STATISTICA MULTIVARIATA

Corso di laurea in matematica — Anno accademico 2004/05 — Numero 8

L'immagine 2-dimensionale

Situazione 29.1. $X \in \mathbb{R}^n_m$ sia la nostra matrice di dati con $n \geq 2$. Come nella nota 24.7 e quando non indicato diversamente, siano $\lambda_1, \dots, \lambda_m$ gli autovalori di X^{COM} con $\lambda_1 \geq \dots \geq \lambda_m$ ed $e^1, \dots, e^m$ una base ortonormale di $\mathbb{R}_m$ tale che

$$\varphi_X e^1 = \lambda_1 e_1, \dots, \varphi_X e^m = \lambda_m e_m$$

La funzione `Smp` definita a pagina 25 calcola la matrice le cui colonne sono le componenti principali di X:

$$(X_{e^1}, \dots, X_{e^m})$$

La modifichiamo aggiungendo un secondo argomento facoltativo con cui possiamo calcolare determinate colonne di questa matrice:

```
Smp = function (X,j)
{XE=Smg.cen(X)%*%Smp.autovettori(X)
if (missing(j)) XE
else XE[,j]}
```

In questo modo con `Smp(X,1:2)` otteniamo le coordinate (cioè le lunghezze con segno rispetto al baricentro $\overline{X}$) delle proiezioni ortogonali dei punti X^i sul piano $\overline{X} + \mathbb{R}e^1 + \mathbb{R}e^2$.

X sia la matrice dei dati per i 15 comuni visti a pagina 16. Per caricare i dati e per ottenere la matrice numerica X usiamo le istruzioni

```
Db(2)
X=Db.matrice()
```

Adesso con `XE=Smp(X)` otteniamo la matrice delle componenti principali e, siccome i coefficienti sono piuttosto grandi, li possiamo stampare con `print(round(XE,0))`:

```
-282 -216  215  49
  55  -48  -72 -11
-213 -252   76 -32
-215  187    5 -22
  53  -81  -92 -14
 296   83  -66  45
 973   29   23 -18
-113  -78 -140  24
-163   36  -14 -43
-239    1 -106  27
-228  420  114  -6
 580 -100   96   7
-215 -124   61 -26
 -78  257   17  19
-210 -114 -118  -1
```

Con

```
XE12=Smp(X,1:2)
print(round(XE12,0))
```

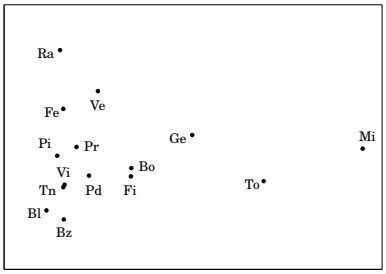
otteniamo quindi

```
-282 -216
  55  -48
-213 -252
-215  187
  53  -81
 296   83
 973   29
-113  -78
-163   36
-239    1
-228  420
 580 -100
-215 -124
 -78  257
-210 -114
```

La prima colonna è uguale ad X_{e^1}, la seconda ad X_{e^2}. Se riportiamo questi valori in un sistema cartesiano piano, otteniamo una proiezione

$$\begin{aligned} \mathbb{R}_4 &\longrightarrow \mathbb{R}_2 \\ X^i &\longmapsto (X^i_{e^1}, X^i_{e^2}) \end{aligned}$$

ottimale nel senso della nota 28.1.



Prima di ogni ragionamento matematico o statistico, proviamo a capire se questa proiezione può essere considerata convincente. E in effetti alcune configurazioni possono essere già intravviste: Milano si distingue fortemente dagli altri comuni, e i comuni più vicini sono le altre grandi città, soprattutto Torino e Genova e poi Bologna e Firenze. Non è un caso che andiamo verso sinistra perché è appunto l'asse orizzontale quello con la varianza maggiore. Vediamo che seguono verso sinistra Venezia, Parma e Padova, e poi gli altri comuni, con quelli più vicini al mare (in particolare Ravenna e Venezia) più in alto, e le città di montagna (Bolzano, Belluno, Trento) più in basso. La rappresentazione 2-dimensionale che abbiamo ottenuto dalle componenti principali è quindi già piuttosto soddisfacente.

