Biometry. Lecture 22

Alexey Shipunov

Minot State University

April 27, 2015

Outline

1. Two-dimensional statistics
   - ANalysis Of VAriation (ANOVA)
2. Multivariate statistics, or Data Mining
   - Generic methods
   - Principal Component Analysis (PCA)
   - Correspondence analysis
   - Similarity
   - Multi-dimensional scaling
   - Cluster analysis
   - Classification (machine learning)
   - Linear Discriminant Analysis (LDA)
   - Regression trees (recursive partitioning)
   - Advanced methods of classification

Outline

1. Two-dimensional statistics
    - ANalysis Of VAriation (ANOVA)
2. Multivariate statistics, or Data Mining
    - Generic methods
    - Principal Component Analysis (PCA)
    - Correspondence analysis
    - Similarity
    - Multi-dimensional scaling
    - Cluster analysis
    - Classification (machine learning)
    - Linear Discriminant Analysis (LDA)
    - Regression trees (recursive partitioning)
    - Advanced methods of classification

Starting...

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

or

“`Change dir`”

in menu!

On Mac, be sure that startup option is working: `getwd()` (`getwd()` checks if R is in working folder, `dir()` checks the folder content)

Two-dimensional statistics

ANalysis Of VAriation (ANOVA)

Post-hoc tests and plots

```
> pairwise.t.test(hwc$HEIGHT, hwc$COLOR)
> pairwise.wilcox.test(hwc$HEIGHT, hwc$COLOR)
> w.c <- aov(lm(WEIGHT ~ COLOR, data=hwc))
> (w.c.hsd <- TukeyHSD(w.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=as.numeric(iris[,5]))
```

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.

RGL: real 3D plots

```
> library(rgl)
> plot3d(iris[,1:3], col=as.numeric(iris[,5]))
```

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

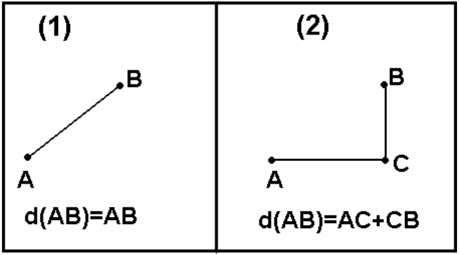
Library `vegan` contains more advanced methods and graphs, represented in particular by functions `cca()` and `decorana()`

Multivariate statistics, or Data Mining

Similarity

Distance and similarity

- Distance is simple a numerical measure of similarity
- Euclidean distance is (1) hypothenuse; manhattan distance (2) is a sum of legs



Closeness and distance

```
> ma <- data.frame(V1=c(7,3,5), V2=c(7,5,3))
> row.names(ma) <- c("A","B","C")
> dist(ma) # Euclidean is default
> dist(ma, method="manhattan")
> iris.d <- dist(iris[,1:4])
> library(cluster)
> iris.dist <- daisy(iris[,1:4], metric="manhattan")
```

`daisy()` function is more universal since it can work with both binary and measurement variables.

Multivariate statistics, or Data Mining

Multi-dimensional scaling

Multi-dimensional scaling

- Multi-dimensional scaling may be seen as making a geographic map from all pairs of distances
- Results are often similar to PCA but axes are not connected with any particular character

Scaling examples

```
> example(cmdscale)
> eurodist
> iris.c <- cmdscale(iris.dist)
> plot(iris.c[,1:2], type="n", xlab="Dim. 1",
+ ylab="Dim. 2")
> text(iris.c[,1:2], labels=abbreviate(iris[,5], 1,
+ method="both.sides"))
```

Multivariate statistics, or Data Mining

Cluster analysis

Cluster analysis

- Clusterization is the making groups
- Hierarchical clusterization makes trees (dendrograms)

Hierarchical clustering

```
> plot(hclust(dist(ma)))
# We will choose every fifth row
> iriss <- iris[seq(1,nrow(iris), 5),]
> iriss.dist <- daisy(iriss[, 1:4])
> iriss.h <- hclust(iriss.dist, method="ward")
> plot(iriss.h, labels=abbreviate(iriss[,5], 1,
+ method="both.sides"))
```

Support for branches

```
> library(pvclust)
> irisst <- t(iriss[, 1:4])
> colnames(irisst) <- paste(abbreviate(iriss[,5], 3),
+ colnames(irisst))
> iriss.pv <- pvclust(irisst, method.dist="manhattan",
+ method.hclust="ward", nboot=100)
> plot(iriss.pv, col.pv=c(1, 0, 0))
```

