UNIVERSITÀ
DEGLI STUDI
DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA (B)

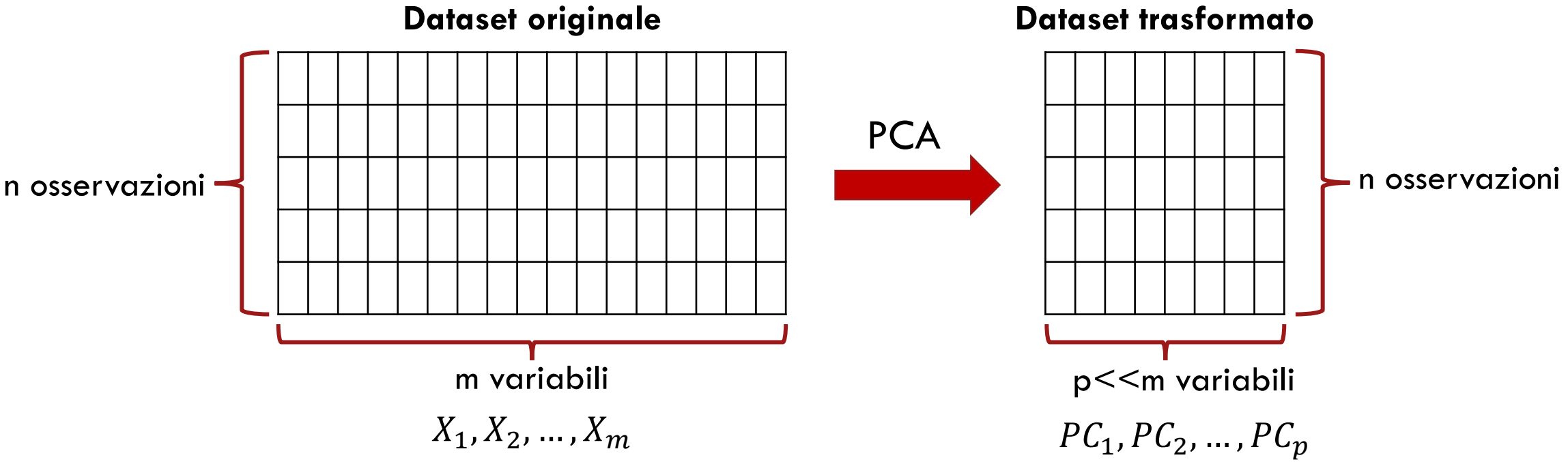
PARTE 18: PRINCIPAL COMPONENT ANALYSIS (PCA)

A.A. 2024-2025

Prof. Martina Vettoretti

PRINCIPAL COMPONENT ANALYSIS (PCA)

- **Principal Component Analysis (PCA)**, o analisi delle componenti principali: tecnica di apprendimento non supervisionato per **ridurre la dimensionalità dei dati.**
- Obiettivo: rappresentare un set di dati di dimensione n x m (n osservazioni, m variabili) in uno spazio a dimensionalità ridotta con p<<m variabili.



LE COMPONENTI PRINCIPALI

- Le p nuove variabili sono dette **componenti principali** (o *principal components*).
- Caratteristiche delle componenti principali:
  - Le componenti principali sono **nuove variabili**, create artificialmente, nessuna di loro coincide con alcuna delle m variabili di partenza.
  - Ogni componente principale è una **combinazione lineare** delle m variabili originali.
  - Le componenti principali sono tali da **riassumere quanta più informazione possibile** sulle m variabili originali.
  - Le componenti principali sono ordinate in base a quanta informazione del dataset originale racchiudono (PC_1 è la componente più informativa, PC_2 la seconda più informativa ecc.)
  - Le componenti principali per costruzione sono **tra loro scorrelate**.

PERCHE' PUO' ESSERE UTILE LA PCA?

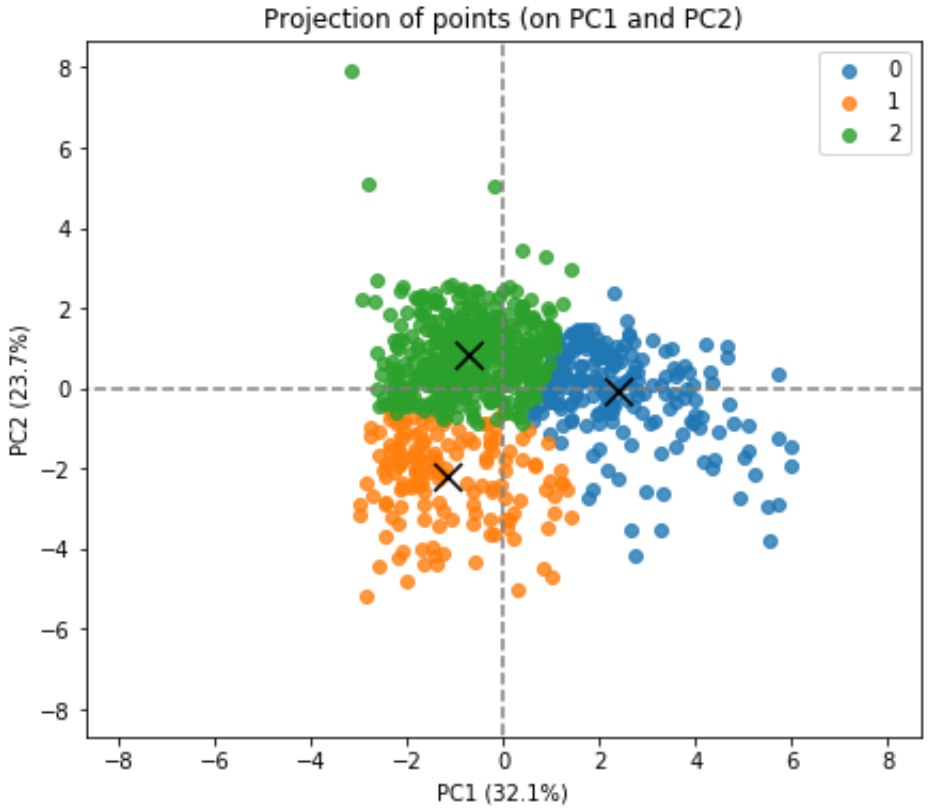
- Visualizzazione dei dati
- Compressione dei dati
- Eliminare la correlazione tra le variabili di ingresso in un modello di regressione lineare multipla (o di altro tipo)

VISUALIZZAZIONE DEI DATI

- Quando abbiamo dataset con tante variabili (m grande) diventa difficile rappresentare graficamente i dati per analizzarli dal punto di vista visivo.
  - Tipicamente si visualizzano le distribuzioni delle singole variabili o al più gli scatterplot di coppie di variabili.
  - Problemi:
    - Lo scatterplot di due variabili rappresenta solo una piccola quantità dell'informazione contenuta nei dati.
    - Con m variabili dovremmo fare $m \cdot (m-1)/2$ scatterplot. Se m = 10 → 10 x $9/2 = 45$ scatterplot!
- Possiamo sfruttare la PCA per riassumere l'informazione contenuta nei dati usando poche variabili, più semplici da rappresentare.
  - Potremmo realizzare lo scatterplot delle prime 2 componenti principali, PC_1 e PC_2, ovvero quelle più informative!

