Università degli studi di Padova

Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Corso di Laurea Triennale in

Scienze Psicologiche dello sviluppo, della personalità e delle relazioni interpersonali

RELAZIONE FINALE

# Separare la tipologia di errori in un modello CFA mediante mistura di covarianze d'errore

Relatore: Prof. Antonio Calcagnì

Laureando: Matteo Scortegagna
(Matricola N° 2056839)

Anno Accademico 2024/2025

# Abstract

Il trattamento dell'errore di misurazione rappresenta un problema comune a molte discipline. L'assenza di un modello adeguato a tale trattamento può condurre ad analisi dei dati con risultati inaccurati. Il presente elaborato si propone di fornire una panoramica generale dell'errore di misurazione all'interno dell'analisi fattoriale confermativa, evidenziando alcune delle sue basi teoriche oltre ai contesti in cui viene modellato, proponendo infine un modello, ancora in via di sviluppo, per trattare le problematiche che l'inadeguata rappresentazione dell'errore può indurre. Il modello in questione distingue due diverse tipologie di errore trattandole tramite mistura di covarianze.

# Indice

# Introduzione

L'obiettivo della presente tesi è quello di formulare un modello di analisi fattoriale confermativa che possa adeguatamente modellare le diverse fonti alla base dell'errore di misurazione. La scelta del framework dell'analisi fattroriale confermativa trova giustificazione nel fatto che essa rappresenti ancora un cardine della ricerca in psicologia, permettendo la creazione di nuovi test e il loro miglioramento continuo.

Ho cercato di scrivere questa tesi con un linguaggio accessibile, sopratutto per quanto riguarda il capitolo primo in cui vengono esposte le basi statistiche dell'analisi fattoriale confermativa e il suo funzionamento generale. Il secondo capitolo racchiude invece diverse tipologie di analisi degli errori, ponendo l'accento proprio sulle diverse tecniche utilizzate per migliorare i dati che analizziamo, non solo all'interno dell'ambito psicometrico ma anche in altre discipline di studio, come l'econometria che condivide con noi la necessità di avere dati empirici accurati. L'ultimo capitolo presenta invece il modello proposto, con tutte le sue caratteristiche, limitazioni e riflessioni. Questo tenta di rappresentare la matrice di covarianza degli errori scomponendola e analizzandola tramite mistura di covarianze, con l'obiettivo di distinguere gli errori in due categorie: *within-factor* e *across-factor*.

Il modello proposto è ancora in uno stato embrionale e, allo stato attuale, è ancora in fase di sviluppo, necessitando di ulteriori approfondimenti teorici.

# Capitolo 1

# Fondamenti teorici

## 1.1 Cos'è la CFA

Da sempre, in psicologia, risulta necessario disporre di dati concreti da poter considerare e valutare efficacemente, in modo tale da aiutare la comunità scientifica nella continua ricerca e sviluppo di strumenti sempre più all'avanguardia. Tuttavia, se parliamo di trasformare qualcosa di astratto, in qualcosa di misurabile e tangibile, la sfida può diventare ardua. L'attuale strumento per eccellenza che viene utilizzato per risolvere questo problema è proprio l'Analisi Fattoriale Confermativa (CFA).

Supponiamo di voler misurare la depressione di una persona. Come possiamo misurarla? Potremmo chiedere a questa di dirci quanto si sente "depressa" su una scala da 1 a 10, ma tale domanda presupporrebbe che la persona in questione sappia che cosa sia effettivamente la depressione. Inoltre la sua risposta sarebbe estremamente soggettiva, e basata sulle valutazioni che secondo lei sono necessarie per rispondere alla nostra domanda, le quali potrebbero essere incomplete o imprecise - ad esempio potrebbe ritenere che dormire un'ora in meno una notte sia sintomo di depressione. Inoltre, pur immaginando che questa persona sia ben informata e che conosca bene i sintomi della depressione, nemmeno la risposta finale sarebbe soddisfacente - supponiamo che abbia risposto "6 punti su 10" - in quanto, il punteggio da lei espresso cosa può significare effettivamente? E sopratutto, il "6" di questa persona ha lo stesso valore del "6" di qualcun altro che si sente allo stesso modo?

La strategia migliore risulta dunque quella di sfruttare delle variabili osservabili in maniera diretta, per andare a misurare quelle che vengono definite variabili latenti[1]. Nel-

[1] Variabili che non sono direttamente misurabili.

l'esempio precedente potremmo dire che, sfruttando un set di domande specifiche[2] (ovvero le nostre variabili osservabili) avremo la possibilità di studiare la nostra variabile latente (la depressione). Tutto questo si riflette in una serie di vantaggi decisivi, tra cui ad esempio la possibilità di indagare un costrutto in tutte le sue caratteristiche[3] tramite domande standardizzate, le quali potranno restituirci dei punteggi anch'essi standardizzati che faciliteranno il confronto tra diverse persone.[4]

### 1.1.1 Modello unidimensionale e multidimensionale

Ecco che dunque si parla di CFA, un modello psicometrico-statistico che si propone di spiegare la variabilità tra le variabili osservate $Y_1, \ldots, Y_p$ introducendo una collezione di variabili definite "latenti" $\eta_1, \ldots, \eta_Q$ e ritenute antecedenti rispetto alle variabili osservate. La figura 1.1 mostra un modello CFA unidimensionale, il quale è definito tale dalla presenza in una singola variabile latente (Q=1), stocasticamente antecedente alle quattro variabili osservabili $Y_1, \ldots, Y_4$.

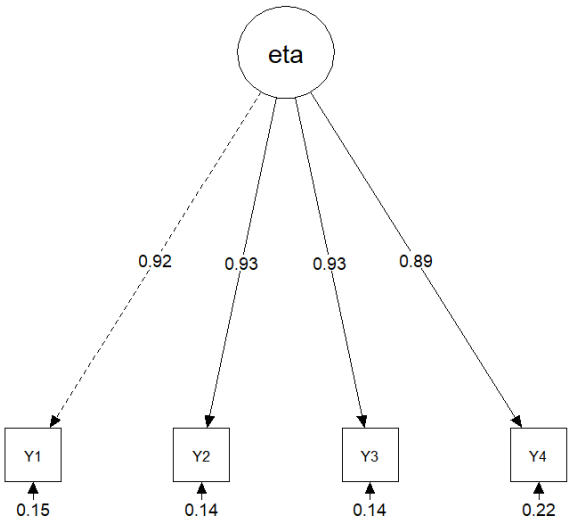



Figura 1.1: Modello unidimensionale

I valori presenti sulle frecce orientate nel grafico in figura, tra la variabile latente (definita anche *fattore*) e quelle osservabili (0.92,0.93,0.93,0.89), rappresentano i *factor-loadings*, ovvero quei numeri reali che quantificano il legame tra il fattore $\eta$ e le variabili

[2]Un altro modo di chiamare le domande presenti all'interno di un test psicometrico è tramite il termine *Item*, i quali riflettono indicatori concreti dei sintomi di un costrutto.

[3]Questa tipologia di validità viene definita *validità di contenuto*, la quale valuta quanto un test analizzi un costrutto in tutte le sue caratteristiche (ad esempio, un test che indaga la depressione dovrebbe includere item che riflettono diverse manifestazioni della depressione, come cambiamenti nel sonno, nell'umore, perdita di interesse, ecc..

[4]In psicologia questa viene definita proprio *validità di costrutto*. Un test possiede validità di costrutto se misura accuratamente il costrutto teorico non osservabile per cui è stato creato. Serve ad assicurarci che, ad esempio, un test per la depressione non stia misurando invece ansia o stress.

osservate. Il numero posto sotto ogni osservata viene invece definito $\delta$ e rappresenta la componente d'errore di ognuna.

Di fatto, la CFA, presuppone che ogni variabile osservata dipenda da un fattore $\eta$ e da un coefficiente reale $\lambda_j$ che quantifica il grado con cui la variabile latente è legata alla $j$-esima variabile osservata, a cui si aggiunge una componente di errore residuo $\delta_j$. Dato il nostro insieme di variabili osservabili $(Y_1, \ldots, Y_j, \ldots, Y_p)$ il modello CFA assume la seguente forma di equazioni lineari, tante quante sono il numero di variabili osservate:

$$
\begin{aligned}
Y_1 &= \tau_1 + \eta\lambda_1 + \delta_1 \\
&\vdots \\
Y_j &= \tau_j + \eta\lambda_j + \delta_j \\
&\vdots \\
Y_p &= \tau_p + \eta\lambda_p + \delta_p
\end{aligned}
$$

Qui sopra possiamo vedere $\tau$, il quale rappresenta l'intercetta del modello (potremmo dire che rappresenta la difficoltà dell'item).

