Analisi Fattoriale

- è un'estensione dell'analisi in componenti principali che si pone come obiettivo di descrivere, se possibile, le molte variabili osservate in funzione di poche variabili sottostanti non osservabili (latenti) cui si dà il nome di "fattori"
- affonda le sue radici nei tentativi di inizio XX secolo, ad opera di Karl Pearson e Charles Spearman, volti a definire e misurare l'intelligenza

© 17 maggio 2005 Luca La Rocca

Stock-price data (Johnson & Wichern, 2002, Chapter 8)

Week	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
1	0.000000	0.000000	0.000000	0.039473	−0.000000
2	0.027027	−0.044855	−0.003030	−0.014466	0.043478
3	0.122807	0.060773	0.088146	0.086238	0.078124
4	0.057031	0.029948	0.066808	0.013513	0.019512
5	0.063670	−0.003793	−0.039788	−0.018644	−0.024154
6	0.003521	0.050761	0.082873	0.074265	0.049504
7	−0.045614	−0.033007	0.002551	−0.009646	−0.028301
8	0.058823	0.041719	0.081425	−0.014610	0.014563
...	...	...	...	...	...
93	−0.039457	−0.029297	−0.065844	−0.015837	−0.045758
94	0.039568	0.024145	−0.006608	0.028423	−0.009661
95	−0.031142	−0.007941	0.011080	0.007537	0.014634
96	0.000000	−0.020080	−0.006579	0.029925	−0.004807
97	0.021429	0.049180	0.006622	−0.002421	0.028985
98	0.045454	0.046375	0.074561	0.014563	0.018779
99	0.050167	0.036380	0.004082	−0.011961	0.009216
100	0.019108	−0.033303	0.008362	0.033898	0.004566

Weekly rates of return for five stocks on the New York Stock Exchange

© 17 maggio 2005 Luca La Rocca

Le componenti principali

Si è visto come si costruiscono

$$
\begin{aligned}
C_1 &= +0.56 \cdot AC + 0.47 \cdot DP + 0.55 \cdot UC + 0.29 \cdot E + 0.28 \cdot T \\
C_2 &= +0.74 \cdot AC - 0.09 \cdot DP - 0.65 \cdot UC - 0.11 \cdot E + 0.07 \cdot T \\
C_3 &= -0.13 \cdot AC - 0.47 \cdot DP - 0.11 \cdot UC + 0.61 \cdot E + 0.62 \cdot T \\
C_4 &= +0.28 \cdot AC - 0.69 \cdot DP + 0.50 \cdot UC - 0.44 \cdot E + 0.06 \cdot T \\
C_5 &= -0.21 \cdot AC + 0.28 \cdot DP - 0.10 \cdot UC - 0.58 \cdot E + 0.72 \cdot T
\end{aligned}
$$

combinazioni lineari delle variabili osservate
ognuna delle quali spiega la massima varianza possibile
sotto il vincolo di non correlazione con le precendenti

© 17 maggio 2005 Luca La Rocca

Un altro punto di vista

Si può mostrare che

$$
\begin{aligned}
AC &= +0.56 \cdot C_1 + 0.74 \cdot C_2 - 0.13 \cdot C_3 + 0.28 \cdot C_4 - 0.21 \cdot C_5 \\
DP &= +0.47 \cdot C_1 - 0.09 \cdot C_2 - 0.47 \cdot C_3 - 0.69 \cdot C_4 + 0.28 \cdot C_5 \\
UC &= +0.55 \cdot C_1 - 0.65 \cdot C_2 - 0.11 \cdot C_3 + 0.50 \cdot C_4 - 0.10 \cdot C_5 \\
E &= +0.29 \cdot C_1 - 0.11 \cdot C_2 + 0.61 \cdot C_3 - 0.44 \cdot C_4 - 0.58 \cdot C_5 \\
T &= +0.28 \cdot C_1 + 0.07 \cdot C_2 + 0.62 \cdot C_3 + 0.06 \cdot C_4 + 0.72 \cdot C_5
\end{aligned}
$$

ovvero le variabili osservate si esprimono come
combinazioni lineari delle componenti principali,
mediante gli stessi pesi

© 17 maggio 2005 Luca La Rocca

Spiegazione semplificata dei dati

Tenere solamente le prime (es. due) componenti principali

$$\begin{aligned} AC &\simeq +0.56 \cdot C_1 + 0.74 \cdot C_2 \\ DP &\simeq +0.47 \cdot C_1 - 0.09 \cdot C_2 \\ UC &\simeq +0.55 \cdot C_1 - 0.65 \cdot C_2 \\ E &\simeq +0.29 \cdot C_1 - 0.11 \cdot C_2 \\ T &\simeq +0.28 \cdot C_1 + 0.07 \cdot C_2 \end{aligned}$$

fornisce un'approssimazione delle variabili osservate
che permette di considerarle misurazioni affette da errore
di un fenomeno sottostante essenzialmente più semplice

© 17 maggio 2005 Luca La Rocca

Che spiegazione vorremmo?

