Multivariate Statistics in Ecology and Quantitative Genetics
Correspondence analysis

Dirk Metzler & Martin Hutzenthaler

http://evol.bio.lmu.de/_statgen

Summer semester 2012

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

Contents

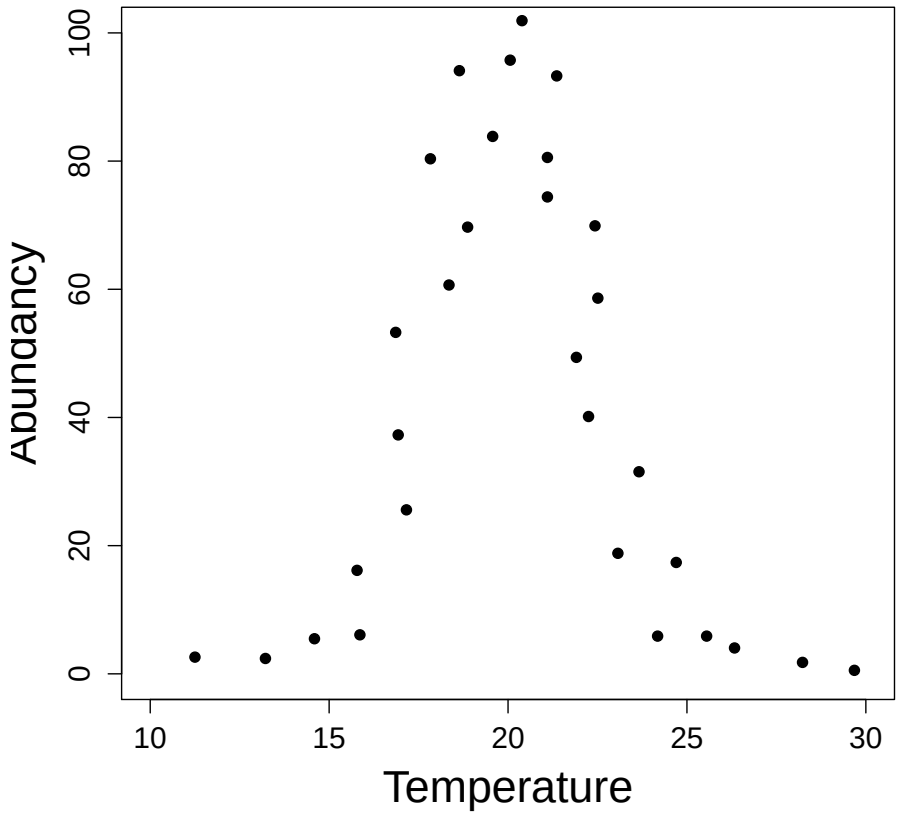
1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

The dependence of species on environmental variables is often not linear
(often not even increasing/decreasing)

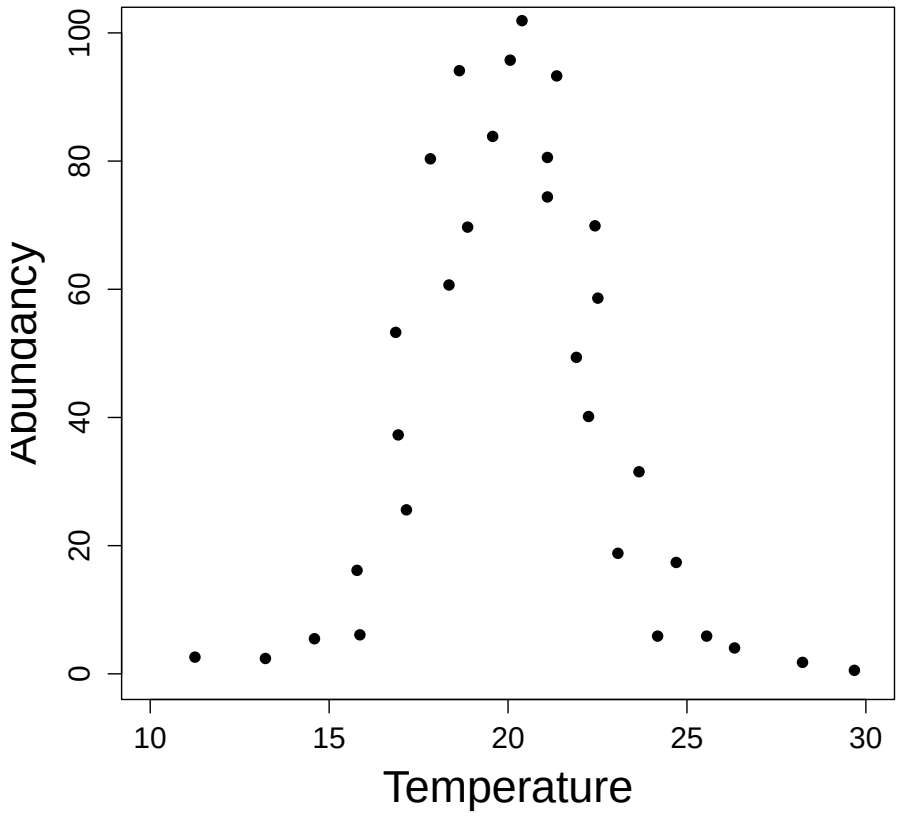
The dependence of species on environmental variables
is often not linear
(often not even increasing/decreasing)

