ANALISI DELLE COMPONENTI PRINCIPALI

PROBLEMI:

Date p variabili tra loro correlate
è possibile
a) sostituirle con un numero inferiore di variabili ottenute come trasformazione lineare delle variabili originarie e che sono tra loro non correlate?

b) esprimere in forma sintetica l'azione complessiva dell'insieme riducendo il numero di variabili, seppure con una perdita d'informazione?

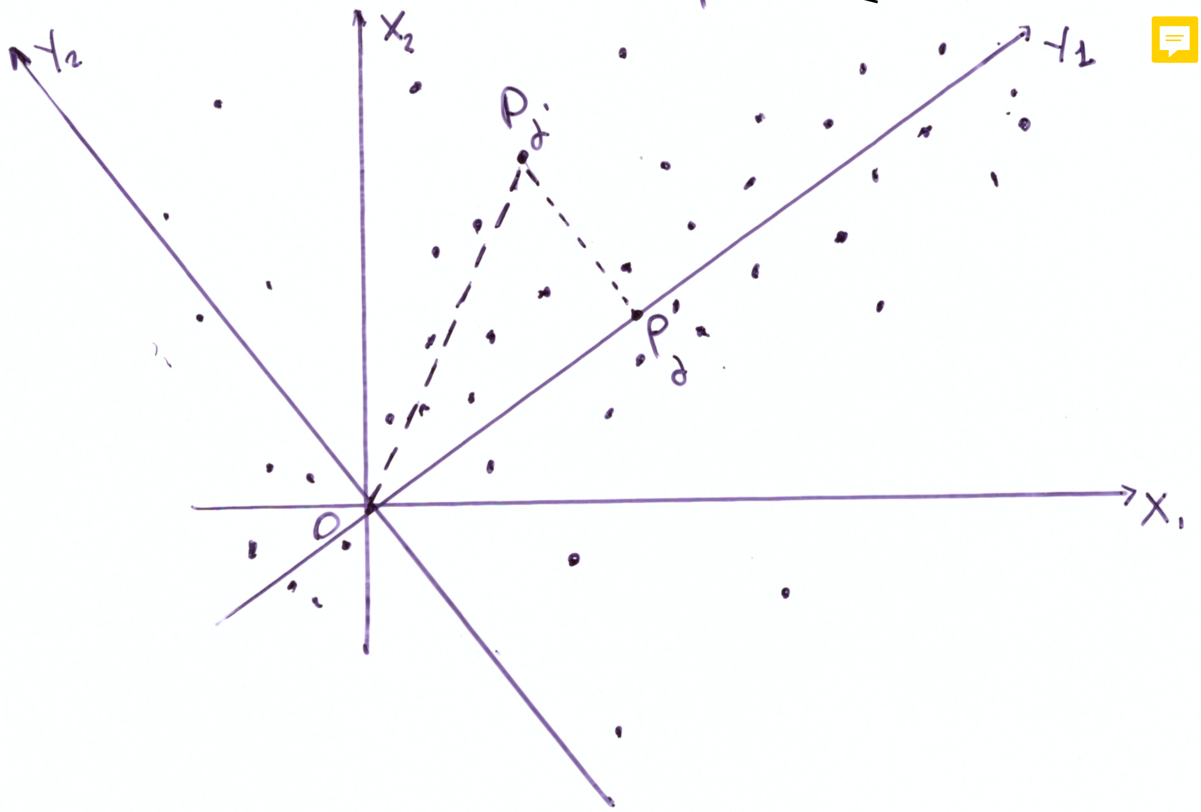
SOLUZIONI

* Galton (1889), Edgeworth (1891)
Tentano di ricombinare le misure antropometriche in strutture lineari indipendenti che consentissero di mettere in risalto le diverse proporzioni con cui i caratteri osservati possono associarsi. Date le misure: statura, lunghezza avambraccio, lunghezza gamba (campione adulti maschi) Edgeworth determina tre funzioni lineari tra loro indipendenti (indici antropometrici)

* Pearson (1901) voleva determinare rette e piani che meglio approssimassero un insieme di punti in uno spazio p-dimensionale

SOLUZIONE PER UN C.P. $p=2$

Si consideri un campione n-dimensionale su cui si sono rilevati: peso X_2, statura X_1 (Scarti dalla media



Si consideri un nuovo sistema di assi Y_1, Y_2 la forma della nube resta immutata cambiano le coordinate dei punti.

