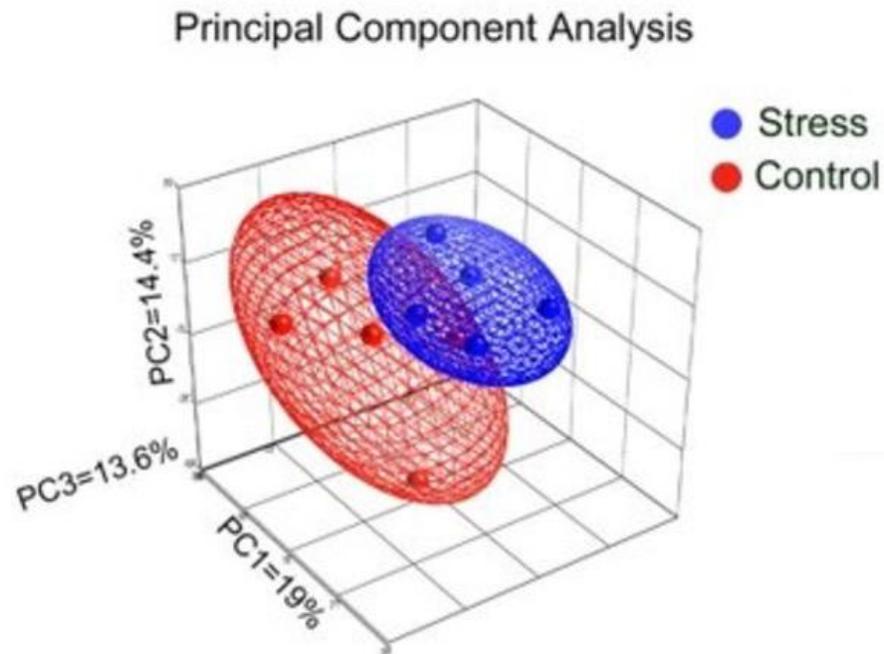


# PRINCIPAL COMPONENT ANALYSIS - PCA



## Cos'è l'Analisi delle Componenti Principali (PCA)?

- Tecnica statistica utilizzata per la **riduzione della dimensionalità**.
- Obiettivo: trasformare un set di dati con molte variabili correlate in un set più piccolo **di variabili non correlate** (componenti principali).
- Ogni componente principale rappresenta una combinazione lineare delle variabili originali.

**Lo scopo principale della PCA è ridurre la complessità dei dati, mantenendo quanta più informazione possibile. In particolare, permette di:**

- **Ridurre la dimensionalità:** Permette di rappresentare i dati con meno variabili, facilitando l'analisi ed eliminando informazioni ridondanti.
- **Migliorare l'efficienza:** Riducendo il numero di variabili, si riducono i tempi computazionali e si migliorano le prestazioni di modelli di machine learning.
- **Facilitare la visualizzazione:** Proietta i dati multidimensionali in 2 o 3 dimensioni, rendendo possibile una visualizzazione più intuitiva.

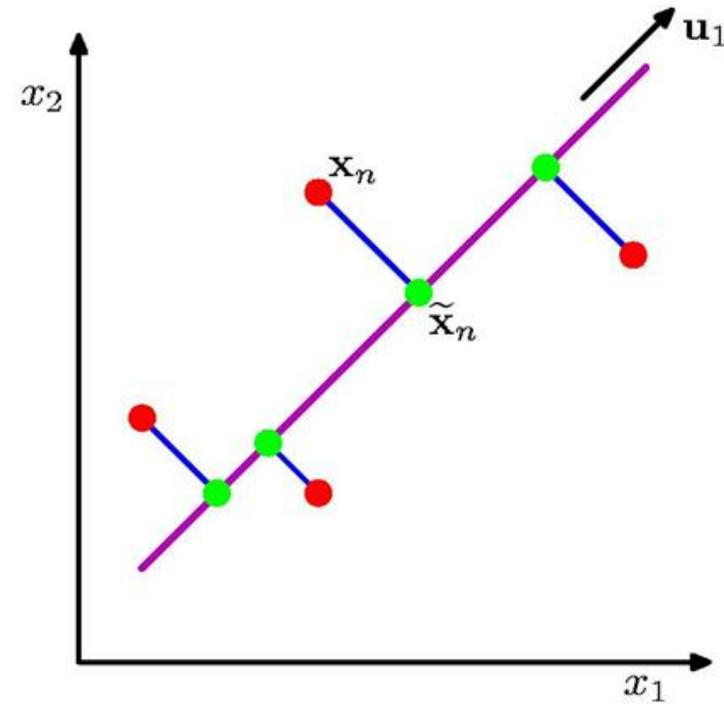
La PCA è una tecnica largamente usata per applicazioni quali: riduzione di dimensioni, compressione di dati, estrazioni di componenti, visualizzazione di dati

- Può essere definita come la **proiezione ortogonale** di dati su uno **spazio lineare** di dimensione inferiore (al numero di dati) in modo tale da **massimizzare la varianza** dei dati proiettati (Hotelling, 1933).
- Equivalentemente, può essere definita come la **proiezione lineare** che minimizza una **funzione costo** rappresentata dalla **media del quadrato della distanza tra i dati e le loro proiezioni** (Pearson, 1901).

In altre parole tramite PCA si può applicare una trasformazione lineare ad un data set composto da N dati, ottenere così M ( $M < N$ ) variabili, considerare solamente P delle M variabili basandosi sul fatto nelle P variabili è contenuta una buona parte dell'informazione disponibile. In questo modo, ad esempio, si riesce ad eliminare componenti non desiderate nei dati o, al contrario, evidenziare componenti interessanti presenti nei dati

**Nota: PCA non è un metodo appropriato se la relazione tra le variabili/dati non è lineare o approssimabile ad essere lineare**

A livello visivo si può fare l'esempio in figura dove i dati originali  $x_n$  vengono proiettati lungo la retta in modo da **massimizzare la varianza dei dati proiettati** (linea viola)



Supponiamo di avere a disposizione un data set composto da N osservazioni:

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N \quad \text{con} \quad \mathbf{x}_n \in \mathcal{R}^{D \times 1}$$

Ad esempio: D=# soggetti, N=# variabili (misure fisiologiche diverse: pressione, altezza, ecc..)

Per procedere con la PCA:

- 1. Si deve sottrarre ad ogni  $x_j$  la sua media (centramento dei dati)**
- 2. Calcolare la matrice delle covarianze S**
- 3. Calcolare autovalori e autovettori di S**
- 4. Scegliere le componenti principali e costruire la nuova rappresentazione dei dati**

Per capire la base teorica della PCA, supponiamo di voler proiettare le variabili correlate su uno spazio monodimensionale e di indicare con

$$\mathbf{u}_1 \in \mathcal{R}^{D \times 1}$$

il vettore che proietta ciascun  $\mathbf{x}_j$  in uno scalare:

$$\mathbf{y}_n = \mathbf{u}_1^T \mathbf{x}_n \quad \mathbf{y}_n \in \mathcal{R}^{1 \times 1}$$

la varianza dei dati così proiettati è pari a:

$$\frac{1}{N} \sum_{n=1}^N \left\{ \mathbf{u}_1^T \mathbf{x}_n - \mathbf{u}_1^T \bar{\mathbf{x}} \right\}^2 = \mathbf{u}_1^T \cdot \mathbf{S} \cdot \mathbf{u}_1$$

dove:

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \quad \bar{\mathbf{x}} \in \mathcal{R}^{D \times 1}$$

media del data set originale

$$\mathbf{S} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

matrice di covarianza

Vogliamo massimizzare la varianza proiettata  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  rispetto a  $\mathbf{u}_1$ . Chiaramente, bisogna sottoporre a qualche vincolo  $\mathbf{u}_1$  altrimenti la soluzione sarebbe  $\mathbf{u}_1 \rightarrow \infty$ . Il vincolo che si pone è pari alla condizione di normalizzazione:

$$\mathbf{u}_1^T \mathbf{u}_1 = 1$$

Per minimizzare una funzione  $f(x)$  sottoposta ad un vincolo  $g(x)=0$ , si ricorre all'uso dei moltiplicatori di Lagrange. In particolare, nei problemi di ottimizzazione, quello dei moltiplicatori di Lagrange è un metodo per trovare i massimi e i minimi di una funzione di più variabili soggetta a una o più vincoli. Nel nostro caso il problema diventa quello di massimizzare la funzione nonvincolata:

$$\mathbf{u}_1^T \cdot \mathbf{S} \cdot \mathbf{u}_1 + \lambda_1 (1 - \mathbf{u}_1^T \mathbf{u}_1)$$

Derivando rispetto a  $\mathbf{u}_1$  e ponendo il tutto = 0 si ottiene:

$$\mathbf{S} \cdot \mathbf{u}_1 = \lambda_1 \mathbf{u}_1$$

quindi, per definizione,  $\mathbf{u}_1$  deve essere un autovettore di  $\mathbf{S}$ .