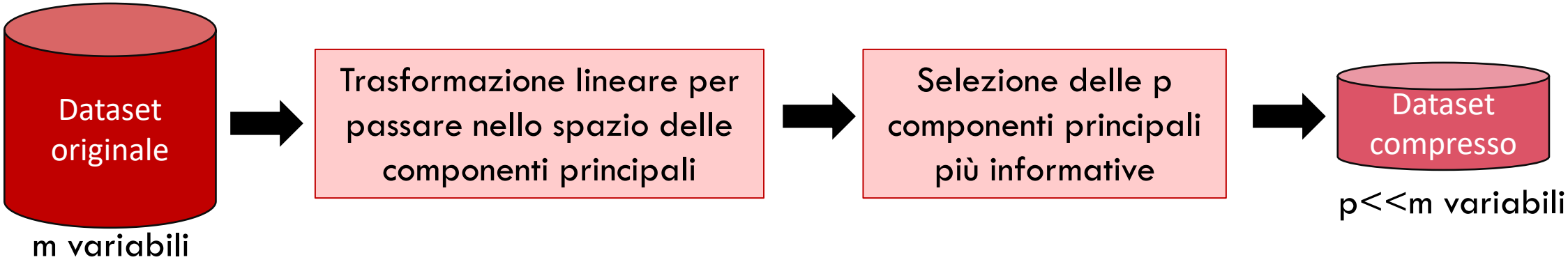
ESEMPIO

- Abbiamo un dataset con 50 variabili e realizziamo un clustering K-means per suddividere le osservazioni in 3 gruppi. Vogliamo verificare dal punto di vista visivo quanto sono separati i 3 cluster.
- Impossibile rappresentare i dati nello spazio a 50 dimensioni del dataset originale!
- Soluzione: applichiamo la PCA e rappresentiamo i cluster nello spazio bidimensionale delle prime 2 componenti principali, le più informative.

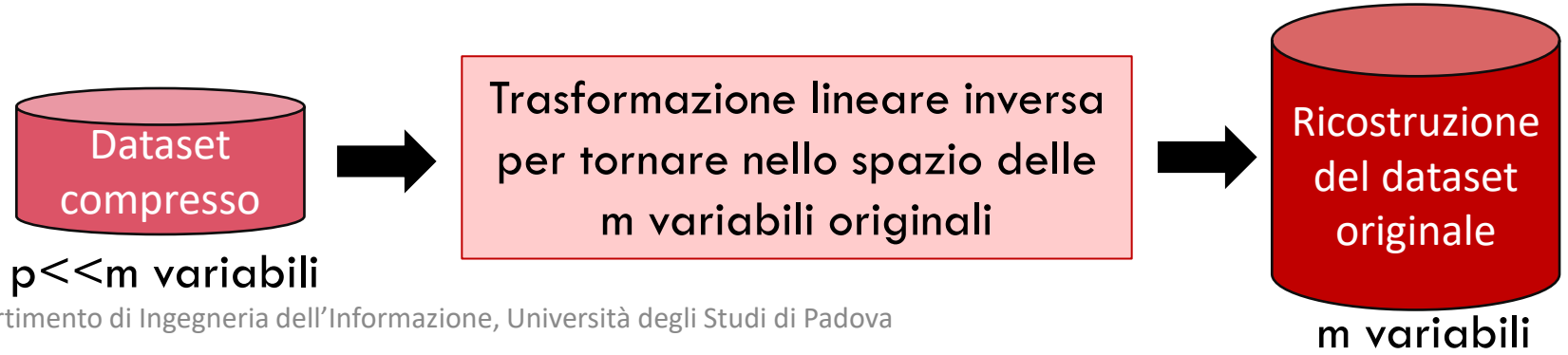


COMPRESSIONE DEI DATI

- **Compressione**: al posto di archiviare le m variabili di partenza, archivio le p componenti principali, con un risparmio in memoria.



- **Decompressione**: utilizzando le p componenti principali ricostruisco le m variabili di partenza. La ricostruzione non sarà perfetta, l'errore introdotto dalla compressione dipende da quanto informative erano le p componenti principali selezionate.



ESEMPIO

- Compressione di un'immagine di dimensione 256 x 256 (matrice di dati 256 x 256) mediante PCA.

INPUT IMAGE

RESTORED IMAGE-10 COMPONENTS

RESTORED IMAGE-50 COMPONENTS

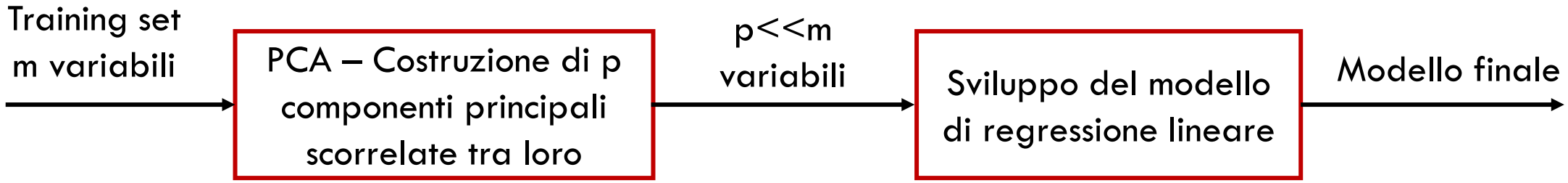
RESTORED IMAGE-100 COMPONENTS

RESTORED IMAGE-150 COMPONENTS

RESTORED IMAGE-256 COMPONENTS

ELIMINARE LA CORRELAZIONE TRA LE VARIABILI IN INGRESSO IN UN MODELLO DI REGRESSIONE LINEARE MULTIPLA

- La multicollinearità tra le variabili di ingresso è problematica per i modelli di regressione lineare multipla → possiamo eliminare la multicollinearità mediante PCA.
- Idea: allenare il modello sulle p componenti principali anziché sulle m variabili originali.

