CdL in Scienze Statistiche ed Economiche - Università degli Studi di Milano-Bicocca

Esercitazione : Analisi delle componenti principali e similaritá

Esercitatrice: Laura Belloni

Esercizio 1.

1. Determinare le componenti principali relative alla matrice di varianze e covarianze

$$\Sigma = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix} \tag{1}$$

2. Si calcoli la proporzione di varianza spiegata dalla prima componente principale

Soluzione:
1. L'equazione caratteristica associata a Σ risulta $\lambda^2 - 7\lambda + 6 = 0$ e ha soluzione $\lambda_1 = 6$ e $\lambda_2 = 1$. Gli autovettori normalizzati risultano rispettivamente

$$v_1' = (\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}}) \qquad v_2' = (-\frac{1}{\sqrt{5}}, \frac{2}{\sqrt{5}})$$

La prima componente principale é pari a :

$$Y_1 = \tilde{X} v_1 = \frac{2}{\sqrt{5}} \tilde{X_1} + \frac{1}{\sqrt{5}} \tilde{X_2}$$

e la seconda componente principale:

$$Y_2 = \tilde{X} v_2 = -\frac{1}{\sqrt{5}} \tilde{X_1} + \frac{2}{\sqrt{5}} \tilde{X_2}$$

2. Si ottiene la seguente proporzione di varianza spiegata :

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.86 \tag{2}$$

Proprietá: Sia R la matrice di varianze e covarianze di Z, matrice dei dati standardizzati:

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{3}$$

Se $r > 0$ due autovalori di R sono

$$\lambda_1 = 1 + r \qquad \lambda_2 = 1 - r$$

I pesi associati sono pari a

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{4}$$

Infine i punteggi delle componenti principali risultano

$$y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} + z_{i,2}) \qquad y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} - z_{i,2})$$

Dimostrazione. Gli autovalori di R soddisfano l'equazione caratteristica $(1-\lambda)^2 - r^2 = 0$ da cui $\lambda_1 = 1 + r$ e $\lambda_2 = 1 - r$.
E' facile verificare che gli autovettori v_1 e v_2 normalizzati sono rapportabili a quanto enunciato da cui segue:

$$Y_1 = \tilde{Z}v_1 = \frac{1}{\sqrt{2}}\tilde{Z}_1 + \frac{1}{\sqrt{2}}\tilde{Z}_2$$
$$Y_2 = \tilde{X}v_2 = \frac{1}{\sqrt{2}}\tilde{Z}_1 - \frac{1}{\sqrt{2}}\tilde{Z}_2$$

Se $r < 0$ si inverte l'ordine degli autovalori e quindi delle componenti principali. □

Esercizio 2.
Convertire la matrice di varianze e covarianze dell'esercizio precedente nella matrice di correlazione ρ.
1. Determinare le componenti principali a partire da ρ e la proporzione di varianza spiegata dalla prima componente.
2. Confrontare i risultati con quelli ottenuti nell'esercizio precedente e motivare la risposta
3. Si calcoli la correlazione tra i punteggi Y_k k=1,2 e Z_j j=1,2.

Soluzione: 1. Considerata Σ si ottiene la seguente matrice di correlazione

$$\begin{bmatrix} 1 & \sqrt{\frac{2}{5}} \\ \sqrt{\frac{2}{5}} & 1 \end{bmatrix} \tag{5}$$

Per la proprietá sopra dimostrata risulta:

$$\lambda_1 = 1 + \sqrt{\frac{2}{5}} = 1.63 \qquad \lambda_2 = 1 - \sqrt{\frac{2}{5}} = 0.36$$

$$v_1' = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}) \qquad v_2' = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$$

$$y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} + z_{i,2}) \qquad y_{i,1} = \frac{1}{\sqrt{2}}(z_{i,1} - z_{i,2})$$

La varianza spiegata é pari a:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.81 \tag{6}$$

2. I risultati non coincidono in quanto l'analisi delle componenti principali non é invariante rispetto trasformazioni di scala pertanto quando si effettua l'analisi é necessario valutare se farla a partire da $\tilde{X}$ o da Z.

3. Ricordando che la correlazione tra Z_j e i punteggi Y_k é pari a $v_{j,k}\sqrt{\lambda_k}$
Segue:

$$cor_{Y_1,Z_1} = \frac{\sqrt{(}0.36)}{\sqrt{2}} \qquad cor_{Y_1,Z_2} = \frac{\sqrt{(}1.63)}{\sqrt{2}} \qquad cor_{Y_2,Z_1} = -\frac{\sqrt{(}0.36)}{\sqrt{2}}$$

Esercizio 3.
1. Data la matrice di dati si trasformino le variabili in variabili binarie e si calcolino il coefficiente di similaritá semplice e quello di Jaccard i presidenti Nixon e Johnson e tra Nixon e Kennedy.

Presidente	Luogo di Nascita	Eletto	Partito	Esperienze pregresse al congresso	Vicepresidente
Nixon	ovest	si	rep.	si	si
Kennedy	est	si	dem.	si	no
Johnson	sud	no	dem.	si	si

Soluzione:
Definendo

$$X_1 = \begin{cases} 1 \;\; se \;\; sud \\ 0 \;\; altrimenti \end{cases} \tag{7}$$

$$X_i = \begin{cases} 1 \;\; se \;\; si \\ 0 \;\; altrimenti \end{cases} \tag{8}$$

$$X_3 = \begin{cases} 1 \;\; se \;\; repubblicano \\ 0 \;\; altrimenti \end{cases} \tag{9}$$

Si ottiene

Presidente	X_1	X_2	X_3	X_4	X_5
Nixon	0	1	1	1	1
Kennedy	0	1	0	1	0
Johnson	0	0	0	1	0

I coefficienti di similaritá tra Nixon e Johnson valgono:

$$s_c = s_j = \frac{2}{5}$$

Mentre i coefficienti di similaritá semplice tra Nixon e Kennedy risultano:

$$s_c = \frac{3}{5} \qquad s_j = \frac{3}{4}$$