Università degli studi di Ferrara
Dipartimento di Matematica
A.A. 2019/20 – I semestre

STATISTICA MULTIVARIATA

SSD MAT/06

LEZION 10 - PCA: introduzione

Docente: Valentina MINI

valentina.mini@unife.it

RICEVIMENTO: lunedi pomeriggio su appuntamento previa mail

INTRODUZIONE

L'Analisi fattoriale e la PCA

L'analisi fattoriale consiste in un insieme di tecniche statistiche che permettono di ottenere una **riduzione della complessità del numero di fattori che spiegano un fenomeno.**

Si propone quindi di **determinare un certo numero di variabili "latenti"** (fattori non direttamente misurabili nella realtà) più ristretto e riassuntivo rispetto al numero di variabili di partenza.

ES:

Si pensi, ad esempio, all'insieme dei voti di una popolazione di studenti di una certa scuola.

I voti riguardano il rendimenti degli stessi nelle diverse materie (italiano, matematica, scienze, geografia, storia, ecc.). È lecito supporre che le abilità di apprendimento possano distinguersi in **due fattori**: ***abilità nelle materie scientifiche* e *abilità nelle materie umanistiche*.**

Con l'analisi fattoriale è possibile misurare queste due abilità attraverso la costruzione di due variabili latenti di sintesi (combinazione lineare) delle variabili originarie (i voti nelle diverse materie) ognuna pesata sulla base dell'importanza "*u*" (del contributo) nel discriminare gli individui sulla base delle loro abilità scientifiche e umanistiche.

introduzione

Spesso interesse scientifico in **fenomeni non direttamente misurabili** → **variabile latente**

Es. in psicologia: la patologia del *burnout* :

- non misurabile direttamente
- Tuttavia possiamo misurarne molti aspetti (motivazione, livello di stress…)
- Domanda centrale: tutte queste variabili ci guidano verso lo stesso fenomeno/variabile di interesse?

introduzione

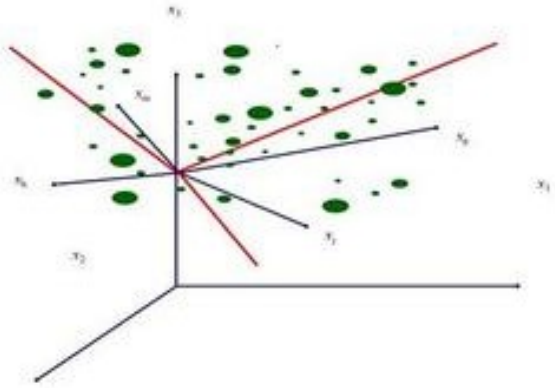
Metodi di Analisi fattoriale

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.



Rappresentazione grafico di un approccio fattoriale.

L'Analisi delle Componenti Principali

L'Analisi delle Componenti Principali (ACP) consente di ridurre la dimensionalità dell'insieme dei dati eliminando la ridondanza di informazioni risultato di p variabili altamente correlate e di sostituire a queste ultime un minor numero h (con $h < p$) di nuove variabili tra loro **non correlate** e **legate linearmente alle variabili di partenza**.

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

Analisi per componenti principali (ACP o PCA)

Analisi fattoriale che ci porta ad identificare gruppi di variabili

3 utilizzi principali:

- 1- capire struttura di un set di variabili
- 2- costruire un questionario per comprendere/cogliere una variabile sottesa (latente)
- 3- ridurre dataset a dimensione più gestibile (es. per procedere con RLM), pur mantenendo la maggior parte delle informazioni iniziali

introduzione

PCA:

- Tra i più storici e comuni metodi di analisi statistica multivariata
- Proposta da Pearson nel 1901 e successivamente (indipendentemente) da Hotelling nel 1933
- Anche nota come trasformazione statistica di **Hotelling** ed espansione di Karhunen-Loeve
- Un metodo effettivo per la rappresentazione di dati multivariati in uno spazio a dimensioni ridotte (*parsimonious summarization of data*), per **semplificare l'analisi statistica**
- Metodo utile sia per analisi **esplorative** che per modelli di **previsione**

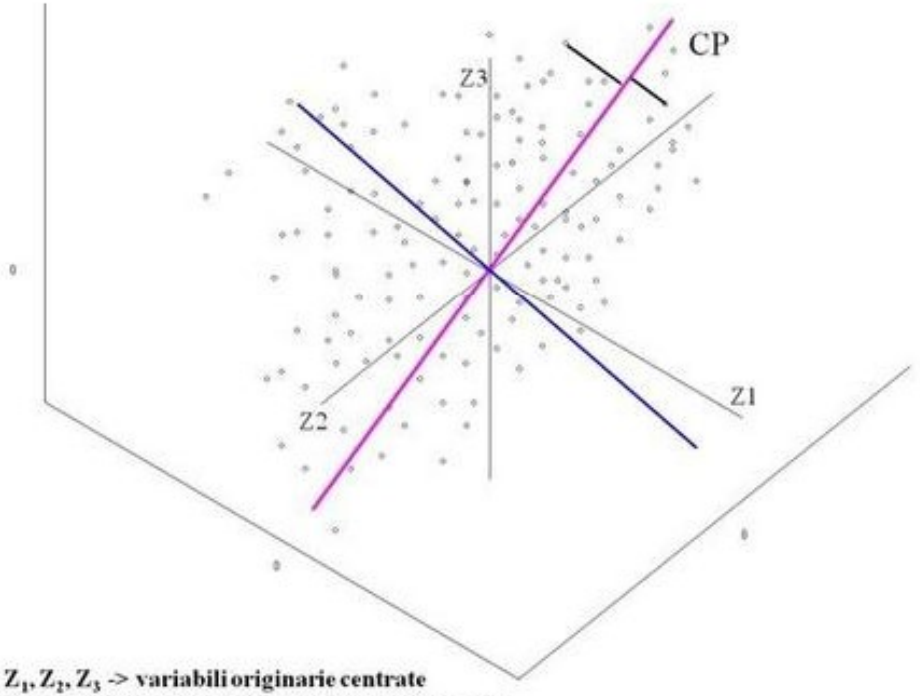
introduzione

Ruolo chiave nell'analisi PCA: I FATTORI O COMPONENTI

- Cosa sono i fattori o COMPONENTI?
- Come li identifichiamo?
- Cosa "raccontano" sulla possibile relazione tra variabili misurate?