Più precisamente, la varianza proiettata nel nuovo spazio  $\mathbf{u}_1^T \mathbf{S} \mathbf{u}_1$  sarà massima quando  $\mathbf{u}_1$  è pari all'autovettore con l'autovalore  $\lambda$  massimo.

## Definizione di Autovettore e Autovalore

Dato uno spazio vettoriale e una matrice quadrata  $A$  di dimensione  $n \times n$ , un vettore  $v \neq 0$  è detto **autovettore** di  $A$  se esiste uno scalare  $\lambda$  (chiamato **autovalore**) tale che:

$$Av = \lambda v$$

In altre parole, moltiplicando la matrice  $A$  per il vettore  $v$ , si ottiene un vettore parallelo a  $v$ , scalato di un fattore  $\lambda$ .

## Proprietà Principali degli Autovettori e Autovalori

### 1. Linearità della trasformazione:

Se  $v$  è un autovettore di  $A$  con autovalore  $\lambda$ , moltiplicando  $v$  per uno scalare  $c \neq 0$ , il vettore  $cv$  sarà ancora un autovettore associato allo stesso autovalore  $\lambda$ .

### 2. Esistenza e numero di autovalori:

Una matrice  $n \times n$  ha al massimo  $n$  autovalori distinti (reali o complessi). Gli autovalori sono le soluzioni del **polinomio caratteristico** di  $A$ :

$$\det(A - \lambda I) = 0$$

dove  $I$  è la matrice identità.

### 3. **Ortogonalità degli autovettori** (per matrici simmetriche):

Se  $A$  è una matrice simmetrica reale ( $A^T = A$ ), gli autovettori corrispondenti ad autovalori distinti sono **ortogonali** tra loro.

Le matrici simmetriche reali come  $S$  hanno proprietà speciali riguardanti i loro autovalori e autovettori. Queste proprietà derivano dalla loro struttura particolare ( $A = A^T$ )

## **Proprietà fondamentali**

### 1. **Autovalori reali:**

Tutti gli autovalori di una matrice simmetrica reale sono numeri reali. Questo è uno dei risultati più importanti e semplifica molto lo studio delle trasformazioni lineari simmetriche.

### 2. **Ortogonalità degli autovettori:**

Gli autovettori corrispondenti ad autovalori distinti di una matrice simmetrica reale sono **ortogonali** tra loro.

Se  $v_i$  e  $v_j$  sono autovettori associati agli autovalori  $\lambda_i$  e  $\lambda_j$  ( $\lambda_i \neq \lambda_j$ ), allora:

$$v_i \cdot v_j = 0 \quad (\text{prodotto scalare})$$

QUINDI:

- essendo  $S$  una matrice simmetrica, gli autovalori associati sono reali
- Il rango della matrice  $S$  coincide con il numero di autovalori non nulli

E considero che:

1. Ad ogni autovettore corrisponde un autovalore
2. Gli autovettori forniscono le direzioni (componenti principali) lungo cui i dati hanno la massima varianza

**IDEA:** proietto i dati  $\mathbf{X}$  nello spazio delle componenti principali (autovettori)

Definiamo ora in modo più compatto:

$$X = U * L^T$$

$$\begin{array}{c}
 D \times N \\
 \begin{bmatrix} x_{11} & x_{12} & \vdots & x_{1N} \\ x_{21} & x_{22} & \vdots & x_{2N} \\ x_{31} & x_{32} & \vdots & x_{3N} \\ \dots & \dots & \dots & \dots \\ x_{D1} & x_{D2} & \vdots & x_{DN} \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 D \times N \\
 \begin{bmatrix} u_{11} & u_{12} & \vdots & u_{1N} \\ u_{21} & u_{22} & \vdots & u_{2N} \\ u_{31} & u_{32} & \vdots & u_{3N} \\ \dots & \dots & \dots & \dots \\ u_{D1} & u_{D2} & \vdots & u_{DN} \end{bmatrix}
 \end{array}
 *
 \begin{array}{c}
 N \times N \\
 \left( \begin{bmatrix} l_{11} & \dots & l_{1N} \\ \vdots & \ddots & \vdots \\ l_{N1} & \dots & l_{NN} \end{bmatrix} \right)^T
 \end{array}$$

$u_1$   
autovettore 1

$u_2$   
autovettore 2

$u_N$   
autovettore N

Definiamo ora in modo più compatto:

$$X = U * L^T$$

$$\begin{array}{c}
 D \times N \qquad \qquad \qquad D \times N \qquad \qquad \qquad N \times N \\
 \begin{bmatrix} x_{11} & x_{12} & \vdots & x_{1N} \\ x_{21} & x_{22} & \vdots & x_{2N} \\ x_{31} & x_{32} & \vdots & x_{3N} \\ \dots & \dots & \dots & \dots \\ x_{D1} & x_{D2} & \vdots & x_{DN} \end{bmatrix} = \begin{bmatrix} u_{11} & u_{12} & \vdots & u_{1N} \\ u_{21} & u_{22} & \vdots & u_{2N} \\ u_{31} & u_{32} & \vdots & u_{3N} \\ \dots & \dots & \dots & \dots \\ u_{D1} & u_{D2} & \vdots & u_{DN} \end{bmatrix} * \left( \begin{bmatrix} l_{11} & \dots & l_{1N} \\ \vdots & \ddots & \vdots \\ l_{N1} & \dots & l_{NN} \end{bmatrix} \right)^T
 \end{array}$$

$u_1$   
 $\downarrow$   
 $\lambda_1$

$u_2$   
 $\downarrow$   
 $\lambda_2$

$u_N$   
 $\downarrow$   
 $\lambda_N$

$\lambda = \text{autovalori}$

Quindi:

$$X = U * L^T + \bar{X} = \text{score} * \text{coef}^T + \bar{X} \quad \mathbf{X} \in \mathcal{R}^{D \times N}$$

Matrice di partenza dei dati

Autovettori /componenti

loadings

Matrice dei valori medi

In generale:

Le nuove coordinate (U) dei vettori corrispondenti alle osservazioni nella base delle componenti principali prendono il nome di **scores = autovettori**

I coefficienti delle combinazioni lineari che definiscono le componenti principali sono detti loadings

Il loading quindi fornisce una misura del contributo di ogni osservabile alle componenti principali



*Le nuove variabili (proiezioni dei dati originari) godono delle seguenti proprietà:*

$U_N$  sono scorrelate (ortogonali), quindi la matrice di covarianza di  $U$  è una matrice diagonale

$U_1$  (= PC1) è l'autovettore relativo all'autovalore massimo, è la componente che è capace di rappresentare da sola il massimo della varianza dei dati

$U_2$  (=PC2) è l'autovettore relativo al secondo autovalore, spiega parte della varianza residua

ecc...

