Analisi dei dati: Analisi multivariata

Domenico De Stefano

a.a. 2022/2023

Indice

Premessa

Cosa si intende con l'aggettivo multivariata?

- La statistica più semplice ha a che fare con una singola variabile (in inferenza si parla di variabile aleatoria):
  - possiamo raccogliere molti dati (in termini tecnici effettuare delle realizzazioni della variabile aleatoria)
  - (1) possiamo sintetizzare questi dati calcolando media, mediana, varianza, ecc.
  - (2) possiamo anche confrontare due diversi sottoinsiemi dei dati (ad es. confrontare la media dei voti per maschi e femmine) o "incrociare" due variabili (diagramma di dispersione o tabella di contingenza)
  - queste sono tutte analisi statistiche di tipo descrittivo (analisi esplorativa dei dati) univariate (1) o bivariate

Premessa (2)

- Sappiamo però che quando si analizzano fenomeni complessi (come quelli sociali) le caratteristiche raccolte per ogni unità statistica sono molteplici e di solito più di 2
  - Per ogni variabile, ovviamente, è possibile fare analisi statistiche separate.
  - Tuttavia la presenza di tante informazioni suggerisce che ci possano essere altri modi di analisi per mettere in luce la dipendenza reciproca tra le variabili.
  - Spesso inoltre si vogliono usare i dati per classificare le unità statistiche in categorie (ad es. la classificazione che i medici fanno in base al rapporto peso/altezza tra sottopeso, normale, sovrappeso, obeso).
  - In tali casi c'è bisogno di un nuovo approccio di analisi statistica quando le variabili sono $> 2 \Rightarrow$ Statistica Multivariata

Statistica multivariata: confermativa vs esplorativa

La statistica multivariata si usa quando **il numero delle variabili rilevate sullo stesso soggetto aumentano ed il problema è gestirle tutte e capirne le relazioni**.
Si divide in esplorativa (o descrittiva) e confermativa (Modelli Statistici)

- Regressione multipla
  - modello statistico (devono valere le assunzioni di base del modello)
  - analisi asimmetrica
  - obiettivo: formulare opportuni modelli interpretativi della realtà (analisi confermativa)
  - si fa inferenza del modello stimato sul campione alla popolazione

Statistica multivariata: confermativa vs esplorativa (2)

- Analisi fattoriale
  - tecnica esplorativa dei dati (non c'è nessun modello)
  - analisi (prevalentemente) simmetrica
  - obiettivo: rappresentare un numero elevato di variabili attraverso un numero inferiore di variabili non osservate (o latenti), i cosiddetti fattori
  - (per come la vedremo noi) non si fa inferenza, i risultati valgono solo per il campione!

Statistica multivariata: confermativa vs esplorativa (3)

Una peculiarità della statistica multivariata rispetto alla statistica univariata è la ricchezza di tecniche per la sintesi dei dati.

- Nella statistica univariata, una variabile si sintetizza con (ad esempio) i concetti di media e varianza (solo variabili quantitative) o moda (tutte i tipi di variabili), ma non vi è molto altro....
- Nel caso multivariato l'esigenza di sintesi diventa molto più forte: in genere ci si trova di fronte a notevoli quantità di dati che presentano un numero spesso elevato di variabili (tipico dei dati di questionari ma anche dei dati estratti dal web)
- le semplici tecniche grafiche usate per 'visualizzare' la distribuzione dei dati univariati vengono meno: se le variabili sono solo 2 o 3 si può ricorrere dei grafici (ad es. diagramma di dispersionet)

Statistica multivariata e correlazione

Ricordiamo che introdurre in un modello tante variabili in un'analisi non ha molto senso, né al livello statistico né al livello di interpretazione di un fenomeno.

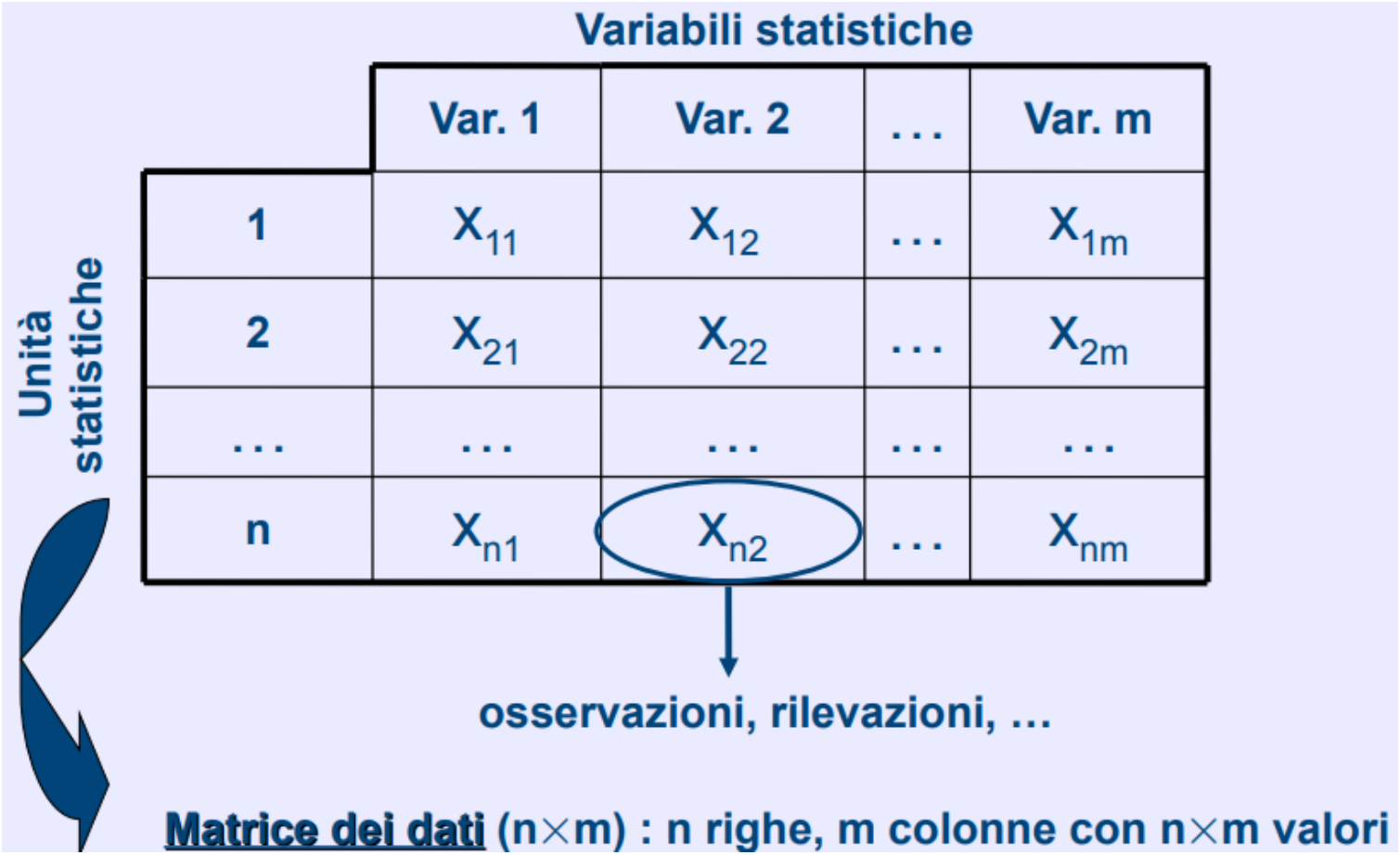
- Il modello diventa troppo complesso.
- Diventa difficile interpretare i risultati.
- Le stime dei parametri diventano molto instabili.
- Più parametri inseriamo (i β associati a ciascuna variabile indipendente), più osservazioni ci vogliono per stimarli.

