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

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

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

Outline

Trasformazioni lineari

La *trasformazione lineare* di $\underset{n\times p}{X}$

$$\boxed{\underset{n\times q}{Y} = \underset{n\times p}{X}\underset{p\times q}{A'} + \underset{n\times 1}{1}\underset{1\times q}{b'}}$$

è definita da

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

Trasformazioni lineari: vettore delle medie

Il vettore delle medie $\underset{q\times 1}{\bar{y}}$ delle trasformazione lineare $\underset{n\times q}{Y} = \underset{n\times p}{X}\underset{p\times q}{A'} + \underset{n\times 1}{1}\underset{1\times q}{b'}$ è dato da

$$\boxed{\underset{q\times 1}{\bar{y}} = \underset{q\times p}{A}\underset{p\times 1}{\bar{x}} + \underset{q\times 1}{b}}$$

Dimostrazione

$$\underset{q\times 1}{\bar{y}} = \frac{1}{n}\underset{q\times n}{Y'}\underset{n\times 1}{1} = \frac{1}{n}\underset{q\times p}{A}\underset{p\times n}{X}\underset{n\times 1}{1} + \frac{1}{n}\underset{q\times 1}{b}\underset{1\times n}{1'}\underset{n\times 1}{1} = \underset{q\times p}{A}\underset{p\times 1}{\bar{x}} + \underset{q\times 1}{b}$$

Trasformazioni lineari: matrice di varianze/covarianze

La matrice di varianze/covarianze $\underset{q\times q}{S^Y}$ della trasformazione lineare $\underset{n\times q}{Y} = \underset{n\times p}{X}\,\underset{p\times q}{A'} + \underset{n\times 1}{1}\,\underset{1\times q}{b'}$ è data da

$$\boxed{\underset{q\times q}{S^Y} = \underset{q\times p}{A}\,\underset{p\times p}{S}\,\underset{p\times q}{A'}}$$

Dimostrazione:

$$\underset{q\times q}{S^Y} = \frac{1}{n}\underset{q\times n}{\tilde{Y}'}\,\underset{n\times q}{\tilde{Y}} = \frac{1}{n}\underset{q\times p}{A}\,\underset{p\times n}{\tilde{X}'}\,\underset{n\times p}{\tilde{X}}\,\underset{p\times q}{A'} = \underset{q\times p}{A}\,\underset{p\times p}{S}\,\underset{p\times q}{A'}$$

dove

$$\underset{n\times q}{\tilde{Y}} = \underset{n\times n}{H}\,\underset{n\times q}{Y} = \underset{n\times n}{H}\,\underset{n\times p}{X}\,\underset{p\times q}{A'} + \underset{n\times n}{H}\,\underset{n\times 1}{1}\,\underset{1\times q}{b'} = \underset{n\times n}{H}\,\underset{n\times p}{X}\,\underset{p\times q}{A'} = \underset{n\times p}{\tilde{X}}\,\underset{p\times q}{A'}$$

Trasformazioni lineari note

q	$\underset{q\times p}{A}$	$\underset{q\times 1}{b}$	$\underset{n\times q}{Y} = \underset{n\times p}{X}\underset{p\times q}{A'} + \underset{n\times 1}{1}\underset{1\times q}{b'}$
p	$\underset{p\times p}{I}$	$-\underset{p\times 1}{\bar{x}}$	$\underset{n\times p}{\tilde{X}}$
p	$\underset{p\times p}{D}^{-1/2}$	$-\underset{p\times p}{D}^{-1/2}\underset{p\times 1}{\bar{x}}$	$\underset{n\times p}{Z}$
p	$\underset{p\times p}{S}^{-1/2}$	$-\underset{p\times p}{S}^{-1/2}\underset{p\times 1}{\bar{x}}$	$\underset{n\times p}{\tilde{Z}}$

Combinazioni lineari

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

$$\underset{n\times 1}{y} = \underset{n\times p}{X}\underset{p\times 1}{a} = \begin{bmatrix} \sum_{j=1}^{p} a_j x_{1j} \\ \vdots \\ \sum_{j=1}^{p} a_j x_{ij} \\ \vdots \\ \sum_{j=1}^{p} a_j x_{nj} \end{bmatrix}$$

è un caso particolare di trasformazione lineare con $q=1$, $\underset{q\times p}{A} = \underset{1\times p}{a'}$ e $\underset{q\times 1}{b} = 0$

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

Combinazioni lineari

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

$$\underset{n\times 1}{y} = \underset{n\times p}{\tilde{X}}\underset{p\times 1}{a}$$

- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 0$
- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \underset{1\times p}{a'}\underset{p\times p}{S}\underset{p\times 1}{a}$

Qual è il vettore $\underset{p\times 1}{a}$ che massimizza la varianza $\underset{1\times p}{a'}\underset{p\times p}{S}\underset{p\times 1}{a}$?

