione di due diverse tipologie di variabili latenti: fattori continui, volti a misurare dimensioni sottostanti non direttamente osservabili, e variabili categoriali, finalizzate all'individuazione di sottopopolazioni nascoste. In questo quadro si inseriscono i modelli a mistura fattoriale che, integrando la logica dell'analisi fattoriale con quella delle misture finite, consentono di combinare riduzione della dimensionalità e individuazione di gruppi latenti.

La trattazione si propone di sistematizzare una letteratura nota per essere frammentata, esaminando sia il filone statistico, orientato alla parsimonia parametrica e alla modellazione di dati ad alta dimensionalità, sia quello psicometrico, più attento al significato sostantivo dei costrutti latenti e alla comparabilità delle misure tra gruppi. Pur essendo nati da esigenze metodologiche distinte ed essendo spesso presentati con nomenclature e finalità differenti, questi modelli possono essere ricondotti a una cornice probabilistica comune, differenziandosi per la struttura dei

vincoli imposti e per le domande di ricerca a cui intendono rispondere. L'elaborato è organizzato come segue:

- Il primo capitolo introduce i modelli di base su cui si fonda l'intera trattazione: l'Analisi Fattoriale e i modelli a mistura finita. Per entrambi vengono presentate la specificazione del modello, la stima di massima verosimiglianza dei parametri mediante l'algoritmo EM, i criteri di selezione del modello e i principali limiti teorici e operativi.

- Il secondo capitolo è dedicato al *Factor Mixture Modelling*. Dopo aver introdotto la logica generale e un modello generale di riferimento, vengono discussi separatamente il filone statistico e quello psicometrico e i principali modelli che si sono diffusi nelle comunità scientifiche di riferimento. A dispetto della disorganicità della letteratura, si procederà ad una sua sistematizzazione volta a mettere in luce analogie e differenze.

- Il terzo capitolo propone uno studio di simulazione Monte Carlo finalizzato a confrontare il comportamento di diversi approcci in condizioni controllate. In particolare, vengono analizzati l'Analisi Fattoriale, i modelli a mistura gaussiana e due modelli a mistura fattoriale, esponenti del filone statistico e del filone psicometrico, con l'obiettivo di valutare in quali condizioni ciascun modello riesca a recuperare correttamente la struttura latente generatrice e quanto sia robusto in presenza di misspecificazione.

- Il quarto capitolo presenta infine un'applicazione a dati reali, provenienti dal contesto psicometrico e noti per il frequente utilizzo di modelli fattoriali per l'individuazione di fattori latenti dei tratti della personalità, ma consente anche di indagare l'eventuale presenza di profili latenti di risposta. L'analisi empirica permette quindi di verificare se, accanto alla struttura continua dei tratti di personalità, emergano forme di eterogeneità categoriale interpretabili.

La tesi intende quindi mostrare come l'integrazione tra fattori continui e classi latenti possa offrire una rappresentazione più flessibile dei dati multivariati complessi. Tale prospettiva consente infatti di distinguere tra variabilità continua all'interno della popolazione e differenze tra sottogruppi, mantenendo al tempo stesso una lettura parsimoniosa e interpretabile della struttura osservata.

Capitolo 1

Modelli Base

1.1 Analisi Fattoriale

1.1.1 Introduzione e cenni storici

L'Analisi Fattoriale classica costituisce una metodologia statistica multivariata fondamentale per descrivere e spiegare la struttura sottostante a un insieme di variabili osservate. Il paradigma teorico di riferimento postula che le variabili osservate siano una manifestazione rumorosa di un segnale non direttamente osservabile, ovvero latente, espresso in uno spazio a dimensionalità ridotta. L'obiettivo del modello fattoriale classico consiste, dunque, nell'identificare tali fattori e nel definire la loro relazione con le variabili osservate, operando una sintesi dell'informazione che riduca la dimensionalità senza compromettere la ricchezza informativa dei dati originari. In termini pratici, questo processo consente di isolare il segnale dal rumore di fondo, facilitando l'interpretazione di fenomeni complessi che risulterebbero altrimenti difficilmente analizzabili attraverso la mera osservazione delle singole variabili.

Dal punto di vista storico, l'Analisi Fattoriale affonda le sue radici nei primi anni del XX secolo con i lavori di Spearman (1904), il quale cercava di fornire una base matematica rigorosa alla teoria secondo cui le prestazioni in vari test cognitivi fossero tutte riconducibili a una singola dimensione latente, che egli definì il fattore di "intelligenza generale" o fattore g. Il successivo superamento del modello monofattoriale si deve a Thurstone (1931), il quale, introducendo l'Analisi Fattoriale Multipla, ammise la possibilità che le variabili osservate potessero dipendere da un insieme di molteplici fattori interagenti. Tale intuizione, nata in ambito psicometrico, si è rivelata nel tempo una metodologia estremamente versatile, trovando applicazioni cruciali in ambiti eterogenei come l'economia, la biologia e le scienze sociali dove ancora oggi rappresenta uno strumento essenziale per l'analisi dei dati.

1.1.2 Specificazione del modello

Per formalizzare il modello, si consideri un vettore $Y = (Y_1, \ldots, Y_p)'$ di variabili osservate spesso denominate *item* nella letteratura psicometrica. Tale vettore è esprimibile come combinazione lineare di un insieme di fattori latenti $F = (F_1, \ldots, F_q)'$, con $q < p$, più un termine di errore specifico $\epsilon = (\epsilon_1, \ldots, \epsilon_p)$.

Formalmente, il modello può essere espresso come segue:

$$\begin{aligned} Y_1 &= \mu_1 + \lambda_{11}F_1 + \lambda_{12}F_2 + \cdots + \lambda_{1q}F_q + \epsilon_1 \\ Y_2 &= \mu_2 + \lambda_{21}F_1 + \lambda_{22}F_2 + \cdots + \lambda_{2q}F_q + \epsilon_2 \\ &\vdots \\ Y_p &= \mu_p + \lambda_{p1}F_1 + \lambda_{p2}F_2 + \cdots + \lambda_{pq}F_q + \epsilon_p \end{aligned}$$

e in forma matriciale:

$$Y = \mu + \Lambda F + \epsilon. \tag{1.1}$$

In tale rappresentazione, $\mu \in \mathbb{R}^{p\times 1}$ è il vettore delle medie delle variabili osservate; $\Lambda \in \mathbb{R}^{p\times q}$ è la matrice dei pesi fattoriali (o *factor loadings*), i cui elementi esprimono l'intensità e la direzione della relazione tra ciascuna variabile osservata e i fattori latenti; $F \in \mathbb{R}^{q\times 1}$ rappresenta il vettore dei fattori latenti; infine, $\epsilon \in \mathbb{R}^{p\times 1}$ è il vettore aleatorio degli errori specifici (o fattori unici), associati alle singole variabili osservate.

Al fine di garantire l'identificabilità del modello definito nell'Equazione (1.1) e di ridurre il numero di parametri liberi da stimare, si introducono le seguenti assunzioni standard:

1. $\mathbb{E}[F] = 0,\ \mathbb{Cov}[F] = \mathbb{E}[FF'] = I_q$
2. $\mathbb{E}[\epsilon] = 0,\ \mathbb{Cov}[\epsilon] = \mathbb{E}[\epsilon\epsilon'] = \Psi = diag(\psi_1, \ldots, \psi_p)$
3. $F \perp \epsilon \implies \mathbb{Cov}[F, \epsilon] = \mathbb{E}[F\epsilon'] = 0$

Tali ipotesi rivestono un ruolo centrale nell'interpretazione del modello. In primo luogo, il vincolo sulla covarianza latente $\mathbb{Cov}[F] = I_q$ (Assunzione 1) definisce il quadro del modello fattoriale ortogonale, nel quale si assume l'assenza di correlazione tra i costrutti latenti. Sebbene in alcuni contesti applicativi si ricorra a modelli obliqui (per i quali $\mathbb{Cov}[F] = \Phi$), l'impostazione ortogonale è ampiamente adottata in letteratura grazie ai suoi vantaggi in termini di parsimonia parametrica e trattabilità analitica.

In secondo luogo, la struttura diagonale della matrice delle covarianze degli errori Ψ (Assunzione 2) formalizza il principio di indipendenza locale (o condizionata). Essa implica che, condizionatamente ai fattori latenti, le variabili osservate risultino tra loro incorrelate: l'intera covariazione tra gli *item* è dunque spiegata unicamente dai fattori F.

Infine, l'incorrelazione esplicita tra fattori ed errori (Assunzione 3) garantisce che la componente specifica di ciascuna variabile non contenga alcuna informazione legata ai costrutti comuni.

A queste ipotesi di natura strutturale si affianca una fondamentale assunzione distribuzionale: si suppone che sia i fattori latenti sia gli errori idiosincratici seguano una distribuzione normale multivariata, ossia $F \sim \mathcal{N}_q(0, I_q)$ ed $\epsilon \sim \mathcal{N}_p(0, \Psi)$.

Per le proprietà di chiusura rispetto alle combinazioni lineari della distribuzione normale, ne consegue che anche il vettore delle variabili osservate Y segue una

distribuzione normale multivariata, con vettore delle medie pari a μ e matrice di covarianza data da:

$$\begin{aligned}\mathbb{Cov}[Y] &= \mathbb{E}[(Y-\mu)(Y-\mu)'] \\ &= \mathbb{E}[(\Lambda F+\epsilon)(\Lambda F+\epsilon)'] \\ &= \Lambda\mathbb{E}[FF']\Lambda' + \Lambda\mathbb{E}[F\epsilon'] + \mathbb{E}[\epsilon F']\Lambda' + \mathbb{E}[\epsilon\epsilon'] \\ &= \Lambda\Lambda' + \Psi = \Sigma.\end{aligned}$$

Questa scomposizione rappresenta il fulcro dell'Analisi Fattoriale, in quanto evidenzia come la variabilità totale dei dati osservati possa essere scomposta in una componente comune ($\Lambda\Lambda'$), governata dalla struttura latente, e una componente specifica (Ψ), attribuibile al rumore.

Infine, la matrice di covarianza incrociata tra variabili osservate e fattori latenti è data da:

$$\begin{aligned}\mathbb{Cov}[Y,F] &= \mathbb{E}[(Y-\mu)F'] = \mathbb{E}[(\Lambda F+\epsilon)F'] \\ &= \Lambda\mathbb{E}[FF'] + \mathbb{E}[\epsilon F'] = \Lambda.\end{aligned}$$

1.1.3 Stima dei parametri

Sotto l'assunzione di normalità multivariata, il metodo d'elezione per la stima dei parametri del modello fattoriale è il metodo della Massima Verosimiglianza (Bartholomew *et al.*, 2011). A partire dai dati osservati $y_1, \ldots, y_n$, intesi come realizzazioni indipendenti del vettore aleatorio Y, è possibile definire la funzione di Verosimiglianza per l'insieme di parametri $\Omega = (\mu, \Lambda, \Psi)$ come segue:

$$\begin{aligned}L(\Omega|Y) &= \prod_{i=1}^{n} \phi_p(y_i;\mu,\Sigma) \\ &= \prod_{i=1}^{n}\left[(2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y_i-\mu)'\Sigma^{-1}(y_i-\mu)\right)\right] \\ &= \prod_{i=1}^{n}\left[(2\pi)^{-\frac{p}{2}}|\Lambda\Lambda'+\Psi|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y_i-\mu)'(\Lambda\Lambda'+\Psi)^{-1}(y_i-\mu)\right)\right],\end{aligned}$$

dove ϕ_p denota la funzione di densità della distribuzione normale multivariata di dimensione p.

La rispettiva funzione di log-verosimiglianza marginale si ottiene applicando il logaritmo naturale:

$$\begin{aligned}\ell(\Omega|Y) &= \ln[L(\Omega|Y)] \\ &= \sum_{i=1}^{n}\ln\left[(2\pi)^{-\frac{p}{2}}|\Lambda\Lambda'+\Psi|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(y_i-\mu)'(\Lambda\Lambda'+\Psi)^{-1}(y_i-\mu)\right)\right] \\ &= -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln|\Lambda\Lambda'+\Psi| - \frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)'(\Lambda\Lambda'+\Psi)^{-1}(y_i-\mu). \quad (1.2)\end{aligned}$$

La massimizzazione diretta della funzione (1.2) rispetto a Λ e Ψ risulta estremamente complessa dal punto di vista analitico. Questo accade perché i parametri

da stimare sono legati tra loro sia all'interno del determinante che nell'inversa della matrice di covarianza Σ. Tale struttura rende impossibile ricavare delle soluzioni in forma chiusa attraverso le normali derivate prime, richiedendo l'uso di procedure iterative.

Un approccio alternativo ed elegante per ottenere le stime di Massima Verosimiglianza si basa sull'impiego dell'algoritmo *Expectation-Maximization* (EM, Dempster *et al.* 1977). Si tratta di una procedura iterativa sviluppata specificamente per determinare le stime di massima verosimiglianza dei parametri di un modello in contesti caratterizzati da dati mancanti o variabili latenti non osservate. Concettualmente, il metodo supera l'impossibilità di osservare l'intero campione massimizzando il valore atteso della funzione di log-verosimiglianza (calcolato rispetto alla distribuzione condizionata dei dati mancanti), alternando iterativamente un passo di calcolo del valore atteso (*Expectation*) e un passo di massimizzazione (*Maximization*).

In quest'ottica, i fattori latenti f_i (intesi come realizzazioni del vettore casuale F per l'osservazione i-esima, con $i = 1, \ldots, n$) vengono trattati a tutti gli effetti come dati mancanti e, insieme alle variabili osservate Y, costituiscono i cosiddetti *dati completi*. Assumendo per semplicità che le variabili osservate siano centrate (ovvero $\mu = 0$; in caso contrario, la stima di massima verosimiglianza per μ coinciderebbe semplicemente col vettore delle medie campionarie $\bar{y}$ prima di applicare l'algoritmo EM), il modello per ogni osservazione si riduce a $y_i = \Lambda f_i + \epsilon_i$. Sotto tali assunzioni, è possibile definire la distribuzione dei dati completi come la distribuzione congiunta di Y ed F:

$$\begin{bmatrix} Y \\ F \end{bmatrix} \sim \mathcal{N}_{p+q} \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Lambda\Lambda' + \Psi & \Lambda \\ \Lambda' & I_q \end{bmatrix} \right). \tag{1.3}$$

A partire da tale distribuzione congiunta, la densità di probabilità dei dati completi per la singola osservazione può essere fattorizzata come $p(y_i, f_i) = p(y_i|f_i)p(f_i)$. Di conseguenza, sotto l'ipotesi di indipendenza tra le n osservazioni, la funzione di Verosimiglianza dei dati completi è la seguente:

$$L^c(\Lambda, \Psi|Y, F) = \prod_{i=1}^{n} p(y_i, f_i) = \prod_{i=1}^{n} p(y_i|f_i)p(f_i)$$

e la relativa funzione di log-verosimiglianza:

$$\ell^c(\Lambda, \Psi|Y, F) = \ln L^c(\Lambda, \Psi|Y, F) = \sum_{i=1}^{n} \ln p(y_i|f_i) + \sum_{i=1}^{n} \ln p(f_i). \tag{1.4}$$

Tuttavia, poiché la distribuzione marginale dei fattori $p(f_i)$ è una normale standard $\mathcal{N}_q(0, I_q)$ indipendente dai parametri da stimare (Λ e Ψ), il suo contributo si riduce a una costante additiva C in fase di massimizzazione logaritmica.

Di conseguenza, l'inferenza si concentra unicamente sulla distribuzione condizionata $y_i|f_i$. Per derivare tale distribuzione, si consideri l'equazione strutturale del modello centrato:

$$y_i = \Lambda f_i + \epsilon_i.$$

Condizionando rispetto a f_i, il termine fattoriale Λf_i agisce a tutti gli effetti come una quantità deterministica. Poiché l'unica componente stocastica residua è l'errore

specifico ϵ_i, il quale è distribuito per assunzione come $\mathcal{N}(0, \Psi)$ e risulta indipendente da f_i, si ottiene:

$$\mathbb{E}[y_i|f_i] = \mathbb{E}[\Lambda f_i + \epsilon_i|f_i] = \Lambda f_i + \mathbb{E}[\epsilon_i] = \Lambda f_i$$

$$\mathbb{V}\mathrm{ar}(y_i|f_i) = \mathbb{V}\mathrm{ar}(\Lambda f_i + \epsilon_i|f_i) = \mathbb{V}\mathrm{ar}(\epsilon_i) = \Psi.$$

Essendo una trasformazione lineare di una variabile normale, si conclude formalmente che la distribuzione condizionata è anch'essa normale:

$$y_i|f_i \sim \mathcal{N}(\Lambda f_i, \Psi). \tag{1.5}$$

Sostituendo l'espressione analitica di questa densità normale nella funzione di log-verosimiglianza definita in (1.4), e inglobando il termine marginale dei fattori nella costante C, si ottiene la formula finale:

$$\begin{aligned}\ell^c(\Lambda, \Psi|Y, F) &= \sum_{i=1}^{n} \ln\left[(2\pi)^{-\frac{p}{2}}|\Psi|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y_i - \Lambda f_i)'\Psi^{-1}(y_i - \Lambda f_i)\right)\right] + C \\ &= c - \frac{n}{2}\ln|\Psi| - \frac{1}{2}\sum_{i=1}^{n}(y_i - \Lambda f_i)'\Psi^{-1}(y_i - \Lambda f_i),\end{aligned}$$

dove $c = C - \frac{np}{2}\ln(2\pi)$ raggruppa le costanti rispetto ai parametri. Sviluppando la forma quadratica all'interno della sommatoria otteniamo:

$$(y_i - \Lambda f_i)'\Psi^{-1}(y_i - \Lambda f_i) = y_i'\Psi^{-1}y_i - 2y_i'\Psi^{-1}\Lambda f_i + f_i'\Lambda'\Psi^{-1}\Lambda f_i. \tag{1.6}$$

Poiché F non è osservato, non è possibile calcolare e massimizzare direttamente ℓ^c. L'algoritmo EM supera questo problema alternando iterativamente due fasi: il passo E (*Expectation*) e il passo M (*Maximization*), fino a convergenza.

Il Passo E. All'iterazione t, l'obiettivo del passo E è calcolare il valore atteso della log-verosimiglianza dei dati completi, condizionatamente ai dati osservati Y e alle stime correnti dei parametri $\Omega^{(t)} = (\Lambda^{(t)}, \Psi^{(t)})$. Tale valore atteso, noto in letteratura come funzione Q, è formalmente definito come:

$$Q(\Omega|\Omega^{(t)}) = \mathbb{E}[\ell^c(\Lambda, \Psi|Y, F)|Y, \Omega^{(t)}].$$

Sostituendo l'espansione della forma quadratica ricavata nell'Equazione (1.6) e sfruttando la linearità dell'operatore di valore atteso (in cui i dati osservati y_i agiscono come costanti), si ottiene l'espressione analitica della funzione Q che si esplicita come segue:

$$\begin{aligned}Q = c - \frac{n}{2}\ln|\Psi| - \sum_{i=1}^{n}\Bigg(&\frac{1}{2}y_i'\Psi^{-1}y_i - y_i'\Psi^{-1}\Lambda\mathbb{E}[f_i|y_i, \Omega^{(t)}] \\ &+ \frac{1}{2}\mathrm{tr}\left(\Lambda'\Psi^{-1}\Lambda\mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}]\right)\Bigg),\end{aligned} \tag{1.7}$$

dove l'ultimo termine all'interno della sommatoria si ottiene applicando la proprietà ciclica della traccia ($\mathrm{tr}(f_i'\Lambda'\Psi^{-1}\Lambda f_i) = \mathrm{tr}(\Lambda'\Psi^{-1}\Lambda f_i f_i')$).

L'Equazione (1.7) evidenzia chiaramente che, per definire completamente la funzione Q in vista della successiva massimizzazione, è sufficiente calcolare esclusivamente i primi due momenti condizionati dei fattori latenti. Sfruttando le proprietà della distribuzione normale multivariata, il valore atteso condizionato del fattore per l'individuo i è dato da:

$$\mathbb{E}[f_i|y_i, \Omega^{(t)}] = \mathbb{C}\mathrm{ov}[f_i, y_i]\mathbb{V}\mathrm{ar}[y_i]^{-1}y_i = \Lambda^{(t)\prime}(\Lambda^{(t)}\Lambda^{(t)\prime} + \Psi^{(t)})^{-1}y_i = \beta^{(t)}y_i,$$

dove $\beta^{(t)} = \Lambda^{(t)\prime}(\Lambda^{(t)}\Lambda^{(t)\prime} + \Psi^{(t)})^{-1}$ rappresenta la matrice dei coefficienti di regressione. Tale formulazione corrisponde al noto metodo di Thomson (1939) per la stima dei *factor scores*. Analogamente, la matrice di covarianza condizionata si ottiene:

$$\begin{aligned}\mathbb{V}\mathrm{ar}[f_i|y_i, \Omega^{(t)}] &= \mathbb{V}\mathrm{ar}[f_i] - \mathbb{C}\mathrm{ov}[f_i, y_i]\mathbb{V}\mathrm{ar}[y_i]^{-1}\mathbb{C}\mathrm{ov}[y_i, f_i] \\ &= I_q - \Lambda^{(t)\prime}(\Lambda^{(t)}\Lambda^{(t)\prime} + \Psi^{(t)})^{-1}\Lambda^{(t)} = I_q - \beta^{(t)}\Lambda^{(t)}.\end{aligned}$$

Di conseguenza, il secondo momento condizionato non centrato (necessario per il calcolo della traccia nell'ultimo termine dell'Equazione 1.7) risulta essere:

$$\begin{aligned}\mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}] &= \mathbb{V}\mathrm{ar}[f_i|y_i, \Omega^{(t)}] + \mathbb{E}[f_i|y_i, \Omega^{(t)}]\mathbb{E}[f_i|y_i, \Omega^{(t)}]' \\ &= I_q - \beta^{(t)}\Lambda^{(t)} + \beta^{(t)}y_i y_i'\beta^{(t)\prime}.\end{aligned}$$

Il Passo M. Nel passo M, si aggiornano le stime dei parametri massimizzando la funzione Q calcolata nel passo E. Per trovare il punto di massimo, è necessario calcolare le derivate parziali della funzione Q rispetto ai parametri di interesse Λ e Ψ e porle pari a zero.

Per ricavare l'equazione di aggiornamento della matrice dei pesi fattoriali Λ, si calcola la derivata parziale di Q rispetto a Λ. Sfruttando le regole del calcolo matriciale per la traccia e le forme quadratiche, otteniamo:

$$\frac{\partial Q}{\partial \Lambda} = \sum_{i=1}^{n} \left(\Psi^{-1}y_i\mathbb{E}[f_i|y_i, \Omega^{(t)}]' - \Psi^{-1}\Lambda\mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}]\right) = 0.$$

Pre-moltiplicando ambo i membri per Ψ (che assumiamo definita positiva e quindi invertibile) e isolando Λ, otteniamo la stima aggiornata per l'iterazione $(t+1)$:

$$\Lambda^{(t+1)} \sum_{i=1}^{n} \mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}] = \sum_{i=1}^{n} y_i\mathbb{E}[f_i|y_i, \Omega^{(t)}]'$$

$$\Lambda^{(t+1)} = \left(\sum_{i=1}^{n} y_i\mathbb{E}[f_i|y_i, \Omega^{(t)}]'\right)\left(\sum_{i=1}^{n} \mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}]\right)^{-1}. \tag{1.8}$$

Per quanto riguarda la matrice delle varianze d'errore Ψ, risulta analiticamente più agevole derivare la funzione Q rispetto alla sua inversa, ovvero la matrice Ψ^{-1}. La derivata, posta a zero, è:

$$\begin{aligned}\frac{\partial Q}{\partial \Psi^{-1}} = \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n}\Bigg(&y_i y_i' - \Lambda^{(t+1)}\mathbb{E}[f_i|y_i, \Omega^{(t)}]y_i' \\ &- y_i\mathbb{E}[f_i|y_i, \Omega^{(t)}]'\Lambda^{(t+1)\prime} + \Lambda^{(t+1)}\mathbb{E}[f_i f_i'|y_i, \Omega^{(t)}]\Lambda^{(t+1)\prime}\Bigg) = 0.\end{aligned}$$

Sostituendo $\Lambda^{(t+1)}$ trovata in precedenza, gli ultimi due termini della sommatoria si semplificano. Moltiplicando per $\frac{2}{n}$ e ricordando l'assunzione di indipendenza degli errori specifici ϵ, è necessario imporre che la matrice Ψ rimanga strettamente diagonale. Si applica dunque l'operatore diag$(\cdot)$ che azzera gli elementi fuori dalla diagonale principale:

$$\Psi^{(t+1)} = \frac{1}{n}\text{diag}\left(\sum_{i=1}^{n} y_i y_i' - \Lambda^{(t+1)} \sum_{i=1}^{n} \mathbb{E}[f_i|y_i, \Omega^{(t)}] y_i'\right). \tag{1.9}$$

La sequenza iterativa definita dai passi E ed M gode di una fondamentale proprietà analitica, formalmente dimostrata nel lavoro seminale di Dempster *et al.* (1977). Ogni iterazione dell'algoritmo EM produce un aggiornamento dei parametri che incrementa il valore della funzione di log-verosimiglianza marginale (o incompleta). Formalmente:

$$\ell(\Omega^{(t+1)}|Y) \geq \ell(\Omega^{(t)}|Y).$$

Questa garanzia di incremento monotono assicura che l'algoritmo converga sempre verso un punto stazionario della funzione di verosimiglianza (tipicamente un massimo locale). Sulla base di tale proprietà, il processo iterativo viene arrestato quando l'incremento della log-verosimiglianza tra due iterazioni successive scende al di sotto di una soglia di tolleranza arbitrariamente piccola (es. 10^{-6}), oppure quando la norma della differenza tra i vettori dei parametri stimati in due step consecutivi risulta trascurabile. Tuttavia, è fondamentale sottolineare che l'algoritmo EM garantisce la convergenza verso un massimo locale, non necessariamente globale. Pertanto, la bontà della soluzione finale dipende fortemente dai valori iniziali dei parametri $\Omega^{(0)}$. Nella prassi operativa, si ovvia a questo limite eseguendo l'algoritmo con multipli avvii.