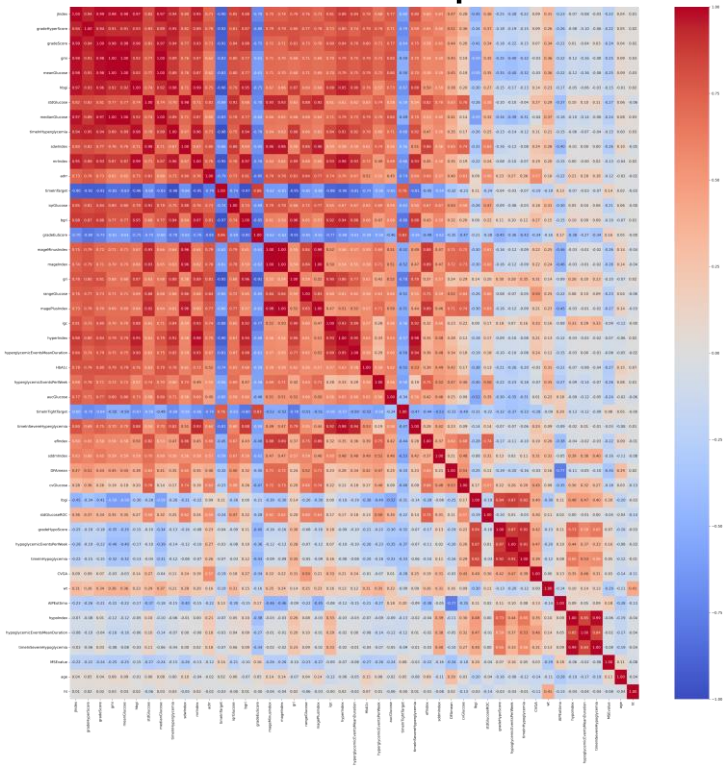


- Attenzione: le p componenti principali sono nuove variabili che non hanno un significato fisico riconducibile alle m variabili di partenza. → Stimando i coefficienti del modello di regressione lineare sulle componenti principali, perdiamo la possibilità di interpretare i coefficienti del modello.

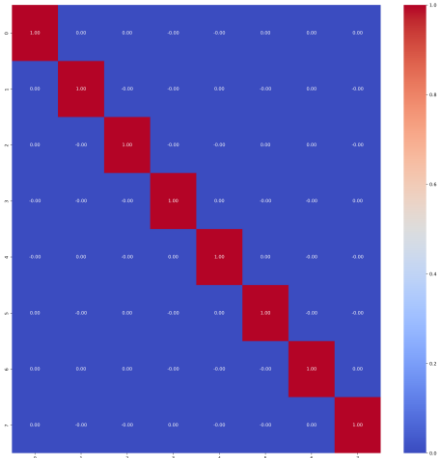
ESEMPIO

- Modello di regressione logistica per la classificazione dei pazienti diabetici vs non diabetici sulla base di indici di variabilità glicemica.

Matrice di correlazione delle 47 variabili di partenza



Matrice di correlazione delle 8 prime componenti principali



Modello di regressione logistica allenato sulle 8 componenti principali anziché sulle 47 variabili originali

PCA, COME SI REALIZZA?

- Dataset originale: m variabili, $X_1, X_2, \ldots, X_m$, a media nulla.
- Nota: se le variabili non sono a media nulla, prima di realizzare la PCA, i dati vanno centrati → ad ogni variabile va sottratta la sua media.
- **Trasformazione lineare normalizzata** delle m variabili → m nuove variabili, $PC_1, PC_2, \ldots, PCm$, dette componenti principali tra loro scorrelate

$$PC_1 = v_{11}X_1 + v_{21}X_2 + \ldots + v_{m1}X_m$$
$$PC_2 = v_{12}X_1 + v_{22}X_2 + \ldots + v_{m2}X_m$$
$$\ldots$$
$$PC_m = v_{1m}X_1 + v_{2m}X_2 + \ldots + v_{mm}X_m$$

- I coefficienti v_{ik} consentono di trasformare le variabili di partenza nelle nuove variabili, le componenti principali.

I LOADINGS

- I coefficienti $v_{ik}, i = 1, \dots, m$ relativi alla componente principale k-esima, PC_k, si dicono ***loadings*** relativi alla componente k-esima:

$$\boldsymbol{v_k} = (v_{1k}, v_{2k}, \dots, v_{mk})^T$$

essi rappresentano il **contributo di ciascuna delle variabili originali** alla componente principale k-esima.

- I loadings di ciascuna componente principale hanno **norma 1**:

$$\sum_{i=1}^{m} v_{ik}^2 = 1 \quad \forall\, k$$

Per questo la trasformazione lineare effettuata dalla PCA si dice normalizzata.

- I vettori di loadings delle diverse componenti principali sono tra loro **ortogonali**:

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

VARIANZA DELLE COMPONENTI PRINCIPALI

- I loadings devono essere tali che **le prime componenti principali riassumano quanta più varianza possibile** delle variabili di partenza:

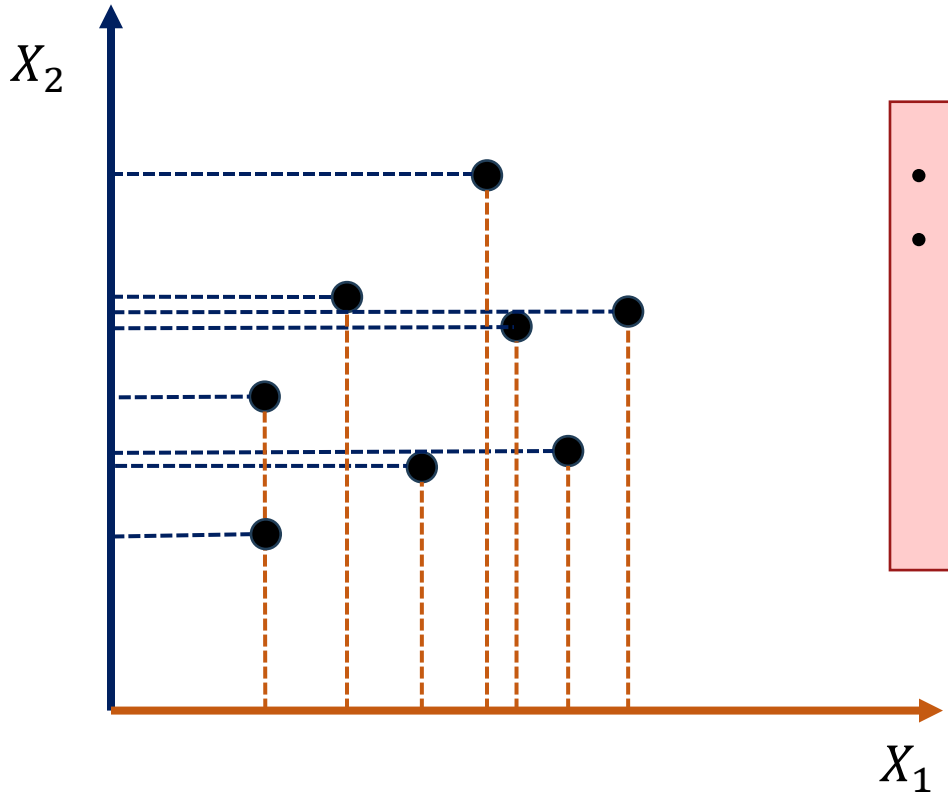
$$var(PC_1) > var(PC_2) > var(PC_3) > \cdots > var(PC_m)$$

