

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 $\{x_i\}_{i=1,\dots,n}$ dove ciascuna $x_i \in \mathbb{R}^d$ (d caratteristiche/features per individuo).

Curse of Dimensionality

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

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

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

$$x_i = (x_{i,j})_{j=1 \dots d}$$

$$y_i = (y_{i,j})_{j=1 \dots d}$$

$$\|x_i - y_i\|^2 = \sum_{j=1}^d (x_{i,j} - y_{i,j})^2$$

"Signal" è in k componenti

"rumore" sono le rimanenti σ^2

$$\|x_i - y_i\|^2 \approx \text{signal} + (d-k)\sigma^2$$

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 = 10x2, dimensione d = 10"
##
##      classi originarie
## cluster individuati  0  1
##      1  0 10
##      2 10  0
## [1] ""
## [1] "Taglia n = 10x2, 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:

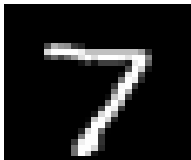
- ▶ in scala di grigio (valori da 0 a 255)
- ▶ ciascuna composta da $28 \times 28 = 784$ pixels $d = 784$

Esempio MNIST

Il dataset MNIST contiene immagini di cifre scritte a mano:

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

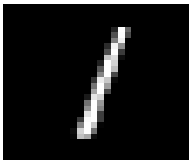
Label: 7



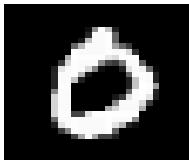
Label: 2



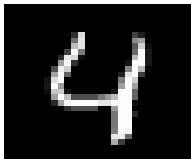
Label: 1



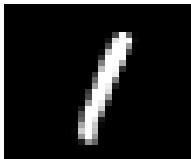
Label: 0



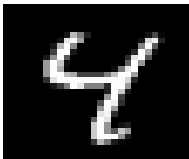
Label: 4



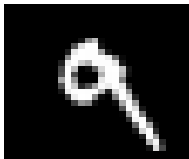
Label: 1



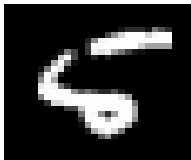
Label: 4



Label: 9



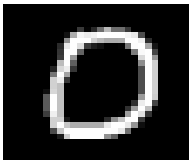
Label: 5



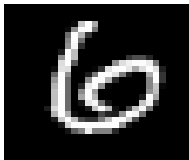
Label: 9



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

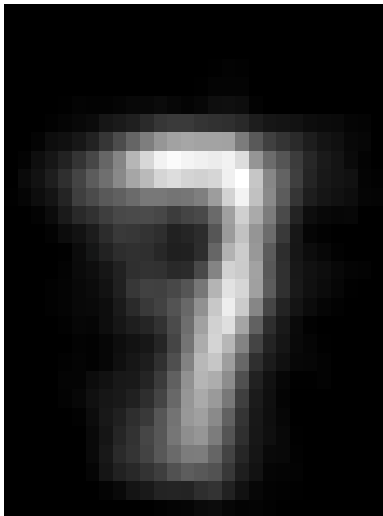
Consideriamo un sottoinsieme di MNIST contenente

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

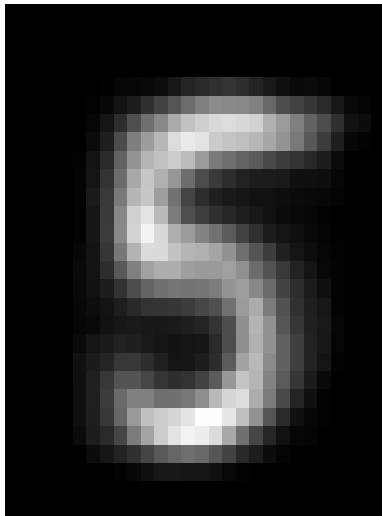
```
##                                classi originarie
## cluster individuati    5    7
##                        1  21 100
##                        2  79   0
```

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

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

Varianza campionaria

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

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

- 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

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

- 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 $(n-1) \text{var}(x)$.

