

Unsupervised Learning and Principal Component Analysis - 03/12

Apprendimento supervisionato vs. Apprendimento non supervisionato

La maggior parte di questo corso si concentra sui metodi di Supervised Learning come la Regressione e la Classificazione.

In quel contesto osserviamo sia un insieme di features X_1, X_2, \dots, X_p per ogni oggetto, che un insieme di variabili di outcome Y . L'obiettivo è poi predire Y usando X_1, X_2, \dots, X_p .

Ora ci concentriamo invece sull'unsupervised learning, nel quale osserviamo solo le features X_1, X_2, \dots, X_p . Non abbiamo interesse nella predizione, perché non abbiamo una variable response Y associata.

Apprendimento non supervisionato

L'obiettivo è scoprire cose interessanti sulle misure: c'è un modo informativo per visualizzare i dati? Possiamo scoprire sottogruppi tra le variabili o tra le osservazioni?

L'Unsupervised Learning è più soggettivo del Supervised Learning, in quanto non c'è un semplice obiettivo per l'analisi, come può essere la predizione della risposta (response).

Le tecniche per l'Unsupervised Learning sono di crescente importanza in un certo numero di campi:

- raggruppare sottogruppi di pazienti con il cancro al seno per le misure della loro gene expression;
- raggruppare gruppi di acquirenti in base alla loro cronologia di browsing e acquisti;
- raggruppare film per il loro rating assegnato dagli spettatori.

Principal Component Analysis (analisi dei componenti principali)

Ci concentriamo sull'Analisi dei Componenti Principali (PCA).

Uno strumento principalmente utilizzato per la visualizzazione di dati o per il pre-processing dei dati prima dell'applicazione delle tecniche supervisionate.

PCA produce una rappresentazione low-dimensional (a bassa dimensionalità) di un dataset.

Trova una sequenza di combinazioni lineari delle variabili che hanno varianza massimale e che sono mutualmente non correlate.

First Principal Component

Il first principal component di un insieme di features X_1, X_2, \dots, X_p è la combinazione lineare normalizzata delle features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \dots + \phi_{p1}X_p$$

che ha la più grande varianza.

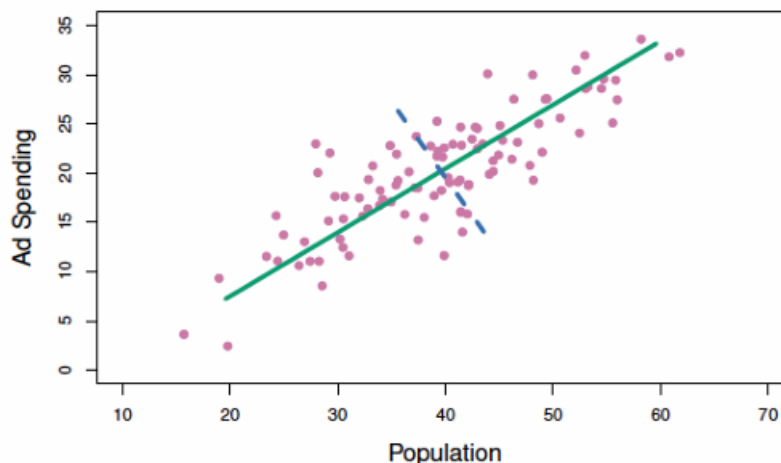
Per normalizzata, intendiamo che $\sum_{j=1}^p \phi_{j1}^2 = 1$.

Ci riferiamo agli elementi $\phi_{11}, \dots, \phi_{p1}$ con il nome di **loadings** (caricamenti ma penso non si traduca) del primo componente principale; insieme, i loadings formano il **principal component loading vector**:

$$\phi_1 = (\phi_{11} \ \phi_{21} \ \dots \ \phi_{p1})^T$$

Limitiamo i loadings in modo che la somma dei loro quadrati sia pari ad 1, visto che se non fosse così settare questi elementi a valori arbitrariamente grandi in valore assoluto potrebbe risultare in una varianza arbitrariamente grande.

Esempio



La population size (pop) e la spesa in pubblicità (ad) per 100 diverse città sono mostrate come cerchietti viola.

La linea continua verde indica la direzione del first principal component.

La linea tratteggiata blu indica la direzione del second principal component.

Computation of principal components (calcolo dei componenti principali)

Supponiamo di avere un data set $n \times p$ chiamato X . Visto che ci interessa solo la varianza, assumiamo che ognuna delle variabili in X sia stata centrata in modo che abbia media pari a 0 (cioè, le column means di X sono zero).

▼ Cos'è la column mean? by ChatGPT

La **column mean** di un dataset è semplicemente la media aritmetica calcolata per una colonna specifica del dataset. In un contesto di analisi dei dati, i dataset sono spesso organizzati in righe e colonne, dove ogni colonna rappresenta una variabile o una caratteristica.

La **column mean** viene calcolata sommando tutti i valori numerici in una colonna e dividendo per il numero di valori validi (non nulli o non mancanti). Questo valore è utilizzato frequentemente per analizzare i dati, riempire valori mancanti (imputazione), o normalizzare i dati.

Se x_1, x_2, \dots, x_n sono i valori della colonna C allora la media della colonna è:

$$\text{Media della colonna} = \frac{\sum_{i=1}^n x_i}{n}$$

Dove n è il numero di valori validi (escludendo i valori mancanti).

Guardiamo poi alla combinazione lineare dei valori dei sample delle features nella forma

$$z_{i1} = \phi_{11}x_{i1} + \phi_{21}x_{i2} + \dots + \phi_{p1}x_{ip} \quad (1)$$

per $i = 1, \dots, n$ che ha la più grande sample variance, condizionatamente al constraint che $\sum_{j=1}^p \phi_{j1}^2 = 1$.

