Biometry. Lecture 26

Alexey Shipunov

Minot State University

April 25, 2012

Outline

1. Questions and answers
2. Multivariate statistics, or Data Mining
    - Generic methods
    - Principal Component Analysis (PCA)
    - Correspondence analysis

Outline

1. Questions and answers
2. Multivariate statistics, or Data Mining
   - Generic methods
   - Principal Component Analysis (PCA)
   - Correspondence analysis

Starting...

```
> setwd("<working folder>")
```

or

“`Change dir`”

in menu!

Previous final question: the answer

What are the null and alternative hypotheses for ANOVA?

Previous final question: the answer

What are the null and alternative hypotheses for ANOVA?

- Null: no one is different from all others
- Alternative: at least one is different from all others

Are people with different types of hair color differ also by their height?

```
> hwc <- read.table("http://ashipunov.info/data/hwc.txt",
+ h=T)
> str(hwc)
> boxplot(HEIGHT ~ COLOR, data=hwc)
> anova(lm(HEIGHT ~ COLOR, data=hwc))
> h.c <- aov(lm(HEIGHT ~ COLOR, data=hwc))
> (h.c.hsd <- TukeyHSD(h.c))
> plot(w.c.hsd)
```

Multivariate statistics, or Data Mining

Generic methods

Multivariate statistics, or Data Mining

- Multivariate statistics is mostly the **analysis of structure**
- Majority of multivariate methods are just visualization via reduction of dimensions (projection)
- Inferential methods also exist

Generic methods

- These methods simply take into account more than two variables
- Conditional (trellis) plots and 3D cubes and surfaces are most common representatives of this group in R

Trellis plots

From now on, we will frequently use embedded "`iris`" data from Fisher.

```
> coplot(Sepal.Length ~ Petal.Length | Species, data=iris)
> library(lattice)
> xyplot(Sepal.Length ~ Petal.Length + Petal.Width
+ | Species, data=iris)
```

Matrix graph

```
> pairs(iris[1:4], pch=21, bg=1:3[iris$Species])
```

Pictographs

```
> stars(mtcars[1:9,1:7])
# install.packages("TeachingDemos") if you do not have it
> library(TeachingDemos)
> faces(mtcars[1:9,1:7])
```

`mtcars` is a data about different design and performance of 32 cars (1973–74 models) like: (1) mpg, (2) Number of cylinders, (3) Displacement, (4) Gross horsepower, (5) Rear axle ratio, (6) Weight (lb/1000) (7) qsec 1/4 mile time.

Parallel coordinates plot

```
> eq8 <- read.table("http://ashipunov.info/data/eq8.txt",
+ h=T)
> library(MASS)
> parcoord(eq8[,-1], col=rep(rainbow(8), table(eq8[,1])))
> legend("top", names(table(eq8[,1])), fill=rainbow(8),
+ ncol=4)
```

Multivariate statistics, or Data Mining

Principal Component Analysis (PCA)

Principal Component Analysis

- Principal Component Analysis tries to achieve the best projection of multivariate cloud, taking into account as many characters (dimensions) as possible
- All characters are transformed into components; first component is the most important, second and third are also significant

PCA for `iris` data

```
> iris.pca <- princomp(scale(iris[,1:4]))
> plot(iris.pca, main="") # this is technical screeplot
> iris.p <- predict(iris.pca)
> plot(iris.p[,1:2], type="n", xlab="PC1", ylab="PC2")
> text(iris.p[,1:2], labels=abbreviate(iris[,5], 1,
+ method="both"))
> loadings(iris.pca)
```

Inferential PCA (library `ade4`)

```
> library(ade4)
> iris.d <- dudi.pca(iris[,1:4], scannf=FALSE)
> s.class(iris.d$li, iris[,5])
> randtest(bca(iris.d, iris[,5], scannf=FALSE))
```

Multivariate statistics, or Data Mining

Correspondence analysis

Correspondence analysis

- You may think about PCA as a multivariate derivative of correlation analysis, and correspondence analysis may be imagined as a derivative of contingency tables analysis
- Unique feature of correspondence analysis is an ability to show **both** rows and columns on one graph

Simple example of correspondence visualization

```
> library(MASS)
> caith
> biplot(corresp(caith, nf=2))
```

`caith` is the embedded data on the cross-classification of people in Caithness, Scotland, by eye and hair colour

Library `vegan` contains more advanced methods and graphs, represented in particular by functions `cca()` and `decorana()`

Finishing...

```
> savehistory("20120427.r")
```

Final question (2 points)

Final question (2 points)

Why biologists need multivariate analysis?

Summary: most important commands

- `coplot()`—make a conditional plot
- `parcoord()`—make a parallel coordinates plot
- `princomp()`—make PCA

For Further Reading

📄
A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access: `http://ashipunov.info/shipunov/school/biol_299`