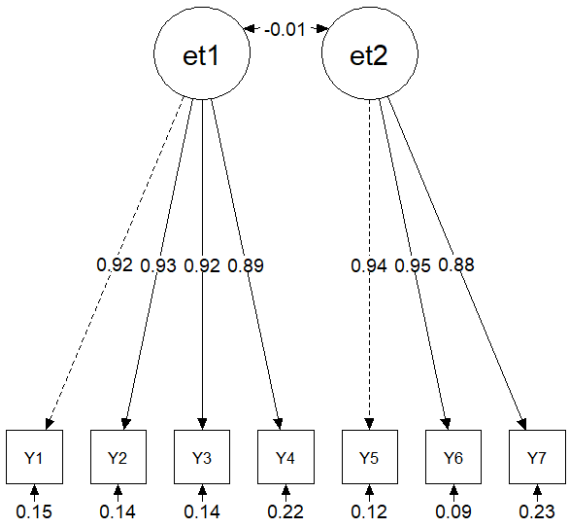



Figura 1.2: Modello multidimensionale

Nel caso riportato nella figura 1.2 parliamo invece di modello multidimensionale dal momento in cui abbiamo a che fare con un modello CFA composto da più di un fattore latente[5] (Q>1). In questo caso, rispetto l'equazione lineare mostrata nel caso del modello unidimensionale, si deve tenere conto di più di un fattore latente. La nuova equazione

[5]Importante ricordare che il vincolo per poter continuare con un modello fattoriale confermativo rimane q<p.

assume dunque la seguente forma:

$$Y_{1xp} = \tau_{1xp} + \eta_{1xQ}\Lambda^{T}_{pxQ} + \delta_{1xp}$$

In cui $Y, \tau, \eta, \delta$ risultano essere vettori, al contrario di $\Lambda$ che rappresenta la matrice contenente i *factor-loadings* del modello di riferimento. Vediamo nel caso a due fattori visto in figura 1.2 la forma matriciale completa della nostra equazione:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \tau_3 \\ \tau_4 \\ \tau_5 \\ \tau_6 \\ \tau_7 \end{bmatrix} + \begin{bmatrix} \lambda_{11} & 0 \\ \lambda_{21} & 0 \\ \lambda_{31} & 0 \\ \lambda_{41} & 0 \\ 0 & \lambda_{52} \\ 0 & \lambda_{62} \\ 0 & \lambda_{72} \end{bmatrix} \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} + \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \\ \delta_4 \\ \delta_5 \\ \delta_6 \\ \delta_7 \end{bmatrix}$$

$$\begin{aligned} Y_1 &= \tau_1 + \lambda_{11}\eta_1 + \delta_1 \\ &\vdots \\ Y_7 &= \tau_7 + \lambda_{72}\eta_2 + \delta_7 \end{aligned}$$

All'interno della matrice dei carichi fattoriali, notiamo alcuni zeri, i quali rappresentano la mancanza di una connessione tra il fattore latente e la variabile osservata. Notiamo infatti che nella prima colonna della matrice in questione sono rappresentate tutte le connessioni del primo fattore latente, il quale è collegato ai primi 4 item ma non agli ultimi 3. Queste ultime connessioni risultano infatti apparire con il valore 0, proprio per segnalare la mancanza di una connessione tra le due parti.

Da un punto di vista pratico, il modello di Analisi Fattoriale Confermativa in figura 1.2 può essere rappresentato come riporta la figura 1.3. La scala AB non è altro che un test psicometrico che contiene due scale diverse, la scala A e la scala B. Potremmo dire che la scala A è il risultato di 4 item (domande) diversi i quali potrebbero, teoricamente parlando, indagare un costrutto - per tornare all'esempio iniziale potremmo dire quello della depressione - mentre la scala B si propone di indagarne uno differente - supponiamo l'ansia - tramite 3 item [6]. Da notare che, nella figura in questione vi è una connessione

[6] Dal momento in cui stiamo parlando di modelli teorici, ci terrei a sottolineare che, da un punto di vista pratico, analizzare il costrutto di ansia e depressione con una scala del genere sarebbe sbagliato sotto ogni aspetto. Per maggiori informazioni potete fare riferimento a [3] [12]

anche tra $\eta_1$ e $\eta_2$ la quale indica una correlazione tra i due fattori (gli approfondimenti saranno presenti nel capitolo successivo). La CFA, parte dunque da un modello iniziale che, secondo noi, potrebbe spiegare la struttura latente del nostro dataset. Lo si cerca di adattare a quest'ultimo e tramite gli *indici di fit* possiamo capire se l'adattamento del modello è da considerare accettabile o meno.

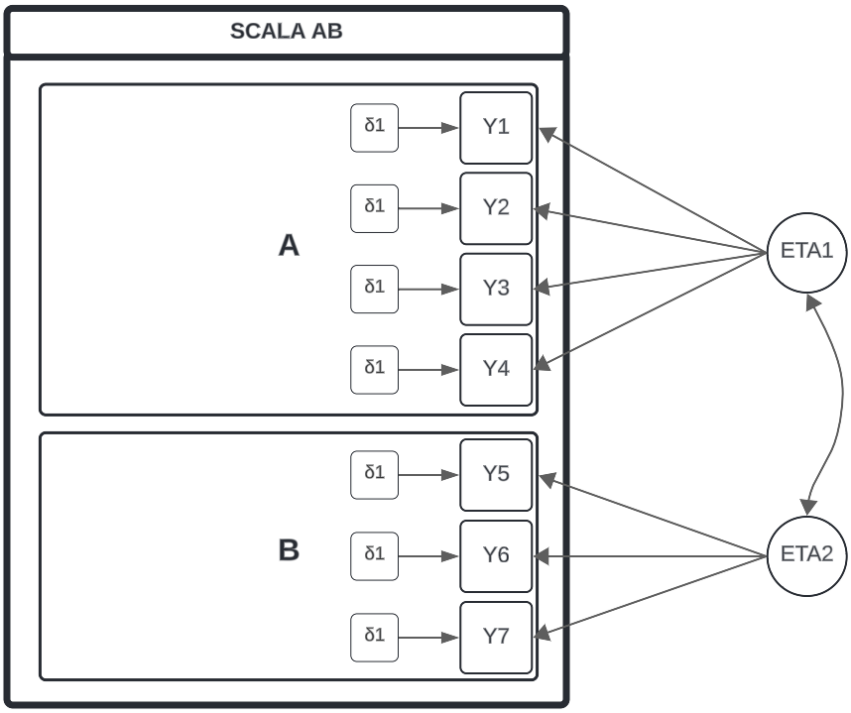



Figura 1.3: Nota: questa figura non è completa, essa rappresenta, in parte e in maniera teorica, quello che potrebbe essere un modello reale.

### 1.1.2 Struttura probabilistica del modello

La struttura probabilistica del modello, spiega come le variabili osservate Y dipendano in maniera probabilistica dai fattori latenti $\eta$ e dall'errore $\delta$. Nello specifico, partendo dalla formula vista in precedenza:

$$Y_{px1} = \tau_{px1} + \Lambda_{pxQ}\eta_{Qx1} + \delta_{px1}$$

Possiamo aggiungere che:

$$\eta_{Qx1} = (\eta_1, \ldots, \eta_Q)^T \sim N_Q(;\mu_\eta, \Phi)$$

Dove $\mu_\eta$ risulta essere il vettore delle medie latenti e $\Phi$ la matrice QxQ delle correlazioni tra le variabili latenti. Da sottolineare che, essendo il vettore colonna delle medie dei fat-

tori, di dimensione Qx1, questo risulterà ampio tanto quanto il numero di fattori presenti nel nostro modello. Avendo poi a che fare con la matrice $\Phi$, la quale rappresenta una matrice di correlazione, ci si aspetta che questa sia quadrata, simmetrica e che lungo la diagonale sia formata da soli uni. Nell'immagine 1.3 avevamo notato una connessione tra i due eta, la quale è visibile all'interno della matrice $\Phi$ del modello in questione. Oltre a questo, la struttura probabilistica del modello aggiunge:

$$\delta_{px1} = (\delta_1, \ldots, \delta_p)^T \sim N_p(;0_p, \Theta_\delta)$$

Dal momento in cui $|E[\delta] = 0_m$ ci si aspetta che l'errore presente all'interno del nostro modello sia di tipo casuale e non sistematico (gli approfondimenti relativi alla loro distinzione saranno presenti nel prossimo capitolo). La matrice $\Theta_\delta$ rappresenta una matrice di dimensione mxm delle covarianze (o correlazioni) tra gli errori delle misurazioni. Solitamente questa risulta essere una matrice diagonale, in quanto ci si aspetta che gli errori di un modello CFA non correlino tra loro (anche se non è sempre così). Un'ultima cosa da precisare prima di andare avanti con il successivo capitolo è che:

$$Y_{mx1} \sim N_m(;\mu_y, \Sigma_y)$$

In cui $\mu_Y$ rappresenta il vettore di dimensione mx1 delle medie riprodotte dal modello e $\Sigma_y$ la matrice di covarianza (o di correlazione) di dimensioni mxm che viene sempre riprodotta dal modello - potremmo definirla la matrice attesa, indotta dal modello - . Entrambe sono quantità teoriche indotte dal modello che noi proponiamo, non sono quantità osservate.