Example:
The reproduction rate of bacteria depends on temperature.
Low and high temperatures are not good or even lethal.

The following figure shows an artificial example of abundancies of some species along the environmental variable 'temperature'.



The following figure shows an artificial example of abundancies of some species along the environmental variable 'temperature'.



In this artificial example, the niche around 20° is prefered.

A good and simple model for niches is the so-called Gaussian response model

$$Z_i = C \exp\left(-\frac{(X_i - \mu)^2}{2t^2}\right)$$

where C, μ, t are parameters.

A good and simple model for niches is the so-called Gaussian response model

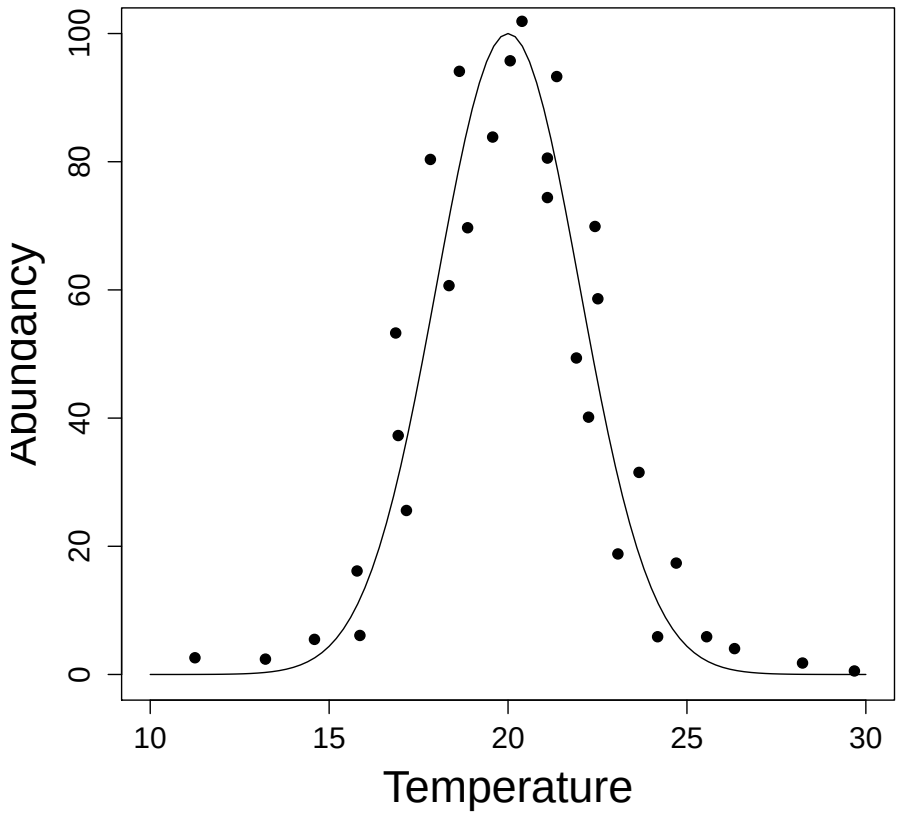
$$Z_i = C \exp\left(-\frac{(X_i - \mu)^2}{2t^2}\right)$$

where C, μ, t are parameters.