Varianza campionaria

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

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

- ▶ 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 $(n-1) \text{var}(x)$.
- ▶ **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 $v \in \mathbb{R}^d$

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

esempio

$$e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad x: e_1 = x_{:,1}$$

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

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

- la varianza (campionaria) direzionale è

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

$$\overline{xv} = \frac{1}{n} \sum_{i=1}^n x_i v = \left(\frac{1}{n} \sum_{i=1}^n x_i \right) v = \bar{x} v$$

↑
centroide

Matrice delle covarianze

► Vale

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

$$(n-1) \text{var}(xv) = \sum_{i=1}^n (x_i v - \bar{x} v)^2 = \sum_{i=1}^n \left[(x_i v)^2 + (\bar{x} v)^2 - 2(x_i v)(\bar{x} v) \right]$$

$$\begin{aligned} (x_i v)^2 &= (x_i v)^T (x_i v) = V^T \underbrace{x_i^T x_i}_{\text{matrice } d \times d} V & (A^T)_{ij} &= A_{ji} \\ &= \sum_{i=1}^n V^T x_i^T x_i V - V^T \bar{x}^T \bar{x} V - V^T x_i^T \bar{x} V - V^T \bar{x}^T x_i V & (AB)^T &= B^T A^T \\ &= V^T \text{var}(x) V (n-1) \end{aligned}$$

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

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

- 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
 - *quadrata* di taglia $d \times d$

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

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

Se traspongo rimane uguale

- $\text{var}(x)$ è una matrice

- quadrata di taglia $d \times d$

- simmetrica

$$(\text{var}(x))^T = \text{var}(x)$$

Matrice delle covarianze

- Vale

$$\overline{xv} = \bar{x}, \quad \text{var}(xv) = v^T \text{var}(x) v \geq 0$$

- 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

- *quadrata* di taglia $d \times d$
- simmetrica
- **semidefinita positiva.**

$$: \quad \forall v \in \mathbb{R}^d \quad v^T \text{var}(x) v \geq 0$$

$\uparrow$
 $\text{var}(x)$

Covarianza

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

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

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

Covarianza

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

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

- ▶ 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

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

Covarianza

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

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

- ▶ 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

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

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

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

Bilinearità della covarianza

- La covarianza è *bilineare*:

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

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

Bilinearità della covarianza

$$\text{cov}(x, y) = \text{cov}(y, x)$$

- La covarianza è *bilineare*:

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

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

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

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

$$\text{var}(x + x') = \text{cov}(x + x', x + x')$$

|

=



Bilinearità della covarianza

- La covarianza è *bilineare*:

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

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

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

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

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

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

Coefficiente di correlazione di Pearson

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

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

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

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

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

$$\boxed{\text{sd}(x_1) \text{sd}(x_2) \geq |\text{cov}(x_1, x_2)|}$$

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

Coefficiente di correlazione di Pearson

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

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

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

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

- Definiamo il **coefficiente di correlazione**:

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

Coefficiente di correlazione di Pearson

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

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

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

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

- ▶ Definiamo il **coefficiente di correlazione**:

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

- ▶ 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**

$$\text{cor}(x) = \text{SD}(x)^{-1} \overset{\text{Var}(x)}{\text{cov}(x)} \text{SD}(x)^{-1} \in \mathbb{R}^{d \times d}$$

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

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

$$\text{SD}(x) = \begin{pmatrix} \text{sd}(x_1) & & 0 \\ & \ddots & \\ 0 & & \text{sd}(x_d) \end{pmatrix}$$

$$\text{SD}^{-1}(x) = \begin{pmatrix} \text{sd}(x_1)^{-1} & & 0 \\ & \ddots & \\ 0 & & \text{sd}(x_d)^{-1} \end{pmatrix}$$

