Note sull'analisi in componenti principali implementata in R

Giovanna Jona Lasinio

Queste note sono da intendersi per uso interno e non vanno divulgate senza l'autorizzazione dell'autore.

1 Introduzione

Lo scopo dell'analisi in componenti principali (ACP in italiano, PCA in inglese) è di rappresentare un insieme multivariato di dati in uno spazio a dimensione ridotta. In altre parole, 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 l'oggetto $\mathbf{X}$ facilmente, o di sintetizzare agevolmente l'informazione in esso contenuta. 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.*

1.1 Visualizzazione dei dati. Un primo esempio

Prendiamo un insieme di misure del carapace di un gruppo di tartarughe, le misure sono registrate in centimetri, il dataset è presente in R nel pacchetto `ade4`:

```
> library(ade4)
> data(tortues)
> names(tortues)

[1] "long" "larg" "haut" "sexe"
```

per praticità rinominiamo in inglese le colonne del dataset:

```
> pturtles <- tortues
> names(pturtles) <- c("length", "width", "height", "sex")
```

Poniamoci ora il problema di come rappresentare queste variabili tutte insieme. Innanzitutto vogliamo distinguere tra maschi e femmine nei grafici, quindi costruiamo una variabile che assegna il colore blu ai maschi e il rosso alle femmine:

```
> sex <- pturtles$sex
> sexcol <- ifelse(sex == "M", "blue", "red")
```

Poi ci costruiamo, solo per mantenere l'ordine, un nuovo dataset con le variabili quantitative continue e rappresentiamo le variabili su grafici a dispersione a coppie:

```
> measures <- pturtles[, 1:3]
> plot(measures, col = sexcol, pch = 19)
```

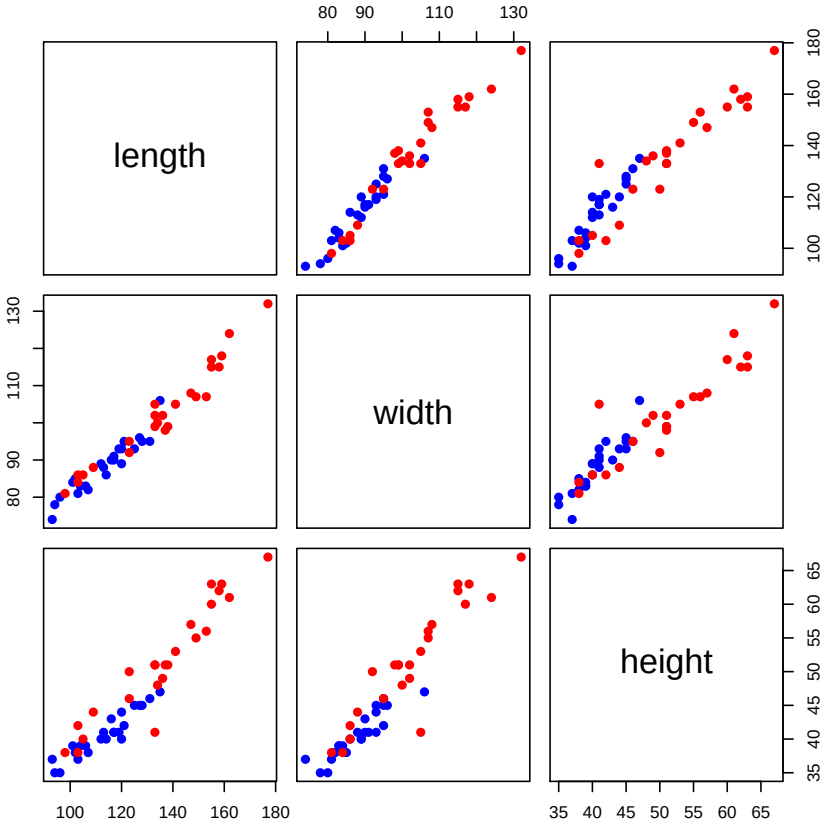


Figura 1: Grafici a dispersione delle misure del carapace delle tartarughe, in blu i maschi in rosso le femmine

Da questi grafici vediamo che le variabili considerate sono molto correlate tra loro, vediamo inoltre che le femmine presentano valori più elevati per tutte le misure.

Visualizziamo ora in 3 dimensioni queste misure, usiamo dei grafici interattivi che sono implementati nel pacchetto `rgl` di R, oppure usando la funzione `scatterplot3d` contenuta nel pacchetto omonimo:

```
> library(rgl)
> plot3d(measures, type = "s", col = sexcol)
```

```
> require(scatterplot3d)
> scatterplot3d(measures, highlight.3d = F, col.axis = "blue", col.grid = "lightblue",
+     main = "", pch = 20, color = sexcol)
```

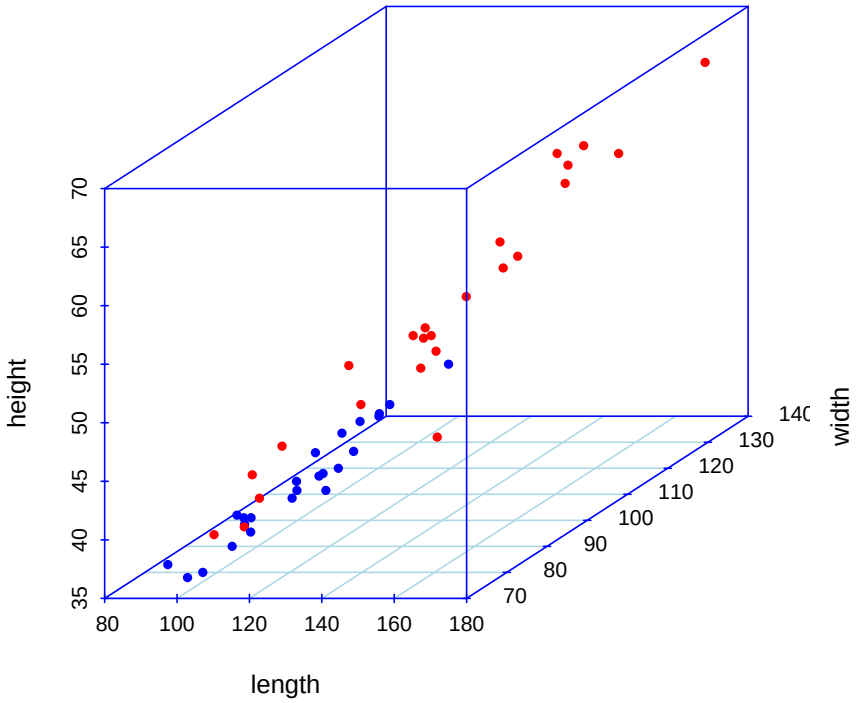


Figura 2: Rappresentazione tridimensionale delle misure del carapace delle tartarughe, in blu i maschi, in rosso le femmine