Possiamo dunque notare che le relazioni tra le variabili osservate $Y_1, \ldots, Y_p$ sono modellate in termini di $\Sigma_y$ che è scritta in funzione dei seguenti parametri:

$$\Sigma_y = \Lambda\Phi\Lambda^T + \Theta_\delta$$

$\Lambda_{pxQ}$ : matrice che contiene informazioni sulle connessioni tra variabili osservabili e fattori.

$\Phi_{QxQ}$ : matrice che contiene informazioni sulle relazioni tra le variabili latenti.

$\Theta_{\delta pxp}$ : matrice che contiene informazioni sulle relazioni tra gli errori di misura (solitamente diagonale).

## 1.2 L'errore di misurazione

Durante il XVIII e XIX secolo, grazie al contributo di alcuni studiosi come Carl Friedrich Gauss e Pierre Simon-Laplace, cominciò ad emergere il concetto di distribuzione degli errori casuali. Questo concetto diventa fondamentale, dal momento in cui la statistica diventa il mezzo necessario a valutare, comprendere e misurare l'errore nelle misurazioni. Successivamente, all'interno delle scienze esatte, il concetto di errore comincia a diventare un concetto portante. Si cominciò dunque a fare una prima distinzione tra errore casuale ed errore sistematico.

Un errore viene definito sistematico nel momento in cui influisce costantemente il punteggio derivante da una misurazione[7] muovendolo in una certa direzione. Ad esempio, nel nostro caso, un errore sistematico presente all'interno di un test potrebbe metodicamente sottovalutare o sopravvalutare ciò che il test si propone di misurare.

Un errore viene invece definito casuale quando il punteggio di una misurazione viene influenzato in maniera stocastica. Nel nostro caso un errore casuale potrebbe derivare da fattori situazionali come l'umore del soggetto, la stanchezza o la distrazione (assunti questi essere eventi non sistematici o strutturali al processo misurativo).

### 1.2.1 TCT ed EIV

Una delle teorie principali, a cui fare riferimento in psicologia, risulta essere la Teoria Classica dei Test (TCT) la quale ci suggerisce che, dato un insieme di variabili aleatorie (di cardinalità m) che codificano item a cui un insieme m di rispondenti reagisce, il risultato sarà $Y_j = T_j + E_j$; dove $Y_j$ è la j-esima variabile aleatoria che potremmo codificare come lo stimolo che vogliamo utilizzare. $T_j$ è la j-esima variabile aleatoria che rappresenta il misurando (quello che vogliamo misurare) ed $E_j$ che è la j-esima variabile aleatoria rappresentante l'errore casuale di misurazione.

Come mostra la figura 1.4 potremmo dire che, nel momento in cui sfruttiamo uno stimolo $Y_j$, otterremo sia ciò che volevamo effettivamente misurare con quello stimolo ($T_j$), ma anche una parte di errore casuale ($E_j$). Per tornare all'esempio iniziale diremmo che, sfruttando un test psicometrico che si propone di analizzare la depressione, all'interno dei dati, otterremo sia una parte rappresentante la depressione, che una parte di rumore (metafora comune per l'errore casuale). Da questa teoria nasce dunque la necessità di misurare, nella maniera più accurata possibile, l'errore in questione. Ecco che si cominciò a sviluppare il concetto di attendibilità.

[7] In psicologia la misurazione più comune avviene tramite la somministrazione di un test.

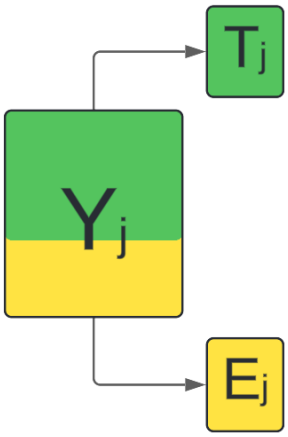



Figura 1.4: Teoria Classica dei Test

Date due misure parallele[8] $Y_j$ ed $Y_j^I$ l'attendibilità $\rho_{YT}$ viene definita come una quantità reale che esprime l'intensità con cui $Y_j$ quantifica $T$. La quantità $\rho_{YT} \in [0,1]$ è legata alla precisione con cui $Y_j$ esprime $T$:

$$\rho_{TY} = \frac{\mathbb{V}(T)}{\mathbb{V}(Y_j)}$$

In questa formula, $\mathbb{V}(T)$ rappresenta la varianza del punteggio vero, mentre $\mathbb{V}(Y_j)$ la varianza del punteggio osservato, che include sia la varianza dell'errore che quella del punteggio vero. Quando $\rho_{YT} = 1$, allora la coppia di misurazioni $Y_j$ ed $Y_j^I$ è massimamente attendibile nel quantificare $T$ (massimamente precisa).

Quindi per stimare la nostra attendibilità necessitiamo di misure parallele (caso poco frequente) che possiamo ottenere, mediante approccio statistico, agendo sul disegno di misurazione, ad esempio misurando $Y_j$ ed $Y_j^I$ sullo stesso campione ma in tempi diversi (metodo delle forme, metodo test-retest) oppure sullo stesso campione con strumenti equivalenti (split-half, alpha di cronbach [9][6]).

In sintesi, secondo la TCT, l'errore risulta essere una parte di dato inevitabile all'interno delle misurazioni. Tuttavia, il compito della psicometria sta proprio nel cercare di ridurlo il più possibile tramite l'utilizzo di tecniche statistiche e un'accurata progettazio-

[8] Due misure, sono definite parallele, nel momento in cui ci permettono di ottenere gli stessi risultati. Nel nostro caso potrebbero essere due test differenti che si propongono di indagare lo stesso costrutto, oppure lo stesso test ripetuto due volte (test-retest). Da due misure parallele ci si aspetta di ottenere la stessa media e stessa varianza. Per esclusione sappiamo che qualunque differenza trovata nei risultati tra la misurazione $Y_j$ e $Y_j^I$ sarà dovuta esclusivamente a errore casuale.

[9] L'alpha di Cronbach è un indice di misurazione dell'attendibilità. spesso definito di coerenza interna di un test in quanto misura quanto, gli item che compongono il test, misurino effettivamente lo stesso costrutto.

ne dei test, garantendo così che i dati ottenuti, riflettano in maniera attendibile il costrutto che si intende misurare.

Un'altra teoria interessante (usata principalmente in econometria) risulta essere quella relativa agli Errors-in-Variables models (EIV), la quale considera, al contrario della TCT, l'errore presente all'interno delle variabili che giocano il ruolo di predittori[10] che comportano a distorsioni e a conclusioni errate come componente da ridurre. La TCT è invece specifica per la psicologia e si focalizza nella stima della vera abilità o caratteristica latente di un individuo partendo dalla variabile osservata e considerandola come affetta da errore.

Nei modelli EIV l'errore è considerato come un problema da correggere in quanto comporterebbe mal interpretazioni delle relazioni tra variabili, mentre la TCT si occupa di quantificarlo. Entrambi riconoscono l'importanza dell'errore e il suo impatto sui dati, ma i modelli EIV si preoccupano della presenza di errori nelle variabili indipendenti all'interno di regressioni lineari o modelli econometrici più complessi. Dunque un modello di regressione assume che, se $X$ è la variabile indipendente e $Y$ è la variabile dipendente[11], allora la prima verrà correttamente misurata. I modelli Errors-in-Variables sostengono invece che $X$ venga osservata con errore e che dunque l'effettiva variabile $X^*$ non sia direttamente accessibile:

$$X = X^* + E_X$$

Dove $E_X$ rappresenta l'errore di misura nella variabile indipendente. Tutto questo comporta a stime distorte dei coefficienti e a varianza aumentata degli stimatori. Al di la dell'aspetto teorico di questo modello, una delle soluzioni proposte risulta essere interessante. Il metodo delle variabili strumentali (IV), ad esempio, cerca di identificare una variabile definita strumentale, che è correlata con quella indipendente $X^*$ ma non con l'errore di misurazione $E_X$ o quello del modello stesso. Di fatto la variabile strumentale risulta utile per isolare la parte di dato effettivo, non effetto da errore. Questa viene dunque sfruttata per stimare i parametri del modello, tuttavia trovare delle buone variabili strumentali è spesso complicato e il successo dipende in gran parte proprio da questo[15].

[10] Queste tipologie di variabili vengono definite anche variabili indipendenti e sono quelle che si suppone influenzino o siano correlate con le variabili di risposta

[11] Questa tipologia viene definita anche variabile di risposta e di fatto rappresenta il comportamento o il valore che si vuole spiegare o prevedere

### 1.2.2 Errori nella matrice di covarianza

Nel primo paragrafo abbiamo visto come un modello di analisi fattoriale confermativa riesca a legare le differenti variabili analizzando la struttura probabilistica dello stesso. Ma per andare avanti abbiamo bisogno di approfondire alcuni aspetti. In parte riassumendo e in parte aggiungendo, possiamo dire che:

1. Si propone un modello CFA che potrebbe spiegare la variabilità delle nostre variabili osservabili appartenenti a un dataset tramite l'aggiunta di una struttura latente.

