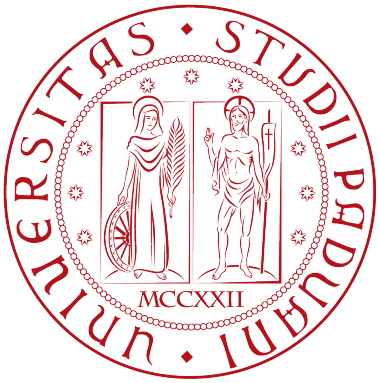
Università degli Studi di Padova
Dipartimento di Filosofia, Sociologia, Pedagogia e Psicologia Applicata
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Magistrale in
Psicologia Clinico-dinamica



Tesi di Laurea Magistrale

Un modello di mistura CFA per campioni eterogenei: uno studio di simulazione

Mixture CFA model for heterogeneous samples:
A simulation study

Relatore:
Prof. Antonio Calcagnì

Laureando: Niccolò Cao
Matricola: 2015613

Anno Accademico 2021/2022

Indice

Introduzione

In psicometria, si possono distinguere due approcci per adattare modelli statistici ai dati: l'approccio variabile-centrico e l'approccio persona-centrico (Muthén & Muthén, 2000). L'approccio variabile-centrico assume che tutte le unità campionarie raccolte siano estratte da una singola popolazione, consentendo di stimare un solo vettore di parametri per l'intero campione (Morin et al., 2018). Lo scopo di questo approccio consiste nello spiegare le relazioni tra variabili di interesse in una popolazione scelta (Howard & Hoffman, 2018). Alcuni metodi analitici tipici di questo approccio sono i seguenti: la regressione lineare, l'analisi fattoriale (Factor Analysis, FA) ed i modelli ad equazioni strutturali (Structural Equation Modeling, SEM;Howard e Hoffman, 2018). Al contrario, l'approccio persona-centrico assume che i soggetti possano derivare da differenti sotto-popolazioni, caratterizzate da distinti insiemi di parametri (Morin et al., 2018). Questo approccio è focalizzato sull'identificare sottogruppi di unità campionarie, che emergono dalla totalità del campione per le loro caratteristiche peculiari, in modo da dar loro risalto nel processo di analisi dei dati (Howard & Hoffman, 2018). Alcune tecniche statistiche, che rientrano in questo approccio, sono: l'analisi delle classi latenti (Latent Class Analysis, LCA), la Cluster Analysis ed i modelli di mistura (Finite Mixture Models) (Howard & Hoffman, 2018).

L'approccio persona-centrico sta crescendo di popolarità negli ultimi anni, in quanto propone di analizzare un campione in modo più particolareggiato rispetto all'approccio variabile-centrico e gestirne le specificità (Howard & Hoffman, 2018). Tra i modelli che rientrano nel solco dell'approccio persona-centrico vi sono i Factor Mixture Models (FFM). I FMM vengono sempre più adottati nelle scienze psicologiche e sociali (Y. Wang et al., 2021), in particolare per gestire il problema l'eterogeneità campionaria, che rappresenta un rischio per la validità dei risultati (Becker et al., 2013). Secondo Lubke e Muthén (2005), I FFM rappresentano una combinazione tra il Common Factor Model (Thurstone, 1947) ed un modello a classi latenti (Lazarsfeld & Henry, 1968). In generale, i FMM risultano una strategia analitica ottimale per affrontare il problema

dell'eterogeneità campionaria, soprattutto nel caso in cui non sia nota in anticipo al/alla ricercatore/ricercatrice (Sawatzky et al., 2009).

Nel presente lavoro, verrà proposto un modello di mistura della famiglia FMM: il Mix-CFA-EFA, un modello di mistura che combina un modello di CFA con un modello di EFA, per gestire l'eterogenità campionaria non nota *a priori*.

In particolare, il presente documento è articolato come segue: nel Capitolo 1 si provvede un'introduzione al problema dell'eterogeneità non-osservata e ai metodi proposti per la sua gestione; nel Capitolo 2 vengono definiti i modelli CFA ed EFA utilizzati nel modelli di mistura, con la stima tramite l'algoritmo Expectation-Maximization; nel Capitolo 3 il modello Mix-CFA-EFA viene specificato con la stima dei parametri tramite l'algoritmo Expectation-Maximization; nel Capitolo 4 si presenta uno studio di simulazione ed un'applicazione del modello Mix-CFA-EFA su dati reali.

Capitolo 1

Eterogeneità campionaria nei modelli SEM e FA

1.1 L'eterogeneità campionaria: definizione e problematiche

1.1.1 Eterogeneità e validità di misurazione

Nelle scienze sociali e psicologiche, l'applicazione di modelli statistici ai dati richiede, di solito, che il ricercatore assuma, implicitamente, l'appartenenza di tutte le unità campionarie ad una stessa popolazione di origine (Kiefer et al., 2022). Tuttavia, tale assunzione è spesso irrealistica, in quanto gli individui tendono a mostrare una certa eterogeneità nelle loro percezioni e valutazioni di costrutti latenti (Ansari et al., 2000), che può dipendere da molteplici fattori: personali, contestuali, demografici e così via (Sawatzky et al., 2009).

Recentemente è stato sviluppato il quadro teorico Draper-Lindley-de Finetti (DLD) per la validità di misurazione in psicologia (Zumbo, 2006). Secondo il DLD, la validità di misurazione può essere messa in discussione da due ordini di problemi: problemi di misurazione e problemi di campionamento (Zumbo, 2006). I problemi di misurazione riguardano la scambiabilità delle variabili osservate (i.e., items di una sotto-scala), mentre i problemi di campionamento di riferiscono al grado in cui la struttura di misurazione sia appropriata per tutti i soggetti (Sawatzky et al., 2018). Per non incorrere in eventuali problemi di campionamento è necessario assicurarsi che le unità campionate e non-campionate dalla popolazione siano scambiabili tra loro (Zumbo, 2006). In particolare,

la scambiabilità dipende dalla presenza di modalità di risposta ad un test, che siano simili tra unità statistiche (Zumbo, 2006). L'assunto di scambiabilità è cruciale, in quanto rappresenta il fondamento per produrre inferenze che siano generalizzabili alla popolazione di interesse (Sawatzky et al., 2009). Qualora, la scambiabilità tra osservazioni non sia mantenuta, si parla di un campione eterogeneo rispetto ai parametri del modello di misurazione (Sawatzky et al., 2009).

In psicologia, svariati lavori empirici hanno mostrato come l'eterogeneità campionaria possa non essere identificata rispetto a determinati costrutti psicologici, qualora ci si affidi esclusivamente a variabili osservate che implicano differenze campionarie note *a priori* (e.g., Cohen e Bolt, 2005; De Ayala et al., 2002; Sawatzky et al., 2009). Una possibile soluzione è rappresentata dall'uso di modelli statistici per individuare differenze tra osservazioni campionarie di origine sconosciuta, come i modelli FMM o, più in generale, modelli di mistura (Sawatzky et al., 2009).

1.1.2 Il problema dell'eterogeneità campionaria

In psicometria, il problema dell'eterogeneità campionaria è stato affrontato per la prima volta in modo sistematico da Muthén (1989). In particolare, Muthén (1989) pone l'accento sulla frequente implausibilità dell'assumere che tutti gli individui di un campione condividano lo stesso insieme di valori parametrici. Pertanto, qualora non si possa assumere omogeneità tra le osservazioni campionarie rispetto ai parametri del modello di misurazione, si parla di eterogeneità campionaria (Muthén, 1989; Sawatzky et al., 2009). Ad esempio, si immagini che, per valutare le conoscenze relative a temi avanzati di matematica, sia raccolto un campione di studenti presenti alle lezioni di un corso universitario della facoltà di statistica. Si può supporre che chi conduce la ricerca possa essere portato/a a pensare che i presenti in aula siano solo studenti di statistica, assumendo, di conseguenza, omogeneità tra le osservazioni campionarie. Tuttavia, quel particolare corso universitario, per motivi ignoti ai/alle ricercatori/ricercatrici, viene frequentato da una proporzione, ristretta ma non indifferente, di studenti di psicologia. A questo punto, sostenere l'omogeneità delle unità campionarie risulterebbe piuttosto difficile, considerando che, solitamente, i programmi in psicologia e statistica richiedono differenti gradi di conoscenze matematiche. Al contrario, potrebbe essere maggiormente verosimile che l'assunzione di omogeneità possa essere mantenuta con minor incertezza, applicando separatamente un modello di analisi fattoriale a ciascuno dei due gruppi.

L'esempio appena visto è molto banale, ma può essere generalizzato a situazioni più sfumate, in cui la presenza di più popolazioni nei dati può risultare pressoché

imprevedibile (Wedel & Kamakura, 2000). Pertanto, l'eterogeneità campionaria può rappresentare un problema insidioso. Difatti, è possibile che i/le ricercatori/ricercatrici possano non essere consapevoli di come alcune unità statistiche si aggreghino in sottogruppi significativamente diversi dal resto dei partecipanti.

Dalla letteratura scientifica sul tema, emergono due differenti tipi di eterogeneità campionaria: l'eterogeneità osservata e non-osservata (Becker et al., 2013). L'eterogeneità osservata è definita da una conoscenza *a priori* da parte del/della ricercatore/ricercatrice rispetto a come la popolazione di interesse, da cui deriva il campione, si suddivida in sotto-popolazioni (Lubke & Muthén, 2005). Il termine sotto-popolazione si riferisce alle differenti popolazioni che vengono rappresentate nel campione attraverso i relativi sotto-campioni e che sono definite dalla presenza di specifiche variabili osservate (Lubke & Muthén, 2005). Il termine "sotto-campione", invece, indica un *cluster* di osservazioni omogenee all'interno di un campione eterogeneo. Le variabili osservate, che partizionano la popolazione del campione, determinano distinti valori parametrici stimabili su ogni "sotto-campione" per ogni sotto-popolazione di riferimento (Lubke & Muthén, 2005; Muthén, 1989). Solitamente, tali variabili osservate, o covariate, consistono in variabili demografiche e variabili psicografiche (Sawatzky et al., 2009; Wedel & Kamakura, 2000). Generalmente, questo tipo di eterogeneità viene gestita sviluppando uno modello specifico per ogni sotto-campione che deriva da ciascuna sotto-popolazione nota (Jöreskog, 1971).

L'eterogeneità non-osservata si riferisce, invece, alla mancata conoscenza *a priori* delle sotto-popolazioni, da cui deriva il campione complessivo ed, anche, delle relative variabili che partizionano la popolazione oggetto di studio (Becker et al., 2013). Inoltre, non solo non è nota la proporzione delle sotto-popolazioni nel campione, ma è parimenti sconosciuta l'appartenenza delle singole osservazioni alle sotto-popolazioni (Lubke & Muthén, 2005). Difatti, la presenza di eterogeneità non-osservata rappresenta un rischio per il/la ricercatore/ricercatrice, in quanto le origini di tale eterogeneità non sono manifeste e potrebbero essere imprevedibili (Wedel & Kamakura, 2000). Fattori, noti o non noti, legati alla cultura di appartenenza, allo sviluppo, alle differenze nella personalità o al contesto di vita possono indurre i singoli soggetti ad interpretare uno stesso item in modo differente, generando un campione caratterizzato da soggetti con modalità di risposta eterogenee (Sawatzky et al., 2009). Inoltre, Julian (2001) ha ipotizzato che i campioni di convenienza, nei termini di costi di ricerca, possano essere affetti da eterogeneità campionaria, in particolare se la correlazione intra-classe è alta. Un'altra possibile fonte di eterogeneità è data dal faking, ovvero uniformare le proprie risposte a un test o un questionario descrivendosi in modo da facilitare il raggiungimento di certi obiettivi personali (Ziegler et al., 2011), come mostrato in Ziegler et al. (2015).

Per questa tipologia di eterogeneità, la strategia maggiormente adottata consiste nel classificare simultaneamente le osservazioni in differenti sotto-popolazioni, ciascuna descritta da un modello SEM o FA, tramite un approccio basato sui FMM (Y. Wang et al., 2021; Yung, 1997). Ad esempio, tramite i FMM, nell'ambito della psicopatologia, sono state individuate 3 sotto-popolazioni qualitativamente differenti che caratterizzano la sensibilità all'ansia (Bernstein et al., 2013).

1.1.3 Il bias indotto da eterogeneità campionaria

Nel caso dei modelli SEM e FA, le evidenze mostrano che non tener conto delle sorgenti di eterogeneità può distorcere le analisi statistiche e condurre a inferenze e conclusioni errate (Ansari et al., 2000; Becker et al., 2013; Yung, 1997). In particolare, l'eterogeneità campionaria può condurre a: stime gonfiate di attendibilità di misurazione, segni delle covarianze fattoriali errati, attenuazioni nell'adattamento del modello e negli errori standard (Ansari et al., 2000, 2002); distorsioni nelle stime dei parametri; stime non significative a livello del sotto-campione, ma significative nel campione complessivo; segni opposti nelle stime tra sotto-campioni che annullano effetti significativi nel campione totale; ridotto potere predittivo del modello ed errori di I e II tipo (Becker et al., 2013). Becker et al. (2013) sottolineano come la presenza di eterogeneità non inclusa nelle analisi SEM e FA può rappresentare una minaccia per:

1. la validità interna del modello, in quanto, non considerando la differenza tra sotto-popolazioni, il modello è incompleto;
2. la validità delle conclusioni statistiche, a causa di un aumento degli errori standard delle stime e una riduzione delle dimensioni dell'effetto medie;
3. l'attendibilità del modello, se i sotto-campioni sono caratterizzati da differenti pattern di correlazione o varianze negli errori;
4. la validità esterna, siccome i risultati ottenuti dal campione complessivo non sono rappresentativi delle sotto-popolazioni e non possono essere, quindi, generalizzati ad un'unica popolazione.

Inoltre, è stato dimostrato che se viene applicato un modello SEM o FA ad un campione senza gestirne l'eterogeneità, gli indici di adattamento tradizionali non sono utili per individuare l'eterogeneità non-osservata (Jedidi et al., 1997b). Pornprasertmanit et al. (2014) hanno dimostrato che ignorare le sotto-popolazioni da cui provengono i dati diminuisce l'adattamento dei modelli CFA, specialmente quando la correlazione intra-classe è alta.

1.2 Metodi di gestione dell'eterogeneità campionaria

Nell'ambito dei SEM e della FA, sono stati sviluppati metodi che permettono di considerare nelle analisi l'eterogeneità campionaria osservata e non-osservata. Nel seguito viene presentata una rassegna delle principali soluzioni avanzate per trattare entrambe le tipologie di eterogeneità, con maggior attenzione rispetto all'eterogeneità non-osservata.

1.2.1 Metodi per l'eterogeneità osservata

Rispetto alla gestione dell'eterogeneità osservata, i primi modelli statistici sviluppati sono i SEM multi-gruppo (Jöreskog, 1971; Sörbom, 1974). Tali modelli consentono di suddividere il campione in un ristretto numero di sotto-campioni in base ad una o più covariate e stimare un modello SEM separato per ogni sotto-campione. I modelli SEM multi-gruppo permettono di stimare differenze nelle strutture di covarianza (i.e., i parametri e le matrici di varianza-covarianza dei fattori e degli errori del modello) tra gruppi (Ansari et al., 2000). L'approccio multi-gruppo, inoltre, può essere anche applicato per esplorare l'invarianza di misurazione tra molteplici gruppi (Van de Schoot et al., 2012).

Un altro approccio si basa sull'uso di una famiglia di test derivati dalla funzione punteggio (Merkle & Zeileis, 2013). Per verificare la presenza di eterogeneità campionaria, si effettua un test di questa famiglia a partire dai risultati di un modello SEM o FA già stimato, che permetta di attribuire tale eterogeneità all'effetto di una o più variabili osservate (Merkle & Zeileis, 2013; Merkle et al., 2014; T. Wang et al., 2014). Tuttavia, questa famiglia di test permette di verificare l'ipotesi di eterogeneità campionaria, ma non di quantificare direttamente tale eterogeneità.

Infine, Brandmaier et al. (2013) hanno proposto il metodo SEM trees per individuare covariate che predicono differenze nella stima dei parametri del modello strutturale. Il metodo SEM trees si basa su un partizionamento ricorsivo dei dati in sottoinsiemi definiti da variabili osservate, in modo da individuare sottogruppi omogenei nelle stime dei parametri ottenute dal modello ipotizzato (Arnold et al., 2020; Brandmaier et al., 2013). Recentemente, sono stati confrontati i Factor Mixture Models (si veda la Sottosezione 1.2.2) con i SEM trees, rilevandone le rispettive complementarietà (Jacobucci et al., 2017).

1.2.2 Metodi per l'eterogeneità non-osservata

Rispetto all'analisi dell'eterogeneità non-osservata, l'approccio multi-gruppo o, più in generale, un approccio basato su covariate esterne è difficilmente applicabile, in quanto i sotto-gruppi presenti nel campione devono essere conosciuti *a priori*. Una possibile soluzione potrebbe prevedere una metodologia a due passi con, in un primo tempo, l'applicazione di una Cluster Analysis e successivamente di un modello SEM multi-gruppo. Ad ogni modo, considerazioni teoriche e studi simulativi hanno mostrato che questa metodologia è inadeguata (Görz et al., 2000; Jedidi et al., 1997b).

Inoltre, una limitazione dell'approccio basato su una definizione *a priori* delle sorgenti di eterogeneità si ravvisa nel fatto che, spesso, l'eterogeneità non viene catturata adeguatamente solo dalle variabili demografiche o psicografiche (Sawatzky et al., 2009; Wedel & Kamakura, 2000). Infatti, la conoscenza relativa alle variabili che producono eterogeneità nei dati, di solito, non è disponibile o è incompleta (Moore, 1980; Wedel & Kamakura, 2000).

Factor Mixture Models

I modelli più utilizzati nell'ambito dell'eterogeneità non-osservata si fondano su un approccio Finite Mixture, ovvero assumono che l'eterogeneità del campione possa essere attribuita alla mistura di un numero limitato di sotto-popolazioni omogenee, che vengono modellate come componenti di mistura (o classi latenti) (Becker et al., 2013; Temme et al., 2002). Solitamente, fissato il numero di sotto-popolazioni ipotizzate dal ricercatore e definito un unico modello SEM o FA, l'approccio mistura consente di stimare la proporzione in cui le sotto-popolazioni sono presenti nel dataset e di assegnare empiricamente ciascun soggetto ad una di queste sotto-popolazioni, contemporaneamente stimando il modello SEM o FA per ogni sotto-popolazione (Lubke & Luningham, 2017). Pertanto, per ogni sotto-popolazione individuata si ottengono differenti valori parametrici stimati (Jedidi et al., 1997b). Come già accennato, questa classe di modelli, di solito, viene genericamente definita con il termine Factor Mixture Models (FMM) o Structural Equation Mixture Model (SEMM) (Jedidi et al., 1997b; Lubke & Muthén, 2005).

Come riportato da Lubke e Muthén (2005), i FMM possono essere considerati una combinazione tra il Common Factor Model (Thurstone, 1947) ed un modello a classi latenti (Lazarsfeld & Henry, 1968). Il vantaggio principale dei FMM e dei SEMM consiste nella loro capacità di individuare sotto-campioni di soggetti che presentano modalità di risposta simili agli items di un questionario o di un test psicologico (Lubke & Muthén, 2005).

Relativamente a questo approccio sono stati sviluppati modelli SEM (Jedidi et al., 1997b, 1997a) e FA confermativa (CFA; Yung, 1997), che considerano l'eterogeneità nelle intercette, nelle medie dei fattori, nelle matrici dei coefficienti fattoriali e nelle strutture di covarianza del modello. In particolare, Yung (1997) ha sviluppato quattro classi di modelli di mistura di modelli CFA: una prima classe di modelli considera le intercette e le matrici dei coefficienti fattoriali invarianti tra le componenti della mistura; la seconda classe prevede che solo le matrici dei coefficienti fattoriali siano invarianti, mentre le intercette variano e, per mantenere l'identificabilità del modello, viene posto a zero il vettore delle medie del modello; infine, una terza una classe di modelli assume che la matrice di varianza-covarianza dei fattori sia una matrice di identità ed il vettore delle medie dei fattori sia posto a zero, mentre si permette alle matrici dei coefficienti fattoriali di spiegare l'eterogeneità nei dati. Inoltre, Yung (1997) ha impiegato l'algoritmo Expectation-Maximiztion (EM; Dempster et al., 1977) e, parallelamente, il metodo di Approximate-Scoring per la stima dei parametri. Jedidi et al. (1997b), invece, hanno sviluppato un modello Finite Mixture SEM generale, che può essere ridotto ad un modello di cluster analysis come ad un modello di CFA multi-gruppo o ad un modello SEM multi-gruppo. La procedura di stima dei parametri per il modello sviluppato si basa sull'algoritmo EM con l'aggiunta di una procedura iterativa ulteriore nello step di massimizzazione (Jedidi et al., 1997b). Allo stesso modo, Dolan e van der Maas (1998) hanno presentato un modello di mistura di distribuzioni Normali multivariate generale, permettendo di sottoporre i parametri del modello a vincoli di uguaglianze, vincoli non-lineari o vincoli relativi ai limiti. La stima dei parametri di questo modello avviene tramite due approcci: un algoritmo quasi-Newton ed un metodo a due passi, in cui, nel primo, si utilizza l'algoritmo EM e, nel secondo, le stime di ogni componente vengono trattate come indipendenti in un'analisi multi-gruppo standard (Dolan & van der Maas, 1998). Inoltre, Arminger e Stein (1997) hanno proposto un modello generale di mistura per variabili Normali condizionate a variabili esplicative, specificando casi speciali come la mistura di modelli CFA. La procedura di stima dei parametri prevede due passi (Arminger & Stein, 1997): in primo luogo viene applicato l'algoritmo EM e, in secondo luogo, la stima viene ultimata tramite minimi quadrati generalizzati o tramite l'approccio della distribuzione asintoticamente libera. Similmente, Muthén e Shedden (1999) hanno specificato un modello di mistura di distribuzioni Normali multivariate con variabili risposta categoriali che utilizza covariate esterne per predire la variabile risposta, le medie dei fattori delle classi latenti e l'appartenenza alle classi latenti. Tale modello consente di predire anche variabili risposta categoriali sulla base dell'appartenenza alla classe latente (Muthén & Shedden, 1999). Inoltre, Muthén e Shedden (1999) ha

permesso il successivo sviluppo di classi di Growth Mixture Models, per studiare l'eterogeneità di sotto-popolazioni che variano nei pattern di cambiamento nel tempo (e.g., Dolan et al., 2005; Li et al., 2001; Muthén, 2001). Successivamente, Arminger et al. (1999) hanno definito una classe di modelli SEM per misture di densità di probabilità Normali multivariate condizionate a variabili esplicative con tre differenti algoritmi EM per la stima dei parametri dei suddetti modelli.

Per sistematizzare sotto un unica classe i modelli precedentemente definiti, Lubke e Muthén (2005) ha provveduto un'introduzione generale ai FMM e, nello stesso articolo, ha proposto una metodologia di analisi "step-by-step" per esplorare l'eterogeneità campionaria. I FMM sono stati riadattati anche per studi specifici come nel caso di DeSarbo et al. (2006), che hanno sviluppato un FMM su misura per i loro obiettivi di ricerca nell'ambito del management, o di Leite e Cooper (2010), che tramite l'uso di un FMM identificano: quali soggetti riportano risposte affette da bias indotto da desiderabilità sociale, quali items tendono ad elicitare risposte più distorte e quali covariate predicono tale bias. Diversamente dalla letteratura precedente, Bauer (2005) hanno mostrato come applicazioni indirette di modelli di mistura di SEM possano modellare relazioni non-lineari tra variabili latenti, approssimando la funzione di regressione latente. In linea con il precedente articolo, Wall et al. (2012) hanno utilizzato un FMM per approssimare fattori latenti continui con distribuzione non-Normale in presenza di covariate continue e dicotomiche. Inoltre, An e Bentler (2011) hanno esteso un FMM per l'inclusione di covariate al fine di modellare variabili risposta miste continue e binarie. Un altro contributo interessate è dato da Cai et al. (2011), che hanno proposto un modello di mistura generalizzato a variabili latenti per dati misti (continui, ordinali, di conteggio e nominali) in presenza di eterogeneità campionaria con stima dei parametri tramite un algoritmo Reversible Jump Markov Chain Monte Carlo (MCMC). Dall'ambito *data mining* è stato proposto un modello di mistura SEM, il SubgroupSEM, che prevede un approccio a 4 passi (Kiefer et al., 2022): (1) viene specificato un SEM a due gruppi, (2) viene scelto uno spazio di covariate che contiene descrittori potenziali dei gruppi, (3) viene calcolata una misura per la rilevanza dei gruppi e (4) viene applicato un algoritmo o un'euristica per estrarre sotto-gruppi rilevanti.

Relativamente alla stima dei parametri nell'ambito FMM e SEMM, Hoshino (2001) ha impiegato il Gibbs sampling per stimare i parametri in modelli di mistura di CFA. Zhu e Lee (2001) hanno utilizzato un approccio Bayesiano con metodi MCMC per analizzare misture di modelli SEM definiti in LISREL. Cai e Song (2010) hanno sviluppato un modello di mistura di SEM con dati mancanti non-trascurabili. Cai et al. (2010) hanno esteso il precedente modello all'inclusione di covariate. Yuan e Bentler (2010) hanno

sviluppato un approccio di stima di Maximum Likelihood (ML) a due stadi per i modelli di mistura di SEM: nel primo stadio si calcolano le stime ML del modello saturo tramite l'algoritmo EM e le loro matrici di covarianza asintotiche e, nel secondo stadio, si adattano SEM convenzionali alle medie e alle covarianze per ogni componente ottenuta nel primo stadio. Più recentemente, Liu e Song (2017) hanno sviluppato un approccio Bayesiano che tramite uso di un metodo Reversible Jump MCMC adatta modelli di mistura SEM.

Parallelamente, anche nella comunità scientifica che si occupa di Machine Learning e di statistica computazionale sono stati costruiti modelli di mistura di FA con procedure di stima dei parametri basati sulla stima di ML, approcci Bayesiani o approcci misti: rispetto all'approccio della ML è stato impiegato l'algoritmo EM (Ghahramani & Hinton, 1997), estensioni dell'EM come lo Split-and-Merge-EM (Ueda et al., 2000), l'algoritmo l'Expected-Conditional-Maximization (Zhao & Philip, 2008), l'Alternating Expectation-Conditional Maximization (Meng & Van Dyk, 1997), applicato da McLachlan et al., 2003, ed una derivazione degli stimatori di massima verosimiglianza diretti per le misture di FA (Montanari & Viroli, 2011); rispetto all'approccio Bayesiano sono stati sviluppati un'approssimazione variazionale della distribuzione a posteriori per fare inferenza (Ghahramani & Beal, 1999) e metodi MCMC (Fokoué, 2005; Fokoué & Titterington, 2003; Utsugi & Kumagai, 2001); relativamente agli approcci misti è stato utilizzato un algoritmo Monte Carlo Expectation-Maximization (An & Bentler, 2011). Anche in questo frangente, sono stati sviluppati modelli di mistura di FA che permettono l'inclusione di covariate per indagare il legame di variabili osservate con le variabili latenti del modello, ovvero con la variabile aleatoria che governa l'appartenenza al gruppo e con la variabile aleatoria da cui sono generati i fattori latenti (e.g., An e Bentler, 2011; Fokoué, 2005; Zhou e Liu, 2008). Montanari e Viroli (2010) hanno proposto un modello di mistura di FA basato sull'assunto che i fattori latenti siano governati da una mistura di differenti distribuzioni, la cui stima dei parametri avviene tramite algoritmo EM. Successivamente, Cagnone e Viroli (2012, 2014) hanno esteso il precedente modello per lo studio di, rispettivamente, variabili risposta binarie e miste binarie, ordinali e di conteggio. Rispetto al modello di Cagnone e Viroli (2014), Amiri et al. (2018) hanno incluso anche il modellamento dell'effetto di covariate.

Rispetto agli studi di simulazione con modelli della famiglia FMM, Williams et al. (2002) hanno condotto una simulazione Monte Carlo per analizzare le performance dell'approccio MECOSA (MEan and COvariance Structure Analysis; Arminger et al., 1996) a modelli SEM condizionati a covariate: MECOSA ha funzionato al meglio quando le proporzioni delle componenti di mistura erano uguali nei dati e in presenza di ampia

differenza tra parametri delle singole componenti; al contrario, le prestazioni peggioravano quando le proporzioni erano sbilanciate e i parametri erano molto simili tra gruppi. Lubke e Muthén (2007) hanno studiato le prestazioni del FMM definito in Muthén e Shedden (1999) tramite uno studio di simulazione, che ha rilevato: stime dei parametri veri soddisfacenti, anche in caso di una ridotta separazione tra le classi latenti; un assegnamento insoddisfacente delle osservazioni alle classi latenti in presenza di ridotta separazione tra le classi e in assenza di covariate, che migliora se aumentano la separazione e/o gli effetti delle covariate. Tueller e Lubke (2010) hanno mostrato come i modelli di mistura SEM permettano di individuare classi latenti di soggetti che differiscono in base alla loro struttura di covarianza. Inoltre, è stato osservato che stime accurate e tassi di convergenza soddisfacenti sono ottenibili in presenza di campioni anche con $n \geq 100$ e quando la separazione tra classi è alta (Tueller & Lubke, 2010). Rispetto a modelli con variabili risposta binarie o con più categorie, il tempo computazionale aumenta drasticamente e diminuiscono i tassi di convergenza (Tueller & Lubke, 2010). Infine, l'assegnazione alla classe latente corretta da parte di modelli di mistura SEM risulta generalmente insoddisfacente, in particolare per i modelli più complessi. Inoltre, Henson et al. (2007) hanno dimostrato come i modelli FMM siano strumenti validi per stimare misture di sotto-popolazioni associate a differenze nel modello strutturale. In aggiunta, è stato osservato che anche in presenza di campioni con $n = 500$ le stime tendono ad essere accurate (Henson et al., 2007). Recentemente, Y. Wang et al. (2021) hanno riportato che l'inclusione di covariate permette di ridurre l'ampiezza campionaria richiesta dai FMM: per covariate con un effetto medio si raccomandano campioni con $n = [750, 4000]$, mentre per covariate con un effetto più consistente si possono utilizzare campioni con $n = [250, 1000]$. D'altro canto, Buzick (2010) ha osservato che un'ampiezza campionaria almeno pari a $n = 800$ con due classi latenti di uguale proporzione permette di ottenere risultati adeguati.

Riguardo alla selezione delle componenti della mistura, Lee e Song (2003) hanno utilizzato il Bayes factor per modelli di mistura di SEM attraverso una procedura di path sampling. Henson et al. (2007) hanno comparato numerosi indici di fit, rilevando: rispetto agli indicatori basati sulla verosimiglianza, il sample-size adjusted Bayesian Information Criterion (ssBIC) ha dimostrato ottime capacità di classificazione di modelli con due componenti, seguito da il Vuong–Lo–Mendell–Rubin likelihood ratio test (VLMR) e il Vuong–Lo–Mendell–Rubin adjusted likelihood ratio test (aVLMR); relativamente alle statistiche basate sulla classificazione, il Classification Likelihood Information Criterion (CLC) e l'Integrated Classification Likelihood BIC (ICL-BIC) hanno mostrato capacità molto soddisfacenti per identificare modelli con due e tre componenti. Henson et al.

(2007) hanno rilevato che per campioni con $n = 500$ le statistiche di fit studiate non hanno sufficiente potenza per differenziare modelli a due componenti da modelli con una componente. Più recentemente è stato proposto il metodo CHull (Ceulemans & Kiers, 2006) per misture di FA (Bulteel et al., 2013). Secondo un approccio tipico dell'ambito del Machine Learning, Grimm et al. (2017) hanno proposto una procedura standard di k-fold cross-validation per FMM: un modello viene adattato separatamente a $k - 1$ differenti partizioni del dataset, i modelli risultanti sono applicati alla k-esima porzione dei dati e vengono confrontati gli indici di fit. Inoltre, come recentemente hanno dimostrato Y. Wang et al. (2021), l'inclusione di covariate con effetto non nullo migliora notevolmente la capacità dei FMM di selezionare il numero corretto di componenti di mistura.

Riguardo alle criticità dei FMM, in particolare dei più generali STEMM (Jedidi et al., 1997b), Bauer e Curran (2004) hanno individuato tre condizioni che possono condurre alla stima di classi latenti spurie: un'errata specificazione del modello SEM o FA della mistura, che può portare ad una sovrastima del numero delle classi latenti; violazione dell'assunto di Normalità della variabile risposta (e.g., distribuzione asimmetrica, curtosi elevata, multi-modale), che, in presenza di un modello SEM o FA correttamente specificato e di una sola sotto-popolazione, può indurre la stima di classi latenti addizionali; relazioni non-lineari tra variabili osservate e/o latenti, che possono implicare la stima diclassi latenti spurie.

Modelli SEM multilivello

In alternativa ai FMM sono stati progettati ed applicati i modelli SEM multilivello, per gestire l'eterogeneità a livello delle medie dei fattori e delle intercette del modello strutturale (Longford & Muthén, 1992; Muthén, 1989; Muthén & Satorra, 1989). Tali modelli consentono di trattare dati caratterizzati da una struttura gerarchica (e.g., un campione casuale di studenti universitari può provenire da differenti classi, percorsi accademici, facoltà, etc.; Ansari et al., 2002). Successivamente, (Ansari et al., 2000) ha proposto una generalizzazione dei modelli SEM multilivello (Longford & Muthén, 1992; Muthén, 1989), ovvero i modelli SEM gerarchici Bayesiani. Questi ultimi sono stati sviluppati in modo da stimare l'eterogeneità, oltre che nelle intercette e nelle medie, anche nelle matrici di varianza-covarianza dei fattori e degli errori del modello (Ansari et al., 2000). In linea con il precedente studio, Ansari et al. (2002) hanno sviluppato una classe di modelli multilivello di FA, la cui stima si basa sull'approccio Bayesiano.

Vermunt (2003) e Vermunt (2008) hanno proposto, rispettivamente, due modelli

multilivello a classi latenti: nel primo, l'eterogeneità al livello superiore è stata modellata usando effetti casuali continui, mentre, nel secondo, è stata modellata assumendo che sia unità al livello inferiore che al livello superiore possano appartenere a classi latenti.

Più recentemente, Varriale e Vermunt (2012) hanno sviluppato un modello multilivello di mistura di EFA, in cui le unità di livello superiore appartengono a classi latenti, che differiscono in termini dei parametri dei modelli EFA stimati sulle unità di livello inferiore. L'algoritmo di stima prevede la massimizzazione della verosimiglianza marginale (Varriale & Vermunt, 2012). Successivamente, De Roover et al. (2017) hanno presentato un modello multilivello di mistura simultanea di EFA, che applica un modello EFA alle unità di livello inferiore ed un modello mistura per raggruppare in classi le unità di livello superiore, in base alla somiglianza della loro struttura fattoriale. La stima dei parametri avviene in due stadi: vengono realizzate un certo numero di iterazioni dell'algoritmo EM e, una volta vicino alla soluzione, si utilizza un approccio Newton-Raphson per velocizzare la convergenza (De Roover et al., 2017). In un recente articolo, De Roover et al. (2020) hanno sviluppato un modello di mistura di FA multi-gruppo, che assegna le unità al livello superiore a classi latenti in base ad uno specificato livello di invarianza di misurazione. Gli autori hanno utilizzato un approccio di ML per la stima, basato sull'algoritmo Expectation-Conditional Maximization (De Roover et al., 2020).