Notiamo che i tre assi sono su scale abbastanza diverse, per poter confrontare bene le tre variabili dovremmo cercare di riportarle tutte su di una stessa scala. Cominciamo semplicemente portando i tre assi ad avere gli stessi limiti:

```
> lims <- c(min(measures), max(measures))
> scatterplot3d(measures, highlight.3d = F, col.axis = "blue", col.grid = "lightblue",
+     main = "", pch = 20, color = sexcol, xlim = lims, ylim = lims,
+     zlim = lims)
```

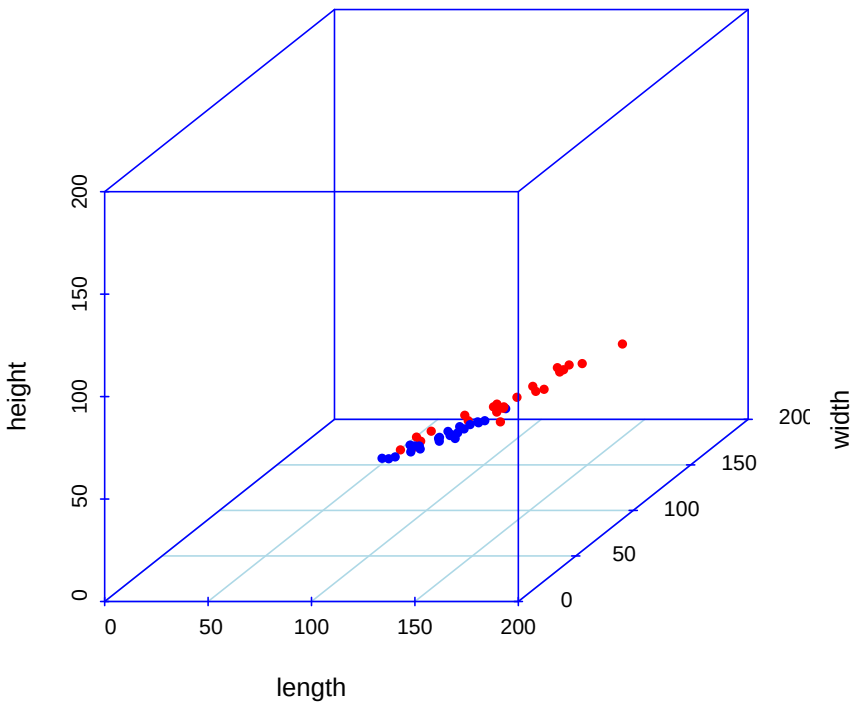


Figura 3: Rappresentazione tridimensionale delle misure del carapace delle tartarughe, in blu i maschi, in rosso le femmine, gli assi sono riportati tutti sulla stessa scala

In questo plot le variabili sono riportate su scale simili ma il risultato è poco leggibile, ciò è dovuto alla differenza esistente tra le medie delle variabili:

```
> sapply(measures, mean)
```

```
   length     width    height
124.68750  95.50000  46.14583
```

A causa di questa differenza vediamo i punti tutti schiacciati verso la base del grafico in 3-D. Quindi centriamo le variabili rispetto alla media. In questo ci aiuta la funzione `scale`,

l'opzione `center` permette di sottrarre la media di colonna dalle stesse, mentre l'opzione `scale` permette di dividere i valori per le deviazioni standard delle colonne:

```
> measures.c <- scale(measures, center = TRUE, scale = FALSE)
> lims <- c(min(measures.c), max(measures.c))
> scatterplot3d(measures.c, highlight.3d = F, col.axis = "blue", col.grid = "lightblue",
+     main = "", pch = 20, color = sexcol, xlim = lims, ylim = lims,
+     zlim = lims)
```

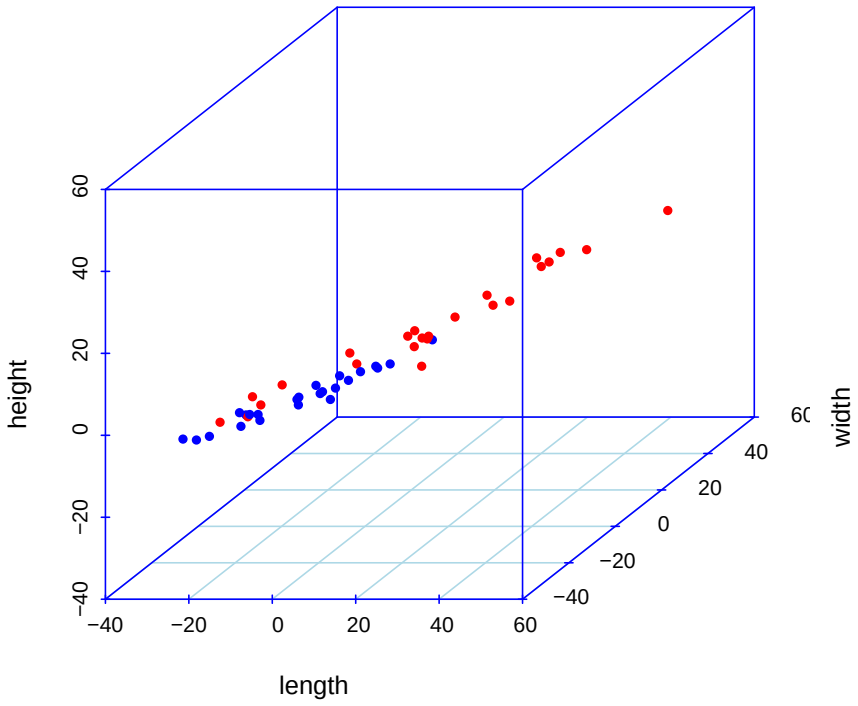


Figura 4: Rappresentazione tridimensionale delle misure del carapace delle tartarughe, in blu i maschi, in rosso le femmine, gli assi sono riportati tutti sulla stessa scala e le variabili centrate rispetto alle rispettive medie.

ora il grafico è più leggibile. Però la diversa varianza (scala) delle variabili influenza ancora molto la sua leggibilità, dobbiamo quindi dividere anche per la deviazione standard, questo al fine di ottenere delle variabili standardizzate e ben confrontabili tra loro.

```
> measures.cr <- data.frame(scale(measures))
> lims <- c(min(measures.cr), max(measures.cr))
> scatterplot3d(measures.cr, highlight.3d = F, col.axis = "blue",
+     col.grid = "lightblue", main = "", pch = 20, color = sexcol,
+     xlim = lims, ylim = lims, zlim = lims)
```