La posizione di un individuo lungo OY_1 dà indicazioni sulla sua "taglia"

La posizione di un individuo lungo OY_2 dà indicazioni sulla "forma"

Letti rispetto al nuovo sistema di riferimento

i punti consentono "forse" di ottenere maggiori

informazioni rispetto al vecchio sistema

Fissati gli assi OY_1, OY_2 si immagini che i

punti vengono compressi prima lungo l'asse OY_1

e poi lungo OY_2.

* i pti vicini all'estremità dx di OY_1

↓

individui con peso e statura elevati

i pti vicini all'estremità sin di OY

↓

individui con basso peso e statura

* i pti vicini all'estremità dx di OY_2 (individui bassi)

↓

individui alti con peso basso

i pti vicini all'estremità sin di OY_2 (individui alti e sin)

↓

individui di peso elevato ma bassa statura

Date x_1, x_2
La relazione fra i due sistemi di riferimento:

$$y_1 = x_1 \cos\alpha + x_2 \operatorname{sen}\alpha$$
$$y_2 = -x_1 \operatorname{sen}\alpha + x_2 \cos\alpha$$

α : angolo tra OX_1 e OY_1
" " " OX_2 e OY_2

Se la dispersione lungo l'asse Y_1 è molto elevata mentre quella lungo l'asse Y_2 è ridotta $\Rightarrow$ gli individui hanno "taglie" differenti ma "forme" simili.

$\Rightarrow$ la rappresentazione unidimensionale nell'asse Y_1 si può ritenere una buona approssimazione della nube bidimensionale.
Invece di x_1 "statura" e "peso" x_2

$\Downarrow$

"taglia" Y_1

$\neq$ valori di α generano $\neq Y_1$ tra tutte una rappresenta la migliore approssimazione dello spazio bidimensionale:
Derivare del valore α che rende minimo lo "spostamento" dei punti (dalla posizione originale) nel processo di proiezione

P_j : generico punto
P'_j : proiezione di P_j sull'asse OY_1 ortogonale

La retta migliore è quella che rende minima

$$\sum_{j=1}^{n} (\overline{P_j P'_j})^2$$

via teorema di Pitagora

$$(\overline{OP_j})^2 = (\overline{OP'_j})^2 + (\overline{P_j P'_j})^2$$

$$\sum_{j=1}^{n} (\overline{OP_j})^2 = \sum_{j=1}^{n} (\overline{OP'_j})^2 + \sum_{j=1}^{n} (\overline{P_j P'_j})^2$$

$$\underbrace{\frac{1}{n-1} \sum_{j=1}^{n} (\overline{OP_j})^2}_{(1)} = \underbrace{\frac{1}{n-1} \sum_{j=1}^{n} (\overline{OP'_j})^2}_{(2)} + \underbrace{\frac{1}{n-1} \sum_{j=1}^{n} (\overline{P_j P'_j})^2}_{(3)}$$

Scegliere OY_1 in modo da minimizzare (3) equivale a scegliere OY_1 in modo da massimizzare (2) $\Rightarrow$ poiché O è il baricentro della nube di p.ti (2) è la varianza delle proiezioni delle unità campionarie sull'asse Y_1

$\Rightarrow$ Per trovare la retta OY_1 che rende minima la somma dei quadrati delle distanze (perpendicolari) basta trovare la retta OY_1 tale che le proiezioni dei punti su di essa abbiano massima varianza.