  - PC_1 da sola deve essere in grado di spiegare quanta più varianza possibile dei dati di partenza.
  - PC_2 da sola deve essere in grado di spiegare quanta più varianza possibile della porzione di varianza non spiegata da PC_1.
  - …

- **Considerando solo le prime p componenti** principali possiamo efficacemente ridurre la dimensionalità dei dati → nuovo set di p<<m variabili che riassume la maggior parte della varianza delle m variabili di partenza.

INTERPRETAZIONE GEOMETRICA

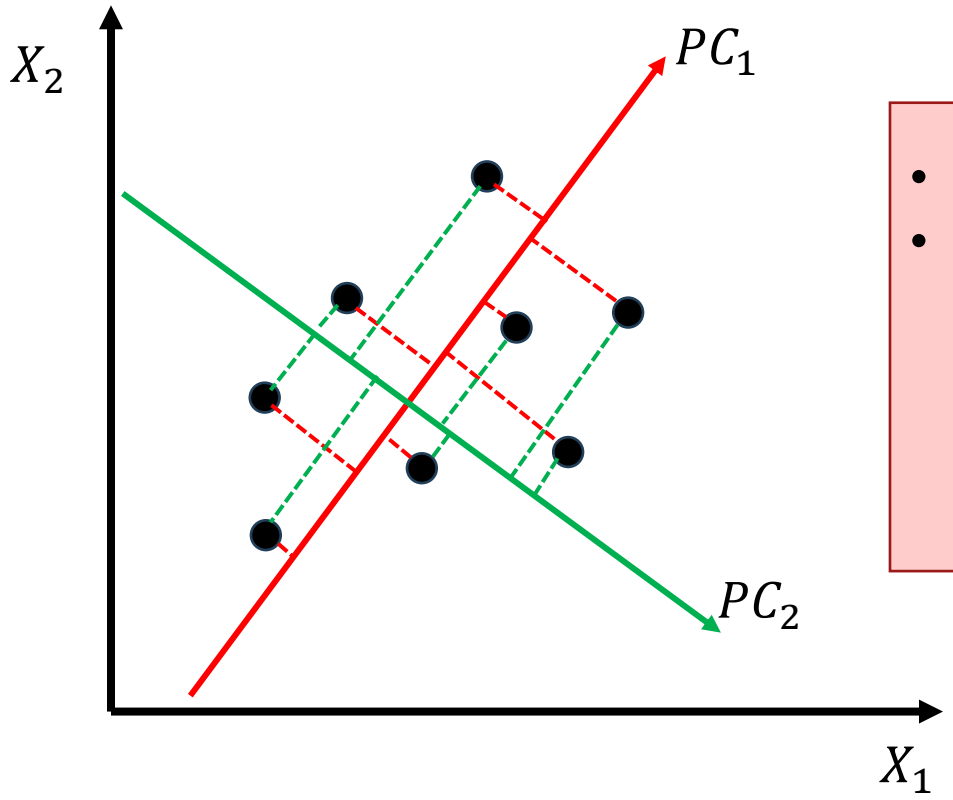
- In pratica la PCA effettua una **proiezione ortogonale** dei dati su uno spazio definito da nuove dimensioni dette componenti principali. Queste sono tali per cui la varianza delle coordinate dei dati proiettati sulle nuove dimensioni è massima per le prime dimensioni.



- Ogni osservazione è un punto.
- Le linee tratteggiate rappresentano le coordinate dei punti nello spazio definito dalle variabili di partenza X_1 e X_2.

INTERPRETAZIONE GEOMETRICA

- In pratica la PCA effettua una **proiezione ortogonale** dei dati su uno spazio definito da nuove dimensioni dette componenti principali. Queste sono tali per cui la varianza delle coordinate dei dati proiettati sulle nuove dimensioni è massima per le prime dimensioni.



- Ogni osservazione è un punto.
- Le linee tratteggiate rappresentano le coordinate dei punti nello spazio definito dalle componenti principali PC_1 e PC_2.

PASSARE DALLO SPAZIO ORIGINALE ALLO SPAZIO DELLE COMPONENTI PRINCIPALI

- Per passare dalle coordinate dei punti nello spazio originale a quelle nello spazio definito dalle componenti principali, basta applicare la trasformazione lineare definita dai loadings.

- Dataset originale: n osservazioni x m variabili (a media nulla).
  - $\boldsymbol{x_i}$ → vettore colonna contenente le n osservazioni per la variabile X_i
  - $\boldsymbol{X}$ matrice dei dati originali (n righe, m colonne).

$$\boldsymbol{X} = [\boldsymbol{x_1}\ \boldsymbol{x_2}\ \dots\ \boldsymbol{x_m}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

  - La media campionaria di ciascuna colonna di X deve essere 0.

COORDINATE DEI DATI NEL SISTEMA DELLE COMPONENTI PRINCIPALI

- Calcoliamo le nuove coordinate dei dati riferite al sistema di riferimento delle componenti principali, $PC_1, PC_2, \ldots, PCm$:

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \cdot \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m1} & v_{m2} & \cdots & v_{mm} \end{bmatrix}$$

Y (n x m)

Matrice dei dati trasformati, detti scores. La riga i-esima contiene le coordinate dell'osservazione i-esima nello spazio delle componenti principali.

X (n x m)

Matrice dei dati originali. La riga i-esima contiene le coordinate dell'osservazione i-esima nello spazio delle variabili originali.

V (m x m)

Matrice dei loadings, i coefficienti che definiscono la combinazione lineare. La colonna k-esima rappresenta i loadings della componente principale k-esima.

