Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/2020 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZIONE 11 – Analisi Fattoriale : introduzione per applicazioni

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: Lunedì su appuntamento (previa mail)

Passaggi per implementare un AF

Verificare che una AF possa essere eseguita:

1) Dati quantitativi standardizzabili
2) Verificare che seguano una distribuzione normale
3) Esclusione di outliers di disturbo
4) Ampiezza sufficiente del campione (n=>5*k)
5) Analizzare le correlazioni attraverso la matrice di correlazione

4*Nota su ampiezza del campione

- Più soggetti ci sono meglio è
- Più casi che variabili (si capirà con le componenti principali)
- Da 10 a 20 casi per ogni variabile
- Come minimo 5 casi per variabile ma non meno di 100 casi in totale (Gorsuch, 1983)
- ogni fattore deve avere almeno 4 saturazioni superiori a .60, non importa l'ampiezza del campione (Guadagnoli e Velicer, 1988)
- ogni fattore deve avere almeno 10 saturazioni superiori a .40 e il campione almeno di 150 casi (idem)
- un campione di almeno 300 casi (idem)
- bastano anche 60 casi se tutte le comunalità sono maggiori di .60 (MacCallum, Widaman, Zhang e Hong, 1999)
- con comunalità attorno a .50 , il campione dev'essere tra 100 e 200 casi (idem)

5- matrice di correlazione

L'analisi fattoriale esplorativa si applica ad una matrice di correlazione (o meglio ad una matrice di associazione fra variabili). Gli indici che si possono usare sono:

- **correlazione di Pearson**: è la correlazione più usata in assoluto (il default in quasi tutti i software statistici). Implica variabili misurate a livello intervallo/rapporto
- **correlazione di Spearman**: è una correlazione per variabili di tipo ordinale oppure quando si presume che le variabili intervallo non siano "normali"
- **varianze/covarianze**: è la scelta preferita nelle analisi fattoriali confermative ed è utilizzabile quando le variabili hanno varianza simile

5-osserviamo la matrice di correlazione

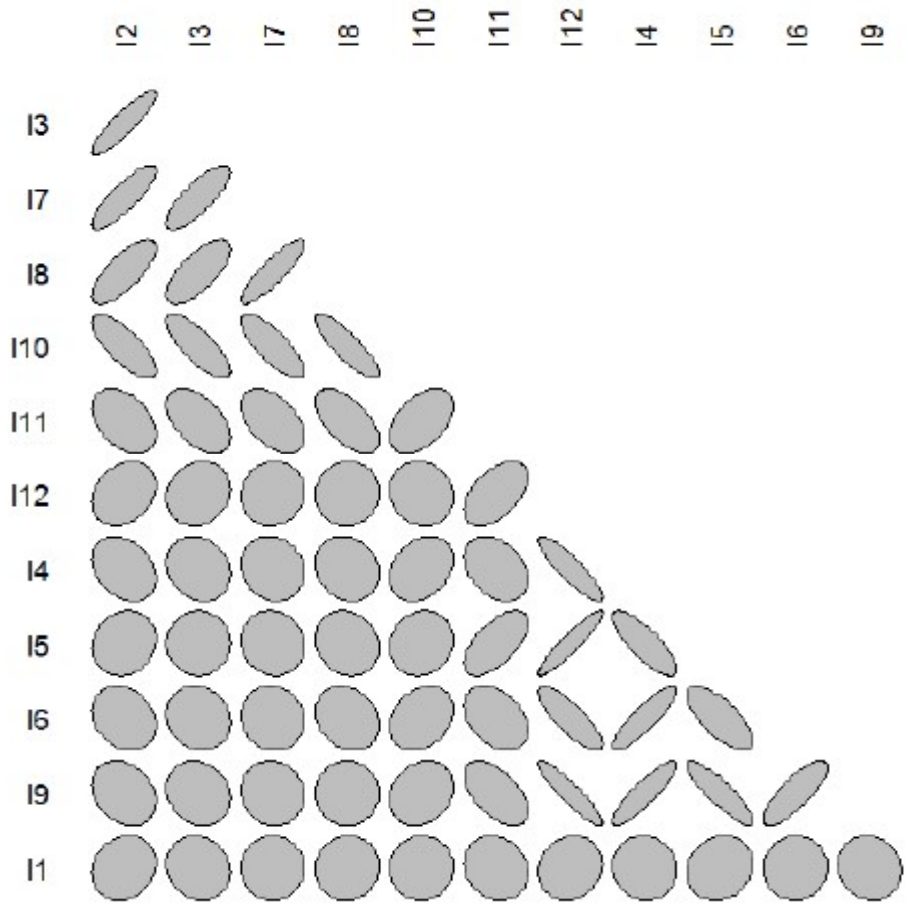
- Non tutte le correlazioni sono inferiori a |.30|
- Ispezione visiva della matrice di correlazione per vedere che abbia blocchi di correlazioni alte fra loro e basse con le alte
- Determinante: se è nullo non si può fare l'analisi;
- KMO (Test di adeguatezza campionaria di Kaiser-Meyer-Olkin)
- Test della Sfericità di Bartlett [tende a sovrastimare]