1.1.4 Selezione del modello e scelta del numero di fattori

Nell'Analisi Fattoriale, un problema cruciale risiede nella determinazione della complessità ottimale del modello, ovvero la scelta del numero appropriato di fattori latenti q. Poiché la funzione di log-verosimiglianza $\ell(\Omega|Y)$ è monotonamente non decrescente rispetto al numero di parametri stimati, basarsi unicamente sulla sua massimizzazione porterebbe inevitabilmente all'*overfitting*, spingendo a estrarre fattori spuri che modellano il rumore campionario piuttosto che la reale struttura latente. Per ovviare a questo problema, la prassi statistica ricorre ai Criteri di Informazione, i quali introducono una penalizzazione proporzionale al numero di parametri liberi del modello. Tra i più diffusi e robusti spicca il *Bayesian Information Criterion* (BIC, Schwarz 1978), originariamente definito come:

$$BIC = -2\ell(\hat{\Omega}|Y) + \nu \ln(n), \tag{1.10}$$

dove $\hat{\Omega}$ rappresenta la stima di Massima Verosimiglianza dei parametri, ν è il numero totale di parametri indipendenti stimati nel modello fattoriale e n è la numerosità campionaria. Nella pratica operativa, si stimano modelli sequenziali per diversi valori di q e si seleziona quello che restituisce il valore minimo del BIC, garantendo così il miglior compromesso teorico tra bontà di adattamento ai dati e parsimonia.

1.1.5 Limiti del modello fattoriale classico

Nonostante la flessibilità del modello fattoriale, almeno in linea di principio, l'approccio classico si fonda su assunzioni teoriche stringenti che ne limitano l'applicabilità in determinati contesti reali. Di seguito vengono analizzate le principali criticità.

Indeterminazione dei fattori e rotazione

Il modello fattoriale espresso nell'Equazione (1.1) presenta un problema intrinseco di identificabilità noto come indeterminazione rotazionale. Data una qualsiasi matrice ortogonale T di dimensioni $q \times q$ (tale per cui $TT' = I_q$), è possibile definire una nuova matrice dei pesi $\Lambda^* = \Lambda T$ e un nuovo vettore di fattori $F^* = T'F$, tali per cui la matrice di covarianza implicata dal modello rimane invariata:

$$\mathbb{Cov}[Y] = (\Lambda T)(\Lambda T)' + \Psi = \Lambda T T' \Lambda' + \Psi = \Lambda^*(\Lambda^*)' + \Psi.$$

Ciò implica l'esistenza di infinite combinazioni di parametri capaci di riprodurre la stessa struttura di covarianza osservata. Per ovviare a tale limite, la letteratura statistica e psicometrica hanno sviluppato diversi approcci metodologici.

L'Analisi Fattoriale Esplorativa affronta il problema a posteriori tramite tecniche di rotazione degli assi (ortogonali o oblique), cercando di trasformare la matrice Λ per approssimare una "struttura semplice" che faciliti l'interpretazione dei costrutti latenti. Il concetto di struttura semplice, formalizzato originariamente da Thurstone (1931), implica che ciascuna variabile osservata presenti un peso fattoriale elevato su un singolo fattore e valori prossimi allo zero su tutti gli altri. Nella pratica operativa, ciò si ottiene applicando algoritmi di rotazione ortogonale (come la *Varimax*, che preserva l'incorrelazione tra i fattori massimizzando la varianza dei pesi) o obliqua (come la *Promax*, che ammette costrutti latenti correlati qualora teoricamente plausibile).

Nell'Analisi Fattoriale Confermativa, invece, l'indeterminazione viene risolta a priori. Basandosi su una solida teoria di riferimento, il ricercatore impone dei vincoli di identificabilità fissando esplicitamente a zero determinati elementi della matrice Λ, stabilendo in anticipo quali variabili siano indicatori di quali fattori. Inoltre, per risolvere definitivamente l'indeterminazione di scala, si fissa la metrica dei costrutti latenti imponendo la varianza del fattore pari a uno, oppure vincolando a uno il peso di una variabile indicatrice di riferimento. Queste restrizioni rendono il modello unicamente identificabile senza alcuna necessità di rotazione empirica, permettendo di testare rigorosamente la validità della struttura ipotizzata tramite indici di adattamento globale.

Tuttavia, in specifici contesti analitici, quali gli studi di simulazione Monte Carlo o la valutazione della stabilità fattoriale nel tempo, il ricercatore dispone di una matrice generatrice o *target* Λ vera e nota a priori. In tali casistiche, l'obiettivo non è la ricerca esplorativa di una struttura interpretabile, bensì la misurazione rigorosa della discrepanza tra le stime ottenute e il parametro reale.

Per aggirare il problema dell'indeterminazione rotazionale in questo scenario, senza dover ricorrere a criteri euristici di struttura semplice, si applica l'Analisi di Procruste Ortogonale (Schönemann, 1966). Formalmente, data la matrice dei pesi stimata $\hat{\Lambda}$ e la matrice reale Λ, il problema consiste nel determinare la matrice di

trasformazione ortogonale T che minimizza la distanza tra i due spazi, quantificata tramite la norma di Frobenius:

$$\min_T ||\hat{\Lambda}T - \Lambda||_F^2 \quad \text{sotto il vincolo} \quad T'T = I_q. \tag{1.11}$$

La soluzione analitica si ottiene attraverso la Decomposizione ai Valori Singolari della matrice $M = \hat{\Lambda}'\Lambda$. Posto $M = U\Sigma V'$, la matrice di rotazione ottimale è definita come $\hat{T} = UV'$. Questa procedura permette di allineare lo spazio delle stime a quello dei parametri reali, eliminando l'arbitrarietà della rotazione.

L'assunzione di normalità e la natura dei dati

Un ulteriore limite del modello fattoriale classico riguarda l'ipotesi di normalità multivariata. In particolare, il modello assume che le variabili osservate siano continue e distribuite normalmente. Tuttavia, in numerosi ambiti applicativi, come la psicometria, i dati raccolti tramite questionari sono spesso di natura discreta, ad esempio dicotomici come vero/falso oppure ordinali come le scale Likert.

L'applicazione dell'analisi fattoriale classica a tali dati può generare distorsioni rilevanti. Poiché il modello si basa sulla correlazione lineare, esso tende a raggruppare tra loro gli *item* sulla base della similarità delle loro distribuzioni marginali, piuttosto che sulla base del costrutto latente sottostante. Questo fenomeno può condurre all'identificazione di fattori spuri, nei quali gli *item* vengono aggregati semplicemente in funzione della loro probabilità di risposta. Inoltre, l'impiego di un modello lineare su variabili discrete risulta teoricamente incoerente: le previsioni del modello non sono vincolate allo spazio delle categorie osservabili e possono assumere valori continui non interpretabili nel contesto dei dati.

Per affrontare tali criticità, la letteratura psicometrica ha sviluppato la Teoria della Risposta all'*Item*. Questo approccio sostituisce la struttura lineare con modelli basati su funzioni di collegamento non lineari (ad esempio *logit* o *probit*), che permettono di modellare direttamente la probabilità di risposta in funzione del tratto latente.

Infine, l'analisi fattoriale classica si limita a definire il cosiddetto "modello di misura": essa esplora o conferma come le variabili manifeste si raggruppino attorno ai costrutti latenti, ma non consente di testare relazioni direzionali o di dipendenza tra i costrutti stessi. Quando l'obiettivo di ricerca richiede di validare teorie complesse o analizzare reti di nessi causali tra i fattori, la sola analisi fattoriale è insufficiente. Per colmare questa lacuna è necessario ricorrere ai Modelli di Equazioni Strutturali (Bollen 1989), i quali superano i limiti dell'analisi fattoriale integrandola al loro interno: uniscono un modello di misura (tipicamente un'Analisi Fattoriale Confermativa) a un "modello strutturale", ovvero un sistema di equazioni di regressione simultanee che agisce direttamente sulle variabili latenti, al netto degli errori di misurazione.

Il limite dei dati omogenei

Un ulteriore limite dell'approccio classico riguarda l'assunzione di omogeneità della popolazione. Il modello presuppone, infatti, che il vettore dei parametri Ω (medie μ,

pesi fattoriali Λ e varianze d'errore Ψ) sia costante e valido per tutte le unità statistiche che compongono il campione. In altre parole, si assume implicitamente che il costrutto latente misurato operi con le stesse dinamiche per ogni singolo individuo.

Tuttavia, nell'analisi dei dati reali, è frequente imbattersi in campioni eterogenei, composti da sottopopolazioni non osservate (classi latenti) caratterizzate da parametri distributivi differenti. L'applicazione di un modello fattoriale globale a una popolazione intrinsecamente eterogenea, ignorando tale stratificazione latente, può condurre a gravi *bias* metodologici: distorsioni sistematiche nelle stime, errata valutazione dell'adattamento del modello o l'estrazione di fattori spuri. Il superamento di questa limitazione rende necessaria l'integrazione della modellistica a variabili latenti continue (i fattori) con metodologie capaci di identificare differenze qualitative non osservate tra gli individui. A tal fine, risulta essenziale procedere con l'introduzione dei modelli a mistura.

1.2 Modelli a mistura finita

1.2.1 Specificazione del modello

I modelli a mistura finita, la cui trattazione è ampiamente discussa da McLachlan e Peel (2000), costituiscono una classe di modelli probabilistici utilizzati per analizzare dati nei quali si ipotizza la presenza di sottopopolazioni distinte (o *cluster*), la cui appartenenza non è osservabile direttamente per ciascuna unità statistica. In altri termini, quando un fenomeno non si manifesta in modo omogeneo, si può supporre che tale eterogeneità derivi dalla presenza di più gruppi latenti, ciascuno caratterizzato da una propria distribuzione di probabilità.

Formalmente, denotiamo con $y_1, \ldots, y_n$ un campione di dati osservati, con $y_i \in \mathbb{R}^p$, intesi come realizzazioni indipendenti di un vettore aleatorio Y e ipotizziamo che tale campione provenga da una popolazione suddivisa in G classi non direttamente osservabili. Sia Z una variabile latente che identifica l'appartenenza alle classi e indichiamo con $z_i = (z_{i1}, \ldots, z_{iG})$ il vettore di variabili indicatrici che ne rappresenta la realizzazione per l'osservazione i-esima, tale che $z_{ig} = 1$ se l'osservazione i-esima appartiene alla sottopopolazione g, e $z_{ig} = 0$ altrimenti. La variabile di interesse Y, a seconda della sua natura, assume una funzione di probabilità o di densità diversa in base alla sottopopolazione di appartenenza. Condizionatamente all'appartenenza alla classe g, la distribuzione per la singola osservazione i è:

$$y_i | z_{ig} = 1 \sim p_g(y_i; \theta_g), \quad g = 1, \ldots, G.$$

Marginalizzando rispetto alla variabile latente Z, si ottiene la funzione di densità marginale tipica dei modelli a mistura:

$$\begin{aligned} p(y_i) &= \sum_{g=1}^{G} p(y_i, z_{ig} = 1) \\ &= \sum_{g=1}^{G} p(z_{ig} = 1) p_g(y_i | z_{ig} = 1, \theta_g) = \sum_{g=1}^{G} \pi_g p_g(y_i | \theta_g), \end{aligned} \tag{1.12}$$

dove le proporzioni o pesi della mistura $\pi_g = p(z_{ig} = 1)$ rappresentano le probabilità a priori che un'unità appartenga al g-esimo *cluster*, con i vincoli $0 \leq \pi_g \leq 1$ e $\sum_{g=1}^{G} \pi_g = 1$. Le funzioni $p_g(y_i|\theta_g)$ costituiscono le densità componenti della mistura.

A partire dai dati osservati, intesi come realizzazioni indipendenti della distribuzione definita in (1.12), la funzione di verosimiglianza per il vettore di tutti i parametri del modello $\Omega = (\pi, \Theta) = (\pi_1, \ldots, \pi_G, \theta_1, \ldots, \theta_G)$ è:

$$L(\Omega|Y) = \prod_{i=1}^{n} p(y_i|\Omega) = \prod_{i=1}^{n} \sum_{g=1}^{G} \pi_g p_g(y_i|\theta_g)$$

e la rispettiva funzione di log-verosimiglianza marginale risulta essere:

$$\ell(\Omega|Y) = \ln[L(\Omega|Y)] = \sum_{i=1}^{n} \ln \left[\sum_{g=1}^{G} \pi_g p_g(y_i|\theta_g) \right]. \tag{1.13}$$

Come illustrato nell'Equazione (1.13), la massimizzazione analitica diretta di tale funzione risulta ardua a causa della presenza di una sommatoria all'interno dell'operatore logaritmico. Per superare tale ostacolo computazionale e ottenere le stime di Massima Verosimiglianza, si ricorre nuovamente all'algoritmo EM, assegnando alle etichette di gruppo Z il ruolo di dati mancanti.

1.2.2 Stima dei parametri

Prima di introdurre i passi specifici dell'algoritmo EM, è necessario ridefinire il problema nei termini della verosimiglianza dei dati completi. Si assume che, oltre alle osservazioni manifeste $y_1, \ldots, y_n$, siano teoricamente disponibili anche i vettori latenti $z_1, \ldots, z_n$, che indicano l'effettiva etichetta di classe per ciascuna unità i. Tali vettori latenti si assumono indipendenti e identicamente distribuiti secondo una legge Multinomiale con singola prova:

$$z_i \sim \text{Mult}_{\text{G}}(1, \pi), \quad \text{con } \pi = (\pi_1, \ldots, \pi_G).$$

A partire dai dati completi, rappresentati dall'insieme delle coppie (y_i, z_i) per $i = 1, \ldots, n$, si può definire la funzione di verosimiglianza completa come la distribuzione congiunta delle variabili osservate e latenti:

$$\begin{aligned} L^C(\Omega|Y, Z) &= \prod_{i=1}^{n} p(y_i, z_i|\Omega) = \prod_{i=1}^{n} p(z_i|\pi) p(y_i|z_i, \Theta) \\ &= \prod_{i=1}^{n} \left(\prod_{g=1}^{G} \pi_g^{z_{ig}} \right) \left(\prod_{g=1}^{G} p_g(y_i|\theta_g)^{z_{ig}} \right) = \prod_{i=1}^{n} \prod_{g=1}^{G} (\pi_g p_g(y_i|\theta_g))^{z_{ig}}. \end{aligned}$$

Applicando il logaritmo naturale, si ottiene la rispettiva log-verosimiglianza completa:

$$\begin{aligned}\ell^C(\Omega|Y,Z) &= \ln\left[\prod_{i=1}^{n}\prod_{g=1}^{G}(\pi_g p_g(y_i|\theta_g))^{z_{ig}}\right] \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln(\pi_g p_g(y_i|\theta_g)) \\ &= \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\ln p_g(y_i|\theta_g). \qquad (1.14)\end{aligned}$$

Come si evince dall'Equazione (1.14), l'introduzione delle variabili latenti indicatrici z_{ig} consente di estrarre la sommatoria sulle classi dal logaritmo. Inoltre, la funzione risulta ora additivamente scomponibile in due termini distinti: il primo dipendente esclusivamente dai pesi della mistura π, e il secondo dipendente unicamente dai parametri delle distribuzioni componenti Θ.

Il Passo E. Nel passo E all'iterazione t, si calcola il valore atteso della log-verosimiglianza dei dati completi ℓ^C, condizionatamente ai dati osservati Y e alle stime correnti dei parametri $\Omega^{(t)} = (\pi^{(t)}, \Theta^{(t)})$. Tale valore atteso definisce la funzione Q:

$$Q(\Omega|\Omega^{(t)}) = E[\ell^C(\Omega|Y,Z)|Y,\Omega^{(t)}].$$

Sostituendo l'espressione della log-verosimiglianza completa ricavata nell'Equazione (1.14), si nota che l'unica componente non osservata rispetto alla quale calcolare l'aspettativa è la variabile indicatrice latente z_{ig} (McLachlan e Peel, 2000):

$$Q(\Omega|\Omega^{(t)}) = \sum_{i=1}^{n}\sum_{g=1}^{G} E[z_{ig}|y_i,\Omega^{(t)}]\ln\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G} E[z_{ig}|y_i,\Omega^{(t)}]\ln p_g(y_i|\theta_g).$$

Poiché z_{ig} è una variabile dicotomica che assume valore 1 se l'unità i appartiene alla classe g e 0 altrimenti, il suo valore atteso condizionato coincide con la probabilità a posteriori che l'unità i appartenga a tale classe. Questa quantità, comunemente denotata con $\tau_{ig}^{(t)}$, può essere calcolata analiticamente applicando il Teorema di Bayes:

$$\begin{aligned}\tau_{ig}^{(t)} = p(z_{ig}=1|y_i,\Omega^{(t)}) &= \frac{p(z_{ig}=1|\Omega^{(t)})p(y_i|z_{ig}=1,\Omega^{(t)})}{p(y_i|\Omega^{(t)})} \\ &= \frac{\pi_g^{(t)}p_g(y_i|\theta_g^{(t)})}{\sum_{j=1}^{G}\pi_j^{(t)}p_j(y_i|\theta_j^{(t)})}.\end{aligned}$$

La quantità $\tau_{ig}^{(t)}$ rappresenta il "peso" o la "responsabilità" che la componente g-esima assume nello spiegare l'osservazione y_i all'iterazione corrente. Sostituendo $\tau_{ig}^{(t)}$ nella funzione Q, si ottiene la forma esplicita da massimizzare nel passo successivo:

$$Q(\Omega|\Omega^{(t)}) = \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(t)}\ln\pi_g + \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(t)}\ln p_g(y_i|\theta_g). \qquad (1.15)$$

Il Passo M. Nel passo M, si aggiornano le stime massimizzando la funzione $Q(\Omega|\Omega^{(t)})$ appena definita rispetto al vettore dei parametri Ω. Come si evince dalla struttura additiva dell'Equazione (1.15), la massimizzazione può essere condotta separatamente per i pesi della mistura π e per i parametri delle componenti Θ.

Per ricavare le stime aggiornate dei pesi $\pi_g^{(t+1)}$, si massimizza il primo termine di Q sotto il vincolo di probabilità $\sum_{g=1}^{G} \pi_g = 1$. Introducendo un moltiplicatore di Lagrange λ, la funzione obiettivo diviene:

$$H(\pi, \lambda) = \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(t)} \ln \pi_g - \lambda \left(\sum_{g=1}^{G} \pi_g - 1 \right).$$

Calcolando la derivata parziale rispetto al generico peso π_g e ponendola pari a zero, si ottiene:

$$\frac{\partial H}{\partial \pi_g} = \frac{1}{\pi_g} \sum_{i=1}^{n} \tau_{ig}^{(t)} - \lambda = 0 \implies \pi_g = \frac{1}{\lambda} \sum_{i=1}^{n} \tau_{ig}^{(t)}.$$

Sommando ambo i membri su $g = 1, \ldots, G$ e ricordando che $\sum_{g=1}^{G} \pi_g = 1$, si ricava $\lambda = \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(t)}$. Poiché la somma delle probabilità a posteriori per un singolo individuo su tutte le classi è sempre unitaria ($\sum_{g=1}^{G} \tau_{ig}^{(t)} = 1$), ne consegue che $\lambda = n$. Pertanto, l'equazione di aggiornamento in forma chiusa per i pesi della mistura è data dalla media campionaria delle probabilità a posteriori:

$$\pi_g^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ig}^{(t)}.$$

Per quanto riguarda il vettore dei parametri specifici di classe Θ_g, l'aggiornamento si ottiene massimizzando il secondo termine della funzione Q. Questo porta a risolvere, per ciascuna classe g, il seguente sistema di equazioni di verosimiglianza pesata:

$$\sum_{i=1}^{n} \tau_{ig}^{(t)} \frac{\partial \ln p_g(y_i|\theta_g)}{\partial \theta_g} = 0.$$

Le proprietà analitiche e di convergenza dell'algoritmo EM, introdotte precedentemente nel contesto dell'Analisi Fattoriale, si estendono naturalmente ai modelli a mistura. Anche in questo caso, l'alternanza iterativa dei passi E ed M garantisce un incremento monotono della funzione di log-verosimiglianza marginale, guidando l'algoritmo verso un punto stazionario. Il processo viene arrestato quando la differenza tra le log-verosimiglianze in due iterazioni successive risulta inferiore a una soglia prestabilita.

Tuttavia, la superficie della funzione di verosimiglianza nei modelli a mistura è notoriamente complessa, multimodale e ricca di massimi locali. Di conseguenza, la sensibilità dell'algoritmo rispetto ai valori di partenza $\Omega^{(0)}$ risulta ancora più critica. Per mitigare il rischio di convergere verso soluzioni sub-ottimali, nella prassi operativa si ricorre a tecniche di *pre-clustering* gerarchico o partizionale, tra cui spicca l'algoritmo *K-means*.

1.2.3 Modelli a mistura gaussiana

Il caso applicativo di gran lunga più diffuso all'interno della famiglia dei modelli a mistura finita si ottiene assumendo che le distribuzioni componenti p_g siano normali multivariate; in tal senso si parla di *Gaussian Mixture Models* (GMM). Sotto tale assunzione, la componente g-esima è completamente caratterizzata dal vettore dei parametri $\theta_g = (\mu_g, \Sigma_g)$, dove μ_g è il vettore delle medie di dimensione $p \times 1$ e Σ_g è la matrice di covarianza di dimensione $p \times p$. La densità della mistura diviene pertanto:

$$p(y_i|\Omega) = \sum_{g=1}^{G} \pi_g \phi_p(y_i; \mu_g, \Sigma_g).$$

Nel contesto dei GMM, la massimizzazione della funzione Q nel Passo M dell'algoritmo EM conduce a stimatori in forma chiusa estremamente intuitivi. Isolandone la parte dipendente da θ_g, la funzione da massimizzare per la singola classe g è:

$$Q(\theta_g) = \sum_{i=1}^{n} \tau_{ig}^{(t)} \ln \phi_p(y_i; \mu_g, \Sigma_g).$$

Sostituendo l'espressione analitica della log-densità normale multivariata, si ottiene:

$$Q(\mu_g, \Sigma_g) = \sum_{i=1}^{n} \tau_{ig}^{(t)} \left(-\frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma_g| - \frac{1}{2} (y_i - \mu_g)' \Sigma_g^{-1} (y_i - \mu_g) \right).$$

Per trovare l'equazione di aggiornamento del vettore delle medie μ_g, si calcola la derivata parziale di Q rispetto a μ_g e la si pone pari a zero:

$$\frac{\partial Q}{\partial \mu_g} = \sum_{i=1}^{n} \tau_{ig}^{(t)} \Sigma_g^{-1} (y_i - \mu_g) = 0.$$

Pre-moltiplicando ambo i membri per Σ_g (assunta non singolare) e isolando μ_g, si ottiene la stima aggiornata all'iterazione $(t+1)$:

$$\sum_{i=1}^{n} \tau_{ig}^{(t)} y_i = \left(\sum_{i=1}^{n} \tau_{ig}^{(t)} \right) \mu_g \implies \mu_g^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ig}^{(t)} y_i}{\sum_{i=1}^{n} \tau_{ig}^{(t)}}.$$

Analogamente, per ottenere l'aggiornamento della covarianza risulta analiticamente più agevole derivare rispetto alla sua inversa, ovvero la matrice Σ_g^{-1}. La derivata, posta a zero, è:

$$\frac{\partial Q}{\partial \Sigma_g^{-1}} = \frac{1}{2} \sum_{i=1}^{n} \tau_{ig}^{(t)} \Sigma_g - \frac{1}{2} \sum_{i=1}^{n} \tau_{ig}^{(t)} (y_i - \mu_g^{(t+1)})(y_i - \mu_g^{(t+1)})' = 0.$$

Risolvendo l'equazione rispetto a Σ_g, si ricava la formula di aggiornamento:

$$\Sigma_g^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ig}^{(t)} (y_i - \mu_g^{(t+1)})(y_i - \mu_g^{(t+1)})'}{\sum_{i=1}^{n} \tau_{ig}^{(t)}} = \frac{1}{N_g^{(t)}} \sum_{i=1}^{n} \tau_{ig}^{(t)} (y_i - \mu_g^{(t+1)})(y_i - \mu_g^{(t+1)})'.$$

Come si evince dalle dimostrazioni, tali espressioni corrispondono alle classiche formule della media e della matrice di covarianza campionaria, in cui però il contributo di ciascuna osservazione y_i è ponderato per la rispettiva probabilità a posteriori $\tau_{ig}^{(t)}$, e dove il denominatore $N_g^{(t)} = \sum_{i=1}^{n} \tau_{ig}^{(t)}$ rappresenta il numero effettivo di unità assegnate alla classe g .

1.2.4 Selezione del modello e scelta del numero di classi

Analogamente a quanto avviene per la scelta dei fattori nell'Analisi Fattoriale, l'applicazione dei modelli a mistura richiede la determinazione a priori del numero ottimale di componenti latenti G. Anche in questo contesto, l'incremento di G porta a un aumento automatico e fittizio della log-verosimiglianza marginale $\ell(\Omega|Y)$, rendendo inefficaci i test basati unicamente sull'adattamento ai dati per evitare la sovra-parametrizzazione. Nella letteratura del *model-based clustering*, lo strumento d'elezione per guidare la selezione del modello è nuovamente il Bayesian Information Criterion (BIC). Adattando il principio introdotto in precedenza, il criterio valuta il bilanciamento tra la verosimiglianza della mistura e la penalità derivante dal numero totale di parametri liberi ν (che, come vedremo, nei GMM non vincolati cresce molto rapidamente) secondo l'espressione (1.10).

1.2.5 Limiti dei GMM in Alta Dimensionalità

Nonostante l'eleganza formale e l'estrema flessibilità nel modellare densità complesse, l'impiego dei GMM classici incontra severe limitazioni pratiche e teoriche all'aumentare del numero di variabili p, problematiche ampiamente documentate nella letteratura del *model-based clustering* (Bouveyron *et al.*, 2019). I limiti principali derivano dall'esplosione parametrica e dalla geometria degli spazi iper-dimensionali.

Omoschedasticità vs eteroschedasticità

In un modello GMM generale, si assume che ciascuna componente latente possieda una propria matrice di covarianza Σ_g definita positiva e non vincolata (mistura *eteroschedastica*). Il numero totale di parametri liberi da stimare per l'intero modello, indicato con ν, è dato da:

$$\nu = (G-1) + Gp + G\frac{p(p+1)}{2}. \tag{1.16}$$

Risulta evidente come il numero di parametri cresca con un ordine quadratico $\mathcal{O}(p^2)$ rispetto alla dimensione dello spazio. Per mitigare questa sovra-parametrizzazione, si ricorre spesso a modelli ristretti, imponendo ad esempio l'assunzione di *omoschedasticità* ($\Sigma_g = \Sigma$ per ogni g). Sebbene ciò riduca il numero di parametri a $(G-1) + Gp + p(p+1)/2$, tale vincolo costringe geometricamente tutti i *cluster* a condividere la medesima forma, lo stesso orientamento e lo stesso volume nello spazio p-dimensionale, un'assunzione quasi sempre irrealistica nelle applicazioni su dati reali, motivo per cui si preferisce il ricorso a componenti eteroschedastiche.

