Schede di analisi multivariata: L'analisi fattoriale

1. Introduzione

L'analisi fattoriale si pone l'obiettivo di riassumere l'informazione contenuta in una matrice di correlazione o di varianza/covarianza, cercando di individuare statisticamente la dimensioni latenti e non direttamente osservabili (Stevens, 1986). In sintesi si può dire che se due variabili hanno una forte correlazione con uno stesso fattore, una parte non trascurabile della correlazione tra le due variabili si spiega col fatto che esse hanno quel fattore in comune. Fornendo, quindi, un principio di identificazione di questi fattori comuni, l'analisi fattoriale fornisce una descrizione in forma semplice, della complessa rete di interpolazioni esistente nell'ambito di un insieme di variabili associate. Questa descrizione consente di definire, all'interno della matrice di correlazione, un limitato numero di componenti indipendenti l'una dall'altra e identificate nei fattori: esse spiegano il massimo possibile di varianza delle variabili contenute nella matrice d'informazione originaria. Data, pertanto, una matrice *nxp* contenente *p* variabili rilevate su *n* unità, si tratta di verificare in che misura ciascuna variabile costituisce una ripetizione della descrizione effettuata dalle rimanenti *p-1* e, quindi, se esiste la possibilità di raggiungere la stessa efficacia descrittiva con un numero minore di variabili non osservate dette, appunto, fattori.

Un esempio può essere rappresentato dall'analisi di un settore produttivo utilizzando dati tratti da archivi fiscali, in questo caso ci troviamo di fronte al problema della scelta delle variabili da considerare.

Da un lato, si potrebbero usare le 5/6 macro-variabili (fatturato, consumi intermedi, risultato lordo di gestione, ecc.), anche se probabilmente scaturirebbe un'immagine indistinta delle imprese. Dall'altro lato, scendendo al livello dei righi delle dichiarazioni dei redditi utilizzeremmo un'informazione ridondante, e quindi di più difficile lettura, sull'attività aziendale. Con l'analisi fattoriale si trova una soluzione intermedia, in quanto partendo da un insieme anche ampio di variabili essa consente una semplificazione del patrimonio informativo mettendo in luce le caratteristiche distintive latenti delle contabilità aziendali.

2. Metodi di estrazione dei fattori

Le dimensioni latenti possono essere determinate in vari modi grazie alle svariate tecniche di estrazione dei fattori di cui l'analisi dei fattori si avvale. Tra le più utilizzate ricordiamo: l'analisi

delle componenti principali, l'analisi dei fattori principali, l'analisi fattoriale canonica che per le caratteristiche dei loro algoritmi vengono definite "Variance-oriented" (Kim e Mueller, 1978).

1. Analisi delle componenti principali. Il metodo delle componenti principali si propone di sostituire le *p* variabili date con un certo numero di variabili (tra loro non interdipendenti), ottenute come trasformazione lineare delle variabili originarie, riducendo così il numero di variabili necessarie a descrivere un certo ambito. Si tratta cioè di ricercare una serie di trasformate della matrice originaria dette, appunto, componenti principali, che spieghino quanta più parte possibile della varianza delle variabili originarie e che siano tra loro ortogonali. É possibile estrarre tante componenti quante sono le variabili originarie, quando però lo scopo è quello di conseguire un'economia nella descrizione in termini quantitativi di un certo fenomeno il risultato fornito dall'applicazione del metodo è tanto più utile quanto minore è il numero di componenti prese in considerazione. In genere il processo viene arrestato non appena la parte di varianza delle *p* variabili estratte dalle prime *q* componenti è sufficientemente grande. A tal pro un test comunemente usato per la scelta del numero di componenti da considerare, che utilizza la matrice della varianza e covarianza tra le variabili standardizzate, è quello di Bartlett (1950).
2. Analisi dei fattori principali. Il modello fattoriale classico invece prevede *p* fattori comuni a tutte le *p* variabili più un fattore specifico per ogni variabile. L'obiettivo è quello di trasformare la matrice di variabili originaria in una che contenga il più possibile di fattori allo scopo di eliminare eventuali ridondanze presenti nelle variabile. Poiché i fattori comuni sono assunti essere variabili normalmente distribuite con varianza unitaria, essi non sono determinabili univocamente, e di conseguenza non lo sono neanche i vettori. Questo metodo estrae $q<p$ di fattori comuni considerati è sufficiente a racchiudere l'informazione fornita dalle *p* variabili originarie è quella di fissare una certa quota *V* di varianza delle variabili e considerare solo i primi *q* fattori se la quota di varianza cumulata estratta da questi è maggiore di *V*.
3. Analisi fattoriale canonica. Il principio che guida questa analisi è di trovare una soluzione fattoriale nella quale la correlazione tra il set di ipotetici fattori e il set di variabili è massimizzata. Il metodo parte dalla considerazione di due serie di variabili *Z* e *X*, la prima contiene una batteria di *p* variabili osservate e *X* contiene invece *q* variabili ortogonali incognite, le cui trasformate, opportunamente ridotte in forma standardizzata, costituiscono le colonne della matrice dei fattori da determinare.

