Multivariate Statistics in Ecology and Quantitative Genetics
Canonical correspondence analysis

Dirk Metzler & Noémie Becker

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

Summer semester 2016

1. Canonical correspondence analysis
   - Setting
   - Mathematical background
   - The CCA triplot
   - Example: Mexican plant data
   - When to use PCA, RDA, CA or CCA

Contents

Given: Data frames/matrices Y and X

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

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

X are the explanatory variables

Given: Data frames/matrices Y and X
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`
X are the explanatory variables

Goal: Find associations of species abundancies and sites with each environmental condition on a site being a linear combination of the environmental variables of X.

Given: Data frames/matrices Y and X
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`
X are the explanatory variables

Goal: Find associations of species abundancies and sites with each environmental condition on a site being a linear combination of the environmental variables of X.

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

Given: Data frames/matrices Y and X
$Y[i, \cdot]$ are the observations of species `i`
$Y[\cdot, j]$ are the observations at site `j`
X are the explanatory variables

Goal: Find associations of species abundancies and sites with each environmental condition on a site being a linear combination of the environmental variables of X.

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

Contents

Let Y have M rows and N columns and let X have M rows and Q columns. The site scores $(l_1, l_2, \cdots, l_M)$ in the Gaussion response model

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

are now assumed to be a linear combination of our environmental variables, that is,

$$l_i = \sum_{p=1}^{Q} X[i,p]\alpha[p].$$

Let Y have M rows and N columns and let X have M rows and Q columns. The site scores $(l_1, l_2, \cdots, l_M)$ in the Gaussion response model

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

are now assumed to be a linear combination of our environmental variables, that is,

$$l_i = \sum_{p=1}^{Q} X[i,p]\alpha[p].$$

Canonical correspondence analysis is realized by a correspondence analysis in which weighted multiple regression is used to represent the axes as linear combination of the explanatory variables.

So CCA is a CA with the axes being linear combinations of the explanatory variables.

Contents

The species scores, the site scores and the environmental scores are plotted in a figure called a triplot (confer with triplots in RDA). These triplots are the biplots from CA with additionally the explanatory variables plotted as lines.
Again 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 species scores, the site scores and the environmental scores are plotted in a figure called a triplot (confer with triplots in RDA). These triplots are the biplots from CA with additionally the explanatory variables plotted as lines.
Again 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.

In addition: Species can be projected perpendicular (=orthogonally) on the lines showing the species optima of the respective explanatory variables (in the respective scaling). Projecting sites perpendicular on the lines results in the values of the respective environmental variable at those sites.

The angle between lines does NOT represent correlation between the variables. Instead if the tip of a line is projected on another line or an axis then the resulting value represents a weighted correlation.

