L'Analisi in Componenti Principali

(S. Terzi)

Data una matrice dei dati riferiti ad n individui e k variabili **quantitative,** si sintetizzano i dati nel senso di pervenire ad una riduzione delle colonne della matrice dei dati X, definendo un numero q (q<p) di variabili artificiali.
La riduzione del numero delle variabili consente alle volte piu' agevoli sintesi interpretative.

Dal punto di vista geometrico, la matrice dei dati $X_{n,k}$ e' rappresentabile come n punti nello spazio R^k. Si tratta di proiettare gli n punti in un sottospazio R^q, individuato in modo tale che la nuvola degli n punti in R^k sia deformata il meno possibile.

1.Introduzione

Supponiamo di voler effettuare un'indagine sulle capacità sportive di un gruppo di giovani. Ciascun giovane viene sottoposto ad una serie di prove fisiche (salto in alto, 100mt, salto in lungo, 1500 mt, prove di nuoto,ecc.) rilevandone le "performances . A partire dalle variabili osservate vogliamo costruire una variabile, "artificiale", che possa considerarsi una misura globale delle capacità atletiche di una persona.
Si assume che tale variabile "artificiale" (denominata y) sia una combinazione lineare delle variabili osservate ("originali"), cioe' di x_1 ,x_2, …. x_k.

Problema: CON QUALI COEFFICIENTI ?

Si desidera che la variabile artificiale y abbia varianza massima. Si ritiene infatti che una elevata variabilità equivalga ad avere un elevato contributo informativo.

La variabile artificiale y si chiama **componente principale**, e desideriamo che:

1. sia combinazione lineare delle variabili x_1 ,x_2, …. x_k, originarie ;
2. abbia varianza massima.

Il discorso si puo' ampliare, nel senso che trovata una prima componente principale, potremmo andare a cercarne altre che riassumano ulteriori aspetti connessi con le capacità atletiche (per esempio "resistenza", oppure "scatto").

Supponiamo che per lo studio delle capacità atletiche si disponga di (tante o tantissime) variabili relative a diversi tipi di attività fisiche. Appare importante cercare di rappresentare il fenomeno con un numero piccolo di variabili, ottenute partendo dalle variabili originarie osservate, e che conservino quanta piu' informazione possibile sul fenomeno oggetto di studio.

Tecnicamente il metodo delle componenti principali risolve (o cerca di risolvere) questo problema costruendo un insieme di variabili (y_1 , y_2, …, y_q), che siano combinazioni lineari delle variabili osservate x_1 ,x_2, …. x_k, e tali che:

1) siano tra loro incorrelate
2) abbiano, ciascuna, varianza massima.

2. Determinazione analitica delle componenti principali.

Si definisca una prima componente principale y_1 , come combinazione lineare di x_1 ,x_2, …. x_k. Essendo le variabili x_1 ,x_2, …. x_k, dei vettori contenenti n osservazioni, anche y_1 sarà un vettore di n osservazioni, il cui generico elemento puo' essere indicato con y_{i1} .

Abbiamo quindi:

$$y_{i1} = a_{11}\, x_{i1} + a_{21}\, x_{i2} + \ldots + a_{k1}\, x_{ik} = \Sigma_j\, a_{j1}\, x_{ij} \qquad i=1,\ldots n$$

$$y_1 = \mathbf{X a_1}$$

dove abbiamo definito con a_{11} ,a_{21}, …., a_{k1} i coefficienti della combinazione lineare, e con a_1 il vettore *kx1* che li contiene;
dove x_{ij} e' il generico i-esimo elemento di $\mathbf{x}_j$, e $\mathbf{x_1}$,$\mathbf{x_2}$, …. $\mathbf{x_k}$ sono i vettori colonna della matrice dei dati **X.**

Parentesi: quanto vale la varianza di y_1 ?

Si dimostra che e':

$$\text{var}(y_1) = a_1^T S a_1.$$

Occorre pero' porre un vincolo sul vettore dei coefficienti. Supponiamo infatti di aver trovato un vettore a_1 che massimizzi la varianza di y_1. Tale varianza potrà essere ulteriormente incrementata utilizzando anziché il vettore $\mathbf{a_1}$ appena trovato, un nuovo vettore $c\mathbf{a}_1$, con c>1.
In sostanza si hanno un'infinità di soluzioni, note a meno di un fattore di proporzionalità c.
Per avere un'unica soluzione e' necessario porre un vincolo sugli elementi del vettore $\mathbf{a_1}$. Il vincolo che usualmente si pone e' che sia:

$$\mathbf{a_1}^T\mathbf{a_1} = 1$$

ovvero che il vettore $\mathbf{a}_1$ abbia norma unitaria (cioe' che la somma dei quadrati dei suoi elementi sia =1).