Riassumendo:

$$X = U * L^T$$

*poiché:*

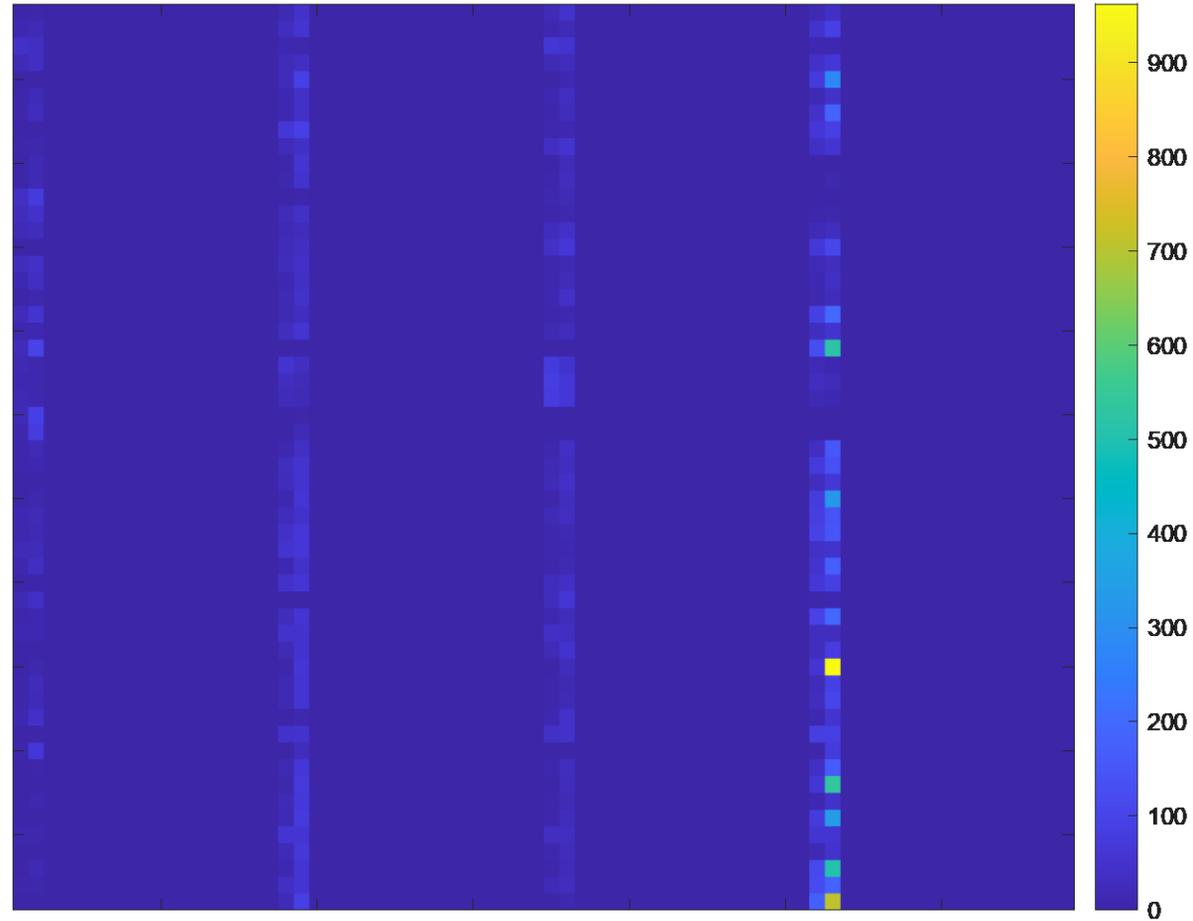
$$L * L^T = L^T * L = I$$

*Si ottiene anche che:*

$$X * L = U$$

68 variabili

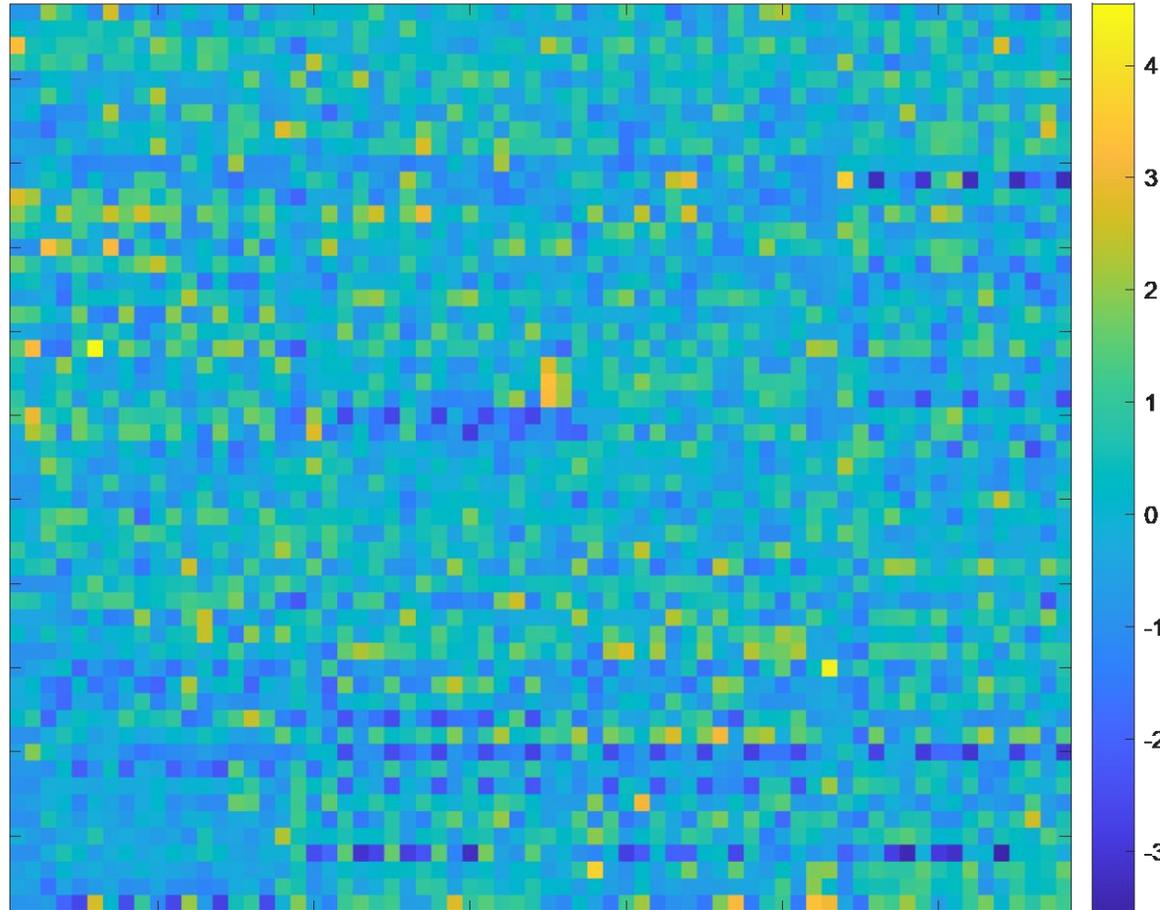
54 soggetti



54 soggetti x 68 variabili (ogni variabile ha una unità di misura diversa) = Normalizzo tramite zscore

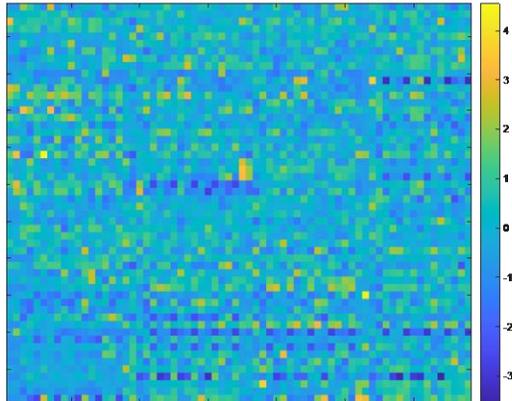
68 variabili ognuna con media = 0 e  
varianza pari a 1

