Note sull'analisi delle corrispondenze implementata in R

Giovanna Jona Lasinio

1 Introduzione

L'analisi delle componenti principali viene utilizzata quando le variabili coinvolte sono quantitative, meglio se continue. Il punto fondamentale è che tutta la PCA è basata sul concetto di distanza tra osservazioni o tra variabili secondo una metrica euclidea. Quando le variabili in studio sono di tipo qualitativo (nomi di specie ad esempio associate a misure di abbondanza o di sola presenza assenza), non è naturale utilizzare una metrica euclidea per definire la distanza tra gli oggetti presenti nel dataset. Proprio per risolvere questo aspetto e comunque rendere possibile uno studio su di uno spazio a dimensione ridotta di insiemi di dati multidimensionali, è stata introdotta l'Analisi delle Corrispondenze (semplice o Multipla-ACM).

Il principio su cui si basa l'ACM è lo stesso su cui si basa l'analisi in componenti principali, solo che invece di usare come misura di variabilità la correlazione o la covarianza, utilizza una distanza tra righe (o colonne) della matrice dei dati, basata sulla statistica χ^2, cioè sulle frequenze. Fine dell'analisi è quello di spiegare perché la matrice dei dati si scosta da una situazione di omogeneità che si presenta quando le righe (o le colonne) sono proporzionali, portando alla luce l'intreccio di legami, le corrispondenze, tra le righe, tra le colonne e tra righe e colonne della matrice dei dati e perciò tra le diverse caratteristiche dell'insieme dei dati in esame. Vediamo prima un esempio e poi i dettagli.

2 Esempio: Doubs

Riprendiamo i dati dell'esempio del fiume Doubs. Nel dataset abbiamo una tabella (`doubs$poi`) con sulle righe i siti di monitoraggio del fiume e sulle colonne le specie di pesci. In ogni cella è riportato il numero di individui della specie osservati nel luogo. Questa tabella può essere vista come una tabella di frequenza doppia, dove si incrociano i siti di monitoraggio con le specie ittiche presenti. Esploriamo un po' questi dati.
Vogliamo rispondere ad alcune domande:
Quali sono le specie più numerose?

```
> require(ade4)
> data(doubs)
> sort(apply(doubs$poi, 2, sum), decreasing = TRUE)
```

```
LOC VAI GAR TRU ABL CHE GOU TAN VAN BAR BRO GRE PER BOU BBO PSO SPI ANG TOX BCO CAR
 73  68  63  57  57  56  55  45  43  43  40  38  36  33  31  29  27  27  26  26  25
ROT BLA HOT PCH CHA OMB
 21  19  18  18  15  15
```

la funzione `apply` permette di sommare i valori nelle colonne, ordinando questi in ordine decrescente (funzione `sort` con l'opzione `decreasing=TRUE`) otteniamo ai primi posti le sigle delle specie a più elevata numerosità.
In quali siti si trovano il maggior numero di individui?

```
> sort(apply(doubs$poi, 1, sum), decreasing = TRUE)
```

```
30 29 22 28 27 21 20 19 17 26 18 16  5 15 14  4  6 13 12  3  7 24  9 10  2 11 25 23
89 87 72 70 63 62 56 46 44 43 42 40 34 33 28 21 21 19 18 16 16 15 14 14 12 11 11  4
 1  8
 3  0
```

questa informazione la otteniamo applicando la stessa procedura del punto precedente alle righe della matrice dei dati.

