

# **Analisi delle Componenti Principali**

## **Analisi Esplorativa**

Aldo Solari



- ① Trasformazioni lineari
- ② Analisi delle componenti principali
- ③ Interpretazione geometrica
- ④ PCA per dati standardizzati



# Riduzione della dimensionalità

$$\underset{n \times p}{X} \mapsto \underset{n \times q}{Y} \quad q \leq p$$

- Vogliamo che questa trasformazione preservi il più possibile la struttura dei dati originali
- Considereremo *trasformazioni lineari*



# Outline

- ① Trasformazioni lineari
- ② Analisi delle componenti principali
- ③ Interpretazione geometrica
- ④ PCA per dati standardizzati



# Trasformazioni lineari

La *trasformazione lineare* di  $X$   
 $n \times p$

$$Y = X A' + 1 b'$$

$n \times q$     $n \times p$   $p \times q$     $n \times 1$   $1 \times q$

è definita da

- la matrice  $A$   
 $q \times p$
- il vettore  $b$   
 $q \times 1$



# Trasformazioni lineari: vettore delle medie

Il vettore delle medie  $\bar{y}$  della trasformazione lineare

$$Y = X A' + \frac{1}{n} b' \quad \text{è dato da}$$

$n \times q$     $n \times p$   $p \times q$     $n \times 1$   $1 \times q$     $q \times 1$

$$\boxed{\bar{y}_{q \times 1} = A_{q \times p} \bar{x}_{p \times 1} + b_{q \times 1}}$$

*Dimostrazione*

$$\bar{y}_{q \times 1} = \frac{1}{n} Y' \frac{1}{n \times n \times 1} = \frac{1}{n} A_{q \times p} X \frac{1}{n \times p \times n \times 1} + \frac{1}{n} b_{q \times 1} \frac{1' \quad 1}{1 \times n \times 1} = A_{q \times p} \bar{x}_{p \times 1} + b_{q \times 1}$$



# Trasformazioni lineari: matrice di varianze/covarianze

La matrice di varianze/covarianze  $S^Y_{q \times q}$  della trasformazione lineare

$Y_{n \times q} = X_{n \times pp \times q} A' + 1_{n \times 1} b'$  è data da

$$S^Y_{q \times q} = A_{q \times pp \times pp \times q} S A'$$

*Dimostrazione:*

$$S^Y_{q \times q} = \frac{1}{n} \tilde{Y}'_{n \times q \times nn \times q} \tilde{Y}_{n \times q \times nn \times q} = \frac{1}{n} A_{q \times pp \times nn \times pp \times q} \tilde{X}'_{nn \times pp \times q} \tilde{X}_{nn \times pp \times q} A'_{pp \times q} = A_{q \times pp \times pp \times q} S_{pp \times pp \times pp \times q} A'_{pp \times q}$$

dove

$$\tilde{Y}_{n \times q} = H_{n \times nn \times q} Y_{n \times q} = H_{n \times nn \times q} X_{n \times nn \times pp \times q} A'_{pp \times q} + H_{n \times nn \times q} 1_{n \times 1} b'_{1 \times q} = H_{n \times nn \times q} X_{n \times nn \times pp \times q} A'_{pp \times q} = \tilde{X}_{n \times pp \times q} A'_{pp \times q}$$



# Trasformazioni lineari note

$$\begin{array}{ccccc} q & A & b & Y & = X A' + 1 b' \\ & q \times p & q \times 1 & n \times q & n \times p \quad p \times q \quad n \times 1 \quad 1 \times q \end{array}$$


---

$$\begin{array}{ccccc} p & I & - \bar{x} & & \tilde{X} \\ & p \times p & p \times 1 & & n \times p \end{array}$$

$$\begin{array}{ccccc} p & D^{-1/2} & - D^{-1/2} \bar{x} & & Z \\ & p \times p & p \times p & p \times 1 & n \times p \end{array}$$

$$\begin{array}{ccccc} p & S^{-1/2} & - S^{-1/2} \bar{x} & & \tilde{Z} \\ & p \times p & p \times p & p \times 1 & n \times p \end{array}$$


