PSP6075525 - Testing psicologico
Modelli e metodi statistici per la misurazione in psicologia

ESERCIZI

Versione: 4 gennaio 2023

Esercizio 1. Si consideri il dataset `Chen` della libreria `psych`. Dopo aver compreso la natura del dataset e il suo formato (si esegua il comando `?psych::Chen` per ulteriori dettagli sul dataset in questione):

1. Si rappresenti graficamente la matrice di correlazione associata al dataset.
2. Sulla base dei risultati precedenti, si definisca un modello CFA unidimensionale ($n = 403$) e lo si adatti mediante `lavaan` utilizzando la metrica UVI.
3. Si estraggano le matrici dei parametri (standardizzati) del modello adattato al punto precedente e le si interpretino (per quanto concerne la matrice dei coefficienti $\widehat{\mathbf{\Lambda}}$ si utilizzi la regola euristica $\lambda_{jk} \geq 0.35$).
4. Sulla base dei risultati al punto 1, si definisca un modello CFA ortogonale a tre dimensioni ($n = 403$) e lo si adatti mediante `lavaan` utilizzando la metrica UVI. Si ricordi che nel caso ULI un modello CFA è di tipo ortogonale quando $\mathbf{\Phi} = \mathbf{I}$.
5. Si interpreti il risultato del modello adattato mediate utilizzo dei parametri stimati standardizzati e si effettui una rappresentazione grafica dello stesso.
6. Si effettuino, se necessario, una serie di considerazioni teoriche circa l'uso di un modello CFA di tipo ortogonale nella quantificazione di misurandi utilizzati nella ricerca psicosociale.

Esercizio 2. Si consideri il dataset `attitude` della libreria `datasets`. Dopo aver escluso la prima variabile (si lavori dunque su un dataset ridotto), si faccia quanto segue:

1. Si esegua un'analisi descrittiva del dataset e si valuti per ciascuna variabile gli indici di asimmetria e curtosi, interpretandone i risultati. Si consideri l'uso della funzione `describe([nome del dataset])` della libreria `psych`.
2. Si valuti mediante procedura dello split-half l'attendibilità della scala formata dalle variabili del dataset ridotto. Tale calcolo può essere effettuato mediante l'utilizzo della funzione `split_half(data = [nome del dataset])` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.
3. Si calcoli il punteggio totale $\tilde{\mathbf{y}}_{\text{tot}}$ della scala formata dalle variabili del dataset ridotto applicando la formula
$$\hat{y}^{(i)}_{\text{tot}} = \rho y^{(i)}_{\text{grezzo}} + (1 - \rho)\overline{y}_{\text{grezzo}}$$
dove $\mathbf{y}_{\text{grezzo}}$ è il punteggio totale grezzo ottenuto per somma sugli item che compongono la scala, $\overline{y}_{\text{grezzo}}$ è la media di quest'ultimo, ρ è l'attendibilità della scala ottenuta mediante procedura di split-half.

4. Si calcoli il punteggio totale $\tilde{\mathbf{y}}_{\text{tot}}$ della scala formata dalle variabili del dataset ridotto (si veda punto precedente) dove ρ è calcolato mediante alpha di Cronbach. Nota: l'alpha di Cronbach può essere calcolato mediante la funzione `alpha([nome del dataset])` della libreria `psych` oppure mediante la funzione`coef_alpha([nome del dataset])` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

5. Si usi una rappresentazione grafica idonea per visualizzare la distribuzione di $\tilde{\mathbf{y}}_{\text{tot}}$ ottenuto ai punti 3-4. Si confrontino le distribuzioni grafiche ottenute con quella del punteggio totale semplice $\mathbf{y}_{\text{grezzo}}$.

6. Si commenti il risultato del punto precedente anche alla luce della matrice di correlazione delle variabili del dataset ridotto. Un'altra rappresentazione grafica della matrice di correlazione può essere ottenuta mediante la funzione `pairs([nome del dataset])`.

Esercizio 3. La cartella compressa `mach.zip`, disponibile nella cartella *Datasets* della pagina Moodle del corso, contiene il dataset `mach.Rdata` e il file testuale `codebook.txt` che fornisce informazioni sul contenuto delle variabili di quest'ultimo. Dopo aver decompresso la cartella, si importi il dataset in `R` e si eseguano le analisi seguenti:

1. Si esegua un'analisi descrittiva delle variabili presenti nel dataset (a tal fine si può utilizzare la semplice funzione `summary()`).

