Analisi Multivariata
Corso di laurea in Statistica

Carla Rampichini

1 Distribuzione Normale multivariata

L'utilizzo di computer sempre più potenti consente oggi di considerare distribuzioni campionare approssimate diverse dalla Normale basate su tecniche di ricampionamento, tipo *bootstrap*, che sono utili in situazioni molto generali.

La distribuzione normale multuvariata gioca un ruolo fondamentale nella Statistica perchè può essere vista come approssimazione e distribuzione limite di molte altre distribuzioni multivariate. Questo risultato deriva sostanzialmente dal teorema limite centrale.

La distribuzione normale multivariata con vettore delle medie $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$ ha densità congiunta

$$f(x) = |2\pi\boldsymbol{\Sigma}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{1}$$

e si indica con $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Il vettore delle medie (centroide) e la matrice di covarianza di una distribuzione multivariata, forniscono molta informazione sulla relazione tra variabili, ma rappresentano solo una parte delle informazioni desumibili da una distribuzione multivariata. Una v.c. multivariata in generale è descritta dalla sua distribuzione congiunta e dalle distribuzioni marginali e condizionate da questa derivabili.

La relazione tra la distribuzione normale multivariata con media $\boldsymbol{\mu}$ e covarianza $\boldsymbol{\Sigma}$ e la distribuzione multinormale standard $N_p(\mathbf{0}, \mathbf{I}_p)$ si ricava attraverso una trasformazione lineare in base al teorema seguente.

Poichè $\boldsymbol{\Sigma}$ è semidefinita positiva allora possiamo scrivere $\boldsymbol{\Sigma} = \mathbf{E}\Lambda\mathbf{E}'$, dove Λ è la matrice diagonale degli autovalori di $\boldsymbol{\Sigma}$ e $\mathbf{E}$ è la matrice degli

autovettori ortonormali corrispondenti. La potenza α di $\mathbf{\Sigma}$ è definita come $\mathbf{\Sigma}^{\alpha} = \mathbf{E}\Lambda^{\alpha}\mathbf{E}'$. In particolare, $\mathbf{\Sigma}^{1/2}$ è la matrice semidifinita positiva unica tale che $\mathbf{\Sigma}^{1/2} = \mathbf{E}\Lambda^{1/2}\mathbf{E}'$, $\Lambda^{1/2} = diag\{\lambda^{1/2}\}$. Si veda l'Appendice 3.1 per maggiori dettagli sulle potenze di una matrice.

Teorema 1 *Sia* $\mathbf{X} \sim N_p(\boldsymbol{\mu}.\mathbf{\Sigma})$ *e* $\mathbf{Y} = \mathbf{\Sigma}^{-1/2}(\mathbf{X} - \boldsymbol{\mu})$ *(Trasformazione di Mahalanobis). Allora* $Y \sim N_p(\mathbf{0}, \mathbf{I}_p)$*, cioè le componenti* $Y_j \in \mathbb{R}$ *di* $\mathbf{Y}$ *sono v.c. univariate* $N(0,1)$ *indipendenti.*

Infatti $(\mathbf{X} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{Y}'\mathbf{Y}$, Inoltre $\mathcal{J} = \mathbf{\Sigma}^{-1}$, dove $\mathcal{J} = \frac{\partial x_i}{\partial y_j}$ è lo Jacobiano della trasformazione, da cui si ottiene $f_Y(\mathbf{y}) = (2\pi)^{-p/2} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{y}\right)$ che è la pdf di una $N_p(\mathbf{0}, \mathbf{I}_p)$.

Si noti che la trasformazione di Mahalanobis che compare nel teorema precedente porta alla v.c. $\mathbf{Y} = (Y_1, \ldots, Y_p)$ le cui componenti sono v.c univariate indipendenti e identicamente distribuite, $Y_j \sim N(0,1)$. Infatti:

$$\begin{aligned} f_Y(\mathbf{y}) &= \frac{1}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}\mathbf{y}'\mathbf{y}\right) \\ &= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y_j^2\right) \\ &= \prod_{j=1}^{p} f_{Y_j}(y_j) \end{aligned}$$

Come possiamo creare una $N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$ data la multinormale standard $\mathbf{Y} \sim N_p(\mathbf{0}, \mathbf{I}_p)$? Possiamo utilizzare la trasformazione lineare inversa:

$$\mathbf{X} = \mathbf{\Sigma}^{1/2}\mathbf{Y} + \boldsymbol{\mu}$$

Il teorema seguente è utile perchè fornisce la distribuzione congiunta di una v.c. multivariata dopo che è stata linearmente trasformata.