This model is

- simple: only 3 parameters
- good: the Gaussian response model is chosen (among all unimodal distributions) because of the normal approximation (central limit theorem) which loosly speaking says that if many independent (or at most weakly dependent) effects contribute to a quantity and no effect is dominating, then the quantity is distributed normally.

Fitted Gaussian response model



Optimum $\mu = 20$, maximum $C = 100$, tolerance $t = 2$.

Contents

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

Correspondence analysis

Given: Data frame/matrix Y

$Y[i, \cdot]$ are the observations of site `i`

$Y[\cdot, j]$ are the observations at species `j`

Correspondence analysis

Given: Data frame/matrix Y
$Y[i, \cdot]$ are the observations of site `i`
$Y[\cdot, j]$ are the observations at species `j`

Goal: Find associations of species and sites

Correspondence analysis

Given: Data frame/matrix Y
$Y[i, \cdot]$ are the observations of site `i`
$Y[\cdot, j]$ are the observations at species `j`

Goal: Find associations of species and sites

Assumption: There is a niche dependence of the species on the environmental variables

Correspondence analysis

Given: Data frame/matrix Y
$Y[i, \cdot]$ are the observations of site `i`
$Y[\cdot, j]$ are the observations at species `j`

Goal: Find associations of species and sites

Assumption: There is a niche dependence of the species on the environmental variables

The setting is formulated here in terms of species and sites.
If you have measured quantities (variables) of some objects, then replace 'site' by 'object' and 'species' by 'variable'.

Contents

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

Let Y have M rows and N columns.
The Gaussian regression from the motivation subsection would lead to finding species scores $(u_1, u_2, \cdots, u_N)$ and site scores $(l_1, l_2, \cdots, l_M)$ such that

$$Y[i,k] \approx C_k \exp\left(-\frac{(l_i - u_k)}{2t_k^2}\right)$$

In fact, correspondence analysis does not use this approach but a weighted PCA approach which results in a similar representation.

The approach of correspondence analysis is based on the Chi-square statistic which is used for testing the nullhypothesis that the species do not depend on the sites. In that case we would have

$$\frac{Y[i,k]}{n} \approx \frac{Y[i,+]}{n} \cdot \frac{Y[+,k]}{n}$$

where

- $Y[i,+] = \sum_k Y[i,k]$ is the row sum,
- $Y[+,k] = \sum_i Y[i,k]$ is the column sum and
- $n = Y[+,+] = \sum_i \sum_k Y[i,k]$ is the total sum.

The approach of correspondence analysis is based on the Chi-square statistic which is used for testing the nullhypothesis that the species do not depend on the sites. In that case we would have

$$\frac{Y[i,k]}{n} \approx \frac{Y[i,+]}{n} \cdot \frac{Y[+,k]}{n}$$

where

- $Y[i,+] = \sum_k Y[i,k]$ is the row sum,
- $Y[+,k] = \sum_i Y[i,k]$ is the column sum and
- $n = Y[+,+] = \sum_i \sum_k Y[i,k]$ is the total sum.

The Chi-square test statistic is given by

$$X^2 = \sum\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

$$= \sum_i \sum_k \frac{\left(Y[i,k]/n - Y[i,+]Y[+,k]/n^2\right)^2}{Y[i,+]Y[+,k]/n^2}$$

Instead of frequencies we now consider probabilities

$$p[i,k] := Y[i,k]/n$$

and define a matrix Q with entries

$$Q[i,k] := \frac{p[i,k] - p[i,+] \cdot p[+,k]}{\sqrt{p[i,+]p[+,k]}}$$

Now all further steps are just as in PCA with the centred/normalized matrix Y replaced by the association matrix Q. Again we get a distance biplot and a correlation biplot.

