Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in
Statistica per l'Economia e l'Impresa

RELAZIONE FINALE

Modello Fattoriale Robusto

Relatore Prof. Erlis Ruli

Dipartimento di Scienze Statistiche

Laureando Daniele Bozzoli

Matricola 1226852

Anno Accademico 2022/2023

Indice

Prefazione

Il problema di ricerca affrontato in questa tesi riguarda l'analisi di dataset complessi, spesso caratterizzati da dati mancanti o outlier, e l'individuazione delle relazioni tra le variabili presenti nel dataset. L'obiettivo è quello di esplorare l'efficacia del Modello Fattoriale Robusto rispetto al Modello Fattoriale Classico nell'identificare le componenti sottostanti ai dati in presenza di contaminazione e nell'interpretazione dei risultati.
Il Modello Fattoriale è un metodo di analisi dei dati multivariati che cerca di spiegare le relazioni tra le variabili osservate attraverso un insieme ridotto di fattori latenti, non osservabili. In dettaglio, esso assume che ogni variabile sia influenzata da una combinazione lineare di questi fattori latenti, cui si aggiungono degli errori.
Il Modello Fattoriale Robusto, invece, è una variante che tiene conto della presenza di outlier o di dati mancanti. Esso cerca di minimizzare l'effetto di questi dati anomali nell'analisi dei dati attraverso l'utilizzo di metodi di stima robusti. In questo modo, è in grado di fornire stime più accurate dei parametri della procedura classica, anche in presenza di dati non conformi.
In sintesi, l'obiettivo di questa tesi è di confrontare l'efficacia dei due diversi metodi nell'analisi di dataset complessi, mettere a confronto i risultati e valutare quale performa meglio.
Quello che ci aspettiamo di osservare è che l'analisi del Modello Fattoriale Robusto sarà meno influenzata dai valori anomali e dagli errori di misura, inoltre ci aspettiamo di vedere una migliore adattabilità dei dati rispetto alla controparte, il che dovrebbe tradursi in una maggiore stabilità delle stime dei fattori e dei loro pesi fattoriali.

Capitolo 1

Modello Fattoriale Classico

Il modello fattoriale classico è un approccio ampiamente utilizzato nell'analisi dei dati per comprendere la struttura sottostante a un insieme di variabili osservate. Si basa sull'idea che le variabili misurate siano influenzate da un numero più limitato di fattori latenti che rappresentano costrutti non osservabili.
L'obiettivo del modello fattoriale classico è identificare questi fattori latenti e descrivere la loro relazione con le variabili osservate. In altre parole, cerca di spiegare la covarianza o la correlazione tra le variabili attraverso la presenza di fattori comuni che le influenzano.
Nel modello fattoriale classico, si assume che le variabili osservate siano misurate senza errori e che i fattori latenti siano non correlati tra loro. Inoltre, si suppone che le relazioni tra i fattori latenti e le variabili osservate siano lineari. Attraverso metodi come la massima verosimiglianza o i metodi dei minimi quadrati, è possibile stimare i parametri del modello, compresi i loading fattoriali che rappresentano le relazioni tra i fattori e le variabili, nonchè le varianze dei fattori e degli errori.
Il modello fattoriale è ampiamente utilizzato in diverse discipline, tra cui la psicologia, la sociologia, l'economia e la ricerca di mercato, per esplorare la struttura sottostante ai dati e comprendere le relazioni nascoste tra le variabili. Fornisce una base solida per l'analisi dei dati multidimensionali e può essere utilizzato per scopi di descrizione, previsione e inferenza.

Sia il vettore casuale $X = (X_1, ..., X_p)$ con $X \sim \mathcal{N}_p(\mu, \Sigma)$, il modello fattoriale esprime X come trasformazione lineare di variabili latenti $F_1, F_2, ..., F_k$, i fattori comuni, sommato ad altre p variabili $\epsilon_1, \epsilon_2, ..., \epsilon_p$, i fattori specifici o errori.

Formalmente, il modello fattoriale può essere scritto come

$$X_1 - \mu_1 = \lambda_{11}F_1 + \lambda_{12}F_2 + ... + \lambda_{1m}F_m + \epsilon_1$$

$$X_2 - \mu_2 = \lambda_{21}F_1 + \lambda_{22}F_2 + ... + \lambda_{2m}F_m + \epsilon_2$$

$$...$$

$$X_p - \mu_p = \lambda_{p1}F_1 + \lambda_{p2}F_2 + ... + \lambda_{pm}F_m + \epsilon_p$$

In forma matriciale diventa

$$X - \mu = L\ F + \epsilon_{p\times 1} \tag{1}$$

dove $L = [\lambda_{ij}]$ è una matrice $p \times m$, detta matrice dei *Loadings* e λ_{ij}, l'elemento della riga i e colonna j di L, rappresenta il peso della i-esima variabile casuale nel j-esimo fattore.
I fattori specifici ϵ_i e quelli comuni F_j sono variabili latenti o non osservabili. Pertanto, la situazione è diversa da quella della regressione multivariata dove le covarianze (nel caso nostro le F_i) sono note.

Il modello (1) non è pratico, poichè sono presenti troppi parametri ignoti. È usuale allora imporre le seguenti restrizioni:

- (A1) $\mathbf{E}(F) = 0_{m\times 1}$, $\mathbf{C}(F) = \mathbf{E}(FF^T) = I_{m\times m}$
- (A2) $\mathbf{E}(\epsilon) = 0_{p\times 1}$, $\mathbf{C}(\epsilon) = \mathbf{E}(\epsilon\epsilon^T) = \Psi_{p\times p} = diag(\psi_1, \psi_2, ..., \psi_p)$
- (A3) F e ϵ sono indipendenti.

La (A1) impone restrizioni sui fattori comuni: media nulla, varianze unitarie e assenza di correlazione. La (A2) impone restrizioni sui fattori specifici: media nulla e assenza di correlazione. La (A3) impone indipendenza tra tutti i fattori. Risulta allora che $\mathbf{C}(F, \epsilon) = 0$.
L'equazione (1) assieme ad (A1)-(A3) danno luogo al cosiddetto modello a fattori ortogonali, in breve, FA (Factor Analysis).

1.1 Proprietà

- (i) $\Sigma_{FA} = \mathbf{C}(X|FA) = \mathbf{E}[(X - \mu)(X - \mu)^T|FA] = LL^T + \Psi$
- (ii) $\mathbf{C}(X, F|FA) = L$

dove $\mathbf{C}(X|FA)$ significa covarianza di X condizionatamente alle assunzioni della FA. Quest'ultimo quindi assume che la matrice di covarianza di X abbia una struttura additiva del tipo $LL^T + \Psi$, dove la matrice dei loadings

L è anche matrice di cross-correlazione tra X e F.
Con $\Sigma_{FA} = [\sigma_{ij}^{FA}]$, le proprietà (i) - (ii) affermano che sotto la FA:

$$\sigma_{ii}^{FA} = \mathbf{V}(X_i|FA) = \lambda_{i1}^2 + \lambda_{i2}^2 + ... + \lambda_{im}^2 + \psi_i$$

$$\sigma_{ik}^{FA} = \mathbf{C}(X_i, X_k|FA) = \lambda_{i1}\lambda_{k1} + \lambda_{i2}\lambda_{k2} + ... + \lambda_{im}\lambda_{km}$$

$$\mathbf{C}(X_i, F_j|FA) = \lambda_{ij}$$

La quantità $h_i^2 = \lambda_{i1}^2 + \lambda_{i2}^2 + ... + \lambda_{im}^2$ è chiamata comunality della i-esima variabile e costituisce la parte della varianza di X_i attribuita agli m fattori comuni. La ψ_i è chiamata *varianza specifica* o *specificity*.
Pertanto si ha:

$$\sigma_{ii}^{FA} = h_i^2 + \psi_i$$

Il vettore casuale X ha matrice di covarianza (incondizionata) $\Sigma = [\sigma_{ij}]$ la quale ha $p(p+1)/2$ parametri ignoti.
Il modello fattoriale impiega, nella peggiore delle ipotesi, $p(m+1)$ parametri ignoti, pertanto la matrice di covarianza di X sotto il modello fattoriale $\Sigma_{FA} = LL^T + \Psi$ è solo una approssimazione di Σ.
La bontà dell'approssimazione dipende sia dalla scelta di $m \in \{1, ..., p\}$ e sia dal fatto che Σ ammetta effettivamente una decomposizione del tipo $LL^T + \Psi$.
Per $m = p$, $\Sigma_{FA} = \Sigma$. Infatti, data la decomposizione spettrale $\Sigma = \Gamma\Lambda\Gamma^T$, basta porre $L = \Gamma\Lambda^{1/2}$ e $\Psi = 0$.
La FA è però utile se $m << p$, dove il costo della perdita di accuratezza è superato dal beneficio di modello parsimonioso ed interpretabile.

Una delle maggiori critiche avanzata alla FA è la non unicità di L.
Se $m > 1$ e E è una matrice $m \times m$ e ortogonale, allora

$$X - \mu = LF + \epsilon = LEE^T + \epsilon = L^*F^* + \epsilon$$

con $L^* = LE$ e $F^* = E^TF$.
I *loadings* L e i *fattori comuni* F sono quindi determinati a meno di rotazioni ortogonali, mentre le communalities sono invarianti alle rotazioni ortogonali.

1.2 Stima dei parametri

La non unicità della soluzione rispetto alle rotazioni ortogonali rappresenta un problema, ma anche un vantaggio.
Una volta stimato L e Ψ, la rotazione di L tipicamente agevola l'interpretazione dei fattori.
Siano $X_1, X_2, ..., X_n$ n realizzazioni del vettore aleatorio X. Dai dati possiamo ottenere S, la stima campionaria di Σ, e successivamente usarla per stimare L, Ψ ed F.

La stima L e Ψ può essere affrontata con due metodi diversi: il metodo delle componenti principali ed il metodo della massima verosimiglianza. Questi metodi possono dare soluzioni molto diverse, ma se il modello scelto è ben supportato dai dati, le soluzioni non dovrebbero divergere in maniera sostanziale. È consigliabile, quindi, applicare più metodi e confrontarne i risultati.

