Analisi in componenti principali

Principal Component Analysis PCA

IL PRESENTE MATERIALE È RISERVATO AL PERSONALE DELL'UNIVERSITÀ DI BOLOGNA E NON PUÒ ESSERE UTILIZZATO AI TERMINI DI LEGGE DA ALTRE PERSONE O PER FINI NON ISTITUZIONALI

PCA: utilizzo dell'analisi spettrale per analizzare la forma e la dimensionalita' di un insieme di dati sparsi

IL PRESENTE MATERIALE È RISERVATO AL PERSONALE DELL'UNIVERSITÀ DI BOLOGNA E NON PUÒ ESSERE UTILIZZATO AI TERMINI DI LEGGE DA ALTRE PERSONE O PER FINI NON ISTITUZIONALI

Analisi delle Componenti Principali

- PCA è un trattamento matematico dei dati il cui scopo è rappresentare la variazione presente nelle tante variabili utilizzando un numero molto più piccolo di “componenti principali”
- Si costruisce un nuovo spazio su cui rappresentare i campioni ridefinendo gli assi utilizzando le componenti principali (PC) al posto delle variabili originali.
- L’uso di questi nuovi assi – le PC – permette di rappresentare la vera natura multivariata dei dati in un numero relativamente piccolo di dimensioni e di usare questa rappresentazione per identificare la struttura dei dati stessi

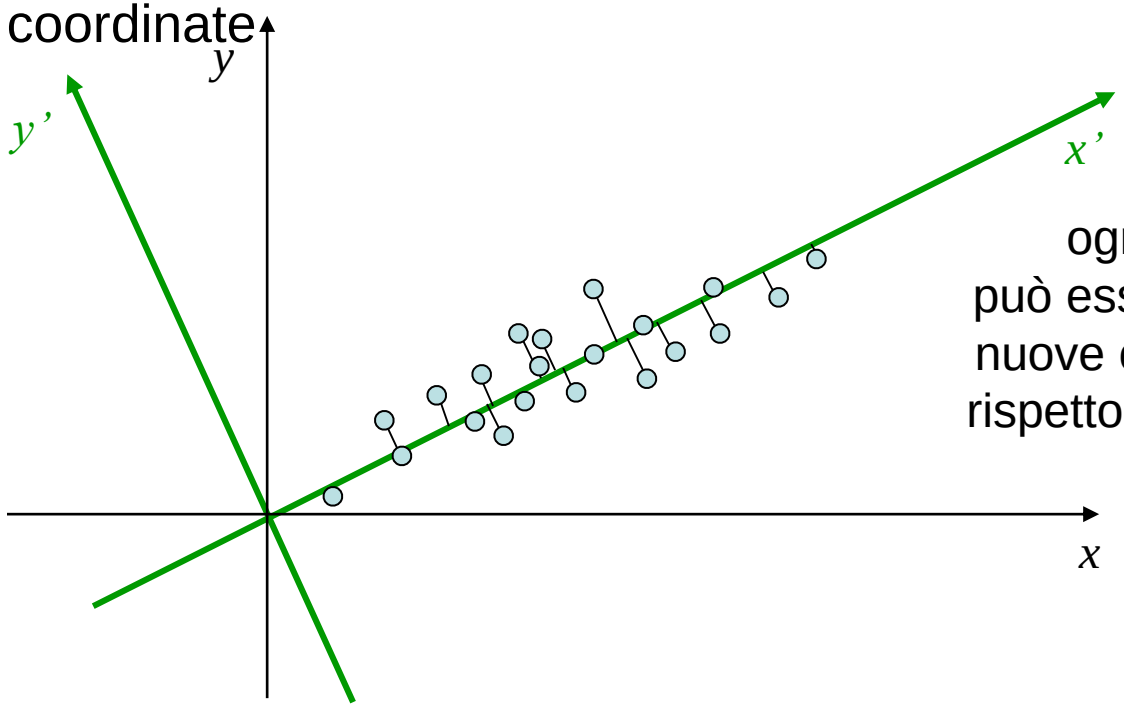
PCA e riduzione di dimensionalità

- La PCA è particolarmente utile quando la dimensionalità dello spazio delle misure è particolarmente elevata ma i campioni si trovano in un sottospazio di dimensioni significativamente ridotte
- In molti fenomeni la **dimensionalità intrinseca** del problema è infatti significativamente più piccola del numero di variabili misurate
- La dimensionalità intrinseca del problema è il numero di PC necessarie per spiegare la variabilità
- Uno degli obiettivi della PCA quindi è quello di determinare il numero di componenti principali significative

PCA – idea generale

- Immaginiamo di aver misurato il valore di due variabili (x,y) su 21 campioni
- la PCA costituisce una rappresentazione alternativa dei dati in un altro sistema di coordinate

Dati 2D
->2 PC



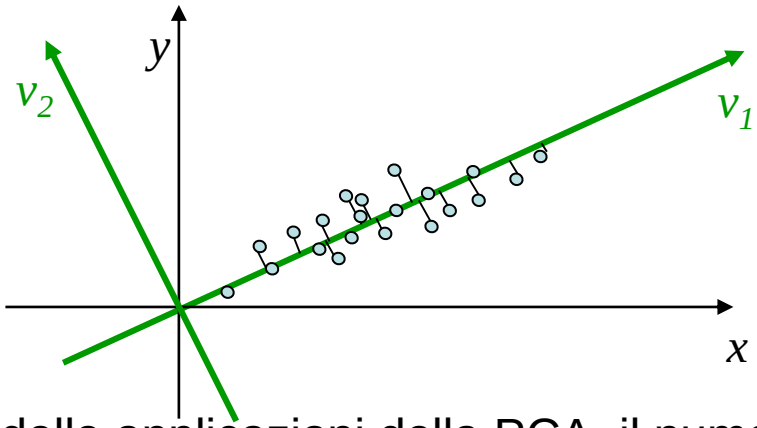
ogni campione
può essere descritto da
nuove coordinate (x',y')
rispetto allo spazio delle
PC

- PCA determina una base ortogonale che meglio rappresenta il data set
- La prima componente principale x' è la direzione lungo cui si registra la massima dispersione dei dati (calcolata attraverso la varianza)

