Statistica II (750AA)

Lezione 3

Dario Trevisan

06/10/2025

Riduzione della dimensionalità

Obiettivi della lezione

- Concetto di **Curse of Dimensionality**

Obiettivi della lezione

- Concetto di **Curse of Dimensionality**
- Varianza e covarianza **campionarie**

Obiettivi della lezione

- Concetto di **Curse of Dimensionality**
- Varianza e covarianza **campionarie**
- PCA (Analisi delle Componenti Principali)

Obiettivi della lezione

- Concetto di **Curse of Dimensionality**
- Varianza e covarianza **campionarie**
- PCA (Analisi delle Componenti Principali)
- EFA (Analisi Fattoriale Esplorativa)

Curse of Dimensionality

- Supponiamo di aver raccolto un campione $\{x_i\}_{i=1,\ldots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).

Curse of Dimensionality

- Supponiamo di aver raccolto un campione $\{x_i\}_{i=1,\ldots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).
- Organizziamo in una matrice (*dataframe* in R):

$$x \in \mathbb{R}^{n \times d}$$

Curse of Dimensionality

- Supponiamo di aver raccolto un campione $\{x_i\}_{i=1,\dots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).
- Organizziamo in una matrice (*dataframe* in R):

$$x \in \mathbb{R}^{n \times d}$$

- La **Curse of Dimensionality** si riferisce alle problematiche che sorgono quando il numero di features d è grande.

Curse of Dimensionality

- Supponiamo di aver raccolto un campione $\{x_i\}_{i=1,\dots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).
- Organizziamo in una matrice (*dataframe* in R):

$$x \in \mathbb{R}^{n \times d}$$

- La **Curse of Dimensionality** si riferisce alle problematiche che sorgono quando il numero di features d è grande.
- Se $d \gg 1$, la *densità dell'informazione nei dati* diminuisce, rendendo difficile l'analisi.

Curse of Dimensionality

- Supponiamo di aver raccolto un campione $\{x_i\}_{i=1,\dots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).
- Organizziamo in una matrice (*dataframe* in R):

$$x \in \mathbb{R}^{n \times d}$$

- La **Curse of Dimensionality** si riferisce alle problematiche che sorgono quando il numero di features d è grande.
- Se $d \gg 1$, la *densità dell'informazione nei dati* diminuisce, rendendo difficile l'analisi.
- In particolare le distanze tipiche (Euclidea, Minkowski ecc.) non distinguono il *segnale* dal *rumore*.

Spiegazione euristica

Esempio generato

- La prima componente indica una classe -1 o 1

Esempio generato

- La prima componente indica una classe -1 o 1
- Le altre $d-1$ componenti contengono rumore casuale in $[-1/2, 1/2]$

Esempio generato

- La prima componente indica una classe -1 o 1
- Le altre $d-1$ componenti contengono rumore casuale in $[-1/2, 1/2]$
- Eseguiamo clustering con k-means

Esempio generato

- La prima componente indica una classe -1 o 1
- Le altre $d-1$ componenti contengono rumore casuale in $[-1/2, 1/2]$
- Eseguiamo clustering con k-means
- Confrontiamo con la classe indicata

```
## [1] "Taglia n =  10 , dimensione d =  10"
##                      classi originarie
## cluster individuati  0  1
##                   1  0 10
##                   2 10  0
## [1] ""
## [1] "Taglia n =  10 , dimensione d =  100"
##                      classi originarie
## cluster individuati 0 1
##                   1 8 3
##                   2 2 7
## [1] ""
## [1] "Taglia n =  10 , dimensione d =  1000"
##                      classi originarie
## cluster individuati 0 1
##                   1 6 6
##                   2 4 4
## [1] ""
```

Esempio MNIST

Il dataset MNIST contiene immagini di cifre scritte a mano:

- in scala di grigio (valori da 0 a 255)

Esempio MNIST

Il dataset MNIST contiene immagini di cifre scritte a mano:

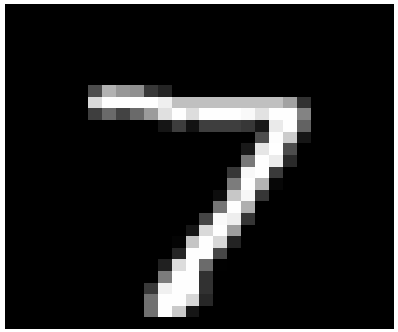
- in scala di grigio (valori da 0 a 255)
- ciascuna composta da 28*28 = 784 pixels

Esempio MNIST

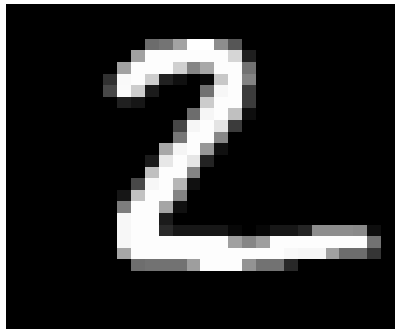
Il dataset MNIST contiene immagini di cifre scritte a mano:

- in scala di grigio (valori da 0 a 255)
- ciascuna composta da 28*28 = 784 pixels
- e una componente che indica la cifra effettivamente scritta.

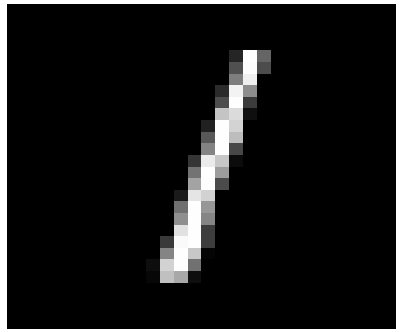
Label: 7



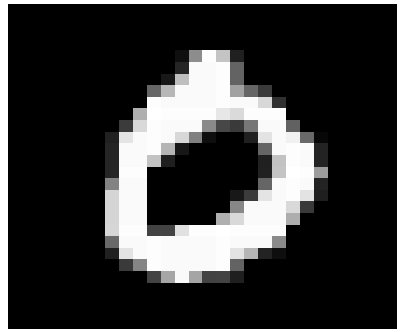
Label: 2



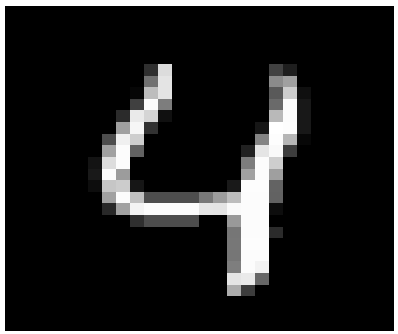
Label: 1



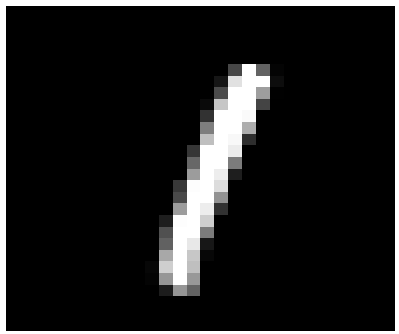
Label: 0



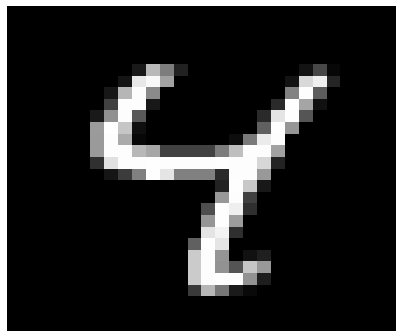
Label: 4



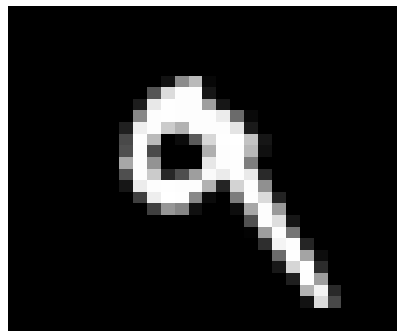
Label: 1



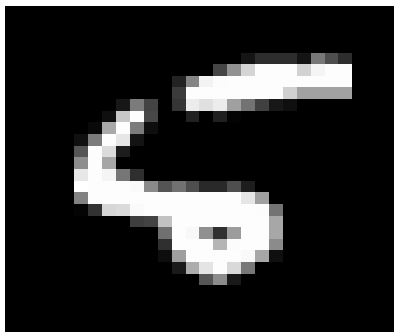
Label: 4



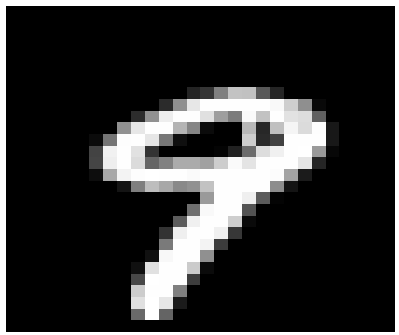
Label: 9



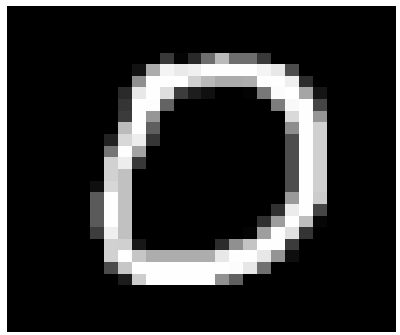
Label: 5



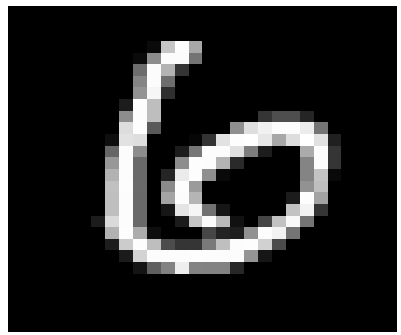
Label: 9



Label: 0



Label: 6



CLustering con k-means

Consideriamo un sottoinsieme di MNIST contenente

- 100 immagini della cifra 7

```
##                       classi originarie
## cluster individuati    5   7
##                   1   21 100
##                   2   79   0
```

