L'ANALISI IN COMPONENTI PRINCIPALI (PCA)

La PCA è una tecnica statistica utilizzata per ridurre la dimensionalità di un dataset, mantenendo quanta più variabilità possibile. Questo si ottiene trasformando le variabili originali in un nuovo sistema di coordinate, le **componenti principali** (PC), che sono combinazioni lineari delle variabili iniziali. Ogni componente principale è costruita per essere ortogonale alle altre e spiegare la massima variabilità residua nei dati.

Un esempio semplice per illustrare la PCA è un **dataset sulle diete di paesi europei**, con variabili come il consumo di cereali, zucchero, verdure, carne, latte e uova. Il dataset viene rappresentato in una matrice di dimensioni $n \times p$, dove n è il numero di osservazioni (paesi) e p è il numero di variabili (alimenti). Ad esempio:

Paese	Cereali	Zucchero	Verdure	Carne	Latte	Uova
Danimarca	70.5	39.5	50.0	105.8	145.2	14.3
Grecia	109.8	30.0	229.5	77.1	63.1	11.3
Italia	110.2	27.9	181.9	88.0	65.0	11.1
...	...	...	...	...	...	...

Per applicare la PCA è necessario:

1. **Centrare la matrice dei dati** sottraendo la media di ciascuna variabile, per garantire che il baricentro della nube di punti coincida con l'origine degli assi. Ad esempio, i valori centrati sono calcolati come: $x_{ij} = y_{ij} - \mu_j$, dove μ_j è la media della variabile j.
2. **Standardizzare i dati** (se necessario) per rendere le variabili comparabili in caso di scale diverse. La standardizzazione si calcola dividendo i valori centrati per la deviazione standard della rispettiva variabile: $z_{ij} = \frac{y_{ij} - \mu_j}{\sigma_j}$.
3. **Calcolo delle proiezioni ortogonali** che riducono la dimensionalità dei dati mantenendo il massimo dell'informazione possibile (varianza). Questa trasformazione geometrica consente di rappresentare i dati in un nuovo spazio semplificato, ideale per l'interpretazione o per ulteriori analisi

Questi passaggi trasformano i dati in modo che la PCA possa concentrarsi sulle relazioni tra le variabili, eliminando gli effetti della scala e semplificando la complessità dei dati mantenendo la struttura fondamentale.

DEFINIZIONE DEL MODELLO

La PCA trasforma il dataset in un nuovo insieme di variabili (le componenti principali), ordinate per importanza in termini di variabilità spiegata. L'obiettivo è identificare un sottospazio di di mensioni ridotte (q), con $q<p$, che mantenga la massima informazione possibile.

Ottica analitica:

la PCA individua un numero limitato di **nuove variabili** che combinano linearmente quelle di partenza, eliminando ridondanze. Queste nuove variabili sono le componenti principali.

Ottica geometrica:

la matrice dei dati può essere interpretata come una configurazione di punti nello spazio R^p. La PCA cerca un sottospazio R^q, con $q<p$, che massimizzi la variabilità spiegata dalle proiezioni dei punti.

TRASFORMAZIONI SULLA MATRICE DEI DATI

Le trasformazioni iniziali dei dati sono cruciali per garantire un'applicazione corretta della PCA.

1. **Centratura della matrice dei dati**: ogni variabile è centrata sottraendo la sua media:

$$x_{ij} = y_{ij} - \mu_j ,$$

dove μ_j è la media della variabile j. Geometricamente, la centratura sposta il baricentro della nube di punti nell'origine.

$$\mathbf{X} = \begin{pmatrix} y_{1,1} & y_{1,2} & \cdots & y_{1,p} \\ y_{2,1} & y_{2,2} & \cdots & y_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ y_{n,1} & y_{n,2} & \cdots & y_{n,p} \end{pmatrix} - \begin{pmatrix} \mu_1 & \mu_2 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_p \\ \cdots & \cdots & \cdots & \cdots \\ \mu_1 & \mu_2 & \cdots & \mu_p \end{pmatrix} =$$

$$= \begin{pmatrix} (y_{1,1} - \mu_1) & (y_{1,2} - \mu_2) & \cdots & (y_{1,p} - \mu_p) \\ (y_{2,1} - \mu_1) & (y_{2,2} - \mu_2) & \cdots & (y_{2,p} - \mu_p) \\ \cdots & \cdots & \cdots & \cdots \\ (y_{n,1} - \mu_1) & (y_{n,2} - \mu_2) & \cdots & (y_{n,p} - \mu_p) \end{pmatrix}$$

2. **Standardizzazione (opzionale)**: quando le variabili hanno scale diverse, è necessario standardizzare dividendo ogni valore per la deviazione standard della rispettiva variabile:

$$z_{ij} = \frac{y_{ij} - \mu_j}{\sigma_j} .$$

dove:

- dove σ_j è la deviazione standard della variabile j, definita come:

$$\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \mu_j)^2}.$$

Queste trasformazioni garantiscono che la PCA si concentri sulle relazioni tra variabili piuttosto che sulla loro scala.