Ovviamente, la scelta di un metodo comporta assunzioni ed obiettivi diversi.

L'analisi delle componenti principali genera soltanto uno spostamento del sistema di riferimento in corrispondenza del baricentro, in pratica viene cambiato esclusivamente il punto di osservazione del collettivo allo studio.

L'analisi dei fattori principali implica la specificazione di un modello di stima delle variabili latenti. Si apre perciò la necessità di esplorare le conoscenze disponibili a *priori* sul settore artigiano per definire il modello.

L'*analisi fattoriale canonica* si scosta poco dalla precedente, ma essa opera sulla matrice di correlazione parziale invece che sulla matrice di correlazione totale delle variabili. In presenza di un ristretto ventaglio di variabili osservate sulle unità, essa consente di legare più nitidamente i fattori latenti ad esse.

3. Il problema della rotazione dei fattori

Il problema della rotazione si pone perché le variabili possono venir saturate in modo pressoché uguale da diversi fattori: la rotazione si sostanzia nella riduzione dei pesi fattoriali che nella prima fase erano già relativamente piccoli e nell'incremento, sia positivo che negativo, dei valori dei pesi fattoriali che erano preponderanti nella prima fase. Infatti, la matrice delle saturazioni non presenta un'unica soluzione e, attraverso la sua trasformazione matematica, si possono ottenere infinite matrici dello stesso ordine. È per questo che i fattori vengono trasformati o analizzati mediante un procedimento di rotazione degli assi. In una soluzione non ruotata, infatti, ogni variabile è spiegata da due o più fattori comuni, mentre in una soluzione ruotata ogni variabile è spiegata da un singolo fattore comune.

Si dispone attualmente di vari metodi di rotazione che possono essere suddivisi essenzialmente in due gruppi: quelli che producono "rotazioni ortogonali dei fattori" e quelli che invece considerano "rotazioni oblique". Il vincolo dell'ortogonalità è stato criticato, in quanto si sostiene che le dimensioni fondamentali possono essere tra loro interdipendenti sono stati, quindi, messi a punto una serie di metodi di rotazione obliqua nei quali gli assi, presi a due a due, sono lasciati liberi di disporsi in modo da formare un angolo o maggiore o minore di 90 gradi. La pluralità di queste tecniche di rotazione dei fattori provoca una indeterminatezza nella soluzione fattoriale, poiché non è possibile stabilire quale delle rotazioni sia migliore in assoluto; e questo non solo per la scelta tra rotazione obliqua e rotazione ortogonale, ma anche all'interno dei due tipi di rotazione. Un criterio di sicuro effetto è quello di confrontare tra loro i risultati di diverse applicazioni e scegliere quella che meglio si adatta ai risultati osservati.

4. Determinatezza della soluzione fattoriale

Una soluzione fattoriale è determinata se i fattori comuni che si adattano al modello sono unici. La condizione di indeterminatezza implica che insiemi contraddittori di punteggi fattoriali

risultano ugualmente plausibili e che la scelta di una soluzione piuttosto che di un'altra è arbitraria. Nell'analisi fattoriale l'indeterminatezza si verifica a due livelli:

(i.) nell'accettazione della soluzione che soddisfa il modello in senso statistico;

(ii.) nella ricerca di una soluzione più facilmente interpretabile di quella ottenuta in prima istanza (Guilford, Hoepfner 1971).