CLustering con k-means

Consideriamo un sottoinsieme di MNIST contenente

- 100 immagini della cifra 7
- 100 immagini della cifra 5

```
##                       classi originarie
## cluster individuati    5   7
##                   1   21 100
##                   2   79   0
```

CLustering con k-means

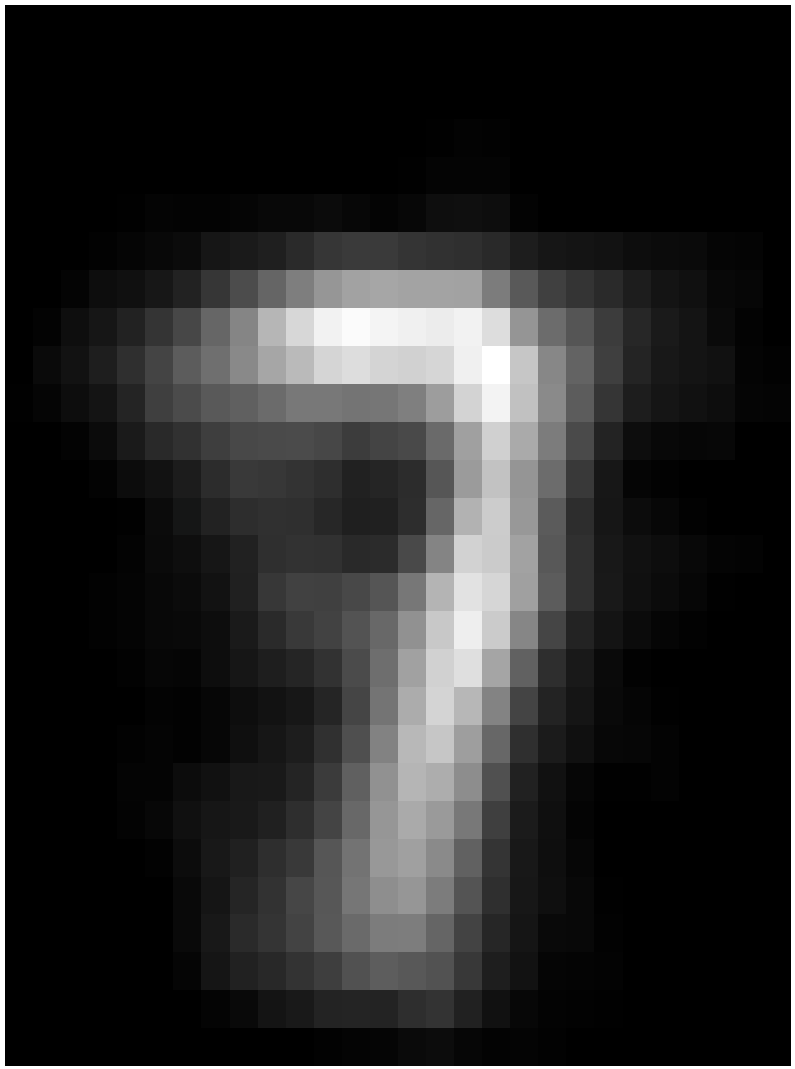
Consideriamo un sottoinsieme di MNIST contenente

- 100 immagini della cifra 7
- 100 immagini della cifra 5
- e applichiamo k-means.

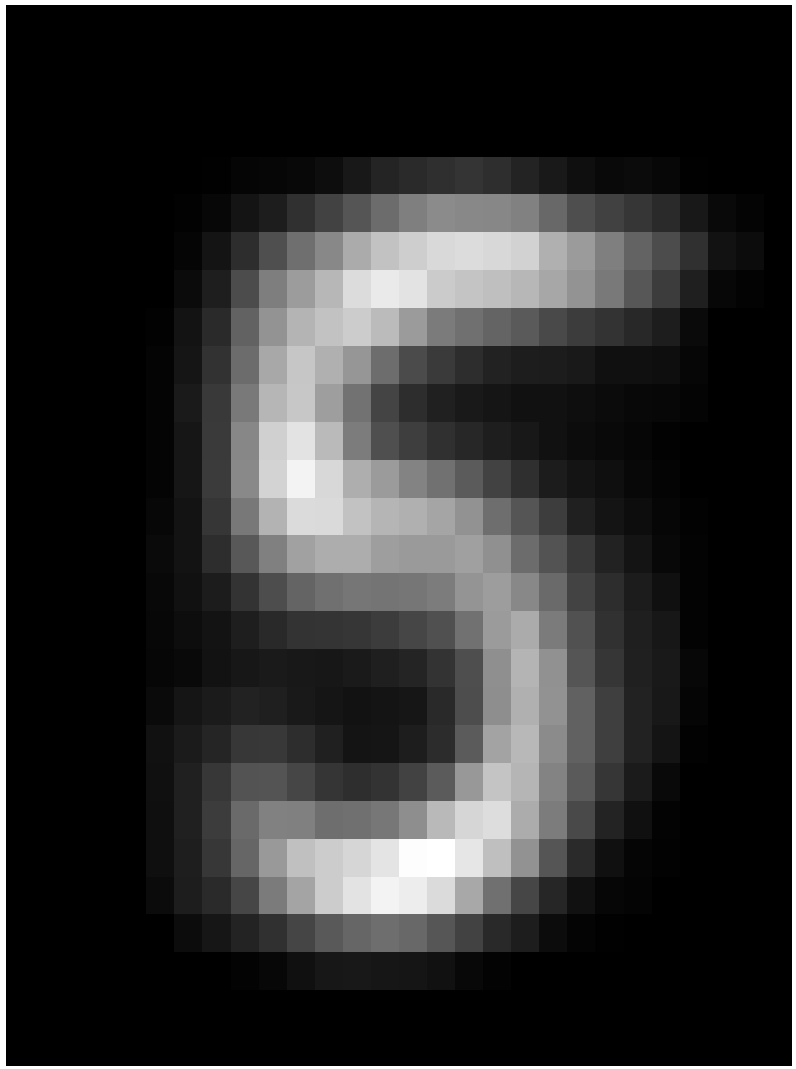
```
##                       classi originarie
## cluster individuati    5   7
##                   1   21 100
##                   2   79   0
```

Plottiamo i centri trovati: riconosciamo delle *cifre medie*.

Label: Centroide cluster 1



Label: Centroide cluster 2



Varianza campionaria

- La varianza di un campione **univariato** $x = \{x_i\}_{i=1,\dots,n} \subseteq \mathbb{R}$ è

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \mathrm{sd}(x) = \sqrt{\mathrm{var}(x)},$$

Varianza campionaria

- La varianza di un campione **univariato** $x = \{x_i\}_{i=1,\dots,n} \subseteq \mathbb{R}$ è

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \mathrm{sd}(x) = \sqrt{\mathrm{var}(x)},$$

- La media campionaria $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ minimizza il MSE:

$$\bar{x} \in \arg\min_{h\in\mathbb{R}} \sum_{i=1}^{n}(x_i - h)^2$$

Varianza campionaria

- La varianza di un campione **univariato** $x = \{x_i\}_{i=1,\dots,n} \subseteq \mathbb{R}$ è

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \mathrm{sd}(x) = \sqrt{\mathrm{var}(x)},$$

- La media campionaria $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ minimizza il MSE:

$$\bar{x} \in \arg\min_{h\in\mathbb{R}} \sum_{i=1}^{n}(x_i - h)^2$$

- Il rischio minimo è $(n-1)\,\mathrm{var}(x)$.

Varianza campionaria

- La varianza di un campione **univariato** $x = \{x_i\}_{i=1,\dots,n} \subseteq \mathbb{R}$ è

$$\mathrm{var}(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad \mathrm{sd}(x) = \sqrt{\mathrm{var}(x)},$$

- La media campionaria $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ minimizza il MSE:

$$\bar{x} \in \arg\min_{h\in\mathbb{R}} \sum_{i=1}^{n}(x_i - h)^2$$

- Il rischio minimo è $(n-1)\,\mathrm{var}(x)$.
- **Problema**: estendere la varianza nel caso multivariato $x_i \in \mathbb{R}^d$.

Varianza direzionale

Dato un campione multivariato $x = (x_i)_{i=1,\dots,n} \in \mathbb{R}^{n \times d}$

- consideriamo la sua *proiezione* lungo una direzione $v \in \mathbb{R}^d$

$$(x_i v)_{i=1,\dots,n} = xv \in \mathbb{R}^n.$$

Varianza direzionale

Dato un campione multivariato $x = (x_i)_{i=1,\dots,n} \in \mathbb{R}^{n \times d}$

- consideriamo la sua *proiezione* lungo una direzione $v \in \mathbb{R}^d$

$$(x_i v)_{i=1,\dots,n} = xv \in \mathbb{R}^n.$$

- la varianza (campionaria) direzionale è

$$\mathrm{var}(xv) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i v - \overline{xv})^2 \geq 0$$

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \text{var}(xv) = v^T \text{var}(x) v$$

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \text{var}(xv) = v^T \text{var}(x) v$$

- dove la **matrice delle covarianze** campionarie è

$$\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}^{d \times d}.$$

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \text{var}(xv) = v^T \text{var}(x) v$$

- dove la **matrice delle covarianze** campionarie è

$$\text{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}^{d \times d}.$$

- $\text{var}(x)$ è una matrice

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \mathrm{var}(xv) = v^T \mathrm{var}(x) v$$

- dove la **matrice delle covarianze** campionarie è

$$\mathrm{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}^{d \times d}.$$

- $\mathrm{var}(x)$ è una matrice
  - *quadrata* di taglia $d \times d$

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \mathrm{var}(xv) = v^T \mathrm{var}(x) v$$

- dove la **matrice delle covarianze** campionarie è

$$\mathrm{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}^{d \times d}.$$

- $\mathrm{var}(x)$ è una matrice
  - *quadrata* di taglia $d \times d$
  - simmetrica

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \operatorname{var}(xv) = v^T \operatorname{var}(x) v$$

- dove la **matrice delle covarianze** campionarie è

$$\operatorname{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^T (x_i - \bar{x}) \in \mathbb{R}^{d \times d}.$$

- $\operatorname{var}(x)$ è una matrice
  - *quadrata* di taglia $d \times d$
  - simmetrica
  - semidefinita positiva.

