Analisi in componenti principali

Carla Rampichini

rampichini@ds.unifi.it
Dipartimento di Statistica “G. Parenti”
Università di Firenze

Indice

Analisi in componenti principali

- Date p v.c. quantitative $\mathbf{x}' = [X_1, X_2, \ldots, X_p]$, con vettore delle medie $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$, si sostituisce alle variabili originali *correlate* un nuovo insieme di variabili, dette *componenti principali* che hanno le seguenti proprietà:
  - sono **incorrelate** (ortogonali)
  - sono elencate in ordine decrescente rispetto alla loro **varianza**
- La prima componente principale Y_1 è la *combinazione lineare* delle p variabili di partenza avente la *massima varianza*.
- La seconda componente principale Y_2 è la *combinazione lineare* delle p variabili con la varianza immediatamente inferiore alla varianza di Y_1 e ad essa *incorrelata*.
- ecc. fino alla p-esima componente.
- Se le p variabili originali sono molto correlate, un numero $q < p$ tiene conto di una quota elevata di varianza totale per cui le prime q componenti forniscono una buona descrizione della struttura dei dati.

Come si determina la prima CP?

Sarà necessario trovare il vettore di dimensione p dei coefficienti della combinazione lineare che rende massima la varianza di Y_1.

- $Y_1 = \mathbf{a}_1'\mathbf{x}$, $\mathbf{a}_1' = [a_{11}, \ldots, a_{1p}]$
- Y_1 deve avere varianza massima, cioè si dovrà trovare $\mathbf{a}_1$: $var(Y_1) = var(\mathbf{a}_1'\mathbf{x}) = \mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$ sia massima.
- Le soluzioni di questo problema di massimo sono infinite proporzionali perchè la combinazione lineare contiene un fattore di scala arbitrario.
- Si impone il *vincolo* che il vettore dei coefficienti abbia lunghezza unitaria: $\mathbf{a}_1'\mathbf{a}_1 = 1$
- Il problema di massimo vincolato si può risolvere attraverso il metodo dei moltiplicatori di Lagrange
  - si esprime il vincolo come eguaglianza a zero $\mathbf{a}_1'\mathbf{a}_1 - 1 = 0$
  - si calcolano le derivate prime parziali di $\mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a}_1 - 1)$
  - si eguagliano a zero le derivate prime.

Determinazione del massimo vincolato

$max_{\mathbf{a}_1}$: $\mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1$, s.t. $\mathbf{a}_1'\mathbf{a}_1 = 1$

$$\begin{aligned} \frac{\partial[\mathbf{a}_1'\mathbf{\Sigma}\mathbf{a}_1 - \lambda(\mathbf{a}_1'\mathbf{a}_1 - 1)]}{\partial \mathbf{a}_1} &= \mathbf{0} \\ 2\mathbf{\Sigma}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 &= 2[\mathbf{\Sigma} - \lambda\mathbf{I}]\mathbf{a}_1 = \mathbf{0} \\ (\mathbf{\Sigma} - \lambda\mathbf{I})\mathbf{a}_1 &= \mathbf{0} \end{aligned}$$

- Questo è un sistema lineare che ammette soluzioni non tutte nulle se il suo determinante è diverso da zero
- equazione caratteristica di $\mathbf{\Sigma}$: $\mid \mathbf{\Sigma} - \lambda\mathbf{I} \mid = 0$
- è un polinomio di grado p il λ, che ammette p soluzioni (non necessariamente distinte): $\lambda_1 \geq \ldots \geq \lambda_p$
- poichè la matrice di covarianza è definita semipositiva gli autovalori sono tutti non negativi.