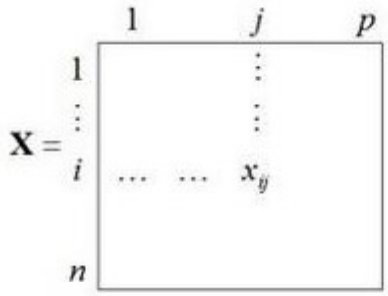
I FATTORI/COMPONENTI

Rappresentazione grafica dell'ACP



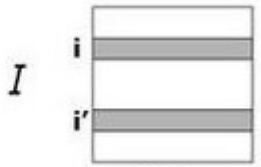
Z_1, Z_2, Z_3 -> variabili originarie centrate
CP -> Variabile latente (componente principale)

Approccio geometrico all'ACP

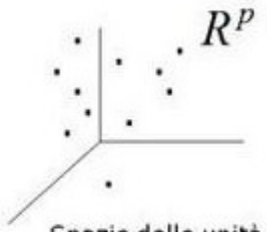


Matrice dei dati

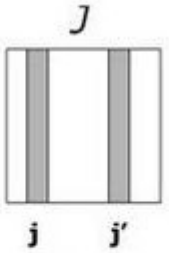
- I vettori riga di **X** sono **punti-unità** nello spazio $\Re^p$ generato dalle variabili.
- I vettori colonna di **X** sono **punti-variabile** nello spazio $\Re^n$ generato dalle unità.



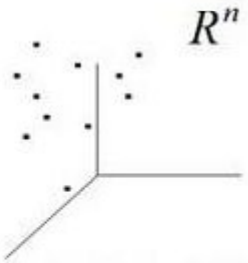
Profili riga



Spazio delle unità



Profili colonna



Spazio delle variabili

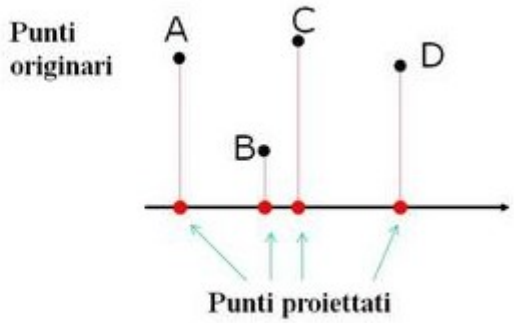
Lettura geometrica della matrice dei dati.

Sintesi dell'informazione

Seguendo un approccio geometrico, *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.



Proiezione dei punti su un asse.

Obiettivo e finalità operativa dell'ACP

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

Finalità operativa:
Ricercare di un sistema di assi fattoriali (le componenti principali) ortogonali che generi il sottospazio di "migliore" approssimazione tale da deformare il meno possibile le distanze tra i punti.

Punto di partenza per individuare le componenti principali: la matrice di covarianza

La matrice di varianze e covarianze

Se x è una matrice dei dati "unità per variabili" di dimensioni *n-k*. la matrice Σ ("Sigma") di varianze e covarianze è: (vedi figura). La variabilità del sistema *k*-variato viene sintetizzato con la *traccia* della matrice di var-cov.

Essa esprime al contempo la **variabilità** delle singole variabili (sulla diagonale) e la **co-variazione** tra le stesse, prese due a due (elementi non diagonali).

La variabilità del sistema *k*-variato viene sintetizzata con la *traccia* della matrice var-voc.

Punto di partenza per individuare le componenti principali: la matrice di covarianza

Esempio di matrice di varianze e covarianze

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

Una generica **Componente Principale** (CP) si definisce come una combinazione lineare delle *p* variabili originarie pesate per un vettore di pesi *u*.

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 (variabili centrate):

$$CP_1 = \tilde{X} u_1$$

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

$$VAR(\tilde{X} u_1) = u_1^{'} \sum u_1$$

Posta tale relazione, il vettore u_1 è ricercato in modo tale da massimizzare la $VAR(\tilde{X} u_1)$ secondo il vincolo $u_1^{'} u_1$.

La ricerca delle CP si concretizza in un problema massimo vincolato.

Si ricercano i pesi u che massimizzano la varianza delle componenti con i vincoli:

- che i vettori *u* siano unitari (il loro prodotto è pari a 1);
- che, per le componenti successive alla prima, i vettori siano a due a due ortogonali ($u_i u_j = 0$ per ogni i≠j)

Il problema di massimo si risolve attraverso l'utilizzo del moltiplicatore di Lagrange che porta alla seguente soluzione:

Per la prima componente $u_1^{'} \sum u_1 = u_1^{'} \lambda_1 u_1 = \lambda_1$

Pari anche a $\tilde{X}' \tilde{X} u_1 = \lambda_1 u_1$

Dove la matrice di varianze e covarianze $\sum = \tilde{X}' \tilde{X}$ è ottenuta come prodotto della matrice dei dati centrati per se stessa.

Dal problema di massimo agli autovettori e autovalori

Autovalori e autovettori

$\tilde{X}'\tilde{X}u_1 = \lambda_1 u_1$

Dalla soluzione del problema di massimo si evince come u_1 rappresenta il primo **autovettore** della matrice $\tilde{X}'\tilde{X}$ mentre λ_1 è invece il corrispondente **autovalore**.

L'autovalore *j-esimo* può anche essere interpretato come la varianza della j-esima componente principale:

$\lambda_j = u_j'\tilde{X}'\tilde{X}u_j = (CP_j)^2 = VAR(CP_j)$

Quindi siccome l'obiettivo è quello di identificare le variabili latenti che spiegano quanta più informazione (variabilità) della nube originaria, allora **La prima componente principale sarà quella con λ maggiore, e a seguire la seconda sarà quella con λ maggiore dopo la prima e cosi via....**

L’analisi sulla 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 differiscono unicamente in media.

Nella realtà il ricercatore si trova ad analizzare variabili con scale di misurazione differenti che quindi, prima dell’analisi, devono essere rese omogenee.

Nell’ACP, per superare tale difficoltà, si considerano le variabili espresse in termini di scostamenti standardizzati, quindi il punto di partenza dell’analisi diviene la ***matrice di correlazione***.

Infatti, essendo $\tilde{X}$ la matrice delle variabili standardizzate, il prodotto $\tilde{X}'\tilde{X} = R$ sarà pari alla matrice di correlazione R.

