Analisi in componenti principali: introduzione

Giovanna Jona Lasinio

Dipartimento di Scienze Statistiche
Università di Roma "La Sapienza"
email: giovanna.jonalasinio@uniroma1.it
materiale didattico: **(STATECO)**
http://elearning2.uniroma1.it/course/index.php?categoryid=641

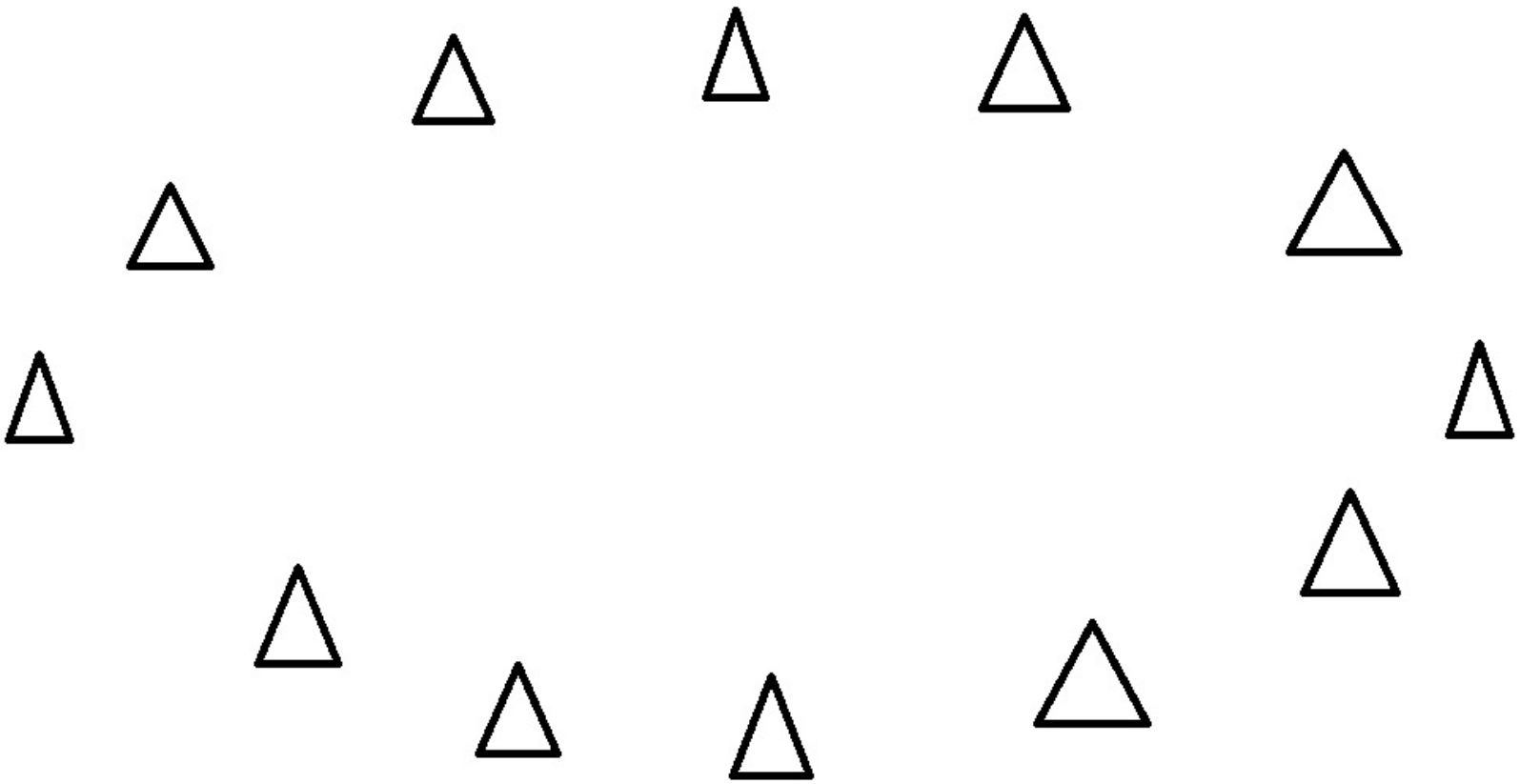
Ecobiologia e Scienze Naturali

Quanto segue è ripreso dal sito di George Dallas[1]

