

Giovagnoli 29 marzo

Analisi fattoriale II parte

Ripasso

L'analisi fattoriale nasce in ambito medico-psicologico, ed è utilizzata in particolare per validare i questionari perché fornisce info sulla struttura dei dati. E' una procedura matematica e spesso viene sovrastimata dagli autori, che sperano di trovare una 'formula magica' in essa; in realtà non ci permette di arrivare a confermare l'H di partenza, ma è utilizzata perché fornisce info difficilmente ritrovabili altrimenti.

Lavora con var gaussiane analizzate allo stesso livello e correlate. Tutto ciò che studiamo in un'unica popolazione di persone non può essere studiato come slegato dal resto che rappresenta una persona. Es. studio della memoria, non possiamo pensare sia slegato dalla personalità della persona. Un questionario va ad analizzare un campo specifico, es. mobbing in campo medico-infermieristico (vedi ultima lezione Giovagnoli), vediamo se il questionario va veramente ad analizzare l'argomento e quindi se è costruito bene. L'argomento generale si può dividere in sottoargomenti, che formano insieme l'argomento che stiamo studiando, e dobbiamo vedere se le info raccolte ci danno un quadro completo di ciò che studiamo. Le operazioni matematiche compiute sono diverse: si usa una matrice di correlazione che correla tutti gli item tra loro e ci dice quanto sono legati, da qui poi estraiamo i fattori (nuove var che esaminiamo, che raccolgono le info comuni agli item di partenza); l'estrazione avviene partendo dalla matrice. Infine ci sono le tecniche di rotazione: nascono dalle critiche mosse all'analisi fattoriale, soprattutto il fatto di dare nome alle var che estraiamo. Attraverso la rotazione è possibile rendere più interpretabili i fattori 08' per diminuire la difficoltà nell'interpretazione e renderla più oggettiva possibile, in modo che le interpretazioni che si possono dare dei fattori possano essere più simile possibile tra i ricercatori. Scopi: riduzione var di studio ma non dell'info; trasformazione var in var dip, individuazione delle sorgenti di var e assegnazione di significato reale a queste var. Le ultime due operazioni sono le più criticabili.

Critiche:

- Trova var artificiali, aleatorie, ne possiamo trovare un numero infinito e per questo criticiamo l'estrazione dei fattori, possiamo creare sempre nuove var e per questo è soggettiva l'interpretazione delle var.
- Le var sono indipendenti tra loro, e non rispecchiano la realtà psicologica in cui non esistono var del genere.
- Validazione dei questionari.
-
- Vantaggi:

- Creazione di fattori indipendenti, perché rende più facile l'operazione matematica, non ridondante.
- Non vengono considerate le differenze tra le diverse variabili.

Tecnica della max verosimiglianza: non funziona a cascata, cerca di distribuire equamente fra i fattori l'info rappresentata dagli item di partenza. La distribuzione delle info è più equa=cerca di essere più reale possibile, più simile possibile alla realtà che stiamo studiando. Se la realtà che studiamo raggruppa le info in un unico sottoargomento è ovvio che il primo fattore sarà quello con più info. Ma se studiamo un fattore diviso in molti sottoargomenti, la distribuzione dell'info è divisa tra questi. I nostri fattori sono un riassunto degli item di partenza.

Standardizzazione dei dati raccolti: i punteggi grezzi vengono trasformati in punti z con media 0 e varianza 1, si fa per rendere più semplici le operazioni matem successive. Lo si fa perchè standardizzando i punteggi grezzi possiamo confrontare tutte le var che stiamo studiando. Es. se un questionario utilizza sia scale Likert a 5 punti che a 7 o a 3, o risposte dicotomiche, la distribuzione dei punteggi è differente e non si può confrontare, per cui li si porta alla stessa scala (vedi formula lucidi). Varianza: quantità di info che viene trasportata in ogni item, è la variabilità. Quando studiamo una popolazione, ci interessa la differenza tra le persone. Tutti i dati quindi saranno punti z.