• Esempio ACP: i consumi alimentari

Matrice di correlazione

	Cere	Riso	Pata	Zucc	Verd	Vino	Carr	Latt	Burr	Uova
Cere	1,00									
Riso	0,13	1,00								
Pata	0,06	0,23	1,00							
Zucc	-0,41	-0,69	-0,28	1,00						
Verd	0,56	0,57	0,07	-0,64	1,00					
Vino	0,29	0,42	-0,13	-0,62	0,54	1,00				
Carn	-0,07	-0,15	0,29	-0,19	0,22	0,39	1,00			
Latt	-0,34	-0,39	-0,04	0,58	-0,75	-0,69	-0,41	1,00		
Burr	-0,52	-0,34	-0,19	0,43	-0,46	-0,06	0,29	0,10	1,00	
Uova	-0,34	-0,31	-0,10	0,02	0,07	0,11	0,60	-0,22	0,45	1,00

- Elementi su diagonale di matrice R =1 (ogni variabile è perfettamente correlata con se stessa)
- Elementi al di fuori della diagonale = coefficienti di correlazione tra coppie di variabili

I fattori/componenti

Possibili aree con alti coefficienti di correlazione:

- Gruppo di variabili che sta misurando **aspetti diversi della stessa dimensione/fenomeno sotteso**
- Tale fenomeno/dimensione sottesa è definito **FATTORE o COMPONENTE** (variabile latente)
- Riduzione di database da gruppi di variabili correlate tra loro a un set più piccolo ci permette di avere una minore dimensione, (usare il minor numero di concetti esplorativi) pur mantenendo la massima quantità di varianza comune.

I fattori/componenti

ESEMPI COMUNI: psicologia studi personalità (Eysenck,1953); economia; sociologia; marketing ecc.

Es . Studio sulla popolarità di una persona

coefficienti di correlazione per ogni coppia di variabili e si crea una matrice R

Obiettivo = misurare popolarità delle persone attraverso varie caratteristiche

- social skills
- selfish (egoismo)
- interest (interesse per gli altri)
- talk1 (tempo impiegato in una conversazione a parlare di altri)
- talk2 (tempo impiegato in una conversazione a parlare di sé)

I fattori/componenti

	Talk 1	Social Skills	Interest	Talk 2	Selfish	Liar
Talk 1	1.000					
Social Skills	0.772	1.000				
Interest	0.646	0.879	1.000			
Talk 2	0.074	–0.120	0.054	1.000		
Selfish	–0.131	0.031	–0.101	0.441	1.000	
Liar	0.068	0.012	0.110	0.361	0.277	1.000

Factor 1

Factor 2

Emergono chiaramente **due GRUPPI** (dimensione comune)

Generalmente nell'analisi fattoriale e in PCA):

- cerchiamo di ridurre questa matrice
- guardando a quali variabili sembrano raggrupparsi
-ma non sono correlate con variabili fuori dal gruppo

Nell'esempio: 2 cluster = 1) fattore definito: socialità
2) fattore definito: considerazione

DALLA IDENTIFICAZIONE ALLA RAPPRESENTAZIONE GRAFICA DI FATTORI/COMPONENTI

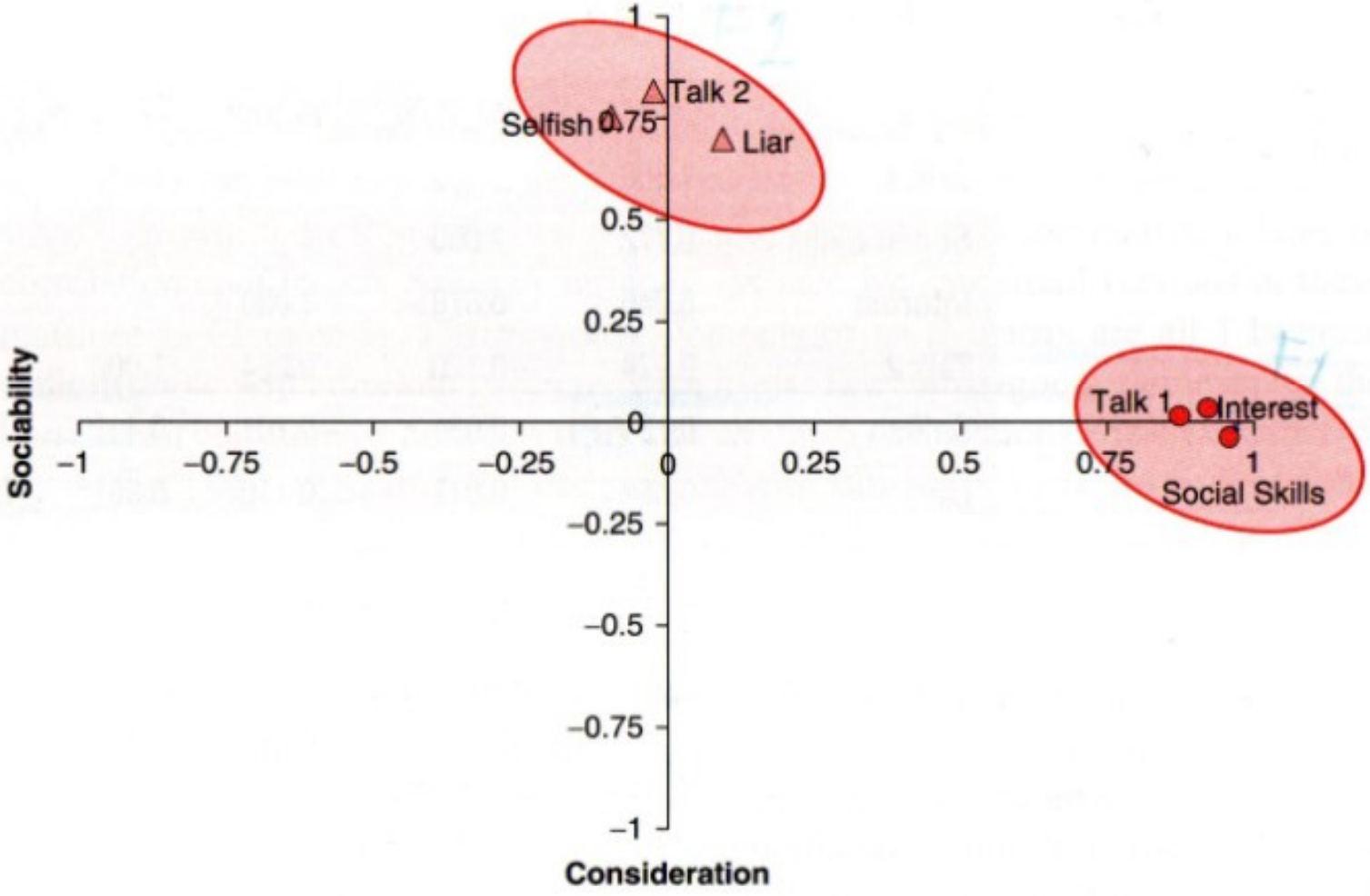
Rappresentazione grafica dei fattori/componenti