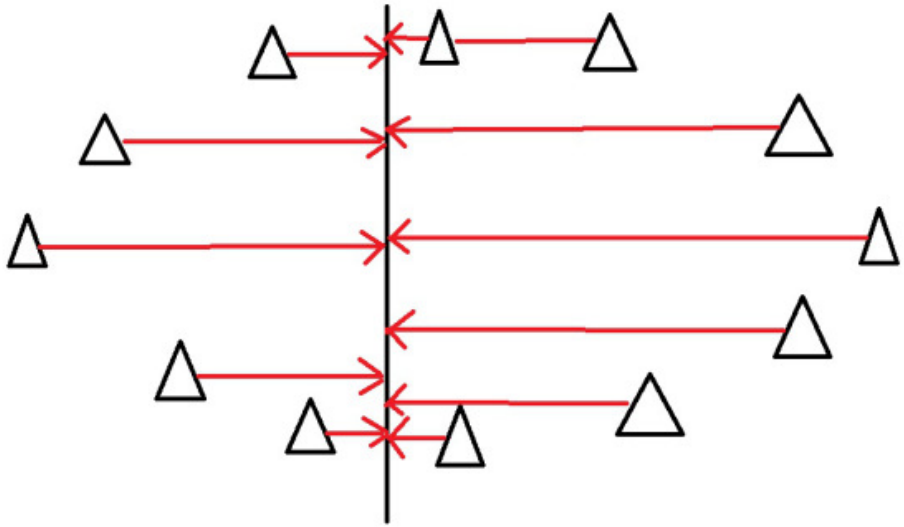
- **Cosa è l'Analisi in Componenti Principali?:** L'analisi in sé fa esattamente quel che il suo nome dice, trova le componenti principali, i pattern più rilevanti, che caratterizzano i nostri dati.
- le componenti principali (CP) sono la struttura insita nei dati, sono le direzioni di massima varianza, quelle in cui i dati sono più dispersi.

[1] (https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummie-eigenvectors-eigenva

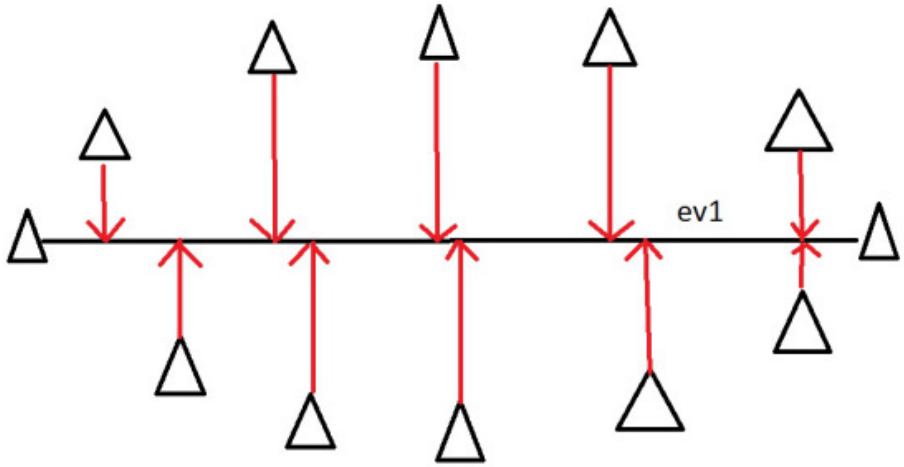
Vediamo meglio con un esempio, nella figura riportiamo dei triangoli che descrivono un ovale:



Asse verticale, poca variazione su questo, distinguiamo male i triangoli



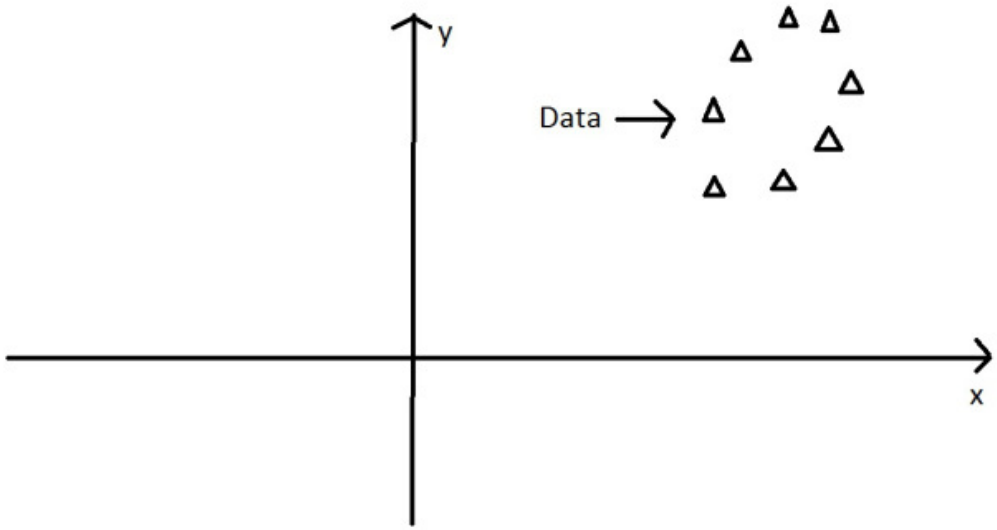
Asse orizzontale, molta più variazione su questo, riusciamo a vedere un pattern, in realtà non esiste un altro asse sul quale sia visibile maggiore variabilità, questo asse è il primo *autovettore* dei dati rappresentati con i triangoli



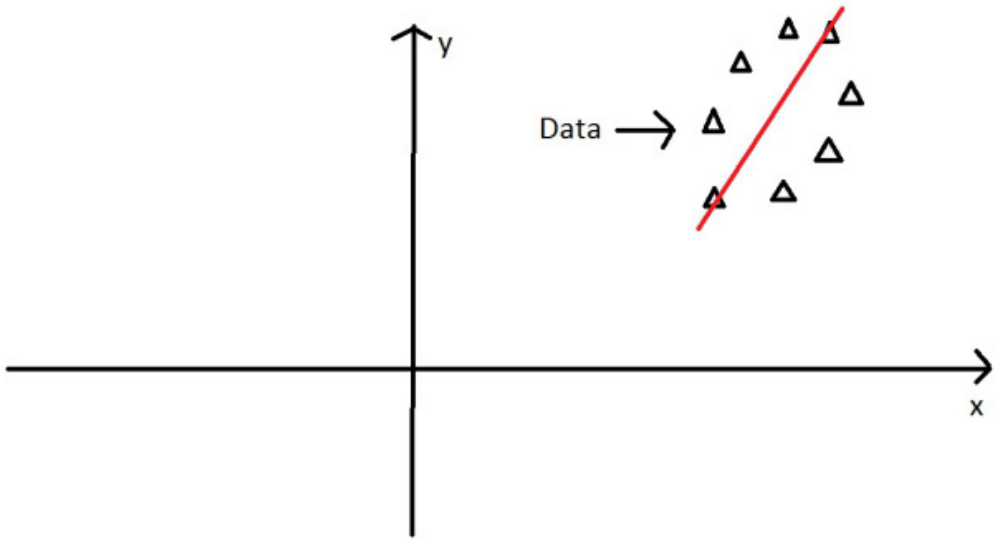
- Ogni insieme di dati può essere decomposto in *autovettori* ed *autovalori*.
- Autovettore e autovalori esistono in coppie: ogni autovettore ha un autovalore associato, il primo descrive una direzione (nell'esempio quella orizzontale), il secondo è uno scalare e misura quanta variabilità è presente in quella direzione.
- Nell'esempio dei triangoli l'autovalore ci dice quanto i dati siano dispersi sulla linea.
- L'autovettore con l'autovalore più grande è quindi la *prima componente principale*.