2. Indipendentemente dai dati, otterremo la matrice di covarianza attesa $\Sigma_y$, che dunque il modello riproduce sempre, senza considerare i dati.

3. Successivamente, considerando i dati, si relaziona la matrice di covarianza osservata (che dunque considera i dati) alla matrice di covarianza attesa, mediante opportuna metrica (vale a dire d($S_y^{oss}, \Sigma_y$)).

4. Si sceglie poi, la matrice $\Sigma_y$ che meglio si avvicina alla matrice $S_y^{oss}$ grazie all'analisi di determinati indici di fit come RMSEA, SRMR, CFI, AIC etc.

Il problema cruciale però, risiede proprio nel momento in cui calcoliamo la nostra matrice di covarianza osservata. Questa infatti proviene dall'insieme di dati che stiamo analizzando i quali sappiamo essere affetti da errore. Come visto in precedenza, secondo la TCT, gli errori di misurazione sono infatti sempre presenti e questi possono derivare da diversi aspetti, tra cui, ad esempio, imprecisione degli strumenti di misura, variazioni individuali non legate alle variabili latenti ed errori casuali. Questi andranno poi a riflettersi sulla matrice di covarianza osservata comportando distorsioni nei risultati della CFA e dunque anche a possibili conclusioni errate[10]. Gli errori presenti nella matrice $S_y^{oss}$ portano ad un adattamento non accurato del modello proposto (rispetta la matrice $\Sigma_y$), il che comporta ad un'erronea interpretazione degli indici di fit (interpretati come troppo alti o troppo bassi) e alla conseguente mal identificazione delle relazioni tra variabili latenti.
Un'altra assunzione importante della CFA è che gli errori che distorcono la relazione tra variabili osservate e latenti siano di tipologia simmetrica[12] e distribuita casualmente. Tuttavia, ciò non è sempre vero. Gli errori presenti all'interno di tale modello, possono infatti essere sia di tipo casuale che sistematico. Il fatto che ci siano errori casuali all'interno di un modello non esclude la presenza di quelli sistematici e viceversa.

[12]Per errori simmetrici si intende errori che tendono a seguire una distribuzione simmetrica (come una Normale) e che dunque si distribuiscono attorno a un valore centrale, che spesso è lo zero. Di fatto sono interpretati come errori casuali e dunque non influenzino sistematicamente i dati, sottostimandoli o sovrastimandoli.

# Capitolo 2

# Soluzioni per il modellamento dell'errore di misura

Nel capitolo precedente è stata analizzata la storia e la teoria dell'errore di modo tale da poter fornire una succinta base circa l'importanza e le difficoltà riscontrate nel tentativo di riduzione, o anche solo di quantificazione, dell'errore di misurazione. In questo paragrafo andremo invece a vedere alcune tecniche utilizzate in differenti campi di studio che si propongono di trattarlo.

## 2.1 Stochastic Frontier Analysis

Un approccio utile, che si occupa della distinzione tra le diverse tipologie di rumore, risulta essere quello alla base della c.d. Stochastic Frontier Analysis[13] (SFA). L'approccio che essa adopera si focalizza sulla separazione delle diverse tipologie di errore, considerando l'errore sistematico come spostato verso una determinata direzione - ne potrebbero far parte alcuni bias o alcune inefficienze legate allo strumento di misura - e dunque considerato asimmetrico, al contrario di quello casuale che rimane simmetrico.
Immaginiamo di dover analizzare un dataset contentente bias di risposta[1], se non si riesce a distinguere tra errori casuali e bias sistematici, il fit del modello potrebbe distorcersi comportando conclusioni errate. Ecco che dunque, la tecnica in questione, cerca di distinguerli migliorando, seppur complessificando, l'attenibilità del modello.
Come suggerisce l'articolo di Aigner, Lovell e Schmidt (1977)[1], la teoria alla base è che vi siano due disturbi, con caratteristiche differenti (dunque più facilmente distinguibili) e

[1] Il bias di risposta è una distorsione sistematica che si verifica quando i rispondenti forniscono risposte non accurate o non rappresentative delle loro vere opinioni. Un esempio può essere la desiderabilità sociale.

che l'errore totale sia decomponibile in due componenti principali:

$$\varepsilon = v_i + \mu_i$$

Dove $\varepsilon$ rappresenta l'errore (totale), $v_i$ rappresenta la componente di rumore simmetrico e si presuppone sia distribuita in modo indipendente e identico, spesso come una distribuzione Normale $N(0, \sigma_i^2)$. Al contrario, $\mu_i$ rappresenta la parte di errore distribuita indipendentemente da $v_i$ e soddisfa la condizione $\mu_i \leq 0$. Questa componente non segue una distribuzione simmetrica, bensì asimmetrica. L'errore in questione viene definito $\mu_i \leq 0$ in quanto riflette un aumento nei costi di produzione previsti per la massima efficienza (è importante tenere a mente che stiamo parlando di modelli econometrici). L'errore casuale ha media 0 e non distorce sistematicamente (va dunque a riflettere fattori imprevedibili e neutri); l'errore sistematico influenza in maniera asimmetrica il risultato, rappresentando un errore non casuale e direzionale (che tende verso la perdita di efficienza).
In altre parole, si considera il primo come errore casuale e rappresenta dunque la parte di errore definita a variazioni casuali, come condizioni esterne imprevedibili, e il secondo come errore sistematico dovuto a inefficienze durante la valutazione (o produzione se parliamo dell'ambiente econometrico).

La SFA è un metodo econometrico utilizzato per stimare l'efficienza delle unità decisionali. La teoria di fondo suggerisce che i dati osservabili siano guidati non solo da un processo strutturale sottostante ma anche da disturbi casuali (in maniera simile alla CFA). Riconosce di fatto la presenza di molteplici fonti di variazione dei dati, tra cui fattori casuali e inefficienze. La funzione di produzione di frontiera rappresenta la quantità massima di output che può essere prodotta da un insieme di input. La distanza effettiva tra l'output osservato di un'impresa e la funzione di produzione di frontiera viene dunque utilizzata per stimare l'inefficienza dell'impresa. Dato il suo potenziale, la SFA viene utilizzata in molteplici campi, che spaziano dall'agricoltura ai trasporti ed è tipicamente applicata per stimare l'inefficienza delle unità decisionali. I risultati della stessa vengono sfruttati per identificare le aree in cui è possibile migliorare (aree in cui si è ottenuta una eccessiva distanza tra valore ottimale calcolato tramite SFA e valore reale). Il punto forte della SFA risiede nella capacità di calcolare la massima efficienza ottenibile e dunque, per esclusione, di arrivare a capire quali sono le fonti di errore che non permettono il raggiungimento di tale traguardo.

Gli attuali sviluppi di questa analisi permettono di rilevare: nuove distribuzioni per l'inefficienza, per i disturbi casuali e di considerare l'errore come più complesso e composto da diverse fonti. Tuttavia, una delle questioni cruciali dell'efficienza è il controllo di

altre fonti di eterogeneità non osservata. L'inefficienza è una componente non osservata dell'errore ed è dunque necessario controllarne gli effetti individuali fissi[2]. Per migliorare questo problema è importante modellare la considerazione di nuove distribuzioni sia per l'inefficienza che per l'errore casuale. Come sostiene l'articolo sopra citato, la scelta di distribuzioni più flessibili può portare a una migliore rappresentazione della variazione reale dei dati e, quindi, a stime di inefficienza più accurate.
Oltre a questo, per ottimizzare l'errore di misurazione risulta necessario distinguere tra inefficienza produttiva ed altri disturbi che sono fuori dal controllo dell'impresa. Ad esempio, un raccolto agricolo ridotto a causa di intemperie presentatesi durante il periodo di crescita dello stesso, dovrebbe essere considerato sfortuna e non inefficienza dell'impresa. Tenere a mente questi fattori, fuori dal controllo dell'impresa, durante l'analisi può comportare a stime più precise della reale inefficienza. Infatti, pur avendo subito perdite per cause esterne, la stessa azienda agricola potrebbe essere considerata altamente performante - e con scarse perdite dovute all'inefficienza - in quanto si considerano anche i fattori relative alle perdite non volute. Una assunzione di base di questa analisi, sostiene che l'inefficienza produttiva sia distribuita in modo identico per ogni osservazione, il che potrebbe non essere sempre veritiero. L'articolo [9] discute dei modelli SFA come mezzo per affrontare questa limitazione consentendo alle variabili esogene[2] di influenzare il processo di inefficienza, permettendo di tener conto di alcuni aspetti dell'eterogeneità che verrebbero altrimenti attribuiti all'inefficienza.

In sintesi, attualmente esistono diverse strategie per ridurre al minimo l'impatto dell'errore nell'analisi stocastica di frontiera, modellando il processo di errore o separando l'inefficienza da altri disturbi. Attualmente l'SFA non può eliminare completamente l'errore, tuttavia seguendo i principi discussi precedentemente i ricercatori possono migliorare significativamente l'accuratezza nelle loro stime.
Sebbene le fonti analizzate non trattino direttamente l'utilizzo di tale tecnica nel campo della CFA, è bene considerare alcuni degli approcci proposti come potenzialmente utili anche nel nostro caso. La modellazione esplicita dell'errore, affrontare l'eterogeneità non osservata e considerare diversi modelli CFA che presentano differenti specifiche relative all'errore e alle combinazioni dei risultati, potrebbero migliorare la gestione degli errori anche nel nostro ambito.