•**Fattori** = entità statistiche che possono essere visualizzate come assi di classificazione sui quali le variabili misurate possono essere rappresentate

•Se fattori come **assi** di un grafico → possiamo rappresentare le variabili su questi assi

•**Coordinate delle variabili = forza della relazione tra ogni variabile e il fattore rappresentato**

Rappresentazione grafica dei fattori/componenti



Rappresentazione grafica dei fattori/componenti

Note alla figura :

• per ogni fattore la linea degli assi va da -1 a +1 = limiti dei coefficienti di correlazione

•La posizione di ogni variabile dipende dalla sua correlazione con i due fattori

•Cerchi = variabili correlate con medesimo fattore e bassa corr. con l'altro fattore

•La figura rappresenta ciò che avevamo visto in matrice R

•Variabili = relate molto bene solo con un fattore (non sempre così)

•Variabili con alte coordinate su stesso asse: stanno misurando diversi aspetti di stessa dimensione

•Le coordinate di una variabile su un fattore sono note come peso fattoriale (factor loading)

Rappresentazione grafica dei fattori/componenti

PESO FATTORIALE =

- **correlazione di Pearson** tra un fattore e una variabile

- info contenute nel **coefficiente di correlazione**

- se eleviamo al 2 il peso fattoriale → misura dell'importanza di una particolare variabile per un fattore

ESPRESSIONE LINEARE DEI FATTORI/COMPONENTI

Rappresentazione matematica dei fattori/componenti

FATTORI = assi = linee

→Riconducibili matematicamente ad una linea retta

→Descritti da equazione che richiama il modello lineare

$$Y_i = b_1x_1 + b_2x_2 + ... + b_nx_n + E_i$$

Lo applichiamo allo scenario di descrizione di un fattore

$$Fattore_i = b_1 * variabile_1 + b_2 * variabile_2 + ... + b_n * variabile_n + E_i$$

Si noti che:

-Non c'è intercetta (b_0), in quanto intercetta degli assi è 0

-b_s rappresenta il peso fattoriale

Rappresentazione matematica dei fattori/componenti

ESEMPIO

2 fattori: socialità e considerazione → equazione descrive ogni fattore in termini delle variabili misurate:

$$Y_i = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n + \varepsilon_i$$

$$\begin{aligned}\text{Sociability}_i = {} & b_1\,\text{Talk 1}_i + b_2\,\text{Social Skills}_i + b_3\,\text{Interest}_i \\ & + b_4\,\text{Talk 2}_i + b_5\,\text{Selfish}_i + b_6\,\text{Liar}_i + \varepsilon_i\end{aligned}$$

$$\begin{aligned}\text{Consideration}_i = {} & b_1\,\text{Talk 1}_i + b_2\,\text{Social Skills}_i + b_3\,\text{Interest}_i \\ & + b_4\,\text{Talk 2}_i + b_5\,\text{Selfish}_i + b_6\,\text{Liar}_i + \varepsilon_i\end{aligned}$$

Rappresentazione matematica dei fattori/componenti

NOTE:

-Le equazioni sono identiche nella forma (tutte variabili in entrambe)

-La differenza dei valori di b = importanza relativa di ogni variabile su fattori

-Possibile sostituire ogni b con coordinate dal grafico (ovvero peso fattoriale)

-Le equazioni risultanti sono:

$$Y_i = b_1 X_1 + b_2 X_2 + \cdots + b_n X_n + \varepsilon_i$$

$$\begin{aligned}\text{Sociability}_i = {} & .87\ \text{Talk 1}_i + .96\ \text{Social Skills}_i + .92\ \text{Interest}_i \\ & + .00\ \text{Talk 2}_i - .10\ \text{Selfish}_i + .09\ \text{Liar}_i + \varepsilon_i\end{aligned}$$

$$\begin{aligned}\text{Consideration}_i = {} & .01\ \text{Talk 1}_i - .03\ \text{Social Skills}_i + .04\ \text{Interest}_i \\ & + .82\ \text{Talk 2}_i + .75\ \text{Selfish}_i + .70\ \text{Liar}_i + \varepsilon_i\end{aligned}$$

Rappresentazione matematica dei fattori/componenti

Coefficienti = Pearson correlation

Coefficiente correlazione al quadrato = R^2 → quanta della variabilità di una variabile è spiegata da un'altra

NOTE:

1. Per ogni fattore: alcune variabili valore alto Vs altre (importanza della variabile sul fattore)
2. Grafico ed equazioni rappresentano la stessa situazione
3. In mondo ideale: per ogni fattore le variabili hanno valore b_s molto alto rispetto ad un fattore e molto basso rispetto agli altri

Rappresentazione matematica dei fattori/componenti

4- i pesi fattoriali individuati possono essere organizzati in una matrice:

- colonne = fattori
- righe = peso di ogni variabile per un fattore

(es. : 2 fattori; 6 righe)

Matrice = A

$$A = \begin{pmatrix} .87 & .01 \\ .96 & -.03 \\ .92 & .04 \\ .00 & .82 \\ -.10 & .75 \\ .09 & .70 \end{pmatrix}$$

Rappresentazione matematica dei fattori/componenti

-Relazioniamo matrice A con equazioni fattoriali

- definizione di A = (in PCA) matrice dei componenti o (in analisi fattoriale) matrice dei fattori

-ASSUNZIONI alla base della matrice A:

-I fattori algebrici rappresentano le dimensioni del mondo reale

- la natura dei fattori ci permette di capire quali variabili hanno più importanza rispetto ad ogni fattore

Rappresentazione matematica dei fattori/componenti

BOX APPROFONDIMENTO SUI PESI FATTORIALI

matrice di struttura VS matrice di disposizione

Fino a qui pesi fattoriali indicati come: sia coefficienti di correlazione, sia coefficienti di regressione.

