Analisi in Componenti Principali (ACP)

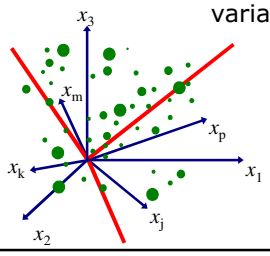
Metodi di analisi fattoriale

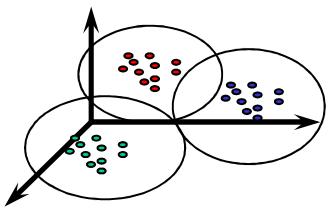
Obiettivo*:* individuazione di variabili di sintesi = dimensioni = variabili latenti = variabili non osservate

Approccio*:* Ordinamenti tra variabili/mutabili

Metodi:

- Analisi in Componenti Principali (ACP) per variabili quantitative
- Analisi delle Corrispondenze Binarie (ACB) per tabelle di contingenza
- Analisi delle Corrispondenze Multiple (ACM) per variabili qualitative

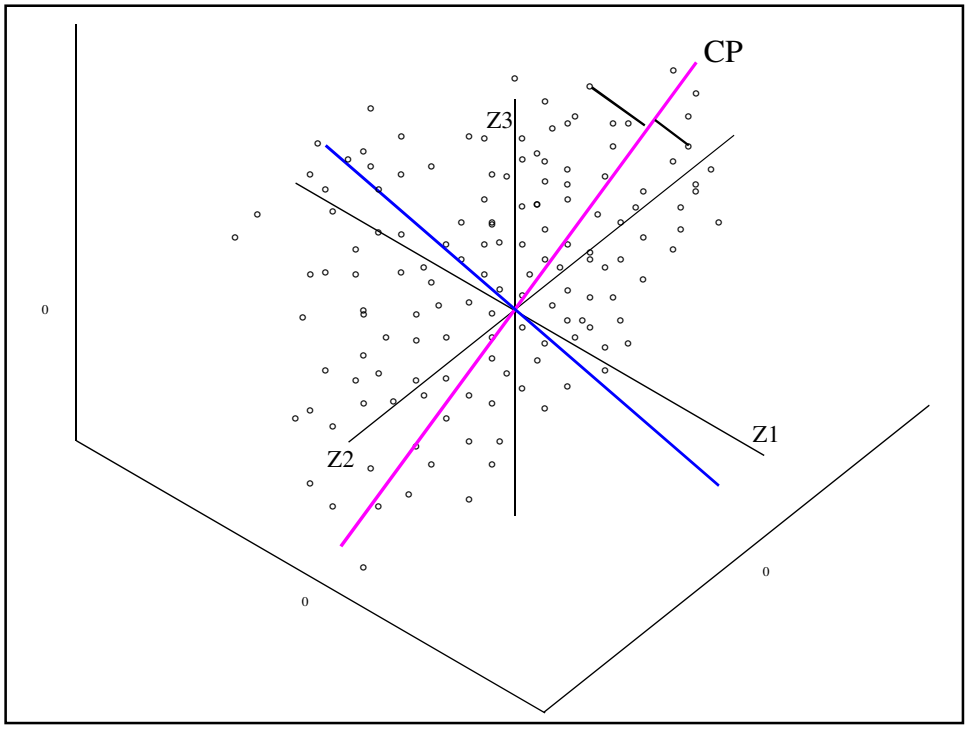




Analisi in Componenti Principali (ACP)

- L'obiettivo:

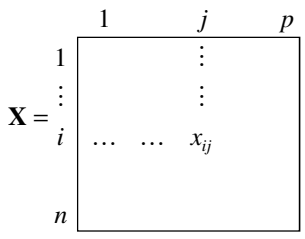
Ridurre la dimensionalità dell'insieme dei dati eliminando la ridondanza di informazioni risultato di p variabili altamente correlate e sostituendo a queste ultime ***un minor numero h (h < p) di nuove variabili tra loro non correlate e legate linearmente alle variabili di partenza***.