54 soggetti



54 soggetti

68 variabili ognuna con  
media = 0 e varianza  
pari a 1

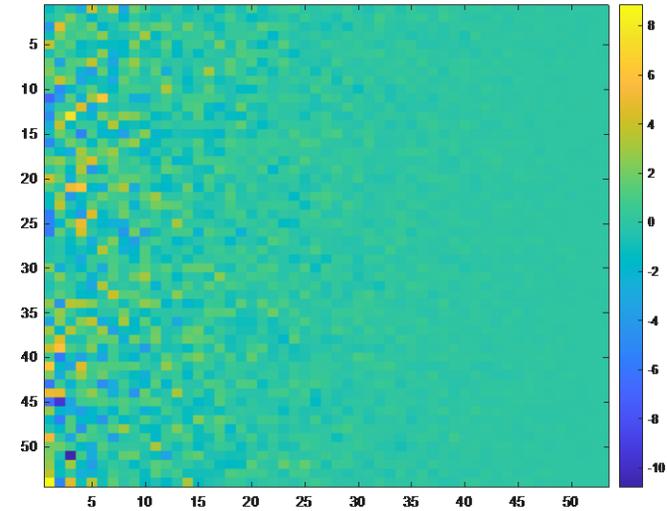


PCA



54 soggetti

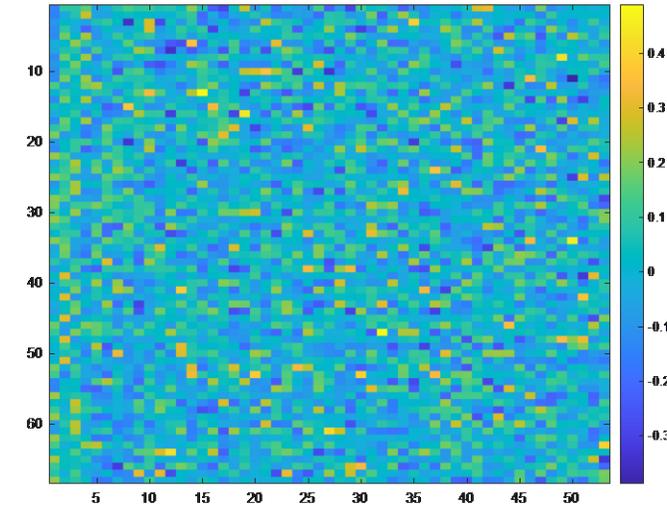
Score = autovettori



53 componenti = 53 PC

Coeff = Loadings

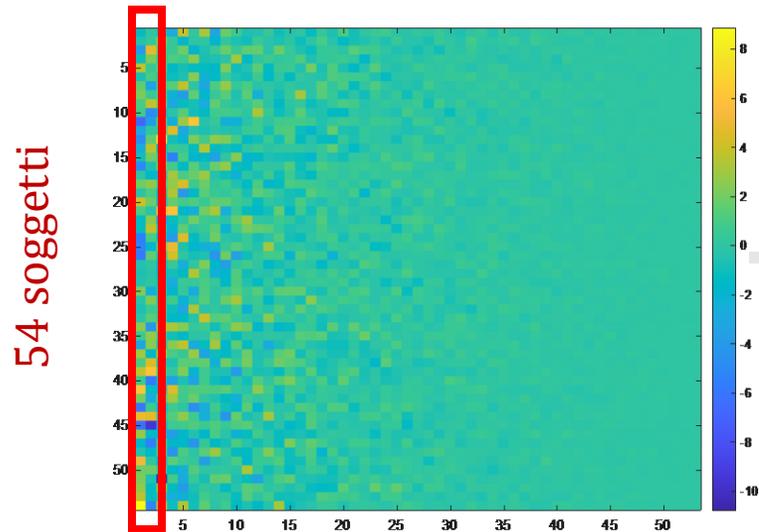
68 variabili



53 componenti = 53 PC

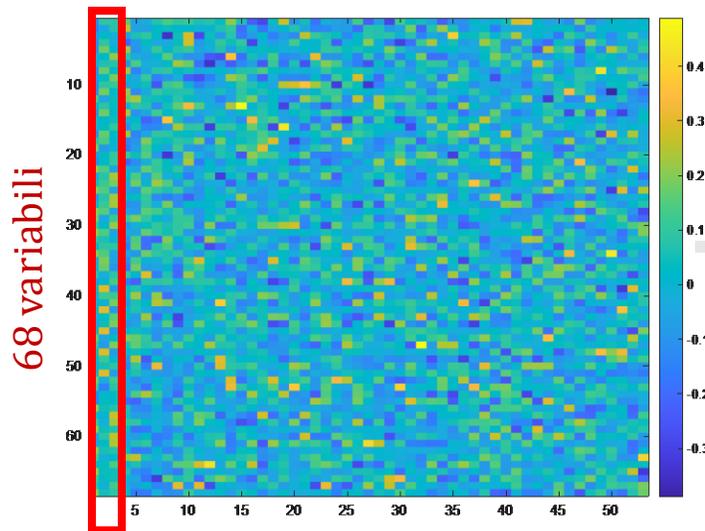
Ricordiamoci: Valori di partenza (a cui è stata tolta la media) = SCORE \* COEFF<sup>T</sup>