COORDINATE DEI DATI NEL SISTEMA DELLE COMPONENTI PRINCIPALI

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{V}$$

Dati trasformati o scores

Dati originali

Coefficienti o loadings

- La matrice **V** è di fatto una **matrice di rotazione** che consente di passare dalle coordinate nel sistema di riferimento originale, alle coordinate nel sistema di riferimento delle componenti principali.

COME TORNARE NELLO SPAZIO DELLE VARIABILI ORIGINALI

- Poiché le colonne di **V** sono tra loro ortogonali e hanno norma 1, si ha che:

$$\boldsymbol{V} \cdot \boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{V}^{\boldsymbol{T}} \cdot \boldsymbol{V} = \boldsymbol{I}$$

- Calcolo delle coordinate nello spazio delle variabili originali partendo dalle coordinate nello spazio delle componenti principali:

$$\boldsymbol{Y} = \boldsymbol{X} \cdot \boldsymbol{V}$$

$$\boldsymbol{Y} \cdot \boldsymbol{V}^{\boldsymbol{T}} = \boldsymbol{X} \cdot \boldsymbol{V} \cdot \boldsymbol{V}^{\boldsymbol{T}}$$

$$\boldsymbol{X} = \boldsymbol{Y} \cdot \boldsymbol{V}^{\boldsymbol{T}}$$

RIDUZIONE DELLA DIMENSIONALITA' (1/2)

- La PCA ci fornisce dunque una nuova rappresentazione dei dati nello spazio delle componenti principali, PC_1, PC_2, …,PCm, che ha 2 caratteristiche fondamentali:
  - Le componenti principali sono scorrelate
  - La varianza dei dati proiettati lungo le componenti principali decresce all'aumentare delle componenti, la maggior parte della varianza complessiva è concentrata nelle prime componenti principali

$$var(PC_1) > var(PC_2) > var(PC_3) > \cdots > var(PC_m)$$

- Per ridurre la dimensionalità possiamo considerare solo le prime p componenti principali.

RIDUZIONE DELLA DIMENSIONALITA' (2/2)

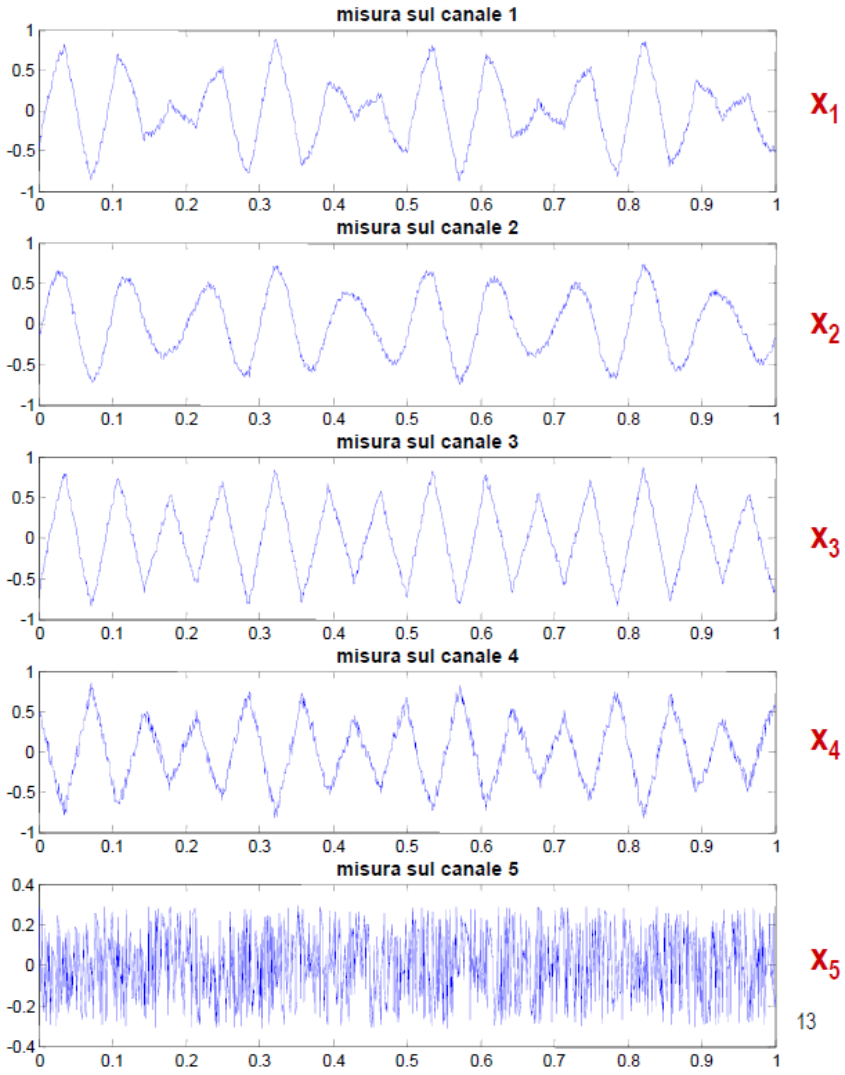
- Invece che memorizzare $\boldsymbol{Y}$, memorizziamo solo le prime p colonne di $\boldsymbol{Y}$.
- p viene scelto in modo da garantire che le prime p colonne di $\boldsymbol{Y}$ contengano la maggior parte della varianza dei dati originali.
- Se la PCA viene utilizzata per realizzare una compressione dei dati, utilizzando la matrice $\boldsymbol{V^T}$ possiamo ricostruire, con un certo errore dovuto alla compressione, i dati originali $\boldsymbol{X}$.