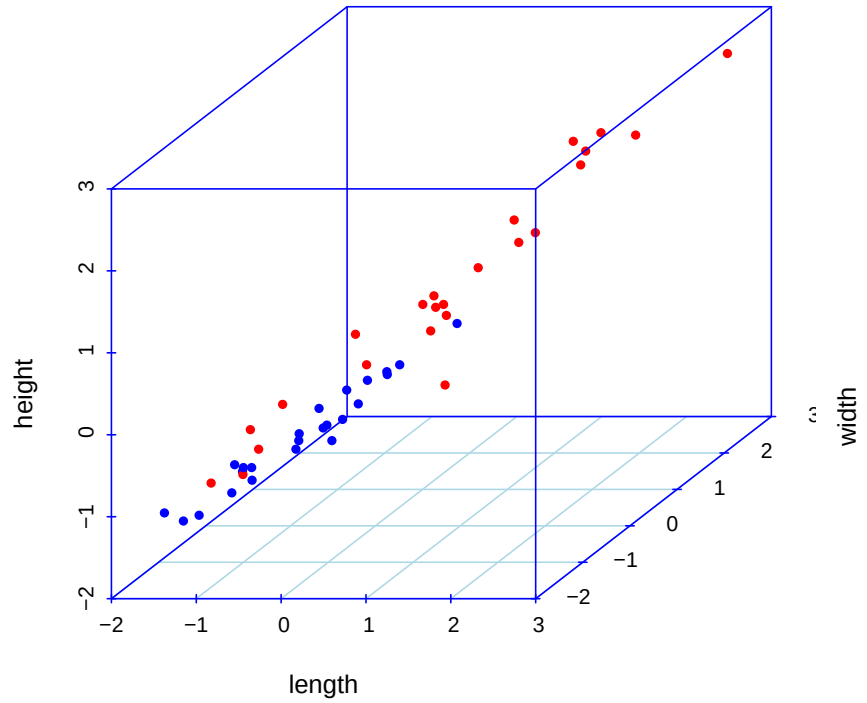


Figura 5: Rappresentazione tridimensionale delle misure del carapace delle tartarughe, in blu i maschi, in rosso le femmine, gli assi sono riportati tutti sulla stessa scala e le variabili sono standardizzate in modo da avere media 0 e varianza 1.

L'Analisi in componenti principali (PCA) cerca di costruire dei sistemi di riferimento in cui rappresentare più di tre variabili su scale confrontabili. Fa questo cercando di ridurre la dimensione dello spazio di rappresentazione, conservando al contempo la *maggior parte dell'informazione*. Per misurare l'informazione viene usata, in questo contesto, *l'inerzia della nuvola dei punti*, ovvero la sua variabilità (varianza). Come si trova questa rappresentazione? vediamo un caso semplicissimo in cui consideriamo solo due variabili, la lunghezza e l'altezza del carapace delle tartarughe. Come primo passo costruiamo la retta di regressione

tra lunghezza e altezza usando le variabili standardizzate:

```
> yy1 <- lm(length ~ height, data = measures.cr)
> summary(yy1)

Call:
lm(formula = length ~ height, data = measures.cr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.58719 -0.17823  0.01505  0.21013  0.99131

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.876e-17  4.378e-02    0.00        1
height      9.539e-01  4.424e-02   21.56   <2e-16 ***
---
Signif. codes:  0
```

l'intercetta è praticamente 0 mentre il coefficiente angolare circa 1, quindi la retta di regressione in questo caso coincide con la bisettrice. Ora, usando il pacchetto `ade4`, applichiamo la PCA a queste due variabili:

```
> pca0 = dudi.pca(cbind(measures.cr$length, measures.cr$height), scann = F,
+     scale = T)
```

riportiamo su di un grafico i valori osservati, la retta di regressione e la prima componente principale:

```
> plot(measures.cr$length, measures.cr$height, pch = 20, col = sexcol,
+     xlab = "height", ylab = "length")
> abline(h = 0)
> abline(v = 0)
> abline(coefficients(yy1), col = 2, lty = 5)
> abline(0, (pca0$c1[2, 1]/pca0$c1[1, 1]), col = 3, lty = 2)
> legend(-1.5, 2, c("retta di regressione", "primo asse fattoriale"),
+     lty = c(5, 6), col = c(2, 3), cex = 0.7)
```

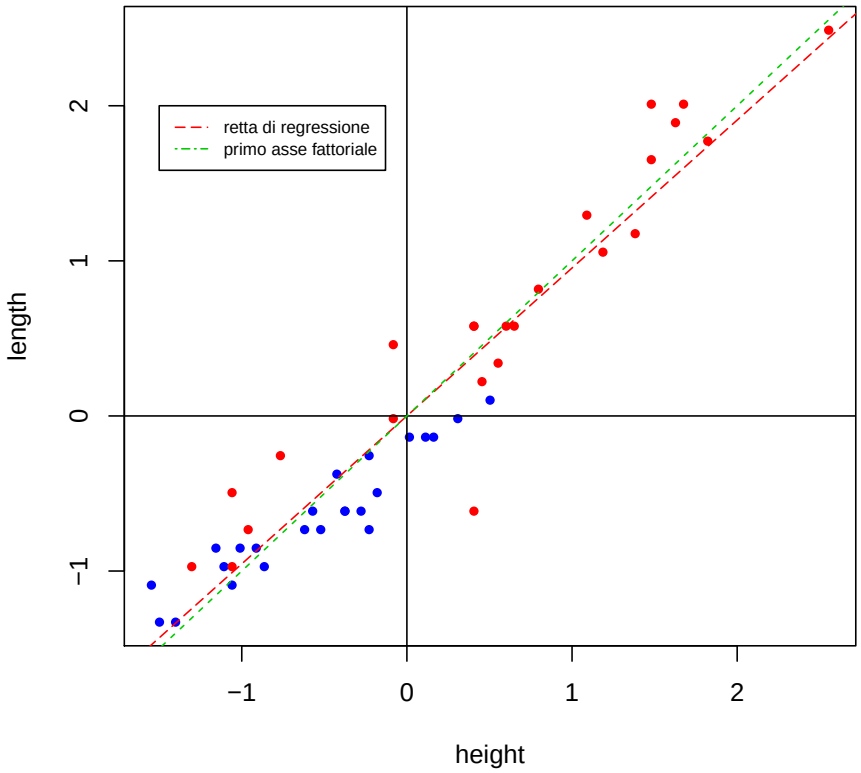


Figura 6: Retta di regressione e primo asse fattoriale per lunghezza e altezza del carapace delle tartarughe, dati standardizzati

A meno di approssimazioni numeriche, il primo asse fattoriale coincide con la retta di regressione. Se consideriamo ora la PCA di tutte e tre le variabili possiamo rappresentarle su di un piano cartesiano:

```
> pca1 = dudi.pca(measures, scann = F, scale = T)
```