Riduzione della dimensionalità

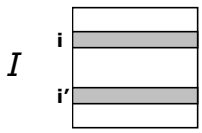
- Le nuove variabili oltre ad essere non correlate sono ordinate rispetto alla percentuale di variabilità presente nei dati originali.

Lettura geometrica della matrice dei dati

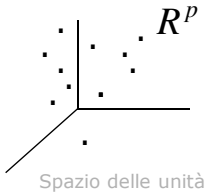


Matrice dei dati

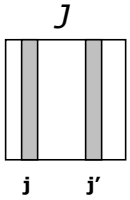
- I vettori riga di **X** sono **punti-unità** nello spazio $\Re^p$ generato dalle variabili.
- I vettori colonna di **X** sono **punti-variabile** nello spazio $\Re^n$ generato dalle unità.



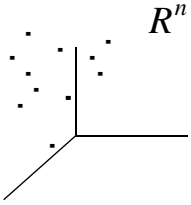
Profili riga



Spazio delle unità



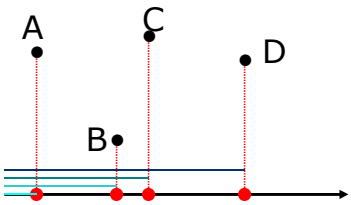
Profili colonna



Spazio delle variabili

Sintesi delle informazioni

Per sintetizzare le relazioni esistenti tra più punti-unità (o punti-variabile) in uno spazio multidimensionale è possibile proiettare tutti i punti su un asse e studiare le distanze tra i vari punti.



Tali proiezioni costituiscono un'approssimazione delle relazioni esistenti tra i vari punti in quanto le distanze originarie risultano deformate.

Obiettivo e finalità operativa

Obiettivo: Sintetizzare le informazioni a disposizione garantendo la minima perdita di informazione (in termini di relazioni tra i dati).

Finalità operativa: Ricerca di un sistema di assi fattoriali ortogonale che generi il sottospazio di "migliore" approssimazione tale da deformare il meno possibile le distanze tra i punti.

… La variabilità

Nella statistica univariata, la variabilità è rappresentata dagli indici di variabilità:

Varianza, deviazione standard, devianza, ecc.

Nella statistica multivariata, la variabilità è definita a partire dalla matrice di varianze e covarianze:

… La matrice di varianze e covarianze

Se **X** è una matrice dei dati unità per variabili di dimensioni *n,k*, la matrice di varianze e covarianze è:

	x_1	x_2	x_3	x_4	x_j	x_k
x_1	Var_{x1}					
x_2	$Cov_{1,2}$	Var_{x2}				
x_3	$Cov_{1,3}$	$Cov_{2,3}$	Var_{x3}			
x_4	$Cov_{1,4}$	$Cov_{2,4}$	$Cov_{3,4}$	Var_{x4}		
x_j	$Cov_{1,j}$	$Cov_{2,j}$	$Cov_{3,j}$	$Cov_{4,j}$	Var_{xj}	
x_k	$Cov_{1,k}$	$Cov_{2,k}$	$Cov_{3,k}$	$Cov_{4,k}$	$Cov_{j,k}$	Var_{xk}

La variabilità del sistema *k*-variato viene sintetizzato con la *traccia* della matrice di var-cov

A	B	C	D
7,51	4,90	4,05	75,49
9,12	12,92	7,70	14,51
5,28	9,60	1,99	86,61
5,69	17,51	6,96	8,01
0,06	13,36	5,87	35,28
7,02	3,30	5,72	60,48
7,36	21,65	1,74	19,27
0,34	15,54	2,26	69,93
2,00	29,05	6,94	52,14
4,39	26,25	0,44	37,23
6,84	13,25	1,87	32,70
4,15	21,63	0,03	29,77
7,60	11,57	3,90	76,20

Matrice dei dati

	A	B	C	D
A	7,65			
B	-7,99	54,35		
C	0,53	-3,77	6,32	
D	-8,28	-81,86	-11,23	617,84

Matrice di varianze e covarianze

Variabilità = 686,17

Gli autovalori della matrice di var-cov sono:

4,54
6,25
45,60
629,78

La cui somma è 686,17!!!