[2]In econometria una variabile esogena rappresenta una variabile che influisce sull'equilibrio del modello ma non è influenzata dall'equilibrio stesso.

## 2.2 Proprietà di basso rango delle matrici unità x variabili

Un ulteriore caratteristica, sfruttata per semplificare l'analisi di un modello, può essere presa dalle proprietà di basso rango (c.d. low-rank properties)[14].

In algebra lineare, il rango, definisce alcune caratteristiche importanti di una matrice e altro non è che un numero intero non negativo associato alla matrice stessa, identificando il numero di componenti lineari indipendenti. Una matrice è definita di basso rango se il suo rango è inferiore alla sua dimensione totale. In maniera più formale diremmo che, se una matrice $A$ è di dimensione mxn (con m righe ed n colonne) il rango della matrice in questione equivarrà al numero massimo di vettori linearmente indipendenti[3].

Se dunque, la matrice $A$ fosse di dimensione 3x3 ci si aspetterebbe un rango massimo equivalente a 3 ($rank(A) = 3$) nel caso in cui abbia tutte le colonne indipendenti. Ma, se la matrice in questione dovesse presentare rango pari a 1 o 2, sarebbe per definizione una matrice di basso rango. Il che implica che una o più colonne possono essere espresse come combinazioni lineari delle altre. Ecco che quindi, in questa casistica, potremmo lavorare con una matrice di dimensionalità ridotta rispetto a quella iniziale ($A_{3x3}$) riuscendo comunque a ricavare le colonne (o righe) che sono dipendenti linearmente[4] dalle altre tramite queste ultime. Questo ultimo punto si riflette in un vantaggio notevole, in quanto definisce le c.d. low-rank properties. Nel contesto delle matrici di covarianza, in particolare, possiamo sfruttare queste proprietà per ridurre la dimensionalità delle variabili osservabili. Dal momento in cui le matrici di covarianza riflettono le relazioni tra le variabili osservabili, avere a che fare con una matrice in questione che presenta le caratteristiche del basso rango, significherebbe lavorare con una matrice che contiene variabili descrivibili in termini di altre oppure che esiste una dipendenza lineare tra loro. Se contiamo inoltre, il concetto di dipendenza lineare tra variabili osservate possiamo assumere che, all'interno di modelli latenti come lo è la CFA, vi siano dei fattori latenti che sappiamo essere in grado di spiegare la correlazione tra variabili osservabili che coincidono con le dipendenze lineari trovate.

Risulta dunque possibile osservare matrici di covarianze osservate con struttura a bas-

[3]Le colonne o le righe di una matrice, per definizione, sono linearmente indipendenti nel momento in cui il determinante è diverso da 0. Il determinante di una matrice quadrata esprime alcune caratteristiche geometriche ed algebriche della matrice.

[4]Il termine linearmente dipendenti suggerisce che questo vettore può essere ricostruito come una combinazione lineare degli altri vettori.

so rango, specialmente quando i dati mostrano alta collinearità[5] (dunque un gran numero di variabili sono potenzialmente ridondanti) oppure ci sono forti dipendenze latenti.
Da un punto di vista pratico, l'utilizzo di questa tecnica si riflette in:

1. Riduzione della complessità; come visto in precedenza è possibile ridurre il numero di variabili o fattori e riuscire ugualmente a spiegare la variabilità dei dati.

2. Stima più stabile; riducendo il numero di parametri si riduce anche la variabilità delle stime, riducendo l'influenza di rumore casuale.

3. Maggiore interpretabilità; avendo a che vedere con un minor numero di variabili ci possiamo concentrare sulle relazioni più rilevanti, riducendo la complessità e migliorando l'identificazione delle relazioni latenti.

Fondamentalmente, questa tecnica permette di ridurre (quando possibile) le dimensioni dei nostri dati, riflettendosi in una serie di vantaggi decisivi all'analisi dei dati, anche quando si tratta con matrici di covarianza.

Un'altra tecnica alla base dell'esistenza di proprietà di basso rango è la Principal Component Analysis (PCA). Anch'essa propone di ridurre la dimensionalità dei dati, mantenendo la maggior parte delle informazioni presenti. Nella prima parte della CFA può essere utilizzata (nello specifico all'interno dell'EFA[6]), sopratutto nel momento in cui si trattano dataset molto ampi e che racchiudono un numero di osservazioni importante, consentendo di visualizzare i dati all'interno di uno spazio a dimensione ridotta o migliorare le prestazioni di algoritmi di apprendimento. La PCA riduce la numerosità del dataset, di modo tale da renderlo più facilmente analizzabile durante l'analisi confermativa del modello, permettendo di spiegare la varianza totale dei dati tramite $k$ componenti principali, ritenendo che questi siano linearmente combinati con le variabili originali. In breve, questa si occupa di preservare la varianza dei dati senza però ipotizzare relazioni causali tra le variabili.

Concentrandoci nuovamente sulle low-rank properties, possiamo dire che queste facciano riferimento alla struttura intrinseca di una matrice, che può essere espressa grazie ad un numero ridotto di componenti principali. Da un punto di vista teorico queste proprietà possono essere molto utili, sia per quanto riguarda le interpretazioni dei dati che per la semplificazione degli stessi. Tuttavia, il processo matematico che effettivamente ci permette di suddividere tale matrice nelle sue componenti principali viene definito

[5] Per alta collinearità si intende una alta o perfetta correlazione tra due variabili o più (multicollinearità)

[6] EFA è l'acronimo di Analisi Fattoriale Esplorativa e risulta essere la prima parte di analisi dei dati nella CFA, utile per supporre una possibile struttura latente degli stessi tramite tecniche specifiche come il clustering gerarchico.

fattorizzazione. Le low-rank properties, assieme ad un'adeguata fattorizzazione, ci permettono di migliorare i nostri dati riducendone l'errore. Come riporta l'articolo proposto da Chelsea Zang del 2020 [5] partendo da una matrice di risposte $R$ di dimensioni $n$x$k$ (con $n$ rispondenti e $k$ variabili) possiamo sfruttare la decomposizione a basso rango per stimare i valori mancanti all'interno di un questionario. Questo comporta un miglioramento della qualità dei dati, mostrando valori maggiormente attendibili tramite il calcolo di due matrici $U$ e $V$ che rappresentano rispettivamente i punteggi fattoriali stimati dei rispondenti e i carichi fattoriali stimati, tali che $Z \approx UV^T$ (in cui $Z$ rappresenta un'approssimazione di $R$). Questo approccio ci permette di ottenere una matrice di dati completa nel caso in cui avessimo a che fare con una matrice che non lo è, sfruttando un metodo alternativo a quelli classici e più attendibile. Anche nel caso in cui non avessimo matrici iniziali con tale caratteristica, la decomposizione a basso rango risulterebbe ugualmente utile, permettendoci di compiere *denoising* (che consiste nella riduzione del rumore presente nei dati), dal momento in cui la decomposizione filtra le componenti con varianza bassa mantenendo quelle principali che sono maggiormente interpretabili dai fattori latenti.

Dunque, partendo da un dataset iniziale, a cui vogliamo applicare la CFA, possiamo pensare di utilizzare il metodo di completamento delle matrici visto in precedenza, trattando il dataset come se fosse la nostra matrice $R$ e calcolando la sua approssimazione $Z$ tramite i punteggi fattoriali stimati dei rispondenti $U$ e da un'approssimazione iniziale dei factor-loadings rappresentata da $V$. Tutto questo si rifletterebbe in una migliore qualità dei dati e precisione delle stime.

## 2.3 Blind Source Separation

La Blind Source Separation (BSS) è un approccio che si focalizza sulla separazione delle fonti di errore o di segnali sovrapposti quando le sorgenti non sono note a priori. Viene sfruttata principalmente nelle analisi dei sensori e nel trattamento di segnali misti (biologici, acustici o elettromagnetici) per separare le componenti nei dati.
Questa tecnica diventa rilevante nella misura in cui, il rumore, contribuisce all'insieme di dati osservati, distorcendo le misurazioni per differenti cause o componenti, ciascuna delle quali dovrebbe essere trattata separatamente. La BSS[4] sfrutta diverse tecniche (tra cui la PCA) per suddividere le tipologie di segnali senza conoscere a priori le caratteristiche delle loro sorgenti, ma la più utilizzata è l'Indipendent Component Analysis (ICA).