Statistica multivariata e correlazione (2)

Ricordiamo che in un modello vanno eliminate le variabili che sono molto correlate tra di loro.

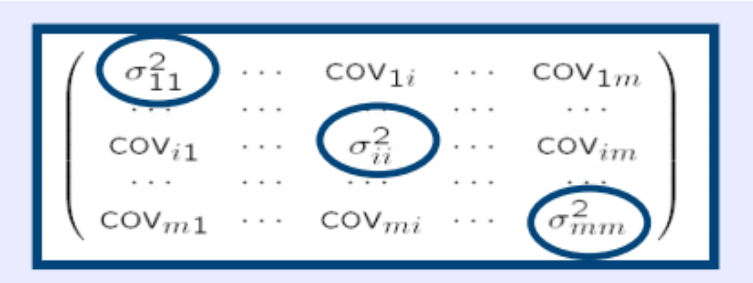
- concetto di multicollinearità
- Se due variabili sono molto correlate, allora l'informazione di una è contenuta quasi completamente nell'altra (alcune tecniche di sintesi sfruttano proprio questo!)

Matrice dati come... matrice



Dalla matrice dati alla matrice di varianza covarianza

Se calcoliamo la varianza e la covarianza tra tutte le variabili quantitative contenute in una matrice dati possiamo costruire la matrice di varianza-covarianza (o analogamente la matrice di correlazione)



dove:

- La diagonale contiene la varianza dell'i-ma variabile quantitativa
- le altre celle la covarianza tra la j-ma e la i-ma variabile
- IMPORTANTE: la matrice var-covar è quadrata (numero di righe = numero di colonne) e simmetrica (il valore nella riga i e la colonna j è lo stesso di quello nella riga j e la colonna i)

Dalla matrice dati alla matrice di varianza covarianza (2)

Se le variabili sono qualitative o miste (qualitative/quantitative) si parla di associazione (anziché di correlazione) e si calcolano altri indici per valutare la relazione tra queste (ad esempio l'indice del χ^2)

Metodi di analisi multivariata esplorativa

Due grandi famiglie di metodi:

- Metodi di riduzione della dimensionalità dei dati (o complessità) $\Rightarrow$ Analisi fattoriale (Analisi in componenti principali e analisi delle corrispondenze)
- Metodi di classificazione $\Rightarrow$ Clustering

Indice

Metodi di riduzione della dimensionalità: la maledizione della dimensionalità

Intuitivamente (per adesso):

- Quante più caratteristiche conosciamo di un individuo o di un oggetto tanto più ci appare diverso dal collettivo oggetto di studio
- inoltre troppe informazioni sono impossibili da visualizzare e anche da immaginare

In statistica le dimensioni sono le variabili, i punti proiettati in esse sono le unità statistiche...

L'analisi in componenti principali (ACP): intro

L'analisi delle componenti principali (ACP) è una delle più vecchie tecniche multivariate di analisi esplorativa dei dati e riduzione della dimensionalità

- L'ACP si applica al caso di dati quantiativi oppure dati ordinali trasformati in numeri (ad esempio le scale Likert di misurazione degli atteggiamenti)
- L'obiettivo che si pone è di costruire nuove variabili, ottenute come sintesi delle variabili originarie in modo che un numero ridotto di queste nuove variabili sia in grado di spiegare una porzione rilevante della varianza totale dei dati (che ricordiamo è l'informazione nei nostri dati)
- È una particolare tecnica fattoriale perchè tali nuove variabili si dicono fattori (da non confondere con i fattori in R!) o componenti principali
- Le componenti principali sono combinazioni lineari delle variabili di partenza non correlate tra di loro

L'analisi in componenti principali (ACP): intro (2)

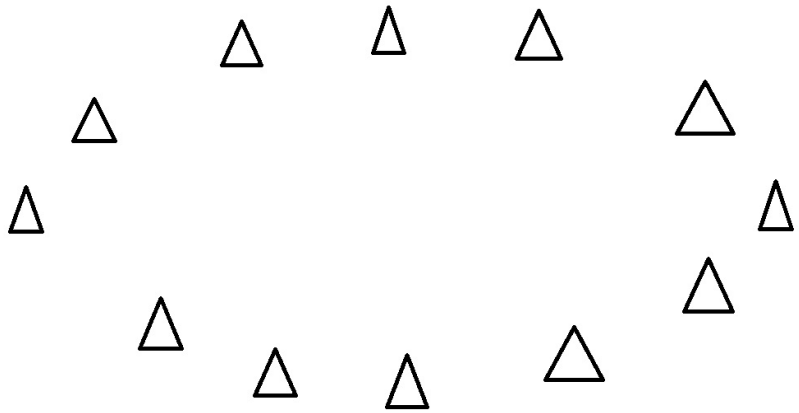
- In questo modo si tenta di risolvere a un tempo e il problema della dimensionalità del campione e quello della multicollinearità
- Le componenti principali possono essere il risultato finale dell'analisi o essere un risultato intermedio: ad es. nella regressione multipla, anziche' usare le variabili esplicative originali, si possono ottenere le componenti principali e usare queste come esplicative (con l'ovvio vantaggio che tra loro la correlazione è nulla)

L'analisi in componenti principali (ACP): intro (3)

- Il metodo ha una forte connotazione "geometrica" e ha la sua giustificazione teorica nella teoria delle matrici simmetriche
- La matrice di varianza-covarianza e' simmetrica!