Covarianza

Per j, $\ell \in \{1, \ldots, d\}$, la componente $\mathrm{var}(x)_{j,\ell}$ è

$$\mathrm{var}(x)_{j,\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \bar{x_j})(x_{i,\ell} - \bar{x_\ell}) =: \mathrm{cov}(x_j, x_\ell).$$

- Se $v = e_j$, $w = e_\ell \in \mathbb{R}^d$ (vettori di base canonica) troviamo $xv = x_j$, $xw = x_\ell$.

Covarianza

Per j, $\ell \in \{1, \ldots, d\}$, la componente $\mathrm{var}(x)_{j,\ell}$ è

$$\mathrm{var}(x)_{j,\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \bar{x_j})(x_{i,\ell} - \bar{x_\ell}) =: \mathrm{cov}(x_j, x_\ell).$$

- Se $v = e_j$, $w = e_\ell \in \mathbb{R}^d$ (vettori di base canonica) troviamo $xv = x_j$, $xw = x_\ell$.
- Dati v, $w \in \mathbb{R}^d$, definiamo

$$\mathrm{cov}(xv, xw) := v^T \mathrm{var}(x) w = \frac{1}{n-1} \sum_{i=1}^{n} (x_i v - \bar{x} v)(x_i w - \bar{x} w).$$

Covarianza

Per j, $\ell \in \{1, \ldots, d\}$, la componente $\mathrm{var}(x)_{j,\ell}$ è

$$\mathrm{var}(x)_{j,\ell} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{i,j} - \bar{x}_j)(x_{i,\ell} - \bar{x}_\ell) =: \mathrm{cov}(x_j, x_\ell).$$

- Se $v = e_j$, $w = e_\ell \in \mathbb{R}^d$ (vettori di base canonica) troviamo $xv = x_j$, $xw = x_\ell$.
- Dati v, $w \in \mathbb{R}^d$, definiamo

$$\mathrm{cov}(xv, xw) := v^T \mathrm{var}(x) w = \frac{1}{n-1} \sum_{i=1}^{n} (x_i v - \bar{x} v)(x_i w - \bar{x} w).$$

- Più in generale, se $V \in \mathbb{R}^{d \times k_1}$, $W \in \mathbb{R}^{d \times k_2}$, si pone

$$\mathrm{cov}(xV, xW) = V^T \mathrm{var}(x) W \in \mathbb{R}^{k_1 \times k_2}, \quad \mathrm{var}(xV) = V^T \mathrm{var}(x) V.$$

Bilinearità della covarianza

- La covarianza è *bilineare*:

$$\operatorname{cov}(x + x', y) = \operatorname{cov}(x, y) + \operatorname{cov}(x', y)$$

$$\operatorname{cov}(x, y + y') = \operatorname{cov}(x, y) + \operatorname{cov}(x, y')$$

Bilinearità della covarianza

- La covarianza è *bilineare*:

$$\operatorname{cov}(x + x', y) = \operatorname{cov}(x, y) + \operatorname{cov}(x', y)$$

$$\operatorname{cov}(x, y + y') = \operatorname{cov}(x, y) + \operatorname{cov}(x, y')$$

- Ne segue la regola per la *varianza della somma*:

$$\operatorname{var}(x + x') = \operatorname{var}(x) + \operatorname{var}(x') + 2\operatorname{cov}(x, x').$$

Bilinearità della covarianza

- La covarianza è *bilineare*:

$$\operatorname{cov}(x + x', y) = \operatorname{cov}(x, y) + \operatorname{cov}(x', y)$$

$$\operatorname{cov}(x, y + y') = \operatorname{cov}(x, y) + \operatorname{cov}(x, y')$$

- Ne segue la regola per la *varianza della somma*:

$$\operatorname{var}(x + x') = \operatorname{var}(x) + \operatorname{var}(x') + 2\operatorname{cov}(x, x').$$

- Se x, x' **non sono correlate** $\operatorname{cov}(x, x') = 0$ allora

$$\operatorname{var}(x + x') = \operatorname{var}(x) + \operatorname{var}(x')$$

Coefficiente di correlazione di Pearson

Consideriamo il caso $d = 2$, $x = (x_1, x_2)$:

$$\mathrm{var}(x) = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix}$$

- La matrice è semidefinita positiva $\Rightarrow \det(\mathrm{var}(x)) \geq 0$:

$$\mathrm{var}(x_1)\,\mathrm{var}(x_2) \geq \mathrm{cov}(x_1, x_2)^2$$

Coefficiente di correlazione di Pearson

Consideriamo il caso $d = 2$, $x = (x_1, x_2)$:

$$\mathrm{var}(x) = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix}$$

- La matrice è semidefinita positiva $\Rightarrow \det(\mathrm{var}(x)) \geq 0$:

$$\mathrm{var}(x_1)\,\mathrm{var}(x_2) \geq \mathrm{cov}(x_1, x_2)^2$$

- Definiamo il **coefficiente di correlazione**:

$$\mathrm{cor}(x_1, x_2) := \frac{\mathrm{cov}(x_1, x_2)}{\mathrm{sd}(x_1)\,\mathrm{sd}(x_2)} \in [-1, 1]$$

Coefficiente di correlazione di Pearson

Consideriamo il caso $d = 2$, $x = (x_1, x_2)$:

$$\mathrm{var}(x) = \begin{pmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1, x_2) \\ \mathrm{cov}(x_2, x_1) & \mathrm{var}(x_2) \end{pmatrix}$$

- La matrice è semidefinita positiva $\Rightarrow \det(\mathrm{var}(x)) \geq 0$:

$$\mathrm{var}(x_1)\,\mathrm{var}(x_2) \geq \mathrm{cov}(x_1, x_2)^2$$

- Definiamo il **coefficiente di correlazione**:

$$\mathrm{cor}(x_1, x_2) := \frac{\mathrm{cov}(x_1, x_2)}{\mathrm{sd}(x_1)\,\mathrm{sd}(x_2)} \in [-1, 1]$$

- Valori estremi (vicini a 1 o -1) indicano una *forte relazione lineare* tra x_1 e x_2:

$$x_1 \approx x_2 a + b$$

Matrice di correlazione

- Per d generale definiamo la **matrice di correlazione**

$$\mathrm{cor}(x) = \mathrm{SD}(x)^{-1}\,\mathrm{cov}(x)\,\mathrm{SD}(x)^{-1} \in \mathbb{R}^{d\times d}$$

dove $SD(x)$ è una matrice *diagonale* contenente le deviazioni standard delle singole features:

$$SD(x)_{jj} = \mathrm{sd}(x_j).$$

Matrice di correlazione

- Per d generale definiamo la **matrice di correlazione**

$$\mathrm{cor}(x) = \mathrm{SD}(x)^{-1}\,\mathrm{cov}(x)\,\mathrm{SD}(x)^{-1} \in \mathbb{R}^{d\times d}$$

  dove $SD(x)$ è una matrice *diagonale* contenente le deviazioni standard delle singole features:

$$SD(x)_{jj} = \mathrm{sd}(x_j).$$

- Vale $\mathrm{cor}(x)_{j,\ell} = \mathrm{cor}(x_j, x_\ell) \in [-1, 1]$.

Matrice di correlazione

- Per d generale definiamo la **matrice di correlazione**

$$\mathrm{cor}(x) = \mathrm{SD}(x)^{-1}\,\mathrm{cov}(x)\,\mathrm{SD}(x)^{-1} \in \mathbb{R}^{d\times d}$$

  dove $SD(x)$ è una matrice *diagonale* contenente le deviazioni standard delle singole features:

$$SD(x)_{jj} = \mathrm{sd}(x_j).$$

- Vale $\mathrm{cor}(x)_{j,\ell} = \mathrm{cor}(x_j, x_\ell) \in [-1, 1]$.
- la matrice di correlazione è pure simmetrica, semidefinita positiva e vale sempre 1 sulla diagonale.

Matrice di correlazione

- Per d generale definiamo la **matrice di correlazione**

$$\mathrm{cor}(x) = \mathrm{SD}(x)^{-1}\,\mathrm{cov}(x)\,\mathrm{SD}(x)^{-1} \in \mathbb{R}^{d\times d}$$

  dove $SD(x)$ è una matrice *diagonale* contenente le deviazioni standard delle singole features:

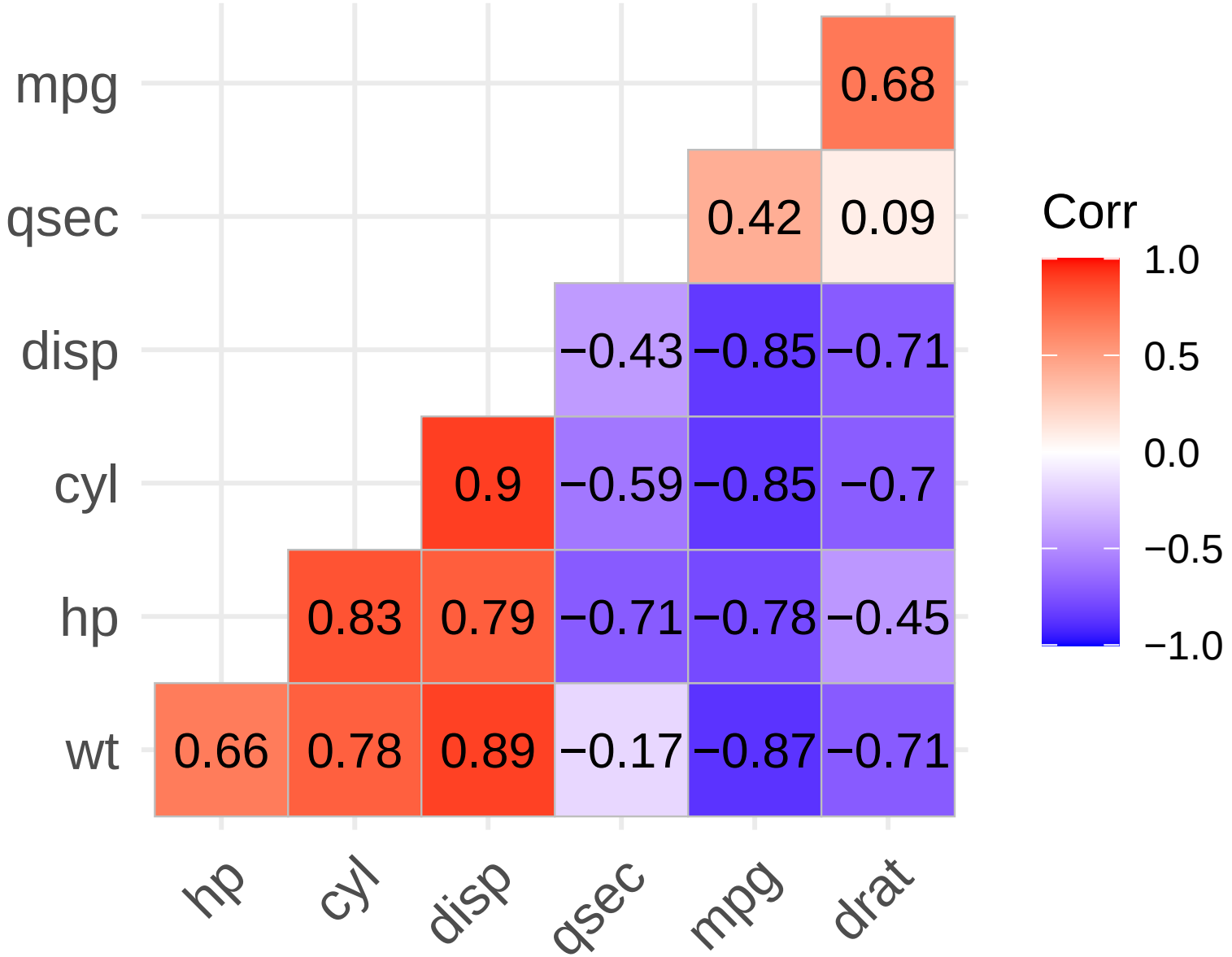
$$SD(x)_{jj} = \mathrm{sd}(x_j).$$

- Vale $\mathrm{cor}(x)_{j,\ell} = \mathrm{cor}(x_j, x_\ell) \in [-1, 1]$.
- la matrice di correlazione è pure simmetrica, semidefinita positiva e vale sempre 1 sulla diagonale.
- La matrice di correlazione è uguale alla matrice di covarianza delle features **standardizzate** (riscalate):

$$\mathrm{scale}(x) := (x - \bar{x})SD(x)^{-1} \quad \Rightarrow \quad \mathrm{cov}(\mathrm{scale}(x)) = \mathrm{cor}(x)$$

Correlation heatmap (mappa di calore)

Esempio: per il dataset `mtcars`:



Riduzione della dimensionalità

Obiettivo: dato $x \in \mathbb{R}^{n \times d}$, trovarne una *buona approssimazione* $z \in \mathbb{R}^{n \times k}$ con $k \ll d$.

- Due approcci:

Riduzione della dimensionalità

Obiettivo: dato $x \in \mathbb{R}^{n \times d}$, trovarne una *buona approssimazione* $z \in \mathbb{R}^{n \times k}$ con $k \ll d$.

- Due approcci:
  - Selezione/**proiezione** delle features:

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, \quad z_i = \Pi(x_i)$$

Riduzione della dimensionalità

Obiettivo: dato $x \in \mathbb{R}^{n \times d}$, trovarne una *buona approssimazione* $z \in \mathbb{R}^{n \times k}$ con $k \ll d$.

- Due approcci:
  - Selezione/**proiezione** delle features:

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, \quad z_i = \Pi(x_i)$$

  - **Generazione**/parametrizzazione (spazio latente):

$$g : \mathbb{R}^k \to \mathbb{R}^d, \quad x_i \approx g(z_i)$$

Riduzione della dimensionalità

Obiettivo: dato $x \in \mathbb{R}^{n \times d}$, trovarne una *buona approssimazione* $z \in \mathbb{R}^{n \times k}$ con $k \ll d$.

- Due approcci:
  - Selezione/**proiezione** delle features:

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, \quad z_i = \Pi(x_i)$$

  - **Generazione**/parametrizzazione (spazio latente):

$$g : \mathbb{R}^k \to \mathbb{R}^d, \quad x_i \approx g(z_i)$$

- Si possono anche combinare (*autoencoders*):

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, g : \mathbb{R}^k \to \mathbb{R}^d, \quad g \circ \Pi(x_i) \approx x_i$$

Riduzione della dimensionalità

Obiettivo: dato $x \in \mathbb{R}^{n \times d}$, trovarne una *buona approssimazione* $z \in \mathbb{R}^{n \times k}$ con $k \ll d$.

- Due approcci:
  - Selezione/**proiezione** delle features:

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, \quad z_i = \Pi(x_i)$$

  - **Generazione**/parametrizzazione (spazio latente):

$$g : \mathbb{R}^k \to \mathbb{R}^d, \quad x_i \approx g(z_i)$$

- Si possono anche combinare (*autoencoders*):

$$\Pi : \mathbb{R}^d \to \mathbb{R}^k, g : \mathbb{R}^k \to \mathbb{R}^d, \quad g \circ \Pi(x_i) \approx x_i$$

- Ci limitiamo a mappe *lineari*:

$$\Pi(x_i) = x_i U + b, \quad g(z_i) = z_i L + \varepsilon$$

Analisi delle componenti principali (PCA)

Consideriamo il caso $k = 1$.

- cerchiamo una proiezione $\Pi(x_i) = x_i v$ dove $v \in \mathbb{R}^d$ con $\|v\| = 1$

Analisi delle componenti principali (PCA)

Consideriamo il caso $k = 1$.

- cerchiamo una proiezione $\Pi(x_i) = x_i v$ dove $v \in \mathbb{R}^d$ con $\|v\| = 1$
- determiniamo $v \rightarrow$ la varianza direzionale sia *massima*:

$$v_1 \in \arg \max_{\|v\|=1} \text{var}(xv)$$

Analisi delle componenti principali (PCA)

Consideriamo il caso $k = 1$.

- cerchiamo una proiezione $\Pi(x_i) = x_i v$ dove $v \in \mathbb{R}^d$ con $\|v\| = 1$
- determiniamo $v \rightarrow$ la varianza direzionale sia *massima*:

$$v_1 \in \arg \max_{\|v\|=1} \text{var}(xv)$$

- Poiché $\text{var}(xv) = v^T \text{var}(x) v$, abbiamo che v_1 è *autovettore* di $\text{var}(x)$ con *autovalore* λ_1 (massimo):

$$\text{var}(x) v_1 = \lambda_1 v_1, \quad \Rightarrow \quad \text{var}(x v_1) = \lambda_1.$$

Analisi delle componenti principali (PCA)

Consideriamo il caso $k = 1$.

- cerchiamo una proiezione $\Pi(x_i) = x_i v$ dove $v \in \mathbb{R}^d$ con $\|v\| = 1$
- determiniamo $v \rightarrow$ la varianza direzionale sia *massima*:

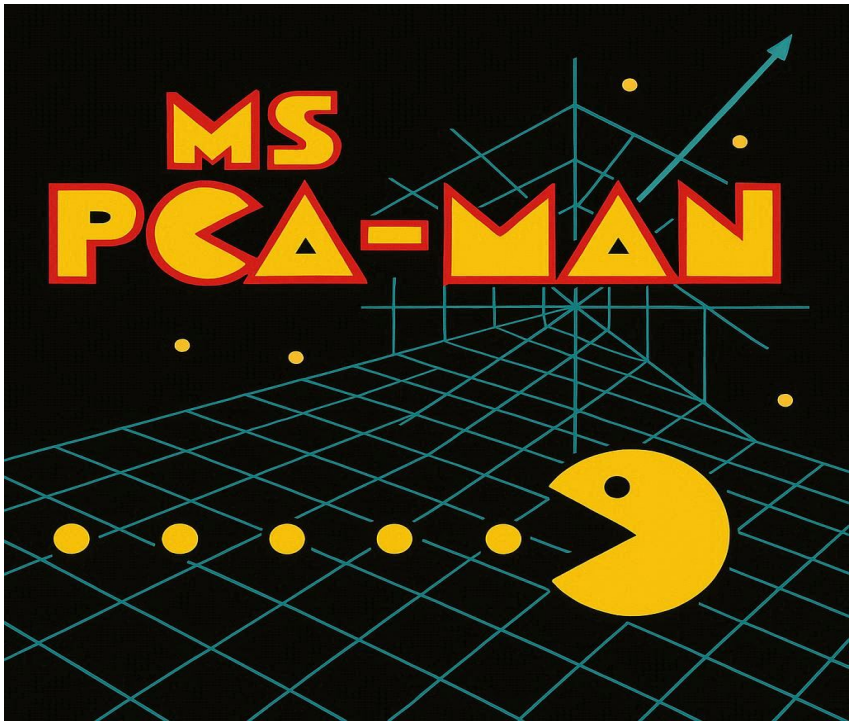
$$v_1 \in \arg \max_{\|v\|=1} \text{var}(xv)$$

- Poiché $\text{var}(xv) = v^T \text{var}(x) v$, abbiamo che v_1 è *autovettore* di $\text{var}(x)$ con *autovalore* λ_1 (massimo):

$$\text{var}(x) v_1 = \lambda_1 v_1, \quad \Rightarrow \quad \text{var}(x v_1) = \lambda_1.$$