Contents

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

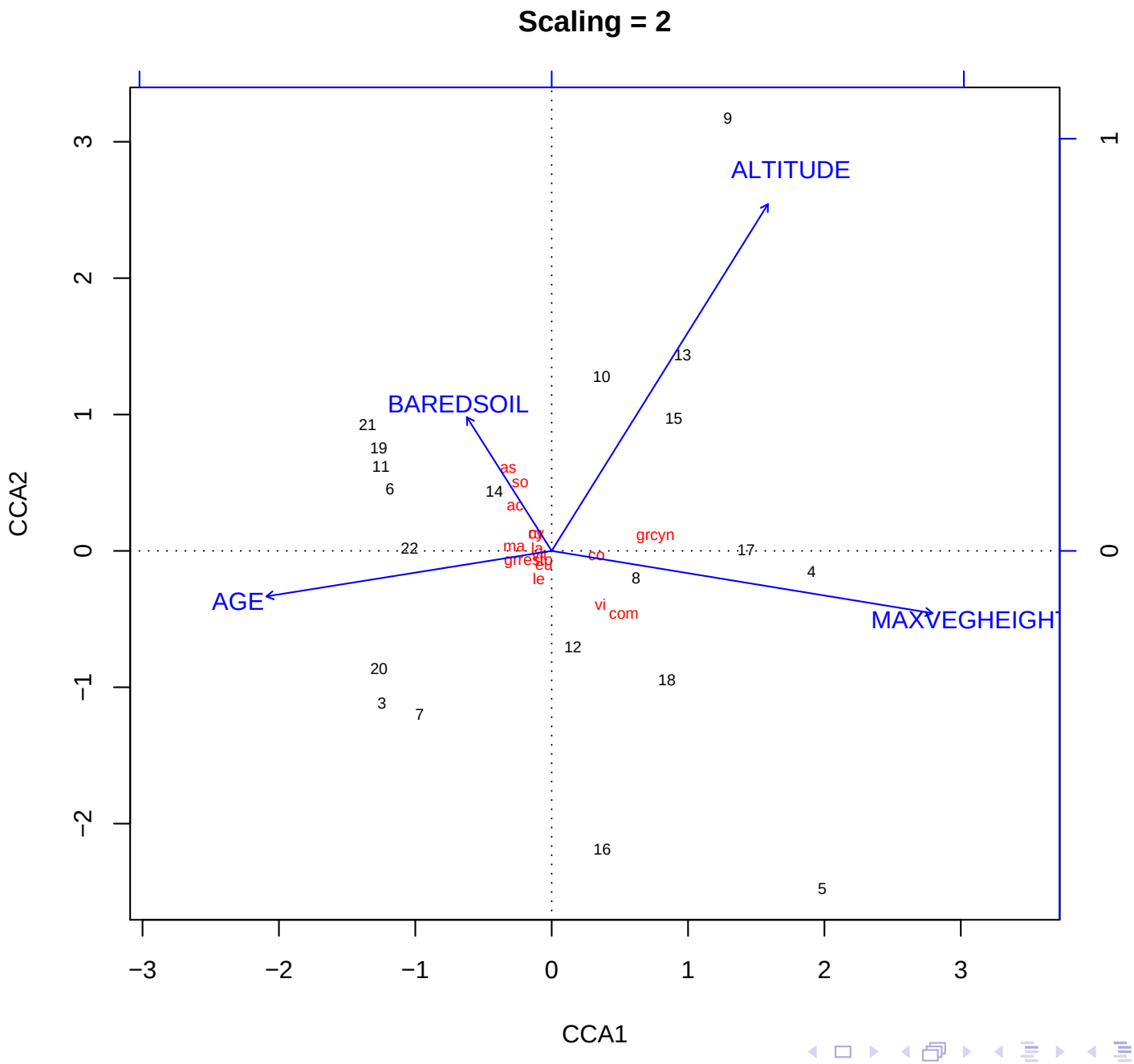
BLOCK Nominal: 1=grama pastures in Balzapote, 2=star pastures in Balzapote, 3=grama pastures in La Palma, 2=star pastures in La Palma