Score = autovettori



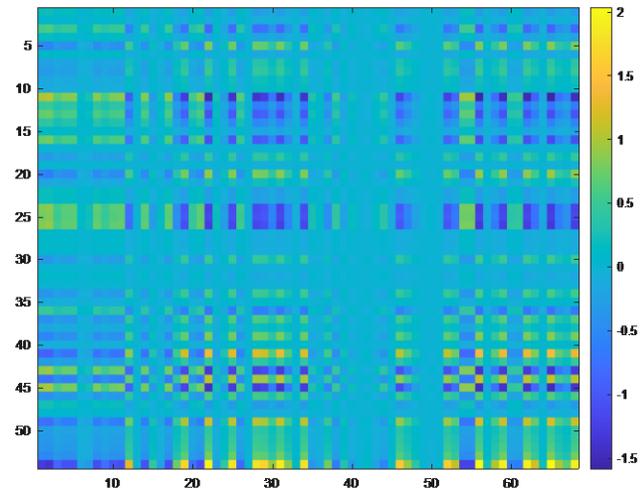
53 componenti = 53 PC

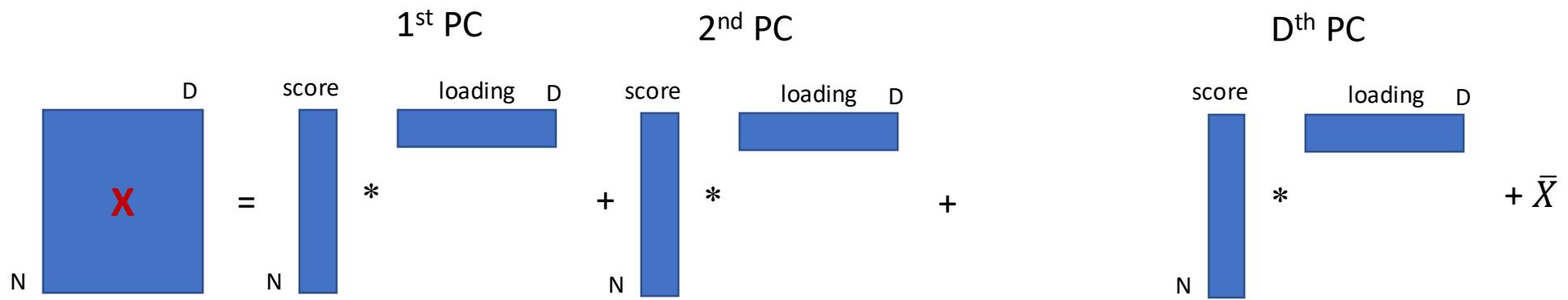
Coeff = Loadings



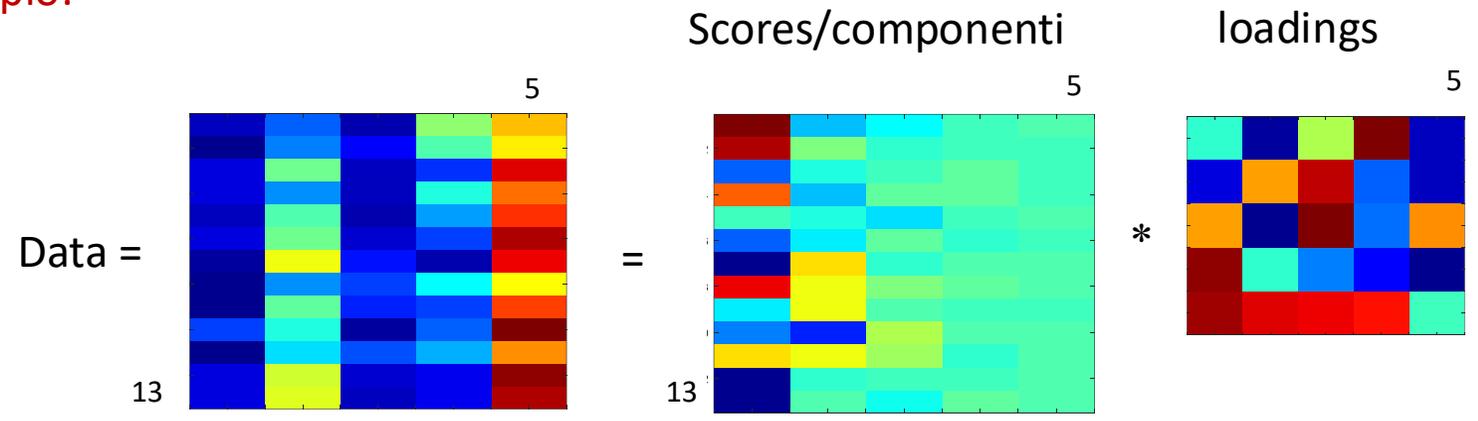
53 componenti = 53 PC

$$(53 \times 1) * (1 \times 68) \rightarrow$$

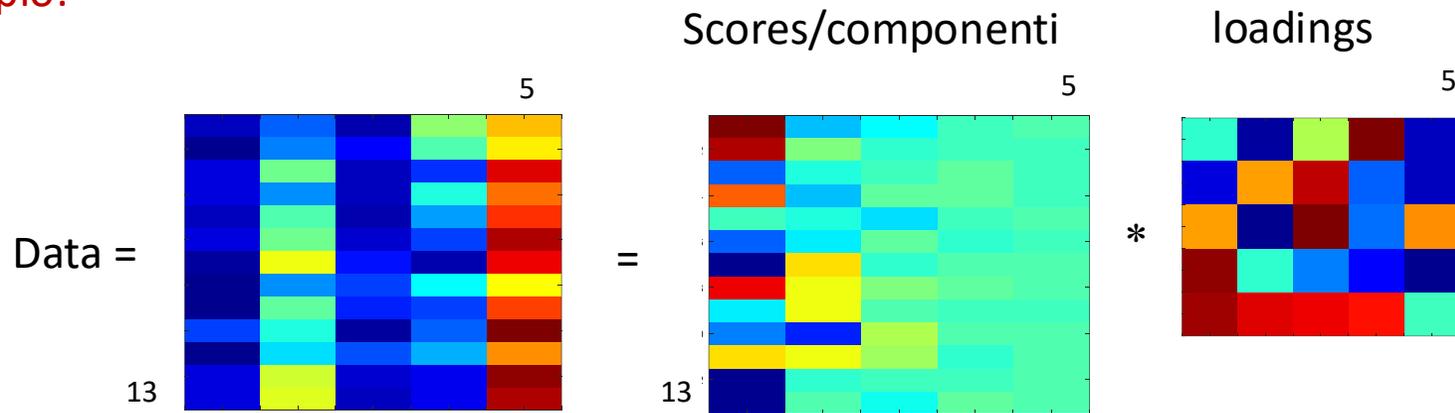




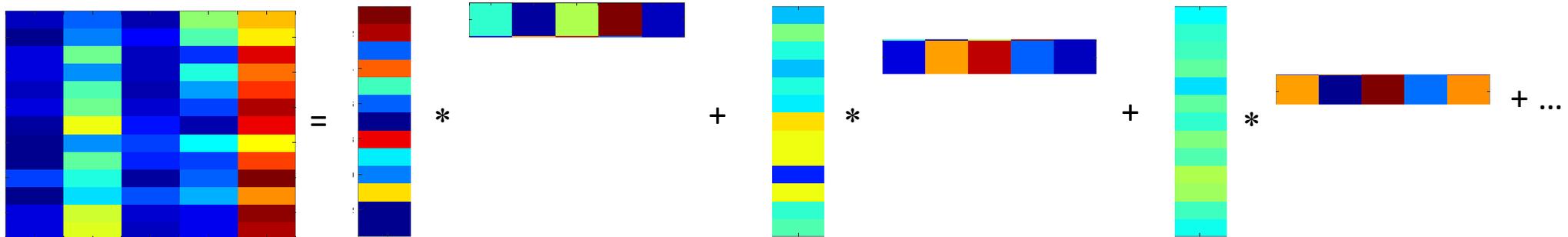
Esempio:

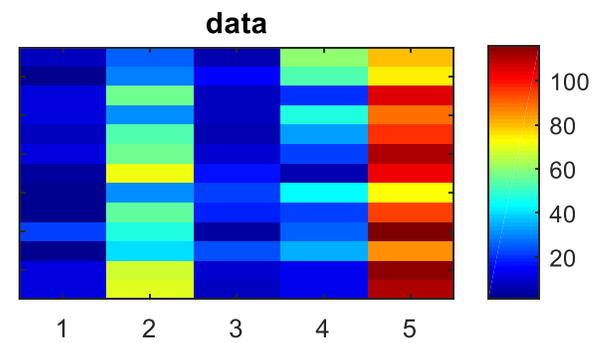
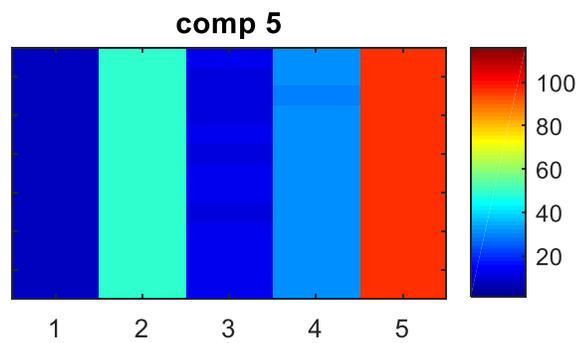
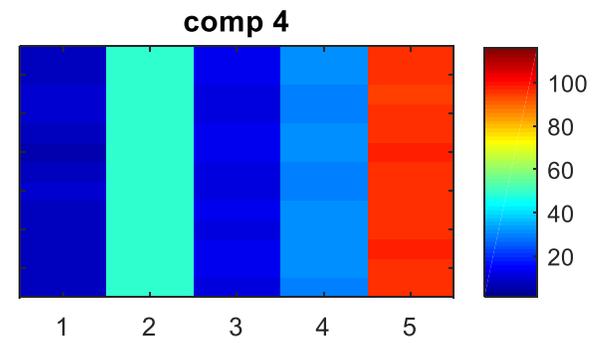
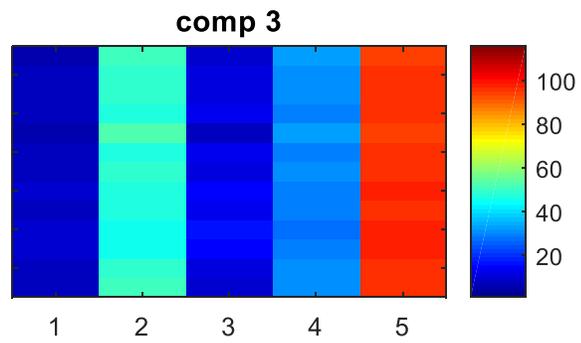
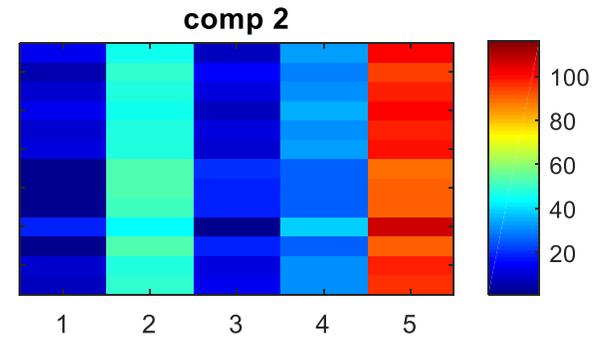
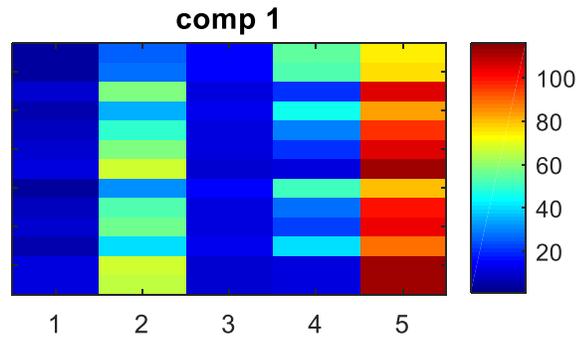


Esempio:



Ossia per le prime tre componenti:





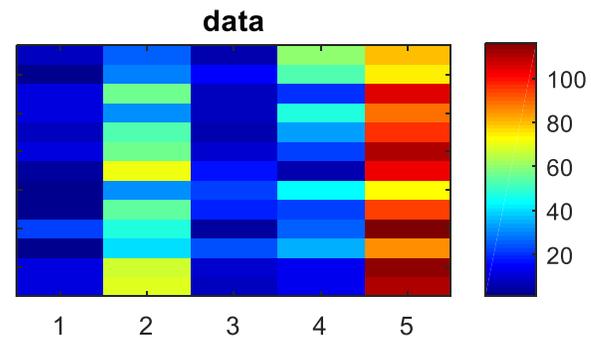
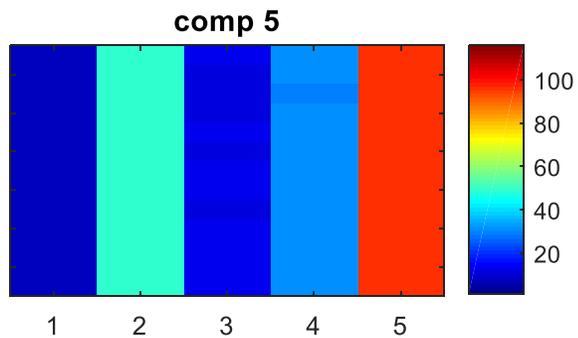
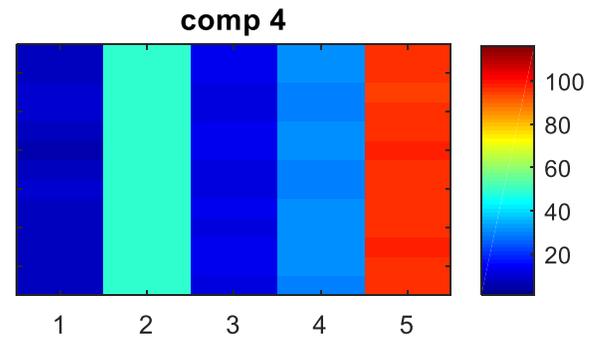
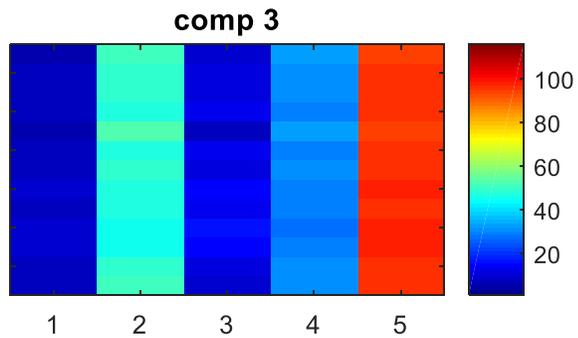
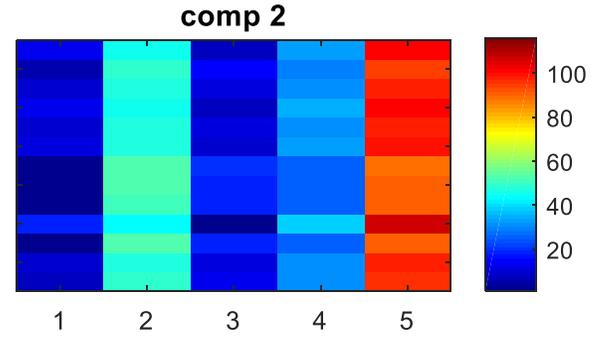
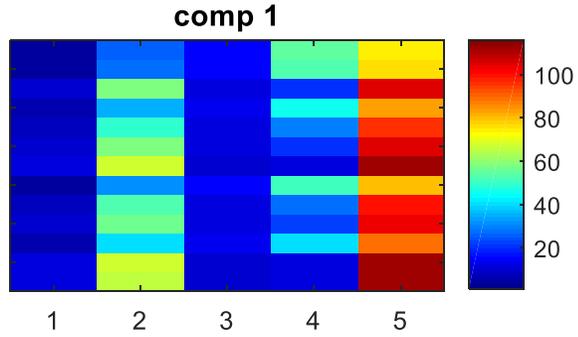
## IMPORTANTE:

I dati proiettati ( $U$ ) hanno varianza uguale ai loro autovalori

$$\text{VAR}(u_i) = \lambda_i$$

La varianza % spiegata da ciascuna componente è data da:

$$\lambda_i / \sum_{i=1}^n \lambda_i$$



$\text{var}\%_{(\text{comp } 1)} = 84.36\%$   
 $\text{var}\%_{(\text{comp } 2)} = 13.56\%$   
 $\text{var}\%_{(\text{comp } 3)} = 1.86\%$   
 $\text{var}\%_{(\text{comp } 4)} = 0.18\%$   
 $\text{var}\%_{(\text{comp } 5)} = 0.04\%$

Dal punto di vista numerico:

*scores*

39.8331	-9.8213	-5.2115	-0.5073	0.3642
36.4980	3.6952	-1.6029	-0.8316	-0.4699
-15.9135	-3.0512	-0.2464	1.9006	-0.9318
23.8891	-9.8666	0.9838	1.4567	-0.2345
-0.9111	-3.1142	-6.8314	-0.5559	0.1112
-16.6198	-6.0836	1.7519	-2.1020	-0.3823
-31.0147	15.2749	-1.5894	0.6155	0.1687
31.7817	11.6528	3.6351	1.6734	0.5103
-6.1745	11.5187	0.8831	-0.5327	-0.5966
-14.3400	-20.7331	6.6827	0.1999	0.3059
14.5601	11.9010	5.4483	-1.4408	0.2723
-31.2632	-1.9752	-0.3458	-0.8512	0.3506
-30.3253	0.6025	-3.5575	0.9757	0.5318

*Loadings/coefficienti*

-0.0951	-0.5735	0.0611	0.6179	-0.5259
-0.4881	0.2673	0.5376	-0.3368	-0.5366
0.2635	-0.5939	0.6228	-0.3239	0.2917
0.5967	-0.0807	-0.2904	-0.4550	-0.5883
0.5721	0.4904	0.4848	0.4391	-0.0659

→ Sono ortonormali

Ossia la componente principale  $score_i$  sarà maggiormente caratterizzata dalle variabili  $X_j$  a cui corrispondono i loadings/coefficienti  $ij$  più grandi in valore assoluto.