L'output della pca più interessante è il cosidetto *biplot* (che verrà spiegato in dettaglio in seguito) nel quale, sul piano fattoriale (cioè il piano definito dalle componenti principali) si rappresentano insieme le variabili, disegnate come vettori, e i punti osservati:

```
> scatter(pca1)
```

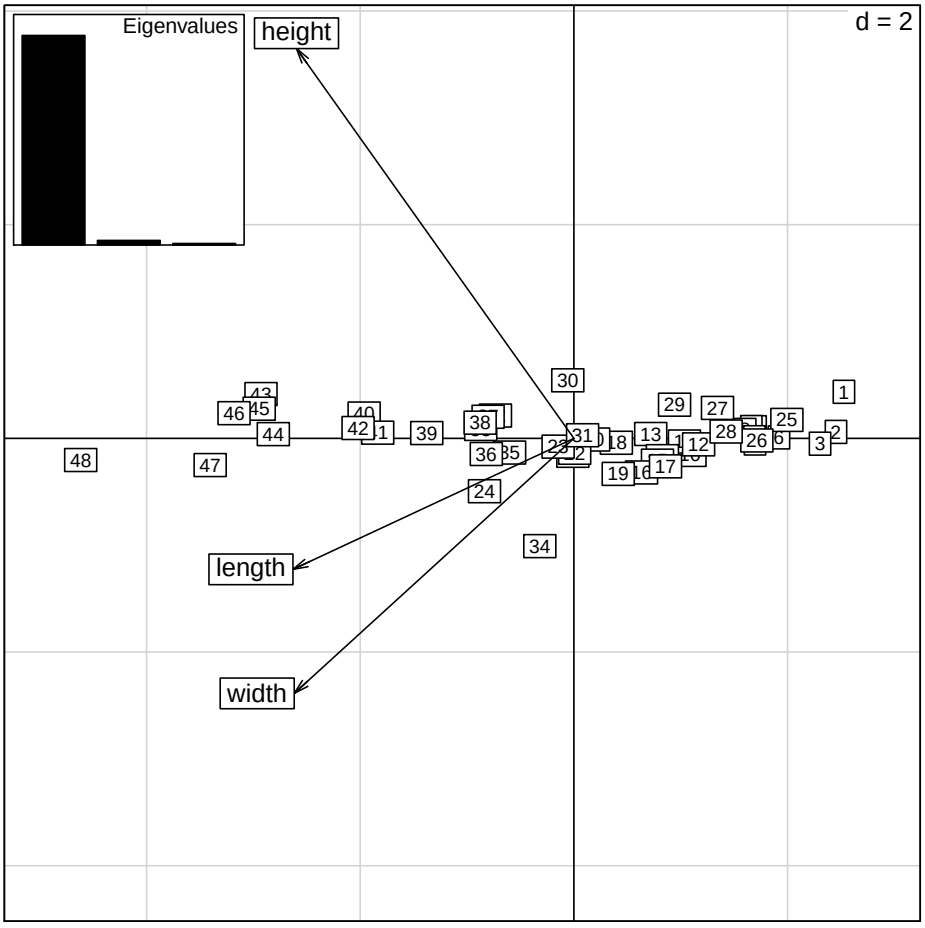


Figura 7: Biplot del primo piano fattoriale (componenti 1 e 2) per i dati delle tartarughe

la figura 7 fornisce diverse informazioni: la posizione dei vettori delle variabili rispetto agli assi permette di caratterizzare questi ultimi secondo la correlazione delle variabili con gli stessi; i quadratini rappresentano le singole osservazioni proiettate sul piano fattoriale e permette di verificare se esistono gruppi di osservazioni con comportamenti simili etc. Inoltre, il barplot in alto a sinistra mostra gli *autovalori* della matrice di correlazione, questi hanno un ruolo determinante nella PCA, infatti *la loro somma è una misura della variabilità complessiva dei dati* (*inerzia*), gli autovalori sono riportati in ordine decrescente e ciascuno è legato ad uno degli assi fattoriali, più precisamente il primo autovalore al primo asse, il secondo autovalore al secondo asse e così via. Dunque *ciascun autovalore rapresenta*

una porzione della variabilità totale. Per vedere quanta variabilità viene spiegata dalle prime due componenti, rappresentate nel grafico, calcoliamo la somma dei primi due autovalori e la rapportiamo alla somma totale:

```
> 100 * sum(pca1$eig[1:2])/sum(pca1$eig)

[1] 99.28835
```

Considerando solo 3 variabili è abbastanza naturale che la maggior parte (la quasi totalità in questo caso) della variabilità sia spiegata da due soli assi.

2 Un po' di teoria e un po' di pratica

2.1 Un esempio per cominciare

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.

```
> data(doubs)
> plot(doubs$xy, type = "l", lwd = 4, col = "blue")
> s.label(doubs$xy, add.plot = T)
```

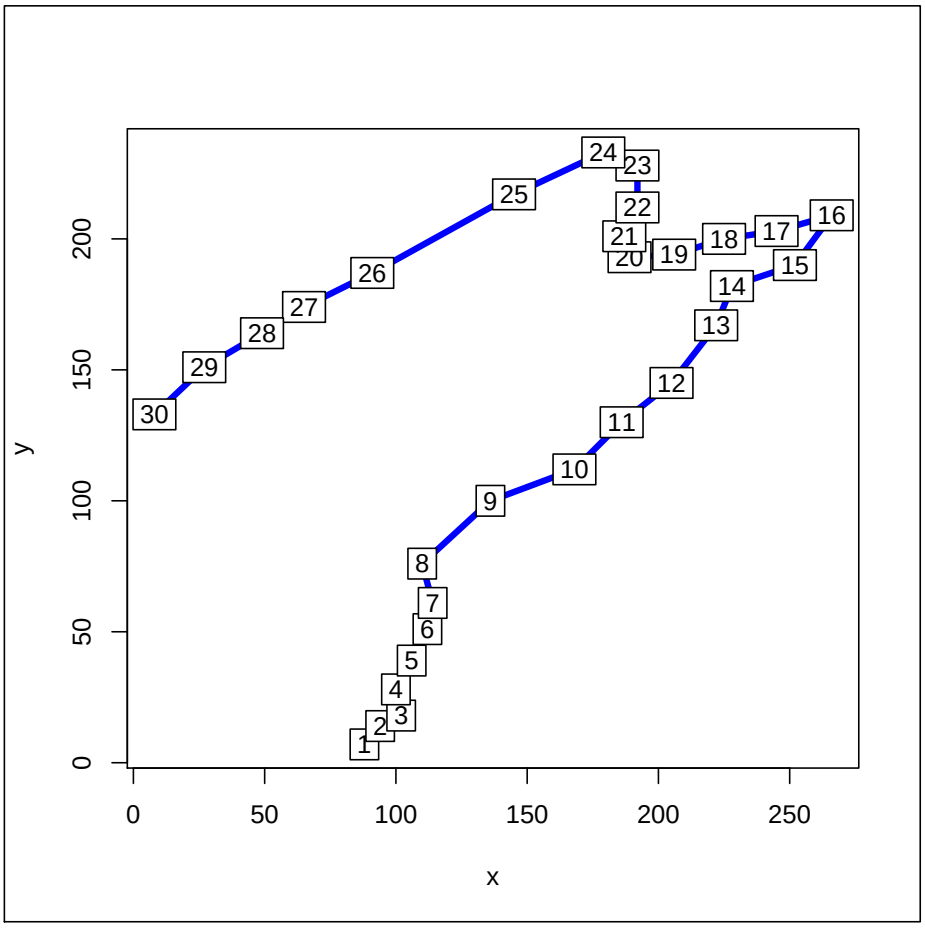


Figura 8: Dataset doubs, stazioni di rilevazione lungo il Doubs

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 posso riportare su grafico? Come posso sintetizzare l'informazione? e sopratutto **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:

```
> boxplot(doubs$mil, col = rainbow(11))
```