- Esistono infinite direzioni in cui puntare una retta nel piano dei triangoli, ma esistono pochissime direzioni associate agli autovettori (e quindi con proprietà ottimali)
- Il numero massimo di autovettori (e autovalori) è definito dalle dimensioni dei dati.
- Le direzioni definite dagli autovettori sono tra loro ortogonali (autovettori incorrelati). Gli autovalori associati sono > 0.
- Se abbiamo una tabella di dati $\mathbf{X}$ con n righe e p colonne, con $n > p$ potremo avere al massimo p autovettori ortogonali e autovalori positivi.

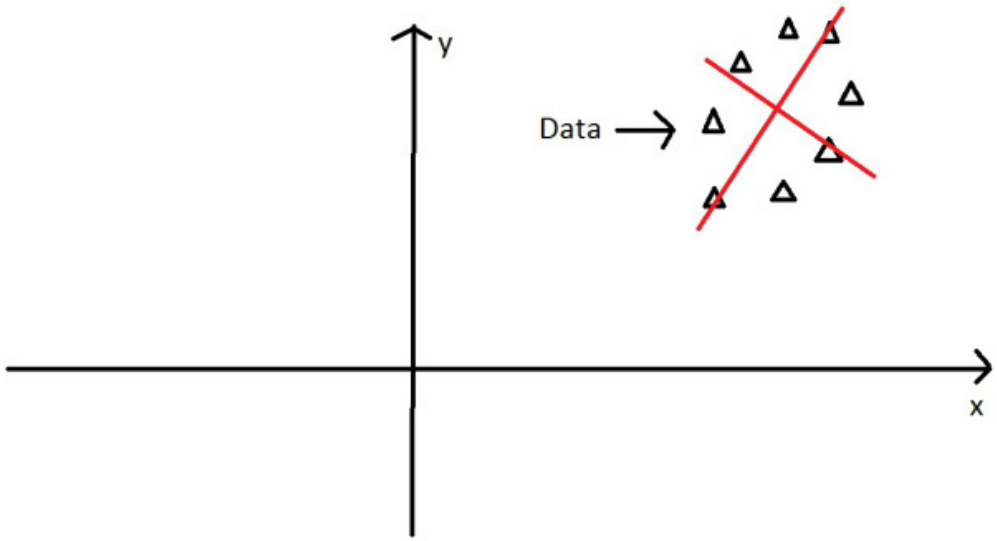
Supponiamo che i triangoli siano descritti da 3 variabili: età, ore trascorse su internet e tempo di uso del cellulare. Le prime due grandezze sono piuttosto variabili, mentre per la terza non si rilevano differenze rilevanti, ovvero il tempo di uso del cellulare è quasi costante.
Osserviamo i dati nel sistema di assi definito dalle prime due variabili



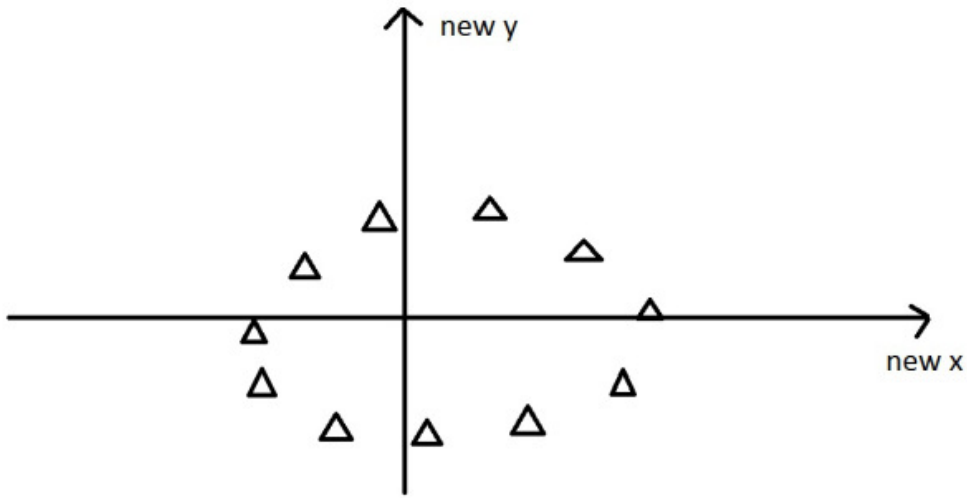
Gli autovettori ci forniscono un sistema di assi più interessante