In generale, avendo osservato p variabili $X_1, X_2, \ldots, X_p$,
vorremmo rappresentarle come sovrapposizione delle loro
medie $\mu_1, \mu_2, \ldots, \mu_p$, di $m < p$ fattori latenti $F_1, F_2, \ldots, F_m$,
a media nulla, varianza unitaria e fra loro non corrrelati,
e di p termini di errore $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$, anch'essi a media nulla
e non correlati, né fra loro né con i fattori latenti:

$$\begin{aligned} X_1 &= \mu_1 + L_{11} \cdot F_1 + L_{12} \cdot F_2 + \cdots + L_{1m} \cdot F_m + \epsilon_1 \\ X_2 &= \mu_2 + L_{21} \cdot F_1 + L_{22} \cdot F_2 + \cdots + L_{1m} \cdot F_m + \epsilon_2 \\ &\cdots \\ X_p &= \mu_p + L_{p1} \cdot F_1 + L_{p2} \cdot F_2 + \cdots + L_{pm} \cdot F_m + \epsilon_p \end{aligned}$$

Questo è noto come modello fattoriale ortogonale;
il coefficiente L_{jk} si dice loading del fattore k sulla variabile j

© 17 maggio 2005 Luca La Rocca

La spiegazione delle componenti principali

Si può senz'altro supporre che le variabili osservate abbiano media nulla; dopo di che basta prendere, per $j = 1, \ldots, p$ e $k = 1, \ldots, m$,

$$\begin{aligned} L_{jk} &= \sqrt{\lambda_k} \cdot W_{jk} \\ F_k &= \frac{1}{\sqrt{\lambda_k}} \cdot C_k \\ \epsilon_j &= W_{j(m+1)} \cdot C_{m+1} + \ldots + W_{jp} \cdot C_p \end{aligned}$$

dove

- W_{jk} è il peso di X_j in C_k (ovvero di C_k in X_j)
- λ_k è la varianza di C_k (varianza spiegata da C_k)

e questa è nota come analisi fattoriale
con il metodo delle componenti principali

© 17 maggio 2005 Luca La Rocca

È una buona spiegazione?

- I termini di errore risultano fra loro correlati, infatti per es.

  $$\mathrm{Cov}[\epsilon_1, \epsilon_2] = W_{1(m+1)} \cdot W_{2(m+1)} \cdot \lambda_{m+1} + \cdots + W_{1p} \cdot W_{2p} \cdot \lambda_p$$

  tuttavia tale correlazione è trascurabile se $\lambda_{m+1} + \cdots + \lambda_p$ è "piccola", ovvero se la prime m componenti principali riescono a spiegare la maggior parte della variabilità osservata

- Poi si vorrebbe che il ruolo giocato dai fattori sia preminente rispetto a quello giocato dai termini di errore, il che si traduce nella stessa condizione, dal momento che (per es.)

  $$\mathrm{Var}[\epsilon_1] = W_{1(m+1)}^2 \cdot \lambda_{m+1} + \cdots + W_{1p}^2 \cdot \lambda_p$$

. . . dunque si, nella misura in cui le p variabili osservate si lasciano rappresentare dalle loro prime m componenti principali. . .

© 17 maggio 2005 Luca La Rocca

Communalities

Nel modello fattoriale ortogonale (anche con termini di errore correlati) si ha la decomposizione delle varianze osservate

$$\begin{aligned}
\mathrm{Var}[X_1] &= L_{11}^2 + \cdots + L_{1m}^2 + \mathrm{Var}[\epsilon_1] &= h_1^2 + \psi_1 \\
\mathrm{Var}[X_2] &= L_{21}^2 + \cdots + L_{2m}^2 + \mathrm{Var}[\epsilon_2] &= h_2^2 + \psi_2 \\
&\cdots \\
\mathrm{Var}[X_p] &= L_{p1}^2 + \cdots + L_{pm}^2 + \mathrm{Var}[\epsilon_p] &= h_p^2 + \psi_p
\end{aligned}$$

dove si dice che

- h_j^2 è la communality della j-esima variabile
- ψ_j è la varianza specifica della j-esima variabile

e l'interpretazione è facilitata se le variabili osservate sono state standardizzate, come si conviene generalmente di fare...