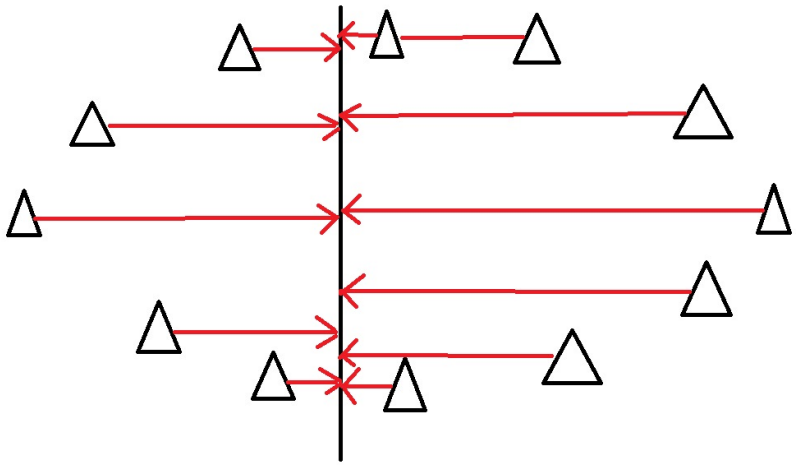
Ricerca delle componenti principali: spiegazione intuitiva

- Le componenti principali possono essere intanto immaginate come delle direzioni nello spazio dei dati orientate nel verso della maggior varianza che da tali dati emerge
- Ma guardiamo un esempio... nella figura sotto ci sono dei triangoli disposti secondo una ellisse (immaginiamo essere tali triangoli le nostre unità statistiche rappresentate in un piano cioè mediante due variabili quantitative)



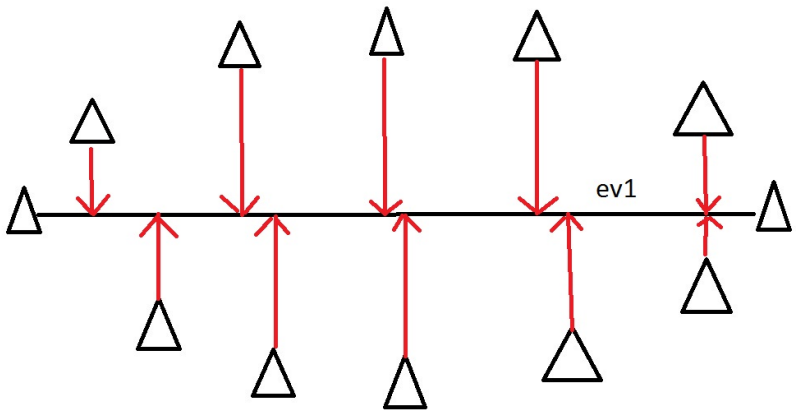
Ricerca delle componenti principali: spiegazione intuitiva (2)

- Cerchiamo quella direzione (una retta) che va nel verso della maggior varianza (cioè dove vi è la loro maggiore estensione) e poi proiettiamo i nostri dati su di essa
- Una retta verticale che passa tra i punti e sulla quale questi vengono proiettati apparirebbe così:



Ricerca delle componenti principali: spiegazione intuitiva (3)

- Tuttavia la retta non è orientata nel verso della maggiore estensione nei dati ossia dove vi è la maggior varianza, probabilmente non è la componente principale
- Una retta orizzontale che passa tra i punti e sulla quale questi vengono proiettati apparirebbe invece così:



Ricerca delle componenti principali: spiegazione intuitiva (4)

- Questa retta è ora invece orientata nel verso in cui vi è la maggiore estensione nei dati, cioè laddove vi è maggiore varianza
- Non vi è una migliore retta di questa orizzontale che approssimi la nuovola dei punti nel verso della maggiore varianza, questa è la prima componente principale
- ovviamente vi sono dei metodi matematici (di algebra lineare che ci consentono di individuare la migliore retta che approssima i punti nel verso della loro maggiore variabilità)
- qui entrano in gioco autovettori ed autovalori (in inglese: eigenvector ed eigenvalues, rispettivamente)

Ricerca delle componenti principali: spiegazione intuitiva (5)

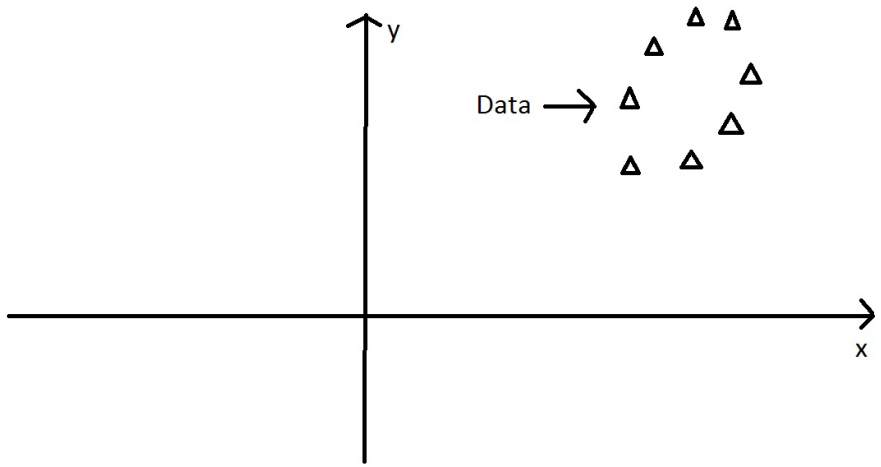
- Qunado abbiamo a che fare con una matrice dati possiamo "**scomporla**" in un insieme di autovettori ed autovalori
- autovettori ed autovalori esistono in coppia: ogni autovettore ha un corrispondente autovalore
- Pensate all'autovettore come ad una "direzione": nell'esempio di sopra l'autovettore è la direzione della retta individuata (verticale, orizzontale, a 45 gradi, ecc.)
- Un autovalore è un numero che ci dice invece quanta varianza nei dati è presente in quella direzione nell'esempio l'autovalore èp un numero che ci dice qual tra le due direzioni 'riassume' la maggiore varianza nei dati
- L'autovettore associato al più alto autovalore è la componente principale

Ricerca delle componenti principali: spiegazione intuitiva (6)