*autovalori*

695.3761  
111.7525  
15.3443  
1.5555  
0.2211

*autovalori %*

84.3648  
13.5581  
1.8616  
0.1887  
0.0268

*In generale:*

La prima componente principale spiega la massima percentuale della variabilità presente nei dati rappresentabile in una sola dimensione

Messa in un'altra maniera: è la direzione lungo cui si registra la massima dispersione dei dati.

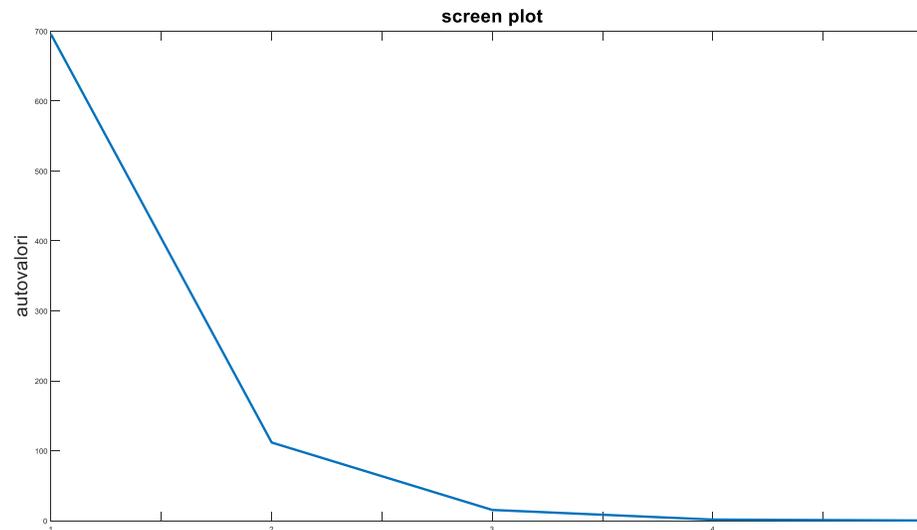
Nell'esempio descritto, la prima componente principale cattura praticamente tutta la variabilità presente nei dati (84.36%)

La seconda descrive la rimanente variazione (13.56%). Questa considerazione può essere generalizzata: le componenti principali successive spiegano una sempre minore percentuale della variabilità originale.

Seguendo questo principio è possibile dire che le ultime componenti principali descrivono principalmente “rumore” ovvero il contributo degli errori di misura o informazioni irrilevanti

### *Scelta del numero di componenti:*

- 1) Prendere solo quelle componenti che rappresentano l' 80-90% della variabilità complessiva;
- 2) Seguire la “Regola di Kaiser”: prendere solo quelle componenti che hanno un autovalore maggiore o uguale ad uno, oppure, equivalentemente, le componenti che hanno varianza maggiore di quella media
- 3) La scelta del numero di componenti può essere fatta attraverso il grafico degli autovalori o “Screen Plot”. All’interno del grafico si sceglie il numero di componenti corrispondente al punto di “gomito” della spezzata



# PCA

## Standardizzazione delle variabili d'origine

Nel caso in cui la matrice su cui applicare PCA sia composta da dati con unità di misura non paragonabili e/o dati che si riferiscono a varianze molto diverse tra gli elementi della matrice, non si applica PCA sulla matrice originale ma è spesso necessario prima standardizzare le variabili in essa contenute.

Uno dei modi più diffusi per standardizzare le variabili ( $X_i$ ) consiste nel rimuovere la media e dividere per la deviazione standard:

$$Z_1 = \frac{X_1 - \mu_1}{\sigma_1}; Z_2 = \frac{X_2 - \mu_2}{\sigma_2}; \dots Z_N = \frac{X_N - \mu_N}{\sigma_N}; \quad \text{Z-scores}$$

ed ottenere così un nuovo campione ( $Z_i$ ) che avrà media nulla e varianza unitaria.

Per questo nuovo campione, si avrà che la matrice di covarianza delle  $Z_i$  coincide con la matrice di correlazione del campione d'origine  $X_i$ :

$$COV (Z) = CORR (X)$$

Pertanto, quando si ha un campione  $X_i$  “non omogeneo” e si passa in  $Z_i$  si sta operando sulla matrice di correlazione,  $R$ .

Questa operazione può essere vista come una pesatura dei dati:

$$Z_1 = \omega_1(X_1 - \mu_1); Z_2 = \omega_2(X_2 - \mu_2); \dots Z_N = \omega_N(X_N - \mu_N);$$

Attenzione: gli autovalori ottenuti con la matrice di correlazione R sono diversi da quelli della matrice di covarianza  $\Sigma$  relativa al campione d'origine (X).

---

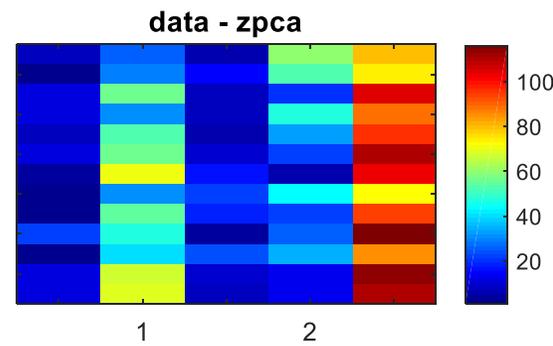
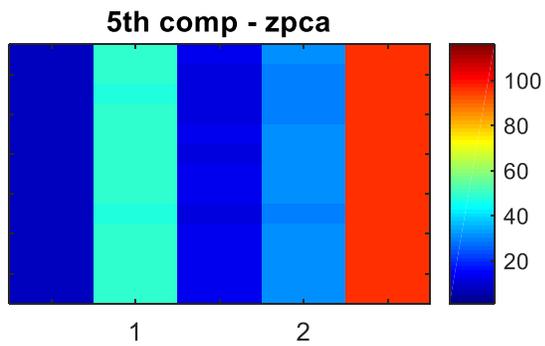
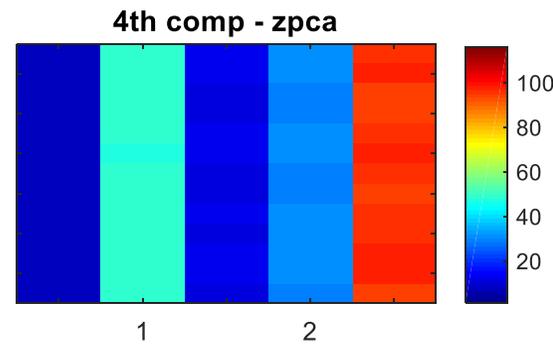
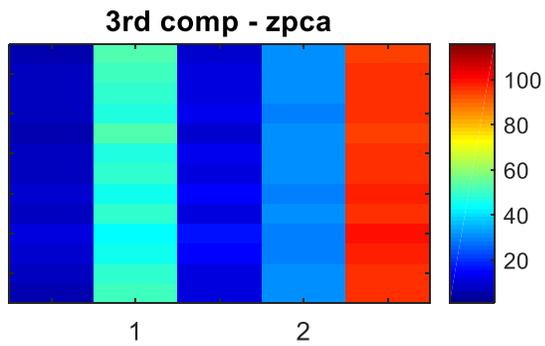
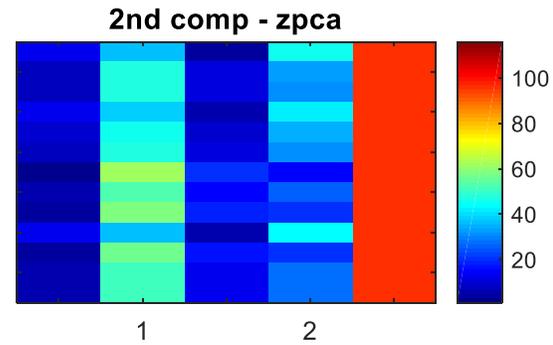
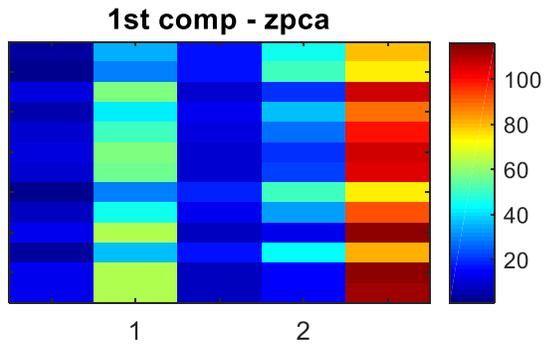
Quindi possiamo distinguere due modi di applicare PCA: sulla matrice di covarianza e su quella di correlazione:

#### *Matrice di covarianza*

le variabili devono avere le stesse unità di misura  
enfattizza le variabili che hanno varianza maggiore  
media degli autovalori  $\neq 1.0$

#### *Matrice di correlazione*

le variabili sono standardizzate (media 0, SD 1)  
le variabili di partenza possono avere unità di misura diverse  
tutte le variabili hanno lo stesso peso nell'analisi  
media degli autovalori = 1.0



$$\text{var}\%_{(\text{comp } 1)} = 64.23\%$$

$$\text{var}\%_{(\text{comp } 2)} = 31.52\%$$

$$\text{var}\%_{(\text{comp } 3)} = 3.98\%$$

$$\text{var}\%_{(\text{comp } 4)} = 0.23\%$$

$$\text{var}\%_{(\text{comp } 5)} = 0.04\%$$

# Singular Value Decomposition

## SVD

SVD è una scomposizione matriciale che permette di lavorare con matrici anche singolari (matrice singolare = matrice quadrata a determinante nullo. Una matrice singolare non è invertibile) o numericamente molto prossime ad essere singolari.

SVD è applicabile ad ogni matrice  $m \times n$  ( $m \geq n$ )

Remark:

$$\mathbf{A} = \mathbf{V} \cdot \mathbf{W} \cdot \mathbf{U}^T$$

$$\left. \begin{array}{l} \mathbf{V}^T \mathbf{V} = \mathbf{I}^{n \times n} \\ \mathbf{U}^T \mathbf{U} = \mathbf{I}^{p \times p} \end{array} \right\} \rightarrow \text{U e V sono ortogonali}^*$$

$\mathbf{W}$   $\rightarrow$  matrice diagonale

*\*una matrice ortogonale è una matrice quadrata invertibile la cui trasposta coincide con la sua inversa*

$$\left. \begin{array}{l} \mathbf{V}^T \mathbf{V} = \mathbf{I}^{n \times n} \\ \mathbf{U}^T \mathbf{U} = \mathbf{I}^{p \times p} \end{array} \right\} \rightarrow \text{U e V sono ortogonali}$$

$$\mathbf{W} \rightarrow \text{matrice diagonale}$$

Calcolare la SVD significa trovare gli autovalori e autovettori di  $\mathbf{A}^T \mathbf{A}$  e  $\mathbf{A} \mathbf{A}^T$ :

autovettori di  $\mathbf{A}^T \mathbf{A}$  = colonne di  $\mathbf{U}$

autovettori di  $\mathbf{A} \mathbf{A}^T$  = colonne di  $\mathbf{V}$

radice quadrata degli autovalori di  $\mathbf{A}^T \mathbf{A}$  = matrice diagonale  $\mathbf{W}$

Nota: in Matlab si può usare il comando `[V W U]=svd(A)`. Se si vogliono ottenere autovalori e autovettori di una matrice A il comando è `[autovet autoval]=eig(A)`

## ROTAZIONE DEGLI ASSI FATTORIALI (Loadings)

Per meglio interpretare i fattori stimati, possiamo applicare una rotazione dei loading fattoriali.

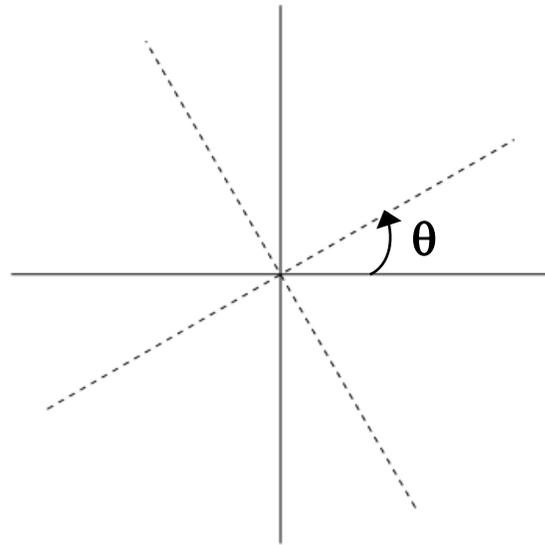
Esistono due classi di metodi di rotazione fattoriale: rotazioni ortogonali e oblique. Le prime presuppongono che i fattori siano tra loro non correlati. Le seconde che i fattori siano tra lo correlati.

Una rotazione ortogonale molto popolare è la VARIMAX, finalizzata a produrre fattori con elevata correlazione con un insieme piccolo di variabili e correlazioni piccole o nulle con gli altri insiemi.

In particolare, scegliendo una rotazione opportuna si può ottenere una matrice dei loadings più semplice da interpretare. I loadings rappresentano le covarianze (correlazioni) tra le PC e le variabili originali (standardizzate), quindi ha senso cercare le rotazioni che portano a dei fattori che hanno la massima correlazione con vari gruppi di variabili originali.

## ■ Rotazione VARIMAX

- La rotazione VARIMAX ha come obiettivo quello di trovare una soluzione in cui una variabile pesa molto su una componente particolare e pesa il meno possibile sulle altre.



$$\mathbf{T} = \begin{bmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{bmatrix}$$



**Rotation matrix**

$$\mathbf{T} * \mathbf{T}^T = \mathbf{I}$$

- L'algoritmo massimizza l'indice VARIMAX,  $V$ , la somma delle varianze della matrice dei loadings
- Dopo la rotazione varimax ciascuna delle variabili originarie tenderà ad essere associata con una o poche componenti ed ogni componente sarà espressione di un numero ridotto di variabili
- In generale non si vuole ruotare tutti i loadings ma solo quelli di interesse

$$\begin{array}{c}
 D \times N \quad D \times N \quad N \times N \\
 X = U * L^T
 \end{array}$$

Si seleziona un numero di  $L$  pari al numero di PC selezionate (con uno dei vari metodi) =  $L_K$

$$\begin{array}{c}
 L_{new} = \text{coefficienti ruotati} = LK * T \\
 N \times K \qquad \qquad \qquad N \times K \quad K \times K
 \end{array}$$

$$\begin{array}{c}
 U_{new} = X * L_{new} \\
 D \times K \quad D \times N \quad N \times K
 \end{array}$$