Uno svantaggio di questa classe di modelli riguarda la necessità di dover raccogliere molteplici osservazioni per ciascun livello oggetto di studio, al fine di stimare l'eterogeneità presente a tale livello. Tuttavia, è stato dimostrato come l'applicazione di modelli CFA a dati multilivello, senza considerare la dipendenza tra le osservazioni, porta ad un peggioramento generale dell'adattamento del modello: sovrastima dei parametri, sottostima dei relativi errori standard ed incremento del valore della statistica χ^2 (Julian, 2001; Pornprasertmanit et al., 2014).

Modelli MIMIC

Parallelamente, Muthén (1989) ha proposto un approccio basato sul modellamento MIMIC (Multiple Indicators, Multiple Causes) (Jöreskog & Goldberger, 1975). Un modello MIMIC cattura l'eterogeneità presente nei dati tramite la specificazione di un insieme di predittori e permette alle intercette e alle medie dei fattori latenti di variare tra gruppi senza la necessità di campioni molto ampi, come nel caso dei SEM multi-gruppo (Muthén, 1989). I modelli MIMIC consentono di individuare eterogeneità non-osservata testando l'effetto di determinate covariate (Muthén, 1989).

Approccio Partial Least Squares

Una quarta soluzione è rappresentata dall'approccio Partial Least Squares (PLS) applicato ai modelli SEM, che consente di indagare e gestire l'eterogeneità non-osservata e, in certi casi, quella osservata (Rigdon et al., 2010). In particolare, il modello Finite Mixture Partial Least Squares (FIMIX-PLS; Hahn et al. (2002)) è il primo e più diffuso approccio a classi latenti nell'ambito PLS (Sarstedt, 2008). Il modello FIMIX-PLS stima simultaneamente i path coefficients specifici per ogni sotto-popolazione, assegnando ciascuna osservazione ad una specifica sotto-popolazione (Rigdon et al., 2010). Tuttavia, il modello FIMIX-PLS non spiega l'eterogeneità presente nella covarianza dei fattori e nella varianza del modello strutturale (Sarstedt & Ringle, 2010). Ad ogni modo, una recente revisione conferma l'utilità dei modelli a classi latenti offerti dall'approccio PLS-SEM nell'identificazione e nel trattamento dell'eterogeneità non-osservata (Sarstedt et al., 2022). Inoltre, Becker et al. (2013) ha proposto una metodologia denominata "Unobserved Heterogeneity Discovery", che ha lo scopo di guidare il/la ricercatore/ricercatrice nell'applicazione dei metodi statistici, per assicurarsi la validità dei risultati e per produrre una teoria che converta l'eterogeneità non-osservata in eterogeneità osservata.

1.2.3 Applicazioni dei Factor Mixture Models

L'applicazione di modelli di mistura di SEM o FA per individuare eterogeneità campionaria non-osservata ha riscosso ampio successo in numerosi ambiti di studio: nell'ambito della psicologia cognitiva (Reynolds et al., 2010), della psicologia dello sport (Lämmle et al., 2013), del turismo (Assaf et al., 2016), dell'econometria (Cappozzo & Greselin, 2019), della psicologia del lavoro (Suárez & Muñiz, 2018), della psicologia clinica (Ulbricht et al., 2018) e della psicologia sociale (Friehs et al., 2022). Inoltre, i FMM sono stati impiegati con successo per indagare gli stili di risposta nei questionari self-report (McIntyre, 2011) e il bias indotto da desiderabilità sociale (Leite & Cooper, 2010). Difatti, i FMM sono in grado di identificare i soggetti che rispondendo ad un questionario attuando comportamenti di faking, anche in contesti non-sperimentali (Ziegler et al., 2015).

In particolare, i FFM hanno rappresentato un importante contributo alla metodologia della ricerca in psicologia clinica e in psichiatria. Di seguito è riportata una breve rassegna degli sviluppi avvenuti in questi due ambiti di ricerca grazie all'introduzione dei FMM, al fine di illustrarne le potenzialità.

Factor Mixture Models in Psicologia Clinica e Psichiatria

Secondo Van Dam et al. (2017) un grave limite della disciplina psichiatrica è dovuto all'ampia eterogeneità all'interno delle categorie diagnostiche. Conseguentemente, approcci data-driven che distinguano sotto-popolazioni omogenee, come i FMM, possono risultare rilevanti per rivelare la struttura latente della psicopatologia nelle sue sfaccettature (Whalen, 2017). A tal proposito, Muthén (2006) ha sviluppato una metodologia per individuare la rappresentazione più adatta per una specifica psicopatologia tramite tre modelli psicometrici, che indicano se sia categoriale (i.e. LCA), dimensionale (i.e. FA) o un ibrido tra le due (i.e. FMM). Successivamente, Masyn et al. (2010) hanno presentato un unico quadro teorico per utilizzare i FMM nell'esplorazione della struttura latente dei costrutti psicologici. Clark et al. (2013) hanno rilevato i FMM come strumenti adeguati per contribuire alla concettualizzazione della struttura sottostante ai disturbi psicologici tramite un'esemplificazione relativa ai disturbi della condotta. Successivamente, i FMM sono stati applicati per studiare l'eterogeneità relativa al disturbo da deficit dell'attenzione/iperattività (Lubke et al., 2007), al disturbo di panico (Roberson-Nay & Kendler, 2011), alla schizofrenia (Picardi et al., 2012), a pazienti con diagnosi di disturbo della personalità (Yun et al., 2013), alla depressione maggiore (Sunderland et al., 2013; Ten Have et al., 2016), al disturbo da uso di alcol (Jackson et al., 2014), alla psicopatia (Yildirim & Derksen, 2015), al tratto della rabbia (Lubke et al., 2015), ai disturbi di comportamento dirompente in infanzia (Bolhuis et al., 2017), alle costellazioni della qualità della relazione diadica nelle famiglie adottive (Jensen, 2017) e alla struttura latente del disturbo post-traumatico da stress (Redican et al., 2022).

Recentemente, Miettunen et al. (2016) hanno presentato un quadro metodologico che si basa sulla selezione del modello psicometrico maggiormente adattato ai dati, per determinare quale sia la miglior caratterizzazione della struttura latente di un costrutto psichiatrico, in linea con la proposta di Muthén (2006). Tale metodologia ha trovato un discreto successo applicativo, come nel caso dell'eterogeneità delle caratteristiche sensoriali del disturbo dello spettro autistico Tillmann et al. (2020) e dei sintomi della depressione Divers et al. (2022).

Capitolo 2

Analisi fattoriale confermativa ed esplorativa

2.1 Analisi fattoriale confermativa

2.1.1 Modelli a variabili latenti

Un modello di analisi fattoriale confermativa è un modello statistico multivariato per lo studio simultaneo di più variabili osservate (i.e., items o domande di un test psicologico; Bollen, 1989). Tale modello appartiene alla casse più generale dei modelli a variabili latenti. I modelli a variabili latenti consentono di ridurre la dimensionalità di dati multivariati, in quanto l'informazione presente nelle relazioni tra molteplici variabili può essere descritta da un insieme più ristretto di variabili (Bollen, 1989). Questa riduzione della dimensionalità aumenta l'interpretabilità delle relazioni presenti nei dati (Bartholomew et al., 2011).

Per definire un modello a variabili latenti è possibile riferirsi ad un modello statistico che specifica la distribuzione congiunta di un insieme di variabili casuali, in cui alcune di queste variabili sono osservabili o manifeste e altre sono inosservabili o latenti (Bartholomew et al., 2011). Formalmente, sia $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ un insieme di variabili aleatorie continue osservate e sia $\boldsymbol{E} = (E_1, \ldots, E_q)$ un insieme di variabili aleatorie continue latenti[1], considerando solo il caso in cui $q < p$. Inoltre, assumiamo che le realizzazioni $i = \{1, \ldots, n\}$ delle variabili aleatorie osservate e latenti siano indipendenti

[1]Per non confondere variabili osservate e latenti, si utilizzano le lettere greche per queste ultime. Si noti che E corrisponde alla lettera maiuscola η.

e con identica distribuzione, ottenendo rispettivamente: $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{ip})$, quale vettore i-esimo di realizzazioni di variabili osservate e $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iq})$, quale vettore i-esimo di variabili latenti. Quindi, la distribuzione congiunta delle realizzazioni delle due variabili casuali consiste in $P(\boldsymbol{y}_i, \boldsymbol{\eta}_i \mid \boldsymbol{\theta})$, in cui $\boldsymbol{\theta}$ indica un vettore di parametri non noti da cui dipendono i due vettori di variabili. Tale distribuzione congiunta viene studiata condizionando il vettore delle variabili osservate al vettore delle variabili latenti (Bartholomew et al., 2011), in modo che le variabili osservate non possano essersi realizzate senza che le variabili latenti si siano precedentemente verificate. Conseguentemente, si induce una gerarchia tra le variabili, per cui nel meccanismo generatore dei dati le variabili latenti $\boldsymbol{\eta}_i$ sono generate prima delle variabili osservate $\boldsymbol{y}_i$, come di seguito:

1. $\boldsymbol{\eta}_i \sim P(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}_1)$;
2. $\boldsymbol{y}_i \mid \boldsymbol{\eta}_i \sim P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}_2)$.

dove, $\boldsymbol{\theta}_1 \subset \mathbb{R}^q$ è il vettore $q \times 1$ di parametri che governa la densità di probabilità delle variabili latenti, mentre $\boldsymbol{\theta}_2 \subset \mathbb{R}^p$ è il vettore $p \times 1$ di parametri della densità di probabilità delle variabili osservate. Questo schema di condizionamento dei dati è detto schema riflessivo, in contrasto con lo schema formativo in cui le variabili latenti sono condizionate alle variabili osservate (Bollen, 1989).

La Tabella 2.1 mostra come nei modelli a variabili latenti si possono ottenere differenti classi di modelli, modificando la tipologia di variabili latenti e variabili osservate implicate. In pratica, si agisce sulle densità di probabilità $P(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}_1)$ e $P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}_2)$, scegliendo distribuzioni di probabilità che descrivano le caratteristiche dei dati a disposizione. Come si osserva, nel caso della CFA la variabili aleatorie sia latenti che osservate sono continue.

		Variabili osservate	
		Continue	Categoriali
Variabili latenti	Continue	Analisi Fattoriale	Analisi del Tratto Latente
	Categoriali	Analisi del Profilo Latente	Analisi della Classe Latente

Tabella 2.1: Classi di modelli a variabili latenti in base alla tipologia di variabile aleatoria manifesta e latente.

2.1.2 Analisi Fattoriale Confermativa

Introduzione

Storicamente l'analisi fattoriale trova le sue origini in Spearman (1904) ed è stata estesa ampiamente da Thurstone (1935, 1947). Lo scopo principale della FA, come accennato nella precedente Sezione, è descrivere le relazioni di covarianza osservata tra molteplici variabili in termini di un ridotto numero di costrutti latenti comuni (Bollen, 1989). Tali costrutti latenti vengono rappresentati da variabili aleatorie latenti denominate fattori comuni (Bollen, 1989). Pertanto, un fattore comune è definibile come un costrutto non-osservabile, ovvero non direttamente misurabile, che influenza le variabili osservate e che ne spiega la covariazione (Kline, 2015). Infatti, la FA assume che tutte le covarianze o correlazioni tra un insieme di variabili osservate riflettano la presenza di fattori latenti continui con un'influenza comune sulle variabili (Bauer & Curran, 2004). Ad esempio, se uno studio su un campione di bambini/bambine mostrasse una correlazione positiva tra ore passate a giocare da soli/sole e il numero di amicizie, tramite la FA si potrebbe ipotizzare una dimensione latente comune relativa all'introversione (Bauer & Curran, 2004). Utilizzando il fattore latente dell'introversione, infatti, la covarianza residuale tra le ore passate a giocare da soli/sole e il numero di amicizie potrebbe ridursi notevolmente fino ad annullarsi (Bauer & Curran, 2004). Questo esempio introduce l'assioma dell'indipendenza locale, per cui le correlazioni residuali tra variabili osservate devono essere vicine a zero dopo aver stimato i fattori latenti comuni (Bauer & Curran, 2004). Tuttavia, si osserva frequentemente l'assenza di una correlazione perfetta tra variabili osservate. Ciò è stato interpretato come dovuto alla presenza di fattori unici associati separatamente a ciascuna variabile osservata, ovvero la varianza unica (Mulaik, 2009). In linea con la Teoria Classica del Test, la varianza unica è decomponibile in una parte di varianza data dall'errore casuale nel processo di misurazione delle variabili osservate e in una parte detta "varianza specifica", ovvero la varianza vera della variabile osservata (Mulaik, 2009). Tuttavia, in questa sede ci riferiremo alla varianza unica senza decomporla e considerandola, primariamente, come errore casuale dovuto al processo di misurazione delle osservate.

La FA si divide in due metodologie differenti: l'analisi fattoriale confermativa o CFA e l'analisi fattoriale esplorativa o EFA (Bollen, 1989). Nella CFA il/la ricercatore/ricercatrice può includere nel modello le proprie ipotesi su quali variabili osservate sono influenzate da quali fattori latenti, in modo da testarle direttamente (Bollen, 1989). Invece, nel caso dell'EFA, come suggerisce il nome, si conduce un'analisi del tutto

esplorativa, in cui non ci sono ipotesi da testare e tutte le associazioni possibili tra fattori latenti e variabili osservate vengono stimate (Bollen, 1989).

Specificazione del modello CFA

Come osservato in Sottosezione 2.1.1, la CFA può essere derivata come caso particolare di un modello a variabili latenti. Nello specifico, la CFA assume una densità di probabilità Normale multivariata sia per $P(\boldsymbol{\eta}_i \mid \boldsymbol{\theta}_1)$ che per $P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}_2)$. Quindi, le variabili latenti saranno distribuite secondo un modello Normale q-variato:

$$\boldsymbol{\eta}_i \sim \mathcal{N}_q(\boldsymbol{\mu}_i, \boldsymbol{\Phi}) \tag{2.1}$$

dove, $\boldsymbol{\mu}_i$ è il vettore $q \times 1$ delle medie e $\boldsymbol{\Phi}$ è la matrice di covarianza di ordine $q \times q$. Pertanto, è possibile ottenere differenti tipi di modelli CFA sulla base dell'ordine della matrice $\boldsymbol{\Phi}_{q \times q}$:

- Modelli unidimensionali ($q = 1$):
  $\boldsymbol{\Phi}_{q \times q} = \phi$ contiene solo la varianza ϕ^2 dell'unica variabile latente η_i.
- Modelli bidimensionali ($q = 2$):
  $\boldsymbol{\Phi}_{q \times q} = \boldsymbol{\Phi}_{2 \times 2}$ contiene le varianze $\{\phi_1^2, \phi_2^2\}$ delle variabili latenti $\{\eta_{i1}, \eta_{i2}\}$ con anche il termine di covarianza ϕ_{12} (o ϕ_{21}).
- Modelli multidimensionali ($q > 2$):
  $\boldsymbol{\Phi}_{q \times q}$ contiene q varianze sulla diagonale relative alle variabili latenti $\frac{1}{2}(q \times q - q)$ termini di covarianza.

A questo punto, è necessario definire la relazione tra variabili latenti e variabili osservate tramite un modello lineare:

$$\boldsymbol{y}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i + \boldsymbol{\delta}_i \tag{2.2}$$

dove, $\boldsymbol{\tau}_i$ il vettore $p \times 1$ delle intercette, $\boldsymbol{\Lambda}_1$[2] è la matrice $p \times q$ di coefficienti reali, ovvero i coefficienti fattoriali, e $\boldsymbol{\delta}_i$ è il vettore $p \times 1$ di errore del modello. Nello specifico, la matrice $\boldsymbol{\Lambda}_1$ contiene quei coefficienti che indicano la presenza e l'intensità dell'associazione tra le p variabili osservate e le q variabili latenti. Più in dettaglio, la struttura della matrice $\boldsymbol{\Lambda}_1$ nella CFA prevede che alcune associazioni tra variabili osservate, o indicatori, e variabili latenti, o misurandi, siano fissate ed altre libere di essere stimate. Quindi, con $\lambda = 0$ si indica che l'indicatore non è legata al misurando (parametro fissato), mentre $\lambda \neq 0$

[2]Il pedice della matrice $\boldsymbol{\Lambda}_1$ serve per distinguere la matrice dei coefficienti fattoriali del modello CFA rispetto a quella del modello EFA definito successivamente.

indica che il legame tra indicatore e misurando è da stimare (parametro libero). Un esempio di matrice $\mathbf{\Lambda}_1$:

$$\begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ 0 & \lambda_{42} \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \end{bmatrix} \tag{2.3}$$

Inoltre, viene assunto un modello Normale p-variato per l'errore nelle misurazioni delle variabili osservate:

$$\boldsymbol{\delta}_i \sim \mathcal{N}_p(\mathbf{0}_p, \mathbf{\Theta}_{\boldsymbol{\delta}}) \tag{2.4}$$

dove, $\mathbf{0}_p$ è un vettore $p \times 1$ di zeri, che indica la presenza di medie nulle, e $\mathbf{\Theta}_\delta$ è la matrice $p \times p$ di varianze-covarianze. Di solito, la matrice $\mathbf{\Theta}_\delta$ è diagonale, ovvero con valori non-nulli solo sulla diagonale e zero nelle altre posizioni.

Per convenzione, si applica il seguenti assunto (Bollen, 1989):

$$\mathbb{C}ov[\boldsymbol{\eta}_i \boldsymbol{\delta}_i^T] = \mathbf{0}_{q \times p} \tag{2.5}$$

per cui gli errori casuali di misurazione delle variabili osservate $\boldsymbol{\delta}_i$ sono incorrelati con i fattori latenti $\boldsymbol{\eta}_i$ e il simbolo T posto all'apice di un vettore o di una matrice indica l'operatore di trasposizione.

Sulla base dell'Equazione 2.2 è possibile derivare il modello probabilistico marginale delle variabili aleatorie osservate tramite il calcolo del valore atteso e della varianza:

$$\begin{aligned} \mathbb{E}[\boldsymbol{y}_i] &= \mathbb{E}[\boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \boldsymbol{\eta}_i + \boldsymbol{\delta}_i] \\ &= \boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \mathbb{E}[\boldsymbol{\eta}_i] + \mathbb{E}[\boldsymbol{\delta}_i] \\ &= \boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \boldsymbol{\mu}_i \end{aligned} \tag{2.6}$$

$$\begin{aligned} \mathbb{V}ar[\boldsymbol{y}_i] &= \mathbb{V}ar[\boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \boldsymbol{\eta}_i + \boldsymbol{\delta}_i] \\ &= \mathbb{V}ar[\mathbf{\Lambda}_1 \boldsymbol{\eta}_i] + \mathbb{V}ar[\boldsymbol{\delta}_i] \\ &= \mathbf{\Lambda}_1 \mathbb{V}ar[\boldsymbol{\eta}_i] \mathbf{\Lambda}_1^T + \mathbf{\Theta}_{\boldsymbol{\delta}} \end{aligned} \tag{2.7}$$

dove $\mathbb{V}ar[\boldsymbol{\eta}_i] = \mathbf{\Phi}$. Pertanto, si ottiene:

$$\boldsymbol{y}_i \sim \mathcal{N}_p(\boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \boldsymbol{\mu}_i, \mathbf{\Lambda}_1 \mathbf{\Phi} \mathbf{\Lambda}_1^T + \mathbf{\Theta}_{\boldsymbol{\delta}}) \tag{2.8}$$

dove, $\boldsymbol{\tau}_i + \mathbf{\Lambda}_1 \boldsymbol{\mu}_i$ è il vettore $p \times 1$ delle medie marginali delle variabili osservate e $\mathbf{\Lambda}_1 \mathbf{\Phi} \mathbf{\Lambda}_1^T + \mathbf{\Theta}_{\boldsymbol{\delta}}$ è la matrice di covarianza di ordine $p \times p$.

Per concludere la derivazione della CFA come modello a variabili latenti definiamo $P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, \boldsymbol{\theta}_2)$ come segue, sulla base dei risultati notevoli del condizionamento di una distribuzioni Normale mutlivariata a ad un'altra distribuzione Normale multivariata (Pace, Salvan et al., 2001):

$$\boldsymbol{y}_i \mid \boldsymbol{\eta}_i \sim N_p(\boldsymbol{\tau}_i + \boldsymbol{\Lambda}_1 \boldsymbol{\mu}_i, \boldsymbol{\Theta}_{\boldsymbol{\delta}}) \tag{2.9}$$

Infine, è possibile scrivere il modello CFA in termini di modello lineare a variabili latenti:

$$\boldsymbol{\eta}_i \sim \mathcal{N}_q(\boldsymbol{\mu}_i, \boldsymbol{\Phi}) \tag{2.10}$$

$$\boldsymbol{\delta}_i \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Theta}_{\boldsymbol{\delta}}) \tag{2.11}$$

$$\boldsymbol{y}_i = \boldsymbol{\tau}_i + \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i + \boldsymbol{\delta}_i \tag{2.12}$$

dove possiamo osservare che: la relazione tra variabili osservate $\boldsymbol{y}_i$ e fattori latenti $\boldsymbol{\eta}_i$ è lineare, l'intensità della relazione è regolata dalla matrice dei coefficienti fattoriali $\boldsymbol{\Lambda}_1$ e tale relazione viene perturbata dalle componenti di errore casuale $\boldsymbol{\delta}_i$. Figura 2.1 mostra un esempio di modello CFA con $p = 6$ e $q = 2$. I nodi circolari indicano le variabili latenti, mentre i nodi rettangolari indicano le variabili osservate. Le frecce unidirezionali indicano la relazione tra i fattori latenti $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\}$, le frecce bidirezionali indicano correlazione se tra nodi diversi (come tra i due fattori latenti), ma se associate allo stesso nodo indicano la varianza (come nel caso del fattore latente $\boldsymbol{\eta}_1$).

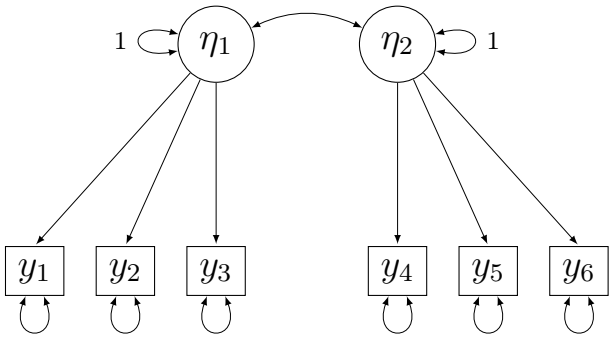


Figura 2.1: Esempio di modello CFA con $p = 6$ e $q = 2$. I nodi circolari indicano le variabili latenti, mentre i nodi rettangolari indicano le variabili osservate. Le frecce unidirezionali indicano la relazione tra i fattori latenti $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\}$, le frecce bidirezionali indicano correlazione se tra nodi diversi, ma se associate allo stesso nodo indicano la varianza (come nel caso del fattore latente $\boldsymbol{\eta}_1$).

Inoltre, di solito si assume che le medie delle variabili latenti e le medie delle variabili osservate siano nulle:

$$\mathbb{E}[\boldsymbol{\eta}_i] = \mathbf{0}_q$$
$$\mathbb{E}[\boldsymbol{y}_i] = \mathbf{0}_p$$

per cui $\boldsymbol{\mu}_i = \mathbf{0}_q$ e $\boldsymbol{\tau}_i = \mathbf{0}_p$. Di conseguenza, il modello marginale sarà:

$$\boldsymbol{y}_i \sim N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}}) \tag{2.13}$$

per cui, gli unici parametri del modello delle variabili osservate sono le componenti della matrice di varianze-covarianze, oggetto delle procedure di stima e di inferenza statistica.

A questo punto è possibile tradurre in equazioni la decomposizione della varianza della CFA. In particolare, la varianza della j-esima variabile osservata nel caso $\boldsymbol{\Phi} = \boldsymbol{I}_{q \times q}$ (matrice identità), per fattori latenti incorrelati con varianze pari a 1, corrisponde a:

$$\mathbb{V}ar[y_{ij}] = \boldsymbol{\Lambda}_{1j} \boldsymbol{\Lambda}_{1j}^T + \boldsymbol{\Theta}_{\boldsymbol{\delta} jj} \tag{2.14}$$

Nel caso in cui la correlazione tra fattori latenti sia ammessa, si ottiene:

$$\mathbb{V}ar[y_{ij}] = \boldsymbol{\Lambda}_{1j} \boldsymbol{\Phi} \boldsymbol{\Lambda}_{1j}^T + \boldsymbol{\Theta}_{\boldsymbol{\delta} jj} \tag{2.15}$$

dove possiamo individuare le due componenti della varianza: la varianza comune o comunalità ($\boldsymbol{\Lambda}_{1j} \boldsymbol{\Phi} \boldsymbol{\Lambda}_{1j}^T$) e la varianza unica ($\boldsymbol{\Theta}_{\boldsymbol{\delta} jj}$), che consiste nella parte di varianza dell'osservata non spiegata dalla presenza delle variabili latenti $\boldsymbol{\eta}_i$.

A questo punto, si presenta la funzione di log-verosimiglianza condizionata delle osservazioni $\boldsymbol{y}$ ai fattori latenti $\boldsymbol{\eta}$:

$$\ell\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\eta}\}\right) = \log \left(\prod_{i=1}^{n} \frac{\exp\left(-1/2 \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i\right)^T \boldsymbol{\Theta}_{\delta}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i\right)\right)}{\sqrt{(2\pi)^n \mid \boldsymbol{\Theta}_{\delta} \mid}} \right) \tag{2.16}$$

Mentre, la funzione di log-verosimiglianza dei fattori latenti $\boldsymbol{\eta}$:

$$\ell\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}\right) = \log \left(\prod_{i=1}^{n} \frac{\exp\left(-1/2 \left(\boldsymbol{\eta}_i - \mathbf{0}\right)^T \boldsymbol{\Phi}^{-1} \left(\boldsymbol{\eta}_i - \mathbf{0}\right)\right)}{\sqrt{(2\pi)^n \mid \boldsymbol{\Phi} \mid}} \right) \tag{2.17}$$

Pertanto, si può definire la log-verosimiglianza completa del modello CFA come:

$$\ell\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \boldsymbol{\Phi} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}\right) = \ell\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\eta}\}\right) \ell\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}\right) \tag{2.18}$$

$$= \log \left[\prod_{i=1}^{n} \frac{\exp\left(-1/2 \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i\right)^T \boldsymbol{\Theta}_{\delta}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_i\right)\right)}{\sqrt{(2\pi)^n \mid \boldsymbol{\Theta}_{\delta} \mid}} \right] + \tag{2.19}$$

$$+ \log \left[\prod_{i=1}^{n} \frac{\exp\left(-1/2 \left(\boldsymbol{\eta}_i - \mathbf{0}\right)^T \boldsymbol{\Phi}^{-1} \left(\boldsymbol{\eta}_i - \mathbf{0}\right)\right)}{\sqrt{(2\pi)^n \mid \boldsymbol{\Phi} \mid}} \right] \tag{2.20}$$

Specificazione del modello EFA

In generale, la differenza tra un modello EFA ed un modello CFA è piuttosto sfocata (Bollen, 1989). Pertanto, per definire un modello EFA è possibile far riferimento ai risultati raggiunti durante la specificazione del modello CFA, cambiando le lettere identificative dei parametri ed apportando modifiche mirate.

Il modello EFA viene definito come segue:

$$\boldsymbol{\xi}_i \sim \mathcal{N}_K(\boldsymbol{\alpha}_i, \boldsymbol{\Omega}) \tag{2.21}$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\varepsilon}}) \tag{2.22}$$

$$\boldsymbol{y}_i = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_2 \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \tag{2.23}$$

dove $\boldsymbol{\xi}_i$ è il vettore $K \times 1$ di fattori latenti, $\boldsymbol{\alpha}_i$ il vettore $K \times 1$ delle medie dei fattori latenti, $\boldsymbol{\Omega}$ è la matrice di varianze-covarianze $K \times K$ dei fattori latenti; $\boldsymbol{\varepsilon}_i$ sono gli errori del modello e $\boldsymbol{\Psi}_{\boldsymbol{\varepsilon}}$ è la matrice di varianze-covarianze degli errori; infine, $\boldsymbol{\nu}_i$ sono le intercette del modello e $\boldsymbol{\Lambda}_2$ è la matrice dei coefficienti fattoriali.

La caratteristica peculiare della EFA consiste nel fatto che un modello preciso che associ le variabili latenti a quelle osservate non è specificato in anticipo, per cui tutti i coefficienti fattoriali possono essere stimati (non ci sono parametri fissati nella matrice $\boldsymbol{\Lambda}_2$; Bollen, 1989). Inoltre, anche il numero di fattori latenti della EFA non è specificato *a priori* (Bollen, 1989). Un altro assunto riguarda la matrice $\boldsymbol{\Psi}_{\boldsymbol{\varepsilon}}$, in cui gli errori casuali di misurazione non possono covariare tra loro, dando forma ad una matrice diagonale (Bollen, 1989). In aggiunta, vengono fissati a zero sia il vettore delle intercette del modello sia il vettore delle medie dei fattori latenti: $\boldsymbol{\nu}_i = \mathbf{0}_p$ e $\boldsymbol{\alpha}_i = \mathbf{0}_K$. Infine, si assume che i fattori latenti siano incorrelati ($\mathbb{C}ov[\boldsymbol{\xi}_i, \boldsymbol{\xi}_i] = \mathbf{0}_{K \times K}$) e le varianze dei fattori siano unitarie (Caso 1 in Rubin e Thayer, 1982). Difatti, otteniamo che la matrice di varianze-covarianze sia una matrice identità:

$$\boldsymbol{\xi}_i \sim \mathcal{N}_K(\mathbf{0}_K, \boldsymbol{I}_K) \tag{2.24}$$

dove $\boldsymbol{I}_K$ indica la matrice di varianza-covarianza dei fattori latenti $\boldsymbol{\xi}_i$ nella forma di matrice identità $K \times K$. La Figura 2.2 mostra un esempio di modello EFA, in cui osserviamo la correlazioni tra fattori latenti fissata a 0, le varianze unitarie dei fattori latenti e la presenza di coefficienti fattoriali, indicati dalle frecce unidirezionali che partono dai fattori latenti per arrivare alle variabili osservate, che vengono stimati per tutte le osservate.

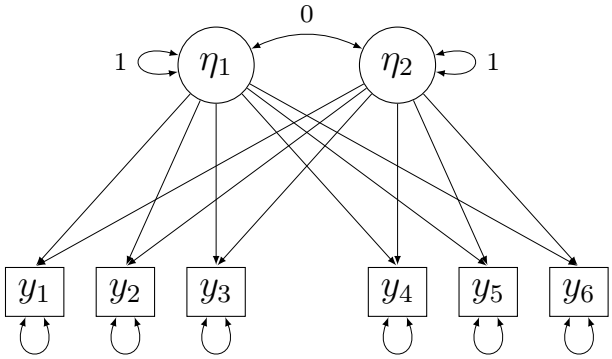


Figura 2.2: Esempio di modello EFA con $p = 6$ e $K = 2$. I nodi circolari indicano le variabili latenti, mentre i nodi rettangolari indicano le variabili osservate. Le frecce unidirezionali indicano la relazione tra i fattori latenti $\{\boldsymbol{\eta}_1, \boldsymbol{\eta}_2\}$, le frecce bidirezionali indicano correlazione se tra nodi diversi, ma se associate allo stesso nodo indicano la varianza (come nel caso del fattore latente $\boldsymbol{\eta}_1$). Lo 0 posto nella relazione tra variabili latenti indica che non viene stimata la correlazione.

In particolare, il modello EFA si dimostra una tecnica utile nel caso in cui l'oggetto di indagine sia pressoché sconosciuto ai/alle ricercatori/ricercatrici rispetto alla sua struttura di associazione indicatori-misurandi, in quanto permette di scoprire possibili schemi latenti nei dati (Bollen, 1989).

Pertanto, il modello ottenuto sarà:

$$\boldsymbol{\xi}_i \sim \mathcal{N}_K(\mathbf{0}_K, \boldsymbol{\Omega}), \quad \text{con} \quad \boldsymbol{\Omega} = \boldsymbol{I}_K \tag{2.25}$$

$$\boldsymbol{\varepsilon}_i \sim \mathcal{N}_p(\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\delta}}) \tag{2.26}$$

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}_2 \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_i \tag{2.27}$$

A questo punto, si presenta la funzione di log-verosimiglianza condizionata delle osservazioni $\boldsymbol{y}$ ai fattori latenti $\boldsymbol{\xi}$:

$$\ell\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}\right) = \log \left[\prod_{i=1}^{n} \frac{\exp\left(-1/2 \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2 \boldsymbol{\xi}_i\right)^T \boldsymbol{\Psi}_{\delta}^{-1} \left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2 \boldsymbol{\xi}_i\right)\right)}{\sqrt{(2\pi)^n \mid \boldsymbol{\Psi}_{\delta} \mid}} \right] \tag{2.28}$$

Mentre, la funzione di log-verosimiglianza dei fattori latenti $\boldsymbol{\xi}$:

$$\ell\left(\boldsymbol{\Omega} \mid \boldsymbol{\xi}\right) = \log\left[\prod_{i=1}^{n} \frac{\exp\left(-1/2\left(\boldsymbol{\xi}_i - \mathbf{0}\right)^T \boldsymbol{\Omega}^{-1}\left(\boldsymbol{\xi}_i - \mathbf{0}\right)\right)}{\sqrt{\left(2\pi\right)^n \mid \boldsymbol{\Omega} \mid}}\right] \tag{2.29}$$

Pertanto, si può definire la log-verosimiglianza completa del modello CFA come:

$$\ell\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}}, \boldsymbol{I} \mid \{\boldsymbol{y}, \boldsymbol{\xi}\}\right) = \ell\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}\right) \ell\left(\boldsymbol{I} \mid \boldsymbol{\xi}\right) \tag{2.30}$$

$$= \log\left[\prod_{i=1}^{n} \frac{\exp\left(-1/2\left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i\right)^T \boldsymbol{\Psi}_{\delta}^{-1}\left(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i\right)\right)}{\sqrt{\left(2\pi\right)^n \mid \boldsymbol{\Psi}_{\delta} \mid}}\right] + \tag{2.31}$$

$$+ \log\left[\prod_{i=1}^{n} \frac{\exp\left(-1/2\left(\boldsymbol{\xi}_i - \mathbf{0}\right)^T \boldsymbol{I}^{-1}\left(\boldsymbol{\xi}_i - \mathbf{0}\right)\right)}{\sqrt{\left(2\pi\right)^n \mid \boldsymbol{I} \mid}}\right] \tag{2.32}$$

2.2 Stima dei parametri

2.2.1 L'algoritmo Expectation-Maximization

L'algoritmo Expectation-Maximization (EM) è stato sviluppato da Dempster et al. (1977), per calcolare stime di ML in presenza di dati incompleti, come nel caso di variabili latenti. Come di seguito approfondiremo brevemente, l'algoritmo EM si basa su una procedura di stima dei parametri iterativa, in cui ogni iterazione è composta da due passi: l'*Expectation* o E-step, ovvero il calcolo dei valori attesi sulla distribuzione condizionata dei dati completi per ogni parametro da stimare, e la *Maximization* o M-step, ovvero il calcolo delle stime di ML tramite derivate parziali su ciascun parametro (McLachlan & Krishnan, 2007).

L'algoritmo EM permette di avere svariati vantaggi nella stima dei parametri nel contesto della ML (McLachlan & Krishnan, 2007): ogni iterazione dell'EM garantisce un incremento montòno dei valori della funzione di verosimiglianza; l'implementazione dell'EM è realizzabile tramite calcoli basati sulla distribuzione completa dei dati, il che può semplificare i calcoli manuali; l'EM può esser utilizzato per provvedere valori stimati di dati mancanti, come variabili latenti. Tuttavia, l'algoritmo EM non è privo di criticità (McLachlan & Krishnan, 2007): la convergenza nell'EM tende ad essere lenta; l'algoritmo EM non garantisce la convergenza ad un massimo globale quando ci sono punti di massimo molteplici e la stima, in questo caso, dipende dai valori iniziali.