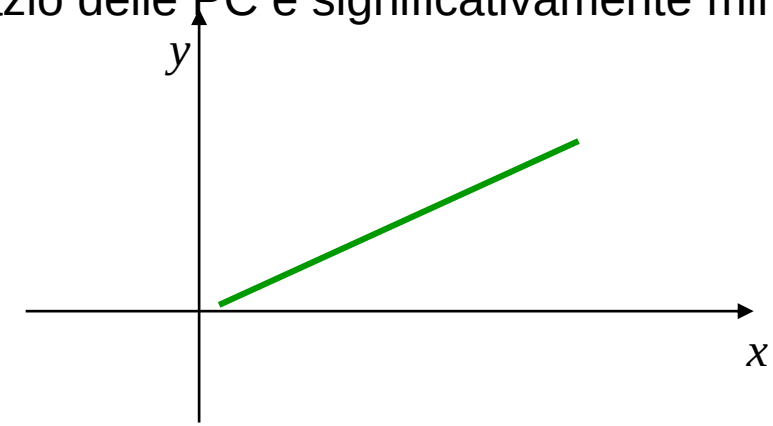
PCA: riduzione dimensionalità

Dati 2D

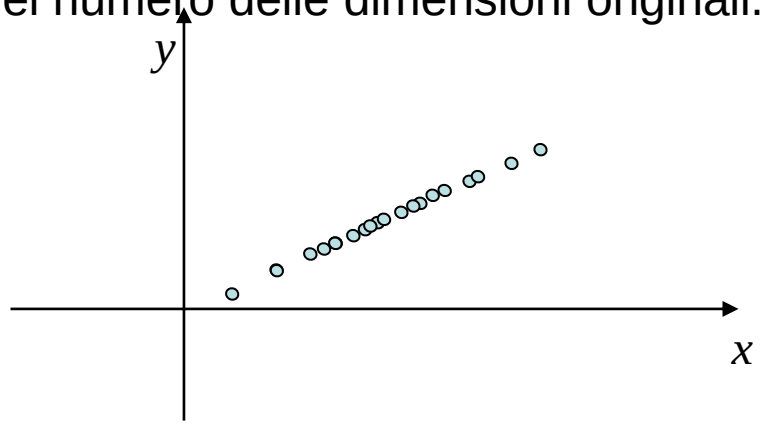
->1 PC



Nella maggior parte delle applicazioni della PCA, il numero di dimensioni dello spazio delle PC è significativamente minore del numero delle dimensioni originali.



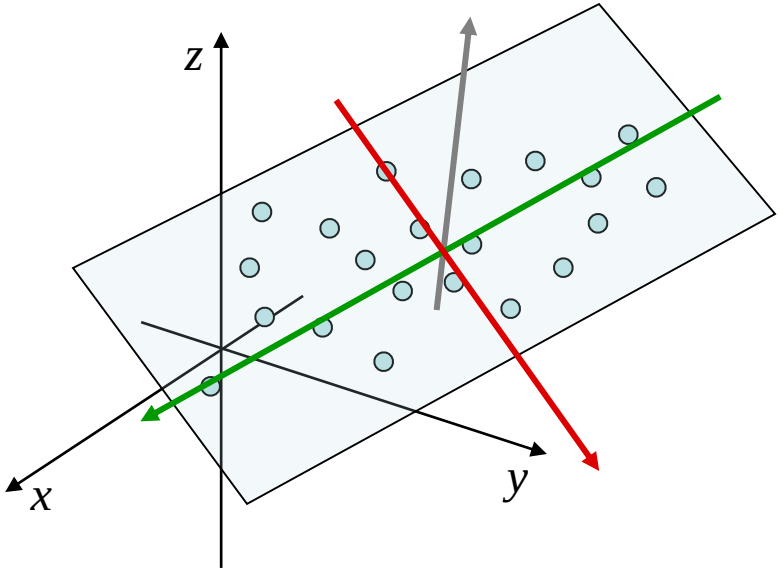
Questo segmento di retta approssima I dati originali



Il data set proiettato 1D approssima il data set originale 2D

PCA – idea generale

- PCA determina una base ortogonale che meglio rappresenta il data set.



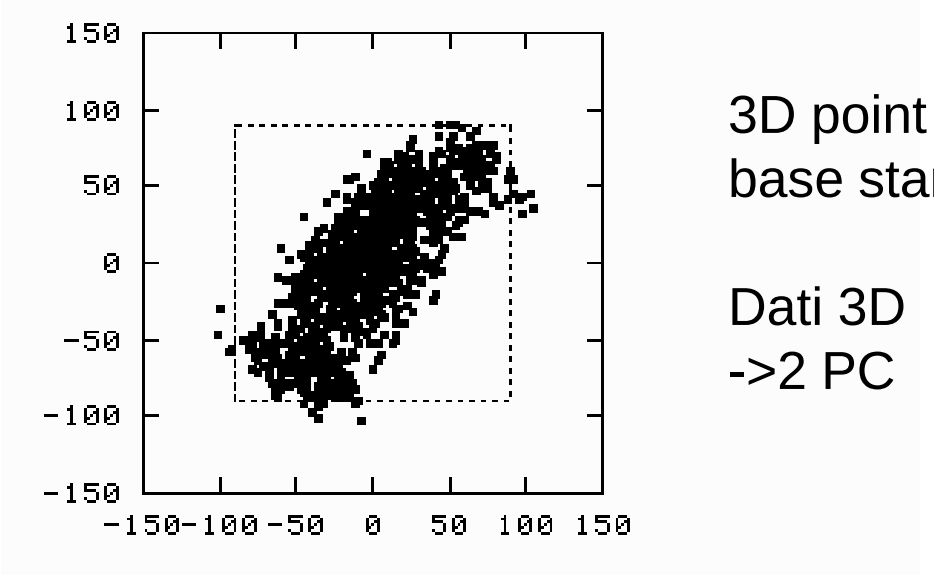
3D point set nella base standard

Dati 3D
->2 PC

- PCA trova il miglior piano approssimante (in termini di $\Sigma distances^2$)

PCA – idea generale

- PCA determina una base ortogonale che meglio rappresenta il data set.