L'idea alla base della tecnica è che i dati osservati siano rappresentati da miscele lineari[7] di diverse fonti indipendenti e cerca dunque di invertire questo processo per risalire alle sorgenti che ne hanno dato origine. Il fulcro della tecnica risiede proprio nella massimizzazione dell'indipendenza statistica delle componenti osservate.
In termini matematici assume che;

$$x = As$$

Dove $s$ è il vettore colonna che raccoglie i segnali sorgente, $x$, in maniera simile, raccoglie i segnali osservati (n), mentre $A$ rappresenta una matrice quadrata ($A_{nxn}$) di miscelazione, la quale contiene i coefficienti di miscelazione. La presente matrice è sconosciuta e rappresenta come le sorgenti indipendenti si combinino per formare i dati osservati ($x$). Data la formula appena vista possiamo assumere che, ogni variabile osservata sia data dalla miscela delle diverse sorgenti indipendenti.
Dal momento in cui anche $s$ ci risulta sconosciuta (in quanto sappiamo rappresentare le sorgenti indipendenti che vogliamo stimare), l'ICA cerca di invertire il processo con l'intento di ottenere le differenti sorgenti indipendenti. Per fare ciò, necessitiamo di una matrice di separazione $B$ che permetta di ottenere $y$, definito come una stima del vettore $s$ dei segnali sorgenti:

$$y = Bx$$

Di fatto, con questa tecnica, è possibile ottenere una stima dei valori osservati che tuttavia non sempre combacia con gli stessi a causa di rumore nei dati o limitazioni nelle informazioni.

Tutto questo diventa possibile solamente sotto le seguenti assunzioni:

1. Indipendenza statistica; le componenti che vogliamo estrarre devono essere indipendenti, in termini statistici, l'una dall'altra.

2. Non gaussianità; l'idea è che i segnali affetti da errori siano meno gaussiani rispetto a quelli di partenza.

3. Miscele lineari; i dati osservati sono ricavati da miscele lineari delle sorgenti indipendenti.

4. Numero di componenti; il numero di sorgenti indipendenti deve essere inferiore o uguale al numero di osservazioni.

[7] Con il termine miscele lineari, si intende una combinazione di diverse sorgenti, ognuna delle quali contribuisce con una proporzione specifica.

Uno dei limiti principali è rappresentato proprio dall'assunzione spiegata al punto due. In alcuni casi potremmo ritrovarci ad avere a che fare con variabili che risultano comunque gaussiane pur essendo affette da errore, ed è proprio in questo caso che la tecnica dell'ICA non può essere utilizzata. Un altro problema risalta nel momento in cui utilizziamo dataset di grandi dimensioni, in cui l'utilizzo della tecnica computazionale rischia di diventare eccessivamente oneroso.

La presente esposizione evidenzia l'efficacia di questa tecnica anche nel nostro caso. Si potrebbe infatti utilizzare la stessa, per dividere sorgenti multiple di errore che si riflettono all'interno della matrice di covarianza, permettendo di capire meglio quali componenti la distorcono maggiormente e conseguentemente, di ridurre l'errore stesso migliorando il fit del modello; il tutto, senza avere informazioni a priori circa le loro cause.

# Capitolo 3

# Il modello proposto

Come già riferito nella parte introduttiva, l'attuale modello si trova ancora in uno stato iniziale. Lo studio dello stesso dovrà essere migliorato e successivamente testato dal momento in cui presenta ancora diversi limiti. Attualmente, questo progetto è ancora in fase di studio, ma mi auguro che la lettura di questo modello possa ispirare altri ad approfondirlo e migliorarlo.

## 3.1 Introduzione al modello

Il problema che più affligge l'analisi della matrice di covarianza $\Theta_\delta$ è che questa viene considerata dall'analisi fattoriale confermativa come matrice diagonale, con lo scopo di facilitarne l'interpretazione da parte del modello latente proposto. Come abbiamo ben visto nel capitolo 1, un'assunzione del genere esclude a priori che un qualsivoglia modello possa avere errori che correlino tra loro. Escludere a priori che due o più errori possano correlare tra loro può comportare risultati imprecisi e interpretazioni inesatte dei modelli, proprio a causa del forzato tentativo di adattare un modello a fattori latenti ai dati osservabili [7]. Al contempo, introducendo degli errori al di fuori della diagonale della suddetta matrice, assumiamo che il modello latente non sia in grado di spiegare una parte di variabilità solamente grazie a $\eta$ e la rimanente grazie a $\delta$, ma che dobbiamo necessariamente considerare altre fonti di errore che accomunano due o più variabili osservate, che sono ancora attive e che il nostro modello non riesce a considerare.

Il tentativo del modello proposto è dunque quello di andare a fornire un approccio basato su modelli, atto ad analizzare le diverse fonti di errore presenti in $\Theta_\delta$, per rendere gli elementi non più solo nulli o manualmente modificati, ma consapevoli degli errori di misurazione. Il modello tenta di distinguere due diversi tipi di errore, le diverse fonti che

li causano (legate alle difficoltà dell'item, variazioni tra misurazioni ripetute o altri fattori) e di modellare le situazioni in cui questi coesistono.

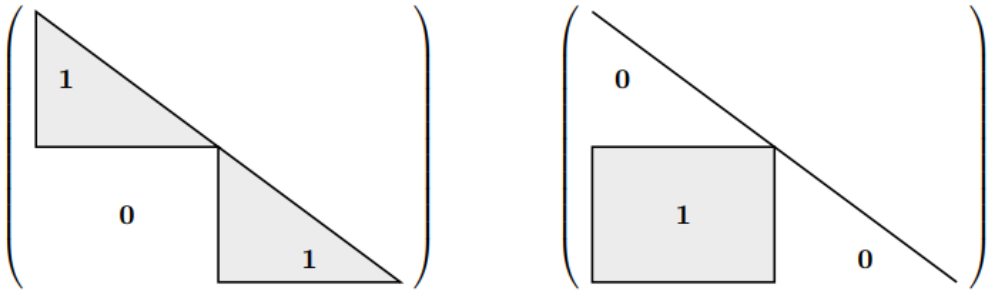
## 3.2 Descrizione del modello

Lo spazio dei dati del modello proposto è caratterizzato da variabili osservabili rappresentate come variabili casuali $y_j \in \mathbb{R}^J$ dove $J$ corrisponde al numero di indicatori. Secondo i modelli a struttura latente, come sottolineato in precedenza, le variabili osservabili sono affette da errori di misurazione additivi $\delta_i$ e a monte da variabili latenti $\eta_i \in \mathbb{R}^Q$, dove $Q$ rappresenta il numero di fattori latenti (con $Q < J$) che modellano le variabili sottostanti. La nostra matrice $\Lambda$ è la matrice che contiene le connessioni tra fattore latente e variabile osservata. Ogni elemento di questa matrice $\lambda_{mq}$ ci dice quanto una variabile osservata dipenda dal nostro fattore. Considerando una versione booleanizzata[1] della matrice in questione $\Lambda^*$=bool($\Lambda$) possiamo trasformare l'informazione quantitativa che questa mostra, in qualitativa. Diciamo dunque che se $\lambda_{mq} = 0$ allora $\lambda^*_{mq} = 0$ (non esiste nessuna connessione tra le due variabili osservate) e che se $\lambda_{mq} > 0$ allora $\lambda^*_{mq} = 1$ (esiste una connessione tra le due variabili osservate).

Per studiare meglio come le variabili osservate interagiscano tra di loro e non solo con i fattori, andiamo a creare due nuove matrici, partendo proprio da $\Lambda^*$. La prima è la matrice triangolare inferiore $H_1$, la quale rappresenta i percorsi diretti tra le variabili osservate e che è possibile calcolare tramite $H_1 = tril\triangle(\Lambda^*\Lambda^{*T})$. La seconda è detta $H_2$, è data da $H_2 = 1_{JxJ} - H_1$ e rappresenta i percorsi complementari (indiretti). Se un elemento $H_1 = 1$, allora lo stesso elemento sarà 0 in $H_2$. Queste due matrici si dimostrano utili nel rappresentare visivamente le strutture interne del modello come i percorsi *within-factor* (fig. 2.1-a) e *across-factor* (fig. 2.1-b) [8]. La prima tipologia di errori si verifica all'interno dello stesso fattore latente, questa può derivare da fonti di rumore specifiche accomunate dallo stesso fattore latente. Si riflettono in una parte di variabilità residua, non spiegata dal fattore latente. Ad esempio, un questionario che si propone di misurare ansia e depressione, sarebbe affetto da errori *within-factor* nel momento in cui, due domande, siano accomunate non solo dal fattore latente ansia, ma anche da qualcos'altro. Ad esempio, entrambe le domande potrebbero indagare stati fisiologici specifici dell'ansia[2]. Se si parla invece di errori *across-factor*, allora facciamo riferimento a correlazioni

[1] La booleanizzazione di una matrice serve a trasfomare una matrice di modo tale che contenga solamente due numeri: 0 e 1. Questa operazione viene utilizzata per semplificare l'interpretazione dei dati. I valori 0 o 1 vengono assegnati sulla base di determinate condizioni.