Resta un unico elemento di indeterminatezza:
se $\mathbf{a}_1$ e' soluzione (cioe' rende massima la varianza della prima componenete principale), anche $-\mathbf{a}_1$ sarà soluzione.

Per individuare la prima componenet principale bisognerà risolvere il seguente problema di massimo vincolato:

$$\text{var}(y_1) = a_1^T S a_1 \equiv \max$$
$$\mathbf{a_1}^T\mathbf{a_1} \equiv 1$$

La funzione lagrangiana da massimizzare sarà quindi data da:

$$L = a_1^T S a_1 - \lambda(\mathbf{a_1}^T \mathbf{a_1} - 1)$$

Dove λ e' il moltiplicatore di Lagrange.

Trattandosi di un problema di massimo vincolato, la soluzione si trova uguagliando a zero la derivata , rispetto al vettore $\mathbf{a}_1$, della funzione Lagrangiana.

$$= 2\mathbf{S}\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 2(\mathbf{S} - \lambda\mathbf{I})\mathbf{a}_1 \equiv \mathbf{0} \qquad \textbf{(1)}$$

La (1) individua un sistema lineare omogeneo che ammette soluzioni se e solo se:

$$\det(\mathbf{S} - \lambda\mathbf{I}) = 0 \qquad (2)$$

Le soluzioni della (2) sono gli autovalori della matrice S, Poiche' S ha dimensione *kxk*, si avranno, in linea di massima, k soluzioni.
Presa una di queste, poniamo λ_1, troveremo a_1 risolvendo:

$$(\mathbf{S} - \lambda_1\mathbf{I})\mathbf{a}_1 \equiv \mathbf{0}$$

ovvero:

$$\mathbf{S}\,\mathbf{a}_1 = \lambda_1\mathbf{a}_1 \qquad (3)$$

Ovvero, il vettore dei coefficienti a1, che stiamo cercando, sarà un autovettore (di norma unitaria) della matrice S.
Se la matrice S ha k autovalori, per determinare a_1 dobbiamo sceglierne uno tra i k.

Quale autovalore scegliere?
Riprendiamo la (3) e premoltiplichiamo entrambi i membri per $\mathbf{a}_1^T$:

$$a_1^T S a_1 = \lambda_1\, \mathbf{a_1}^T\mathbf{a_1}$$

ma anche, essendo $\mathbf{a}_1$ un vettore normalizzato:

$$a_1^T S a_1 = \lambda_1 = \mathrm{var}(y_1)$$

Poiche' vogliamo che che la varianza della prima componenete principale sia massima, sceglieremo per λ_1 il piu' grande degli autovalori di S.

Abbiamo ottenuto il seguente risultato:
la prima componente principale $\mathbf{y}_1$ e' una combinazione lineare delle k colonne della matrice **X**, con coefficienti uguali alle componenti dell'autovettore $\mathbf{a}_1$, associato al massimo autovalore della matrice S.

La matrice S e' una matrice simmetrica. Cosa sappiamo sugli autovalori?

- $\text{tr}(\mathbf{A}) = \Sigma_i \lambda_i$
- $\det(\mathbf{A}) = \Pi_i \lambda_i$

delle matrici simmetriche?

- Gli autovalori di una matrice simmetrica sono reali
- Gli autovettori di una matrice simmetrica sono a due a due ortogonali.
- Il rango di una matrice simmetrica risulta uguale al numero dei suoi autovalori non nulli.

Se la matrice S ha rango pieno (=k) possiamo trovare fino a k autovalori ed autovettori (ad essi associati), cioe' possiamo trovare fino a k componenti principali (purche', come vedremo tra poco, gli autovalori – i quali rappresentano le varianze delle C.P. -siano non negativi).

Inizialmente abbiamo detto che desideriamo che le componenti principali siano a due a due incorrelate. Si puo' dimostrare (lo vediamo forse alla lavagna) che l'ortogonalità tra coppie di autovettori di S, implica l'incorrelazione tra le corrispondenti componenti principali. Pertanto, formalmente, cercare una j-esima componente principale :

$$\mathbf{y}_j = \mathbf{X}\mathbf{a}_j$$

significa trovare il massimo della seguente funzione Lagrangiana:

$$L = a_j^T S a_j - \lambda_j(\mathbf{a}_j^T \mathbf{a}_j - 1)$$

Il vettore $\mathbf{a}_j$ sarà l'autovettore associato al j-esimo (in ordine decrescente) autovalore della matrice **S**.