Teorema 2 *Sia* $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$*,* $\mathbf{A}$ *una matrice non singolare di dimensione* $p \times p$ *e* $\mathbf{c} \in \mathbb{R}^p$*. Allora* $\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{c}$ *è ancora una v.c. Normale p-variata, cioè* $Y \sim N_p(\mathbf{A}\boldsymbol{\mu} + \mathbf{c}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}')$*.*

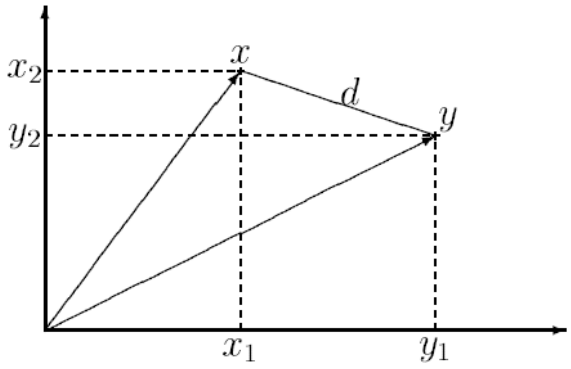

Figure 1: Distanza Euclidea d in $\mathbb{R}^2$, $\mathbf{A} = \mathbf{I}$

2 Geometria della Normale multivariata

Ricordiamo prima di tutto la definizione di distanza Euclidea d tra due punti $\mathbf{x}$ e $\mathbf{y}$ in $\mathbb{R}^p$

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})' \mathbf{A} (\mathbf{x} - \mathbf{y}) \tag{2}$$

dove $\mathbf{A}$ è una matrice definita positiva. $\mathbf{A}$ è detta *metrica*. Un caso particolare si ha quando $\mathbf{A} = \mathbf{I}$

$$d^2(x, y) = \sum_{j=1}^{p} (x_j - y_j)^2$$

La Figura 1 mostra un esempio di questa definizione in $\mathbb{R}^2$.

Si noti che gli insiemi $E_d = \{x \in \mathbb{R}^p : (\mathbf{x} - \mathbf{x}_0)'(\mathbf{x} - \mathbf{x}_0) = d^2\}$, sono le curve di iso-distanza dal punto $\mathbf{x}_0$, cioè le sfere di raggio d e centro $\mathbf{x}_0$. Un esempio per $p = 2$ è riportato in Fig. 2. La distanza più generale definita nella (2) ha curve di iso-distanza

$$E_d = \left\{ x \in \mathbb{R}^p : (\mathbf{x} - \mathbf{x}_0)' \mathbf{A} (\mathbf{x} - \mathbf{x}_0) = d^2 \right\}$$

cioè ellissoidi centrati in $\mathbf{x}_0$, come si vede in Fig. 3 nel caso di $p = 2$.

La densità congiunta della distribuzione $N_p(\boldsymbol{\mu}; \boldsymbol{\Sigma})$ è costante sugli elissoidi di forma $(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = d^2$, con centro in $\boldsymbol{\mu}$.

Indichiamo con $\gamma_1, \ldots, \gamma_p$ gli autovettori ortogonali e normalizzati (ortonormali) di $\boldsymbol{\Sigma}^{-1}$ corrispondenti agli autovalori $\nu_1, \ldots, \nu_p$. Allora

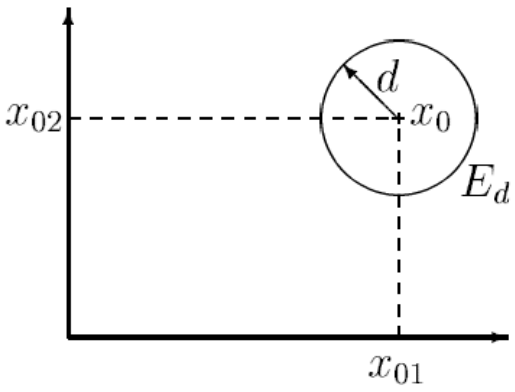

Figure 2: Sfera di iso-distanza

Teorema 3

(i) *Gli assi principali dell'elissoide E_d sono nella direzione degli autovettori γ_j, $j = 1, \ldots, p$.*

(ii) *La semi-lunghezza degli assi è data dagli autovalori $\sqrt{d^2/\nu_j}$, dove $\nu_j = \frac{1}{\lambda_j}$ sono gli autovalori di $\mathbf{\Sigma}^{-1}$ e λ_j sono gli autovalori di $\mathbf{\Sigma}$,$j = 1, \ldots, p$.*

(iii) *Per $p = 2$, il rettangolo circoscritto all'ellisse ha lati di lunghezza $2\sigma_{jj}$ e risulta quindi proporzionale alle deviazioni standard delle X_j, $j = 1, 2$.*

La distribuzione della forma quadratica che compare nella (1) è data dal teorema seguente.