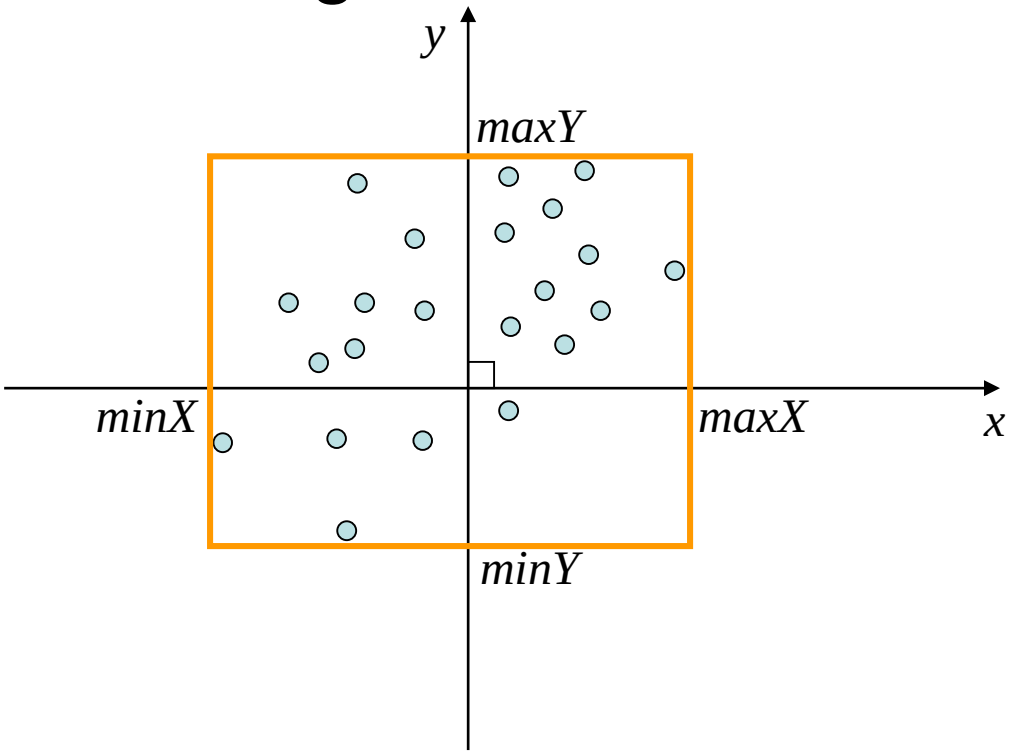
3D point set nella base standard

Dati 3D
->2 PC

- PCA trova il miglior piano approssimante (in termini di $\Sigma distances^2$)

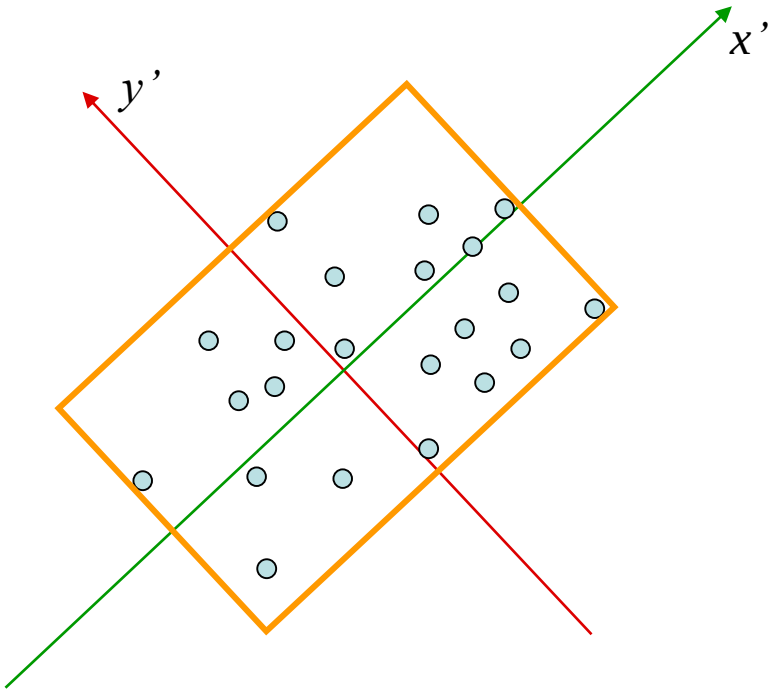
Applicazione: determinare una tight bounding box

- Una axis-aligned bounding box: allineamento agli assi



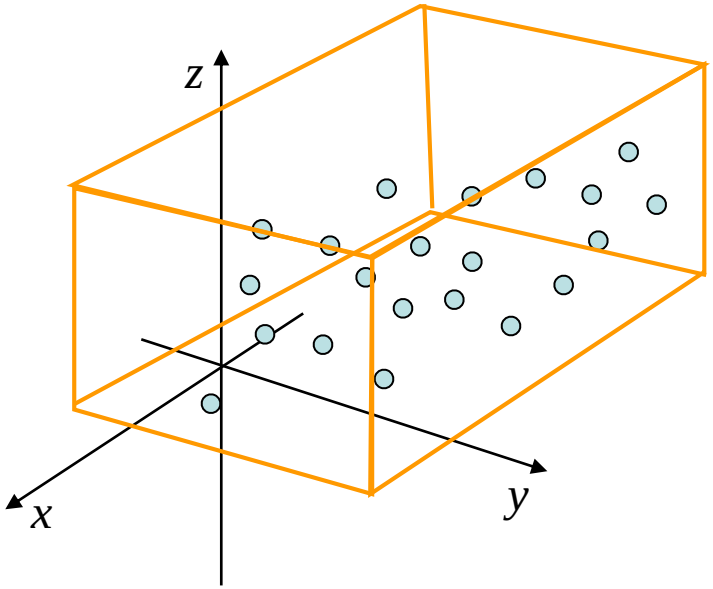
Applicazione: determinare una tight bounding box

- Oriented bounding box: trovare gli assi migliori!



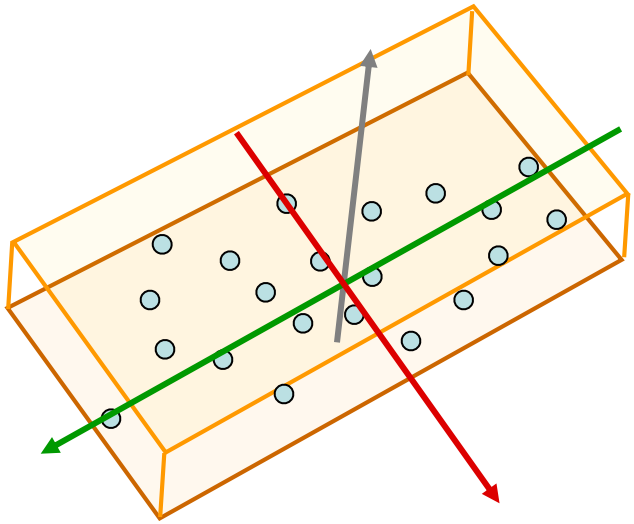
Applicazione: determinare una tight bounding box

- Questa non è una bounding box ottimale!



Applicazione: determinare una tight bounding box

- Oriented bounding box: trovare gli assi migliori!



Calcolo PCA

- Denotiamo i punti del data set da

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in R^d$$

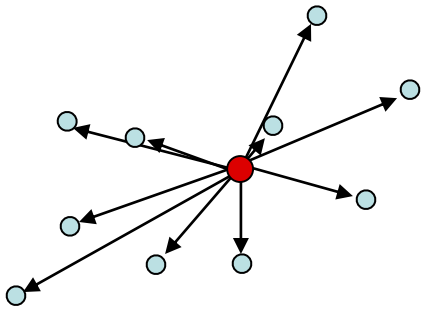
- Ciascuno definito da d elementi (caratteri)

$$\boldsymbol{x}_1 = \begin{pmatrix} x_1^1 \\ x_1^2 \\ \vdots \\ x_1^d \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} x_2^1 \\ x_2^2 \\ \vdots \\ x_2^d \end{pmatrix}, \quad \ldots, \quad \boldsymbol{x}_n = \begin{pmatrix} x_n^1 \\ x_n^2 \\ \vdots \\ x_n^d \end{pmatrix}$$