PROIEZIONI ORTOGONALI

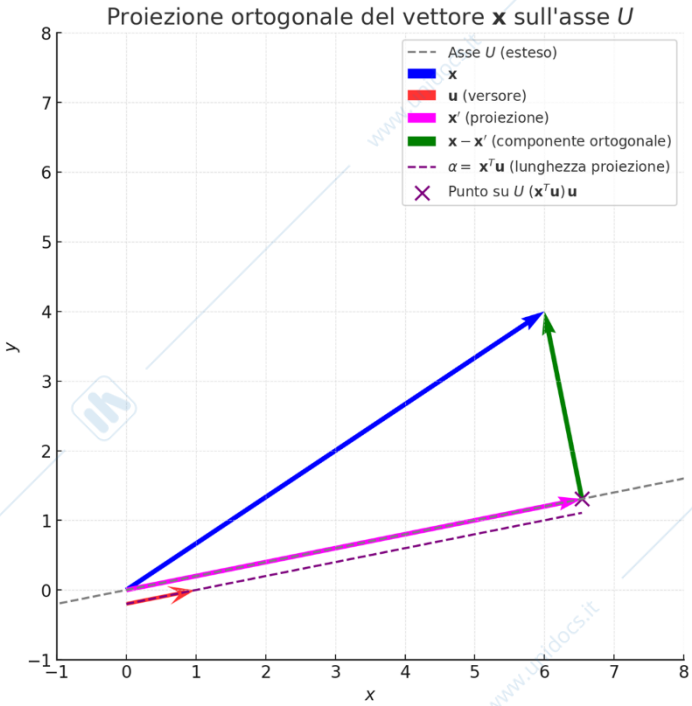
La PCA utilizza le proiezioni ortogonali per identificare i sottospazi che massimizzano la variabilità dei dati. La proiezione ortogonale di un punto o vettore su un asse U corrisponde a trovare il punto più vicino su U, seguendo una direzione ortogonale.

1. **Proiezione di un vettore su un asse**: La proiezione di un vettore x sul versore u che definisce l'asse di proiezione U è:

$\alpha = x^T u$, dove α è la coordinata della proiezione. Il vettore proiettato è: $x' = \alpha u = (x^T u)u$.

- Il **versore** è semplicemente il vettore che rappresenta la direzione dell'asse, ma con lunghezza unitaria, cioè norma pari a 1. La proiezione ortogonale di un vettore x su una direzione U dipende solo dalla direzione di U, non dalla sua lunghezza

quindi usiamo la trasposta del vettore da proiettare, perchè cosi sono definiti i prodotti scalari (sennò non sarebbe possibile fare il prodotto tra due vettori colonna) per ottenere α che ci esprime la lunghezza lungo U. Da qui ci troviamo x' che è un vettore quindi ci indica anche la direzione oltre la lunghezza della proiezione e combinandoli possiamo identificare il punto nello spazio.



Esempio di proiezione ortogonale di vettori su un asse

Si consideri di voler proiettare ortogonalmente i vettori $\mathbf{v}_1 = [-7, 6]$ e $\mathbf{v}_2 = [3, 5]$ sull'asse individuato dal vettore $U_2 = [-9, 2]$.

- Il versore che identifica l'asse U_2 è $u_2 = [-0.9762, 0.2169]$
- Le coordinate dei vettori $\mathbf{v}_1$ e $\mathbf{v}_2$ sono

$$\alpha_1 = \begin{pmatrix} -7 & 6 \end{pmatrix} \times \begin{pmatrix} -0.9762 \\ 0.2169 \end{pmatrix} = 8.13$$

$$\alpha_2 = \begin{pmatrix} 3 & 5 \end{pmatrix} \times \begin{pmatrix} -0.9762 \\ 0.2169 \end{pmatrix} = -1.84$$

- Le coordinate dei punti che individuano i vettore proiettati $\hat{\mathbf{v}}_1$ e $\hat{\mathbf{v}}_2$ sono rispettivamente $\hat{\mathbf{v}}_1 = [-7.94, 1.76]$ e $\hat{\mathbf{v}}_2 = [1.8, -0.4]$

2. **Proiezione di una matrice**: Se X è la matrice dei dati centrati, le coordinate delle osservazioni nel sottospazio sono:

$$\overline{\alpha} = XU$$

dove *U* è la matrice degli autovettori.

$$\overline{\alpha} = \mathbf{Xu} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \times \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_p \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \dots \\ \alpha_p \end{pmatrix}$$

α è un vettore colonna di lunghezza *n*, contenente le coordinate delle proiezioni.