Sia $\boldsymbol{Y} = (Y_1, \ldots, Y_p)$ un vettore casuale che si realizza nel vettore osservato $\boldsymbol{y} = (y_1, \ldots, y_p)$ con una funzione di densità di probabilità $P(\boldsymbol{y} \mid \boldsymbol{\theta})$ e spazio parametrico

$\mathcal{Y}$, in cui $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ è un vettore di parametri ignoti con spazio parametrico $\boldsymbol{\Omega}$ (McLachlan & Krishnan, 2007). La funzione di log-verosimiglianza per $\boldsymbol{\theta}$ per i dati osservati $\boldsymbol{y}$ corrisponde a:

$$\ell_{\text{incompleti}}(\boldsymbol{\theta}) = \log\left(P(\boldsymbol{y} \mid \boldsymbol{\theta})\right) \tag{2.33}$$

Sia $\boldsymbol{Z} = (Z_1, \ldots, Z_q)$ il vettore casuale dei dati incompleti o non-osservabili che si realizza nel vettore $\boldsymbol{z} = (z_1, \ldots, z_q)$ con una funzione di densità di probabilità $P(\boldsymbol{z} \mid \boldsymbol{\theta})$. Sia $\boldsymbol{X} = (Y_1, \ldots, Y_p, Z_1, \ldots, Z_q)$ il vettore casuale dei dati completi che si realizza nel vettore $\boldsymbol{x} = (y_1, \ldots, y_p, z_1, \ldots, z_q)$ con una funzione di densità di probabilità $P(\boldsymbol{x} \mid \boldsymbol{\theta})$ e spazio parametrico $\mathcal{X}$. In questo contesto, abbiamo due spazi campionari, $\mathcal{X}$ e $\mathcal{Y}$, e una associazione molti-a-uno dal dominio $\mathcal{X}$ al codominio $\mathcal{Y}$ (McLachlan & Krishnan, 2007). Anziché osservare il vettore dei dati completi $\boldsymbol{x}$ in $\mathcal{X}$, osserviamo il vettore di dati incompleti, considerato una funzione osservabile dei dati completi $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{x})$, in $\mathcal{Y}$ (Dempster et al., 1977). In pratica, i dati completi non sono osservabili direttamente, ma solo tramite i dati osservati.

Nel caso in cui $\boldsymbol{x}$ sia completamente osservabile, sia definita la funzione di log-verosimiglianza dei dati completi come:

$$\ell_{\text{completi}}(\boldsymbol{\theta}) = \log P(\boldsymbol{x} \mid \boldsymbol{\theta}) \tag{2.34}$$

L'algoritmo EM permette di ottenere le stime di ML basate sull'Equazione 2.33, indirettamente, tramite una procedura iterativa basata sulla funzione di log-verosimiglianza dei dati completi $\ell_{\text{completi}}(\boldsymbol{\theta})$ (McLachlan & Krishnan, 2007). Nello specifico, essendo $\ell_{\text{completi}}(\boldsymbol{\theta})$ inosservabile, essa viene sostituita dal suo valore atteso condizionato da $\boldsymbol{y}$, usando le stime correnti dei parametri $\hat{\boldsymbol{\theta}}^{(k)}$, con k che indica l'iterazione corrente (McLachlan & Krishnan, 2007).

Sia $\hat{\boldsymbol{\theta}}^{(0)}$ un valore iniziale per $\boldsymbol{\theta}$. Per ogni iterazione, l'E-step richiede di calcolare:

$$\mathbb{Q}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(0)}) = \mathbb{E}\left[\ell_{\text{completi}}(\boldsymbol{\theta}) \mid \boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(0)}\right] \tag{2.35}$$

Invece, il M-step richiede di massimizzare $\mathbb{Q}(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}^{(0)})$ rispetto a $\boldsymbol{\theta}$ sullo spazio dei parametri $\boldsymbol{\Omega}$ (McLachlan & Krishnan, 2007), in modo da aggiornare il vettore di parametri iniziale $\hat{\boldsymbol{\theta}}^{(0)}$ a $\hat{\boldsymbol{\theta}}^{(1)}$. In particolare, si sceglie $\hat{\boldsymbol{\theta}}^{(1)}$ tale che:

$$\hat{\boldsymbol{\theta}}^{(1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(0)}) \tag{2.36}$$

A questo punto, l'E-step e il M-step vengono reiterati con $\hat{\boldsymbol{\theta}}^{(1)}$ al posto di $\hat{\boldsymbol{\theta}}^{(0)}$ e successivamente, più in generale, per ogni k-esima iterazione. L'algoritmo procede nelle

iterazioni finché non soddisfa il criterio di convergenza basato sulla differenza tra la log-verosimiglianza completa condizionata all'iterazione numero $k-1$ e all'iterazione numero k:

$$\mid \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) - \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k-1)}) \mid < \varepsilon \tag{2.37}$$

dove ε rappresenta un valore numerico di tolleranza.

Di seguito è proposto un riassunto dell'algoritmo EM nei suoi specifici passaggi (Bishop & Nasrabadi, 2006):

1. Scelta del vettore di parametri iniziali $\hat{\boldsymbol{\theta}}^{(k=0)}$;
2. $k = k + 1$
3. **E-step**: calcolo di $\mathbb{E}[\hat{\boldsymbol{\theta}} \mid \boldsymbol{y}, \hat{\boldsymbol{\theta}}^{(k-1)}]$;
4. **M-step**: calcolo di $\hat{\boldsymbol{\theta}}^{(k)} = \arg\max_{\boldsymbol{\theta}} \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k-1)})$;
5. Verifica della convergenza per $\mid \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k)}) - \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}^{(k-1)}) \mid < \varepsilon$;
6. Se il criterio non è soddisfatto si torna al secondo passaggio.

2.2.2 Stima dei parametri del modello CFA tramite EM

Fin dalle origini, l'algoritmo EM è stato applicato a modelli di analisi fattoriale come mostrato in Dempster et al. (1977). Successivamente, Rubin e Thayer (1982) hanno sviluppato 3 modelli di analisi fattoriale ed implementato una procedura di EM per ciascuno di essi. In particolare, gli autori hanno elencato tre casi: il primo con $\boldsymbol{\Phi} = \boldsymbol{I}$ e $\boldsymbol{\Lambda}_1$ senza parametri fissati, il secondo con $\boldsymbol{\Phi} = \boldsymbol{I}$ e $\boldsymbol{\Lambda}_1$ con alcuni parametri fissati e il terzo con $\boldsymbol{\Phi}$ libera di essere stimata e $\boldsymbol{\Lambda}_1$ con alcuni parametri fissi. Inoltre, già Dempster et al. (1977) e successivamente Rubin e Thayer (1982), mostrano un ulteriore vantaggio dell'EM: la possibilità di utilizzare le statistiche sufficienti dei parametri nelle equazioni dell'algoritmo, che ne riducono i costi computazionali. Nel seguito di questa sezione proponiamo i risultati ottenuti nello sviluppo di un algoritmo EM per il modello CFA specificato in Sottosezione 2.1.2.

E-step

Sia $\boldsymbol{\theta} = \{\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \boldsymbol{\Phi}\}$ il vettore di parametri da stimare della CFA e $\hat{\boldsymbol{\theta}}$ è il vettore delle stime. Di seguito viene proposto il calcolo dal valore atteso della log-verosimiglianza

completa dopo aver osservato $\boldsymbol{y}_i$:

$$\begin{aligned}
\mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}}) &= \mathbb{E}\left[\ell\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \boldsymbol{\Phi} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}\right)\right] \\
&= \mathbb{E}\left[\log\left(\prod_{i=1}^{n} \frac{\exp\left(-1/2(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i)^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i)\right)}{\sqrt{(2\pi)^p \mid \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid}} \cdot\right.\right. \\
&\quad \left.\left. \cdot \frac{\exp\left(-1/2(\boldsymbol{\eta}_i - \mathbf{0}_i)^T\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \mathbf{0}_i)\right)}{\sqrt{(2\pi)^q \mid \boldsymbol{\Phi} \mid}}\right) \mid \boldsymbol{y}_i\right] \\
&= \mathbb{E}\left[-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left((\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i)^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i)\right)\right. \\
&\quad \left. -\frac{nq}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Phi} \mid) - \sum_{i=1}^{n}\frac{1}{2}\boldsymbol{\eta}_i^T\boldsymbol{\Phi}^{-1}\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right] \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right] - \right. \\
&\quad -\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i + \left.\mathbb{E}\left[\boldsymbol{\eta}_i^T\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right]\right) - \frac{nq}{2}\log(2\pi) - \\
&\quad - \frac{n}{2}\log(\mid \boldsymbol{\Phi} \mid) - \frac{1}{2}\sum_{i=1}^{n} trace\left(\boldsymbol{\Phi}^{-1}\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\right) \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i] - \right. \\
&\quad -\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i + trace\left(\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\right)\left.\right] - \frac{nq}{2}\log(2\pi) - \\
&\quad - \frac{n}{2}\log(\mid \boldsymbol{\Phi} \mid) - \frac{1}{2}\sum_{i=1}^{n} trace\left(\boldsymbol{\Phi}^{-1}\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\right) \qquad (2.38)
\end{aligned}$$

Si presenta il calcolo dei valori attesi condizionati tramite le proprietà della distribuzione Normale multivariata condizionata (Pace, Salvan et al., 2001):

$$\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i] = \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \mathbf{0}_i) \qquad (2.39)$$

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i] &= \mathbb{V}ar[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i] + \mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T \\
&= \boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi} + \\
&\quad + \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \mathbf{0}_i)[\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \mathbf{0}_i)]^T
\end{aligned} \qquad (2.40)$$

M-step

Di seguito, si calcolano le stime di ML per il vettore dei parametri $\boldsymbol{\theta}$.

Per $\boldsymbol{\Lambda}_1$ si ottiene:

$$\begin{aligned}\frac{\partial \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}_1} &= -\frac{1}{2}\sum_{i=1}^{n} -2\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T + \frac{\partial \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Lambda}_1}\left\{trace\left(\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\right)\right\} \\ &= -\frac{1}{2}\sum_{i=1}^{n}(-2\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T + 2\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]) = 0 \qquad (2.41)\end{aligned}$$

Quindi,

$$\hat{\boldsymbol{\Lambda}}_1 = \frac{\sum_{i=1}^{n}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T}{\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]} \qquad (2.42)$$

Per $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ si ottiene:

$$\begin{aligned}\frac{\partial \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}} &= \frac{\partial \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}}\Bigg\{-\frac{n}{2}\log(\mid \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\Big[\boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right] - \\ &\qquad -\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T\boldsymbol{\Lambda}_1^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}y_i + trace\left(\boldsymbol{\Lambda}^T\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\right)\Big]\Bigg\} \\ &= \frac{n}{2}\boldsymbol{\Theta}_{\boldsymbol{\delta}} - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right]\boldsymbol{y}_i^T + \boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right]\boldsymbol{\Lambda}_1^T\right) \qquad (2.43)\end{aligned}$$

Quindi,

$$\begin{aligned}\hat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}} &= \frac{1}{n}diag\left[\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_1\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}\right]\boldsymbol{y}_i^T + \boldsymbol{y}_i\mathbb{E}\left[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_1^T\right)\right] \qquad (2.44) \\ &= \frac{1}{n}diag\left[\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - \boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T\right)\right] \qquad (2.45)\end{aligned}$$

Per $\boldsymbol{\Phi}$ si ottiene:

$$\frac{\partial \mathbb{Q}(\boldsymbol{\theta} \mid \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\Phi}^{-1}} = \frac{n}{2}\boldsymbol{\Phi} - \sum_{i=1}^{n}\frac{1}{2}\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right] = 0 \qquad (2.46)$$

Quindi,

$$\hat{\boldsymbol{\Phi}} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i\right] \qquad (2.47)$$

2.2.3 Stima dei parametri del modello EFA tramite EM

In questa sezione proponiamo i risultati ottenuti nello sviluppo di un algoritmo EM per il modello EFA specificato in Sottosezione 2.1.2.

E-step

Sia $\boldsymbol{\zeta} = \{\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}}\}$ il vettore di parametri da stimare del modello EFA. Si noti che $\boldsymbol{\Omega}$ non viene stimata, in quanto trattasi di una matrice di identità. Di seguito viene proposto il calcolo dal valore atteso della log-verosimiglianza completa dopo aver osservato $\boldsymbol{y}_i$:

$$
\begin{aligned}
\mathbb{Q}(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}}) &= \mathbb{E}\left[\ell\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}}, \boldsymbol{I} \mid \{\boldsymbol{y}, \boldsymbol{\xi}\}\right)\right] \\
&= \mathbb{E}\left[\log\left(\prod_{i=1}^{n} \frac{\exp\left(-1/2(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i)^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i)\right)}{\sqrt{(2\pi)^p \mid \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid}} \cdot\right.\right. \\
&\quad \left.\left. \cdot \frac{\exp\left(-1/2(\boldsymbol{\xi}_i - \boldsymbol{0}_i)^T\boldsymbol{I}^{-1}(\boldsymbol{\xi}_i - \boldsymbol{0}_i)\right)}{\sqrt{(2\pi)^K \mid \boldsymbol{I} \mid}}\right) \mid \boldsymbol{y}_i\right] \\
&= \mathbb{E}\left[-\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left((\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i)^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i)\right)\right. \\
&\quad \left. -\frac{nK}{2}\log(2\pi) - \sum_{i=1}^{n}\frac{1}{2}\boldsymbol{\xi}_i^T\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right] \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right] -\right. \\
&\quad -\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i + \left.\mathbb{E}\left[\boldsymbol{\xi}_i^T\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right]\right) - \frac{nK}{2}\log(2\pi) - \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right] \\
&= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log(\mid \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i] -\right. \\
&\quad -\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i + \left.trace\left(\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right]\right)\right] - \frac{nK}{2}\log(2\pi) - \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right] \qquad (2.48)
\end{aligned}
$$

Si presenta il calcolo dei valori attesi condizionati tramite le proprietà della distribuzione Normale multivariata condizionata (Pace, Salvan et al., 2001):

$$
\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i] = \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{0}_i) \qquad (2.49)
$$

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i] &= \mathbb{V}ar[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i] + \mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T \\
&= \boldsymbol{I} - \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2 + \\
&\quad + \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{0}_i)[\boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{0}_i)]^T \qquad (2.50)
\end{aligned}
$$

M-step

Vengono calcolate le stime di ML per il vettore dei parametri $\boldsymbol{\zeta}$. Per $\boldsymbol{\Lambda}_2$ si ottiene:

$$\begin{aligned}\frac{\partial \mathbb{Q}(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})}{\partial \boldsymbol{\Lambda}_2} &= -\frac{1}{2}\sum_{i=1}^{n} -2\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T + \frac{\partial \mathbb{Q}(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})}{\partial \boldsymbol{\Lambda}_2}\left\{trace\left(\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right]\right)\right\} \\ &= -\frac{1}{2}\sum_{i=1}^{n}(-2\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T + 2\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}_2\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i]) = 0 \qquad (2.51)\end{aligned}$$

Quindi,

$$\hat{\boldsymbol{\Lambda}}_2 = \frac{\sum_{i=1}^{n}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T}{\sum_{i=1}^{n}\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i]} \qquad (2.52)$$

Per $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$ si ottiene:

$$\begin{aligned}\frac{\partial \mathbb{Q}(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})}{\partial \boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}} &= \frac{\partial \mathbb{Q}(\boldsymbol{\zeta} \mid \hat{\boldsymbol{\zeta}})}{\partial \boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}}\Bigg\{-\frac{n}{2}\log(\mid \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid) - \frac{1}{2}\sum_{i=1}^{n}\left[\boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y}_i - \boldsymbol{y}_i^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right] - \right. \\ &\qquad \left. -\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T\boldsymbol{\Lambda}_2^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}y_i + trace\left(\boldsymbol{\Lambda}^T\boldsymbol{\Psi}_{\boldsymbol{\delta}}^{-1}\boldsymbol{\Lambda}\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right]\right)\right]\Bigg\} \\ &= \frac{n}{2}\boldsymbol{\Psi}_{\boldsymbol{\delta}} - \frac{1}{2}\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right]\boldsymbol{y}_i^T + \boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i\right]\boldsymbol{\Lambda}_2^T\right) \qquad (2.53)\end{aligned}$$

Quindi,

$$\begin{aligned}\hat{\boldsymbol{\Psi}}_{\boldsymbol{\delta}} &= \frac{1}{n}diag\left(\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_2\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}\right]\boldsymbol{y}_i^T + \boldsymbol{y}_i\mathbb{E}\left[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i\right]^T\boldsymbol{\Lambda}_2^T\right)\right) \\ &= \frac{1}{n}diag\left(\sum_{i=1}^{n}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - \boldsymbol{\Lambda}_2\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T\right)\right) \qquad (2.54)\end{aligned}$$

Capitolo 3

Modello di mistura CFA-EFA

3.1 I modelli di mistura

3.1.1 I modelli di mistura nella letteratura scientifica

La letteratura scientifica relativa ai modelli di mistura di funzioni di densità di probabilità o Finite Mixture Models ha una storia piuttosto lunga, iniziata nella seconda metà dell'Ottocento (Holmes, 1892; Pearson, 1894). Attualmente, il numero di pubblicazioni sui modelli di mistura aumenta di anno in anno (McLachlan et al., 2019), testimoniando la rilevanza del contributo che i modelli di mistura stanno dando alla ricerca scientifica. Infatti, i modelli di mistura di funzioni di densità di probabilità permettono rappresentazioni computazionalmente convenienti per il modellamento di distribuzioni complesse di dati (e.g., asimmetriche, con curtosi elevata, multi-modali; McLachlan et al., 2019). Di seguito alcuni degli ambiti scientifici in cui i modelli di mistura hanno trovato una proficua applicazione: le neuroscienze, la bioinformatica, la psicologia, la medicina, l'ingegneria, le scienze sociali, l'economia, la fisica astronomica, ecc. (McLachlan et al., 2019). In queste discipline, in questa classe di modelli sono state generalizzate molteplici tecniche, come la Cluster Analysis, la Latent Class Analysis, la FA, come visto nel Capitolo 1, e molte altre (McLachlan & Peel, 2000). In particolare, i modelli di mistura sono rilevanti per la loro capacità di individuare l'eterogeneità nei dati, che può essere indotta da molteplici distribuzioni di probabilità note o non note. Tuttavia, questa classe di modelli si è rivelata utile anche nell'approssimazione di distribuzioni continue o nel fornire modalità semiparametriche per modellare forme distribuzionali sconosciute (McLachlan & Peel, 2000).

3.1.2 Introduzione formale ai modelli di mistura

Definizione

Sia $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n)$ un campione casuale di n vettori di variabili aleatorie, dove $\boldsymbol{Y}_i$ è un vettore di variabili aleatorie d-dimensionale. Sia $\boldsymbol{y}_i$ il vettore di valori osservati di $\boldsymbol{Y}_i$ e $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ il campione osservato con $\boldsymbol{y} \in \mathbb{R}^d$. La funzione di densità di probabilità di $\boldsymbol{Y}_i$ è uguale a:

$$P(\boldsymbol{y}_i) = \sum_{g=1}^{G} \pi_g P_g(\boldsymbol{y}_i) \tag{3.1}$$

in cui π_g è il parametro che indica la proporzione di mistura con $\pi_g \geq 0$ e $\sum_{g=1}^{G} \pi_g = 1$, mentre $P_g(\boldsymbol{y}_i)$ identifica la componente di densità di probabilità della mistura per l'i-esimo vettore di osservazioni (McLachlan & Peel, 2000). Nello specifico, la densità $P(\boldsymbol{y}_i)$ prende il nome di distribuzione di mistura di G componenti, laddove il numero delle componenti G viene fissato *a priori*. Dall'Equazione 3.1 si può osservare come il contributo della densità di probabilità $P_g(\boldsymbol{y})$ alla forma della distribuzione complessiva $P(\boldsymbol{y})$ sia pesato tramite il parametro di mistura π_g. Inoltre, si assume che le densità componenti siano note sulla base di un vettore di parametri $\boldsymbol{\theta}_g$, per cui si scrive (McLachlan et al., 2019):

$$P(\boldsymbol{y}_i \mid \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g P_g(\boldsymbol{y}_i \mid \boldsymbol{\theta}_g) \tag{3.2}$$

dove, $\boldsymbol{\Psi} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_g, \pi_1, \ldots, \pi_{g-1})$ denota il vettore di parametri ignoti. Spesso, si assume che le densità componenti abbiano una distribuzione proveniente dalla stessa famiglia parametrica, come la distribuzione Normale multivariata (McLachlan et al., 2019).

Esempio di mistura di due distribuzioni Normali univariate

Di seguito viene presentato un esempio di modello di mistura a 2 componenti, in cui entrambe le componenti appartengono alla famiglia parametrica di distribuzioni Normali univariate.

Sia $\boldsymbol{Y}^{(1)} = (Y_1^{(1)}, \ldots, Y_m^{(1)})$ il campione casuale di variabili aleatorie univariate provenienti dalla prima distribuzione Normale, dove $\mathcal{T} = \{1, \ldots, m\}$ e $t \in \mathcal{T}$. Mentre, sia $\boldsymbol{Y}^{(2)} = (Y_1^{(2)}, \ldots, Y_l^{(2)})$ il campione causale generato dalla seconda distribuzione Normale con $\mathcal{H} = \{1, \ldots, l\} \setminus \mathcal{T}$ e $h \in \mathcal{H}$. Siano, quindi, $\boldsymbol{y}^{(1)} = (y_1^{(1)}, \ldots, y_m^{(1)})$ e

$\boldsymbol{y}^{(2)} = (y_1^{(2)}, \ldots, y_l^{(2)})$ i due campioni casuali dei valori osservati rispettivamente della prima e della seconda distribuzione Normale. Formalmente:

$$Y_t^{(1)} \sim \mathcal{N}(\mu_1, \sigma_1^2) \tag{3.3}$$
$$Y_h^{(2)} \sim \mathcal{N}(\mu_2, \sigma_2^2) \tag{3.4}$$

dove, μ_1 e σ_1^2 rappresentano, rispettivamente, la media e la varianza ignote della prima distribuzione Normale, mentre μ_2 e σ_2^2 indicano la media e la varianza della seconda. Assumiamo, inoltre, che $\boldsymbol{Y}^{(1)}$ e $\boldsymbol{Y}^{(2)}$ siano indipendenti tra loro. In continuità con il sottocapitolo precedente, per $g = \{1, 2\}$ siano le due densità componenti tali che:

$$P_1(y_t^{(1)} \mid \boldsymbol{\theta}_1) = \mathcal{N}(y_t^{(1)} \mid \mu_1, \sigma_1^2) \tag{3.5}$$
$$P_2(y_h^{(2)} \mid \boldsymbol{\theta}_2) = \mathcal{N}(y_h^{(2)} \mid \mu_2, \sigma_2^2) \tag{3.6}$$

dove $\mathcal{N}(y_t^{(1)} \mid \mu_1, \sigma_1^2)$ e $\mathcal{N}(y_h^{(2)} \mid \mu_2, \sigma_2^2)$ corrisponde a:

$$\mathcal{N}(y_t^{(1)} \mid \mu_1, \sigma_1^2) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{y_t^{(1)} - \mu_1}{2\sigma_1}\right)^2 \tag{3.7}$$
$$\mathcal{N}(y_h^{(2)} \mid \mu_2, \sigma_2^2) = \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{y_h^{(2)} - \mu_2}{2\sigma_2}\right)^2 \tag{3.8}$$

Definite le due densità componenti della mistura, si sviluppa un modo banale per generare il vettore casuale di mistura. Pertanto, sia $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ un vettore di variabili aleatorie discrete con realizzazioni campionarie $\boldsymbol{z} = (z_1, \ldots, z_n)$ e supporto $z_i \in \{0, 1\}$, con $n \leq m + l$ e $i \in \{1, \ldots, n\}$. Tale vettore casuale discreto avrà una distribuzione Bernoulliana:

$$Z_i \sim \mathcal{B}ern(\pi_1) \tag{3.9}$$

dove, π_1 è la proporzione di mistura della prima densità componente[1] e si noti che $P(z_i = 1) = \pi_1$. Quindi, il vettore $\boldsymbol{z}$ contiene tutti gli indici associati alla prima densità componente, rappresentati dal numero 1, e gli indici associati alla seconda densità componente, rappresentati dal numero 0.

Di conseguenza, si possono ottenere le densità componenti precedentemente definite come densità condizionate dal valore delle variabili osservate z_i nel vettore $\boldsymbol{z}$:

$$P_1(y_t^{(1)} \mid z_i = 1, \mu_1, \sigma_2^2) = \mathcal{N}(y_t^{(1)} \mid \mu_1, \sigma_1^2) \tag{3.10}$$
$$P_2(y_h^{(2)} \mid z_i = 0, \mu_2, \sigma_2^2) = \mathcal{N}(y_h^{(2)} \mid \mu_2, \sigma_2^2) \tag{3.11}$$

[1] π_2 non compare in quanto equivale a $(1 - \pi_1)$.

Sia $\boldsymbol{Y} = (Y_t^{(1)}, \ldots, Y_m^{(1)}, Y_h^{(2)}, \ldots, Y_l^{(2)}) = (Y_1, \ldots, Y_n)$ il vettore casuale che contiene una mistura di variabili aleatorie della prima e della seconda distribuzione Normale in una proporzione definita da π_1, per la prima, e π_2, per la seconda. Sia $\boldsymbol{y} = (y_t^{(1)}, \ldots, y_m^{(1)}, y_h^{(2)}, \ldots, y_l^{(2)}) = (y_1, \ldots, y_n)$ il vettore di realizzazioni di $\boldsymbol{Y}$. Perciò, il campionamento che porta alla generazione di ciascuna variabile di mistura procede nel seguente modo:

$$Y_i \mid Z_i = 1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \tag{3.12}$$

$$Y_i \mid Z_i = 0 \sim \mathcal{N}(\mu_2, \sigma_2^2) \tag{3.13}$$

dove si nota come la variabile indicatrice selezioni per ogni osservazione i-esima se la provenienza deve essere dalla prima o dalla seconda distribuzione Normale.

A questo punto, si può definire il modello di mistura come somma di densità condizionate che permettono di ottenere la densità marginale di y_i, quale mistura di due vettori di realizzazioni campionarie, ovvero $y_t^{(1)}$ e $y_h^{(2)}$:

$$P(y_l \mid \pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2) = \pi_1 \mathcal{N}(y_t^{(1)} \mid \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(y_h^{(2)} \mid \mu_2, \sigma_2^2) \tag{3.14}$$

dove si ricordi che $\pi_2 = (1 - \pi_1)$. Nell'Equazione 3.14 è evidente come la proporzione di mistura determini quanto una delle due densità componenti prevarrà sull'altra nei termini della forma della distribuzione finale.

Tuttavia, non è noto se una generica osservazione y_i appartenga alla prima o alla seconda distribuzione Normale. Infatti, la distribuzione che viene osservata dal/dalla ricercatore/ricercatrice è una distribuzione in cui il campione di realizzazioni $\boldsymbol{y}^{(1)}$ e $\boldsymbol{y}^{(2)}$ sono già mischiati tra loro. Per chiarire ulteriormente, la Figura 3.1 mostra in viola le realizzazioni campionarie y_i di una distribuzione di mistura generata artificialmente con $n = 500$, rappresentata tramite stima kernel di densità[2]. Si può facilmente ipotizzare che la distribuzione in viola sia una mistura di due densità Normali, la prima in blu e la seconda in rosso. Difatti, la distribuzione di mistura composta dalle osservazioni y_i in viola è stata generata campionando dalla distribuzione Normale blu con $\mu_{\text{blu}} = -2$ e $\sigma^2_{\text{blu}} = 1.5$, selezionata da $z_i = 1$, e dalla distribuzione Normale rossa $\mu_{\text{rossa}} = 2$ e $\sigma^2_{\text{rossa}} = 1$, selezionata da $z_i = 0$. Come si osserva anche in figura, le due distribuzioni sono state mischiate in modo piuttosto bilanciato e, infatti, il parametro di mistura è stato fissato con $\pi_{\text{blu}} = 0.5$.

[2]Il codice del grafico si trova nell'Appendice A.1.

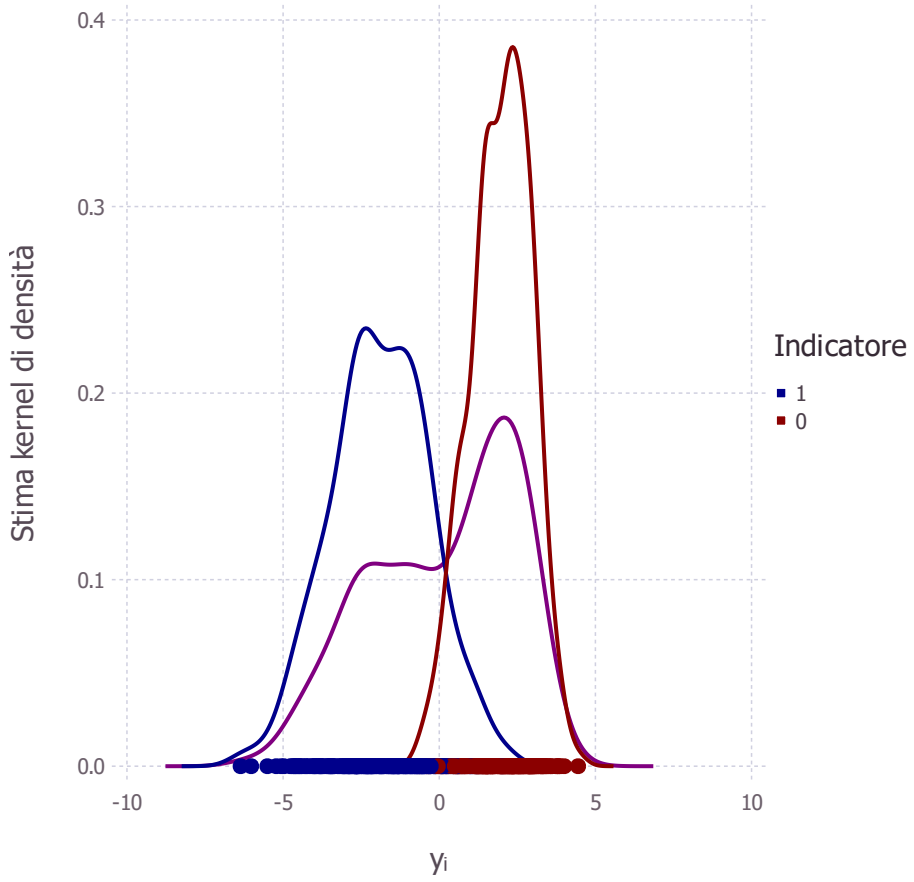


Figura 3.1: Densità kernel di una mistura di due distribuzioni Normali con $n = 500$. La distribuzione blu indica una distribuzione Normale con $\mu_{\text{blu}} = -2$ e $\sigma^2_{\text{blu}} = 1.5$ e la distribuzione in rosso indica una Normale con parametri $\mu_{\text{rosso}} = 2$ e $\sigma^2_{\text{blu}} = 1$. La distribuzione viola indica la distribuzione risultante dalla mistura delle precedenti.

Si applichi, adesso, l'esempio proposto in Figura 3.1 ad un contesto di ricerca, in cui i/le ricercatori/ricercatrici hanno compiuto un campionamento con lo scopo di raccogliere dati da una popolazione di interesse. Si immagini che i/le ricercatori/ricercatrici abbiano ottenuto come distribuzione empirica dei dati la distribuzione di mistura in viola. Ciò rappresenta un problema, che, se non gestito adeguatamente, può diventare grave a tal punto da rendere invalidi i risultati di una ricerca. Si supponga che i/le ricercatori/ricercatrici dell'esempio abbiano, inconsapevolmente, campionato osservazioni generate da due popolazioni parametriche differenti, rendendo impossibile fare inferenza sulla base di un siffatto campione. Ad esempio, potrebbero aver somministrato un questionario su temi relativi alla sfera della sessualità ad adolescenti che provengono

da due scuole differenti, non avendo saputo che, precedentemente, in una delle due scuole era stato condotto un programma di interventi sulla sessualità e l'affettività da parte di psicologi formati. In tal caso, applicare un modello di mistura per analizzare le risposte al questionario, verosimilmente, potrebbe rilevare l'eterogeneità presente nel campione dei partecipanti, in modo da adottare strategie di analisi adatte al caso.

Pertanto, l'utilità dei modelli di mistura risiede nella loro capacità di concettualizzare situazioni in cui il campione può essere composto da più popolazioni, ovvero nel caso di campioni eterogenei. In particolare, le potenzialità di questa classe di modelli si esprimono al meglio nell'individuare e separare i sotto-campioni in presenza di eterogeneità campionaria non-osservata, come visto nella Sottosezione 1.2.2. Nel concreto, lo scopo dei modelli di mistura consiste nell'effettuare una stima separata sia dei parametri, che contraddistinguono le singole popolazioni, sia della proporzione in cui tali popolazioni sono presenti nei dati. Come osservato nell'esempio in Figura 3.1 e delle due scuole, i modelli di mistura possono fare la differenza nel comprendere più in dettaglio le caratteristiche del campione raccolto e nell'evitare di trarre da esso conclusioni errate.

3.2 Il modello di Mix-CFA-EFA

Quanto visto nell'esempio di modello di mistura sviluppato nella Sottosezione 3.1.2 ricalca concettualmente quanto verrà proposto in questa Sezione. In primo luogo, anziché definire un modello di mistura a due componenti tramite due densità Normali univariate, verranno utilizzare le densità del modello CFA e del modello EFA. In secondo luogo, per definire il meccanismo statistico, che determina la modalità in cui le due densità del modello CFA e del modello EFA si mischiano tra loro, si definisce una variabile latente che indica per ogni osservazione se questa provenga dalla densità componente del modello CFA o del modello EFA.

Il modello di mistura qui proposto, denominato Mix-CFA-EFA, è stato sviluppato per fornire ai/alle ricercatori/ricercatrici una tecnica statistica, che consenta di esplorare più in dettaglio le caratteristiche di dati raccolti tramite somministrazioni di test psicologici o questionari. Quando un/una ricercatore/ricercatrice ha intenzione di testare le proprie ipotesi sul legame tra un insieme di items ed uno o più fattori latenti, il primo passo è definire un modello CFA per una popolazione di interesse, fissando o liberando la stima dei coefficienti fattoriali della matrice $\mathbf{\Lambda}_1$. In secondo luogo, è necessario raccogliere un campione dalla popolazione di interesse. Ottenuto il campione, il modello CFA può essere adattato ai dati. Nel processo di stima dei parametri, la matrice di covarianza stimata dal modello tenderà ad approssimare la matrice di covarianza osservata. Tuttavia, se la

matrice di covarianza osservata è stata alterata dalla presenza nel campione di soggetti che non appartengono alla popolazione di interesse, le stime non saranno né generalizzabili alla popolazione di interesse né valide di per sé (Becker et al., 2013). Si ricorda che il rumore nella misurazione indotto dalla presenza di eterogeneità campionaria non gestita può condurre a molteplici conseguenze indesiderate, come visto nella Sottosezione 1.1.3.