Calcolo **origine** dei nuovi assi

- L'origine è l'approssimazione di ordine zero del data set (un punto)
- Sarà il centro di massa:

$$\boldsymbol{m} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$$



- Si può provare che:

$$\boldsymbol{m} = \underset{\boldsymbol{x}}{\operatorname{argmin}}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - \boldsymbol{x}\right\|^2$$

Calcolo matrice delle covarianze

- Centriamo i dati:

  denotiamo $\boldsymbol{y}_i = \boldsymbol{x}_i - \boldsymbol{m}, \qquad i = 1, 2, \ldots, n$

- Costruiamo la matrice delle covarianze

$$X = \begin{bmatrix} \vdots & \vdots & \vdots & \vdots \\ y_1 & y_2 & \vdots & y_n \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \quad \text{matrice} \quad \text{dxn}$$

$$C = XX^T = \sum_{k=1}^{n} \boldsymbol{y}_k \boldsymbol{y}_k^{\mathrm{T}}$$

Matrice delle covarianze

- In statistica multivariata, la **matrice delle covarianze** rappresenta la variazione di ogni variabile rispetto alle altre (inclusa se stessa). È una matrice **simmetrica** di elementi:

$$c_{ij} = \mathrm{cov}(x^i, x^j) = \sum_{k=1}^{n} (\boldsymbol{x}_k^i - \boldsymbol{m}_i)(\boldsymbol{x}_k^j - \boldsymbol{m}_j) = \sum_{k=1}^{n} \boldsymbol{y}_k^i \boldsymbol{y}_k^j$$

- Ogni elemento sulla diagonale c_{ii} è la **varianza** del carattere $\mathbf{x}^i$. Ogni elemento c_{ij} (con $i \neq j$) è la **covarianza** tra i caratteri $\mathbf{x}^i$ e $\mathbf{x}^j$.
- Nel caso in cui questo valore sia positivo, significa che al crescere di una carattere, cresce anche l'altro. Nel caso in cui questo valore sia negativo, accade il contrario.
- Se i caratteri i e j sono statisticamente indipendenti, allora $c_{ij} = 0$.

Matrice delle covarianze

- $\boldsymbol{C} = XX^T$, X is $d \times n$ matrix

$$d \times d$$

$$\sum_{k=1}^{n} \boldsymbol{y}_k \boldsymbol{y}_k^{\mathrm{T}} = \sum_{k=1}^{n} \begin{pmatrix} y_k^1 \\ y_k^2 \\ \vdots \\ y_k^d \end{pmatrix} \begin{pmatrix} y_k^1 & y_k^2 & \cdots & y_k^d \end{pmatrix} = \sum_{k=1}^{n} \begin{pmatrix} y_k^1 y_k^1 & y_k^1 y_k^2 & & y_k^1 y_k^d \\ y_k^2 y_k^1 & y_k^2 y_k^2 & & y_k^2 y_k^d \\ \vdots & \vdots & \vdots & \vdots \\ y_k^d y_k^1 & y_k^d y_k^2 & & y_k^d y_k^d \end{pmatrix}$$

$$= \begin{pmatrix} y_1^1 & y_2^1 & \cdots & y_n^1 \\ y_1^2 & y_2^2 & & y_n^2 \\ \vdots & \vdots & & \vdots \\ y_1^d & y_2^d & \cdots & y_n^d \end{pmatrix} \begin{pmatrix} y_1^1 & y_1^2 & & y_1^d \\ y_2^1 & y_2^2 & & y_2^d \\ \vdots & \vdots & & \vdots \\ y_n^1 & y_n^2 & & y_n^d \end{pmatrix} = XX^T$$

Esempio matrice C

In un'indagine su n=10 soggetti si riporta per ciascuno:
- voto medio (da 0 a 33),
- intelligenza (da 0 a 10),
- provenienza
- media ore studiate in un giorno (che varia da 1 a 3)

matrice di correlazione		Zscore (VotoMedio)	Zscore (Intelligenza)	Zscore (Provenienza)	Zscore (OreMed Studio)
Correlation	Zscore(VotoMedio)	1,000	,600	-,838	,788
	Zscore(Intelligenza)	,600	1,000	-,222	,022
	Zscore(Provenienza)	-,838	-,222	1,000	-,918
	Zscore(OreMedStudio)	,788	,022	-,918	1,000

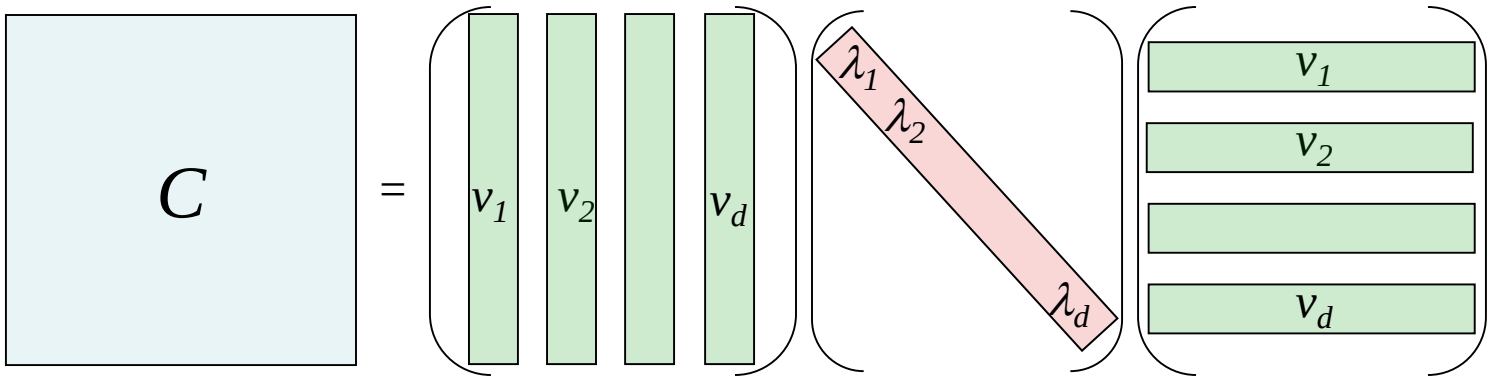
Vediamo come ci sia un forte legame tra voto, media ore studio e intelligenza.

