Analisi in componenti principali

Giovanna Jona Lasinio
giovanna.jonalasinio@uniroma1.it

Dipartimento di Scienze Statistiche - Università *Sapienza* Roma

Lo scopo dell'analisi in componenti principali (ACP in italiano, PCA in inglese) è:

> *rappresentare un insieme multivariato di dati in uno spazio a dimensione ridotta.*

Se consideriamo una tabella di dati $\mathbf{X}$ con n righe e p colonne, questa può essere vista come un oggetto in uno spazio a n dimensioni (se guardiamo le colonne) o a p dimensioni (se guardiamo le righe).
Se $n, p > 3$ non siamo in grado di visualizzare o trattare l'oggetto $\mathbf{X}$ facilmente,. Quindi

> *ci si propone di cercare uno spazio di dimensione $k < n, p$ in cui rappresentare i nostri dati in modo "ottimale", ovvero conservando la maggior quantità d'informazione possibile.*

Per capire cosa si intende per “non facilmente” apriamo R e lavoriamo su di un esempio.

Esempio delle misure del carapace delle tartarughe (library(ade4))

Prendiamo il dataset `doubs` contenuto nel pacchetto `ade4`. Questo insieme di dati riguarda osservazioni di variabili ambientali (11) e abbondanze di specie di pesci (27) in 30 siti lungo il Doubs, un fiume che percorre Francia e Svizzera. In particolare

- doubs$mil contiene le seguenti variabili ambientali: das - distanza dalla sorgente (km * 10), alt - altitudine (m), pen ($\log(x+1)$ dove x è l'inclinazione (per mil * 100), deb - minimum average debit (m3/s * 100), pH (* 10), dur - durezza dell'acqua (mg/l di Calcio), pho - fosfati (mg/l * 100), nit - nitrati (mg/l * 100), amm - ammonia nitrogeno (mg/l * 100), oxy - ossigeno disciolto (mg/l * 10), dbo - domanda biologica di ossigeno (mg/l * 10).
- doubs$poi contiene le abbondanze delle seguenti specie di pesci: *Cottus gobio* (CHA), *Salmo trutta fario* (TRU), *Phoxinus phoxinus* (VAI), *Nemacheilus barbatulus* (LOC), *Thymallus thymallus* (OMB), *Telestes soufia agassizi* (BLA), *Chondrostoma nasus* (HOT), *Chondostroma toxostoma* (TOX), *Leuciscus leuciscus* (VAN), *Leuciscus cephalus cephalus* (CHE), *Barbus barbus* (BAR), *Spirlinus bipunctatus* (SPI), *Gobio gobio* (GOU), *Esox lucius* (BRO), *Perca fluviatilis* (PER), *Rhodeus amarus* (BOU), *Lepomis gibbosus* (PSO), *Scardinius erythrophtalmus* (ROT), *Cyprinus carpio* (CAR), *Tinca tinca* (TAN), *Abramis brama* (BCO), *Ictalurus melas* (PCH), *Acerina cernua* (GRE), *Rutilus rutilus* (GAR), *Blicca bjoerkna* (BBO), *Alburnus alburnus* (ABL), *Anguilla anguilla* (ANG).
- doubs$xy contiene le coordinate dei 30 siti di rilevazione.