Teorema 4 *Se $X \sim N_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, allora la variabile $U = (\mathbf{x} - \boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$ ha distribuzione χ^2 con p gradi di libertà.*

3 Alcune proprietà della distribuzione normale multivariata

La distribuzione Normale multivariata ha molte proprietà interessanti: è stabile ripetto a trasformazioni di tipo lineare, correlazione nulla corrisponde a

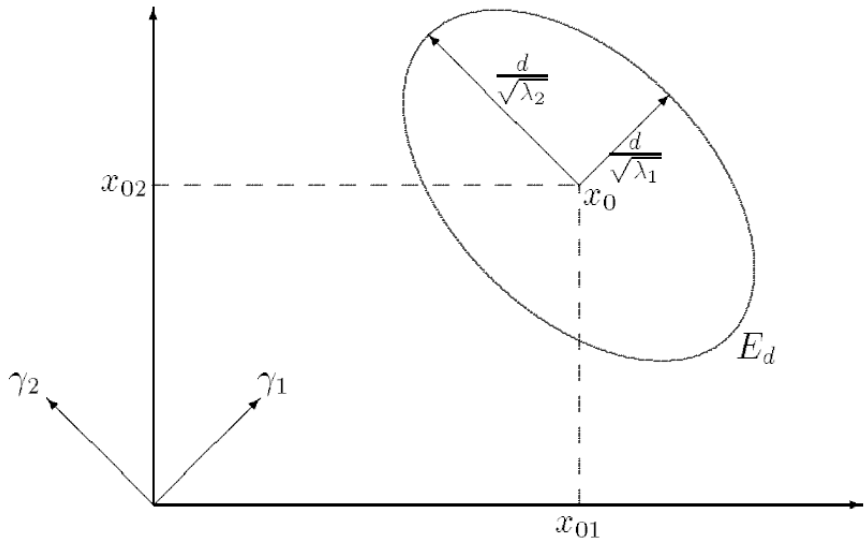

Figure 3: Ellisse di iso-distanza

indipendenza tra le componenti, le distribuzioni marginali e condizionate sono ancora distribuzioni normali, ecc. Le proprietà analitiche della normale multivariata sono tali da rendere più semplici le analisi condotte e la derivazione delle proprietà.

Risultato 1

(a) Se $\underset{(q_1\times 1)}{\mathbf{X}_1}$ e $\underset{(q_2\times 1)}{\mathbf{X}_2}$ sono indipendenti, allora $\mathrm{Cov}(\mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$, dove $\mathbf{0}$ è una matrice $(q_1 \times q_2)$ di zero.

(b) Se $\begin{bmatrix} \mathbf{X}_1 \\ -- \\ \mathbf{X}_2 \end{bmatrix} \sim N_{q_1+q_2}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & | & \boldsymbol{\Sigma}_{12} \\ - & - & - \\ \boldsymbol{\Sigma}_{21} & | & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$ 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:

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

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

(a) $(\mathbf{X} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ è distribuita come una v.c. χ^2_p, dove χ^2_p 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 $\{\mathbf{x} : (\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \leq \chi^2_p(\alpha)\}$, dove $\chi^2_p(\alpha)$ è il (100α)-esimo percentile destro della distribuzione χ^2_p.

Dimostrazione. Sappiamo che χ^2_p è definita come la distribuzione della somma $Z_1^2 + \ldots + Z_p^2$, dove $Z_1, \ldots, Z_p$ sono p v.c. indipendenti e identicamente distribuite con distribuzione $N(0, 1)$. Consideriamo la scomposizione spettrale dell'inversa della matrice di covarianza: $\boldsymbol{\Sigma}^{-1} = \sum_{j=1}^p \frac{1}{\lambda_j}\mathbf{e}_j\mathbf{e}_j'$, dove $\boldsymbol{\Sigma}\mathbf{e}_j = \lambda_j\mathbf{e}_j$ e quindi $\boldsymbol{\Sigma}^{-1}\mathbf{e}_j = (1/\lambda_j)\mathbf{e}_j$. Di conseguenza possiamo scrivere:

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

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

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

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

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

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

$$(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \tag{3}$$

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 contribuito 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.