1.2.1 Metodo delle componenti principali

Per capire il metodo partiamo prima con il caso teorico. Sia Σ la matrice di covarianza del vettore aleatorio, con le coppie di autovalori e autovettori (λ_1, e_1), (λ_2, e_2), ... , (λ_p, e_p), dove $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p \geq 0$, allora:

$$\begin{aligned}\Sigma &= \lambda_1 e_1 e_1^T + \lambda_2 e_2 e_2^T + ... + \lambda_p e_p e_p^T \\ &= (\Gamma \Lambda^{1/2})(\Lambda^{1/2} \Gamma^T) \\ &= [\sqrt{\lambda_1} e_1 | \sqrt{\lambda_2} e_2 | ... | \sqrt{\lambda_p} e_p] \begin{bmatrix} \sqrt{\lambda_1} e_1^T \\ \sqrt{\lambda_2} e_2^T \\ ... \\ \sqrt{\lambda_p} e_p^T \end{bmatrix} \end{aligned} \tag{2}$$

Per un modello fattoriale con $m = p$ fattori basta porre $L = \Gamma \Lambda^{1/2}$ e $\psi_i = 0$. I loadings del fattore F_j sono i coefficienti della j-esima componente principale per $\sqrt{\lambda_j}$.
Per ottenere un modello fattoriale parsimonioso, l'idea è di scartare gli ultimi $p - m$ contributi in (2) se gli ultimi $p - m$ autovalori ≈ 0.
Scartando gli ultimi $p - m$ contributi in (2) si ottiene l'approssimazione di Σ

$$[\sqrt{\lambda_1} e_1 | \sqrt{\lambda_2} e_2 | ... | \sqrt{\lambda_m} e_m] \begin{bmatrix} \sqrt{\lambda_1} e_1^T \\ \sqrt{\lambda_2} e_2^T \\ ... \\ \sqrt{\lambda_m} e_m^T \end{bmatrix} = LL^T. \tag{3}$$

Poichè nella (3) abbiamo scartato solo i contributi per i quali gli autovalori sono nulli o trascurabili, ci si aspetta che la differenza tra Σ e LL^T sia nulla o trascurabile.
Tale differenza trascurabile però ci permette di includere nella (3) anche i fattori specifici. Posto $\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2$, $i = 1, 2, ..., p$, allora

$$\Sigma \approx LL^T + diag(\psi_1, \psi_2, ..., \psi_p) = LL^T + \Psi. \tag{4}$$

Sostituendo Σ con S abbiamo le stime di L e Ψ che cercavamo.
Riassumendo, il metodo delle componenti principali per la stima del modello fattoriale è il seguente:

- Dai dati X si ottiene S e si ricava poi la sua decomposizione spettrale, con matrice di autovettori $\hat{\Gamma}$ e matrice di autovalori $\hat{\Lambda}$.

- Fissato m il numero di fattori comuni, la stima dei loadings è

$$\tilde{L} = [\sqrt{\hat{\lambda}_1}\hat{e}_1|\sqrt{\hat{\lambda}_2}\hat{e}_2|...|\sqrt{\hat{\lambda}_m}\hat{e}_m]$$

  con $(\hat{\lambda}_i, \hat{e}_i), i \in \{1, ..., m\}$ i-esima coppia autovalore-autovettore di S.

- La stima delle varianze specifiche è $\tilde{\psi}_i = s_{ii} - \sum_{j=1}^{m} \tilde{l}_{ij}^2$, e la stima della matrice di covarianza dei fattori specifici è $\tilde{\Psi} = diag(\tilde{\psi}_1, \tilde{\psi}_2, ..., \tilde{\psi}_p)$.

È evidente che, se da un modello con m fattori si passa ad uno con $m+1 \leq p$, le stime dei primi m fattori di questi due modelli sono identiche.
Il modello fattoriale così stimato riproduce esattamente gli elementi della diagonale di S.

1.2.2 Metodo dei fattori principali

Questo metodo è una variazione del precedente metodo delle componenti principali. Descriviamo il ragionamento in termini di un'analisi fattoriale con la matrice di correlazione R, anche se la procedura è appropriata anche per la matrice di covarianza S. Se il modello dei fattori $\rho = LL^T + \Psi$ è correttamente specificato, i m fattori comuni dovrebbero tener conto degli elementi che si trovano fuori dalla diagonale di ρ, così come le porzioni di comunalità degli elementi sulla diagonale

$$\rho_{ii} = 1 = h_i^2 + \psi_i$$

Se viene rimosso il contributo specifico del fattore ψ_i dalla diagonale oppure, equivalentemente, l'1 viene sostituito da h_i^2, la matrice risultante sarà $\rho - \Psi = LL^T$. Supponiamo ora che siano disponibili le stime iniziali ψ_i^* delle varianze specifiche. Sostituendo l'i-esimo elemento diagonale di R con $h_i^{*2} = 1 - \psi_i^*$, otteniamo una matrice di correlazione "ridotta" del campione

$$R_r = \begin{bmatrix} h_1^{*2} & r_{12} & ... & r_{1p} \\ r_{12} & h_2^{*2} & ... & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & ... & h_p^{*2} \end{bmatrix}$$

Oltre alla variazione campionaria, tutti gli elementi della matrice di correlazione "ridotta" del campione R_r dovrebbero essere spiegati dai m fattori comuni. In particolare, R_r viene scomposto come segue

$$R_r = L_r^* L_r^{*T}$$

dove $L_r^* = \{l_{ij}^*\}$ sono i loadings stimati.
Il metodo dei fattori principali utilizza le stime

$$L_r^* = [\sqrt{\hat{\lambda}_1^*}\hat{e}_1^*|\sqrt{\hat{\lambda}_2^*}\hat{e}_2^*| \ ... \ |\sqrt{\hat{\lambda}_m^*}\hat{e}_m^*]$$

$$\psi_i^* = 1 - \sum_{j=1}^{m} l_{ij}^{*2}$$

dove $(\hat{\lambda}_i^*, \hat{e}_i^*)$, $i = 1, 2, ..., m$ sono le coppie autovalore-autovettore (le più grandi) determinate da R_r. A loro volta, le comunalità sarebbero quindi (ri)stimate mediante

$$\tilde{h}_i^{*2} = \sum_{j=1}^{m} l_{ij}^{*2} \tag{5}$$

La soluzione data dal metodo dei fattori principali può essere ottenuta in modo iterativo, con le stime delle comunalità di (5) che diventano le stime iniziali per la fase successiva di iterazione. La considerazione degli autovalori stimati $\hat{\lambda}_1^*, \hat{\lambda}_2^*, ..., \hat{\lambda}_p^*$ aiuta a determinare il numero di fattori comuni da mantenere. Una complicazione aggiuntiva è che ora alcuni autovalori possono essere negativi a causa dell'uso di stime iniziali delle comunalità. Idealmente, dovremmo prendere il numero di fattori comuni uguale al rango della matrice ridotta della popolazione. Purtroppo, questo rango non è sempre ben determinato da R_r, e quindi è necessario fare delle valutazioni. Sebbene ci siano molte scelte per le stime iniziali delle varianze specifiche, la scelta più popolare, quando si lavora con una matrice di correlazione, è $\psi_i^* = 1/r^{ii}$, dove r^{ii} è i-esimo elemento diagonale di R^{-1}. Le stime iniziali delle comunalità diventano quindi

$$h_i^{*2} = 1 - \psi_i^* = 1 - \frac{1}{r^{ii}}$$

che è uguale al quadrato del coefficiente di correlazione multipla tra X_i e le altre $p - 1$ variabili. La relazione con il coefficiente di correlazione multipla significa che h_i^{*2} può essere calcolato anche quando R non ha rango pieno. Per la fattorizzazione di S, le stime iniziali delle varianze specifiche utilizzano s^{ii}, gli elementi diagonali di S^{-1}.
Sebbene il metodo dei componenti principali per R possa essere considerato come un metodo dei fattori principali con stime iniziali delle comunalità unitarie o varianze specifiche nulle, i due metodi sono filosoficamente e geometricamente differenti (vedi Harmon (1976)). Tuttavia, nella pratica, i due producono spesso pesi fattoriali comparabili se il numero di variabili è grande e il numero di fattori comuni è piccolo.

Scelta del numero m di fattori

Spesso la scelta di m è determinata da considerazioni teoriche contestuali all'applicazione. In assenza di tali considerazioni:
Sia $\Upsilon = S - \tilde{L}\tilde{L}^T - \tilde{\Psi}$ la matrice dei residui. Se $diag(\Upsilon) = 0$ e se anche gli altri elementi di Υ fossero zero, allora il modello FA scelto è supportato.
Per Υ, la seguente uguaglianza è valida

$$||\Upsilon||_2^2 \leq \hat{\lambda}_{m+1}^2 + \hat{\lambda}_{m+2}^2 + ... + \hat{\lambda}_p^2$$

Si noti che $||\Upsilon||_2^2$ è la somma dei residui al quadrato e può essere considerato come una misura di **errore totale**. La bontà del modello fattoriale scelto dipende dall'entità degli autovalori di S che sono stati scartati. Alla varianza campionaria della i-esima variabile s_{ii}, il primo fattore contribuisce con $\tilde{l}_{i1}^2$. Pertanto, il contributo del primo fattore alla varianza totale $tr(S) = s_{11} + s_{22} + ... + s_{pp}$ è $\sum_{i=1}^{p} \tilde{l}_{i1}^2 = \hat{\lambda}_1$.
Usando la proporzione della varianza totale dovuta al fattore j

$$\hat{\lambda}_j / tr(S)$$

abbiamo i seguenti criteri per fissare m:

- includere i fattori fino all'eventuale punto di gomito dello screeplot
- scegliere tanti fattori quanti servono per ottenere una certa proporzione della varianza spiegata (e.g. 85%)
- includere tutti i fattori con autovalori della matrice di correlazione campionaria maggiori di uno.

1.2.3 Metodo della massima verosimiglianza

Si assume che F e ϵ siano distribuiti normalmente, quindi

$$X_j \sim \mathcal{N}_p(\mu, \Sigma), \qquad j = 1, 2, ..., n$$

dove $\Sigma = LL^T + \Psi$. La funzione di log-verosimiglianza di $\theta = \mu, L, \Psi$ è

$$\begin{aligned} l(\theta) = & \qquad \log \prod_{j=1}^{n} \phi_p(x_j; \mu, \Sigma) \\ = & \qquad -\frac{n}{2} \log|\Sigma| - \sum_{j=1}^{n} (x_j - \mu)^T \Sigma^{-1} (x_j - \mu)/2 \end{aligned} \tag{6}$$

dove ϕ_p è la funzione di densità della distribuzione normale multivariata.
Il modello è tuttavia soggetto al problema della non unicità rispetto a rotazioni ortogonali di L.
Pertanto si considera il **vincolo di unicità**

$$L^T \Psi^{-1} L = \Delta, \qquad \text{con } \Delta \text{ matrice diagonale}$$

La stima dei parametri tramite il metodo della massima verosimiglianza (MV), $\hat{\mu}, \hat{L} = [\hat{l}_{ij}]$, $\hat{\Psi} = diag(\hat{\psi}_1, ..., \hat{\psi}_p)$ è definita come

$$argmax_{(\theta \in \Theta, L^T \Psi^{-1} L = \Delta)} l(\theta)$$

La stima $\hat{\mu} = \bar{x}$ è analitica, mentre $\hat{L}$ e $\hat{\Psi}$ vanno calcolati numericamente. Le comunalità stimate MV sono

$$\hat{h}_i^2 = \sum_{j=1}^{m} \hat{l}_{ij}^2$$

e la proporzione della varianza spiegata dal fattore j è

$$\sum_{i=1}^{p} \hat{l}_{ij}^2 / tr(S)$$

Scelta del numero m di fattori

Sia $\hat{\Sigma} = \hat{L}\hat{L}^T + \hat{\Psi}$. Le ipotesi del test per l'adeguatezza del modello a m fattori sono:

$$H_0 : \Sigma = LL^T + \Psi \qquad \text{contro} \qquad H_1 : \Sigma \neq LL^T + \Psi$$

La statistica test log-rapporto di verosimiglianza è

$$-2\log\Lambda_n = -2\log\left(\frac{|\hat{\Sigma}|}{|S|}\right)^{-n/2} + n[tr(\hat{\Sigma}^{-1}S) - p]$$

per cui vale

$$-2\log\Lambda_n \sim \chi^2_\nu, \qquad n \to \infty$$

con $\nu = [(p-m)^2 - p - m]/2$. Poichè $n[tr(\hat{\Sigma}^{-1}S) - p] = 0$ quando $\hat{\Sigma} = \hat{L}\hat{L}^T + \hat{\Psi}$, si ha allora

$$-2\log\Lambda_n = n\log\left(\frac{|\hat{\Sigma}|}{|S|}\right). \tag{7}$$

La correzione di Bartlett con $n - 1 - (2p + 4m + 5)/6$ al posto di n in (7) fornisce un'approssimazione più accurata.