[2] Per approfondire questa nota potremmo dire che questo errore si verificherebbe nel momento in cui le due (o più) domande presenti nel questionario condividono caratteristiche specifiche relative al contesto o




*(a)* $\mathbf{H}_1$ matrix *(b)* $\mathbf{H}_2$ matrix

Figura 3.1: Rappresentazione grafica delle matrici $H_1$ e $H_2$

significative, non spiegate dal nostro modello, da due o più variabili che fanno parte di fattori latenti distinti. Nell'esempio precedente, si parlerebbe di questa tipologia di errore nel momento in cui, una domanda relativa all'ansia, fosse significativamente correlata con un'altra relativa alla depressione. Queste correlazioni possono essere dovute da similarità del contenuto e da sovrapposizione di sintomi tra i due fattori soggetti di indagine[3] .

Distinguere queste due tipologie di errore risulta un punto cruciale, in quanto ci permette di capire se l'eccessiva variabilità causata dall'errore, deriva da problematiche tra-fattore o intra-fattore, permettendoci dunque di modificare gli item del test in questione, affinché l'analisi dei dati subisca un miglioramento decisivo. Distinguere queste tipologie di errore, risulta di fatto un punto di riflessione particolarmente rilevante. Gli errori di tipo *within* possono indicare problemi specifici nel design delle domande del test soggetto di analisi, al contrario quelli di tipo *across* possono segnalare sovrapposizioni di tipo concettuali (ad esempio item che indagano più fattori). La matrice $H_1$ funge dunque da rete di analisi delle connessioni dirette tra variabili che fanno parte dello stesso fattore. Al contrario la matrice $H_2$, completa l'informazione fungendo da rete di salvataggio, mappando le connessioni complementari (tra diversi fattori). É vero dire che, $H_2$ risulta essere il contrario di $H_1$, ma considerarla superflua sarebbe un grave errore. Questa facilita la gestione delle varianze rimanenti tra variabili facenti parte di fattori latenti distinti e il modellamento separato delle due tipologie di errore.

Per modellare in maniera più efficace la nostra matrice degli errori $\Theta_\delta$ la si scompone

al linguaggio, ad esempio: "Mi sento nervoso senza motivo apparente" e "Ho difficoltà a rilassarmi durante la giornata". Se esiste una correlazione significativa tra le due domande abbiamo a che fare con un errore *within-factor*.

[3]Se uno degli item di indagine dell'ansia fosse "Mi sento sopraffatto da situazioni quotidiane" e uno relativo all'indagine della depressione fosse "Mi sento incapace di affrontare i miei problemi" e questi fossero significativamente correlati all'interno di $\Theta_\delta$ a causa della loro similarità nei sintomi tra ansia e depressione, potremmo dire di avere a che fare con un errore *across-factor*

in termini di varianze ($\Sigma_\delta$) e di correlazioni ($R$):

$$\begin{aligned}\Theta_\delta &= \Sigma_\delta R \Sigma_\delta \\ &= \Sigma_\delta (LL^T) \Sigma_\delta\end{aligned}$$

con $L$ che rappresenta il triangolo inferiore scomposto tramite Cholesky delle correlazioni $R$. Questo passaggio risulta utile per rappresentare in maniera più compatta le correlazioni tra errori, facilitando l'analisi e la correzione delle loro fonti, migliorando la robustezza del modello. Si noti in questo contesto la parametrizzazione sferica del fattore di Cholesky, in cui ogni elemento di $L$ viene parametrizzato in termini di angoli sferici ($\Theta_L$)[11], permettendo a $R$ di rimanere sempre positiva semi-definita, simmetrica e riducendo il numero di parametri liberi, dal momento in cui si lavora con più dimensioni:

$$\begin{aligned}\Theta_{l_{j1}} &= \arccos(l_{j1}) && j = 2, \ldots, J \\ \Theta_{l_{jh}} &= \left( \frac{\arccos(l_{jh})}{\prod_{k=1}^{h-1} \sin(\arccos(l_{jk}))} \right) && 2 \leq h < j \leq J\end{aligned}$$

Questo passaggio garantisce che $L$ sia sempre valida per generare una matrice $R$ simmetrica e positiva semi-definita. Poichè $\Theta_{l_{jh}} \in [0,1]$, una sua versione non vincolata $\tilde{\Theta}_L$ può essere calcolata sfruttando i logaritmi:

$$\tilde{\Theta}_L = log((\Theta_{l_{jh}})/(1-\Theta_{l_{jh}}))$$

La suddetta tecnica ci permette di descrivere le correlazioni senza vincoli, semplificando il lavoro matematico e lavorando su tutto il dominio dei numeri reali. Nel primo caso $f(\Theta_L, \Omega)$ viene modellata tramite una distribuzione di probabilità consueta, mentre nel secondo $f(\tilde{\Theta}_L, \Omega)$ tramite una distribuzione di probabilità reale simmetrica (ad esempio una normale o t). Lavorare con una versione non vincolata ci permette di modellare $\tilde{\Theta}_L$ tramite una distribuzione appropriata (e più semplice) per incorporare incertezze e ottimizzare i calcoli.

Si raggiunge alla fine una rappresentazione probabilistica della varianza d'errore, sia

flessibile che interpretabile nel contesto dei modelli lineari latenti:

$$Z_k \sim Bern(z;\pi_k) \tag{3.1}$$
$$vech(H_1 \circ \tilde{\Theta}_L)_k | Z_k = 0 \sim N_Q(\tilde{\Theta}_L;\mu_1,\Sigma_1) \tag{3.2}$$
$$vech(H_2 \circ \tilde{\Theta}_L)_k | Z_k = 1 \sim N_{Q-1}(\tilde{\Theta}_L;\mu_2,\Sigma_2) \tag{3.3}$$
$$\pi_k = logit^{-1}(x_k\beta) \tag{3.4}$$

$$k = 1,\ldots,K = 1_J^T H_{(.)} 1_J$$

in cui la distribuzione di probabilità del fattore di Cholesky sferico non vincolato è una miscela di distribuzioni Normali multivariate e ponderate da $\pi$. Un ulteriore aspetto da notare è che, il modello sopra riportato consente di districare due caratteristiche dell'errore di misurazione, ovvero il tipo di errore (equazione 3.2 e 3.3) e la fonte di errore tramite covariate esterne $x$ (equazione 3.4) come i tempi di risposta, le difficoltà dell'item e i metodi di misurazione. In questo modello vengono dunque considerati due tipi di errore, ovvero *within-factor* e *acros-factor*, anche se il caso limite di $\pi_k = 0.5$ consentirebbe di modellare quelle situazioni in cui entrambi i tipi di errore sono presenti.

L'elemento $Z_k$ è una variabile casuale Bernoulliana che determina a quale tipologia di errore appartiene l'elemento k ($Z_k = 0$ errore di tipo *within*, $Z_k = 1$ errore di tipo *across*). Mentre, l'elemento $\pi_k$, ci dice la probabilità con cui l'errore sia di tipo *across* ($\pi_k$) o *within* ($1 - \pi_k$). A seconda dell'errore ottenuto si sfrutterà l'equazione 3.2 o 3.3 che vettorizza (prendendo solo il triangolo inferiore) il prodotto riga colonna di $H_1$ o $H_2$ (a seconda dell'errore) con $\tilde{\Theta}_L$, semplificando il calcolo matematico e consentendo una modellazione chiara e flessibile che permetta la riduzione degli errori.
Detto ciò possiamo affermare che, lo spazio parametrico del modello proposto, presenta le seguenti caratteristiche:

1. Matrice dei coefficienti fattoriali; $\Lambda \in \mathbb{R}^{JxQ}$, i parametri di questa matrice definiscono la forza delle connessioni tra variabili osservabili e fattori. In alcuni casi questi sono fissati a zero per rappresentare l'assenza di connessioni e ridurre la complessità dello spazio parametrico.

2. Matrice di correlazione dei fattori; $\Phi \in \mathbb{R}^{QxQ}$, questa matrice rappresenta le correlazioni (e la varianza) tra fattori latenti. Deve essere simmetrica e positiva semidefinita e solitamente ci si aspetta che $diag(\Phi) = 1_Q$. Dunque è normalizzata per garantire coerenza delle stime delle varianze.

3. Matrice degli errori di misurazione; $\Theta_\delta \in \mathbb{R}^{JxJ}$, è una matrice che descrive le covarianze degli errori (e la varianza). Questa deve essere positiva semi-definita. La scomposizione utile a modellare errori correlati proposta dal modello $\Theta_\delta = \Sigma_\delta R \Sigma_\delta$, è formata dalle varianza d'errore $\Sigma_\delta$ e dalle correlazioni $R$. Per quanto riguarda $\Sigma_\delta$ i valori lungo la diagonale rappresentano le varianze degli errori, che devono essere positivi.

4. Probabilità di classificazione; $\pi$ rappresenta la soglia di probabilità che l'errore sia di tipo *within-* ($\pi < 0.5$) o *across-factor* ($\pi > 0.5$) o di entrambi i tipi ($\pi = 0.5$). Questa è legata alla parte di mistura del modello e risulta utile a suddividere gli errori in due categorie differenti.