FORMALIZZAZIONE DEL METODO (VARIANZA)

In PCA, Si vuole trovare il versore *u* (direzione dell'asse *U*) che ottimizza la rappresentazione dei dati proiettati lungo *U*. Questo può essere fatto minimizzando la **somma dei quadrati delle componenti ortogonali** (le distanze tra i punti e l'asse *U*) oppure massimizzando la **varianza** delle proiezioni (Maggiore è la varianza, maggiore è la diversità tra i dati lungo una certa direzione) perché in questo modo facciamo si che la proiezione dei dati sia il più simile possibile ai dati originali.
Questo si formalizza come un problema di **ottimizzazione vincolata**.

1. **Definizione del problema**:
   a. Si cerca un versore u che massimizzi la varianza delle proiezioni: $\max_u \quad Var(Xu)\ soggetto\ a\ \| u \| = 1.$
   b. La varianza delle proiezioni è: $Var(Xu) = u^T X^T Xu$

$X^T X$ sarebbe la matrice di covarianza (o correlazione, per dati standardizzati) dove abbiamo eliminato il fattore $\frac{1}{n-1}$ per semplicità di calcolo e perchè ininfluente sul calcolo dei autovettori

2. **Metodo dei moltiplicatori di Lagrange**:
   a. La funzione obiettivo con vincolo *//u1//=1 (ovvero $u^T u=1$)* diventa:

$$L(u, \lambda) = u^T X^T Xu - \lambda(u^T u - 1)$$

   b. Ponendo il gradiente a zero:

$$\frac{\partial L}{\partial u} = 2X^T Xu - 2\lambda u = 0$$

   c. Si ottiene l'equazione caratteristica: $X^T Xu = \lambda u$ dove:
      i. λ è l'autovalore (varianza spiegata dall'asse),
      ii. *u* è l'autovettore (direzione della componente principale).
3. **Iterazione per componenti successive:**

a. Dopo aver trovato il primo asse, si cercano gli assi successivi imponendo vincoli di **ortogonalità** con quelli già calcolati ($u_1^T u_2 = 0$).

$$L(u_2, \lambda_2, \mu) = u_2^T X^T X u_2 - \lambda_2(u_2^T u_2 - 1) - \mu(u_1^T u_2)$$

4. **Calcolo delle coordinate dei dati proiettati sui nuovi assi principali (sottospazio):** una volta trovati gli **autovettori** *U (*le colonne che rappresentano gli assi principali), questi devono essere **moltiplicati** con la matrice *X* dei dati originali (centrati) per ottenere la matrice C_U che utilizzeremo per ottenere una **rappresentazione ridotta dei dati** (selezionando i primi k autovettori con k<p poichè sono quelli che spiegano più variabilità, in modo decrescente)

$$C_U = XU$$

Dove:

- *X* è la matrice $n \times p$ dei dati centrati (con *n* osservazioni e *p* variabili).
- *U* è la matrice $p \times p$ degli autovettori (ciascun autovettore rappresenta un asse principale).
- C_U è la matrice $n \times p$ delle coordinate principali (proiezioni dei dati sui nuovi assi).

Ogni **riga** di C_U rappresenta un'osservazione (un individuo) espresso nello spazio dei componenti principali.

Ogni **colonna** di C_U rappresenta un **componente principale**:

- PC_1 : La direzione con la massima varianza spiegata.
- PC_2 : La seconda direzione con la massima varianza, ortogonale a *PC1*.
- PC_k : Altri componenti principali.

FORMALIZZAZIONE DEL METODO (SOMMA DEI QUADRATI DELLE COMPONENTI ORTOGONALI)

1. **Definizione del problema**:
   a. Si cerca un versore v_α che **minimizzi la somma dei quadrati delle componenti ortogonali**:

$$\min_{v_\alpha} \quad \sum_{j=1}^{n} \| x_j - (x_j^T v_\alpha) v_\alpha \|^2 \quad soggetto\ a \ \| v_\alpha \| = 1$$

dove:

- x_j è la *j*-esima osservazione nello spazio originale.
- $(x_j^T v_\alpha) v_\alpha$ è la proiezione di x_j sul versore v_α .