---





# Combinazioni lineari

La *combinazione lineare* di  $X_{n \times p}$

$$y_{n \times 1} = X_{n \times p} a_{p \times 1} = \begin{bmatrix} \sum_{j=1}^p a_j x_{1j} \\ \vdots \\ \sum_{j=1}^p a_j x_{ij} \\ \vdots \\ \sum_{j=1}^p a_j x_{nj} \end{bmatrix}$$

è un caso particolare di trasformazione lineare con  $q = 1$ ,  $A_{q \times p} = a'_{1 \times p}$  e  $b_{q \times 1} = 0$

- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = a'_{1 \times p} \bar{x}$
- $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = a'_{1 \times p} S_{p \times p} a_{p \times 1}$



# Combinazioni lineari

La *combinazione lineare* di  $\tilde{X}_{n \times p}$

$$y_{n \times 1} = \tilde{X}_{n \times pp} a_{p \times 1}$$

- $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = 0$
- $\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = a'_{1 \times pp} S_{pp \times pp} a_{p \times 1}$

Qual è il vettore  $a_{p \times 1}$  che massimizza la varianza  $a'_{1 \times pp} S_{pp \times pp} a_{p \times 1}$  ?



# Vincolo sulla lunghezza del vettore

- La varianza di  $y_{n \times 1}$  dipende dalla lunghezza del vettore  $a_{p \times 1}$ :

$$a' S a = \|a\|^2 \cdot v' S v$$

$1 \times pp \times pp \times 1 \quad 1 \times pp \times pp \times 1$

dove  $v_{p \times 1} = \frac{a_{p \times 1}}{\|a\|}$  ha lunghezza unitaria  $\|v\| = 1$

- Di conseguenza, la varianza di una combinazione lineare  $y_{n \times 1} = X_{n \times pp} a_{p \times 1}$  può essere resa grande/piccola a piacere cambiando la lunghezza di  $a_{p \times 1}$
- Per questo motivo andremo a considerare solo vettori  $v_{p \times 1}$  di lunghezza unitaria  $\|v\| = 1$ , e diremo che  $y_{n \times 1} = \tilde{X}_{n \times pp} v_{p \times 1}$  è una *combinazione lineare normalizzata*



# Teorema: prima componente principale

Sia  $S_{p \times p}$  la matrice di varianze/covarianze di  $\tilde{X}_{n \times p}$ .

Il vettore  $v_{p \times 1}$  di lunghezza unitaria  $\|v\| = 1$  che massimizza  $v'Sv$  è l'autovettore normalizzato  $v_1$  (con segno  $+$  o  $-$ ) di  $S$

$$\pm v_1 = \arg \max_{v: \|v\|=1} v' S v$$

e il massimo di  $v'Sv$  è pari all'autovalore più grande  $\lambda_1$  di  $S$

$$\max_{v: \|v\|=1} v'Sv = v_1'Sv_1 = (-v_1)'S(-v_1) = \lambda_1.$$

La *combinazione lineare normalizzata*

$$y_1 = \tilde{X} v_1$$

(oppure  $-y_1$  con  $-v_1$ ) è detta *prima componente principale* di  $\tilde{X}$ .



# Dimostrazione

Vedi lavagna.



# Outline

- ① Trasformazioni lineari
- ② Analisi delle componenti principali
- ③ Interpretazione geometrica
- ④ PCA per dati standardizzati



# Analisi delle componenti principali

- Sia  $\tilde{X}_{n \times p}$  con  $\text{rango}(\tilde{X}) = p$ .
- Le  $p$  componenti principali di  $\tilde{X}_{n \times p}$  sono le  $p$  colonne della trasformazione lineare

$$\begin{bmatrix} y_1 & y_2 & \dots & y_p \\ n \times 1 & n \times 1 & & n \times 1 \end{bmatrix} = Y_{n \times p} = \tilde{X}_{n \times p} V_{p \times p}$$

dove le colonne di  $V_{p \times p}$  sono gli autovettori normalizzati di  $S$

- Per ridurre la dimensionalità di  $\tilde{X}_{n \times p}$  basta considerare le prime  $q < p$  componenti principali

$$\begin{bmatrix} y_1 & y_2 & \dots & y_q \\ n \times 1 & n \times 1 & & n \times 1 \end{bmatrix} = Y_q_{n \times q} = \tilde{X}_{n \times p} V_q_{p \times q}$$