La prima componente principale

- Si sceglie come prima CP la combinazione lineare $Y_1 = \mathbf{e}_1' \mathbf{x}$, dove $\mathbf{e}_1$ è l'autovettore associato al primo autovalore λ_1 di $\mathbf{\Sigma}$
- Ricordando che vale la seguente eguaglianza $\mathbf{\Sigma e}_1 = \lambda_1 \mathbf{e}_1$, si ha:

$$var(Y_1) = \mathbf{e}_1' \mathbf{\Sigma e}_1 = \mathbf{e}_1' \lambda_1 \mathbf{e}_1 = \lambda_1 \mathbf{e}_1' \mathbf{e}_1 = \lambda_1$$

La j-ma componente principale

- La j-ma componente principale del vettore casuale $x = [X_1, \ldots, X_p]$, la cui matrice di covarianza $\mathbf{\Sigma}$ ha autovalori-autovettori $(\lambda_j, \mathbf{e}_j)$, $j = 1, \ldots, p$, è data da: $Y_j = \mathbf{e}_j' \mathbf{x} = e_{j1} X_1 + \ldots e_{jp} X_p$,
- $Var(Y_j) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_j = \lambda_j$ (ricordare: $\mathbf{\Sigma} \mathbf{e}_j = \lambda_j \mathbf{e}_j$)
- $Covar(Y_j, Y_k) = \mathbf{e}_j' \mathbf{\Sigma} \mathbf{e}_k = \mathbf{e}_j' \lambda_k \mathbf{e}_k = \lambda_k \mathbf{e}_j' \mathbf{e}_k = 0$

Varianza totale

La varianza totale delle p componenti principali è uguale alla varianza totale delle $X_1, \ldots, X_p$:

$$\sum_{j=1}^{p} Var(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} Var(Y_j)$$

Infatti:

- $\sum_{j=1}^{p} \sigma_{jj} = tr(\mathbf{\Sigma})$
- poichè $\mathbf{\Sigma} = \mathbf{P\Lambda P'}$, con $\mathbf{\Lambda} = diag(\lambda_j)$ e $\mathbf{P} = [\mathbf{e}_1, \ldots, \mathbf{e}_p]$, si ha
- $tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P'})$
- ricordare $tr(\mathbf{AB}) = tr(\mathbf{BA})$, posto $\mathbf{A} = \mathbf{P}$ e $\mathbf{B} = \mathbf{\Lambda P'}$, si ha infine
- $tr(\mathbf{\Sigma}) = tr(\mathbf{P\Lambda P'}) = tr(\mathbf{\Lambda P'P}) = tr(\mathbf{\Lambda I}) = tr(\mathbf{\Lambda})$

Correlazione tra CP e X

La correlazione tra le CP e le X originali è data da

$$\rho(Y_j, X_k) = \frac{e_{jk}\sqrt{\lambda_j}}{\sqrt{\sigma_{kk}}}$$

Infatti:

- posto $\mathbf{a}'_k = [0, \ldots, 0, 1, 0, \ldots]$, $X_k = \mathbf{a}'_k \mathbf{x}$
- $Cov(X_k, Y_j) = Cov(\mathbf{a}'_k \mathbf{x}, \mathbf{e}'_j \mathbf{x}) = \mathbf{a}'_k \mathbf{\Sigma} \mathbf{e}_j$
- ricordando $\mathbf{\Sigma} \mathbf{e}_j = \lambda_j \mathbf{e}_j$ si ha
- $Cov(X_k, Y_j) = \mathbf{a}'_k \lambda_j \mathbf{e}_j = \lambda_j \mathbf{a}'_k \mathbf{e}_j = \lambda_j e_{jk}$
- poichè $Var(X_k) = \sigma_{kk}$ e $Var(Y_j) = \lambda_j$ si ottiene
- $\rho(Y_j, X_k) = \frac{Cov(X_k, Y_j)}{\sqrt{Var(X_k) Var(Y_j)}} = \frac{\lambda_j e_{jk}}{\sqrt{\sigma_{kk} \lambda_j}}$

Componenti principali campionarie

- date n osservazioni indipendenti $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ da una popolazione p-dimensionale con centroide μ e matrice di covarianza $\mathbf{\Sigma}$
- $\bar{\mathbf{x}}$ vettore delle medie campionarie
- $\mathbf{S}$ e $\mathbf{R}$ matrici di covarianza e correlazione campionarie

▶ le *componenti principali* campionarie sono definite come le *combinazioni lineari* delle $\mathbf{x}$ con la varianza campionaria massima.

obiettivo: trovare $k < p$ combinazioni lineari incorrelate che tengano conto il più possibile della variabilità presente nei dati.

