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

Esercitazione : Analisi delle Componenti Principali

Esercitatrice: Chiara Gaia Magnani

Example 0.1. *1. Determinare le componenti principali relative alla matrice di varianze e covarianze*

$$S = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

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

Dimostrazione. 1. L'equazione caratteristica associata a S 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 :

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.86 \tag{1}$$

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

$$R = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \tag{2}$$

Se $r > 0$ due autovalori di R sono

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

I pesi associati sono pari a

$$v_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} v_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \tag{3}$$

Infine i punteggi delle componenti principali risultano

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

□

Example 0.2. *Partendo dalla matrice di varianze e covarianze dell'esercizio precedente, calcolare nella matrice di correlazione R.*
1. Determinare le componenti principali a partire da R e la proporzione di varianza spiegata dalla prima componente.
2. Confrontare i risultati con quelli ottenuti nell'esercizio precedente e commentare la risposta
3. Si calcoli la correlazione tra il k-simo vettore dei punteggi $\underset{n\times 1}{y_k}$, per $k = 1, 2$, e la j-sima colonna di Z, i.e. $\underset{n\times 1}{z_j}$, per $j = 1, 2$.

Dimostrazione. 1. Partendo da S dell'esercizio precedente si ottiene la seguente matrice di correlazione

$$\begin{bmatrix} 1 & \sqrt{\frac{2}{5}} \\ \sqrt{\frac{2}{5}} & 1 \end{bmatrix} \tag{4}$$

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:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.81 \tag{5}$$

2. I risultati non coincidono in quanto l'analisi delle componenti principali non é invariante ri-

spetto 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 u_k é pari a $v_{j,k}\sqrt{\lambda_k}$
Segue:

$$cor_{y_1,z_1} = \frac{\sqrt{(}0.36)}{\sqrt{2}} \qquad cor_{y_1,z_2} = \frac{\sqrt{(}1.63)}{\sqrt{2}} \qquad cor_{y_2,z_1} = -\frac{\sqrt{(}0.36)}{\sqrt{2}}$$

□

Example 0.3. *Si supponga che la matrice dei dati X consista di due colonne $\underset{n\times 1}{x_1}$ e $\underset{n\times 1}{x_2}$ 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

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

quindi gli autovalori sono $\lambda_1 = 2$ e $\lambda_2 = 0$. I corrispondenti autovalori $\underset{2\times 1}{v_1} = (v_{11}, v_{21})'$ e $\underset{2\times 1}{v_2} = (v_{12}, v_{22})'$ sono la soluzione di

$$\begin{bmatrix} 1-\lambda_j & 1 \\ 1 & 1-\lambda_j \end{bmatrix} \begin{bmatrix} v_{1j} \\ v_{2j} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Per $j = 1$ otteniamo $v_{11} = v_{21}$, e considerando il vincolo di lunghezza unitaria $\| \underset{2\times 1}{v_1} \|^2 = v_{11}^2 + v_{21}^2 = 1$, segue $v_{11} = 1/\sqrt{2}$ (oppure $v_{11} = -1/\sqrt{2}$).
Per $j = 2$ otteniamo $v_{12} + v_{22} = 0$ e quindi $v_{12} = -v_{22}$. Considerando il vincolo di lunghezza unitaria $\| \underset{2\times 1}{v_2} \|^2 = 1$ otteniamo $v_{12} = 1/\sqrt{2}$ (oppure $v_{12} = -1/\sqrt{2}$). Riassumendo

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

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

Example 0.4. *Supponiamo che alla matrice dei dati* $\underset{n\times p}{X}$ *sia associata la varianze/covarianze* $S = diag(s_{11}, \ldots, s_{pp})$ *con* $s_{11} \geq \ldots \geq s_{pp} > 0$*. Per questa particolare matrice di varianze/covarianze, ha senso effettuare l'analisi delle componenti principali? E l'analisi delle componenti principali basata sulla corrispondente matrice di correlazione* R*? Giustificare le risposte.*

Dimostrazione. Gli autovalori di S sono la soluzione di

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

quindi $\lambda_1 = s_{11}, \ldots, \lambda_p = s_{pp}$.
Per determinare il j-simo autovettore di S bisogna risolvere

$$Sv_j = \lambda_j v_j$$

Si osservi che vale

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

per $\underset{p\times 1}{v_j^*} = (v_{1j}^*, \ldots, v_{pj}^*)'$ 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. $\underset{n\times 1}{y_j} = \tilde{X}v_j^* = \underset{n\times 1}{\tilde{x}_j}$, con matrice dei punteggi $\underset{n\times p}{Y} = \tilde{X}V^* = \tilde{X}I = \tilde{X}$.
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 0.5. *Supponiamo di aver ortogonalizzato i dati attraverso la trasformazione di Mahalanobis, ottenendo così la matrice* $\tilde{Z}$*. Ha senso effettuare l'analisi delle componenti principali sui dati ortogonalizzati* $\tilde{Z}$*? Giustificare 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

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

quindi le componenti calcolate su $\tilde{Z}$ principali coincidono con $\tilde{Z}$. □