3.La scelta del numero di componenti

Siamo partiti da k variabili $\mathbf{x_1}$,$\mathbf{x_2}$, …. $\mathbf{x_k}$ (i vettori colonna della matrice dei dati **X),** con l'obiettivo di sintetizzarle in un numero inferiore di variabili artificiali**.** A seconda del rango della matrice S, potremmo trovare **fino a k** componenti principali….

Come decidere quante variabili artificiali prendere?

Parentesi sul rango della matrice S.

La matrice di varianze e covarianze S ha lo stesso rango della matrice dei dati X. Inoltre la matrice S e' simmetrica, per cui il suo rango = al numero di autovalori non nulli.
Si puo' anche dimostrare che la matrice S e' semi-definita positiva; cio' comporta che i suoi autovalori sono sempre o positivi o nulli.
Quindi:

- se le colonne della matrice dei dati sono linearmente indipendenti (possibile solo se $n \geq k$), S avrà rango k, ed i suoi autovalori saranno tutti positivi.

- Se qualche autovalore risulta nullo (poniamo k-r autovalori nulli), S avrà rango r, e potremo trovare r autovalori positivi.

Tutto cio' ci porta a concludere che se S ha rango k (ovvero se e' definita positiva) troveremo k autovalori positivi, k autovettori corrispondenti e quindi k componenti principali.

Se S e' semi-definita positiva e di rango r, troveremo r autovalori non nulli ed r componenti principali.

Come fare per scegliere "poche" componenti principali?

Ciascuna componente principale ha varianza (massima) pari all'autovalore che le corrisponde.

$\text{var}(y_1) = \lambda_1$; $\text{var}(y_2) = \lambda_2$; …..$\text{var}(y_r) = \lambda_r$;

Inoltre sappiamo che $\lambda_1 \geq \lambda_2 \ldots . \geq \lambda_k$.

Se prendiamo tutte le componenti principali ricavabili, la loro varianza complessiva sarà data da:

$$\text{tr}(S) = \Sigma_j \lambda_j$$

Inoltre, per definizione di traccia, sarà anche

$$\text{tr}(S) = \Sigma_j (\text{var } x_j)$$

Volendo, possiamo calcolare il contributo della prima componente principale alla variabilità complessiva come:

$$\lambda_1 / \Sigma_j \lambda_j$$

Seguendo questo ragionamento possiamo calcolare il contributo delle prime t componenti principali alla variabilità complessiva come:

$$(\lambda_1 + \lambda_2 + \ldots . + \lambda_t) / \text{tr}(S) = I_t$$

L'idea di fondo e' che se la varianza delle prime q componenti principali e' di poco inferiore alla varianza dell k variabili originarie (le x) allora q componenti principali sono una buona sintesi delle k variabili da cui siamo partiti.
Possiamo calcolare i rapporti I_t per t= 2,3,….

Se per qualche q (q= 1,…..,k) tale rapporto risulta circa pari a 0,8 possiamo decidere di sintetizzare le k variabili di partenza mediante le prime q componenti principali, in quanto le prime q componenti principali riassumono l'80% della variabilità totale del fenomeno.

Alternativamente possiamo domandarci se l'aggiunta di una ulteriore componente principale continua a far aumentare in maniera significativa il rapporto tra varianza "spiegata" dalle prime C.P. e la varianza totale.
Sappiamo che e' : $\lambda_1 \geq \lambda_2 \ldots . \geq \lambda_k$
L'idea e' di costruire un grafico (scree-plot) in cui in ascissa si indicano i numeri d'ordine degli autovalori (1,2,…,k), ed in ordinata i λ_1 , λ_2 ,…..,λ_k ad essi corrispondenti. I punti di coordinata (j, λ_j) (j=1,…,k) vengono uniti con segmenti.

Il numero di C.P. da utilizzare sarà dato dal piu' piccolo q tale che:

1. a sinistra di q l'andamento dei λ_j sia fortemente decrescente;
2. a destra di λ_j l'andamento sia pressoche' costante o, comunque, debolmente decrescente.

E' come dire che, finche' $\lambda_j - \lambda_{j+1}$ e' "elevato" scegliamo di utilizzare almeno le prime j+1 componenti principali.
Viceversa se $\lambda_q - \lambda_{q+1}$ e' "piccolo", ci fermiamo alle prime q C.P.