Matrice di correlazione

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

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

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

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

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

Matrice di correlazione

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

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

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

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

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

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

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

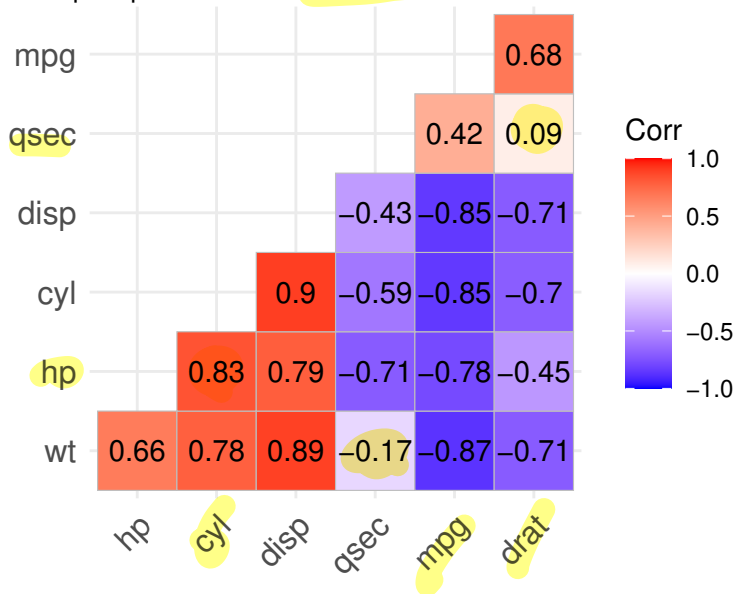
$$\text{SD}(x)_{jj} = \text{sd}(x_j).$$

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

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

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:

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

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:

PCA ► Selezione/**proiezione** delle features:

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

EFA ► **Generazione**/parametrizzazione (spazio latente):

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

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:

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

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

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

► Si possono anche combinare (**autoencoders**):

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

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:

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

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

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

► Si possono anche combinare (*autoencoders*):

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

► Ci limitiamo a mappe *lineari*:

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

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):

$$\text{var}(x) v_1 = \lambda_1 v_1, \quad \Rightarrow \quad \text{var}(xv_1) = \lambda_1.$$

$$v_1^T \underbrace{\text{var}(x) v_1}_{\lambda_1 v_1} = v_1^T (\lambda_1 v_1) = \lambda_1 \underbrace{v_1^T v_1}_{\|v_1\|^2 = 1} = \lambda_1$$

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*:

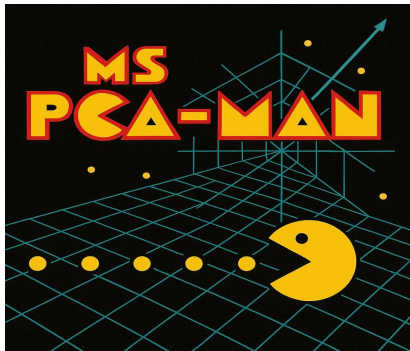
$$v_1 \in \arg \max_{\|v\|=1} \text{var}(xv)$$

- ▶ 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):

$$\text{var}(x)v_1 = \lambda_1 v_1, \quad \Rightarrow \quad \text{var}(xv_1) = \lambda_1.$$

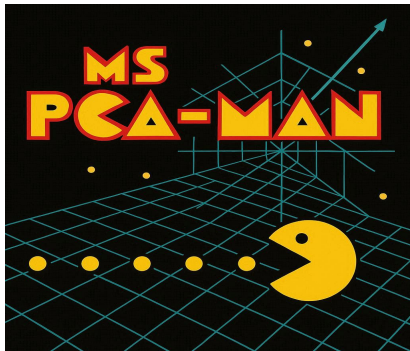
- ▶ La direzione v_1 è detta **prima componente principale** di x .

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** $v_1, v_2, \dots, v_{k-1}$, 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 :

$$\lambda_1 = \text{var}(xv_1) \geq \lambda_2 \geq \dots \geq \lambda_k = \text{var}(xv_k) \geq 0$$

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 $\text{var}(x)$.