Entrambi concetti = relazione tra una variabile e un modello lineare

→concetto chiave: il peso fattoriale da informazioni su contributo relativo che una variabile ha su un fattore

→Interpretazione di peso fattoriale = rappresenta entrambi (coeff.corr e coef.regg)

→2 distinzioni: Rotazione

→ ORTOGONALE (fattori indipendenti; peso corr e reg)

→OBLIQUA (reg =pattern matrix; corr = structure matrix)

I VALORI FATTORIALI

I valori fattoriali

es. *descrivere una persona*

$$
\begin{aligned}
\text{Sociability} &= .87\,\text{Talk 1} + .96\,\text{Social Skills} + .92\,\text{Interest} \\
&\quad + .00\,\text{Talk 2} - .10\,\text{Selfish} + .09\,\text{Liar} \\
\text{Sociability} &= (.87 \times 4) + (.96 \times 9) + (.92 \times 8) + (.00 \times 6) \\
&\quad - (.10 \times 8) + (.09 \times 6) \\
&= 19.22 \\
\text{Consideration} &= .01\,\text{Talk 1} - .03\,\text{Social Skills} + .04\,\text{Interest} \\
&\quad + .82\,\text{Talk 2} + .75\,\text{Selfish} + .70\,\text{Liar} \\
\text{Consideration} &= (.01 \times 4) - (.03 \times 9) + (.04 \times 8) + (.82 \times 6) \\
&\quad + (.75 \times 8) + (.70 \times 6) \\
&= 15.21
\end{aligned}
$$

I valori fattoriali

Descrizione di un fattore:

- b = in termini di variabili che hanno un peso su di esso
- in termini di media ponderata per ogni caso

(es. per descrivere una persona e confrontarla con altre)

Punto critico = molto sensibile alla scala di misura delle variabili (se domande diverse con scale diverse → non paragonabili)

Metodo molto elementare
vediamo metodi più sofisticati

I valori fattoriali: metodo di regressione

Diverse tecniche per definire il punteggio fattoriale

- equazione $Y_i = b_1x_1 + b_2x_2 + \ldots + b_nx_n + E_i$

- b_s coefficiente di valore fattoriale

calcolato con metodo di regressione:

- il peso fattoriale è “aggiustato” per considerare l’iniziale correlazione tra variabili

- la differenza nelle scale di misura è neutralizzata

I valori fattoriali: metodo di regressione

METODO:

-B = matrice dei coefficienti di valore fattoriale

$= R^{-1} * A$

dove: R^{-1} = inversa di matrice di correlazione

A = matrice dei pesi fattoriali

Concetto = dividiamo i pesi fattoriali per i coefficienti di correlazione

B = matrice risultante (relazione tra ogni variabile e ogni fattore, considerando la relazione originaria tra coppie di variabili)

➔ B rappresenta una misura della relazione unica tra variabili e fattori

I valori fattoriali: metodo di regressione

Es

$$B = R^{-1}A$$

$$B = \begin{pmatrix} 4.76 & -7.46 & 3.91 & -2.35 & 2.42 & -.49 \\ -7.46 & 18.49 & -12.42 & 5.45 & -5.54 & 1.22 \\ 3.91 & -12.42 & 10.07 & -3.65 & 3.79 & -.96 \\ -2.35 & 5.45 & -3.65 & 2.97 & -2.16 & .02 \\ 2.42 & -5.54 & 3.79 & -2.16 & 2.98 & -.56 \\ -.49 & 1.22 & -.96 & .02 & -.56 & 1.27 \end{pmatrix} \begin{pmatrix} .87 & .01 \\ .96 & -.03 \\ .92 & .04 \\ .00 & .82 \\ -.10 & .75 \\ .09 & .70 \end{pmatrix}$$

$$= \begin{pmatrix} .343 & .006 \\ .376 & -.020 \\ .362 & .020 \\ .000 & .473 \\ -.037 & .437 \\ .039 & .405 \end{pmatrix}$$

- Le matrici riportano stesso tipo informazioni date dai pesi fattoriali e dai coefficienti di punteggio
- Cambia il peso (perché consideriamo la relazione tra variabili)

I valori fattoriali: metodo di regressione

I valori ottenuti da B possono essere inseriti nelle equazioni :

Es

$$\begin{aligned}
\text{Sociability} &= .343\,\text{Talk 1} + .376\,\text{Social Skills} + .362\,\text{Interest} \\
&\quad + .000\,\text{Talk 2} - .037\,\text{Selfish} + .039\,\text{Liar} \\
\text{Sociability} &= (.343 \times 4) + (.376 \times 9) + (.362 \times 8) + (.000 \times 6) \\
&\quad - (.037 \times 8) + (.039 \times 6) \\
&= 7.59 \\
\text{Consideration} &= .006\,\text{Talk 1} - .020\,\text{Social Skills} + .020\,\text{Interest} \\
&\quad + .473\,\text{Talk 2} + .437\,\text{Selfish} + .405\,\text{Liar} \\
\text{Consideration} &= (.006 \times 4) - (.020 \times 9) + (.020 \times 8) + (.473 \times 6) \\
&\quad + (.437 \times 8) + (.405 \times 6) \\
&= 8.768
\end{aligned}$$

Con tali equazioni è possibile descrivere una persona rispetto alla domanda di ricerca iniziale.

-Tale tecnica per ottenere valori fattoriali ha come risultato valori con : media = 0 ; varianza = ($\widehat{\text{score}}$ – score$_i$)^2

-Punto debole = i valori possono essere correlati con più fattori

I valori fattoriali: altri metodi

Per ovviare alla correlazione possibile nel metodo di regressione

→Diversi aggiustamenti

2 i principali :

1) metodo Barlett (fattori; punteggi fattoriali – multicollinearità)

2) metodo Anderson-Rubin (punteggi fattoriali non correlati e standardizzati. Media = 0; std=1)

quale metodo utilizzare?

Tabachnick e Fidell (2011)

- se necessità di non correlazione = A.R.