La Maledizione della dimensionalità e il fenomeno dello spazio vuoto

Quando la dimensionalità p è elevata, i modelli eteroschedastici subiscono gli effetti deleteri della "maledizione della dimensionalità" (*curse of dimensionality*, coniato da Bellman 1954). All'aumentare di p, il volume dello spazio dei dati cresce in modo esponenziale, rendendo qualsiasi campione di taglia n intrinsecamente "sparso". Questo genera il cosiddetto fenomeno dello spazio vuoto: in alta dimensione,

i dati tendono a posizionarsi vicino ai bordi dello spazio o a distanziarsi uniformemente, annullando il concetto di "vicinanza" euclidea. Di conseguenza, diventa estremamente arduo per l'algoritmo EM individuare regioni di alta densità locale che giustifichino la presenza di *cluster*.

Singolarità della funzione di Verosimiglianza

Una conseguenza diretta dell'alta dimensionalità e della sovra-parametrizzazione nei modelli GMM eteroschedastici è il problema dell'illimitatezza della funzione di verosimiglianza, ampiamente discusso da McLachlan e Peel (2000). Affinché la matrice di covarianza empirica Σ_g sia invertibile, è necessario che il numero effettivo di osservazioni assegnate al *cluster* sia superiore al numero di variabili ($N_g > p$). Tuttavia, se durante l'algoritmo EM una componente si contrae fino a centrare la propria media μ_g esattamente su una singola osservazione y_i (o su un numero esiguo di punti coplanari), la matrice Σ_g diviene singolare e il suo determinante tende a zero ($|\Sigma_g| \to 0$). Poiché la densità normale multivariata presenta il termine $|\Sigma_g|^{1/2}$ al denominatore, questo collasso spinge il valore della funzione di verosimiglianza verso l'infinito. Tale singolarità corrompe l'ottimizzazione, portando l'algoritmo a convergere verso soluzioni degeneri, note come *spurious maximizers*, prive di qualsiasi utilità pratica.

Capitolo 2

I modelli a mistura fattoriale

2.1 Un modello generale di riferimento

Come discusso nel Capitolo 1, l'Analisi Fattoriale, pur rimanendo ancora oggi uno strumento di riferimento per la riduzione della dimensionalità, è soggetta a diversi limiti teorici. Tra questi, l'ipotesi di normalità multivariata costituisce un vincolo che nella pratica empirica risulta spesso difficile da rispettare: i dati reali manifestano frequentemente asimmetrie, code pesanti o, più in generale, deviazioni dalla forma gaussiana che il modello fattoriale classico non è in grado di accogliere. A questa criticità si affianca, come evidenziato, l'assunzione di omogeneità della popolazione, raramente plausibile in presenza di sottopopolazioni latenti caratterizzate da dinamiche distinte.

I modelli a mistura finita si pongono come una naturale generalizzazione in grado di superare entrambe le limitazioni. Da un lato, essi contemplano il caso gaussiano come caso particolare, ma consentono al contempo di descrivere fenomeni la cui distribuzione complessiva è asimmetrica o comunque non normale, approssimandola attraverso una combinazione di componenti elementari. Dall'altro, attraverso l'introduzione di una variabile latente categoriale, permettono di modellare popolazioni eterogenee, suddividendole in classi non osservabili a priori. Tuttavia, come discusso nella Sezione 1.2.5, l'applicazione dei modelli a mistura gaussiana in spazi ad alta dimensionalità è fortemente penalizzata dalla proliferazione dei parametri nelle matrici di covarianza delle singole componenti.

I modelli a mistura fattoriale (*Factor Mixture Modelling*, FMM) nascono dalla volontà di mettere insieme queste due classi di modelli, l'Analisi Fattoriale e i modelli a mistura finita, integrando all'interno di un unico modello probabilistico due diverse tipologie di variabili latenti: una variabile categoriale che identifica le classi latenti della popolazione e un insieme di fattori continui che descrivono la variabilità individuale all'interno di ciascuna classe. In questo modo, la popolazione viene rappresentata come un insieme di sottogruppi caratterizzati da specifiche strutture latenti, consentendo di distinguere la variabilità tra gruppi dalla variabilità interna ai gruppi stessi. La caratteristica distintiva degli FMM risiede nella stima congiunta delle componenti categoriali e continue: le classi latenti e i fattori non vengono determinati in modo sequenziale, ma stimati simultaneamente attraverso la massimizzazione della funzione di verosimiglianza, evitando così le potenziali distorsioni

derivanti dall'applicazione separata di tecniche di riduzione della dimensionalità e di *clustering*.

Nonostante l'apparente unitarietà della formulazione, la letteratura sui modelli a mistura fattoriale si è sviluppata lungo due percorsi metodologici simili ma caratterizzati da differenze anche rilevanti, nati per rispondere a esigenze diverse: il filone statistico, orientato al *machine learning* e alla parsimonia parametrica in contesti ad alta dimensionalità, e il filone psicometrico, basato sui modelli a equazioni strutturali e focalizzato sul significato sostantivo dei costrutti e sulla loro comparabilità tra sottopopolazioni.

Per comprendere in modo unitario questa classe di modelli, è utile partire dalla formulazione generale proposta da Lubke e Muthén (2005), che fornisce un quadro di riferimento entro cui collocare le diverse specificazioni introdotte nel seguito del capitolo. Assumendo l'assenza di covariate esterne, coerentemente con l'obiettivo del presente lavoro, rivolto all'individuazione dell'eterogeneità non osservata, l'equazione strutturale per un'osservazione y_i appartenente alla classe g è:

$$y_i \mid (z_{ig} = 1) = \nu_g + \Lambda_g f_{ig} + \epsilon_{ig}, \tag{2.1}$$

dove ν_g è un vettore di parametri specifici di classe che concorre a determinare la media delle variabili osservate per la classe g, Λ_g è la matrice dei pesi fattoriali specifica di classe, e gli errori specifici si distribuiscono come $\epsilon_{ig} \sim \mathcal{N}_p(0, \Psi_g)$. A differenza della parametrizzazione fattoriale classica, in cui i fattori hanno media nulla e covarianza unitaria, qui i fattori latenti seguono una distribuzione gaussiana con parametri specifici di classe:

$$f_{ig} \mid (z_{ig} = 1) \sim \mathcal{N}_q(\alpha_g, \Phi_g),$$

dove α_g è il vettore delle medie fattoriali, che cattura il livello medio del tratto latente nel gruppo g, e Φ_g è la matrice di covarianza latente. Questo implica che, a livello di popolazione marginale, la distribuzione dei fattori sia essa stessa modellata come una mistura finita di gaussiane:

$$f_i \sim \sum_{g=1}^{G} \pi_g \mathcal{N}_q(\alpha_g, \Phi_g).$$

Di conseguenza, il vettore delle medie osservate e la matrice di covarianza per la classe g risultano strutturati come:

$$\mu_g = \nu_g + \Lambda_g \alpha_g, \qquad \Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Psi_g.$$

Tale scomposizione esplicita delle medie e delle covarianze osservate nelle loro controparti latenti costituisce l'elemento centrale della formulazione, in quanto permette di distinguere la struttura del modello di misura, rappresentata dai pesi Λ_g e dalle intercette ν_g, dalle differenze distribuzionali tra sottopopolazioni, codificate nelle medie e nelle covarianze latenti α_g e Φ_g.

È importante sottolineare che il modello (2.1) non rappresenta il precursore cronologico delle specificazioni che saranno discusse nel resto del capitolo: la sua formulazione, di matrice psicometrica, è infatti successiva a diversi modelli del filone

statistico. Esso costituisce piuttosto un quadro concettuale di riferimento, rispetto al quale le diverse proposte presenti in letteratura possono essere lette, almeno idealmente se non cronologicamente, come specificazioni derivate. Va tuttavia osservato che il modello (2.1), nella sua forma più generale con tutti i parametri liberi di variare tra le classi, non è identificabile. La media marginale $\mu_g = \nu_g + \Lambda_g \alpha_g$ non consente di separare il contributo dell'intercetta da quello della media fattoriale, mentre la covarianza $\Sigma_g = \Lambda_g \Phi_g \Lambda_g' + \Psi_g$ è invariante rispetto a trasformazioni lineari della metrica dei fattori, riproponendo in forma aggravata l'indeterminazione rotazionale già discussa nella Sezione 1.1.5.

Per ottenere un modello stimabile è dunque necessario imporre opportune restrizioni, ed è proprio nella diversa natura di tali restrizioni che i due filoni si differenziano. Da un lato, il filone statistico mantiene i pesi fattoriali liberi di variare tra le classi, normalizzando la distribuzione dei fattori e scaricando l'eterogeneità sul modello di misura; dall'altro, il filone psicometrico impone una struttura di misura comune tra le classi, in particolare attraverso il vincolo $\Lambda_g = \Lambda$, riconducendo l'eterogeneità sostantiva alla distribuzione dei fattori latenti, così da preservare la comparabilità dei costrutti tra sottopopolazioni. Le due sezioni seguenti sviluppano in dettaglio questi due regimi di vincolo.

2.2 L'approccio statistico

2.2.1 Mixtures of factor analyzers

Il filone statistico si sviluppa con un'enfasi specifica sugli aspetti computazionali e sulla gestione della complessità parametrica in contesti ad alta dimensionalità. Come discusso nel Capitolo 1 (Sezione 1.2.5), l'impiego dei modelli a mistura gaussiana in spazi ad alta dimensionalità è fortemente limitato dalla proliferazione dei parametri nelle matrici di covarianza delle singole componenti. In questo contesto, la struttura fattoriale viene introdotta come strumento di regolarizzazione, con l'obiettivo di ridurre il numero di parametri necessari a descrivere la variabilità locale senza compromettere eccessivamente la flessibilità del modello. L'idea centrale consiste nello scomporre la covarianza di ciascuna componente in una parte a basso rango, spiegata da un insieme ridotto di fattori latenti, e in una componente residua specifica, ottenendo per il generico *cluster* g una struttura del tipo:

$$\Sigma_g = \Lambda_g \Lambda_g' + \Psi_g.$$

Tale scomposizione consente di preservare la capacità del modello di adattarsi a strutture complesse nei dati, riducendo al contempo il rischio di instabilità numerica e di singolarità delle matrici di covarianza. In termini del modello generale (2.1), questa strategia corrisponde a normalizzare la distribuzione dei fattori latenti — ponendo $\alpha_g = 0$ e $\Phi_g = I_q$ — e a lasciare liberi i pesi fattoriali Λ_g. È proprio l'aver fissato la metrica del fattore a $\Phi_g = I_q$ a far sì che il termine a basso rango assuma la forma $\Lambda_g \Lambda_g'$ anziché quella più generale $\Lambda_g \Phi_g \Lambda_g'$. Sotto questo regime di vincolo, l'eterogeneità tra le classi non risiede nella distribuzione dei fattori, che mantiene una metrica standard comune a tutti i gruppi, ma viene interamente assorbita dal

modello di misura, ovvero dai parametri μ_g, Λ_g e Ψ_g. È entro questo quadro che si collocano i modelli presentati nel seguito della sezione.

La prima formalizzazione dei modelli a mistura di fattori nel filone statistico si deve a Ghahramani *et al.* (1996), che hanno introdotto le *Mixtures of Factor Analyzers* (MFA). Da un punto di vista formale, si assume che, per ogni componente g ($g = 1, \ldots, G$), la relazione tra le variabili osservate e i costrutti latenti sia descritta dalla seguente equazione strutturale:

$$y_i|(z_{ig} = 1) = \mu_g + \Lambda_g f_{ig} + \epsilon_i, \tag{2.2}$$

dove, condizionatamente all'appartenenza al *cluster* g (ovvero $z_{ig} = 1$), μ_g è il vettore delle medie della componente, Λ_g è la matrice dei pesi fattoriali di dimensione $p \times q$ (con $q \ll p$), e f_{ig} è il vettore dei fattori latenti distribuito come $f_{ig} \sim \mathcal{N}_q(0, I_q)$. Il termine ϵ_i rappresenta il vettore degli errori specifici.

Nella formulazione originaria del modello, si assume che gli errori siano distribuiti come $\epsilon_i \sim \mathcal{N}_p(0, \Psi)$, dove Ψ è una matrice diagonale comune a tutte le componenti della mistura. Questa restrizione implica che la quota di variabilità non spiegata dai fattori latenti sia identica in tutti i gruppi della mistura. Da un punto di vista interpretativo, gli errori specifici vengono trattati come una sorgente di rumore comune all'intera popolazione, indipendente dalla struttura locale dei *cluster*. Sotto tali assunzioni, poiché f_{ig} ed ϵ_i sono assunti indipendenti, il modello osservazionale per ogni componente condizionato ai fattori latenti risulta:

$$y_i \mid (f_{ig}, z_{ig} = 1) \sim \mathcal{N}_p(\mu_g + \Lambda_g f_{ig}, \Psi).$$

Integrando rispetto alla distribuzione dei fattori latenti f_{ig}, la distribuzione marginale di y_i condizionata all'appartenenza al gruppo g diviene una normale multivariata:

$$y_i|z_{ig} = 1 \sim \mathcal{N}_p(\mu_g, \Lambda_g\Lambda_g' + \Psi).$$

Marginalizzando ulteriormente rispetto alla variabile indicatrice di classe z_{ig} (che segue una distribuzione multinomiale con probabilità π_g), la densità marginale per l'osservazione y_i diviene una mistura finita di gaussiane:

$$p(y_i) = \sum_{g=1}^{G} \pi_g \phi_p(y_i; \mu_g, \Lambda_g\Lambda_g' + \Psi),$$

dove la matrice di covarianza locale è $\Sigma_g = \Lambda_g\Lambda_g' + \Psi$.

Successivamente, McLachlan *et al.* (2003) hanno generalizzato il modello MFA, rilassando l'assunzione di uguaglianza degli errori e permettendo a ciascun *cluster* di avere la propria matrice di unicità Ψ_g. In questo modo, la covarianza locale diventa $\Sigma_g = \Lambda_g\Lambda_g' + \Psi_g$, rendendo il modello molto più flessibile per catturare l'eteroschedasticità dei dati.

Da un punto di vista geometrico e interpretativo, il modello MFA, invece di proiettare l'intero *set* di dati su un unico iperpiano a bassa dimensionalità, approssima la distribuzione dei dati attraverso G iperpiani orientati diversamente nello spazio originario a p dimensioni. Le matrici Λ_g definiscono l'orientamento e l'inclinazione di tali iperpiani per ciascun *cluster*, mentre i vettori μ_g ne determinano la posizione

spaziale. Questa caratteristica rende l'MFA uno strumento estremamente potente per modellare relazioni non lineari complesse, approssimandole in modo flessibile tramite una combinazione di strutture lineari locali.

La stima dei parametri del modello mediante massima verosimiglianza sarà discussa nel seguito, nel contesto della più ampia famiglia dei *Parsimonious Gaussian Mixture Models* (PGMM), che includono l'MFA come caso particolare.

2.2.2 Mixtures of Probabilistic Principal Component Analyzers

Un'altra importante declinazione per la riduzione della dimensionalità è stata introdotta da Tipping e Bishop (1999a) con le *Mixtures of Probabilistic Principal Component Analyzers* (MPPCA). Come suggerisce la nomenclatura, tale approccio modella l'eterogeneità della popolazione attraverso una mistura in cui le singole componenti sono definite *Probabilistic Principal Component Analyzers* (PPCA).

Condizionatamente all'appartenenza al *cluster* g (ovvero $z_{ig} = 1$), l'equazione generativa del modello PPCA coincide con la medesima struttura dell'Analisi Fattoriale e delle MFA descritte in (2.2). La differenza fondamentale che separa il PPCA dalla classica Analisi Fattoriale risiede interamente nella specificazione della covarianza degli errori idiosincratici. Mentre l'Analisi Fattoriale utilizza una matrice diagonale libera, il modello PPCA assume che l'errore locale sia strettamente isotropico o sferico:

$$\Psi_g = \psi_g I_p,$$

dove ψ_g è un parametro scalare e I_p è la matrice identità di dimensione p. Questo vincolo implica che la varianza residua non spiegata dai fattori sia identica in tutte le direzioni dello spazio osservato. La matrice di covarianza marginale per il *cluster* g assume quindi la forma:

$$\Sigma_g = \Lambda_g \Lambda_g' + \psi_g I_p.$$

È esattamente questa assunzione di isotropia dell'errore a creare il collegamento teorico con l'Analisi delle Componenti Principali (PCA) classica. Tale legame diventa molto più intuitivo se analizzato da una prospettiva geometrica. Nella PCA tradizionale, metodo puramente deterministico, i dati vengono proiettati su un sottospazio lineare individuato dagli autovettori associati ai maggiori autovalori della matrice di covarianza campionaria. Nel modello PPCA, il vincolo di isotropia ($\Psi_g = \psi_g I_p$) equivale ad assumere che l'incertezza attorno al sottospazio dei fattori formi una "sfera" perfetta. Poiché questo rumore sferico non distorce lo spazio verso alcuna direzione privilegiata, l'unico modo che il modello ha per massimizzare la verosimiglianza è allineare il sottospazio latente esattamente lungo le direzioni di massima varianza dei dati. Di conseguenza, le colonne della matrice dei caricamenti Λ_g finiscono per generare lo stesso identico sottospazio individuato dai primi q autovettori della PCA.

In sintesi, il PPCA conferisce una rigorosa base probabilistica alla PCA classica, sostituendo la proiezione ortogonale rigida con una proiezione probabilistica. Da un punto di vista analitico, tale proiezione si concretizza nel calcolo del valore atteso dei fattori latenti condizionatamente ai dati osservati, che rappresenta l'analogo probabilistico dei punteggi delle componenti principali nella PCA classica. Poiché

la distribuzione congiunta dell'osservazione y_i e del fattore latente f_{ig} (all'interno del gruppo g) è gaussiana, dalle proprietà della distribuzione normale multivariata e sfruttando l'identità di Woodbury, si ottiene:

$$\begin{aligned}\mathbb{E}[f_{ig} \mid y_i, z_{ig} = 1] &= \Lambda_g'(\Lambda_g\Lambda_g' + \psi_g I_p)^{-1}(y_i - \mu_g) \\ &= M_g^{-1}\Lambda_g'(y_i - \mu_g),\end{aligned}$$

dove $M_g = (\Lambda_g'\Lambda_g + \psi_g I_q)$ ha dimensione ridotta $q \times q$.

L'inclusione del termine di rumore ψ_g nella matrice M_g agisce come un fattore di regolarizzazione. In termini geometrici, l'equazione stima la posizione latente "attirando" il dato osservato y_i verso l'origine del sottospazio in misura direttamente proporzionale al livello di rumore.

Questo legame teorico si salda definitivamente nel caso limite in cui la varianza del rumore tenda a zero ($\psi_g \to 0$): la matrice M_g si riduce a $\Lambda_g'\Lambda_g$ e la sfera di incertezza si restringe fino a svanire. In questa condizione asintotica, l'operatore $M_g^{-1}\Lambda_g'$ collassa esattamente nell'operatore di proiezione ortogonale $(\Lambda_g'\Lambda_g)^{-1}\Lambda_g'$, e il valore atteso recupera formalmente la proiezione rigida sui componenti principali tipica della PCA standard. Per una trattazione dettagliata delle proprietà teoriche del modello PPCA e delle stime di massima verosimiglianza si rimanda a Tipping e Bishop (1999b).

Rispetto al modello MFA, il vantaggio principale del MPPCA risiede nella maggiore parsimonia parametrica. Vincolando la matrice di unicità ad essere proporzionale all'identità, il numero di parametri necessari per modellare l'errore idiosincratico in ciascun *cluster* si riduce da p (gli elementi diagonali nel caso MFA) a un singolo parametro scalare ψ_g. Tale restrizione risulta particolarmente efficace in contesti ad alta dimensionalità, in cui si ricerca una rappresentazione estremamente compatta della struttura di covarianza; tuttavia, viene ottenuta al prezzo di una minore flessibilità nella rappresentazione delle dipendenze specifiche tra le variabili osservate, che risultano interamente attribuite alla struttura fattoriale.

2.2.3 L'evoluzione della Parsimonia: i modelli PGMM ed EPGMM

Al fine di sistematizzare i diversi modelli proposti in letteratura e superare il problema dell'iper-parametrizzazione, McNicholas e Murphy (2008) hanno formalizzato la famiglia dei *Parsimonious Gaussian Mixture Models (PGMM)*. L'imposizione strategica di vincoli di uguaglianza sulle matrici Λ_g e Ψ_g permette di navigare in modo continuo tra modelli molto flessibili e modelli estremamente parsimoniosi, applicando tre vincoli binari:

1. Pesi fattoriali: comuni tra le classi ($\Lambda_g = \Lambda$) oppure liberi.
2. Varianza degli errori: comune tra le classi ($\Psi_g = \Psi$) oppure libera.
3. Isotropia degli errori: sferica ($\Psi_g = \psi_g I_p$) oppure unicamente diagonale.

La combinazione di queste restrizioni genera una famiglia di 8 modelli, identificati da una nomenclatura a tre lettere (C = *Constrained*, U = *Unconstrained*), dove

Tabella 2.1: Famiglia dei modelli PGMM (McNicholas e Murphy, 2008)

Modello	**Pesi Fattoriali (Λ_g)**	**Errore (Ψ_g)**	**Isotropia**	**N. Parametri di Covarianza**
CCC	Vincolati (Λ)	Vincolati (Ψ)	Sferici (ψI_p)	$[pq - q(q-1)/2] + 1$
CCU	Vincolati (Λ)	Vincolati (Ψ)	Diagonali	$[pq - q(q-1)/2] + p$
CUC	Vincolati (Λ)	Liberi (Ψ_g)	Sferici ($\psi_g I_p$)	$[pq - q(q-1)/2] + G$
CUU	Vincolati (Λ)	Liberi (Ψ_g)	Diagonali	$[pq - q(q-1)/2] + Gp$
UCC	Liberi (Λ_g)	Vincolati (Ψ)	Sferici (ψI_p)	$G[pq - q(q-1)/2] + 1$
UCU	Liberi (Λ_g)	Vincolati (Ψ)	Diagonali	$G[pq - q(q-1)/2] + p$
UUC	Liberi (Λ_g)	Liberi (Ψ_g)	Sferici ($\psi_g I_p$)	$G[pq - q(q-1)/2] + G$
UUU	Liberi (Λ_g)	Liberi (Ψ_g)	Diagonali	$G[pq - q(q-1)/2] + Gp$

ciascuna posizione descrive rispettivamente i tre vincoli imposti sui caricamenti fattoriali, sulle unicità e sulla struttura isotropica degli errori. Da un punto di vista geometrico, tali vincoli controllano il grado di eterogeneità ammesso tra i *cluster* sia nella struttura dei sottospazi latenti sia nella dispersione residua delle osservazioni attorno ad essi. Come possiamo notare dalla Tabella 2.1 che ne riassume le caratteristiche, il modello UUU corrisponde alla formulazione MFA di McLachlan e Peel, il modello UCU alla MFA originaria di Ghahramani, e il modello UUC al modello MPPCA.

Sebbene i PGMM comprendano già numerose configurazioni della struttura di covarianza, non rappresentano il punto finale di questo percorso evolutivo. Nel tentativo di ampliare ulteriormente lo spettro dei compromessi possibili tra flessibilità e parsimonia parametrica, McNicholas e Murphy (2010) hanno successivamente introdotto la famiglia degli *Expanded Parsimonious Gaussian Mixture Models* (EPGMM). Ispirandosi alla decomposizione spettrale delle matrici di covarianza utilizzata nei modelli gaussiani parsimoniosi, gli autori hanno scomposto la matrice di errore diagonale ponendo:

$$\Psi_g = \omega_g \Delta_g,$$

dove $\omega_g \in \mathbb{R}^+$ è uno scalare che controlla il volume globale del rumore per il gruppo g, e Δ_g è una matrice diagonale con determinante pari a 1 ($|\Delta_g| = 1$) che ne controlla la forma, ovvero definisce le esatte proporzioni con cui il rumore globale si distribuisce tra le p variabili osservate, senza alterarne il volume complessivo. Vincolando in modo indipendente i pesi fattoriali Λ_g, i volumi ω_g e le matrici di forma Δ_g, la famiglia si espande a 12 modelli (Tabella 2.2) ottenendo una gamma più ampia di compromessi tra flessibilità e parsimonia parametrica.