Gli autovalori ricostruiscono la variabilità della matrice dei dati

Definizione delle Componenti Principali

La prima CP è la combinazione lineare delle *p* variabili di partenza avente massima varianza; la seconda CP è la combinazione lineare delle *p* variabili con varianza immediatamente inferiore, soggetta al vincolo di essere ortogonale alla componente precedente, e così via.

La determinazione della prima CP richiede l'individuazione del vettore *p*-dimensionale u_1 dei coefficienti della seguente combinazione lineare delle *p* variabili espresse in termini degli scostamenti dalle loro medie:

$$c_1 = Xu_1$$

La varianza totale di una trasformazione lineare di X è esprimibile in funzione della matrice di Varianza-Covarianza S

$VAR(Xu_1) = u_1' S u_1$

Posto $u_1' u_1 = 1$ il vettore dei coefficienti u_1 deve essere tale da massimizzare questa espressione

Definizione delle Componenti Principali …segue

Si applica la formula dei moltiplicatori di Lagrange:

$$L = u_1' S u_1 - \lambda_1 (u_1' u_1 - 1) = \max$$

si deriva rispetto ad u_1 e si annullano le derivate parziali, ottenendo:

$$(S - \lambda_1 I) = 0$$

L'obiettivo è la massimizzazione della varianza della prima CP, si sceglie come λ_1 il massimo autovalore, in quanto:

$$u_1' S u_1 = u_1' \lambda_1 u_1 = \lambda_1$$

Cioè il primo autovalore è uguale alla varianza della prima CP

Definizione delle Componenti Principali …segue

La prima CP di *p* variabili, espresse in termini di *scostamenti dalla media*, è la combinazione lineare

$$c_1 = X u_1$$

in cui u_1 è l'autovettore corrispondente all'autovalore λ_1 più grande della matrice di varianza-covarianza S

$c_2 = X u_2$ Seconda CP

$u_2' u_2 = 1$ Vincolo di normalità

$u_1' u_2 = 0$ Vincolo di ortogonalità

La somma degli autovalori delle CP ricostruisce l'intera variabilità dello spazio originario.

La matrice di correlazione

Le CP ottenute dalla matrice di varianza-covarianza (combinazioni lineari degli scostamenti dalla media delle variabili originarie) sono lecite se le variabili sono espresse tutte nella stessa unità di misura

E' necessario, nel caso contrario, superare tale difficoltà considerando le variabili espresse in termini di scostamenti standardizzati, cioè il punto di partenza dell'ACP diviene la *matrice di correlazione.*

Criteri di scelta del numero di assi

1. **Variabilità spiegata**

   si fissa una soglia minima di variabilità spiegata
2. **Eigenvalue-one** (per variabili standardizzate)

   Poiché le variabili hanno varianza unitaria si scelgono solo gli autovalori maggiori di uno
3. **Scree-Test**

   si considerano i fattori i cui autovalori precedono il salto massimo di variabilità spiegata.

Decomposizione in valori singolari

- Sia **X** una matrice di rango **r**

$$\underset{n,p}{\mathbf{X}} = \underset{n,r}{\mathbf{V}}\,\underset{r,r}{\mathbf{T}}\,\underset{r,p}{\mathbf{U}'} \qquad \mathbf{U}'\mathbf{U} = \mathbf{I} \qquad \mathbf{V}'\mathbf{V} = \mathbf{I}$$

  – dove **U** e **V** sono matrici le cui colonne sono ortonormali e **T** è diagonale con elementi positivi.

↓

Cioè di norma unitaria e ortogonali

Principio: una matrice rettangolare può essere ricostruita nei suoi valori a partire dalla somma di matrici delle stesse dimensioni ma di rango 1, cioè come prodotto di un vettore colonna (n,1) per un vettore riga (1,p)

Modello di rango ridotto

- Obiettivo dei metodi fattoriali è quello di fornire una descrizione parsimoniosa e grafica della struttura latente (presente ma non osservata) della matrice dei dati.
- La decomposizione in valori singolari definisce il seguente modello, in forma scalare

$$x_{ij} = \tau_1 v_{i1} u_{j1} + \ldots + \tau_r v_{ir} u_{jr} = \sum_{\alpha=1}^{r} \tau_\alpha v_{i\alpha} u_{j\alpha}$$

- Il modello si dice di rango ridotto se si considerano le prime h dimensioni,
  con $h < r = \min(p,n)$.