Vincolo sulla lunghezza del vettore

- La varianza di $\underset{n\times 1}{y}$ dipende dalla lunghezza del vettore $\underset{p\times 1}{a}$:

$$\underset{1\times p}{a'}\underset{p\times p}{S}\underset{p\times 1}{a} = \|a\|^2 \cdot \underset{1\times p}{v'}\underset{p\times p}{S}\underset{p\times 1}{v}$$

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

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

Teorema: prima componente principale

Sia $\underset{p\times p}{S}$ la matrice di varianze/covarianze di $\underset{n\times p}{\tilde{X}}$.
Il vettore $\underset{p\times 1}{v}$ di lunghezza unitaria $\|v\|=1$ che massimizza $v'Sv$ è l'autovettore normalizzato $\underset{p\times 1}{v_1}$ (con segno $+$ o $-$) di S

$$\pm \underset{p\times 1}{v_1} = \underset{v:\|v\|=1}{\arg\max} \underset{1\times p}{v'} \underset{p\times p}{S} \underset{p\times 1}{v}$$

e il massimo di $v'Sv$ è pari all'autovalore più grande λ_1 di S

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

La *combinazione lineare normalizzata*

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

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

Dimostrazione

Vedi lavagna.

Outline

Analisi delle componenti principali

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

$$[\underset{n\times 1}{y_1} \;\; \underset{n\times 1}{y_2} \;\; \ldots\ldots \;\; \underset{n\times 1}{y_p}] = \underset{n\times p}{Y} = \underset{n\times p}{\tilde{X}}\underset{p\times p}{V}$$

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

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

$$[\underset{n\times 1}{y_1} \;\; \underset{n\times 1}{y_2} \;\; \ldots \;\; \underset{n\times 1}{y_q}] = \underset{n\times q}{Y_q} = \underset{n\times p}{\tilde{X}}\underset{p\times q}{V_q}$$

- La soluzione $-\underset{n\times p}{Y} = \underset{n\times p}{\tilde{X}}(-\underset{p\times p}{V})$ è equivalente a $\underset{n\times p}{Y}$

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

$$\underset{p\times 1}{v_1} = \underset{v:\|v\|=1}{\arg\max}\ \underset{1\times p}{v'}\underset{p\times p}{S}\underset{p\times 1}{v}$$

dove $\underset{p\times 1}{v_1}$ è l'autovettore normalizzato di S associato a λ_1

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

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

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

$$\lambda_1 = \underset{1\times p}{v_1'}\underset{p\times p}{S}\underset{p\times 1}{v_1}$$

Seconda componente principale

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

$$\underset{p\times 1}{v_2} = \underset{\substack{v:\ \|v\|=1,\\ v'v_1=0}}{\arg\max} \underset{1\times p}{v'} \underset{p\times p}{S} \underset{p\times 1}{v}$$

dove $\underset{p\times 1}{v_2}$ è l'autovettore normalizzato di S associato a λ_2

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

$$\underset{n\times 1}{y_2} = \underset{n\times p}{\tilde{X}} \underset{p\times 1}{v_2}$$

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

$$\lambda_2 = \underset{1\times p}{v_2'} \underset{p\times p}{S} \underset{p\times 1}{v_2}$$

j-sima componente principale

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

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

  dove $\underset{p\times 1}{v_j}$ è l'autovettore normalizzato di S associato a λ_j

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

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

- 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:

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

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

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

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

Proprietà delle componenti principali

- Varianza totale di S^Y:

$$\mathrm{tr}(S^Y) = \mathrm{tr}(\underset{p \times p}{\Lambda}) = \sum_{j=1}^{p} \lambda_j = \mathrm{tr}(S)$$

  coincide con la varianza totale di S

- Proporzione di varianza spiegata dalla j-sima componente principale

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

- Varianza generalizzata di S^Y:

$$\det(S^Y) = \det(\underset{p \times p}{\Lambda}) = \prod_{j=1}^{p} \lambda_j = \det(S)$$

  coincide con la varianza generalizzata di S

Proprietà delle componenti principali

- La correlazione tra la j-sima colonna $\underset{n\times 1}{\tilde{x}_j}$ di $\tilde{X}$ e i punteggi $\underset{n\times 1}{y_k} = \tilde{X}v_k$ della k-sima componente principale di $\tilde{X}$ è pari a

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

Dimostrazione:
Possiamo scrivere $\underset{n\times 1}{\tilde{x}_j} = \underset{n\times p}{\tilde{X}}\underset{p\times 1}{a_j}$ dove $\underset{p\times 1}{a_j}$ ha valore 1 in posizione 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 $Sv_k = V\Lambda V'Va_k = V\Lambda a_k = V\lambda_k a_k = \lambda_k v_k$.
La correlazione risulta quindi $\frac{v_{jk}\lambda_k}{\sqrt{\lambda_k}\sqrt{s_{jj}}} = \frac{v_{jk}\sqrt{\lambda_k}}{\sqrt{s_{jj}}}$.
□

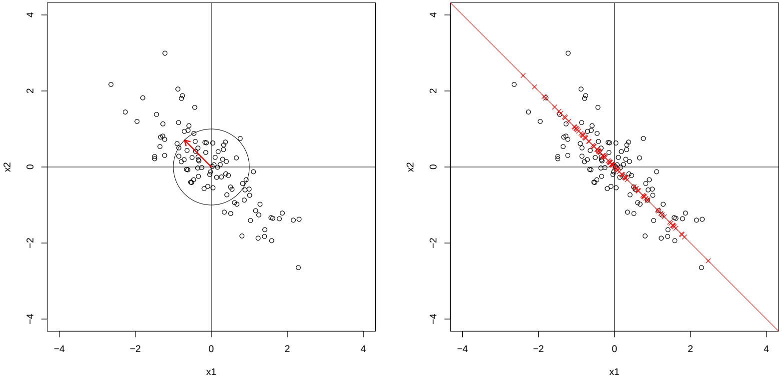
Outline

Proiezione su v_1

- La proiezione delle righe di $\underset{n\times p}{\tilde{X}}$ sul vettore $\underset{p\times 1}{v_1}$ è

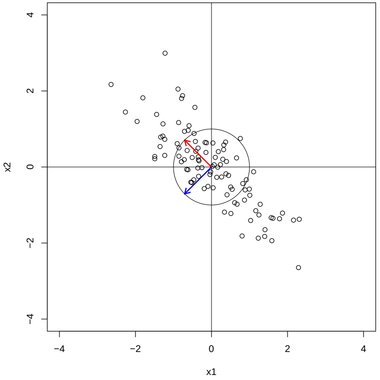
$$\underset{n\times p}{\tilde{X}}\underset{p\times 1}{v_1}\underset{1\times p}{v_1'} = \underset{n\times 1}{y_1}\underset{1\times p}{v_1'}$$

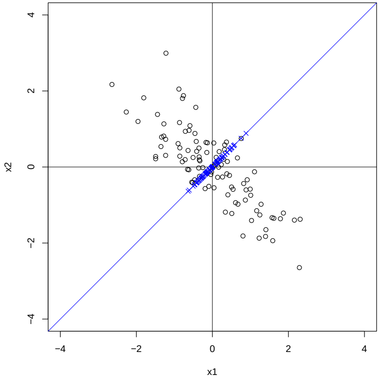
Proiezione su v_1



$p = 2$: 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 $v_1, \ldots, v_q$