Visto che ognuno degli x_{ij} ha media zero, allora lo stesso vale per z_{i1} (per ogni valore di ϕ_{j1}). Di conseguenza la sample variance (varianza campionaria) di z_{i1} può essere scritta come $\frac{1}{n} \sum_{i=1}^n z_{i1}^2$ ($n \gg 1$).

Pluggin in (1) (mettere nella formula (1)) il first principal component loading vector (vettore dei loading del primo componente principale) risolve il problema di ottimizzazione

$$\underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \frac{1}{n} \sum_{i=1}^n (z_{i1})^2 = \underset{\phi_{11}, \dots, \phi_{p1}}{\text{maximize}} \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_{ij} \right)^2$$

condizionatamente a $\sum_{j=1}^p \phi_{j1}^2 = 1$. Che è la massimizzazione della varianza campionaria.

Questo problema può essere risolto tramite una decomposizione a singolo valore (singular-value decomposition) della matrice X , una tecnica standard in algebra lineare.

▼ Cos'è la singular-value decomposition? by ChatGPT

La **Singular Value Decomposition (SVD)** è una tecnica fondamentale dell'algebra lineare usata per decomporre una matrice in componenti più semplici, utili per molte applicazioni in matematica, statistica, elaborazione di segnali, apprendimento automatico, e altro.

Definizione Matematica

Data una matrice A di dimensioni $m \times n$, la SVD la rappresenta come il prodotto di tre matrici:

$$A = U \Sigma V^T$$

dove:

- U : è una matrice $m \times m$ ortogonale (o unitario se la matrice è complessa), i cui vettori colonna sono chiamati **autovettori sinistri** di A .

Una **matrice ortogonale** è una matrice quadrata A che soddisfa la seguente proprietà fondamentale:

$$A^T A = I$$

- Σ : è una matrice diagonale $m \times n$, in cui i valori sulla diagonale (detti **singular values** o valori singolari) sono non negativi e ordinati in ordine decrescente.
- V^T : è la trasposta (o la coniugata trasposta, se i numeri sono complessi) di una matrice $n \times n$ ortogonale, i cui vettori colonna sono chiamati **autovettori destri** di A .

Intuizione

La SVD permette di scomporre la matrice A in tre parti:

1. U : descrive una rotazione (o riflessione) nello spazio di input.
2. Σ : fornisce informazioni sulla scala, cioè su quanto si allungano o si comprimono i vettori nelle varie direzioni.
3. V^T : rappresenta una rotazione (o riflessione) nello spazio delle colonne.

Proprietà Chiave

- I valori singolari (σ_i) nella matrice Σ sono le radici quadrate degli autovalori della matrice $A^T A$.
- La SVD è definita per qualsiasi matrice A , quadrata o rettangolare.
- La matrice A può essere approssimata in modo ottimale usando un numero ridotto di valori singolari (compressione).

Applicazioni

1. **Compressione delle immagini**: SVD può essere usata per approssimare una matrice di pixel riducendo i ranghi (valori singolari più grandi).
2. **Elaborazione del segnale**: per separare il segnale dal rumore.
3. **Raccomandazioni (ad esempio in sistemi come Netflix)**: per ridurre la dimensionalità dei dati.
4. **Risoluzione di sistemi lineari**: in particolare, quelli sovradeterminati o sottodeterminati.
5. **PCA (Principal Component Analysis)**: SVD è alla base del calcolo delle componenti principali in analisi statistica.

Ci riferiamo a Z_1 con il nome di first principal component, con valori realizzati z_{11}, \dots, z_{n1} .

Il loading vector ϕ_1 con elementi $\phi_{11}, \phi_{21}, \dots, \phi_{p1}$ definisce una direzione nel feature space (spazio delle features) lungo il quale i dati variano di più.

Se proiettiamo gli n data points x_1, \dots, x_n su questa direction, i valori proiettati sono gli stessi principal component scores z_{11}, \dots, z_{n1} .

Second principal component

Il **second principal component** è la combinazione lineare di X_1, \dots, X_p che ha varianza massimale tra tutte le combinazioni lineari che sono non correlate a Z_1 .

I second principal component scores $z_{12}, z_{22}, \dots, z_{n2}$ prendono la forma

$$z_{i2} = \phi_{12}x_{i1} + \phi_{22}x_{i2} + \dots + \phi_{p2}x_{ip}$$

dove ϕ_2 è il second principal component loading vector che ha elementi $\phi_{12}, \phi_{22}, \dots, \phi_{p2}$.

Emerge che limitare Z_2 ad essere incorrelato con Z_1 è equivalente a limitare la direzione ϕ_2 ad essere ortogonale (perpendicolare) alla direction ϕ_1 . E così via.

Le principal component directions $\phi_1, \phi_2, \phi_3, \dots$ sono la sequenza ordinata di right singular vectors della matrice X , e le varianze dei componenti sono $\frac{1}{n}$ volte i quadrati dei singular values. Ci sono al massimo $\min(n-1, p)$ componenti principali.

▼ Cosa sono i right singular vectors? by ChatGPT

I **right singular vectors** di una matrice sono una parte fondamentale della decomposizione ai valori singolari (**Singular Value Decomposition**, o **SVD**) di una matrice.

Definizione dei right singular vectors

Ricordando che data una matrice A di dimensioni $m \times n$, la SVD la rappresenta come il prodotto di tre matrici:

$$A = U\Sigma V^T$$

I **right singular vectors** di A sono i vettori colonna di V . Essi rappresentano una base ortonormale dello spazio delle righe di A . In altre parole:

- Ogni right singular vector è associato a un valore singolare, che rappresenta il contributo di quel vettore nel descrivere l'azione di A in termini di stretching e rotazioni.
- Sono i vettori propri della matrice $A^T A$. Formalmente, se $A^T A$ è simmetrica e definita positiva, allora esistono autovalori λ tali che:

$$A^T A v = \lambda v$$

dove λ sono i quadrati dei valori singolari di A , e v sono i right singular vectors.