Caso studio: indicatori di performance aziendali

- Le variabili:
  - ECON.PRO -> *economic profit* , differenziale tra rendimento del capitale investito ed il suo costo
  - CASH -> *cash flow* sul fatturato in %
  - LAVOR.VA -> costo del lavoro sul valore aggiunto, in%
  - ROE -> *return on equity*, utile netto sul patrimonio, in%
  - INDE.CAP -> indebitamento sul capitale proprio
  - FATTURATO

Caso studio: indicatori di performance aziendali

Il dataset

Azienda	ECON.PRO	CASH	LAVOR.VA	ROE	INDE.CAP	FATTURATO
Barilla	-25,40	7,39	59,54	4,20	0,83	2867
Eridania	-141,00	4,00	68,99	4,20	0,83	1693
Ferrero	65,80	9,61	53,70	21,12	-0,02	3031
Galbani	-71,90	8,40	56,32	2,66	-0,02	2136
Kraft	-32,00	5,88	72,11	3,20	0,35	1563
Lavazza	-28,90	4,96	39,08	5,29	-0,05	1117
Nestlè	-98,80	2,72	81,25	0,00	1,69	3463
Parmalat	-145,10	5,96	38,51	2,23	2,91	1664
Plasmon	31,70	27,76	31,35	24,60	1,35	858
Star	2,4	6,47	62,49	10,60	0,00	811

Data la disomogeneità delle variabili si procede standardizzando le stesse

Le 5000 società leader, supplemento a Milano Finanza, 1998; Zani, 2000

Matrice dei dati standardizzati

Azienda	ECON.PRO	CASH	LAVOR.VA	ROE	INDE.CAP	FATTURATO
Barilla	0,285	-0,137	0,210	-0,452	0,047	1,072
Eridania	-1,456	-0,639	0,830	-0,452	0,047	-0,257
Ferrero	1,659	0,192	-0,173	1,665	-0,878	1,257
Galbani	-0,415	0,013	-0,001	-0,644	-0,878	0,244
Kraft	0,186	-0,360	1,035	-0,577	-0,475	-0,404
Lavazza	0,232	-0,496	-1,132	-0,315	-0,910	-0,909
Nestlè	-0,821	-0,828	1,634	-0,977	0,982	1,746
Parmalat	-1,518	-0,348	-1,169	-0,698	2,309	-0,290
Plasmon	1,145	2,877	-1,639	2,100	0,612	-1,202
Star	0,704	-0,273	0,404	0,349	-0,856	-1,256

MATRICE DES CORRELATIONS

	ECON	CASH	LAVO	ROE	INDE	FATT
ECON	1.00					
CASH	0.53	1.00				
LAVO	-0.27	-0.62	1.00			
ROE	0.79	0.80	-0.51	1.00		
INDE	-0.57	0.08	-0.17	-0.20	1.00	
FATT	-0.09	-0.36	0.51	-0.24	0.11	1.00
	ECON	CASH	LAVO	ROE	INDE	FATT

L'osservazione della matrice di correlazione è una fase importante:

- *se tutte le variabili fossero non correlate tra di loro non avrebbe senso procedere con un metodo fattoriale, infatti si avrebbero tante componenti quante variabili osservate*

Autovalori della matrice di correlazione

0,097
0,150
0,341
0,919
1,491 ←
3,003 ←

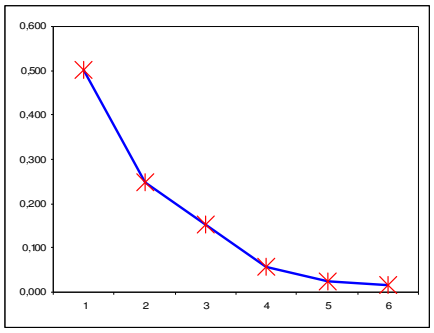
Li ordiniamo in ordine decrescente:

