L'Analisi in Componenti Principali con metriche diverse

Analisi dei Dati[1]

[1]Corso di Laurea in Scienze Statistiche e Attuariali
Dipartimento di Diritto, Economia, Management e Metodi Quantitativi (DEMM)
Università degli Studi del Sannio

Prof. Pietro Amenta

Fonte: *Pietro Amenta. Appunti di Analisi dei Dati Multidimensionali*

Indice

- Uno studio multidimensionale nello spazio $\Re^p$ deve considerare, oltre alla matrice dei dati, anche un sistema di *pesi* legato alle unità statistiche e un criterio per il calcolo delle distanze tra le unità rappresentato dalla *metrica*.
- L'insieme dei pesi relativi alle singole unità può essere riportato in una matrice diagonale **D** di dimensione ($n \times n$):

$$\mathbf{D} = \begin{pmatrix} p_1 & 0 & 0 & 0 \\ 0 & p_1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & p_n \end{pmatrix}$$

- I pesi (o masse) delle unità sono generalmente equiparati a delle frequenze relative ed hanno quindi la caratteristica che $\sum_{i=1}^{n} p_i = 1$.

- Quando le unità sono costituite da osservazioni casuali o quando non vi sia motivo di assegnare pesi differenti alle diverse unità statistiche, i pesi risultano tutti uguali e pari a $1/n$, cioè:
$$p_i = \frac{1}{n}, \forall i \in \{1, \dots, n\}$$
e la matrice $\mathbf{D}$ può essere scritta come $\mathbf{D} = \frac{1}{n} \times \mathbf{I}_n$.
- Il vettore del baricentro $\mathbf{g}$ della nube dei punti in $\Re^p$ è dato da
$$\mathbf{g} = \mathbf{Y}^T \mathbf{D} \mathbf{1}$$
dove $\mathbf{1}$ è un vettore colonna $(n \times 1)$ con tutte le componenti pari a 1.
- La matrice di varianza-covarianza è data da
$$\mathbf{V} = \mathbf{Y}^T \mathbf{D} \mathbf{Y} - \mathbf{g}\mathbf{g}^T = \mathbf{X}^T \mathbf{D} \mathbf{X}$$

Disuguaglianza di Cauchy-Schwarz

Si consideri i vettori $\mathbf{x}$ e $\mathbf{y}$ e il prodotto scalare $< \mathbf{x}, \mathbf{y} >_{\mathbf{D}} = \mathbf{x}^T\mathbf{D}\mathbf{y}$
Sia

$$\mathbf{a} = [\frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|}] = [\mathbf{x}(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}]$$

allora si consideri la disuguaglianza $< \mathbf{a}, \mathbf{a} >_{\mathbf{D}} \geqslant 0$

$$\begin{aligned}
\mathbf{a}^T\mathbf{D}\mathbf{a} &= [\mathbf{x}(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}]^T\mathbf{D}[\mathbf{x}(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}] \\
&= \underbrace{(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}}\mathbf{x}^T\mathbf{D}\mathbf{x}(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}}}_{=1} - (\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}}\mathbf{x}^T\mathbf{D}\mathbf{y}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}} + \\
&- (\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}\mathbf{y}^T\mathbf{D}\mathbf{x}(\mathbf{x}^T\mathbf{D}\mathbf{x})^{-\frac{1}{2}} + \underbrace{(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}\mathbf{y}^T\mathbf{D}\mathbf{y}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{-\frac{1}{2}}}_{=1} \\
&= 2 - 2\frac{\mathbf{x}^T\mathbf{D}\mathbf{y}}{(\mathbf{x}^T\mathbf{D}\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{D}\mathbf{y})^{\frac{1}{2}}}
\end{aligned}$$

Disuguaglianza di Cauchy-Schwarz

Si consideri i vettori $\mathbf{x}$ e $\mathbf{y}$ e il prodotto scalare $\mathbf{x}^T\mathbf{y}$

Sia

$$\mathbf{a} = [\frac{\mathbf{x}}{\|\mathbf{x}\|} - \frac{\mathbf{y}}{\|\mathbf{y}\|}] = [\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}]$$

allora si consideri la disuguaglianza $\mathbf{a}^T\mathbf{a} = \sum_{i=1}^{n} a_i^2 \geqslant 0$

$$\begin{aligned}
\mathbf{a}^T\mathbf{a} &= [\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}]^T[\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}} - \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}] \\
&= \underbrace{(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}}\mathbf{x}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}}}_{=1} - (\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}}\mathbf{x}^T\mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}} + \\
&- (\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}\mathbf{y}^T\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}} + \underbrace{(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}\mathbf{y}^T\mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}}_{=1} \\
&= (2 - 2\frac{\mathbf{x}^T\mathbf{y}}{(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}}) \geqslant 0
\end{aligned}$$

Disuguaglianza di Cauchy-Schwarz

quindi

$$(2 - 2\frac{\mathbf{x}^T\mathbf{y}}{(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}}) \geqslant 0 \implies \mathbf{x}^T\mathbf{y} \leqslant (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}$$

Se, invece,

$$\mathbf{a} = [\frac{\mathbf{x}}{\|\mathbf{x}\|} + \frac{\mathbf{y}}{\|\mathbf{y}\|}] = [\mathbf{x}(\mathbf{x}^T\mathbf{x})^{-\frac{1}{2}} + \mathbf{y}(\mathbf{y}^T\mathbf{y})^{-\frac{1}{2}}]$$

abbiamo

$$\mathbf{a}^T\mathbf{a} = 2 + 2\frac{\mathbf{x}^T\mathbf{y}}{(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}}$$

e quindi

$$(2 + 2\frac{\mathbf{x}^T\mathbf{y}}{(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}}) \geqslant 0 \implies -(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}} \leqslant \mathbf{x}^T\mathbf{y}$$

Disuguaglianza di Cauchy-Schwarz

Unendo i due risultati otteniamo la **Disuguaglianza di Cauchy-Schwarz**

$$-(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}} \leqslant \mathbf{x}^T\mathbf{y} \leqslant (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}$$

Se, inoltre, dividiamo tutto per $(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}$, abbiamo

$$\boxed{-1 \leqslant \frac{\mathbf{x}^T\mathbf{y}}{(\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}(\mathbf{y}^T\mathbf{y})^{\frac{1}{2}}} \leqslant 1}$$

cioè

$$\boxed{-1 \leqslant \frac{\mathbf{x}^T\mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \leqslant 1} \text{ o anche } \boxed{-1 \leqslant \frac{\mathbf{x}^T\mathbf{D}\mathbf{y}}{\|\mathbf{x}\|_{\mathbf{D}}\|\mathbf{y}\|_{\mathbf{D}}} \leqslant 1}$$

Disuguaglianza di Cauchy-Schwarz

Se $\mathbf{x}$ e $\mathbf{y}$ sono centrati rispetto alle medie $\bar{\mathbf{x}} = \mathbf{1}\mathbf{1}^T\mathbf{x}/n$ e $\bar{\mathbf{y}} = \mathbf{1}\mathbf{1}^T\mathbf{y}/n$

$$\hat{\mathbf{x}} = [\mathbf{x} - \bar{\mathbf{x}}] \qquad \hat{\mathbf{y}} = [\mathbf{y} - \bar{\mathbf{y}}]$$

allora

$$\hat{\mathbf{x}}^T\hat{\mathbf{y}} = [\mathbf{x} - \bar{\mathbf{x}}]^T[\mathbf{y} - \bar{\mathbf{y}}] = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = Codev(x, y)$$

$$\|\hat{\mathbf{x}}\| = \sqrt{\hat{\mathbf{x}}^T\hat{\mathbf{x}}} = \sqrt{[\mathbf{x} - \bar{\mathbf{x}}]^T[\mathbf{x} - \bar{\mathbf{x}}]} = \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \sqrt{Dev(x)}$$

$$\|\hat{\mathbf{y}}\| = \sqrt{\hat{\mathbf{y}}^T\hat{\mathbf{y}}} = \sqrt{[\mathbf{y} - \bar{\mathbf{y}}]^T[\mathbf{y} - \bar{\mathbf{y}}]} = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \sqrt{Dev(y)}$$