Decomposizione spettrale di C

$\boldsymbol{C}$ è matrice simmetrica, reale, semi-definita positiva,

$\Rightarrow$ $\boldsymbol{C}$ è diagonalizzabile e ha decomposizione:

$$\boldsymbol{C} = V \Lambda V^{\mathrm{T}}$$



Gli autovettori (colonne di V) formano
una base ortogonale
V è matrice di dimensione dxd

C è semi-definita positiva

- Nel nostro caso infatti $<\boldsymbol{Cv}, \boldsymbol{v}> \geq 0 \quad \forall \boldsymbol{v}$
- Poichè C può essere rappresentata come

$$\boldsymbol{C} = XX^T, \quad X \quad \text{è una} \quad d \times n \text{ matrice}$$

- Quindi:

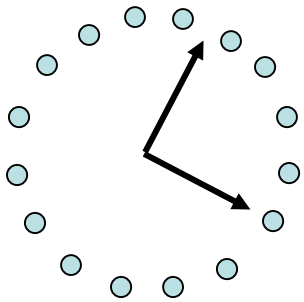
$$\boxed{\begin{aligned} &< Cv, v > = < XX^T v, v > = v^T XX^T v = \\ &= (X^T v)^T (X^T v) = < X^T v, X^T v > = \| X^T v \|^2 \geq 0 \end{aligned}}$$

Componenti Principali (PC)

- $\boldsymbol{C}$ misura la dispersione o variabilità che descrive la diffusione dei valori in un insieme di dati
- Gli **autovettori** che corrispondono agli **autovalori più grandi** rappresentano le direzioni (PC) nelle quali i dati presentano le componenti più pronunciate.
- Se gli **autovalori** sono più o meno gli stessi – non c'è direzione preferibile.

Componenti principali

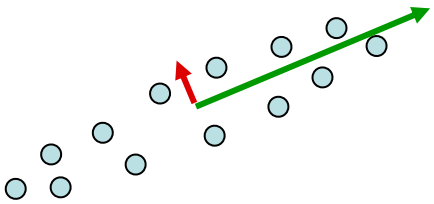
- Non c'è direzione preferibile



- $\boldsymbol{C}$ è della forma:

$$\begin{pmatrix} \lambda & \\ & \lambda \end{pmatrix}$$

- Ogni vettore è un autovettore



- C'è una chiara direzione preferibile
- $\boldsymbol{C}$ è della forma:

$$\boldsymbol{V}\begin{pmatrix} \lambda & \\ & \mu \end{pmatrix}\boldsymbol{V}^T$$

- μ è vicino a zero, molto più piccolo di λ.

Calcolo PC in caso *d* >> *n*

- Se d è grande (es. per immagini) allora la matrice di covarianza è molto grande!
- Se d>>n , è più conveniente decomporre con SVD la matrice X rettangolare (dxn) senza calcolare la matrice di covarianza.
- Infatti, si calcola la SVD di X:

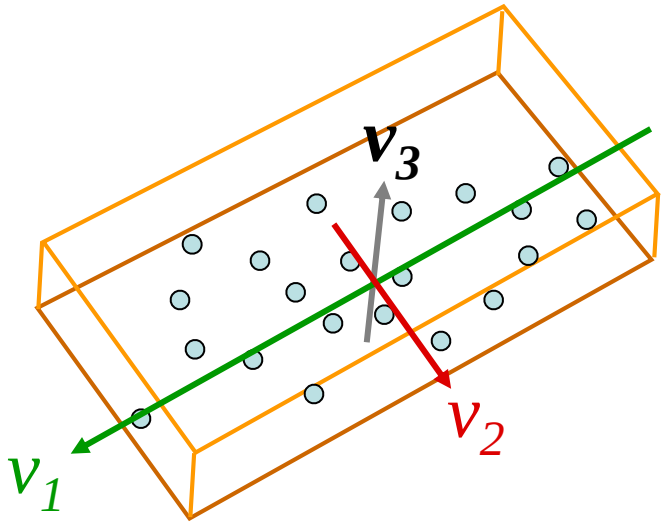
$$X = U\Gamma V^T, \quad U \in \mathbb{R}^{\mathrm{dxn}}, V \in \mathbb{R}^{\mathrm{nxn}} \quad \text{ortonormali,} \quad \Gamma \in \mathbb{R}^{\mathrm{nxn}}$$

$$C = XX^T = U\Gamma V^T V \Gamma U^T = U\Gamma^2 U^T$$

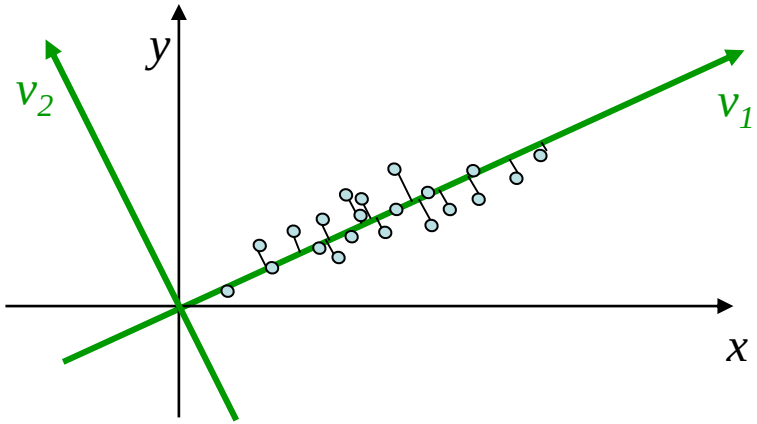
- Autovettori e autovalori di C possono essere ottenuti dalle colonne di U (vettori singolari sinistri di X) e corrispondenti elementi diagonali di Γ al quadrato (valori singolari al quadrato di X)

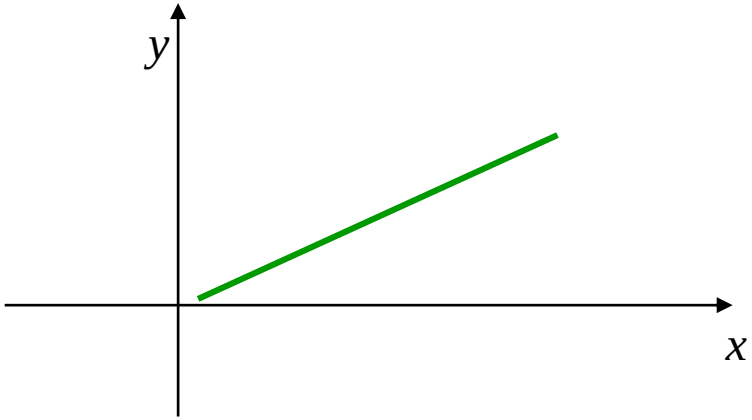
Applicazione 1

- Per trovare la bounding box orientata - calcoliamo semplicemente la box definita dagli assi definiti dagli autovettori. L'origine è il punto medio **m**.