3,003
1,491
0,919
0,341
0,150
0,097

Calcoliamo la percentuale di variabilità spiegata da ognuno di essi:

percentuale	percentuale cumulata
0,501	0,501
0,249	0,749
0,153	0,902
0,057	0,959
0,025	0,984
0,016	1,000

Scree plot



Analisi dei punti-unità in $\Re^p$

- Il vettore **c**1 delle proiezioni degli n punti-unità sul primo asse fattoriale **u**1 (di norma unitaria) è dato da:

  c1=**Xu**1 (prima componente principale)

- La funzione obiettivo da massimizzare è:

  L1=(**Xu**1)2=**u'**1**X'Xu**1

  sotto il vincolo

$$\mathbf{u}_1'\mathbf{u}_1 = \sum_{j=1}^{p} \mathbf{u}_{1j}^2 = 1$$

… secondo asse principale

- Il secondo asse fattoriale $\mathbf{u}_2$ è un asse ortogonale al primo ($\mathbf{u}_1$) e di norma unitaria che massimizza la variabilità dei punti proiettati:

  $\mathbf{c}_2=\mathbf{X}\mathbf{u}_2$ (seconda componente principale)

- La funzione obiettivo è:

  $L_2=\mathbf{u'}_2\mathbf{X'X}\mathbf{u}_2-\lambda(\mathbf{u'}_2\mathbf{u}_2-1)=\max$

  sotto i vincoli: $\mathbf{u}_2'\mathbf{u}_2=\sum_{j=1}^{p}\mathbf{u}_{2j}^2=1$ e $\mathbf{u}_1'\mathbf{u}_2=0$

Importanza degli autovalori

Vale la relazione:

$$tr(\mathbf{X}'\mathbf{X})=\sum_{\alpha=1}^{p}\lambda_\alpha$$

Lo spazio *p*-dimensionale definito dagli assi fattoriali ricostruisce esattamente la variabilità della nube dei punti nello spazio originario R^p.

λ_α rappresenta la **varianza spiegata dalla α-ma componente principale**.

Coordinate dell'i-mo punto-unità

Una volta determinato il sottospazio ottimale ad h dimensioni individuato dagli h autovettori $\{\mathbf{u}_1,\mathbf{u}_2,\ldots,\mathbf{u}_\alpha,\ldots,\mathbf{u}_l\}$ le coordinate dell'i-mo punto-unità sull' α-mo asse fattoriale saranno:

$$c_\alpha(i) = \mathbf{x}_i'\mathbf{u}_\alpha$$

Autovettori (u)…

0,655	0,356	0,116	-0,134	-0,460	-0,448
0,178	-0,669	-0,409	-0,251	0,195	-0,503
0,043	0,101	-0,792	-0,099	-0,429	0,408
-0,674	0,391	-0,169	-0,270	-0,105	-0,529
0,269	0,443	-0,169	-0,464	0,684	0,140
-0,101	-0,256	0,368	-0,788	-0,296	0,284

0,097	0,150	0,341	0,919	1,491	3,003

… associati agli autovalori (λ)

Coordinate degli individui sul primo piano fattoriale

Matrice dei dati standardizzati (*10*x*6*) x Autovettore (*6*x*1*)

0,285	-0,137	0,210	-0,452	0,047	1,072
-1,456	-0,639	0,830	-0,452	0,047	-0,257
1,659	0,192	-0,173	1,665	-0,878	1,257
-0,415	0,013	-0,001	-0,644	-0,878	0,244
0,186	-0,360	1,035	-0,577	-0,475	-0,404
0,232	-0,496	-1,132	-0,315	-0,910	-0,909
-0,821	-0,828	1,634	-0,977	0,982	1,746
-1,518	-0,348	-1,169	-0,698	2,309	-0,290
1,145	2,877	-1,639	2,100	0,612	-1,202
0,704	-0,273	0,404	0,349	-0,856	-1,256

