

Lezione : Approfondimenti ed esercizi

Docente: Aldo Solari

1 La matrice dei dati

Example 1.1. (a) Calcolare la traccia della matrice di centramento $H_{n \times n}$ (b) Calcolare $H \mathbf{1}_{n \times 1}$ (c) Si supponga che $a_{n \times 1}$ è un vettore i cui elementi sommano 0. Calcolare $H a_{n \times 1}$

Dimostrazione. (a)

$$\text{tr}(H) = \text{tr}\left(I - \frac{1}{n} \mathbf{1} \mathbf{1}'\right) = \text{tr}(I) - \frac{1}{n} \text{tr}(\mathbf{1} \mathbf{1}') = n - \frac{1}{n} n = n - 1$$

(b)

$$H \mathbf{1} = \left(I - \frac{1}{n} \mathbf{1} \mathbf{1}'\right) \mathbf{1} = \mathbf{1} - \frac{1}{n} \mathbf{1} \mathbf{1}' \mathbf{1} = \mathbf{1} - \frac{1}{n} n \mathbf{1} = \mathbf{0}_{n \times 1}$$

(c)

$$H a = \left(I - \frac{1}{n} \mathbf{1} \mathbf{1}'\right) a = a - \frac{1}{n} \mathbf{1} \mathbf{1}' a = a - \frac{1}{n} \mathbf{1} \sum_{i=1}^n a_i = a_{n \times 1}$$

□

Example 1.2. Sia $J_{n \times n} = \frac{1}{n} \mathbf{1} \mathbf{1}'$, quindi $H = I - J$.(a) Calcolare $J a_{n \times 1}$ per un generico vettore a .(b) Si dimostri che $J_{n \times n}$ è una matrice idempotente.

Dimostrazione. (a)

$$J a = \frac{1}{n} \mathbf{1} \mathbf{1}' a = \frac{1}{n} \mathbf{1} \sum_{i=1}^n a_i = \begin{bmatrix} \bar{a} \\ \bar{a} \\ \vdots \\ \bar{a} \end{bmatrix}$$

(b) J è una matrice simmetrica, i.e. $J' = \left(\frac{1}{n} \mathbf{1} \mathbf{1}'\right)' = \frac{1}{n} \mathbf{1} \mathbf{1}' = J$. Inoltre

$$J J = \frac{1}{n} \mathbf{1} \mathbf{1}' \frac{1}{n} \mathbf{1} \mathbf{1}' = \frac{1}{n^2} \mathbf{1} \mathbf{1}' \mathbf{1} \mathbf{1}' = \frac{1}{n^2} n \mathbf{1} \mathbf{1}' = \frac{1}{n} \mathbf{1} \mathbf{1}' = J$$

□

2 Analisi delle componenti principali

Proposition 2.1 (Caso particolare dell'analisi delle componenti principali). *Per casi particolari di matrici di varianze/covarianze S e di correlazione R , le componenti principali si possono esprimere in forme semplificate. Supponiamo che la varianze/covarianze sia $S = \text{diag}(s_{11}, \dots, s_{pp})$ con $s_{11} \geq \dots \geq s_{pp} > 0$.*

Gli autovalori di S sono la soluzione di

$$\det(S - \lambda I) = (s_{11} - \lambda)(s_{22} - \lambda) \cdots (s_{pp} - \lambda) = 0$$

quindi $\lambda_1 = s_{11}, \dots, \lambda_p = s_{pp}$.

Per determinare il j -simo autovettore di S bisogna risolvere

$$Sv_j = \lambda_j v_j$$

Si osservi che vale

$$\text{diag}(s_{11}, \dots, s_{pp})v_j^* = s_{jj}v_j^*$$

per $v_j^ = (v_{1j}^*, \dots, v_{pj}^*)'$ dove $v_{jj}^* = 1$ e $v_{kj}^* = 0$ per $k \neq j$.*

Segue che (s_{jj}, v_j^) è la j -sima coppia di autovalori-autovettori di S .*

Concludiamo che le p componenti principali corrispondono alle variabili originali, i.e. $y_j =$

$$\tilde{X}v_j^* = \tilde{x}_j, \text{ con matrice dei punteggi } Y = \tilde{X}V^* = \tilde{X}I = \tilde{X}.$$