- Potreste pensare che le direzioni sono infinite, in realtà non ci sono infiniti autovettori ed autovalori mediante i quali scomporre la nostra matrice dati
- il numero di autovettori e autovalori è uguale al numero di dimensioni nei dati (numero di variabili)
- immaginiamo che ad esempio abbiamo rilevato due variabili: l'età e il numero di ore passato su internet
- Ci sono 2 variabili, le dimensioni sono 2, ci sono 2 autovettori/valori
- La ragione di ciò è che in realtà gli autovettori proiettano i dati in un nuovo riferimento spaziale che ha lo stesso numero di dimensioni di quello originario

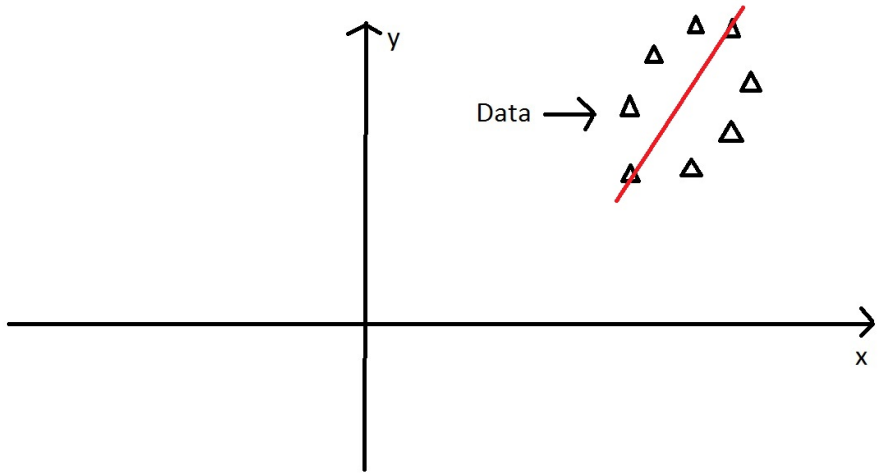
Ricerca delle componenti principali: spiegazione intuitiva (7)

- ritorniamo al nostro esempio: immaginiamo che sull'asse x vi sa la variabile età e sull'asse y quella delle ore passate su internet.



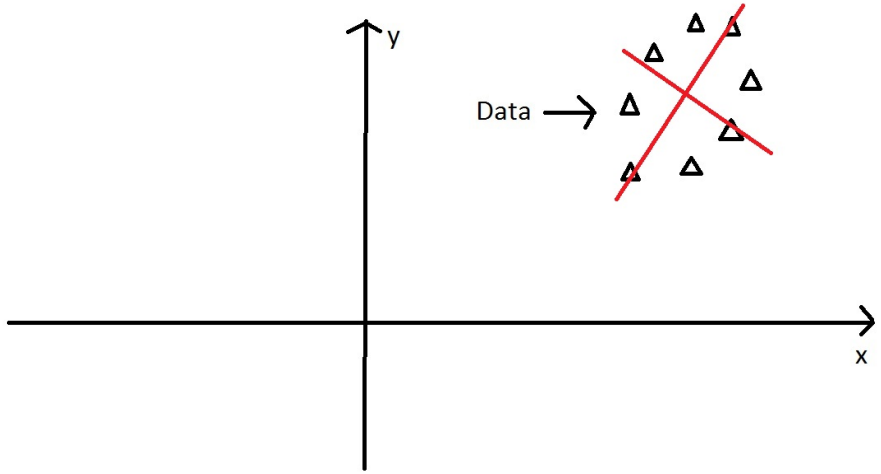
Ricerca delle componenti principali: spiegazione intuitiva (8)

- Il primo autovettore passa lungo la direzione della maggior varianza: questa è la prima componente principale



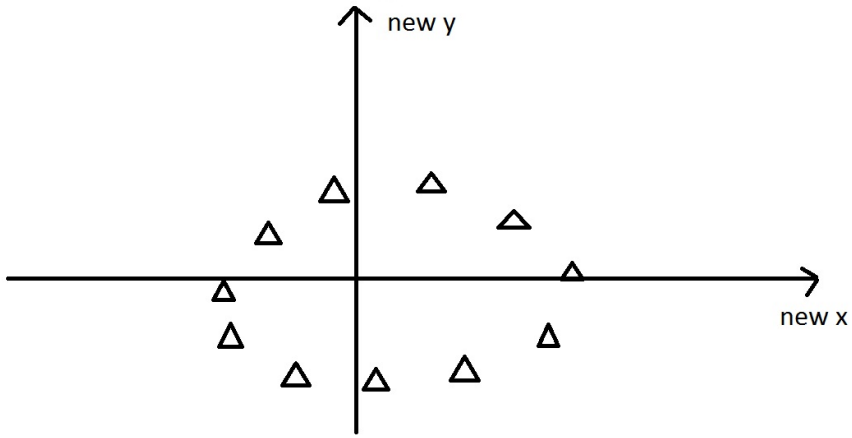
Ricerca delle componenti principali: spiegazione intuitiva (9)

- Ma abbiamo detto che i possibili autovettori sono 2
- l'altra componente principale con il metodo dell'ACP è rappresentata dall'autovettore perpendicolare al primo (forma un angolo di 90 gradi)
- quindi la seconda c.p. sarà:



Ricerca delle componenti principali: spiegazione intuitiva (10)

- Gli autovettori sono assi più utili (vedremo in seguito) di quelli orginari (che sono le variabili) perché legati e derivanti dalla variabilità intrinseca nei dati
- È possibile poi rappresentare i nostri dati nel nuovo riferimento spaziale come dalla seguente immagine:



Ricerca delle c.p.: spiegazione un po' più tecnica

- Costruiamo la matrice di varianza-covarianza, ovvero una matrice $m \times m$ (dove m è il numero di variabili nella matrice dati), il cui elemento σ_{jk} è la covarianza della variabile X_j con la variabile X_k.
- Per semplicità centriamo le variabili (cioè usiamo gli scarti dalla media anzichè i valori originari). Indichiamo le variabili centrate come $\widetilde{X}_j$ e $\widetilde{X}_k$ (che avranno media=0)
- Pertanto la covarianza tra la variabile j-ma e k-ma sarà: $\sigma_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{n}$

Ricerca delle c.p.: spiegazione un po' più tecnica (2)

- La matrice di covarianza ci permette di esprimere la varianza di una nuova variabile Y come combinazione lineare delle variabili originarie X, ossia:

$$Y_h = \sum_{i=1}^{m} \alpha_i X_i = \alpha_1 X_1 + \alpha_2 X_2 + \ldots + \alpha_m X_m$$