2.2.4 Stima dei parametri

La stima dei parametri per le famiglie PGMM ed EPGMM richiede di formalizzare il problema in termini di verosimiglianza dei dati completi $\{Y, F, Z\}$. Al fine di fornire una trattazione chiara, nel prosieguo verrà esplicitata la derivazione analitica dell'algoritmo per il modello UUU (con parametri liberi e indipendenti per ciascun gruppo), il quale rappresenta la configurazione più generale. Tale formulazione deve necessariamente tenere conto della duplice natura dei dati inosservati che caratterizza questi modelli: le etichette di classe z_{ig} e i fattori latenti f_{ig}. Posto il vettore dei parametri $\Omega = \{\pi_g, \mu_g, \Lambda_g, \Psi_g\}$, la funzione di verosimiglianza completa si ottiene combinando la densità marginale delle etichette, la densità marginale dei fattori e

Tabella 2.2: Famiglia dei modelli EPGMM (McNicholas e Murphy, 2010)

Modello	**Pesi** (Λ_g)	**Volume** (ω_g)	**Forma** (Δ_g)	**Struttura Covarianza Errore**
CCC	Λ	ω	I_p	$\Sigma_g = \Lambda\Lambda' + \omega I_p$
CCU	Λ	ω	Δ	$\Sigma_g = \Lambda\Lambda' + \omega\Delta$
CUC	Λ	ω_g	I_p	$\Sigma_g = \Lambda\Lambda' + \omega_g I_p$
CUU	Λ	ω_g	Δ_g	$\Sigma_g = \Lambda\Lambda' + \omega_g\Delta_g$
UCC	Λ_g	ω	I_p	$\Sigma_g = \Lambda_g\Lambda_g' + \omega I_p$
UCU	Λ_g	ω	Δ	$\Sigma_g = \Lambda_g\Lambda_g' + \omega\Delta$
UUC	Λ_g	ω_g	I_p	$\Sigma_g = \Lambda_g\Lambda_g' + \omega_g I_p$
UUU	Λ_g	ω_g	Δ_g	$\Sigma_g = \Lambda_g\Lambda_g' + \omega_g\Delta_g$
CCUC*	Λ	ω_g	Δ	$\Sigma_g = \Lambda\Lambda' + \omega_g\Delta$
CUCU*	Λ	ω	Δ_g	$\Sigma_g = \Lambda\Lambda' + \omega\Delta_g$
UCUC*	Λ_g	ω_g	Δ	$\Sigma_g = \Lambda_g\Lambda_g' + \omega_g\Delta$
UUCU*	Λ_g	ω	Δ_g	$\Sigma_g = \Lambda_g\Lambda_g' + \omega\Delta_g$

* Modelli aggiuntivi introdotti dalla scomposizione EPGMM rispetto ai PGMM classici.

la densità condizionata delle osservazioni:

$$\begin{aligned} L^c(\Omega \mid Y, F, Z) &= \prod_{i=1}^{n}\prod_{g=1}^{G}\left[p(z_{ig}=1)p(f_{ig})p(y_i \mid f_{ig}, z_{ig}=1)\right]^{z_{ig}} \\ &= \prod_{i=1}^{n}\prod_{g=1}^{G}\left[\pi_g\phi_q(f_{ig}; 0, I_q)\phi_p(y_i; \mu_g + \Lambda_g f_{ig}, \Psi_g)\right]^{z_{ig}} \end{aligned}$$

e la relativa funzione di log-verosimiglianza completa risulta:

$$\begin{aligned} \ell^c(\Omega \mid Y, F, Z) = \sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\Big[&\ln\pi_g + \ln\phi_q(f_{ig}; 0, I_q) \\ &+ \ln\phi_p(y_i; \mu_g + \Lambda_g f_{ig}, \Psi_g)\Big]. \end{aligned} \tag{2.3}$$

Come illustrato per i modelli precedenti, la procedura di stima mira a massimizzare il valore atteso di tale funzione, condizionato ai dati osservati e alle stime correnti:

$$Q(\Omega \mid \Omega^{(t)}) = \mathbb{E}\big[\ell^c(\Omega \mid Y, F, Z) \mid Y, \Omega^{(t)}\big].$$

Per gestire questa doppia struttura latente in modo organizzato e agevolare il calcolo analitico delle stime, McNicholas e Murphy (2008) adottano l'algoritmo *Alternating Expectation-Conditional Maximization* (Meng e Van Dyk, 1997). Tale approccio semplifica la procedura suddividendo ogni iterazione in due stadi e alternando la definizione di quali variabili debbano essere considerate come dati mancanti.

Stadio 1: Aggiornamento dei pesi della mistura e delle medie

Nel primo stadio, l'insieme dei dati mancanti è limitato esclusivamente alle etichette di appartenenza z_{ig}. Poiché i fattori latenti f_{ig} vengono momentaneamente integrati fuori dalla verosimiglianza, la distribuzione condizionata delle osservazioni collassa nella densità marginale $y_i \mid z_{ig} = 1 \sim \mathcal{N}_p(\mu_g, \Sigma_g^{(t)})$, dove $\Sigma_g^{(t)} = \Lambda_g^{(t)}\Lambda_g^{(t)\prime} + \Psi_g^{(t)}$.

La funzione obiettivo da massimizzare, denotata con Q_1, corrisponde al valore atteso della log-verosimiglianza rispetto alla distribuzione a posteriori di Z:

$$\begin{aligned}
Q_1(\Omega \mid \Omega^{(t)}) &= \mathbb{E}\big[\ell^c(\Omega \mid Y, Z) \mid Y, \Omega^{(t)}\big] \\
&= \mathbb{E}\Bigg[\sum_{i=1}^{n}\sum_{g=1}^{G} z_{ig}\big(\ln \pi_g + \ln \phi_p(y_i; \mu_g, \Sigma_g^{(t)})\big) \;\Bigg|\; Y, \Omega^{(t)}\Bigg] \\
&= \sum_{i=1}^{n}\sum_{g=1}^{G} \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t)}]\Bigg[\ln \pi_g - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_g^{(t)}| \\
&\qquad\qquad - \frac{1}{2}(y_i - \mu_g)'(\Sigma_g^{(t)})^{-1}(y_i - \mu_g)\Bigg] \\
&= C + \sum_{i=1}^{n}\sum_{g=1}^{G} \tau_{ig}^{(t)}\Bigg[\ln \pi_g - \frac{1}{2}\ln|\Sigma_g^{(t)}| \\
&\qquad\qquad - \frac{1}{2}(y_i - \mu_g)'(\Sigma_g^{(t)})^{-1}(y_i - \mu_g)\Bigg],
\end{aligned}$$

dove la costante C raggruppa i termini additivi indipendenti dai parametri oggetto di massimizzazione.

Nel Passo E, le quantità $\tau_{ig}^{(t)}$ calcolate nell'equazione precedente corrispondono alle consuete probabilità a posteriori:

$$\tau_{ig}^{(t)} = \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t)}] = \frac{\pi_g^{(t)}\phi_p(y_i; \mu_g^{(t)}, \Sigma_g^{(t)})}{\sum_{h=1}^{G}\pi_h^{(t)}\phi_p(y_i; \mu_h^{(t)}, \Sigma_h^{(t)})}.$$

Nel Passo CM, si massimizza l'aspettativa della log-verosimiglianza dei dati completi per aggiornare i pesi della mistura e i vettori delle medie:

$$\pi_g^{(t+1)} = \frac{N_g^{(t)}}{n}, \quad \mu_g^{(t+1)} = \frac{1}{N_g^{(t)}}\sum_{i=1}^{n}\tau_{ig}^{(t)} y_i, \tag{2.4}$$

dove $N_g^{(t)} = \sum_{i=1}^{n}\tau_{ig}^{(t)}$. Tali formule di aggiornamento risultano identiche per tutti i modelli della famiglia EPGMM.

Stadio 2: Aggiornamento dei parametri di covarianza

Nel secondo stadio, la definizione di dati incompleti viene ampliata, includendo oltre alle etichette di classe z_{ig} anche i fattori latenti f_{ig}. La funzione obiettivo da ottimizzare, denotata con Q_2, corrisponde al valore atteso della log-verosimiglianza rispetto alla distribuzione congiunta a posteriori di Z ed F, valutata mantenendo fissati i parametri π_g e μ_g ai valori appena aggiornati nel primo stadio secondo l'espressione (2.4) (denotando tale vettore intermedio con $\Omega^{(t+1/2)}$). Sostituendo l'espressione (2.3) della log-verosimiglianza completa e sfruttando la legge delle aspettative

iterate, si ottiene:

$$
\begin{aligned}
Q_2(\Omega \mid \Omega^{(t+1/2)}) &= \mathbb{E}\big[\ell^c(\Omega \mid Y, F, Z) \mid Y, \Omega^{(t+1/2)}\big] \\
&= \mathbb{E}\Bigg[\sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \Big(\ln \pi_g^{(t+1)} + \ln \phi_q(f_{ig}; 0, I_q) \\
&\quad + \ln \phi_p(y_i; \mu_g^{(t+1)} + \Lambda_g f_{ig}, \Psi_g) \Big) \;\Bigg|\; Y, \Omega^{(t+1/2)} \Bigg] \\
&= C_2 + \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t+1/2)}] \\
&\quad \times \mathbb{E}\Big[\ln \phi_p(y_i; \mu_g^{(t+1)} + \Lambda_g f_{ig}, \Psi_g) \;\Big|\; y_i, z_{ig} = 1, \Omega^{(t+1/2)} \Big] \\
&= C_2 + \sum_{i=1}^{n} \sum_{g=1}^{G} \tau_{ig}^{(t)} \Bigg[-\frac{1}{2} \ln |\Psi_g| \\
&\quad - \frac{1}{2} \mathbb{E}\Big[(y_i - \mu_g^{(t+1)} - \Lambda_g f_{ig})' \Psi_g^{-1} (y_i - \mu_g^{(t+1)} - \Lambda_g f_{ig}) \;\Big|\; y_i, z_{ig} = 1, \Omega^{(t+1/2)} \Big] \Bigg],
\end{aligned}
$$

dove C_2 rappresenta una costante che raggruppa tutti i termini additivi indipendenti dai parametri Λ_g e Ψ_g (tra cui i pesi della mistura aggiornati $\pi_g^{(t+1)}$, i termini della densità marginale dei fattori e le costanti della densità gaussiana).

Nel Passo E, si calcolano i momenti condizionati dei fattori latenti necessari per valutare analiticamente l'espressione di Q_2. Poiché la distribuzione congiunta di y_i e f_{ig} è gaussiana, sfruttando le proprietà della normale multivariata, si ottiene:

$$
\begin{aligned}
\mathbb{E}[f_{ig} \mid y_i, z_{ig} = 1, \Omega^{(t+1/2)}] &= \beta_g^{(t)} (y_i - \mu_g^{(t+1)}) \\
\mathbb{E}[f_{ig} f_{ig}' \mid y_i, z_{ig} = 1, \Omega^{(t+1/2)}] &= I_q - \beta_g^{(t)} \Lambda_g^{(t)} + \beta_g^{(t)} (y_i - \mu_g^{(t+1)})(y_i - \mu_g^{(t+1)})' \beta_g^{(t)'},
\end{aligned}
$$

dove $\beta_g^{(t)} = \Lambda_g^{(t)'} (\Lambda_g^{(t)} \Lambda_g^{(t)'} + \Psi_g^{(t)})^{-1}$ rappresenta l'operatore di regressione lineare ottimale che proietta le osservazioni nello spazio dei fattori latenti del gruppo g.

Nel Passo CM, per ricavare le stime di aggiornamento per Λ_g e Ψ_g, si ottimizza la funzione del valore atteso della log-verosimiglianza condizionata limitata al secondo stadio di dati incompleti (McNicholas e Murphy, 2008). Sviluppando il termine quadratico della densità normale e omettendo le costanti additive indipendenti dai parametri in esame, la funzione assume la forma:

$$
Q_2(\Lambda_g, \Psi_g) = C + \sum_{g=1}^{G} \frac{N_g^{(t)}}{2} \left[-\ln |\Psi_g| - \operatorname{tr}\left\{ \Psi_g^{-1} \left(S_g - 2\Lambda_g \beta_g^{(t)} S_g + \Lambda_g \Theta_g \Lambda_g' \right) \right\} \right],
$$

dove $S_g = \frac{1}{N_g^{(t)}} \sum_{i=1}^{n} \tau_{ig}^{(t)} (y_i - \mu_g^{(t+1)})(y_i - \mu_g^{(t+1)})'$ è la matrice di covarianza campionaria pesata del gruppo g, e $\Theta_g = I_q - \beta_g^{(t)} \Lambda_g^{(t)} + \beta_g^{(t)} S_g \beta_g^{(t)'}$.

Per il modello non vincolato UUU, le stime ottimali si ricavano azzerando le derivate prime della funzione Q_2. Derivando rispetto a Λ_g si ottiene:

$$
\frac{\partial Q_2}{\partial \Lambda_g} = \sum_{g=1}^{G} N_g^{(t)} \left(\Psi_g^{-1} S_g \beta_g^{(t)'} - \Psi_g^{-1} \Lambda_g \Theta_g \right) = 0.
$$

Pre-moltiplicando per Ψ_g e isolando la matrice dei pesi fattoriali, l'aggiornamento risulta:

$$\Lambda_g^{(t+1)}\Theta_g = S_g\beta_g^{(t)\prime} \implies \Lambda_g^{(t+1)} = S_g\beta_g^{(t)\prime}\Theta_g^{-1}. \tag{2.5}$$

Procedendo similmente, si deriva rispetto alla matrice Ψ_g^{-1}:

$$\frac{\partial Q_2}{\partial \Psi_g^{-1}} = \sum_{g=1}^{G} \frac{N_g^{(t)}}{2}\left[\Psi_g - \left(S_g - 2\Lambda_g^{(t+1)}\beta_g^{(t)}S_g + \Lambda_g^{(t+1)}\Theta_g\Lambda_g^{(t+1)\prime}\right)\right] = 0.$$

Risolvendo per Ψ_g e applicando l'operatore $\text{diag}\{\cdot\}$ per preservare l'assunzione di indipendenza locale delle unicità, si ricava:

$$\Psi_g^{(t+1)} = \text{diag}\left\{S_g - 2\Lambda_g^{(t+1)}\beta_g^{(t)}S_g + \Lambda_g^{(t+1)}\Theta_g\Lambda_g^{(t+1)\prime}\right\}.$$

Sostituendo l'identità $\Lambda_g^{(t+1)}\Theta_g = S_g\beta_g^{(t)\prime}$ all'interno della traccia, il termine quadratico si riduce esattamente a un contributo lineare, ottenendo la nota espressione di aggiornamento della matrice delle unicità:

$$\Psi_g^{(t+1)} = \text{diag}\left\{S_g - \Lambda_g^{(t+1)}\beta_g^{(t)}S_g\right\}.$$

Per i modelli con vincoli inter-classe, le espressioni di aggiornamento derivate per il caso generale UUU vengono opportunamente modificate imponendo le restrizioni strutturali previste su Λ_g e Ψ_g. In altre parole, il modello UUU funge da riferimento generale, mentre le altre specificazioni della famiglia PGMM si ottengono come casi particolari vincolati.

Ad esempio, nei modelli con pesi fattoriali comuni (come CCC, CUC, CCU e CUU), la matrice Λ_g viene sostituita da una matrice condivisa Λ. La formula di aggiornamento (2.5) viene sostituita da:

$$\Lambda^{(t+1)} = \left(\sum_{g=1}^{G} N_g^{(t)} S_g\beta_g^{(t)\prime}\right)\left(\sum_{g=1}^{G} N_g^{(t)}\Theta_g\right)^{-1}.$$

Analogamente, nei modelli con matrice degli errori comune (come CCC, CCU, UCC e UCU), si impone $\Psi_g = \Psi$, ottenendo una stima condivisa come combinazione pesata delle quantità specifiche di gruppo.

Nel caso di ulteriori vincoli, come l'isotropia ($\Psi_g = \psi_g I_p$ oppure $\Psi = \psi I_p$), l'aggiornamento viene ottenuto proiettando la stima generale sulla classe delle matrici ammissibili, ad esempio imponendo che tutti gli elementi diagonali coincidano.

In questo modo, l'intera famiglia dei modelli PGMM può essere stimata all'interno di un unico schema AECM, adattando opportunamente i passi di massimizzazione alle specifiche restrizioni imposte dal modello considerato. Per maggiori dettagli, si rimanda alla trattazione estesa presente nei lavori fondativi di McNicholas e Murphy (2008) e McNicholas e Murphy (2010).

L'iterazione tra lo Stadio 1 e lo Stadio 2 continua monotonamente fino alla convergenza della funzione di log-verosimiglianza. Poiché l'algoritmo AECM eredita le medesime criticità dell'EM standard in merito ai massimi locali, l'inizializzazione dei parametri viene gestita analogamente a quanto descritto nel Capitolo 2, ricorrendo a partizioni preliminari ottenute, ad esempio, tramite *K-means*.

Infine, come ampiamente discusso nel Capitolo 2 per i modelli base, la selezione della complessità ottimale per i modelli della famiglia PGMM ed EPGMM avviene tramite il BIC. In questo contesto multidimensionale, tale criterio determina simultaneamente tre elementi: il numero ottimale di classi latenti G, il numero di fattori q e la specifica struttura dei vincoli sulle matrici di covarianza. Conformemente alla prassi consolidata, verrà selezionata l'architettura che, ottimizzando il valore del BIC, garantisce il miglior compromesso tra adattamento ai dati e parsimonia parametrica.

2.3 L'approccio psicometrico

2.3.1 Il Principio dell'Invarianza di Misurazione

A differenza del filone statistico, l'approccio psicometrico allo sviluppo dei *Factor Mixture Models* è guidato da un'esigenza interpretativa radicalmente diversa: garantire il significato sostantivo dei fattori latenti. Nelle scienze sociali, in psicologia e in altre discipline, i fattori non sono semplici strumenti operativi volti alla riduzione dei parametri, ma rappresentano veri e propri costrutti teorici di interesse, quali ad esempio abilità cognitive o tratti di personalità. In questo contesto, l'obiettivo del FMM non è solo identificare i *cluster*, ma richiede che i fattori estratti mantengano una coerenza interpretativa tale da permettere il confronto tra le diverse sottopopolazioni rispetto ai tratti misurati.

Questa esigenza si traduce in una diversa strategia di trattamento del modello generale (2.1). Mentre il filone statistico normalizza la distribuzione dei fattori e lascia liberi i pesi Λ_g, l'approccio psicometrico mira a ricondurre l'eterogeneità tra le sottopopolazioni ai soli parametri della distribuzione latente (le medie α_g e le covarianze Φ_g dei fattori) mantenendo per quanto possibile invariante il modello di misura tra i gruppi. Se infatti i parametri di misura variassero liberamente tra le classi, non vi sarebbe alcuna garanzia che il costrutto latente individuato in un gruppo coincida con quello individuato in un altro, rendendo priva di significato qualsiasi comparazione. L'invarianza di misurazione (*measurement invariance*, Meredith 1993) assicura che, a parità di punteggio nel tratto latente, individui appartenenti a *cluster* diversi abbiano lo stesso valore atteso sulle variabili osservate. Essa si articola in quattro livelli.

1. **Invarianza configurazionale:** rappresenta il livello base e costituisce il modello di riferimento. Richiede che la struttura concettuale fondamentale del modello fattoriale sia la stessa in tutti i *cluster*. In termini pratici, questo significa che il numero di fattori latenti q deve essere identico tra i gruppi e che lo stesso insieme di variabili manifeste deve fungere da indicatore per i medesimi fattori. Matematicamente, si impone che la matrice dei pesi Λ_g presenti lo stesso *pattern* di elementi nulli e non nulli in tutte le classi, pur permettendo che il valore numerico esatto dei parametri liberi vari da un gruppo all'altro. Se questo requisito preliminare non è soddisfatto, significa che le diverse sottopopolazioni stanno concettualizzando il tratto latente in modo strutturalmente diverso, rendendo del tutto priva di significato qualsiasi successiva comparazione.

2. **Invarianza metrica (o debole):** richiede che la matrice dei pesi fattoriali sia identica tra le classi ($\Lambda_g = \Lambda$). Questo vincolo garantisce che la metrica, ovvero l'unità di misura del fattore latente, sia coerente in tutta la popolazione. Se Λ_g variasse, il significato stesso del fattore cambierebbe da un *cluster* all'altro, rendendo impossibile qualsiasi confronto diretto.

3. **Invarianza scalare (o forte):** oltre a Λ, impone l'uguaglianza delle intercette tra le classi ($\nu_g = \nu$). Sotto questo vincolo, il vettore delle medie osservate diventa $\mu_g = \nu + \Lambda\alpha_g$. Questa restrizione è cruciale nei FMM: implica che qualsiasi differenza osservata nelle medie dei *cluster* (μ_g) sia imputabile *esclusivamente* a reali differenze nelle medie dei tratti latenti (α_g). Senza l'invarianza scalare, le differenze di media potrebbero derivare da *bias* specifici del gruppo verso determinati indicatori, portando a identificare *cluster* spuri.

4. **Invarianza stretta:** aggiunge infine l'uguaglianza delle matrici di covarianza degli errori ($\Psi_g = \Psi$). Sotto questa ipotesi, l'intera differenza nella struttura di covarianza osservata tra i gruppi è da attribuirsi esclusivamente alle differenze nelle matrici di covarianza dei fattori latenti (Φ_g).

Come argomentano Lubke e Muthén (2005), nei *Factor Mixture Models* l'invarianza metrica e scalare rappresenta il requisito minimo affinché l'estrazione delle classi latenti abbia validità sostantiva.

Guardando a posteriori la famiglia dei modelli PGMM discussa nella sezione precedente (Tabella 2.1), risulta evidente come i modelli contrassegnati dalla lettera iniziale 'C' (come CUC o CCC), imponendo $\Lambda_g = \Lambda$, rispettino intrinsecamente il secondo livello di invarianza (metrica). Al contrario, modelli statistici flessibili come l'MFA originale o i modelli 'U' (es. UUC, UUU) non garantiscono l'invarianza, privilegiando la flessibilità modellistica rispetto all'interpretabilità psicometrica.

2.3.2 Heteroscedastic Factor Mixture Analysis

Come evidenziato nel paragrafo precedente, il requisito di comparabilità più stringente conduce a un modello in cui i parametri di misura sono comuni a tutte le sottopopolazioni. Specializzando il modello generale (2.1) con $\Lambda_g = \Lambda$ e $\Psi_g = \Psi$, e assumendo l'assenza di covariate, l'equazione per un'osservazione y_i appartenente alla classe g diviene

$$y_i \mid (z_{ig} = 1) = \nu_g + \Lambda f_{ig} + \epsilon_{ig}, \qquad f_{ig} \mid (z_{ig} = 1) \sim \mathcal{N}_q(\alpha_g, \Phi_g), \qquad \epsilon_{ig} \sim \mathcal{N}_p(0, \Psi),$$

da cui discendono media e covarianza osservate della classe g:

$$\mu_g = \nu_g + \Lambda\alpha_g, \qquad \Sigma_g = \Lambda\Phi_g\Lambda' + \Psi.$$

Poiché i pesi Λ e gli errori Ψ sono comuni a tutte le classi, ogni differenza osservata tra i gruppi è attribuibile esclusivamente ai parametri latenti α_g e Φ_g. Il modello che formalizza analiticamente questo requisito è l'*Heteroscedastic Factor Mixture Analysis* (HFMA, Montanari e Viroli 2010), che ne adotta la versione su

dati centrati. Sia dunque y_i un vettore p-dimensionale di variabili osservate centrate; l'equazione strutturale di base dell'HFMA è un classico modello fattoriale:

$$y_i = \Lambda f_i + \epsilon_i, \tag{2.6}$$

dove Λ è la matrice dei pesi fattoriali di dimensione $p \times q$, globale e invariante per tutta la popolazione, ed $\epsilon_i \sim \mathcal{N}_p(0, \Psi)$ è il vettore degli errori specifici, anch'esso caratterizzato da una matrice di covarianza diagonale Ψ omogenea per tutti gli individui. L'eterogeneità della popolazione è modellata esclusivamente attraverso la distribuzione dei fattori latenti f_i che seguono una mistura di G componenti gaussiane:

$$f_i \sim \sum_{g=1}^{G} \pi_g \mathcal{N}_q(\alpha_g, \Phi_g),$$

dove π_g è il peso della g-esima componente (con $\sum_g \pi_g = 1$), α_g è il vettore delle medie fattoriali del *cluster* g, e Φ_g è la relativa matrice di covarianza dei fattori $(q \times q)$, la quale è libera di variare catturando complesse strutture di dipendenza latente. In questo modo, tutta la variabilità tra gruppi è trasferita nello spazio latente, mentre il modello di misura rimane invariato.

Sotto questa specificazione, la distribuzione marginale delle osservazioni y_i si ottiene integrando rispetto alla distribuzione dei fattori, risultando in una mistura di gaussiane multivariate nello spazio osservato:

$$p(y_i) = \sum_{g=1}^{G} \pi_g \phi_p(y_i; \Lambda\alpha_g, \Lambda\Phi_g\Lambda' + \Psi). \tag{2.7}$$

Da questa equazione emerge chiaramente che l'HFMA è un modello di *clustering* a tutti gli effetti, ma la cui matrice di covarianza specifica per il singolo gruppo, $\Sigma_g = \Lambda\Phi_g\Lambda' + \Psi$, è altamente strutturata e parsimoniosa.

Per risolvere l'indeterminazione del modello generale discussa nella Sezione 2.1, l'HFMA impone che la distribuzione marginale dei fattori abbia media nulla e covarianza identità:

$$\begin{aligned} E[f] &= \sum_{g=1}^{G} \pi_g \alpha_g = 0 \\ Var(f) &= \sum_{g=1}^{G} \pi_g (\Phi_g + \alpha_g \alpha_g') = I_q. \end{aligned} \tag{2.8}$$

Nella prassi estimativa, l'algoritmo EM converge a una soluzione non identificata rispetto a rotazioni e traslazioni dello spazio latente, che viene successivamente standardizzata mediante trasformazioni ortogonali per soddisfare le condizioni di identificabilità.

2.3.3 Stima dei parametri

La stima di massima verosimiglianza del modello HFMA, come tutti i modelli a mistura fattoriale presentano due livelli di dati mancanti: l'etichetta di classe latente z_{ig} e i fattori latenti continui f_i.

Sia $\Omega = \{\pi_g, \alpha_g, \Phi_g, \Lambda, \Psi\}$ l'intero set di parametri da stimare. Sfruttando la struttura gerarchica del modello, la funzione di verosimiglianza dei dati completi si scompone nel prodotto della densità dei dati osservati (condizionata ai soli fattori) per la probabilità congiunta dei fattori e delle classi latenti. Formalmente:

$$\begin{aligned} L^c(\Omega \mid Y, F, Z) &= \prod_{i=1}^{n} \left\{ p(y_i \mid f_i) \prod_{g=1}^{G} [p(z_{ig} = 1) p(f_i \mid z_{ig} = 1)]^{z_{ig}} \right\} \\ &= \prod_{i=1}^{n} \left\{ \phi_p(y_i; \Lambda f_i, \Psi) \prod_{g=1}^{G} [\pi_g \phi_q(f_i; \alpha_g, \Phi_g)]^{z_{ig}} \right\} \end{aligned}$$

e la rispettiva funzione di log-verosimiglianza:

$$\ell^c(\Omega \mid Y, F, Z) = \sum_{i=1}^{n} \ln \phi_p(y_i; \Lambda f_i, \Psi) + \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \Big[\ln \pi_g + \ln \phi_q(f_i; \alpha_g, \Phi_g) \Big]. \quad (2.9)$$

In questa formulazione, la componente relativa alla densità condizionata di y_i è esterna alla sommatoria sui *cluster* g, poiché la matrice dei pesi Λ e la varianza d'errore Ψ sono globali e non dipendono dall'appartenenza di classe. L'algoritmo *Expectation-Maximization* itererà sui seguenti passi.