Si osservi che in questo caso l'analisi delle componenti principali non comporta alcun vantaggio. Un risultato analogo si ottiene considerando la matrice dei dati standardizzati Z e la relativa matrice di correlazione $R = I$: abbiamo $Rv_j^ = 1v_j^*$ e quindi $(1, v_j^*)$ è la j -sima coppia di autovalori-autovettori di R . Segue che le p componenti principali y_j corrispondono alle variabili standardizzate z_j .*

Example 2.2. *Si consideri la seguente matrice dei dati*

$$X = \begin{bmatrix} x_1 & x_2 \end{bmatrix} = \begin{bmatrix} -3 & 1 \\ 1 & -0.5 \\ 0 & -1 \\ -1 & -0.5 \\ 3 & 1 \end{bmatrix}$$

Calcolare il vettore dei punteggi y_1 relativo alla prima componente principale basata sulla matrice di varianza/covarianza di X e la corrispondente proporzione di varianza spiegata.

Dimostrazione. Il vettore delle medie di X è $\bar{x} = (0, 0)'$, quindi $\tilde{X} = X$. Svolgendo i calcoli si ottiene

$$S = \frac{1}{n} \tilde{X}' \tilde{X} = \begin{bmatrix} 4 & 0 \\ 0 & 0.7 \end{bmatrix}$$

quindi sfruttando il risultato precedente gli autovalori di S sono $\lambda_1 = 4$ e $\lambda_2 = 0.7$ con rispettivi autovettori $v_1 = (1, 0)'$ e $v_2 = (0, 1)'$. I punteggi della prima componente principale sono $y_1 = \frac{x_1}{5 \times 1} = \frac{x_1}{5 \times 1}$ e la varianza spiegata è $\lambda_1/(s_{11} + s_{22}) = 4/(4 + 0.7) = 85.1\%$ \square

Proposition 2.3 (Caso particolare dell'analisi delle componenti principali). *Supponiamo che la matrice di varianze/covarianze sia*

$$S_{p \times p} = \begin{bmatrix} s & sr & \cdots & sr \\ sr & s & \cdots & sr \\ \cdots & \cdots & \cdots & \cdots \\ sr & sr & \cdots & s \end{bmatrix}$$

per $s > 0$ e $0 < r < 1$ e quindi la corrispondente matrice di correlazione risulta

$$R_{p \times p} = \begin{bmatrix} 1 & r & \cdots & r \\ r & 1 & \cdots & r \\ \cdots & \cdots & \cdots & \cdots \\ r & r & \cdots & 1 \end{bmatrix}$$

Questa matrice di correlazione descrive p variabili ugualmente correlate. Per $r > 0$, gli autovalori-autovettori di R risultano

$$\lambda_1 = 1 + (p-1)r, \quad v_1 = (1/\sqrt{p}, \dots, 1/\sqrt{p})'$$

e

$$\lambda_j = 1 - r, \quad v_j = \left(\underbrace{\frac{1}{\sqrt{(j-1)j}}, \dots, \frac{1}{\sqrt{(j-1)j}}}_{j-1}, \frac{-(j-1)}{\sqrt{(j-1)j}}, \underbrace{0, \dots, 0}_{p-j} \right)', \quad j = 2, \dots, p$$

La prima componente principale è proporzionale alla somma delle p variabili standardizzate:

$$y_1 = Zv_1 = \frac{1}{\sqrt{p}} \left(\sum_{j=1}^p z_{1j}, \dots, \sum_{j=1}^p z_{nj} \right)'$$

e spiega una proporzione di varianza pari a

$$\frac{\lambda_1}{p} = \frac{1 + (p-1)r}{p} = r + \frac{1-r}{p}$$

quindi $\lambda_1/p \approx r$ per r prossimo a 1 oppure p molto grande. Ad esempio, se $r = 0.8$ e $p = 5$, la prima componente principale spiega 84% della variabilità.

Example 2.4. Si supponga che la matrice dei dati X consista di due colonne x_1 e x_2 tali che $x_2 = 2x_1$. Determinare autovalori e autovettori della matrice di correlazione R di X . Qual è la percentuale di varianza spiegata dalla prima componente principale?