Procedimento:

1. Creazione della matrice di correlazione R: correla tutti gli item tra di loro, ha tante caselle di correlazione quanti sono gli item di partenza, es. 43x43. Possiamo dividerla in due parti: una rispecchia l'altra perchè le correlazioni vengono fatte con tutti gli item. Due var sono statisticamente correlate quando al variare di una varia anche l'altra (positiva aumentano insieme, negativa una aumenta e l'altra diminuisce). Quando estraiamo i fattori, la prima colonna è con un numero di fattori pari agli item di partenza con varianza differente da 1, mentre noi teniamo i fattori con varianza $>$ di 1.
2. Estrazione dei fattori: la tecnica più usata è quella delle componenti principali, poi quella della max verosimiglianza, serve per archiviare i dati, perchè ci permette di eliminare gli item che sono ridondanti e non ci servono per quello studio. Il metodo delle componenti principali permette di creare variabili artificiali, fra loro ortogonali (correlazione=0), cercano di non trasportare le stesse info. I fattori che noi estraiamo, le nuove var che creiamo, sono operazioni matematiche, quindi dobbiamo considerarle tali. Sono combinazioni lineari che si ottengono dalla somma dei prodotti delle var sperimentali per i loro coefficienti. Il segno dei coefficienti indicano come il singolo fattore sia legato alle var sperimentali. Metodo delle componenti principali si avvale del calcolo di autovalori e autovettori della matrice di correlazione. Il primo fattore ha la maggior parte dell'info, e

possiamo capire quanta parte dell'info è raggruppata nel primo fattore...Gli autovettori e autovalori vengono calcolati da R attraverso un processo algebrico. Gli autovalori sono la quantità di varianza di un fattore (comunalità), mentre gli autovettori sono i fattori (componenti). Formula: $R \cdot \text{autovettore} = \text{autovettore} \cdot \text{autovalore}$. $\text{Autovettore} \cdot R$ dev'essere uguale a se stesso a meno di una costante che si chiama autovalore.

Decidere il numero di fattori tenere in base a: posso decidere un numero preciso da tenere (sottoscale), posso decidere la percentuale di info spiegata che voglio tenere, oppure decido di tenere solo i fattori con varianza maggiore o uguale ad uno, perchè standardizzando i dati gli item di partenza hanno tutti media 0 e varianza 1, quindi la quantità di info di ognuno è pari ad 1, ma quando noi estraiamo i fattori vogliamo raggruppare l'info e quindi decidiamo di tenere una quantità di item che trasportano una quantità di info o pari agli item di partenza o maggiori, altrimenti non serve a niente. Posso guardare lo scree plots, che è un grafico che fa vedere sul piano cartesiano la distribuzione dei fattori estratti (tutti), come si suddivide la quantità di info.

Le nuove variabili servono a:

- compattare i dati
- tenerne un numero inferiore perdendo una parte accettabile dell'info di partenza
- guardo il legame tra fattori e var di partenza (per colonna sulla matrice)
- ecc

E' possibile calcolare il valore di ogni fattore per ogni soggetto.

Metodo dei minimi quadrati: minimizza la somma dei quadrati tra gli scarti tra i dati osservati e la R prodotta dal modello -> noi non lo studiamo perchè è poco usato. Usa tecniche matematiche differenti.

Metodo della max verosimiglianza: parte dalle componenti principali, cioè fa estrarre i fattori e li prende modificando la quantità di info trasportata da ognuno. E' un metodo a cascata, quindi il primo fattore ha la maggiore quantità di info, mentre il secondo ne avrà meno ma più del terzo ecc ecc, fino ad arrivare ad una quantità \geq ad 1. Lavora per approssimazioni successive. Il nostro scopo non è quello di archiviare i nostri dati o compattare le var con meno perdita di info possibile, ma di rappresentare al meglio la realtà che studiamo. Vogliamo creare fattori più simili possibile alla realtà che studiamo. Chi ci dice che il primo sottoargomento è quello con maggiori info? Ogni sottoargomento avrà una quantità differente ma non è detto che sia minore di quello precedente. Con questa tecnica è come se assumessimo che il primo sottoargomento è il più importante perchè trasporta più info. Ogni volta che modifica la qtt di fattori (qtt di varianza per ognuno) fa un chi quadro, e va a vedere la differenza tra i dati grezzi e quelli riportati, e guarda se si somigliano,