Ispezione visiva della matrice di correlazione

I2	I3	I7	I8	I10	I11	I12	I4	I5	I6	I9
1										
0,89	1									
0,84	0,80	1								
0,75	0,70	0,88	1							
-0,76	-0,79	-0,79	-0,85	1						
-0,41	-0,50	-0,48	-0,60	0,47	1					
0,26	0,14	0,08	0,04	-0,11	0,51	1				
-0,30	-0,18	-0,13	-0,16	0,28	-0,34	-0,88	1			
0,13	0,00	-0,09	-0,15	0,11	0,61	0,93	-0,80	1		
-0,22	-0,12	-0,07	-0,21	0,27	-0,38	-0,82	0,88	-0,72	1	
-0,26	-0,15	-0,08	0,00	0,20	-0,55	-0,91	0,89	-0,88	0,80	1
0,16	-0,12	0,00	0,05	0,02	-0,15	0,12	-0,05	0,14	0,02	-0,10

Guardare quali sono le aree di correlazione
elevata fra variabili che non sono correlate con altre

Ispezione visiva della matrice di correlazione



KMO

- KMO (Test di adeguatezza campionaria di Kaiser-Meyer-Olkin)

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j p_{ij}^2}$$

  dove r_{ij} sono le correlazioni e p_{ij} sono le correlazioni parzializzate su tutte le altre
- se le correlazioni parzializzate sono piccole tende a 1, quindi (secondo Kaiser, 197)
- se > 0.90 è eccellente
- fra .80 e .90 buono;
- fra .70 e .80 accettabile
- fra .60 e .70 mediocre
- inferiore a .60, meglio non fare l'analisi
- N.B. la dicitura "campionaria" non si riferisce al campione

Sfericità di Bartlett

- Il test della sfericità di Bartlett verifica l'ipotesi $H_0 : \mathbf{R} = \mathbf{I}$ tramite la formula:

$$\chi^2 = -\left[n - 1 - \frac{1}{6}(2p + 5)\right] ln \mid \mathbf{R} \mid$$

  in cui n è il numero dei soggetti, p il numero delle variabili e $\mid \mathbf{R} \mid$ il determinante della matrice di correlazione.
- Si distribuisce con $(p^2 - p)/2$ gradi di libertà
- Se è significativo, significa che $\mathbf{R}$ ha correlazioni sufficientemente elevate da non essere paragonabili a 0; se è non significativo le correlazioni sono basse e non si distinguono da 0
- ma questo test dipende dal numero delle variabile e dalla numerosità del campione, quindi tende ad essere significativo all'aumentare del campione e del numero delle variabili anche se ci sono correlazioni basse.

Quanti fattori estrarre

- Teoria (analisi della letteratura)
- Rango della matrice (solo teorico)
- Criterio di Kaiser (1959) o di Guttman (1954) ovvero autovalori maggiori di 1 (sovrastima)
- Almeno il 60-75% di varianza spiegata
- Scree-test di Cattell (1966)
- Test statistici di alcuni metodi di estrazione
- Il buon senso
- Analisi parallela

Gorsuch (1983) ritiene necessario effettuare più analisi e tenere quei fattori che si mantengono nelle varie soluzioni

Come Eseguiamo un'AF completa in R

```
#1. verificare che un'AF possa essere eseguita

#1.a: i dati sono quantitativi e standardizzabili?
#se si --> continua

#analizziamo la struttura dei dati
getwd()
#cambia directory
torte=read.csv2("torte2.csv")
str(torte)
View(torte)

 #da qui vediamo com'è composto il database e
 la natura dei dati in esso contenuti
 #ricordarsi la regola empirica: n=>5*k
```

Come Eseguiamo un'AF completa in R

#1.b: i dati seguono una distribuzione normale?

