Biometry. Lecture 27

Alexey Shipunov

Minot State University

May 2, 2012

Outline

Outline

Minot State UNIVERSITY

Starting...

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

or

“Change dir”

in menu!

Previous final question: the answer

Why biologists need multivariate analysis?

Previous final question: the answer

Why biologists need multivariate analysis?

- To plan research
- To find order and structure
- To assemble large amounts of data

Multivariate statistics, or Data Mining

Principal Component Analysis (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

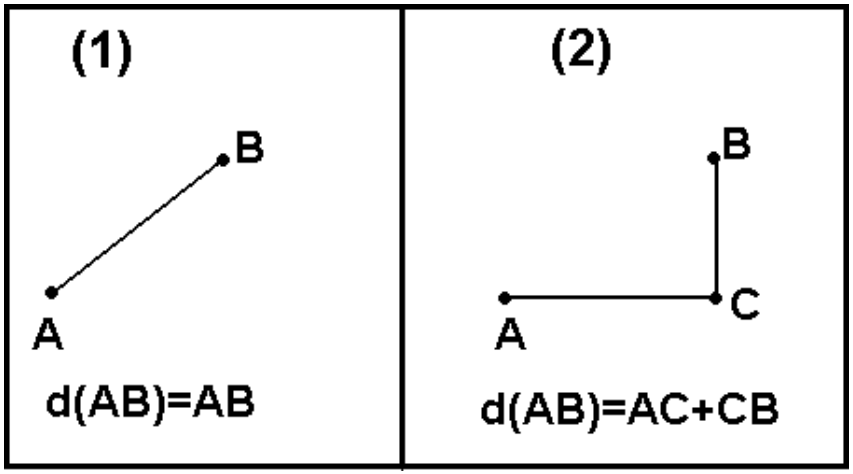
Library `vegan` contains more advanced methods and graphs, represented in particular by functions `cca()` and `decorana()`

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))
```

Finishing...

```
> savehistory("20120502.r")
```

Final question (2 points)

Final question (2 points)

What is a difference between euclidean and manhattan distances?

Summary: most important commands

- `dist()`—calculates distance
- `hclust()`—performs hierarchical clusterization

For Further Reading

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