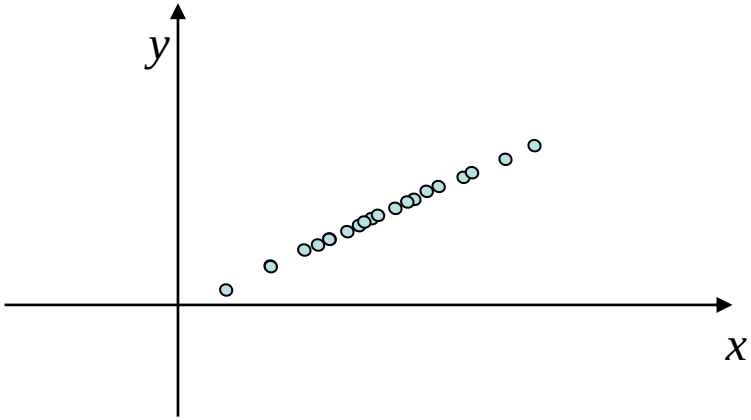


Applicazione 2: riduzione dimensionalità





Questo segmento di retta approssima I dati originali



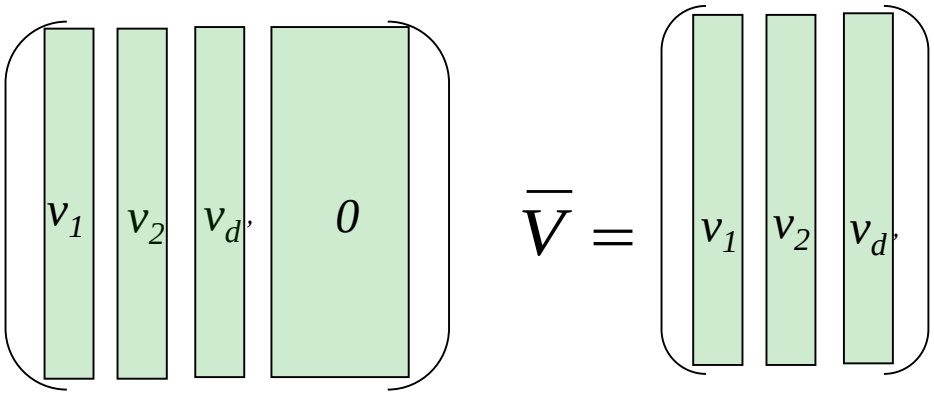
Il data set proiettato approssima il data set originale

Riduzione dimensionalità

- Per una dimensione generica d, gli autovalori sono ordinati in ordine decrescente:

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d$$

- Gli autovettori rispettano lo stesso ordine.



$$\overline{V} \in \mathbb{R}^{dxd'}$$

- Per ottenere un'approssimazione di dimensione $d' < d$, consideriamo solo i primi d' autovettori;
- Questi generano un sottospazio

($d' = 1$ è una linea, $d' = 2$ è un piano…)

- $x \in \mathrm{R}^{d'}$
- $\overline{V}x \in \mathrm{R}^{d}$
- Variando x, $\overline{V}x$ varia nel sottospazio lineare di R^{d} generato dalle colonne di $\overline{V}$

$$\overline{V}x = \begin{pmatrix} v_1 & v_2 & & v_{d'} \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ \alpha_{d'} \end{pmatrix} = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_{d'} v_{d'}$$

dx1 = (d x d') * (d' x 1)

Riduzione dimensionalità

Proiezione dei dati originali $\boxed{y_i = Vx_i \quad x_i, y_i \in \mathbb{R}^d}$

$$\boldsymbol{x}_i = \boldsymbol{m} + \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_{d'} \boldsymbol{v}_{d'} \underbrace{+ \ldots + \alpha_d \boldsymbol{v}_d}_{0}$$

nel sottospazio scelto: $\boxed{y_i' = \overline{V} x_i \quad y_i' \in \mathbb{R}^{d'}, x_i \in \mathbb{R}^d}$

$$\boldsymbol{x}_i' = \boldsymbol{m} + \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_{d'} \boldsymbol{v}_{d'} + 0 \cdot \boldsymbol{v}_{d'+1} + \ldots + 0 \cdot \boldsymbol{v}_d$$

PCA riduce la dimensionalità preservando al massimo l'informazione dei dati originali

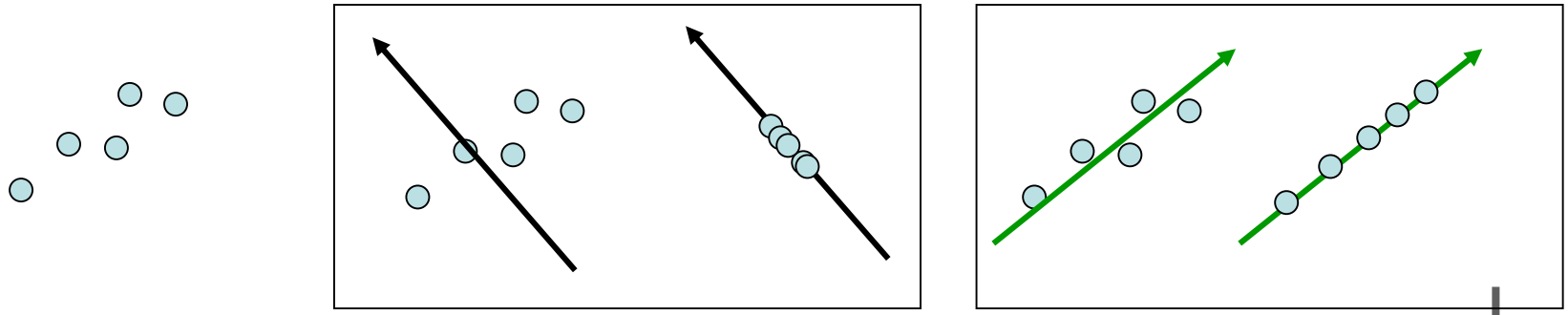
Ottimalità dell'approssimazione

PCA proietta i dati su uno spazio a dimensione inferiore tale che

1) la distanza tra il punto e la sua proiezione sia minima.
L'approssimazione è ottimale nel senso dei minimi quadrati.