Per valutare l'affidabilità matematica calcoliamo, con `Smp.autovalori(X)`, gli autovalori di X^{COM}, ottenendo dopo arrotondamento i valori

$$\begin{aligned} \lambda_1 &= 1791717 \\ \lambda_2 &= 450423 \\ \lambda_3 &= 141728 \\ \lambda_4 &= 10903 \end{aligned}$$

per cui

$$\begin{aligned} &\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} \\ &= \frac{1791717 + 450423}{1791717 + 450423 + 141728 + 10903} \\ &= \frac{2242140}{2394771} = 0.936 \end{aligned}$$

In questo numero

Il rapporto di varianza

Ricordiamo dalla nota 24.13 che la varianza di X_{e^k} è uguale a λ_k per ogni k e quindi la somma $\lambda_1 + \dots + \lambda_m$ (nel caso generale) può essere considerata la *varianza totale* dei nostri dati; siccome la *traccia* di una matrice quadratica è uguale alla somme dei suoi autovalori, la varianza totale è uguale alla traccia di X^{COM}. A questo punto è naturale, in una proiezione 2-dimensionale sui primi due assi principali, considerare il quoziente

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \dots + \lambda_m}$$

detto secondo *rapporto (cumulativo) di varianza*, come indice della bontà statistica della proiezione, da interpretare con molta precauzione, come vedremo, soprattutto quando si confrontano standardizzazioni diverse. Nell'esempio presentato su questa pagina il rapporto di varianza è uguale a 0.936 e quindi le prime due componenti principali rappresentano più del 93% della varianza totale.

La differenza

$$1 - \frac{\lambda_1 + \lambda_2}{\lambda_1 + \dots + \lambda_m}$$

è una misura invece della *profondità* dei dati rappresentati; a una profondità maggiore corrisponde un rischio maggiore che punti vicini nella proiezione sul piano siano invece lontane nella realtà, cioè in $\mathbb{R}_m$.

„The central idea of principal component analysis is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set ... Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix. Thus, the definition and computation of principal components are straightforward but, as will be seen, this apparently simple technique has a wide variety of different applications, as well as a number of different derivations ... Despite the apparent simplicity of the technique, much research is still being done in the general area of PCA, and it is very widely used.“ (Jolliffe, ix, 9)

La standardizzazione X^{NG}

Definizione 30.1. Poniamo

$$X^{NG} := (X_1^{NG}, \dots, X_m^{NG})$$

Sappiamo dall'osservazione 7.9 che

$$\overline{X_j^{NG}} = 0$$

per ogni j e quindi

$$(X_j^{NG})^{CE} = X_j^{NG}$$

Ciò a sua volta implica che

$$(X^{NG})^{COM} = (X^{NG})^t X^{NG}$$

Sostituendo la matrice X con X^{NG}, possiamo perciò applicare la teoria finora sviluppata a questa nuova matrice.

Poniamo $Y := X^{NG}$. Allora

$$(Y^t Y)^i_j = \|X_i^{NG}, X_j^{NG}\| = r_{X_i X_j}$$

Per questa ragione la matrice $(X^{NG})^{COM}$ si chiama anche la *matrice di correlazione* di X. Essa in R può essere ottenuta semplicemente con `cor(X)`. Per trovare X^{NG} possiamo definire la funzione

```
Smg.ng = function (X)
apply(X,2,Sg.ng)
```

Sia adesso X la matrice dei 15 comuni; definiamo $Y := X^{NG}$ come sopra e procediamo come a pagina 29, sostituendo X con Y.

```
Db(2)
X=Db.matrice()
Y=Smg.ng(X)
print(round(Y,2))
```

ottenendo prima $Y = X^{NG}$:

```
-0.22  0.65  0.08 -0.12
 0.04 -0.10  0.05 -0.13
-0.17  0.37  0.47 -0.27
-0.15 -0.20 -0.10  0.30
 0.03 -0.10  0.08 -0.19
 0.23 -0.17 -0.35  0.03
 0.73  0.06  0.28 -0.06
-0.09 -0.19 -0.22 -0.20
-0.12 -0.09  0.17  0.07
-0.18 -0.21 -0.30 -0.05
-0.14 -0.21 -0.32  0.71
 0.43  0.32  0.26 -0.14
-0.17  0.22  0.29 -0.10
-0.04 -0.22 -0.36  0.38
-0.17 -0.13 -0.04 -0.22
```