Commenti

- Per n elevato e $m << p$, il test log-rapporto di verosimiglianza tende a rigettare H_0 a favore di modelli con più fattori.
- Tuttavia, spesso l'aggiunta di ulteriori fattori al modello non apporta miglioramenti praticamente apprezzabili
- Si ripresenta un problema ricorrente: distinguere la significatività statistica dalla significatività pratica.

- Il test log-rapporto di verosimiglianza va quindi usato con cautela, come un'indicazione di massima, e va sempre affiancato agli altri criteri di analisi statistica.

- Il più utile fra tutti è forse l'interpretazione dei fattori: è meglio un modello semplice e interpretabile che uno molto flessibile, ma difficilmente interpretabile

- La rotazione dei loadings gioca un ruolo cruciale nella loro interpretazione.

1.3 Rotazione dei Loadings

Se $\hat{L}$ e $\hat{\Psi}$ sono delle stime dei parametri e E è una matrice ortogonale, allora

$$\hat{L}\hat{L}^T + \hat{\Psi} = \hat{L}EE^T\hat{L}^T + \hat{\Psi} = \hat{L}^*\hat{L}^* + \hat{\Psi}$$

$\hat{L}$ e la sua trasformazione ortogonale $LE = \hat{L}^* = l^*_{ij}$, danno luogo alla stessa matrice di covarianza e dei residui.
Questa trasformazione viene chiamata **rotazione** dei loadings o dei fattori.
L'obiettivo della rotazione è semplificare e chiarire la struttura dei dati.
La rotazione non può migliorare gli aspetti fondamentali dell'analisi, come la quantità di varianza estratta dagli elementi. Per quanto riguarda il metodo di estrazione, ci sono diverse scelte disponibili.
La rotazione Varimax è di gran lunga la scelta più comune. Varimax, quartimax ed equamax sono metodi ortogonali comuni di rotazione (Costello and Osborne (2005)); direct oblimin, quartimin e promax sono invece metodi obliqui.
Le rotazioni ortogonali producono fattori non correlati; i metodi obliqui consentono ai fattori di essere correlati. Comunemente si consiglia ai ricercatori di utilizzare la rotazione ortogonale perchè produce risultati più facilmente interpretabili, ma questo è un argomento fallace.
Nelle scienze sociali, per esempio, ci aspettiamo generalmente una certa correlazione tra i fattori, poichè il comportamento raramente è suddiviso in unità nettamente separate che funzionano in modo indipendente l'una dall'altra.
Pertanto, l'utilizzo della rotazione ortogonale comporta una perdita di informazioni preziose se i fattori sono correlati, e teoricamente la rotazione obliqua dovrebbe fornire una soluzione più accurata e forse più riproducibile. Se i fattori sono veramente non correlati, la rotazione ortogonale e obliqua producono risultati quasi identici.
(a) : gli assi fattoriali sono ruotati per migliorare l'interpretazione dei fattori, preservandone l'ortogonalità.
(b) : gli assi fattoriali sono ruotati per passare il più vicino possibile ai

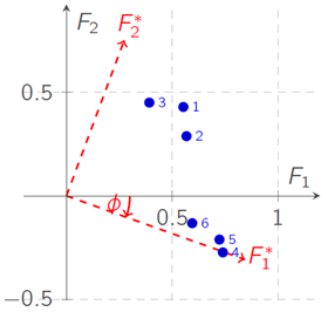

Figura 1.1: a) Esempio di rotazione ortogonale

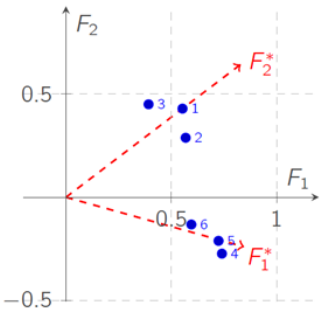

Figura 1.2: b) Esempio di rotazione obliqua

loadings. Si ottiene una migliore interpretazione dei fattori, ma si perde il vincolo di ortogonalità.

Scelta della rotazione

Come già prima menzionato, tra i vari tipi di rotazione obliqua e ortogonale, la **varimax** (ortogonale) e la **promax** (obliqua) sono le più note.
Siano $\tilde{l}_{ij}^* = \hat{l}_{ij}^*/\hat{h}_i$, i loadings stimati, ruotati e scalati per la radice quadrata delle comunalità. Il criterio varimax mira a selezionare la matrice ortogonale E che massimizza

$$V = \sum_{j=1}^{m} \left[\sum_{i=1}^{p} \tilde{l}_{ij}^{*\,4} - \left(\sum_{i=1}^{p} \tilde{l}_{ij}^{*\,2} \right)^2 \Big/ p \right] \Big/ p$$

La quantità V non è altro che la varianza dei quadrati di $\tilde{l}_{ij}^*$
Massimizzare V rispetto alla matrice ortogonale E vuol dire quindi individuare quella particolare E che ottiene la massima dispersione dei quadrati di $\tilde{l}_{ij}^*$.

1.4 Stima dei fattori comuni

La stima dei fattori comuni F è importante per varie finalità, tra cui:

- verificare l'ipotesi di normalità di F
- confrontare due soggetti o gruppi di soggetti

Una stima dei fattori può essere ottenuta partendo dalla stima di L e di Ψ.
Le stime dei fattori comuni vengono chiamati **punteggi fattoriali**.
I metodi principali per il calcolo dei punteggi fattoriali sono:

1. metodo dei minimi quadrati pesati
2. metodo della regressione

1.4.1 Metodo dei minimi quadrati pesati

Trattando $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_p)$ come residui, la stima dei minimi quadrati pesati è data dall'f che minimizza

$$\sum_{i=1}^{p} \frac{\epsilon_i^2}{\psi_i} = \epsilon^T \Psi^{-1} \epsilon = (x - \mu - Lf)^T \Psi^{-1} (x - \mu - Lf).$$

Il minimo è raggiunto in

$$f_B = (L^T \Psi^{-1} L)^{-1} L^T \Psi^{-1} (x - \mu).$$

I parametri ignoti μ, L, Ψ in f_B vengono sostituiti con le stime ed x è una generica osservazione.
Con L e Ψ stimati con MV, dove $\hat{\Delta} = \hat{L}^T \hat{\Psi}^{-1} \hat{L}$, la stima dei fattori per il j-esimo soggetto quindi diventa

$$\begin{aligned} \hat{f}_j^B &= (\hat{L}^T \hat{\Psi}^{-1} \hat{L})^{-1} \hat{L}^T \hat{\Psi}^{-1} (x_j - \hat{\mu}) \\ &= \hat{\Delta}^{-1} \hat{L}^T \hat{\Psi}^{-1} (x_j - \overline{x}) \end{aligned}$$

1.4.2 Metodo della regressione

Per μ, L, Ψ fissati e sotto l'ipotesi di normalità di F e ϵ, ne consegue che $X - \mu = LF + \epsilon \sim \mathcal{N}_p(0, LL^T + \Psi)$.
Inoltre, la distribuzione di $X - \mu$ e F è $\sim \mathcal{N}_{p+m}(0, \Sigma^*)$, con

$$\Sigma^*_{(p+m)\times(p+m)} = \begin{bmatrix} LL^T + \Psi & L \\ L^T & I_{m\times m} \end{bmatrix},$$

Usando proprietà della normale multivariata ne consegue che
$F|X = x \sim N_p(\mu_f, \Sigma_F)$, pertanto

$$\mu_f = L^T (LL^T + \Psi)^{-1} (x - \mu), \qquad \Sigma_F = I - L^T (LLT + \Psi)^{-1} L.$$

F si stima tramite $\hat{f}^T = \mathrm{E}(F|X)$. I punteggi fattoriali con il metodo della regressione, con L, Ψ stimati con MV sono quindi

$$\hat{f}_j^T = \hat{L}^T (\hat{L}\hat{L}^T + \hat{\Psi})^{-1} (x_j - \overline{x}), \qquad j = 1, 2, ..., n. \tag{8}$$

I punteggi fattoriali ruotati sono $\hat{f}_j^{*T} = E^T \hat{f}_j^T$.

1.5 Covarianza o correlazione?

La scelta tra l'utilizzare la matrice di covarianza o la matrice di correlazione dipende dal contesto e dagli obiettivi dell'analisi.
Entrambe le matrici forniscono informazioni sulle relazioni tra le variabili,

ma differiscono nella scala di misura utilizzata. La matrice di covarianza riflette le relazioni lineari tra le variabili originali, tenendo conto delle loro unità di misura. D'altra parte, la matrice di correlazione standardizza le variabili dividendole per le deviazioni standard, eliminando così le unità di misura. Questo rende le variabili comparabili direttamente in termini di relazioni lineari e ne facilita il confronto.
La matrice di correlazione è particolarmente utile quando si desidera esaminare le relazioni tra le variabili in modo comparativo o quando le unità di misura delle variabili sono diverse o non rilevanti per l'analisi.
La scelta tra i due metodi dipende quindi dagli obiettivi specifici dell'analisi e dalle considerazioni teoriche e/o pratiche del contesto di studio.

Capitolo 2

Modello Fattoriale Robusto

Il modello fattoriale robusto è un'estensione dell' analisi fattoriale che tiene conto della presenza di dati atipici o devianti nell'analisi dei dati. Mentre l'analisi fattoriale classica assume l'omogeneità e la normalità dei dati, il modello fattoriale robusto mira a gestire i dati non conformi a tali assunzioni, fornendo stime più robuste e affidabili dei parametri del modello.
L'obiettivo è identificare e interpretare le relazioni sottostanti tra le variabili osservate attraverso la stima dei fattori latenti e dei loadings. La differenza principale rispetto all'approccio classico è l'utilizzo di stime robuste della matrice di covarianza o correlazione, che sono meno sensibili agli effetti di dati atipici o devianti.
Nelle prossime sezioni affronteremo alcuni metodi di stima e concetti chiave necessari per progettare procedure robuste efficaci per scopi come la stima della matrice di covarianza o di correlazione.

2.1 Introduzione alla robustezza

Un primo requisito degli stimatori robusti è l'efficienza nella distribuzione specificata del modello, il che significa che ci si aspetta solo una perdita limitata e eventualmente controllabile di efficienza in assenza di contaminazione. La perdita di efficienza può essere misurata come il rapporto tra la varianza delle procedure robuste e classiche.
Un secondo requisito degli stimatori robusti è che abbiano una funzione di influenza limitata. Il bias dello stimatore non può essere arbitrariamente grande.
Un terzo requisito è un alto punto di rottura: gli stimatori robusti devono resistere a una grande quantità di contaminazione prima di diventare inaffidabili. Ciò dovrebbe essere possibilmente combinato con un massimo bias relativamente piccolo. Le stime robuste sono definite da un insieme di pesi volti a sottopesare quei punti che si discostano dalle assunzioni del model-

lo. Dopo l'utilizzo degli stimatori robusti, le osservazioni anomale possono essere rilevate come outlier osservando la loro distanza dall'adattamento robusto.