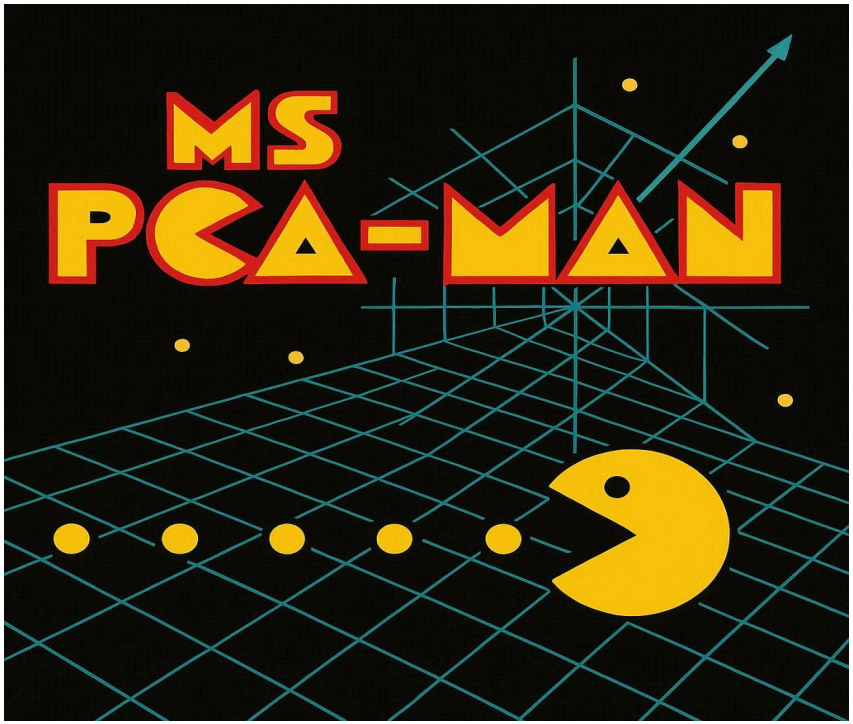
- La direzione v_1 è detta **prima componente principale** di x.

Ms PCA-Man



- Giocate voi! https://dario-trevisan.shinyapps.io/MsPCAMan/

Ms PCA-Man



- Giocate voi! https://dario-trevisan.shinyapps.io/MsPCAMan/
- Codice (R shiny) disponibile nel team e nella pagina del corso

Caso generale

Procediamo *ricorsivamente*: avendo già trovate le prime $k-1$ **componenti principali** $v_1, v_2, \ldots, v_{k-1}$, cerchiamo

$$v_k \in \mathbb{R}^d$$

con $\|v_k\| = 1$ che sia *ortogonale* a tutte le precedenti:

$$v_j^T v_k = 0 \quad \forall j < k$$

e tale che

$$v_k \in \arg \max_{v:\, v_j^T v = 0\, \forall j<k} \text{var}(xv).$$

- Si ha che v_k è autovettore di $\text{var}(x)$ con autovalore λ_k:

$$\lambda_1 = \text{var}(xv_1) \geq \lambda_2 \ldots \geq \lambda_k = \text{var}(xv_k).$$

Teorema spettrale

La procedura termina a $k = d$:

- si trova una **base ortonormale** di $\mathbb{R}^d$ da componenti principali.

Teorema spettrale

La procedura termina a $k = d$:

- si trova una **base ortonormale** di $\mathbb{R}^d$ da componenti principali.
- I vettori $(v_j)_{j=1,\dots,d}$ sono *autovettori* di $\mathrm{var}(x)$.

Teorema spettrale

La procedura termina a $k = d$:

- si trova una **base ortonormale** di $\mathbb{R}^d$ da componenti principali.
- I vettori $(v_j)_{j=1,\dots,d}$ sono *autovettori* di $\mathrm{var}(x)$.
- Questo dimostra il (già noto) *teorema spettrale* per matrici simmetriche.

Teorema spettrale

La procedura termina a $k = d$:

- si trova una **base ortonormale** di $\mathbb{R}^d$ da componenti principali.
- I vettori $(v_j)_{j=1,\dots,d}$ sono *autovettori* di $\mathrm{var}(x)$.
- Questo dimostra il (già noto) *teorema spettrale* per matrici simmetriche.
- Vale

$$\mathrm{cov}(xv_j, xv_k) = v_j^T \,\mathrm{var}(x) v_k = v_j^T v_k \lambda_k = \begin{cases} \lambda_k & \text{se } j = k \\ 0 & \text{altrimenti.} \end{cases}$$

Teorema spettrale

La procedura termina a $k = d$:

- si trova una **base ortonormale** di $\mathbb{R}^d$ da componenti principali.
- I vettori $(v_j)_{j=1,\dots,d}$ sono *autovettori* di $\mathrm{var}(x)$.
- Questo dimostra il (già noto) *teorema spettrale* per matrici simmetriche.
- Vale

$$\mathrm{cov}(xv_j, xv_k) = v_j^T \,\mathrm{var}(x) v_k = v_j^T v_k \lambda_k = \begin{cases} \lambda_k & \text{se } j = k \\ 0 & \text{altrimenti.} \end{cases}$$

- In forma matriciale $V = (v_1, v_2, \dots, v_d)$ (colonne della matrice):

$$\mathrm{var}(xV) = V^T \,\mathrm{var}(x) V = \Lambda.$$

dove Λ è diagonale.

Algoritmo PCA:

- **standardizzare i dati** $x \leftarrow \text{scale}(x)$

Algoritmo PCA:

- **standardizzare i dati** $x \leftarrow \text{scale}(x)$
- calcolare la matrice $\text{var}(x) = \text{cor}(x)$

Algoritmo PCA:

- **standardizzare i dati** $x \leftarrow \text{scale}(x)$
- calcolare la matrice $\text{var}(x) = \text{cor}(x)$
- determinare una base ortonormale $V = (v_1, \ldots, v_d)$ di autovettori (**componenti principali**) con rispettivi autovalori decrescenti

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$$

Algoritmo PCA:

- **standardizzare i dati** $x \leftarrow \text{scale}(x)$
- calcolare la matrice $\text{var}(x) = \text{cor}(x)$
- determinare una base ortonormale $V = (v_1, \ldots, v_d)$ di autovettori (**componenti principali**) con rispettivi autovalori decrescenti
$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$$
- **scegliere** $1 \leq k \leq d$ e porre $U = V_{\leq k} := (v_1, \ldots, v_k) \in \mathbb{R}^{d \times k}$ la matrice delle prime k componenti principali

Algoritmo PCA:

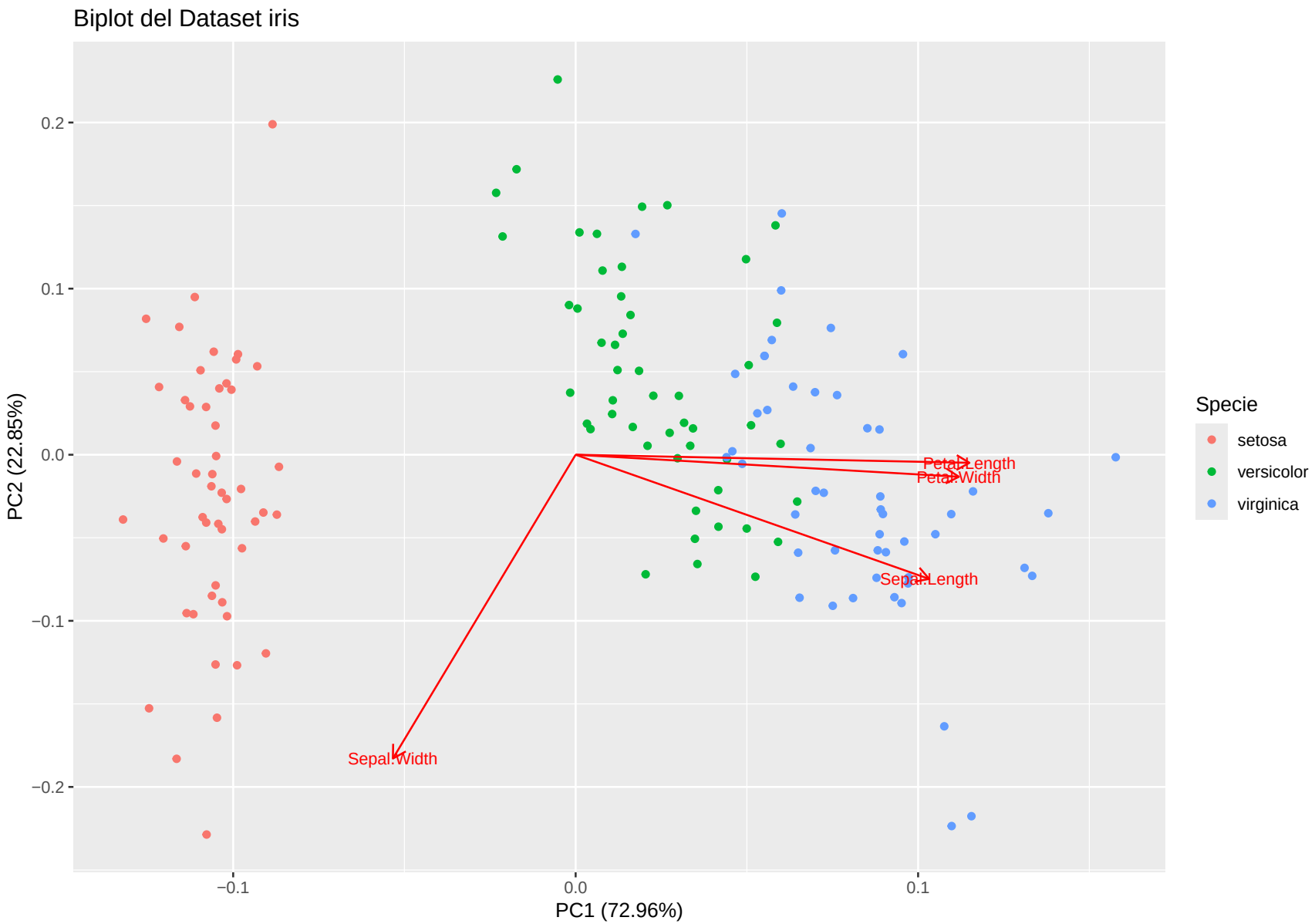
- **standardizzare i dati** $x \leftarrow \text{scale}(x)$
- calcolare la matrice $\text{var}(x) = \text{cor}(x)$
- determinare una base ortonormale $V = (v_1, \ldots, v_d)$ di autovettori (**componenti principali**) con rispettivi autovalori decrescenti