2. Si estragga il dataset corrispondente alle sole unità statistiche per le quali `country=="IT"` (partecipanti italiani).

3. Sul dataset ottenuto al punto precedente, si definisca e si adatti un modello CFA unidimensionale.

4. Si valuti l'adattamento del modello unidimensionale ai dati e si interpretino i parametri stimati del modello.

5. Si calcoli l'indice di attendibilità ω per il modello unidimensionale adattato e se ne interpreti il risultato. Nota: l'indice ω può essere calcolato mediante la funzione `reliability([nome modello adattato])` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

6. Sulla base del contenuto semantico delle variabili osservate (si veda a tal fine il file `codebook.txt`), si definisca un secondo modello CFA con $q > 1$ e lo si adatti ai dati.

7. Si confronti in termini di adattamento complessivo il modello multidimensionale ottenuto al punto precedente con quello unidimensionale e si effettui la scelta del modello migliore. Di quest'ultimo si valuti l'attendibilità mediante indice ω e se ne interpretino i parametri anche alla luce del contenuto semantico degli item.

Esercizio 4. Il file `data_ex4.csv` contiene i dati relativi ad un questionario composto da $p = 15$ variabili osservate costruito per rilevare la soddisfazione complessiva dei clienti di una nota compagnia area. Ciascuna delle variabili osservate è rilevata mediante scala categoriale ordinale a cinque livelli (`1: completamente insoddisfatto`, ..., `5: completamente soddisfatto`). Dopo aver importato il file in `R`, si svolga quanto segue:

1. Si definisca e si adatti ai dati un modello CFA unidimensionale.

2. Si definisca e si adatti ai dati un modello CFA a tre dimensioni a struttura semplice ($\Phi = \mathbf{I}$) secondo la seguente assegnazione:

   - $\eta_1 \rightarrow$ `item1-item5`
   - $\eta_2 \rightarrow$ `item6-item10`
   - $\eta_3 \rightarrow$ `item11-item15`

3. Si definisca e si adatti ai dati un modello CFA a tre dimensioni non ortogonali secondo l'assegnazione precedente.

4. Si valutino i modelli adattati ai punti precedenti e si scelga il migliore.

5. Si interpreti il modello finale scelto e si dica se è attendibile o meno per la quantificazione della soddisfazione complessiva.

6. Si valuti se il modello finale scelto possa essere ulteriormente migliorato mediante una procedura statistica razionale.

Esercizio 5. Il file `data_ex5.Rdata` contiene la statistica $\mathbf{S}_{10\times 10}$ relativa alla somministrazione di un questionario ad un campione casuale semplice di $n = 1250$ unità. Con l'obiettivo di studiare la dimensionalità del questionario somministrato, si importi il dataset in `R` e si svolga quanto segue:

1. Si definisca e si adatti ai dati un modello CFA a due dimensioni ortogonali secondo la seguente assegnazione:

   - $\eta_1 \rightarrow$ `item1-item5`
   - $\eta_2 \rightarrow$ `item6-item10`

   Nell'adattare il modello ai dati si scelga la metrica ULI.

2. Si valuti l'adattamento complessivo del modello e in caso lo si migliori utilizzando una procedura razionale.

3. Si interpreti il modello finale migliorato (si utilizzino le stime standardizzate) e si calcoli l'attendibilità delle scale.

Esercizio 6. Si riprenda l'Esercizio 4. Dopo aver definito e adattato ai dati i tre modelli CFA li si valutino sulla base del loro errore di previsione e si scelga il modello con errore di previsione minore (si veda la lezione di didattica integrativa: `07-cross-validation`). Nota: si consideri la validazione incrociata con due soli insiemi (stima/verifica). Si verifichi se la scelta del modello effettuato mediante convalida incrociata differisca da quella effettuata mediante adattamento complessivo ai dati (Esercizio 4) e, nel caso, si giustifichi tale differenza.