Interpretazione

1. **Geometrica:** I right singular vectors indicano le direzioni principali dello spazio in cui A agisce, viste dallo spazio delle righe.
2. **Pratica:** Nelle applicazioni (ad esempio, nella compressione dati o nell'elaborazione del linguaggio naturale), i right singular vectors sono usati per determinare i "modelli latenti" o le componenti principali nei dati rappresentati da A .

In sintesi, i right singular vectors costituiscono una base ortonormale per lo spazio delle righe, e insieme ai valori singolari e ai left singular vectors, descrivono completamente il comportamento di una matrice A .

▼ Cosa sono i singular values? by ChatGPT

I **valori singolari** di una matrice sono numeri non negativi che emergono dalla decomposizione ai valori singolari (**Singular Value Decomposition**, o **SVD**) e descrivono le proprietà geometriche della trasformazione lineare associata alla matrice.

Definizione dei right singular vectors

Ricordando che data una matrice A di dimensioni $m \times n$, la SVD la rappresenta come il prodotto di tre matrici:

$$A = U\Sigma V^T$$

Definizione dei valori singolari

I **valori singolari** di A :

- Sono i numeri non negativi $\sigma_1, \sigma_2, \dots, \sigma_r$, dove r è il rango della matrice A .
- Sono i valori della diagonale della matrice Σ nella SVD.
- Si calcolano come le radici quadrate degli autovalori della matrice $A^T A$ (o AA^T): $\sigma_i = \sqrt{\lambda_i}$ dove λ_i sono gli autovalori di $A^T A$.

Interpretazione geometrica

I valori singolari rappresentano lo "stretching" che la matrice A applica alle direzioni definite dai **right singular vectors**. Geometricamente:

1. Una matrice A può essere vista come una trasformazione lineare che:
 - Ruota i vettori nello spazio \mathbb{R}^n ,
 - Li scala lungo direzioni specifiche (i valori singolari),
 - E poi li ruota di nuovo in \mathbb{R}^m .
2. I valori singolari indicano quanto ogni direzione principale (definita dai vettori singolari) viene dilatata o compressa da A .

Proprietà principali

1. **Non negativi:** I valori singolari sono sempre $\sigma_i \geq 0$.

2. **Ordinati:** I valori singolari sono in genere ordinati in ordine decrescente:

$$\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$$

3. **Rango:** Il numero dei valori singolari positivi di A è uguale al rango di A

4. **Norma:**

- La norma di A (norma 2) è data dal massimo valore singolare: $\|A\|_2 = \sigma_1$.
- La **traccia** di A è la somma dei valori singolari: $Tr(A) = \sum \sigma_i$.

Applicazioni

I valori singolari sono fondamentali in molte applicazioni, come:

1. **Compressione dei dati:** Nella riduzione dimensionale (ad esempio, PCA), solo i valori singolari più grandi e i corrispondenti vettori sono utilizzati per rappresentare i dati.
2. **Approssimazione di matrici:** La matrice di rango più basso che approssima A può essere ottenuta troncando i valori singolari più piccoli.
3. **Risoluzione di sistemi lineari:** Nella regressione lineare o nella pseudoinversa di una matrice.

In sintesi, i valori singolari rappresentano i fattori di scala fondamentali di una trasformazione lineare e forniscono informazioni importanti sulla struttura intrinseca della matrice.

Un'altra interpretazione dei componenti principali

Il first principal component loading vector ha una proprietà molto speciale: definisce la retta nello spazio p -dimensionale che è più vicina alle n osservazioni (usando come misura di vicinanza la average squared Euclidian distance, cioè distanza Euclidea quadrata mediata).

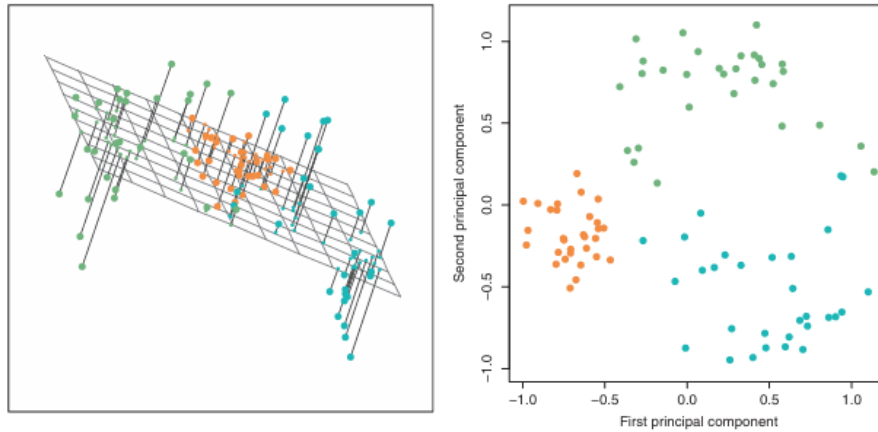
La **distanza euclidea quadrata mediata** tra due vettori \mathbf{a} e \mathbf{b} di dimensione n è definita come:

$$\frac{1}{n} \sum_{i=1}^n (a_i - b_i)^2$$

In pratica, è la media delle distanze euclidee quadrate calcolate elemento per elemento.

La nozione di principal components come dimensioni che sono più vicine alle n osservazioni si estende oltre il solo first principal component.

Per esempio, i primi due principal components di un data set generano (span) il piano che è il più vicino possibile alle n osservazioni, sempre in termini di average squared Euclidian distance.



In figura: 90 osservazioni simulate in tre dimensioni.