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \geq 0$$

- **scegliere** $1 \leq k \leq d$ e porre $U = V_{\leq k} := (v_1, \ldots, v_k) \in \mathbb{R}^{d \times k}$ la matrice delle prime k componenti principali
- proiettare: $z := xU$ sono nuove features dette **scores** (punteggi):

$$z_{i,\ell} = \sum_{j=1}^{d} x_{i,j} v_{\ell,j} \quad i = 1, \ldots, n, \ell = 1, \ldots, k$$

Biplot


Biplot del Dataset iris
PC2 (22.85%)
PC1 (72.96%)
0.2
0.1
0.0
−0.1
−0.2
−0.1
0.0
0.1
Sepal.Width
Petal.Length
Petal.Width
Sepal.Length
Specie
setosa
versicolor
virginica

Covarianze features/scores, matrice dei loadings

Partendo da $z = xU = xV_{\leq k}$,

- gli scores z sono *non correlati*:

$$\text{var}(z) = V_{\leq k}^T \, \text{var}(x) \, V_{\leq k} = \Lambda_{\leq k}$$

Covarianze features/scores, matrice dei loadings

Partendo da $z = xU = xV_{\leq k}$,

- gli scores z sono *non correlati*:

$$\text{var}(z) = V_{\leq k}^T \, \text{var}(x) \, V_{\leq k} = \Lambda_{\leq k}$$

- **Osservazione**: $\hat{z} := z\Lambda_{\leq k}^{-1/2} = xV_{\leq k}\Lambda_{\leq k}^{-1/2}$ sono standardizzate ma non provengono più da componenti principali!

Covarianze features/scores, matrice dei loadings

Partendo da $z = xU = xV_{\leq k}$,

- gli scores z sono *non correlati*:

$$\text{var}(z) = V_{\leq k}^T \, \text{var}(x) \, V_{\leq k} = \Lambda_{\leq k}$$

- **Osservazione**: $\hat{z} := z\Lambda_{\leq k}^{-1/2} = xV_{\leq k}\Lambda_{\leq k}^{-1/2}$ sono standardizzate ma non provengono più da componenti principali!
- Ciascuna $v_{\ell,j}$ indica l'*associazione* tra z_ℓ e x_j:

$$\text{cov}(x, z) = \text{var}(x) V_{\leq k} = \Lambda_{\leq k} V_{\leq k} = (\lambda_1 v_1, \lambda_2 v_2 \ldots, \lambda_k v_k)$$

Covarianze features/scores, matrice dei loadings

Partendo da $z = xU = xV_{\leq k}$,

- gli scores z sono *non correlati*:

$$\text{var}(z) = V_{\leq k}^T \, \text{var}(x) \, V_{\leq k} = \Lambda_{\leq k}$$

- **Osservazione**: $\hat{z} := z\Lambda_{\leq k}^{-1/2} = xV_{\leq k}\Lambda_{\leq k}^{-1/2}$ sono standardizzate ma non provengono più da componenti principali!
- Ciascuna $v_{\ell,j}$ indica l'*associazione* tra z_ℓ e x_j:

$$\text{cov}(x, z) = \text{var}(x) V_{\leq k} = \Lambda_{\leq k} V_{\leq k} = (\lambda_1 v_1, \lambda_2 v_2 \ldots, \lambda_k v_k)$$

- x è standardizzato $\rightarrow \text{sd}(x_j) = 1$:

$$\text{cor}(x_j, z_\ell) = \frac{\lambda_\ell v_{\ell,j}}{\sqrt{\lambda_\ell}} = \sqrt{\lambda_\ell} v_{\ell,j} = L_{\ell,j}$$

Covarianze features/scores, matrice dei loadings

Partendo da $z = xU = xV_{\leq k}$,

- gli scores z sono *non correlati*:

$$\text{var}(z) = V_{\leq k}^T \, \text{var}(x) V_{\leq k} = \Lambda_{\leq k}$$

- **Osservazione**: $\hat{z} := z\Lambda_{\leq k}^{-1/2} = xV_{\leq k}\Lambda_{\leq k}^{-1/2}$ sono standardizzate ma non provengono più da componenti principali!
- Ciascuna $v_{\ell,j}$ indica l'*associazione* tra z_ℓ e x_j:

$$\text{cov}(x, z) = \text{var}(x) V_{\leq k} = \Lambda_{\leq k} V_{\leq k} = (\lambda_1 v_1, \lambda_2 v_2 \ldots, \lambda_k v_k)$$

- x è standardizzato $\rightarrow \text{sd}(x_j) = 1$:

$$\text{cor}(x_j, z_\ell) = \frac{\lambda_\ell v_{\ell,j}}{\sqrt{\lambda_\ell}} = \sqrt{\lambda_\ell} v_{\ell,j} = L_{\ell,j}$$

- $L = \sqrt{\Lambda_{\leq k}} V_{\leq k}^T \in \mathbb{R}^{k \times d}$ è detta *matrice dei loadings*.

Criteri per determinare k

Come per il clustering, non c'è un numero *giusto* di componenti principali da selezionare: dipende dalla applicazione (intepretabilità, applicabilità di algoritmi, ecc.)

Vediamo alcuni criteri:

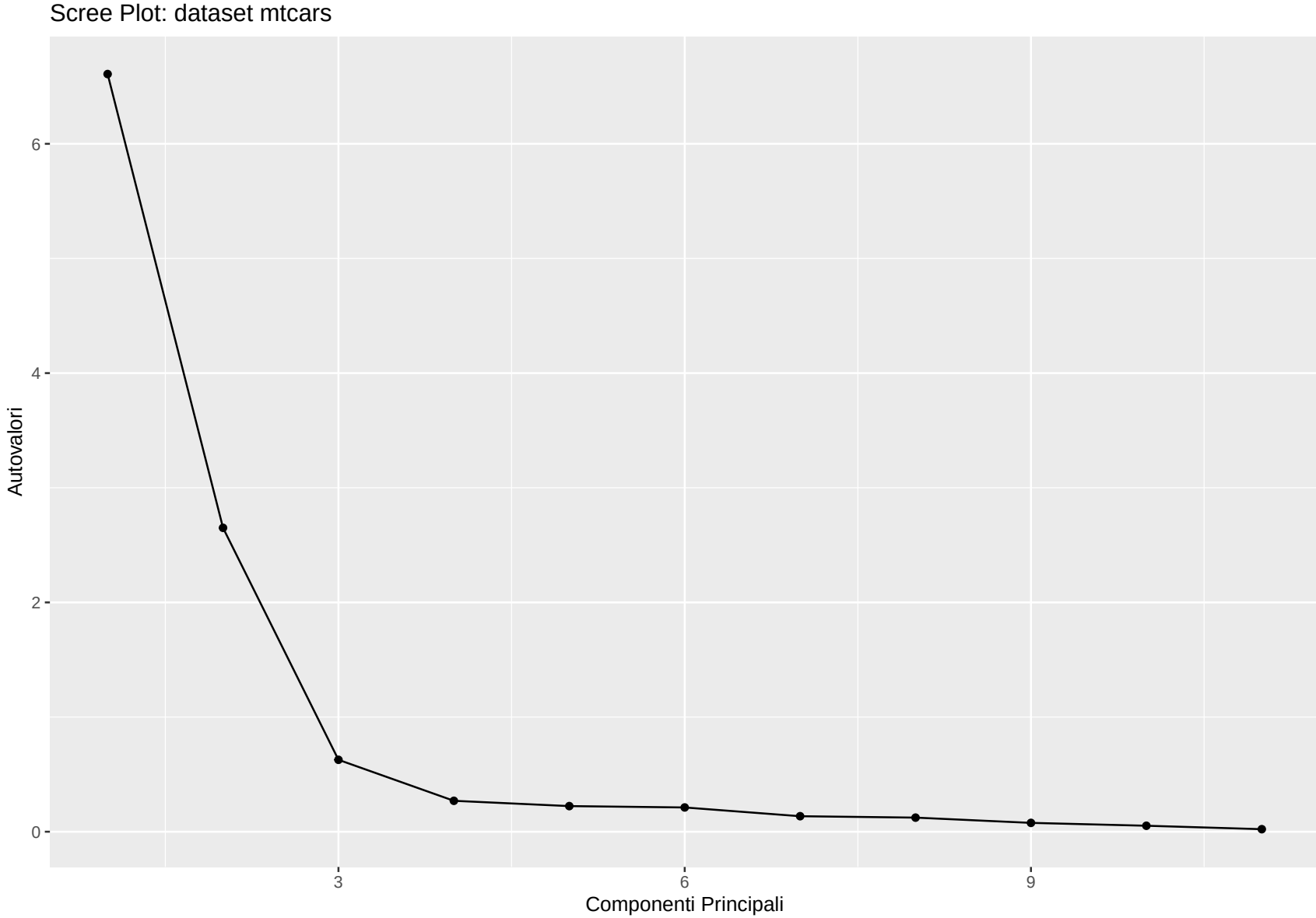
- **Scree plot**: è *metodo elbow* applicato al grafico che che mostra gli autovalori delle componenti principali in ordine decrescente.

Criteri per determinare k

Come per il clustering, non c'è un numero *giusto* di componenti principali da selezionare: dipende dalla applicazione (intepretabilità, applicabilità di algoritmi, ecc.)