Instead of frequencies we now consider probabilities

$$p[i,k] := Y[i,k]/n$$

and define a matrix Q with entries

$$Q[i,k] := \frac{p[i,k] - p[i,+] \cdot p[+,k]}{\sqrt{p[i,+]p[+,k]}}$$

Now all further steps are just as in PCA with the centred/normalized matrix Y replaced by the association matrix Q. Again we get a distance biplot and a correlation biplot.

Correspondence analysis assesses
the association between species and sites
(or objects and variables)

Contents

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

- Is there a difference in species community amoung four groups of pastures?
- Data of vegetation on pastures in Mexico from the dry season of 1992.
- see also

Zuur, Ieno, Smith (2007) *Analysing Ecological Data.* Springer

Meaning of the Variables

Variable names with small letters are shortcuts for plant families.

ALTITUDE Altitude above sea level in meters

FIELDSLOPE Field slope in degrees

AGE Time since forest clearing; index from 1-8 representing ages from 6 to 40 years

CATTLEINTENSITY number of cattle per hectar

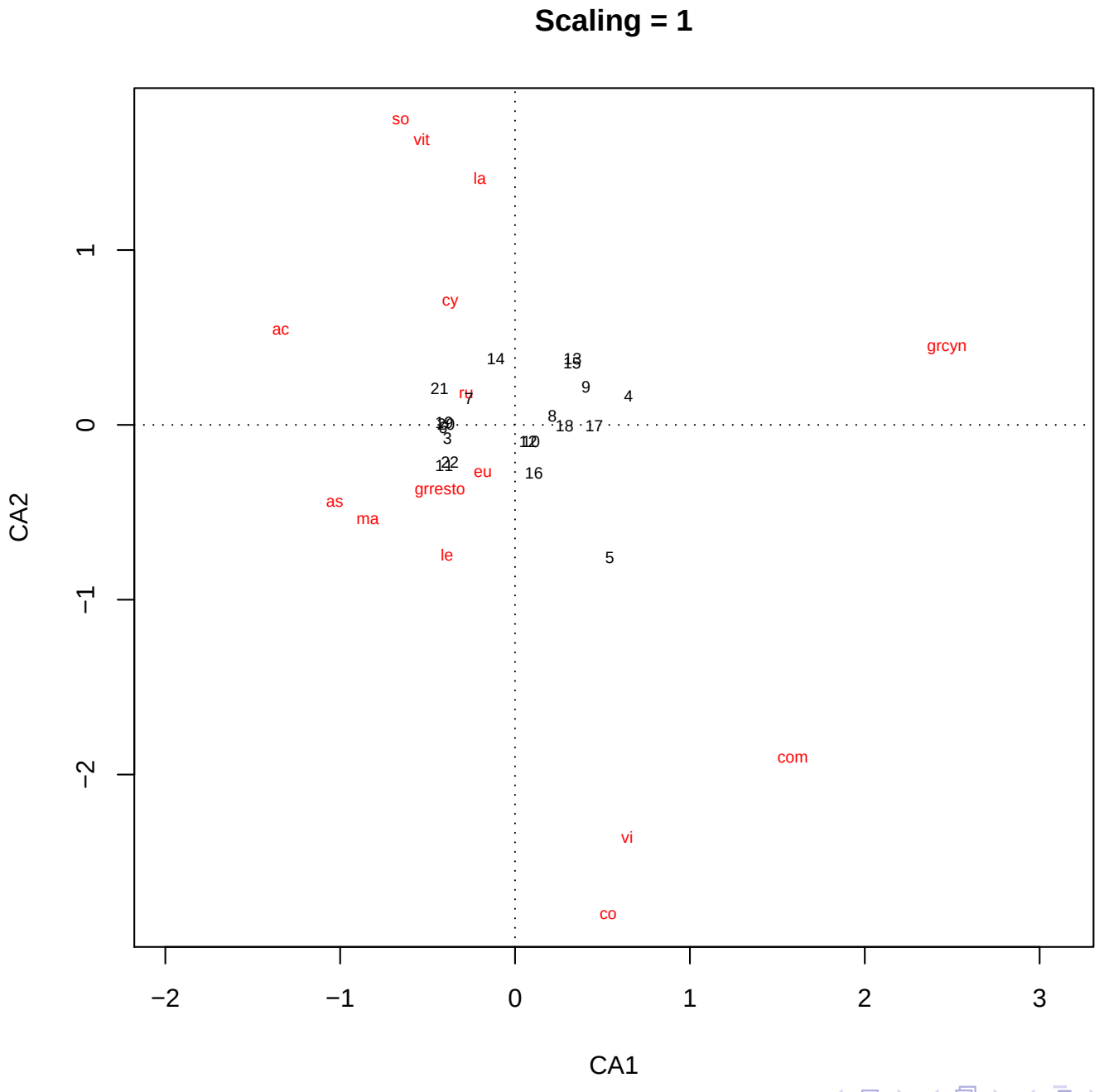
PLAGUE Nominal: 0=no plague, 1=plague (herbivore insects) in the year before sampling