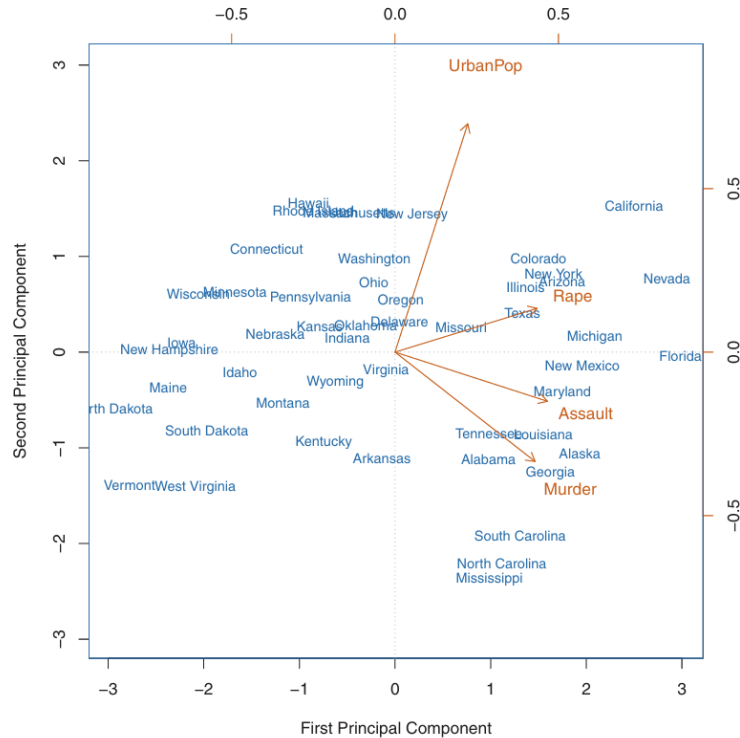
A sinistra: i primi due componenti principali generano (span) il piano che meglio fitta i dati. Minimizza la somma delle distanze al quadrato di ogni punto dal piano.

A destra: i primi due principal component score vectors danno le coordinate della proiezione delle 90 osservazioni sul piano. La varianza nel piano è massimizzata.

Esempio di PCA: USAarrests

USAarrests data:

- per ognuno dei 50 stati degli Stati Uniti il data set contiene il numero di arresti per 100000 residenti per ognuno dei tre crimini: Aggressione, Omicidio, Violenza sessuale;
- UrbanPop è la percentuale di popolazione in ogni stato che vive in aree urbane.



I Principal Component Score Vectors hanno lunghezza $n = 50$.

I Principal Component Loading Vectors hanno lunghezza $p = 4$.

PCA è stato eseguito standardizzando ogni variabile in modo che avesse media 0 e deviazione standard 1.

In figura sono plottati i primi due principal component scores e i loading vectors in un singolo biplot, i principal component loading vectors ϕ_1 e ϕ_2 sono anche presentati nella tabella seguente.

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

TABLE 10.1. The principal component loading vectors, ϕ_1 and ϕ_2 , for the **USArrests** data. These are also displayed in Figure 10.1.

Interpretare il biplot

Il primo loading vector piazza approssimativamente lo stesso peso ad Assault, Murder e Rape con invece molto meno peso a UrbanPop. Di conseguenza questo componente corrisponde più o meno ad una misura della frequenza generale di crimini gravi.

Il secondo loading vector piazza la maggior parte del suo peso su UrbanPop e molto meno peso sulle altre tre features. Di conseguenza questo componente corrisponde più o meno al livello di urbanizzazione dello stato.

In generale vediamo che le variabili legate ai crimini sono vicine tra loro mentre UrbanPop è lontano. Questo indica che le variabili relative ai crimini sono correlate tra loro, gli stati con alti rate di omicidi tendono ad avere anche alti rate di aggressioni e violenze sessuali. UrbanPop è meno correlata con le altre tre.

Possiamo esaminare differenze tra stati tramite i due principal component score vectors. Da quello che abbiamo detto prima possiamo concludere che stati con grandi scores positivi sulla prima componente, come la California, il Nevada e la Florida, hanno alto tasso di criminalità; invece stati come il Nord Dakota, con punteggi negativi sulla prima componente, hanno basso tasso di criminalità.

La California ha uno score alto anche sul secondo componente, il che indica un alto livello di urbanizzazione, mentre il contrario è vero per stati come il Mississippi.

Gli stati vicini a 0 su entrambe le componenti, come l'Indiana, hanno approssimativamente livelli medi sia di crimine che di urbanizzazione.

Scaling the Variables

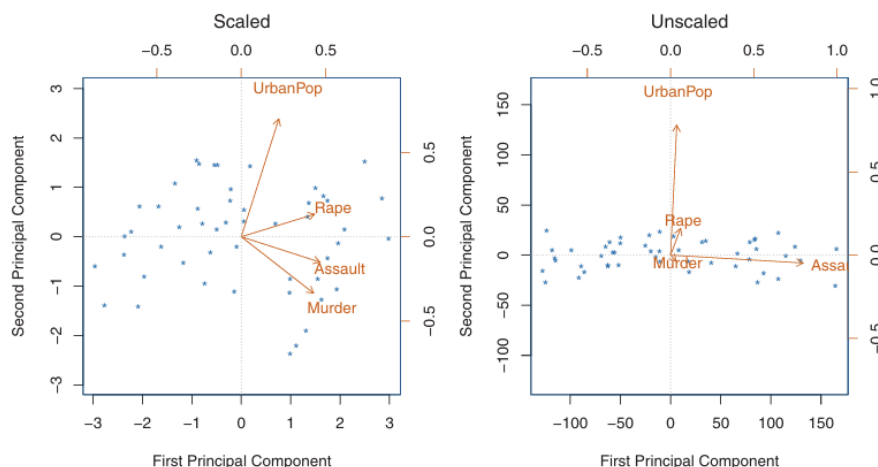
Come detto quando si effettua PCA le variabili dovrebbero essere centrate per avere media nulla. Inoltre, i risultati ottenuti quando eseguiamo PCA dipenderanno da se le variabili sono state individualmente scalate (cioè ognuna è stata moltiplicata per una diversa costante).