- La varianza di questa nuova variabile è funzione delle varianze e covarianze delle variabili orginarie X pesate per i coefficienti α della combinazione lineare:

$$VAR(Y_h) = \sum_{j=1}^{m} \sum_{k=1}^{m} \alpha_j \alpha_k \sigma_{jk}$$

Ricerca delle c.p.: spiegazione un po' più tecnica (3)

- Ora il punto è: tra tutte le possibili infinite combinazioni delle variabili originarie (cioè per tutti i possibili valori degli α) cerchiamo quella per cui la varianza della nuova variabile Y sia la massima possibile (perché?)
- Quello posto è un problema di massimo vincolato dove mediante il metodo dei moltiplicatori di Lagrange si ottiene il massimo con l'autovalore più grande $\lambda_{(1)}$ (**la quantità di varianza "riassunta"**) a cui corrisponde l'autovettore che 'indica' **la direzione nella quale si individua la massima varianza**

Ricerca delle c.p.: spiegazione un po' più tecnica (4)

- Poi si ricerca (sempre un massimo vincolato in cui si garantisce la non correlazione con la prima c.p.) l'autovalore immediatamente successivo in ordine di grandezza al primo. Questo secondo autovalore avrà associato un secondo autovettore α_2 che sarà la seconda componente principale
- Procedendo così si trovano m componenti principali (tante quante sono le variabili originarie) corrispondenti agli m autovalori ordinati in senso decrescente (da quello più importante, perché riassume più varianza, a quello meno importante)
- Y_1 e l'insieme degli α rappresentano la c.d. prima componenente principale!

$$\lambda_{(1)} > \lambda_{(2)} > \ldots > \lambda_{(m)}$$

Scelta delle componenti principali

La derivazione proposta sopra individua, a partire da m variabili, m componenti principali, quindi non si è ottenuta alcuna riduzione nel numero di variabili!

- In realtà qualcosa si è ottenuto, comunque, poiché le nuove variabili sono ortogonali.
- La riduzione del numero di variabili si ottiene invece considerando solo alcune delle c.p.
- per la precisione le prime q con $q < m$
- **Ricordiamo che le c.p. hanno importanza (varianza) decrescente**
- Sapendo che la somma degli autovalori è uguale alla varianza totale nelle variabili originarie allora va da se che il rapporto tra autovalore associato ad una data c.p. diviso la somma degli autovalori è uguale alla % di varianza riassunta da quella componente

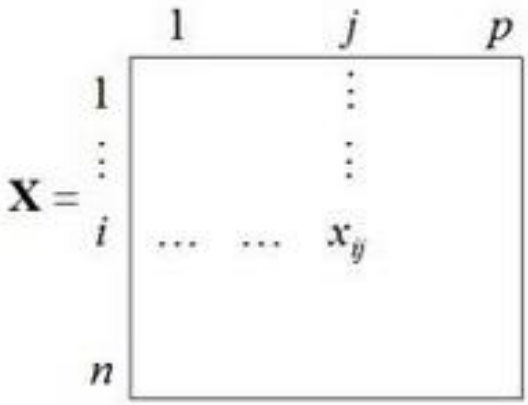
Scelta delle componenti principali (2)

- ad esempio: $P_{(1)} = \frac{\lambda_{(1)}}{\sum_{i=1}^{m} \lambda_{(i)}}$ è la percentuale di varianza totale riassunta dalla prima componente
- sommando le frazioni di varianza cosi ottenute sappiamo ad esempio quanta varianza nelle nostre variabili viene riassunta dalle prima q componenti principali
- Le percentuali di varianza riassunte dalle componenti principali possono essere usate come **linea guida** per la scelta del numero q di componenti. Di seguito alcune regole:
  - in genere si scelgono q componenti mediante le quali riusciamo a riassumere almeno il 70% della varianza totale (la soglia può essere diminuita qualora il numero m di variabili sia molto grande);
  - ignorare in genere le ultime c.p. o che comunque contribuiscono per un ammontare di varianza inferiore a un livello prefissato;
  - Se le condizioni sopra dette sono verificate scegliere al massimo 3 c.p. per ottenere una rappresentazione grafica dei dati

Scelta delle componenti principali

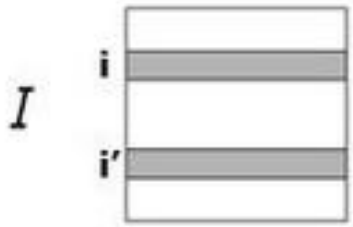
- Le combinazioni tra le variabili cosi ottenute sono ora pensate come nuove variabili, che tuttavia non sono, in genere, calcolabili direttamente: per questa ragione sono spesso dette variabili latenti.
- Oltre alla proprietà geometrica della perpendicolarità (ortogonalità), le componenti principali godono di una proprietà ancora più significativa. Infatti sono statisticamente indipendenti, ovvero la covarianza di due qualsiasi di queste variabili è nulla.
- In altri termini se ricalcoliamo la matrice di covarianza per queste nuove variabili otteniamo una matrice diagonale (ovvero una matrice in cui tutti gli elementi fuori la diagonal sono nulli).
- Considerando l'insieme delle variabili latenti al posto delle variabili originali abbiamo quindi una descrizione completa della totalità dei nostri dati in termini di variabili statisticamente indipendenti, ognuna delle quali sintetizza una informazione sul campione del tutto assente nelle altre variabili.

Interpretazione risultati

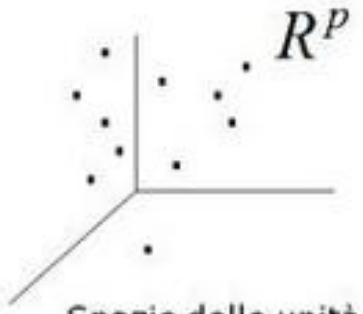


Matrice dei dati

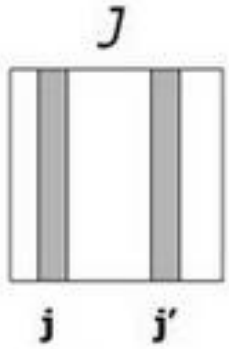
- I vettori riga di **X** sono **punti-unità** nello spazio $\Re^p$ generato dalle variabili.
- I vettori colonna di **X** sono **punti-variabile** nello spazio $\Re^n$ generato dalle unità.