2.1.1 Funzione d'influenza

La funzione di influenza (Hampel (1974); Huber and Ronchetti (1981)) descrive l'effetto delle deviazioni dal modello parametrico specificato F all'interno di un intorno $P(F;\epsilon)$. La funzione di influenza è data dal cambiamento relativo in $T(F)$, come ad esempio una stima puntuale o la potenza di una statistica test, causato da una piccola proporzione ϵ di valori anomali, tutti uguali a un valore arbitrario $x \in \mathbb{R}^p$.
(da Farcomeni et al. (2015))

Sia $F_\epsilon = (1-\epsilon)F + \epsilon\delta_x$, dove δ_x rappresenta un'osservazione contaminata in x. La funzione di influenza per una contaminazione infinitesima di punto ϵ in posizione x nel modello Fe con una statistica $T(X)$, esprimibile anche nella forma $T(F_n)$ è

$$IF(x;T,F) = \lim_{\epsilon\to 0} \frac{T(F_\epsilon - T(F)}{\epsilon} = \frac{\partial}{\partial\epsilon} T(F_\epsilon)_{|\epsilon=0}.$$

La quantità $\epsilon IF(x;T;F)$ fornisce un'approssimazione del bias di $T(F_\epsilon)$ determinato dalla contaminazione infinitesimale. Una funzione di influenza limitata significa che il cambiamento relativo nella funzione non è arbitrariamente grande. Pertanto, se $IF(x;T;F)$ è limitato, un'osservazione contaminata situata in x non può essere troppo dannosa per le proprietà statistiche di T nel modello F.
Se $IF(x;T;F)$ è limitato per tutti i valori di x, si dimostra formalmente la robustezza locale di T.
L'effetto relativo di un' osservazione contaminazione, misurato dalla IF, può essere riassunto dalla *sensibilità totale degli errori*

$$\gamma^* = \sup_{x\in\mathbb{R}^p} ||IF(x;T,F)||.$$

Un γ^* limitato identifica una statistica robusta.
Un secondo indice che può essere ottenuto dall'IF è mirato a misurare gli effetti di piccoli cambiamenti nei dati, come quelli derivanti dall'arrotondamento e dal raggruppamento, ad esempio. Supponiamo che l'osservazione x sia sostituita da un nuovo valore y.
L'effetto dello spostamento può essere misurato dalla sensibilità locale allo spostamento

$$\lambda^* = \sup_{x\neq y, x, y\in\mathbb{R}^p} \frac{||IF(y;T,F) - IF(x;T,F)||}{||y-x||}.$$

Se $y = x + \epsilon_x$, con $\epsilon_x \to 0$, allora la sensibilità allo spostamento locale corrisponde alla pendenza della funzione di influenza nel punto x.
Un terzo valore di sintesi direttamente correlato all'IF è il punto di rigetto, che identifica i punti in cui la funzione di influenza si annulla e quindi la contaminazione non ha effetto; si consulti Hampel et al. (1986) e Huber and Ronchetti (2009) per una discussione più approfondita.

2.1.2 Punto di rottura

Il punto di rottura (*Breakdown Point, BP*) di una statistica $T(F_n)$ è la frazione massima dei dati (ovvero la quantità di contaminazione) che può essere arbitrariamente sostituita da outlier mentre $T(F_n)$ rimane limitata.
Fino a un certo tasso di contaminazione arbitraria, ci si aspetta che una statistica robusta rimanga limitata lontano dal confine dello spazio parametrico. Questo non accade ad esempio con la media campionaria: una singola osservazione infinita porta a una media infinita, indipendentemente dalla dimensione del campione.
La quantità massima di contaminazione tollerabile (almeno asintoticamente) da qualsiasi statistica robusta è del 50%, mentre molte stime classiche (come la media campionaria) non possono sopportare alcuna contaminazione. Esistono diverse definizioni possibili del punto di rottura, basate su argomenti asintotici (Hampel (1971)), sul comportamento in campioni finiti (Hodges Jr (1967); Donoho and Huber (1983)) di una statistica, o sulla natura della contaminazione.
(da Farcomeni et al. (2015))

Il punto di rottura asintotico di una funzione statistica $T(F)$ è il più grande tasso di contaminazione ϵ tale che $T(F_\epsilon)$ sia limitato e distante dal confine dello spazio parametrico, cioè

$$\epsilon^* = \max \{\epsilon : T(F_\epsilon) \in K \subset \Theta, \forall G\},$$

dove K è un insieme limitato e chiuso che non contiene i punti di confine dello spazio parametrico.
(da Farcomeni et al. (2015)):

Sia $X_r \in \mathcal{X}_r$, dove $\mathcal{X}_r$ è la collezione di tutti i data set X_r di dimensione n che hanno $(n - r)$ elementi (ovvero intere righe) in comune con i dati originali X. Sia K un insieme limitato e chiuso che non contiene i punti di confine dello spazio parametrico. Il punto di rottura nel campione finito è

$$\epsilon^{(i)} = \max \left\{ \frac{r}{n} : \sup_{\mathcal{X}_r} ||T(X) - T(X_r)|| \in K \right\}.$$

In parole semplici, $\epsilon^{(i)}$ misura la frazione più grande di outlier strutturali arbitrari che possiamo includere senza ottenere una divergenza di $T(X)$.

In modo simile e in termini semplici, ϵ^* è il punto di rottura di una procedura man mano che la dimensione del campione cresce all'infinito.
Se un valore di rottura è infinitesimale, possiamo dichiarare formalmente che la procedura non è robusta.

2.2 Stima robusta della matrice di covarianza

2.2.1 Stimatori M multivariati

Sotto l'assunzione che i dati derivino da una distribuzione $\mathcal{N}_p(\mu, \Sigma)$, dove ora $\mu = (\mu_1, \mu_2, ..., \mu_p)^T$ e Σ è una matrice $p \times p$ simmetrica e definita positiva, lo stimatore di massima verosimiglianza $(\overline{x}, S)$ è la soluzione alle equazioni di massima verosimiglianza

$$\begin{cases} \sum_{i=1}^n (x_i - \mu) = 0 \\ \sum_{i=1}^n [(x_i - \overline{x})(x_i - \overline{x})^T - \Sigma] = 0. \end{cases}$$

Il problema di trovare una stima robusta per μ, Σ può essere affrontato sostituendo le solite funzioni score con diverse funzioni ϕ che conducono a stimatori a influenza limitata.
Una prima soluzione è data dalla classe di stimatori M multivariati (Hampel et al. (1986)). In modo simile all'ambiente univariato, le stime $M(\hat{\mu}_M, \hat{\Sigma}_M)$ sono un vettore medio ponderato e una matrice di covarianza, rispettivamente.
Formalmente, una stima M multivariata di posizione e covarianza $M(\hat{\mu}_M, \hat{\Sigma}_M)$, ora con $\hat{\mu}_M = (\hat{\mu}_{M;1}, \hat{\mu}_{M;2}, ..., \hat{\mu}_{M;p})^T$, (Maronna (1976); Huber and Ronchetti (1981)) è definita come la soluzione delle seguenti equazioni

$$\begin{cases} \sum_{i=1}^n \phi_\mu(x_i; \mu, \Sigma) = \sum_{i=1}^n (x_i - \mu) w_{i1} = 0 \\ \sum_{i=1}^n \phi_\Sigma(x_i; \mu, \Sigma) = \sum_{i=1}^n [(x_i - \mu)(x_i - \mu)^T w_{i2} - \Sigma] = 0, \end{cases}$$

dove i due vettori di peso w_{i1} e w_{i2} dipendono da $d_i = d(x_i; \hat{\mu}_M, \hat{\Sigma}_M)$.
Le funzioni di peso non sono necessariamente uguali. I pesi sono progettati per essere piccoli per quelle osservazioni la cui distanza robusta di Mahalanobis è troppo grande rispetto alla maggioranza dei dati. La distanza di Mahalanobis proveniente da un adattamento robusto è comunemente indicata come *distanza robusta*. Le distanze robuste sono lo strumento più importante per la rilevazione di outlier multivariati. Le proprietà di robustezza e l'efficienza dello stimatore del vettore di posizione $\hat{\mu}_M$ e dello stimatore di covarianza $\hat{\Sigma}_M$ possono essere ricercate mediante le loro funzioni di influenza. In particolare, la funzione di influenza dello stimatore M di posizione (Hampel et al. (1986)) è

$$IF(x, \hat{\mu}_M, F) = \frac{1}{pM} \phi_\mu(X; 0, I_p) \tag{9}$$

con $M = \mathbf{E}_F \left[-\frac{\partial}{\partial_\mu^T} \phi_\mu(x_i; \mu, \Sigma) \right] = \mathbf{E}_F[w_i(p-1) + \phi'_\mu(x_i; \mu, \Sigma)]$,
quando $p = 1$, otteniamo nuovamente la funzione di influenza univariata degli stimatori M.

2.2.2 Stimatori S multivariati

Gli stimatori S (Davies (1987); Lopuha (1989)) sono definiti come la soluzione $(\hat{\mu}_S, \hat{\Sigma}_S$ al problema di minimizzare il determinante di $|\hat{\Sigma}|$ soggetto al vincolo

$$\frac{1}{n} \sum_{i=1}^{n} \rho(d_i) = const, \qquad 0 < const < \sup_d \rho(d)$$

dove $\rho(d)$ è una funzione opportunamente specificata: una scelta popolare è la funzione biquadratica di Tukey (Farcomeni et al. (2015), par 2.13).
Ancora una volta, tra le possibili scelte per *const*, nella pratica viene utilizzata solo una singola funzione, quella che garantisce la consistenza nel modello normale. Questo corrisponderà al valore atteso di $\rho(d_i)$ nel modello normale.
La costante di robustezza c può essere scelta come una funzione del BP una volta che il BP è stato fissato, attraverso (Farcomeni et al. (2015), par 2.14).
Nella distribuzione normale standard multivariata $\Phi_p = \mathcal{N}(0; I_p)$ (Riani et al. (2014)),

$$\mathbf{E}_{\Phi_p}[\rho_T(X)] = \frac{H_c(2)}{2} - \frac{H_c(4)}{2c^2} + \frac{H_c(6)}{6c^4} + \frac{c^2}{6p}[1 - F_{\chi^2_p}(c^2)],$$

dove

$$H_c(k) = \prod_{j=0}^{k/2-1} (p + 2j) F_{\chi^2_{p+k}}(c^2).$$

Gli stimatori S possono anche essere definiti come la soluzione delle equazioni di stima di tipo M e le loro proprietà di robustezza ed efficienza possono essere valutate secondo (9) e Lemma 2.1 (Lopuha (1989); Croux and Haesbroeck (1999)).
La costante di robustezza c determina non solo il punto di rottura, ma anche l'efficienza asintotica degli stimatori di posizione e covarianza (Riani et al. (2014)).
Pertanto, è necessaria una certa attenzione una volta che abbiamo deciso di fissare c in base a considerazioni di efficienza o rottura. Vale a dire, se scegliamo c per ottenere un'efficienza prestabilita, dobbiamo verificare il corrispondente BP, e viceversa.