x

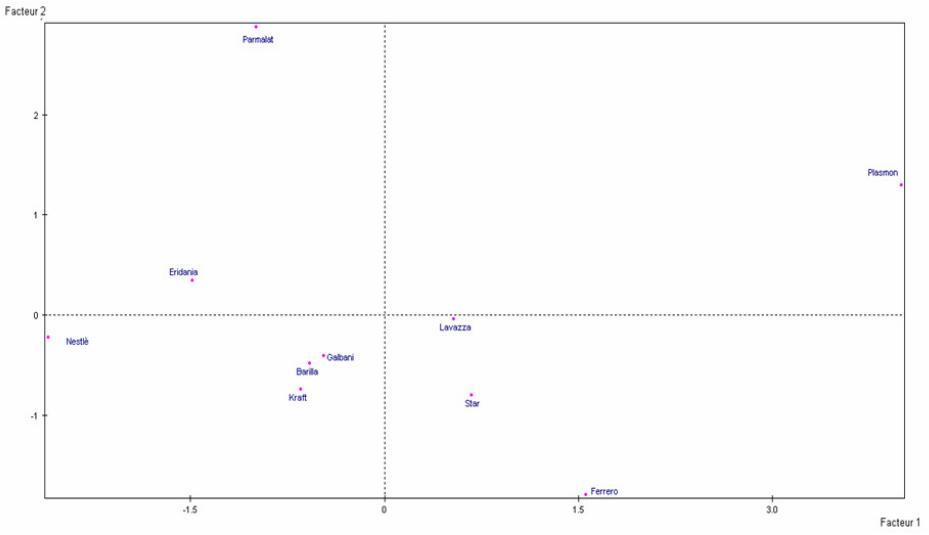
-0,448
-0,503
0,408
-0,529
0,140
0,284

=

-0,576
-1,485
1,557
-0,467
-0,644
0,535
-2,601
-0,988
3,995
0,674

Coordinate (*10*x*1*)

La rappresentazione grafica: il primo piano fattoriale



Analisi dei punti-variabili in $\Re^n$

Obiettivo:

$$\begin{cases} \max_{(v)}[\mathbf{v}'\mathbf{X}\mathbf{X}'\mathbf{v}] \\ \mathbf{v}'\mathbf{v}=1 \end{cases}$$

L'equazione agli autovalori è la seguente:

$$\mathbf{X}\mathbf{X}'\mathbf{v}_\alpha = \mu_\alpha \mathbf{v}_\alpha \quad \alpha = 1,\ldots,h$$

$\mu_\alpha = \alpha$-mo autovalore

$\mathbf{v}_\alpha = \alpha$-mo autovettore

Coordinate della *j*-ma variabile (*j=1,…,p*)

$$c_\alpha^*(j) = \mathbf{x}_j'\mathbf{v}_\alpha$$

Dualità tra R^n ed R^p

Si dimostra che $\lambda_\alpha = \mu_\alpha$ per ogni $\alpha=1,\ldots,h$.

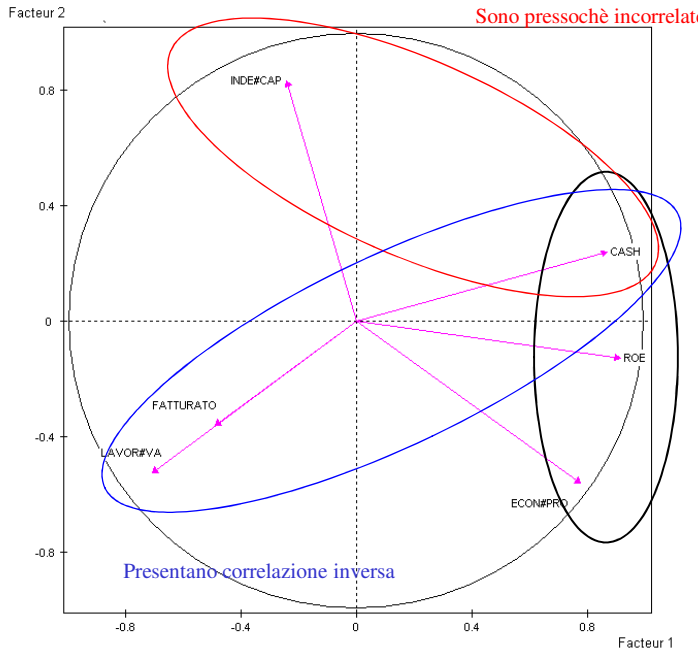
Formule di transizione:

$$\begin{cases} \mathbf{v}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\mathbf{u}_\alpha \\ \mathbf{u}_\alpha = \dfrac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}'\mathbf{v}_\alpha \end{cases}$$