Combinazioni lineari

Dati n valori campionari, per una qualsiasi combinazione lineare delle p variabili $\mathbf{x}$:

$$\mathbf{a}_1'\mathbf{x} = \sum_{j=1}^{p} a_{1j}x_{ij}, \;\; i = 1, \ldots, n$$

- la media campionaria della combinazione lineare è $\mathbf{a}_1'\overline{\mathbf{x}}$
- la varianza campionaria della combinazione lineare è $\mathbf{a}_1'\mathbf{S}\mathbf{a}_1$
- le due combinazioni lineari $(\mathbf{a}_1'\mathbf{x}, \mathbf{a}_2'\mathbf{x})$ hanno covarianza $\mathbf{a}_1'\mathbf{S}\mathbf{a}_2$

Come trovare le CP campionarie

- Matrice di covarianza campionaria **S**, con autovalori e autovettori $(\widehat{\lambda}_j, \widehat{\mathbf{e}}_j)$, $j = 1, \ldots, p$
- $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \ldots \geq \widehat{\lambda}_p \geq 0$
- la j-ma CP è

$$\widehat{y}_j = \widehat{\mathbf{e}}_j' \mathbf{x} = \widehat{e}_{j1} x_1 + \ldots + \widehat{e}_{jp} x_p$$

- $var(\widehat{y}_j) = \widehat{\lambda}_j$
- $covar(\widehat{y}_j, \widehat{y}_k) = 0$
- varianza campionaria totale $\sum_{j=1}^{p} s_{jj} = \sum_{j=1}^{p} \widehat{\lambda}_j$
- $corr(\widehat{y}_j, \mathbf{x}_k) = \frac{\widehat{e}_{jk} \sqrt{\widehat{\lambda}_j}}{\sqrt{s_{kk}}}$

Componenti principali campionarie: osservazioni

- Le CP campionarie possono essere ricavate da **S** o da **R**
- Le CP ricavate da **S** sono diverse dalle CP ricavate da **R**
- Di solito le **x** vengono *centrate*: $\mathbf{x} - \overline{\mathbf{x}}$
- La *centratura* non modifica **S** e quindi nemmeno autovalori e autovettori
- CP *centrate* $\widehat{y}_j = \mathbf{e}_j'(\mathbf{x} - \overline{\mathbf{x}})$
- In tal modo la media campionaria delle CP è zero. Infatti, considerando le n osservazioni campionarie $\mathbf{x}_i$ (vettore $p \times 1$ i-ma oss.), $i = 1, \ldots, n$,

$$\overline{\widehat{y}_j} = \frac{1}{n}\mathbf{e}_j' \left(\sum_{i=1}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}) \right) = \frac{1}{n}\mathbf{e}_j'\mathbf{0} = 0$$

CP campionarie nel caso di distribuzione normale

Se $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- le CP campionarie $\widehat{y}_j = \widehat{\mathbf{e}}_j'(\mathbf{x} - \overline{\mathbf{x}})$ sono *realizzazioni* delle CP della popolazione $Y_j = \mathbf{e}_j'(\mathbf{x} - \boldsymbol{\mu})$
- $\mathbf{Y} \sim N_p(\mathbf{0}, \boldsymbol{\Lambda})$, $\boldsymbol{\Lambda} = diag\{\lambda_j\}$, $j = 1, \ldots, p$
- Il profilo che si ottiene considerando per tutti i vettori $\mathbf{x}$ di dimensione $p \times 1$ che soddisfano