MAXVEGHEIGHT in centimeter

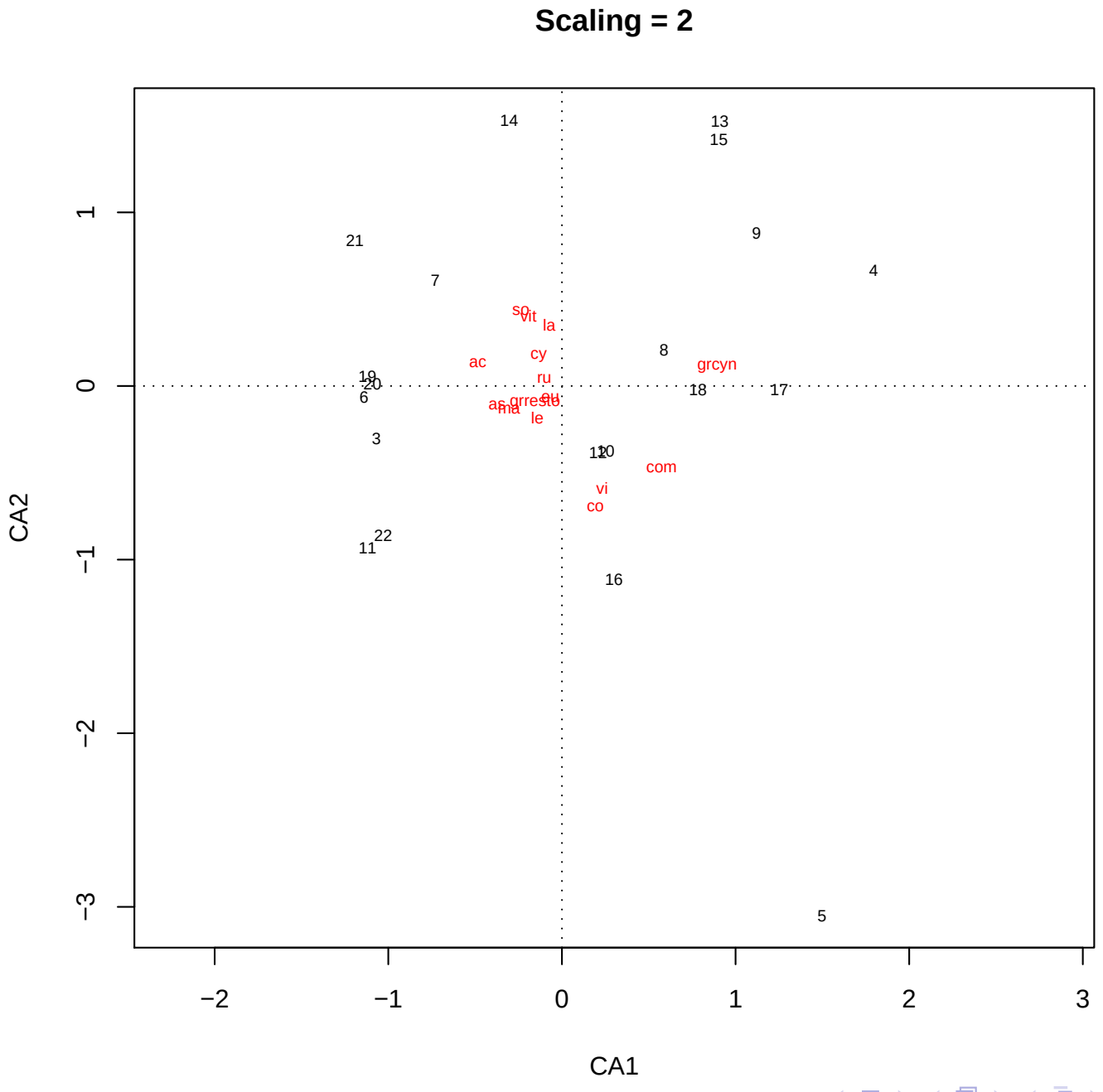
BLOCK Nominal: 1=grama pastures in Balzapote, 2=star pastures in Balzapote, 3=grama pastures in La Palma, 2=star pastures in La Palma

