Analisi Multivariata

Corso di laurea in Statistica. A.A. 2008/09

Carla Rampichini

Alcune proprietà della distribuzione normale multivariata

Risultato 1

(a) Se $\mathbf{X}_1$ e $\mathbf{X}_2$ sono indipendenti, allora $\text{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, dove $\mathbf{0}$ è una matrice $(q_1 \times q_2)$ di zero.

(b) Se $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$ allora $\mathbf{X}_1$ e $\mathbf{X}_2$ sono indipendenti se e solo se $\boldsymbol{\Sigma}_{12} = \mathbf{0}$

(c) Se $\mathbf{X}_1 \sim N_{q_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ e $\mathbf{X}_2 \sim N_{q_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$ sono indipendenti, allora $\begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}$ ha una distribuzione normale multivariata:

$$N_{q_1+q_2} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

Risultato 2 **Elissoidi di isodensità.** Sia $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con $|\boldsymbol{\Sigma}| > 0$. Allora:

(a) $(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu})$ è distribuita come una v.c. χ_p^2 , dove χ_p^2 è la distribuzione chi-quadrato con p gradi di libertà.

(b) La distribuzione $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ assegna probabilità $1 - \alpha$ all'ellissoide solido $\{\mathbf{x} : (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)\}$, dove $\chi_p^2(\alpha)$ è il (100α) -esimo percentile destro della distribuzione χ_p^2 .

Dimostrazione. Sappiamo che χ_p^2 è definita come la distribuzione della somma $Z_1^2 + \dots + Z_p^2$, dove $Z_1, \dots, Z_p$ sono p v.c. indipendenti e identicamente distribuite con distribuzione $N(0, 1)$. Consideriamo la scomposizione

spettrale dell'inversa della matrice di covarianza: $\Sigma^{-1} = \sum_{j=1}^p \frac{1}{\lambda_j} \mathbf{e}_j \mathbf{e}_j'$, dove $\Sigma \mathbf{e}_j = \lambda_j \mathbf{e}_j$ e quindi $\Sigma^{-1} \mathbf{e}_j = (1/\lambda_j) \mathbf{e}_j$. Di conseguenza possiamo scrivere:

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) &= \sum_{j=1}^p (1/\lambda_j) (\mathbf{X} - \boldsymbol{\mu})' \mathbf{e}_j \mathbf{e}_j' (\mathbf{X} - \boldsymbol{\mu}) = \\ &= \sum_{j=1}^p (1/\lambda_j) (\mathbf{e}_j' (\mathbf{X} - \boldsymbol{\mu}))^2 = \sum_{j=1}^p [(1/\sqrt{\lambda_j}) (\mathbf{e}_j' (\mathbf{X} - \boldsymbol{\mu}))]^2 = \sum_{j=1}^p Z_j^2 \end{aligned}$$

Possiamo infatti scrivere $\mathbf{Z} = \mathbf{A}(\mathbf{X} - \boldsymbol{\mu})$, con

$$\mathbf{Z}_{(p \times 1)} = \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{bmatrix}, \mathbf{A}_{(p \times p)} = \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \mathbf{e}_1' \\ \frac{1}{\sqrt{\lambda_2}} \mathbf{e}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}} \mathbf{e}_p' \end{bmatrix}$$

e con $(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \Sigma)$, e quindi $\mathbf{Z} = \mathbf{A}(\mathbf{X} - \boldsymbol{\mu}) \sim N_p(\mathbf{0}, \mathbf{A}\Sigma\mathbf{A}')$, con

$$\begin{aligned} \mathbf{A}_{(p \times p)} \Sigma_{(p \times p)} \mathbf{A}'_{(p \times p)} &= \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \mathbf{e}_1' \\ \frac{1}{\sqrt{\lambda_2}} \mathbf{e}_2' \\ \vdots \\ \frac{1}{\sqrt{\lambda_p}} \mathbf{e}_p' \end{bmatrix} \left[\sum_{j=1}^p \lambda_j \mathbf{e}_j \mathbf{e}_j' \right] \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \mathbf{e}_1 | \frac{1}{\sqrt{\lambda_2}} \mathbf{e}_2 | \cdots | \frac{1}{\sqrt{\lambda_p}} \mathbf{e}_p \end{bmatrix} \\ &= \begin{bmatrix} \sqrt{\lambda_1} \mathbf{e}_1' \\ \sqrt{\lambda_2} \mathbf{e}_2' \\ \vdots \\ \sqrt{\lambda_p} \mathbf{e}_p' \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{\lambda_1}} \mathbf{e}_1 | \frac{1}{\sqrt{\lambda_2}} \mathbf{e}_2 | \cdots | \frac{1}{\sqrt{\lambda_p}} \mathbf{e}_p \end{bmatrix} \end{aligned}$$

In virtù del Risultato 1, $Z_1, Z_2, \dots, Z_p$ sono v.c. *indipendenti* con distribuzione normale standard, e possiamo quindi concludere che $(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$ ha una distribuzione χ_p^2 .

Per dimostrare il punto (b) possiamo osservare che $P[(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \leq c^2]$ è la probabilità assegnata all'ellissoide $(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu})$ dalla densità $N_p(\boldsymbol{\mu}, \Sigma)$. Ma dal punto (a) si ha che $P[(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \leq \chi_p^2(\alpha)] = 1 - \alpha$ e pertanto il punto (b) vale.

Osservazione (interpretazione della distanza statistica). Il Risultato 2 fornisce un'interpretazione della distanza statistica al quadrato. Quando $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$,

$$(\mathbf{X} - \boldsymbol{\mu})' \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) \tag{1}$$

è la distanza statistica al quadrato di $\mathbf{X}$ dal vettore delle medie $\boldsymbol{\mu}$ della popolazione. Se una componente ha una varianza molto più grande delle altre darà un contributo più piccolo alla distanza al quadrato. Inoltre, due componenti con un valore elevato della covarianza daranno un contributo inferiore rispetto a due componenti incorrelate. In sostanza, l'utilizzo dell'inversa della matrice di covarianza (detta anche matrice di precisione) nel calcolo della distanza:

1. standardizza tutte le variabili;
2. elimina gli effetti della correlazione.

Dalla dimostrazione del Risultato 2 si ha che

$$(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) = Z_1^2, Z_2^2, \dots, Z_p^2$$

Osserviamo infine che il vettore casuale¹ $\mathbf{Z} = \boldsymbol{\Sigma}^{-1/2} (\mathbf{X} - \boldsymbol{\mu})$ ha distribuzione normale multivariata $N_p(\mathbf{0}, \mathbf{I}_p)$ e:

$$\begin{aligned} (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) &= (\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} (\mathbf{X} - \boldsymbol{\mu}) \\ &= \mathbf{Z}' \mathbf{Z} = Z_1^2 + Z_2^2 + \dots, Z_p^2 \end{aligned}$$

Quindi la distanza statistica al quadrato (1) è calcolata come se: (i) il vettore casuale $\mathbf{X}$ fosse trasformato in p v.c. normali standardizzate indipendenti e (ii) quindi venisse calcolata la distanza al quadrato nel modo usuale, cioè come somma dei quadrati delle variabili.

¹Considerando $4\boldsymbol{\Sigma}^{-1/2}$, si ricordi che: $\boldsymbol{\Sigma}^{-1/2} \boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-1/2} = \sum_{j=1}^p (1/\sqrt{\lambda_j}) \mathbf{e}_j \mathbf{e}_j'$, con $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{1/2})^{-1}$