- altrimenti regressione (per interpretazione)

I valori fattoriali: utilizzo

Definizione =

Il valore fattoriale è un punteggio composito per ogni individuo su un particolare fattore (o dimensione)

UTILIZZI:

-Riduzione del database

- condurre analisi con fattori (non con variabili originarie)

-Risolvere la collinearità per la regressione multipla

Individuare i fattori/componenti

Punto situazione:

1. Concetto di fattore
2. Sua rappresentazione grafica
3. Sua rappresentazione algebrica
4. Costruzione di un valore/punteggio composito che rappresenta la “performance” di un caso (individuo) in un singolo fattore

COME INDIVIDUARE I FATTORI?

Individuare i fattori/componenti: scelta del metodo

Tinsley & Tinsley (1987): 2 elementi cruciali nella scelta del metodo

a) Se vogliamo generalizzare o no

b) Esplorazione dati o conferma ipotesi

(es. confirmatory factor analysis : Pedhazur e Shmelkin 1991, cap.23)

	Risultati al campione (metodi descrittivi)	Generalizzazione (metodi inferenziali)
Esplorare dati	PCA base	PCA ⇔ -ripetute analisi su diversi campioni rivelano stessa struttura fattoriale, oppure -Casi selezionati casualmente e variabili rappresentano la popolazione di variabili di interesse (solo var. misurate) (es. Max. Likelihood method –Harman, 1976; Kaiser Method su dipendenza dei fattori).
Confermare ipotesi	Confirmatory factor analysis	Confirmatory factor analysis

Scelta del metodo dipende dall'obiettivo

Individuare i fattori/componenti: le comunalità

Definizione di comunalità:

- La comunalità esprime la proporzione della varianza di ogni variabile riprodotta da un certo numero di componenti. Essendo una proporzione, essa varia tra zero e uno. Quindi ci dice quanta varianza perdiamo di ciascuna variabile, tenendo conto delle componenti che abbiamo deciso di utilizzare.

- Le comunalità indicano la parte di varianza spiegata di ogni indicatore, considerando il modello fattoriale stimato

- Vanno tendenzialmente tenuti in considerazione item o variabili che abbiano un valore di comunalità di almeno .500

Logica della comunalità:

Nella matrice R è possibile calcolare il valore della variabilità (ovvero la varianza) per ogni variabile.

Individuare i fattori/componenti: le comunalità

Logica della comunalità:
Varianza per ogni variabile

$$S^2 = \frac{\sum_{i=1}^{n} (X_j - \overline{X})^2}{n-1}$$

La varianza totale di ogni variabile ha due componenti:
• una parte è specifica di quella variabile (varianza unica, con una componente di errore definita varianza random)
• una parte è condivisa con le altre variabili (varianza comune)

→**La proporzione di varianza comune presente in una variabile è nota come COMUNALITA'**

Varia da 1 a 0:

- 0 = variabile che non condivide varianza con altre variabili
- 1 = variabile che non ha varianza unica

Noi siamo interessati a trovare dimensioni comuni sottostanti nei dati raccolti = = siamo interessati alla varianza in comune (comunalità)

Individuare i fattori: le comunalità

Prima di eseguire la PCA dovremmo conoscere la varianza in comune – empasse

Risoluzione del problema: 2 opzioni

- assumiamo che tutta VARIANZA sia varianza comune (comunalità =1)
- stimiamo la comunalità per ogni variabile

Possiamo usare le informazioni contenute in R^2 per ogni variabile

Es. Multipla lm usando "selfish" come Y e tutte le altre variabili come Xi R^2 associato è una misura di comunalità di "selfish". Si ripete per tutte le variabili.

Da qui procediamo con L'ESTRAZIONE FATTORIALE

Criteri di scelta del numero di componenti

1. **Variabilità spiegata**
   si fissa una soglia minima di variabilità spiegata (in percentuale rispetto alla variabilità totale che è pari alla traccia della matrice di correlazione/di varianze e covarianze);
2. **Eigenvalue-one** (per variabili standardizzate)
   Poiché le variabili originarie standardizzate hanno varianza unitaria si scelgono solo gli autovalori maggioridi uno (i quali esprimono CP che , essendo la varianza maggiore di uno, sintetizzano maggiore informazione rispetto alle singole variabili originarie);
3. **Scree-Test**
   si considerano le CP i cui autovalori precedono il salto massimo di variabilità spiegata.

Letture consigliate

- Field, A. (2002) Discovering statistics using SPSS (2° ed.) Sage publication. Chap 15.
- Lattin, J. Et al. (2005) Analyzing Multivariate Data. Chap 4.

Caso studio

UNO STUDIO DI CASO

Un caso studio: lo stato di salute di alcune aziende

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:

È quella di individuare la *migliore sintesi degli indici di bilancio* che consenta di ordinare le aziende sulla base dei due fattori ipotizzati.

Essendo le variabili tutte di natura numerica, si utilizza l'**Analisi delle Componenti Principali.**

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

Le variabili considerate nel dataset. Fonte: Zani S. (2000). Analisi dei dati statistici, volume II, Editore Giuffrè.

Un caso studio: 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

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

Il dataset. Data la disomogeneità delle varibili si procede standardizzando le stesse.

Un caso studio: la 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 dei dati standardizzati.

Un caso studio: la matrice di correlazione

L'osservzione 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.

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 caso studio: scelta delle componenti

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

I due spazi dell'analisi

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**.
  Si ricercano gli autovalori e gli autovettori della matrice $\tilde{X}'\tilde{X}$
- *dei punti-variabile nello* **spazio degli individui**
  Si ricercano gli autovalori e gli autovettori della matrice trasposta $\tilde{X}\tilde{X}'$