Il Passo E. Nel passo E, si definisce la funzione obiettivo $Q(\Omega \mid \Omega^{(t)})$, ovvero il valore atteso della log-verosimiglianza dei dati completi rispetto alla distribuzione congiunta a posteriori delle variabili latenti (Z, F), condizionata ai dati osservati Y e alle stime correnti $\Omega^{(t)}$:

$$Q(\Omega \mid \Omega^{(t)}) = \mathbb{E}\big[\ell^c(\Omega \mid Y, F, Z) \mid Y, \Omega^{(t)}\big].$$

Sostituendo l'espressione (2.9) della log-verosimiglianza completa e sfruttando la linearità dell'operatore valore atteso, la funzione si scompone additivamente in tre componenti indipendenti da ottimizzare separatamente:

$$Q(\Omega \mid \Omega^{(t)}) = Q_\pi + Q_{\alpha,\Phi} + Q_{\Lambda,\Psi},$$

dove, applicando la legge delle aspettative iterate, si ottiene:

$$\begin{aligned} Q_\pi &= \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t)}] \ln \pi_g \\ Q_{\alpha,\Phi} &= \sum_{i=1}^{n} \sum_{g=1}^{G} \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t)}] \, \mathbb{E}\Big[\ln \phi_q(f_i; \alpha_g, \Phi_g) \;\Big|\; y_i, z_{ig} = 1, \Omega^{(t)} \Big] \\ Q_{\Lambda,\Psi} &= \sum_{i=1}^{n} \mathbb{E}\Big[\ln \phi_p(y_i; \Lambda f_i, \Psi) \;\Big|\; y_i, \Omega^{(t)} \Big]. \end{aligned}$$

Per valutare analiticamente tali espressioni, è necessario calcolare i momenti condizionati di classe e di fattore. Innanzitutto, si calcola la probabilità a posteriori che l'osservazione i appartenga al *cluster* g (necessaria per Q_π e $Q_{\alpha,\Phi}$):

$$\tau_{ig}^{(t)} = \mathbb{E}[z_{ig} \mid y_i, \Omega^{(t)}] = \frac{\pi_g^{(t)} \phi_p(y_i; \Lambda^{(t)} \alpha_g^{(t)}, \Lambda^{(t)} \Phi_g^{(t)} \Lambda^{(t)\prime} + \Psi^{(t)})}{\sum_{h=1}^{G} \pi_h^{(t)} \phi_p(y_i; \Lambda^{(t)} \alpha_h^{(t)}, \Lambda^{(t)} \Phi_h^{(t)} \Lambda^{(t)\prime} + \Psi^{(t)})}. \quad (2.10)$$

In secondo luogo, applicando il Teorema di Bayes alla congiunta di (y_i, f_i), si deriva la distribuzione dei fattori condizionata sia all'osservazione y_i che all'appartenenza al *cluster* g. Tale distribuzione è anch'essa gaussiana: $f_i \mid (y_i, z_{ig} = 1) \sim \mathcal{N}_q(\rho_{ig}^{(t)}, \xi_g^{(t)})$, con matrice di covarianza e vettore delle medie definiti come:

$$\xi_g^{(t)} = \left(\Lambda^{(t)\prime}\Psi^{(t)^{-1}}\Lambda^{(t)} + \Phi_g^{(t)^{-1}}\right)^{-1}$$
$$\rho_{ig}^{(t)} = \xi_g^{(t)}\left(\Lambda^{(t)\prime}\Psi^{(t)^{-1}}y_i + \Phi_g^{(t)^{-1}}\alpha_g^{(t)}\right).$$

Da questa distribuzione, estraiamo i primi due momenti condizionati necessari per la valutazione della componente $Q_{\alpha,\Phi}$:

$$\mathbb{E}[f_i \mid y_i, z_{ig} = 1] = \rho_{ig}^{(t)}$$
$$\mathbb{E}[f_i f_i' \mid y_i, z_{ig} = 1] = \xi_g^{(t)} + \rho_{ig}^{(t)}\rho_{ig}^{(t)\prime}.$$

Infine, poiché la componente $Q_{\Lambda,\Psi}$ dipende dalla densità incondizionata rispetto alla classe, occorrono i momenti marginali dei fattori. Applicando la legge dell'aspettativa totale, si ponderano i momenti condizionati appena trovati per le probabilità a posteriori $\tau_{ig}^{(t)}$:

$$\mathbb{E}[f_i \mid y_i] = \sum_{g=1}^{G}\tau_{ig}^{(t)}\rho_{ig}^{(t)}$$
$$\mathbb{E}[f_i f_i' \mid y_i] = \sum_{g=1}^{G}\tau_{ig}^{(t)}\left(\xi_g^{(t)} + \rho_{ig}^{(t)}\rho_{ig}^{(t)\prime}\right).$$

Il Passo M. Nel passo M, si aggiornano i parametri massimizzando la funzione Q, procedura che equivale ad azzerare le sue derivate parziali rispetto a ciascun gruppo di parametri. L'aggiornamento dei pesi si ottiene massimizzando Q_π sotto il vincolo $\sum_g \pi_g = 1$, risultando in:

$$\pi_g^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\tau_{ig}^{(t)}.$$

Per i parametri locali del *cluster* (medie e covarianze dei fattori), si ottimizza la componente $Q_{\alpha,\Phi}$, che sviluppando il logaritmo della densità gaussiana e trascurando le costanti additive assume la forma:

$$Q_{\alpha,\Phi} = \sum_{i=1}^{n}\sum_{g=1}^{G}\tau_{ig}^{(t)}\left[-\frac{1}{2}\ln|\Phi_g| - \frac{1}{2}\mathrm{tr}\left\{\Phi_g^{-1}\mathbb{E}\left[(f_i - \alpha_g)(f_i - \alpha_g)' \mid y_i, z_{ig} = 1\right]\right\}\right].$$

Derivando rispetto al vettore delle medie fattoriali α_g e ponendo a zero, si ottiene:

$$\frac{\partial Q_{\alpha,\Phi}}{\partial \alpha_g} = \sum_{i=1}^{n}\tau_{ig}^{(t)}\Phi_g^{-1}\Big(\mathbb{E}[f_i \mid y_i, z_{ig} = 1] - \alpha_g\Big) = 0,$$

da cui si ricava l'equazione di aggiornamento:

$$\alpha_g^{(t+1)} = \frac{\sum_{i=1}^{n}\tau_{ig}^{(t)}\mathbb{E}[f_i \mid y_i, z_{ig} = 1]}{\sum_{i=1}^{n}\tau_{ig}^{(t)}}.$$

Similmente, derivando rispetto alla matrice di precisione latente Φ_g^{-1} e ponendo a zero:

$$\frac{\partial Q_{\alpha,\Phi}}{\partial \Phi_g^{-1}} = \frac{1}{2}\sum_{i=1}^{n} \tau_{ig}^{(t)} \left[\Phi_g - \mathbb{E}\left[(f_i - \alpha_g^{(t+1)})(f_i - \alpha_g^{(t+1)})' \mid y_i, z_{ig} = 1\right]\right] = 0.$$

Sostituendo i momenti condizionati, si giunge all'aggiornamento delle covarianze dei fattori:

$$\Phi_g^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{ig}^{(t)} \left(\mathbb{E}[f_i f_i' \mid y_i, z_{ig} = 1] - \alpha_g^{(t+1)} \alpha_g^{(t+1)'}\right)}{\sum_{i=1}^{n} \tau_{ig}^{(t)}}.$$

I parametri globali di misurazione vengono, invece, aggiornati massimizzando la componente marginale $Q_{\Lambda,\Psi}$:

$$Q_{\Lambda,\Psi} = -\frac{n}{2}\ln|\Psi| - \frac{1}{2}\sum_{i=1}^{n} \operatorname{tr}\left\{\Psi^{-1}\mathbb{E}\left[(y_i - \Lambda f_i)(y_i - \Lambda f_i)' \mid y_i\right]\right\}.$$

Calcolando la derivata rispetto alla matrice dei pesi fattoriali Λ:

$$\frac{\partial Q_{\Lambda,\Psi}}{\partial \Lambda} = \sum_{i=1}^{n} \left(\Psi^{-1} y_i \mathbb{E}[f_i' \mid y_i] - \Psi^{-1}\Lambda \mathbb{E}[f_i f_i' \mid y_i]\right) = 0.$$

Pre-moltiplicando per Ψ e isolando Λ, l'equazione fornisce lo stimatore:

$$\Lambda^{(t+1)} = \left(\sum_{i=1}^{n} y_i \mathbb{E}[f_i' \mid y_i]\right)\left(\sum_{i=1}^{n} \mathbb{E}[f_i f_i' \mid y_i]\right)^{-1}.$$

Infine, derivando rispetto a Ψ^{-1}:

$$\begin{aligned}\frac{\partial Q_{\Lambda,\Psi}}{\partial \Psi^{-1}} = \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n} \Big(&y_i y_i' - \Lambda^{(t+1)}\mathbb{E}[f_i \mid y_i] y_i' \\ &- y_i \mathbb{E}[f_i' \mid y_i]\Lambda^{(t+1)'} + \Lambda^{(t+1)}\mathbb{E}[f_i f_i' \mid y_i]\Lambda^{(t+1)'}\Big) = 0.\end{aligned}$$

Semplificando i termini incrociati e applicando l'operatore $\operatorname{diag}\{\cdot\}$ per rispettare il vincolo di indipendenza locale delle unicità, si ottiene:

$$\Psi^{(t+1)} = \operatorname{diag}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(y_i y_i' - \Lambda^{(t+1)}\mathbb{E}[f_i \mid y_i] y_i'\right)\right\}.$$

2.4 Il modello MCFA e il confronto con l'HFMA

Come evidenziato nella Sezione 2.1, il modello generale di riferimento (2.1) ammette, nella sua formulazione più flessibile, parametri di misura specifici per ciascuna classe latente. In particolare, la matrice dei pesi fattoriali Λ_g può variare liberamente tra i gruppi, consentendo a ciascuna componente della mistura di descrivere una propria struttura locale di dipendenza tra le variabili osservate. Tuttavia, questa flessibilità

introduce anche una criticità rilevante: se ogni *cluster* possiede una propria matrice Λ_g, allora ciascun gruppo viene rappresentato rispetto a un diverso sistema di coordinate latenti.

Tale aspetto diventa particolarmente problematico in contesti applicativi caratterizzati da elevata dimensionalità, come ad esempio l'analisi di dati di espressione genica, dove spesso il numero di variabili p supera ampiamente il numero di osservazioni n. In questi casi, la stima di una matrice dei pesi fattoriali distinta per ciascuna componente aumenta sensibilmente il numero di parametri liberi, aggravando i problemi di instabilità numerica e di possibile singolarità delle matrici di covarianza locali. Inoltre, l'assenza di uno spazio fattoriale comune rende difficile confrontare direttamente i punteggi latenti di individui appartenenti a sottopopolazioni diverse, poiché tali punteggi non sono espressi rispetto allo stesso sistema di riferimento. Di conseguenza, viene meno la possibilità di rappresentare l'intero insieme dei dati in un unico spazio a bassa dimensionalità, utile per la visualizzazione e l'esplorazione congiunta dei *cluster*.

Per superare questi limiti, è necessario imporre restrizioni alla formulazione generale. L'esigenza di definire uno spazio latente condiviso in cui rappresentare tutte le osservazioni si traduce matematicamente nell'assunzione di una struttura di misura comune tra le classi. In particolare, vincolando la matrice dei pesi fattoriali e la matrice di covarianza degli errori a essere costanti tra i gruppi,

$$\Lambda_g = \Lambda, \qquad \Psi_g = \Psi,$$

e lavorando su dati centrati, così da porre a zero il termine di intercetta, l'equazione (2.1) si specializza nella forma

$$\begin{aligned} y_i \mid (z_{ig} = 1) &= \Lambda f_{ig} + \epsilon_{ig}, \\ f_{ig} \mid (z_{ig} = 1) &\sim \mathcal{N}_q(\alpha_g, \Phi_g), \\ \epsilon_{ig} &\sim \mathcal{N}_p(0, \Psi). \end{aligned} \tag{2.11}$$

Sotto questo regime di vincoli, il modello di misura rimane identico per tutte le classi, mentre l'eterogeneità tra le sottopopolazioni viene ricondotta alla distribuzione dei fattori latenti. Integrando rispetto ai fattori, il vettore delle medie e la matrice di covarianza per la classe g assumono la forma

$$\mu_g = \Lambda\alpha_g, \qquad \Sigma_g = \Lambda\Phi_g\Lambda' + \Psi. \tag{2.12}$$

Il modello che formalizza questa specifica parametrizzazione nel filone statistico è il *Mixtures of Common Factor Analyzers* (MCFA), introdotto da Baek *et al.* (2010). La caratteristica distintiva dell'MCFA consiste dunque nell'imposizione di una matrice dei pesi fattoriali comune, che consente di ottenere un unico spazio latente condiviso per tutte le componenti della mistura. In questo senso, l'MCFA conserva la capacità dei modelli a mistura di descrivere sottopopolazioni eterogenee, ma evita che ciascun gruppo venga rappresentato attraverso un sistema fattoriale indipendente. La riduzione del numero di parametri e la possibilità di visualizzare i dati in uno spazio a bassa dimensionalità costituiscono quindi due motivazioni centrali del modello.

Osservando attentamente le equazioni appena derivate, emerge un risultato strutturale notevole: l'MCFA presenta una coincidenza formale nella struttura marginale con l'HFMA, discusso nella Sezione 2.3.2. Entrambi i modelli conducono alla medesima struttura di covarianza marginale $\Sigma_g = \Lambda\Phi_g\Lambda' + \Psi$.

Questa coincidenza formale non implica però una piena sovrapposizione concettuale. Al contrario, MCFA e HFMA mostrano come una medesima specializzazione del modello generale possa derivare da esigenze metodologiche differenti.

- L'esigenza statistico-computazionale (MCFA): in scenari con centinaia di variabili, l'uso di una matrice globale Λ per garantire stabilità e agevolare la rappresentazione grafica. Un Λ condiviso riduce i parametri ed evita matrici singolari, fornendo al contempo un unico spazio di coordinate in cui mappare e visualizzare facilmente tutti i *cluster*.

- L'esigenza psicometrica (HFMA): in ambito psicologico o sociale, l'uso di matrici Λ e Ψ globali risulta essere un vincolo teorico rigoroso. Serve a garantire l'invarianza di misurazione: solo se il modello di misura funziona allo stesso modo per tutti, possiamo essere certi che le differenze trovate tra i *cluster* riflettano veri scostamenti nei tratti latenti misurati (α_g e Φ_g) e non semplici distorsioni degli indicatori.

Infine, questa divergenza di prospettive si riflette in modo lampante nella strategia adottata per garantire l'identificabilità del modello. Nell'HFMA, come visto in (2.8), il modello viene identificato imponendo che la distribuzione marginale dei fattori per l'intera popolazione abbia media nulla e matrice di covarianza identità. Questa è una scelta tipicamente psicometrica: garantisce che il tratto latente misurato assuma la forma di un classico "punteggio standardizzato" globale, permettendo di interpretare le deviazioni dei singoli *cluster* rispetto alla media generale. Al contrario, nell'MCFA, per garantire l'unicità della soluzione, una volta raggiunta la convergenza dell'algoritmo si applica la scomposizione di Cholesky alla matrice dei pesi Λ. Trasformando Λ in una matrice ortonormale e aggiustando di conseguenza i parametri latenti α_g e Φ_g, il modello fissa univocamente la rotazione dello spazio geometrico. Anche in questo caso, la scelta metodologica ribadisce la natura del modello: nell'MCFA i fattori non richiedono una standardizzazione globale per essere interpretati, ma necessitano solo di un ancoraggio matematico per garantire l'unicità della soluzione computazionale e la successiva rappresentazione grafica.

Capitolo 3

Studio di simulazione

3.1 Obiettivi dello studio

Alla luce del quadro teorico delineato nei capitoli precedenti, il presente capitolo illustra uno studio di simulazione Monte Carlo finalizzato a valutare l'efficacia e la robustezza dei principali approcci discussi nell'individuazione della struttura latente sottostante ai dati. In particolare, lo studio intende confrontare il comportamento dell'Analisi Fattoriale classica, dei Modelli a Mistura Gaussiana (GMM), dei Modelli a Mistura Gaussiana Parsimoniosi (PGMM) e dell'Analisi a Mistura Fattoriale Eteroschedastica (HFMA) al variare di diverse condizioni sperimentali. In particolare, le condizioni considerate riguardano:

- natura del segnale latente, articolata in tre configurazioni principali:
  - segnale continuo e categoriale (LCC), nel quale coesistono una struttura fattoriale latente continua e una componente categoriale di appartenenza a gruppo;
  - segnale esclusivamente continuo (LCo), nel quale la variabilità osservata è generata da una struttura fattoriale continua latente comune, in assenza di gruppi categoriali;
  - segnale esclusivamente categoriale (LCa), nel quale la variabilità osservata è determinata dalla presenza di gruppi latenti, senza una struttura fattoriale continua condivisa a bassa dimensionalità.
- forza del segnale, considerata a due livelli, debole e forte. Questa condizione permette di distinguere scenari nei quali la struttura latente è chiaramente identificabile da scenari più complessi, nei quali il segnale risulta meno netto o maggiormente alterato dal rumore;
- dimensionalità dello spazio osservato (p), considerata a due livelli, bassa e alta dimensionalità. Tale aspetto consente di valutare l'effetto dell'aumento del numero di variabili osservate sulla complessità parametrica, sulla stabilità della stima e sulla capacità dei modelli di recuperare la struttura generatrice;
- numerosità campionaria (n), anch'essa articolata in due livelli, campioni piccoli e campioni ampi. Questa condizione permette di analizzare il comportamento

dei modelli in funzione dell'informazione disponibile e della stabilità delle stime in campioni finiti.

Lo studio è quindi costruito per rispondere a due domande principali. In primo luogo, si vuole verificare se i modelli riescano a recuperare correttamente la struttura latente quando il processo generatore è coerente con le loro assunzioni teoriche. In secondo luogo, si intende valutare la loro robustezza in condizioni di misspecificazione, cioè quando i dati presentano una struttura diversa da quella implicitamente assunta dal modello stimato.

3.2 Disegno di simulazione

Il disegno di simulazione è stato strutturato a partire da un unico Modello Generatore dei dati, formulato in modo sufficientemente generale da rappresentare le diverse configurazioni del segnale latente considerate nello studio. Il modello è definito dalle seguenti relazioni:

$$\begin{aligned} y_i &= \Lambda f_i + \epsilon_i \\ \epsilon_i &\sim \mathcal{N}_p(0, \Psi) \\ f_i \mid z_{ig} = 1 &\sim p_g(\cdot), \end{aligned}$$

dove y_i rappresenta il vettore delle p variabili osservate, Λ è una matrice di pesi fattoriali di dimensione $p \times q$, e ϵ_i è un vettore di errori idiosincratici indipendenti con matrice di covarianza diagonale Ψ. La variabile latente categoriale z_i identifica l'appartenenza dell'osservazione alla componente g-esima, con $g = 1, \ldots, G$, e probabilità a priori π_g. La distribuzione condizionata dei fattori latenti è indicata in forma generale tramite $p_g(\cdot)$, così da consentire differenti specificazioni del segnale latente.

Imponendo specifiche restrizioni sui parametri strutturali del modello generatore e sulle assunzioni distribuzionali, il modello si declina nelle seguenti configurazioni.

Nel caso di LCC, si assume la presenza simultanea di una struttura fattoriale a dimensione ridotta ($q < p$) e di due componenti latenti ($G = 2$), con probabilità a priori uguali ($\pi_g = 0.5$). I fattori latenti sono generati secondo

$$f_i \mid z_{ig} = 1 \sim \mathcal{N}_q(\alpha_g, \Phi_g),$$

così che la distribuzione dei dati osservati, condizionatamente all'appartenenza al gruppo g, risulti

$$y_i \mid z_{ig} = 1 \sim \mathcal{N}_p(\Lambda\alpha_g, \Lambda\Phi_g\Lambda^T + \Psi).$$

Questa configurazione corrisponde quindi alla struttura generativa di un modello HFMA, in cui la dipendenza tra le variabili osservate è spiegata da un sottospazio fattoriale comune, mentre l'eterogeneità tra gruppi è introdotta attraverso distribuzioni latenti diverse dei fattori. La matrice dei pesi fattoriali Λ viene generata imponendo una struttura semplice: i carichi principali, associati a blocchi di p/q variabili per fattore, sono estratti da una distribuzione uniforme $\mathcal{U}(0.6, 0.9)$, mentre i carichi trasversali, corrispondenti alle saturazioni sulle variabili non di pertinenza,

sono estratti da $\mathcal{U}(-0.25, 0.25)$. Le unicità in Ψ sono generate indipendentemente da $\mathcal{U}(0.1, 0.3)$.

Per garantire l'identificabilità teorica del modello, discussa nell'Equazione (2.8), i parametri latenti — medie α_g e matrici di covarianza Φ_g — sono stati vincolati in modo tale che la distribuzione marginale dei fattori presentasse media nulla e covarianza identità. A tal fine, i centroidi sono stati fissati in posizione speculare rispetto all'origine ($\alpha_2 = -\alpha_1$), così da annullare la media marginale della mistura. La covarianza marginale unitaria è stata ottenuta scompondo la variabilità complessiva in una componente *within-cluster* e una componente *between-cluster*:

$$Var(f) = \underbrace{\sum_{g=1}^{G} \pi_g \Phi_g}_{\text{Varianza Within}} + \underbrace{\sum_{g=1}^{G} \pi_g \alpha_g \alpha_g^T}_{\text{Varianza Between}} = I_q.$$

In una prima fase, sono stati definiti dei parametri grezzi (asteriscati) volti a modulare la forza del segnale nello spazio latente. Nella condizione di segnale debole ("vicini"), la distanza grezza tra i centroidi è stata fissata a $1.5/\sqrt{q}$, così da mantenere costante la distanza euclidea complessiva al variare della dimensionalità latente. Le matrici di covarianza intra-gruppo grezze sono state poste pari a $\Phi^*_{1,v} = 0.7 I_q$ e $\Phi^*_{2,v} = 1.4 I_q$. Nella condizione di segnale forte ("lontani"), per indurre una netta separazione spaziale, la distanza grezza è stata drasticamente aumentata a $5.0/\sqrt{q}$. In parallelo, le matrici di covarianza grezze sono state ampliate a $\Phi^*_{1,l} = 2.0 I_q$ e $\Phi^*_{2,l} = 4.0 I_q$.

In seguito, è stato calcolato un fattore di scala specifico S^2 derivante dalla scomposizione della varianza della mistura grezza:

$$S^2 = 0.5 \cdot \text{diag}(\Phi^*_1) + 0.5 \cdot \text{diag}(\Phi^*_2) + (\alpha^*_1)^2.$$

I parametri latenti definitivi utilizzati nella generazione dei dati sono stati infine ottenuti normalizzando i valori grezzi tramite il fattore di scala: $\alpha_g = \alpha^*_g / S$ e $\Phi_g = \Phi^*_g / S^2$.

Nel caso LCo, si assume una popolazione omogenea ($G = 1$), mantenendo attiva la riduzione dimensionale ($q < p$). In tale configurazione il modello generatore dei dati coincide con la struttura generativa dell'Analisi Fattoriale classica (FA), in quanto la variabilità condivisa è interamente catturata dal sottospazio fattoriale e non sono presenti raggruppamenti discreti. Per garantire coerenza e comparabilità, le matrici Λ e Ψ sono mantenute invariate.

In questo contesto, i due livelli del fattore sperimentale non corrispondono a differenti gradi di separazione spaziale, assente per costruzione, bensì al livello di conformità rispetto all'assunzione di normalità multivariata. Nella condizione favorevole alla FA, identificata come segnale forte poiché priva di distorsioni, si assume

$$f_i \sim \mathcal{N}_q(0, I_q),$$

ovvero la distribuzione teorica prevista dal modello. Nella condizione di segnale debole, invece, viene introdotta una deviazione dalle assunzioni strutturali generando i fattori da una distribuzione χ^2 con tre gradi di libertà, introducendo di fatto "rumore" distribuzionale. Per preservare i vincoli di identificabilità relativi alla scala,

le variabili vengono standardizzate mediante la trasformazione

$$f_{ij} = \frac{\chi^2_{(3)} - 3}{\sqrt{6}},$$

ottenendo fattori latenti con media nulla e varianza unitaria, ma caratterizzati da una marcata asimmetria.

Nel caso LCa, si assume la presenza di due gruppi latenti ($G = 2$) in assenza di un costrutto fattoriale condiviso a basso rango. Tale configurazione è ottenuta imponendo $q = p$, fissando la matrice dei carichi pari alla matrice identità ($\Lambda = I_p$) e assumendo nulle le varianze idiosincratiche ($\Psi = 0$). L'equazione generativa si riduce pertanto a

$$y_i = f_i,$$

così che i dati osservati coincidano con le variabili latenti. Assumendo inoltre

$$f_i \mid z_{ig} = 1 \sim \mathcal{N}_p(\mu_g, \Sigma_g),$$

il modello generatore diviene formalmente equivalente a un *Gaussian Mixture Model* nello spazio osservato. La forza del segnale viene definita attraverso il grado di separazione tra i vettori medi μ_g, traslando i centroidi μ_1 e μ_2. Per garantire coerenza con il disegno sperimentale e compensare l'effetto della maledizione della dimensionalità, la distanza tra i centroidi è stata calibrata in funzione della dimensionalità osservata: in bassa dimensionalità è pari a 2.5 nella condizione di segnale debole ("vicini") e 5.0 nella condizione di segnale forte ("lontani"); in alta dimensionalità le distanze sono state incrementate rispettivamente a 6.0 e 10.0. Le matrici di covarianza Σ_g sono generate dense e senza imporre vincoli di parsimonia.

L'incrocio delle tre nature del segnale latente, LCC, LCo e LCa, con i due livelli di forza appena descritti dà luogo a sei differenti modelli generativi. Ciascun modello viene ulteriormente specificato secondo due livelli di complessità dimensionale e due livelli di numerosità campionaria. Nelle configurazioni che prevedono una struttura fattoriale a dimensione ridotta, ossia LCC e LCo, la bassa dimensionalità corrisponde a $p = 10$ variabili osservate e $q = 2$ fattori latenti, mentre l'alta dimensionalità corrisponde a $p = 50$ variabili osservate e $q = 10$ fattori latenti. La configurazione LCa è invece definita in assenza di un sottospazio fattoriale ridotto, ponendo formalmente $q = p$. Le numerosità campionarie considerate sono $n = 100$ e $n = 1000$, così da confrontare il comportamento dei modelli in campioni piccoli e in campioni ampi. Il disegno sperimentale complessivo è quindi costituito da $3 \times 2 \times 2 \times 2 = 24$ scenari, e per ciascuna configurazione sono state generate 100 repliche Monte Carlo.