Minimizzare le componenti ortogonali equivale a **massimizzare la somma delle lunghezze delle proiezioni al quadrato**, cioè:

$$\max_{v_\alpha} \quad \sum_{j=1}^{n} (x_j^T v_\alpha)^2$$

2. **Matrice XX^T:**
   In forma matriciale, considerando X come la matrice dei dati centrati ($n \times p$), la somma delle proiezioni al quadrato può essere scritta come:

$\sum_{j=1}^{n}(x_j^T v_\alpha)^2 = v_\alpha^T X X^T v_\alpha.$

   Pertanto, il problema di ottimizzazione diventa:

$$\max_{v_\alpha} \quad v_\alpha^T X X^T v_\alpha \ soggetto\ a\ \| v\alpha \| = 1.$$

XX^T è una matrice con gli stessi autovalori della matrice di covarianza ($X^T X$) e quindi con autovettori correlati e ricavabili dalla matrice di covarianza (e viceversa) perchè anche se lavorano in spazi diversi (spazio delle osservazioni n / spazio delle variabili p) entrambi condividono la matrice X. Di solito è utilizzata quando vogliamo analizzare le osservazioni oppure quando le osservazioni sono di meno delle variabili

3. **Metodo dei moltiplicatori di Lagrange**:
   a. La funzione obiettivo con vincolo *//u1//=1 (ovvero $u^T u=1$)* diventa:
   $$L(v_\alpha, \lambda_\alpha) = v_\alpha^T X X^T v_\alpha - \lambda_\alpha (v_\alpha^T v_\alpha - 1).$$
   b. Ponendo il gradiente a zero:
   $$\frac{\partial L}{\partial v_\alpha} = 2XX^T v_\alpha - 2\lambda_\alpha v_\alpha$$
   c. Si ottiene l'equazione caratteristica:
   $$XX^T v_\alpha = \lambda_\alpha v_\alpha.$$

dove:

   i. λ_α è l'autovalore (varianza spiegata dall'asse),
   ii. v_α è l'autovettore (direzione della componente principale).

4. **Iterazione per componenti successive:**
   a. Dopo aver trovato il primo asse, si cercano gli assi successivi imponendo vincoli di **ortogonalità** con quelli già calcolati ($v_1^T v_2 = 0$).
   $$L(v_2, \lambda_2, \mu) = v_2^T X X^T v_2 - \lambda_2 (v_2^T v_2 - 1) - \mu (v_1^T v_2).$$

5. **Calcolo delle coordinate dei dati proiettati sui nuovi assi principali (sottospazio):**
   Una volta determinati gli autovettori v_α, possiamo calcolare le coordinate principali per ciascuna osservazione x_j lungo il componente α usando la formula:
   ○ $C_\alpha(j) = x_j^T v_\alpha.$
   Qui:

• $C_\alpha(j)$ è la j-esima osservazione proiettata lungo l'asse α (componente principale).

Queste coordinate rappresentano le **proiezioni dei dati lungo gli assi principali**. Questa formula a differenza della matrice $\boldsymbol{C_U}$ calcola solo un singolo elemento della matrice ($C_\alpha(j) = C_U[j, \alpha]$) . Per ridurre la dimensionalità dobbiamo calcolarci la matrice $\boldsymbol{C_U}$ (con la formula generale oppure calcolando ogni elemento per il versore di ogni asse principale)

SCELTA DEL NUMERO DI DIMENSIONI

Uno degli aspetti fondamentali della PCA è scegliere il numero di dimensioni (q) in cui proiettare i dati. L'obiettivo è trovare un equilibrio tra la riduzione della dimensionalità e la conservazione della variabilità presente nei dati originali. Questo processo richiede un'analisi accurata della **variabilità spiegata** dalle componenti principali.

Gli autovalori ($\lambda_1, \lambda_2, \ldots, \lambda_p$) della matrice di covarianza (o correlazione, nel caso di dati standardizzati) indicano quanta variabilità ciascuna componente principale spiega. La variabilità totale è la somma di tutti gli autovalori e corrisponde alla traccia (somma degli elementi diagonali) della stessa matrice

$$tr(X^T X) = \sum_{\alpha=1}^{p} \lambda_\alpha$$

La percentuale di variabilità spiegata da una componente k associato all'autovalore λk è calcolata come:

$$inertia_k = \frac{\lambda_k}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100.$$

Se si considerano le prime q componenti principali, la **variabilità cumulativa spiegata** è data da:

$$inertia\ cumulativa = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{\alpha=1}^{p} \lambda\alpha} \times 100.$$

Un buon modello di riduzione della dimensionalità cerca di spiegare almeno l'80%-90% della variabilità totale.