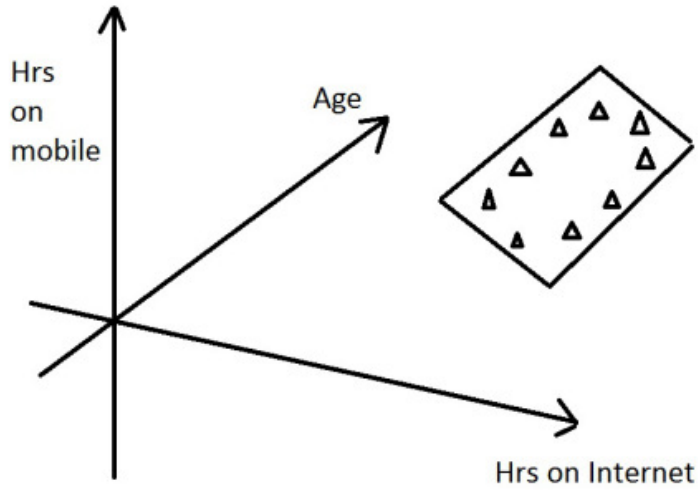
Gli autovettori ci forniscono un sistema di assi più interessante,



in particolare i primi due. Il centro del nuovo sistema di assi è il baricentro dei dati. Va notato che non abbiamo fatto nulla ai dati in questo caso, abbiamo solo cambiato sistema di riferimento. C'è ora da chiedersi cosa rappresentino ora questi due nuovi assi rispetto ai precedenti. Questa è la parte più complessa dell'analisi in componenti principali, ovvero la *caratterizzazione degli assi*. Vedremo in seguito con degli esempi.

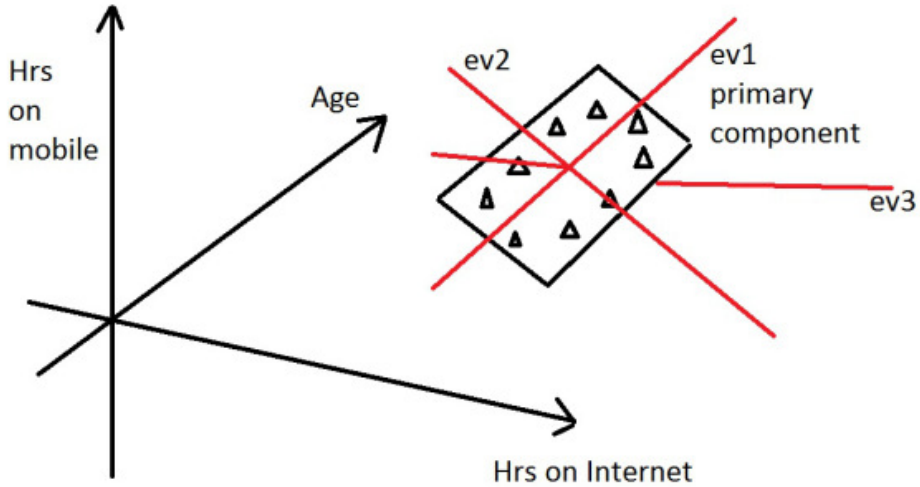


La PCA serve anche a rappresentare i dati usando meno dimensioni, o meglio usando solo le dimensioni davvero rilevanti. Nell'esempio dei triangoli, usando tutte e tre le variabili rilevati abbiamo:

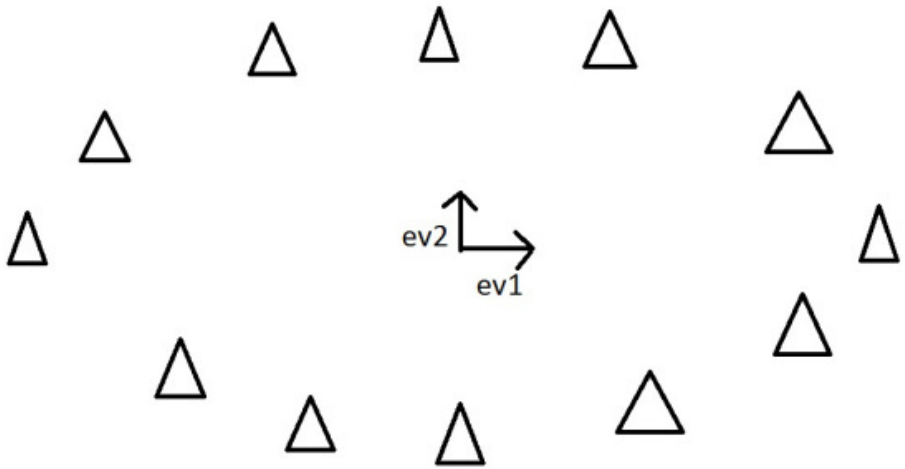


Dato che il tempo passato usando il cellulare è quasi uguale per tutti, non c'è variazione apprezzabile lungo quell'asse.

Quindi nel sistema degli autovettori avremo due autovalori positivi e un terzo molto piccolo, quasi nullo. Ovvero il terzo autovettore non contiene informazioni utili. Posso rappresentare tutta l'informazione usando solo due dimensioni



Posso rappresentare tutta l'informazione usando solo due dimensioni



Introduciamo un concetto utile

Definizione

Siano $X_1, \ldots, X_p$ p variabili distinte e siano $a_1, \ldots, a_p$ delle costanti non nulle. Chiameremo **combinazione lineare** *delle variabili X la variabile Y ottenuta come:*

$$Y = a_1 \cdot X_1 + \ldots + a_p \cdot X_p = \sum_{i=1}^{p} a_i X_i$$

Le combinazioni lineari hanno alcune proprietà utili:

Definizione

- **Media** *Sia $E(X_i)$ il valor medio della variabile X_i, allora*

$$E(Y) = \sum_{i=1}^{p} a_i E(X_i)$$

- **Varianza***Sia $Var(X_i)$ la varianza della variabili X_i e sia $Cor(X_i, X_j) = 0$ per ogni coppia i, j, allora*

$$Var(Y) = \sum_{i=1}^{p} a_i^2 Var(X_i)$$

Ora sia $\mathbf{X}$ una matrice di dati con n righe e p colonne, per misurare l'informazione in essa contenuta usiamo una *distanza* o tra le righe della matrice o tra le colonne. In realtà la distanza tra le righe e quella tra le colonne sono tra loro legate secondo uno schema algebrico (*schema duale*) ed hanno gli stessi autovalori ed autovettori.
Possiamo definire molti tipi di distanza, ad esempio se standardizziamo le colonne di $\mathbf{X}$ la distanza tra le colonne è la matrice di correlazione. Se invece semplicemente centriamo le colonne rispetto alla loro media, la distanza tra di esse è data dalla matrice di covarianza.

Definizione

Quando scegliamo una distanza diremo che stiamo fissando una metrica nello spazio dei nostri dati.

Indichiamo con $\mathbf{R}$ la matrice di distanza ad esempio la matrice di correlazione, allora potremo scrivere $\mathbf{R} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ con $\mathbf{Q}$ matrice degli autovettori e $\mathbf{\Lambda}$ matrice degli autovalori (una matrice che ha gli autovalori sulla diagonale principale e tutti zeri fuori da questa).

Definizione

Le componenti principali sono ottenute come combinazioni lineari delle variabili originali con coefficienti gli elementi degli autovettori:

$$C_i = q_{1i}X_1 + \ldots + q_{pi}X_p$$

Definizione

Gli autovalori misurano quanta variabilità è spiegata da ciascuna componente principale, quindi se λ_i è l'autovalore associato alla componente i-esima $\frac{\lambda_i}{\sum_i \lambda_i} * 100$ è la percentuale di variabilità spiegata dalla componente i

Inoltre per $k \leq p$

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} * 100$$

misura la percentuale di variabilità spiegata dalle componenti da 1 a k.

Possiamo applicare la PCA a qualsiasi matrice di distanze

Ora apriamo R e vediamo degli esempi