Analisi in Componenti Principali

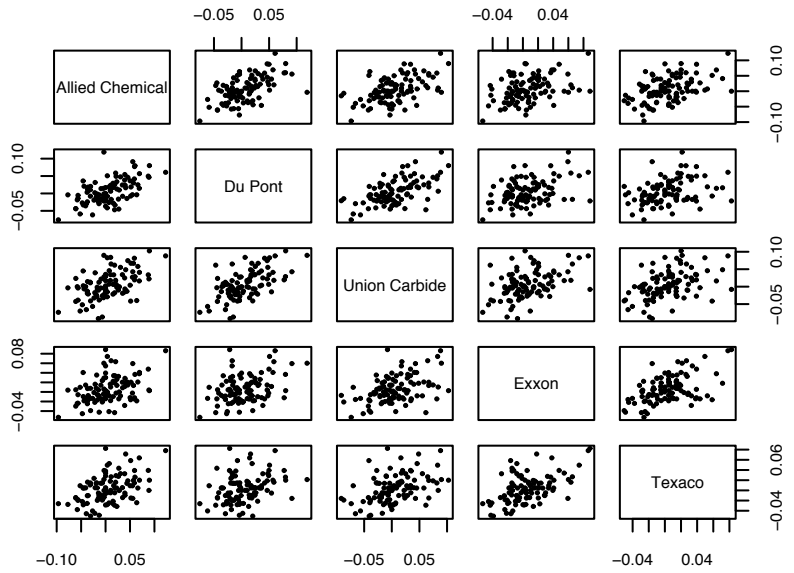
- tecnica di riduzione e interpretazione dei dati
- spesso gioca un ruolo ausiliario rispetto ad altre tecniche (es. analisi fattoriale, cluster analysis)
- adatta per variabili quantitative

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

Prospetto dei diagrammi di dispersione



Matrice delle correlazioni

	Allied Chemical	Du Pont	Union Carbide	Exxon	Texaco
Allied.Chemical	1.000	0.577	0.509	0.387	0.462
Du.Pont	0.577	1.000	0.598	0.390	0.322
Union.Carbide	0.509	0.598	1.000	0.436	0.426
Exxon	0.387	0.390	0.436	1.000	0.524
Texaco	0.462	0.322	0.426	0.524	1.000

Vettori delle medie e delle deviazioni standard

	Mean	Standard Deviation
Allied Chemical	0.005434	0.040
Du Pont	0.004827	0.035
Union Carbide	0.005654	0.039
Exxon	0.006291	0.028
Texaco	0.003709	0.028

La prima componente principale

È una combinazione lineare delle variabili osservate

$$PC1 = w_1^{AC} \cdot AC + w_1^{DP} \cdot DP + w_1^{UC} \cdot UC + w_1^{E} \cdot E + w_1^{T} \cdot T$$

ottenuta scegliendo il vettore dei pesi w_1 in modo da

- massimizzare la deviazione standard di PC1
- sotto il vincolo $(w_1^{AC})^2 + (w_1^{DP})^2 + (w_1^{UC})^2 + (w_1^{E})^2 + (w_1^{T})^2 = 1$

cosicché si trova

$$PC1 = 0.56 \cdot AC + 0.47 \cdot DP + 0.55 \cdot UC + 0.29 \cdot E + 0.28 \cdot T$$

con deviazione standard di PC1 pari a 0.060

La seconda componente principale

È una combinazione lineare delle variabili osservate

$$PC2 = w_2^{AC} \cdot AC + w_2^{DP} \cdot DP + w_2^{UC} \cdot UC + w_2^{E} \cdot E + w_2^{T} \cdot T$$

ottenuta scegliendo il vettore dei pesi w_2 in modo da

- massimizzare la deviazione standard di PC2
- sotto il vincolo che PC2 non sia correlata con PC1
- e inoltre $(w_2^{AC})^2 + (w_2^{DP})^2 + (w_2^{UC})^2 + (w_2^{E})^2 + (w_2^{T})^2 = 1$

cosicché si trova

$$PC2 = 0.74 \cdot AC - 0.09 \cdot DP - 0.65 \cdot UC - 0.11 \cdot E + 0.07 \cdot T$$

con deviazione standard di PC2 pari a 0.028

Le ulteriori componenti principali

Vi sono in tutto tante componenti principali quante sono le variabili osservate e ognuna si ottiene come combinazione lineare a varianza massima sotto il vincolo di non correlazione con tutte le precedenti

cosicché si trovano

$$
\begin{aligned}
PC3 &= -0.13 \cdot AC - 0.47 \cdot DP - 0.11 \cdot UC + 0.61 \cdot E + 0.62 \cdot T \\
PC4 &= +0.28 \cdot AC - 0.69 \cdot DP + 0.50 \cdot UC - 0.44 \cdot E + 0.06 \cdot T \\
PC5 &= -0.21 \cdot AC + 0.28 \cdot DP - 0.10 \cdot UC - 0.58 \cdot E + 0.72 \cdot T
\end{aligned}
$$

con deviazioni standard di PC3, PC4 e PC5 rispettivamente pari a 0.027, 0.023 e 0.019

Proprietà delle componenti principali

In definitiva si sono rimpiazzate le variabili osservate AC, DP, UC, E e T con le componenti principali PC1, PC2, PC3, PC4 e PC5...

