Dispense sull'analisi delle componenti principali

Giorgio Marrubini e Camillo Melzi

Indice

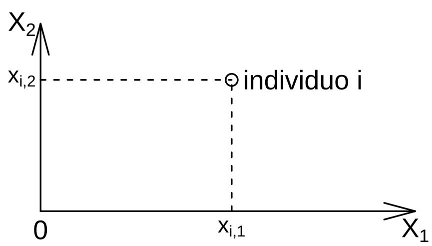
Capitolo 1

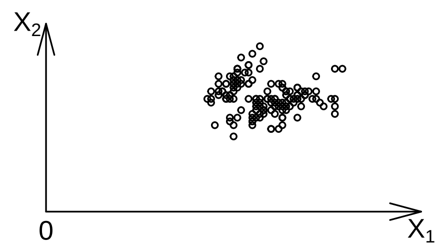
Dati multidimensionali

1.0.1 Rappresentazione matriciale e geometrica

Tabella 1.1: Rappresentazione matriciale

Indiv.	X_1	X_2	$\dots$	X_p
1	x_{11}	x_{12}	$\dots$	x_{1p}
2	x_{21}	x_{22}	$\dots$	x_{2p}
...				
m	x_{m1}	x_{m2}	$\dots$	x_{mp}





1.0.2 Trasformazione delle variabili: centratura e standardizzazione

Indichiamo con $\bar{x_1}, \ldots, \bar{x_p}$ le medie delle variabili $X_1, \ldots, X_p$, cioè le p medie delle p colonne della Tabella 1.1, e con $\sigma_1^2, \ldots, \sigma_p^2$ le rispettive varianze.
Il vettore $\bar{x} = (\bar{x_1}, \ldots, \bar{x_p})$ viene chiamato **baricentro**.

Centratura: semplice traslazione del baricentro nell'origine

$$x_{ij}^{'} = x_{ij} - \bar{x_j} \tag{1.1}$$

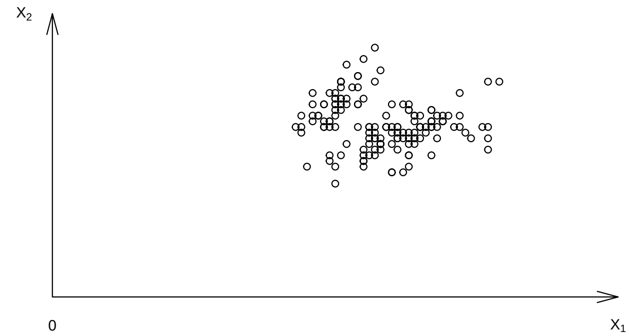
- non perdo informazione sulla distanza tra i punti (la geometria della nuvola di punti rimane invariata)
- perdo solo informazione sul baricentro
- semplifica formule e conti (da ora in poi useremo sempre dati centrati)

Standardizzazione: questa trasformazione porta ogni variabile ad avere varianza 1 (in generale questa trasformazione viene fatta insieme alla centratura)

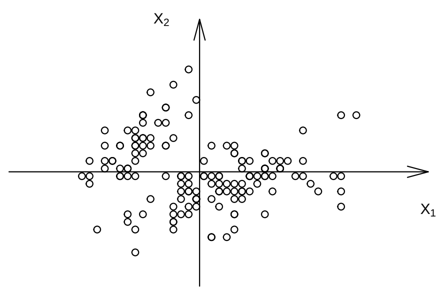
$$x_{ij}^{'} = \frac{x_{ij} - \bar{x_j}}{\sigma_j} \tag{1.2}$$

- questa trasformazione rende le variabili degli scalari (numeri puri)
- questa trasformazione è necessaria quando si vogliono confrontare variabili con differenti unità di misura (le variabili devono essere omogenee per essere confrontabili)
- tutte le variabili hanno lo stesso "peso"
- cambia la distanza (la geometria) tra i punti. E' una dilatazione o contrazione.