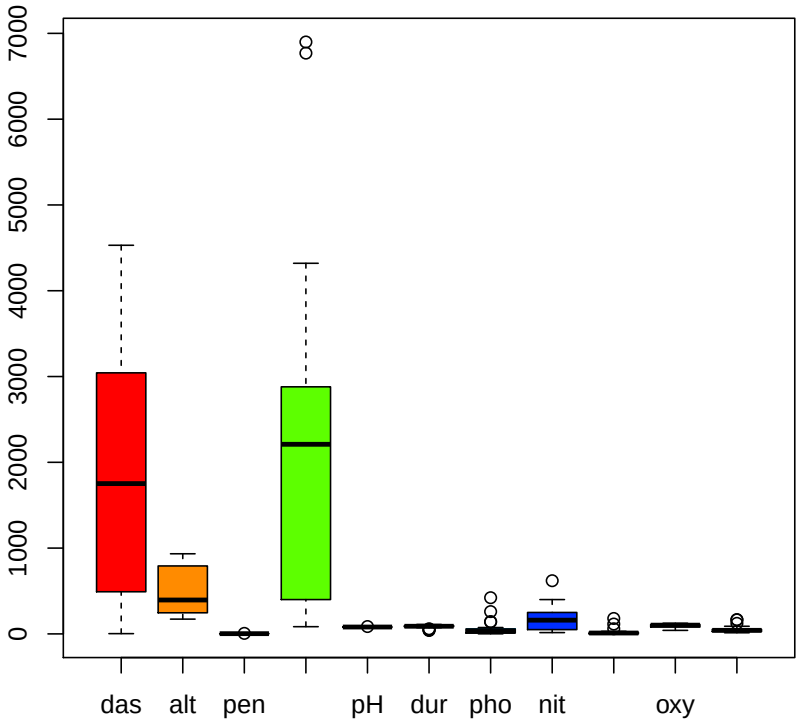


Figura 9: Dataset doubs: rappresentazione delle variabili ambientali (doubs$mil)

Dai boxplot in figura 9 vediamo che le variabili ambientali sono misurate su scale molto diverse tra loro, hanno medie molto diverse e deviazioni standard altrettanto diverse:

```
> print("medie")
```

```
[1] "medie"
```

```
> round(apply(doubs$mil, 2, mean), 3)
```

```
     das      alt      pen      deb       pH      dur      pho      nit      amm
1879.033  481.500    2.758 2220.100   80.500   86.100   55.767  165.400   20.933
     oxy      dbo
  93.900   51.167
```

```
> print("deviazioni standard")
```

```
[1] "deviazioni standard"
```

```
> round(apply(doubs$mil, 2, sd), 3)
```

```
     das      alt      pen      deb       pH      dur      pho      nit      amm
1396.508  271.387    1.080 1810.186    1.737   16.865   87.645  141.338   37.905
     oxy      dbo
  22.151   38.641
```

la funzione `apply` permette di applicare alle righe (opzione 1) o alle colonne (opzione 2) una qualsiasi funzione, nelle istruzioni riportate sopra la usiamo per calcolare le medie e le deviazioni standard delle variabili ambientali.
Per misurare l'informazione complessiva possiamo usare due strumenti:

- *La matrice delle varianze e covarianze*: questa matrice 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. Quindi sono una "buona" rappresentazione dell'informazione contenuta nei dati vista come variazione attorno alla media. l'unico problema risiede nel fatto che questi oggetti sono tra loro poco confrontabili dato che sono calcolati su scale molto diverse tra loro.

- *La matrice di correlazione*: questa matrice 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.

La matrice di correlazione è forse lo strumento più idoneo a misurare l'informazione complessiva contenuta in una tabella di dati, diamo uno sguardo alle due matrici nel caso dei dati ambientali doubs:

```
> print("matrice di varianze e covarianze")
```

```
[1] "matrice di varianze e covarianze"
```

```
> round(var(doubs$mil), 2)
```

```
           das        alt      pen        deb     pH      dur       pho       nit
das 1950234.93 -356641.84 -1140.22 2399119.86  11.47 16437.38  58492.87 147387.06
alt -356641.84   73650.88   223.94 -427037.88 -17.57 -3409.02 -10514.40 -29172.45
pen   -1140.22     223.94     1.17   -1399.72  -0.51   -11.91    -38.23    -93.28
deb 2399119.86 -427037.88 -1399.72 3276772.85  64.22 21272.33  61126.16 155318.10
pH       11.47     -17.57    -0.51      64.22   3.02     2.60    -12.67    -12.00
dur   16437.38   -3409.02   -11.91   21272.33   2.60   284.44    537.75   1217.44
pho   58492.87  -10514.40   -38.23   61126.16 -12.67   537.75   7681.56   9913.10
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
oxy  -15786.96    2175.53    11.10  -14350.33   6.81  -142.85  -1404.96  -1969.44
dbo   21354.82   -3542.88   -13.24   17710.95 -10.19   224.81   2999.01   3507.76
         amm       oxy      dbo
das 21632.80 -15786.96 21354.82
alt -3922.69   2175.53 -3542.88
pen   -14.39     11.10   -13.24
deb 20235.42 -14350.33 17710.95
pH     -8.17      6.81   -10.19
dur   185.87   -142.85   224.81
pho  3220.95  -1404.96  2999.01
nit  4273.58  -1969.44  3507.76
amm  1436.82   -605.21  1297.43
oxy  -605.21    490.64  -721.64
dbo  1297.43   -721.64  1493.11
```

```
> print("matrice di correlazione")
```

```
[1] "matrice di correlazione"
```

```
> round(cor(doubs$mil), 2)
```