- La soluzione  $-Y_{n \times p} = \tilde{X}_{n \times p} (-V_{p \times p})$  è equivalente a  $Y_{n \times p}$



# Analisi delle componenti principali

La derivazione delle componenti principali avviene sequenzialmente:

- si cerca la combinazione lineare normalizzata con varianza massima
- poi si cerca una seconda combinazione lineare normalizzata con varianza massima con il vincolo che sia incorrelata con la precedente;
- poi si cerca una terza combinazione lineare normalizzata con varianza massima e che sia incorrelata con le precedenti;
- e così via, determinando un numero di componenti principali pari al rango di  $\tilde{X}$





# Prima componente principale

- I *pesi* (*loadings* in inglese) della prima componente principale di  $\tilde{X}$  sono gli elementi di

$$v_1 = \arg \max_{v: \|v\|=1} v' S v$$

$p \times 1 \quad 1 \times p \quad p \times p \quad p \times 1$

dove  $v_1$  è l'autovettore normalizzato di  $S$  associato a  $\lambda_1$

- I *punteggi* (*scores* in inglese) della prima componente principale di  $\tilde{X}$  sono i valori della combinazione lineare normalizzata

$$y_1 = \tilde{X} v_1$$

$n \times 1 \quad n \times p \quad p \times 1$

- La *varianza spiegata* dalla prima componente principale di  $\tilde{X}$  è

$$\lambda_1 = v_1' S v_1$$

$1 \times p \quad p \times p \quad p \times 1$



## Seconda componente principale

- I *pesi* della seconda componente principale di  $\tilde{X}$  sono gli elementi di

$$v_2 = \arg \max_{\substack{v: \|v\|=1, \\ v'v_1=0}} v' S v$$

$p \times 1 \quad 1 \times p \quad p \times p \quad p \times 1$

dove  $v_2$  è l'autovettore normalizzato di  $S$  associato a  $\lambda_2$

$p \times 1$

- I *punteggi* della seconda componente principale di  $\tilde{X}$  sono i valori della combinazione lineare normalizzata

$$y_2 = \tilde{X} v_2$$

$n \times 1 \quad n \times p \quad p \times 1$

- La *varianza spiegata* dalla seconda componente principale di  $\tilde{X}$  è

$$\lambda_2 = v_2' S v_2$$

$1 \times p \quad p \times p \quad p \times 1$



## $j$ -sima componente principale

- I *pesi* della  $j$ -sima componente principale di  $\tilde{X}$  sono gli elementi di

$$v_j = \underset{\substack{v: \|v\|=1, \\ v'v_k=0, k=1,\dots,j-1}}{\arg \max} \quad \underset{1 \times p}{v'} \underset{p \times p}{S} \underset{p \times 1}{v}$$

dove  $v_j$  è l'autovettore normalizzato di  $S$  associato a  $\lambda_j$

- I *punteggi* della  $j$ -sima componente principale di  $\tilde{X}$  sono i valori della combinazione lineare normalizzata

$$y_j = \underset{n \times 1}{\tilde{X}} \underset{n \times p}{v_j}$$

- La *varianza spiegata* dalla  $j$ -sima componente principale di  $\tilde{X}$  è

$$\lambda_j = \underset{1 \times p}{v_j'} \underset{p \times p}{S} \underset{p \times 1}{v_j}$$



# Proprietà delle componenti principali

- Il vettore delle medie di  $Y = \tilde{X}V$  è nullo:

$$\frac{1}{n} Y' \mathbf{1}_{n \times 1} = \frac{1}{n} V' \tilde{X}' \mathbf{1}_{n \times 1} = \underbrace{V' \mathbf{0}_{p \times p \times 1}}_{p \times 1} = \mathbf{0}_{p \times 1}$$

- La matrice di varianze/covarianze di  $Y = \tilde{X}V$  è

$$S^Y_{p \times p} = \frac{1}{n} Y' Y = \frac{1}{n} V' \tilde{X}' \tilde{X} V = V' S V = V' V \Lambda V' V = \Lambda_{p \times p}$$

dove  $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_p)$ , ovvero  $\underbrace{y_1}_{n \times 1}, \dots, \underbrace{y_p}_{n \times 1}$  hanno varianze pari a  $\lambda_1 \geq \dots \geq \lambda_p$  e sono tra loro incorrelati