A questo punto con

```
YE=Smp(Y)
print(round(YE,2))
```

calcoliamo le componenti principali di Y:

```
-0.43 -0.39  0.24  0.31
-0.05  0.05 -0.14 -0.06
-0.60 -0.29  0.05 -0.11
 0.35 -0.07  0.11 -0.13
-0.10  0.05 -0.19 -0.08
 0.28  0.27 -0.12  0.20
-0.37  0.68  0.10 -0.05
 0.15 -0.05 -0.32  0.08
 0.00 -0.09  0.03 -0.22
 0.31 -0.12 -0.23  0.09
 0.71 -0.04  0.40 -0.05
-0.50  0.32  0.14  0.09
-0.32 -0.23  0.07 -0.10
 0.55  0.04  0.13  0.07
 0.01 -0.14 -0.27 -0.03
```

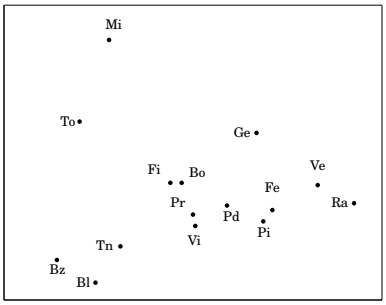
oppure, con

```
YE12=Smp(Y,1:2)
print(round(YE12,2))
```

le prime due componenti principali di Y:

```
-0.43 -0.39
-0.05  0.05
-0.60 -0.29
 0.35 -0.07
-0.10  0.05
 0.28  0.27
-0.37  0.68
 0.15 -0.05
 0.00 -0.09
 0.31 -0.12
 0.71 -0.04
-0.50  0.32
-0.32 -0.23
 0.55  0.04
 0.01 -0.14
```

Riportiamo anche stavolta questi valori in un sistema cartesiano piano:



Sicuramente la risoluzione in questo caso è migliore che prima della standardizzazione; anche i gruppi che possiamo formare, ad esempio

Milano, Torino
Genova, Venezia, Ravenna
Ferrara, Padova, Pisa
Firenze, Bologna, Parma, Vicenza
Trento, Bolzano, Belluno

sono abbastanza convincenti. Forso l'unico dubbio potrebbe riguardare la vicinanza tra Ferrara e Pisa (bisogna però anche tener conto dei dati che avevamo a disposizione) e la notevole distanza tra Trento e Vicenza molto vicine nella prima proiezione. Calcoliamo anche qui gli autovalori con `Smp.autovalori(Y)`, ottenendo

$$\lambda_1 = 2.18$$
$$\lambda_2 = 0.98$$
$$\lambda_3 = 0.58$$
$$\lambda_4 = 0.26$$

Nonostante la favorevole impressione, stavolta il secondo rapporto di variazione è uguale a 0.789 e perciò più basso di quello ottenuto a pagina 29; ma siamo partiti da standardizzazioni diverse. Anche qui, come quando si osserva un oggetto tridimensionale in natura, è utile osservarlo da prospettive diverse.

Esercizio 30.2. Definendo

$$X^{NS} := (X_1^{NS}, \dots, X_m^{NS})$$

si ha

$$(X^{NS})^t X^{NS} = (n-1) \cdot (X^{NG})^t X^{NG}$$

Se definiamo quindi $Y := X^{NG}$, $Z := X^{NS}$, allora le componenti principali di Z si distinguono da quelle di Y solo per un fattore $\sqrt{n-1}$, per cui otteniamo risultati sostanzialmente equivalenti.

La standardizzazione X^{01}

Proviamo adesso ad applicare il metodo generale alla matrice X^{01} che si ottiene da X mediante proiezione su $[0,1]$. Per i quindici comuni X^{01} è già stata calcolata a pagina 16. Con

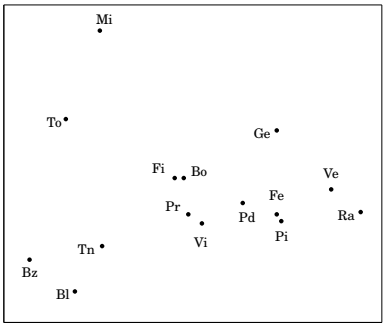
```
Db(2)
X=Db.matrice()
X01=Sm.tra01(X)

X01E=Smp(X01)
print(round(X01E,2))

X01E12=Smp(X01,1:2)
print(round(X01E12,2))
```