In questo contesto, lo scopo principale del modello Mix-CFA-EFA consiste nello stimare la proporzione di unità campionarie, che mostrano modalità di risposta differenti da quelle ipotizzate dai/dalle ricercatori/ricercatrici nel modello CFA, per assicurarsi che la validità di misurazione sia preservata (Zumbo, 2006). Considerando che tali differenze nel processo di risposta possono riflettersi nella matrice di covarianza osservata degli items, l'applicazione simultanea di un modello CFA ed un modello EFA dovrebbe indirizzare il modello EFA a catturare tutte quelle osservazioni che non riproducono la matrice di covarianza ipotizzata nel modello CFA (i.e., $\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}}$). Più in concreto, la quota di variabilità nei dati non spiegata dalla CFA sarebbe catturata e stimata tramite un modello EFA. Pertanto, la presenza di un modello EFA, come componente di mistura, può avere due funzioni, differenti ma complementari, a seconda degli scopi applicativi: (1) epurare la stima del modello CFA ipotizzato di tutte quelle osservazioni che, per motivi di eterogeneità campionaria, potrebbero rappresentare una quota di "rumore" nell'applicazione di un modello CFA standard; (2) identificare quale gruppo di specifiche osservazioni campionarie diverge rispetto al modello CFA ipotizzato, permettendo di sviluppare ipotesi sull'origine di tale eterogeneità, ad esempio, sulla base dell'uso di indici di classificazione rispetto a covariate esterne[3].

La prima funzione del modello Mix-CFA-EFA trova il proprio contesto applicativo ottimale, nel caso in cui i/le ricercatori/ricercatrici abbiano il dubbio che un campione, in una proporzione ristretta, sia contaminato da osservazioni provenienti da altre popolazioni. In questo contesto, il modello Mix-CFA-EFA potrebbe stimare i parametri della componente CFA sulla maggior parte dei dati, fornendo stime migliori, ed individuare in modo più netto le unità statistiche fonte di "rumore" nella misurazione, gestendole separatamente. Si ipotizzano tre casi specifici in cui il modello Mix-CFA-EFA potrebbe essere applicato con profitto: in presenza di più dataset simili combinati tra loro, anche provenienti da ricerche differenti; nel caso di campionamento non-probabilistico condotto su temi di ricerca particolarmente complessi; in presenza di faking o di possibili bias dati da stili di risposta (Ziegler et al., 2015).

La seconda funzione del presente modello di mistura permette di indagare più in

[3]Come illustrato nell'applicazione su dati reali in Sezione 4.2.

dettaglio le origini dell'eterogeneità, in quanto il modello Mix-CFA-EFA consente di classificare le osservazioni nella densità componente CFA o nella EFA. Ciò potrebbe risultare utile in presenza di covariate, in quanto il vettore delle osservazioni classificate come appartenenti al modello CFA o EFA può indicare se tale classificazione sia o meno sovrapponibile a determinati valori nelle covariate (come si vedrà nell'applicazione in Sezione 4.2). Pertanto, il modello Mix-CFA-EFA potrebbe rappresentare una strategia per indagare l'eterogeneità tramite l'ispezione del vettore delle osservazioni classificate.

3.2.1 Specificazione del modello

Come anticipato precedentemente, il presente lavoro intende proporre un modello di mistura che abbia come densità componenti un modello CFA ed un modello EFA, ovvero il modello Mix-CFA-EFA. Al fine di specificare il modello di mistura, saranno riprese le definizioni precedentemente delineate del modello CFA e del modello EFA, descritte nella Sottosezione 2.1.2. Di seguito, vengono riproposte sinteticamente le definizioni del modello CFA e del modello EFA:

CFA	EFA
$\underset{p\times 1}{\boldsymbol{y}_t^{\text{CFA}}} = \underset{p\times qq\times 1}{\boldsymbol{\Lambda}_1\,\boldsymbol{\eta}_t} + \underset{p\times 1}{\boldsymbol{\delta}_t}$	$\underset{p\times 1}{\boldsymbol{y}_h}^{\text{EFA}} = \underset{p\times KK\times 1}{\boldsymbol{\Lambda}_2\,\boldsymbol{\xi}_h} + \underset{p\times 1}{\boldsymbol{\varepsilon}_h}$
$\boldsymbol{\eta}_t \sim N_q(\mathbf{0}_q, \boldsymbol{\Phi})$	$\boldsymbol{\xi}_h \sim N_K(\mathbf{0}_K, \boldsymbol{\Omega}), \quad \text{con} \quad \boldsymbol{\Omega} = \boldsymbol{I}_K$
$\boldsymbol{\delta}_t \sim N_p(\mathbf{0}_p, \boldsymbol{\Theta}_{\boldsymbol{\delta}})$	$\boldsymbol{\varepsilon}_h \sim N_p(\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\delta}})$
$\boldsymbol{y}_t^{\text{CFA}} \sim N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})$	$\boldsymbol{y}_h^{\text{EFA}} \sim N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})$
$\boldsymbol{y}_t^{\text{CFA}} \mid \boldsymbol{\eta}_t \sim N_p(\mathbf{0}_p, \boldsymbol{\Theta}_{\boldsymbol{\delta}})$	$\boldsymbol{y}_h^{\text{EFA}} \mid \boldsymbol{\xi}_h \sim N_p(\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\delta}})$

dove utilizziamo y_t^{CFA} e y_h^{EFA} per indicare le realizzazioni campionarie derivanti dai rispettivi modelli CFA ed EFA, con, rispettivamente, $\mathcal{T} = \{1, \ldots, m\}$ e $t \in \mathcal{T}$ per y_t^{CFA} e con $\mathcal{H} = \{1, \ldots, l\} \setminus \mathcal{T}$ e $h \in \mathcal{H}$ per y_h^{EFA}.

Il meccanismo statistico alla base della generazione della mistura si assume che sia governato da un vettore di variabili indicatrici. Pertanto, sia $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$, con $n \leq m + l$, un vettore di variabili aleatorie latenti indicatrici e $\boldsymbol{z} = (z_1, \ldots, z_n)$ il vettore delle sue realizzazioni campionarie con $z_i \in \{0, 1\}$. Le realizzazioni z_i della variabile aleatoria latente hanno il fine di selezionare la presenza di un vettore di valori osservati dalla CFA $\boldsymbol{y}_t^{\text{CFA}}$ o dalla EFA $\boldsymbol{y}_h^{\text{CFA}}$. Si assume, inoltre, che le variabili aleatorie latenti Z_i si distribuiscano come segue:

$$Z_i \sim \mathcal{B}ern(\kappa_1) \tag{3.15}$$

dove $\mathcal{B}ern$ indica una distribuzione Bernoulliana e κ_1 rappresenta il parametro della proporzione di mistura a favore delle osservazioni generate dal modello CFA. Per $g = 1, 2$ sia $0 \geq \kappa_g \geq 1$ e $\sum_{g=1}^{G} \kappa_g = 1$[4]. Si noti che la funzione di verosimiglianza di z_i è definita come segue:

$$L(\kappa \mid \boldsymbol{z}) = \prod_{i=1}^{n} \kappa^{z_i}(1-\kappa)^{1-z_i} \tag{3.16}$$

Pertanto, per ogni $z_i = 1$ si seleziona un'osservazione derivante dal modello CFA, mentre per ogni $z_i = 0$ si seleziona un'osservazione derivante dal modello EFA.

Sia $\boldsymbol{Y} = (Y_t^{\text{CFA}}, \ldots, Y_m^{\text{CFA}}, Y_h^{\text{EFA}}, \ldots, Y_l^{\text{EFA}}) = (Y_1, \ldots, Y_n)$ la matrice casuale contenente i vettori di variabili aleatorie campionate dalle due densità della CFA e della EFA. Inoltre, $\boldsymbol{y} = (y_t^{\text{CFA}}, \ldots, y_m^{\text{CFA}}, y_h^{\text{EFA}}, \ldots, y_l^{\text{EFA}}) = (y_1, \ldots, y_n)$ indica la matrice casuale che raccoglie i vettori delle realizzazioni della matrice $\boldsymbol{Y}$. La matrice $\boldsymbol{y}$ è costituita da righe che provengono dal modello EFA e dal modello CFA in proporzioni complementari.

Più formalmente, si descrive il processo di campionamento governato dal vettore di variabili indicatrici latenti $\boldsymbol{Z}$ per ottenere la distribuzione di mistura come:

$$\boldsymbol{Y}_i \mid Z_i = 1 \sim N_p(\mathbf{0}_i, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}}) \tag{3.17}$$

$$\boldsymbol{Y}_i \mid Z_i = 0 \sim N_p(\mathbf{0}_i, \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}}) \tag{3.18}$$

dove, $\boldsymbol{Y}_i \mid Z_i = 1$ corrisponde a $\boldsymbol{Y}_l^{\text{CFA}}$ e $\boldsymbol{Y}_i \mid Z_i = 0$ a $\boldsymbol{Y}_i^{\text{EFA}}$; di qui in poi verrà adottata questa nuova notazione per indicare le variabili osservate provenienti dal modello CFA e dal modello EFA.

Per rendere più evidente che il modello ottenuto sia composto da due variabili latenti poste in ordine gerarchico dal punto di vista del meccanismo generatore dei dati $\boldsymbol{y}_i$, si noti che la probabilità marginale delle osservazioni $\boldsymbol{y}_i$ del modello Mix-CFA-EFA è definita come:

$$\begin{aligned} P(\boldsymbol{y}_i \mid \boldsymbol{\Gamma}) = &\int_{\mathbb{R}} P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, z_i = 1)\, P(\boldsymbol{\eta}_i \mid z_i = 1)\, P(z_i = 1)\, d\boldsymbol{\eta}_i + \\ &+ \int_{\mathbb{R}} P(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i, z_i = 0)\, P(\boldsymbol{\xi}_i \mid z_i = 0)\, P(z_i = 0)\, d\boldsymbol{\xi}_i \end{aligned} \tag{3.19}$$

dove $\boldsymbol{\Gamma} = \{\boldsymbol{\eta}_i, \boldsymbol{\xi}_i, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2, \boldsymbol{\Theta}_\delta, \boldsymbol{\Psi}_\delta, \boldsymbol{\Phi}, \kappa\}$ rappresenta il vettore di parametri ignoti, da cui dipende la realizzazione $\boldsymbol{y}_i$ e si noti che $P(z_i = 0) = (1 - P(z_i = 1)) = (1 - \kappa)$. Mentre la probabilità congiunta di $\{\boldsymbol{y}_i, \boldsymbol{y}_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i, z_i\}$ è definita come segue:

$$\begin{aligned} P(\boldsymbol{y}_i, \boldsymbol{y}_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i, z_i) = &[P(\boldsymbol{y}_i \mid \boldsymbol{\eta}_i, z_i = 1)\, P(\boldsymbol{\eta}_i \mid z_i = 1)\, P(z_i = 1)]^{z_i} \cdot \\ &\cdot [P(\boldsymbol{y}_i \mid \boldsymbol{\xi}_i, z_i = 0)\, P(\boldsymbol{\xi}_i \mid z_i = 0)\, P(z_i = 0)]^{1-z_i} \end{aligned} \tag{3.20}$$

[4]Di qui in poi, κ_1 verrà indicato più semplicemente con κ e κ_2 verrà indicato con $(1 - \kappa)$.

A questo punto, si presenta l'equazione della log-verosimiglianza completa del modello Mix-CFA-EFA corrisponde a:

$$
\begin{aligned}
\ell(\boldsymbol{\Gamma}) &= \log\Bigg(\prod_{i=1}^{n}\left[L\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}, z_i = 1\right) L\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}, z_i = 1\right) \kappa\right]^{z_i} \cdot \\
&\qquad \cdot \prod_{i=1}^{n}\left[L\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}, z_i = 0\right) L(\boldsymbol{\Omega} \mid \boldsymbol{\xi}, z_i = 0)(1-\kappa)\right]^{1-z_i}\Bigg) \\
&= \log\Bigg(\left[L\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}, z_i = 1\right) L\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}, z_i = 1\right) \kappa\right]^{\sum_{i=1}^{n} z_i} \cdot \\
&\qquad \cdot \left[L\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}, z_i = 0\right) L(\boldsymbol{\Omega} \mid \boldsymbol{\xi}, z_i = 0)(1-\kappa)\right]^{\sum_{i=1}^{n}(1-z_i)}\Bigg) \\
&= \sum_{i=1}^{n} z_i \left[\ell\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}, z_i = 1\right) + L\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}, z_i = 1\right) + \log(\kappa)\right] + \\
&\qquad + \sum_{i=1}^{n}(1-z_i)\left[L\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}, z_i = 0\right) L(\boldsymbol{\Omega} \mid \boldsymbol{\xi}, z_i = 0) + \log(1-\kappa)\right] \\
&= \sum_{i=1}^{n} z_i \left[\log(\kappa) + \log\left(\frac{\exp(-1/2(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i)^T\boldsymbol{\Theta}_{\delta}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_1\boldsymbol{\eta}_i))}{\sqrt{(2\pi)^n \mid \boldsymbol{\Theta}_{\delta} \mid}}\right) + \right. \\
&\qquad \left. + \log\left(\frac{\exp(-1/2(\boldsymbol{\eta}_i - \mathbf{0})^T\boldsymbol{\Phi}^{-1}(\boldsymbol{\eta}_i - \mathbf{0}))}{\sqrt{(2\pi)^n \mid \boldsymbol{\Phi} \mid}}\right)\right] + \\
&\qquad + \sum_{i=1}^{n}(1-z_i)\left[\log(1-\kappa) + \log\left(\frac{\exp(-1/2(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i)^T\boldsymbol{\Psi}_{\delta}^{-1}(\boldsymbol{y}_i - \boldsymbol{\Lambda}_2\boldsymbol{\xi}_i))}{\sqrt{(2\pi)^n \mid \boldsymbol{\Psi}_{\delta} \mid}}\right) + \right. \\
&\qquad \left. + \log\left(\frac{\exp(-1/2(\boldsymbol{\xi}_i - \mathbf{0})^T\boldsymbol{\Omega}^{-1}(\boldsymbol{\xi}_i - \mathbf{0}))}{\sqrt{(2\pi)^n \mid \boldsymbol{\Omega} \mid}}\right)\right] \qquad (3.21)
\end{aligned}
$$

dove $\{L\left(\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y}, \boldsymbol{\eta}\}, z_i = 1\right), L\left(\boldsymbol{\Phi} \mid \boldsymbol{\eta}, z_i = 1\right)\}$ corrispondono alle log-verosimiglianze definite per il modello CFA, rispettivamente, nella Equazione 2.16 e nell'Equazione 2.17, così come $\{L\left(\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}} \mid \{\boldsymbol{y} \mid \boldsymbol{\xi}\}, z_i = 0\right), L(\boldsymbol{\Omega} \mid \boldsymbol{\xi}, z_i = 0)\}$ corrispondono alle log-verosimiglianze definite per il modello EFA, rispettivamente, nell'Equazione 2.28 e nell'Equazione 2.29.

La Figura 3.2 illustra un modello Mix-CFA-EFA con $q = 2$ e $K = 2$. In particolare, si osservino in alto a sinistra e a destra le rappresentazioni grafiche di un modello CFA, in blu, e di un modello EFA, in rosso e al di sotto le rispettive equazioni definitorie. Si noti, più in basso, un cerchio che indica l'insieme delle osservazioni campionarie, rappresentate da figure umane stilizzate. Le frecce che partono dalle equazioni del modello CFA, in blu, e del modello EFA, in rosso, intendono mostrare l'estrazione della specifica osservazione campionaria dalla popolazione descritta dai parametri assunti dal modello statistico

CFA ed EFA, la cui derivazione è simboleggiata dal colore blu o rosso. Inoltre, si ponga attenzione sul fatto che il colore del cerchio, che circoscrive il campione raccolto, è il viola, colore che risulta dalla mistura tra blu e rosso, ed indica l'eterogeneità presente nei dati.

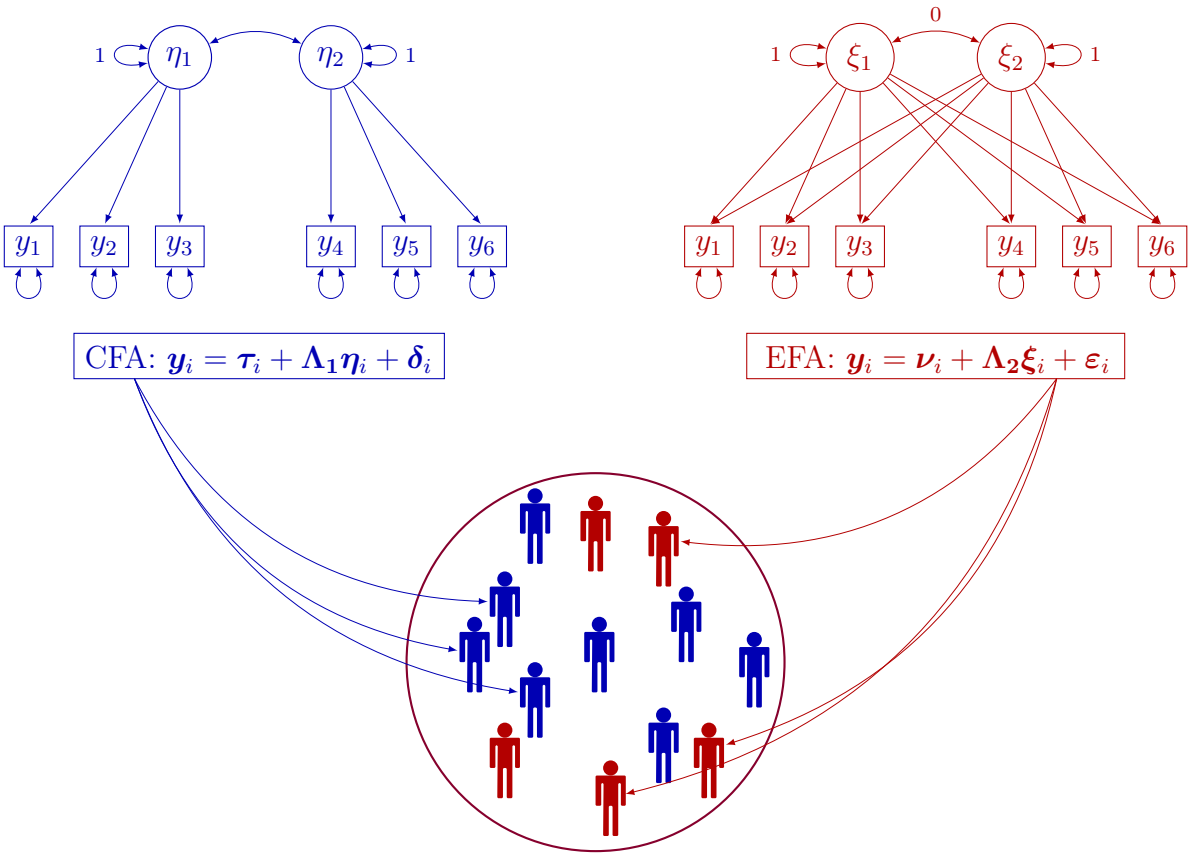


Figura 3.2: Illustrazione del modello di Mix-CFA-EFA con $q = 2$ e $K = 2$. In alto a sinistra e a destra sono disposte le rappresentazioni grafiche di un modello CFA, in blu, e di un modello EFA, in rosso, e al di sotto le rispettive equazioni definitorie. Più in basso il cerchio indica l'insieme delle osservazioni campionarie, rappresentate da figure umane stilizzate. Le frecce che partono dalle equazioni del modello CFA, in blu, e del modello EFA, in rosso, mostrano l'estrazione della specifica osservazione campionaria dalla popolazione parametrizzata dal modello statistico CFA ed EFA, la cui derivazione è, appunto, simboleggiata dal colore blu o rosso. Il colore del cerchio è il viola, colore che risulta dalla mistura tra blu e rosso, ed indica l'eterogeneità presente nei dati.

3.2.2 Stima dei parametri del modello Mixture CFA-EFA tramite EM

Il problema della stima dei parametri del modello Mix-CFA-EFA è dato da tre livelli di informazioni non note. Riprendendo la Figura 3.2, i tre livelli di informazioni ignote si possono rappresentare come segue: i parametri presenti nelle equazioni definitorie dei modelli CFA ed EFA, la proporzione di osservazioni di colore blu e di colore rosso nel campione e se la freccia unidirezionale per ciascuna osservazione parta dal modello CFA o del modello EFA. Più formalmente, le tre informazioni sono, rispettivamente: i parametri dei singoli modelli CFA (i.e., $\{\boldsymbol{\Lambda}_1, \boldsymbol{\Theta}_{\boldsymbol{\delta}}, \boldsymbol{\Phi}\}$) ed EFA (i.e., $\{\boldsymbol{\Lambda}_2, \boldsymbol{\Psi}_{\boldsymbol{\delta}}\}$, la proporzione in cui le osservazioni generate dal modello CFA e dal modello EFA sono presenti nei dati, descritta dal parametro di mistura κ, e l'assegnazione di ogni specifica osservazione ad una o all'altra componente, indicata dalla variabile indicatrice z_i.

Più nello specifico, il problema di stima dei parametri del modello Mix-CFA-EFA riguarda la presenza delle variabili latenti dei fattori $\{\boldsymbol{\eta}, \boldsymbol{\xi}\}$ e del vettore di indicatori $\boldsymbol{z}$. Per condurre la stima dei parametri è stato scelto l'algoritmo EM, presentato nella Sottosezione 2.2.1. Difatti, l'algoritmo EM consente di risolvere problemi di stima, in cui vi sono variabili latenti o dati mancanti.

Tuttavia, l'algoritmo EM, nel contesto dei modelli di mistura di FA o SEM, richiede di utilizzare numerosi set diversi di starting points (valori di inizializzazione) per evitare la non-convergenza dell'algoritmo (Tueller & Lubke, 2010), in quanto la convergenza viene fortemente influenzata dai valori iniziali (Hipp & Bauer, 2006). Inoltre, siccome in questi casi la funzione di log-verosimiglianza non è limitata, è possibile che determinati starting points conducano ad una convergenza in più punti di massimo locale (McLachlan & Peel, 2000), restituendo differenti stime finali (Dolan & van der Maas, 1998). Pertanto, è consigliabile stimare molteplici modelli per individuare il massimo locale raggiunto con maggior frequenza (Dolan & van der Maas, 1998). L'algoritmo EM è a tal punto sensibile ai valori iniziali, che Hoshino (2001) raccomanda di cambiare starting points finché il modello non converge. Durante l'ottimizzazione è anche possibile che le matrici di covarianza del modello CFA o EFA non siano positive, portando a non-convergenza dell'algoritmo (Dolan & van der Maas, 1998).

Expectation-Maximization

In primo luogo, si calcola le probabilità delle variabili indicatrici latenti $z_i = 1$ e $z_i = 0$ condizionate ai dati $\boldsymbol{y}_i$:

$$\begin{aligned} P(z_i = 1 \mid \boldsymbol{y}_i) &= \frac{P(z_i = 1)P(\boldsymbol{y}_i \mid z_i = 1)}{P(z_i = 1)P(\boldsymbol{y}_i \mid z_i = 1) + P(\boldsymbol{y}_i \mid z_i = 0) - P(z_i = 1)P(\boldsymbol{y}_i \mid z_i = 0)} \\ &= \frac{\kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_\delta)}{\kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_\delta) + N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_\delta) - \kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_\delta)} \end{aligned} \tag{3.22}$$

$$P(z_i = 0 \mid \boldsymbol{y}_i) = 1 - P(z_i = 1 \mid \boldsymbol{y}_i) \tag{3.23}$$

E-step

Si presenta il valore atteso della funzione di log-likelihood completa (Equazione 3.21) condizionato ai dati $\boldsymbol{y}_i$:

$$\begin{aligned} \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}}) &= \mathbb{E}[\ell(\boldsymbol{\Gamma}) \mid \boldsymbol{y}_i] \\ &= \sum_{i=1}^{n} \mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] \Big\{ \log(\kappa) - \frac{p}{2}\log(2\pi) - \frac{1}{2}\log(\mid \boldsymbol{\Theta}_\delta \mid) - \\ &\quad - \frac{1}{2}\left(\boldsymbol{y}_i^T \boldsymbol{\Theta}_\delta^{-1} \boldsymbol{y}_i - 2\boldsymbol{y}_i^T \boldsymbol{\Theta}_\delta^{-1} \boldsymbol{\Lambda}_1 \mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i] + trace\left(\boldsymbol{\Lambda}_1^T \boldsymbol{\Theta}_\delta^{-1} \boldsymbol{\Lambda}_1 \mathbb{E}[\boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\right)\right) - \frac{q}{2}\log(2\pi) - \\ &\quad - \frac{1}{2}\log(\mid \boldsymbol{\Phi} \mid) - \frac{1}{2} trace\left(\boldsymbol{\Phi}^{-1} \mathbb{E}[\boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\right) \Big\} + \sum_{i=1}^{n} \mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i] \Big\{ \log(1-\kappa) - \frac{p}{2}\log(2\pi) - \\ &\quad - \frac{1}{2}\log(\mid \boldsymbol{\Psi}_\delta \mid) - \frac{1}{2}\left(\boldsymbol{y}_i^T \boldsymbol{\Psi}_\delta^{-1} \boldsymbol{y}_i - 2\boldsymbol{y}_i^T \boldsymbol{\Psi}_\delta^{-1} \boldsymbol{\Lambda}_2 \mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i] + trace\left(\boldsymbol{\Lambda}_2^T \boldsymbol{\Psi}_\delta^{-1} \boldsymbol{\Lambda}_2 \mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i]\right)\right) - \\ &\quad - \frac{K}{2}\log(2\pi) - \frac{1}{2}\mathbb{E}[\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i] \Big\} \end{aligned} \tag{3.24}$$

dove $\boldsymbol{\Gamma}$ indica il vettore di parametri veri del modello e $\hat{\boldsymbol{\Gamma}}$ il vettore di stime di ML dei parametri. Si presenta il calcolo dei valori attesi delle variabili latenti $\{z_i, \boldsymbol{\eta}_i, \boldsymbol{\xi}_i\}$ condizionati ai dati $\boldsymbol{y}_i$ secondo le proprietà della distribuzione Normale multivariata condizionata (Pace, Salvan et al., 2001):

$$\begin{aligned} \mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] &= \frac{\kappa P(\boldsymbol{y}_i \mid z_i = 1)}{\kappa P(\boldsymbol{y}_i \mid z_i = 1) + P(\boldsymbol{y}_i \mid z_i = 0) - \kappa P(\boldsymbol{y}_i \mid z_i = 0)} \\ &= \frac{\kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_\delta)}{\kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_\delta) + N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_\delta) - \kappa N_p(\mathbf{0}_p, \boldsymbol{\Lambda}_2 \boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_\delta)} \end{aligned}$$

$$\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i] = 1 - \mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] \tag{3.25}$$

$$\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i] = \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T (\boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}) \tag{3.26}$$

$$\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i] = \boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi} + \mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T \tag{3.27}$$

$$\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i] = \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{\alpha}) \tag{3.28}$$

$$\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i] = \boldsymbol{I}_K - \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2 + \mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T \tag{3.29}$$

M-step

Vengono calcolate le stime di ML per il vettore di parametri $\boldsymbol{\Gamma}$.
Per κ:

$$\frac{\partial\mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial\kappa} = \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\frac{1}{\kappa} + \sum_{i=1}^{n}(1 - \mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i])\frac{1}{(1-\kappa)} = 0 \tag{3.30}$$

Quindi,

$$\hat{\kappa} = \frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]}{n} \tag{3.31}$$

Per $\boldsymbol{\Lambda}_1$:

$$\frac{\partial\mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial\boldsymbol{\Lambda}_1} = \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\left\{\boldsymbol{\Theta}_\delta^{-1}\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T - \boldsymbol{\Theta}_\delta^{-1}\boldsymbol{\Lambda}\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\right\} = \boldsymbol{0} \tag{3.32}$$

Quindi,

$$\hat{\boldsymbol{\Lambda}}_1 = \frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T}{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]} \tag{3.33}$$

Analogamente per $\boldsymbol{\Lambda}_2$:

$$\hat{\boldsymbol{\Lambda}}_2 = \frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T}{\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \mid \boldsymbol{y}_i]} \tag{3.34}$$

Per $\boldsymbol{\Theta}_\delta$:

$$\begin{aligned}
\frac{\partial\mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial\boldsymbol{\Theta}_\delta^{-1}} &= \sum_{i=1}^{n}\mathbb{E}\left[z_i = 1 \mid \boldsymbol{y}_i\right]\sum_{i=1}^{n}\left[\frac{1}{2}\boldsymbol{\Theta}_\delta - \frac{1}{2}\left(\boldsymbol{y}_i\boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]y_i^T + \boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1^T\right)\right] \\
&= \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\frac{1}{2}\boldsymbol{\Theta}_\delta - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T + \frac{1}{2}\sum_{i=1}^{n}2\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T - \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1^T \\
&= \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\frac{1}{2}\boldsymbol{\Theta}_\delta - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T + \frac{1}{2}\sum_{i=1}^{n}2\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T -
\end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T\boldsymbol{\Lambda}_1^T \\
&= \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\frac{1}{2}\boldsymbol{\Theta}_\delta - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T + \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T = \mathbf{0}
\end{aligned}
\tag{3.35}
$$

Quindi,

$$
\hat{\boldsymbol{\Theta}}_\delta = diag\left\{\frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i](\boldsymbol{y}_i\boldsymbol{y}_i^T - \boldsymbol{\Lambda}_1\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T)}{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]}\right\} \tag{3.36}
$$

Analogamente per $\boldsymbol{\Psi}_\delta$:

$$
\hat{\boldsymbol{\Psi}}_\delta = diag\left\{\frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i](\boldsymbol{y}_i\boldsymbol{y}_i^T - \boldsymbol{\Lambda}_2\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]\boldsymbol{y}_i^T)}{\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]}\right\} \tag{3.37}
$$

Per $\boldsymbol{\Phi}$:

$$
\frac{\partial\mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial\boldsymbol{\Phi}^{-1}} = \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\left\{\frac{1}{2}\boldsymbol{\Phi} - \frac{1}{2}\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]\right\} = \mathbf{0} \tag{3.38}
$$

Quindi,

$$
\hat{\boldsymbol{\Phi}} = \frac{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \mid \boldsymbol{y}_i]^T}{\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]} \tag{3.39}
$$

3.2.3 Expectation-Maximization con le statistiche sufficienti

Seguendo la procedura per ottenere le statistiche sufficienti presente in Bartholomew et al. (2011), si riscrive la funzione di log-verosimiglianza:

$$
\begin{aligned}
\ell(\boldsymbol{\Gamma}) = {} & \log(\kappa)\sum_{i=1}^{n}z_i - \frac{p}{2}\log(2\pi)\sum_{i=1}^{n}z_i - \frac{1}{2}\log(\mid\boldsymbol{\Theta}_\delta\mid)\sum_{i=1}^{n}z_i - \frac{n}{2}trace\Bigg(\Theta_\delta^{-1}\Bigg(\frac{1}{n}\sum_{i=1}^{n}z_i\boldsymbol{y}_i\boldsymbol{y}_i^T - \\
& - 2\boldsymbol{\Lambda}_1\frac{1}{n}\sum_{i=1}^{n}z_i\boldsymbol{y}_i\boldsymbol{\eta}_i^T + \frac{1}{n}\sum_{i=1}^{n}z_i\boldsymbol{\Lambda_1}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\boldsymbol{\Lambda_1}^T\Bigg)\Bigg) - \frac{q}{2}\log(2\pi)\sum_{i=1}^{n}z_i - \frac{1}{2}\log(\mid\boldsymbol{\Phi}\mid)\sum_{i=1}^{n}z_i - \\
& - \frac{n}{2}trace\left(\boldsymbol{\Phi}^{-1}\frac{1}{n}\sum_{i=1}^{n}z_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T\right) + \log(1-\kappa)\sum_{i=1}^{n}(1-z_i) - \frac{p}{2}\log(2\pi)\sum_{i=1}^{n}(1-z_i) - \\
& - \frac{1}{2}\log(\mid\boldsymbol{\Psi}_\delta\mid)\sum_{i=1}^{n}(1-z_i) - \frac{n}{2}trace\Bigg[\Psi_\delta^{-1}\Bigg(\frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{y}_i\boldsymbol{y}_i^T - \\
& - 2\boldsymbol{\Lambda}_2\frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{y}_i\boldsymbol{\xi}_i^T + \frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{\Lambda_2}\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T\boldsymbol{\Lambda_2}^T\Bigg)\Bigg]
\end{aligned}
\tag{3.40}
$$

e, successivamente, si calcolano le funzioni punteggio per l'insieme di parametri ignoti $\boldsymbol{\Gamma}$.