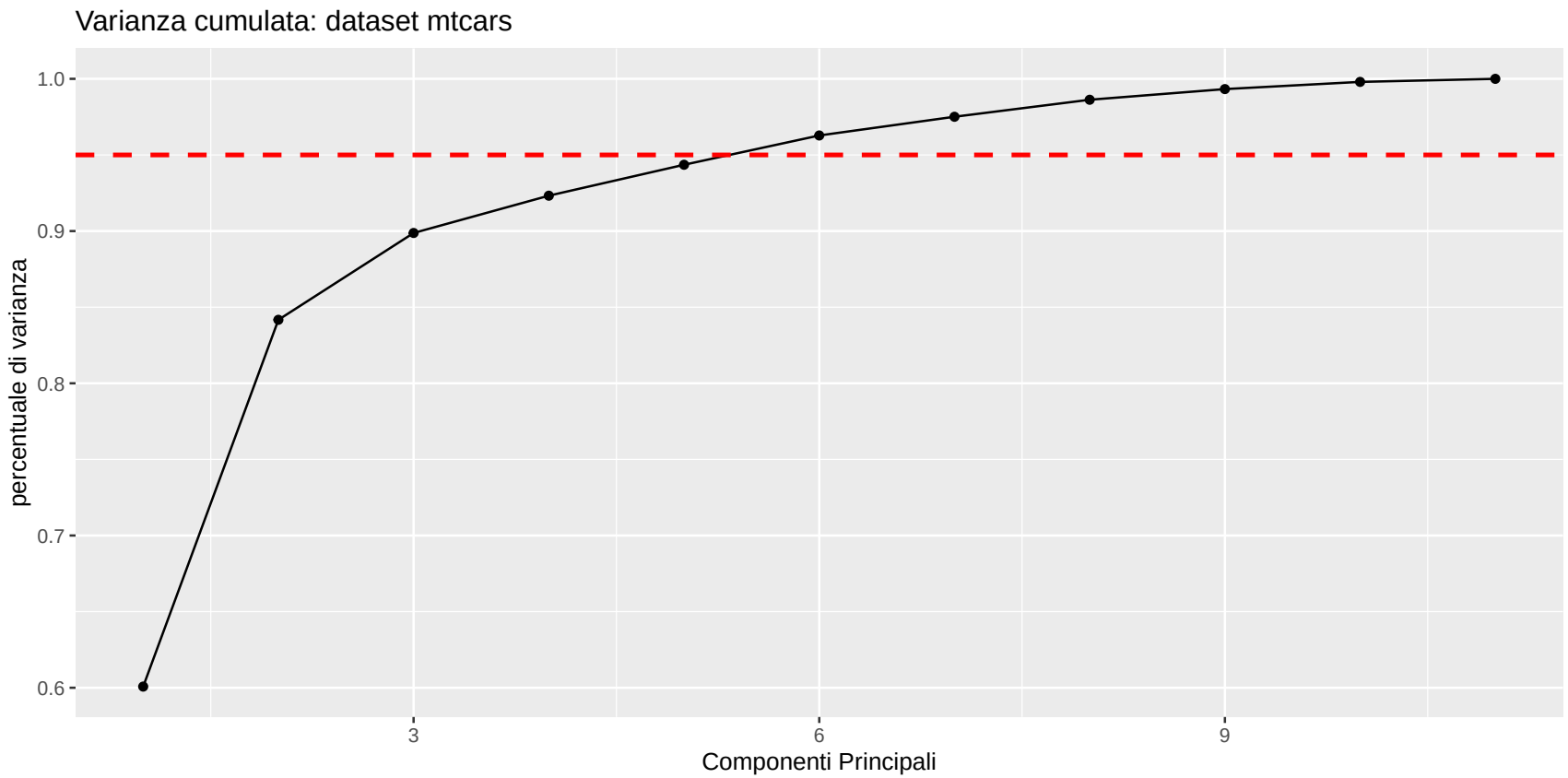
Vediamo alcuni criteri:

- **Scree plot**: è *metodo elbow* applicato al grafico che che mostra gli autovalori delle componenti principali in ordine decrescente.
- **Criterio di Kaiser**: selezionare solo le componenti con autovalori > 1.


Scree Plot: dataset mtcars
Autovalori
6
4
2
0
3
6
9
Componenti Principali

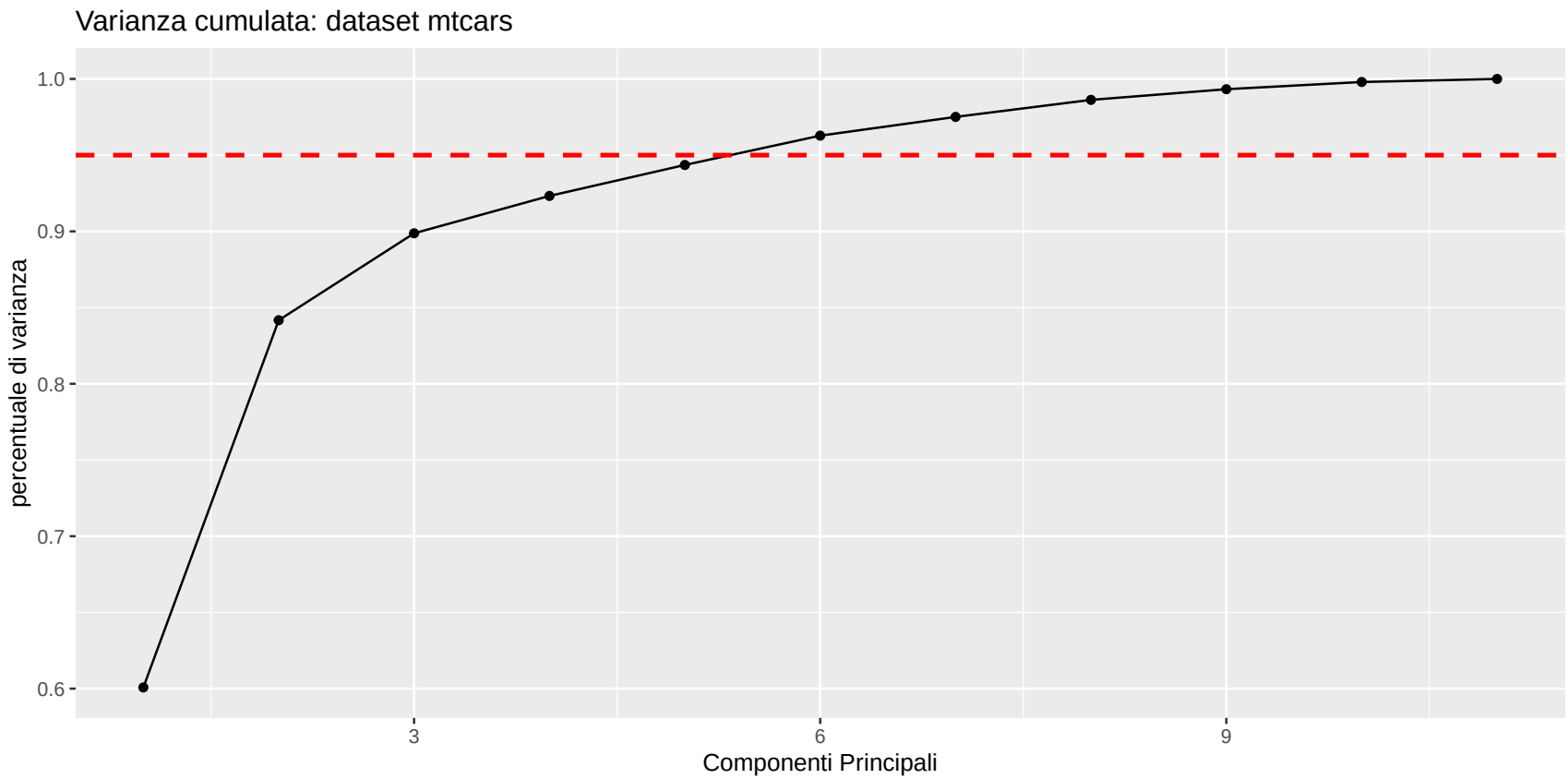
Criterio della varianza cumulata

- Calcolare la percentuale cumulativa della varianza *spiegata* dalle componenti principali.



Criterio della varianza cumulata

- Calcolare la percentuale cumulativa della varianza *spiegata* dalle componenti principali.
- Scegliere il numero di componenti che spiegano una percentuale predeterminata di varianza (es. 95%).



Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:
  1. $z \leftarrow$ fattori (variabili latenti),

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:
  1. $z \leftarrow$ fattori (variabili latenti),
  2. $L \leftarrow$ matrice dei loadings (saturazioni) dei fattori

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:
  1. $z \leftarrow$ fattori (variabili latenti),
  2. $L \leftarrow$ matrice dei loadings (saturazioni) dei fattori
  3. $\varepsilon \leftarrow$ errori o residui

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:
  1. $z \leftarrow$ fattori (variabili latenti),
  2. $L \leftarrow$ matrice dei loadings (saturazioni) dei fattori
  3. $\varepsilon \leftarrow$ errori o residui
- Nel caso $k = d$, possiamo invertire $U^{-1} = V^T$ e troviamo esattamente

$$x = zV^T$$

Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

$$z = xU, \quad U = V_{\leq k}$$

- cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

- *Terminologia*:
  1. $z \leftarrow$ fattori (variabili latenti),
  2. $L \leftarrow$ matrice dei loadings (saturazioni) dei fattori
  3. $\varepsilon \leftarrow$ errori o residui
- Nel caso $k = d$, possiamo invertire $U^{-1} = V^T$ e troviamo esattamente

$$x = zV^T$$

- **Problema**: come fare se $k < d$?

$$x = zV^T$$

- Decomponiamo $V = (V_{\leq k}, V_{>k})$.