$$\tilde{\boldsymbol{X}} = \boldsymbol{Y}_{\boldsymbol{p}}\boldsymbol{V^T}$$

Ricostruzione dei dati in **X**

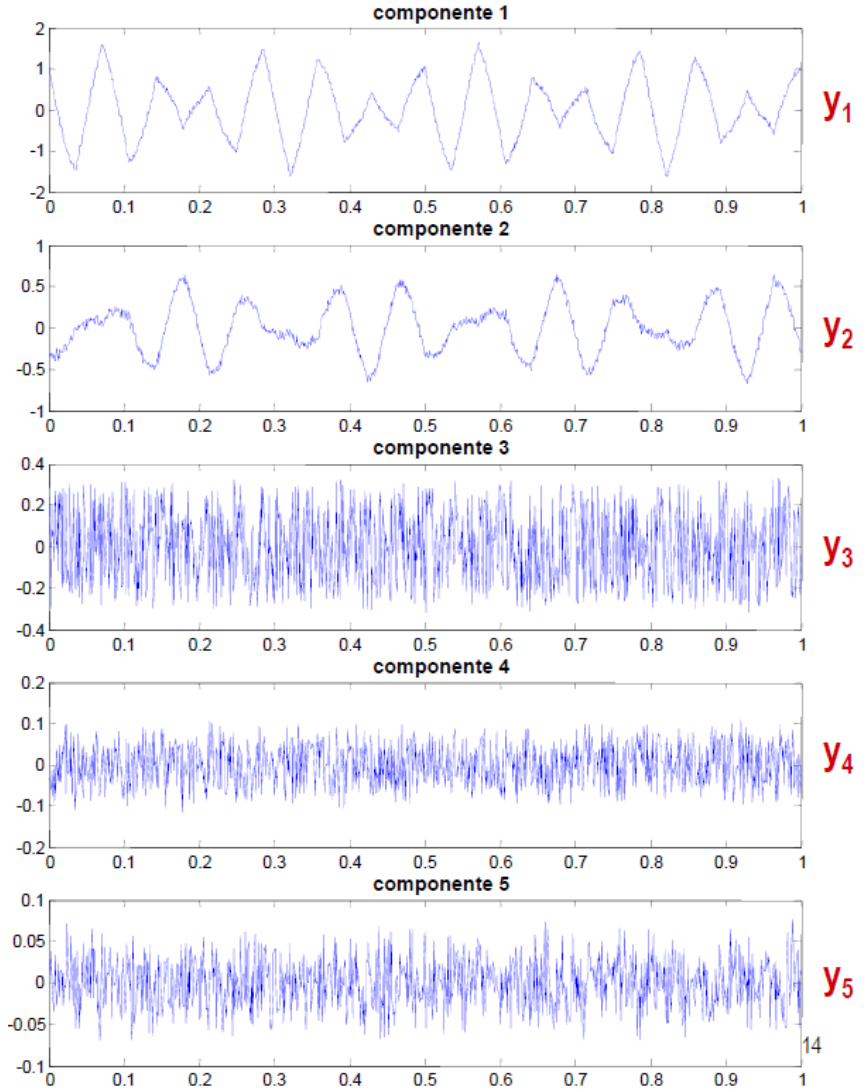
Prime p colonne di **Y**

ESEMPIO (m=5, n=1000)



x_1 x_2 x_3 x_4 x_5

PCA

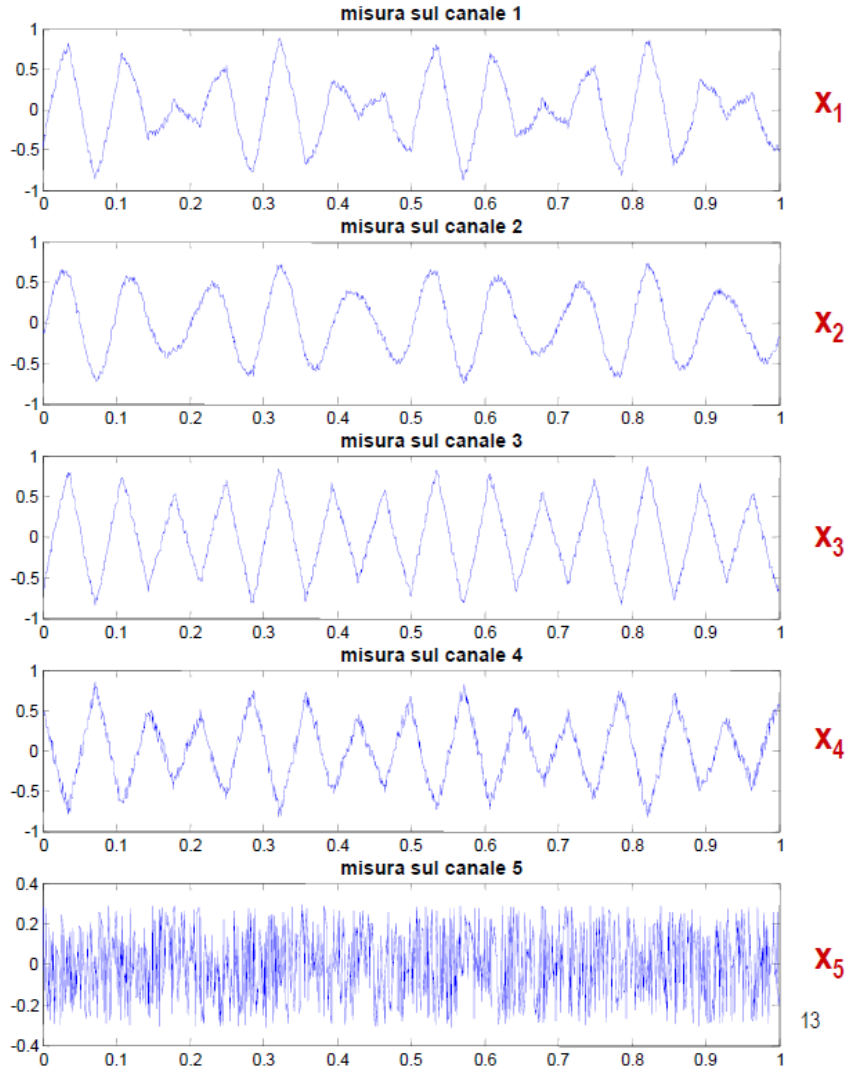


y_1 y_2 y_3 y_4 y_5

Esempio tratto dal corso di Analisi di Dati Biologici del prof. G. Sparacino.