...con quali vantaggi?

Intanto osserviamo che per costruzione le componenti principali

- sono fra loro non correlate (ortogonali)
- sono ordinate in ordine decrescente di varianza

inoltre si può mostrare che

- la varianza totale (somma delle varianze) si conserva nel passaggio dalle variabili osservate alle componenti principali

Vantaggi delle componenti principali

Alla luce delle loro proprietà possiamo dire che le componenti principali forniscono una spiegazione alternativa della variabilità osservata con il pregio di descrivere il fenomeno oggetto di studio mediante dimensioni fra loro non correlate e ordinate in termini della loro importanza nella spiegazione

Questo permette (con maggiore o minore successo nei vari casi) di

- ridurre il numero di variabili da considerare, scartando le ultime componenti principali (laddove si ritenga trascurabile il loro contributo alla spiegazione della variabilità osservata)
- interpretare il fenomeno oggetto di studio, mediante un'opportuna interpretazione delle componenti principali che non sono state scartate

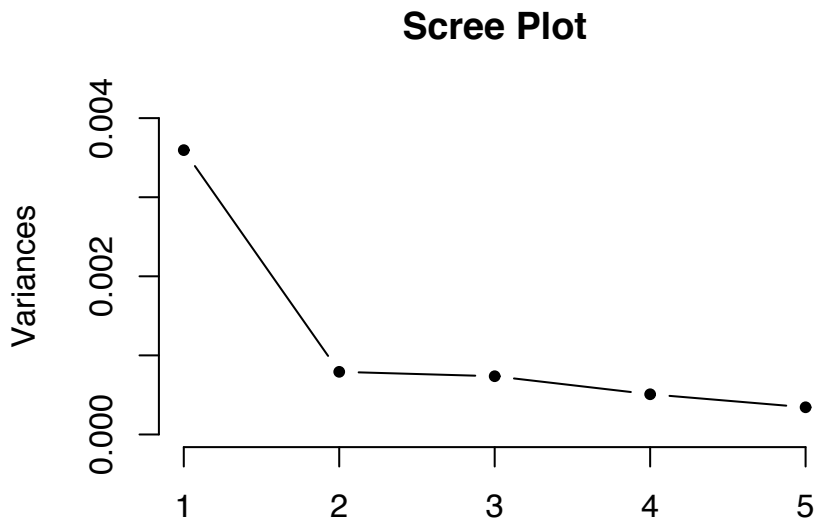
Quante componenti principali tenere?

Non vi è una risposta definitiva a questa domanda; occorre valutare caso per caso

- la percentuale di varianza totale spiegata da ogni componente principale, anche in relazione a quella spiegata dalle altre
- la conoscenza specifica del fenomeno oggetto di studio
- l'uso che delle componenti principali si vuole fare (per esempio, se si vogliono rappresentare graficamente i dati, è giocoforza accontentarsi delle prime due o tre componenti)

In ogni caso ci si può aiutare con il cosiddetto scree plot, che letteralmente è il "diagramma della falda detritica"...

Alla ricerca del "gomito"



Varianza spiegata

	Standard Deviation	Variance	Percentage
PC1	0.060	0.003595	60.2%
PC2	0.028	0.000792	13.3%
PC3	0.027	0.000736	12.3%
PC4	0.023	0.000509	8.5%
PC5	0.019	0.000344	5.8%
TOTAL		0.005976	100%

La prima componente principale spiega da sola la maggior parte della varianza totale; i contributi delle altre componenti principali sono nettamente inferiori, come mostra il "gomito" dello scree plot

Interpretazione

Ricordando che la prima componente principale si scrive

$$PC1 = 0.56 \cdot AC + 0.47 \cdot DP + 0.55 \cdot UC + 0.29 \cdot E + 0.28 \cdot T$$

si vede che essa risulta dal contributo positivo di tutte le variabili osservate e pertanto PC1 può interpretarsi come una componente generale del mercato azionario di riferimento

Nella misura in cui si accetta di descrivere i cinque titoli AC, DP, UC, E e T mediante la sola PC1 si può dire che il loro andamento è essenzialmente quello del mercato e che pertanto questi titoli da soli non consentono una strategia di differenziazione

Ma facciamo un passo indietro...

Cosa accade se...

...prima di costruire le componenti principali si centrano le variabili?

Si ottengono le stesse componenti principali, solo centrate (questo perché non cambia la matrice delle covarianze)

...prima di costruire le componenti principali si standardizzano le variabili?