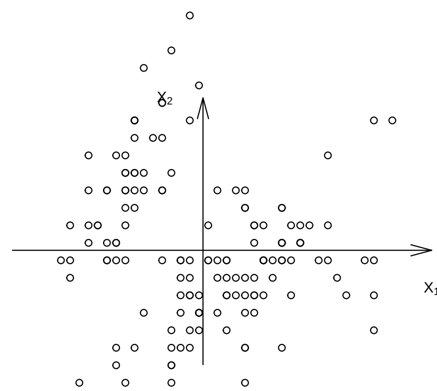
Si veda la seguente figura per una rappresentazione grafica di dati centarti e scalati per una matrice di dati di 2 variabili



Dati originali



Dati centrati



Dati centrati e scalati

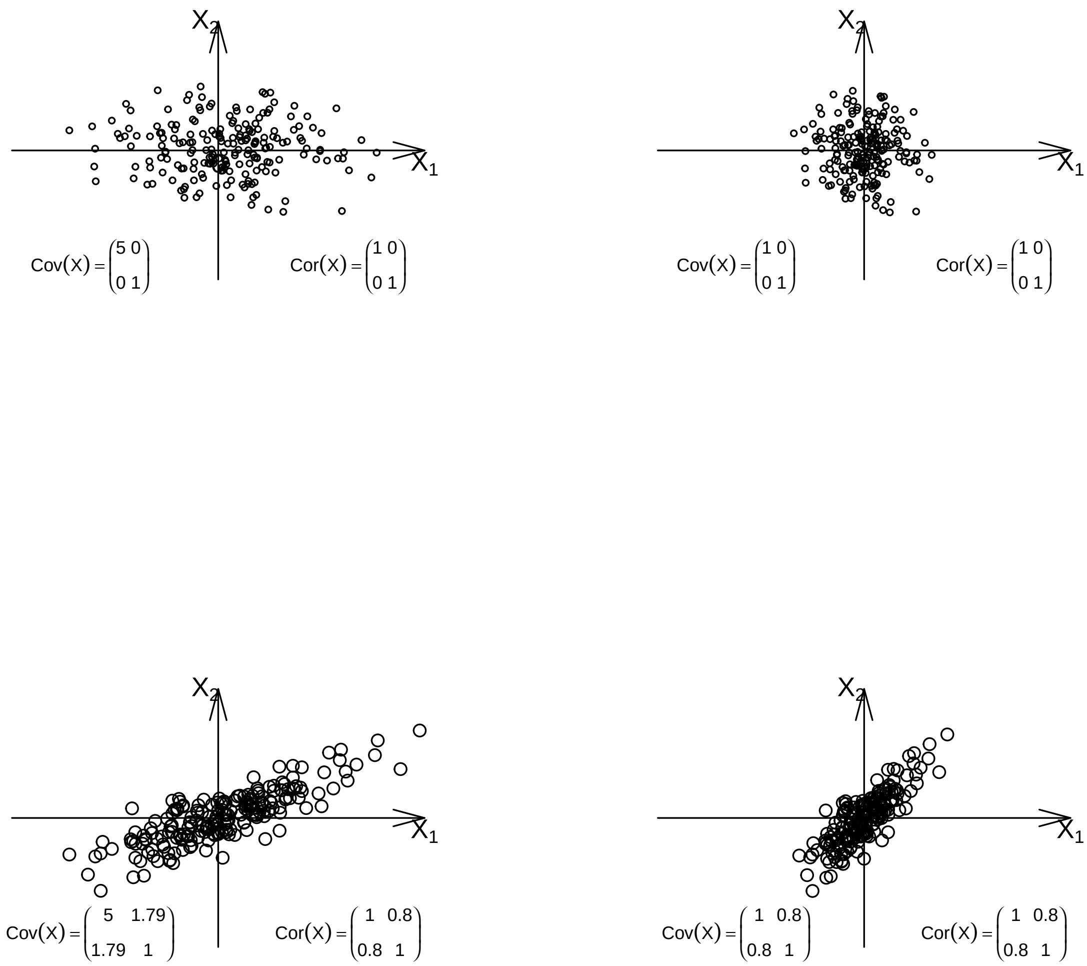
1.0.3 Matrice di covarianza e correlazione

$$Cov(X) = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1p} \\ \vdots & & \vdots \\ \sigma_{m1} & \dots & \sigma_{pp} \end{pmatrix}, \tag{1.3}$$

