METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 19: NOTE SULLA PCA

A.A. 2024-2025

Prof. Martina Vettoretti

NOTE SULLA PCA

- Standardizzazione delle variabili
- Singular value decomposition (SVD)

STANDARDIZZAZIONE DELLE VARIABILI

- Se le variabili originali $X_1, X_2, \dots, X_m$ presentano scale diverse, ovvero varianze diverse, il risultato della PCA potrebbe risultare polarizzato dalle variabili a varianza maggiore.
- In questi casi conviene **standardizzare le variabili** prima di applicare la PCA. In pratica, a ciascuna colonna di X sottraiamo la sua media e dividiamo il risultato per la sua deviazione standard campionaria:

$$\boldsymbol{z_k} = \frac{\boldsymbol{x_k} - \bar{x}_k}{s_k}$$

$\boldsymbol{x_k}$: colonna k-esima di **X**
$\bar{x}_k$: media campionaria di $\boldsymbol{x_k}$
s_k: deviazione standard campionaria di $\boldsymbol{x_k}$

- Matrice dei dati standardizzati:

$$\boldsymbol{Z} = [\boldsymbol{z_1}\ \boldsymbol{z_2}\ \dots\ \boldsymbol{z_m}]$$

PCA SULLE VARIABILI STANDARDIZZATE

- Realizziamo quindi la PCA sulle variabili standardizzate → Matrice **V** costruita a partire dagli **autovettori e autovalori della matrice di covarianza di Z**.

- Si può dimostrare che la matrice di covarianza di Z è equivalente alla matrice di correlazione di X (dati originali, non standardizzati).

- Possiamo quindi realizzare la PCA costruendo la matrice V a partire dagli **autovalori e autovettori della matrice di correlazione di X**.

SINGULAR VALUE DECOMPOSITION (SVD)

- **X** matrice qualsiasi di dimensione n x m.
- La matrice **X** può essere scomposta nel prodotto di 3 matrici:

$$X = U \cdot D \cdot V^T$$

- **U:** matrice n x n le cui colonne sono ortogonali e a norma 1.
  - Le colonne di **U** sono gli autovettori di $X \cdot X^T$.
- **D:** matrice n x m diagonale, con valori non negativi decrescenti sulla diagonale, detti **valori singolari di X**.
  - Gli elementi sulla diagonale sono la radice quadrata degli autovalori di $X^T \cdot X$ ordinati dal più grande al più piccolo.
- **V:** matrice m x m le cui colonne sono ortogonali e a norma 1.
  - Le colonne di **V** sono gli autovettori di $X^T \cdot X$ corrispondenti agli autovalori sulla diagonale di **D**.

CORRISPONDENZA TRA AUTOVETTORI DI $X^T X$ E S

- La matrice di covarianza campionaria di **X**, **S**, si può scrivere come:

$$S = \frac{1}{n-1} X^T X$$

- Gli autovettori di $X^T X$ sono anche autovettori di **S**.
- Chiamiamo $\lambda_{X^T X}$ il vettore degli autovalori di $X^T X$, ordinato in ordine decrescente (elementi sulla diagonale di D elevati al quadrato).
- Chiamiamo λ_S il vettore degli autovalori di S, ordinato in ordine decrescente.
- Tra i due abbiamo la relazione:

$$\lambda_S = \frac{\lambda_{X^T X}}{n-1}$$

- L'autovettore relativo al più grande autovalore di $X^T X$ è uguale all'autovettore corrispondente al più grande autovalore di **S**.

SVD PER REALIZZARE LA PCA

- **X**: matrice dei dati centrati o standardizzati.
- Fattorizziamo **X** tramite SVD:

$$\boldsymbol{X} = \boldsymbol{U} \cdot \boldsymbol{D} \cdot \boldsymbol{V}^{\boldsymbol{T}}$$

- La matrice di rotazione che mi consente di passare alle coordinate nello spazio delle componenti principali è **V.**
- Dati trasformati: $\mathbf{Y} = \boldsymbol{X} \cdot \boldsymbol{V}$
- Autovalori della matrice di covarianza di **X**:

$$\lambda_k = \frac{d_{kk}^2}{n-1}, k = 1, \dots, m$$

dove d_{kk} è l'elemento in posizione k,k della matrice **D.**

PERCHE' PCA CON LA SVD?

- Quando la dimensionalità dei dati è elevata la SVD mi consente di risolvere il problema in maniera più efficiente rispetto al calcolo di autovettori e autovalori di **S**.