Infine, è opportuno specificare un limite imposto all'architettura PGMM. A fini computazionali, per gestire l'eccessiva onerosità della stima dell'intera griglia di modelli parsimoniosi all'aumentare dello spazio osservato, la stima è stata limitata in modo esclusivo alla parametrizzazione della famiglia 'CCU' (matrici di covarianza latente specifiche per gruppo ma diagonali, ed errore idiosincratico isotropo). Tale scelta garantisce tempi di esecuzione gestibili preservando auspicabilmente la validità comparativa dello studio.

3.3 Valutazione dei risultati

Per garantire una valutazione robusta delle performance algoritmiche, l'analisi dei risultati è stata condotta secondo due distinte prospettive metodologiche:

- Valutazione Oracolo: agli algoritmi vengono forniti in input i parametri strutturali reali che governano la generazione dei dati, ossia il numero corretto di componenti della mistura G e il numero corretto di dimensioni latenti q. Questa procedura consente di isolare la capacità inferenziale del modello nella stima delle matrici dei pesi fattoriali e nella classificazione delle unità statistiche, eliminando gli effetti dovuti a eventuali errori di specificazione della dimensionalità strutturale.

- Selezione del Modello: questa prospettiva riproduce le condizioni tipiche dell'analisi empirica, in cui la reale struttura generatrice dei dati è ignota. Gli algoritmi esplorano autonomamente una griglia predefinita di possibili valori per G e q, selezionando la combinazione ottimale tramite la minimizzazione del BIC, descritto nell'Equazione (1.10). Tale procedura permette di valutare l'affidabilità pratica dei modelli nel recuperare la corretta complessità del processo generatore.

Per entrambe le procedure, al fine di quantificare in modo rigoroso le prestazioni e rispondere agli obiettivi della simulazione, lo studio si avvale di due metriche principali. Tali indici sono stati selezionati per valutare due dimensioni distinte e complementari dell'analisi: l'accuratezza della partizione in classi e la precisione nel recupero della vera struttura fattoriale.

La capacità dei modelli di identificare la corretta partizione delle unità statistiche è stata valutata impiegando l'*Adjusted Rand Index* (ARI). L'utilizzo di tale metrica è stato circoscritto esclusivamente alle configurazioni LCC e LCa, essendo gli unici contesti sperimentali in cui il processo generatore prevede una reale eterogeneità di gruppo ($G = 2$). A differenza del semplice tasso di classificazione corretta, l'ARI confronta la partizione vera e la partizione stimata correggendo per l'accordo dovuto unicamente al caso. Data una tabella di contingenza che incrocia le assegnazioni vere con quelle stimate, sia n_{ij} il numero di osservazioni appartenenti simultaneamente al gruppo vero i e al gruppo stimato j. Siano inoltre $a_i = \sum_j n_{ij}$ i totali marginali di riga e $b_j = \sum_i n_{ij}$ i totali marginali di colonna. L'ARI è definito come:

$$ARI = \frac{\sum_{i,j} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}. \tag{3.1}$$

L'indice assume un valore massimo pari a 1 quando la classificazione stimata coincide perfettamente con la realtà, e valori prossimi a 0 quando la partizione non è migliore di un'assegnazione puramente aleatoria.

Per valutare invece la fedeltà con cui i modelli recuperano la vera matrice dei pesi fattoriali Λ, è stata impiegata la *Cosine Similarity*. Tale metrica è applicata unicamente alle configurazioni LCC e LCo, ove il costrutto latente continuo è effettivamente presente. A causa dell'indeterminatezza rotazionale dei modelli fattoriali,

prima del calcolo della congruenza è stato applicato un allineamento spaziale tramite Analisi di Procruste Ortogonale. Individuata la matrice di rotazione ottimale $\hat{T}$ tramite l'Equazione (1.11), si ottiene la matrice stimata riallineata $\tilde{\Lambda} = \hat{\Lambda}\hat{T}$. Successivamente, viene eseguito un *matching* iterativo fattore per fattore. La Similarità del Coseno tra l'i-esimo fattore vero (λ_i) e il j-esimo fattore stimato riallineato ($\tilde{\lambda}_j$) è calcolata come:

$$CosSim(\lambda_i, \tilde{\lambda}_j) = \frac{|\lambda_i^T \tilde{\lambda}_j|}{\|\lambda_i\|_2 \|\tilde{\lambda}_j\|_2}, \tag{3.2}$$

dove l'impiego del valore assoluto al numeratore ignora le arbitrarie inversioni di segno. La performance complessiva è estratta tramite un algoritmo che individua iterativamente la coppia di fattori con similarità più elevata, rimuovendoli fino a esaurimento delle dimensioni latenti vere q. Infine, qualora il modello sottostimi la dimensionalità ($q_{est} < q_{true}$), la similarità associata ai fattori reali non estratti viene posta a zero, penalizzando l'incapacità dell'algoritmo di recuperare interamente lo spazio latente.

Per sintetizzare la mole di dati generata dalle iterazioni Monte Carlo, è stata implementata una specifica procedura di aggregazione. Le metriche continue (ARI e *Cosine Similarity*) sono state riassunte calcolando la media aritmetica e la deviazione standard sulle iterazioni convergenti per ciascuno dei 24 scenari. In particolare, per la *Cosine Similarity* è stato adottato un approccio di *Micro-Averaging*: tutti i valori di congruenza ottenuti nelle repliche Monte Carlo sono stati aggregati in un'unica distribuzione globale da cui sono state estratte media e varianza, catturando così l'instabilità algoritmica sul recupero dei singoli fattori. Infine, l'attenzione valutativa si concentra sui parametri discreti di complessità strutturale estratti nella prospettiva di Selezione del Modello: il numero stimato di *cluster* ($\hat{G}$) e di fattori ($\hat{q}$). Trattandosi di conteggi, tali stime sono state sintetizzate riportando il valore modale affiancato dalla frequenza di selezione, espressa come percentuale di iterazioni in cui l'algoritmo ha selezionato tale valore.

3.4 Risultati

I risultati sono riportati nelle Tabelle 3.1-3.6. In condizioni di bassa dimensionalità ($p = 10$), le performance dei modelli risultano complessivamente equiparabili e coerenti con le rispettive assunzioni teoriche. Quando la separazione tra i gruppi è netta ("Lontani") e supportata da una numerosità campionaria adeguata ($n = 1000$), il GMM e le architetture a mistura fattoriale (PGMM e HFMA) individuano correttamente la partizione, raggiungendo valori medi di ARI compresi tra 0.94 e 0.98, indipendentemente dal fatto che i raggruppamenti risiedano in un sottospazio latente a basso rango (segnale continuo e categoriale, Tabelle 3.1 e 3.2) o direttamente nello spazio osservato (segnale esclusivamente categoriale, Tabelle 3.3 e 3.4). In quest'ultimo contesto esplorativo ristretto, i modelli a mistura fattoriale, pur in assenza di veri fattori generativi, tendono a sovrastimare la dimensionalità latente (estraendo tra 4 e 5 dimensioni) per sintetizzare la matrice di correlazione, senza tuttavia compromettere l'accuratezza del *clustering*.

Non sorprendentemente, la variabilità delle stime aumenta in presenza di campioni di dimensioni ridotte ($n = 100$). Questa condizione permette di osservare

chiaramente la differenza tra la valutazione oracolo e la selezione autonoma, mettendo in luce il duplice comportamento del criterio BIC. Da una parte, il BIC agisce come un filtro protettivo: vincolare l'algoritmo EM su parametri fissi a priori, come nella modalità oracolo, lo porta a convergere frequentemente su massimi locali subottimali; l'esplorazione autonoma su griglia permette invece di scartare queste stime instabili. Questo effetto è evidente per l'HFMA in bassa dimensionalità (condizione "Lontani", $n = 100$ per il segnale LCC): la selezione autonoma riduce la deviazione standard dell'errore (da 0.29 a 0.15) e migliora l'ARI medio (da 0.86 a 0.92). Dall'altra parte, passando all'alta dimensionalità ($p = 50, n = 100$), la selezione autonoma mostra i propri limiti quando i parametri da stimare sono troppi a fronte di una numerosità campionaria ridotta. I modelli a mistura fattoriale devono stimare matrici dei carichi molto dense (500 parametri per $\hat{q} = 10$), un valore sproporzionato rispetto alle sole 100 osservazioni disponibili. Questa mancanza di dati causa una penalizzazione estrema da parte del BIC, che costringe il PGMM e l'HFMA a ripiegare sulla soluzione più semplice, ovvero l'assenza di partizioni ($\hat{G} = 1$, con ARI nullo). In questa specifica situazione il GMM, non dovendo stimare il sottospazio latente, richiede un numero inferiore di parametri e riesce paradossalmente a mantenere una buona capacità di classificazione (ARI 0.84).

Il comportamento dei modelli si differenzia nettamente analizzando campioni ampi in alta dimensionalità ($p = 50, n = 1000$). In questo regime, l'incremento delle variabili manifeste genera un elevato livello di rumore che incide sulle prestazioni in base alla natura del segnale latente. Se i gruppi sono separati esclusivamente nel sottospazio latente (LCC), il GMM, operando solo nello spazio osservato, subisce un severo mascheramento e non riesce a isolare il segnale: la stima converge sistematicamente sull'assenza di partizioni ($\hat{G} = 1$ al 100%, ARI 0.00). I modelli a mistura fattoriale, al contrario, traggono pieno vantaggio dalla propria architettura ibrida. Estraendo i fattori latenti, PGMM e HFMA riescono a filtrare il rumore manifesto, recuperano con precisione la matrice dei carichi (*Cosine Similarity* tra 0.97 e 1.00) e raggiungono livelli di classificazione ottimali (ARI 0.98 e 0.99).

L'utilità delle misture fattoriali in alta dimensionalità trova conferma anche in assenza di un reale costrutto latente condiviso (segnale LCa). In tale contesto strutturale, il GMM godrebbe di un netto vantaggio teorico, ma la necessità di stimare matrici di covarianza intere per 50 variabili impone un onere parametrico tale da saturare l'algoritmo, inducendolo a unificare i gruppi ($\hat{G} = 1$ in oltre il 59% dei casi) e abbattendo l'ARI a valori limitati ($0.08 - 0.39$). Di fronte a questa criticità, l'estrazione latente assume una vitale funzione di regolarizzazione. Proiettando l'informazione su un sottospazio ridotto (con estrazioni modali comprese tra 5 e 9 fattori), PGMM e HFMA contraggono drasticamente lo spazio dei parametri liberi, eludono la penalizzazione imposta dal BIC e riescono a isolare fedelmente la separazione reale, mantenendo un ARI compreso tra 0.96 e 1.00.

Infine, il comportamento algoritmico a fronte di un segnale esclusivamente continuo (Tabelle 3.5 e 3.6) definisce i limiti legati all'eccessiva flessibilità dei modelli in popolazioni omogenee ($G = 1$). Sotto la condizione di *baseline* normale, tutte le architetture identificano correttamente l'assenza di *cluster*. L'introduzione di distribuzioni a code pesanti (χ^2) in bassa dimensionalità porta ad identificare un numero di componenti superiori a uno per approssimare la non-gaussianità dei dati: il GMM stima frequentemente due componenti, mentre il PGMM e l'HFMA ne identificano

tre in oltre l'80% delle simulazioni. In alta dimensionalità ($p = 50$), il peso parametrico del *clustering* compensa parzialmente questa distorsione per il GMM e il PGMM, riportandoli a selezionare $\hat{G} = 1$ in oltre il 92% dei casi. L'HFMA, per via della rigorosa parsimonia imposta alla matrice degli errori, mantiene invece una flessibilità eteroschedastica che lo porta a selezionare tre componenti nel 70% delle prove, verosimilmente come adattamento alla non gaussianità della distribuzione. Ciononostante, i risultati restituiscono una fondamentale conferma metodologica per quanto concerne il modello di misura. L'Analisi Fattoriale, il PGMM e l'HFMA, nonostante la selezione di componenti aggiuntive, estraggono sistematicamente il numero corretto di dimensioni latenti ($\hat{q} = 2$ e $\hat{q} = 10$) e preservano valori di *Cosine Similarity* pressoché perfetti (tra 0.98 e 1.00). Tale evidenza suggerisce che l'asimmetria distribuzionale può indurre i modelli a mistura a introdurre componenti aggiuntive per approssimare meglio la forma della distribuzione, anche in assenza di vere classi generatrici. Questo comportamento non deve quindi essere interpretato necessariamente come un errore di classificazione, ma piuttosto come una diversa funzione delle componenti di mistura: non tanto l'individuazione di gruppi sostantivi, quanto la rappresentazione flessibile di una distribuzione non gaussiana. Al tempo stesso, il recupero del sottospazio fattoriale continuo rimane stabile, come mostrano la corretta selezione del numero di fattori e i valori di *Cosine Similarity*.

I risultati dello studio Monte Carlo mostrano che le prestazioni dei modelli dipendono da un equilibrio tra capacità descrittiva, complessità parametrica e criterio di selezione utilizzato. Un modello più complesso non è automaticamente migliore: la sua efficacia dipende dal tipo di dati analizzati, dalla dimensionalità e dalla quantità di osservazioni disponibili.

L'Analisi Fattoriale si conferma il metodo più stabile per il recupero della struttura latente. Pur non essendo progettata per classificare le unità statistiche in gruppi, riesce quasi sempre a identificare correttamente il numero di fattori e mostra una notevole robustezza sia rispetto al rumore sia rispetto alle deviazioni distribuzionali.

Il GMM, invece, è molto efficace quando la dimensionalità è contenuta e i gruppi sono separati direttamente nello spazio osservato. Tuttavia, il modello soffre rapidamente la crescita del numero di variabili, poiché la stima delle matrici di covarianza complete richiede un numero elevato di parametri. In alta dimensionalità, questo porta il BIC a preferire soluzioni troppo semplici, spesso incapaci di recuperare i *cluster* reali.

I modelli ibridi PGMM e HFMA rappresentano quindi il compromesso più flessibile. In scenari a bassa dimensionalità offrono stime paragonabili al GMM, ma il loro vero vantaggio emerge in spazi ad alta dimensionalità, declinandosi in due direttrici. In presenza di strutture latenti, l'estrazione simultanea dei fattori agisce come un filtro che isola il segnale e smaschera i *cluster*. In contesti densi in cui la separazione è puramente spaziale, la riduzione dimensionale diviene invece uno strumento di regolarizzazione parametrica. Contraendo lo spazio di stima, la sintesi fattoriale aggira gli ostacoli imposti dalla maledizione della dimensionalità e assicura il corretto recupero dei gruppi lì dove l'approccio puramente multivariato fallisce.

Tuttavia, modelli complessi e flessibili come il PGMM e l'HFMA richiedono particolare cautela interpretativa. In presenza di distribuzioni asimmetriche o non gaussiane, la selezione di più componenti non indica necessariamente l'esistenza di gruppi latenti sostantivamente distinti, ma può riflettere la capacità della mistura di

approssimare forme distributive più complesse. Ne consegue che il numero di componenti selezionato deve essere interpretato alla luce dell'obiettivo dell'analisi: se l'interesse è la ricostruzione della densità, più componenti possono costituire una soluzione adeguata; se invece l'obiettivo è individuare sottopopolazioni interpretabili, è necessario verificare che le classi ottenute abbiano anche un significato sostantivo e non siano soltanto il risultato dell'adattamento alla forma della distribuzione.

Nel complesso, l'integrazione tra *clustering* e variabili latenti rappresenta un importante avanzamento metodologico, soprattutto in contesti ad alta dimensionalità. Tuttavia, la scelta del modello deve sempre essere guidata da un compromesso tra flessibilità, interpretabilità e parsimonia, evitando di applicare automaticamente l'approccio più sofisticato indipendentemente dalla struttura reale dei dati.

Tabella 3.1: Valori medi (deviazione standard) al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale continuo e categoriale (LCC): valutazione oracolo.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Vicini	Lontani	Vicini	Lontani	Vicini	Lontani	Vicini	Lontani
FA	CosSim	0.96 (0.01)	0.89 (0.02)	0.96 (0.00)	0.89 (0.01)	0.96 (0.01)	0.89 (0.02)	0.98 (0.00)	0.91 (0.01)
GMM	ARI	0.47 (0.18)	0.96 (0.04)	0.63 (0.05)	0.97 (0.01)	0.16 (0.30)	0.78 (0.39)	0.00 (0.00)	0.00 (0.00)
PGMM	ARI	0.47 (0.22)	0.92 (0.12)	0.58 (0.14)	0.94 (0.02)	0.44 (0.16)	0.92 (0.08)	0.53 (0.12)	0.98 (0.01)
	CosSim	0.69 (0.34)	0.65 (0.33)	0.99 (0.01)	0.63 (0.41)	0.92 (0.17)	0.90 (0.21)	0.96 (0.16)	1.00 (0.00)
HFMA	ARI	0.50 (0.24)	0.86 (0.29)	0.58 (0.22)	0.95 (0.14)	0.11 (0.12)	0.20 (0.24)	0.68 (0.23)	0.91 (0.26)
	CosSim	0.96 (0.01)	0.88 (0.02)	0.96 (0.00)	0.88 (0.01)	0.95 (0.01)	0.87 (0.02)	0.97 (0.01)	0.89 (0.01)

Tabella 3.2: Valori medi (deviazione standard) dell'ARI e della Cosine Similarity, e valori modali (frequenza % di selezione) di $\hat{q}$ e $\hat{G}$ al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale continuo e categoriale (LCC): selezione del modello.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Vicini	Lontani	Vicini	Lontani	Vicini	Lontani	Vicini	Lontani
FA	$\hat{q}$	2 (100%)	2 (100%)	2 (100%)	2 (100%)	10 (100%)	10 (99%)	10 (100%)	10 (100%)
	CosSim	0.96 (0.01)	0.89 (0.02)	0.96 (0.00)	0.89 (0.01)	0.96 (0.01)	0.89 (0.03)	0.98 (0.00)	0.91 (0.01)
GMM	$\hat{G}$	1 (97%)	3 (76%)	2 (100%)	2 (100%)	1 (83%)	3 (70%)	1 (100%)	1 (100%)
	ARI	0.01 (0.09)	0.76 (0.18)	0.63 (0.05)	0.97 (0.01)	0.10 (0.23)	0.84 (0.15)	0.00 (0.00)	0.00 (0.00)
PGMM	$\hat{G}$	3 (38%)	3 (50%)	2 (93%)	3 (68%)	1 (97%)	1 (98%)	1 (100%)	2 (100%)
	$\hat{q}$	1 (43%)	1 (89%)	2 (100%)	1 (35%)	10 (99%)	10 (83%)	10 (100%)	10 (100%)
	ARI	0.22 (0.24)	0.78 (0.19)	0.58 (0.13)	0.79 (0.14)	0.02 (0.10)	0.02 (0.14)	0.00 (0.00)	0.98 (0.01)
	CosSim	0.74 (0.39)	0.51 (0.46)	1.00 (0.01)	0.82 (0.38)	0.95 (0.03)	0.85 (0.11)	0.97 (0.01)	1.00 (0.00)
HFMA	$\hat{G}$	1 (81%)	2 (88%)	2 (84%)	2 (97%)	1 (100%)	1 (100%)	2 (89%)	2 (99%)
	$\hat{q}$	2 (100%)	2 (100%)	2 (100%)	2 (100%)	10 (100%)	10 (80%)	10 (100%)	10 (93%)
	ARI	0.13 (0.26)	0.92 (0.15)	0.65 (0.08)	0.97 (0.01)	0.00 (0.00)	0.00 (0.00)	0.67 (0.24)	0.99 (0.03)
	CosSim	0.96 (0.01)	0.88 (0.02)	0.96 (0.00)	0.88 (0.01)	0.95 (0.01)	0.85 (0.13)	0.97 (0.01)	0.89 (0.01)

Tabella 3.3: Valori medi (deviazione standard) al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale esclusivamente categoriale (LCa): valutazione oracolo.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Vicini	Lontani	Vicini	Lontani	Vicini	Lontani	Vicini	Lontani
GMM	ARI	0.21 (0.13)	0.94 (0.06)	0.76 (0.03)	0.98 (0.01)	0.01 (0.09)	0.20 (0.40)	0.04 (0.19)	0.22 (0.42)

Tabella 3.4: Valori medi (deviazione standard) dell'ARI e della Cosine Similarity, e valori modali (frequenza % di selezione) di $\hat{q}$ e $\hat{G}$ al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale esclusivamente categoriale (LCa): selezione del modello.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Vicini	Lontani	Vicini	Lontani	Vicini	Lontani	Vicini	Lontani
FA	$\hat{q}$	2 (55%)	1 (49%)	5 (88%)	5 (100%)	1 (100%)	1 (100%)	6 (24%)	7 (31%)
GMM	$\hat{G}$	2 (58%)	2 (72%)	2 (100%)	2 (100%)	1 (97%)	1 (60%)	1 (92%)	1 (59%)
	ARI	0.18 (0.15)	0.88 (0.13)	0.76 (0.03)	0.98 (0.01)	0.03 (0.16)	0.40 (0.49)	0.08 (0.26)	0.39 (0.47)
PGMM	$\hat{G}$	1 (67%)	2 (79%)	2 (99%)	2 (92%)	1 (88%)	2 (100%)	2 (95%)	2 (82%)
	$\hat{q}$	1 (86%)	1 (87%)	4 (63%)	5 (55%)	1 (100%)	1 (100%)	7 (32%)	5 (33%)
	ARI	0.07 (0.12)	0.89 (0.12)	0.75 (0.04)	0.96 (0.05)	0.12 (0.32)	1.00 (0.00)	0.98 (0.05)	0.96 (0.10)
HFMA	$\hat{G}$	1 (100%)	2 (99%)	2 (82%)	2 (72%)	2 (100%)	2 (100%)	2 (92%)	2 (94%)
	$\hat{q}$	1 (56%)	1 (70%)	5 (75%)	5 (78%)	1 (100%)	1 (100%)	9 (25%)	6 (16%)
	ARI	0.00 (0.00)	0.91 (0.12)	0.53 (0.12)	0.89 (0.09)	0.98 (0.03)	1.00 (0.00)	0.98 (0.04)	1.00 (0.03)

Tabella 3.5: Valori medi (deviazione standard) al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale esclusivamente continuo (LCo): valutazione oracolo.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Norm.	χ^2	Norm.	χ^2	Norm.	χ^2	Norm.	χ^2
FA	CosSim	0.99 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.97 (0.01)	1.00 (0.00)	1.00 (0.00)
PGMM	CosSim	0.65 (0.40)	0.65 (0.40)	0.97 (0.14)	0.96 (0.16)	0.97 (0.08)	0.98 (0.01)	0.99 (0.08)	0.98 (0.11)
HFMA	CosSim	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)

Tabella 3.6: Valori medi (deviazione standard) dell'ARI e della Cosine Similarity, e valori modali (frequenza % di selezione) di $\hat{q}$ e $\hat{G}$ al variare delle repliche Monte Carlo. I risultati si riferiscono al macroscenario segnale esclusivamente continuo (LCo): selezione del modello.

Modello	**Metrica**	**Bassa Dimensionalità** ($p = 10$)				**Alta Dimensionalità** ($p = 50$)			
		$n = 100$		$n = 1000$		$n = 100$		$n = 1000$	
		Norm.	χ^2	Norm.	χ^2	Norm.	χ^2	Norm.	χ^2
FA	$\hat{q}$	2 (100%)	2 (100%)	2 (100%)	2 (100%)	10 (100%)	10 (100%)	10 (100%)	10 (100%)
	CosSim	0.99 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.97 (0.01)	1.00 (0.00)	1.00 (0.00)
GMM	$\hat{G}$	1 (98%)	1 (49%)	1 (100%)	2 (83%)	1 (100%)	1 (98%)	1 (100%)	1 (92%)
PGMM	$\hat{G}$	3 (47%)	3 (63%)	1 (98%)	3 (80%)	1 (100%)	1 (100%)	1 (99%)	1 (97%)
	$\hat{q}$	1 (35%)	1 (50%)	2 (95%)	2 (63%)	10 (99%)	10 (100%)	10 (100%)	10 (100%)
	CosSim	0.80 (0.37)	0.72 (0.42)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)
HFMA	$\hat{G}$	1 (100%)	2 (71%)	1 (100%)	3 (100%)	1 (100%)	1 (100%)	1 (100%)	3 (70%)
	$\hat{q}$	2 (100%)	2 (100%)	2 (100%)	2 (100%)	10 (100%)	10 (100%)	10 (100%)	10 (100%)
	CosSim	1.00 (0.00)	0.99 (0.00)	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.98 (0.01)	1.00 (0.00)	1.00 (0.00)

Capitolo 4

Dati reali

4.1 Descrizione del dataset

In questo capitolo si presenta l'applicazione dei modelli discussi nei capitoli precedenti a un insieme di dati reali noto come BFI (*Big Five Inventory*, Goldberg 1992), ampiamente utilizzato in psicometria per lo studio della struttura della personalità umana. Il questionario è costruito sulla base del modello teorico dei *Big Five*, secondo il quale la personalità può essere descritta attraverso cinque dimensioni latenti principali: Amicalità, Coscienziosità, Estroversione, Nevroticismo e Apertura all'esperienza. I dati sono stati raccolti nell'ambito del progetto *Synthetic Aperture Personality Assessment* (SAPA) mediante un questionario di autovalutazione. A seconda della rilevazione, di tali dati esistono diverse versioni. In questa tesi si utilizza una versione leggermente ridotta, indicata come `bfi`, disponibile nel pacchetto R `psych` (William Revelle, 2026).

La matrice dei dati è composta da 28 variabili complessive, di cui 3 raccolgono informazioni demografiche ausiliarie (genere, livello di istruzione ed età dei rispondenti) e le restanti 25 corrispondono agli *item* del questionario psicometrico basato sul modello teorico dei *Big Five*. Ciascun dominio è misurato da cinque *item* specifici, valutati mediante una scala Likert ordinale a sei modalità con valori compresi tra 1, corrispondente a "Molto inaccurato", e 6, corrispondente a "Molto accurato".