2.2.3 MCD

Il metodo MCD (*Minimum Covariance Determinant*), introdotto da Rousseeuw (1985), stima $(\mu; \Sigma)$ basandosi su un sottoinsieme di $n(1-\epsilon)$ punti dei dati. Questi punti vengono scelti prendendo l'insieme il cui determinante della matrice di covarianza è il più basso. Informalmente, possono essere considerati come i punti dei dati più vicini tra loro, e quindi meno propensi ad essere anomali. La stima finale della media corrisponderà semplicemente alla media del sottoinsieme, mentre la stima finale della matrice di covarianza sarà proporzionale alla matrice di covarianza del sottoinsieme. La costante di proporzionalità viene scelta per garantire la consistenza dello stimatore. Il metodo è molto popolare grazie alle sue buone proprietà asintotiche, sia per la componente di posizione (Butler et al. (1993)) che per la componente di covarianza (Croux and Haesbroeck (1999); Cator and Lopuhaä (2012)), e grazie alla disponibilità di algoritmi veloci ed efficienti, come il FASTMCD (Rousseeuw and Driessen (1999)).
I livelli di troncamento tipici sono " $\epsilon = .5$" o "$\epsilon = .25$".
Formalmente, sia z un vettore binario tale che $\sum z_i = n(1-\epsilon)$, dove zero indica un'osservazione troncata.
Lo stimatore MCD di posizione e covarianza è

$$\hat{\mu}_{MCD} = \frac{1}{\sum_i z_i \sum_i z_i x_i}$$

$$\hat{\Sigma}_{MCD} = \frac{c(p,\epsilon)}{\sum_i z_i - 1} \sum_i z_i (x_i - \hat{\mu}_{MCD})(x_i - \hat{\mu}_{MCD})^T,$$

dove z è tale che

$$|\hat{\Sigma}_{MCD}(z)| \leq |\hat{\Sigma}_{MCD}(z')|, \forall z'.$$

Il fattore

$$c(p,\epsilon) = \frac{1-\epsilon}{F_{\chi^2_{p+2}}}(q_{p,1-\epsilon}) \tag{10}$$

rende l' MCD consistente al modello Normale inflazionando la stima di covarianza basata sul sottoinsieme selezionato.
In (10), $q_{p,1-\epsilon}$ indica il quantile di livello $1-\epsilon$ di una distribuzione χ^2_p con p gradi di libertà.
È importante notare che un fattore di correzione per campioni di piccole dimensioni può ridurre l'errore quadratico medio di $\hat{\Sigma}_{MCD}$. Questo viene ottenuto in Pison et al. (2002) mediante metodi di approssimazione numerica.
Le proprietà del punto di rottura dello stimatore MCD dipendono dal tasso di *trimming* (tasso di troncatura). Si può dimostrare che l' MCD ha un punto di rottura asintotico $\epsilon^* = \epsilon$. Quando $\epsilon = 50\%$, questo diventa il punto di breakdown (asintotico) più grande possibile per gli stimatori affini equivarianti (Lopuhaa and Rousseeuw (1991); (Agulló et al. (2008)).

Le proprietà di robustezza locale del MCD sono discusse in (Croux and Haesbroeck (1999)).

2.3 Stima robusta della matrice di correlazione

2.3.1 Il coefficiente di correlazione τ-rank di Kendall

Sia $(x_1, y_1), ..., (x_n, y_n)$ un campione di osservazioni delle variabili casuali congiunte X e Y, in cui tutti i valori di x_i e y_i sono unici.
Due coppie di osservazioni (x_i, y_i) e (x_j, y_j) si dicono *concordanti* se i loro ranghi coincidono: cioè se sia $x_i > x_j$ e $y_i > y_j$ o sia $x_i < x_j$ e $y_i < y_j$.
Si dicono discordanti se $x_i > x_j$ e $y_i < y_j$ o $x_i < x_j$ e $y_i > y_j$.
Se $x_i = x_j$ o $y_i = y_j$, la coppia non è nè concordante nè discordante.

Indichiamo con n_+ e n_- il numero di coppie concordanti e discordanti, rispettivamente. Il coefficiente di correlazione di ranghi di Kendall τ (τ-*rank*) è definito come segue (Kendall (1938)):

$$\tau = \frac{n_+ - n_-}{n(n-1)/2}.$$

Poichè il denominatore rappresenta il numero totale di combinazioni di coppie, allora $-1 \leq \tau \leq 1$.
Un'altra modalità per definire τ è data dalla seguente costruzione:

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j).$$

La perfetta concordanza tra due ranghi, cioè quando sono identici, implica $\tau = 1$.
Dalla perfetta discordia tra due ranghi, cioè quando un rango è l'inverso dell'altro, segue che $\tau = -1$.
Se X e Y sono indipendenti, ci aspetteremmo che τ sia approssimativamente zero.

2.3.2 Il coefficiente di correlazione per ranghi di Spearman

Ricordando che il rango $R(x_i)$ di un'osservazione x_i è definito come la posizione della corrispondente statistica d'ordine nella serie variabile $x_{(1)} \leq ... \leq x_{(n)}$, ovvero, se un'osservazione x_i dal campione iniziale $x_1, ..., x_n$ viene trasformata nella statistica d'ordine $x_{(j)}$, allora $R(x_i) = j, 1 \leq j \leq n$.
Questa misura di correlazione (Spearman (1904)) è il coefficiente di correlazione di Pearson tra i ranghi delle osservazioni $\{R(x_i)\}_1^n$ e $\{R(y_i)\}_1^n$.

$$\hat{\rho}_S = \frac{\sum_{i=1}^n [R(x_i) - \overline{R}_x][R(y_i) - \overline{R}_y]}{\left(\sum_{i=1}^n [R(x_i) - \overline{R}_x]^2 \sum_{i=1}^n [R(y_i) - \overline{R}_y]^2\right)^{1/2}}. \tag{11}$$

Poichè i ranghi sono numeri interi, per il calcolo è più conveniente utilizzare la versione trasformata dell'equazione (11) (Kendall and Stuart (1961))

$$\hat{\rho}_S = 1 - \frac{6S(d^2)}{n(n^2-1)}, \qquad S(d^2) = \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2,$$

da ciò segue che per calcolare il coefficiente di correlazione di Spearman è sufficiente ottenere le deviazioni quadratiche dei ranghi.
Poichè il coefficiente di correlazione di Spearman è un caso particolare del coefficiente di Pearson, eredita tutte le sue proprietà, ad esempio, $-1 \leq \hat{\rho}_S \leq 1$.
Se si osserva una corrispondenza diretta tra i ranghi, allora $d_i = 0, i = 1, ..., n$, e quindi $\hat{\rho}_S = 1$.
Al contrario, se si verifica una corrispondenza inversa tra i ranghi, allora $R(x_i) = n + 1 - R(y_i), i = 1, ..., n$, e $\hat{\rho}_S = -1$.
Nella pratica può accadere che due o più elementi campionari non possano essere distinti l'uno dall'altro. In questo caso, a ciascuno di questi elementi viene assegnato il rango corrispondente alla media dei loro numeri di ordine iniziali (Kendall and Stuart (1961)).

2.4 Analisi Fattoriale Robusta

Ora che abbiamo affrontato diversi metodi robusti di stima per le matrici di covarianza e di correlazione, possiamo quindi introdurre la versione robusta dell'Analisi Fattoriale.
Come precedentemente introdotto nell'apertura di questo capitolo, nel Modello Fattoriale Robusto, i parametri vengono stimati utilizzando una matrice di covarianza o correlazione **robusta**, che tiene conto degli effetti di dati anomali e non normalmente distribuiti, a differenza della variante classica, che si basa su una matrice di covarianza o correlazione calcolata in modo standard.
Questa è la principale differenza tra i due approcci, il problema rimane nel capire quale metodo di stima robusto sia più opportuno al nostro dataset, quindi, quale metodo di stima porta a risultati più interpretabili e adeguati dell'analisi fattoriale.

Capitolo 3

Applicazioni

In questo capitolo vedremo nello specifico due applicazioni dei due approcci all'Analisi Fattoriale, la prima su un dataset generato da parametri prefissati, per verificare l'efficacia dei diversi metodi di stima per le matrici di covarianza / correlazione. La seconda applicazione invece riguarda il dataset "Aircraft", preso da (Pison et al. (2003)).

3.1 Simulazione Monte Carlo

L'obiettivo di questa analisi è verificare la sensibilità del modello fattoriale agli outliers, confrontando i risultati ottenuti dall'approccio classico con quelli ottenuti dall'approccio robusto.

3.1.1 Dataset senza outliers

Il nostro dataset è composto da $n = 100$ osservazioni, $p = 10$ parametri, ed è stimato da $m = 2$ fattori.
Partiamo da una matrice dei loadings $L_{p\times k}$ prefissata

$$L_{p\times k} = \begin{bmatrix} 0.6 & 0 \\ 0.3 & 0 \\ 0.4 & 0 \\ 0.6 & 0 \\ 0.2 & 0 \\ 0 & 0.6 \\ 0 & 0.5 \\ 0 & 0.4 \\ 0 & 0.3 \\ 0 & 0.7 \end{bmatrix}$$

e da una matrice diagonale $\Psi_{p\times p}$ con valori generati da una variabile casuale Uniforme sull'intervallo $[0.1, 1.5]$.

Assumendo che i punteggi fattoriali contenuti nella matrice F siano generati da una $\mathcal{N}(0,1)$, i valori presenti nel dataset sono quindi generati da una $\mathcal{N}(0,\Sigma)$, con $\Sigma = LL^T + \Psi$.
Effettuiamo quindi un analisi fattoriale passando alla funzione `factanal` la matrice di covarianza Σ. Memorizziamo la "vera" matrice dei loadings risultataci e continuiamo con un processo Monte Carlo di $B = 10000$ simulazioni. Qui, per ogni simulazione, generiamo un dataset $\mathbf{X}_{n\times p}$, privo di dati anomali, con osservazioni provenienti da una $\mathcal{N}_p(0_p, S)$.
Stimiamo quindi le matrici di covarianza / correlazione con il metodo standard e con i vari metodi robusti discussi nel capitolo precedente (*Stimatore M, Stimatore S, MCD*[1]*)* per la stima della matrice di covarianza e metodi di *Kendall* e *Spearman* per la stima della matrice di correlazione).
Applichiamo la factor analysis con le diverse matrici ottenute e memorizziamo in una serie di sette matrici (una per diverso metodo di stima) la somma dei valori contenuti nelle varie matrici dei loadings per metodo. Una volta terminate tutte le simulazioni, dividiamo il contenuto di ogni matrice per B, ottenendo così una matrice contenente i valori medi dei loadings su B simulazioni.
Poi, creiamo una misura di errore totale prendendo una matrice differenza tra la matrice di loadings medi e la "vera" matrice dei loadings calcolata inizialmente, sommandoci, in valore assoluto, ogni elemento presente nella matrice differenza.
Con questa misura di errore, possiamo ora mettere a confronto i vari metodi di stima delle matrici di covarianza / correlazione in un grafico a barre, distinguendo il valore prodotto dalla stima standard non-robusta e gli altri risultati robusti.

[1] h = *3/4*, come descritto in (Pison et al. (2003)) come migliore h per l'MCD.