Esercizio 7. Il file `rse.RData` contiene i dati raccolti su $n = 1000$ partecipanti relativi al test *Rosenberg Self-Esteem Scale* formato da $p = 10$ item definiti mediante scala ordinale a quattro livelli. Dopo aver importato il dataset in `R` si svolga quanto di seguito riportato:

1. Si dividi il dataset in due parti secondo la seguente proporzione: 60% sottoinsieme A, 40% sottoinsieme B. Si utilizzi per semplicità la funzione `split_dataset()` con `prop=0.60` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

2. Sul sottoinsieme A si faccia un'analisi esplorativa che includa:

   (a) l'esplorazione grafica della matrice di correlazione (quest'ultima opportunamente calcolata rispetto al tipo di variabili osservate)

   (b) l'individuazione dei raggruppamenti delle variabili osservate mediante clustering gerarchico. A tal fine, si suggerisce l'utilizzo della funzione `hclust(...,method = "ward.D2")`.

3. Sul sottoinsieme B si conduca un'analisi confermativa. Si definiscano e si adattino ai dati tre modelli di analisi fattoriale confermativa - idonei rispetto al tipo di variabili osservate - definiti come segue:

- modello 1: definito sulla base dei risultati del punto 2.b
- modello 2: congenerico secondo la TCT
- modello 3: generalizzazione del modello 1 che includa un fattore di secondo ordine

4. Dopo aver valutato la convergenza dell'algoritmo di stima, si individui il modello al punto 3 che meglio si adatta ai dati (a tal fine si utilizzino idonee misure di adattamento globale).

5. Si valuti quale dei tre modelli al punto 3 presenta minore errore di previsione. In particolare si utilizzi la validazione incrociata con 10 sottoinsiemi e calcolo dell'errore mediante metodo Monte Carlo.
   Si suggerisce l'utilizzo della funzione `kFold_validation(...,nfold=10,error="montecarlo")`).

6. Si scelga il modello finale sulla base di attente valutazioni dei punti 4-5. Si commenti estensivamente il modello finale (anche mediante l'ausilio della standardizzazione dei parametri stimati).

7. Si calcoli l'attendibilità del modello scelto al punto precedente (si valuti anche l'attendibilità totale del modello) e si interpreti il risultato ottenuto.

Esercizio 8. Si consideri l'esercizio precedente. Si proponga l'analisi della dimensionalità dello stesso dataset mediante la definizione di un modello CFA basato sull'analisi semantica degli item. Si consulti la pagina web `https://github.com/cddesja/hemp/blob/master/man/rse.Rd` per ulteriori dettagli circa gli item utilizzati nel test. Si confronti il modello CFA così ottenuto rispetto a quello scelto al punto precedente e si valuti se questo presenti (i) una migliore capacità predittiva, (ii) una maggiore o minore attendibilità.

Esercizio 9. Il dataset `dataex9.csv` contiene i dati relativi alla somministrazione di un test composto da $p = 6$ variabili osservate (continue) ad un campione di $n = 3500$ unità statistiche. Il dataset inoltre contiene la variabile categoriale `z` che codifica il genere del partecipante al test. Dopo aver opportunamente importato il dataset in `R`, si svolga quanto segue:

1. Si definiscano tre modelli CFA come di seguito specificato:
   (a) $\eta \rightarrow$ `item1-item6`
   (b) $\eta_1 \rightarrow$ `item1-item4`, $\eta_2 \rightarrow$ `item5-item6`
   (c) modello *bi-factor* sulla base del modello (b)

2. Si valuti quale dei tre modelli si adatta meglio ai dati osservati e lo si scelga per le successive analisi.

3. Si valuti se il modello scelto al punto precedente è anche quello che presenta minore errore di previsione rispetto agli altri modelli adattati.

4. Sulla base del modello scelto al punto precedente si consideri se una sua generalizzazione come modello sovraordinato - qualora tale generalizzazione abbia senso - possa adattarsi meglio ai dati.

5. Si valuti se il modello sovraordinato e il modello scelto al punto 3 presenta maggiore o minore errore di previsione e si scelga un modello per le analisi successive.