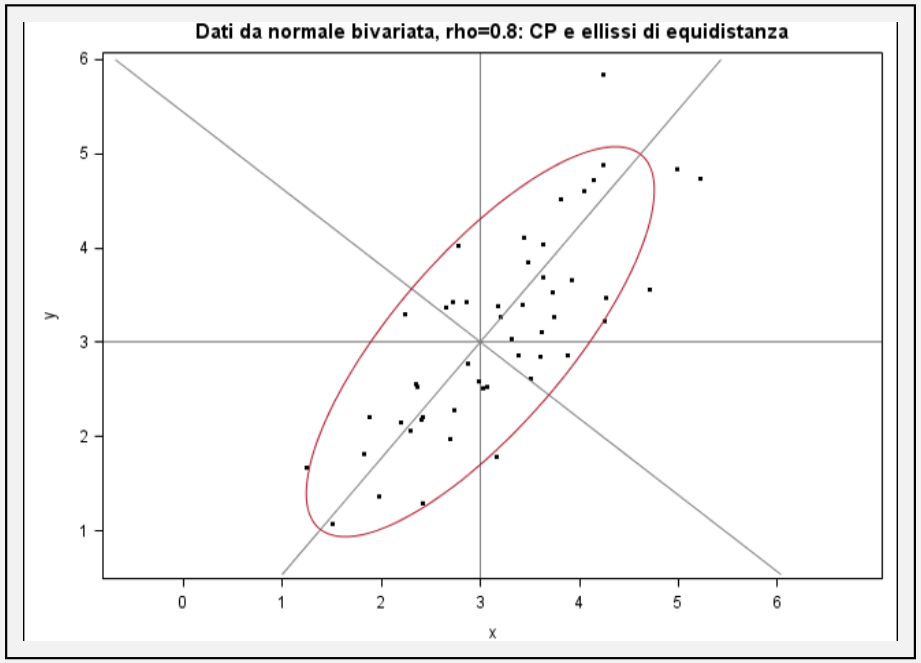
$$(\mathbf{x} - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) = c^2$$

  è una stima del profilo a densità costante $(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2$ della densità normale sottostante.
- I profili $(\mathbf{x} - \overline{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}}) = c^2$ sono ellissoidi centrati in $\overline{\mathbf{x}}$ e con assi pari agli autovettori di $\mathbf{S}^{-1}$ (equivalenti a quelli di $\mathbf{S}$)
- la lunghezza degli assi è proporzionale a $\sqrt{\widehat{\lambda}_j}$, ($\widehat{\lambda}_j$ autovalori di $\mathbf{S}$)

CP campionarie e distribuzione normale

- I profili $(\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) = c^2$ posssono essere sovrapposti allo scatterplot della nuvola di punti per evidenziare la distribuzione normale generatrice dei dati
- le CP campionarie $\widehat{\mathbf{y}}_j$ possono essere costruite anche quando la distribuzione delle $\mathbf{X}$ non è normale