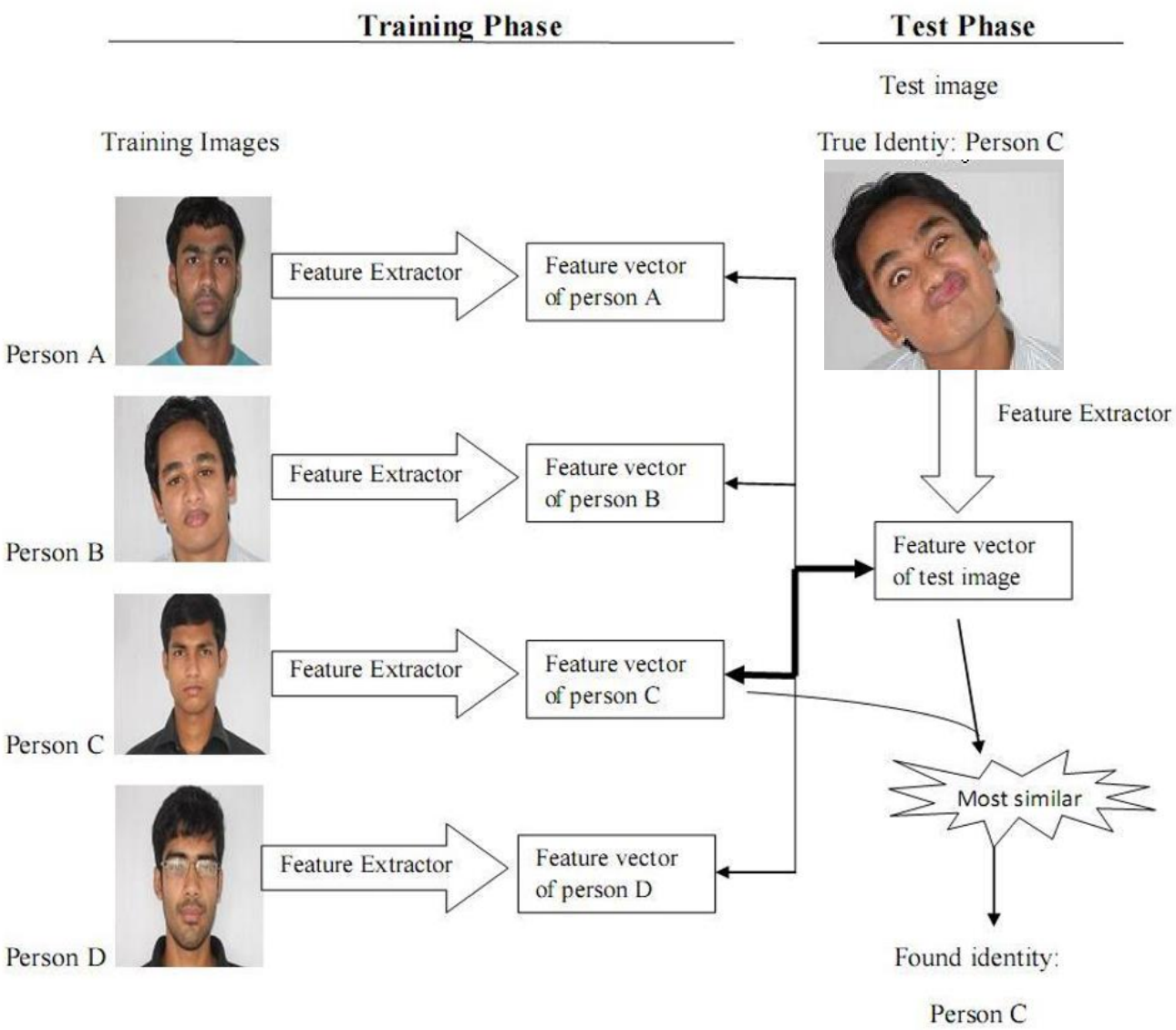
$$\min \sum_{k=1}^{n} \left\| \boldsymbol{x}_k - \boldsymbol{x}'_k \right\|^2$$

2) la varianza dei dati proiettati sia massima;



Data set Originale Proiezione su una linea arbitraria Proiezione sull'asse v_1

Esempio: Face Recognition (FR) con feature PCA



PCA per FR

- A partire dalla matrice di covarianza C del training set la diagonalizzazione di C permette di ottenere una matrice di autovettori V chiamati eigenfaces.
- Ogni immagine nel training set è rappresentata come combinazione lineare degli autovettori pesati.

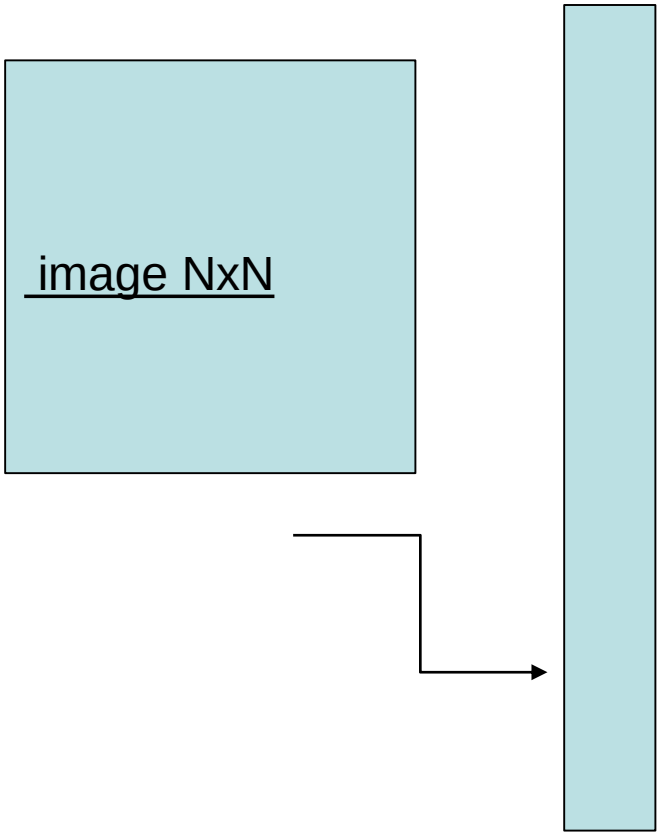
$$\boxed{y_i = Vx_i \qquad x_i, y_i \in \mathbb{R}^d}$$

- L'estrazione di una feature PCA per ogni face del training set si ottiene proiettando la face image nel sottospazio generato dagli autovettori (eigenfaces) più rilevanti (con autovalori più grandi)

$$\boxed{y_i^{'} = \overline{V}x_i \qquad y_i^{'} \in \mathbb{R}^{d'}, x_i \in \mathbb{R}^d}$$

- Il riconoscimento di una test face è elaborato proiettandola nel medesimo sottospazio
- La classificazione è effettuata misurando la minima distanza Euclidea del test face da ogni elemento del training set.

Conversione di $N \times N$ image in $N^2 \times 1$ vettore



N^2 x1 vettore

Matrice di covarianza C per M immagini test è $N^2 \times N^2$
Dalla quale si ottengono N^2 autovalori e N^2 autovettori
Per un'imagine 256x256 significa 65,536 eigenfaces

Se $M << N^2$ si applica la SVD di X (di dimensione N^2 x M) piuttosto che la diagonalizzazione di C

-> riduzione in un sottospazio di dimensione 10, 20 o 30..

Esempio

Cinque potenziali indicatori di benessere nelle diverse regioni italiane:

- X1 = PLIC (posti letto in istituti di cura)
- X2 = SC (spese complessive per famiglia)
- X3 = SA.SC (proporzione di SC dedicata agli alimentari)
- X4 = TD (tasso di disoccupazione)
- X5 = TMI (tasso di mortalità infantile)