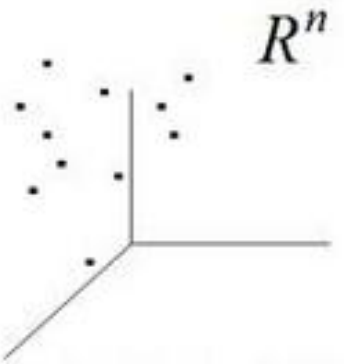
Profili riga



Spazio delle unità



Profili colonna



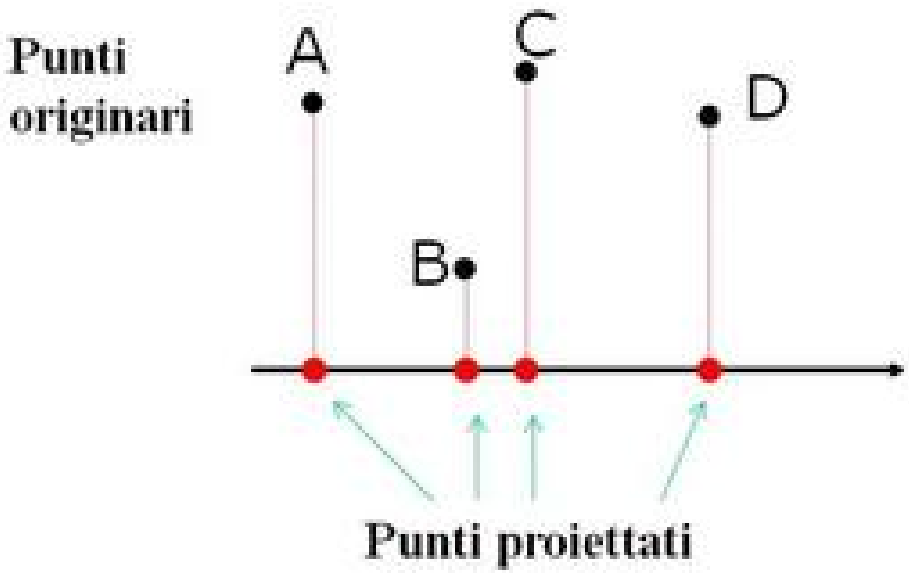
Spazio delle variabili

Lettura geometrica della matrice dei dati.

Interpretazione risultati (2)

La matrice dei dati X può essere vista come una nube dei punti in uno spazio multidimensionale.
Obiettivo dell'ACP di individuare una o più variabili latenti si concretizza, in un'ottica geometrica, nell'individuare uno spazio di dimensione ridotta su cui proiettare la nube dei punti originari e studiare le distanze tra i vari punti (proiettati).
Tali proiezioni costituiscono un'approssimazione delle relazioni esistenti tra i vari punti in quanto le distanze originarie risultano deformate.

Interpretazione risultati (3)



Proiezione dei punti su un asse.

Interpretazione risultati (4)

Punto di partenza è la matrice di varianza-covarianza (oppure se standardizziamo le variabili la matrice di correlazione)

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

Un esempio. La performance di un campione di imprese

- Ipotesi della ricerca: Gli indicatori di bilancio, pur essendo molteplici, rappresentano l'espressione di due fattori latenti:
  - la performance economica dell'azienda;
  - l'equilibrio finanziario dell'azienda;

Obiettivo dell'analisi:

- individuare la migliore sintesi degli indici di bilancio che consenta di ordinare le aziende sulla base dei due fattori ipotizzati

Essendo le variabili rilevate sulle unità statistiche (le aziende nel campione) tutte di natura numerica, si utilizza l'Analisi delle Componenti Principali.

Un esempio. La performance di un campione di imprese (2)

La variabili

- ECON.PRO: Economic Profit (differenza tra capitale e costi)
- CASH: Cash flow (liquidità dell'impresa misurato come % sul fatturato)
- LAVOR.VA: costo del lavoro sul valore aggiunto in %
- ROE: Return on equity (rendimento del capitale)
- INDE.CAP: indebitamento sul capitale
- FATTURATO

Un esempio. La performance di un campione di imprese (3)

la matrice dati (i valori sono su scale tanto diverse) $\Rightarrow$ conviene standardizzare la variabili

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

Le 5000 società leader, supplemento a Milano Finanza, 1998.

Un esempio. La performance di un campione di imprese (4)

la matrice 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 dei dati standardizzati.

Un esempio. La performance di un campione di imprese (5)

L'osservazione della matrice di correlazione (la matrice di varianza-covarianza per dati standardizzati è direttamente la 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 sono le variabili osservate!

Un esempio. La performance di un campione di imprese (6)

MATRICE DI CORRELAZIONE

	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

Un esempio. La performance di un campione di imprese (7)

La scelta delle componenti principali con le quali ridurre la dimensionalità dei dati (ricordate che l'ipotesi inziale è che vi siano due dimensioni nelle quali si può 'scomporre' la performance delle imprese)

Un esempio. La performance di un campione di imprese (8)

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

Si selezionano le prime due CP:

- spiegano il 74% della variabilità totale
- hanno autovalori superiori a 1

Un esempio. La performance di un campione di imprese (9)

Nota: nel caso di ACP sulla matrice di correlazione la componenti piu importanti sono quelle associate agli autovalori $\lambda > 1$

Un esempio. La performance di un campione di imprese (10)

- La ricerca dello spazio di dimensioni ridotte che sintetizzi nella maniera più efficiente la struttura informativa contenuta nella matrice dei dati originari può essere effettuata sia rispetto agli individui sia rispetto alle variabili.
- Si parla così di analisi:
  - dei punti-unità nello spazio delle variabili.
  - dei punti-variabile nello spazio degli individui
- Si può dimostrare che gli autovalori ottenuti nelle due analisi coincidono.
- Ciò implica che le CP individuate sono le stesse anche se differiscono nei due spazi per la diversa unità di misura delle colonne di X rispetto alle righe (standardizzate le prime, non le seconde).
- L'analisi nello spazio degli individui permette di 'interpretare' il significato delle variabili latenti selezionate.