Dimostrazione. La matrice di correlazione è

$$R = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

di rango 1, quindi un autovalore deve essere pari a 0. Gli autovalori si possono ottenere risolvendo

$$0 = |R - \lambda I| = \begin{vmatrix} 1 - \lambda & 1 \\ 1 & 1 - \lambda \end{vmatrix} = (1 - \lambda)^2 - 1 = \lambda^2 - 2\lambda = \lambda(\lambda - 2)$$

quindi gli autovalori sono $\lambda_1 = 2$ e $\lambda_2 = 0$. I corrispondenti autovalori $v_1 = (v_{11}, v_{21})'$ e

$v_2 = (v_{12}, v_{22})'$ sono la soluzione di

$$\begin{bmatrix} 1 - \lambda & 1 \\ 1 & 1 - \lambda \end{bmatrix} \begin{bmatrix} v_{1j} \\ v_{2j} \end{bmatrix} = \begin{bmatrix} v_{1j} \\ v_{2j} \end{bmatrix} \lambda_j$$

Per $j = 1$ otteniamo $v_{11} + v_{21} = 2v_{11} = 2v_{21}$, quindi $v_{11} = v_{21}$, e considerando il vincolo di lunghezza unitaria $\|v_1\|_{2 \times 1}^2 = v_{11}^2 + v_{21}^2 = 1$, segue $v_{11} = \pm 1/\sqrt{2}$.

Per $j = 2$ otteniamo $v_{12} + v_{22} = 0$ e quindi $v_{12} = -v_{22}$. Considerando il vincolo di lunghezza unitaria $\|v_2\|_{2 \times 1}^2 = 1$ otteniamo $v_{12} = \pm 1/\sqrt{2}$. Riassumendo

$$v_1 = \pm \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}, \quad v_2 = \pm \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Si noti che il segno degli autovalori non è univocamente determinato. La percentuale di varianza spiegata dalla prima componente è $\lambda_1/p = 100\%$. \square

Example 2.5. Supponiamo di aver ortogonalizzato i dati attraverso la trasformazione di Mahalanobis, ottenendo così la matrice \tilde{Z} . Dire se potrebbe essere utile oppure no considerare l'analisi delle componenti principali sui dati ortogonalizzati Z , motivando la risposta.

Dimostrazione. No, non è utile. La motivazione è che la matrice di varianze/covarianze dei dati ortogonalizzati $\tilde{Z} = \tilde{X}S^{-1/2}$ è

$$S^{\tilde{Z}} = \frac{1}{n} \tilde{Z}' \tilde{Z} = \frac{1}{n} S^{-1/2} \tilde{X}' \tilde{X} S^{-1/2} = S^{-1/2} S S^{-1/2} = I$$

quindi le componenti principali coincidono con le variabili originali, i.e. $Y = \tilde{Z}I = \tilde{Z}$ \square

Example 2.6. Si consideri l'analisi delle componenti principali sui dati standardizzati $Z_{n \times 2}$ con la seguente matrice di correlazione

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix}$$

dove $0 < r < 1$. Si consideri ora una trasformazione di scala: $y_1 = c z_1$ e $y_2 = z_2$ per $c > 0$.

Determinare la matrice di varianze/covarianze di $Y_{n \times 2} = [y_1 \ y_2]$ e i relativi autovalori.

Dimostrazione. Abbiamo $Y = ZA'$ per

$$A' = \begin{bmatrix} c & 0 \\ 0 & 1 \end{bmatrix} = A$$

quindi

$$S^Y = ARA' = ARA = \begin{bmatrix} c^2 & cr \\ cr & 1 \end{bmatrix}$$

Gli autovalori si possono ottenere risolvendo

$$|S^Y - \lambda I| = \begin{vmatrix} c^2 - \lambda & cr \\ cr & 1 - \lambda \end{vmatrix} = 0$$

quindi $(c^2 - \lambda)(1 - \lambda) - c^2r^2 = \lambda^2 - \lambda(1 + c^2) + c^2(1 - r^2) = 0$ ha come soluzione

$$\lambda = \frac{(1 + c^2) \pm \sqrt{(1 + c^2)^2 - 4c^2(1 - r^2)}}{2} = \frac{(1 + c^2) \pm \sqrt{(1 - c^2)^2 - 4c^2r^2}}{2}$$

□