rappresentando la realtà all'astrazione. Se il confronto è sufficientemente buono si ferma, vuol dire che rappresenta al meglio la realtà che stiamo studiando. [noi tutte le volte andiamo a studiare i fattori che abbiamo estratto e li confrontiamo con la realtà, confronta i dati grezzi di partenza con le var estratte, per vedere se rappresentano al meglio la realtà che studiamo] Se io decido di tenere un fattore in meno o in più la max verosimiglianza deve adattare la quantità di info che ognuno trasporta, quindi devo rifare l'analisi ogni volta. Con le componenti principali invece posso modificare senza rifare l'analisi perchè è un metodo lineare, non è necessario rifare l'analisi ogni volta. Ogni volta che si fa una modifica si applica il chi quadro, per misurare la bontà dell'adattamento dei dati. La significatività di chi² mi dice se la modifica che ho fatto è significativamente della precedente. Il chi² confronta le situazioni relative agli stessi dati e ci dice quanto è reale la rappresentazione attuale.

Anche con questo metodo ottengo var ortogonali, indipendenti e non correlate tra loro, quindi l'info non è ripetuta, ogni fattore trasporta una info differente rispetto agli altri. C'è una distribuzione più equa (più simile possibile alla realtà dell'info). Per creare fattori non ridondanti non è necessario creare un fattore principale con effetto a cascata sulla quantità di varianza degli altri. Ogni domanda ottiene info diverse da un'altra, in un questionario.

[vedi slide per confronto tra componenti principali e max verosimiglianza]

Confrontando le due tecniche, vedo che maggiore è il num di var di partenza e maggiore darà la differenza dei risultati delle due analisi. Di solito, più sono gli item di partenza e più sono i fattori che decido di tenere.

3. Rotazione dei fattori: modifica la rel tra gli item di partenza e fattori estratti attraverso una rotazione. Su un piano cartesiano, la x rappresenta il fattore 1 e la y il 2. Su questo piano cartesiano mettiamo tutti i nostri item e il coefficiente ci è dato dalla R che abbiamo. Possiamo rappresentare dunque su un piano cartesiano i due fattori, e in base al coefficiente che abbiamo, c'è il valore che ci permette di inserire i nostri item sul piano cartesiano. Per rendere più interpretabili i fattori che ho estratto, con la tecnica delle rotazioni io posso decidere di spostare i miei assi in modo da avvicinare, ad es. l'asse y ad alcuni item. E' una tecnica puramente matematica, io devo essere sicuro che il mio fattore che verrà modificato rappresenti bene solo un certo numero di item. Io vado a vedere il primo fattore con quali item correla, e vedo che correla con i primi 20. Fatto questo posso decidere come chiamare il mio fattore a seconda di cosa chiedono le 20 domande. Se chiedono della working memory, io posso chiamarlo così. Ma se il primo fattore oltre a questi 20 item è legato anche ad altri 5, io posso spostare il mio asse y per avvicinarlo di più ai 20 item allontanandolo dai 5 che sono vicini ma non così legati. Allora posso dire che il primo fattore è legato con i primi 20 e non con gli altri 5. Queste rotazioni si chiamano ortogonali perchè mantengono i

fattori indipendenti, ma si chiamano anche oblique, perchè con queste i fattori estratti non sono più indipendenti. La rotazione obliqua rende i fattori maggiormente interpretabili (è comunque molto pericolosa). Nasce da due critiche: la prima è che vengono scoperte variabili latenti indipendenti, la seconda è che vengono dati dei nomi ai fattori. E' molto criticata perchè io vado ad inventare qualcosa per interpretare meglio i dati. Una diversa interpretazione dei fattori va a modificare l'interpretazione e il significato della ricerca... La rotazione ortogonale mantiene l'indipendenza tra i fattori, l'obliqua no. La prima ruota rigidamente gli assi, che restano perpendicolari e quindi la somma delle varianze resta uguale a quella di partenza, mentre con la seconda gli assi possono diventare obliqui.

Tecniche di rotazione (non dobbiamo sapere quali operazioni comportano, solo sapere che esistono tecniche diverse):

- varimax method: ortogonale, minimizza il num di var correlate con ogni fattore, interpretazione dei fattori è semplificata
 - quartimax method: funziona in modo opposto
 - equamax m.: combinazioni delle prime due
 - oblimin m.: obliqua, correla
 - promax m.: obliqua, metodo più diretto che fa una calcolo matematico diretto
1. 4. si possono usare i punteggi fattoriali per un'ulteriore analisi. E' possibile verificare la correlazione tra i fattori e le sottoscale.