Dove si trovano i siti più densamente popolati?

```
> plot(doubs$xy, pch = 20)
> points(doubs$xy[c(30, 29, 22, 28, 27), ], pch = 20, col = 2)
```

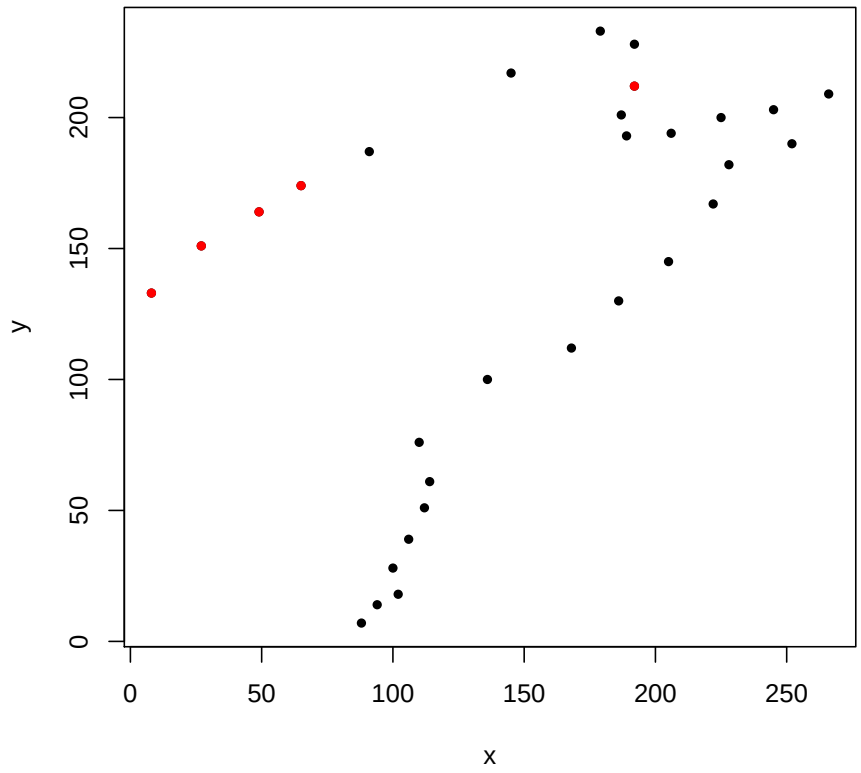

Figura 1: Esempio Doubs: localizzazione dei siti campionati, in rosso i 4 siti con la più alta numerosità di individui osservati

Come si caratterizzano i siti rispetto alla presenza di specie? Per rispondere a questa domanda abbiamo bisogno di poter rappresentare siti e specie sullo stesso spazio/grafico, quindi utilizziamo l'AC per ottenere uno spazio "ottimale" di dimensione 2:

```
> ac = dudi.coa(doubs$poi, scann = F)
> biplot(ac, pos = "bottomright")
```