Coordinate dei punti-variabile

COORDONNEES DES VARIABLES SUR LES AXES 1 A 5
VARIABLES ACTIVES

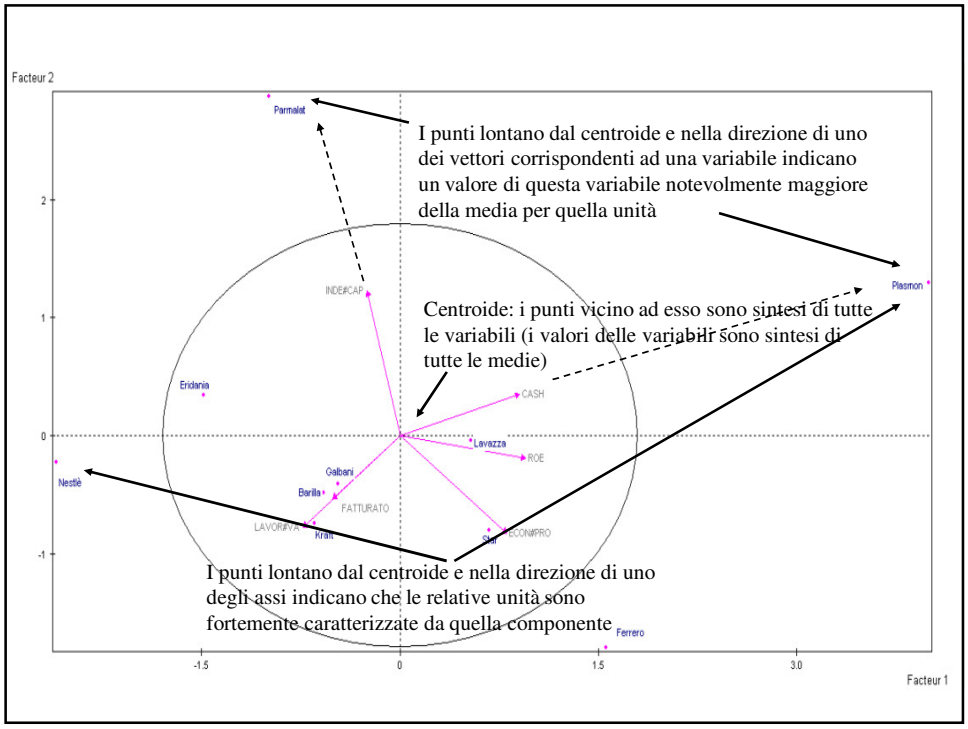
VARIABLES	COORDONNEES				
IDEN - LIBELLE COURT	1	2	3	4	5
ECON - ECON#PRO	0.78	-0.56	-0.13	0.07	-0.14
CASH - CASH	0.87	0.24	-0.24	-0.24	0.26
LAVO - LAVOR#VA	-0.71	-0.52	-0.10	-0.46	-0.04
ROE - ROE	0.92	-0.13	-0.26	-0.10	-0.15
INDE - INDE#CAP	-0.24	0.84	-0.44	-0.10	-0.17
FATT - FATTURATO	-0.49	-0.36	-0.76	0.21	0.10