$$x = zV^T$$

- Decomponiamo $V = (V_{\leq k}, V_{>k})$.
- Decomponiamo

$$(z_{\leq k}, z_{>k}) = (xV_{\leq k}, xV_{>k})$$

troviamo

$$x = (z_{\leq k}, z_{>k})(V_{\leq k}, V_{>k})^T = z_{\leq k} V_{\leq k}^T + z_{>k} V_{>k}^T = zL + \varepsilon$$

$$x = zV^T$$

- Decomponiamo $V = (V_{\leq k}, V_{>k})$.
- Decomponiamo

$$(z_{\leq k}, z_{>k}) = (xV_{\leq k}, xV_{>k})$$

troviamo

$$x = (z_{\leq k}, z_{>k})(V_{\leq k}, V_{>k})^T = z_{\leq k}V_{\leq k}^T + z_{>k}V_{>k}^T = zL + \varepsilon$$

- Se definiamo i fattori, loadings ed errori

$$z := z_{\leq k}\Lambda_{\leq k}^{-1/2}, \quad L := \Lambda_{\leq k}^{1/2}V_{\leq k}^T, \quad \varepsilon := z_{>k}V_{>k}^T = xV_{>k}V_{>k}^T$$

Proprietà della EFA mediante PCA

La decomposizione $x = zL + \varepsilon$ è tale che

- $\text{var}(z) = I_{k \times k}$

Proprietà della EFA mediante PCA

La decomposizione $x = zL + \varepsilon$ è tale che

- $\text{var}(z) = I_{k \times k}$
- $\text{cov}(z, \varepsilon) = \text{cov}(z_{\leq k} \Lambda_{\leq k}^{-1/2}, z_{>k} V_{>k}^T) = 0_{k \times d}$

Proprietà della EFA mediante PCA

La decomposizione $x = zL + \varepsilon$ è tale che

- $\text{var}(z) = I_{k \times k}$
- $\text{cov}(z, \varepsilon) = \text{cov}(z_{\leq k} \Lambda_{\leq k}^{-1/2}, z_{>k} V_{>k}^T) = 0_{k \times d}$
- $\text{var}(\varepsilon) = \text{var}(z_{>k} V_{>k}^T, z_{>k} V_{>k}^T) = V_{>k} \Lambda_{>k} V_{>k}^T$

Proprietà della EFA mediante PCA

La decomposizione $x = zL + \varepsilon$ è tale che

- $\text{var}(z) = I_{k \times k}$
- $\text{cov}(z, \varepsilon) = \text{cov}(z_{\leq k} \Lambda_{\leq k}^{-1/2}, z_{>k} V_{>k}^T) = 0_{k \times d}$
- $\text{var}(\varepsilon) = \text{var}(z_{>k} V_{>k}^T, z_{>k} V_{>k}^T) = V_{>k} \Lambda_{>k} V_{>k}^T$
- $\text{cov}(z, x) = \text{cov}(z, zL + \varepsilon) = L$

Proprietà della EFA mediante PCA

La decomposizione $x = zL + \varepsilon$ è tale che

- $\text{var}(z) = I_{k \times k}$
- $\text{cov}(z, \varepsilon) = \text{cov}(z_{\leq k} \Lambda_{\leq k}^{-1/2}, z_{>k} V_{>k}^T) = 0_{k \times d}$
- $\text{var}(\varepsilon) = \text{var}(z_{>k} V_{>k}^T, z_{>k} V_{>k}^T) = V_{>k} \Lambda_{>k} V_{>k}^T$
- $\text{cov}(z, x) = \text{cov}(z, zL + \varepsilon) = L$
- Inoltre minimizza un rischio empirico:

$$L_{PCA} \in \arg\min_{L} \sum_{i=1}^{n} \min_{z_i \in \mathbb{R}^k, \|z_i\|=1} \|x_i - z_i L\|^2$$

Definizione precisa di EFA

La rappresentazione $x = zL + \varepsilon$ trovata tramite PCA *non* è propriamente quella dell'analisi fattoriale esplorativa.

- Si richiede per la EFA:

Definizione precisa di EFA

La rappresentazione $x = zL + \varepsilon$ trovata tramite PCA *non* è propriamente quella dell'analisi fattoriale esplorativa.

- Si richiede per la EFA:
  - $\text{var}(z) = I_{k \times k}$

Definizione precisa di EFA

La rappresentazione $x = zL + \varepsilon$ trovata tramite PCA *non* è propriamente quella dell'analisi fattoriale esplorativa.

- Si richiede per la EFA:
  - $\text{var}(z) = I_{k \times k}$
  - $\text{cov}(z, \varepsilon) = 0_{k \times d}$

Definizione precisa di EFA

La rappresentazione $x = zL + \varepsilon$ trovata tramite PCA *non* è propriamente quella dell'analisi fattoriale esplorativa.

- Si richiede per la EFA:
  - $\text{var}(z) = I_{k \times k}$
  - $\text{cov}(z, \varepsilon) = 0_{k \times d}$
  - $\text{var}(\varepsilon) = \Psi \in \mathbb{R}^{d \times d}$ **diagonale**: gli *errori* di features diverse sono *non correlati*

Definizione precisa di EFA

La rappresentazione $x = zL + \varepsilon$ trovata tramite PCA *non* è propriamente quella dell'analisi fattoriale esplorativa.

- Si richiede per la EFA:
  - $\text{var}(z) = I_{k \times k}$
  - $\text{cov}(z, \varepsilon) = 0_{k \times d}$
  - $\text{var}(\varepsilon) = \Psi \in \mathbb{R}^{d \times d}$ **diagonale**: gli *errori* di features diverse sono *non correlati*
- **Non** si richiede che le features x siano standardizzate (ma si possono centrare lavorando con $x - \bar{x}$). Vale comunque

$$L = \text{cov}(z, x)$$

Equazione fondamentale della EFA

Partendo dalle ipotesi della EFA troviamo l'equazione matriciale

$$\text{var}(x) = \text{var}(zL + \varepsilon) = L^T L + \Psi.$$

- sono $d(d+1)/2$ equazioni *non lineari* in $kd + d$ incognite.

Equazione fondamentale della EFA

Partendo dalle ipotesi della EFA troviamo l'equazione matriciale

$$\text{var}(x) = \text{var}(zL + \varepsilon) = L^T L + \Psi.$$

- sono $d(d+1)/2$ equazioni *non lineari* in $kd + d$ incognite.
- In *pratica* si trovano soluzioni approssimate ammettendo Ψ non esattamente diagonale (come ad esempio quella trovata con la PCA).

Equazione fondamentale della EFA

Partendo dalle ipotesi della EFA troviamo l'equazione matriciale

$$\text{var}(x) = \text{var}(zL + \varepsilon) = L^T L + \Psi.$$

- sono $d(d+1)/2$ equazioni *non lineari* in $kd + d$ incognite.
- In *pratica* si trovano soluzioni approssimate ammettendo Ψ non esattamente diagonale (come ad esempio quella trovata con la PCA).
- **Definizioni**: Per ciascuna feature x_j, $j = 1, \ldots, d$:

Equazione fondamentale della EFA

Partendo dalle ipotesi della EFA troviamo l'equazione matriciale

$$\text{var}(x) = \text{var}(zL + \varepsilon) = L^T L + \Psi.$$

- sono $d(d+1)/2$ equazioni *non lineari* in $kd + d$ incognite.
- In *pratica* si trovano soluzioni approssimate ammettendo Ψ non esattamente diagonale (come ad esempio quella trovata con la PCA).
- **Definizioni**: Per ciascuna feature x_j, $j = 1, \ldots, d$:
  - $\text{var}\,\varepsilon_j = \Psi_{jj}$ è detta *unicità* (uniqueness)

Equazione fondamentale della EFA

Partendo dalle ipotesi della EFA troviamo l'equazione matriciale

$$\operatorname{var}(x) = \operatorname{var}(zL + \varepsilon) = L^T L + \Psi.$$

- sono $d(d+1)/2$ equazioni *non lineari* in $kd + d$ incognite.
- In *pratica* si trovano soluzioni approssimate ammettendo Ψ non esattamente diagonale (come ad esempio quella trovata con la PCA).
- **Definizioni**: Per ciascuna feature x_j, $j = 1, \ldots, d$:
  - $\operatorname{var} \varepsilon_j = \Psi_{jj}$ è detta *unicità* (uniqueness)
  - $\operatorname{var}(x_j) - \operatorname{var}(\varepsilon_j) = \sum_{\ell=1}^{k} L_{\ell,j}^2$ è detta *comunalità* (communality)

Rotazioni

Trovata una rappresentazione EFA: $x = zL + \varepsilon$, ve ne sono *infinite altre* (matematicamente) equivalenti.

- Data una matrice ortogonale (rotazione) $R \in \mathbb{R}^{k \times k}$, $R^T R = I_{k \times k}$ trasformiamo i fattori e i loadings

$$\hat{z} := zR \quad \hat{L} := R^T L$$

Rotazioni

Trovata una rappresentazione EFA: $x = zL + \varepsilon$, ve ne sono *infinite altre* (matematicamente) equivalenti.

- Data una matrice ortogonale (rotazione) $R \in \mathbb{R}^{k \times k}$, $R^T R = I_{k \times k}$ trasformiamo i fattori e i loadings

$$\hat{z} := zR \quad \hat{L} := R^T L$$

- continuano a valere: $x = \hat{z}\hat{L} + \varepsilon$ e

$$\text{var}(\hat{z}) = R^T \text{var}(z) R = R^T R = I_{k \times k}$$

Rotazioni

Trovata una rappresentazione EFA: $x = zL + \varepsilon$, ve ne sono *infinite altre* (matematicamente) equivalenti.

- Data una matrice ortogonale (rotazione) $R \in \mathbb{R}^{k \times k}$, $R^T R = I_{k \times k}$ trasformiamo i fattori e i loadings

$$\hat{z} := zR \quad \hat{L} := R^T L$$

- continuano a valere: $x = \hat{z}\hat{L} + \varepsilon$ e

$$\text{var}(\hat{z}) = R^T \text{var}(z) R = R^T R = I_{k \times k}$$

- Per agevolare l'**interpretazione** dei fattori come *caratteristiche latenti* cerchiamo una R in modo che i fattori ruotate $R^T L$ siano una **matrice sparsa** (con tanti 0). Ci sono diversi **criteri**: *varimax*, *quartimax*. . . .