Disuguaglianza di Cauchy-Schwarz

allora

$$\boxed{-1 \leqslant \frac{\hat{\mathbf{x}}^T \hat{\mathbf{y}}}{\|\hat{\mathbf{x}}\| \|\hat{\mathbf{y}}\|} \leqslant 1}$$

risulta essere pari a

$$\boxed{-1 \leqslant \frac{Codev(x, y)}{\sqrt{Dev(x) Dev(y)}} \leqslant 1}$$

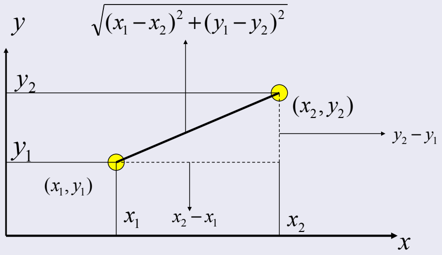
equivalente al coefficiente di correlazione lineare r tra **x** e **y**

$$\boxed{-1 \leqslant r(\mathbf{x}, \mathbf{y}) \leqslant 1}$$

Problema

Posto che ciascuna delle n unità statistiche (righe) può essere come un punto nello spazio $\Re^p$ generato dalle p variabili osservate il cui baricentro è rappresentato dal vettore $\mathbf{g}$, *come si misura la distanza tra due righe* ?

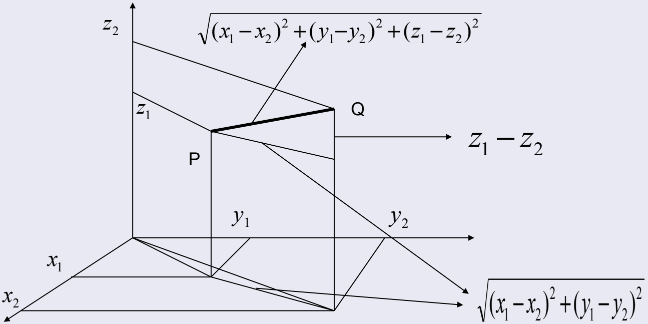
Distanza euclidea nel piano
(teorema di Pitagora)



Distanza Euclidea nello spazio $\Re^3$

- La distanza di due punti P e Q, di coordinate rispettivamente (x_1, y_1, z_1) e (x_2, y_2, z_2) è pari a

$$d(P,Q) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$



Distanza Euclidea nello spazio $\Re^p$

Il concetto di distanza con $p > 3$ può essere facilmente generalizzato. La distanza euclidea di due punti P e G appartenenti a $\Re^p$, di coordinate rispettivamente $(P_1, \ldots, P_p)$ e $(G_1, \ldots, G_p)$ risulta allora pari a

$$d(P, G) = \sqrt{(P_1 - G_1)^2 + (P_2 - G_2)^2 + \ldots + (P_p - G_p)^2}$$

$$= \sqrt{\sum_{i=1}^{p} (P_i - G_i)^2} = \sqrt{(\mathbf{P} - \mathbf{G})^T (\mathbf{P} - \mathbf{G})}$$

$$= < (\mathbf{P} - \mathbf{G}), (\mathbf{P} - \mathbf{G}) >^{\frac{1}{2}}$$

dove $\mathbf{P}^T = (P_1, \ldots, P_p)$ e $\mathbf{G}^T = (G_1, \ldots, G_p)$

Distanza Euclidea nello spazio $\Re^p$

- Il concetto di distanza con $p > 3$ può essere adesso facilmente generalizzato per l'utilizzo con qualunque metrica **Q**.
- La distanza euclidea di due punti P e Q appartenenti a $\Re^p$, di coordinate rispettivamente $(P_1, \ldots, P_p)$ e $(G_1, \ldots, G_p)$ risulta allora pari a

$$d(P, G) = \sqrt{(\mathbf{P}-\mathbf{G})^T\mathbf{Q}(\mathbf{P}-\mathbf{G})} =< (\mathbf{P}-\mathbf{G}), (\mathbf{P}-\mathbf{G}) >_{\mathbf{Q}}^{\frac{1}{2}}$$

  dove **Q** è una metrica in $\Re^p$ di dimensione $(p \times p)$ simmetrica e definita positiva.
- **Q** sarà diagonale se la ponderazione riguarda solo le variabili, mentre conterrà elementi extra-diagonali non nulli qualora si considerino anche le interazioni.

Metriche più utilizzate

- La metrica $\mathbf{Q} = \mathbf{I}_p$ non comporta alcuna ponderazione e viene utilizzata su dati omogenei.
- La metrica $\mathbf{Q} = \mathbf{D}_{1/\sigma}$ equivale a dividere ciascun elemento per lo scarto quadratico medio della variabile corrispondente e, quindi, considerando che i dati sono centrati, a standardizzare le variabili iniziali. E' una operazione molto utile quando si analizzano variabili eterogenee o espresse in unità di misura molto differenti.
- Le due situazioni possono essere equivalenti. L'analisi può essere effettuata o sulla matrice dei dati iniziali utilizzando la metrica $\mathbf{Q} = \mathbf{D}_{1/\sigma}$ oppure sulla matrice dei dati trasformati $\mathbf{X}^* = \mathbf{X}\mathbf{D}_{1/\sigma}$ utilizzando la metrica euclidea $\mathbf{Q} = \mathbf{I}_p$.

- Si dimostra che per ogni matrice simmetrica definita positiva $\mathbf{Q}$ esiste una matrice $\mathbf{T}$ tale che $\mathbf{Q} = \mathbf{T}^T\mathbf{T}$. Indicati con $\mathbf{e}_1$ e $\mathbf{e}_2$ i vettori contenenti i valori delle variabili osservate su due unità, il prodotto scalare è pari a

$$< \mathbf{e}_1, \mathbf{e}_2 >_{\mathbf{Q}} = \mathbf{e}_1^T\mathbf{Q}\mathbf{e}_2 = \mathbf{e}_1^T\mathbf{T}^T\mathbf{T}\mathbf{e}_2 =$$

$$= (\mathbf{T}\mathbf{e}_1)^T(\mathbf{T}\mathbf{e}_2) = < \mathbf{T}\mathbf{e}_1, \mathbf{T}\mathbf{e}_2 >_{\mathbf{I}_p}$$

- Effettuare il prodotto scalare tra due vettori $\mathbf{e}_1$ e $\mathbf{e}_2$ utilizzando la metrica $\mathbf{Q}$ equivale quindi ad effettuare il prodotto scalare ordinario, con metrica $\mathbf{Q} = \mathbf{I}$, dopo aver sostituito la matrice iniziale $\mathbf{X}$ con la matrice trasformata $\mathbf{X}^* = \mathbf{X}\mathbf{T}^T$.

L'Analisi in Componenti Principali

Pearson, 1901; Hotelling, 1930

Interpretazione analitica

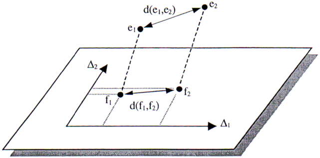
- L'Analisi in Componenti Principali (ACP) si propone di individuare i fattori latenti che costituiscono la struttura di fondo delle relazioni osservate.
- L'ipotesi è che i fattori siano legati linearmente alle variabili originarie e di numero inferiore. L'economia nella descrizione del sistema non viene ottenuta riducendo il numero delle variabili di partenza ma eliminando la ridondanza di informazioni che deriva dall'aver osservato variabili tra loro correlate

L'Analisi in Componenti Principali

Pearson, 1901; Hotelling, 1930

Interpretazione geometrica

- L'ACP determina nello spazio delle *p* variabili delle nuove variabili, combinazioni lineari delle originarie, in grado di rappresentare al meglio l'informazione strutturale del sistema rispetto ad un criterio di ottimizzazione. Si ricerca la migliore rappresentazione, in termini di proiezione, delle distanze originarie tra i punti osservati.