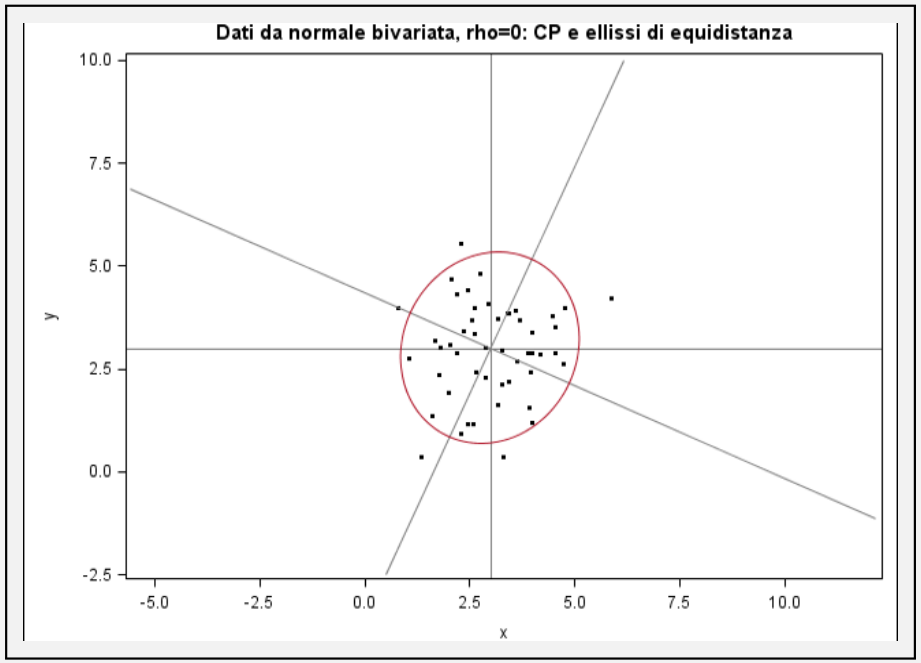
- Le CP sono le proiezioni ortogonali delle $\mathbf{x}$ centrate sugli autovettori $\widehat{\mathbf{e}}_j \Rightarrow$ (*i*) *traslazione* dell'origine degli assi in $\bar{\mathbf{x}}$; (*ii*) *rotazione* degli assi fino a passare per la nuvola di punti seguendo la direzione di massima varianza (gli assi dell'ellissoide)



CP campionarie per variabili incorrelate

- Se $\widehat{\lambda}_1 \doteq \widehat{\lambda}_2$ la direzione degli assi dell'ellissi non è univocamente determinata e gli assi possono giacere in qualsiasi direzione, compresa quella degli assi originali.

- quando il profilo di equidistanza è approx circolare, la variabilità campionaria è omogenea lungo qualsiasi direzione
- non è possibile rappresentare i dati in un spazio di dimensione ridotta



(dati e grafici generati con cp_bivariata.sas, SAS v9.2)

Standardizzazione delle variabili

In generale le CP campionarie *non sono invarianti* per trarsformazioni di scala.

- La standardizzazione delle variabili si ottiene come

$$\mathbf{z}_i = \mathbf{D}^{-1/2}(\mathbf{x}_i - \overline{\mathbf{x}}) = \left[\frac{x_{i1} - \overline{x}_1}{\sqrt{s_{11}}} \dots \frac{x_{ip} - \overline{x}_p}{\sqrt{s_{pp}}}\right]'$$

  $i = 1, \dots, n$, $\mathbf{D}^{-1/2} = diag\{\frac{1}{\sqrt{s_{jj}}}\}$, $j = 1, \dots, p$.

- La matrice dei dati standardizzati è data da $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]'$
- con vettore delle medie
  $\overline{\mathbf{z}} = \frac{1}{n}(\mathbf{1}'\mathbf{Z})' = \frac{1}{n}\mathbf{Z}'\mathbf{1} = \left[\sum_{i=1}^{n} \frac{x_{i1}-\overline{x}_1}{\sqrt{s_{11}}} \dots \sum_{i=1}^{n} \frac{x_{ip}-\overline{x}_p}{\sqrt{s_{pp}}}\right]'$
- e matrice di covarianza $\frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \mathbf{R}$

CP campionarie sulle variabili standardizzate

- La varianza totale delle variabili standardizzate è $tr(\mathbf{R}) = p = \sum_{j=1}^{p} \widehat{\lambda}_j$
- Autovalori e autovettori di **R** sono diversi da quelli di **S**
- $corr(\mathbf{z}_j, \widehat{\mathbf{y}}_k) = \widehat{e}_{jk}\sqrt{\widehat{\lambda}_k}$
- regola empirica: considerare solo le CP che spiegano una proporzione pari almeno a $1/p$ della varianza totale, cioè con $var(\widehat{y}_j) = \widehat{\lambda}_j > 1$