6. Si calcoli l'attendibilità del modello scelto al punto precedente e si commenti tale risultato.

7. Si stimino i punteggi fattoriali $\hat{\boldsymbol{\eta}}$ del modello scelto al punto 5. Se il modello scelto è almeno bidimensionale, si calcolino i seguenti punteggi trasformati:

$$\hat{\boldsymbol{\zeta}} = (\hat{\boldsymbol{\eta}}_{n\times q}\hat{\boldsymbol{\Psi}}_{q\times q})\mathbf{1}_{q\times 1}$$

   dove $\mathbf{1}_{q\times 1}$ è un vettore di lunghezza q di tutti 1. Se il modello scelto è unidimensionale si stimino allora solo i punteggi fattoriali del modello adattato.

8. Si definisca e si adatti ai dati un modello di regressione lineare per valutare se la variabile $\hat{\boldsymbol{\zeta}}$ vari in funzione della variabile `z`. Si commenti il risultato ottenuto.

Esercizio 10. Il dataset `taylor_test.rda` contiene i dati relativi alla somministrazione del Taylor Anxiety Test[1] a un campione di $n = 4468$ unità statistiche. Il test è originariamente composto da $p = 50$ variabili manifeste di tipo Booleano (il valore 1 indica che il partecipante ha raggiunto il livello critico di ansia, 0 altrimenti). L'obiettivo è quello di valutare la dimensionalità del test di Taylor. Dopo aver importato in `R` il dataset, si svolga quanto di seguito riportato:

1. Si faccia una divisione a metà del test e si ottengano un dataset di *training* ($\mathbf{Y}_A$) e uno di *test* ($\mathbf{Y}_B$). Il dataset $\mathbf{Y}_A$ deve contenere il 60% di unità statistiche del dataset originale.

2. Si consideri il dataset $\mathbf{Y}_A$. Si calcoli la matrice di distanza idonea per il tipo di dato a disposizione e si esegua su di essa un clustering gerarchico con metodo `ward.D2`.[2]. Si analizzi l'output del clustering con l'obiettivo di individuare un numero congruo di raggruppamenti. Si utilizzino questi ultimi per definire un primo modello CFA.[3]

3. Si esegua un clustering gerarchico mediante metodo dei centroidi (si ottiene con la sintassi: `hclust(...,method="centroid")`) sulla matrice di distanze calcolata al punto precedente e si valuti l'output alla ricerca di un numero congruo di raggruppamenti. Si utilizzino questi ultimi per definire un secondo modello CFA.

4. Si adattino ai dati i due modelli CFA definiti ai punti 3-4 sulla metà rimanente del dataset originale (i.e., $\mathbf{Y}_B$). Nota: le variabili dicotomiche a disposizione possono essere considerate categoriali.

5. Si valuti (a) l'adattamento ai dati e la (b) capacità predittica dei due modelli adattati al punto precedente e si faccia una scelta razionale rispetto al modello migliore.

6. Si interpretino le matrici stimate del modello scelto al punto precedente.[4] Si calcoli infine l'attendibilità e la si interpreti.

7. Si calcolino i *factor scores* $\hat{\boldsymbol{\eta}}$ del modello scelto al punto 5. Successivamente si calcoli la matrice di distanze di $\hat{\boldsymbol{\eta}}$ e la si utilizzi come input per un clustering gerarchico con metodo `ward.D2`. Si individuino un numero congruo di raggruppamenti delle unità statistiche rappresentate sulle righe di $\hat{\boldsymbol{\eta}}$ e si usino tali gruppi per distinguere le unità statistiche tra loro in una idonea rappresentazione grafica.[5] Si interpreti il risultato alla luce dell'informazione espressa da $\hat{\boldsymbol{\eta}}$.

Esercizio 11. Si consideri l'esercizio 9. Dopo aver scelto un opportuno modello CFA per il dataset `dataex9.csv` (punto 5), si valuti quale livello di invarianza il modello di misura raggiunge rispetto alla variabile categoriale `z` (genere).

[1] `https://en.wikipedia.org/wiki/Taylor_Manifest_Anxiety_Scale`. Informazioni sugli item sono reperibili al seguente link: `https://github.com/cran/edmdata/blob/master/man/items_taylor_manifest_anxiety_scale.Rd`

[2] Per il calcolo della distanza per dati dicotomici si può specificare la seguente sintassi: `dist(...,method="binary")`.

[3] Per trasformare l'output del clustering gerarchico in un modello lavaan direttamente si può utilizzare la funzione `hclust2lavaan(three,ngroups)` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso (il parametro `tree` indica l'output della funzione `hclust(..)`, `ngroups` il numero di raggruppamenti desiderati).