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

```
> panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor) {
+     usr <- par("usr")
+     on.exit(par(usr))
+     par(usr = c(0, 1, 0, 1))
+     r <- abs(cor(x, y))
+     txt <- format(c(r, 0.123456789), digits = digits)[1]
+     txt <- paste(prefix, txt, sep = "")
+     if (missing(cex.cor))
+         cex.cor <- 0.8/strwidth(txt)
+     text(0.5, 0.5, txt, cex = cex.cor * r)
+ }
> panel.hist <- function(x, ...) {
+     usr <- par("usr")
+     on.exit(par(usr))
+     par(usr = c(usr[1:2], 0, 1.5))
+     h <- hist(x, plot = FALSE)
+     breaks <- h$breaks
+     nB <- length(breaks)
+     y <- h$counts
+     y <- y/max(y)
+     rect(breaks[-nB], 0, breaks[-1], y, col = "cyan", ...)
+ }
> pairs(doubs$mil, lower.panel = panel.cor, diag.panel = panel.hist,
+     upper.panel = panel.smooth)
```

Figura 10: Dataset doubs, variabili ambientali. Grafici a dispersione, istogrammi, correlazioni e in rosso le curve di regressione tra le coppie di variabili

Nonostante la figura 10 contenga molte informazioni, non è certo di facile lettura!

Esercizio Scrivere una pagina di commento alla figura 10.

2.2 Formalizzazione della teoria

Riassumendo un po' quel che abbiamo detto finora abbiamo

- Sia $\mathbf{X}$ 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).
- $\mathbf{X}$ 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.

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

Cosa può essere interessante analizzare?

- Quali siano le similitudini tra le variabili (p punti)
- Quali relazioni legano tra di loro gli individui o i siti osservati (n punti)
- Visualizzare i dati in unico grafico al massimo tridimensionale.
- Rispondere a queste questioni possibilmente in spazi di dimensione ridotta k ($k < p$ e $k < n$) senza perdere informazione.

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 $\mathbb{R}^n$

[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/.`

- $\mathbf{Q}$ 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 $x_{.j}$ e con $x_{i.}$ 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 $\mathbf{Q} = diag(x_{.1}, \ldots, x_{.p})$ 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 <u>incorrelate</u>. Questo significa che ogni componente contiene una parte di informazione diversa dalle altre.

2.3 Torniamo all'esempio

Per i nostri dati ambientali la cosa migliore da fare è procedere ad una PCA usando la matrice di correlazione. ***Usare questa matrice significa calcolare le distanze euclidee medie tra le variabili standardizzate, ovvero centrate rispetto alla media e divise per la deviazione standard***. Invece ***usare la matrice di covarianza implica il calcolo delle distanze euclidee medie tra le variabili la centrate solo rispetto alla media***. Vediamo praticamente che differenza c'è tra la PCA basata sulla matrice di covarianza e la stessa analisi basata sulla matrice di correlazione:

```
> pca1.cov = dudi.pca(doubs$mil, scale = F, scann = F)
> biplot(pca1.cov)
> title("covarianza")
```

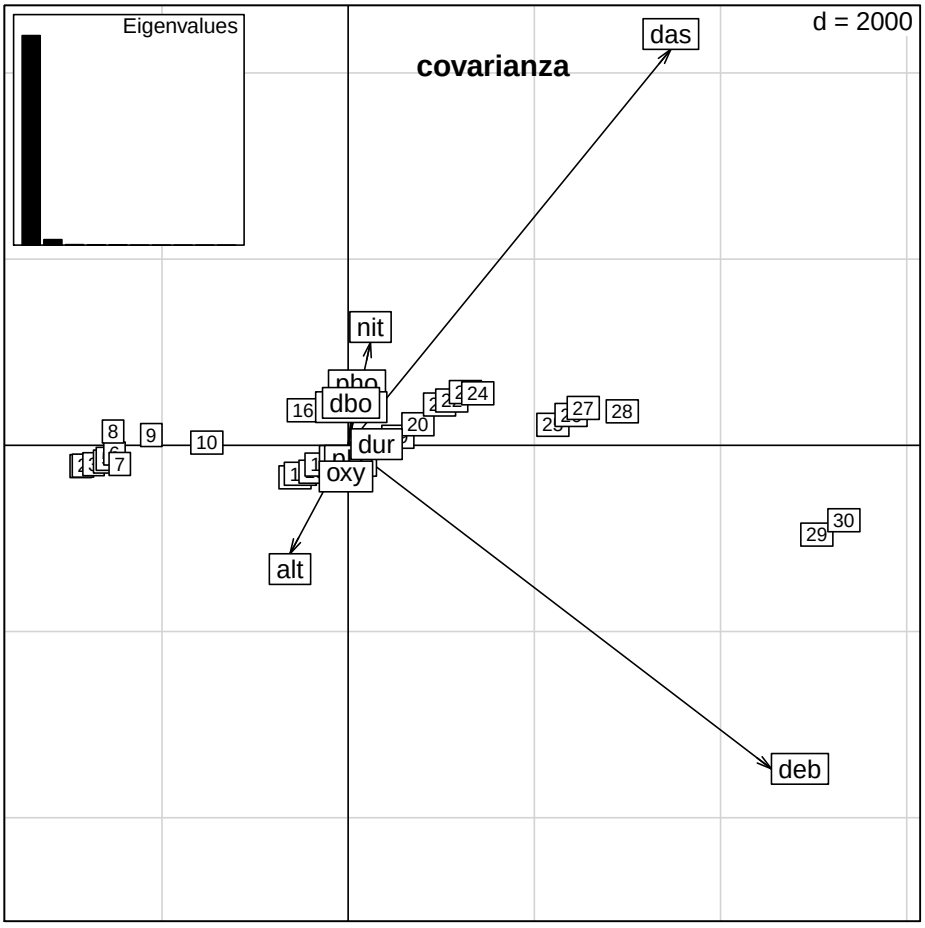


Figura 11: Dati doubs: variabili ambientali biplot dell'analisi in componenti principali basata sulla matrice di covarianza

```
> pca1.cor = dudi.pca(doubs$mil, scale = T, scann = F)
> biplot(pca1.cor)
> title("correlazione")
```