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 $\text{var}(x)$.
- ▶ 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 $\text{var}(x)$.
- ▶ Questo dimostra il (già noto) *teorema spettrale* per matrici simmetriche.
- ▶ Vale

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

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 $\text{var}(x)$.
- ▶ Questo dimostra il (già noto) *teorema spettrale* per matrici simmetriche.
- ▶ Vale

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

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

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

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 $V = (v_1, \dots, v_d)$ di autovettori (**componenti principali**) con rispettivi autovalori decrescenti

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$$

Algoritmo PCA:

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

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$$

- ▶ **scegliere** $1 \leq k \leq d$ e porre $U = V_{\leq k} := (v_1, \dots, v_k) \in \mathbb{R}^{d \times k}$
la matrice delle prime k componenti principali

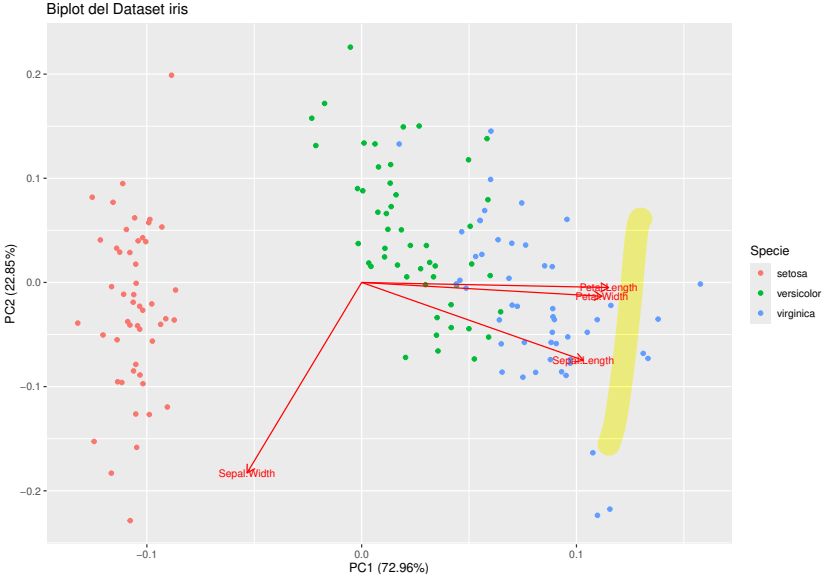
Algoritmo PCA:

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

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_d \geq 0$$

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

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



Covarianze features/scores, matrice dei loadings

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

- ▶ gli scores z sono *non correlati*:

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

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 :

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

$$V = (v_1 v_2 \dots v_k)$$

$$v_{\ell,j}$$

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 :

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

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

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

colonna ℓ riga 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 :

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

- ▶ 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 (interpretabilità, applicabilità di algoritmi, ecc.)

Vediamo alcuni criteri:

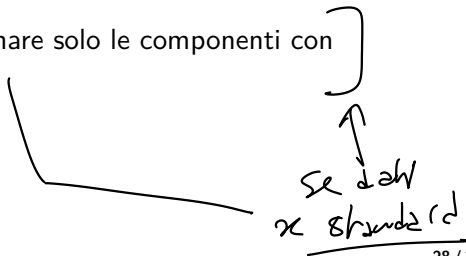
- ▶ **Scree plot:** è *metodo elbow* applicato al grafico che mostra gli autovalori delle componenti principali in ordine decrescente.