Di seguito si riporta l'elenco delle 25 variabili, suddivise per dominio di appartenenza, con una sintesi del costrutto indagato:

- **Amicalità**:
  - `A1`: Indifferenza verso i sentimenti altrui.
  - `A2`: Interesse per il benessere degli altri.
  - `A3`: Capacità di confortare le persone.
  - `A4`: Affetto verso i bambini.
  - `A5`: Capacità di mettere gli altri a proprio agio.
- **Coscienziosità**:
  - `C1`: Esigenza e rigore nel proprio lavoro.
  - `C2`: Tendenza a continuare fino al raggiungimento della perfezione.

  - `C3`: Attitudine a seguire una pianificazione.
  - `C4`: Tendenza a fare le cose a metà.
  - `C5`: Tendenza a sprecare il proprio tempo.

- **Estroversione**:
  - `E1`: Tendenza a parlare poco.
  - `E2`: Difficoltà ad approcciarsi agli altri.
  - `E3`: Capacità di affascinare o coinvolgere le persone.
  - `E4`: Facilità nel fare amicizia.
  - `E5`: Tendenza a prendere l'iniziativa e ad assumere il comando.

- **Nevroticismo**:
  - `N1`: Tendenza ad arrabbiarsi facilmente.
  - `N2`: Tendenza a irritarsi con facilità.
  - `N3`: Frequenti sbalzi d'umore.
  - `N4`: Tendenza a sentirsi spesso malinconici o giù di morale.
  - `N5`: Tendenza al panico.

- **Apertura all'esperienza**:
  - `O1`: Ricchezza di idee.
  - `O2`: Tendenza a evitare letture o argomenti difficili.
  - `O3`: Capacità di elevare il livello della conversazione.
  - `O4`: Tendenza a trascorrere tempo riflettendo sulle cose.
  - `O5`: Tendenza a non approfondire i soggetti.

Trattandosi di variabili misurate su scala Likert, questi dati non rappresentano, in linea di principio, un riferimento ideale per i modelli discussi nella tesi. Tuttavia, a dispetto della palese violazione dell'ipotesi di normalità, i dati sono noti per la frequente applicazione di tecniche di analisi fattoriale. Sembra pertanto utile valutare se, oltre alla già provata presenza di tratti latenti, sia presente anche una struttura di gruppo, ovvero se tali componenti si manifestino in modo eterogeneo tra gli individui.

4.2 Analisi esplorativa

In via preliminare rispetto alla stima dei modelli, è stata condotta un'analisi esplorativa per valutare la qualità dei dati, indagare la presenza di valori mancanti, esaminare la distribuzione delle risposte e delineare la struttura di dipendenza tra le variabili.

Come primo passo, sono state isolate le 25 variabili psicometriche del questionario. Al fine di mantenere un'elevata qualità dell'informazione, le osservazioni caratterizzate da almeno un valore mancante su tali domande sono state rimosse. Questa

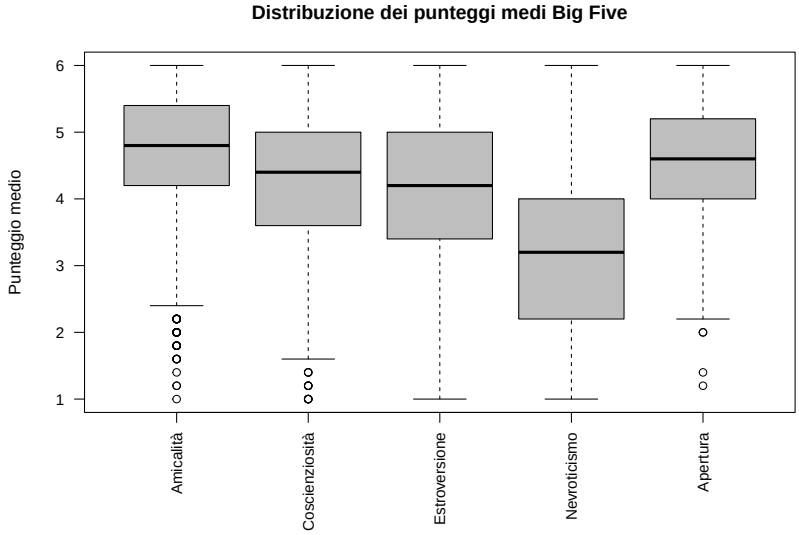

Figura 4.1: Distribuzione dei punteggi medi nelle cinque dimensioni dei *Big Five*.

operazione ha portato a un campione finale di 2436 unità complete, comportando un'esclusione di 364 individui, pari al 13% del campione originario.

Successivamente, poiché alcuni *item* risultano formulati in direzione opposta rispetto al tratto di riferimento – in particolare `A1`, `C4`, `C5`, `E1`, `E2`, `O2` e `O5` – si è proceduto alla loro ricodifica in direzione inversa, garantendo che a punteggi più elevati corrispondesse sempre una maggiore intensità del tratto latente. Per operare su scale confrontabili ed evitare che differenze spurie di posizione o dispersione potessero distorcere l'adattamento dei modelli multivariati, tutte le variabili sono state poi standardizzate.

A livello descrittivo, sono stati calcolati i punteggi medi per ciascuna delle cinque dimensioni teoriche dei *Big Five*. La distribuzione di tali valori, illustrata nella Figura 4.1, evidenzia come l'amicalità e l'apertura all'esperienza tendano ad assumere valori centrali più elevati; di contro, il nevroticismo fa registrare i punteggi più bassi, accompagnati però da una maggiore variabilità.

Per esplorare la struttura di dipendenza, è stata calcolata la matrice di correlazione sui dati ricodificati e standardizzati (Figura 4.2). Il grafico restituisce un *pattern* fortemente allineato all'organizzazione teorica del questionario: si osserva infatti una correlazione decisamente più marcata tra le variabili appartenenti al medesimo dominio rispetto a quelle associate a dimensioni distinte. Tale evidenza costituisce un primo forte indizio a favore di una struttura latente fattoriale.

Infine, l'idoneità della matrice di correlazione all'analisi fattoriale è stata valutata attraverso l'indice di Kaiser-Meyer-Olkin (Kaiser, 1974). Tale indice ha restituito un valore complessivo pari a 0.85, denotando un'ottima fattorizzabilità della base di dati, in quanto evidenzia come la proporzione di varianza comune tra gli *item* sia ampiamente sufficiente per giustificare l'estrazione di fattori latenti.

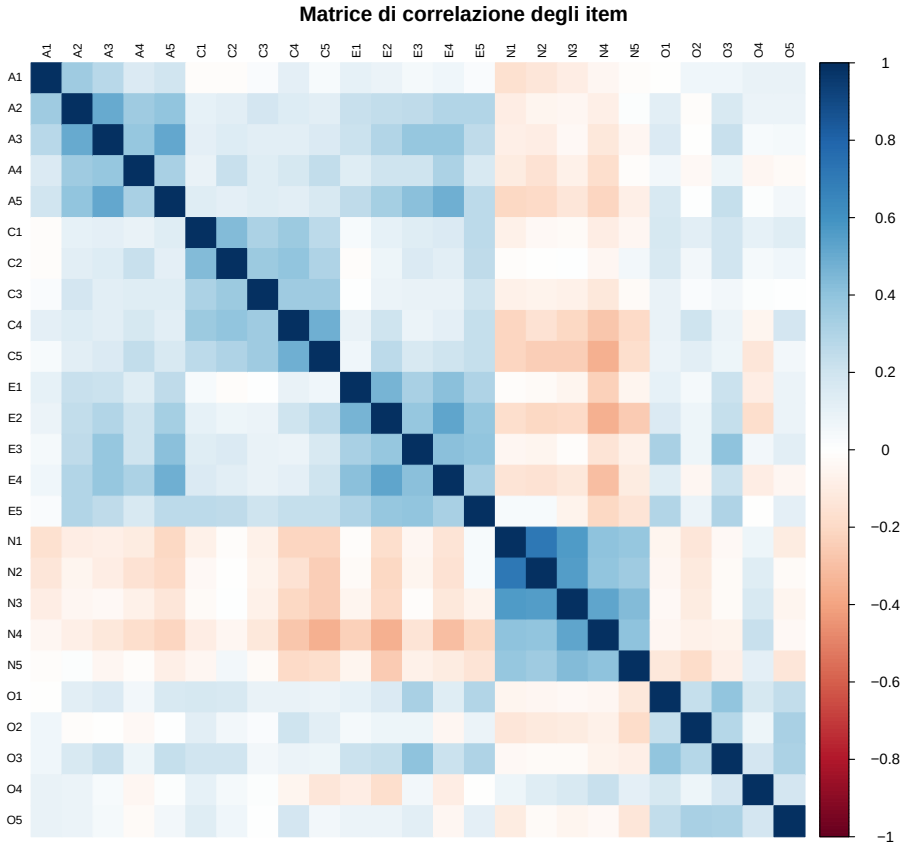

Figura 4.2: Matrice di correlazione tra i 25 *item* ricodificati e standardizzati.

4.3 Stima dei modelli

4.3.1 Analisi fattoriale

L'analisi fattoriale rappresenta il primo modello stimato sui dati reali e costituisce il riferimento per descrivere la struttura latente continua delle variabili. Coerentemente con l'impianto teorico di riferimento basato sul modello dei *Big Five*, si è deciso di adottare un approccio interpretativo e di fissare a priori il numero di dimensioni latenti a cinque. Questa specifica scelta metodologica verrà mantenuta costante anche per tutti i successivi modelli a fattori latenti analizzati nel capitolo (PGMM e HFMA), così da garantire un confronto coerente tra le diverse specificazioni.

Per completezza, è opportuno segnalare incidentalmente che criteri puramente *data-driven* fornirebbero indicazioni in parte divergenti: l'andamento dello *scree plot* degli autovalori suggerirebbe infatti l'estrazione di sei componenti, mentre il BIC tenderebbe a favorirne otto. Tuttavia, poiché l'obiettivo dell'analisi è ottenere risultati interpretabili e fedeli alla struttura psicometrica del questionario, l'analisi si focalizza esclusivamente sull'architettura a cinque fattori.

La stima dei parametri è avvenuta tramite il metodo della massima verosimiglianza. Inoltre, poiché non è realistico imporre che i tratti della personalità siano del tutto indipendenti tra loro, è stata applicata una rotazione obliqua di tipo *oblimin*. Questa procedura, ampiamente consolidata in ambito psicometrico, consente ai

Item	F1	F2	F3	F4	F5
E1	0.56				
E2	0.66				
E3	0.34			0.33	
E4	0.54				0.36
E5	0.39				
N1		0.87			
N2		0.82			
N3		0.66			
N4	-0.44	0.41			
N5		0.44			
C1			0.54		
C2			0.64		
C3			0.57		
C4			0.67		
C5			0.58		
O1				0.54	
O2				0.46	
O3				0.64	
O4	-0.36			0.37	
O5				0.52	
A1					0.39
A2					0.63
A3					0.69
A4					0.47
A5					0.58

Tabella 4.1: Principali carichi fattoriali della soluzione a cinque fattori (rotazione *oblimin*). Sono riportati soltanto i carichi con valore assoluto superiore a 0.30.

fattori di risultare correlati tra loro, spingendo al contempo le saturazioni secondarie verso lo zero per approssimare il principio della struttura semplice.

La Tabella 4.1 riporta i principali carichi fattoriali della soluzione ruotata. Per maggiore chiarezza, sono mostrati solo i coefficienti con valore assoluto superiore a 0.30. La conformazione della matrice dei *loadings* restituisce un quadro ampiamente compatibile con il questionario: ogni dimensione latente estratta riflette infatti in maniera definita uno specifico dominio. In particolare, il primo fattore riflette l'Estroversione, il secondo il Nevroticismo, il terzo la Coscienziosità, il quarto l'Apertura all'esperienza e il quinto l'Amicalità. Sebbene alcuni *item* (in particolare E3, E4, N4 e O4) esibiscano fisiologici carichi secondari, l'architettura generale permette di isolare i domini dei *Big Five* senza ambiguità.

Questa prima analisi conferma la presenza di una solida struttura latente continua, che fungerà da asse interpretativo per i modelli successivi che andranno a verificare se coesistono forme di eterogeneità categoriale.

4.3.2 Modelli a mistura gaussiani

Per valutare se l'eterogeneità osservata possa essere descritta solo in termini categoriali, sono stati stimati i *Gaussian Mixture Models* sui 25 *item* standardizzati. Il numero di componenti è stato fatto variare tra 1 e 12, utilizzando il criterio BIC per selezionare sia il numero di gruppi sia la struttura della matrice di covarianza.

L'implementazione di *default* dell'algoritmo EM nel pacchetto R `mclust` (Scrucca *et al.*, 2023) prevede un'inizializzazione basata su *clustering* gerarchico; tuttavia, per campioni ampi, tale procedura iniziale viene applicata su un sotto-campione estratto casualmente. Di conseguenza, la procedura è stata ripetuta con diverse inizializzazioni. La struttura di covarianza è risultata molto stabile: il BIC ha sempre selezionato il modello `VVE` (componenti ellissoidali con volume e forma variabili, ma orientamento comune). Al contrario, il numero di classi selezionato ha manifestato una notevole variabilità, oscillando tra 4 e 8. Questo risultato evidenzia la sensibilità dei GMM ai massimi locali della funzione di verosimiglianza.

Tra i risultati ottenuti, è stata scelta come riferimento la soluzione `VVE` a 8 componenti, poiché associata al miglior valore del BIC e ricorrente al variare delle inizializzazioni. I gruppi individuati non devono essere interpretati come categorie psicologiche rigide, ma "profili latenti di risposta", ossia sottogruppi di individui con *pattern* simili nelle risposte al questionario.

Per interpretare le classi individuate, sono stati calcolati i profili medi sulle cinque dimensioni dei *Big Five*, ottenuti come media degli *item* ricodificati appartenenti a ciascun dominio teorico. I risultati sono riportati nella Tabella 4.2; tra parentesi sono indicate le deviazioni standard, utili per valutare il grado di variabilità interna a ciascun gruppo.

La soluzione individua una segmentazione articolata del campione. La Classe 6 si distingue per i valori medi più elevati di Amicalità, Coscienziosità ed Estroversione, insieme al livello medio più basso di Nevroticismo. Inoltre, le deviazioni standard relativamente contenute indicano una maggiore omogeneità interna rispetto ad altre classi. La Classe 8 presenta valori elevati di Amicalità, Estroversione e Apertura all'esperienza, associati a un Nevroticismo relativamente contenuto; anche in questo caso, la variabilità interna risulta moderata, soprattutto per Amicalità e Apertura. La Classe 7 si caratterizza invece per il valore medio più elevato di Apertura all'esperienza, ma presenta valori più bassi di Amicalità ed Estroversione e una variabilità più marcata in alcune dimensioni, in particolare Coscienziosità, Estroversione e Nevroticismo. Infine, la Classe 2 registra il livello medio più elevato di Nevroticismo e uno dei valori medi più bassi di Estroversione, accompagnati da deviazioni standard relativamente elevate, segnalando una maggiore eterogeneità interna. Nel complesso, il GMM evidenzia la presenza di una possibile struttura categoriale nei dati, ma la descrive attraverso un numero relativamente elevato di componenti. Le differenze tra i profili medi consentono di individuare alcune configurazioni ricorrenti, mentre le deviazioni standard mostrano che la separazione tra i gruppi non è netta e che una quota di variabilità individuale permane anche all'interno delle classi.

Classe	Numerosità	Amicalità	Coscienziosità	Estroversione	Nevroticismo	Apertura
1	280	4.89 (0.42)	4.55 (0.80)	4.59 (0.76)	3.16 (1.15)	4.34 (0.66)
2	345	4.12 (0.94)	3.94 (1.00)	3.43 (1.10)	3.83 (1.22)	4.34 (0.71)
3	227	4.01 (0.97)	4.05 (1.05)	3.90 (1.04)	3.18 (1.23)	3.87 (0.85)
4	490	4.36 (0.59)	4.06 (0.86)	4.13 (0.90)	3.23 (1.03)	4.29 (0.64)
5	279	5.62 (0.32)	4.36 (0.89)	4.54 (0.78)	3.01 (1.09)	4.49 (0.65)
6	154	5.46 (0.41)	5.09 (0.53)	5.18 (0.59)	1.76 (0.57)	4.78 (0.62)
7	335	3.99 (0.92)	4.13 (1.05)	3.36 (1.12)	3.38 (1.25)	5.42 (0.47)
8	326	5.32 (0.40)	4.52 (0.89)	4.58 (0.79)	2.99 (1.06)	5.29 (0.45)

Tabella 4.2: Profili medi dei *Big Five* nelle classi individuate dal GMM. Tra parentesi sono riportate le deviazioni standard.

4.3.3 Parsimonious Gaussian Mixture Models

In modo analogo all'approccio descritto precedentemente, anche per i *Parsimonious Gaussian Mixture Models* è stata condotta una ricerca per determinare la configurazione più adeguata della struttura di gruppo. Coerentemente con la scelta metodologica discussa nel paragrafo, il numero di dimensioni latenti è stato mantenuto fisso pari a cinque ($q = 5$), in modo da preservare l'aderenza al costrutto teorico dei *Big Five*.

La stima è stata effettuata mediante il pacchetto `R pgmm` (McNicholas *et al.*, 2025), considerando le diverse specificazioni della matrice di covarianza e facendo variare il numero di classi tra 1 e 8. Per ridurre la dipendenza della procedura dai valori iniziali, aspetto rilevante nei modelli stimati tramite algoritmo AECM, le stime sono state ripetute utilizzando due diverse strategie di inizializzazione: una basata sull'algoritmo *k-means* e una basata sul raggruppamento gerarchico con metodo di Ward. Tra i modelli stimati, la combinazione associata al valore ottimale del criterio d'informazione BIC corrisponde alla specificazione `CUU` con $G = 4$ classi e $q = 5$ fattori latenti. La Tabella 4.3 riporta i profili medi e le deviazioni standard utilizzati per descrivere le quattro classi individuate. Rispetto alla soluzione ottenuta con i GMM, il modello parsimonioso restituisce una classificazione più compatta. La Classe 1 si distingue per i valori medi più elevati di Amicalità, Coscienziosità ed Estroversione e per il livello medio più basso di Nevroticismo. Le deviazioni standard risultano complessivamente contenute rispetto alle altre classi, suggerendo una maggiore omogeneità interna del gruppo. La Classe 2, che rappresenta il raggruppamento più numeroso del campione, presenta il valore medio più elevato di Apertura all'esperienza e un livello medio di Nevroticismo superiore rispetto alle altre classi; le deviazioni standard indicano inoltre una variabilità interna relativamente marcata, soprattutto per Estroversione, Nevroticismo e Coscienziosità. La Classe 3 si colloca su valori intermedi per gran parte dei domini, con valori relativamente elevati di Amicalità e Apertura. Infine, la Classe 4 presenta i valori medi più contenuti di Amicalità e Apertura all'esperienza, mantenendosi su livelli intermedi nelle altre dimensioni.

Al fine di valutare visivamente la distribuzione delle classi nello spazio latente, le 2436 unità statistiche sono state proiettate sullo spazio continuo generato dai *factor scores* dell'analisi fattoriale, utilizzando l'assegnazione alle quattro classi del PGMM come variabile di raggruppamento. La Figura 4.3 illustra le distribuzioni marginali dei punteggi latenti tramite *boxplot*, mentre la Figura 4.4 fornisce un focus bidimensionale sulle dimensioni del Nevroticismo e dell'Amicalità, arricchito dalle

ellissi di densità al 95%.

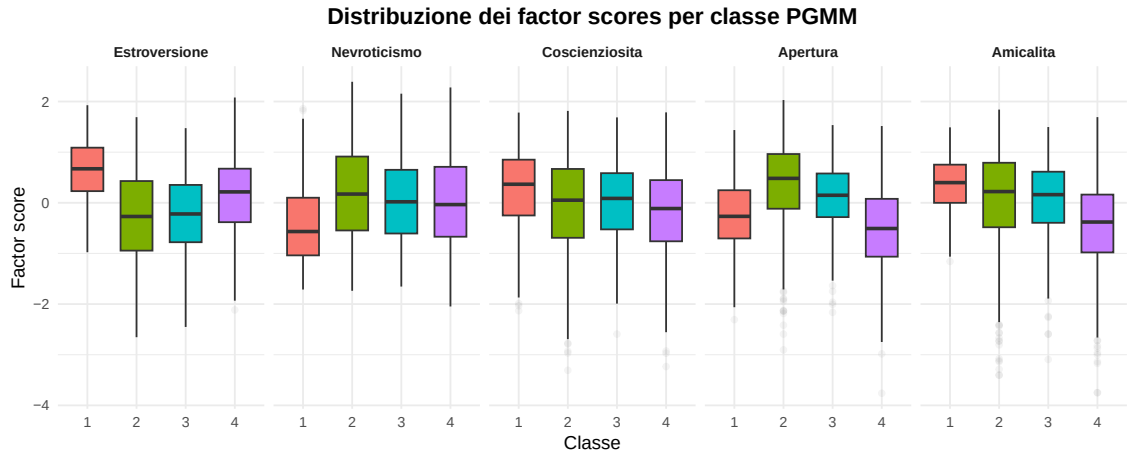

Figura 4.3: Distribuzione marginale dei *factor scores* per le quattro classi individuate dal modello PGMM.

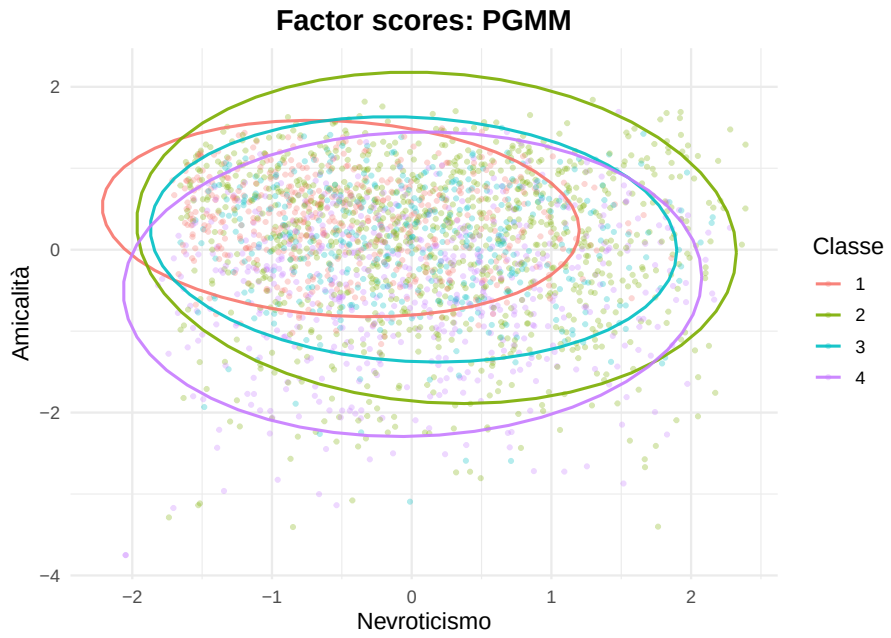

Figura 4.4: Proiezione bivariata dei *factor scores* di Nevroticismo e Amicalità. Le ellissi rappresentano le regioni di densità al 95% per le quattro classi PGMM.

Classe	Numerosità	Amicalità	Coscienziosità	Estroversione	Nevroticismo	Apertura
1	476	5.11 (0.56)	4.51 (0.85)	4.71 (0.75)	2.52 (1.00)	4.28 (0.62)
2	962	4.64 (0.99)	4.23 (1.03)	4.00 (1.19)	3.44 (1.25)	5.03 (0.72)
3	431	4.81 (0.73)	4.31 (0.88)	3.94 (0.96)	3.31 (1.08)	4.77 (0.56)
4	567	4.14 (0.90)	4.10 (0.95)	4.01 (0.99)	3.15 (1.14)	4.04 (0.78)

Tabella 4.3: Profili medi dei *Big Five* nelle classi individuate dal PGMM. Tra parentesi sono riportate le deviazioni standard.

Le rappresentazioni grafiche mostrano che le classi PGMM presentano differenze nei livelli medi dei fattori, ma anche una sovrapposizione non trascurabile tra le distribuzioni. Questo risultato è coerente con la natura dei dati analizzati: i tratti di personalità sono dimensioni continue e, di conseguenza, le classi latenti individuate dal modello devono essere interpretate come profili probabilistici di risposta, più che come gruppi nettamente separati nello spazio fattoriale.

4.3.4 Heteroscedastic Factor Mixture Analysis

L'ultimo modello esaminato è l'*Heteroscedastic Factor Mixture Analysis* (HFMA). Mantenendo la piena coerenza metodologica con le analisi precedenti, il numero di fattori latenti è stato fissato a priori a cinque per preservare l'aderenza all'impianto teorico dei *Big Five*, lasciando libero di variare unicamente il numero di gruppi. La stima dei parametri è stata effettuata avvalendosi del pacchetto R `FactMixtAnalysis` (Viroli, 2012) e, in maniera analoga agli approcci precedenti, sono stati considerati diversi valori iniziali. Il criterio d'informazione BIC ha condotto a una soluzione a due classi. Il primo gruppo comprende 858 osservazioni (pari al 35.2% del campione), mentre il secondo raggruppamento risulta maggioritario, includendo 1578 unità (64.8%). I relativi profili medi, accompagnati dalle deviazioni standard, sono riportati nella Tabella 4.4.

La soluzione HFMA restituisce una rappresentazione più sintetica dell'eterogeneità rispetto ai modelli precedenti. La Classe 1 presenta valori medi più bassi di Amicalità, Coscienziosità ed Estroversione, associati a un livello medio più elevato di Nevroticismo e a un valore leggermente più alto di Apertura all'esperienza. Le deviazioni standard indicano una variabilità interna non trascurabile, soprattutto per Nevroticismo, Estroversione e Coscienziosità.

La Classe 2 presenta invece valori medi più elevati di Amicalità, Coscienziosità ed Estroversione e un livello medio più basso di Nevroticismo. L'Apertura all'esperienza risulta leggermente inferiore rispetto alla Classe 1, pur mantenendosi su valori medi elevati. Rispetto alla Classe 1, le deviazioni standard appaiono generalmente più contenute per Amicalità, Coscienziosità ed Estroversione, suggerendo una maggiore omogeneità interna su queste dimensioni.

Oltre alle partizioni, è stata analizzata la matrice dei *loadings* stimata dal modello che, dopo rotazione obliqua di tipo *oblimin*, riproduce una struttura coerente con quella emersa nell'analisi fattoriale. La Tabella 4.5 mostra i principali carichi fattoriali, omettendo per chiarezza espositiva i valori inferiori a 0.30 in valore assoluto.