Funzione punteggio per κ:

$$\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \kappa} = \frac{1}{\kappa}\sum_{i=1}^{n} z_i \frac{1}{(1-\kappa)}\sum_{i=1}^{n}(1-z_i) \tag{3.41}$$

Funzione punteggio per $\boldsymbol{\Lambda}_1$ e $\boldsymbol{\Lambda}_2$:

$$\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \boldsymbol{\Lambda}_1} = -\frac{n}{2} trace\left[\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\left(-2\frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{y}_i \boldsymbol{\eta}_i^T + 2\boldsymbol{\Lambda}_1 \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T\right)\right] \tag{3.42}$$

$$\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \boldsymbol{\Lambda}_2} = -\frac{n}{2} trace\left[\boldsymbol{\Theta}_{\boldsymbol{\delta}}^{-1}\left(-2\frac{1}{n}\sum_{i=1}^{n} (1-z_i) \boldsymbol{y}_i \boldsymbol{\xi}_i^T + 2\boldsymbol{\Lambda}_2 \frac{1}{n}\sum_{i=1}^{n} (1-z_i) \boldsymbol{\xi}_i \boldsymbol{\xi}_i^T\right)\right] \tag{3.43}$$

Funzione punteggio per $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ e $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$:

$$\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \boldsymbol{\Theta}_{\boldsymbol{\delta}}} = \frac{1}{2}\boldsymbol{\Theta}_{\boldsymbol{\delta}} \sum_{i=1}^{n} z_i - \frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{y}_i \boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_1 \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{y}_i \boldsymbol{\eta}_i^T + \boldsymbol{\Lambda}_1 \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T \boldsymbol{\Lambda}_1^T\right) \tag{3.44}$$

$$\begin{aligned}\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \boldsymbol{\Psi}_{\boldsymbol{\delta}}} = \frac{1}{2}\boldsymbol{\Psi}_{\boldsymbol{\delta}} \sum_{i=1}^{n}(1-z_i) - \frac{n}{2}\Bigg(&\frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{y}_i \boldsymbol{y}_i^T - 2\boldsymbol{\Lambda}_2 \frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{y}_i \boldsymbol{\xi}_i^T + \\ &+ \boldsymbol{\Lambda}_2 \frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{\xi}_i \boldsymbol{\xi}_i^T \boldsymbol{\Lambda}_2^T\Bigg)\end{aligned} \tag{3.45}$$

Funzione punteggio per $\boldsymbol{\Phi}$:

$$\frac{\partial \mathbb{Q}(\boldsymbol{\Gamma} \mid \hat{\boldsymbol{\Gamma}})}{\partial \boldsymbol{\Phi}} = \frac{1}{2}\boldsymbol{\Phi} \sum_{i=1}^{n} z_i - \frac{n}{2}\left(\frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{\eta}_i \boldsymbol{\eta}_i^T\right) \tag{3.46}$$

Le statistiche sufficienti, individuabili dalle funzioni punteggio dei parametri, sono le seguenti:

$$S_{\boldsymbol{z}=1} = \sum_{i=1}^{n} z_i \tag{3.47}$$

$$S_{\boldsymbol{z}=0} = \sum_{i=1}^{n}(1-z_i) \tag{3.48}$$

$$S_{\boldsymbol{y}\boldsymbol{y}^T} = \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{y}_i \boldsymbol{y}_i^T \tag{3.49}$$

$$S_{\boldsymbol{y}\boldsymbol{\eta}^T} = \frac{1}{n}\sum_{i=1}^{n} z_i \boldsymbol{y}_i \boldsymbol{\eta}_i^T \tag{3.50}$$

$$S_{\boldsymbol{y}\boldsymbol{\xi}^T} = \frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{y}_i\boldsymbol{\xi}_i^T \tag{3.51}$$

$$S_{\boldsymbol{\eta}\boldsymbol{\eta}^T} = \frac{1}{n}\sum_{i=1}^{n} z_i\boldsymbol{\eta}_i\boldsymbol{\eta}_i^T \tag{3.52}$$

$$S_{\boldsymbol{\xi}\boldsymbol{\xi}^T} = \frac{1}{n}\sum_{i=1}^{n}(1-z_i)\boldsymbol{\xi}_i\boldsymbol{\xi}_i^T \tag{3.53}$$

Si calcolano, quindi, i valori attesi delle statistiche sufficienti[5]:

$$\mathbb{E}[S_{\boldsymbol{z}=1} \mid \boldsymbol{y}_i] = \sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] \tag{3.54}$$

$$\mathbb{E}[S_{\boldsymbol{z}=0} \mid \boldsymbol{y}_i] = \sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i] \tag{3.55}$$

$$\mathbb{E}[S^{(\boldsymbol{z}=1)}_{\boldsymbol{y}\boldsymbol{y}^T} \mid \boldsymbol{y}_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T \tag{3.56}$$

$$\mathbb{E}[S^{(\boldsymbol{z}=1)}_{\boldsymbol{y}\boldsymbol{y}^T} \mid \boldsymbol{y}_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T \tag{3.57}$$

$$\begin{aligned}
\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\eta}^T} \mid \boldsymbol{y}_i] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i(\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}))^T \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_{\mathbf{1}}\boldsymbol{\Phi} \\
&= \mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T 1} \mid \boldsymbol{y}_i](\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_{\mathbf{1}}\boldsymbol{\Phi}
\end{aligned} \tag{3.58}$$

$$\begin{aligned}
\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\xi}^T} \mid \boldsymbol{y}_i] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]\boldsymbol{y}_i(\boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}(\boldsymbol{y}_i - \boldsymbol{\alpha}))^T \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T 0} \mid \boldsymbol{y}_i](\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2
\end{aligned} \tag{3.59}$$

$$\begin{aligned}
\mathbb{E}[S_{\boldsymbol{\eta}\boldsymbol{\eta}^T} \mid \boldsymbol{y}_i] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\left(\boldsymbol{\Phi} - \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi} + \mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\eta}_i \mid \boldsymbol{y}_i]^T\right) \\
&= \boldsymbol{\Phi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] - \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] + \\
&\quad + \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]\boldsymbol{y}_i\boldsymbol{y}_i^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}
\end{aligned}$$

[5]Si ricorda che il valore atteso $\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]$ è stato calcolato nell'Equazione 3.25.

$$= \boldsymbol{\Phi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i] - \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 1 \mid \boldsymbol{y}_i]$$
$$+ \boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T(\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T1} \mid \boldsymbol{y}_i](\boldsymbol{\Lambda}_1\boldsymbol{\Phi}\boldsymbol{\Lambda}_1^T + \boldsymbol{\Theta}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_1\boldsymbol{\Phi} \tag{3.60}$$

$$\mathbb{E}[S_{\boldsymbol{\xi}\boldsymbol{\xi}^T} \mid \boldsymbol{y}_i] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]\left(\boldsymbol{I}_K - \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2 + \mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]\mathbb{E}[\boldsymbol{\xi}_i \mid \boldsymbol{y}_i]^T\right)$$
$$= \boldsymbol{I}_K\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i] - \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[z_i = 0 \mid \boldsymbol{y}_i]$$
$$+ \boldsymbol{\Lambda}_2^T(\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T0} \mid \boldsymbol{y}_i](\boldsymbol{\Lambda}_2\boldsymbol{\Lambda}_2^T + \boldsymbol{\Psi}_{\boldsymbol{\delta}})^{-1}\boldsymbol{\Lambda}_2 \tag{3.61}$$

M-step

A questo punto, si calcolano le stime ML basate sulle statistiche sufficienti. Per κ:

$$\hat{\kappa} = \frac{\mathbb{E}[S_{\boldsymbol{z}=1} \mid \boldsymbol{y}_i]}{n} \tag{3.62}$$

Per $\boldsymbol{\Lambda}_1$ e $\boldsymbol{\Lambda}_2$:

$$\hat{\boldsymbol{\Lambda}}_1 = \frac{\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\eta}^T} \mid \boldsymbol{y}_i]}{\mathbb{E}[S_{\boldsymbol{\eta}\boldsymbol{\eta}^T} \mid \boldsymbol{y}_i]} \tag{3.63}$$

$$\hat{\boldsymbol{\Lambda}}_2 = \frac{\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\xi}^T} \mid \boldsymbol{y}_i]}{\mathbb{E}[S_{\boldsymbol{\xi}\boldsymbol{\xi}^T} \mid \boldsymbol{y}_i]} \tag{3.64}$$

$$\tag{3.65}$$

Per $\boldsymbol{\Theta}_{\boldsymbol{\delta}}$ e $\boldsymbol{\Psi}_{\boldsymbol{\delta}}$:

$$\hat{\boldsymbol{\Theta}}_\delta = \frac{\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T}^{(\boldsymbol{z}=1)} \mid \boldsymbol{y}_i] - \boldsymbol{\Lambda}_1\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\eta}^T} \mid \boldsymbol{y}_i]}{\mathbb{E}[S_{\boldsymbol{z}=1} \mid \boldsymbol{y}_i]/n} \tag{3.66}$$

$$\hat{\boldsymbol{\Psi}}_\delta = \frac{\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{y}^T}^{(\boldsymbol{z}=0)} \mid \boldsymbol{y}_i] - \boldsymbol{\Lambda}_2\mathbb{E}[S_{\boldsymbol{y}\boldsymbol{\xi}^T} \mid \boldsymbol{y}_i]}{\mathbb{E}[S_{\boldsymbol{z}=0} \mid \boldsymbol{y}_i]/n} \tag{3.67}$$

Per $\boldsymbol{\Phi}$:

$$\hat{\boldsymbol{\Phi}} = \frac{\mathbb{E}[S_{\boldsymbol{\eta}\boldsymbol{\eta}^T}]}{\mathbb{E}[S_{\boldsymbol{z}=1} \mid \boldsymbol{y}_i]/n} \tag{3.68}$$

Capitolo 4

Studio di simulazione ed applicazione su dati reali

4.1 Studio di simulazione

Il presente studio di simulazione ha tre obiettivi: (1) testare la capacità del modello Mix-CFA-EFA di ottenere stime vicine ai valori veri dei parametri, in contesti con differenti proporzioni di osservazioni generate dalla componente CFA ed EFA; (2) verificare la capacità di un insieme di indici di fit, nel selezionare il modello vero a confronto con un modello specificato in modo errato; (3) valutare la capacità del modello di classificare correttamente le osservazioni generate dal modello CFA e dal modello EFA.

Il presente studio di simulazione è stato condotto su una macchina ThinkPad-L15 con un processore Intel Core i5-10210U CPU 1.60 GHz e 8x2 GB di RAM. L'implementazione dello studio è stata effettuata tramite il software open-source `Julia` versione 1.8.0 (Bezanson et al., 2017), il codice completo viene presentato nell'Appendice A.2. Si anticipa che non è stato possibile portare a termine integralmente lo studio di simulazione, a causa di costi computazionali eccessivi per la macchina a disposizione. Pertanto, si riportano i risultati solo di una parte del disegno sperimentale.

4.1.1 Disegno dello studio di simulazione

Il disegno dello studio prevede tre fattori: (a) K-false $= \{2, 4\}$ rappresentano i numeri di fattori stimati dal modello sulla componente EFA, che può non coincidere con il numero di fattori veri K dei dati generati; (b) $K = \{2, 4\}$ sono i numeri di fattori veri dei

dati generati; (c) $\kappa = \{0.80, 0.60, 0.20\}$ sono le proporzioni di componente CFA presente nei dati. I fattori sono stati combinati in modo da avere un disegno fattoriale completo di $2 \times 2 \times 3 = 12$ scenari, come si vede in Tabella 4.1. Per ogni combinazione,sono stati generati $B = 500$ campioni. I restanti valori per la generazione dei dati sono stati mantenuti fissi: la numerosità campionaria prevede $n = 1000$, le variabili osservate (o items) corrispondono a $p = 20$, il numero di fattori della componente CFA è $q = 1$.

Design N.	K-false	K	κ
1	2	2	0.80
2	4	2	0.80
3	2	4	0.80
4	4	4	0.80
5	2	2	0.60
6	4	2	0.60
7	2	4	0.60
8	4	4	0.60
9	2	2	0.20
10	4	2	0.20
11	2	4	0.20
12	4	4	0.20

Tabella 4.1: Il design dello studio di simulazione.

4.1.2 Generazione dei dati

I parametri veri del modello Mix-CFA-EFA per ciascun disegno sono stati generati come segue:

CFA	EFA
$\boldsymbol{\Lambda}_{1j} \sim U(0,1), \quad j \in \{1, \ldots, p\}$	$\boldsymbol{\Lambda}_{2jk} \sim U(0,1), \quad k \in 1, \ldots, K$
$\boldsymbol{\Phi} = \boldsymbol{I}_{q=1}$	$\boldsymbol{\Psi_\delta} = 0.85\boldsymbol{I}_p$
$\boldsymbol{\Theta_\delta} = \mathbf{1} - diag(\boldsymbol{\Lambda}_1 \boldsymbol{\Phi} \boldsymbol{\Lambda}_1^T)$	$\boldsymbol{\Omega} = \boldsymbol{I}_K$

dove: $U(0,1)$ indica una distribuzione Uniforme tra 0 ed 1; I_K indica una matrice di identità di ordine $K \times K$; $\boldsymbol{\Psi_\delta}$ corrisponde ad una matrice scalare, ovvero una matrice diagonale con singolo valore numerico sulla diagonale diverso da 1 (0.85); invece, $\boldsymbol{\Theta_\delta}$ è ottenuta fissando la covarianza delle variabili osservate ad 1 (si veda l'Equazione 2.15).

I dati $\boldsymbol{Y}_{ib}^{*} \in \mathbb{R}^{p}$ per ogni $i \in \{1, \ldots, n\}$ ed ogni $b \in \{1, \ldots, B\}$ sono stati generati come segue:

$$\begin{array}{cc} \text{CFA} & \text{EFA} \\ \boldsymbol{\eta}_{tb} \sim N_q(\mathbf{0}_q, \boldsymbol{\Phi}) & \boldsymbol{\xi}_{hb} \sim N_K(\mathbf{0}_K, \boldsymbol{\Omega}) \\ \boldsymbol{\delta}_{tb} \sim N_p(\mathbf{0}_p, \boldsymbol{\Theta}_{\boldsymbol{\delta}}) & \boldsymbol{\varepsilon}_{hb} \sim N_p(\mathbf{0}_p, \boldsymbol{\Psi}_{\boldsymbol{\delta}}) \\ \boldsymbol{y}_{tb}^{\text{CFA}} = \boldsymbol{\Lambda}_1 \boldsymbol{\eta}_{tb} + \boldsymbol{\delta}_{tb} & \boldsymbol{y}_{hb}^{\text{EFA}} = \boldsymbol{\Lambda}_2 \boldsymbol{\xi}_{hb} + \boldsymbol{\varepsilon}_{hb} \end{array}$$

dove gli indici consistono in $t \in \{1, \ldots, pop\}$ e $h \in \{1, \ldots, pop\}$ con $pop = 50000$. Successivamente, è stato generato il vettore di indicatori latenti per ogni combinazione:

$$z_{ib} \sim \mathcal{B}ern(\kappa) \tag{4.1}$$

Infine, sono stati creati B campioni di mistura per ogni disegno:

$$\boldsymbol{y}_{ib}^{*} \mid z_i = 1 = \boldsymbol{y}_{tb}^{\text{CFA}}$$
$$\boldsymbol{y}_{ib}^{*} \mid z_i = 0 = \boldsymbol{y}_{hb}^{\text{EFA}}$$

4.1.3 Misure di performance

Misure di errore nelle stime

Per ogni scenario del disegno sperimentale è stata valutata la capacità del modello di produrre stime vicine al valore vero dei parametri simulati. Le misure di performance utilizzate sono: la misura RMSE ed una misura di Proportion of Agreeement (PA; Timmerman e Kiers, 2002) modificata:

1. La misura RMSE (Root Mean Squared Error; Morris et al., 2019) è definita, rispettivamente per parametri scalari e marticiali, come:

$$RMSE(\hat{\theta}) = \sqrt{\mathbb{E}(\hat{\theta} - \theta)^2} \tag{4.2}$$

$$RMSE(\hat{\theta}) = \sqrt{\mathbb{E}\left((\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right)} \tag{4.3}$$

   dove, $RMSE(\hat{\theta}) \approx 0$ indica che l'errore di stima è pressoché nullo. Inoltre, la misura RMSE è particolarmente sensibile ad eventuali bias nelle stime (Willmott & Matsuura, 2006).

2. La PA-mod è calcolata come:

$$PA\text{-}mod = \frac{\|\hat{\theta} - \theta\|^2}{\|\theta\|^2} \tag{4.4}$$

   dove, PA-mod≈ 0 indica eccellente accuratezza nella stima dei parametri.

Indici di fit

Gli indici di fit, valutati nella loro abilità di selezionare il modello correttamente specificato, sono i seguenti: l'Akaike Information Criterion (AIC; Akaike, 1974), l'Akaike Information Criterion corrected (AICc;Hurvich e Tsai, 1989), il consistent Akaike Information Criterion (CAIC; Bozdogan, 1987), il Bayesian Information Criterion (BIC; Schwarz, 1978), il sample-adjusted BIC (ssBIC; Sclove, 1987), il Classification Likelihood Information Criterion (CLC; McLachlan e Peel, 2000) e l'Integrated Classification Likelihood (ICL-BIC; McLachlan e Peel, 2000).

Si riportano le formule di calcolo:

$$AIC = 2d - 2\ell \tag{4.5}$$

$$AICc = 2d - 2\ell + 2\frac{d(d+1)}{n-d-1} \tag{4.6}$$

$$CAIC = 2d\left(\log(n) + 1\right) \tag{4.7}$$

$$BIC = 2d\log(n) - 2\ell \tag{4.8}$$

$$ssBIC = \log\left(\frac{n+2}{24}\right)d - 2\ell \tag{4.9}$$

$$CLC = 2E(k) - 2\ell \tag{4.10}$$

$$ICL\text{-}BIC = 2E(k) + \log(n)d - 2\ell \tag{4.11}$$

dove d è il numero di parametri da stimare ed ℓ il valore di log-verosimiglianza, mentre la statistica Ek è stata calcolata come segue:

$$E(k) = -\sum_{i=1}^{n} P(z_i = 1 \mid \boldsymbol{y}_i)\log\left(P(z_i = 1 \mid \boldsymbol{y}_i)\right) + \sum_{i=1}^{n} P(z_i = 0 \mid \boldsymbol{y}_i)\log\left(P(z_i = 0 \mid \boldsymbol{y}_i)\right) \tag{4.12}$$

dove $P(z_i = 1 \mid \boldsymbol{y}_i)$ è definito nell'Equazione 3.22.

Misure di classificazione

Le seguenti misure di classificazione binaria sono state utilizzate per valutare la capacità di ricostruire correttamente la componente di origine (CFA o EFA) di ciascuna osservazione: l'accuratezza (ACC; Metz, 1978) e F_1-score (Powers, 2007):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.13}$$

$$F_1\text{-score} = \frac{2TP}{2TP + FP + FN} \tag{4.14}$$

dove TP indica i veri positivi, TN i veri negativi, FP i falsi positivi e FN i falsi negativi. L'accuratezza è una misura molto semplice ed intuitiva, tuttavia non è una misura ottimale in presenza di classi sbilanciate ($\kappa \neq 0.50$) (Prati et al., 2009). F_1-score, invece, consiste nella media armonica della precisione e del recupero e consente di valutare più adeguatamente la presenza di classi sbilanciate (Powers, 2007).

4.1.4 Risultati

Per ciascuna combinazione del disegno sperimentale sono state escluse le replicazioni b che hanno portato a non convergenza ed è stata calcolata una media tra le restanti replicazioni. Nella Tabella 4.2 sono presentati i risultati parziali dello studio di simulazione, in cui sono presentati i due disegni portati completamente a compimento.

Nella condizione con $n = 1000$, $\kappa = 0.80$ e $K = K\text{-false} = 2$, il modello Mix-CFA-EFA dimostra di possedere eccellenti capacità nella stima dei parametri veri per quanto riguarda tutti i parametri, eccetto $\hat{\mathbf{\Lambda}}_2$, che, al contrario, non viene stimato in maniera soddisfacente. Ciò può essere dovuto alla bassa proporzione di osservazioni generate dalla componente EFA (0.2). Ad ogni modo, i coefficienti fattoriali della matrice $\mathbf{\Lambda}_2$ non sono di per sé interpretabili né interessanti, in quanto, nell'applicazione del modello a contesti applicativi, si utilizza la componente EFA con l'obiettivo di raccogliere il rumore presente nei dati. Invece, nella condizione con $K = 4$ e $K\text{-false} = 2$, le stime $\{\hat{\mathbf{\Lambda}}_1, \hat{\mathbf{\Psi}}_{\boldsymbol{\delta}}\}$ peggiorano lievemente, mentre le restanti mostrano un leggero miglioramento.

Riguardo al confronto tra modelli, gli indici sembrano non individuare con successo il modello correttamente specificato.

Come atteso, le capacità di classificazione del modello Mix-CFA-EFA vengono valutate piuttosto differentemente dall'indice ACC e F_1-score. In particolare, l'ACC indica una performance del modello nel range medio basso, mentre l'indice F_1-score rileva una prestazione nettamente migliore, ma non ottimale. Anche in questo caso, entrambi gli indici di performance nella condizione con $K = 4$ e $K\text{-false} = 2$ presentano un lieve miglioramento.

K-false	K	κ	$\hat{\mathbf{\Lambda}}_1$		$\hat{\mathbf{\Theta}}_\delta$		$\hat{\mathbf{\Phi}}$		$\hat{\mathbf{\Lambda}}_2$		$\hat{\mathbf{\Psi}}_\delta$		$\hat{\kappa}$		ℓ	AIC	AICc	CAIC	BIC	ssBIC	CLC	ICL-BIC	ACC	F_1-score
			RMSE	PA-mod	RMSE	PA-mod	RMSE	PA-mod	RMSE	PA-mod	RMSE	PA-mod	RMSE	PA-mod										
		0.80	0.028	0.035	0.017	0.046	0	0	0.157	0.412	0.028	0.031	0.06	0.056	-25195.6	50593.2	51786.5	50768.1	50616.1	51988.5	50376.9	51074.6	0.654	0.757
	2	0.60																						
		0.20																						
2		0.80	0.0326	0.055	0.013	0.029	0	0	-	-	0.034	0.037	0.039	0.03	-24711.2	49624.5	50817.8	49799.4	49647.4	51019.8	49418.0	50115.7	0.673	0.782
	4	0.60																						
		0.20																						
		0.80																						
	2	0.60																						
		0.20																						
4		0.80																						
	4	0.60																						
		0.20																						

Tabella 4.2: Risultati dello studio di simulazione: (1) valutazione dell'accuratezza nelle stime tramite misura RMSE e PA-mod, si noti che $\hat{\mathbf{\Lambda}}_1$ non può avere misure di accuratezza qualora $K \neq K$-false; (2) confronto tra indici di fit nella selezione del modello corretto, in cui il modello selezionato presenta il valore più basso; (3) verifica della capacità di corretta classificazione del modello Mix-CFA-EFA tramite ACC e F_1-score.

4.2 Applicazione su dati reali

Nella presente Sezione si propone un'applicazione su dati reali, al fine di provvedere un'esemplificazione dell'uso e dell'utilità del modello Mix-CFA-EFA nell'analisi dei dati nella ricerca in psicologia. Il dataset a disposizione è composto da risposte provenienti da due gruppi di soggetti, uno dei quali potrebbe aver attuato comportamenti di faking nel processo di risposta (Anglim et al., 2017).

Gli obiettivi dell'applicazione sono due: (1) identificare la proporzione di campione, la cui eterogeneità è dovuta a bias indotto da faking good e (2) individuare correttamente i soggetti all'origine del bias, verificandone l'appartenenza a uno dei due gruppi. Pertanto, tramite il modello Mix-CFA-EFA si verifica la presenza di un effetto del faking a livello della matrice di covarianza tra le risposte agli items. Sulla base della letteratura precedente (e.g., Birkeland et al., 2006; Ziegler et al., 2015) e delle caratteristiche del modello Mix-CFA-EFA, si ipotizza che il modello identifichi una proporzione di eterogeneità pari a $\kappa = 0.50$. Inoltre, si ipotizza che i soggetti individuati dal modello CFA e dal modello EFA vengano in linea di massima classificati correttamente in uno dei due gruppi.

Le analisi statistiche e l'implementazione del modello Mix-CFA-EFA sono state effettuate tramite il software open-source `Julia` versione 1.8.0 (Bezanson et al., 2017).

4.3 Dataset

Il dataset utilizzato è stato presentato da Anglim et al. (2017)[1]. I dati sono stati raccolti da una società di consulenza in ambito risorse umane con sede in Australia. Il test psicologico utilizzato è HEXACO-Personality Inventory-Revised (HEXACO-PI-R; Ashton et al., 2014). Il test HEXACO-PI-R prevede una versione completa di 200 items, utilizzata nello studio, che può essere ridotta ad una versione di 60 items (HEXACO-PI-R-60). La versione ridotta HEXACO-PI-R-60, rispetto alla versione completa, non consente di indagare le 25 scale facets, ma solo di disporre delle scale dominio: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C) e Openness (O). Ogni scala dominio è composta da 10 items su scala Likert a 5 punti (1 = fortemente in disaccordo, 2 = in disaccordo, 3 = neutrale (né d'accordo né in disaccordo), 4 = d'accordo, 5 = fortemente d'accordo).

[1]Il dataset, con i relativi scripts in `R` e materiali aggiuntivi, è accessibile liberamente al link https://osf.io/9e3a9/.

I partecipanti provengono da due differenti condizioni di raccolta dati. In una condizione, il test HEXACO-PI-R è stato somministrato a candidati all'interno di un processo di selezione per un posto di lavoro (`applicants`), mentre, nell'altra condizione, la somministrazione è avvenuta su soggetti che erano stati invitati a rispondere al test tramite posta elettronica (`non-applicants`; Anglim et al., 2017). I due campioni raccolti sono stati manipolati in modo da ottenere un campione finale più comparabile per distribuzione di età e genere, ottenendo un sotto-campione di `applicants` con $n = 1613$ (46% maschi; età: $Mean = 42.06$ anni, $SD = 10.54$) e un sotto-campione di `non-applicants` con $n = 1613$ (46% maschi; età: $Mean = 42.38$ anni, $SD = 10.56$). Quindi, il dataset finale è composto da $n = 3226$ soggetti (46% maschi; età: $Mean = 42$ anni, $SD = 11$) e $J = 60$ items, ovvero $\boldsymbol{Y}_{60\times3226}$.

Dalla letteratura in psicologia del lavoro, è noto che durante il processo di assunzione i candidati ad un posto lavorativo tendano a distorcere le loro risposte (Birkeland et al., 2006). Inoltre, Anglim et al. (2017) hanno mostrato tramite indici basati sulle medie che si può supporre la presenza di faking nella condizione `applicants`. Tuttavia, Anglim et al. (2017) non stati sviluppati indici in grado di rilevare l'impatto del faking sulla matrice di covarianza osservata tra le rispose agli items, principale oggetto di interesse dei modelli FA. Pertanto, in Figura 4.1, in particolare in 4.1a e 4.1b, si presentano, rispettivamente, le due matrici di covarianza osservata nel gruppo `non-applicants` e `applicants`. Come si nota, le strutture di correlazioni non sono particolarmente differenti, se non per un minimo grado di correlazione negativa tra i primi 10 items ed i successivi 20. Ci si attende che due strutture di covarianza osservata così poco distinguibili tra loro potrebbero rendere strutturalmente difficile discriminare i due gruppi tramite il modello Mix-CFA-EFA.

4.4 Analisi dei dati e risultati

In primo luogo, sono stati definiti 7 modelli Mix-CFA-EFA da adattare ai dati. Tutti i modelli prevedono la presenza di 6 fattori per il modello CFA ($q = 6$), corrispondenti a quelli standard dell'inventario HEXACO-PI-R. Per la componente EFA, invece sono stati sviluppati modelli da $K = 1$ a $K = 7$, in modo da raccogliere le differenze indotte dal faking, che potrebbe implicare una differente struttura latente nella risposta al test. Per migliorare la convergenza del modello Mix-CFA-EFA, sono stati utilizzati dei valori iniziali razionali per la componente CFA, ovvero le stime ottenute precedentemente

dall'applicazione di un singolo modello CFA[2] sono state poste come valori iniziali per l'algoritmo EM del modello Mix-CFA-EFA.

Di seguito si elencano gli indici di fit utilizzati per la selezione del modello[3]: l'Akaike Information Criterion (AIC; Akaike, 1974), l'Akaike Information Criterion corrected (AICc;)Hurvich e Tsai, 1989, il consistent Akaike Information Criterion (CAIC; Bozdogan, 1987), il Bayesian Information Criterion (BIC; Schwarz, 1978), il sample-adjusted BIC (ssBIC; Sclove, 1987), il Classification Likelihood Information Criterion (CLC; McLachlan e Peel, 2000) e l'Integrated Classification Likelihood (ICL-BIC; McLachlan e Peel, 2000). In Tabella 4.3 sono riportati gli indici di adattamento del modello, per individuare il corretto numero di fattori del modello EFA. Complessivamente, gli indici selezionano il modello con $K = 4$.

Modello	ℓ	AIC	AICc	CAIC	BIC	ssBIC	CLC	ICL-BIC
$K = 1$	−265832.571	532687.141	532879.943	540943.878	539921.878	534169.841	531640.675	535769.043
$K = 2$	−265213.254	531568.509	531814.637	**540794.725**	**539652.725**	533225.302	530395.316	535008.424
$K = 3$	−264832.505	530927.009	531234.482	541122.705	539860.705	532757.896	529630.792	534728.64
$K = 4$	−264420.672	**530223.344**	**530600.749**	541388.52	540006.52	**532228.325**	**528800.713**	**534383.301**
$K = 5$	−264481.072	530464.144	530920.693	542598.799	541096.799	532643.218	528917.649	534984.976
$K = 6$	−265345.452	532312.903	532858.497	545417.038	543795.038	534666.072	530645.386	537197.453
$K = 7$	−266513.572	534769.144	535414.439	548842.759	547100.759	537296.406	532982.378	540019.185

Tabella 4.3: Modelli Mix-CFA-EFA con $q = 6$ e numero di K che varia. Secondo tutti gli indici di fit presentati si preferisce il modello con il valore minore, evidenziato per ogni indice in grassetto. Il modello con maggior accordo tra indici di selezione è con $K = 4$.

Nella Tabella 4.4 sono mostrate le stime dei parametri $\{\hat{\mathbf{\Lambda}}_1, \hat{\mathbf{\Theta}}_{\boldsymbol{\delta}}, \hat{\mathbf{\Lambda}}_2, \hat{\mathbf{\Psi}}_{\boldsymbol{\delta}}, \hat{\kappa}\}$ del modello con $K = 4$, dove ϕ indicano i fattori del modello CFA e ω i fattori del modello EFA. Si nota che $\hat{\kappa} = 0.567$ corrisponde alle attese, essendo di poco superiore a 0.50. Il parametro di mistura, pertanto, identifica metà del campione che risponde al test secondo il modello CFA a 6 fattori e l'altra metà secondo il modello EFA a 4 fattori. In grassetto, sono presentati i $\lambda \geq 0.3$, mostrando che entrambe le matrici $\hat{\mathbf{\Lambda}}_1$ e $\hat{\mathbf{\Lambda}}_2$ hanno associazioni item-fattore prevalentemente moderate. Inoltre, si puntualizza che i coefficienti fattoriali non sono stati standardizzati. La matrice degli errori $\hat{\mathbf{\Theta}}_{\boldsymbol{\delta}}$ mostra valori inferiori rispetto a quelli della matrice $\hat{\mathbf{\Psi}}_{\boldsymbol{\delta}}$, indicando una stima più precisa nel

[2]Il modello è stato implementato basandosi sul caso 3 descritto da Rubin e Thayer (1982) ed il codice è presente nell'Appendice A.3.3.

[3]Le corrispondenti equazioni sono nella Sottosezione 4.1.3.

caso del modello CFA. Infine, la matrice $\hat{\mathbf{\Phi}}$ mostra correlazioni tra fattori da basse a moderate, con un maggior associazione tra Emotionality e Conscientiousness ($\phi_{5,2}$).

Per valutare la capacità di classificazione del modello Mix-CFA-EFA in questo contesto empirico sono state calcolate le seguenti misure: le proporzioni di classificazioni corrette ($CC = 0.32$), o accuratezza, le proporzioni di classificazioni errate ($MC = 0.68$), o inaccuratezza, e l'indice F_1-score (F_1-score$= 0.36$). Gli indici segnalano un limite del modello Mix-CFA-EFA nell'individuare correttamente quale soggetto appartiene alla condizione `applicant` o `non-applicant`. Tuttavia, è verosimile che lo stesso campione in esame non sia ottimale per valutare le capacità del modello, in quanto, come visto in Figura 4.1a e in Figura 4.1b, non ci sono differenze particolarmente marcate a livello di matrice di covarianza in entrambi i gruppi. Inoltre, nelle Figure 4.1c e 4.1d sono riportate le matrici di correlazioni osservate per i sotto-campioni predetti dal modello Mix-CFA-EFA. Si può notare come il sotto-campione EFA sembri raccogliere la struttura di correlazione degli items del gruppo `non-applicants` anziché del gruppo `applicants`.

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	$\hat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$	ω_1	ω_2	ω_3	ω_4	$\hat{\boldsymbol{\Psi}}_{\boldsymbol{\delta}}$
Item 62	**0.392**	0	0	0	0	0	0.657	0.237	**−0.404**	0.012	**−0.344**	0.918
Item 2	**0.317**	0	0	0	0	0	0.724	0.09	−0.178	−0.07	**−0.394**	1.022
Item 8	**0.4**	0	0	0	0	0	0.668	−0.116	−0.239	−0.089	**−0.473**	0.909
Item 20	**0.432**	0	0	0	0	0	0.627	0.229	**−0.3**	0.146	**−0.329**	0.984
Item 68	**0.42**	0	0	0	0	0	0.662	0.273	−0.191	0.213	−0.191	1.028
Item 80	**0.373**	0	0	0	0	0	0.613	−0.106	−0.068	−0.012	**−0.439**	1.099
Item 134	**0.415**	0	0	0	0	0	0.652	0.287	−0.193	0.115	**−0.398**	0.933
Item 146	**0.337**	0	0	0	0	0	0.729	0.199	−0.074	−0.088	**−0.353**	1.027
Item 164	**0.443**	0	0	0	0	0	0.619	**0.471**	−0.113	0.26	−0.173	0.932
Item 170	**0.437**	0	0	0	0	0	0.654	0.228	−0.175	−0.085	**−0.447**	0.903
Item 69	0	0.296	0	0	0	0	0.525	**0.506**	0.142	0.11	**0.665**	0.737
Item 3	0	0.217	0	0	0	0	0.487	0.262	0.147	0.022	**0.547**	1.202
Item 33	0	**0.396**	0	0	0	0	0.341	0.027	0.032	0.246	**0.676**	1.148
Item 57	0	**0.549**	0	0	0	0	0.464	0.296	**−0.332**	0.14	**0.529**	0.842
Item 81	0	**0.521**	0	0	0	0	0.543	−0.067	−0.234	0.166	**0.615**	0.825
Item 99	0	**0.355**	0	0	0	0	0.385	**0.397**	0.132	0.185	**0.614**	1.03
Item 111	0	**0.465**	0	0	0	0	0.573	**0.435**	**−0.336**	0.185	0.298	0.903
Item 135	0	**0.363**	0	0	0	0	0.67	**0.423**	−0.278	0.092	**0.376**	0.863
Item 165	0	**0.414**	0	0	0	0	0.362	**0.339**	−0.001	0.237	**0.71**	0.925
Item 171	0	**0.321**	0	0	0	0	0.399	**0.348**	**0.322**	0	**0.719**	0.884
Item 52	0	0	**0.441**	0	0	0	0.545	**0.45**	0.194	0.198	0.281	0.969
Item 4	0	0	**0.47**	0	0	0	0.52	**0.391**	0.263	0.184	**0.312**	0.967
Item 16	0	0	**0.442**	0	0	0	0.638	**0.406**	0.267	0.15	0.095	0.953
Item 46	0	0	**0.496**	0	0	0	0.395	**0.429**	**0.471**	0.125	0.292	0.956
Item 82	0	0	**0.583**	0	0	0	0.402	**0.561**	**0.31**	0.205	0.049	0.903
Item 94	0	0	**0.362**	0	0	0	0.643	0.232	0.182	0.024	0.236	1.158
Item 112	0	0	0.215	0	0	0	0.808	0.239	−0.052	0.068	−0.012	1.134
Item 130	0	0	**0.395**	0	0	0	0.665	**0.315**	0.087	0.229	0.04	1.086
Item 154	0	0	**0.35**	0	0	0	0.802	0.227	0.022	0.076	−0.127	1.058
Item 160	0	0	**0.371**	0	0	0	0.548	**0.366**	**0.399**	0.131	0.204	1.028
Item 138	0	0	0	**0.548**	0	0	0.492	−0.144	−0.094	**0.725**	0.016	0.729
Item 12	0	0	0	**0.462**	0	0	0.563	−0.108	0.007	**0.454**	0.26	0.999
Item 6	0	0	0	**0.574**	0	0	0.423	−0.074	0.109	**0.682**	0.087	0.835
Item 48	0	0	0	0.209	0	0	0.575	−0.064	0.094	**0.487**	**0.327**	1.134
Item 78	0	0	0	**0.549**	0	0	0.507	−0.12	0.032	**0.555**	0.066	0.928
Item 120	0	0	0	0.256	0	0	0.615	−0.167	−0.091	**0.549**	0.212	1.027
Item 162	0	0	0	**0.333**	0	0	0.668	−0.106	−0.242	**0.416**	0.265	0.961
Item 180	0	0	0	**0.49**	0	0	0.575	−0.141	0.205	**0.414**	0.225	0.963
Item 186	0	0	0	**0.423**	0	0	0.629	−0.025	−0.128	**0.522**	0.153	0.929
Item 192	0	0	0	**0.557**	0	0	0.411	−0.159	0.051	**0.686**	0.17	0.849
Item 53	0	0	0	0	**0.563**	0	0.315	0.117	**0.334**	−0.091	0.202	1.318
Item 35	0	0	0	0	**0.348**	0	0.576	−0.005	0.151	0.035	**0.421**	1.178
Item 41	0	0	0	0	**0.562**	0	0.332	0.032	**0.352**	0.02	0.169	1.33
Item 47	0	0	0	0	**0.39**	0	0.458	0.035	**0.619**	0.01	**0.383**	0.974
Item 65	0	0	0	0	**0.333**	0	0.72	−0.025	0.127	−0.094	−0.048	1.209
Item 71	0	0	0	0	**0.433**	0	0.416	0.091	**0.581**	0.045	**0.438**	0.962
Item 89	0	0	0	0	**0.479**	0	0.65	**−0.267**	0.048	−0.102	0.082	1.135
Item 131	0	0	0	0	0.296	0	0.479	0.136	**0.416**	0.14	**0.428**	1.131
Item 173	0	0	0	0	**0.461**	0	0.356	0.088	**0.418**	−0.047	**0.371**	1.227
Item 191	0	0	0	0	**0.332**	0	0.484	0.12	**0.525**	−0.034	0.248	1.173
Item 73	0	0	0	0	0	**0.441**	0.56	0.156	**0.457**	0.171	−0.02	1.049
Item 37	0	0	0	0	0	**0.446**	0.601	0.238	0.265	0.251	−0.05	1.072
Item 7	0	0	0	0	0	0.284	0.43	**0.332**	**0.559**	0.223	0.196	1.11
Item 97	0	0	0	0	0	**0.374**	0.588	0.085	**0.435**	0.186	0.071	1.129
Item 121	0	0	0	0	0	**0.372**	0.599	0.109	0.357	0.158	0.015	1.177
Item 127	0	0	0	0	0	0.251	0.366	0.209	0.44	0.226	0.243	1.377
Item 139	0	0	0	0	0	**0.469**	0.481	0.288	**0.4**	0.223	−0.001	1.093
Item 157	0	0	0	0	0	**0.603**	0.468	0.224	**0.415**	0.297	−0.075	0.953
Item 175	0	0	0	0	0	0.244	0.366	**0.318**	**0.589**	0.186	0.253	1.185
Item 187	0	0	0	0	0	**0.626**	0.35	0.241	**0.491**	0.291	−0.1	0.969