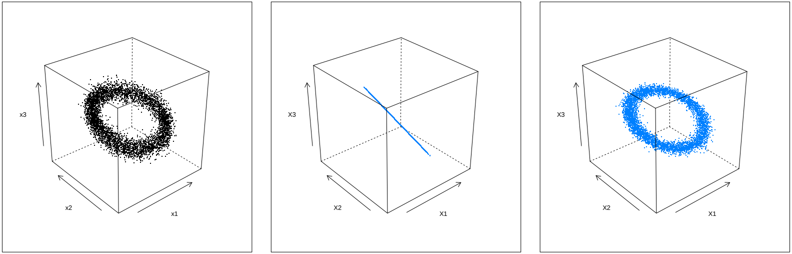
- La proiezione delle righe di $\underset{n\times p}{\tilde{X}}$ sullo spazio generato da $\underset{p\times 1}{v_1}, \ldots, \underset{p\times 1}{v_q}$, con $q \leq p$, è

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

dove

$$\underset{p\times q}{V_q} = \left[\begin{array}{ccc} \underset{p\times 1}{v_1} & \cdots & \underset{p\times 1}{v_q} \end{array} \right]$$

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

La miglior approssimazione (rispetto alla norma di Frobenius) di rango $q \leq \mathrm{rango}(\tilde{X})$ della matrice $\tilde{X}$ è data dalla matrice

$$\underset{n\times p}{A} = \underset{n\times q}{Y_q}\,\underset{q\times p}{V_q'} = \underset{n\times p}{\tilde{X}}\,\underset{p\times q}{V_q}\,\underset{q\times p}{V_q'}$$

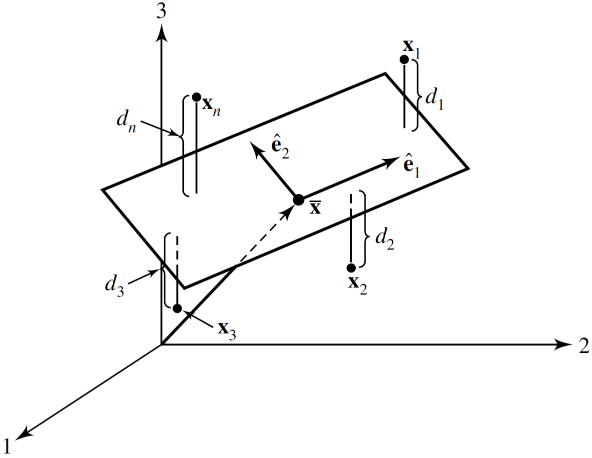
di rango q che minimizza l'errore di approssimazione

$$\|\tilde{X} - A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(\tilde{x}_{ij} - a_{ij})^2 = n \sum_{j=q+1}^{p} \lambda_j$$

rispetto a qualsiasi altra matrice $\underset{n\times p}{B}$ di rango al più q, i.e.

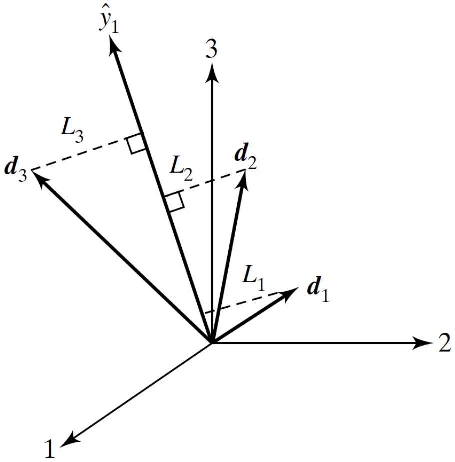
$$\|\tilde{X} - A\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{p}(\tilde{x}_{ij} - a_{ij})^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{p}(\tilde{x}_{ij} - b_{ij})^2 = \|\tilde{X} - B\|_F^2$$

PCA: spazio delle variabili



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

PCA: spazio delle osservazioni



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

Outline

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 $\underset{n\times p}{Y} = \underset{n\times p}{Z}\underset{p\times p}{V}$
- I pesi $\underset{p\times 1}{v_j}$ della j-sima componente principale è il j-simo autovettore normalizzato di R associato al j-simo autovalore λ_j; in generale (v_j, λ_j) di R sono diversi da quelli di S
- I punteggi della j-sima componente principale sono $\underset{n\times 1}{y_j} = \underset{n\times p}{Z}\underset{p\times 1}{v_j}$
- Poichè $\mathrm{tr}(R) = p$, la proporzione di varianza spiegata dalla j-sima componente principale è λ_j/p
- La correlazione tra la j-sima colonna z_j di Z e i punteggi $y_k = Zv_k$ 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

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

- I due autovettori normalizzati di R sono

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

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