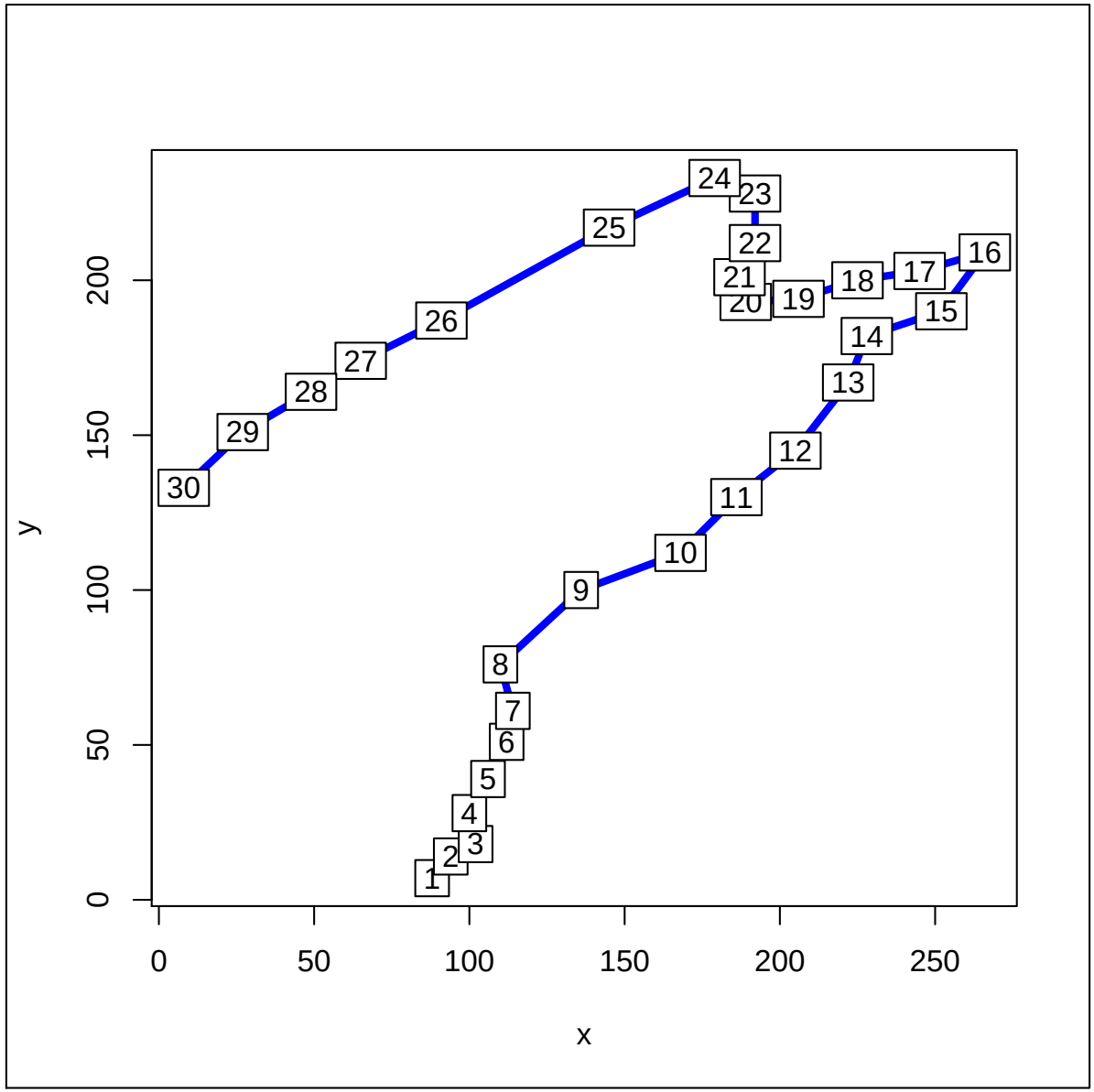

Figura: Dataset doubs, stazioni di rilevazione lungo il Doubs

Le variabili ambientali definiscono un oggetto in 30 dimensioni (numero dei siti campionati)
se invece guardo ai siti ho un oggetto in 11 dimensioni.
La tabella delle specie se vista dal lato delle specie definisce un oggetto in 30 dimensioni,
dal lato dei siti, un oggetto in 27 dimensioni.
Chiaramente non posso rappresentare nessuna delle due tabelle su di un grafico in due o tre dimensioni.

Cosa può essere interessante analizzare?

- Quali siano le similitudini tra le variabili (11 o 27 punti in 30 dimensioni).
- Quali relazioni legano tra di loro gli individui o i siti osservati (30 punti in 11 o 27 dimensioni).
- Visualizzare i dati in unico grafico al massimo tridimensionale.
- Cerco di rispondere a queste questioni possibilmente in spazi di dimensione ridotta k ($k < 11, 27, 30$) **senza perdere informazione**.