Su queste considerazioni il dibattito sviluppatosi tra gli studiosi di analisi fattoriale è stato assai intenso (Morrison, 1976; Diday et al.ii, 1994): se e poiché nelle applicazioni pratiche l'analisi delle componenti e quella dei fattori danno risultati del tutto simili, una volta stabilite le convenzioni della procedura (e cioè gli assunti), l'indeterminatezza del modello fattoriale è solo logica e non più statistica (Fabris, 1983). Nell'ambito delle analisi esplorative condotte per trarre informazioni sulla struttura latente dei dati osservati, ma anche per gettare un ponte verso strutture informative esterne (e collegate) a quelle presenti nel modello, il disporre di più interpretazioni mutuamente consistenti può essere considerata una situazione di privilegio e non di svantaggio.

Una altro problema di notevole interesse è la determinazione del numero probabile dei fattori. Infatti, il rapporto tra numero di fattori *q* e il numero di variabili osservate *p* permette di misurare la determinatezza dei fattori comuni e specifici in termini matematici (Shonemann, Wang 1972). Vale in generale che:

(i) all'aumentare del numero di variabili sotto studio è necessario incrementare il numero di fattori al fine di ottenere una buona interpolazione del modello di analisi;

(ii) all'aumentare del numero dei fattori estratti, cresce il rischio di indeterminatezza della soluzione.

Nella ricerca empirica di tipo esplorativo è ammissibile l'adozione di uno dei seguenti criteri:

i. la varianza spiegata dai fattori. Se la selezione delle variabili non è casuale ma, come si verifica spesso, si inseriscono per prime quelle ritenute più appropriate per il modello che il ricercatore ha in mente, è ragionevole supporre che la comunanza delle variabili aggiunte sia proporzionalmente inferiore a quella delle variabili introdotte in precedenza e che il numero dei fattori richiesti per raggiungere la stessa frazione di varianza spiegata sia relativamente più elevato.

ii. Kaiser (1960) raccomanda di considerare solo i fattori di una matrice di correlazione ai quali è associato *un autovalore maggiore o tuttalpiù uguale a 1*. Il numero di tali fattori dovrebbe variare tra 1/6 e 1/3 del numero di variabili. Si tratta di un criterio permissivo perché considera significativo, e include nel modello, un fattore purché spieghi più varianza di quanto non ne introduca una singola variabile, valore che in una matrice di correlazione è appunto uguale a 1.

iii. *La rappresentazione grafica degli autovalori* (in ordinata) contro l'ordine di estrazione dei fattori (in ascissa) dà un'immagine dell'importanza relativa dei primi autovalori nella

sequenza ricavata. Si escludono i fattori i cui valori appartengono alla spezzata che corre quasi parallela.

iv. *La scelta delle comunanze.* Il sostituire i valori sulla diagonale principale della matrice di correlazione con le comunanze determina l'estrazione di un numero di fattori inferiore al rango della matrice.

v. La verifica dell'ipotesi di significatività dell'adattamento del modello fattoriale ai dati di partenza, e cioè la determinazione del numero q di fattori comuni significativi. Nel caso in cui la stima dei pesi fattoriali sia effettuata in base al criterio della massima verosimiglianza (Joreskog, 1967), si calcola la statistica U_q. La statistica-test U_q ha una distribuzione che approssima quella del $\aleph^2$ e quindi il suo valore lo si confronta con il valore di un $\aleph^2$ e se supera il valore critico si rigetta l'ipotesi di esistenza di *q* fattori, se non lo si supera lo si accetta.