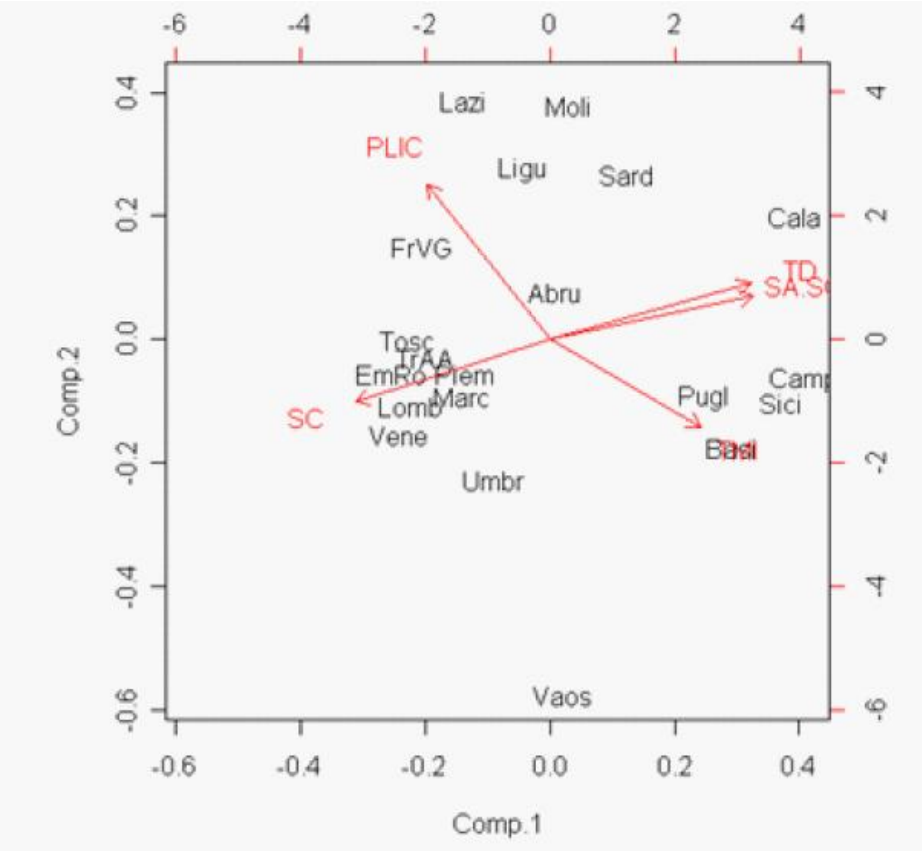
Ad ogni regione italiana R possiamo associare un vettore con cinque coordinate:

$$R = (X1(R)\ X2(R)\ X3(R)\ X4(R)\ X5(R))$$

Abbiamo 20 punti disposti su una nuvoletta simile a un'ellissoide in 5 dimensioni.

	PLIC	SC	SA.SC	TD	TMI
Piem	0.088	0.471	-0.707	-0.607	-0.395
Vaos	-1.545	0.348	-0.642	-0.813	1.578
Lomb	0.202	1.397	-0.836	-0.790	-0.538
TrAA	0.677	0.435	-1.269	-0.966	-0.075
Vene	0.088	1.334	-1.210	-0.848	-0.497
FrVG	0.639	-0.005	-1.028	-0.804	-1.301
Ligu	1.190	-0.247	0.470	-0.429	-0.354
EmRo	0.658	1.177	-1.315	-0.863	-0.347
Tosc	0.126	1.092	-0.795	-0.644	-1.355
Umbr	-1.431	0.675	-0.140	-0.524	-1.287
Marc	0.278	1.090	-0.265	-0.702	-0.0006
Lazi	2.329	0.546	-0.080	-0.113	-0.014
Abru	0.335	-0.373	0.402	-0.456	0.040
Moli	0.658	-1.289	0.065	0.451	-1.151
Camp	-1.811	-1.314	2.031	1.664	0.414
Pugl	-0.766	-0.926	1.038	0.648	1.109
Basi	-0.747	-1.154	0.661	0.844	2.001
Cala	-0.500	-1.727	1.571	2.153	0.632
Sici	-0.918	-1.130	1.332	1.517	1.783
Sard	0.449	-0.403	0.717	1.285	-0.238

	PLIC	SC	SA.SC	TD	TMI
PLIC	1	0.32	-0.41	-0.36	-0.44
SC	0.32	1	-0.84	-0.85	-0.48
SA.SC	-0.41	-0.84	1	0.90	0.51
TD	-0.36	-0.85	0.90	1	0.48
TMI	-0.44	-0.48	0.51	0.48	1



- Proiezione in 2D. Contiene tre elementi: i due nuovi assi, la proiezione dei punti sul piano principale ed alcune frecce rosse corrispondenti alle variabili originarie. Gli assi orizzontale e verticale sono rispettivamente la prima e la seconda componente principale.

- SC, TD e SA.SC sono tutti essenzialmente paralleli, a indicare una forte correlazione tra di loro: potremmo ad esempio leggere la loro direzione comune come un indicatore complessivo di benessere economico.
- Il verso di SC e opposto a quelli di TD e SA.SC, segno che questi indicatori sono correlati negativamente: come ci si aspetta, una maggior disoccupazione media si riflette su una minore spesa complessiva media (a TD alto corrisponde SC basso, e viceversa), mentre se la spesa complessiva media e' molto bassa questa sarà, come e' naturale, in gran parte dedicata agli alimentari (a SC basso corrisponde SA.SC alto, e viceversa). Allo stesso modo, la correlazione positiva tra TD e SA.SC indica che nelle zone di più alta disoccupazione le (poche) spese sono destinate per lo più ai generi alimentari.
- PLIC e TM sono abbastanza paralleli tra loro (in analogia a quanto visto sopra potremmo leggere la loro direzione comune come un indicatore complessivo di salute), ma correlati negativamente, come è naturale.

- PLIC e TM sono abbastanza perpendicolari agli altri indicatori, segno che i due gruppi, e quindi le due direzioni indicatore complessivo di benessere economico e indicatore complessivo di salute, sono abbastanza scorrelati tra loro. Tuttavia notiamo le lievi correlazioni positive nelle direzioni che ci aspettiamo: maggiori posti letto dove ci sono maggiori spese complessive, e maggior mortalità infantile dove c'è più disoccupazione e le spese sono in prevalenza alimentari.
- L'area di maggior benessere e quella nella direzione positiva di SC, con un po' di spostamento verso PLIC. In tale zona si trovano fortemente raggruppate varie regioni (Veneto, Trentino Alto Adige, Lombardia, Piemonte, Emilia Romagna, Marche e Toscana), che pertanto risultano molto simili rispetto agli indicatori considerati.
- Le altre regioni del centro-nord (Liguria, Friuli, Lazio) non eccellono in SC ma eccellono in PLIC, a indicare una buona cura sanitaria nonostante un tenore di vita medio più modesto rispetto al gruppo precedente.

- Particolarmente negativo, sia rispetto all’asse del benessere economico che a quello della salute, risulta il raggruppamento composto da Campania, Sicilia, Basilicata e Puglia, in maniera molto più accentuata rispetto ad altre regioni meridionali o insulari (come Calabria e Sardegna) che nell’immaginario collettivo potremmo invece credere ad esse simili. Questo potrebbe indicare uno sforzo di miglioramento di alcune regioni, e potrebbe ad esempio suggerire l’analisi di altri dati più mirati per averne verifica o smentita.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Serena Morigi

Dipartimento di Matematica

serena.morigi@unibo.it

http://www.dm.unibo.it/~morigi