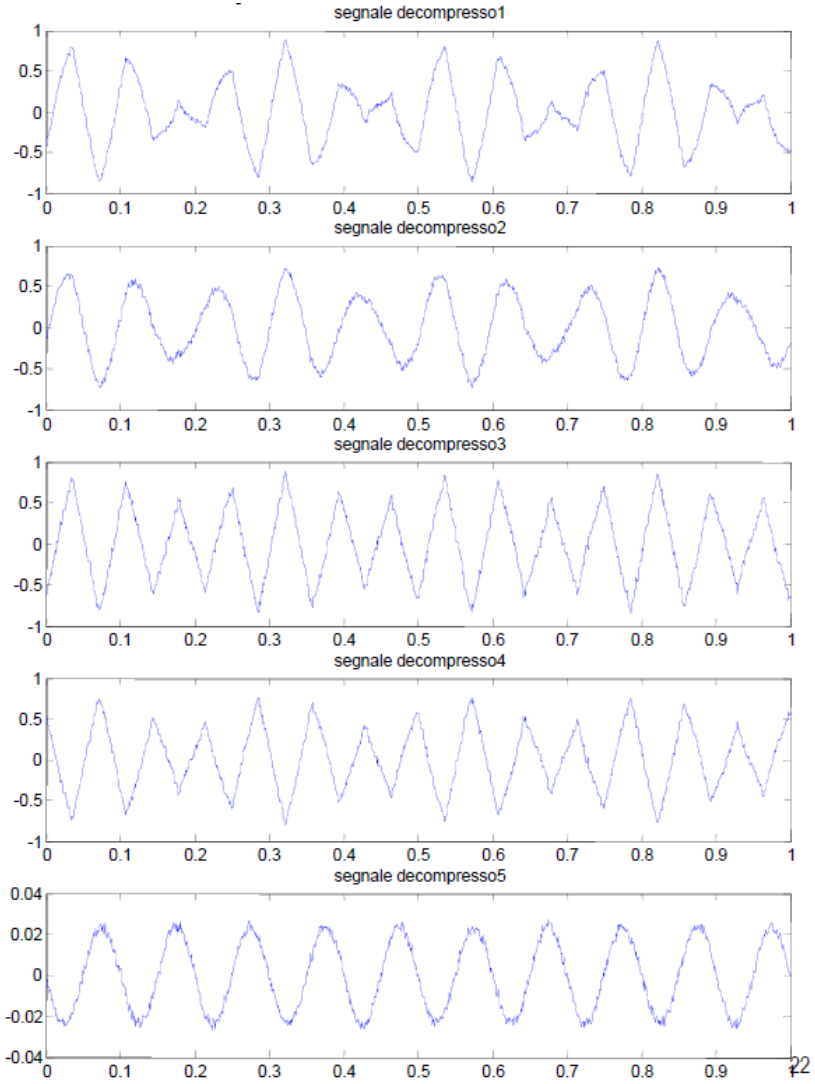
ESEMPIO (m=5, n=1000)

Segnali originali



x_1 x_2 x_3 x_4 x_5

Segnali ricostruiti usando solo le prime 2 componenti principali



COME SI CALCOLANO I LOADINGS?

- Come possiamo calcolare i valori dei loadings, ovvero la matrice **V**, che mi consente di realizzare la PCA?

- Le colonne della matrice **V** sono gli autovettori della matrice di covarianza di **X** ordinati secondo l'ordine decrescente dei rispettivi autovalori.

CALCOLO DEI LOADINGS

- Calcolo della **matrice di covarianza** di **X**: **S**
- **S** è una matrice m x m, reale, e simmetrica, avente sulla diagonale le varianze delle colonne di **X** (le variabili), fuori dalla diagonale le covarianze campionarie tra coppie di colonne di **X**.
  - Elemento in posizione i,j: covarianza tra la colonna i-esima e la colonna j-esima di **X**
- Calcoliamo **autovalori** e **autovettori** di **S**.

AUTOVALORI E AUTOVETTORI

Sia $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ una matrice quadrata di N righe e N colonne.

Se esistono un vettore $\boldsymbol{v} \in \mathbb{R}^{N}$ e uno scalare λ (anche complesso) tali che:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

si dice che $\boldsymbol{v}$ è **autovettore** di $\boldsymbol{A}$ e λ il suo **autovalore** corrispondente.

Proprietà:

- Una matrice di dimensione N x N ha al massimo N autovalori distinti (reali o complessi).
- Gli autovalori sono gli zeri del polinomio caratteristico:

$$\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$$

AUTOVALORI E AUTOVETTORI

- Se la matrice **A** è simmetrica ($\boldsymbol{A = A^T}$) e reale, come la matrice **S**:
  - Gli autovalori sono tutti reali
  - Gli autovettori corrispondenti agli autovalori distinti sono tra loro ortogonali:

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

CALCOLO DEI LOADINGS

- La matrice di covarianza **S**, di dimensione m x m, è reale e simmetrica.
  - Gli autovalori sono reali
  - Gli autovettori distinti sono tra loro sono ortogonali

- **Gli autovettori di S rappresentano i loadings** che definiscono le componenti principali.
- Le componenti principali sono ortogonali tra loro → quindi scorrelate.

- In che ordine vengono considerati gli autovettori per definire le componenti principali? → In base agli autovalori corrispondenti.

DEFINIZIONE DELLA MATRICE **V**

- Ordiniamo i gli autovalori dal più grande al più piccolo:

Autovalori: $\lambda_1 > \lambda_2 > \cdots > \lambda_m$

Autovettori: $\boldsymbol{v_1}$ $\boldsymbol{v_2}$ ... $\boldsymbol{v_m}$

- L'autovettore $\boldsymbol{v_1}$ corrispondente all'autovalore massimo, λ_1, rappresenta il vettore dei loadings della prima componente principale, ovvero la prima colonna di $\boldsymbol{V}$.
- Gli autovettori, $\boldsymbol{v_1}, \boldsymbol{v_2}, \ldots, \boldsymbol{v_m}$, in questo ordine definiscono le colonne di $\boldsymbol{V}$:

$$\boldsymbol{V}=[\boldsymbol{v_1}\ \boldsymbol{v_2}\ \ldots\ \boldsymbol{v_m}]$$

VARIANZA SPIEGATA DALLE COMPONENTI PRINCIPALI

- Gli **autovalori** rappresentano la varianza dei dati proiettati lungo le componenti principali.
  - λ_1 rappresenta la varianza campionaria degli score relativi a PC_1, ovvero la varianza della prima colonna di $\boldsymbol{Y}$
  - λ_2 rappresenta la varianza campionaria degli score relativi a PC_2, ovvero la varianza della seconda colonna di $\boldsymbol{Y}$
  - …

$$\boldsymbol{Y} = [\boldsymbol{y_1}\ \boldsymbol{y_2}\ \cdots\ \boldsymbol{y_m}]$$

$$\lambda_k = s^2_{\boldsymbol{y_k}}, \qquad k = 1, \dots, m$$

Varianza campionaria di $\boldsymbol{y_k}$

FRAZIONE DI VARIANZA SPIEGATA DALLE COMPONENTI PRINCIPALI

- Varianza complessiva dei dati originali:

$$\sum_{k=1}^{m} s_{x_k}^2 = \sum_{k=1}^{m} \lambda_k$$

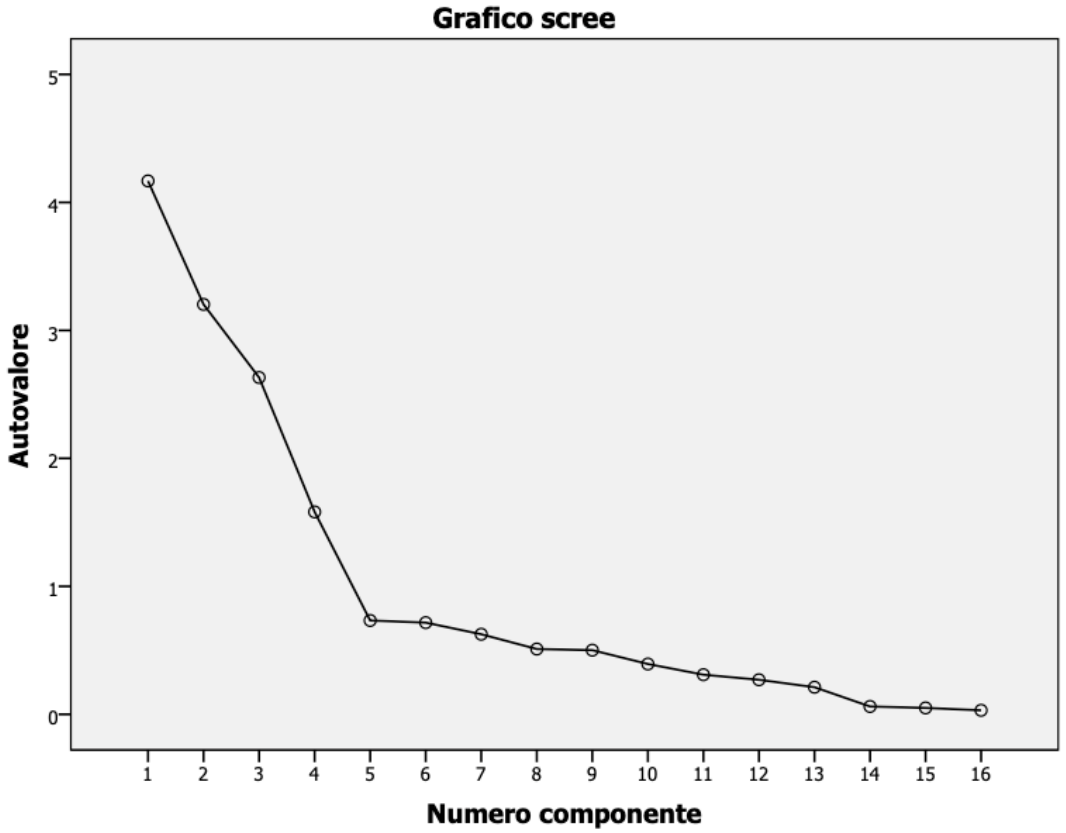
Varianza campionaria della colonna k-esima di **X**

- Frazione della varianza complessiva spiegata dalla componente k-esima:

$$\frac{\lambda_k}{\sum_{k=1}^{m} \lambda_k}$$

SCELTA DEL NUMERO DI COMPONENTI PRINCIPALI

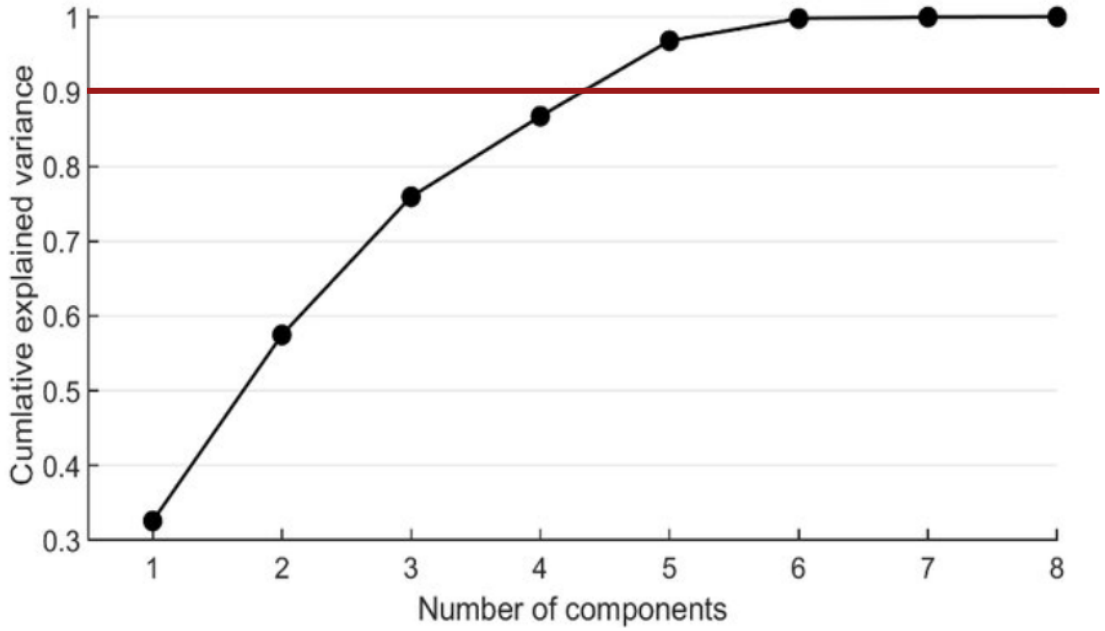
- Rappresentiamo graficamente gli autovalori (*scree plot*) e scegliamo il numero di componenti che corrisponde al punto di gomito.



Scegliamo *5* componenti, perché dalla sesta componente in poi la varianza spiegata dalle singole componenti è parecchio limitata rispetto a quella spiegata dalle prime *5* componenti.

SCELTA DEL NUMERO DI COMPONENTI PRINCIPALI

- In alternativa possiamo rappresentare la frazione di varianza spiegata dalle prime p componenti principali, al variare di p.
- Scegliamo il valore di p che consente di spiegare almeno una certa percentuale (es. 90%) della varianza totale.



Per spiegare almeno il 90% della varianza totale servono 5 componenti.

PCA - RIASSUNTO

1. Centramento dei dati: ad ogni colonna di **X** si sottrae la sua media.
2. Calcolo della matrice di covarianza di **X** → **S** (m x m)
3. Calcolo di autovettori e autovalori di **S.**
   - Gli autovettori rappresentano i coefficienti per definire le componenti principali.
   - L'ordine delle componenti principali è stabilito dall'ordine degli autovalori.
   - L'autovettore corrispondente all'autovalore massimo rappresenta i coefficienti (loadings) della prima componente principale.
4. Trasformazione dei dati passando alle coordinate nello spazio delle componenti principali: $\mathbf{Y}=\mathbf{X}\cdot\mathbf{V}$
5. Scelta delle prime p componenti principali che rappresentano gran parte della varianza delle variabili originali (scree plot o plot della frazione di varianza spiegata dalle prime p componenti).
6. Le coordinate delle n osservazioni lungo le p componenti principali rappresentano il dataset trasformato e ridotto.