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

ha[1] :

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

con $(\mathbf{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) = Z_1^2 + Z_2^2 + \ldots, +Z_p^2$. Quindi la distanza statistica al quadrato (3) 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.

3.1 Appendice: potenze di una matrice

Utilizzeremo spesso espressioni che coinvolgono la potenza di una matrice, per esempio $\mathbf{AA} = \mathbf{A}^2$. Per potenze intere positive queste espressioni possono essere svolte moltiplicando più volte una matrice per se stessa. Ma questo non mostra come sia possibile risolvere un problema del tipo: trovare la matrice $\mathbf{B}$ tale che $\mathbf{B}^2 = \mathbf{A}$, ossia come trovare la radice quadrata di una matrice.

Si consideri la matrice simmetrica definita positiva $\mathbf{A}$ di dimentione p. La soluzione a questo problema si trova utilizzando gli autovalori e gli autovettori di $\mathbf{A}$. Osserviamo prima di tutto che:

$$\begin{aligned}\mathbf{AA} = \mathbf{A}^2 &= (\mathbf{E}\Lambda\mathbf{E}')(\mathbf{E}\Lambda\mathbf{E}') \\ &= \mathbf{E}\Lambda\mathbf{E}'\mathbf{E}\Lambda\mathbf{E}' = \mathbf{E}\Lambda\mathbf{I}\Lambda\mathbf{E}' = \mathbf{E}\Lambda\Lambda\mathbf{E}' = \mathbf{E}\Lambda^2\mathbf{E}', \qquad (4)\end{aligned}$$

dove Λ è la matrice diagonale degli autovalori λ_j di $\mathbf{A}$ e $\mathbf{E}$ è la matrice degli autovettori ortonormali corrispondenti, $j = 1, \ldots, p$. Poichè Λ è una matrice diagonale i cui elementi diversi da zero sono pari a λ_j^2, vale il risultato seguente.

Risultato 3 *Per ogni matrice simmetrica* $\mathbf{A}$*, gli autovalori di* $\mathbf{A}^2$ *sono pari al quadrato degli autovalori di* $\mathbf{A}$*, e gli autovettori sono gli stessi.*

La dimostrazione di questo risultato si ottiene osservando che l'ultima eguaglianza che compare nella (4) è la scomposizione spettrale della matrice $\mathbf{B} = \mathbf{AA}$. Dato che $\mathbf{A}^3 = \mathbf{AA}^2$ e così via, il Risultato 3 si estende a ogni potenza intera positiva di $\mathbf{A}$. Per convenzione, vale $\mathbf{A}^0 = \mathbf{I}$. Pertanto,

[1] $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Sigma}^{-1/2} = \boldsymbol{\Sigma}^{-1}$, $\boldsymbol{\Sigma}^{-1/2} = \sum_{j=1}^{p}(1/\sqrt{(}\lambda_j)\mathbf{e}_j\mathbf{e}_j'$, con $\boldsymbol{\Sigma}^{-1/2} = (\boldsymbol{\Sigma}^{1/2})^{-1}$

Risultato 4 *Per ogni matrice simmetrica* $\mathbf{A}$*, si ha* $\mathbf{A}^\alpha = \mathbf{E}\Lambda^\alpha\mathbf{E}'$*,* $\alpha = 0, 1, \ldots$*. Gli autovalori di* $\mathbf{A}^\alpha$ *sono pari a* λ^α*, mentre gli autovettori sono gli stessi di* $\mathbf{A}$*.*

Il Risultato 3.1 si estende al caso di $\alpha \in \mathbb{R}$.

Risultato 5 *Data una matrice definita positiva* $\mathbf{A}$*,* $\mathbf{A}^\alpha = \mathbf{E}\Lambda^\alpha\mathbf{E}'$ *per* $\forall\alpha \in \mathbb{R}$*.*

Siamo ora in grado di ricavare la radice quadrata di una matrice definita positiva.

$$\mathbf{A}^{1/2} = \mathbf{E}\Lambda^{1/2}\mathbf{E}' = \mathbf{E}\begin{pmatrix} \sqrt{\lambda_1} & 0 & \ldots & 0 \\ 0 & \sqrt{\lambda_2} & \ldots & 0 \\ \vdots & \ldots & \ldots & \vdots \\ 0 & 0 & \ldots & \sqrt{\lambda_p} \end{pmatrix}\mathbf{E}'$$