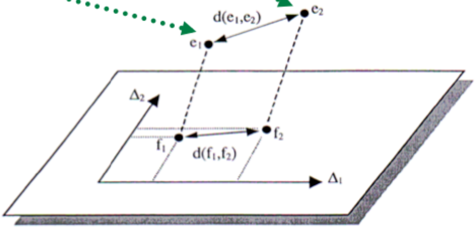


L'Analisi in Componenti Principali

Paesi e consumi alimentari
(16x9) kg pro-capite 1994

Analisi in R^p
(Spazio delle variabili)

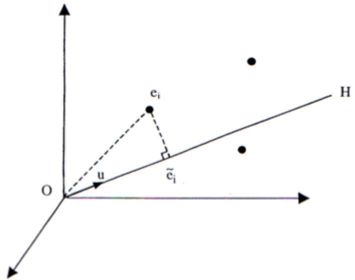
	Cereali	Riso	Patate	Zucchero	Verdure	Carne	Latte	Burro	Uova
Belgio	72,20	4,20	98,80	40,40	103,20	102,00	80,00	7,70	14,20
Italia	110,20	4,80	38,60	27,90	181,90	88,00	65,00	2,40	11,10
....	....	....	....	....	....	....	....	....	....
Grecia	109,80	5,40	90,00	30,00	229,50	77,10	63,10	0,90	11,30
Svezia	69,30	4,30	70,00	37,50	48,50	60,50	154,10	5,70	12,90



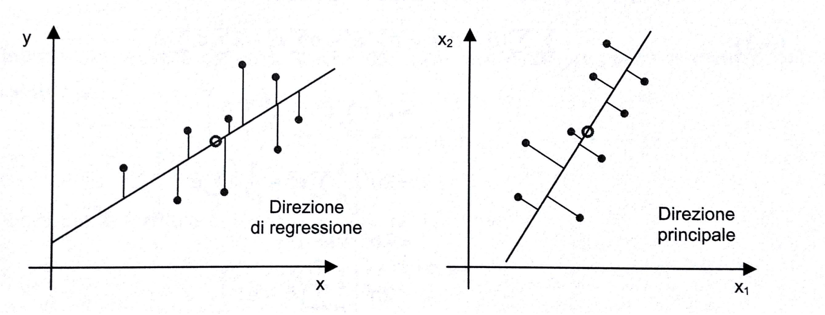
L'Analisi in Componenti Principali

La ricerca degli assi

La distanza euclidea dall' origine di un punto proiettato è sempre minore o al massimo uguale a quello nello spazio originario (cateto ed ipotenusa di un triangolo rettangolo)



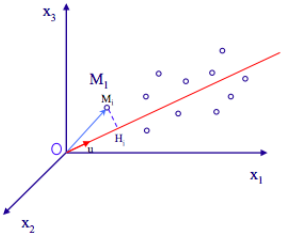
L'Analisi in Componenti Principali



L'Analisi in Componenti Principali

La ricerca degli assi

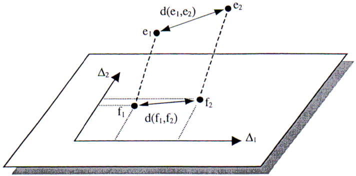
Adattamento degli *n* punti in R^p



$$\boxed{\sum_i (M_i - H_i)^2 = \min \quad \Rightarrow \quad \sum_i (OH_i)^2 = \max}$$

$$\boxed{OH_i = \mathbf{x}_i'\mathbf{u} = \sum_{j=1}^{p} x_{ij} u_j}$$

L'Analisi in Componenti Principali



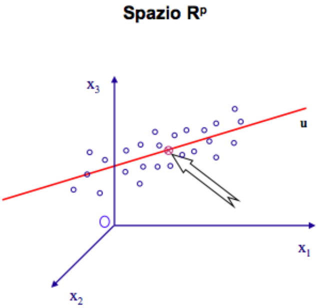
$$
\begin{aligned}
\sum_i \sum_j (f_i - f_j)^2 &= n \sum_i f_i^2 + n \sum_j f_j^2 - 2 \sum_i f_i \sum_j f_j \\
&= 2(n \sum_i f_i^2) - 2(\sum_i f_i)^2 \\
&= 2n^2 \left[\frac{1}{n} \sum_i f_i^2 - \frac{1}{n^2} (\sum_i f_i)^2 \right] \\
&= 2n^2 \left[\frac{1}{n} \sum_i (f_i - \bar{f}_i)^2 \right]
\end{aligned}
$$

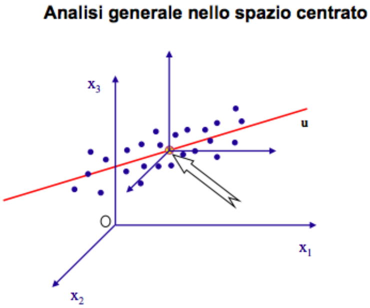
$$
\Delta^2 = \frac{\sum_i \sum_j (f_i - f_j)^2}{n^2} = 2 \left[\frac{1}{n} \sum_i (f_i - \bar{f}_i)^2 \right]
$$

da cui deriva che massimizzare la somma dei quadrati delle distanze tra tutte le coppie di punti proiettati (f_i, f_j) equivale a massimizzare la somma dei quadrati delle distanze dei punti f_i dal baricentro.

L'Analisi in Componenti Principali

Il precedente risultato equivale anche a massimizzare la somma dei quadrati delle distanze dei punti f_i dalla nuova origine ottenuta traslando il sistema il sistema di riferimento nel baricentro dei punti





L'Analisi in Componenti Principali

Lo stesso concetto si può esprimere usando il concetto di **inerzia**. Si definisce **inerzia totale** di una nube di n punti $\mathbf{f}_i$ la media dei quadrati delle distanze degli n punti dal centro di gravità $\mathbf{g}$:

$$\mathcal{I} = \sum_i p_i d^2(\mathbf{f}_i, \mathbf{g})$$

L'inerzia può essere anche espressa rispetto a un punto $\mathbf{h} \neq \mathbf{g}$:

$$\mathcal{I}_{\mathbf{h}} = \sum_i p_i d^2(\mathbf{f}_i, \mathbf{h})$$

Per il teorema di **Christian Huyghens** (1629-1697) abbiamo che

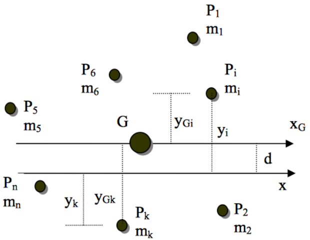
$$\mathcal{I}_{\mathbf{h}} = \mathcal{I} + d^2(\mathbf{g}, \mathbf{h})$$

Tale teorema implica che $\mathcal{I}_{\mathbf{h}}$ è sempre maggiore di $\mathcal{I}$, raggiungendo il valore minimo ($\mathcal{I}_{\mathbf{h}} = \mathcal{I}$) per $\mathbf{g} = \mathbf{h}$.

Teorema di Huyghens (esempio assi paralleli)

Sia $y_i = y_{G_i} + d$ la distanza di un punto P_i rispetto all'asse x:

$$\boxed{d^2(P_i, x) = y_i = y_{G_i} + d = d^2(P_i, G) + d^2(G, x)}$$

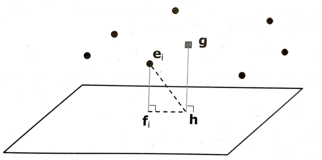


$$\begin{aligned}
\mathcal{I}_x &= \sum_i m_i y_i^2 = \sum_i m_i (y_{G_i} + d)^2 \\
&= \sum_i m_i y_{G_i}^2 + 2d \sum_i m_i y_{G_i} + \sum_i m_i d^2 \\
&= \mathcal{I} + 0 + \sum_i m_i d(G_i, x)^2 \\
&= \mathcal{I} + d(\mathbf{G}, \mathbf{x})^2
\end{aligned}$$

e quindi avremo $\mathcal{I}_x = \mathcal{I}$ solo se $d(\mathbf{G}, \mathbf{x}) = 0$, quindi per $X_G = x$.