(many more variables in original data set)

```
mplants<-read.table("MexicanPlants.txt",h=T,sep="\t")
species<-mplants[,c("ac","as","co","com","cy","eu",
 "grcyn","grresto","la","le","ma","ru","so","vi","vit")]
library(vegan)
mplants_CA<-cca(species)
plot(mplants_CA, scaling=1, cex=2, main="Scaling=1")
round(mplants_CA$CA$tot.chi, 2)
  # 0.33
round(mplants_CA$CA$eig[1:2], 2)
  # CA1  CA2
  # 0.13 0.06
cat("The first two axis explain",
   round(sum(mplants_CA$CA$eig[1:2])
             /mplants_CA$CA$tot.chi*100), "%", "\n")
  # The first two axis explain 57 %
```


Scaling = 1
so
vit
la
cy
ac
grcyn
14
13
15
9
4
21
ru
7
8
18
17
3
12
10
22
11
eu
16
grresto
as
ma
le
5
com
vi
co
CA2
CA1
−2
−1
0
1
2
3

```
> plot(mplants_CA, scaling = 2, main = "Scaling = 2")
```


Scaling = 2
CA1
CA2
−2
−1
0
1
2
3
−3
14
13
15
21
9
4
7
so
vit
la
8
ac
cy
grcyn
19
20
ru
6
18
17
as
grresto
eu
ma
le
3
12
20
com
vi
co
22
11
16
5

Contents

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

The position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

The position of a species represents the optimum value in terms of the Gausian response model (niche) along the first and second axes. For this reason, species scores are represented as labels or points.

Site conditional biplot (scaling=1)

- The sites are the centroids of the species, that is, sites are plotted close to the species which occur at those sites.
- Distances between sites are two-dimensional approximations of their Chi-square distances. So sites close to each other are similar in terms of the Chi-square distance.

Species conditional biplot (scaling=2)

- The species are the centroids of the sites, that is, species are plotted close to the sites where they occur.
- Distances between species are two-dimensional approximations of their Chi-square distances. So species close to each other are similar in terms of the Chi-square distance.

There is also a joint plot of species and site scores (scaling=3). In this plot distances between sites and distances between species can be interpreted as the approximations of the respective Chi-square distances. However the relative positions of sites and frequencies cannot be interpreted. So this biplot is to be used with care if used at all.

Note:

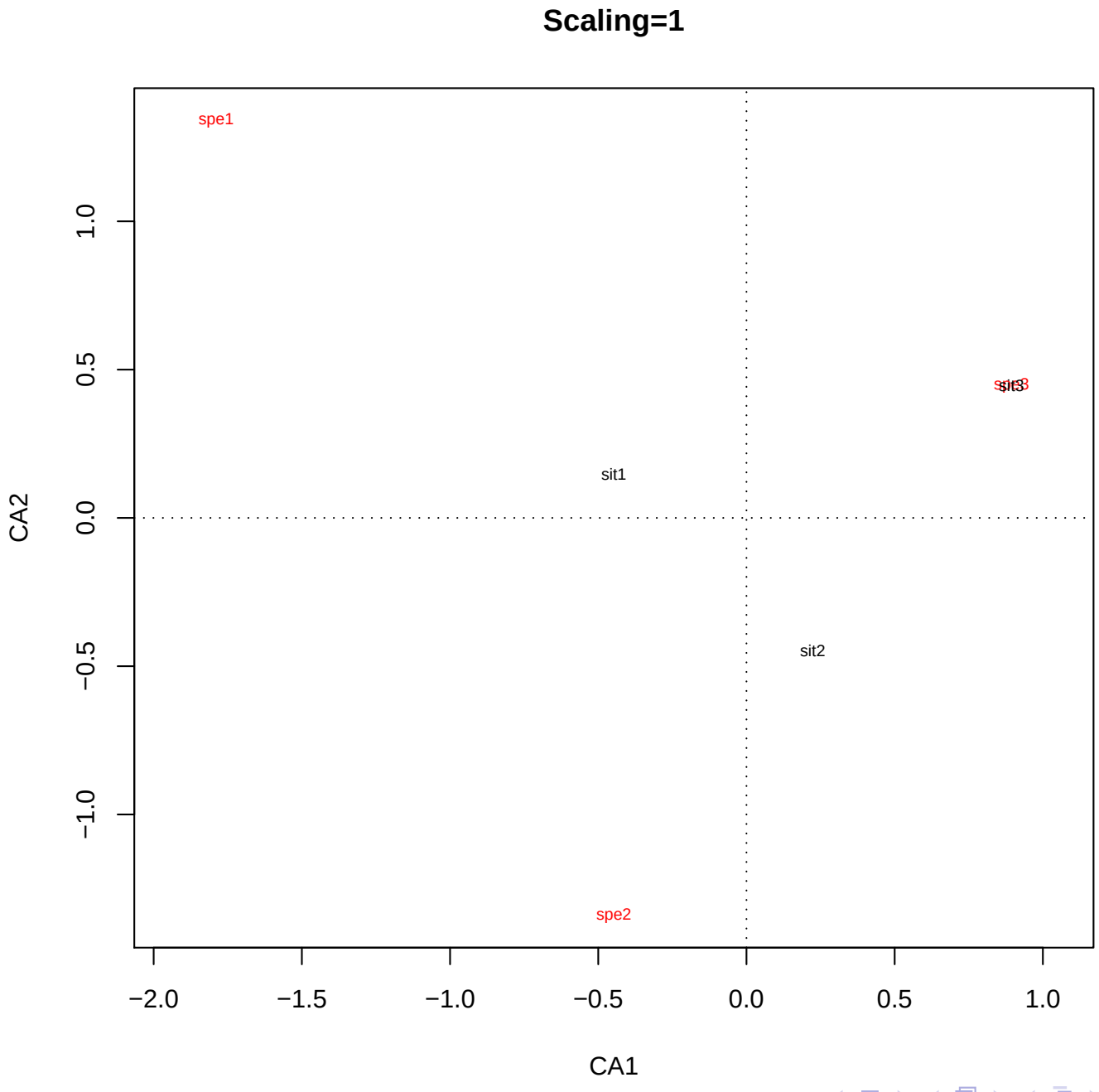
- The total inertia (or total variance) in correspondence analysis is defined as the Chi-square statistic of the site-by-species table divided by the total number of observations.
- Points further away from the origin in a biplot are the most interesting as these points make a relatively high contribution to the Chi-square statistic. So the further away from the origin a site is plotted, the more different it is from the average site.