Criteri per determinare k

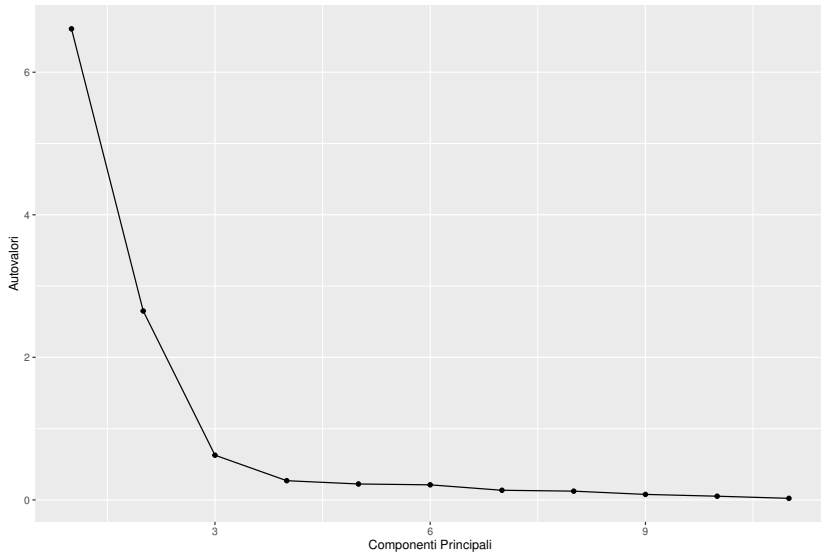
Come per il clustering, non c'è un numero *giusto* di componenti principali da selezionare: dipende dalla applicazione (interpretabilità, applicabilità di algoritmi, ecc.)

Vediamo alcuni criteri:

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

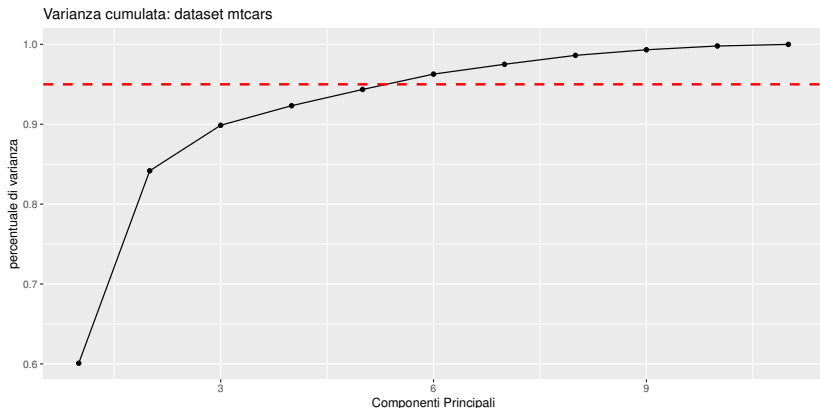


Scree Plot: dataset mtcars



Criterio della varianza cumulata

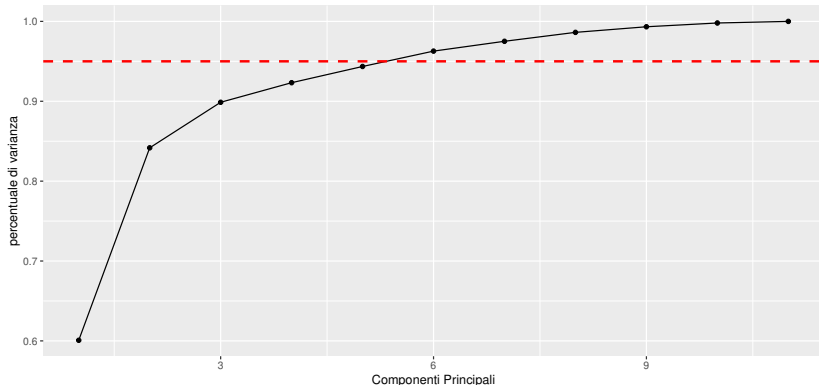
- Calcolare la percentuale cumulativa della varianza *spiegata* dalle componenti principali.



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%).