dove $\sigma_{ij} = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \bar{x_i})(x_{kj} - \bar{x_j})$ è la covarianza tra le variabili X_i e X_j, e in particolare $\sigma_{ii} = \sigma_i^2 = \frac{1}{m-1}\sum_{k=1}^{m}(x_{ki} - \bar{x_i})^2$ è la varianza della variabile X_i.
Nel caso in cui i dati siano centrati $Cov(X) = \frac{1}{m-1}X^tX$

$$Cor(X) = \begin{pmatrix} 1 & \dots & r_{1p} \\ \vdots & & \vdots \\ r_{m1} & \dots & 1 \end{pmatrix}, \tag{1.4}$$

dove $r_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$ è la correlazione tra le variabili X_i e X_j.



1.0.4 Variabili latenti o componenti e proiezioni

Sia T la combinazione lineare delle variabili $X_1, \ldots, X_p$, ossia il vettore (si veda Figura 1.1)

$$T = b_1 X_1 + \cdots + b_p X_p, \tag{1.5}$$

dove $b_1^2 + \cdots + b_p^2 = 1$. Il vettore $\mathbf{b} = (b_1, \ldots, b_p)$ è chiamato versore e indica la direzione della variabile latente T (si veda Figura 1.1).

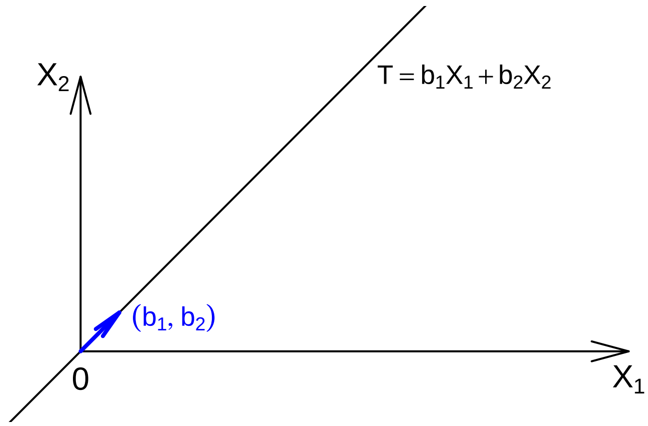


Figura 1.1: Variabile latente T

Sia $\mathbf{x} = (x_1, \dots, x_p)$ un generico punto (vettore) di $\mathbf{R^p}$. Chiamiamo proiezione di $\mathbf{x}$ su T il punto $\mathbf{x'}$ di T la cui distanza da $\mathbf{x}$ è minima (si veda Figura 1.2)

Definiamo *componente* di $\mathbf{x}$ su T la lunghezza del vettore $\|\mathbf{x'}\|$ data da

$$\|\mathbf{x'}\| = b_1 x_1 + \cdots + b_p x_p. \tag{1.6}$$

I valori $b_1, \dots, b_p$ sono chiamati *loading* e la quantità $b_1 x_1 + \cdots + b_p x_p$ *score*.

Si osservi che

$$\|\mathbf{x'}\| = \|\mathbf{x}\| \cos\theta \tag{1.7}$$

ossia al prodotto interno (scalare) tra i vettori $\mathbf{x}$ e $\mathbf{b}$ ($\|\mathbf{b}\| = 1$). Si veda la Figura 1.3.

Proiezione degli m individui della matrice $\mathbf{X}$ sulla variabile latente T

$$\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{m1} & \dots & x_{mp} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_m \end{pmatrix} = \begin{pmatrix} b_1 x_{11} + \cdots + b_p x_{1p} \\ \vdots \\ b_1 x_{m1} + \cdots + b_p x_{mp} \end{pmatrix}. \tag{1.8}$$