#verifichiamolo con un istogramma dei dati

hist(vendita) #questi sono istogrammi di frequenza assoluta
hist(prezzo)
hist(pubb)

hist(vendita,freq=FALSE)

#questo è istogragramma che utilizza frequenze relative

#nel caso non seguano una distribuzione normale possiamo trasformare le variabili
#le trasformazioni più utilizzate sono: ^2 e log() ovvero il logaritmo naturale

Come Eseguiamo un'AF completa in R

#2. analisi delle correlazioni
vedere se la distribuzione è GAUSSIANA O NO
==> SE E' GAUSSIANA --> PEARSON
==> SE NON E' GAUSSIANA --> SPEARMAN O KENDALL

#2.1: indice di Pearson
cor.test(vendita,prezzo)

#Il valore del coefficiente di Pearson è -0.443: è un valore modesto,
che indica una correlazione tra le variabili non molto forte.
Controlliamo ora la significatività di R dal valore della statistica test:
#qt(valore dell'intervallo di confidenza,valore dei gradi di libertà=df)

qt(0.950,13)

#Risulta che t-calcolato < t-tabulato; del resto p-value > 0.05.

Quindi il coefficiente R di Pearson non è statisticamente significativo.

Come Eseguiamo un'AF completa in R

#2.2 Correlazione di Spearman: test Rho di Spearman (variabili ordinali o che si presume non abbiano intervalli uguali)
#cor.test(variabile 1, variabile 2, method="spearman")

#2.3 Correlazione di Kendall: thest Tau di Kendall (quando var non gaussiane)
#cor.test(variabile 1, variabile 2, method="kendall")

#sia per spearman che kendall si guarda il valore (che indica la relazione) e il
confronto tra p-value e alfa di significatività)

#2.4 matrice di correlazione
m=cor(torta)
m

#visualizzazione grafica delle relazioni quando le serie di variabili non sono molte (K limitato)
pairs(x=torta,panel=panel.smooth)

#osservazione e commento della matrice di correlazione

#ricordare: se il determinante della matrice di correlazione è nullo non si può procedere con l'analisi
det(matrice di correlazione)
det(m)

Come Eseguiamo un'AF completa in R

```
3.test dell'adeguatezza campionaria (KMO)
x <- subset(torta, complete.cases(torta))
# Omit missing values
r <- cor(x)
# Correlation matrix
r2 <- r^2
# Squared correlation coefficients
i <- solve(r)
# Inverse matrix of correlation matrix
d <- diag(i)
# Diagonal elements of inverse matrix
p2 <- (-i/sqrt(outer(d, d)))^2
# squared partial correlation coefficients
KMO <- sum(r2)/(sum(r2)+sum(p2))
KMO
#il criterio di adeguatezza campionaria di Kaiser-Meyer-Olkin ha un range di valori
tra 0 e 1, e sono accettabili valori al di sopra di 0.5 per continuare l'analisi fattoriale
```

Come Eseguiamo un'AF completa in R

4. test di sfericità BARTLETT

batlett.test(variabile 1, variabile 2)

--> Attenzione alla versione perché potrebbe dare errore dicendo che ci devono essere almeno due osservazioni ogni gruppo

Come Eseguiamo un'AF completa in R

5.performiamo l'Analisi Fattoriale: stima del modello

f=factanal(torta, factors=1) #dobbiamo noi identificare il numero dei fattori

f #visualizza i valori di unicità e i loadings; l'ultima riga identifica la proporzione di variabilità spiegata. E' accettabile un valore pari o superiore a 0.7

infine si osserva il test di ipotesi che un fattore sia sufficiente

#guardiamo come sono composti i dati:

dim(torte)
summary(torte)

OBIETTIVO: riprodurre e sintetizzare le informazioni contenute in queste variabili mediante un insieme di variabili latenti di dimensione inferiore alle attuali.