L'Analisi in Componenti Principali

Per il teorema di Pitagora, risulta evidente che la distanza fra il punto $\mathbf{f}_i$ e $\mathbf{h}$ è un cateto del triangolo $(\mathbf{f}_i, \mathbf{e}_i, \mathbf{h})$



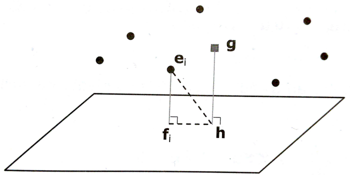
possiamo scrivere allora

$$d^2(\mathbf{e}_i, \mathbf{f}_i) = d^2(\mathbf{e}_i, \mathbf{h}) - d^2(\mathbf{f}_i, \mathbf{h})$$

e quindi, considerando tutti i punti con i rispettivi pesi,

$$\sum_i p_i d^2(\mathbf{e}_i, \mathbf{f}_i) = \sum_i p_i d^2(\mathbf{e}_i, \mathbf{h}) - \sum_i p_i d^2(\mathbf{f}_i, \mathbf{h})$$

L'Analisi in Componenti Principali



Considerando anche il teorema di Huyghens, possiamo scrivere

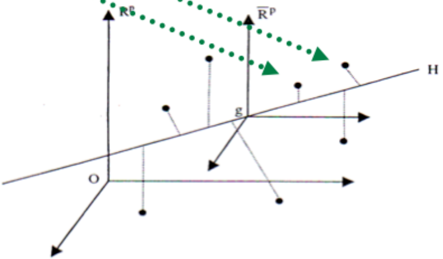
$$\begin{aligned}\sum_i p_i d^2(\mathbf{e}_i, \mathbf{f}_i) &= \sum_i p_i d^2(\mathbf{e}_i, \mathbf{h}) - \sum_i p_i d^2(\mathbf{f}_i, \mathbf{h}) \\ &= \mathcal{I}_{\mathbf{h}} - \sum_i p_i d^2(\mathbf{f}_i, \mathbf{h}) \\ &= \mathcal{I} + d^2(\mathbf{g}, \mathbf{h}) - \sum_i p_i d^2(\mathbf{f}_i, \mathbf{h})\end{aligned}$$

quindi la media dei quadrati delle distanze fra gli $\mathbf{e}_i$ e gli $\mathbf{f}_i$ risulta minima quando $\mathbf{g} = \mathbf{h}$ e quando l'inerzia della nube proiettata $\sum_i p_i d^2(\mathbf{f}_i, \mathbf{h})$ risulta massima.

Gli assi principali passeranno quindi per **g**.

L'Analisi in Componenti Principali

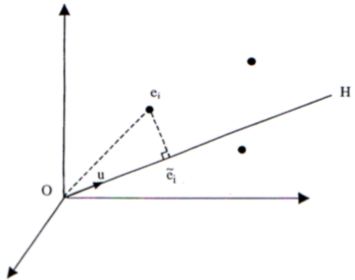
	Cereali	Riso	Patate	Zucchero	Verdure	Carne	Latte	Burro	Uova
Belgio	72,20	4,20	98,80	40,40	103,20	102,00	80,00	7,70	14,20
Grecia	109,80	5,40	90,00	30,00	229,50	77,10	63,10	0,90	11,30
Italia	110,20	4,80	38,60	27,90	181,90	88,00	65,00	2,40	11,10
………	………	………	………	………	………	………	………	………	………
Svezia	69,30	4,30	70,00	37,50	48,50	60,50	154,10	5,70	12,90



L'Analisi in Componenti Principali

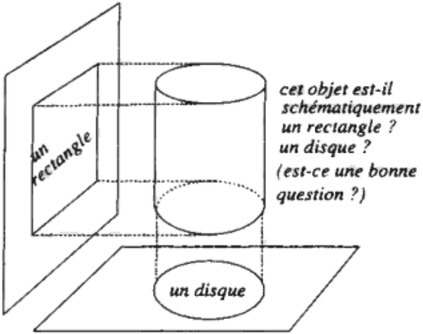
La ricerca degli assi

La distanza euclidea dall' origine di un punto proiettato è sempre minore o al massimo uguale a quello nello spazio originario (cateto ed ipotenusa di un triangolo rettangolo)



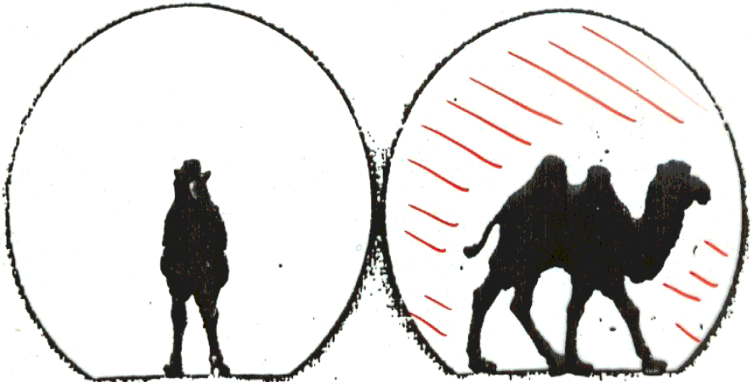
L'Analisi in Componenti Principali

WARNING !!!!



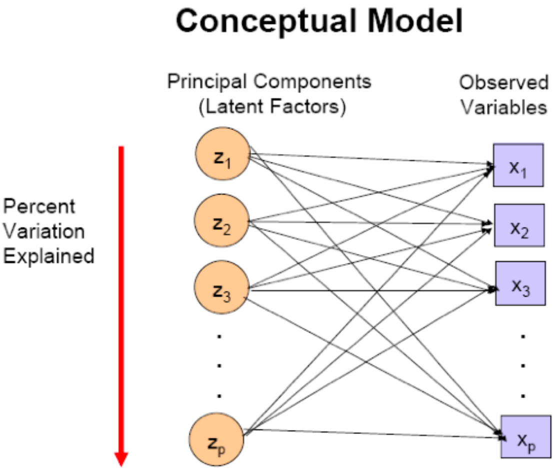
L'Analisi in Componenti Principali

WARNING !!!!

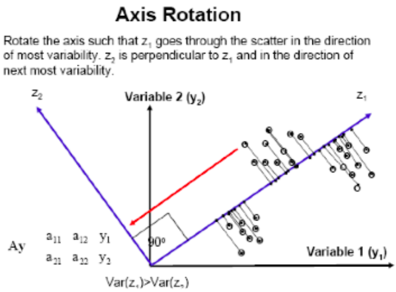


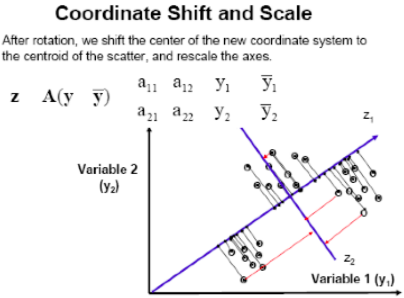
Immagini a due dimensioni da J.P.Fenelon (1981), *Qu'est ce que l'analyse des données ?*, Lefonen.

L'Analisi in Componenti Principali



L'Analisi in Componenti Principali

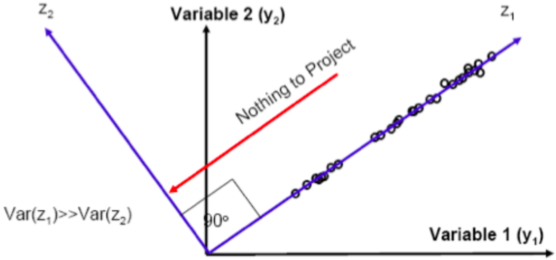




L'Analisi in Componenti Principali

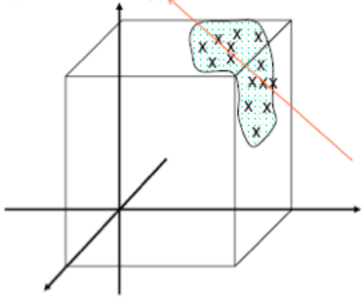
Collinear Data

If y_1 and y_2 are highly correlated (collinear) then the projection z_1 will describe all of the interesting structure in the scatter, leaving nothing for z_2.



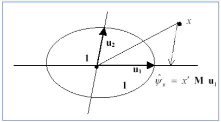
L'Analisi in Componenti Principali

- PCA may not be able to demonstrate all important structures in the original data set. Non-linear relationships (e.g. patterns of points that are arranged along one "wall" of a hyper-cube or which "drape" around a hyper-cube will not be easily seen.