Questa equazione soddisfa le caratteristiche di una radice quadrata, si ha infatti:

$$\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{E}\Lambda^{1/2}\mathbf{E}'\mathbf{E}\Lambda^{1/2}\mathbf{E}' = \mathbf{E}\Lambda\mathbf{E}' = \mathbf{A}$$

3.2 Esempio trasformazione sottovettori

Vogliamo ottenere due sottovettori indipendenti a partire da un vettore multinormale $\mathbf{X} = [X_1, X_2, X_3] \sim N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3]'$ e

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

Definiamo i due sottovettori $\mathbf{X_1} = [X_1, X_2]$ e $\mathbf{X_2} = [X_3]$, da cui si ottiene la matrice di covarianza partizionata $\begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$, dove $\boldsymbol{\Sigma}_{11} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$, $\boldsymbol{\Sigma}_{12} = [\sigma_{13}, \sigma_{23}]'$, $\boldsymbol{\Sigma}_{21} = [\sigma_{31}, \sigma_{23}]$ e $\boldsymbol{\Sigma}_{22} = \sigma_{33}$.

Osserviamo che $\mathbf{X_1} = [\mathbf{I}_2, \mathbf{0}]\mathbf{X} = \left(\begin{array}{cc|c} 1 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right)\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$ e

$$\mathbf{X}_{2.1} = [-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \mathbf{I}_{3-2}]\mathbf{X} = \left[(-\sigma_{31}, -\sigma_{33})\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix}, 1\right]\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix}$$

Pertanto il sottovettore $\mathbf{X}_{2.1}$ indipendente dal sottovettore $\mathbf{X}_1$ è una combinazione lineare delle tre componenti del vettore multinormale $\mathbf{X}$.

4 Verifica della normalità multivariata

I due problemi pù rilevanti nell'analisi dei dati multivariati sono l'individuazione di valori anomali (outliers) e la valutazione della normalità multivariata. I metodi per la verifica della normalità multivariata di un insieme di dati si basano sulle proprietà della distribuzione normale multivariata viste in precedenza. Il processo è complesso e di solito inizia con la verifica della normalità marginale, una variabile alla volta. Se si assume che le osservazioni campionarie siano indipendenti e identicamente distribuite, con distribuzione normale multivariata di media $\boldsymbol{\mu}$ e matrice di covarianza $\boldsymbol{\Sigma}$, per valutare la validità di questa ipotesi si procede di solito come segue.

1. Tutte le distribuzioni marginali devono essere normali, pertanto è necessario verificare la normalità per ciascuna delle variabili considerate. Se anche una sola variabile non è distribuita normalmente, allora l'insieme di variabili non può avere distribuzione normale multivariata.

   - Per valutare la normalità univariata in piccoli campioni, cioè per $n \leq 50$, si può utilizzare il test W di Shapiro e Wilk (1965).
   - Per campioni di dimensione maggiore, cioè quando $n > 50$, conviene utilizzare l'approssimazione proposta da Royston (1982, 1992) e implementata nella procedura UNIVARIATE di SAS.

   La normallità può essere verificata graficamente tramite il Q-Q plot. Si devono avere almeno 20 osservazioni per avere un grafico affidabile. Per costruire il Q-Q plot:

   (a) ordinare le osservazioni dalla più piccola alla più grande $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$, queste costituiscono le statistiche d'ordine della v.c. X e sono una stima dei quantili della distribuzione da cui sono campionate. Il quantile è il valore in corrispondenza del quale una certa proporzione della distribuzione è minore o uguale a quel valore.

(b) Stimare la proporzione della distribuzione che dovrebbe essere minore o uguale a ciascun valore della statsitica d'ordine, p.e. per l'i-ma osservazione nella successione ordinata tale proporzione è pari a

$$p_i = \frac{i - 1/2}{n}$$

dove i è il rango dell'osservazione.

(c) Calcolare il quantile atteso per la distribuzione Normale

$$q_i = \Phi^{-1}\left(\frac{i - 1/2}{n}\right)$$

dove Φ^{-1} è l'inversa della funzione di distribuzione della Normale standard.

(d) Rappresentare in un diagramma cartesiano i quantili osservati $X_{(i)}$ rispetto ai quantili attesi q_i e verificare la linearità del grafico: se la distribuzione osservata è normale i punti devono distribuirsi lungo una retta. Se i punti non si distribuiscno approssimativamente lungo una retta, l'ipotesi di normalità deve essere rifiutata.