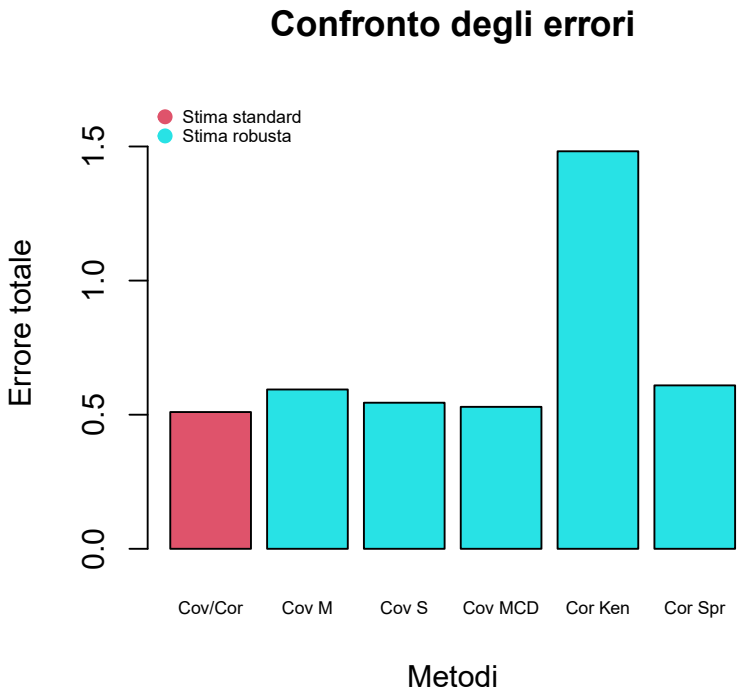

Figura 3.1: Risultati analisi del confronto metodi di stima

Il metodo standard per la stima della matrice di covarianza / correlazione (in entrambi i casi, la factor analysis restituisce ugual valore delle matrici dei loadings, essendo una matrice la trasformazione dell'altra e vice-versa) risulta visibilmente il metodo con minore quantità di errore, per quanto riguarda un dataset privo di outliers. I metodi robusti producono stime con alto valore di errore per il nostro caso, soprattutto i metodi di *Kendall* e *Spearman*.

3.1.2 Dataset con outliers

In questo caso di studio, effettueremo la stessa precedente procedura, ma andremo, per ogni simulazione, a generare dati con presenza di contaminazione.
L'obiettivo è quello di vedere, all'aumentare della percentuale di osservazioni anomale presenti nel dataset, quale metodo di stima produce un errore totale minore, quindi, quale metodo è meno influenzato dalla presenza di outliers nei dati.
Sia il nostro dataset $\mathbf{X}_{n\times p}$, le sue osservazioni $X_1, ..., X_n$, ciascuna composta

da $X_i = (x_{i1}, ..., x_{ip})$, $i = 1, ..., n$ che si distribuice come

$$X_i \sim \begin{cases} \mathcal{N}_p(10_p, S_{p\times p}) & \text{con probabilità } \pi \\ \mathcal{N}_p(0_p, S_{p\times p}) & \text{con probabilità } 1-\pi. \end{cases}$$

Andiamo quindi a replicare 20 scenari di $B = 10000$ simulazioni, andando a aumentare π per ogni replica, partendo da 0.01 e terminando a 0.2.
Analizziamo quindi quale metodo di stima produce matrici di loadings più simili alla matrice "vera" iniziale, tramite la nostra misura di errore, all'aumentare della percentuale di dati contaminati presenti nel nostro dataset generato.

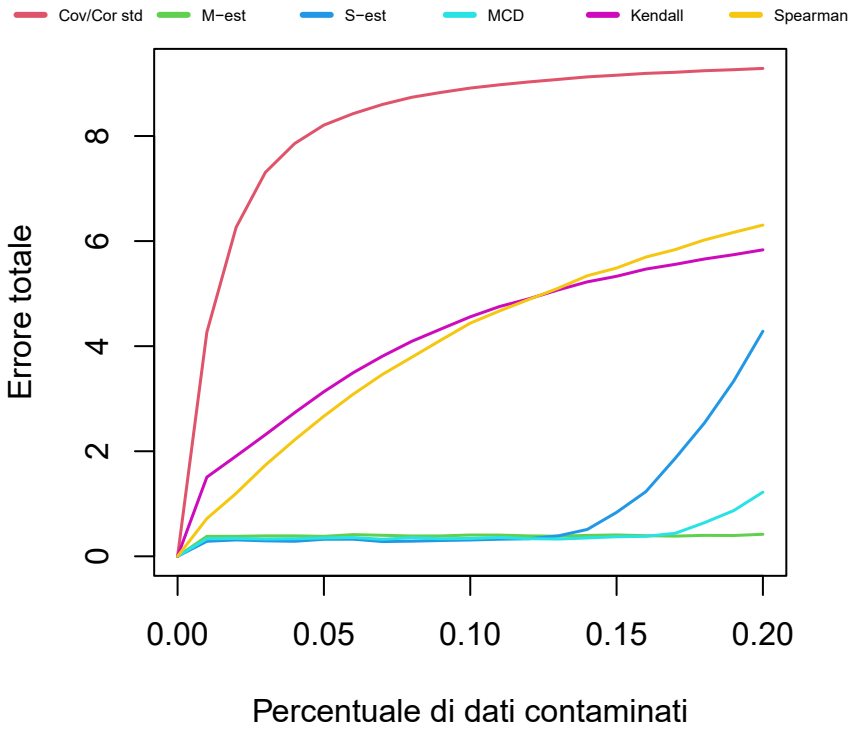

Figura 3.2: Risultati analisi del confronto per diverso metodo di stima, all'aumentare della percentuale di dati contaminati

Come ci si poteva aspettare, gli errori associati ai metodi di stima robusti sembrano variare significativamente meno rispetto al metodo di stima standard non robusto, all'aumentare degli outliers nel dataset. Anche un minimo 1% di dati anomali provoca un aumento massiccio nella misura di errore totale del metodo non robusto.
Notiamo invece come i metodi *Stimatore-M*, *Stimatore-S* e *MCD* siano i più adatti nel nostro scenario, mentre i due metodi robusti di stima della matrice di correlazione producono risultati peggiori, pur rimanendo più

adatti rispetto alla stima tramite metodo standard.
Analizziamo nello specifico come si muovono i tre migliori metodi nella Figura 3.3.

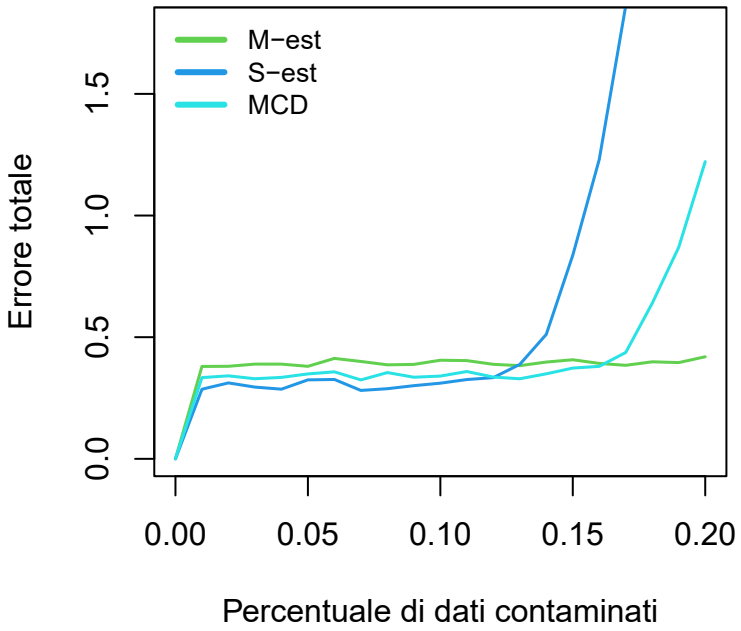

Figura 3.3: Confronto dei tre migliori metodi robusti

Notiamo come, fino a una percentuale di dati contaminati = 12%, la stima prodotta dallo *Stimatore-S* si quella con minore errore totale, seguita dal metodo MCD e infine *Stimatore-M*. Dal 12% di outlier in poi, la misura di errore totale riferita al S-est sale vertiginosamente comparata al suo trend precedente. Stesso discorso vale per l'errore totale del metodo MCD che inizia a salire dal 16% di outlier in poi. Il metodo più solido per lo scenario preso in questione è la stima dello *Stimatore-M*, che non presenta grosse variazioni della sua misura di errore totale dall'inizio alla fine dell'analisi.

3.2 Dataset Aircraft

In questa sezione tratteremo invece un dataset reale, il dataset "Aircraft" (da Pison et al. (2003)).
Il dataset *Aircraft* è composto da $n = 23$ aeromobili monomotore costruiti negli anni 1947-1979. Le $p = 5$ variabili sono il rapporto d'aspetto (*Aspect Ratio*), il rapporto portanza-attrito (*Lift-to-drag ratio*), il peso dell'aereo (*Weight*), la spinta massima (*Maximal thrust*) e il costo (*Cost*). Applicando il MCD (come in Pison et al. (2003)) a questi dati si nota che i casi 14 e 22

sono valori anomali. L'aereo 22 era il modello F-111, costruito per soddisfare contemporaneamente le esigenze dell'Esercito, della Marina e dell'Aeronautica. All'epoca, era l'aereo jet monomotore più sofisticato, veloce, pesante e costoso mai costruito. Tuttavia, presentava molti problemi tecnici. L'aereo 14 era il modello F-104A "Starfighter", caratterizzato da un elevato rapporto portanza-attrito.
Ora procediamo a stimare $m = 2$ fattori sul dataset considerato. Applicando il metodo dei fattori principali (*PFA*) (Johnson and Wichern (2007), pag. 494) alla matrice di correlazione classica, e alla matrice di correlazione robusta basata su MCD, R_n^r, calcolata come:

$$R_n^r = DS_n^r D \qquad \text{con } D = diag(((S_n^r)_{11})^{(1/2)}, ..., (S_n^r)_{pp})^{(1/2)}).$$

La principale differenza tra i due metodi è che nell'analisi fattoriale classica i due outliers influenzano notevolmente S_n^c, L_n^c e F_n^c.
L'analisi fattoriale robusta riduce il peso di questi outliers nell'analisi e fornisce un'immagine più affidabile della maggior parte dei dati.
Confrontiamo i loadings dell'analisi fattoriale classica e quella robusta nella Tabella 3.1.

	L FA classica		*L FA robusta*	
	Fattore 1	**Fattore 2**	**Fattore 1**	**Fattore 2**
Aspect ratio	-0.8462862	0.0655728	-0.1562073	-0.9707605
Lift-to-drag ratio	0.1477898	0.9720913	0.9743873	0.0684479
Weight	0.9247971	0.2518315	0.8576614	0.4764704
Maximal thrust	0.8569184	0.3791438	0.8188761	0.4929290
Cost	0.8708953	0.2072492	0.6168484	0.6939030

Tabella 3.1: Loadings fattoriali dei due diversi metodi di stima

Nel caso classico, il fattore 1 è principalmente una combinazione delle variabili 1 (con coefficiente negativo), 3, 4 e 5, e il fattore 2 è determinato principalmente dalla variabile numero 2.
Nell'analisi fattoriale robusta, il primo fattore è una combinazione positiva delle variabili 2, 3 e 4, mentre il secondo combina principalmente le variabili 1 e 5 (con segni diversi).
In questo esempio, i due metodi forniscono risultati alquanto differenti, lasciando quindi all'analista il lavoro di interpretazione dei due approcci, prendendo una decisione su quale sia quello più ideale, magari confrontandosi con esperti del settore.