Per prima cosa devo decidere **come misurare l'informazione contenuta nei dati**. Intanto diamo uno sguardo complessivo alle variabili ambientali:

Riapriamo R e impariamo ad esplorare questi dati

Per misurare l'informazione complessiva possiamo usare due strumenti:

- *La matrice delle varianze e covarianze*: contiene sulla diagonale principale i valori delle varianze ($Var(X_i)$, $i = 1, \ldots, p$) delle singole variabili e negli elementi extra-diagonali le covarianze ($Cov(X_i, X_j)$ $i, j = 1, \ldots, p$ $i \neq j$) tra le coppie di variabili. Varianze e covarianze sono calcolati centrando le variabili rispetto alle proprie medie e rappresentano le distanze euclidee medie di ciascuna variabile, o coppia di variabili per la covarianza, dal baricentro della variabile o della coppia. Sono una "buona" rappresentazione dell'informazione contenuta nei dati vista come variazione attorno alla media. Sono poco confrontabili tra loro.
- *La matrice di correlazione*: contiene le correlazioni tra le variabili, quindi ha una diagonale composta di soli 1 e gli elementi extra diagonali contenenti le correlazioni tra le variabili ($r(X_i, x_j) = \frac{Cov(X, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$, $i, j = 1, \ldots, p$ $i \neq j$).

  Questa matrice è legata a quella di varianze e covarianze per definizione di correlazione, i valori al suo interno sono ottenuti standardizzando le variabili, infatti le correlazioni sono le distanze euclidee medie dall'origine delle variabili standardizzate.

Matrice di Covarianza

	das	alt	pen	deb	pH	dur	pho	nit	amm
das	1950234.93								
alt	-356641.84	73650.88							
pen	-1140.22	223.94	1.17						
deb	2399119.86	-427037.88	-1399.72	3276772.85					
pH	11.47	-17.57	-0.51	64.22	3.02				
dur	16437.38	-3409.02	-11.91	21272.33	2.60	284.44			
pho	58492.87	-10514.40	-38.23	61126.16	-12.67	537.75	7681.56		
nit	147387.06	-29172.45	-93.28	155318.10	-12.00	1217.44	9913.10	19976.39	
amm	21632.80	-3922.69	-14.39	20235.42	-8.17	185.87	3220.95	4273.58	1436.82
oxy	-15786.96	2175.53	11.10	-14350.33	6.81	-142.85	-1404.96	-1969.44	-605.21
dbo	21354.82	-3542.88	-13.24	17710.95	-10.19	224.81	2999.01	3507.76	1297.43

Matrice di correlazione

	das	alt	pen	deb	pH	dur	pho	nit	amm	oxy	dbo
das	1.00										
alt	-0.94	1.00									
pen	-0.76	0.76	1.00								
deb	0.95	-0.87	-0.72	1.00							
pH	0.00	-0.04	-0.27	0.02	1.00						
dur	0.70	-0.74	-0.65	0.70	0.09	1.00					
pho	0.48	-0.44	-0.40	0.39	-0.08	0.36	1.00				
nit	0.75	-0.76	-0.61	0.61	-0.05	0.51	0.80	1.00			
amm	0.41	-0.38	-0.35	0.29	-0.12	0.29	0.97	0.80	1.00		
oxy	-0.51	0.36	0.46	-0.36	0.18	-0.38	-0.72	-0.63	-0.72	1.00	
dbo	0.40	-0.34	-0.32	0.25	-0.15	0.34	0.89	0.64	0.89	-0.84	1.00

Dalla lettura della matrice di correlazione evinciamo parecchie informazioni, vediamo quali variabili hanno una variazione concorde e quali no, quali hanno una relazione forte tra loro e così via. Va ricordato che il *coefficiente di correlazione misura l'intensità del legame lineare tra le coppie di variabili*, se due variabili hanno una relazione non lineare, questo coefficiente non mi permette di vederla.

Cerchiamo una visualizzazione complessiva delle variabili ambientali, possiamo costruire i grafici a dispersione a coppie e gli istogrammi delle stesse riportando tutto su di un'unica immagine:

- Sia **X** una tabella di dati con n righe e p colonne, sulle righe sono le osservazioni, sulle colonne le variabili osservate
- questa tabella può essere vista secondo due diversi punti di vista [1]
  1. **Oggetto in** $\mathbb{R}^n$: i punti sono le colonne con n coordinate(abbiamo p punti).
  2. **Oggetto in** $\mathbb{R}^p$: i punti sono le righe con p coordinate (abbiamo n punti).
- **X** può essere una matrice floro-faunistica contenente p specie osservate in n luoghi o una tabelle di misure di p variabili ambientali in n siti.

[1] Riferimento Dray S, Dufour AB (2007). The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22(4). URL http://www.jstatsoft.org/v22/i04/.