Classe	Numerosità	Amicalità	Coscienziosità	Estroversione	Nevroticismo	Apertura
1	858	3.94 (0.96)	3.78 (1.05)	3.40 (1.09)	3.50 (1.22)	4.79 (0.81)
2	1578	5.03 (0.61)	4.54 (0.80)	4.53 (0.82)	2.99 (1.15)	4.50 (0.78)

Tabella 4.4: Profili medi dei *Big Five* nelle classi individuate dall'HFMA. Tra parentesi sono riportate le deviazioni standard.

Item	F1	F2	F3	F4	F5
A1					0.401
A2					0.633
A3					0.720
A4					0.481
A5					0.562
C1			0.553		
C2			0.652		
C3			0.571		
C4			0.663		
C5			0.563		
E1	0.567				
E2	0.652				
E3	0.359			0.346	
E4	0.570				0.331
E5	0.405				
N1		0.860			
N2		0.817			
N3		0.671			
N4	-0.427	0.424			
N5		0.450			
O1				0.553	
O2				0.440	
O3				0.637	
O4	-0.345			0.368	
O5				0.507	

Tabella 4.5: Principali *loadings* ruotati della soluzione HFMA con $q = 5$. Sono riportati soltanto i carichi con valore assoluto superiore a 0.30.

La struttura dei carichi risulta fortemente coerente con l'impianto del questionario, riproducendo in modo nitido le cinque dimensioni originali dei *Big Five* pur in presenza di attesi e fisiologici carichi secondari per alcuni *item* (E3, E4, N4 e O4).

Così come è stato fatto per i PGMM, al fine di visualizzare la distribuzione delle due classi nello spazio latente, le unità statistiche sono state proiettate sui *factor scores* dell'analisi fattoriale, utilizzando l'assegnazione HFMA come variabile di raggruppamento. La Figura 4.5 mostra le distribuzioni marginali dei punteggi latenti per le due classi, mentre la Figura 4.6 rappresenta un focus bidimensionale sulle dimensioni del Nevroticismo e dell'Amicalità, arricchito dalle ellissi di densità al 95%. Le rappresentazioni grafiche mostrano che il modello HFMA sintetizza l'eterogeneità del campione in due profili principali. Anche in questo caso, le distribuzioni non risultano completamente separate nello spazio dei *factor scores*, confermando che le classi devono essere interpretate come profili probabilistici all'interno di uno spazio latente continuo, più che come gruppi nettamente disgiunti.

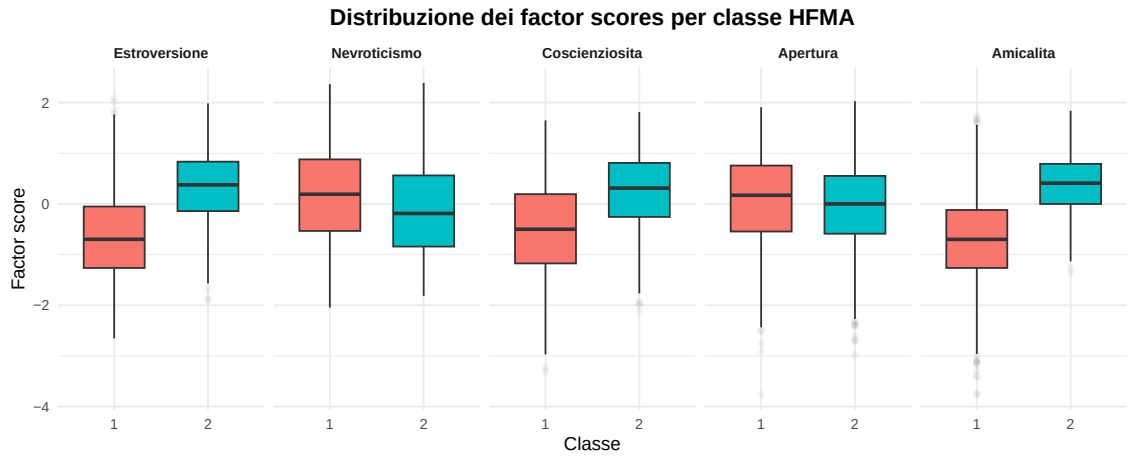

Figura 4.5: Distribuzione marginale dei *factor scores* per le due classi individuate dal modello HFMA.

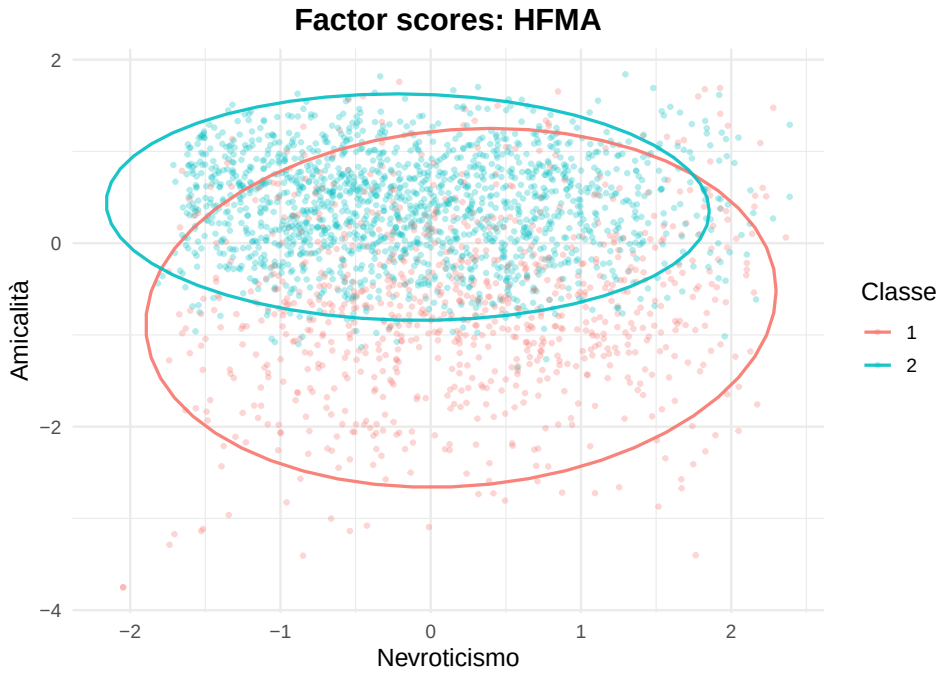

Figura 4.6: Proiezione bivariata dei *factor scores* di Nevroticismo e Amicalità per le due classi HFMA. Le ellissi rappresentano le regioni di densità al 95% per le due classi HFMA

Per valutare la relazione tra la classificazione ottenuta mediante HFMA e quella restituita dal PGMM, è stata costruita una tabella di contingenza tra le due partizioni. La Tabella 4.6 mostra che le due classi HFMA non sembrano corrispondere a una semplice aggregazione delle classi PGMM.

La prima classe HFMA è composta prevalentemente da individui appartenenti alla seconda classe PGMM, che rappresentano il 54.1% del gruppo, ma include anche quote non trascurabili della terza e quarta classe dei PGMM. La seconda classe

	Cluster PGMM			
Cluster HFMA	1	2	3	4
1	18 (2.1%)	464 (54.1%)	147 (17.1%)	229 (26.7%)
2	458 (29.0%)	498 (31.6%)	284 (18.0%)	338 (21.4%)

Tabella 4.6: Confronto tra la classificazione HFMA e la classificazione PGMM. Le percentuali sono calcolate per riga.

HFMA presenta invece una composizione più distribuita, con contributi provenienti da tutte le classi PGMM e, in particolare, dalle prima e dalla seconda.

Questa evidenza indica che i due modelli sintetizzano l'eterogeneità del campione in modo diverso. Il PGMM produce una segmentazione più articolata, individuando quattro profili latenti, mentre l'HFMA riconduce la variabilità tra individui a due profili principali all'interno di una struttura fattoriale comune. In questo senso, la soluzione HFMA non deve essere letta come una mera riduzione del numero di classi PGMM, ma come una diversa rappresentazione della relazione tra eterogeneità categoriale e struttura latente continua.

A conferma di questa lettura, l'ARI tra le due classificazioni è pari a 0.026, valore che segnala una concordanza molto contenuta tra le partizioni. Tale indice deve tuttavia essere interpretato con cautela, poiché i due modelli selezionano un numero diverso di classi; pertanto, più che misurare una mancata corrispondenza in senso assoluto, il confronto suggerisce che PGMM e HFMA mettono in evidenza livelli diversi della stessa struttura latente. Per comprendere appieno il significato di queste diverse stratificazioni, e stabilire quale dei due approcci descriva meglio la reale natura comportamentale dei soggetti analizzati, si renderebbe necessario il parere e l'ausilio di specifiche competenze psicologiche e cliniche.

4.4 Considerazioni finali

L'applicazione dei modelli ai dati reali consente di trarre alcune considerazioni sulla natura dell'eterogeneità latente presente nel *dataset* `bfi`. Il risultato più evidente riguarda la presenza di una struttura continua ben definita: sia l'analisi esplorativa sia la soluzione fattoriale confermano infatti l'esistenza di una configurazione coerente con le cinque dimensioni teoriche dei *Big Five*.

Accanto a questa struttura continua, i modelli a mistura individuano la possibile presenza di una struttura categoriale. La presenza di gruppi latenti emerge in tutte le specificazioni che ammettono una componente discreta, seppure con diverso grado di articolazione. Il GMM restituisce una segmentazione più frammentata, individuando otto componenti; il PGMM, grazie alla struttura fattoriale parsimoniosa, riconduce l'eterogeneità a quattro profili; infine, l'HFMA sintetizza ulteriormente la variabilità del campione in due classi principali, mantenendo una struttura fattoriale comune.

Questa evidenza suggerisce che la variabilità individuale nelle risposte non sia completamente riconducibile a un'unica popolazione omogenea, ma possa essere descritta anche attraverso profili medi differenti. Tuttavia, tali classi non devono essere interpretate come categorie psicologiche rigide o tipi di personalità nettamente

separati. Esse rappresentano piuttosto profili latenti di risposta, utili a sintetizzare configurazioni ricorrenti nei punteggi osservati, senza implicare necessariamente l'esistenza di gruppi naturalmente distinti in senso sostantivo.

Da questo punto di vista, i modelli che combinano mistura e struttura fattoriale risultano particolarmente adatti agli obiettivi del lavoro: essi consentono infatti di mantenere l'interpretazione dei tratti latenti continui e, al tempo stesso, di valutare la presenza di sottopopolazioni. Nel caso dei dati BFI, la componente continua rimane l'elemento interpretativo principale, ma i modelli a mistura mostrano che tale struttura può essere arricchita dalla presenza di profili latenti differenziati.

Conclusioni

Il presente lavoro ha avuto come obiettivo lo studio di modelli statistici in grado di descrivere l'eterogeneità non osservata attraverso l'integrazione di due diverse forme di latenza: una latenza continua, rappresentata dai fattori, e una latenza categoriale, rappresentata dalle classi o gruppi latenti. La tesi si è quindi sviluppata a partire dal confronto tra l'Analisi Fattoriale, i modelli a mistura finita e i modelli che combinano entrambe le prospettive all'interno della famiglia del *Factor Mixture Modelling*.

La trattazione teorica ha messo in evidenza come l'Analisi Fattoriale e i modelli a mistura affrontino due modi diversi di rappresentare la struttura latente dei dati. L'Analisi Fattoriale consente di sintetizzare la covariazione tra molte variabili osservate attraverso un numero limitato di fattori latenti continui, offrendo una rappresentazione parsimoniosa e interpretabile dei costrutti sottostanti. Tuttavia, essa presuppone che la stessa struttura fattoriale sia valida per l'intera popolazione. I modelli a mistura finita introducono invece una variabile latente categoriale, che permette di rappresentare la popolazione come composta da più gruppi non osservati. Questa maggiore flessibilità comporta però un incremento del numero di parametri e può rendere la stima instabile, soprattutto quando la dimensione dello spazio osservato aumenta.

I modelli a mistura fattoriale nascono proprio dall'esigenza di integrare queste due forme di latenza in un'unica cornice probabilistica. Essi permettono infatti di descrivere simultaneamente una struttura fattoriale, legata ai tratti o dimensioni continue sottostanti, e una struttura di classe, legata alla presenza di gruppi latenti. In questo senso, il *Factor Mixture Modelling* non deve essere inteso come una semplice estensione tecnica dell'Analisi Fattoriale o dei modelli a mistura, ma come una classe di modelli capace di combinare riduzione della dimensionalità, interpretazione dei costrutti latenti e rappresentazione della struttura di gruppo.

All'interno di questa famiglia, la distinzione tra filone statistico e filone psicometrico si è rivelata particolarmente rilevante. Il filone statistico nasce soprattutto dall'esigenza di ottenere una modellazione più parsimoniosa della covarianza in contesti multivariati o ad alta dimensionalità. Il filone psicometrico, invece, pone maggiore attenzione al significato dei costrutti latenti e alla loro comparabilità tra gruppi, come emerge nel caso dell'HFMA e del principio di invarianza di misurazione. Modelli formalmente vicini possono quindi rispondere a obiettivi diversi: classificazione e parsimonia parametrica da un lato, interpretabilità dei fattori e confronto tra sottopopolazioni dall'altro.

Lo studio di simulazione ha permesso di valutare il comportamento dei diversi approcci in condizioni controllate. I risultati mostrano che la scelta del modello più adeguato dipende dalla struttura latente che si intende recuperare. Quando il pro-

cesso generatore è descritto principalmente da fattori continui, l'Analisi Fattoriale rappresenta una soluzione efficace, poiché riesce a recuperare correttamente il sottospazio latente e fornisce una descrizione semplice e interpretabile. Quando invece la struttura generatrice è dominata dalla presenza di gruppi, i modelli a mistura gaussiana risultano più coerenti con il meccanismo sottostante, pur mostrando maggiore sensibilità alla dimensionalità e alla separazione tra le classi.

Nei casi in cui sono presenti sia fattori continui sia classi latenti, i modelli a mistura fattoriale risultano i più adeguati dal punto di vista concettuale, poiché consentono di rappresentare entrambe le componenti della latenza. Le simulazioni mostrano tuttavia che anche questi modelli richiedono cautela: la qualità dei risultati dipende dalla forza del segnale, dalla numerosità campionaria, dalla dimensionalità e dalla corretta specificazione della struttura latente. In particolare, quando il segnale è debole o quando le condizioni del modello sono violate, la selezione del numero di fattori e di classi può diventare meno stabile. Questo aspetto evidenzia che l'uso dei criteri automatici di selezione deve essere sempre accompagnato da una valutazione della stabilità e dell'interpretabilità delle soluzioni.

L'applicazione ai dati reali ha consentito di verificare queste considerazioni in un contesto empirico reale. Il *dataset* `bfi`, costruito sulla base del questionario dei *Big Five*, presenta una struttura teoricamente fattoriale e si presta quindi naturalmente all'Analisi Fattoriale. La soluzione a cinque fattori ha confermato la presenza di una struttura continua ben interpretabile, coerente con le dimensioni teoriche di Amicalità, Coscienziosità, Estroversione, Nevroticismo e Apertura all'esperienza. Questo risultato mostra che, nei dati analizzati, la latenza continua costituisce l'elemento interpretativo principale.

L'introduzione dei modelli a mistura ha però mostrato che la sola struttura fattoriale non esaurisce completamente la complessità dei dati. Il GMM ha individuato una segmentazione articolata del campione, descrivendo le risposte attraverso otto componenti. Il PGMM ha fornito una rappresentazione più parsimoniosa, riconducendo la struttura di gruppo a quattro profili latenti all'interno di una rappresentazione fattoriale a bassa dimensionalità. L'HFMA, infine, ha sintetizzato ulteriormente la struttura di classe in due gruppi principali, mantenendo una struttura fattoriale comune. Questi risultati mostrano che modelli diversi possono mettere in evidenza aspetti differenti della struttura latente: il GMM privilegia la componente di classificazione, il PGMM offre un compromesso tra classi e fattori, mentre l'HFMA propone una lettura più sintetica e maggiormente legata alla comparabilità dello spazio fattoriale.

Il confronto tra PGMM e HFMA ha confermato che le due soluzioni non rappresentano semplicemente livelli diversi di dettaglio della stessa classificazione. La partizione HFMA non coincide infatti con una semplice aggregazione delle classi PGMM, ma restituisce una diversa rappresentazione del rapporto tra fattori continui e classi latenti. Questo risultato è rilevante perché mostra che la scelta tra modelli non può dipendere soltanto dal numero di gruppi individuati, ma deve essere guidata dalla domanda di ricerca e dal tipo di latenza che si intende privilegiare.

Nel complesso, i risultati ottenuti suggeriscono che la latenza non debba essere pensata necessariamente in termini esclusivamente continui o esclusivamente categoriali. In molti contesti reali, gli individui possono differire lungo dimensioni latenti continue e, al tempo stesso, essere descritti attraverso profili di classe. I model-

li di *Factor Mixture Modelling* risultano quindi utili proprio perché permettono di integrare queste due letture all'interno di un'unica struttura modellistica.

La principale conclusione del lavoro è che non esiste un modello sempre preferibile, ma piuttosto una famiglia di strumenti da scegliere in funzione dell'obiettivo dell'analisi. L'Analisi Fattoriale rimane appropriata quando l'interesse principale è descrivere costrutti latenti continui in una popolazione considerata omogenea. I modelli a mistura gaussiana sono invece indicati quando l'attenzione è rivolta soprattutto all'individuazione di gruppi latenti. I modelli a mistura fattoriale rappresentano una soluzione più flessibile quando si ritiene plausibile la compresenza di fattori continui e classi latenti.

In conclusione, l'integrazione tra fattori continui e classi latenti offre una prospettiva efficace per analizzare dati multivariati complessi. Essa consente di mantenere l'interpretabilità propria dei modelli fattoriali e, al tempo stesso, di riconoscere la possibile presenza di una struttura di gruppo. Tuttavia, l'interpretazione delle classi deve rimanere prudente: esse non rappresentano necessariamente categorie naturali o tipologie sostantivamente distinte, ma profili probabilistici utili a sintetizzare configurazioni ricorrenti nei dati.

Bibliografia

Baek J.; McLachlan G. J.; Flack L. K. (2010). Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualisation of high-dimensional data. *rts*, **100**, 5.

Bartholomew D. J.; Knott M.; Moustaki I. (2011). *Latent variable models and factor analysis: A unified approach*. John Wiley & Sons.

Bellman R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, **60**(6), 503–515.

Bollen K. A. (1989). A new incremental fit index for general structural equation models. *Sociological methods & research*, **17**(3), 303–316.

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

Dempster A. P.; Laird N. M.; Rubin D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, **39**(1), 1–22.

Ghahramani Z.; Hinton G. E. *et al.* (1996). The em algorithm for mixtures of factor analyzers. Relazione tecnica, Technical Report CRG-TR-96-1, University of Toronto.

Goldberg L. R. (1992). The development of markers for the big-five factor structure. *Psychological assessment*, **4**(1), 26.

Kaiser H. F. (1974). An index of factorial simplicity. *psychometrika*, **39**(1), 31–36.

Lubke G. H.; Muthén B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, **10**(1), 21.

McLachlan G. J.; Peel D. (2000). *Finite mixture models*. John Wiley & Sons.

McLachlan G. J.; Peel D.; Bean R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, **41**(3-4), 379–388.

McNicholas P. D.; Murphy T. B. (2008). Parsimonious gaussian mixture models. *Statistics and Computing*, **18**(3), 285–296.

McNicholas P. D.; Murphy T. B. (2010). Model-based clustering of microarray expression data via latent gaussian mixture models. *Bioinformatics*, **26**(21), 2705–2712.

McNicholas P. D.; ElSherbiny A.; McDaid A. F.; Murphy T. B. (2025). *pgmm: Parsimonious Gaussian Mixture Models*. R package version 1.2.8.

Meng X.-L.; Van Dyk D. (1997). The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **59**(3), 511–567.

Meredith W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, **58**(4), 525–543.

Montanari A.; Viroli C. (2010). Heteroscedastic factor mixture analysis. *Statistical Modelling*, **10**(4), 441–460.

Schwarz G. (1978). Estimating the dimension of a model. *The annals of statistics*, pp. 461–464.

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

Spearman C. (1904). " general intelligence," objectively determined and measured. *The American Journal of Psychology*, **15**(2), 201–292.

Thomson G. H. (1939). The factorial analysis of human ability.

Thurstone L. L. (1931). Multiple factor analysis. *Psychological review*, **38**(5), 406.

Tipping M. E.; Bishop C. M. (1999a). Mixtures of probabilistic principal component analyzers. *Neural computation*, **11**(2), 443–482.

Tipping M. E.; Bishop C. M. (1999b). Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **61**(3), 611–622.

Viroli C. (2012). *FactMixtAnalysis: Factor Mixture Analysis with covariates*. R package version 1.0.

William Revelle (2026). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.6.1.

Ringraziamenti

«Nel mezzo del cammin» del mio percorso di studi, mi sono ritrovato lontano da casa, in una città nuova, con il cuore diviso tra ciò che lasciavo e ciò che speravo di costruire. Da Catania a Padova, questo cammino universitario è stato, a suo modo, una personale Commedia: ci sono stati momenti di smarrimento, giorni simili a una «selva oscura», salite faticose, incontri preziosi e, infine, la luce di un traguardo che per tanto tempo avevo soltanto immaginato.

Questo viaggio non è stato fatto soltanto di esami, lezioni, libri e notti passate a studiare. È stato un percorso di crescita, di distanze, di mancanze, di cadute e ripartenze. E come il Poeta non attraversa i tre regni da solo, anche io non sarei arrivato fin qui senza le persone che, trasformandosi in guide e compagni di viaggio, mi hanno sostenuto e accompagnato.

Desidero innanzitutto ringraziare il mio personale Virgilio, la Professoressa Giovanna Menardi che, più di ogni altra, con il suo contagioso entusiasmo è riuscita a trasferire in me la sua profonda passione per la statistica. Come il «duca» dantesco, mi ha seguito nella realizzazione di questa tesi con infinita disponibilità, attenzione e competenza. La ringrazio per essere stata guida illuminata in questo ultimo tratto del percorso, per i consigli e le correzioni, insegnandomi a scorgere l'ordine e la bellezza anche laddove la materia appariva complessa e dandomi gli strumenti per superare ogni pendenza accademica. E, concedendomi una nota scherzosa, posso assicurarle che non dimenticherò mai il soprannome di "Mr. Probability" che mi ha simpaticamente affibbiato.

In ogni cammino verso la cima, ci sono anime con cui si condivide la fatica della salita. Il mio grazie va a Edoardo, vero e proprio Stazio di questo viaggio, con cui ho condiviso praticamente ogni materia e ogni ansia prima degli esami. A dire il vero, tra i libri sono stati più i momenti dedicati a scambiarci frasi motivazionali che quelli di studio puro, ma quando c'era da fare sul serio siamo sempre stati impeccabili. Studiare insieme ha reso il peso meno gravoso e la meta più vicina. Insieme a lui, ringrazio tutti i compagni di corso incontrati tra le aule: anime tese verso lo stesso obiettivo, che hanno dimostrato come nessun cammino sia davvero individuale, ma sia fatto dei passi di chi ci cammina accanto.

Ogni pellegrino ha bisogno di un'oasi di ristoro prima di riprendere la via. Ringrazio Sharon, che per un tratto importante di questa strada è stata presenza, supporto e ascolto profondo; anche se oggi le nostre strade si sono divise, le sono grato per il bene condiviso e per aver reso più leggero un momento cruciale del percorso. Un pensiero va anche a Noemi e Katia, le mie coinquiline: vi ringrazio per aver condiviso la quotidianità padovana, rendendo le giornate più leggere e la casa un ambiente sereno in cui vivere.

Anche quando l'esilio da fuorisede sembrava farsi più duro, il richiamo della mia terra ha sempre trovato voci pronte a ristorarmi, come l'incontro con Casella sul lido del Purgatorio. Giorgio, sei stato molto più di un amico in questi anni: un punto fermo incrollabile, sempre pronto a chiamarmi ogni sera per rassicurarmi, per dirmi che ce l'avrei fatta e che potevo contare su di te; sei, a tutti gli effetti, di famiglia. Bruno, a te devo la leggerezza e la spensieratezza; nonostante la distanza e la tua vita in America, mi fai sempre sentire accolto, e il fatto che tu sia persino venuto a trovarmi è un gesto che porto nel cuore. Insieme a voi, ringrazio Michele, Alessandro e tutti gli amici storici. Siete stati la certezza di un abbraccio che non conosce confini, la dimostrazione che casa non è un luogo geografico, ma il calore delle persone che ti aspettano e continuano a credere in te, impedendoti di sentirti solo.

Il ringraziamento più profondo va alla mia famiglia, primo e incrollabile punto fermo. A mia mamma, per la forza immensa e l'amore costante capace di annullare ogni distanza. Ai miei fratelli, per il legame indissolubile che ci unisce e che mi ricorda, sempre, chi sono e da dove vengo. Un pensiero colmo d'affetto va anche a zio Salvo e Tiziana, per essere stati una presenza preziosa e rassicurante lungo tutto questo cammino. Un grazie speciale a Nonno Salvatore, mio nobile Cacciaguida: il tuo sostegno in questo ultimo anno, prezioso anche dal punto di vista economico, unito all'incoraggiamento delle tue poesie e al tuo «empireo sempre in movimento», mi ha trasmesso la serenità e la fermezza necessarie per affrontare questa importante tappa della mia vita.

Ma il pensiero più grande, l'emozione più profonda che corona l'intero viaggio, si eleva verso l'alto. Questo traguardo è interamente dedicato a te, papà, mia vera Beatrice.

Sei stato colui che, più di tutti, ha creduto in me, la persona che mi ha accompagnato per tutta la vita con amore e fiducia assoluta. Ti avevo promesso che mi avresti visto su questo palcoscenico, che avresti assistito al frutto di tanti sacrifici, ma il tempo mi ha privato della tua presenza fisica proprio nell'ultimo anno, quando la meta era ormai vicina. Oggi questa laurea non è solo un punto d'arrivo, ma una promessa mantenuta. È il mio modo di dirti che ce l'ho fatta, il traguardo che ti dovevo, la luce che squarcia il buio della tua assenza.

Se alla fine del lungo viaggio si scrive «e quindi uscimmo a riveder le stelle», io quelle stelle oggi le guardo sapendo che tu sei la più luminosa tra loro. Questa tesi e questo giorno sono per te, papà, che continui a guidare i miei passi, da ovunque tu mi stia guardando.