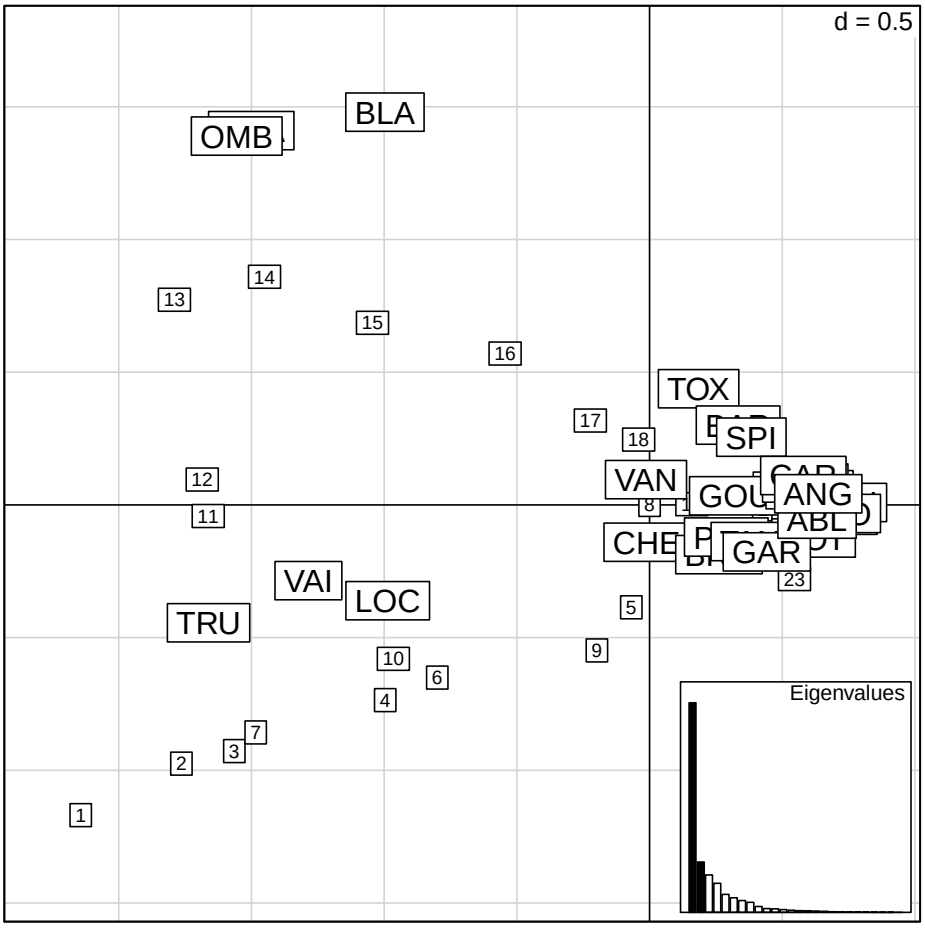

Figura 2: Esempio Doubs: biplot dell'analisi delle corrispondenze per la matrice faunistica

Nel biplot in figura 2 rappresentiamo, come per la PCA, righe (siti) e colonne (specie) sullo stesso piano fattoriale. Le specie giocano il ruolo delle variabili nella PCA e quindi caratterizzano gli assi. In questa figura però la sovrapposizione tra siti e specie è tale da rendere difficile la lettura dello stesso. Produciamo due grafici distinti ed affiancati:

```
> par(mfrow = c(2, 1))
> s.label(ac$co, clab = 0.6, lab = row.names(ac$co))
> s.label(ac$li, clab = 0.6)
> par(mfrow = c(1, 1))
```