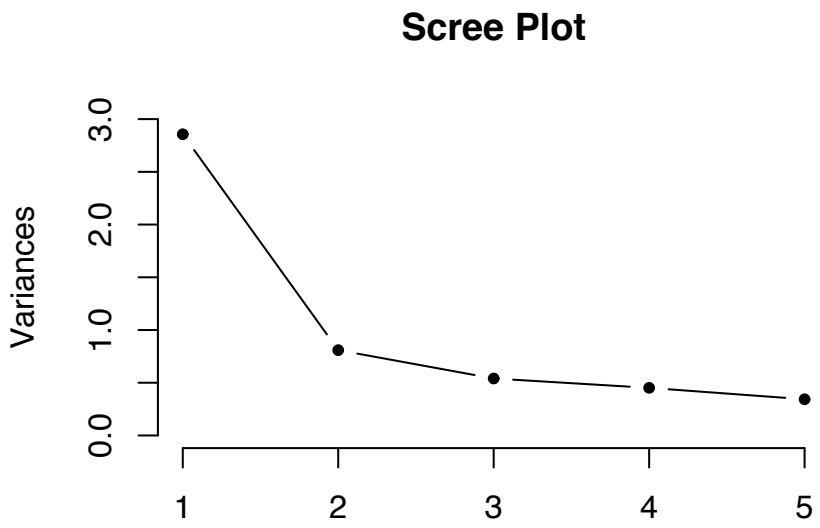
Si ottengono altre componenti principali (perché la matrice delle covarianze è rimpiazzata da quella delle correlazioni)

ACP sulle variabili standardizzate

	PC1	PC2	PC3	PC4	PC5
Allied Chemical	0.46	−0.24	0.61	−0.38	0.45
Du Pont	0.46	−0.51	−0.18	−0.21	−0.67
Union Carbide	0.47	−0.26	−0.34	0.66	0.40
Exxon	0.42	0.53	−0.54	−0.47	0.18
Texaco	0.42	0.58	0.43	0.38	−0.39
VARIANCE	2.86	0.81	0.54	0.45	0.34
PERCENTAGE	57%	16%	11%	9%	7%

La varianza totale delle variabili standardizzate è pari al numero delle variabili stesse (5) e inoltre il valore 1 si offre come valore di riferimento "naturale" per le varianze delle componenti principali

Un altro "gomito"?



Qualcosa è cambiato...

La varianza spiegata da PC1 è diminuita; in compenso

$$PC1 = 0.46 \cdot AC + 0.46 \cdot DP + 0.47 \cdot UC + 0.42 \cdot E + 0.42 \cdot T$$

si interpreta meglio come componente di mercato alla quale tutti i titoli contribuiscono più o meno paritariamente

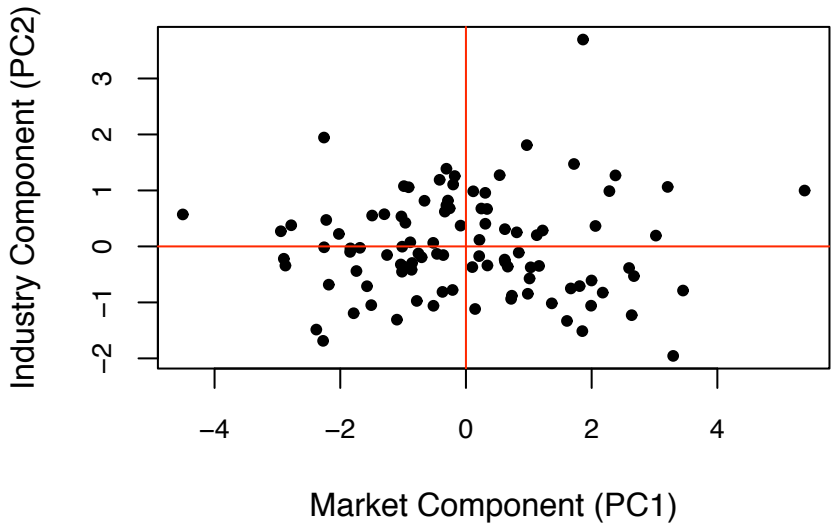
D'altra parte la varianza spiegata da PC2 è aumentata e inoltre

$$PC2 = -0.24 \cdot AC - 0.51 \cdot DP - 0.26 \cdot UC + 0.53 \cdot E + 0.58 \cdot T$$

si lascia interpretare come componente di settore che contrasta gli ultimi due titoli (petrolifero) con i primi tre (chimico)

In definitiva, affiancando PC2 a PC1, si arriva a spiegare il 73% della varianza totale e si introduce un elemento di differenziazione

C'è settimana e settimana



Cosa è successo?

Senza standardizzazione le variabili con varianza maggiore tendono a fare la parte del leone nello spiegare la varianza totale, cosicché i primi tre titoli si "scaricano" completamente su PC1 e questo finisce per oscurare il loro ruolo di PC2

Pertanto, in generale, conviene standardizzare le variabili osservate, a meno che non vi siano ragioni particolari per preservare proprio la scala sulla quale sono state rilevate

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.

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

mailto://larocca.luca@unimore.it