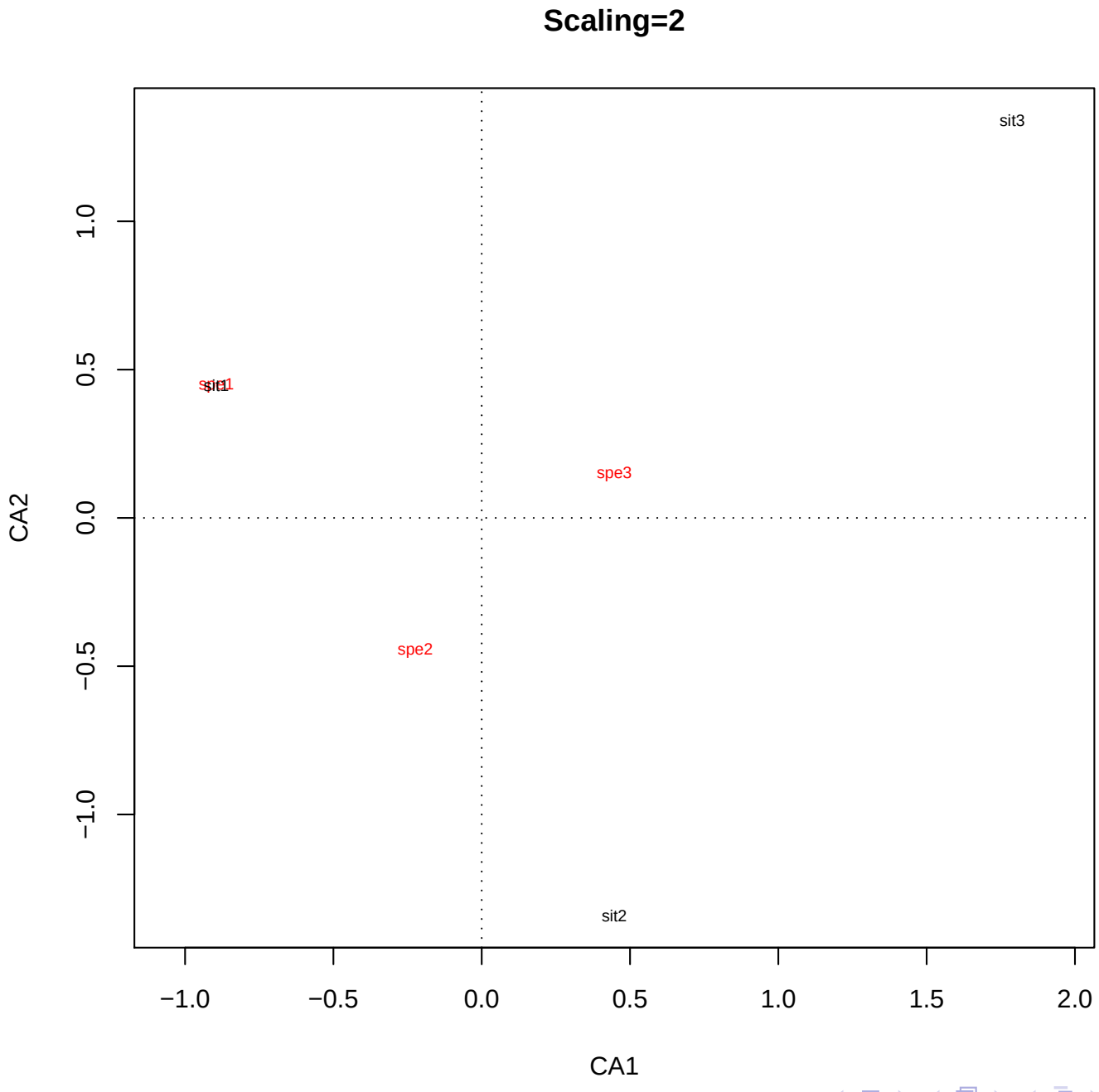
Contents

1. Correspondence analysis
   - Motivation
   - Setting
   - Mathematical background
   - Example: Mexican plant data
   - Site conditional biplot and species conditional biplot
   - Example: An artificial example

Let us look at the following artificial example which is simple enough so that we can infer site-species correspondence 'by eye'.

```
Y <- matrix(c(1,0,0,1,1,0,1,1,1),nrow=3); Y
#      [,1] [,2] [,3]
#      [1,]    1    1    1
#      [2,]    0    1    1
#      [3,]    0    0    1
myca <- cca(Y)
plot(myca,scaling=1)
plot(myca,scaling=2)
p <- Y/sum(Y)
pr <- apply(p,1,sum)
pc <- apply(p,2,sum)
expec <- as.matrix(pr) %*% t( as.matrix(pc) )
sum( (p-expec)^2/expec ) # = myca$tot.chi = 0.3611
sum(myca$CA$eig/myca$tot.chi)  # 1
```


Scaling=1
spe1
sit3
sit1
sit2
spe2
CA2
1.0
0.5
0.0
−0.5
−1.0
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
CA1


Scaling=2
sit3
sit1
spe3
spe2
sit2
CA2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
CA1