Per la scelta di q, si possono applicare tre criteri principali:

1. **Criterio della variabilità spiegata**: Si seleziona q in modo che la variabilità cumulativa superi una soglia predeterminata.
2. **Criterio dell'eigenvalue-one**: Nel caso di dati standardizzati, si scelgono solo le componenti principali i cui autovalori (λk) sono maggiori di 1. Poiché la varianza di ciascuna variabile standardizzata è pari a 1, autovalori minori di 1 non sono informativi.
3. **Criterio dello scree plot**: Questo grafico traccia gli autovalori in ordine decrescente. La scelta di q corrisponde al punto in cui la curva "si appiattisce" (noto come ginocchio del grafico), poiché le successive componenti spiegano variazioni marginali.

Se si decide di mantenere q componenti principali, la perdita di informazione è rappresentata dalla variabilità associata agli autovalori esclusi:

$$perdita = \frac{\sum_{k=q+1}^{p} \lambda_k}{\sum_{\alpha=1}^{p} \lambda_\alpha} \times 100$$

CONTRIBUTI AGLI ASSI E QUALITÀ DELLA RAPPRESENTAZIONE

La PCA analizza il contributo delle osservazioni e delle variabili nel determinare gli assi principali e valuta quanto bene ciascun dato è rappresentato nel sottospazio, nonostante la distorsione dovuta alla proiezione.

Contributi assoluti delle osservazioni:
Il contributo assoluto di un'osservazione *i* al *k*-esimo asse misura quanto questa contribuisce alla determinazione dell'asse stesso. È dato dalla formula:

$$CA_{i,k} = \frac{z_{i,k}^2}{n\,\lambda_k} \ ,$$

dove:

- $z_{i,k}$ è la coordinata dell'osservazione *i* sulla *k*-esima componente principale,
- λ_k è l'autovalore associato alla componente *k*k
- *n* è il numero totale di osservazioni.

Un valore elevato di $CA_{i,k}$ indica che l'osservazione *i* ha un ruolo significativo nella definizione dell'asse *k*.

Qualità della rappresentazione (coseno quadrato):
Il **coseno quadrato** ($cos^2_{i,k}$) misura quanto bene una specifica osservazione è rappresentata da un asse fattoriale (componente principale). È definito come il rapporto tra la norma (lunghezza) della proiezione del vettore e la norma del vettore nello spazio originale:

$$\cos^2_{i,k} = \frac{\| z_{i,k}^2 \|}{\| x_i^2 \|}$$

Dove:

- $\|z_{i,k}\|$ è la lunghezza della proiezione di *i* sull'asse *k*,
- $\|x_i\|$ è la lunghezza del vettore originale.

Se $cos^2_{i,k}$ è vicino a 1, l'asse *k* rappresenta molto bene l'osservazione *i*. Se invece è vicino a 0, l'osservazione è mal rappresentata su quell'asse.

RAPPRESENTAZIONE GRAFICA

Il cerchio delle correlazioni è una rappresentazione grafica che mostra come le variabili originali si proiettano nello spazio definito dai primi due componenti principali della PCA. Ogni freccia rappresenta una variabile originale, e la sua posizione all'interno del cerchio ci aiuta a capire il comportamento e la relazione tra le variabili.
Le variabili vicine al bordo del cerchio sono ben rappresentate dai componenti principali, mentre quelle vicine al centro lo sono meno. La lunghezza della freccia indica quanto la variabile contribuisce alla varianza spiegata dai componenti principali.
L'angolo tra le frecce fornisce informazioni sulla correlazione tra variabili:

- Angoli piccoli indicano una forte correlazione positiva.
- Angoli di 90° indicano una correlazione nulla.
- Angoli vicini a 180° indicano una correlazione negativa.

Ad esempio, nel grafico, "Riso" e "Cereali" sono fortemente correlati (angolo piccolo), mentre "Carne" e "Latte" non hanno correlazione (angolo di circa 90°). Inoltre, "Carne" è meglio rappresentata dal secondo componente principale, mentre "Riso" e "Cereali" contribuiscono maggiormente al primo.

Questa rappresentazione può essere estesa in modo congiunto alle unità (osservazioni), proiettandole nello stesso spazio dei componenti principali. Ciò consente di analizzare non solo le relazioni tra le variabili, ma anche quelle tra le variabili e le unità. In questo contesto, i componenti principali permettono di individuare le variabili più significative per ciascuna unità, fornendo un quadro interpretativo completo.
Il cerchio delle correlazioni è quindi utile per interpretare le relazioni tra le variabili, comprendere come i dati originali si distribuiscono nello spazio dei componenti principali e analizzare la loro interazione con le unità.