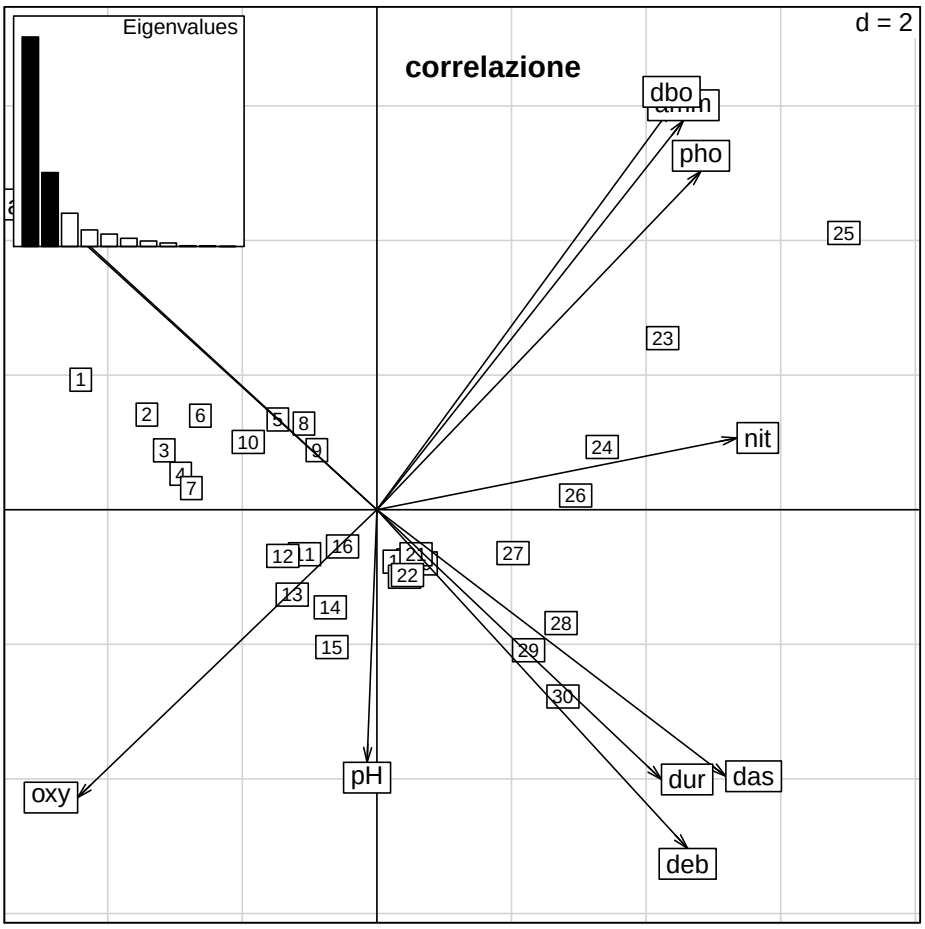


Figura 12: Dati doubs: variabili ambientali biplot dell'analisi in componenti principali basata sulla matrice di correlazione

La differenza fondamentale tra le figure 11 e 12 risiede nel fatto che nella prima il contributo delle variabili agli assi principali è oscurato dalle due variabili a varianza più elevata (das e deb). Nell'analisi basata sulla covarianza il 97% della variabilità viene rappresentato dal primo asse che è "dominato" dalla variabile das. Invece nell'analisi basata sulla correlazione distinguiamo il contributo di ciascuna variabile agli assi. I primi due assi rappresentano circa il 78% della variabilità, dalla correlazione delle variabili con gli assi (figura 13) e dalla lettura dei punteggi (tabella 1) possiamo capire come si caratterizzano gli assi:

```
> s.corcircle(pca1.cor$co)
```

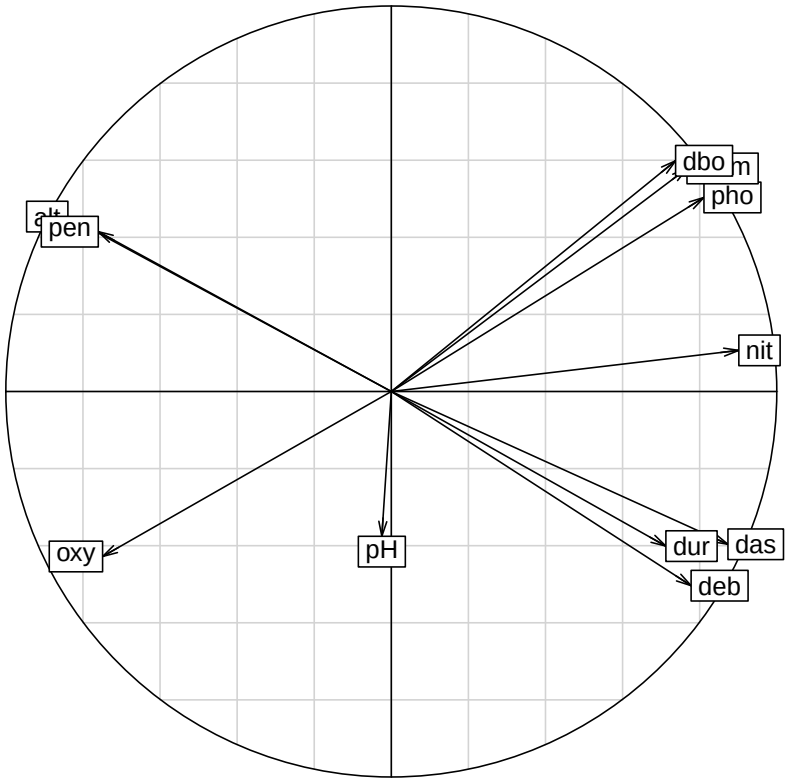


Figura 13: Cerchio di correlazione per i dati ambientali doubs con i primi due assi principali

	Comp1	Comp2
das	0.87	-0.40
alt	-0.84	0.45
pen	-0.76	0.41
deb	0.78	-0.50
pH	-0.02	-0.38
dur	0.71	-0.40
pho	0.81	0.50
nit	0.90	0.11
amm	0.77	0.58
oxy	-0.75	-0.43
dbo	0.74	0.60

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

Il primo asse è correlato positivamente (semi asse positivo) con la domanda di ossigeno, durezza dell'acqua, distanza dalla sorgente, nitrati, fosfati, ammoniaca e debito minimo medio, mentre è correlato negativamente con ossigeno disciolto, altitudine e pendenza. Il ph ha correlazione quasi nulla con il primo asse, mentre è correlato negativamente con il secondo asse. Dai punteggi osserviamo la stessa situazione con maggior dettaglio. Queste considerazioni ci permettono di dire quali variabili danno il contributo maggiore alla costruzione di ciascun asse, i nitrati per il primo asse, la domanda di ossigeno per il secondo e così via. Nel biplot in figura 12 vediamo anche dove si collocano le unità, ovvero i siti, rispetto ai due assi. Questo permette di caratterizzare questi ultimi sulla base delle considerazioni fatte sugli assi. Ad esempio, il sito 1 ha pendenza molto elevata, molto ossigeno disciolto, niente ammoniaca, acqua non molto dura etc. la sua collocazione sul piano riflette questa situazione.