Proprietà delle distanze: Dati x, y due punti in uno spazio *metrico una funzione* $d(x, y)$ *è una* distanza se verifica le seguenti proprietà

- $d(x, y) \geq 0$ e $d(x, y) = 0$ se e solo se $x = y$.
- $d(x, y) = d(y, x)$ simmetria
- dati x, y, z accade che $d(x, y) \leq d(x, z) + d(z, y)$ diseguaglianza triangolare

In uno spazio vettoriale la distanza è definita dal *prodotto scalare* o in termini più generali dalla *norma* associata allo spazio (che poi è proprio una distanza).

Definiamo gli strumenti necessari

- $\mathbf{Q}$ una matrice $p \times p$ che permetta di definire la distanza in $\mathbb{R}^p$ tra le n osservazioni
- $\mathbf{D}$ una matrice $n \times n$ che ha lo stesso ruolo di $\mathbf{Q}$ in $\mathbb{R}^n$
- $\mathbf{Q}$ e $\mathbf{D}$ sono simmetriche e definite positive, ovvero hanno tutti gli autovalori > 0 e $\mathbf{Q} = \mathbf{Q}^T$, $\mathbf{D} = \mathbf{D}^T$
- nella pratica la scelta di $\mathbf{X}, \mathbf{Q}$ e $\mathbf{D}$ dipende strettamente dallo scopo dello studio.
- Indicheremo con $x_{\cdot j}$ e con $x_{i \cdot}$ i totali di colonna e riga rispettivamente.

- Cosa vuol dire definire la distanza in $\mathbb{R}^p$ tra le n osservazioni?
- Significa calcolare $\mathbf{XQX}^T$ in modo tale che questa quantità sia una *distanza*
- Analogamente per calcolare una distanza tra le variabili: $\mathbf{X}^T\mathbf{DX}$.
- Se $\mathbf{X}$ contiene solo misure quantitative
  1. distanza Euclidea tra le osservazioni $\mathbf{Q} = \mathbf{I}_p$ dove $\mathbf{I}_p = diag(1)$
  2. distanza tra le variabili: covarianza, allora $\mathbf{X} = [x_{ij} - m(x^j)]$ con $m(x^j)$ media della colonna j e $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$
  3. se preferiamo usare la correlazione tra le variabili allora $\mathbf{X} = [\frac{x_{ij} - m(x^j)}{sd(x^j)}]$ dove $sd(x^j)$ la deviazione standard della colonna j e $\mathbf{D} = \frac{1}{n}\mathbf{I}_n$.
- Per una matrice di abbondanze può essere molto utile applicare la PCA ai profili di specie $\mathbf{X} = [\frac{x_{ij}}{x_{.j}}]$, in questo modo si rimuove l'effetto delle differenze tra le abbondanze globali delle singole specie.

- Diverse definizioni per $\mathbf{Q}$ e $\mathbf{D}$ permettono di dare un peso diverso alle singole specie o ai singoli siti
- Ad esempio $\mathbf{Q} = diag(x_{\cdot 1}, \ldots, x_{\cdot p})$ permette di pesare ogni specie con la sua abbondanza complessiva quando si calcolano le distanze tra i siti.
- Questa scelta utile, ad esempio, quando si assume che il campionamento usato non sia del tutto rappresentativo per la comunità in studio (specie rare non catturate etc.)

- La definizione di $\mathbf{X}$ è anche di grande importanza, in particolare lo è la scelta su come *centrare i dati*
- La centratura definisce l'origine del sistema di riferimento, se usiamo i dati grezzi senza ad esempio sottrarre la media, il sistema viene centrato sul record di soli zeri.
- Centrare i dati rispetto al totale di specie $\mathbf{X} = [x_{ij} - x_{.j}]$ significa assumere come punto di riferimento un sito ipotetico in cui la composizione delle specie è la composizione media delle specie calcolata su tutti i siti osservati.
- Centrare rispetto al totale di sito $\mathbf{X} = [x_{ij} - x_{i.}]$ implica che il punto di riferimento sia una specie che in ogni sito ha un'abbondanza proporzionale all'abbondanza totale della specie stessa.
- *Ciascuna scelta della tripletta* $\mathbf{X}, \mathbf{Q}, \mathbf{D}$ *corrisponde ad un diverso tipo di analisi*

Ricordiamo che le componenti principali sono tra loro incorrelate.

Questo significa che ogni componente contiene una parte di informazione diversa dalle altre.

Ora riapriamo R ed impariamo ad interpretare l'output della PCA