La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ e l'Analisi in Componenti Principali

- Uno studio di dati multidimensionali nello spazio $\Re^p$ può essere definito, in termini notazionali, da quello che viene denominato **tripletta** statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ che contiene rispettivamente la *matrice dei dati*, la *metrica* (ossia il criterio scelto per il calcolo delle distanze fra le unità) e i *pesi* assegnati alle unità stesse.
- Il vettore $\mathbf{c}_1$ delle proiezioni degli n individui sull'asse $\mathbf{u}_1$ sarà dato da
$$\mathbf{c}_1 = \mathbf{XQu}_1$$
- Il vettore $\mathbf{u}_1$ sarà di norma $\mathbf{Q}$ unitaria: $\mathbf{u}_1^T\mathbf{Qu}_1 = 1$



La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- La quantità da massimizzare sarà data dalla somma dei quadrati delle proiezioni delle n unità, ciascuna ponderata rispetto alla massa p_i

$$\max \sum_{i=1}^{n} p_i c_i^2 = \max \mathbf{u}_1^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_1$$

con il vincolo $\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_1 = 1$

- Il Lagrangiano associato sarà

$$L = \mathbf{u}_1^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_1 - \lambda_1 (\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_1 - 1)$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- Derivando rispetto a $\mathbf{u}_1$ ed annullando le derivate parziali, si ottiene

$$\frac{\partial L}{\partial \mathbf{u}_1} = 2\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1 - 2\lambda_1\mathbf{Q}\mathbf{u}_1 = 0$$

  da cui

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1 = \lambda_1\mathbf{Q}\mathbf{u}_1$$

- Premoltiplicando i due membri per $\mathbf{u}_1^T$, si ottiene

$$\lambda_1 = \mathbf{u}_1^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- Premoltiplicando per $\mathbf{Q}^{-1}$ i due membri dell'identità $\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1 = \lambda_1\mathbf{Q}\mathbf{u}_1$, si ottiene

$$\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$$

  da cui deriva che la soluzione per il primo asse è data dall'autovettore $\mathbf{u}_1$ associato all'autovalore λ_1 più grande derivante alla diagonalizzazione della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$.
- La seconda componente principale è definita dalla combinazione lineare di varianza massima $\mathbf{c}_2 = \mathbf{X}\mathbf{Q}\mathbf{u}_2$ soggetta ai vincoli di normalità ($\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_2 = 1$) e di ortogonalità ($\mathbf{u}_1^T\mathbf{Q}\mathbf{u}_2 = 0$).

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Il Lagrangiano associato sarà

$$L_2 = \mathbf{u}_2^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_2 - \lambda_2 (\mathbf{u}_2^T \mathbf{Q} \mathbf{u}_2 - 1) - \mu (\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_2)$$

Derivando L_2 rispetto a $\mathbf{u}_2$ ed annullando le derivate parziali, si ottiene

$$\frac{\partial L_2}{\partial \mathbf{u}_2} = 2\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2 - 2\lambda_2 \mathbf{Q}\mathbf{u}_2 - \mu \mathbf{Q}\mathbf{u}_1 = 0$$

da cui

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2 = \lambda_2 \mathbf{Q}\mathbf{u}_2 + \frac{\mu}{2}\mathbf{Q}\mathbf{u}_1$$

Premoltiplicando i due membri per $\mathbf{u}_2^T$, si ottiene

$$\lambda_2 = \mathbf{u}_2^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_2$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Si noti che la condizione $\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_1 = 0$ equivale a quella di ortogonalità tra le componenti $\mathbf{c}_2$ e $\mathbf{c}_1$:

$$< \mathbf{c}_2, \mathbf{c}_1 >_{\mathbf{D}} = \mathbf{c}_2^T\mathbf{D}\mathbf{c}_1 = \mathbf{u}_2^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_1 = \mathbf{u}_2^T\lambda_1\mathbf{Q}\mathbf{u}_1 = \lambda_1\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_1 = 0$$

quindi premoltiplicando per $\mathbf{u}_1^T$ i due membri dell'identità

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2 = \lambda_2\mathbf{Q}\mathbf{u}_2 + \frac{\mu}{2}\mathbf{Q}\mathbf{u}_1$$

abbiamo che

$$\boxed{\underbrace{\mathbf{u}_1^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2}_{=0} = \lambda_2\underbrace{\mathbf{u}_1^T\mathbf{Q}\mathbf{u}_2}_{=0} + \frac{\mu}{2}\underbrace{\mathbf{u}_1^T\mathbf{Q}\mathbf{u}_1}_{=1}} \implies \mu = 0$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

quindi si perviene all'identità

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2 = \lambda_2\mathbf{Q}\mathbf{u}_2$$

Premoltiplicando per $\mathbf{Q}^{-1}$ i due membri dell'identità si ottiene

$$\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_2 = \lambda_2\mathbf{u}_2$$

da cui deriva che la soluzione per il secondo asse è data dall'autovettore $\mathbf{u}_2$ associato all'autovalore λ_2 più grande derivante alla diagonalizzazione della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$.

Le soluzioni per gli assi successivi sono ottenute in modo analogo e corrispondono agli autovettori associati agli autovalori della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$ ordinati in modo decrescente.

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$. Terza componente

La terza componente principale è definita dalla combinazione lineare $\mathbf{c}_3 = \mathbf{XQu}_3$ di varianza massima ($\mathbf{c}_3^T\mathbf{Dc}_3$), soggetta ai vincoli di normalità e ortogonalità:

$$\begin{cases} \max\limits_{\mathbf{u}_3} \mathbf{u}_3^T\mathbf{QX}^T\mathbf{DXQu}_3 \\ \mathbf{u}_3^T\mathbf{Qu}_3 = 1 \\ \mathbf{u}_2^T\mathbf{Qu}_2 = 1 \\ \mathbf{u}_1^T\mathbf{Qu}_1 = 1 \\ \mathbf{u}_1^T\mathbf{Qu}_3 = 0 \\ \mathbf{u}_2^T\mathbf{Qu}_3 = 0 \\ \mathbf{u}_1^T\mathbf{Qu}_2 = 0 \end{cases}$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Il Lagrangiano associato sarà

$$\begin{aligned} L_3 = \mathbf{u}_3^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3 &- \lambda_3(\mathbf{u}_3^T \mathbf{Q} \mathbf{u}_3 - 1) - \mu(\mathbf{u}_2^T \mathbf{Q} \mathbf{u}_2 - 1) \\ &- \gamma(\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_1 - 1) - \theta(\mathbf{u}_3^T \mathbf{Q} \mathbf{u}_1) \\ &- \delta(\mathbf{u}_3^T \mathbf{Q} \mathbf{u}_2) - \beta(\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_2) \end{aligned}$$

Derivando L_3 rispetto a $\mathbf{u}_3$ ed annullando le derivate parziali, si ottiene

$$\frac{\partial L_3}{\partial \mathbf{u}_3} = 2\mathbf{Q}\mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3 - 2\lambda_3 \mathbf{Q} \mathbf{u}_3 - \theta \mathbf{Q} \mathbf{u}_1 - \delta \mathbf{Q} \mathbf{u}_2 = 0$$

da cui

$$\mathbf{Q}\mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3 = \lambda_3 \mathbf{Q} \mathbf{u}_3 + \frac{\theta}{2} \mathbf{Q} \mathbf{u}_1 + \frac{\delta}{2} \mathbf{Q} \mathbf{u}_2$$

Premoltiplicando i membri dell'identità per $\mathbf{u}_3^T$, si ottiene

$$\lambda_3 = \mathbf{u}_3^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Si noti che le condizioni $\mathbf{u}_i^T \mathbf{Q} \mathbf{u}_j = 0$ (con $i \neq j$ e $i, j = 1, ...3$) equivalgono a quella di ortogonalità tra le rispettive componenti:

$$< \mathbf{c}_i, \mathbf{c}_j >_{\mathbf{D}} = \mathbf{c}_i^T \mathbf{D} \mathbf{c}_j = \mathbf{u}_i^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_j = \mathbf{u}_i^T \lambda_j \mathbf{Q} \mathbf{u}_j = \lambda_j \mathbf{u}_i^T \mathbf{Q} \mathbf{u}_j = 0$$

quindi premoltiplicando per $\mathbf{u}_1^T$ i due membri dell'identità

$$\mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3 = \lambda_3 \mathbf{Q} \mathbf{u}_3 + \frac{\theta}{2} \mathbf{Q} \mathbf{u}_1 + \frac{\delta}{2} \mathbf{Q} \mathbf{u}_2$$

abbiamo che

$$\boxed{\underbrace{\mathbf{u}_1^T \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{u}_3}_{=0} = \lambda_3 \underbrace{\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_3}_{=0} + \frac{\theta}{2} \underbrace{\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_1}_{=1} + \frac{\delta}{2} \underbrace{\mathbf{u}_1^T \mathbf{Q} \mathbf{u}_2}_{=0}} \implies \theta = 0$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Allo stesso modo, premoltiplicando per $\mathbf{u}_2^T$ i due membri dell'identità