# Proprietà delle componenti principali

- Varianza totale di  $S^Y$ :

$$\text{tr}(S^Y) = \text{tr}\left(\Lambda_{p \times p}\right) = \sum_{j=1}^p \lambda_j = \text{tr}(S)$$

coincide con la varianza totale di  $S$

- Proporzione di varianza spiegata dalla  $j$ -sima componente principale

$$\frac{\lambda_j}{\text{tr}(S)} = \frac{\lambda_j}{\sum_{k=1}^p \lambda_k}$$

- Varianza generalizzata di  $S^Y$ :

$$\det(S^Y) = \det\left(\Lambda_{p \times p}\right) = \prod_{j=1}^p \lambda_j = \det(S)$$

coincide con la varianza generalizzata di  $S$



# Proprietà delle componenti principali

- La correlazione tra la  $j$ -sima colonna  $\tilde{x}_j$  di  $\tilde{X}$  e i punteggi  $n \times 1$

$y_k = \tilde{X} v_k$  della  $k$ -sima componente principale di  $\tilde{X}$  è pari a  $n \times 1$

$$\frac{v_{jk} \sqrt{\lambda_k}}{\sqrt{s_{jj}}}$$

*Dimostrazione:*

Possiamo scrivere  $\tilde{x}_j = \tilde{X} a_j$  dove  $a_j$  ha valore 1 in posizione  $n \times 1$   $n \times p$   $p \times 1$   $j$ -sima e 0 altrove. La covarianza tra  $\tilde{x}_j$  e  $y_k$  è

$$\frac{1}{n} \sum_{i=1}^n \tilde{x}_{ji} y_{ki} = \frac{1}{n} \tilde{x}_j' y_k = \frac{1}{n} a_j' \tilde{X}' \tilde{X} v_k = a_j' S v_k = a_j' \lambda_k v_k = \lambda_k v_{jk}$$

dove abbiamo utilizzato  $S v_k = V \Lambda V' V a_k = V \Lambda a_k = V \lambda_k a_k = \lambda_k v_k$ .

La correlazione risulta quindi  $\frac{v_{jk} \lambda_k}{\sqrt{\lambda_k} \sqrt{s_{jj}}} = \frac{v_{jk} \sqrt{\lambda_k}}{\sqrt{s_{jj}}}$ .



# Outline

- ① Trasformazioni lineari
- ② Analisi delle componenti principali
- ③ Interpretazione geometrica**
- ④ PCA per dati standardizzati



# Proiezione su $v_1$

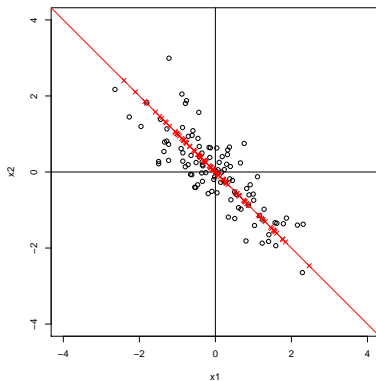
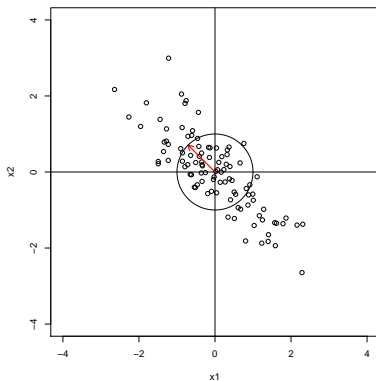
- La proiezione delle righe di  $\tilde{X}$  sul vettore  $v_1$  è

$$\underset{n \times p}{\tilde{X}} \underset{p \times 1}{v_1} \underset{1 \times p}{v_1'} = \underset{n \times 1}{y_1} \underset{1 \times p}{v_1'}$$