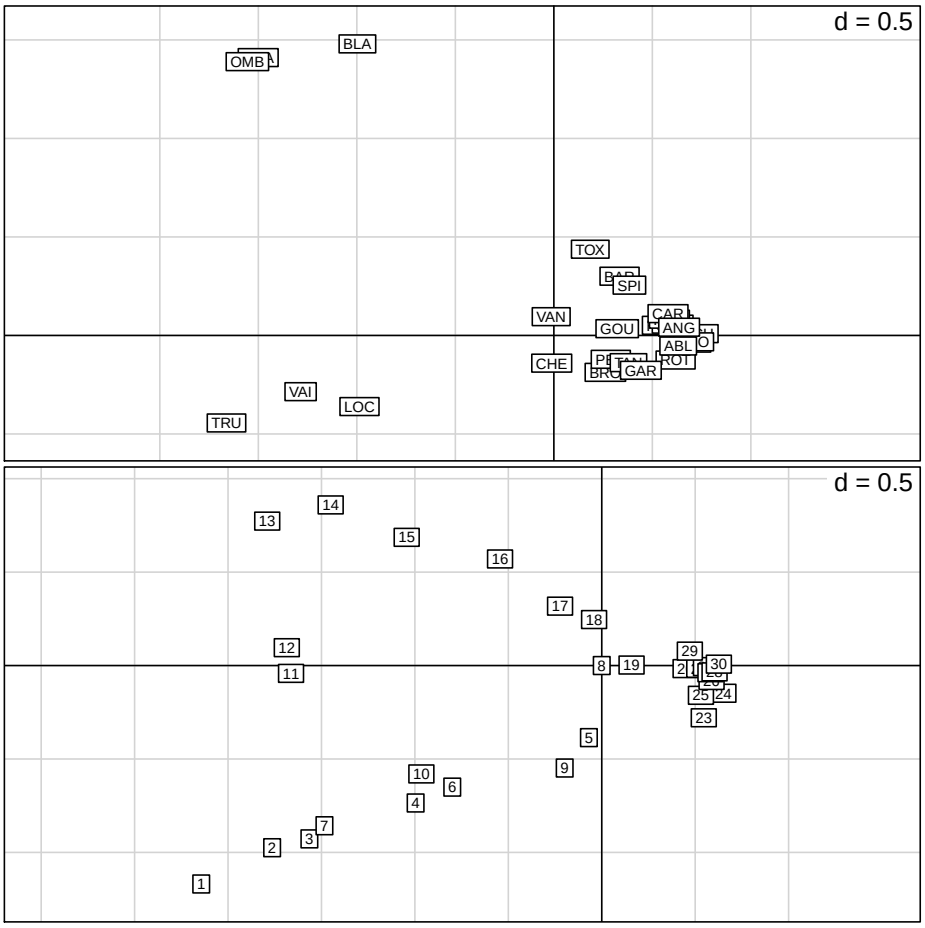

Figura 3: Esempio Doubs: rappresentazione delle specie (in alto) e dei siti sul primo piano fattoriale

Per interpretare questi grafici è utile guardare i punteggi delle specie sugli assi, questo ci permette di caratterizzarli come già visto con la PCA.

	Comp1	Comp2
CHA	-1.50	1.41
TRU	-1.66	-0.44
VAI	-1.29	-0.29
LOC	-0.99	-0.36
OMB	-1.56	1.39
BLA	-1.00	1.48
HOT	0.55	0.05
TOX	0.18	0.44
VAN	-0.01	0.10
CHE	-0.01	-0.14
BAR	0.33	0.30
SPI	0.38	0.26
GOU	0.32	0.03
BRO	0.26	-0.19
PER	0.29	-0.12
BOU	0.60	0.06
PSO	0.59	0.08
ROT	0.62	-0.12
CAR	0.58	0.11
TAN	0.38	-0.14
BCO	0.70	-0.01
PCH	0.73	0.01
GRE	0.69	-0.04
GAR	0.44	-0.18
BBO	0.71	-0.03
ABL	0.63	-0.05
ANG	0.64	0.04

Tabella 1: Esempio DOUBS: Punteggi delle specie sui primi due assi principali

La disposizione delle specie sugli assi permette di capire, osservando la collocazione dei siti sugli stessi, quale sia la composizione delle popolazioni ittiche negli stessi.

3 Un po' di teoria

L'analisi delle corrispondenze semplice (o multipla) si basa, come detto sugli stessi principi della PCA, utilizzando una definizione della matrice dei dati $\mathbf{X}$ e della metrica $\mathbf{D}$ diverse. Vediamo quali.
Per prima cosa consideriamo la matrice dei dati, questa può essere, come nell'esempio precedente, una matrice di occorrenze (frequenze) di un dato fenomeno, nell'esempio, le specie di pesci osservate lungo il fiume Doubs. Per ciascun sito di monitoraggio si contano quanti individui di ogni specie vengono osservati. Abbiamo una sola variabile qualitativa

(la specie di pesci) con un elevato numero di modalità (le singole specie), $\mathbf{X}$ è una tabella di frequenza con sulle righe i luoghi e colonne le singole *modalità*. La matrice della metrica definisce una ponderazione delle righe (o delle colonne), usualmente si prende l'inverso del numero totale di osservazioni per ciascun sito (o per ciascuna specie se ponderiamo le colonne), quindi se n_{ij} è le elemento di riga i e colonna j in $\mathbf{X}$, definiamo $n_{i.} = \sum_j n_{ij}$, quindi $\mathbf{D} = \text{diag}(1/n_{i.})$. Dal prodotto $\mathbf{XD}$ otteniamo la matrice dei *profili riga*, se invece definiamo $n_{.j} = \sum_i n_{ij}$ possiamo costruire $\mathbf{P} = \text{diag}(1/n_{.j})$ e otteniamo $\mathbf{PX}$ la matrice dei *profili colonna*. Schematicamente, possiamo dire che i profili riga sono dati dalle frequenze relative di ciascuna specie in ogni sito, mentre i profili colonna sono costituiti dalle frequenze relative di ogni specie nei diversi siti