2.4 Un breve approfondimento teorico

- Per ottenere la rappresentazione di $\mathbf{X}$ in uno spazio di dimensione ridotta dobbiamo ottenere gli *autovalori* ed *autovettori* della matrice

$$\mathbf{XQX}^T\mathbf{D}$$

  oppure di

$$\mathbf{X}^T\mathbf{DXQ}$$

- Si dimostra che **gli autovalori non nulli di queste due matrici sono uguali**
- Indicheremo con $\mathbf{\Lambda}_{[r]}$ la matrice diagonale degli autovalori non nulli
- il numero r di autovalori non nulli è detto *rango* ed è tale che $r \leq \min(n, p)$
- Esistono inoltre molte simmetrie tra l'analisi per righe e l'analisi per colonne.

Definiamo

- $\mathbf{F}$: $\mathbf{F}^T\mathbf{Q}^{-1}\mathbf{F} = \mathbf{I}_r$ dove $\mathbf{Q}^{-1}$ è l'inversa di $\mathbf{Q}$. Queste r colnne definiscono i ***fattori principali***
- $\mathbf{A}$: $\mathbf{A}^T\mathbf{Q}\mathbf{A} = \mathbf{I}_r$. Le r colonne di $\mathbf{A}$ definiscono gli ***assi principali***.
- $\mathbf{K} : \mathbf{K}^T\mathbf{D}\mathbf{K} = \mathbf{I}_r$ le cui r colonne definiscono le ***componenti principali***.
- $\mathbf{G}$: $\mathbf{G}^T\mathbf{D}^{-1}\mathbf{G} = \mathbf{I}_r$ dove $\mathbf{D}^{-1}$ è l'inversa di $\mathbf{D}$. Le r colonne di $\mathbf{G}$ contengono i cosidetti ***cofattori principali.***
- queste grandezze sono tra loro legate tramite:

$$\mathbf{F} = \mathbf{Q}\mathbf{A}, \mathbf{K} = \mathbf{X}\mathbf{F}\mathbf{\Lambda}_{[r]}^{-(1/2)}$$

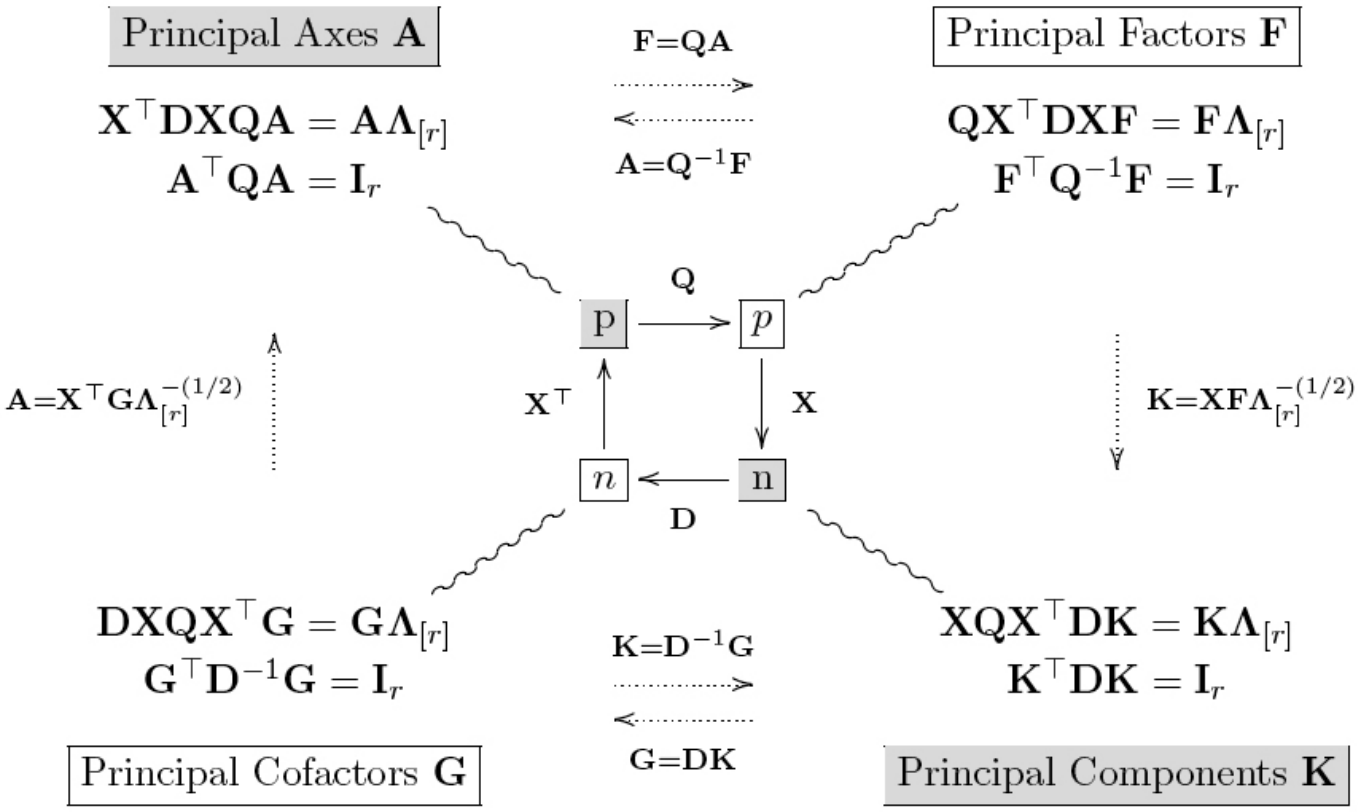
$$\mathbf{G} = \mathbf{D}\mathbf{K}, \mathbf{A} = \mathbf{X}^T\mathbf{G}\mathbf{\Lambda}_{[r]}^{-(1/2)}$$

- In pratica tutte le grandezza ora definite vengono ottenute cercando di conservare quanta più informazione possibile.
- Nello spazio euclideo un modo per misurare l'*informazione* è la distanza (o norma) da un punto di riferimento.
- Ad esempio: la varianza è la distanza (media) dei valori osservati dalla loro media e misura quanto questi siano dispersi attorno ad essa.
- Le grandezze che cerchiamo vengono calcolate in modo *ottimale* rispetto ad un qualche *criterio* che si basi su delle *misure di informazione*.
- Tutte vengono ottenute massimizzando espressioni *quadratiche*, ad esempio il primo degli assi principali si ottiene cercando il vettore $\mathbf{a}_1$ tale che renda massimo il valore di

$$\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{a}$$

.

Lo schema duale:


Principal Axes A
F=QA
Principal Factors F
X⊤DXQA = AΛ[r]
A⊤QA = Ir
A=Q−1F
QX⊤DXF = FΛ[r]
F⊤Q−1F = Ir
Q
p
p
A=X⊤GΛ−(1/2)[r]
X⊤
X
K=XFΛ−(1/2)[r]
n
n
D
DXQX⊤G = GΛ[r]
G⊤D−1G = Ir
K=D−1G
G=DK
XQX⊤DK = KΛ[r]
K⊤DK = Ir
Principal Cofactors G
Principal Components K