Come Eseguiamo un'AF completa in R

il comando factanal effettua la stima di un modello fattoriale con il metodo della massima verosimiglianza

Ipotesi base: le matrici di covarianza dei dati originari e del modello stimato convergono

H0= il modello si adatta perfettamente
H1=il modello non si adatta perfettamente

→p-value se tende a 0 l'ipotesi H0 è rigettata

comando:

g=factanal(torte,factors=n)

summary(g) #elenco degli oggetti calcolati

per estrarre gli oggetti $

Come Eseguiamo un'AF completa in R

g$factors #estraiamo i fattori comuni

uniqueness= vettore delle unicità
g$uniqueness

loadings = matrice dei pesi fattoriali in cui
ogni colonna riporta i pesi di ciascun fattore
g$loadings

per vedere il p-value e la statistica test:
g$PVAL
g$STATISTIC

per vedere i gradi di libertà g$dof

la matrice di correlazione di partenza è:
g$correlation

Come Eseguiamo un'AF completa in R

il criterio utilizzato (o metodo) viene indicato con g$method o g$criteria

il numero di fattori è indicato con g$factors

il numero di osservazioni è g$n.obs

Come Eseguiamo un'AF completa in R

PROPRIETA' DEL MODELLO FATTORIALE

il modello fattoriale è equivalente rispetto a
cambiamenti di scala delle x osservate

Tale proprietà permette di lavorare indifferentemente sulle variabili osservate (matrice di varianza-covarianza)
o sulle standardizzate (matrice di correlazione)

quindi:

dati2=scale(torta,T,T) # standardizziamo i dati

prova1=factanal(dati2,factors=n) # dà risultati equivalenti

Come Eseguiamo un'AF completa in R

ROTAZIONE DEGLI ASSI

il comando "rotation" permette di scegliere il tipo di rotazione da effettuare

la rotazione dei fattori consente di scegliere, tra le possibili trasformazioni della matrice dei pesi fattoriali, quella che facilita l'interpretazione dei fattori comuni in termini delle variabili.

la scelta viene fatta in base al principio per cui l'identificazione dei fattori risulta semplificata se ciascuno di essi è fortemente correlato con un numero limitato di variabili (ed è poco correlato con le altre).

in R è possibile:
- non eseguire nessuna rotazione "none"
-usare la rotazione varimax
-usare la rotazione promax

generalmente se non digitiamo nulla viene eseguita varimax

Come Eseguiamo un'AF completa in R

GLI SCORES O PUNTEGGI FATTORIALI (PER IL DATABASE)

Sono i punteggi che i singoli oggetti-unità-individui hanno sui fattori identificati dall'analisi, ovvero i dati da inserire nel nostro database quali "nuove variabili"

g$scores

se vogliamo calcolarli con un determinato metodo:
gs=factanal(torte,factors=n,scores="regression")

#metodo della regressione multipla
(o metodo di Thomson) che impiega le correlazioni tra le variabili e
le correlazioni delle variabili con il fattore

gs$scores e confrontare

NB: con varimax i punteggi o scores risultano sempre incorrelati!

Come Eseguiamo un'AF completa in R

LE COMUNALITA'

Ogni variabile presenta un'unità di varianza.
questa si distingue in
-specificità= la proporzione della varianza della variabile
che non viene spiegata dalla soluzione fattoriale
-comunalità=il complemento a 1 della unicità
1-g$uniquenesses
le comunalità permettono di valutare:
-in che misura il modello stimato riesce a rendere conto della variabilità di
ogni singola variabile osservata
-costituiscono quindi un elemento per la DIAGNOSTICA della bontà
del modello stimato

Come Eseguiamo un'AF completa in R

INTERPRETAZIONE DEI RISULTATI OTTENUTI

1) dire qual è la variabilità spiegata del modello stimato
gs$loadings
(guardare Cumulative var e Proportion var)

2) dare un'interpretazione ai fattori estratti a partire dalla matrice dei pesi fattoriali ruotata:
gs$loadings[,1:4] #4=n utilizzato

3)dire in che misura il modello stimato riesce a rendere conto della variabilità di ogni singola variabile osservata
1-gs$uniquenesses

4)valutare se il numero di fattori scelto è appropriato
gs$STATISTIC
gs$PVAL

Come Eseguiamo un'AF completa in R

5)confrontare due modelli con numero di fattori differente

(ipotizzando alfa=0.01) e ricordando il principio di parsimonia:

```
f1=factanal(torte,factors=4,scores="regression")
f2=factanal(torte,factors=3,scores="regression")
f1$PVAL>0.01
f2$PVAL>0.01
```