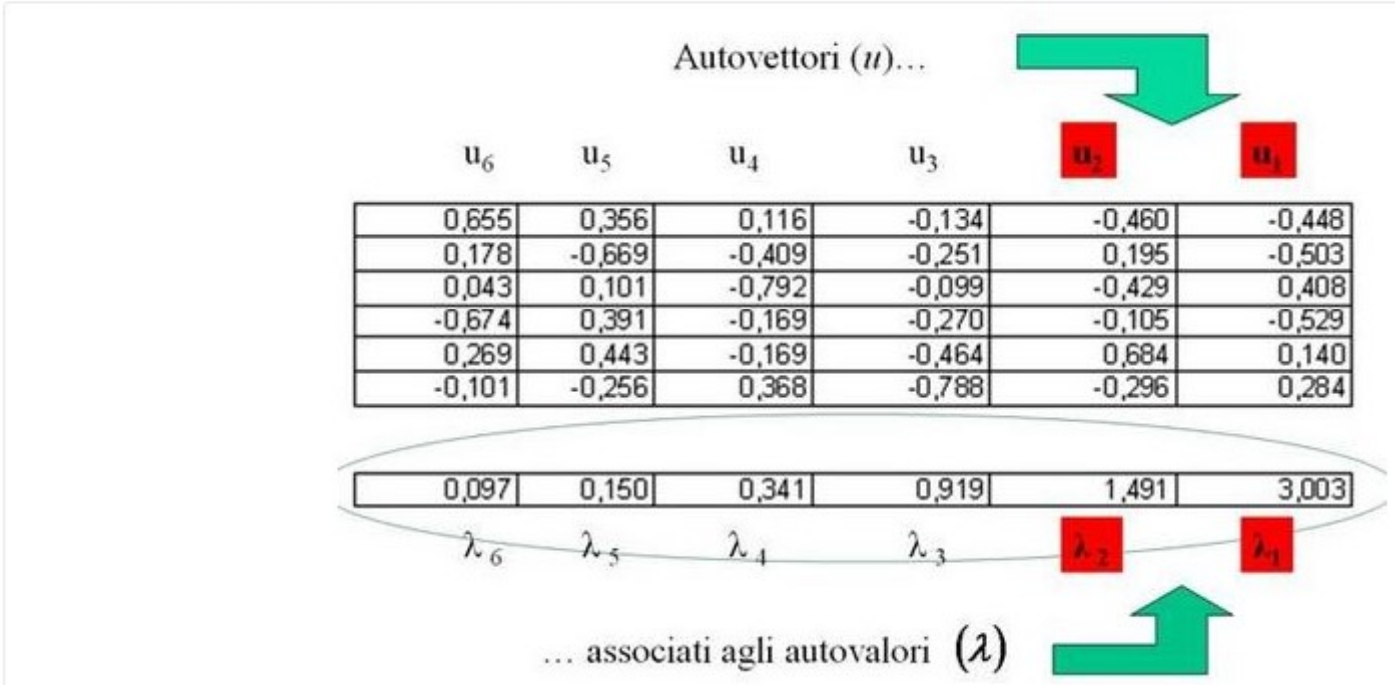
Si può dimostrare che gli autovalori ottenuti nelle due analisi coincidono.

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.

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

Un caso studio: Analisi dei punti-unità



Un caso studio: coordinate delle aziende

Coordinate degli individui sulla prima componente

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

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 $\mathbf{u}_1$ (*6*x*1*)

-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 (*10*x*1*)

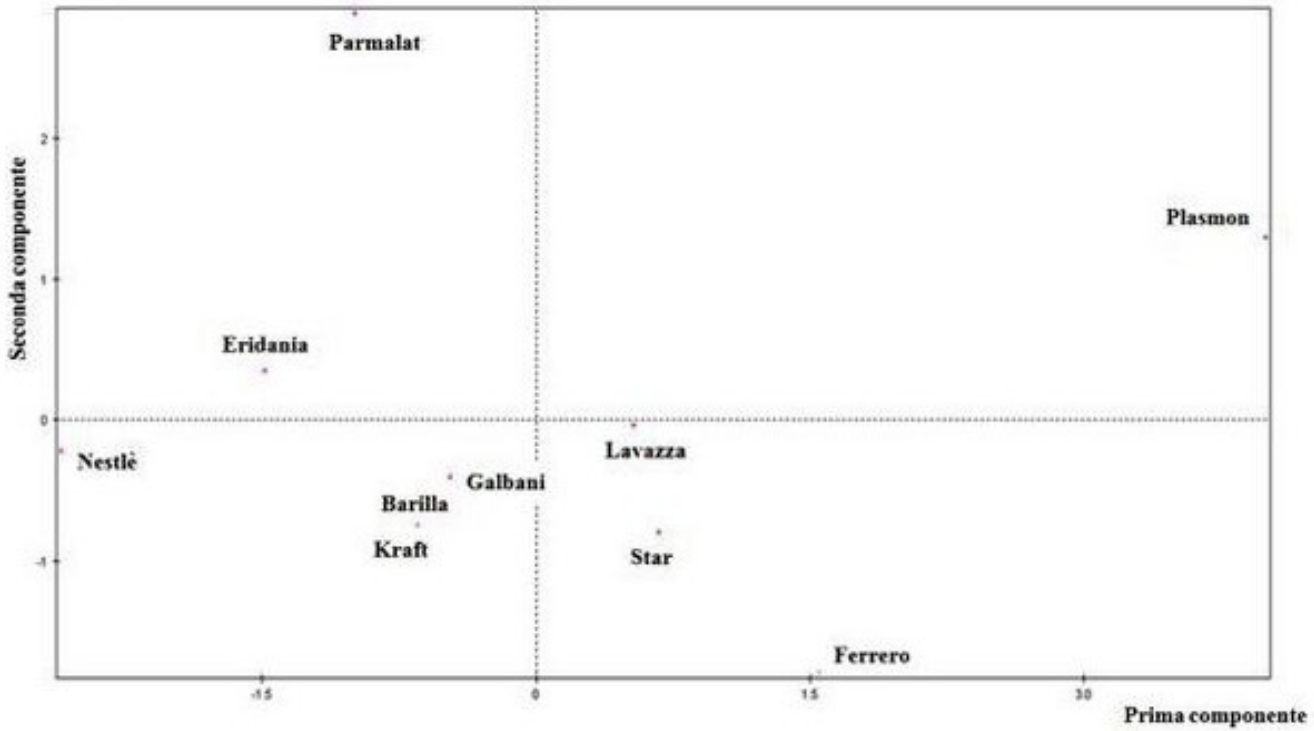
Una volta determinato il sottospazio ottimale ad h dimensioni individuato dagli h autovettori $\{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_j, \ldots, \mathbf{u}_h\}$ le coordinate dell'i-mo punto-unità sull'j-mo asse fattoriale saranno

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

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 caso studio: I° piano fattoriale delle aziende



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

Un caso studio: le coordinate dei punti-variabile

Analogamente all'analisi delle unità, per le variabili li coordinate si calcolano moltiplicando le righe della matrice trasposta X' per il vettore degli autovalori {v1,v2,…,vj,…,vh}.