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_3 = \lambda_3\mathbf{Q}\mathbf{u}_3 + \frac{\theta}{2}\mathbf{Q}\mathbf{u}_1 + \frac{\delta}{2}\mathbf{Q}\mathbf{u}_2$$

abbiamo che

$$\boxed{\underbrace{\mathbf{u}_2^T\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_3}_{=0} = \lambda_3\underbrace{\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_3}_{=0} + \frac{\theta}{2}\underbrace{\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_1}_{=0} + \frac{\delta}{2}\underbrace{\mathbf{u}_2^T\mathbf{Q}\mathbf{u}_2}_{=1}} \Longrightarrow \delta = 0$$

da cui

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_3 = \lambda_3\mathbf{Q}\mathbf{u}_3 + \underbrace{\frac{\theta}{2}\mathbf{Q}\mathbf{u}_1}_{=0} + \underbrace{\frac{\delta}{2}\mathbf{Q}\mathbf{u}_2}_{=0}$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

quindi si perviene all'identità

$$\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_3 = \lambda_3\mathbf{Q}\mathbf{u}_3$$

Premoltiplicando per $\mathbf{Q}^{-1}$ i due membri dell'identità si ottiene

$$\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_3 = \lambda_3\mathbf{u}_3$$

da cui deriva che la soluzione per il terzo asse è data dall'autovettore $\mathbf{u}_3$ associato all'autovalore λ_3 più grande derivante alla diagonalizzazione della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$.

Le soluzioni per gli assi successivi sono ottenute in modo analogo e corrispondono agli autovettori associati agli autovalori della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$ ordinati in modo decrescente.

Diagonalizzazione di un matrice non simmetrica

Risulta evidente che la matrice $\mathbf{X}^T\mathbf{DXQ}$ non è simmetrica.
E' possibile determinare le soluzioni **u** cercate tramite la diagonalizzazione di una matrice equivalente di tipo simmetrico.

$$\boxed{\mathbf{X}^T\mathbf{DXQu}} = \boxed{\lambda\mathbf{u}}$$

$$\mathbf{X}^T\mathbf{DX}\underbrace{\mathbf{Q}^{\frac{1}{2}}\mathbf{Q}^{\frac{1}{2}}}_{=\mathbf{Q}}\mathbf{u} = \lambda\mathbf{u}$$

$$\mathbf{Q}^{\frac{1}{2}}\mathbf{X}^T\mathbf{DXQ}^{\frac{1}{2}}\underbrace{\mathbf{Q}^{\frac{1}{2}}\mathbf{u}}_{=\mathbf{g}} = \lambda\underbrace{\mathbf{Q}^{\frac{1}{2}}\mathbf{u}}_{=\mathbf{g}}$$

$$\boxed{\mathbf{Q}^{\frac{1}{2}}\mathbf{X}^T\mathbf{DXQ}^{\frac{1}{2}}\mathbf{g}} = \boxed{\lambda\mathbf{g}}$$

Gli autovettori **u** si ottengono da quelli **g** tramite la relazione

$$\boxed{\mathbf{u} = \mathbf{Q}^{-\frac{1}{2}}\mathbf{g}}$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

Se indichiamo con $\mathbf{V} = \mathbf{X}^T\mathbf{D}\mathbf{X}$, allora l'autovettore $\mathbf{u}_1$ è associato al più grande autovalore λ_1 della matrice **VQ** ed è di norma **Q**-unitaria ($\mathbf{u}_1^T\mathbf{Q}\mathbf{u}_1 = 1$); risulta essere allora un elemento dello spazio $\Re^p$ (**Spazio delle variabili**) e viene identificato come il primo **asse principale.** Risulta inoltre

$$\boxed{\begin{aligned} \mathbf{V}\mathbf{Q}\mathbf{u}_1 &= \lambda_1 \mathbf{u}_1 \\ \mathbf{Q}\mathbf{V}\underbrace{\mathbf{Q}\mathbf{u}_1}_{=\mathbf{u}_1^*} &= \lambda_1 \underbrace{\mathbf{Q}\mathbf{u}_1}_{=\mathbf{u}_1^*} \\ \mathbf{Q}\mathbf{V}\mathbf{u}_1^* &= \lambda_1 \mathbf{u}_1^* \end{aligned}}$$

A ciascun asse **u** è associata la forma lineare $\mathbf{u}^* = \mathbf{Q}\mathbf{u}$, elemento di quello che viene definito il **duale** dello spazio delle variabili $\Re^p$, che costituisce il primo **fattore principale**

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- I fattori principali $\mathbf{u}^*$ sono quindi gli autovettori della matrice $\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}$ e sono elementi dello spazio $\Re^{p*}$ munito della metrica $\mathbf{Q}^{-1}$ tale che

$$\mathbf{u}_1^{*T}\mathbf{Q}^{-1}\mathbf{u}_1^* = 1$$

- Risulta evidente che quando $\mathbf{Q} = \mathbf{I}_p$ gli assi principali e i fattori principali coincidono.
- Le **componenti principali** sono le variabili $\mathbf{c}_\alpha$, elementi di $\Re^n$, definite a partire dai fattori principali

$$\mathbf{c}_\alpha = \mathbf{X}\mathbf{u}_\alpha^* = \mathbf{X}\mathbf{Q}\mathbf{u}_\alpha$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- Premoltiplicando per **X** i due membri dell'identità $\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{u}_\alpha^* = \lambda_1 \mathbf{u}_\alpha^*$, si ottiene

$$\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{u}_\alpha^* = \lambda_1 \mathbf{X}\mathbf{u}_\alpha^*$$

  da cui

$$\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{c}_\alpha = \lambda_1 \mathbf{c}_\alpha$$

- Le componenti principali $\mathbf{c}_\alpha$ risultano quindi essere a loro volta autovettori di una matrice che definisce le distanze fra le n unità statistiche. Risultano allora essere elementi dello spazio $\Re^n$ (**Spazio degli individui**).

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- Gli assi principali **u**, i fattori principali $\mathbf{u}^*$ e le componenti principali **c** risultano, quindi, essere associati alle relazioni:

$$\begin{array}{ccc}
\text{Assi principali } \mathbf{u} & \Longrightarrow & \mathbf{X}^T\mathbf{DXQu} = \lambda\mathbf{u} \\
\text{Fattori principali } \mathbf{u}^* = \mathbf{Qu} & \Longrightarrow & \mathbf{QX}^T\mathbf{DXu}^* = \lambda\mathbf{u}^* \\
\text{Componenti principali } \mathbf{c} = \mathbf{Xu}^* & \Longrightarrow & \mathbf{XQX}^T\mathbf{Dc} = \lambda\mathbf{c}
\end{array}$$

- Nel caso in cui si utilizzi la metrica $\mathbf{Q} = \mathbf{I}_p$, gli assi principali e i fattori principali coincidono e vengono indicati con **u**.

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- Se le componenti principali $\mathbf{v}_\alpha = \mathbf{c}_\alpha$ risultano essere autovettori della matrice $\mathbf{XQX}^T\mathbf{D}$ allora saranno necessariamente di norma unitaria nello spazio $\Re^n$

$$\mathbf{XQX}^T\mathbf{D}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha \text{ con } \mathbf{v}_\alpha^T\mathbf{D}\mathbf{v}_\alpha = 1$$

- Abbiamo quindi che la struttura della matrice **X** può essere studiata in due spazi

$$\boxed{\Re^p} \qquad \mathbf{X}^T\mathbf{DXQ}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha \text{ con } \mathbf{u}_\alpha^T\mathbf{Q}\mathbf{u}_\alpha = 1$$

$$\boxed{\Re^n} \qquad \mathbf{XQX}^T\mathbf{D}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha \text{ con } \mathbf{v}_\alpha^T\mathbf{D}\mathbf{v}_\alpha = 1$$

Quali sono le relazioni fra i due spazi ? Come è possibile passare da uno spazio all'altro e viceversa ?