Questo è in contrasto con altre tecniche di supervised e unsupervised learning come linear regression, nella quale lo scaling delle variabili non ha effetto.

Ad esempio nel caso di USAarrests il biplot è stato ottenuto dopo aver scalato ogni variabili in modo da farle avere standard deviation 1. Perché è importante?

In questi dati le variabili sono misurate in unità di misura differenti, Murder, Rape ed Assault sono misurate in numero di occorrenze per 100000 persone, invece UrbanPop è la percentuale della popolazione dello stato che vive in un'area urbana.

Le 4 variabili hanno rispettivamente varianze 18.97, 87.73, 6945.16 e 209.5.



Notiamo come conseguenza delle varianze che nel biplot con le variabili non scalate, a destra, Assault ha il più grande loading sul first principal component perché ha la più grande varianza tra le 4 variabili.

Il secondo loading vector invece pone quasi tutto il suo peso su UrbanPop.

In generale è sempre raccomandato scalare le variabili in modo che abbiano standard deviation 1.

Se Assault fosse stato misurato in unità del numero di occorrenze per 100 persone (anziché 100000) allora questo sarebbe stato corrispondente a dividere gli elementi della variabile per 1000, quindi la varianza sarebbe stata piccola e il first principal component loading vector piccolo.

Solo se le variabili sono tutte nella stessa unità di misura si può scegliere di non scalarle.

Proportion of variance explained (proporzione di varianza spiegata)

Per comprendere la forza di ogni componente, siamo interessati alla proporzione di varianza spiegata (PVE) da ognuna.

La varianza totale presente in un dataset (assumendo che le variabili siano state centrate per avere media zero) è:

$$\sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$$

e la varianza spiegata dalla m-esima componente principale è

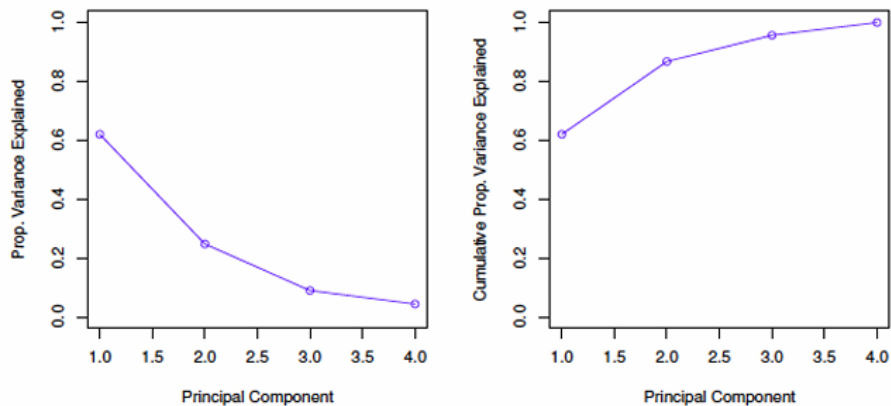
$$\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{jm} x_{ij} \right)^2$$

Si può dimostrare che $\sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2 = \sum_{m=1}^M \text{Var}(Z_m)$ con $M = \min(n-1, p)$.

Di conseguenza la PVE dell'm-esimo componente principale è dato dalla quantità positiva tra 0 e 1:

$$\frac{\sum_{i=1}^n z_{im}^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$$

Le PVE sommano ad 1. Talvolta è utile mostrare le PVE cumulative:



(1317)

Come si sceglie il numero di principal components?

Non c'è una sola (o semplice) risposta a questa domanda, visto che la cross-validation non è disponibile a tale scopo.

Lo scree plot (1317) può essere usato come guida: si cerca un punto per il quale la proporzione di varianza spiegata da ogni componente principale successivo diminuisce bruscamente.

Sfortunatamente non c'è nessun modo oggettivo per decidere quanti componenti principali siano abbastanza.

In pratica, tendiamo a guardare ai primi componenti principali per cercare di capire se ci sono pattern interessanti nei dati.

Secondo pacco di slide

Outline:

- Introduzione: comprendere cos'è PCA
- Notazione e strumenti di algebra lineare
- PCA step-by-step
- Matrice di covarianza dei dati e PCA
- PCA e Singular Value Decomposition (SVD)

Introduzione: cos'è PCA?

PCA è:

- una tecnica usata per analizzare dataset con spazi delle feature grandi
- un metodo di riduzione della dimensionalità
- una trasformazione applicata ad un data set
- una tecnica usata per rendere incorrelate delle data features

- una tecnica usata per spiegare la varianza dei dati

Ma più chiaramente PCA è una trasformazione lineare che proietta un data set su un nuovo spazio che è descritto dalle direzioni principali, che sono un insieme di direzioni ortogonali che meglio "fittano" il data set.



Le Principal Components sono le rappresentazioni dei dati originali nel nuovo spazio, la varianza originale dei dati è ridistribuita sulle componenti principali.

Le Data features sono linearmente non correlate in questo nuovo spazio.

Scegliere qualche componente principale da abbastanza informazioni sui dati originali e permette di eseguire la riduzione della dimensionalità.

PCA: la logica

Rappresentando i dati in un nuovo spazio, PCA punta a:

- preservare la maggior parte della sua varianza in una rappresentazione a bassa dimensionalità;
- dare, nel nuovo spazio, una rappresentazione a minima projection distance (minimum projection distance representation).

I componenti principali sono calcolati come combinazione lineare delle features iniziali; l'interpretazione fisica dei dati viene persa nello spazio trasformato.

Proiettare i dati sul first principal component da la più alta varianza proiettiva (una particolare view dei dati).