3. La scelta dell'unità di misura

I risultati dell'analisi in componenti principali dipendono dall'unità di misura utilizzata per le variabili.
Supponiamo di voler cambiare u.di m. ad una variabile x_j . Formalmente e' come definire una nuova variabile $y_j = cx_j$, in cui c (positiva) e' un'opportuna costante moltiplicativa. P.es. sia $x_j$la variabile "altezza i mt.", sia y_j"altezza in cm". Sarà c=100.

Tale trasformazione altererà la matrice di varianze e covarianze S. Sarà infatti:
$var(y_j) = c^2 var(x_j)$
$cov(y_j\ x_h) = c\ cov(x_j\ x_h)$.

Ne deriva necessariamente che gli autovalori ed autovettori della matrice S verranno alterati in seguito a questo cambiamento di u. di m.

In sostanza i risultati dell'analisi in componenti principali dipendono fortemente dalle unità di misura utilizzate per le variabili oggetto di studio.
Si tratta di un inconveniente non trascurabile perche' modificando le unità di misura potremmo ottenere risultati completamente diversi.

A volte per aggirare il problema si usa condurre l'analisi non piu' sulle variabili x_1 ,x_2, …. x_k. bensi' sulle variabili standardizzate z_1 ,z_2, …. ,z_k.

Una variabile standardizzata e' definita come:

$$\mathbf{(x - M_1(x))/ s.q.m.(x)}$$

Ed ha la caratteristica di avere media zero e varianza unitaria.

La matrice di varianze e covarianze delle variabili standardizzate coincide con la matrice di correlazione (R).

E' quindi possibile effettuare un'analisi in componenti principali cercando gli autovalori (ed autovettori corrispondenti) della matrice R.

E' ovvio che gli autovalori di R saranno diversi dagli autovalori della matrice S. Di conseguenza lo saranno anche le componenti principali ottenute con un metodo o con l'altro.

4. Interpretazione delle componenti principali

Come già accennato l'ACP viene spesso utilizzata per cercare di studiare variabili "latenti". In tal senso le "variabili artificiali" a cui dà luogo rappresenterebbero misurazioni di variabili "nascoste", non osservabili direttamente (capacità sportive, intelligenza, ecc.)
Altre volte l'ACP viene utilizzate come metodo per "riassumere" i dati a disposizione.

Sorge comunque il problema di che significato semantico dare alle "variabili artificiali" componenti principali.

Entrano in gioco capacità del ricercatore, esperienza, sensibilità,…. Una serie di elementi non formalizzabili statisticamente.

Su quali strumenti statistici ci si puo' basare per l'interpretazione?

La j-esima componente principale y_j e' definita come combinazione lineare delle variabili x_1 ,x_2, …. x_k con coefficienti a_{1j} ,a_{2j}, …. a_{kj}. Ovvero:

$$y_{ij} = a_{1j}\, x_{i1} + a_{2j}\, x_{i2} + \ldots.. + a_{kj}\, x_{ik}$$

$$\mathbf{y_j = X a_j}$$

Pertanto il generico coefficiente a_{hj} rappresenta il peso che la variabile x_h ha nella determinazione della componente principale y_j (h=1,….,k). Quanto piu' grande e' a_{hj} (in valore assoluto), tanto maggiore sarà il peso che i valori x_{ih} (i=1,…,n) hanno nel determinare $y_{ij} = a_{1j}\, x_{i1} + a_{2j}\, x_{i2} + \ldots.. + a_{kj}\, x_{ik}$.

Cio' significa che la componente principale y_j sarà maggiormente caratterizzata dalle variabili x_h a cui corrispondano i coefficienti a_{hj} piu' grandi in valore assoluto. In tal modo sono proprio i coefficienti a_{hj} a conferire un significato alla componente principale y_j.

Altre (piu') utili indicazioni sono fornite dai coefficienti di correlazione tra le variabili x_h (h=1,…,k) e la j-esima componente principale y_j .

Si puo' dimostrare che:
$r_{xh,yj} = corr(x_h,y_j) = a_{jh} (\lambda_j / \sigma^2_{hh})^{1/2}$

E' chiaro che quanto piu' elevato $r_{xh,yj}$ (in valore assoluto), tanto maggiore il legame tra x_h ed y_j .
Cio' significa che a determinare il significato di y_j saranno le variabili x_h con cui e' maggiormente correlata.