© 17 maggio 2005 Luca La Rocca

Stock market factors

Stock	F_1	F_2	Communality
Allied Chemical	0.783	−0.217	0.661
Du Pont	0.773	−0.458	0.806
Union Carbide	0.794	−0.234	0.686
Exxon	0.713	0.472	0.731
Texaco	0.712	0.524	0.781
		Total	(73.3%) 3.666

La tabella riporta, per un'analisi fattoriale a due fattori dei dati di esempio, effettuata con il metodo delle componenti principali, sia i factor loading che le communality: i due fattori spiegano complessivamente il 73.3% della varianza totale e in particolare, per esempio, il 66.1% della varianza del primo titolo

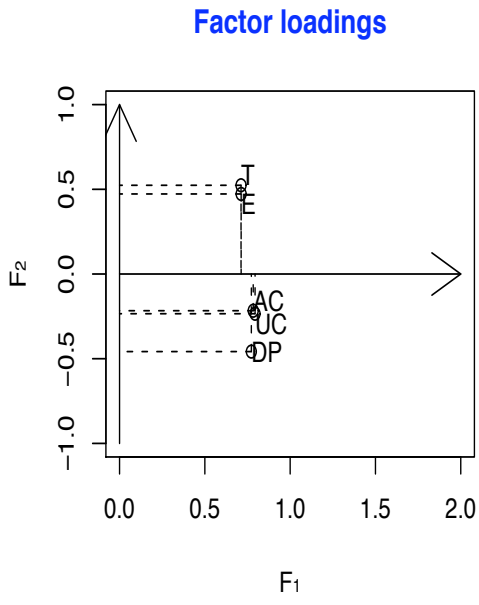
© 17 maggio 2005 Luca La Rocca

Altre possibili spiegazioni

Il metodo delle componenti principali non esaurisce l'analisi fattoriale, per almeno tre ragioni:

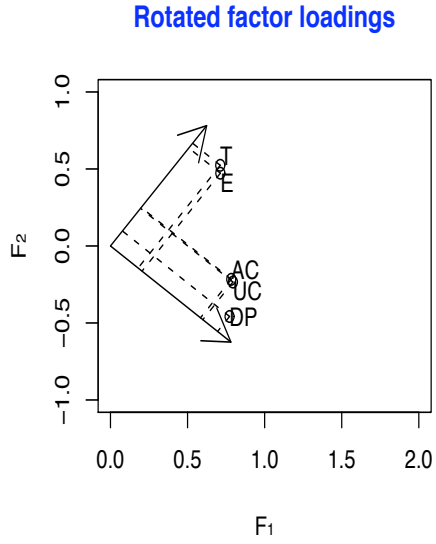
- perché è sempre possibile effettuare una rotazione dei fattori, tecnica che ci si accinge a discutere e ad applicare
- perché una valida alternativa per calcolare i factor loading e le communality è data dal metodo della massima verosimiglianza (maximum likelihood), che applicheremo senza discussione
- perché si può pensare di sostituire il modello fattoriale ortogonale con una altro modello, per esempio pescando nella famiglia dei modelli a equazioni strutturali, opzione per la quale si rimanda senz'altro a Diamantopoulos (1994)

© 17 maggio 2005 Luca La Rocca



L'interpretazione dei fattori dipende dai loro loading...

© 17 maggio 2005 Luca La Rocca



. . . perciò una rotazione degli assi coordinati, facilitando l'associazione delle variabili ai fattori, può essere di aiuto. . .

© 17 maggio 2005 Luca La Rocca

Rotated stock market factors

Stock	F_1'	F_2'	Communality
Allied Chemical	0.747	0.320	0.661
Du Pont	0.889	0.124	0.806
Union Carbide	0.767	0.313	0.686
Exxon	0.262	0.814	0.731
Texaco	0.229	0.854	0.781
		Total	(73.3%) 3.666

. . . al fine di ottenere fattori che rappresentino gruppi di variabili:

- il fattore F_1' rappresenta i titoli del settore chimico
- il fattore F_2' rappresenta i titoli del settore petrolifero

© 17 maggio 2005 Luca La Rocca

Cosa cambia?