[4] Poiché $p = 50$ si suggerisce di utilizzare la funzione `filterLambda(Lambda_est,thr)` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso. Essa permette di estrarre dalla matrice $\hat{\boldsymbol{\Lambda}}$ (`Lambda_est`) quei coefficienti fattoriali la cui magnitudine è maggiore o uguale al valore `thr`.

[5] Per $q = 2$ una possibile rappresentazione grafica è il grafico che si ottiene mediante il comando `plot(Eta[,1],Eta[,2],pch=20,col=gps)` dove `gps=cutree(hc,k)` con `hc` contenente l'output del clustering gerarchico e `k` il numero di raggruppamenti individuati.

Esercizio 12. Il file `finance.Rdata` contiene $p = 10$ item del questionario *Financial Well-Being Scale* che quantifica la qualità percepita della vita sulla base del proprio equilibrio finanziario. Il test è stato somministrato ad un campione di $n = 5000$ partecipanti di cui $n_M = 2668$ di genere maschile. Il dataset contiene gli item del questionario (denotati con il codice iniziale `FWB1` o `FWB2`), il punteggio totale al test (`FWBscore`) e la variabile categoriale per il genere (`PPGENDER` $\in \{1 : M,\ 2 : F\}$). Dopo aver importato il dataset in `R` si svolga quanto segue:

1. Si esegua una divisione del dataset a metà riservando al primo di questi il 30% delle unità statistiche.

2. Sul primo dataset ottenuto al punto precedente si esegua un'analisi esplorativa della matrice di correlazione tra gli item del test. Si faccia un clustering gerarchico con metodo `ward.D2`, si scelga un numero congruo di raggruppamenti e si definisca un modello CFA secondo il risultato ottenuto.

3. Sul secondo dataset ottenuto al punto 1, si valuti se il test FWBS raggiunga il livello di invarianza forte (*scalar invariance*) rispetto alla variabile `PPGENDER`. Si valuti la possibilità di effettuare anche un'analisi dell'invarianza parziale qualora il test non raggiungesse uno dei livelli intermedi. Nota: per facilitare la scelta dei parametri da liberare nell'eventuale invarianza parziale si usi la funzione `evaluate_table(fitted_model,type)`[6] presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

4. Si interpreti il modello CFA invariante per `PPGENDER` ottenuto al punto precedente. Nota: per estrarre facilmente parti dell'output ottenuto tramite la funzione `summary(..)` si può utilizzare la funzione `summary_table(fitted_model,type_summary)`[7] presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

[6] La funzione richiede come primo elemento di input (`fitted_model`) il modello CFA adattato al livello in cui l'invarianza completa si è arrestata e come secondo elemento di input (`type`) il tipo di invarianza parziale che si vuole valutare. Il parametro `type` può assumere i seguenti valori: `"metric"` (invarianza debole), `"scalar"` (invarianza forte) e `"strict"` (invarianza esatta). La funzione restituisce in output una tabella che indica quali parametri devono essere liberati per ciascun gruppo (`group`) utilizzato nel valutare l'invarianza ordinati per maggiore indice di modifica (`mi`). Si veda il file `lab9.R` presente alla pagina Moodle del corso per ulteriori informazioni sulla procedura di invarianza parziale.

[7] La funzione richiede come primo input (`fitted_model`) il modello CFA adattato e come secondo input (`type_summary`) il tipo di parametro da estrarre. Quest'ultimo può assumere uno dei seguenti valori: `"latent"` ($\hat{\boldsymbol{\Lambda}}$), `"covariance"` (tril($\hat{\boldsymbol{\Psi}}$) o tril($\hat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$)), `"intercept"` ($\hat{\boldsymbol{\tau}}$ o $\hat{\boldsymbol{\mu}}$), `"variance"` (diag($\hat{\boldsymbol{\Psi}}$) o diag($\hat{\boldsymbol{\Theta}}_{\boldsymbol{\delta}}$)), `"thresdholds"` (i parametri delle soglie latenti ottenuto quando `cfa(..,estimator="DWLS")`).