Example 0.6. *Si consideri la seguente matrice dei dati*

$$X = \underset{5\times 1}{[x_1} \ \underset{5\times 1}{x_2}] = \begin{bmatrix} -3 & 1 \\ 1 & -0.5 \\ 0 & -1 \\ -1 & -0.5 \\ 3 & 1 \end{bmatrix}$$

Calcolare il vettore dei punteggi $\underset{5\times1}{y_1}$ *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 $\underset{2\times1}{\bar{x}} = (0,0)'$, quindi $X = \tilde{X}$. 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 $\underset{5\times1}{y_1} = \underset{5\times1}{x_1}$ e la varianza spiegata è $\lambda_1/(s_{11}+s_{22}) = 4/(4+0.7) = 85.1\%$. Infine i punteggi delle componenti principali risultano

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

□

Example 0.7. *Alla matrice dei dati standardizzati* $\underset{n\times p}{Z}$ *è associata la seguente matrice di correlazione:*

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

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

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

e

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

Scrivere l'equazione della prima componente principale di Z*. i.e.* $\underset{n\times1}{y_1} = Zv_1$ *e calcolare la proporzione di varianza spiegata. Per quali valori di* r *e* p *la proporzione di varianza spiegata dalla prima componente principale risulta elevata? Fornire un esempio numerico.*

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

$$\underset{n\times1}{y_1} = Zv_1 = \frac{1}{\sqrt{p}}(\sum_{j=1}^{p} z_{1j}, \ldots, \sum_{j=1}^{p} z_{nj})'$$

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 0.8. *Si consideri l'analisi delle componenti principali sui dati standardizzati* $\underset{n\times 2}{Z}$ *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:* $\underset{n\times 1}{y_1} = c\,\underset{n\times 1}{z_1}$ *e* $\underset{n\times 1}{y_2} = \underset{n\times 1}{z_2}$ *per* $c > 0$. *Determinare la matrice di varianze/covarianze di* $\underset{n\times 2}{Y} = [y_1\ y_2]$ *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

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

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

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

□

Example 0.9. *Data la matrice di correlazione*

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

dove r *è il coefficiente di correlazione tra le due variabili* X_1 *e* X_2, *si trovino le componenti principali. Qual è l'ordine delle componenti principali trovate?*

Dimostrazione. Le componenti principali sono le combinazioni lineari

$$Y_i = a_1^i X_1 + a_2^i X_2 = (\mathbf{a^i})'\mathbf{X}$$

con $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ che, sequenzialmente, massimizzano la varianza e sono tra loro incorrelate. Affinchè ciò accada il coefficiente $\mathbf{a^i}$ deve essere l'i-esimo autovettore normalizzato associato a $\mathbf{R}$. Dall'equazione caratteristica

$$(1-\lambda)^2 - r^2 = 0$$

si ricavano gli autovalori $\lambda_1 = 1 + r$ e $\lambda_2 = 1 - r$.
Per trovare i corrispondenti autovettori si risolve il sistema

$$\mathbf{R}\mathbf{a}^i = \lambda_1 \mathbf{a}^i, \;\; i = 1, 2.$$

Per i=1 risolvendo il seguente sistema

$$\begin{cases} a_1^1 + r a_2^1 = (1+r) a_1^1 \\ r a_1^1 + a_2^1 = (1+r) a_2^1 \end{cases}$$

sotto il vincolo di norma unitaria dell'autovettore, troviamo la soluzione

$$\mathbf{a}^1 = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

Procedendo in modo analogo per i=2, troviamo l'autovettore normalizzato

$$\mathbf{a}^2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{bmatrix}$$

Quindi, le due componenti principali sono

$$Y_1 = \frac{1}{\sqrt{2}}(X_1 + X_2) \;\; e \;\; Y_2 = \frac{1}{\sqrt{2}}(X_1 - X_2)$$

La varianza spiegata dalle due componenti principali è data dai rispettivi autovalori $\lambda_1 = 1 + r$ e $\lambda_2 = 1 - r$. Notiamo che esse dipendono dal parametro r, quindi l'ordine delle componenti principali può dipendere da r e, infatti, notiamo che se $r > 0$, $\lambda_1 > \lambda_2$, altrimenti accade il contrario. Se $r = 0$, gli autovalori sono entrambi pari a 1 e le variabili di partenza X_1 e X_2 sono le componenti principali, essendo tra loro incorrelate.

Observation 0.10. *Sappiamo che l'analisi delle componenti principali non è invariante per cambiamenti di scala, quindi la PCA condotta a partire dalle variabili di partenza moltiplicate per qualche coefficiente non darà gli stessi risultati della PCA condotta sulle variabili di partenza. Inoltre, se ci sono grandi differenze nelle varianze delle variabili di partenza, c'è il rischio che la PCA condotta sulla matrice di varianza e covarianza sia influenzata pesantemente dalle variabili ad alta varianza, che risulteranno molto probabilmente tra le prime componenti principali.*
Fare PCA a partire dalla matrice di correlazione equivale a fare PCA sulle variabili di partenza standardizzate (media 0 e varianza 1) e ciò permette di ovviare agli eventuali problemi sopra descritti.

□