Nell'immagine di fianco si riportano 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°.

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 CP_1 ed è incorrelato con CP_2.

L'analisi di queste coordinate consente di interpretare le componenti latenti!!

	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 caso studio: I° piano fattoriale delle variabili

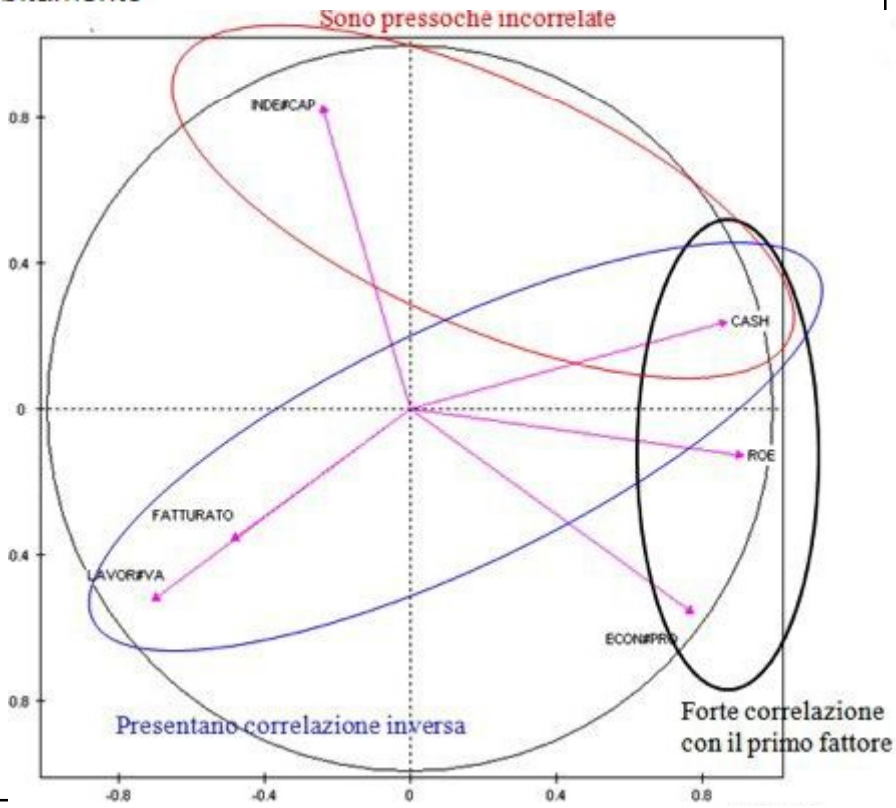
Interpretazione

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.



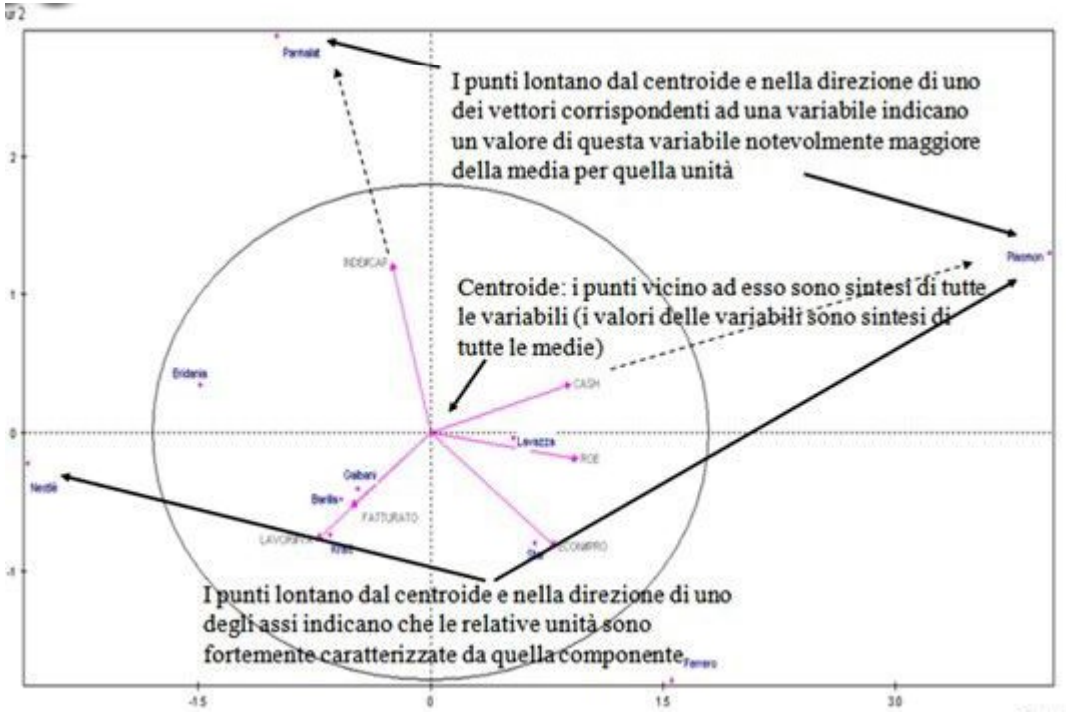
Un caso studio: Interpretazione dei punti-aziende

Alcune osservazioni

La **Plasmon** presenta elevatissimi valori di redditività (CP1) e un indebitamento sopra la media (CP2).

La **Parmalat** presenta scarsi valori di redditività (CP1) e un fortissimo indebitamento sopra la media (CP2).

Le aziende vicino al centro degli assi presentano redditività e indebitamento nella media.



Punti supplementari

Sui piani fattoriali, risultato dell'analisi, è possibile *proiettare alcuni punti in supplementare*.

Punti-unità in supplementare

Non concorrono a determinare la soluzione fattoriale ma sono proiettati sui piani fattoriali per studiare la loro prossimità con i punti-unità che hanno concorso a determinare le componenti principali.

Esempi:

- nuove osservazioni;
- osservazioni "di controllo".

Punti-variabile in supplementare

Non concorrono a determinare la soluzione fattoriale ma sono proiettati sui piani fattoriali per studiare la loro correlazione con le componenti principali.

Esempi:

- variabili socio-demografiche
- variabili ridondanti