Supponiamo di prendere una seconda variabile latente

$$T' = b'_1 X_1 + \cdots + b'_p X_p, \qquad (b'_1)^2 + \cdots + (b'_p)^2 = 1 \tag{1.9}$$

e supponiamo che sia ortogonale a T (i.e **b** e **b'** ortogonali)

$$b_1 b'_1 + \cdots + b_p b'_p = 0. \tag{1.10}$$

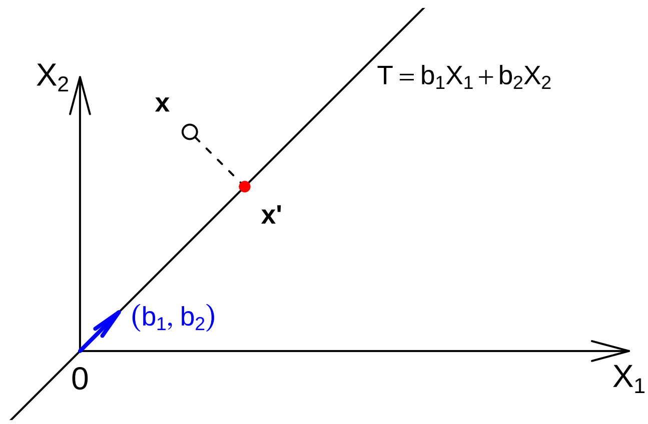


Figura 1.2: Proiezione su T

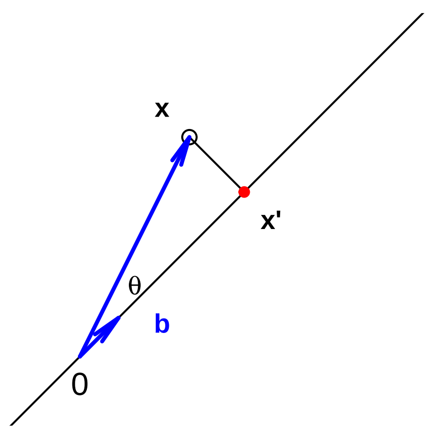


Figura 1.3: Prodotto interno tra **x** e **b**

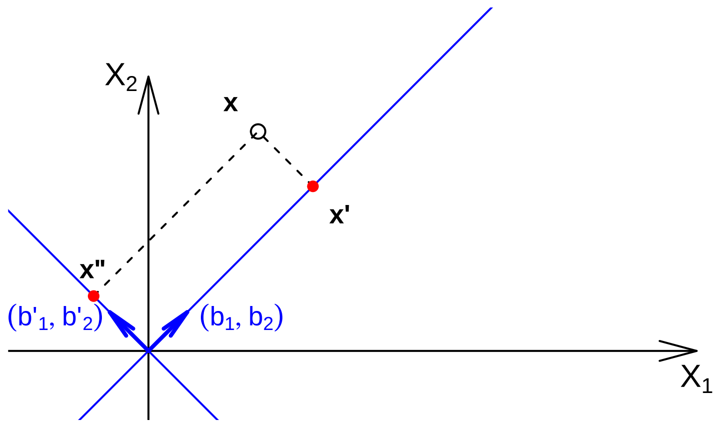


Figura 1.4: Proiezione sul piano TT'

Si veda la Figura 1.4.

Proiezione degli m individui della matrice $\mathbf{X}$ sul piano TT'

$$\begin{pmatrix} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{m1} & \dots & x_{mp} \end{pmatrix} \begin{pmatrix} b_1 & b'_1 \\ \vdots & \vdots \\ b_m & b'_p \end{pmatrix} = \begin{pmatrix} b_1x_{11} + \dots + b_px_{1p} & b'_1x_{11} + \dots + b'_px_{1p} \\ \vdots & \vdots \\ b_1x_{m1} + \dots + b_px_{mp} & b'_1x_{m1} + \dots + b'_px_{mp} \end{pmatrix}. \tag{1.11}$$

E' possibile iterare questo procedimento fino a p variabili latenti, in questo caso otteniamo un cambio di basi (nuove coordinate). Abbiamo semplicemente "cambiato prospettiva" ruotando il sistema di coordinate. Si veda la Figura 1.5.

E' possibile fermarsi prima e proiettare su un iperpiano,

Questo procedimento viene in generale eseguito perchè le variabili latenti hanno certe proprietà desiderate.