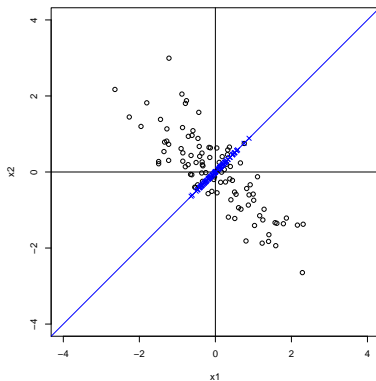
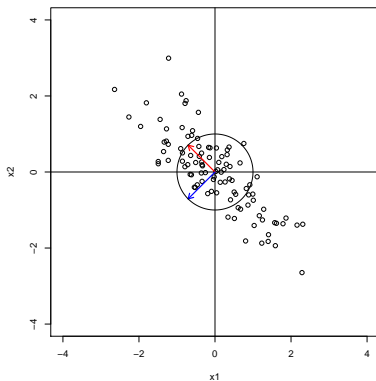
# Proiezione su $v_1$



$p = 2$ : vettore  $v_1$  e proiezione delle righe di  $\tilde{X}$  su  $v_1$



# Proiezione su $v_2$



$p = 2$ : vettore  $v_2$  e proiezione delle righe di  $\tilde{X}$  su  $v_2$



# Proiezione sullo spazio generato da $v_1, \dots, v_q$

- La proiezione delle righe di  $\tilde{X}$  sullo spazio generato da  $v_1, \dots, v_q$ , con  $q \leq p$ , è

$$\tilde{X}_{n \times p} V_q V_q' = Y_q V_q'$$

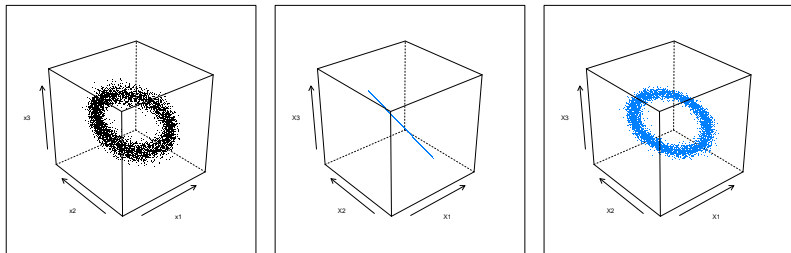
$p \times q \quad q \times p$

dove

$$V_q = \begin{bmatrix} v_1 & \cdots & v_q \\ p \times 1 & & p \times 1 \end{bmatrix}$$



# Proiezione sullo spazio generato da $v_1$ e $v_2$



$p = 3$ : proiezione di  $\tilde{X}$  su  $v_1$  e sullo spazio generato da  $v_1$  e  $v_2$



# Teorema di Eckart-Young

- Miglior approssimazione di rango  $q \leq p$  della matrice  $\tilde{X}_{n \times p}$
- La matrice  $A_{n \times p}$  di rango  $q$  definita da

$$A_{n \times p} = Y_q V_q' = \tilde{X}_{n \times p} V_q V_q' = \arg \min_{\substack{B : \text{rango}(B)=q \\ n \times p}} \sum_{i=1}^n \sum_{j=1}^p (\tilde{x}_{ij} - b_{ij})^2$$

minimizza l'errore di approssimazione

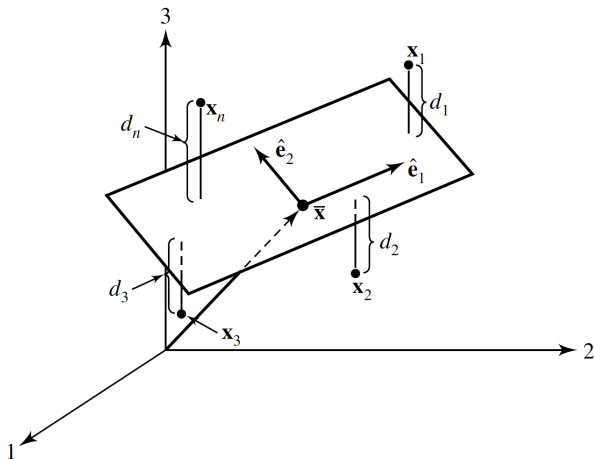
$$\sum_{i=1}^n \sum_{j=1}^p (\tilde{x}_{ij} - a_{ij})^2$$

rispetto a qualsiasi altra matrice  $B_{n \times p}$  di rango  $q$ , i.e.

$$\sum_{i=1}^n \sum_{j=1}^p (\tilde{x}_{ij} - a_{ij})^2 \leq \sum_{i=1}^n \sum_{j=1}^p (\tilde{x}_{ij} - b_{ij})^2$$