5. Gli sviluppi più recenti nell'analisi multi-fattoriale: la matrice a tre vie

L'analisi fattoriale è una metodologia statistica che ha conosciuto, negli anni recenti, consistenti e importanti sviluppi. L'importanza di migliorare le proprietà delle tradizionali tecniche multivariate di analisi dei dati e la necessità pratica di arrivare allo sviluppo di un'analisi multivariata che riduca la perdita di informazioni indotta dalla decomposizione della matrice multiway in molte matrici two-way hanno portato a sviluppare in modo consistente l'analisi multiway (Hayashi, 1988). Nonostante queste metodologie di analisi dei dati siano una branca relativamente nuova della statistica metodologica, è comunque notevole la ricchezza di studi che affrontano i problemi delle analisi multiway e consistente la varietà di tecniche proposte per dare soluzione ai conseguenti problemi computazionali e algebrici (Carrol, Chang, 1970; Kruskal, 1978, 1981; Tucker, 1963, 1964, 1966). Nell'analisi *multiway* un interesse particolare ha avuto, per l'evidente carattere esemplificativo, l'analisi delle matrici a tre vie con tecniche fattoriali. Una grande quantità di studiosi hanno affrontato, negli ultimi anni, il problema delle matrice a tre vie (Kroonenberg, 1983) arrivando a formulare, essenzialmente, la proposta di due tipi di modelli: modelli fissi e modelli stocastici. Infatti, usando la terminologia classica, se denotiamo con x_{ijk} $(i,j,k) \in I * J * K$ il punteggio dell'individuo *i* per la variabile *j* sull'occasione K, possiamo distinguere tra metodi che si propongono una analisi della stabilità attraverso lo studio di una prossimità stabile o media fra gli individui *i* e/o lo studio della relazione tra le variabili *j* dopo avere rimosso l'influenza dell'occasione *k* (Banet, Lebart, 1984; Tehenaus, 1986) e quei metodi che propongono analisi del cambiamento attraverso studi del cambiamento nelle prossimità fra individui *i*, studio del cambiamento delle relazioni fra variabili *j*, studio dei cambiamenti delle variabili o degli individui, eventualmente attraverso una matrice a via media (Escoufier, 1973; Glacon, 1981 e Lavit, 1988).

Bibliografia:

Banet T.A., Lebart L. (1984), *"Local and partial principal component analysis"*, COMPSTAT '84.

Carrol J.D., Chang J.J. (1970), *Analysis of individual differences in multidimensional scaling via an N-way generation of "Eckart-Young" decomposition,* Psychometrica n.35.

Coppi R., Bolasco S. (1989), *Analysis of multiway data matrices*, Elsevier.

Diday et al.ii, (1994), *New approaches in classification and data analysis*, Springer Verlag, New York.

Escoufier Y. (1973), *Le traitement des variables vectorielles*, BIOMETRICS, n.29.

Glacon F. (1981), *Analyse conjointe de plusieurs matrices de donnees*, USM, Grenoble.

Guilford J.P., Hoepfner R. (1971), *The analysis of intelligence*, Mc Graw-Hill, New York.

Lavit C. (1988), *Analyse conjointe de tableaux quantitatifs*, Masson.

Hayashi C. (1988), *New Develepmonents in Multidimensional Data Analysis, recent developments in clustering and data analysis,* Academic Press, New York.

Law H.G., Snyder Jr., Hattie J.A, Mc Donald R.P. (1984), *Research methods for multimode data analysis*, Preager, New York.

Lawley D., Maxwell A.E. (1980), *Factor analysis as statistical method*, Griffin, London.

Joreskog K.G. (1967), *Some contributions to maximum likelihood factor analysis*, Psyconometrica.

Kier H. (1989), *Comparison of anglo-saxon and french three mode methods*, A paratre dans Statistique et Analyse de Donnes.

Kim J.O., Mueller C.W. (1978), *Factor analysis: statistical methods and practical issues*, Sage Pubblications.

Kruskal J.B. (1977), *Three-way arrays: rank and uniqueness of trilinear decomposition with application to arithmetic complexity and statistics*, Linear Algebra and Its Application n. 18.

Kroonenberg P.M. (1983), *Three-mode principal component analysis: theory and application*, Leiden.

Morrison D.F. (1976); *Multivariate statistical method*, Mc Graw-Hill, New York.

Mulaik S.A. (1972), *The foundations of factor analysis*, Mc Graw Hill, New York.

Sadocchi S. (1980), *Manuale di analisi statistica multivariata*, F. Angeli, Milano.

Shonemann P.H., Wang M.M. (1972), *Some new results on factor indeterminacy*, Psycometrika.

Stevens J. (1986), *Applied multivariate statistics for the social sciences*, Hillsdale.

Tenenhaus M. (1986), *Generalized canonical analysis, canonical nalysis and applcations*, Les cahiers de recherche.

Tucker L.R. (1963), *Implication of factor analysis of three-way matrices matrices for measurement of change, problems in Measuring Change.* University of Wisconsin Press.
Tucker L.R. (1964), *The extension of factor analysis to three-dimension matrices. Contributions to Mathematical Psicology,* Frederiksen N., Holt, Rinehart and Winston.
Tucker L.R. (1966), *Some mathematical notes on three-mode factor analysis*, Psychometrica, 31.