- **Matrice dei dati X**

Siti/Specie	X_1	...	X_k	Tot
S_1	n_{11}	...	n_{1k}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_p	n_{p1}	...	n_{pk}	$n_{p.}$
Tot	$n_{.1}$	...	$n_{.k}$	n

- **Matrice dei profili riga**

Siti/Specie	X_1	...	X_k	Tot
S_1	$pr_{11} = \frac{n_{11}}{n_{1.}}$	...	$pr_{1k} = \frac{n_{1k}}{n_{1.}}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_p	$pr_{p1} = \frac{n_{p1}}{n_{p.}}$	...	$pr_{pk} = \frac{n_{pk}}{n_{p.}}$	1
Tot	$pr_{.1} = \frac{n_{.1}}{n}$	...	$pr_{.k} = \frac{n_{.k}}{n}$	-

Questa matrice corrisponde ad una ponderazione delle righe di $\mathbf{X}$ con l'abbondanza totale di individui nei siti campionati. Indicheremo con $\mathbf{pr}_i$ la riga i-esima della matrice dei profili riga (il profilo riga i-esimo).

- **Matrice dei profili colonna**

Siti/Specie	X_1	...	X_k	Tot
S_1	$pc_{11} = \frac{n_{11}}{n_{.1}}$	...	$pc_{1k} = \frac{n_{1k}}{n_{.1}}$	$pc_{.1} = \frac{n_{1.}}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_p	$pc_{p1} = \frac{n_{p1}}{n_{.p}}$	...	$pc_{pk} = \frac{n_{pk}}{n_{.p}}$	$pc_{.p} = \frac{n_{p.}}{n}$
Tot	1	...	1	-

Questa matrice corrisponde ad una ponderazione delle colonne di $\mathbf{X}$ con l'abbondanza totale di individui di ciascuna specie osservati nell'area di studio. Indicheremo con $\mathbf{pc}_j$ la colonna j-esima della matrice dei profili colonna (il profilo colonna j-esimo)

Quello che ci interessa a questo punto è stabilire come valutare la distanza, nello spazio delle specie, tra due specie o tra due siti campionati. La distanza più appropriata tra due profili riga o colonna è la *distanza del* χ^2:

$$\begin{aligned} d(\mathbf{pr}_i, \mathbf{pr}_h) &= \sum_{j=1}^{k} \frac{1}{pr_{.j}} (pr_{ij} - pr_{hj})^2 \\ &= \sum_{j=1}^{k} \frac{1}{\frac{n_{.j}}{n}} \left(\frac{n_{ij}}{n_{.i}} - \frac{n_{hj}}{n_{.h}} \right)^2 \qquad (1) \\ d(\mathbf{pc}_j, \mathbf{pc}_h) &= \sum_{i=1}^{p} \frac{1}{pc_{.i}} (pc_{ij} - pc_{ih})^2 \\ &= \sum_{i=1}^{p} \frac{1}{\frac{n_{i.}}{n}} \left(\frac{n_{ij}}{n_{j.}} - \frac{n_{ih}}{n_{h.}} \right)^2 \qquad (2) \end{aligned}$$

L'espressione (1) è la distanza tra due profili riga mentre l'espressione (2) è la distanza tra due profili colonna. L'analisi delle corrispondenze, utilizzando questa distanza per costruire la matrice di associazione tra profili riga (o colonna), procede ora esattamente come la PCA. In realtà si procede ad una vera e propria PCA sulla matrice dei profili riga (o colonna) usando la distanza del χ^2.