Un esempio. La performance di un campione di imprese (11)

- L'analisi nello spazio delle variabili individua un ordinamento delle unità rispetto alle variabili latenti selezionate

Un esempio. La performance di un campione di imprese (12)

L'analisi dei punti unità

Autovettori (u)…

u_6	u_5	u_4	u_3	u_2	u_1
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
λ_6	λ_5	λ_4	λ_3	λ_2	λ_1

… associati agli autovalori (λ)

Un esempio. La performance di un campione di imprese (13)

Coordinate degli individui sulla prima componente

Matrice dei dati standardizzati (*10x6*)

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

Autovettore $\mathbf{u}_1$ (*6x1*)

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

X =

Azienda	
Barilla	-0,576
Eridania	-1,485
Ferrero	1,557
Galbani	-0,467
Kraft	-0,644
Lavazza	0,535
Nestlè	-2,601
Parmalat	-0,988
Plasmon	3,995
Star	0,674

Coordinate (*10x1*)

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

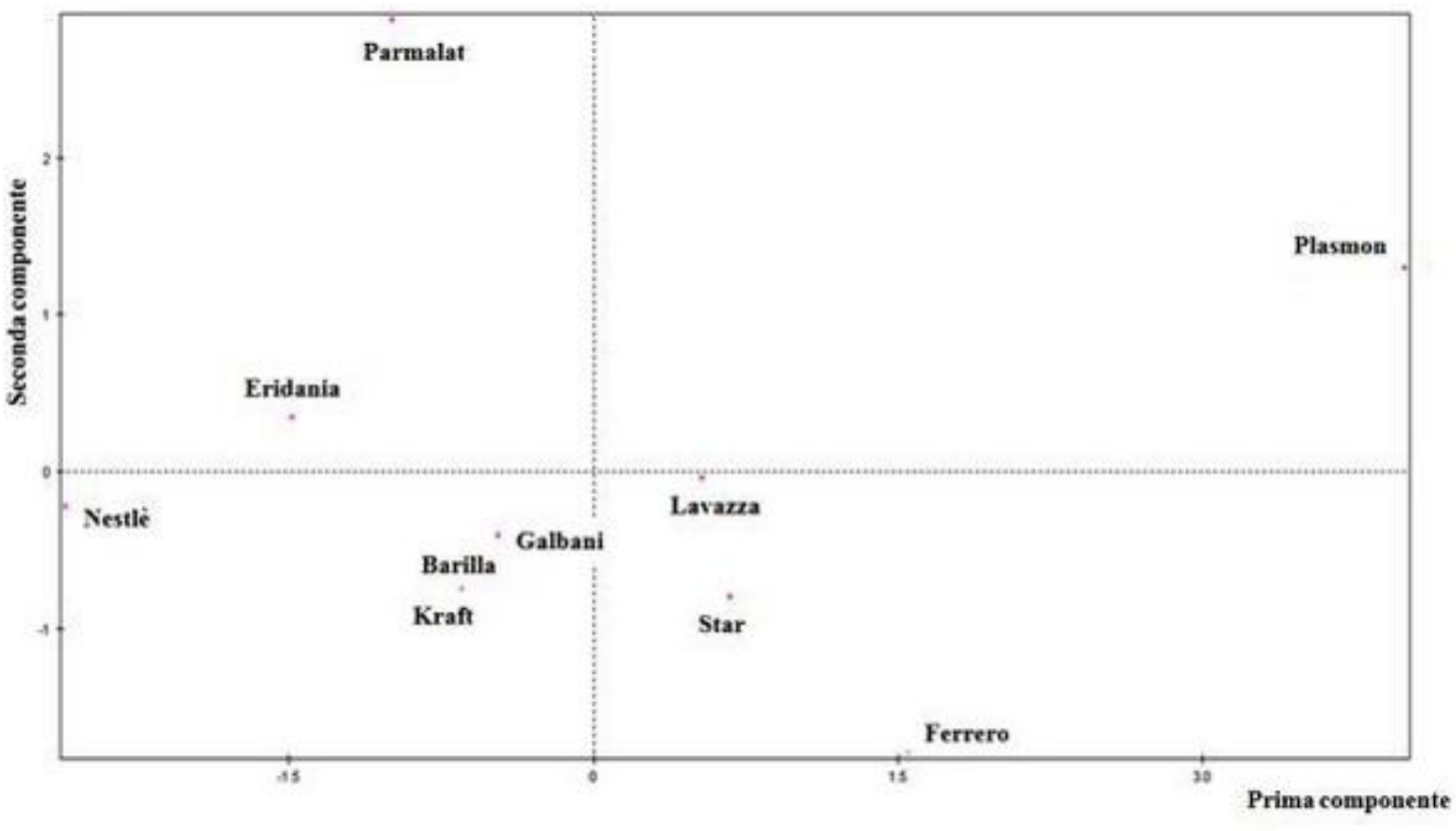
$$CP_j(i) = x_i' \cdot u_j$$

Nel nostro esempio le coordinate dei punti-unità (le aziende) sulla prima componente sono pari al prodotto di ogni riga della matrice dei dati per la colonna dell'autovettore $\mathbf{u}_1$

$$CP_1(i) = x_i' \cdot u_1$$

Un esempio. La performance di un campione di imprese (14)

Lo spazio ridotto delle unità (o individui)



La rappresentazione grafica: *il primo piano fattoriale delle unità* (formato dalla prima e dalla seconda componente).

Un esempio. La performance di un campione di imprese (15)