Esercizio 13. Il file `SCS.Rdata` contiene $p = 10$ item del questionario `Sexual Compulsivity Scale` somministrato ad un campione di $n = 3348$ partecipanti (gli item sono denotati con la lettera `Q`). Le risposte sono state raccolte mediante scale categoriali. Dopo aver importato il dataset in `R` si svolga quanto segue:

1. Si esegua una divisione del dataset a metà riservando al primo di questi il 40% delle unità statistiche (si consiglia di utilizzare il parametro `seedx=90211`).

2. Sul primo dataset ottenuto al punto precedente si esegua un'analisi in componenti principali (PCA: si veda lezione `6-SVD-Algebra File.pdf`) dopo aver centrato e scalato[8] le variabili osservate. Si individui il numero minimo di componenti principali (`PC`) che spieghino almeno il 60% di varianza. Successivamente si individuino quali variabili osservate compongono le componenti principali scelte. Nota: per l'individuazione di tali variabili si vada alla ricerca delle variabili nella matrice `Rotation` dell'output della funzione `prcomp(..)` i cui coefficienti siano maggiori a 0.30 in valore assoluto. Successivamente si utilizzino le variabili individuate per la definizione di un modello CFA idoneo.[9]

3. Sul secondo dataset ottenuto al punto 1, si adatti ai dati il modello scelto al punto 2 aggiungendo il vincolo $\boldsymbol{\Psi} = \mathbf{I}$ (si ricordi di adattare il modello tenendo in considerazione il tipo di variabile a disposizione). Successivamente si adatti ai dati un secondo modello CFA di tipo unidimensionale utilizzando le stesse variabili utilizzate nella definizione del modello definito al punto precedente.

4. Si confrontino i due modelli in termini di adattamento complessivo ai dati.

5. Si valuti l'attendibilità dei due modelli mediante indice ω. Qualora gli indici parziali per entrambi i modelli siano tali che $\omega < 0.75$, si calcoli il numero di item da aggiungere alle scale affinché $\omega \geq 0.80$. Nota: si valuti l'aggiunta di $p \in (1, 3, 5, 7)$ nuovi item alle scale in questione mediante formula profetica di Sperman-Brown.

Esercizio 14. Si considerino i risultati dell'Esercizio 13. Dopo aver scelto al punto 4 il modello CFA che meglio si adatta ai dati, si valuti quale livello di invarianza completa il test *SCS* raggiunge rispetto alla variabile `gender`. Si commentino estensivamente i risultati anche alla luce dei coefficienti stimati standardizzati. Nota: si ricordi di utilizzare il secondo dataset della divisione a metà del dataset originario.

Esercizio 15. Si considerino i risultati dell'Esercizio 13. Dopo aver scelto al punto 4 il modello CFA che meglio si adatta ai dati, si calcolino i punteggi fattoriali $\hat{\boldsymbol{\eta}}$. Si definisca e si adatti ai dati un modello di regressione lineare per valutare se $\hat{\boldsymbol{\eta}}$ vari in funzione di `age`, `gender` e dell'interazione `age:gender`. Si commentino i risultati ottenuti. Nota: si ricordi di utilizzare il secondo dataset della divisione a metà del dataset originario.

Esercizio 16. Il file `mimic.Rdata` contiene i risultati di una rilevazione condotta su un campione di $n = 2000$ unità e $p = 24$ variabili di tipo dicotomico unitamente alla variabile categoriale `z` a due livelli indicante il gruppo di appartenenza. Dopo aver importato il dataset in `R` si svolga quanto segue:

1. Si esegua una suddivisione casuale del dataset in due parti, riservando alla prima di queste il 40% delle unità statistiche (si consiglia di utilizzare il seed di generazione `0661`). Si utilizzi la prima metà del dataset così ottenuta per rispondere ai quesiti successivi.

[8] L'operazione si ottiene impostando i seguenti parametri: `prcomp(..,,center = TRUE,scale. = TRUE)`

[9] È possibile utilizzare la funzione `prcomp2lavaan(prcomp_output,numPC,thr,rotate)` i cui input sono i seguenti: `prcomp_output` (output della funzione `prcomp(..)`), `numPC` (numero di componenti principali scelte), `thr` (valore soglia per l'individuazione delle variabili, default: 0.35)`rotate` (tipo di rotazione della matrice dei coefficienti della PCA. Possibili scelte: `"varimax"`,`"oblimin"`,`"none"`. Valore di default: `"varimax"`). La funzione è presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

2. Si calcoli la matrice di correlazione policorica sugli item che compongono il test.[10]

3. Si esegua un'analisi in componenti principali (PCA) sulla matrice di correlazione policorica. Si individui il numero minimo di componenti principali (`PC`) che spieghino almeno il 70% di varianza. Successivamente si individuino quali variabili osservate compongono le componenti principali scelte e le si utilizzino per definire una prima struttura del test in termini di scale. Si veda l'Esercizio 13, punto 2, per eventuali dettagli implementativi.