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- Nello spazio $\Re^p$, dalla relazione $\mathbf{X}^T\mathbf{DXQu}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$ sappiamo che gli assi sono normati ad 1 ($\mathbf{u}_\alpha^T\mathbf{Qu}_\alpha = 1$) e le componenti principali $\mathbf{c}_\alpha = \mathbf{XQu}_\alpha$ sono normati a λ_α ($\mathbf{c}_\alpha^T\mathbf{Dc}_\alpha = \lambda_\alpha$)
- Nello spazio $\Re^n$ le componenti principali $\mathbf{c}_\alpha$ risultano essere invece di norma unitaria ($\mathbf{c}_\alpha^T\mathbf{Dc}_\alpha = 1$). Poichè sono quindi diversi indichiamo con $\mathbf{v}_\alpha$ la componente $\mathbf{c}_\alpha$ normata ad 1 (asse).
- Dato che $\mathbf{v}_\alpha$ e $\mathbf{c}_\alpha$ sono proporzionali avendo norme diverse, per quale valore di a è vera la relazione $\mathbf{v}_\alpha = a \times \mathbf{c}_\alpha$ tale che $\mathbf{v}_\alpha^T\mathbf{Dv}_\alpha = 1$?

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- Sappiamo che $\|\mathbf{c}_\alpha\|^2 = \mathbf{c}_\alpha^T \mathbf{D} \mathbf{c}_\alpha = \lambda_\alpha$, quindi $a = 1/\sqrt{\lambda_\alpha}$ sarà il valore che normalizza ad 1 il vettore $\mathbf{v}_\alpha$

$$\mathbf{v}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{c}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{XQu}_\alpha$$

  e tale che

$$\mathbf{v}_\alpha^T \mathbf{D} \mathbf{v}_\alpha = \frac{1}{\lambda_\alpha} \mathbf{u}_\alpha^T \mathbf{Q} \mathbf{X}^T \mathbf{DXQu}_\alpha = \frac{1}{\lambda_\alpha} \mathbf{c}_\alpha^T \mathbf{D} \mathbf{c}_\alpha = \frac{1}{\lambda_\alpha} \lambda_\alpha = 1$$

- La formula di transizione $\mathbf{v}_\alpha = \lambda_\alpha^{-1} \mathbf{XQu}_\alpha$ ci consente quindi di passare dallo spazio $\Re^p$ in quello $\Re^n$ $(\mathbf{u}_\alpha \longrightarrow \mathbf{v}_\alpha)$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- Abbiamo visto che la struttura della matrice $\mathbf{X}$, oltre in $\Re^p$ con $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{u}_\alpha = \lambda_\alpha \mathbf{u}_\alpha$, può essere studiata anche in $\Re^n$ trovando gli assi $\mathbf{v}_\alpha$

$$\mathbf{X}\mathbf{Q}\mathbf{X}^T\mathbf{D}\mathbf{v}_\alpha = \lambda_\alpha \mathbf{v}_\alpha \text{ con } \mathbf{v}_\alpha^T\mathbf{D}\mathbf{v}_\alpha = 1$$

- Premoltiplicando per $\mathbf{X}^T\mathbf{D}$ abbiamo

$$\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\underbrace{\mathbf{X}^T\mathbf{D}\mathbf{v}_\alpha}_{=\mathbf{z}_\alpha} = \lambda_\alpha \underbrace{\mathbf{X}^T\mathbf{D}\mathbf{v}_\alpha}_{=\mathbf{z}_\alpha}$$

  Dato che $\mathbf{z}_\alpha$ è autovettore della matrice $\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}$ come $\mathbf{u}_\alpha$ allora dovrà essere di norma unitaria ($\mathbf{z}_\alpha^T\mathbf{Q}\mathbf{z}_\alpha = 1$)

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- Ma la norma di $\mathbf{z}_\alpha$ risulta essere pari a

$$\|\mathbf{z}_\alpha\|^2_{\mathbf{Q}} = \mathbf{z}_\alpha^T \mathbf{Q} \mathbf{z}_\alpha = \mathbf{v}_\alpha^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{v}_\alpha = \lambda_\alpha$$

  i vettori $\mathbf{u}_\alpha$ (normato a 1) e $\mathbf{z}_\alpha$ (normato a λ_α) sono quindi in proporzione. Per quale valore di b è vera la relazione $\mathbf{u}_\alpha = b \times \mathbf{z}_\alpha$ tale che $\mathbf{u}_\alpha^T \mathbf{Q} \mathbf{u}_\alpha = 1$?
- Sappiamo che $\|\mathbf{z}_\alpha\|^2_{\mathbf{Q}} = \lambda_\alpha$, quindi $b = 1/\sqrt{\lambda_\alpha}$ sarà il valore che normalizza ad 1 il vettore $\mathbf{z}_\alpha$

$$\mathbf{u}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{z}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}} \mathbf{X}^T \mathbf{D} \mathbf{v}_\alpha$$

  e tale che

$$\mathbf{u}_\alpha^T \mathbf{Q} \mathbf{u}_\alpha = \frac{1}{\lambda_\alpha} \mathbf{v}_\alpha^T \mathbf{D} \mathbf{X} \mathbf{Q} \mathbf{X}^T \mathbf{D} \mathbf{v}_\alpha = \frac{1}{\lambda_\alpha} \mathbf{z}_\alpha^T \mathbf{Q} \mathbf{z}_\alpha = \frac{1}{\lambda_\alpha} \lambda_\alpha = 1$$

La tripletta statistica $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$: Formule di transizione

- La formula di transizione $\mathbf{u}_\alpha = \lambda_\alpha^{-1}\mathbf{X}^T\mathbf{D}\mathbf{v}_\alpha$ ci consente quindi di passare dallo spazio $\Re^n$ in quello $\Re^p$ ($\mathbf{v}_\alpha \longrightarrow \mathbf{u}_\alpha$)
- La formula di transizione $\mathbf{v}_\alpha = \lambda_\alpha^{-1}\mathbf{X}\mathbf{Q}\mathbf{u}_\alpha$ ci consente quindi di passare dallo spazio $\Re^p$ in quello $\Re^n$ ($\mathbf{u}_\alpha \longrightarrow \mathbf{v}_\alpha$)

Coordinate riga e colonna

- $\mathbf{c}_\alpha = \mathbf{X}\mathbf{Q}\mathbf{u}_\alpha = \lambda_\alpha\mathbf{v}_\alpha$ Componenti principali (coord. riga)
- $\mathbf{z}_\alpha = \mathbf{X}^T\mathbf{D}\mathbf{v}_\alpha = \lambda_\alpha\mathbf{u}_\alpha$ Coordinate colonna nonchè correlazioni fra le p variabili e l'asse $\mathbf{v}_\alpha$ dato che $\|\mathbf{v}_\alpha\|^2_{\mathbf{D}} = 1$ e le variabili sono in genere standardizzate $\mathbf{Q} = \mathbf{D}_{1/\sigma}$.