Indicando con

$$P = \begin{pmatrix} b_1^1 & b_1^2 & \dots & b_1^p \\ \vdots & \vdots & & \vdots \\ b_m^2 & b_m^2 & \dots & b_m^p \end{pmatrix} \tag{1.12}$$

la matrice dei *loading*, si ha

$$T = XP \tag{1.13}$$

e ricordando l'ortonormalità dei vettori $\mathbf{b}_1, \dots, \mathbf{b}_p$ ($P^tP = I$)

$$X = TP^t \tag{1.14}$$

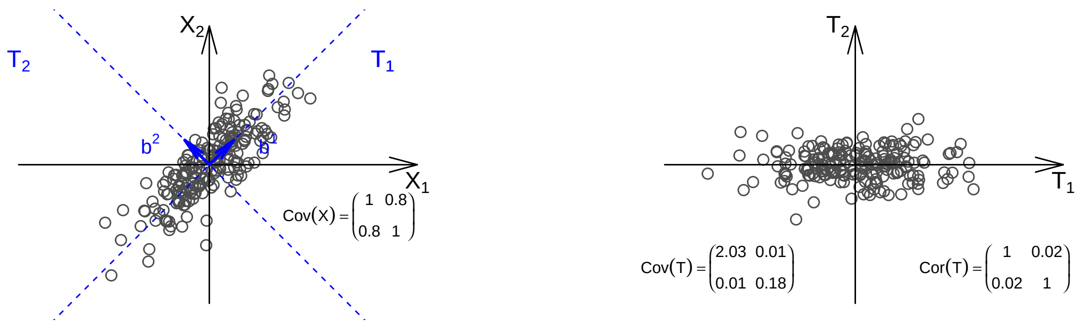


Figura 1.5: Cambio base da X_1X_2 a T_1T_2

```
P=matrix(c(1/sqrt(2),1/sqrt(2),-1/sqrt(2),1/sqrt(2)),ncol=2)
T=X%*%P
head(T)
```

```
##            [,1]         [,2]
## [1,] -0.8819306  0.469438535
## [2,] -1.1965330 -1.084705155
## [3,]  0.8871902  0.024327783
## [4,] -1.0638267  0.008888563
## [5,] -1.3798502 -0.102959280
## [6,] -1.3998573 -0.164815934
```

1.1 Analisi delle componenti principali

Vogliamo costruire le variabili latenti $T_1, \ldots, T_p$ in modo da massimalizzare la distanza tra gli m oggetti in $\mathbb{R}^p$, le cui coordinate sono date dalla matrice X (cf.), nel senso che punti lontani in $\mathbb{R}^p$ siano il più lontano possibile nelle proiezioni su T_1, poi T_2,...... La distanza tra i punti può essere misurata usando il teorema di Pitagora, distanza euclidea, e questa è la formula della varianza delle variabili $X_1, \ldots, X_p$.
Vogliamo massimalizzare la varianza, perchè ad essa è associata l'informazione contenuta nei dati in esame. In definitiva vogliamo massimalizzare l'informazione ricavabile dagli oggetti in esama (varianza).

E' posssibile determinare una variabile latente T_1, che chiameremo *Prima Componente Principale*, in modo tale che

$$Var(T_1) = \mathrm{Max}_T\, Var(T) \tag{1.15}$$

al variare di tutte le direzioni possibili T in $\mathbb{R}^p$.
Tra tutte le variabili latenti perpendicolari alla T_1 è possibile determinare una seconda variabile latente T_2, che chiameremo *Seconda Componente Principale* in modo tale che

$$Var(T_2) = \text{Max}_{T \perp T_1} Var(T) \tag{1.16}$$

Questo procedimento può essere iterato fino alla costruzione di p componenti principali $T_1, T_2, \dots, T_p$.

Per quanto visto nel Paragrafo 1.0.4 abbiamo determinato la matrice P dei *loading*. La matrice degli *score* si ottiene

$$T = XP \tag{1.17}$$

La procedura per determinare P passa attraverso il calcolo degli autovalori $\lambda_1, \dots, \lambda_p$ della matrice di covarianza (di correlazione nel caso in cui i dati fossero stati standardizzati)

$$Cov(X) = X^t X \tag{1.18}$$

e dei relativi autovettori (le p componenti principali).

Uno dei risultati principali di questa costruzione è che nel sistema di coordinate delle componenti principali

$$Cov(T) = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & \lambda_m \end{pmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \tag{1.19}$$

Conseguenze

- $Var(T_i) = \lambda_i$
- varianza totale: $\lambda_1 + \dots + \lambda_p$
- le componenti $T_1, T_2, \dots, T_p$ sono a indipendenti