NOTA: se si lavora con le variabili standardizzate $z_1, z_2, …., z_k$, avremo un diverso valore del coefficiente di correlazione (in quanto gli autovalori della matrice S non coincidono con gli autovalori della matrice R).
Inoltre, poiche' z_h ha varianza unitaria, sarà:
$r_{zh,yj} = corr(z_h,y_j) = a^*_{jh} (\lambda^*_j)^{1/2}$.

5. Il cerchio delle correlazioni
Consideriamo nel piano delle prime due CP, un cerchio di raggio unitario.
Calcoliamo per ognuna delle k variabili della matrice X, il coefficiente di correlazione con la prima e la seconda CP. Rappresentiamo ogni variabile (x_h)come punto (nel cerchio suddetto) di coordinate: ($r_{xh,y1}$, $r_{xh,y2}$).
Avremo cosi' l'indicazione grafica di quali variabili determinino maggiormente l'una, l'altra o entrambe le CP; di quali siano correlate positivamente e quali negativamente e cosi' via.

E' possibile ripetere il procedimento sul piano della terza e quarta CP, ….e cosi' via.
6.Altre considerazioni

Premessa sulle trasformazioni di variabili:

Si abbia una variabile (vettore) x_j, di media $M_1(x_j)$; di varianza σ^2_{jj}.
Si definisca una nuova variabile y_j come $y_j = cx_j$.
Si puo' dimostrare che :
$M_1(y_j) = c\, M_1(x_j)$; $var(y_j) = c^2\, var(xj)$.
Inoltre, data una (diversa) variabile x_h, sara':
$cov(y_j,x_h) = c.cov(x_j,x_h)$.
Quindi se nella matrice dei dati X decidiamo di modificare l'unità di misura di una variabile, risulterà modificata l'intera matrice di varianze e covarianze S.

Prima considerazione
Si abbia una variabile (vettore) y, combinazione lineare di k variabili (vettori) $x_1, x_2, …. x_k$:

$$y_1 = a_1 x_1 + a_2 x_2 + ….. + a_k x_{ik} = \Sigma_j a_j x_j$$

Si puo' dimostrare che:

$$M_1(y) = \Sigma_j a_j M_1(x_j)$$

E che:

$$Var(y) = \Sigma_j a^2_j var(x_j) + \Sigma_j \Sigma_h a_j a_h Cov(x_j , x_h)$$

(Tra l'altro questo e' lo sviluppo della forma quadratica a'Sa).

Se lo scopo e' quello di trovare dei coefficienti a_j che rendano massima la varianza della y, si intuisce che lo scopo verrà raggiunto dando peso "alto" alle variabili con varianza e covarianze "alte".

Tuttavia se questa elevata variabilità e' da attribuirsi ad una "cattiva" scelta dell'unità di misura della variabile in questione, i risultati dell'analisi saranno in una certa misura "falsati".

Seconda considerazione

Supponiamo che x_j, x_h siano incorrelate per ogni j , h = 1,…k, j≠h. Questo significa che tutte le covarianze sono nulle e pertanto:

$$Var(y) = \Sigma_j a^2_j var(x_j)$$

Come trovare i coefficienti a_j, tali da masimizzare la varianza di y?

Supponiamo, per semplicità, che sia:

$var(x_1) > var(x_2) > \ldots\ldots > var(x_k)$

Ricordiamo che il vettore dei coefficienti a e' un vettore normalizzato. Cio' comporta, tra l'altro, $0 \leq a_j \leq 1$ per ogni valore di j.

Var(y) sarà massima se poniamo $a_1 = 1$; $a_j = 0$ per j> 1, ovvero ponendo $y_1 = x_1$.

In maniera simile, andando a cercare la seconda componente principale, si puo' vedere che la sua varianza sarà massima ponendo $a_{22} = 1$; $a_{j2} = 0$ per j≠ 1. La seconda componente principale sarà quindi data da $y_2 = x_2$.

In sostanza in questo caso le componenti principali y_1 ,y_2, …. y_k coincideranno con le variabili originarie x_1 ,x_2, …. x_k . In particolare y_1 coinciderà con la variabile x_j di massima varianza; y_2 con la variabile x_h che ha la seconda massima varianza, e cosi' via.

Formalmente si potrebbe dimostrare che, essendo in questo caso la matrice S una matrice diagonale, i suoi autovalori coincidono con gli elementi che si trovano sulla sua diagonale (cioe' $\lambda_h = \sigma^2_{hh}$), e gli autovettori (di norma unitaria) ad essi corrispondenti coincideranno con i vettori elementari, cioe' $a_h = e_h$.

(Si ricorda che il vettore elementare e_h e' un particolare vettore avente tutti elementi nulli tranne l'h-esimo che vale 1).