$\hat{\boldsymbol{\Phi}}$						
ϕ_1	1					
ϕ_2	−0.227	1				
ϕ_3	−0.194	0.461	1			
ϕ_4	−0.014	0.378	0.189	1		
ϕ_5	0.036	0.558	0.458	0.192	1	
ϕ_6	0.026	0.259	0.457	0.155	0.398	1

$\hat{\kappa} = 0.567$

Tabella 4.4: Parametri stimati dal modello Mix-CFA-EFA con $q = 6$ e $K = 5$ con con $\lambda \geq 0.3$ in grassetto. Nelle celle tra fattori del modello CFA e items si trovano i valori della matrice $\hat{\boldsymbol{\Lambda}}_1$ ed accanto la matrice $\hat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$. Nelle celle tra fattori del modello CFA e items si trovano i valori della matrice $\hat{\boldsymbol{\Lambda}}_2$ con accanto la matrice $\hat{\boldsymbol{\Psi}}_{\boldsymbol{\delta}}$. Sotto si trovano le stime di $\hat{\boldsymbol{\Phi}}$ e $\hat{\kappa}$

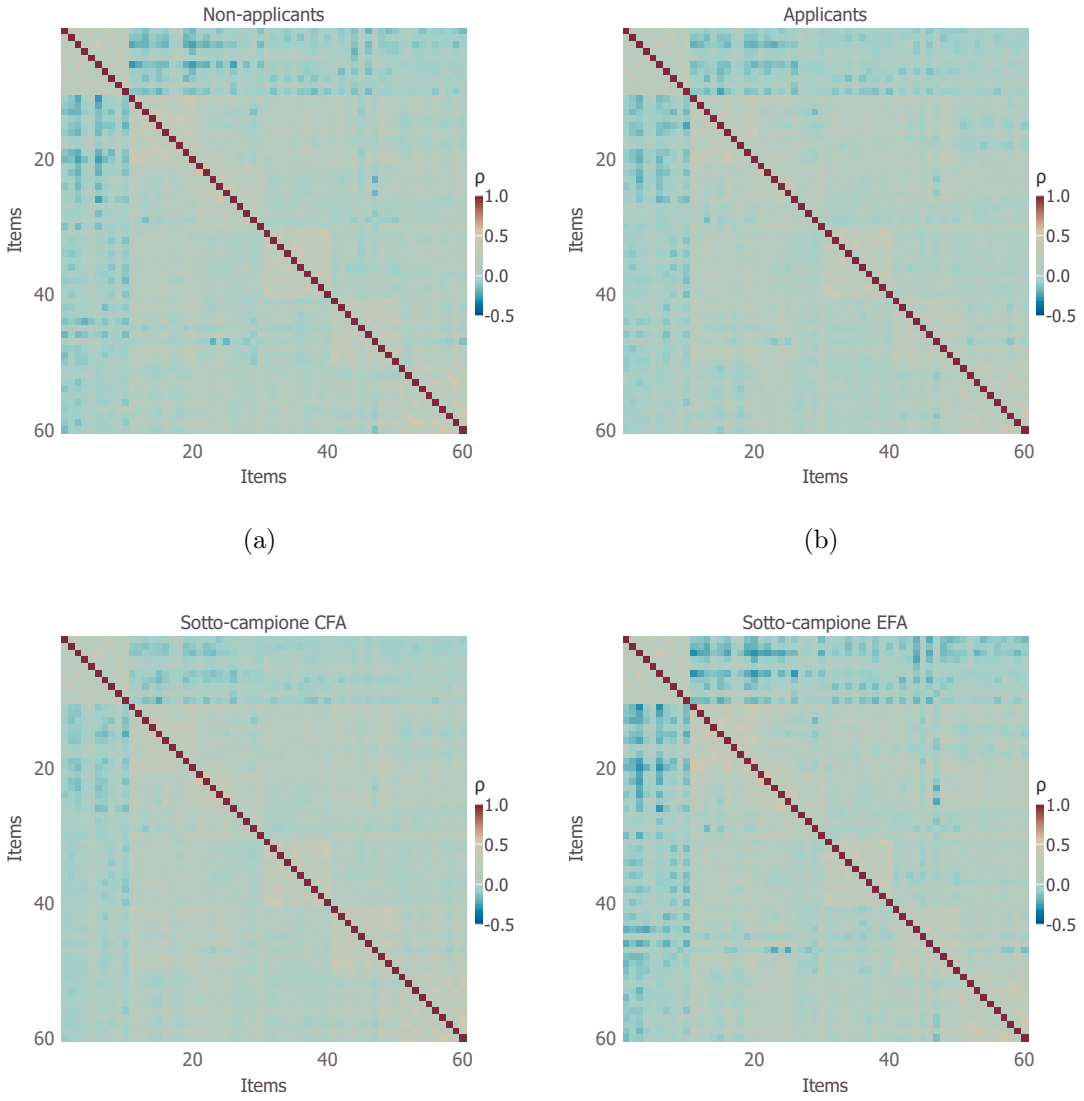


Figura 4.1: Grafici heatmap di matrici di correlazione tra items osservate: (a) matrice di correlazione del gruppo `applicant`; (b) matrice di correlazione del gruppo `non-applicant`; (c) matrice di correlazione del sotto-campione individuato dalla CFA; (d) matrice di correlazione del sotto-campione della EFA. I colori del grafico indicano il grado di correlazione tra items (ρ).

Capitolo 5

Conclusioni

In questa tesi è stato proposto un nuovo Factor Mixture Model, il modello Mix-CFA-EFA. Il modello Mix-CFA-EFA consiste in un modello di mistura che comprende, come componenti di mistura, un modello CFA ed un modello EFA. Come rilevato dalla letteratura scientifica[1], i/le ricercatori/ricercatrici necessitano di strumenti statistici per analizzare campioni di soggetti, che assicurino la validità delle proprie inferenze e la possibilità di generalizzarle alla popolazione oggetto di studio. Per questo scopo, è stata sviluppata la classe di modelli FFM.

Rispetto ai FFM più comuni, il presente modello ha l'obiettivo di gestire l'eterogeneità campionaria, indotta dalla presenza di più popolazioni nei dati, tramite la densità componente EFA, permettendo, al contempo, una stima separata del modello CFA ipotizzato dai/dalle ricercatori/ricercatrici. In particolare, il/la ricercatore/ricercatrice può utilizzare il Mix-CFA-EFA per assicurarsi che il campione, qualora nasconda più sotto-campioni, venga epurato dalle osservazioni appartenenti ad altre popolazioni e che il modello CFA sia stimato sulla porzione di campione corretta, preservando la validità di misurazione. Ad esempio, campioni ottenuti tramite campionamento di convenienza potrebbero richiedere questo tipo di analisi statistica preliminare (Julian, 2001).

Come emerge dai risultati della simulazione, il modello Mix-CFA-EFA permette di ottenere ottime stime dei parametri ed un'accuratezza nella classificazione delle osservazioni nella componente CFA ed EFA piuttosto soddisfacente. L'applicazione con dati alterati da faking, invece, mostra come il modello debba essere migliorato nella classificazione condotta su dati reali. Ciononostante, il Mix-CFA-EFA, ha fornito una stima della proporzione di campione affetta da eterogeneità in linea con le attese. Nel

[1] si veda Capitolo 1.

complesso, i risultati sono incoraggianti e permettono di delineare possibili sviluppi futuri del modello.

Difatti, il modello Mix-CFA-EFA può essere ulteriormente sviluppato seguendo direzioni differenti. Ad esempio, nel contesto di ricerca in ambito psicologico si considera necessaria l'inclusione della possibilità di modellare variabili osservate discrete, per gestire in modo più adeguato i dati di rating. Un altro possibile sviluppo del modello consiste nell'inclusione covariate esterne per lo studio dell'eterogeneità osservata. Infine, siccome il modello Mix-CFA-EFA ha mostrato che in contesti applicativi potrebbe fornire tassi di classificazioni corrette piuttosto bassi, si rileva necessario condurre indagini in merito più approfondite. Tale criticità può sollevare dubbi sulla specificazione stessa del modello. Infatti, il modello CFA ed il modello EFA, come definiti in questo documento, potrebbero essere eccessivamente sovrapponibili. Pertanto, in certe situazioni il modello CFA potrebbe essere sussunto nel modello EFA. Per questo motivo, è necessario condurre studi di simulazione più estesi e particolareggiati.

Bibliografia

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Amiri, L., Khazaei, M., & Ganjali, M. (2018). A mixture latent variable model for modeling mixed data in heterogeneous populations and its applications. *AStA Advances in Statistical Analysis*, *102*(1), 95–115. https://doi.org/10.1007/s10182-017-0294-3

An, X., & Bentler, P. M. (2011). Extended mixture factor analysis model with covariates for mixed binary and continuous responses. *Statistics in medicine*, *30*(21), 2634–2647. https://doi.org/10.1002/sim.4310

Anglim, J., Morse, G., De Vries, R. E., MacCann, C., & Marty, A. (2017). Comparing job applicants to non–applicants using an item–level bifactor model on the HEXACO personality inventory. *European journal of personality*, *31*(6), 669–684. https://doi.org/10.1002/per.2120

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: A Bayesian approach. *Psychometrika*, *67*(1), 49–77. https://doi.org/10.1007/BF02294709

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*, *19*(4), 328–347. https://doi.org/10.1287/mksc.19.4.328.11789

Arminger, G., Wittenberg, J., & Schepers, A. (1996). MECOSA 3 User Guide. Friedrichsdorf. *Additive GmbH*.

Arminger, G., & Stein, P. (1997). Finite mixtures of covariance structure models with regressors: Loglikelihood function, minimum distance estimation, fit indices, and a complex example. *Sociological Methods & Research*, *26*(2), 148–182. https://doi.org/10.1177/0049124197026002002

Arminger, G., Stein, P., & Wittenberg, J. (1999). Mixtures of conditional mean-and covariance-structure models. *Psychometrika*, *64*(4), 475–494. https://doi.org/10.1007/BF02294568

Arnold, M., Oberski, D. L., Brandmaier, A. M., & Voelkle, M. C. (2020). Identifying heterogeneity in dynamic panel models with individual parameter contribution regression. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(4), 613–628. https://doi.org/10.1080/10705511.2019.1667240

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality factors: A review of research and theory. *Personality and Social Psychology Review*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Assaf, A. G., Oh, H., & Tsionas, M. G. (2016). Unobserved heterogeneity in hospitality and tourism research. *Journal of Travel Research*, *55*(6), 774–788. https://doi.org/10.1177/0047287515588593

Bartholomew, D. J., Knott, M., & Moustaki, I. (2011). John Wiley & Sons.

Bauer, D. J. (2005). A semiparametric approach to modeling nonlinear relations among latent variables. *Structural Equation Modeling*, *12*(4), 513–535. https://doi.org/10.1207/s15328007sem1204_1

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological methods*, *9*(1), 3. https://doi.org/10.1037/1082-989X.9.1.3

Becker, J.-M., Rai, A., Ringle, C. M., & Völckner, F. (2013). Discovering unobserved heterogeneity in structural equation models to avert validity threats. *MIS quarterly*, *37*(3), 665–694.

Bernstein, A., Stickle, T. R., & Schmidt, N. B. (2013). Factor mixture model of anxiety sensitivity and anxiety psychopathology vulnerability. *Journal of Affective Disorders*, *149*(1-3), 406–417. https://doi.org/10.1016/j.jad.2012.11.024

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. https://doi.org/10.1137/141000671

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T., & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*(4), 317–335. https://doi.org/10.1111/j.1468-2389.2006.00354.x

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer New York, NY.

Bolhuis, K., Lubke, G. H., van der Ende, J., Bartels, M., van Beijsterveldt, C. E., Lichtenstein, P., Larsson, H., Jaddoe, V. W., Kushner, S. A., Verhulst, F. C., et al. (2017). Disentangling heterogeneity of childhood disruptive behavior problems into dimensions and subgroups. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(8), 678–686. https://doi.org/10.1016/j.jaac.2017.05.019

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons. https://doi.org/10.1002/9781118619179

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological methods*, *18*(1), 71–86. https://doi.org/10.1037/a0030001

Bulteel, K., Wilderjans, T. F., Tuerlinckx, F., & Ceulemans, E. (2013). CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods*, *45*(3), 782–791. https://doi.org/10.3758/s13428-012-0293-y

Buzick, H. M. (2010). Testing for heterogeneous factor loadings using mixtures of confirmatory factor analysis models. *Frontiers in Psychology*, *1*, 165. https://doi.org/10.3389/fpsyg.2010.00165

Cagnone, S., & Viroli, C. (2012). A factor mixture analysis model for multivariate binary data. *Statistical Modelling*, *12*(3), 257–277. https://doi.org/10.1177/1471082X1101200303

Cagnone, S., & Viroli, C. (2014). A factor mixture model for analyzing heterogeneity and cognitive structure of dementia. *AStA Advances in Statistical Analysis*, *98*(1), 1–20. https://doi.org/10.1007/s10182-012-0206-5

Cai, J.-H., & Song, X.-Y. (2010). Bayesian analysis of mixtures in structural equation models with non-ignorable missing data. *British Journal of Mathematical and Statistical Psychology*, *63*(3), 491–508. https://doi.org/10.1348/000711009x475187

Cai, J.-H., Song, X.-Y., & Hser, Y.-I. (2010). A Bayesian analysis of mixture structural equation models with non-ignorable missing responses and covariates. *Statistics in Medicine*, *29*(18), 1861–1874. https://doi.org/10.1002/sim.3915

Cai, J.-H., Song, X.-Y., Lam, K.-H., & Ip, E. H.-S. (2011). A mixture of generalized latent variable models for mixed mode and heterogeneous data. *Computational Statistics & Data Analysis*, *55*(11), 2889–2907. https://doi.org/10.1016/j.csda.2011.05.011

Cappozzo, A., & Greselin, F. (2019). Detecting Wine Adulterations Employing Robust Mixture of Factor Analyzers. In F. Greselin, L. Deldossi, L. Bagnato & V. Maurizio (Cur.), *Statistical Learning of Complex Data* (pp. 13–21). Springer. https://doi.org/10.1007/978-3-030-21140-0

Ceulemans, E., & Kiers, H. A. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British journal of mathematical and statistical psychology*, *59*(1), 133–150. https://doi.org/10.1348/000711005X64817

Clark, S. L., Muthén, B. O., Kaprio, J., D'Onofrio, B. M., Viken, R., & Rose, R. J. (2013). Models and strategies for factor mixture analysis: An example concerning the structure underlying psychological disorders. *Structural equation modeling: a multidisciplinary journal*, *20*(4), 681–703. https://doi.org/10.1080/10705511.2013.824786

Cohen, A. S., & Bolt, D. M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*(2), 133–148. https://doi.org/10.1111/j.1745-3984.2005.00007

De Ayala, R. J., Kim, S.-H., Stapleton, L. M., & Dayton, C. M. (2002). Differential item functioning: A mixture distribution conceptualization. *International Journal of Testing*, *2*(3-4), 243–276. https://doi.org/10.1080/15305058.2002.9669495

De Roover, K., Vermunt, J. K., & Ceulemans, E. (2020). Mixture multigroup factor analysis for unraveling factor loading noninvariance across many groups. *Psychological Methods*, *27*(3), 281–306. https://doi.org/10.1037/met0000355

De Roover, K., Vermunt, J. K., Timmerman, M. E., & Ceulemans, E. (2017). Mixture simultaneous factor analysis for capturing differences in latent variables between higher level units of multilevel data. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(4), 506–523. https://doi.org/10.1080/10705511.2017.1278604

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. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

DeSarbo, W. S., Di Benedetto, C. A., Jedidi, K., & Song, M. (2006). Identifying sources of heterogeneity for empirically deriving strategic types: a constrained finite-mixture structural-equation methodology. *Management Science*, *52*(6), 909–924. https://doi.org/10.1287/mnsc.1060.0529

Divers, R., Robinson, A., Miller, L., Davis, K., Reed, C., & Calamia, M. (2022). Examining heterogeneity in depression symptoms and associations with cognition and everyday function in MCI. *Journal of Clinical and Experimental Neuropsychology*, *44*(3), 185–194. https://doi.org/10.1080/13803395.2022.2102154

Dolan, C. V., Schmittmann, V. D., Lubke, G. H., & Neale, M. C. (2005). Regime switching in the latent growth curve mixture model. *Structural Equation Modeling*, *12*(1), 94–119. https://doi.org/10.1207/s15328007sem1201_5

Dolan, C. V., & van der Maas, H. L. (1998). Fitting multivariage normal finite mixtures subject to structural equation modeling. *Psychometrika*, *63*(3), 227–253. https://doi.org/10.1007/BF02294853

Fokoué, E. (2005). Mixtures of factor analyzers: an extension with covariates. *Journal of Multivariate Analysis*, *95*(2), 370–384. https://doi.org/10.1016/j.jmva.2004.08.004

Fokoué, E., & Titterington, D. (2003). Mixtures of factor analysers. Bayesian estimation and inference by stochastic simulation. *Machine Learning*, *50*(1), 73–94. https://doi.org/10.1023/A:1020297828025

Friehs, M.-T., Masselmann, J., Trautner, M., Kotzur, P. F., & Schmidt, P. (2022). Unobserved heterogeneity between individuals in Group-Focused Enmity. *International Journal of Conflict and Violence*, *16*, 1–17. https://doi.org/10.11576/ijcv-5266

Ghahramani, Z., & Hinton, G. E. (1997). *The EM algorithm for factor analyzers* (rapp. tecn. N. CRG-TR-96-1). Department of Computer Science, University of Toronto Toronto. 6 King's College Road, Toronto, Canada, M5S 1A4.

Ghahramani, Z., & Beal, M. (1999). Variational inference for Bayesian mixtures of factor analysers. In S. Solla, T. Leen & K. Müller (Cur.), *Advances in neural information processing systems* (Vol. 12). MIT Press.

Görz, N., Hildebrandt, L., & Annacker, D. (2000). Analyzing Multigroup Data with Structural Equation Models. In R. Decker & W. Gaul (Cur.), *Classification and Information Processing at the Turn of the Millennium* (pp. 312–319). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-57280-7_34

Grimm, K. J., Mazza, G. L., & Davoudzadeh, P. (2017). Model selection in finite mixture models: A k-fold cross-validation approach. *Structural Equation Modeling: A Multidisciplinary Journal*, *24*(2), 246–256. https://doi.org/10.1080/10705511.2016.1250638

Hahn, C., Johnson, M. D., Herrmann, A., & Huber, F. (2002). Capturing customer heterogeneity using a finite mixture PLS approach. *Schmalenbach Business Review*, *54*(3), 243–269. https://doi.org/10.1007/BF03396655

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(2), 202–226. https://doi.org/10.1080/10705510709336744

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological methods*, *11*(1), 36–53. https://doi.org/10.1037/1082-989X.11.1.36

Holmes, G. K. (1892). Measures of distribution. *Journal of the American Statistical Association*, *3*, 141–157.

Hoshino, T. (2001). Bayesian inference for finite mixtures in confirmatory factor analysis. *Behaviormetrika*, *28*(1), 37–63. https://doi.org/10.2333/bhmk.28.37

Howard, M. C., & Hoffman, M. E. (2018). Variable-centered, person-centered, and person-specific approaches: Where theory meets the method. *Organizational Research Methods*, *21*(4), 846–876. https://doi.org/10.1177/1094428117744021

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307. https://doi.org/10.1093/biomet/76.2.297

Jackson, K., Bucholz, K., Wood, P., Steinley, D., Grant, J., & Sher, K. (2014). Towards the characterization and validation of alcohol use disorder subtypes: Integrating consumption and symptom data. *Psychological medicine*, *44*(1), 143–159. https://doi.org/10.1017/S0033291713000573

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2017). A comparison of methods for uncovering sample heterogeneity: Structural equation model trees and finite mixture models. *Structural equation modeling: a multidisciplinary journal*, *24*(2), 270–282. https://doi.org/10.1080/10705511.2016.1250637

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997a). STEMM: A general finite mixture structural equation model. *Journal of Classification*, *14*(1), 23–50. https://doi.org/10.1007/s003579900002

Jedidi, K., Jagpal, H. S., & DeSarbo, W. S. (1997b). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*(1), 39–59. https://doi.org/10.1287/mksc.16.1.39

Jensen, T. M. (2017). Constellations of dyadic relationship quality in stepfamilies: A factor mixture model. *Journal of Family Psychology*, *31*(8), 1051. https://doi.org/10.1037/fam0000355

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409–426. https://doi.org/10.1007/BF02291366

Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American statistical Association*, *70*(351a), 631–639. https://doi.org/10.1080/01621459.1975.10482485

Julian, M. W. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural equation modeling*, *8*(3), 325–352. https://doi.org/10.1207/s15328007sem0803_1

Kiefer, C., Lemmerich, F., Langenberg, B., & Mayer, A. (2022). Subgroup discovery in structural equation models. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000524

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Lämmle, L., Ziegler, M., Seidel, I., Worth, A., & Bös, K. (2013). Four classes of physical fitness in German children and adolescents: only differences in performance or at-risk groups? *International journal of public health*, *58*(2), 187–196. https://doi.org/10.1007/s00038-012-0427-0

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Houghton Mifflin, New York.

Lee, S.-Y., & Song, X.-Y. (2003). Bayesian model selection for mixtures of structural equation models with an unknown number of components. *British Journal of Mathematical and Statistical Psychology*, *56*(1), 145–165. https://doi.org/10.1348/000711003321645403

Leite, W. L., & Cooper, L. A. (2010). Detecting social desirability bias using factor mixture models. *Multivariate Behavioral Research*, *45*(2), 271–293. https://doi.org/10.1080/00273171003680245

Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling*, *8*(4), 493–530. https://doi.org/10.1207/S15328007SEM0804_01

Liu, H., & Song, X. Y. (2017). Bayesian analysis of mixture structural equation models with an unknown number of components. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 41–55. https://doi.org/10.1080/10705511.2017.1372688

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika*, *57*(4), 581–597. https://doi.org/10.1007/BF02294421

Lubke, G. H., & Luningham, J. (2017). Fitting latent variable mixture models. *Behaviour research and therapy*, *98*, 91–102. https://doi.org/10.1016/j.brat.2017.04.003

Lubke, G. H., Muthen, B. O., Moilanen, I. K., McGOUGH, J. J., Loo, S. K., Swanson, J. M., Yang, M. H., Taanila, A., Hurtig, T., Järvelin, M.-R., et al. (2007). Subtypes versus severity differences in attention-deficit/hyperactivity disorder in the Northern Finnish Birth Cohort. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(12), 1584–1593. https://doi.org/10.1097/chi.0b013e31815750dd

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Lubke, G. H., & Muthén, B. O. (2007). Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(1), 26–47. https://doi.org/10.1080/10705510709336735

Lubke, G. H., Ouwens, K. G., de Moor, M. H., Trull, T. J., & Boomsma, D. I. (2015). Population heterogeneity of trait anger and differential associations of trait anger facets with borderline personality features, neuroticism, depression, Attention Deficit Hyperactivity Disorder (ADHD), and alcohol problems. *Psychiatry Research*, *230*(2), 553–560. https://doi.org/10.1016/j.psychres.2015.10.003

Masyn, K. E., Henderson, C. E., & Greenbaum, P. E. (2010). Exploring the latent structures of psychological constructs in social development using the dimensional–categorical spectrum. *Social Development*, *19*(3), 470–493. https://doi.org/10.1111/j.1467-9507.2009.00573.x

McIntyre, H. H. (2011). Investigating response styles in self-report personality data via a joint structural equation mixture modeling of item responses and response times. *Personality and Individual Differences*, *50*(5), 597–602. https://doi.org/10.1016/j.paid.2010.12.001

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

McLachlan, G. J., Lee, S. X., & Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, *6*, 355–378. https://doi.org/10.1146/annurev-statistics-031017-100325

McLachlan, G. J., & Peel, D. (2000). *Finite Mxiture 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. https://doi.org/10.1016/S0167-9473(02)00183-4

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. https://doi.org/10.1111/1467-9868.00082

Merkle, E. C., Fan, J., & Zeileis, A. (2014). Testing for measurement invariance with respect to an ordinal variable. *Psychometrika*, *79*(4), 569–584. https://doi.org/10.1007/s11336-013-9376-7

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82. https://doi.org/10.1007/s11336-012-9302-4

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in nuclear medicine*, *8*(4), 283–298. https://doi.org/10.1016/S0001-2998(78)80014-2

Miettunen, J., Nordström, T., Kaakinen, M., & Ahmed, A. (2016). Latent variable mixture modeling in psychiatric research–a review and application. *Psychological Medicine*, *46*(3), 457–467. https://doi.org/10.1017/S0033291715002305

Montanari, A., & Viroli, C. (2010). Heteroscedastic factor mixture analysis. *Statistical Modelling*, *10*(4), 441–460. https://doi.org/10.1177/1471082X0901000405

Montanari, A., & Viroli, C. (2011). Maximum likelihood estimation of mixtures of factor analyzers. *Computational statistics & data analysis*, *55*(9), 2712–2723. https://doi.org/10.1016/j.csda.2011.04.001

Moore, W. L. (1980). Levels of aggregation in conjoint analysis: An empirical comparison. *Journal of marketing research*, *17*(4), 516–523. https://doi.org/10.1177/002224378001700410

Morin, A. J., Bujacz, A., & Gagné, M. (2018). Person-centered methodologies in the organizational sciences: Introduction to the feature topic. https://doi.org/10.1177/1094428118773856

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

Mulaik, S. A. (2009). *Foundations of factor analysis*. CRC press.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585. https://doi.org/10.1007/BF02296397

Muthén, B. O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class–latent growth modeling. In L. M. Collins & S. A. G (Cur.), *New methods for the analysis of change* (pp. 291–322). American Psychological Association. https://doi.org/10.1037/10409-010

Muthén, B. O. (2006). Should substance use disorders be considered as categorical or dimensional? *Addiction*, *101*(s1), 6–16. https://doi.org/10.1111/j.1360-0443.2006.01583.x

Muthén, B. O., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental research*, *24*(6), 882–891. https://doi.org/10.1111/j.1530-0277.2000.tb02070.x

Muthén, B. O., & Satorra, A. (1989). 5 - MULTILEVEL ASPECTS OF VARYING PARAMETERS IN STRUCTURAL MODELS. In R. D. Bock (Cur.), *Multilevel Analysis of Educational Data* (pp. 87–99). Academic Press. https://doi.org/10.1016/B978-0-12-108840-8.50009-3

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463–469. https://doi.org/10.1111/j.0006-341X.1999.00463.x

Pace, L., Salvan, A., et al. (2001). *Introduzione alla statistica-II. Inferenza, verosimiglianza, modelli*. Cedam.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, *185*, 71–110. https://doi.org/10.1098/rsta.1894.0003

Picardi, A., Viroli, C., Tarsitani, L., Miglio, R., de Girolamo, G., Dell'Acqua, G., & Biondi, M. (2012). Heterogeneity and symptom structure of schizophrenia. *Psychiatry research*, *198*(3), 386–394. https://doi.org/10.1016/j.psychres.2011.12.051

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate behavioral research*, *49*(6), 518–543. https://doi.org/10.1080/00273171.2014.933762

Powers, D. (2007). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation* (rapp. tecn. N. SIE-07-001). School of Informatics e Engineering, Flinders University of South Australia. Adelaide 5001, South Australia.

Prati, R. C., Batista, G. E., & Monard, M. C. (2009). Data mining with imbalanced class distributions: concepts and methods. *IICAI*, 359–376.

Redican, E., Cloitre, M., Hyland, P., McBride, O., Karatzias, T., Murphy, J., & Shevlin, M. (2022). The latent structure of ICD-11 posttraumatic stress disorder (PTSD) and complex PTSD in a general population sample from USA: A factor mixture

modelling approach. *Journal of Anxiety Disorders*, *85*, 102497. https://doi.org/10.1016/j.janxdis.2021.102497

Reynolds, M. R., Keith, T. Z., & Natasha Beretvas, S. (2010). Use of factor mixture modeling to capture Spearman's law of diminishing returns. *Intelligence*, *38*(2), 231–241. https://doi.org/10.1016/j.intell.2010.01.002

Rigdon, E. E., Ringle, C. M., & Sarstedt, M. (2010). Structural modeling of heterogeneous data with partial least squares. In N. K. Malhotra (Cur.), *Review of marketing research* (pp. 255–296). Emerald Group Publishing Limited, Bingley. https://doi.org/10.1108/s1548-6435(2010)00000070

Roberson-Nay, R., & Kendler, K. (2011). Panic disorder and its subtypes: a comprehensive analysis of panic symptom heterogeneity using epidemiological and treatment seeking samples. *Psychological medicine*, *41*(11), 2411–2421. https://doi.org/10.1017/S0033291711000547

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*(1), 69–76. https://doi.org/10.1007/BF02293851

Sarstedt, M. (2008). A review of recent approaches for capturing heterogeneity in partial least squares path modelling. *Journal of modelling in Management*, *3*(2), 140–161. https://doi.org/10.1108/17465660810890126

Sarstedt, M., Radomir, L., Moisescu, O. I., & Ringle, C. M. (2022). Latent class analysis in PLS-SEM: A review and recommendations for future applications. *Journal of Business Research*, *138*, 398–407. https://doi.org/10.1016/j.jbusres.2021.08.051

Sarstedt, M., & Ringle, C. M. (2010). Treating unobserved heterogeneity in PLS path modeling: a comparison of FIMIX-PLS with different data analysis strategies. *Journal of Applied Statistics*, *37*(8), 1299–1318. https://doi.org/10.1080/02664760903030213

Sawatzky, R., Ratner, P. A., Johnson, J. L., Kopec, J. A., & Zumbo, B. D. (2009). Sample heterogeneity and the measurement structure of the Multidimensional Students' Life Satisfaction Scale. *Social Indicators Research*, *94*(2), 273–296. https://doi.org/10.1007/s11205-008-9423-4

Sawatzky, R., Russell, L. B., Sajobi, T. T., Lix, L. M., Kopec, J., & Zumbo, B. D. (2018). The use of latent variable mixture models to identify invariant items in test construction. *Quality of Life Research*, *27*(7), 1745–1755. https://doi.org/10.1007/s11136-017-1680-8

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

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333–343. https://doi.org/10.1007/BF02294360

Sörbom, D. (1974). A GENERAL METHOD FOR STUDYING DIFFERENCES IN FACTOR MEANS AND FACTOR STRUCTURE BETWEEN GROUPS. *British Journal of Mathematical and Statistical Psychology*, *27*(2), 229–239. https://doi.org/10.1111/j.2044-8317.1974.tb00543.x

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, *15*(2), 201–292. https://doi.org/10.2307/1412107

Suárez, M. J., & Muñiz, C. (2018). Unobserved heterogeneity in work absence. *The European Journal of Health Economics*, *19*(8), 1137–1148. https://doi.org/10.1007/s10198-018-0962-6

Sunderland, M., Carragher, N., Wong, N., & Andrews, G. (2013). Factor mixture analysis of DSM-IV symptoms of major depression in a treatment seeking clinical population. *Comprehensive psychiatry*, *54*(5), 474–483. https://doi.org/10.1016/j.comppsych.2012.12.011

Temme, D., Williams, J., & Hildebrandt, L. (2002). Structural Equation Models for Finite Mixtures — Simulation Results and Empirical Applications. In W. Härdle & B. Rönz (Cur.), *Compstat* (pp. 569–574). Physica-Verlag HD. https://doi.org/10.1007/978-3-642-57489-4_88

Ten Have, M., Lamers, F., Wardenaar, K., Beekman, A., de Jonge, P., van Dorsselaer, S., Tuithof, M., Kleinjan, M., & de Graaf, R. (2016). The identification of symptom-based subtypes of depression: A nationally representative cohort study. *Journal of affective disorders*, *190*, 395–406. https://doi.org/10.1016/j.jad.2015.10.040

Thurstone, L. L. (1935). *The vectors of mind: Multiple-factor analysis for the isolation of primary traits*. University of Chicago Press. https://doi.org/10.1037/10018-000

Thurstone, L. L. (1947). *Multiple-factor analysis*. Chicago: University of Chicago Press.

Tillmann, J., Uljarevic, M., Crawley, D., Dumas, G., Loth, E., Murphy, D., Buitelaar, J., & Charman, T. (2020). Dissecting the phenotypic heterogeneity in sensory features in autism spectrum disorder: a factor mixture modelling approach. *Molecular autism*, *11*(1), 1–15. https://doi.org/10.1186/s13229-020-00367-w

Timmerman, M. E., & Kiers, H. A. L. (2002). Three-way component analysis with smoothness constraints. *Computational Statistics & Data Analysis*, *40*(3), 447–470. https://doi.org/doi.org/10.1016/S0167-9473(02)00059-2

Tueller, S., & Lubke, G. H. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*(2), 165–192. https://doi.org/10.1080/10705511003659318

Ueda, N., Nakano, R., Ghahramani, Z., & Hinton, G. E. (2000). SMEM algorithm for mixture models. *Neural computation*, *12*(9), 2109–2128. https://doi.org/10.1162/089976600300015088

Ulbricht, C. M., Chrysanthopoulou, S. A., Levin, L., & Lapane, K. L. (2018). The use of latent class analysis for identifying subtypes of depression: A systematic review. *Psychiatry Research*, *266*, 228–246. https://doi.org/10.1016/j.psychres.2018.03.003

Utsugi, A., & Kumagai, T. (2001). Bayesian analysis of mixtures of factor analyzers. *Neural Computation*, *13*(5), 993–1002. https://doi.org/10.1162/08997660151134299

Van Dam, N. T., O'Connor, D., Marcelle, E. T., Ho, E. J., Craddock, R. C., Tobe, R. H., Gabbay, V., Hudziak, J. J., Castellanos, F. X., Leventhal, B. L., et al. (2017). Data-driven phenotypic categorization for neurobiological analyses: beyond DSM-5 labels. *Biological psychiatry*, *81*(6), 484–494. https://doi.org/10.1016/j.biopsych.2016.06.027

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European journal of developmental psychology*, *9*(4), 486–492. https://doi.org/10.1080/17405629.2012.686740

Varriale, R., & Vermunt, J. K. (2012). Multilevel mixture factor models. *Multivariate Behavioral Research*, *47*(2), 247–275. https://doi.org/10.1080/00273171.2012.658337

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological methodology*, *33*(1), 213–239. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K. (2008). Latent class and finite mixture models for multilevel data sets. *Statistical methods in medical research*, *17*(1), 33–51. https://doi.org/10.1177/0962280207081238

Wall, M. M., Guo, J., & Yasuo, A. (2012). Mixture Factor Analysis for Approximating a Nonnormally Distributed Continuous Latent Factor With Continuous and Dichotomous Observed Variables. *Multivariate Behavioral Research*, *47*(2), 276–313. https://doi.org/10.1080/00273171.2012.658339

Wang, T., Merkle, E. C., & Zeileis, A. (2014). Score-based tests of measurement invariance: use in practice. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00438

Wang, Y., Hsu, H.-Y., & Kim, E. (2021). Investigating the impact of covariate inclusion on sample size requirements of factor mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(5), 716–724. https://doi.org/10.1080/10705511.2021.1910036

Wedel, M., & Kamakura, W. A. (2000). *Market segmentation: Conceptual and methodological foundations*. Boston, MA: Springer Science & Business Media.