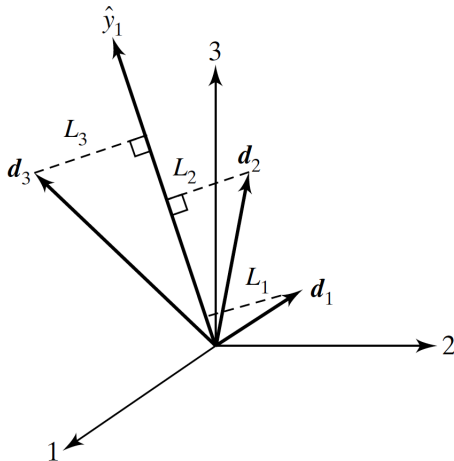
# PCA: spazio delle variabili



$p = 3$ : il piano bidimensionale identificato da  $v_1$  e  $v_2$  minimizza la  
 $p \times 1$   $p \times 1$   
 distanza al quadrato dai punti  $\tilde{x}'_i$  (le righe di  $\tilde{X}$ )  
 $1 \times p$



# PCA: spazio delle osservazioni



$n = 3$ : il vettore  $y_1$   
 $n \times 1$  minimizza le distanze al quadrato dai vettori  
scarto dalla media  $\tilde{x}_j$  (le colonne di  $\tilde{X}$ )  
 $n \times 1$



# Outline

- ① Trasformazioni lineari
- ② Analisi delle componenti principali
- ③ Interpretazione geometrica
- ④ PCA per dati standardizzati





# PCA e trasformazioni lineari

- L'analisi delle componenti principali non è invariante rispetto a trasformazioni lineari, e in particolare di scala
- Essendo le componenti principali costruite sulla base della matrice varianze/covarianze un cambiamento di scala che non sia omogeneo su tutte le variabili produce un cambiamento nelle varianze col risultato di aumentare il peso nelle componenti principali di quelle variabili la cui varianza è aumentata.
- Questo implica, ad esempio, che un cambiamento di unità di misura operato su una sola delle variabili modifica il risultato.
- Queste considerazioni vanno tenute presenti quando si effettua un'analisi per decidere se partire da  $\tilde{X}$  o da  $Z$ ; la scelta andrà fatta caso per caso e non si danno regole generali



# Analisi delle componenti principali di $Z$

- Equivale a considerare la matrice di correlazione:  $S^Z = R$
- Le  $p$  componenti principali sono 
$$Y_{n \times p} = Z_{n \times p} V_{p \times p}$$
- I pesi  $v_j$  della  $j$ -sima componente principale è il  $j$ -simo  $p \times 1$  autovettore normalizzato di  $R$  associato al  $j$ -simo autovalore  $\lambda_j$ ; in generale  $(v_j, \lambda_j)$  di  $R$  sono diversi da quelli di  $S$
- I punteggi della  $j$ -sima componente principale sono 
$$y_j = Z_{n \times 1} v_j_{n \times p_{p \times 1}}$$
- Poichè  $\text{tr}(R) = p$ , la proporzione di varianza spiegata dalla  $j$ -sima componente principale è  $\lambda_j/p$
- La correlazione tra la  $j$ -sima colonna  $z_j$  di  $Z$  e i punteggi  $y_k = Z v_k$  della  $k$ -sima componente principale di  $Z$  è pari a  $v_{jk} \sqrt{\lambda_k}$



## Caso $p = 2$ con dati standardizzati

- Consideriamo i dati standardizzati  $Z$
- Matrice di varianze e covarianze per  $Z$ :

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

con  $r \geq 0$

- I due autovalori di  $R$  sono

$$\lambda_1 = 1 + r, \quad \lambda_2 = 1 - r$$

- I due autovettori normalizzati di  $R$  sono

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



## Caso $p = 2$ con dati standardizzati

- I punteggi delle due componenti principali sono

$$y_{i1} = \frac{1}{\sqrt{2}}(z_{i1} + z_{i2}), \quad y_{i2} = \frac{1}{\sqrt{2}}(z_{i1} - z_{i2})$$

- Se noti che se  $r < 0$ , l'ordine degli autovalori e quindi delle componenti principali è invertito