Capitolo 4

Conclusioni

L'obbiettivo di questo lavoro è stato quello di introdurre, discutere e mettere a confronto i due differenti approcci dell'analisi fattoriale, soffermandoci sul metodo robusto. La statistica robusta è una importante branca della statistica perchè affronta i problemi derivanti dalla violazione delle assunzioni tipiche dei modelli statistici tradizionali. Mentre i metodi statistici classici si basano sull'assunzione di normalità e sull'omogeneità dei dati, la statistica robusta è progettata per essere meno influenzata da valori anomali o deviazioni dalla distribuzione normale. Questi valori anomali possono verificarsi per diversi motivi, come errori di misurazione, dati mancanti, violazioni delle assunzioni sottostanti o addirittura fenomeni strutturali inusuali. Dalla statistica robusta si derivano diversi metodi specifici per tutte le diverse modalità di analisi dei dati, come in questo caso l'analisi fattoriale. Utilizzando tecniche robuste, è possibile ottenere risultati più stabili e coerenti, evitando che valori anomali influenzino in modo significativo i risultati statistici. Ciò è particolarmente cruciale in ambiti in cui la presenza di outlier o deviazioni può essere comune o avere un impatto significativo sulle conclusioni. Come visto nell'esempio di analisi sul dataset *Aircraft* (da Pison et al. (2003)), i diversi metodi producono risultati piuttosto diversi, poichè il metodo robusto va a dare giusto peso alle due osservazioni anomale presenti nel dataset, e quindi a modificare le successive stime dei loadings. Questo sottolinea l'importanza dei metodi robusti all'interno del grande mondo dell'analisi statistica, i quali possono dare risultati significativamente migliori, o diversi in vari scenari di analisi; sta quindi all'analista capire se e quali metodi robusti usare dipendentemente dallo scenario di studio, dal dataset analizzato e dall'obbiettivo di ricerca in questione.

Appendice A

Codice R

Il codice seguente è il codice usato nelle procedure di simulazione Monte Carlo nella sezione 3.1.

```
#### SIMULAZIONI MONTE CARLO PER DIVERSI METODI DI STIMA ############
library(fit.models)
library(robust)
library(robustbase)
library(rrcov)
library(mvtnorm)

#### PARAMETRI ####
{
  p <- 10
  m <- 2
  n <- 100
  L <- matrix(c(0.6,   0,
                0.3,   0,
                0.4,   0,
                0.6,   0,
                0.2,   0,
                0,   0.6,
                0,   0.5,
                0,   0.4,
                0,   0.3,
                0,   0.7), nrow=p, ncol=m, byrow = T)
  psi <- c(runif(p, 0.1, 1.5))

  S <- L %*% t(L) + diag(psi) # S_FA originale dai parametri fissati

  # FA
  f_0 <- factanal(covmat=S, factors=2, n.obs=n, rotation="none")
  # riordinamento delle colonne
  if (f_0$loadings[1,1] < f_0$loadings[1,2]){
    f_0$loadings[, c(1,2)] <- f_0$loadings[, c(2,1)]
  }
```

```
}

######### MC SENZA OUTLIERS #########
{
  # 10 000 replicazioni
  B= 1E4

  # matrici vuote dei loadings medi per metodo
  {
    LM_C_std <- matrix(rep(0, 20), ncol=m, nrow=p)
    LM_M_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
    LM_S_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
    LM_MCD_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
    LM_R_ken <- matrix(rep(0, 20), ncol=m, nrow=p)
    LM_R_spr <- matrix(rep(0, 20), ncol=m, nrow=p)
  }

  # processo Monte Carlo
  for (b in 1:B){

    # stampo ogni 1000 replicazioni
    if (b%%1E3==0) print(b)

    # genero matrice dei dati
    X <- matrix(rmvnorm(n, rep(0, p), S), ncol=p, nrow=n)

    # metodi di stima
    C_std <- cov(X)                                  # COV STD
    M_rob <- CovMest(X)$cov                          # M-EST
    S_rob <- CovSest(X)$cov                          # S-EST
    MCD_rob <- CovMcd(X, alpha = 3/4)$cov            # MCD
    R_ken <- cor(X, method="kendall")                # KEND
    R_spr <- cor(X, method="spearman")               # SPEAR

    # FA
    f_C_std <- factanal(covmat= C_std, factors=2,
                        n.obs=n, rotation="none")
    f_M_rob <- factanal(covmat= M_rob, factors=2,
                        n.obs=n, rotation="none")
    f_S_rob <- factanal(covmat= S_rob, factors=2,
                        n.obs=n, rotation="none")
    f_MCD_rob <- factanal(covmat= MCD_rob, factors=2,
                        n.obs=n, rotation="none")
    f_R_ken <- factanal(covmat= R_ken, factors=2,
                        n.obs=n, rotation="none")
    f_R_spr <- factanal(covmat= R_spr, factors=2,
                        n.obs=n, rotation="none")

    # riordinamento
```

```
if (f_C_std$loadings[1,1] < f_C_std$loadings[1,2]){
  f_C_std$loadings[, c(1,2)] <- f_C_std$loadings[, c(2,1)]
}
if (f_M_rob$loadings[1,1] < f_M_rob$loadings[1,2]){
  f_M_rob$loadings[, c(1,2)] <- f_M_rob$loadings[, c(2,1)]
}
if (f_S_rob$loadings[1,1] < f_S_rob$loadings[1,2]){
  f_S_rob$loadings[, c(1,2)] <- f_S_rob$loadings[, c(2,1)]
}
if (f_MCD_rob$loadings[1,1] < f_MCD_rob$loadings[1,2]){
  f_MCD_rob$loadings[, c(1,2)] <- f_MCD_rob$loadings[, c(2,1)]
}
if (f_R_ken$loadings[1,1] < f_R_ken$loadings[1,2]){
  f_R_ken$loadings[, c(1,2)] <- f_R_ken$loadings[, c(2,1)]
}
if (f_R_spr$loadings[1,1] < f_R_spr$loadings[1,2]){
  f_R_spr$loadings[, c(1,2)] <- f_R_spr$loadings[, c(2,1)]
}

# somma cumulata dei loadings per metodo di stima
LM_C_std <- LM_C_std + f_C_std$loadings
LM_M_rob <- LM_M_rob + f_M_rob$loadings
LM_S_rob <- LM_S_rob + f_S_rob$loadings
LM_MCD_rob <- LM_MCD_rob + f_MCD_rob$loadings
LM_R_ken <- LM_R_ken + f_R_ken$loadings
LM_R_spr <- LM_R_spr + f_R_spr$loadings

# ultimo passaggio
if (b==B) {

  # calcolo matrici dei loadings medi
  LM_C_std <- LM_C_std / B
  LM_M_rob <- LM_M_rob / B
  LM_S_rob <- LM_S_rob / B
  LM_MCD_rob <- LM_MCD_rob / B
  LM_R_ken <- LM_R_ken / B
  LM_R_spr <- LM_R_spr / B

  # calcolo errore per metodo di stima
  E1_C_std <- round(sum(abs(f_0$loadings - LM_C_std)), 4)
  E1_M_rob <- round(sum(abs(f_0$loadings - LM_M_rob)), 4)
  E1_S_rob <- round(sum(abs(f_0$loadings - LM_S_rob)), 4)
  E1_MCD_rob <- round(sum(abs(f_0$loadings - LM_MCD_rob)), 4)
  E1_R_ken <- round(sum(abs(f_0$loadings - LM_R_ken)), 4)
  E1_R_spr <- round(sum(abs(f_0$loadings - LM_R_spr)), 4)

  # risultati
  cat("\nErrore per metodo di stima: \n  Covarianza standard: ",
      E1_C_std,
```

```
          "\n  Covarianza robusta M: ", E1_M_rob,
          "\n  Covarianza robusta S: ", E1_S_rob,
          "\n  Covarianza robusta MCD: ", E1_MCD_rob,
          "\n  Correlazione robusta di Kendall: ", E1_R_ken,
          "\n  Correlazione robusta di Spearman: ", E1_R_spr)
      errori <- c(E1_C_std, E1_M_rob, E1_S_rob,
                  E1_MCD_rob, E1_R_ken, E1_R_spr)
      barplot(errori, xlab="Metodi", ylab="Errore totale",
              cex.names =.6, ylim = c(0, max(errori)+0.2),
              names.arg = c("Cov/Cor", "Cov M", "Cov S", "Cov MCD",
                            "Cor Ken", "Cor Spr"),
              main="Confronto degli errori", col=c(2, rep(5, 5)))
      legend("topleft", cex=.6, bty="n", pt.lwd=5, pch = 20,
             col = c(2, 5), legend = c("Stima standard",
                                       "Stima robusta"))
    }
  }
}

######### MONTE CARLO CON OUTLIERS #########

# funzione di generazione dati contaminati
gen <- function(num){
  if (num==0) rmvnorm(1, rep(0, p), S)
  else rmvnorm(1, rep(10, p), S)
}

{
  # 10 000 replicazioni
  B= 1E4

  # vettori per errori per metodo di stima
  {
    E_C_std <- rep(0, 21)
    E_M_rob <- rep(0, 21)
    E_S_rob <- rep(0, 21)
    E_MCD_rob <- rep(0, 21)
    E_R_ken <- rep(0, 21)
    E_R_spr <- rep(0, 21)
  }

  for (o in c(.01, .02, .03, .04, .05, .06, .07, .08,
              .09, .1, .11, .12, .13, .14, .15, .16,
              .17, .18, .19, .2)){
    cat("\n", o, "\n")

    # matrici vuote dei loadings medi per metodo
    {
      LM_C_std <- matrix(rep(0, 20), ncol=m, nrow=p)
```

```
  LM_M_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
  LM_S_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
  LM_MCD_rob <- matrix(rep(0, 20), ncol=m, nrow=p)
  LM_R_ken <- matrix(rep(0, 20), ncol=m, nrow=p)
  LM_R_spr <- matrix(rep(0, 20), ncol=m, nrow=p)
}

# processo Monte Carlo
for (b in 1:B){

  # stampo ogni 1000 replicazioni
  if (b%%1E3==0) print(b)

  # genero matrice dei dati contaminati
  contam <- rbinom(n, 1, o)
  X <- matrix(sapply(contam, gen), ncol=p, nrow=n, byrow=T)

  # metodi di stima
  C_std <- cov(X)                                  # COV STD
  M_rob <- CovMest(X)$cov                          # M-EST
  S_rob <- CovSest(X)$cov                          # S-EST
  MCD_rob <- CovMcd(X, alpha = 3/4)$cov            # MCD
  R_ken <- cor(X, method="kendall")                # KEND
  R_spr <- cor(X, method="spearman")               # SPEAR

  # FA
  f_C_std <- factanal(covmat= C_std, factors=2,
                      n.obs=n, rotation="none")
  f_M_rob <- factanal(covmat= M_rob, factors=2,
                      n.obs=n, rotation="none")
  f_S_rob <- factanal(covmat= S_rob, factors=2,
                      n.obs=n, rotation="none")
  f_MCD_rob <- factanal(covmat= MCD_rob, factors=2,
                        n.obs=n, rotation="none")
  f_R_ken <- factanal(covmat= R_ken, factors=2,
                      n.obs=n, rotation="none")
  f_R_spr <- factanal(covmat= R_spr, factors=2,
                      n.obs=n, rotation="none")

  # riordinamento
  if (f_C_std$loadings[1,1] < f_C_std$loadings[1,2]){
    f_C_std$loadings[, c(1,2)] <- f_C_std$loadings[, c(2,1)]
  }
  if (f_M_rob$loadings[1,1] < f_M_rob$loadings[1,2]){
    f_M_rob$loadings[, c(1,2)] <- f_M_rob$loadings[, c(2,1)]
  }
  if (f_S_rob$loadings[1,1] < f_S_rob$loadings[1,2]){
    f_S_rob$loadings[, c(1,2)] <- f_S_rob$loadings[, c(2,1)]
  }
```

```
    if (f_MCD_rob$loadings[1,1] < f_MCD_rob$loadings[1,2]){
      f_MCD_rob$loadings[, c(1,2)] <- f_MCD_rob$loadings[, c(2,1)]
    }
    if (f_R_ken$loadings[1,1] < f_R_ken$loadings[1,2]){
      f_R_ken$loadings[, c(1,2)] <- f_R_ken$loadings[, c(2,1)]
    }
    if (f_R_spr$loadings[1,1] < f_R_spr$loadings[1,2]){
      f_R_spr$loadings[, c(1,2)] <- f_R_spr$loadings[, c(2,1)]
    }

    # somma cumulata dei loadings per metodo di stima
    LM_C_std <- LM_C_std + f_C_std$loadings
    LM_M_rob <- LM_M_rob + f_M_rob$loadings
    LM_S_rob <- LM_S_rob + f_S_rob$loadings
    LM_MCD_rob <- LM_MCD_rob + f_MCD_rob$loadings
    LM_R_ken <- LM_R_ken + f_R_ken$loadings
    LM_R_spr <- LM_R_spr + f_R_spr$loadings

    # ultimo passaggio
    if (b==B){

      # calcolo matrici dei loadings medi
      LM_C_std <- LM_C_std / B
      LM_M_rob <- LM_M_rob / B
      LM_S_rob <- LM_S_rob / B
      LM_MCD_rob <- LM_MCD_rob / B
      LM_R_ken <- LM_R_ken / B
      LM_R_spr <- LM_R_spr / B

      # calcolo errore per metodo di stima
      E_C_std[o*100+1] <- round(sum
                                (abs(f_0$loadings - LM_C_std)), 4)
      E_M_rob[o*100+1] <- round(sum
                                (abs(f_0$loadings - LM_M_rob)), 4)
      E_S_rob[o*100+1] <- round(sum
                                (abs(f_0$loadings - LM_S_rob)), 4)
      E_MCD_rob[o*100+1] <- round(sum
                               (abs(f_0$loadings - LM_MCD_rob)), 4)
      E_R_ken[o*100+1] <- round(sum
                                (abs(f_0$loadings - LM_R_ken)), 4)
      E_R_spr[o*100+1] <- round(sum
                                (abs(f_0$loadings - LM_R_spr)), 4)

    }
  }
}

# mat 6x21 degli errori, ogni riga un metodo di stima differente
ERRORI <- matrix(rbind(c(E_C_std, E_M_rob, E_S_rob, E_MCD_rob,
```

```
                          E_R_ken, E_R_spr)),
                  ncol=21, nrow=6, byrow=T)
  rownames(ERRORI) <- c("E_C_std", "E_M_rob", "E_S_rob",
                        "E_MCD_rob", "E_R_ken", "E_R_spr")
  colnames(ERRORI) <- seq(0, 0.2, by=0.01)

  # grafico con tutti i metodi
  {
    plot(seq(0, 0.2, by = 0.01), ERRORI[1,],
         xlab="Percentuale di dati contaminati",
         ylab="Errore totale", type="l", col=2, lwd=1.5,
         ylim=c(0, max(ERRORI[1,])))
    lines(seq(0, 0.2, by = 0.01), ERRORI[2,], col=3, lwd=1.5)
    lines(seq(0, 0.2, by = 0.01), ERRORI[3,], col=4, lwd=1.5)
    lines(seq(0, 0.2, by = 0.01), ERRORI[4,], col=5, lwd=1.5)
    lines(seq(0, 0.2, by = 0.01), ERRORI[5,], col=6, lwd=1.5)
    lines(seq(0, 0.2, by = 0.01), ERRORI[6,], col=7, lwd=1.5)
    legend("topleft", inset = c(-0.24,-0.1),
           legend = c("Cov/Cor std", "M-est", "S-est", "MCD",
                      "Kendall", "Spearman"), xpd = TRUE,
           horiz = TRUE, col = seq(2, 7, by=1), cex=0.5, lty = 1,
           lwd=3, bty = "n")
  }

  # grafico con i migliori tre
  {
    plot(seq(0, 0.2, by = 0.01), ERRORI[1,],
         xlab="Percentuale di dati contaminati",
         ylab="Errore totale", type="l", col=0, lwd=1.5,
         ylim=c(0, max(ERRORI[3,])-2.5))
    lines(seq(0, 0.2, by = .01), ERRORI[2,], col=3, lwd=1.5)
    lines(seq(0, 0.2, by = .01), ERRORI[3,], col=4, lwd=1.5)
    lines(seq(0, 0.2, by = .01), ERRORI[4,], col=5, lwd=1.5)
    abline(h = max(ERRORI[3,])-2.429, lwd=1, col="black")
    legend("topleft", inset = c(0,0), lty = 1, lwd=3, bty = "n",
           cex=0.85, legend = c("M-est", "S-est", "MCD"),
           col = seq(3, 5, by=1))
  }
}
```