4. In maniera simile al punto precedente, si faccia un clustering gerarchico con metodo `Ward.D2` sulla matrice di correlazione policorica e si individui un numero congruo di raggruppamenti. Si usino questi ultimi per definire una seconda struttura del test in termini di scale.

5. Si definisca una terza struttura del test in termini di scale mediante l'assegnazione seguente: `scala1 = item1,..,item8`; `scala2 = item9,..,item16`; `scala3 = item17,..,item24`.

6. Per ciascuna delle tre strutture definite nei punti 3-5, si calcoli un coefficiente di attendibilità congruo al tipo di dato a disposizione (es. indice di Rulon) per ciascuna scala che forma la struttura del test.[11]

7. Delle tre strutture individuate, si scelga quella le cui scale raggiungono almeno un livello di attendibilità pari a 0.40. Dopo aver scelto la struttura che forma il test, si individui la scala con più alto coefficiente di attendibilità e si valuti quanti item occorre aggiungere per avere un'attendibilità almeno pari a 0.90. Nota: si valuti l'aggiunta di $p \in (7, 9, 11, 13, 15)$ nuovi item.

Esercizio 17. Si considerino i risultati ottenuti all'Esercizio 16. Si usi la seconda metà del dataset `mimic.Rdata` per eseguire quanto segue:

1. Si definisca un modello CFA sulla base della struttura del test individuata e lo si adatti ai dati secondo un metodo congruo al tipo di dati a disposizione.

2. Si valutino gli indici di fit complessivi del modello e si interpretino i risulti ottenuti.

3. Si consideri la possibilità di semplificare il modello eliminando gli item con coefficiente fattoriale al di sotto della soglia $\lambda_{jk} \leq 0.30$.

4. Si valuti se il modello ottenuto al punto precedente raggiunga il livello di invarianza totale debole rispetto alla variabile `z`. Si applichi una procedura di invarianza parziale qualora la prima non fosse stabilità. Nota: si consulti la nota 6 all'Esercizio 12 per eventuali dettagli implementativi.

5. Si calcoli il coefficiente di attendibilità ω del modello CFA scelto dall'analisi dell'invarianza al punto precedente. Si interpreti il risultato anche alla luce del coefficiente α. Qualora questi differissero in magnitudine, si giustifichi il motivo mediante appropriata argomentazione.

Esercizio 18. Si considerino i risultati ottenuti all'Esercizio 16. Si usi la seconda metà del dataset `mimic.Rdata` per eseguire quanto segue:

1. Si definiscano tre modelli CFA come di seguito:

   (a) Stesso modello utilizzato al punto 1 dell'Esercizio 17

   (b) Stesso modello utilizzato al punto 1 dell'Esercizio 17 con il vincolo di ortogonalità $\mathbf{\Phi} = \mathbf{I}_{q \times q}$

[10] La matrice di correlazione policorica può essere ottenuta mediante il comando `psych::polychoric(X)$rho` dove `X` è la matrice $n \times p$ di input.

[11] A tal fine si può utilizzare la funzione `split_half()` presente nel file *utilities.R* nella cartella *Utilities* alla pagina Moodle del corso.

(c) Stesso modello utilizzato al punto 1 dell'Esercizio 17 con aggiunta di un fattore sovraordinato η_0

2. Si calcoli l'errore di previsione dei modelli (a)-(c) (si usi il calcolo dell'errore mediante tecnica Monte Carlo con $B = 1000$ repliche).

3. Si scelga quale tra i modelli (a)-(c) produce un minore errore di previsione (si può usare il coefficiente di variazione per scegliere il modello migliore).

4. Si commenti il risultato ottenuto anche alla luce di quanto ottenuto all'Esercizio 17.