otteniamo la matrice delle componenti principali

```
-0.52 -0.43  0.24  0.35
-0.04  0.07 -0.15 -0.07
-0.72 -0.29  0.01 -0.13
 0.37 -0.09  0.12 -0.16
-0.08  0.07 -0.21 -0.09
 0.37  0.28 -0.09  0.23
-0.41  0.72  0.12 -0.05
 0.22 -0.04 -0.33  0.10
-0.02 -0.09  0.01 -0.26
 0.39 -0.12 -0.23  0.11
 0.74 -0.08  0.44 -0.08
-0.56  0.33  0.15  0.11
-0.40 -0.23  0.05 -0.12
 0.61  0.02  0.16  0.07
 0.04 -0.13 -0.30 -0.03
```

e le prime due colonne, che corrispondono alle prime due componenti principali di X^{01} e che poi rappresentiamo in $\mathbb{R}_2$:

```
-0.52 -0.43
-0.04  0.07
-0.72 -0.29
 0.37 -0.09
-0.08  0.07
 0.37  0.28
-0.41  0.72
 0.22 -0.04
-0.02 -0.09
 0.39 -0.12
 0.74 -0.08
-0.56  0.33
-0.40 -0.23
 0.61  0.02
 0.04 -0.13
```



Il risultato è molto simile a quello ottenuto per X^{NG}. Anche il rapporto di variazione 0.79 è praticamente identico.

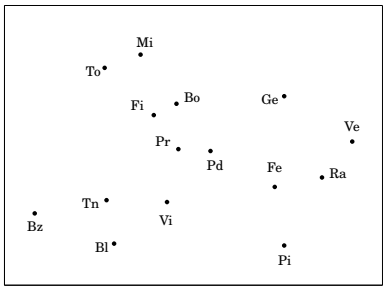
Analisi della matrice dei ranghi

Eseguiamo infine l'analisi delle componenti principali per la matrice dei ranghi. Con

```
Db(2); X=Db.matrice()
XR=Sm.rango(X)
XRE12=Smp(XR,1:2)
print(round(XRE12,2))
```

otteniamo le prime due componenti principali, che riportiamo nel piano:

```
-0.42 -0.46
-0.09  0.28
-0.84 -0.30
 0.43 -0.16
-0.21  0.22
 0.48  0.32
-0.28  0.54
 0.09  0.03
-0.08  0.04
 0.48 -0.47
 0.68 -0.11
-0.47  0.47
-0.46 -0.23
 0.84  0.08
-0.14 -0.24
```



La risoluzione è molto buona e la classificazione in gruppi convincente. Anche qui vediamo che l'uso dei ranghi introduce degli aspetti che sfuggono talvolta all'analisi puramente metrica-lineare. Gli autovalori sono $\lambda_1 = 3.28, \lambda_2 = 1.42, \lambda_3 = 0.81, \lambda_4 = 0.19$, il rapporto di variazione è 0.82.

screeplot

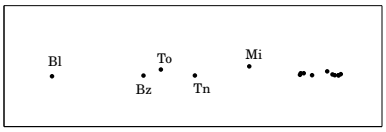
Combinando `princomp` con `screeplot` si possono visualizzare i rapporti tra gli autovalori. Provare dal terminale

```
Db(2); X=Db.matrice(); p=princomp(X)
screeplot(p)
screeplot(p,type='lines')
summary(p)
```

Perché bisogna standardizzare

Siccome le componenti principali dipendono fortemente dalle scale di misura usate per le variabili, i dati devono sempre essere standardizzati, usando X^{NG}, X^{01}, la matrice dei ranghi o un'altra trasformazione per ottenere una forma dei dati che possiede opportune proprietà di invarianza.

Assumiamo di aver misurato le altezze dei 15 comuni in centimetri. Allora nella matrice dei dati la seconda colonna deve essere moltiplicata per 100. Procedendo con la matrice così ottenuta come a pagina 29, otteniamo la seguente figura:



Si vede chiaramente che l'altezza determina in pratica da sola la proiezione cancellando quasi del tutto il significato delle altre variabili. Il rapporto di variazione stavolta è addirittura uguale a 0.9998 ma ciò, come si vede, non garantisce un risultato soddisfacente.