Varianza cumulata: dataset mtcars



Dalla PCA all'analisi fattoriale esplorativa (EFA)

Nella PCA abbiamo *proiettato*

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

► cerchiamo un modello *generativo*:

$$x = zL + \varepsilon$$

$z \in \mathbb{R}^k$
 $\varepsilon \in \mathbb{R}^d$

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 $L \in \mathbb{R}^{k \times d}$

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 $k < d$?

$$\underline{P(A)} \quad k=d \rightarrow \boxed{x = zV^T}$$

► Decomponiamo $V = (V_{\leq k}, V_{> k})$. $\leftarrow$ se $k < d$

$$x = zV^T$$

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

$$z = (z_{\leq k}, z_{> k}) = (\overbrace{xV_{\leq k}}^{\text{target}}, \overbrace{xV_{> k}}^{\text{bucko}})$$

troviamo

$$x = \underbrace{(z_{\leq k}, z_{> k})}_z \underbrace{(V_{\leq k}, V_{> k})^T}_{\begin{pmatrix} V_{\leq k}^T \\ V_{> k}^T \end{pmatrix}} = z_{\leq k} V_{\leq k}^T + z_{> k} V_{> k}^T = zL + \varepsilon$$

$$x = zV^T$$

- ▶ 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

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

↑
 $z_{\leq k}$ standardizzato

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

► $\text{var}(z) = I_{k \times k}$

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

$\uparrow \quad \quad \quad \uparrow$
 $I_k \quad \quad \quad I_{d-k}$

$$\Lambda_{\leq k}^{-1/2} \underbrace{\text{cov}(z_{\leq k}, z_{> k})}_{= 0} V_{> k}^T$$

Proprietà della EFA mediante PCA

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

► $\text{var}(z) = I_{k \times k}$

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

► $\text{var}(\varepsilon) = \text{var}(z_{> k} V_{> k}^T, z_{> k} V_{> k}^T) = V_{> k} \Lambda_{> k} V_{> k}^T$

||

$V_{> k} \underbrace{\text{var}(z_{> k})}_{\Lambda_{> k}} V_{> k}^T$

↑ non diagonale

Proprietà della EFA mediante PCA

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

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

$$\begin{aligned}
 & \text{cov}(z, zL) + \cancel{\text{cov}(z, \varepsilon)} \\
 & \text{var}(z) L \\
 & L
 \end{aligned}$$

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:

► $\text{var}(z) = I_{k \times k} \leftarrow \text{standard}$

► $\text{cov}(z, \varepsilon) = 0_{k \times d}$

► $\text{var}(\varepsilon) = \Psi \in \mathbb{R}^{d \times d}$ **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

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

Se x standard $\leadsto$ $L = \text{cor}(z, x)$

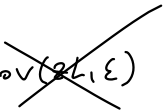
Equazione fondamentale della EFA

$$x = zL + \varepsilon$$

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.

$$\begin{aligned} \text{var}(zL + \varepsilon) &= \text{var}(zL) + \text{var}(\varepsilon) + 2\text{cov}(zL, \varepsilon) \\ &\stackrel{!}{=} L^T \underset{\substack{\text{"} \\ \text{Id}}}{\text{var}(z)} L + \Psi \end{aligned}$$


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 , $j = 1, \dots, d$:

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 , $j = 1, \dots, d$:
 - ▶ $\text{var } \varepsilon_j = \Psi_{jj}$ è detta *unicità* (uniqueness)

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 , $j = 1, \dots, d$:
 - ▶ $\text{var } \varepsilon_j = \Psi_{jj}$ è detta *unicità* (uniqueness)
 - ▶ $\text{var}(x_j) - \text{var}(\varepsilon_j) = \sum_{\ell=1}^k L_{\ell,j}^2$ 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

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

$$\begin{aligned} x &= zL + \varepsilon \\ &= \underbrace{zR}_{\hat{z}} \underbrace{R^T L}_{\hat{L}} + \varepsilon \\ &= \hat{z} \hat{L} + \varepsilon \end{aligned}$$

$$\text{Var}(\hat{z}) = \text{Var}(zR) = R^T \text{Var}(z) R = R^T I_d R = R^T R = I_d$$

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*...