Successivamente, il codice utilizzato nell'analisi del dataset *Aircraft* nella sezione 3.2.

```
##### ANALISI DATASET REALE "AIRCRAFT", FA CLASSICA E ROBUSTA #####

library(FactoMineR)
library(psych)
library(fit.models)
library(robust)
library(robustbase)
library(rrcov)

# dataset

data("aircraft")
air <- as.data.frame(aircraft)
variabili <- c("aspect ratio", "lift-to-drag ratio",
               "weight", "maximal thrust", "cost")
colnames(air) <- variabili

# PFA

# matrice correlazione su MCD
MCD <- CovMcd(air, alpha=3/4)$cov
R_MCD <- diag(diag(MCD)^(-1/2)) %*% MCD %*% diag(diag(MCD)^(-1/2))

# FA su matrici di correlazione
f <- principal(cor(air), n.obs=23, nfactors=2, rotate="varimax")
fR <- principal(R_MCD, n.obs=23, nfactors=2, rotate="varimax")

l <- f$loadings
lR <- fR$loadings
x <- l[,1]
y <- l[,2]
xR <- lR[,1]
yR <- lR[,2]

# grafici di confronto direzione loadings FA e FA robusta
par(mfrow=c(1,2))
plot(0, 0, type = "n", xlim = c(-1.15,1.15), ylim = c(-1, 1.2),
     xlab = "Factor 1", ylab = "Factor 2",
     main="Loadings FA Classica")
arrows(0, 0, x, y, length = 0.1, col="darkred", lwd=2)
text(x[c(-3,-5)], y[c(-3,-5)], variabili[c(-3,-5)], pos = 3,
     cex=.69, col="darkred")
text(x[3]-.1, y[3]+.1, variabili[3], pos = 4, cex=.7, col="darkred")
text(x[5]-.1, y[5]-.1, variabili[5], pos = 4, cex=.7, col="darkred")

plot(0, 0, type = "n", xlim = c(-1.15,1.15), ylim = c(-1, 1.2),
     xlab = "Factor 1", ylab = "Factor 2",
```

```
     main="Loadings FA Robusta")
arrows(0, 0, xR, yR, length = 0.1, col="darkred", lwd=2)
text(xR[c(-2,-4)], yR[c(-2,-4)], variabili[c(-2,-4)], pos = 4,
     cex=.7, col="darkred")
text(xR[2]-.45, yR[2]+.1, variabili[2], pos = 4, cex=.7,
     col="darkred")
text(xR[4]-0.24, yR[4]+.08, variabili[4], pos = 4, cex=.7,
     col="darkred")
```

Bibliografia

Agulló, J., Croux, C., and Van Aelst, S. (2008). The multivariate least-trimmed squares estimator. *Journal of Multivariate Analysis*, 99(3):311–338.

Butler, R., Davies, P., and Jhun, M. (1993). Asymptotics for the minimum covariance determinant estimator. *The Annals of Statistics*, pages 1385–1400.

Cator, E. A. and Lopuhaä, H. P. (2012). Central limit theorem and influence function for the mcd estimators at general multivariate distributions.

Costello, A. B. and Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, 10(1):7.

Croux, C. and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2):161–190.

Davies, P. L. (1987). Asymptotic behaviour of s-estimates of multivariate location parameters and dispersion matrices. *The Annals of Statistics*, pages 1269–1292.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. *A festschrift for Erich L. Lehmann*, 157184.

Farcomeni, A., Greco, L., and LLC., C. P. (2015). *Robust Methods for Data Reduction*. CRC Press, Taylor & Francis Group.

Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *The Annals of Mathematical Statistics*, 42(6):1887 – 1896.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the american statistical association*, 69(346):383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., Stahel, W., and Statistics, R. (1986). The approach based on influence functions. *Robust Statistics. Wiley*.

Harmon, H. (1976). Modern factor analysis (3rd eds). *Chicago, University of Chiacago Process*.

Hodges Jr, J. L. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 163–186.

Huber, P. and Ronchetti, E. (1981). Robust statistics, ser. *Wiley Ser Probab Math Stat New York, NY, USA Wiley-IEEE*, 52:54.

Huber, P. J. and Ronchetti, E. M. (2009). Robust statistics, 2nd edn. hoboken. *NJ: Wiley, doi*, 10:9780470434697.

Johnson, R. A. and Wichern, D. W. (2007). Applied multivariate statistical analysis. 6th. *New Jersey, US: Pearson Prentice Hall*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kendall, M. G. and Stuart, A. (1961). *The advanced theory of statistics: Inference and relationship. Vol. 2*, volume 2. C. Griffin.

Lopuha, H. (1989). On the rdation between s-estimators awl m-estimators ofmultivariali˜ location and covariance. *Annals of Statistics*, 17:l662–168.

Lopuhaa, H. P. and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics*, pages 229–248.

Maronna, R. A. (1976). Robust m-estimators of multivariate location and scatter. *The annals of statistics*, pages 51–67.

Pison, G., Rousseeuw, P. J., Filzmoser, P., and Croux, C. (2003). Robust factor analysis. *Journal of Multivariate Analysis*, 84(1):145–172.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for lts and mcd. *Metrika*, 55:111–123.

Riani, M., Cerioli, A., and Torti, F. (2014). On consistency factors and efficiency of robust s-estimators. *Test*, 23:356–387.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8(283-297):37.

Rousseeuw, P. J. and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.