Lo spazio parametrico del presente modello, necessita di tali vincoli per poter assicurare coerenza e validità matematica. La simmetria di $\Phi$ e $\Theta_\delta$, assieme ai loro autovalori positivi, garantiscono varianze positive e relazioni matematicamente coerenti.

## 3.3 Potenziali legami con modelli d'errore alternativi

Quando si cercano soluzioni comuni è bene trovare problemi comuni. La ricerca di connessioni tra il modello proposto e i tentativi di altre materie di studio nel migliorare le problematiche associate agli errori di misurazione, hanno portato a nuovi spunti di riflessione.

Quanto accennato in precedenza in merito alla SFA, viene riscontrato anche all'interno del modello proposto. Nella Stochastic Frontier Analysis si cerca di distinguere due diverse fonti di errore, quello simmetrico ($v_i$) e quello asimmetrico ($\mu_i$), in maniera simile, il modello scompone la matrice $\Theta_\delta$ in varianze ($\Sigma_\delta$) e correlazioni ($R$) per rappresentare l'entità degli errori e le relazioni tra variabili. Entrambi si propongono di isolare le diverse fonti di errore, con lo scopo di migliorare l'identificabilità dei dati, riducendone la distorsione delle stime. Oltre a ciò, la SFA considera il bias sistematico come una componente asimmetrica che influenza negativamente l'output. Analogamente il modello osservato, mira a riconoscere errori correlati nelle variabili osservabili e differenzia diverse tipologie di errore (*across-* e *within-factor*). Uno spunto interessante, potrebbe essere quello di integrare la separazione dell'inefficienza produttiva da quella per cause esterne, tipica della SFA, per migliorare la distinzione di errori interni da errori esterni nel modello oggetto di questa tesi.

Nelle low-rank properties viene sottolineato come la semplificazione di una matrice di correlazione, possa migliorare l'analisi dei dati riducendone la dimensionalità, migliorando il processo di denoising e rendendo i dati più interpretabili . Allo stesso modo, all'interno del modello, si procede con la scomposizione della matrice $\Theta_\delta$ tramite Cholesky per rappresentare correlazioni tra errori. Le proprietà di basso rango, cercano di identificare variabili ridondanti, mentre il modello proposto distingue tra due diversi tipi di errore. Il primo metodo è utile per mantenere le componenti principali dei dati, diminuendo la variabilità degli stessi. Applicare una tecnica simile al modello proposto, potrebbe migliorare la qualità dei dati osservati e delle stime. É importante non dimenticare che la tecnica di completamento delle matrici potrebbe essere sfruttata per trattare dati mancanti all'interno delle matrici $\Lambda$ e $\Theta_\delta$, pur necessitando di diverse modifiche.

La Blind Source Separation cerca di distinguere diversi segnali senza conoscere le loro fonti; in maniera molto simile noi cerchiamo di distinguere i due diversi errori proposti e la fonte che li causa. All'interno della BSS viene richiesto che il numero di sorgenti sia inferiore al numero di osservazioni, questo si riflette anche nel nostro caso in cui $Q < J$.

Sebbene nessuna di queste tecniche sia stata definita direttamente per modelli tipo CFA, ciascuna offre degli spunti interessanti al miglioramento del modello proposto:

- La SFA fornisce approcci alla modellizzazione degli errori, trattandoli singolarmente e distinguendoli sulla base delle loro caratteristiche.
- Le low-rank properties forniscono strumenti per ridurre la complessità delle matrici e quindi anche l'eccessiva varianza causata dall'errore
- La BSS identifica e separa sorgenti indipendenti di errore migliorando l'analisi dei dati

Questi vantaggi decisivi proposti dai precedenti metodi possono essere sfruttati per:

- Adottare distribuzioni flessibili per rappresentare al meglio gli errori più complessi.
- Sfruttare le tecniche di completamento delle matrici per risolvere problemi derivanti da dati mancanti.
- Integrare la separazione di sorgenti indipendenti per isolare errori correlati

Queste metodologia migliorerebbero indubbiamente il modello, facendogli guadagnare robustezza, flessibilità e interpretabilità.

# Conclusioni

Di seguito sono riportate le riflessioni e i principali limiti del modello oggetto di questa tesi.

Nel caso in cui si lavori con la distribuzione Normale multivariata per la prima componente di misctura ($Z_k = 0$), si noti che il numero di dimensioni equivale al numero delle variabili latenti. Al contrario, quando si lavora con la seconda componente di mistura ($Z_k = 1$), ecco che il numero di dimensioni equivale a Q-1, ovvero al numero di variabili latenti meno uno. Questo requisito non fa altro che aumentare la flessibilità del modello permettendo di catturare in maniera più efficace, le correlazioni di errore tra un numero maggiore di variabili latenti.

Oltre a questo si potrebbe pensare che, l'utilizzo di distribuzioni matrice-variate, possa essere migliorativo per quanto riguarda la modellizzazione di sottoinsiemi di $\Theta_\delta$ (ad esempio tramite Inverse-Wishart, Inverse Gamma, MGIG o Huang-Wand). Inoltre l'utilizzo di questa tipologie di distribuzioni multivariate faciliterebbe la corretta modellizzazione dei sottogruppi di correlazioni, ciascuna con la propria media e varianza. Si sottolinea infine che l'utilizzo di una distribuzione Normale multivariata risulta l'approccio più semplice che si possa adottare per modellare i termini di Cholesky sferici non vincolati.

Alcuni dei limiti riscontrati del modello risiedono nella flessibilità e interpretabilità dello stesso. Si necessita infatti di un modello che sia flessibile, in modo tale che si adatti al meglio a diversi dati con la necessità di mantenere una struttura interpretabile. Sebbene lo spazio parametrico sia flessibile, si necessita di vincoli aggiuntivi, come la scomposizione di $\Theta_\delta$, per adattarsi ai dati reali, evitare overfitting e migliorare l'interpretabilità. Sebbene la scomposizione migliori il modello, una eccessiva correlazione tra variabili si rifletterebbe in una maggiore complessità intrinseca al modello, aumentando il rischio di overfitting e dunque l'inadattabilità del modello a nuovi dati. Al contrario, un modello troppo flessibile rischia di non adattarsi adeguatamente ai dati presentati.

Per assicurarci di poter utilizzare questo modello si necessita di sufficienti validazioni empiriche e di test su dati rappresentativi. Fino ad allora, l'efficacia dello stesso nel distinguere correttamente le due tipologie di errore, viene messa in dubbio. Oltre a questo, le simulazioni offrirebbero la possibilità di dare delle linee guida di interpretazione dei dati, necessarie a capire come modificare i test da cui sono stati originati.

L'elaborato proposto descrive in sintesi le sfide riscontrate durante il processo di studio del modello che richiede ancora revisioni e ulteriori sviluppi per garantire interpretabilità, adattabilità e identificabilità a diversi tipi di dati.

# Bibliografia


[1] Lovell e Schmidt Aigner. formulation and estimation of stochastic frontier production function models. *Journal of Econometrics 6*, 1977.

[2] Carlos Arias Antonio Alvarez. *Economics and Business Letters*, chapter A selection of relevant issues in applied stochastic frontier analysis. EBL, 2014.

[3] Beck. Beck depression inventory–ii (bdi-ii). *APA PsycTests*, 1996.

[4] J.-F. Cardoso. Blind signal separation: statistical principles. *IEEE*, 1998.

[5] Curtiss Cobb Chelsea Zhang, Sean J. Taylor and Jasjeet Sekhon. *Annals of Applied Statistics, Vol. 14, No. 3*, chapter active matrix factorization for surveys. Institute of Mathematical Statistics, 2020.

[6] L.J Cronbach. *Psychometrika 16, 297-334*, chapter Coefficient alpha and the internal structure of tests. University of Illinois, USA, 1951.

[7] David A Cole e Jeffrey A Ciesla e and James H Steiger. The insidious effects of failing to include design-driven correlated residuals in latent-variable covariance structure analysis. *Psychological methods*, 2007.

[8] David W Gerbing and James C Anderson. On the meaning of within-factor correlated measurement errors. *Journal of consumer research*, 1984.

[9] Błażej Mazur Kamil Makieła. Stochastic frontier analysis with generalized errors:inference, model comparison and averaging. *Polish National Science Center*, 2020.

[10] G. Maddala. *Handbook of Statistics v.14*, chapter 17 Errors-in-variables problems in financial models. Elsevier, 1996.

[11] José C Pinheiro and Douglas M Bates. Unconstrained parametrizations for variance-covariance matrices. *Statistics and computing*, 1996.

[12] R. P. Snaith. The clinical anxiety scale: An instrument derived from the hamilton anxiety scale. *The British Journal of Psychiatry*, 1982.

[13] C. A. Knox Lovell Subal C, Kumbhakar. *Stochastic Frontier Analysis*. Cambridge University Press, 2000.

[14] Patrick Desrosiers Vincent Thibeault, Antoine Allard. The low-rank hypothesis of complex systems. *nature physics*, 2024.

[15] J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data, second edition*. The MIT Press, 2010.