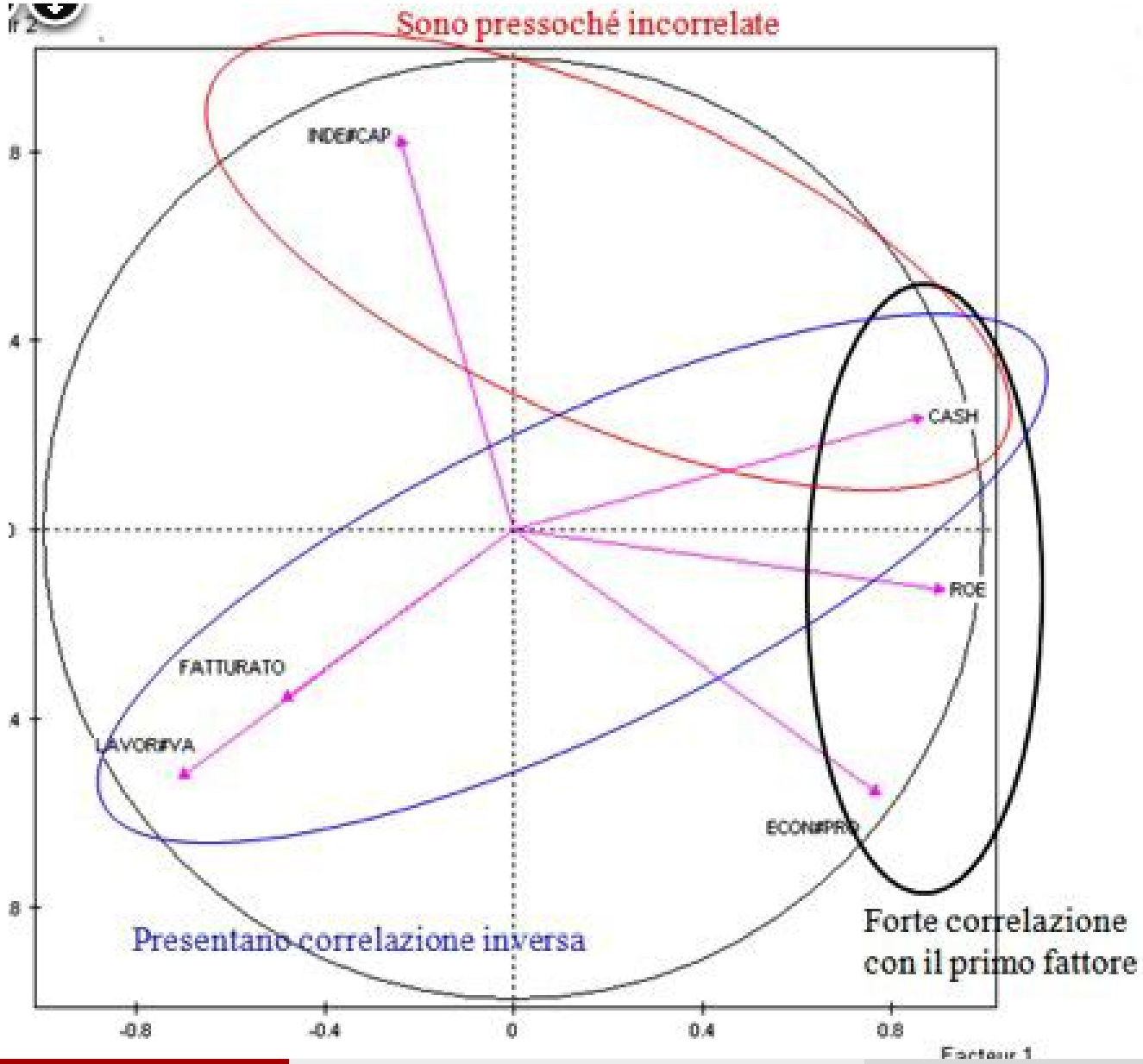
L'analisi per le variabili è analoga
Esistono le coordinate dei punti-variabile sulle prime due componenti.
In generale, la correlazione variabile-componente è data dal coseno dell'angolo tra i due vettori. Più l'angolo è stretto e maggiore sarà la correlazione. La correlazione è nulla per angoli di 90 gradi.
Quando l'analisi è effettuata sulla matrice di correlazione, le coordinate possono essere interpretate come coefficienti di correlazione delle variabili originarie rispetto alle componenti considerate.
Così, nel nostro caso studio, si può affermare che il ROE è fortemente correlato in maniera positiva con CP1 ed è incorrelato con CP2.
L'analisi di queste coordinate consente di interpretare le componenti latenti!

Un esempio. La performance di un campione di imprese (16)

	CP_1	CP_2
ECON - ECON#PRO	0.78	-0.56
CASH - CASH	0.87	0.24
LAVO - LAVOR#VA	-0.71	-0.52
ROE - ROE	0.92	-0.13
INDE - INDE#CAP	-0.24	0.84
FATT - FATTURATO	-0.49	-0.36

Le coordinate dei punti-variabile.

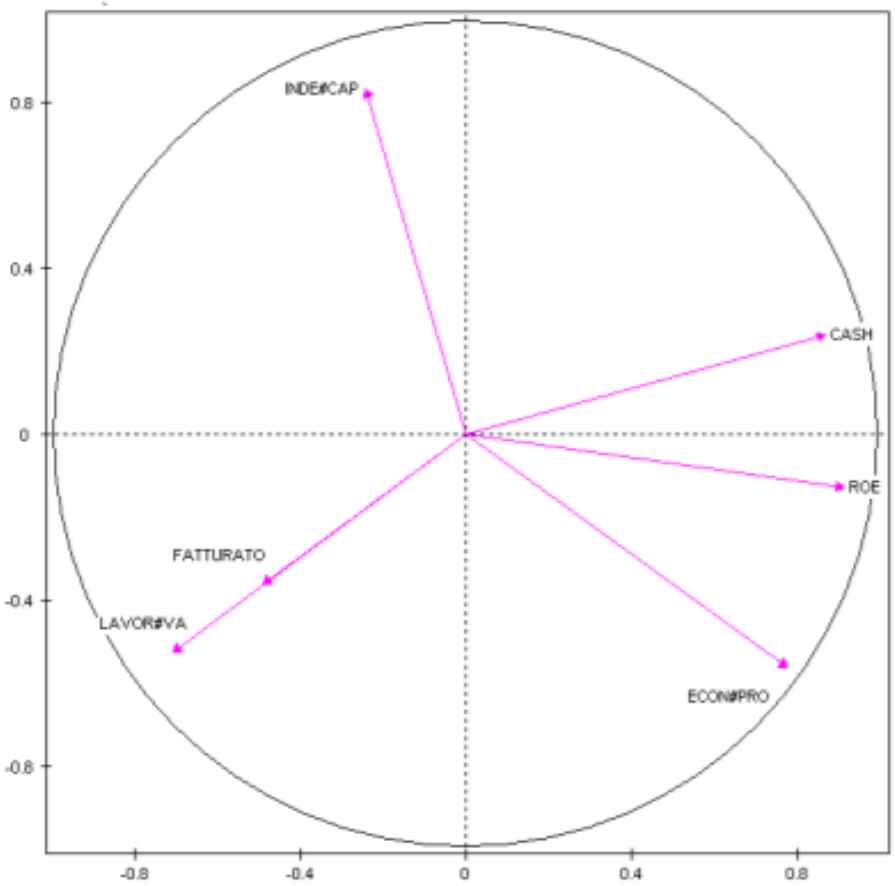
Un esempio. La performance di un campione di imprese (17)



Un esempio. La performance di un campione di imprese

- Le variabili correlate con la prima CP suggeriscono di interpretare lo stesso come una sintesi di redditività;
- a destra vi è una redditività alta;
- a sinistra una redditività bassa;
- La seconda CP discrimina sull'indebitamento:
- in alto si posizioneranno le aziende ad alto tasso di indebitamento
- in basso quelle che sono meno indebitate.

Un esempio. La performance di un campione di imprese



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

Un esempio. La performance di un campione di imprese