Proiettare i dati sul second principal component da la seconda più alta varianza e così via.

▼ Cos'è la varianza proiettiva? by ChatGPT

La **varianza proiettiva** è un concetto legato all'analisi della varianza dei dati in uno spazio multidimensionale, specialmente nel contesto della riduzione dimensionale, come il **Principal Component Analysis** (PCA).

Definizione

La varianza proiettiva rappresenta la quantità di **varianza dei dati** catturata quando i dati vengono proiettati su una determinata direzione o sottospazio. In altre parole, misura quanto i dati si "disperdono" lungo una direzione specifica (ad esempio, un vettore o una componente principale).

Formalmente, data una matrice di dati X (di dimensione $m \times n$, con m osservazioni e n variabili):

1. Si considera una direzione specifica w (un vettore unitario, $\|w\| = 1$).
2. La proiezione dei dati lungo w è data da X_w .
3. La varianza proiettiva è la varianza della proiezione:

$$\text{Varianza proiettiva lungo } w = \frac{1}{m} \|X_w\|^2$$

Contesto

La varianza proiettiva è strettamente legata al concetto di **componenti principali**:

- Nel PCA, le componenti principali sono le direzioni lungo le quali la varianza proiettiva dei dati è massimizzata.
- La varianza proiettiva massima per una matrice di dati X è associata all'autovalore massimo della matrice di covarianza di X .

Interpretazione geometrica

- Ogni proiezione di un insieme di dati su una direzione riduce il problema a una dimensione minore.
- La varianza proiettiva ci dice quanto "informazione" o "dispersione" dei dati viene mantenuta nella proiezione.

Proprietà

1. **Massimizzazione:** Nel PCA, le direzioni che massimizzano la varianza proiettiva (le componenti principali) sono ortogonali tra loro.
2. **Somma delle varianze proiettive:** La somma delle varianze proiettive lungo tutte le direzioni è pari alla traccia della matrice di covarianza (cioè alla varianza totale dei dati).
3. **Proporzione della varianza:** La varianza proiettiva lungo una direzione può essere espressa come una proporzione rispetto alla varianza totale, permettendo di quantificare l'importanza di una direzione specifica.

Applicazioni

- **Riduzione dimensionale:** Per ridurre le dimensioni dei dati, si scelgono le direzioni che catturano la maggiore varianza proiettiva (cioè le prime componenti principali).
- **Analisi dei dati:** La varianza proiettiva aiuta a capire la struttura intrinseca dei dati, evidenziando le direzioni di maggiore variabilità.

In sintesi, la **varianza proiettiva** è una misura chiave per valutare quanto i dati si disperdono lungo una direzione specifica e viene utilizzata principalmente per identificare le direzioni più significative in spazi multidimensionali.

PCA step-by-step

Derivare PCA

Considerato un data set $X' \in \mathbb{R}^{n \times p}$ le cui colonne hanno media zero

ave zero mean

$$\mu = \frac{1}{n} \sum_{i=1}^n x'_i = \left(\frac{1}{n} \sum_{i=1}^n x'_{i1}, \frac{1}{n} \sum_{i=1}^n x'_{i2}, \dots, \frac{1}{n} \sum_{i=1}^n x'_{ip} \right) \quad X = \begin{bmatrix} x'_1 - \mu \\ x'_2 - \mu \\ \vdots \\ x'_n - \mu \end{bmatrix}$$

Il primo passo per effettuare una PCA consiste nell'ottenere una trasformazione lineare $W \in \mathbb{R}^{p \times p}$.

Il ruolo di W è mappare ogni riga x di X su una riga t_i , nel nuovo spazio.

Piazzando le righe t_i in una matrice T possiamo scrivere la trasformazione lineare del PCA come:

$$T = XW$$



Le colonne di W sono le Principal Directions.



L'elemento t_{ij} è la j -esima componente principale della i -esima riga.

Derivare PCA - come costruire W

Le colonne di W sono un insieme di vettori ortonormali.

Il primo vettore colonna $w^{(1)}$ è calcolato come:

$$w^{(1)} = \arg \max_{\omega \in \mathbb{R}^p} \left\{ \frac{1}{n-1} \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} \omega_j \right)^2 \right\} \quad \text{subject to } \|\omega\| = 1$$

Unknown column vector

La condizione imposta è che w deve avere norma unitaria.

La sommatoria

$$\left(\sum_{j=1}^p x_{ij} \omega_j \right)^2$$

è il **Principal Component Score**, è il prodotto scalare tra x_i e w , cioè la componente di x_i sulla direzione definita dal vettore w .

L'intero argomento dell'argmax


$$\frac{1}{n-1} \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} \omega_j \right)^2$$

è la varianza del first principal component rispetto alla variabilità tra le righe (ricordiamo che le righe hanno media 0).

W in forma matriciale

$w^{(1)}$ può essere scritto in forma matriciale.

$$w^{(1)} = \arg \max_{\omega \in \mathbb{R}^p} \left\{ \frac{1}{n-1} \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} \omega_j \right)^2 \right\} \quad \text{subject to } \|\omega\| = 1$$

 Matrix form

$$w^{(1)} = \arg \max_{\omega \in \mathbb{R}^p} \|X\omega\|^2 = \arg \max_{\omega \in \mathbb{R}^p} \omega^\top X^\top X \omega \quad \text{subject to } \|\omega\| = 1$$

Considerando la limitazione della norma unitaria il problema diventa equivalente a


$$w^{(1)} = \underset{w \in \mathbb{R}^p}{\operatorname{argmax}} \frac{w^\top X^\top X w}{w^\top w}$$

La quantità $\frac{w^\top X^\top X w}{w^\top w}$ è chiamata **quoziente di Rayleigh**.

Per le matrici semidefinite positive, come $X^\top X$, il quoziente di Rayleigh è massimizzato quando w è l'autovettore associato al massimo autovalore di $X^\top X$.