L'Analisi in Componenti Principali

- Le coordinate colonna $\mathbf{z}_\alpha$ e quindi le correlazioni fra le p variabili e l'asse $\mathbf{v}_\alpha$ ci aiutano ad apprezzare le distanze tra le variabili
- Siano $\hat{\mathbf{x}}_j$ e $\hat{\mathbf{x}}_{j'}$ due variabili centrate con $\mathbf{Q} = \mathbf{D}_{1/\sigma}$, allora

$$\rightarrow \quad d^2(0,j) = \|\hat{\mathbf{x}}_j\|^2_{\mathbf{D}} = \hat{\mathbf{x}}_j^T \mathbf{D} \hat{\mathbf{x}}_j = 1$$

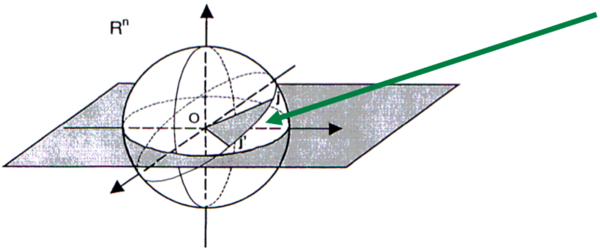
$$\begin{aligned} d^2(j,j') &= \|\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_{j'}\|^2_{\mathbf{D}} = (\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_{j'})^T \mathbf{D} (\hat{\mathbf{x}}_j - \hat{\mathbf{x}}_{j'}) \\ \rightarrow \quad &= \hat{\mathbf{x}}_j^T \mathbf{D} \hat{\mathbf{x}}_j + \hat{\mathbf{x}}_{j'}^T \mathbf{D} \hat{\mathbf{x}}_{j'} - 2\hat{\mathbf{x}}_j^T \mathbf{D} \hat{\mathbf{x}}_{j'} \\ &= \|\hat{\mathbf{x}}_j\|^2_{\mathbf{D}} + \|\hat{\mathbf{x}}_{j'}\|^2_{\mathbf{D}} - 2r(j,j') = 2(1 - r(j,j')) \end{aligned}$$

- $r(j,j') \simeq 1 \Longrightarrow d^2(j,j') \simeq 0$ punti variabili vicini
- $r(j,j') = 0 \Longrightarrow d^2(j,j') = 2$ punti a distanza media
- $r(j,j') \simeq -1 \Longrightarrow d^2(j,j') = 4$ punti molto distanti

L'Analisi in Componenti Principali

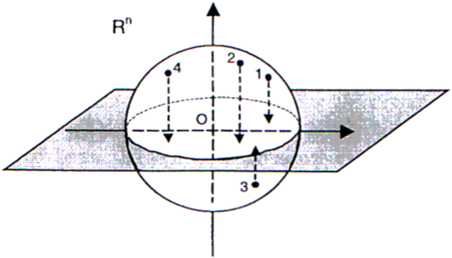
Nello spazio R^n la matrice **X** definisce una nube di p punti e tutti i punti-variabile si trovano su una ipersfera di raggio unitario.

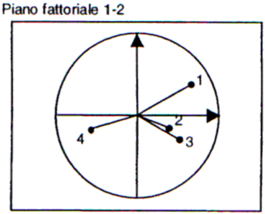
In R^p si osservano le distanze tra le unità, in R^n ci si intresserà agli angoli formati dai punti variabili (correlazioni)



L'Analisi in Componenti Principali

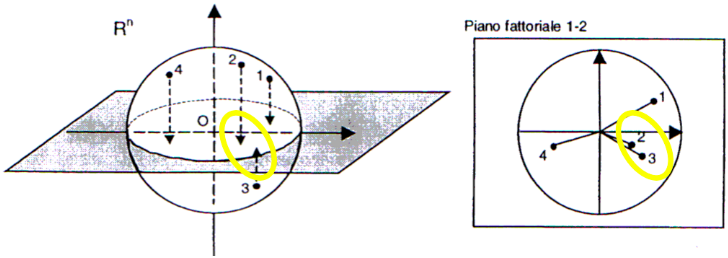
Nella costruzione del piano fattoriale (grafico), si taglia l'ipersfera secondo un cerchio di raggio unitario. Poiché l'operazione di proiezione tende a ridurre le distanze originarie, tutti i punti variabile si trovano all'interno del cerchio





L'Analisi in Componenti Principali

Nella costruzione del piano fattoriale (grafico), si taglia l'ipersfera secondo un cerchio di raggio unitario. Poiché l'operazione di proiezione tende a ridurre le distanze originarie, tutti i punti variabile si trovano all'interno del cerchio

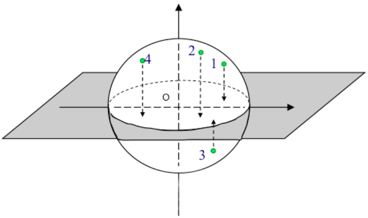


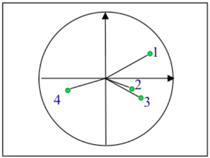
L'Analisi in Componenti Principali

Fattore "Taglia" (spesso l'asse v_1)

Se la maggior parte delle variabili è correlata positivamente tra loro - ciò equivale a dire che $x_{i1}, x_{i2}, \ldots, x_{ip}$ sono simultaneamente forti o simultaneamente deboli per tutti gli i

Spazio R^n : Proiezione delle 4 variabili

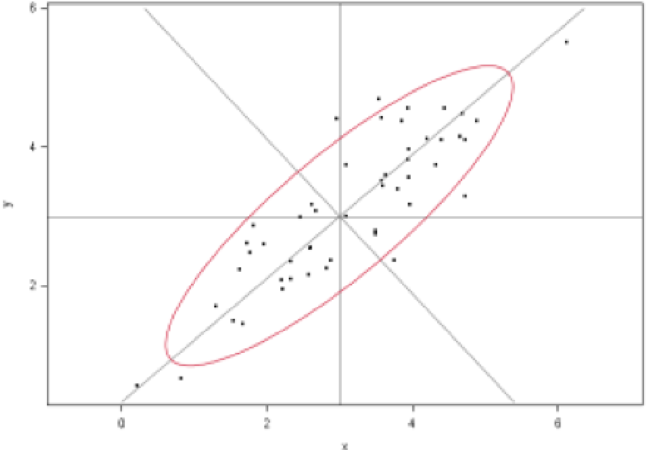




Piano Fattoriale

L'Analisi in Componenti Principali

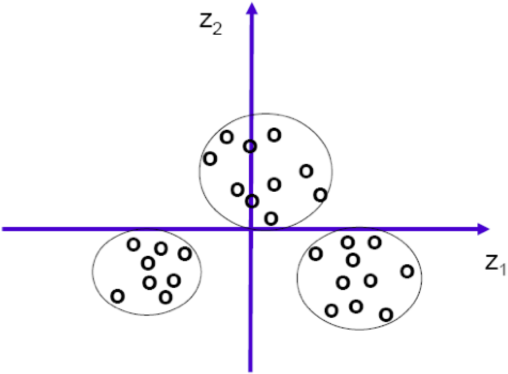
- Le unità statistiche sono contenute invece in un ellissoide di p dimensioni. Gli assi principali dell'ellissoide sono esattamente le rette $\mathbf{c}_\alpha$, cioè le componenti principali



L'Analisi in Componenti Principali

Clustering using PCA

Scatter plots of the first couple of principal components can be used to identify clusters of observations.



L'Analisi in Componenti Principali. Contributi assoluti

- Quali sono le unità che maggiormente hanno contribuito ad orientare gli assi ? Considereremo i contributi assoluti per ogni asse
- $\lambda_\alpha / \sum_{\alpha=1}^{n} \lambda_\alpha$ Peso dell'α-esima componente
- $CR(i) = p_i c_i^2 / \lambda_\alpha$ Contributo assoluto della i-esima unità alla α-esima componente con $\sum_i CR(i) = 1$ dove $\sum_i^n p_i c_i^2 = \mathbf{c}_\alpha^T \mathbf{D} \mathbf{c}_\alpha = \|\mathbf{c}_\alpha\|_{\mathbf{D}}^2 = \lambda_\alpha$.
- Date due unità i e i' con coordinate e pesi (c_i, p_i) e $(c_{i'}, p_{i'})$, potremo avere allora
  - $CR(i) > CR(i')$ se $c_i = c_{i'}$ e $p_i > p_{i'}$
  - $CR(i) > CR(i')$ se $c_i > c_{i'}$ e $p_i = p_{i'}$
  - $CR(i) = CR(i')$ se $c_i > c_{i'}$ e $p_i < p_{i'}$

L’Analisi in Componenti Principali. Contributi relativi

- Una misura della qualità della rappresentazione di un punto sul nuovo asse o sul nuovo piano può essere fornita dal quadrato del coseno dell’angolo formato dai vettori corrispondenti al punto nello spazio originario ed alla sua proiezione: quanto più questo valore si avvicina all’unità tanto più piccolo sarà l’angolongolo formato dai due vettori e quindi tanto migliore la rappresentazione

$$CR(i) = cos^2_{i,\alpha} = \|\mathbf{c}_{i,\alpha}\| / \|\mathbf{c}_i\|$$