Whalen, D. J. (2017). Using hybrid modeling to determine the latent structure of psychopathology. *Biological psychiatry*, *81*(6), e41–e42. https://doi.org/10.1016/j.biopsych.2016.12.017

Williams, J., Temme, D., & Hildebrandt, L. (2002). *A Monte Carlo study of structural equation models for finite mixtures* (SFB 373 Discussion Paper). Berlin, Humboldt University of Berlin, Interdisciplinary Research Project 373: Quantification; Simulation of Economic Processes. http://hdl.handle.net/10419/65330

Willmott, C. J., & Matsuura, K. (2006). On the use of dimensioned measures of error to evaluate the performance of spatial interpolators. *International Journal of Geographical Information Science*, *20*(1), 89–102. https://doi.org/10.1080/13658810500286976

Yildirim, B. O., & Derksen, J. J. (2015). Clarifying the heterogeneity in psychopathic samples: Towards a new continuum of primary and secondary psychopathy. *Aggression and Violent Behavior*, *24*, 9–41. https://doi.org/10.1016/j.avb.2015.05.001

Yuan, K.-H., & Bentler, P. M. (2010). Finite Normal Mixture SEM Analysis by Fitting Multiple Conventional SEM Models. *Sociological methodology*, *40*(1), 191–245. https://doi.org/10.1111/j.1467-9531.2010.01224.x

Yun, R. J., Stern, B. L., Lenzenweger, M. F., & Tiersky, L. A. (2013). Refining personality disorder subtypes and classification using finite mixture modeling. *Personality Disorders: Theory, Research, and Treatment*, *4*(2), 121–128. https://doi.org/10.1037/a0029944

Yung, Y.-F. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika*, *62*(3), 297–330. https://doi.org/10.1007/BF02294554

Zhao, J.-H., & Philip, L. (2008). Fast ML estimation for the mixture of factor analyzers via an ECM algorithm. *IEEE Transactions on Neural Networks*, *19*(11), 1956–1961. https://doi.org/10.1109/TNN.2008.2003467

Zhou, X., & Liu, X. (2008). The EM algorithm for the extended finite mixture of the factor analyzers model. *Computational statistics & data analysis*, *52*(8), 3939–3953. https://doi.org/10.1016/j.csda.2008.01.023

Zhu, H.-T., & Lee, S.-Y. (2001). A Bayesian analysis of finite mixtures in the LISREL model. *Psychometrika*, *66*(1), 133–152. https://doi.org/10.1007/BF02295737

Ziegler, M., Maaß, U., Griffith, R., & Gammon, A. (2015). What Is the Nature of Faking? Modeling Distinct Response Patterns and Quantitative Differences in Faking at the Same Time. *Organizational Research Methods*, *18*(4), 679–703. https://doi.org/10.1177/1094428115574518

Ziegler, M., MacCann, C., & Roberts, R. D. (2011). 3 Faking: Knowns, Unknowns, and Points of Contention. In *New Perspectives on Faking in Personality Assessment*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195387476.003.0011

Zumbo, B. D. (2006). 3 Validity: Foundational Issues and Statistical Methodology. In C. Rao & S. Sinharay (Cur.), *Psychometrics* (pp. 45–79, Vol. 26). Elsevier. https://doi.org/10.1016/S0169-7161(06)26003-6

Appendice A

Codice in Julia

A.1 Codice del grafico del modello di mistura di Normali

```
# Caricamento librerie
using Gadfly, Distributions, DataFrames, Fontconfig, Cairo


# Simulazione di dati da una mistura di Normali univariate
n = 500;
modone = rand(Normal(-2,1.5),n);
modsec = rand(Normal(2,1),n);

z = rand(Binomial(1, 0.5),n);
norm_mix=zeros(Float64,n);
for i=1:n
    if z[i]==1
        norm_mix[i] = modone[i];
    elseif z[i]==0
        norm_mix[i] = modsec[i];
    end
end
df_mix = DataFrame(dat=norm_mix,z=string.(z));


# Grafico
coord = Coord.cartesian(ymin=0, ymax=0.4);
pb = plot(df_mix,Guide.xlabel("yi"), Guide.ylabel("Stime kernel di densita'"), coord, layer(x=:dat, y=zeros(n),
      color=:z, Geom.point, Geom.density, Theme(key_position=:none,line_style=[:solid,:dot], line_width=1.5pt,
     discrete_highlight_color=identity,colorkey_swatch_shape=:circle)), layer(x=:dat, Geom.density, Theme(
     default_color=colorant"purple",line_width=1.5pt)), Scale.color_discrete_manual("darkblue","darkred"),
     Guide.colorkey(title="Indicatore",labels=["1","0"]))
draw(PDF("C:/Users/NC/MEGA/Tesi/Tesi/fig4.pdf", 5inch, 5inch), pb)
```

A.2 Codice dello studio di simulazione e relative funzioni

A.2.1 Codice dello studio di simulazione

```
# Caricamento delle librerie
using StatsKit, Random, SpecialFunctions, Roots, QuadGK, Optim, LinearAlgebra, Distributions, DataFrames, CSV,
      StatsBase, Missings, TimeSeries, HDF5, JLD, DelimitedFiles

# Caricamento delle funzioni
include("C:/Users/NC/MEGA/Tesi/functions.jl")



# Studio di simulazione

# Definizione del design
n = 1000;
p = 20;
q = 1;
Kfalse = [2 4];
K = [2 4];
kappa0 = [0.80 0.60 0.20];
design = vec(collect(Base.product(n,p,q,Kfalse,K,kappa0)));
B = 500;


# Generazione dei parametri veri e dei dati
TruePar = Design_True_Data(design,B,true,false);
Y_sim = Design_True_Data(design,B,TruePar,true);


# Applicaizone del modello ai dati simulati e storing dei risultati
Simulation(design,Y_sim)


# Risulati per singolo disegno fattoriale

#Inizializzazione della matrici dei risultati
d1 = reshape(NORM[1:size(design,1)*B],size(design,1),B); d2 = reshape(NORM[(size(design,1)*B+1):(size(design,1)
     *B*2)],size(design,1),B); d3 = reshape(NORM[(size(design,1)*B*2+1):(size(design,1)*B*3)],size(design,1),B)
     ;d4 = reshape(NORM[(size(design,1)*B*3+1):(size(design,1)*B*4)],size(design,1),B); d5 = reshape(NORM[(
     size(design,1)*B*4+1):(size(design,1)*B*5)],size(design,1),B); d6 = reshape(NORM[(size(design,1)*B*5+1):(
     size(design,1)*B*6)],size(design,1),B);
LLIK = reshape(INDEX[1:size(design,1)*B],size(design,1),B); AIC = reshape(INDEX[(size(design,1)*B+1):(size(
     design,1)*B*2)],size(design,1),B); AICc = reshape(INDEX[(size(design,1)*B*2+1):(size(design,1)*B*3)],size
     (design,1),B); CAIC = reshape(INDEX[(size(design,1)*B*3+1):(size(design,1)*B*4)],size(design,1),B); BIC =
      reshape(INDEX[(size(design,1)*B*4+1):(size(design,1)*B*5)],size(design,1),B); ssBIC = reshape(INDEX[(
     size(design,1)*B*5+1):(size(design,1)*B*6)],size(design,1),B); MIXAIC = reshape(INDEX[(size(design,1)*B
     *6+1):(size(design,1)*B*7)],size(design,1),B); CLC = reshape(INDEX[(size(design,1)*B*7+1):(size(design,1)*
     B*8)],size(design,1),B); ICLBIC = reshape(INDEX[(size(design,1)*B*8+1):(size(design,1)*B*9)],size(design
     ,1),B);
l1 = reshape(RMSE[1:size(design,1)*B],size(design,1),B); l2 = reshape(RMSE[(size(design,1)*B+1):(size(design,1)
     *B*2)],size(design,1),B); l3 = reshape(RMSE[(size(design,1)*B*2+1):(size(design,1)*B*3)],size(design,1),B)
     ;l4 = reshape(RMSE[(size(design,1)*B*3+1):(size(design,1)*B*4)],size(design,1),B); l5 = reshape(RMSE[(
     size(design,1)*B*4+1):(size(design,1)*B*5)],size(design,1),B); l6 = reshape(RMSE[(size(design,1)*B*5+1):(
     size(design,1)*B*6)],size(design,1),B);
l1 = reshape(RMSE[1:size(design,1)*B],size(design,1),B); cc= reshape(RMSE[(size(design,1)*B+1):(size(design,1)*
     B*2)],size(design,1),B); l3 = reshape(RMSE[(size(design,1)*B*2+1):(size(design,1)*B*3)],size(design,1),B);
     l4 = reshape(RMSE[(size(design,1)*B*3+1):(size(design,1)*B*4)],size(design,1),B); l5 = reshape(RMSE[(size
     (design,1)*B*4+1):(size(design,1)*B*5)],size(design,1),B); l6 = reshape(RMSE[(size(design,1)*B*5+1):(size
     (design,1)*B*6)],size(design,1),B);


# Esclusione di valori NaN di non convergenza e calcolo della media delle replicazioni

d1_mean = Vector{Float64}(undef,size(design,1)); d2_mean = Vector{Float64}(undef,size(design,1)); d3_mean =
     Vector{Float64}(undef,size(design,1)); d4_mean = Vector{Float64}(undef,size(design,1)); d5_mean = Vector{
```

```
     Float64}(undef,size(design,1)); d6_mean = Vector{Float64}(undef,size(design,1)); aic_mean = Vector{
     Float64}(undef,size(design,1)); aicc_mean = Vector{Float64}(undef,size(design,1)); bic_mean = Vector{
     Float64}(undef,size(design,1)); caic_mean = Vector{Float64}(undef,size(design,1)); clc_mean = Vector{
     Float64}(undef,size(design,1)); iclbic_mean = Vector{Float64}(undef,size(design,1)); llik_mean = Vector{
     Float64}(undef,size(design,1)); ssbic_mean = Vector{Float64}(undef,size(design,1));
d1 = replace(d1, NaN => missing); d2 = replace(d2, NaN => missing); d3 = replace(d3, NaN => missing); d4 =
     replace(d4, NaN => missing); d5 = replace(d5, NaN => missing); d6 = replace(d6, NaN => missing); AIC =
     replace(AIC, NaN => missing); BIC = replace(BIC, NaN => missing); AICc = replace(AICc, NaN => missing);
     LLIK = replace(LLIK, NaN => missing); CAIC = replace(CAIC, NaN => missing); CLC = replace(CLC, NaN =>
     missing); ICLBIC = replace(ICLBIC, NaN => missing);; ssBIC = replace(ssBIC, NaN => missing);
l1_mean = Vector{Float64}(undef,size(design,1)); l2_mean = Vector{Float64}(undef,size(design,1)); l3_mean =
     Vector{Float64}(undef,size(design,1)); l4_mean = Vector{Float64}(undef,size(design,1)); l5_mean = Vector{
     Float64}(undef,size(design,1)); l6_mean = Vector{Float64}(undef,size(design,1)); aic_mean = Vector{
     Float64}(undef,size(design,1)); aicc_mean = Vector{Float64}(undef,size(design,1)); bic_mean = Vector{
     Float64}(undef,size(design,1)); caic_mean = Vector{Float64}(undef,size(design,1)); clc_mean = Vector{
     Float64}(undef,size(design,1)); iclbic_mean = Vector{Float64}(undef,size(design,1)); llik_mean = Vector{
     Float64}(undef,size(design,1)); ssbic_mean = Vector{Float64}(undef,size(design,1));
l1 = replace(l1, NaN => missing); l2 = replace(l2, NaN => missing); l3 = replace(l3, NaN => missing); l4 =
     replace(l4, NaN => missing); l5 = replace(l5, NaN => missing); l6 = replace(l6, NaN => missing); AIC =
     replace(AIC, NaN => missing); BIC = replace(BIC, NaN => missing); AICc = replace(AICc, NaN => missing);
     LLIK = replace(LLIK, NaN => missing); CAIC = replace(CAIC, NaN => missing); CLC = replace(CLC, NaN =>
     missing); ICLBIC = replace(ICLBIC, NaN => missing); ssBIC = replace(ssBIC, NaN => missing);

for i in 1:size(design,1)
    d1_mean[i] = mean(collect(skipmissing(d1[i,:])))
    d2_mean[i] = mean(collect(skipmissing(d2[i,:])))
    d3_mean[i] = mean(collect(skipmissing(d3[i,:])))
    d4_mean[i] = mean(collect(skipmissing(d4[i,:])))
    d5_mean[i] = mean(collect(skipmissing(d5[i,:])))
    d6_mean[i] = mean(collect(skipmissing(d6[i,:])))

    l1_mean[i] = mean(collect(skipmissing(l1[i,:])))
    l2_mean[i] = mean(collect(skipmissing(l2[i,:])))
    l3_mean[i] = mean(collect(skipmissing(l3[i,:])))
    l4_mean[i] = mean(collect(skipmissing(l4[i,:])))
    l5_mean[i] = mean(collect(skipmissing(l5[i,:])))
    l6_mean[i] = mean(collect(skipmissing(l6[i,:])))

    aic_mean[i] = mean(collect(skipmissing(AIC[i,:])))
    bic_mean[i] = mean(collect(skipmissing(BIC[i,:])))
    ssbic_mean[i] = mean(collect(skipmissing(ssBIC[i,:])))
    aicc_mean[i] = mean(collect(skipmissing(AICc[i,:])))
    caic_mean[i] = mean(collect(skipmissing(CAIC[i,:])))
    clc_mean[i] = mean(collect(skipmissing(CLC[i,:])))
    iclbic_mean[i] = mean(collect(skipmissing(ICLBIC[i,:])))
    llik_mean[i] = mean(collect(skipmissing(LLIK[i,:])))
end
```

A.2.2 Funzione per il calcolo del modello Mix-CFA-EFA tramite statistiche sufficienti

```
# Caricamento librerie
using StatsKit, Random, StatsPlots, SpecialFunctions, Roots, QuadGK, Optim, LinearAlgebra, Distributions,
     DataFrames, CSV, StatsBase, Missings, TimeSeries, HDF5, JLD, DelimitedFiles, Base.Threads



function Mix_CFA_EFA_suf(Y,q,K)
    local Lambda_1, Lambda_2, Theta_delta, Psi_delta, Phi, kappa, llik, t, Ez1

        # Criteri per inizializzare nuovamente l'algoritmo utilizzando nuovi starting points casuali
        diff_llik = 1; dtol = 0.05; re = 0; t = 1; err2 = 1; non_conv = true; sum_diff_tf = true;
             DomainError_warn = 1; retr = 0;
        while err2 == 1 || non_conv || sum_diff_tf || DomainError_warn == 1;

        p = size(Y,2);
        n = size(Y,1);


```

```
17          # CFA
18          # Inizializzazione della matrice coefficienti fattoriali e delle varianze-covarianze dei fattori
19          A = reshape(Diagonal(ones(q)),q,q);
20          L_str = kron(A, ones(Int64(p/q)));
21          Lambda_1 = rand(Uniform(0,1),p,q) .* L_str;
22          Phi = rand(LKJ(q,1))
23
24          # Inizializzazione della matrice degli errori
25          Theta_delta = Diagonal(vec(1 .- sum(Lambda_1,dims=2).^2));
26
27          # EFA
28          # Inizializzazione della matrice coefficienti fattoriali e delle varianze-covarianze dei fattori
29          Lambda_2 = rand(Uniform(0,1),p,K)
30          Psi_delta = Diagonal(ones(p)*0.85)
31
32          # Mistura
33          # Inizializzazione del parametro di mistura
34          kappa = 0.50;
35
36          # Inizializzazioni e impostazioni per il ciclo while
37          t = 1; maxIter = 4000; llik = zeros(Float64, 1, maxIter); oldllik = -Inf; err=zeros(Int64,1,maxIter);
                 err2 = 0; sum_diff = zeros(Float64,1,maxIter); diff_llik = 1; sum_diff_tf = true; DomainError_warn
                  = 0;Ez1=zeros(n)
38
39          # Ciclo while dell'algoritmo Expectation-Maximization
40          while err2 == 0 && diff_llik>dtol && t<=maxIter && sum_diff_tf && DomainError_warn == 0
41
42              # E-step
43              b1 = Phi*Lambda_1'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta);
44              B1 = Phi-Phi*Lambda_1'*inv(Theta_delta+Lambda_1*Phi*Lambda_1')*Lambda_1*Phi
45              b2 = Lambda_2'*inv(Lambda_2*Lambda_2' + Psi_delta)
46              B2 = Diagonal(ones(K))-Lambda_2'*inv(Psi_delta+Lambda_2*Lambda_2')*Lambda_2
47
48              # Calcolo dei valori attesi delle statistiche sufficienti condizionate a y
49              Ez1 = zeros(n)
50              for i=1:n
51                  Ez1[i] = ((kappa .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(
                         sqrt((2 .* pi)^(p) *det(Lambda_1*Phi*Lambda_1'+Theta_delta)))))/ ((kappa .* ((exp((-1/2) *
                         Y[i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1*
                         Phi*Lambda_1'+Theta_delta))))) .+((exp((-1/2) * Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y
                         [i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*Lambda_2'+Psi_delta)))) .-(kappa .*((exp((-1/2)
                         * Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*
                         Lambda_2'+Psi_delta)))))))
52              end
53              Ez0 = 1 .- Ez1;
54              Ek1 = sum(Ez1)
55              Ek0 = sum(Ez0)
56
57              ESyy1 = zeros(p,p); ESyy0 = zeros(p,p);
58              for i in 1:n
59                  ESyy1 = ESyy1 + (Ez1[i]*Y[i,:]*Y[i,:]')/n
60                  ESyy0 = ESyy0 + (Ez0[i]*Y[i,:]*Y[i,:]')/n
61              end
62
63              ESyeta = ESyy1*b1'
64              ESyxi = ESyy0*b2'
65              ESetaeta = B1*(sum(Ez1)/n)+b1*ESyeta
66              ESxixi = (sum(Ez0)/n)*(B2) + b2*ESyxi
67
68
69              # M-step
70              # Stime di ML delle matrici Lambda_1 e Lambda_2
71              Lambda_1 = ESyeta*inv(ESetaeta) .* L_str
72              Lambda_2 = ESyxi*inv(ESxixi)
73
74              # Stime di ML delle matrici degli errori del modello CFA ed EFA
75              Theta_delta = Diagonal((ESyy1 - Lambda_1*ESyeta')/(Ek1/n))
76              Psi_delta = Diagonal((ESyy0 - Lambda_2*ESyxi')/(Ek0/n))
77
78              # Criterio per gestire eventuali valori negativi nella matrice Theta_delta
79              if sum(Theta_delta.<0) > 0
80                  println("Error Theta_delta")
81                  DomainError_warn = 1
82              else
83
```

```
84              # Stime di ML della matrice di correlazione del modello CFA
85              Phi = ESetaeta/(Ek1/n)
86              Phi = sqrt(inv(Diagonal(Phi)))*Phi*sqrt(inv(Diagonal(Phi)))
87
88              # Stima di ML del parametro di mistura
89              kappa = Ek1/n
90
91              # Calcolo della funzione di log-verosimiglianza completa per l'iterazione t
92              old_diff = diff_llik
93              llik[t] = 0
94              for i=1:n
95                  llik[t] = llik[t] + (Ez1[i]*(log(kappa)-(p/2)*log(2*pi)-(1/2)*log(det(Theta_delta))-(1/2)*(Y
                        [i,:]'*inv(Theta_delta)*Y[i,:]-2*Y[i,:]'*inv(Theta_delta)*Lambda_1*(b1*Y[i,:])+tr(
                        Lambda_1'*inv(Theta_delta)*Lambda_1*((Phi - b1*Lambda_1*Phi + b1*Y[i,:]*Y[i,:]'*b1'))))
                        -(q/2)*log(2*pi)-(1/2)*log(det(Phi))-(1/2)*tr(inv(Phi)*((Phi - b1*Lambda_1*Phi + b1*Y[i
                        ,:]*Y[i,:]'*b1'))))+ Ez0[i]*(log(1-kappa)-(p/2)*log(2*pi)-(1/2)*log(det(Psi_delta))
                        -(1/2)*(Y[i,:]'*inv(Psi_delta)*Y[i,:]-2*Y[i,:]'*inv(Psi_delta)*Lambda_2*(b2*Y[i,:])+tr(
                        Lambda_2'*inv(Psi_delta)*Lambda_2*((Diagonal(ones(K))-b2*Lambda_2 + b2*Y[i,:]*Y[i,:]'*
                        b2'))))-(K/2)*log(2*pi)-(1/2)*tr(Diagonal(ones(K)) - b2*Lambda_2 + b2*Y[i,:]*Y[i,:]'*b2
                        )))
96              end
97
98              # Calcolo della differenza tra log-verosimiglianza all'iterazione t-1 e t (criterio di stop)
99              diff_llik=abs(oldllik - llik[t])
100
101
102
103
104             # Criterio di stop per evitare eccessivi incrementi nella differenza tra verosimiglianza
                    all'interazione t-1 e t
105             sum_diff[t] = diff_llik>old_diff
106             err[t] = oldllik>llik[t]
107             err2 = err[t]
108
109
110             oldllik = llik[t]
111
112             t+=1;
113
114             sum_diff_tf = sum(sum_diff) < 70
115         end
116     end
117
118     sum_diff_tf = sum(sum_diff) >= 70
119     non_conv = t>=maxIter
120
121     # Criterio per gestire il raggiungimento del numero massimo di iterazioni con restitutuzione di
            risultati NaN
122     if non_conv && retr == 0
123         retr += 1 #non convergenza ammessa con un solo nuovo tentativo
124     elseif non_conv && retr == 1
125         non_conv = false; err2 == 1; sum_diff_tf == false;
126         Lambda_1 = Array{Float64}(undef,p,q); fill!(Lambda_1,NaN); Lambda_2 = Array{Float64}(undef,p,K);
                fill!(Lambda_2,NaN); Theta_delta = Array{Float64}(undef,p,p); fill!(Theta_delta,NaN); Psi_delta
                 = Array{Float64}(undef,p,p); fill!(Psi_delta,NaN); Phi = Array{Float64}(undef,q,q); fill!(Phi,
                NaN); kappa = Array{Float64}(undef,1,1); fill!(kappa,NaN); llik = Array{Float64}(undef,1,1);
                fill!(llik,NaN); Ez1 = Array{Float64}(undef, 1, n); fill!(Ez1,NaN);
127         return vcat(vec(Lambda_1), vec(Lambda_2), vec(diag(Theta_delta)), vec(diag(Psi_delta)), vec(Phi),
                kappa, llik, vec(Ez1))
128     end
129     re+=1
130 end
131 return vcat(vec(Lambda_1), vec(Lambda_2), vec(diag(Theta_delta)), vec(diag(Psi_delta)), vec(Phi), kappa,
        llik[t-1], Ez1);
132 end
```

A.2.3 Funzioni per la costruzione del design sperimentale

```
1  # Funzione per generare i parametri veri
2  function TrueParams(d,p,q,K,kappa0)
```

```
3
4       # CFA
5       # Generazione di Lambda_10 da una distribuzione Uniforme tra 0 e 1
6       Lambda0 = rand(Uniform(0,1),p,q);
7       A = reshape(Diagonal(ones(q)),q,q);
8       L_str = kron(A, ones(Int64(p/q)));
9       Lambda0 = Lambda0 .* L_str;
10
11      # Generazione di Phi0 da una distribuzione Lewandowski-Kurowicka-Joe
12      Phi0 = rand(LKJ(q,1))
13
14      # Generazione di Theta_delta0 tramite decomposizione della varianza
15      Theta_delta0 = Diagonal(vec(1 .- sum(Lambda0,dims=2).^2));
16
17
18      # Noise o EFA
19
20      # Generazione di Lambda_10 da una distribuzione Uniforme tra 0 e 1
21      Lambda02 = rand(Uniform(0,1),p,K)
22
23      # Generazione della matrice di correlazione della EFA come semplice matrice identità
24      P0 = Diagonal(ones(K));
25
26      # Generazione di Psi_delta
27      Psi_delta0 = Diagonal(ones(p)*0.85);
28
29      # Generazione del vettore della variabile indicatrice
30      z = rand(Binomial(1, kappa0), n);
31
32      # Gestione dei risultati tramite una matrice sparsa
33      npartot=size(vec(Lambda0),1) + size(vec(Lambda02),1) + size(vec(diag(Theta_delta0)),1) + size(vec(diag(
            Psi_delta0)),1) + size(vec(Phi0),1) + size(kappa0,1) + size(z,1) #per avere vettori output con
            dimensioni uniformi nonostante la differenza tra disegni sperimentali
34
35      TruePar = vcat(vec(Lambda0), vec(Lambda02), vec(diag(Theta_delta0)), vec(diag(Psi_delta0)), vec(Phi0),
            kappa0, z,vec(zeros(3000-npartot)))
36      writedlm(string("C:/Users/NC/Desktop/res/true/TruePar_",d,".csv"), TruePar)
37      return TruePar
38  end
39
40
41
42
43
44
45  # Funzione per generare i dati
46  function Data_generation(TruePar,n,p,q,K,pop=50000)
47      Eta0 = rand(MvNormal(zeros(q), reshape(TruePar[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q)), pop)';
48      Delta0 = rand(MvNormal(zeros(p), Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+1):((p*q)+(p*K)+p)]))), pop)
            ';
49
50      # Equazione della CFA
51      Ystar = Eta0*reshape(TruePar[1:(p*q)],p,q)' + Delta0;
52
53      Xi0 = rand(MvNormal(zeros(K), Diagonal(ones(K))), pop)'; #P0 è sempre una matrice identità
54      P_Delta0 = rand(MvNormal(zeros(p), Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+p+1):((p*q)+(p*K)+p+p)]))),
              pop)';
55
56      # Equazione della EFA
57      Ystar2 = Xi0*reshape(TruePar[(p*q+1):((p*q)+(p*K))],p,K)' + P_Delta0
58
59      # Mistura
60      z = TruePar[((p*q)+(p*K)+p+p+q^2+1+1):((p*q)+(p*K)+p+p+q^2+1+n)]
61      Y_mix_cont=zeros(Float64,n,p)
62      for i=1:n
63          if z[i]==1
64              Y_mix_cont[i,:] = vec(Ystar[i,:]);
65          elseif z[i]==0
66          Y_mix_cont[i,:] = vec(Ystar2[i,:]);
67          end
68      end
69      return Y_mix_cont
70  end
71
72
```

```
73
74
75  # Funzione per generare parametri veri e dati o solo i parametri veri
76  function Design_True_Data(design, B,TruePar=nothing,Data=false)
77
78      # Generazione solo dei parametri veri
79      if TruePar == true && Data == false
80          TruePar = Array{Float64}(undef,3000,size(design,1))
81          for d in 1:size(design,1)
82              println(string("Design no.: ",d))
83              n = design[d,1][1]
84              p = design[d,1][2]
85              q = design[d,1][3]
86              K = design[d,1][5]
87              kappa0 = design[d,1][6]
88
89              TruePar[:,d] = TrueParams(d,p,q,K,kappa0)
90          end
91          return TruePar
92
93      #Generazione dei parametri veri e dei dati
94      elseif Data == true && typeof(TruePar) != Bool
95          local Y_sim
96          for d in 1:size(design,1)
97              println(string("Design no.: ",d))
98              n = design[d,1][1]
99              p = design[d,1][2]
100             q = design[d,1][3]
101             Kfalse = design[d,1][4]
102             K = design[d,1][5]
103             kappa0 = design[d,1][6]
104
105             Y_sim = Array{Float64}(undef,n,p,B,size(design,1))
106         end
107         for d in 1:size(design,1)
108             println(string("Design no.: ",d))
109             n = design[d,1][1]
110             p = design[d,1][2]
111             q = design[d,1][3]
112             Kfalse = design[d,1][4]
113             K = design[d,1][5]
114             kappa0 = design[d,1][6]
115
116
117             dat = Array{Float64}(undef,n,p,B)
118             for b in 1:B
119                 dat[:,:,b] = Data_generation(TruePar[:,d],n,p,q,K)
120             end
121
122             Y_sim[:,:,:,d] = dat
123         end
124         return Y_sim
125     end
126 end
127
128
129
130 # Funzione per calcolare il numero dei parametri
131
132 function nparameters(design)
133     npar_D = Vector{Int64}(undef,size(design,1))
134
135     for d in 1:size(design,1)
136         n = design[d,1][1]
137         p = design[d,1][2]
138         q = design[d,1][3]
139         Kfalse = design[d,1][4]
140         K = design[d,1][5]
141         kappa0 = design[d,1][6]
142
143         A = reshape(Diagonal(ones(q)),q,q);
144         L_str = kron(A, ones(Int64(p/q)));
145         if q == 1
146             npar_D[d] = p*q+p*Kfalse+p+p+1
147         elseif q > 1
```

```
148             npar_D[d] = p*q-sum(count(x->x.!=0, L_str, dims=1))+p*Kfalse+p+p+(q*(q-1))+1
149         end
150     end
151     return npar_D
152 end
153 
154 
155 # Funzione per lo studio di simulazione
156 function Simulation(design, Y_sim)
157 
158     #Inizializzazione delle matrici dei risultati
159     F1score = Array{Float64}(undef,size(design,1),B); ACC = Array{Float64}(undef,size(design,1),B);
160     BIC = Array{Float64}(undef,size(design,1),B); ssBIC = Array{Float64}(undef,size(design,1),B); CLC = Array{
            Float64}(undef,size(design,1),B); ICLBIC = Array{Float64}(undef,size(design,1),B); CAIC = Array{
            Float64}(undef,size(design,1),B); LLIK = Array{Float64}(undef,size(design,1),B); AIC = Array{Float64}(
            undef,size(design,1),B); AICc = Array{Float64}(undef,size(design,1),B);
161     d1 = Array{Float64}(undef,size(design,1),B); d2 = Array{Float64}(undef,size(design,1),B); d3 = Array{
            Float64}(undef,size(design,1),B); d4 = Array{Float64}(undef,size(design,1),B); d5 = Array{Float64}(
            undef,size(design,1),B); d6 = Array{Float64}(undef,size(design,1),B);
162     Lambda_1_rmse = Array{Float64}(undef,size(design,1),1); Theta_delta_rmse = Array{Float64}(undef,size(design
            ,1),1); Lambda_2_rmse = Array{Float64}(undef,size(design,1),1); Psi_delta_rmse = Array{Float64}(undef,
            size(design,1),1); Phi_rmse = Array{Float64}(undef,size(design,1),1); kappa_rmse = Array{Float64}(
            undef,size(design,1),1);
163     RMSE = Array{Float64}(undef,5,size(design,1)); NORM = Array{Float64}(undef,5,size(design,1));
164     npar_D = Vector{Float64}(undef,size(design,1))
165     npar_D = nparameters(design);
166 
167 
168     for d in 1:size(design,1)
169         println(string("Design no.: ",d))
170         n = design[d,1][1]
171         p = design[d,1][2]
172         q = design[d,1][3]
173         Kfalse = design[d,1][4]
174         K = design[d,1][5]
175         kappa0 = design[d,1][6]
176 
177         res_B = Array{Float64}(undef,p*q+p*Kfalse+p+p+q^2+1+1+n,B)
178 
179         # Stima del modello tramite multi-thread e storing delle replicazioni dei disegni
180         @threads for b in 1:B println("Iter:",b); res_vec = Mix_CFA_EFA_suf(Y_sim[:,:,b,d],q,Kfalse); writedlm(
                string("C:/Users/NC/Desktop/res/backup/res_",d,"_",b,".csv"),res_vec); res_B[:,b] = res_vec end
181 
182 
183 
184         Lambda_1_B = Array{Float64}(undef,p,q); Lambda_2_B = Array{Float64}(undef,p,Kfalse); Theta_delta_B =
                Array{Float64}(undef,p,p); Psi_delta_B = Array{Float64}(undef,p,p); Phi_B = Array{Float64}(undef,q,
                q); kappa_B = Vector{Float64}(undef,1); Expllik_B = Vector{Float64}(undef,1);
185         Lambda_1_diff = Array{Float64}(undef,size(design,1),B); Lambda_2_diff=Array{Float64}(undef,size(design
                ,1),B);Theta_delta_diff = Array{Float64}(undef,size(design,1),B); Psi_delta_diff = Array{Float64}(
                undef,size(design,1),B); Phi_diff = Array{Float64}(undef,size(design,1),B); kappa_diff = Array{
                Float64}(undef,size(design,1),B);
186 
187 
188         for b in 1:B
189 
190             # Ricostruzione dei parametri stimati
191             Lambda_1_B = reshape(res_B[1:(p*q),b],p,q)
192             Lambda_2_B = reshape(res_B[(p*q+1):((p*q)+(p*Kfalse)),b],p,Kfalse)
193             Theta_delta_B = Diagonal(res_B[((p*q)+(p*Kfalse)+1):((p*q)+(p*Kfalse)+p),b])
194             Psi_delta_B = Diagonal(res_B[((p*q)+(p*Kfalse)+p+1):((p*q)+(p*Kfalse)+p+p),b])
195             Phi_B = reshape(res_B[((p*q)+(p*Kfalse)+p+p+1):((p*q)+(p*Kfalse)+p+p+q^2),b],q,q)
196             kappa_B = reduce(vcat,res_B[((p*q)+(p*Kfalse)+p+p+q^2+1):((p*q)+(p*Kfalse)+p+p+q^2+1),b])
197             Expllik_B = reduce(vcat,res_B[((p*q)+(p*Kfalse)+p+p+q^2+1+1):((p*q)+(p*Kfalse)+p+p+q^2+1+1),b])
198             Ez1 = reduce(vcat,res_B[((p*q)+(p*Kfalse)+p+p+q^2+1+1+1):((p*q)+(p*Kfalse)+p+p+q^2+1+1+n),b])
199             Ez1=round.(Int64,Ez1)
200 
201             # Calcolo della statistica Ek per indici CLC e ICL-BIC
202             b1 = Phi_B*Lambda_1_B'*inv(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B);
203             b2 = Lambda_2_B'*inv(Lambda_2_B*Lambda_2_B' + Psi_delta_B);
204 
205             Y = Y_sim[:,:,b,d]
206             P1 = zeros(n)
207             for i=1:n
208                 P1[i] = ((kappa_B .* ((exp((-1/2) *Y[i,:]'*inv(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)*Y[i
```

```
                ,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)))))/ ((kappa_B
                .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)*Y[i,:]))/(sqrt((2
                 .* pi)^(p) *det(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B))))) .+((exp((-1/2) * Y[i,:]'*
                inv(Lambda_2_B*Lambda_2_B'+Psi_delta_B)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2_B*
                Lambda_2_B'+Psi_delta_B)))) .-(kappa_B .*((exp((-1/2) * Y[i,:]'*inv(Lambda_2_B*Lambda_2_B'+
                Psi_delta_B)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2_B*Lambda_2_B'+Psi_delta_B)))))) *
                log(((kappa_B .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)*Y[i
                ,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)))))/((kappa_B .*
                 ((exp((-1/2) * Y[i,:]'*inv(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)*Y[i,:]))/(sqrt((2
                .* pi)^(p) *det(Lambda_1_B*Phi_B*Lambda_1_B'+Theta_delta_B)))) .+((exp((-1/2) * Y[i,:]'*
                inv(Lambda_2_B*Lambda_2_B'+Psi_delta_B)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2_B*
                Lambda_2_B'+Psi_delta_B)))) .-(kappa_B .*((exp((-1/2) * Y[i,:]'*inv(Lambda_2_B*Lambda_2_B'+
                Psi_delta_B)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2_B*Lambda_2_B'+Psi_delta_B))))))))
209         end
210         P0 = zeros(n)
211         for i=1:n
212             P0[i] = (1-P1[i])*log(1-P1[i])
213         end
214         Ek = - (sum(P1) + sum(P0))
215
216
217         # PA-mod
218         d1[d,b] = norm(Lambda_1_B-reshape(TruePar[1:(p*q)],p,q))^2 / norm(reshape(TruePar[1:(p*q)],p,q))^2;
219         d2[d,b] = norm(Theta_delta_B-Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+1):((p*q)+(p*K)+p)])))^2 /
                norm(Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+1):((p*q)+(p*K)+p)])))^2;
220         d3[d,b] = norm(Phi_B-reshape(TruePar[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q))^2 / norm(
                reshape(TruePar[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q))^2;
221         d4[d,b] = norm(Lambda_1_B-reshape(TruePar[(p*q+1):((p*q)+(p*K))],p,K))^2 / norm(reshape(TruePar[(p*q
                +1):((p*q)+(p*K))],p,K))^2;
222         d5[d,b] = norm(Psi_delta_B- Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+p+1):((p*q)+(p*K)+p+p)])))
                ^2 / norm( Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+p+1):((p*q)+(p*K)+p+p)])))^2;
223         d6[d,b] = (kappa_B.-reduce(vcat,TruePar[((p*q)+(p*K)+p+p+q^2+1):((p*q)+(p*K)+p+p+q^2+1)])).^2 / (
                reduce(vcat,TruePar[((p*q)+(p*K)+p+p+q^2+1):((p*q)+(p*K)+p+p+q^2+1)])).^2;
224
225
226         # RMSE
227         Lambda_1_diff[d,b] = sqrt(mean((Lambda_1_B - reshape(TruePar[1:(p*q)],p,q))*(Lambda_1_B-reshape(
                TruePar[1:(p*q)],p,q))'))
228         Theta_delta_diff[d,b] = sqrt(mean((Theta_delta_B -Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+1):((
                p*q)+(p*K)+p)])))*(Theta_delta_B -Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+1):((p*q)+(p*K)+
                p)])))'))
229         Lambda_2_diff[d,b] = sqrt(mean((Lambda_2_B - reshape(TruePar[(p*q+1):((p*q)+(p*K))],p,K))*(
                Lambda_2_B - reshape(TruePar[(p*q+1):((p*q)+(p*K))],p,K))'))
230         Psi_delta_diff[d,b] = sqrt(mean((Psi_delta_B - Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+p+1):((p
                *q)+(p*K)+p+p)])))*(Psi_delta_B - Diagonal(Array{Float64}(TruePar[((p*q)+(p*K)+p+1):((p*q)+(p*K
                )+p+p)])))'))
231         Phi_diff[d,b] = sqrt(mean((Phi_B -reshape(TruePar[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q))*(
                Phi_B -reshape(TruePar[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q))'))
232         kappa_diff[d,b] =sqrt(mean((kappa_B -reduce(vcat,TruePar[((p*q)+(p*K)+p+p+q^2+1):((p*q)+(p*K)+p+p+q
                ^2+1)])).^2))
233
234
235         # Indici di fit
236         LLIK[d,b] = Expllik_B
237         AICc[d,b] = 2*npar_D[d].-2*Expllik_B .+ 2*(npar_D[d]*(npar_D[d]+1))/(n-npar_D[d]-1)
238         AIC[d,b] = 2*npar_D[d].-2*Expllik_B
239         CAIC[d,b] = 2*npar_D[d]*(log(n)+1).-2*Expllik_B #jedidi 1997
240         BIC[d,b] = -2*Expllik_B.+2*npar_D[d]*log(n)
241         ssBIC[d,b] = -2*Expllik_B.+log((n+2)/24)*npar_D[d] #Detecting Mixtures From Structural Model
242         CLC[d,b] = -2*Expllik_B+2*Ek #Detecting Mixtures From Structural Model
243         ICLBIC[d,b] = -2*Expllik_B+log(n)*npar_D[d]+2*Ek #Detecting Mixtures From Structural Model
244
245
246         # Indici di classificazione
247
248         #Accuratezza
249         ACC[d,b] = sum(Ez1.== TruePar[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)])/n
250
251         #F1-score
252         TruePositives = sum(Ez1.== 1 .&& TruePar[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)].==
                1)
253         FalseNegatives = sum(Ez1.!= 1 .&& TruePar[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)]
                .== 1)
254         FalsePositives = sum(Ez1.== 1 .&& TruePar[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)]
                .!= 1)
```

```
            TrueNegatives = sum(Ez1.!= 1 .&& TruePar[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)]
                 .!= 1)
            Precision = TruePositives/(TruePositives+FalsePositives)
            Recall = TruePositives/(TruePositives+FalseNegatives)
            F1score[d,b] = 2*((Precision*Recall)/(Precision+Recall))
        end
    end

    RMSE = vcat(Lambda_1_rmse,Theta_delta_rmse,Psi_delta_rmse,Phi_rmse,kappa_rmse); NORM = vcat(vec(d1),vec(d2)
         ,vec(d3), vec(d4),vec(d5),vec(d6));
    writedlm("C:/Users/NC/Desktop/res/RMSE.csv", RMSE); writedlm("C:/Users/NC/Desktop/res/NORM.csv", NORM);
    INDEX = vcat(vec(LLIK),vec(AIC),vec(AICc),vec(CAIC),vec(BIC),vec(ssBIC),vec(CLC),vec(ICLBIC));
    writedlm("C:/Users/NC/Desktop/res/INDEX.csv", INDEX)
end
```