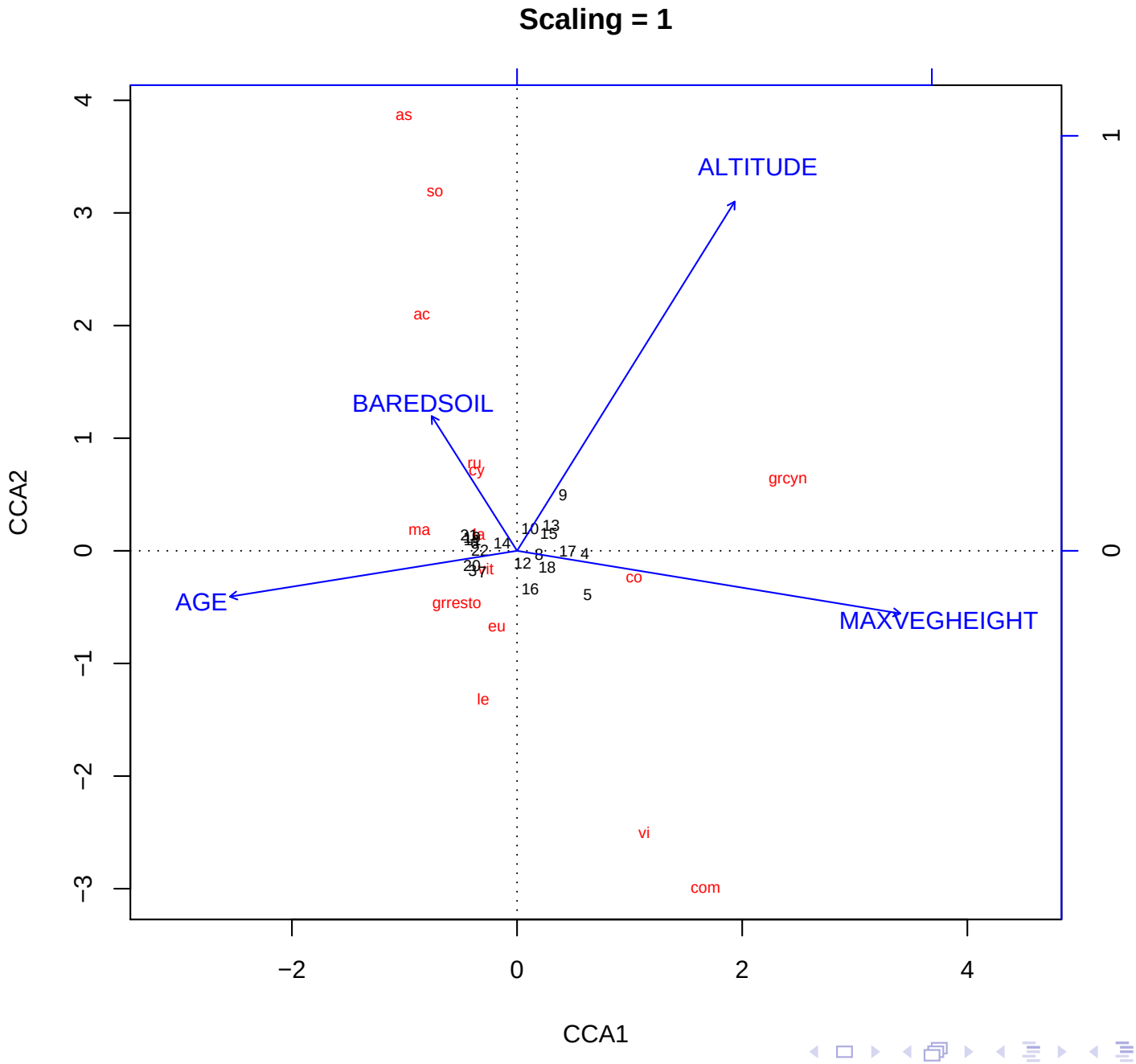
(many more variables in original data set)

```
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")]
env_var<-mplants[,c("MAXVEGHEIGHT","AGE",
                    "ALTITUDE","BAREDSOIL")]
library(vegan)
mplants_CCA<-cca(species, env_var)
```

```
# the total variation in the data is: 0.33
round(mplants_CCA$tot.chi, 2)

# the sum of all canonical eigenvalues: 0.15
round(mplants_CCA$CCA$tot.chi, 2)

# all four explanatory variables explain
cat(round(mplants_CCA$CCA$tot.chi
          /mplants_CCA$tot.chi*100), "% of data", "\n")
# 46% of the total variation in the data

# the first two (canonical) eigenvalues are: 0.10, 0.02
round(mplants_CCA$CCA$eig[1:2], 2)
```

```
# so the first two canonical axes explain:
cat(round(sum(mplants_CCA$CCA$eig[1:2])
            /mplants_CCA$CCA$tot.chi*100), "%", "\n")
# 82% of the variation that can be explained
# with the four environmental variables

# but this is (the first two canonical axes explain):
cat(round(sum(mplants_CCA$CCA$eig[1:2])
            /mplants_CCA$tot.chi*100), "%", "\n")
# 37% of the total variation in the data

plot(mplants_CCA, scaling = 2, main="Scaling2")
plot(mplants_CCA, scaling = 1, main="Scaling1")
```


Scaling = 2
ALTITUDE
BAREDSOIL
AGE
MAXVEGHEIGHT
9
13
10
15
21
19
11
6
14
22
17
4
8
12
20
18
3
7
16
5
as
so
ac
grcyn
co
vi
com
CCA1
CCA2


Scaling = 1
ALTITUDE
BAREDSOIL
AGE
MAXVEGHEIGHT
as
so
ac
ru
cy
ma
grcyn
co
grresto
eu
le
vi
com
CCA1
CCA2

Contents

When PCA, RDA, CA, CCA?

Summary of methods:

- Relationships in PCA and RDA are linear.
- In RDA and CCA two sets of variables are used, and a cause-effect relationship is asssumed.

When PCA, RDA, CA, CCA?

Summary of methods:

- Relationships in PCA and RDA are linear.
- In RDA and CCA two sets of variables are used, and a cause-effect relationship is asssumed.

	Pure ordination	Cause-effect relation
Linear model	PCA	RDA
Unimodal model	CA	CCA