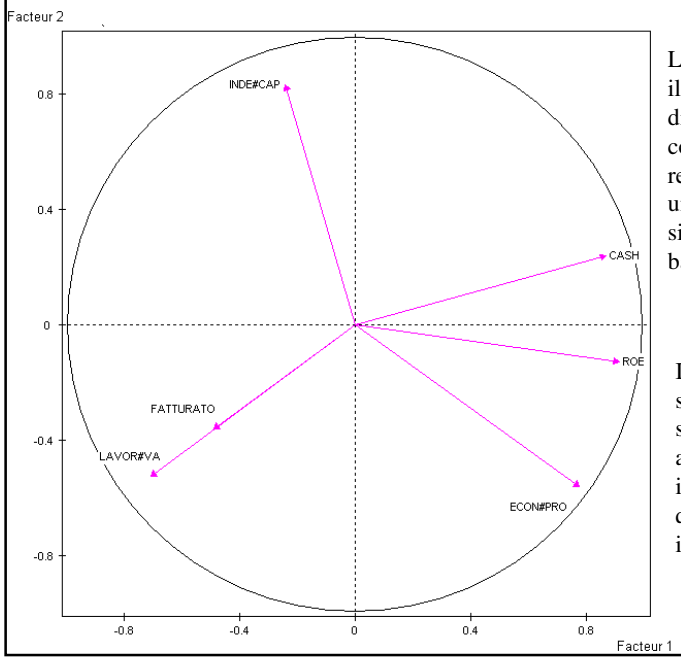
Sono pressochè incorrelate

Forte correlazione con il primo fattore

Presentano correlazione inversa



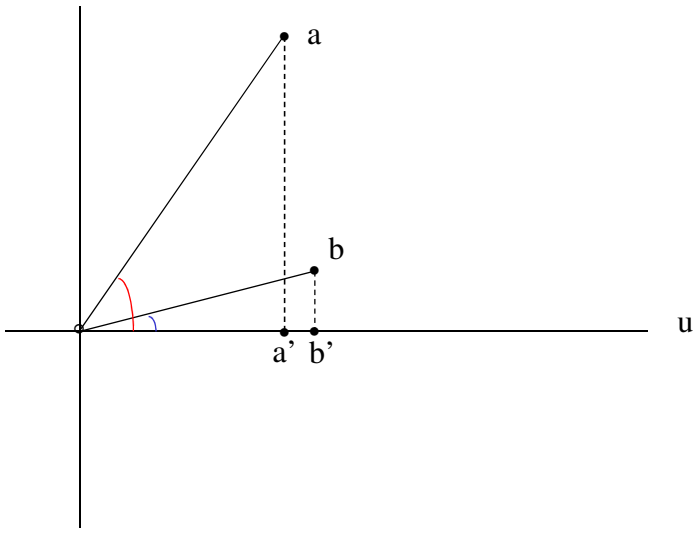
Interpretazione della mappa fattoriale



Le variabili correlate con il primo asse suggeriscono di interpretare lo stesso come una sintesi di redditività: a destra vi è una redditività alta, a sinistra una redditività bassa

Il secondo asse discrimina sull'indebitamento: in alto si posizioneranno le aziende ad alto tasso di indebitamento, in basso quelle che sono meno indebitate

Diagnostica nella ACP



Diagnostica nella ACP

COORDONNEES, CONTRIBUTIONS ET COSINUS CA
AXES 1 A 5

INDIVIDUS			CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5
Barilla	10.00	1.50	1.1	1.6	6.4	4.3	3.1	0.22	0.16	0.39	0.10	0.03
Eridania	10.00	3.49	7.3	0.8	3.6	10.3	0.6	0.63	0.03	0.10	0.10	0.00
Ferrero	10.00	7.94	8.1	21.7	18.0	9.9	10.0	0.31	0.41	0.21	0.04	0.02
Galbani	10.00	1.42	0.7	1.1	2.1	2.5	49.4	0.15	0.12	0.14	0.06	0.52
Kraft	10.00	1.96	1.4	3.7	4.7	11.3	0.4	0.21	0.28	0.22	0.20	0.00
Lavazza	10.00	3.34	1.0	0.0	22.2	29.3	0.0	0.09	0.00	0.61	0.30	0.00
Nestlè	10.00	9.00	22.5	0.3	21.7	4.9	0.1	0.75	0.01	0.22	0.02	0.00
Parmalat	10.00	9.69	3.2	55.2	0.7	7.7	10.6	0.10	0.85	0.01	0.03	0.02
Plasmon	10.00	18.51	53.1	11.2	4.1	12.2	5.3	0.86	0.09	0.02	0.02	0.00
Star	10.00	3.16	1.5	4.3	16.4	7.4	20.4	0.14	0.20	0.48	0.08	0.10