A.3 Codice dell'applicazione e relative funzioni

A.3.1 Codice dell'applicazione

```
# Trattamento dati e analisi preliminari
# Caricamento librerie
using StatsKit, Random, SpecialFunctions, Roots, QuadGK, LinearAlgebra, Distributions, Formatting, HDF5, JLD,
     DelimitedFiles, Latexify, Compose, Gadfly;

#Caricamento funzioni
include("C:/Users/NC/MEGA/Tesi/functions_application_utils.jl");
include("C:/Users/NC/MEGA/Tesi/functions_application_Mix-CFA-EFA.jl");
include("C:/Users/NC/MEGA/Tesi/functions_application_CFA.jl");

# Caricamento dati
Y = CSV.read("C:/Users/NC/MEGA/Tesi/Dati_applicazione/hexaco/data/ccases.csv", DataFrame)
Y_60 =Y[:,Cols(:hexaco62,:hexaco2,:hexaco8,:hexaco20,:hexaco68,:hexaco80,:hexaco134,:hexaco146,:hexaco164,:
     hexaco170,:hexaco69,:hexaco3,:hexaco33,:hexaco57,:hexaco81,:hexaco99,:hexaco111,:hexaco135,:hexaco165,:
     hexaco171,:hexaco52,:hexaco4,:hexaco16,:hexaco46,:hexaco82,:hexaco94,:hexaco112,:hexaco130,:hexaco154,:
     hexaco160,:hexaco138,:hexaco12,:hexaco6,:hexaco48,:hexaco78,:hexaco120,:hexaco162,:hexaco180,:hexaco186,:
     hexaco192,:hexaco53,:hexaco35,:hexaco41,:hexaco47,:hexaco65,:hexaco71,:hexaco89,:hexaco131,:hexaco173,:
     hexaco191,:hexaco73,:hexaco37,:hexaco7,:hexaco97,:hexaco121,:hexaco127,:hexaco139,:hexaco157,:hexaco175,:
     hexaco187)];
Y_60 = Matrix{Float64}(Y_60);

# Inversione di item
reversed_ind = [1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 1,0, 0, 0, 1, 1, 1, 1, 0, 1, 1,
      0, 0, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 0, 0];
reversed = [1, 1, 1, 1, 1, 6, 7, 1, 9, 10, 1, 1, 13, 1, 1, 16, 1, 1, 19, 20, 1, 1, 23, 24, 25, 1, 1, 1, 30,
     1, 1, 33, 34, 1, 1, 1, 38, 39, 40, 1, 1, 43, 44, 1, 46, 1, 48, 49, 50, 1, 1, 53, 54, 1, 1, 57, 58, 59,
     60];
for r in 1:size(Y_60,2)
    if reversed_ind[r] == 0;
        Y_60[:,reversed[r]] = 6 .-Y_60[:, reversed[r]];
    end
end


# Aggiunta di rumore bianco ai dati per motivi di convergenza
for i in 1:size(Y_60,1)
    Y_60[i,:] = Y_60[i,:] + rand(Uniform(-0.02,0.02),size(Y_60,2));
end

# Standardizzazione dei dati
for i in 1:size(Y_60,2)
    Y_60[:,i]=(Y_60[:,i].-mean(Y_60,dims=1)[i])/std(Y_60,dims=1)[i];
end



# Analisi dei dati

```

```
# Inizializzazione del modello CFA con q=6 per starting points razionali
q = 6;
n = size(Y_60,1);
p = size(Y_60,2);
res_60_CFA = Array{Float64}(undef,p*q+p+q^2);


# Adattamento del modello CFA
res_60_CFA = CFA_suf(Y_60,6,nothing,true);


# Ricomposizione delle matrici dei parametri stimate
Lambda_1 = reshape(res_60_CFA[1:(p*q)],p,q);
Theta_delta = Diagonal(res_60_CFA[((p*q)+1):((p*q)+p)]);
Phi = reshape(res_60_CFA[((p*q)+p+1):((p*q)+p+q^2)],q,q);


# Calcolo multi-thread del modello Mix-CFA-EFA da K=1 a K=7
@threads for K in 1:7
            println("K: ", K);
            Lambda_1 = reshape(res_60_CFA[1:(p*q)],p,q);
            Theta_delta = Diagonal(res_60_CFA[((p*q)+1):((p*q)+p)]);
            Phi = reshape(res_60_CFA[((p*q)+p+1):((p*q)+p+q^2)],q,q);
            res_60 = Array{Float64}(undef,p*q+p*K+p+p+q^2+1+1+n);
            res_60 = Mix_CFA_EFA_suf_appl(Y_60,q,K,nothing,true, Lambda_1, Theta_delta, Phi); writedlm(string("
                 C:/Users/NC/Desktop/res/appl_60/res_60_",K,".csv"),res_60);
end

# Calcolo degli indici di fit
idxs = Array{Float64}(undef,7,8);
for K in 1:7
    idxs[K,:] = model_selection(60,6,K,Y_60);
end
indx_names = ["LLIK" "AIC" "AICc" "CAIC" "BIC" "ssBIC" "CLC" "ICLBIC"];
d=Formatting.format.(idxs,precision=3);
da = DataFrame(reshape(d,7,8),vec(Symbol.(indx_names)));
latexify(da,env=:table) |> print # tabella



# Risultati del modello con K=4
for i in 1:60
    Lambda_1[i,:] = round.(Lambda_1[i,:],digits=3);
    Lambda_2[i,:] = round.(Lambda_2[i,:],digits=3);
    Theta_delta[i,i] = round.(Theta_delta[i,i],digits=3);
    Psi_delta[i,i] = round.(Psi_delta[i,i],digits=3);
end

latexify(hcat(Lambda_1,diag(Theta_delta),Lambda_2,diag(Psi_delta)),env=:table, fmt=x->format(round(x,
     sigdigits=4))) |> print #
     tabella
latexify(Phi,env=:table, fmt=x->format(round(x, sigdigits=3))) |> print # tabella



# Grafici applicants, non-applicants, sotto-campione CFA e sotto-campione EFA

z1_veri = Y_mix.sample
z1_veri = replace(z1_veri, "applicant" => 0)
z1_veri = replace(z1_veri, "research" => 1)

Y_cc = Y_60[Ez1 .== 1] # sotto campione CFA
Y_misc =Y_60[Ez1 .!= 1] # sotto campione EFA
Y_1 = Y_60[z1_veri .== 1,:] # non-applicants
Y_0 = Y_60[z1_veri .! =1,:] # applicants

p1=Gadfly.spy(cor(Y_1),Guide.xlabel("Items"),Guide.ylabel("Items"),Guide.title("Non-applicants"),Theme(
     grid_color=colorant"white",bar_spacing=-0.5mm,key_label_font_size=23pt, key_title_font_size=25pt,
     minor_label_font_size=25pt, major_label_font_size=25pt),Guide.colorkey(title="\rho"))

p2=Gadfly.spy(cor(Y_0),Guide.xlabel("Items"),Guide.ylabel("Items"),Guide.title("Applicants"), Theme(grid_color
     =colorant"white",bar_spacing=-0.5mm,key_label_font_size=23pt, key_title_font_size=25pt,
     minor_label_font_size=25pt, major_label_font_size=25pt),Guide.colorkey(title="\rho"));

p3=Gadfly.spy(cor(Y_cc),Guide.xlabel("Items"),Guide.ylabel("Items"),Guide.title("Sotto-campione CFA"), Theme(
```

```
    grid_color=colorant"white",bar_spacing=-0.5mm,key_label_font_size=23pt, key_title_font_size=25pt,
    minor_label_font_size=25pt, major_label_font_size=25pt),Guide.colorkey(title="\rho"));

p4=Gadfly.spy(cor(Y_misc),Guide.xlabel("Items"),Guide.ylabel("Items"),Guide.title("Sotto-campione EFA",),
    Theme(grid_color=colorant"white",bar_spacing=-0.5mm,key_label_font_size=23pt, key_title_font_size=25pt,
    minor_label_font_size=25pt, major_label_font_size=25pt),Guide.colorkey(title="\rho"))

draw(PDF("C:/Users/NC/MEGA/Tesi/Tesi/fig5a.pdf", 10inch, 10inch), p1)
draw(PDF("C:/Users/NC/MEGA/Tesi/Tesi/fig5b.pdf", 10inch, 10inch), p2)
draw(PDF("C:/Users/NC/MEGA/Tesi/Tesi/fig5c.pdf", 10inch, 10inch), p3)
draw(PDF("C:/Users/NC/MEGA/Tesi/Tesi/fig5d.pdf", 10inch, 10inch), p4)
```

A.3.2 Funzione per il calcolo del modello Mix-CFA-EFA tramite statistiche sufficienti

```
function Mix_CFA_EFA_suf_appl(Y,q,K,L_str=nothing,lambdavari=nothing,Lambda_1=nothing,Theta_delta=nothing,Phi=
    nothing)
    local Lambda_2, Psi_delta, kappa, llik, t, Ez1


    # Criteri per inizializzare nuovamente l'algoritmo utilizzando nuovi starting points casuali

    diff_llik = 1; dtol = 0.05; re = 0; t = 1; err2 = 1; non_conv = true; sum_diff_tf = true; DomainError_warn
        = 1; retr = 0;
    while err2 == 1 || non_conv || sum_diff_tf || DomainError_warn == 1;

        p = size(Y,2);
        n = size(Y,1);


        # CFA
        # Inizializzazione della matrice coefficienti fattoriali e delle varianze-covarianze dei fattori

        if L_str === nothing && Lambda_1 === nothing
            A = reshape(Diagonal(ones(q)),q,q);
            L_str = kron(A, ones(Int64(p/q)));
            Lambda_1 = rand(Uniform(0,1),p,q) .* L_str;
            Phi = rand(LKJ(q,1));
        elseif !(L_str === nothing) && !(Lambda_1 === nothing) # Uso di starting points razionali per Lambda e
             Phi
            Lambda_1 = Lambda_1 .* L_str;
        elseif !(L_str === nothing) && Lambda_1 === nothing
            Lambda_1 = rand(Uniform(0,1),p,q) .* L_str;
            Phi = rand(LKJ(q,1));
        elseif L_str === nothing && !(Lambda_1 === nothing) # Uso di starting points razionali solo per Lambda
            A = reshape(Diagonal(ones(q)),q,q);
            L_str = kron(A, ones(Int64(p/q)));
            Lambda_1 = rand(Uniform(0,1),p,q) .* L_str;
            Phi = rand(LKJ(q,1));
        end


        # Inizializzaione della matrice degli errori con tre opzioni per gli starting points

        if lambdavari === nothing && Theta_delta === nothing
            Theta_delta = Diagonal(vec(1 .- sum(Lambda_1,dims=2).^2));
        elseif !(lambdavari === nothing) && Theta_delta === nothing
            Theta_delta = Diagonal(rand(Uniform(0.5,2),p,p));
        end


        # EFA
        # Inizializzazione della matrice coefficienti fattoriali

        Lambda_2 = rand(Uniform(0,1),p,K);


        # Inizializzaione della matrice degli errori

```

```
52          Psi_delta = Diagonal(rand(Uniform(0.5,2),p,p));
53
54
55          # Mistura
56          # Inizializzazione proporzione di mistura
57
58          kappa = 0.50;
59
60
61          # Inizializzazioni e impostazioni per il ciclo while
62
63          t = 1; maxIter = 3000; llik = zeros(Float64, 1, maxIter); oldllik = -Inf; err=zeros(Int64,1,maxIter);
                 err2 = 0; sum_diff = zeros(Float64,1,maxIter); diff_llik = 1; sum_diff_tf = true; DomainError_warn
                  = 0; Ez1=zeros(n);
64
65
66          # Ciclo while dell'algoritmo Expectation-Maximization
67
68          while err2 == 0 && diff_llik>dtol && t<=maxIter && sum_diff_tf && DomainError_warn == 0
69
70              # E-step
71
72              b1 = Phi*Lambda_1'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta);
73              B1 = Phi-Phi*Lambda_1'*inv(Theta_delta+Lambda_1*Phi*Lambda_1')*Lambda_1*Phi;
74              b2 = Lambda_2'*inv(Lambda_2*Lambda_2' + Psi_delta);
75              B2 = Diagonal(ones(K))-Lambda_2'*inv(Psi_delta+Lambda_2*Lambda_2')*Lambda_2;
76
77
78              # Valori attesi della variabile latente z condizionati ad y
79
80              Ez1 = zeros(n);
81              for i=1:n
82                  Ez1[i] = ((kappa .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(
                          sqrt((2 .* pi)^(p) *det(Lambda_1*Phi*Lambda_1'+Theta_delta)))))/((kappa .* ((exp((-1/2) *Y[
                          i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1*
                          Phi*Lambda_1'+Theta_delta))))).+((exp((-1/2) * Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y[
                          i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*Lambda_2'+Psi_delta)))) .-(kappa .*((exp((-1/2) *
                           Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*
                          Lambda_2'+Psi_delta)))));
83              end
84              Ez0 = 1 .- Ez1;
85
86
87              # Valori attesi delle statistiche sufficienti condizionati ad y
88
89              Ek1 = sum(Ez1);
90              Ek0 = sum(Ez0);
91              ESyy1 = zeros(p,p); ESyy0 = zeros(p,p);
92              for i in 1:n
93                  ESyy1 = ESyy1 + (Ez1[i]*Y[i,:]*Y[i,:]')/n;
94                  ESyy0 = ESyy0 + (Ez0[i]*Y[i,:]*Y[i,:]')/n;
95              end
96              ESyeta = ESyy1*b1';
97              ESyxi = ESyy0*b2';
98              ESetaeta = B1*(sum(Ez1)/n)+b1*ESyeta;
99              ESxixi = (sum(Ez0)/n)*(B2) + b2*ESyxi;
100
101
102             # M-step
103
104             # Calcolo stima di ML delle matrici dei coefficienti fattoriali
105
106             Lambda_1 = ESyeta*inv(ESetaeta) .* L_str;
107             Lambda_2 = ESyxi*inv(ESxixi);
108
109             # Calcolo stima di ML delle matrici degli errori
110
111             Theta_delta = Diagonal((ESyy1 - Lambda_1*ESyeta')/(Ek1/n));
112             Psi_delta = Diagonal((ESyy0 - Lambda_2*ESyxi')/(Ek0/n));
113
114
115             # Criterio per gestire eventuali valori negativi nella matrice Theta_delta
116
117             if sum(Theta_delta.<0) > 0
118                 println("Error Theta_delta");
```

```
119             DomainError_warn = 1;
120         else
121
122
123             # Calcolo stima di ML della matrice di correlazione dei fattori
124
125             Phi = ESetaeta/(Ek1/n);
126             Phi = sqrt(inv(Diagonal(Phi)))*Phi*sqrt(inv(Diagonal(Phi)));
127
128
129             # Calcolo stima di ML del parametro di mistura
130
131             kappa = Ek1/n;
132
133
134             #Calcolo della log-verosimiglianza completa
135
136             old_diff = diff_llik;
137             llik[t] = 0;
138             for i=1:n
139                 llik[t] = llik[t] +(Ez1[i]*(log(kappa)-(p/2)*log(2*pi)-(1/2)*log(det(Theta_delta))-(1/2)*(Y[
                        i,:]'*inv(Theta_delta)*Y[i,:]-2*Y[i,:]'*inv(Theta_delta)*Lambda_1*(b1*Y[i,:])+tr(
                        Lambda_1'*inv(Theta_delta)*Lambda_1*((Phi - b1*Lambda_1*Phi + b1*Y[i,:]*Y[i,:]'*b1'))))
                        -(q/2)*log(2*pi)-(1/2)*log(det(Phi))-(1/2)*tr(inv(Phi)*((Phi - b1*Lambda_1*Phi + b1*Y[i
                        ,:]*Y[i,:]'*b1'))))+     Ez0[i]*(log(1-kappa)-(p/2)*log(2*pi)-(1/2)*log(det(Psi_delta))
                        -(1/2)*(Y[i,:]'*inv(Psi_delta)*Y[i,:]-2*Y[i,:]'*inv(Psi_delta)*Lambda_2*(b2*Y[i,:])+tr(
                        Lambda_2'*inv(Psi_delta)*Lambda_2*((Diagonal(ones(K))-b2*Lambda_2 + b2*Y[i,:]*Y[i,:]'*
                        b2'))))-(K/2)*log(2*pi)-(1/2)*tr(Diagonal(ones(K)) - b2*Lambda_2 + b2*Y[i,:]*Y[i,:]'*b2
                        ')));
140             end
141
142
143             # Differenza tra log-verosimiglianza completa all'iterazione t-1 e t
144
145             diff_llik=abs(oldllik - llik[t]);
146
147
148             # Criterio di stop per evitare eccessivi incrementi nella differenza tra verosimiglianza
                    all'interazione t-1 e t
149
150             sum_diff[t] = diff_llik>old_diff ;
151             err[t] = oldllik>llik[t];
152             err2 = err[t];
153
154
155             println([t llik[t] abs(oldllik - llik[t]) err[t] kappa]);
156             oldllik = llik[t];
157
158             t+=1;
159
160             sum_diff_tf = sum(sum_diff) < 70;
161             end
162         end
163
164         sum_diff_tf = sum(sum_diff) >= 70;
165         non_conv = t>=maxIter;
166
167
168         # Criterio per gestire il superamento delle iterazioni massime provvedendo risultati pari a NaN
169
170         non_conv = t>=maxIter;
171         if non_conv && retr == 0
172             retr += 1; #non convergenza ammessa per un solo nuovo tentativo con starting points diversi
173         elseif non_conv && retr == 1
174             non_conv = false; err2 == 1; sum_diff_tf == false;
175             Lambda_1 = Array{Float64}(undef,p,q); fill!(Lambda_1,NaN); Lambda_2 = Array{Float64}(undef,p,K);
                    fill!(Lambda_2,NaN); Theta_delta = Array{Float64}(undef,p,p); fill!(Theta_delta,NaN);
                Psi_delta = Array{Float64}(undef,p,p); fill!(Psi_delta,NaN); Phi = Array{Float64}(undef,q,q
                ); fill!(Phi,NaN); kappa = Array{Float64}(undef,1,1); fill!(kappa,NaN); llik = Array{
                Float64}(undef,1,1); fill!(llik,NaN); Ez1 = Array{Float64}(undef, 1, n); fill!(Ez1,NaN);
176         return vcat(vec(Lambda_1), vec(Lambda_2), vec(diag(Theta_delta)), vec(diag(Psi_delta)), vec(Phi),
                kappa, llik, vec(Ez1))
177     end
178         re+=1;
179 end
```

```
    return vcat(vec(Lambda_1), vec(Lambda_2), vec(diag(Theta_delta)), vec(diag(Psi_delta)), vec(Phi), kappa,
        llik[t-1], Ez1);
end
```

A.3.3 Funzione per il calcolo del modello CFA tramite statistiche sufficienti

```
# Funzione CFA con statistiche sufficienti

function CFA_suf(Y,q,L_str=nothing,lambdavari=nothing)
local Lambda, Theta_delta, Phi, llik, t


    # Criteri per inizializzare nuovamente l'algoritmo utilizzando nuovi starting points casuali

    diff_llik = 1; dtol = 0.05; re = 0; t = 1; err2 = 1; non_conv = true; sum_diff_tf = true; DomainError_warn
        = 1; retr = 0;
    while err2 == 1 || non_conv || sum_diff_tf || DomainError_warn == 1
        p = size(Y,2);
        n = size(Y,1);


        # Inizializzazione della matrice coefficienti fattoriali e delle varianze-covarianze dei fattori

        if L_str === nothing # L_str indica i parametri fissi e liberi
           A = reshape(Diagonal(ones(q)),q,q);
           L_str = kron(A, ones(Int64(p/q)));
           Lambda = rand(Uniform(0,1),p,q) .* L_str;
           Phi = rand(LKJ(q,1));
        elseif !(L_str === nothing)
           Lambda = rand(Uniform(0,1),p,q) .* L_str;
           Phi = rand(LKJ(q,1));
        end


        # Inizializzaione della matrice degli errori con due opzioni per gli starting points

        if lambdavari === nothing
           Theta_delta = Diagonal(vec(1 .- sum(Lambda,dims=2).^2));
           elseif !(lambdavari === nothing)
           Theta_delta = Diagonal(rand(Uniform(0.5,2),p,p));
        end


        # Inizializzazione delle strutture dati per il ciclo while

        C_yy=zeros(p,p); C_yeta=zeros(p,q); C_etaeta=zeros(q,q); C_yet=zeros(p,q);


        # Impostazioni di implementazione
        t = 1; maxIter = 3000; llik = zeros(Float64, 1, maxIter); oldllik = -Inf; err=zeros(Int64,1,maxIter);
             err2 = 0; sum_diff = zeros(Float64,1,maxIter); diff_llik = 1; sum_diff_tf = true; DomainError_warn
              = 0;

        # Ciclo while per la stima dei parametri tramite Expectation-Maximization
        while err2 == 0 && diff_llik>dtol && t<=maxIter && sum_diff_tf && DomainError_warn == 0


            # E-step

            b1 = Phi*Lambda'*inv(Lambda*Phi*Lambda'+Theta_delta);
            B1 = Phi-Phi*Lambda'*inv(Theta_delta+Lambda*Phi*Lambda')*Lambda*Phi;

            # Calcolo dei valori attesi delle statistiche suffcienti dato y

            C_yy = Y'*Y/n;
            C_yeta = C_yy*b1';
            C_etaeta = b1*C_yeta+B1;

```

```
60
61            # M-step
62
63            # Calcolo della stima di ML della matrice dei coefficienti fattoriali
64
65            Lambda = C_yeta*inv(C_etaeta);
66
67
68            # Calcolo della stima di ML della matrice degli errori
69
70            Theta_delta = Diagonal(C_yy-C_yeta*inv(C_etaeta)*(C_yeta)');
71
72            # Calcolo della stima di ML della matrice di correlazione dei fattori
73
74            Phi = sqrt(inv(Diagonal(C_etaeta)))*C_etaeta*sqrt(inv(Diagonal(C_etaeta)));
75
76
77            # Criterio per gestire eventuali valori negativi nella matrice Theta_delta
78            if sum(Theta_delta.<0) > 0
79               println("Error Theta_delta");
80               DomainError_warn = 1;
81            else
82
83               # Calcolo della funzione di log-verosimiglianza completa
84               old_diff = diff_llik;
85               llik[t] = 0;
86               for i=1:n
87                   llik[t] = llik[t] + (-(p/2)*log(2*pi)-(1/2)*log(det(Theta_delta))-(1/2)*(Y[i,:]'*inv(
                          Theta_delta)*Y[i,:]-2*Y[i,:]'*inv(Theta_delta)*Lambda*(b1*Y[i,:])+tr(Lambda'*inv(
                          Theta_delta)*Lambda*((Phi - b1*Lambda*Phi + b1*Y[i,:]*Y[i,:]'*b1'))))-(q/2)*log(2*pi)
                          -(1/2)*log(det(Phi))-(1/2)*tr(inv(Phi)*((Phi - b1*Lambda*Phi + b1*Y[i,:]*Y[i,:]'*b1'))));
88                end
89
90                # Differenza tra log-verosimiglianza completa all'iterazione t-1 e t
91                diff_llik=abs(oldllik - llik[t]);
92
93                # Criterio di stop per evitare eccessivi incrementi nella differenza tra verosimiglianza
                       all'interazione t-1 e t
94                sum_diff[t] = diff_llik>old_diff;
95                err[t] = oldllik>llik[t];
96                err2 = err[t];
97
98                println([t llik[t] abs(oldllik - llik[t]) err[t]])
99                oldllik = llik[t];
100
101               t+=1;
102
103               sum_diff_tf = sum(sum_diff) < 70;
104           end
105           end
106
107           sum_diff_tf = sum(sum_diff) >= 70;
108
109           # Criterio per gestire il superamento delle iterazioni massime provvedendo risultati pari a NaN
110           non_conv = t>=maxIter;
111           if non_conv && retr == 0
112              retr += 1; #non convergenza ammessa per un solo nuovo tentativo con starting points diversi
113           elseif non_conv && retr == 1
114              non_conv = false; err2 == 1; sum_diff_tf == false;
115              Lambda = Array{Float64}(undef,p,q); fill!(Lambda,NaN); Theta_delta = Array{Float64}(undef,p,p);
                      fill!(Theta_delta,NaN); Phi = Array{Float64}(undef,q,q); fill!(Phi,NaN); llik = Array{
                      Float64}(undef,1,1); fill!(llik,NaN);
116              return vcat(vec(Lambda), vec(diag(Theta_delta)), vec(Phi), llik)
117           end
118           re+=1;
119       end
120    return vcat(vec(Lambda), vec(diag(Theta_delta)), vec(Phi), llik[t-1]);
121 end
```

A.3.4 Funzione per il calcolo degli indici di fit

```
# Funzione per la selezione del modello Mix-CFA-EFA tramite calcolo di indici di fit

function model_selection(p,q,K,Y)

    # Caricamento risulati del modello con K selezionato
    res = readdlm(string("C:/Users/NC/Desktop/res/appl_",p,"/res_",p,"_",K,".csv"));

    # Ricomposizioni dei risultati
    Lambda_1 = reshape(res[1:(p*q)],p,q);
    Lambda_2 = reshape(res[(p*q+1):((p*q)+(p*K))],p,K);
    Theta_delta = Diagonal(res[((p*q)+(p*K)+1):((p*q)+(p*K)+p)]);
    Psi_delta = Diagonal(res[((p*q)+(p*K)+p+1):((p*q)+(p*K)+p+p)]);
    Phi = reshape(res[((p*q)+(p*K)+p+p+1):((p*q)+(p*K)+p+p+q^2)],q,q);
    kappa = reduce(vcat,res[((p*q)+(p*K)+p+p+q^2+1):((p*q)+(p*K)+p+p+q^2+1)]);
    Expllik = reduce(vcat,res[((p*q)+(p*K)+p+p+q^2+1+1):((p*q)+(p*K)+p+p+q^2+1+1)]);
    Ez1 = reduce(vcat,res[((p*q)+(p*K)+p+p+q^2+1+1+1):((p*q)+(p*K)+p+p+q^2+1+1+n)]);

    # Calcolo del numero dei parametri
    npar = Vector{Int64}(undef,1);
    if q == 1
       npar = p*q+p*K+p+p+1;
    elseif q > 1
       npar = p*q - sum(count(x->x.!=0, Lambda_1, dims=1)) + p*K+p+p+(q*(q-1))+1;
    end

    # Inizializzazione dei vettori degli indici di fit
    BIC = Array{Float64}(undef,1); ssBIC = Array{Float64}(undef,1); CLC = Array{Float64}(undef,1); ICLBIC =
         Array{Float64}(undef,1); CAIC = Array{Float64}(undef,1); LLIK = Array{Float64}(undef,1); AIC = Array{
         Float64}(undef,1); AICc = Array{Float64}(undef,1);
    INDEX = Array{Float64}(undef,9);

    # Calcolo di Ek per gli indici CLC e ICL-BIC
    b1 = Phi*Lambda_1'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta);
    b2 = Lambda_2'*inv(Lambda_2*Lambda_2' + Psi_delta);


    P1 = zeros(n)
    for i=1:n
        P1[i] = ((kappa .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .*
             pi)^(p) *det(Lambda_1*Phi*Lambda_1'+Theta_delta)))))/((kappa .* ((exp((-1/2) * Y[i,:]'*inv(
             Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1*Phi*Lambda_1'+
             Theta_delta))))) .+((exp((-1/2) * Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y[i,:]))/(sqrt((2 .*
             pi)^(p) .*det(Lambda_2*Lambda_2'+Psi_delta)))) .-(kappa .*((exp((-1/2) * Y[i,:]'*inv(Lambda_2*
             Lambda_2'+Psi_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*Lambda_2'+Psi_delta)))))) * log
             (((kappa .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1*Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .* pi
             )^(p) *det(Lambda_1*Phi*Lambda_1'+Theta_delta)))))/((kappa .* ((exp((-1/2) * Y[i,:]'*inv(Lambda_1*
             Phi*Lambda_1'+Theta_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) *det(Lambda_1*Phi*Lambda_1'+Theta_delta)))
             ).+((exp((-1/2) * Y[i,:]'*inv(Lambda_2*Lambda_2'+Psi_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(
             Lambda_2*Lambda_2'+Psi_delta))))  .-(kappa .*((exp((-1/2) * Y[i,:]'*inv(Lambda_2*Lambda_2'+
             Psi_delta)*Y[i,:]))/(sqrt((2 .* pi)^(p) .*det(Lambda_2*Lambda_2'+Psi_delta)))))));
    end
    P0 = zeros(n);
    for i=1:n
        P0[i] = (1-P1[i])*log(1-P1[i]);
    end
    Ek = - (sum(P1) + sum(P0));


    # Calcolo degli indici

    LLIK = Expllik;
    AICc = 2*npar .-2*Expllik .+ 2*(npar *(npar +1))/(n-npar -1);
    AIC = 2*npar .-2*Expllik;
    CAIC = 2*npar *(log(n)+1).-2*Expllik;
    BIC = -2*Expllik.+2*npar *log(n);
    ssBIC = -2*Expllik.+log((n+2)/24)*npar;
    CLC = -2*Expllik+2*Ek;
    ICLBIC = -2*Expllik+log(n)*npar +2*Ek;

    INDEX = hcat(LLIK, AIC, AICc, CAIC, BIC, ssBIC, CLC, ICLBIC);
    return INDEX
```

```
59 end
```