Per calcolare le colonne rimanenti di W dobbiamo imporre il vincolo di ortogonalità (e ripetere il processo di massimizzazione)

$$X_k = X - X \sum_{i=1}^{k-1} \omega^{(i)} (\omega^{(i)})^\top$$

 Orthogonal projection matrix

In verde:

$$X_k = X - X \sum_{i=1}^{k-1} \omega^{(i)} (\omega^{(i)})^\top$$



La proiezione dei dati sullo spazio dei vettori ortogonali all'insieme delle $k - 1$ direzioni principali trovate.

Dopo aver imposto il vincolo di ortogonalità si ripete il processo di massimizzazione

$$w^{(k)} = \arg \max_{\omega \in \mathbb{R}^p} \frac{\omega^\top X_k^\top X_k \omega}{\omega^\top \omega}$$

Si può dimostrare che i vettori $w^{(k)}$ sono i restanti autovettori di $X^T X$, ordinati seguendo i valori dei corrispondenti autovalori disposti in ordine decrescente.

Matrice di covarianza dei dati e PCA

La matrice di covarianza

La matrice di trasformazione W ha come colonne gli autovettori della matrice $X^T X$.

La matrice di covarianza del data set centrato (penso intenda centrato per avere media 0) X è la matrice $p \times p$

$$C = \frac{1}{n-1} \sum_{i=1}^n x_i^\top x_i = \frac{1}{n-1} X^\top X$$

PCA è basato sulla decomposizione spettrale (eigendecomposition) della matrice di covarianza del dataset.

Gli autovettori (eigenvectors) della matrice di covarianza sono le direzioni principali.

Gli autovalori (eigenvalues) della matrice di covarianza rappresentano la varianza dei dati proiettati sulle direzioni principali.

Consideriamo la prima direzione principale $w^{(1)}$, dall'equazione degli autovalori possiamo scrivere

can write

$$\begin{aligned} X^\top X w^{(1)} &= \lambda_1 w^{(1)} \Rightarrow (w^{(1)})^\top X^\top X w^{(1)} = \lambda_1 \|w^{(1)}\|^2 \Rightarrow \overbrace{(w^{(1)})^\top X^\top X w^{(1)}}^{\|X w^{(1)}\|^2} = \lambda_1 \\ \text{Var}[X w^{(1)}] &= \frac{1}{n-1} \|X w^{(1)}\|^2 \Rightarrow \text{Var}[X w^{(1)}] = \frac{\lambda_1}{n-1} \end{aligned}$$

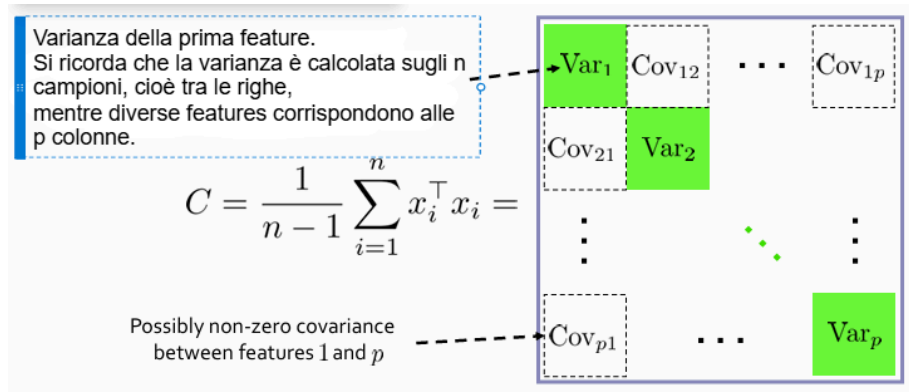
Per definizione essendo $w^{(1)}$ un autovalore di $X^T X$ allora $X^T X w^{(1)} = \lambda_1 w^{(1)}$, ovviamente se λ_1 è autovalore.

Concludiamo che la varianza (tra le righe) del first principal component è proporzionale all'autovalore λ_1 .

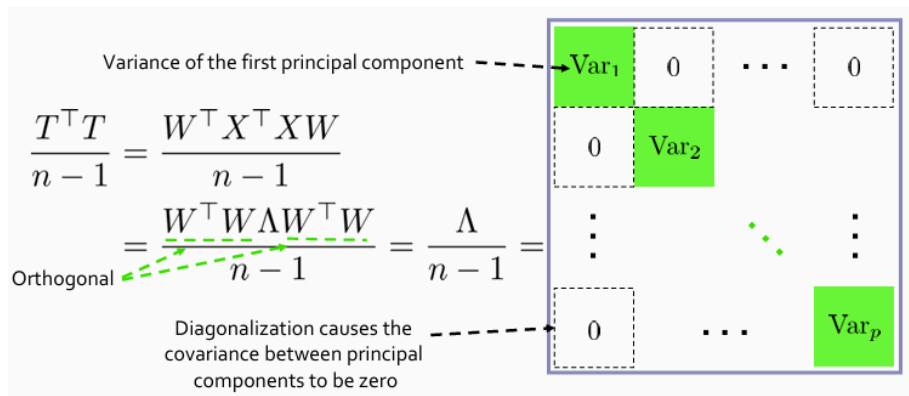
Similmente, la stessa cosa vale per le altre direzioni principali e autovalori.

PCA e correlazione tra le features

La matrice di covarianza del data set contiene informazioni riguardanti la covarianza tra features.



La **matrice di covarianza dei componenti principali** è tale che la covarianza tra componenti principali è 0.



Riduzione della dimensionalità

Un modo per ridurre la dimensionalità è scegliere solo un sottoinsieme degli m componenti principali, m può essere scelto assicurando un dato livello di PVE (ad esempio 95% della varianza totale del data set).