- cambiano i factor loading, auspicabilmente in modo da facilitare l'interpretazione dei fattori
- non cambiano le communality (dunque nemmeno la percentuale di varianza totale spiegata), cosicché la bontà della spiegazione è preservata
- i fattori contribuiscono in modo più paritario alla spiegazione della varianza totale (in quanto non più vincolati a spiegare nell'ordine la massima varianza possibile)

Che poi la sostituzione di un forte fattore di mercato coadiuvato da un contrasto di settore con due fattori di settore più o meno paritari costituisca un vantaggio per l'interpretazione dell'esempio studiato è, ovviamente, un'affermazione discutibile

© 17 maggio 2005 Luca La Rocca

Quale rotazione?

- Nel caso di due soli fattori si può procedere per via grafica...
- ...tuttavia, anche per il semplice esempio studiato, è stato più agevole adottare il cosiddetto criterio VARIMAX, ovvero scegliere la rotazione ortogonale che massimizza

$$V = \sum_{j=1}^{m} \left[\frac{1}{p} \sum_{i=1}^{p} L_{ij}^4 - \left(\frac{1}{p} \sum_{i=1}^{p} L_{ij}^2 \right)^2 \right]$$

Si può migliorare ulteriormente l'interpretazione dei fattori, se si considerano anche rotazioni non rigide; in questo modo si rinuncia all'ortogonalità dei fattori e di fatto si considera il cosiddetto modello fattoriale obliquo (non approfondiamo)

© 17 maggio 2005 Luca La Rocca

Maximum likelihood (rotated) stock market factors

```
Call:
factanal(x = X, factors = 2)

Uniquenesses:
Allied.Chemical         Du.Pont   Union.Carbide           Exxon          Texaco
          0.497           0.252           0.474           0.610           0.176

Loadings:
                Factor1 Factor2
Allied.Chemical 0.601   0.378
Du.Pont         0.849   0.165
Union.Carbide   0.643   0.336
Exxon           0.365   0.507
Texaco          0.207   0.884

               Factor1 Factor2
SS loadings      1.671   1.321
Proportion Var   0.334   0.264
Cumulative Var   0.334   0.598

Test of the hypothesis that 2 factors are sufficient.
The chi square statistic is 0.58 on 1 degree of freedom.
The p-value is 0.448
```

© 17 maggio 2005 Luca La Rocca

Risultati differenti?

- anche il metodo della massima verosimiglianza individua come fattori il settore chimico e quello petrolifero (sebbene l'interpretazione sia meno nitida, in quanto i loading sono meno "sblilanciati")
- questo è un buon segno, perché se davvero vi sono dei fattori sottostanti significativi qualsiasi metodo dovrebbe individuarli
- la percentuale di varianza totale spiegata è scesa al 60%; ciò non sorprende, per come sono costruite le componenti principali

Se si vogliono conoscere i valori dei fattori in corrispondenza delle singole osservazioni (l'analogo delle componenti principali) occorre calcolare i cosiddetti factor score (non approfondiamo)

© 17 maggio 2005 Luca La Rocca

Quando essere soddisfatti?

The WOW criterion

If, while scritinizing the factor analysis,
the investigator can shout "Wow, I understand these factors",
the application is deemed successful.

Questo per dire che l'analisi fattoriale presenta una forte componente soggettiva e rimane più un'arte che una scienza; tra i buoni consigli il più importante è forse quello di ricercare la stabilità dei risultati, anche al variare dei dati analizzati (l'ipotesi di lavoro è che davvero vi siano dei fattori sottostanti, anche se magari non è noto a priori quanti siano)

Si noti che i risultati conseguibili con un modello fattoriale dipendono dalla struttura della matrice delle correlazioni...

© 17 maggio 2005 Luca La Rocca

Riferimenti

R. A. Johnson & D. W. Wichern (2002). Applied Multivariate Statistical Analysis. Prentice-Hall, Upper Saddle River, NJ.

L. Fabbris (1997). Statistica Multivariata. McGraw-Hill, Milano.

L. Molteni & G. Troilo (2003). Ricerche di Marketing. McGraw-Hill, Milano.

A. Diamantopoulos (1994). Modelling with LISREL. Journal of Marketing Management, 10, 105–136.

http://www-dimat.unipv.it/luca/

mailto://larocca.luca@unimore.it

© 17 maggio 2005 Luca La Rocca