È quindi sempre necessario effettuare una standardizzazione. In alcuni casi ci possono essere ragioni per attribuire *pesi* diversi alle variabili, lavorando ad esempio con $(X_1^{NG}, 2X_2^{NG}, 0.4X_3^{NG})$, se la seconda variabile ci sembre più importante della prima e questa a sua volta più importante della terza. Una tale scelta deve però essere giustificata dalle caratteristiche dei dati.

Se più colonne della matrice dei dati esprimono lo stesso fenomeno, esse naturalmente avranno più peso in un'analisi delle componenti principali e questa molteplicità di colonne essenzialmente uguali non è eliminata dalle standardizzazioni finora viste. Ciò mostra che è molto importante pianificare in anticipo quali variabili vogliamo scegliere per l'analisi statistica. Talvolta anche qui può aiutare l'analisi delle componenti principali di X^t.

Analisi di X^t

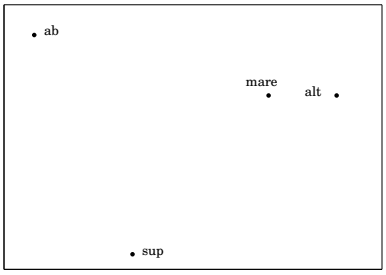
Molto spesso (non solo per scoprire colonne multiple) può essere utile studiare anche la trasposta X^t della matrice dei dati mediante un'analisi delle componenti principali. Usiamo la proiezione su $[0, 1]$ come standardizzazione e procediamo come a pagina 30:

```
Db(2)
X=Db.matrice()
Xt=t(X)
Xt01=Sm.tra01(Xt)

CP12=Smp(Xt01,1:2)
print(round(CP12,2))
```

ottenendo così le prime due delle 15 componenti principali:

```
-1.64  0.88
 1.47  0.24
 0.80  0.26
-0.63 -1.39
```

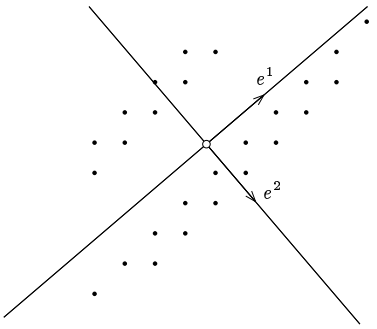
che possiamo riportare anche in questo caso in un sistema cartesiano:



La figura, una proiezione 2-dimensionale del $\mathbb{R}^{15}$, mostra la vicinanza tra i fattori *distanza dal mare* e *altezza*. In un'indagine medica, dove le colonne corrispondono a caratteristiche cliniche e le righe a pazienti oppure le righe a cellule tumorali e le colonne a geni di ciascuno dei quali per ogni cellula è indicata l'intensità di espressione, in questo modo si possono individuare gruppi di fattori o geni con effetti vicini. Una discussione di tecniche multivariate nello studio di microarray di DNA si trova nei libri di Amaratunga/Cabrera, Jagota e Lee.

Un problema di classificazione

Non sempre la prima componente principale è la più adatta nei compiti di classificazione. Guardiamo la seguente figura:



È evidente che la varianza in direzione e^1 è notevolmente maggiore che in direzione e^2; nonostante ciò i dati si distinguono in due gruppi che sono determinati dalla seconda componente principale. Se ciò accade in una proiezione $\mathbb{R}_m \longrightarrow \mathbb{R}_2$ con $m > 2$, una tale divisione in gruppi può sfuggirci.

Bibliografia

16498 **D. Amaratunga/J. Cabrera:** Exploration and analysis of DNA microarray and protein array data. Wiley 2004.

16693 **G. Dunn/B. Everitt:** An introduction to mathematical taxonomy. Dover 2004.

16041 **J. Gentle:** Elements of computational statistics. Springer 2002.

13332 **H. Handels:** Medizinische Bildverarbeitung. Teubner 2000.

15645 **A. Jagota:** Microarray data analysis and visualization. Bay Press 2001.

15993 **I. Jolliffe:** Principal component analysis. Springer 2002.

15524 **K. Mardia/J. Kent/J. Bibby:** Multivariate analysis. Academic Press 2000.

16728 **M. Lee:** Analysis of microarray gene expression data. Kluwer 2004.

17084 **D. Morrison:** Multivariate statistical methods. Thomson 2005.

In meccanica ($m = 3$) la ricerca del primo asse principale (asse con momento inerziale massimo) è importante, perché la rotazione attorno a questo asse gode di stabilità.