Another hierarchical clustering example (very simple)

```
> fences <- read.table(
+ "http://ashipunov.info/data/fences.txt", h=T)
> library(cluster)
> str(fences)
> fences.d <- daisy(fences)
> summary(fences.d)
> plot(hclust(fences.d))
```

Fuzzy clustering

```
> iris.f <- fanny(iris[,1:4], 3)
> plot(iris.f, which=1, main="")
> head(data.frame(sp=iris[,5], iris.f$membership))
```

Multivariate statistics, or Data Mining

Classification (machine learning)

Classification (machine learning)

- Machine learning, or classification is alway based on the example where objects already distributed into groups
- These methods are trying to find a best classification algorithm

Multivariate statistics, or Data Mining

Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA)

- Linear discriminant analysis is based on the idea that classification could be made on a bases of linear equations
- This is a parametric method

LDA example

```
> library(MASS)
> iris.train <- iris[seq(1,nrow(iris),5),]
> iris.unknown <- iris[-seq(1,nrow(iris),5),]
> iris.lda <- lda(Species ~ . , data=iris.train)
> iris.ldap <- predict(iris.lda, iris.unknown[,1:4])$class
> table(iris.ldap, iris.unknown[,5])
```

LDA testing

```
> ldam <- manova(as.matrix(iris.unknown[,1:4]) ~
+ iris.ldap)
> summary(ldam, test="Wilks")
```

“Wilks” value is not only a statistic, it is also a likelihood ratio: for better classifications, Wilks is closer to 0

LDA visualization

```
> iris.lda2 <- lda(scale(iris[,1:4]), iris[,5])
> iris.ldap2 <- predict(iris.lda2, dimen=2)$x
> plot(iris.ldap2, type="n", xlab="LD1", ylab="LD2")
> text(iris.ldap2, labels=abbreviate(iris[,5], 1,
+ method="both.sides"))
```

Multivariate statistics, or Data Mining

Regression trees (recursive partitioning)

Regression trees (recursive partitioning)

- Regression trees, or recursive partitioning are based on the same idea as biological descriptive keys
- On each step, methods searches for the best separation between members of group

Regression tree example I

```
> library(tree)
> iris.tree <- tree(Species ~ ., data=iris)
> plot(iris.tree); text(iris.tree)
```

Regression tree example II

```
> eq <- read.table("http://ashipunov.info/data/eq.txt", h=TRUE)
> eq.tree <- tree(SPECIES ~ ., data=eq)
> plot(eq.tree); text(eq.tree)
```

Multivariate statistics, or Data Mining

Advanced methods of classification

Advanced methods of classification

- “Random Forest” is based on the construction of multiple regression trees
- “Support Vector Machines” try to find a hyperplane which separates objects best

Random Forest example

```
> library(randomForest)
> set.seed(17)
> iris.rf <- randomForest(Species ~ ., data=iris.train)
> iris.rfp <- predict(iris.rf, iris.unknown[,1:4])
> table(iris.rfp, iris.unknown[,5])
```

Random Forest visualization

```
> set.seed(17)
> iris.urf <- randomForest(iris[,1:4])
> MDSplot(iris.urf, iris[,5])
```

SVM example

```
> library(e1071)
> iris.svm <- svm(Species ~ ., data=iris.train)
> iris.svmp <- predict(iris.svm, iris.unknown[,1:4])
> table(iris.svmp, iris.unknown[,5])
```

Finishing...

Save your commands!

(`savehistory(<todaysdate>.r)` or `File -> Save as...` on Mac)

Summary: most important commands

- `coplot()`—make a conditional plot
- `parcoord()`—make a parallel coordinates plot
- `princomp()`—make PCA
- `dist()`—calculates distance
- `hclust()`—performs hierarchical clusterization
- `lda()`—preforms linear discriminate analysis (LDA)
- `tree()`—performs recursive partitioning

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access:
`http://ashipunov.info/shipunov/school/biol_240`

A. Shipunov, and others.
Visual statistics. Use R!
Ongoing translation from Russian.

Short anonymous absolutely voluntary survey

1. What do you **like** most in biometrics course?
2. What do you **dislike** most in biometrics course?
3. **Which lab** do you remember most of all?
4. Please grade (1—bad, 5—excellent):
    1. Lectures
    2. Labs
    3. Final questions
    4. Exams
5. Please recommend something for the Spring 2016 Biometrics.