2. Se le variabili non sono normali univariate, si può procedere ad una trasformazione utilizzando la trasformazione di Box e Cox (1964) o altre trasformazioni, tipo la logistica (si veda Sezione 4.1).

3. Tutte le coppie di variabili devono essere normali bivariate:

   - tracciare gli scatterplot per tutte le coppie di variabili
   - tracciare il *convex-hull*: le regioni di densità devono avere una forma approssimativamente ellittica con relazione lineare tra coppie le variabili .

4. Combinazioni lineari delle variabili devono essere normali. Verificare che somme e differenze delle variabili sono approssimativamente Normali univariate. Verificare che coppie di combinazioni lineari delle variabili sono ancora normali.

5. Le distanze al quadrato rispetto al vettore delle medie della popolazione sono distribuite come un chi-quadrato con p gradi di libertà. Stimare il

vettore delle medie della popolazione $\boldsymbol{\mu}$ con il vettore delle medie campionarie $\overline{\mathbf{x}}$, e la matrice di covarianza della popolazione $\boldsymbol{\Sigma}$ tramite la matrice di covarianza campionaria $\mathbf{S}$. Calcolare la distanza al quadrato di ogni osservazione rispetto al vettore delle medie campionarie e verificare se tali distanze seguono una distribuzione chi-quadrato con p gradi di libertà. Tale verifica può essere fatta coatruendo un Q-Q plot per la distribuzione chi-quadrato:

(a) Calcolare $d_i^2 = (\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \overline{\mathbf{x}})$

(b) Ordinare le distanze al quadrato d_i^2 dalla più piccola alla più grande per ottenere i quantili della distribuzione osservata: $d_{(1)}^2 \leq d_{(2)}^2 \leq \ldots \leq d_{(n)}^2$

(c) Calcolare i quantili attesi dalla distribuzione χ_p^2 corrispondenti a ogni distanza d_i^2, $i = 1, 2, \ldots, n$:

$$q_i = \chi_p^2 \left(\frac{i - 1/2}{n} \right)$$

(d) Rappresentare in un diagramma cartesiano d_i^2 rispetto a q_i per $i = 1, 2, \ldots, n$ e verificare se i punti si dispongono lungo una retta. Se i punti non formano approssimativamente una retta, i quantili osservati non vengono da una distribuzione chi-quadrato con p gradi idi libertà e l'ipotesi di normalità multivariata deve essere rifiutata.

Queste verifiche possono essere ripetute per sottoinsiemi e/o combinazioni lineari delle variabili coinvolte. Questo può essere utile per capire se la non normalità deriva da particolari sottoinsiemi di variabili o da una singola variabile.

Se la normalità (multivariata) viene rifiutata, allora procedere nel modo seguente.

1. Verificare che non ci siano outliers o errori nei dati. Se si trovano degli outliers, utilizzare metodi per il trattamento degli outliers per determinare il loro impatto sull'analisi. Si possono utilizzare anche metodi robusti. Ricordare che un valore anomalo può costituire l'informazione più rilevante nei dati in quanto è un dato diverso da tutti gli altri.

2. Considerare l'oportunità di procedere ad una trasformazione di una o più variabili. Una variabile per esempio può seguire una distribuzione log-normale, e quindi una trasformazione logaritmica può riportarla alla normalità. Si noti che tali trasformazioni modificano anche l'associazione della variabile in questione con le altre varabili.

3. Quando possibile, utilizzare metodi robusti o metodi alternativi, che non presuppongono la normalità della distribuzione. Alcune tecniche di analisi multivariata sono meno sensibili di altre alla presenza di valori anomali e alle ipotesi sulla distribuzione congiunta. Si veda per esempio Silverman (1986).

4. Per fare inferenza sui risultati utilizzare tecniche di ricampionamento (Monte Carlo, bootstrap, o metodi di permutazione, si veda, per esempio Manly, 1997).

4.1 Trasformazioni dei dati

- Matrice dei dati centrata
- Matrice degli scarti standardizzati
- Trasformazione di Box-Cox per la normalità
- Trasformazione logit

Bibliografia

Härdle W.K. , Simar L. (2007). *Applied Multivariate Statistical Analysis.* 2nd Edition, Springer.

Manly, B.F.J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*, second edition, Chapman & Hall, New York.

Johnson R.A. e Wichern D.W. (2007). *Applied Multivariate Statistical Analysis.* Sixth Edition. Pearson Education International.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, New York.