Si tratta di un metodo empirico dipendente dallo specifico caso d'uso.

$$PVE = \frac{\sum_{k=1}^m \left[\frac{1}{n-1} \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} w_{jk} \right)^2 \right]}{\sum_{j=1}^p \frac{1}{n-1} \sum_{i=1}^n x_{ij}^2}$$

Il denominatore è la varianza totale del dataset.

L'argomento della sommatoria a numeratore è la varianza del k -esimo componente principale.

Ruolo degli autovalori nella riduzione della dimensionalità

Il ruolo degli autovalori nella riduzione della dimensionalità può essere reso esplicito

$$PVE = \frac{\sum_{k=1}^m \left[\frac{1}{n-1} \sum_{i=1}^n \left(\sum_{j=1}^p x_{ij} w_{jk} \right)^2 \right]}{\sum_{j=1}^p \frac{1}{n-1} \sum_{i=1}^n x_{ij}^2} = \frac{\sum_{k=1}^m \lambda_k}{\sum_{j=1}^p \lambda_j}$$

Il numero di autovalori non-zero è pari a $r = \text{rank}(X^T X) = \text{rank}(X)$, si nota che $r \leq \min\{n, p\}$.

Possiamo scegliere $m \leq r$ componenti principali non zero.

PCA e Singular Value Decomposition

Le basi della Singular Value Decomposition

SVD è una fattorizzazione matriciale che generalizza la decomposizione spettrale, infatti la eigendecomposition può essere applicata solo alle matrici quadrate, mentre SVD può essere applicata anche a matrici rettangolari.

L'SVD di una matrice reale $M \in \mathbb{R}^{n \times p}$ è definita come $M = U \Sigma V^T$

$$M \in \mathbb{R}^{n \times p} = U \in \mathbb{R}^{n \times n} \Sigma \in \mathbb{R}^{n \times p} V^T \in \mathbb{R}^{p \times p}$$

Le colonne di U sono i Left singular vectors.

Le righe di V^T sono i Right singular vectors.

Sia U che V^T sono matrici ortogonali quindi la loro trasposta è uguale alla loro inversa.

$$M \in \mathbb{R}^{n \times p} = U \in \mathbb{R}^{n \times n} \Sigma \in \mathbb{R}^{n \times p} V^T \in \mathbb{R}^{p \times p}$$

Singular values

Il numero di Singular values è pari al rango di M .

$$M \in \mathbb{R}^{n \times p} = U \in \mathbb{R}^{n \times n} \begin{matrix} \sigma_1 \\ \vdots \\ \sigma_r \\ \text{All zeros!} \end{matrix} \Sigma \in \mathbb{R}^{n \times p} V^T \in \mathbb{R}^{p \times p}$$

Può essere utile considerare la SVD con solo i valori singolari non-zero, rimuovendo le corrispondenti colonne di U e righe di V^T .

$$M \in \mathbb{R}^{n \times p} = U \in \mathbb{R}^{n \times r} \begin{matrix} \sigma_1 \\ \vdots \\ \sigma_r \end{matrix} \Sigma \in \mathbb{R}^{r \times r} V^T \in \mathbb{R}^{r \times p}$$

SVD e PCA

Consideriamo l'SVD del data set $X = U\Sigma V^T$.

Possiamo scrivere la matrice $X^T X$ come

$$X^T X = V \Sigma^T U^T U \Sigma V^T = V \Sigma^T \Sigma V^T$$

$X^T X$ and $\Sigma^T \Sigma$ are similar, therefore they have the same eigenvalues

Squared $\rightarrow \sigma_i^2 = \lambda_i$

The singular values of X are the square roots of the eigenvalues of $X^T X$

I valori singolari di X sono le radici quadrate degli autovalori di $X^T X$.

I Right singular vectors in V sono le direzioni principali (gli autovettori di $X^T X$)

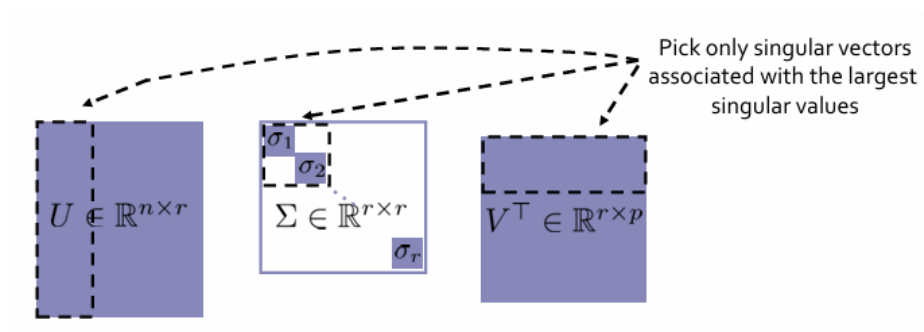
$$T = XV \rightarrow T = U\Sigma V^T V \rightarrow T = U\Sigma$$

Principal components can be obtained from both singular vectors

SVD e riduzione della dimensionalità

Effettuare l'SVD ridotto scegliendo solo un sottoinsieme di m right singular vectors da V , e i singular values associati ad essi, permette di ridurre la dimensionalità.

Ricordiamo che $\sigma_i^2 = \lambda_i$, e che λ_i è proporzionale alla varianza.



Riassunto

PCA rappresenta dati in uno spazio trasformato descritto dalle direzioni principali; le direzioni principali sono gli autovettori della matrice di covarianza dei dati.

- The **principal components** are a linear combination of the original data features
 - The variance of the original data is redistributed over the principal components
 - The eigenvalues of the data covariance matrix are proportional to the variance of the principal components
 - Principal components are uncorrelated
- PCA can be performed through eigendecomposition or, equivalently, through singular value decomposition
- By retaining only a subset of the principal components, based on a variance preservation criterion, one can perform dimensionality reduction