Recap: Classical ordination methods

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Outline

- see workshop gavin if you really want to know this stuff

Ordination

Goodall (1954) introduced the word "ordination"

1) Ordination summarizes data
2) Ordination **embeds** in a low-dimensional space
3) Ordination **orders** samples and species

Gradients

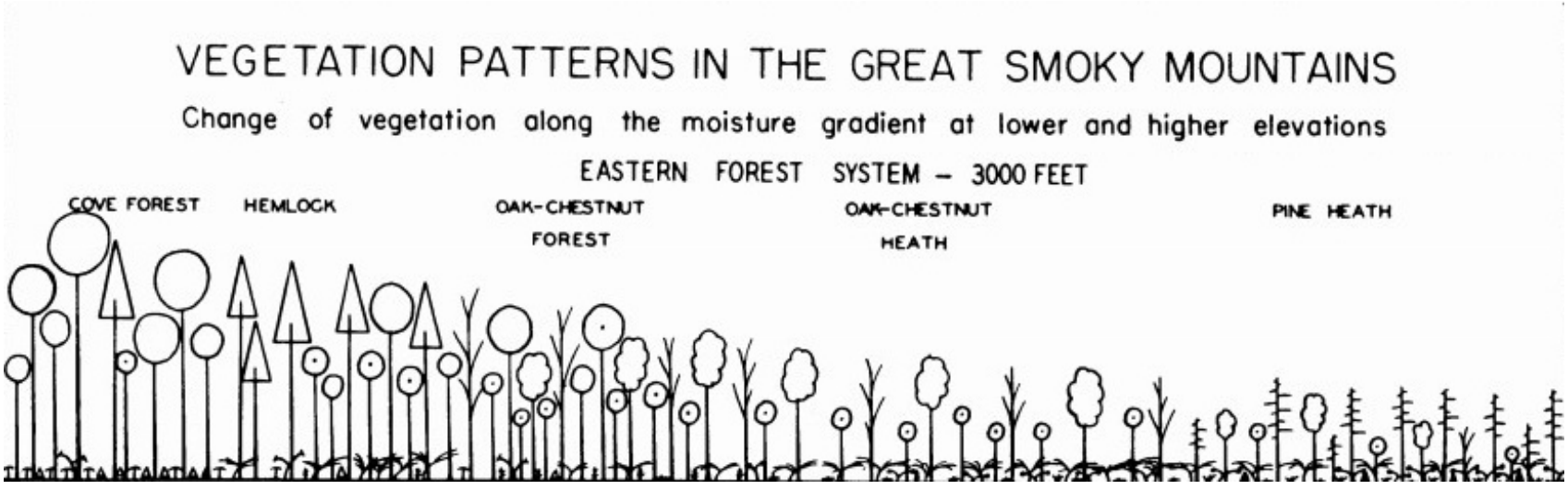
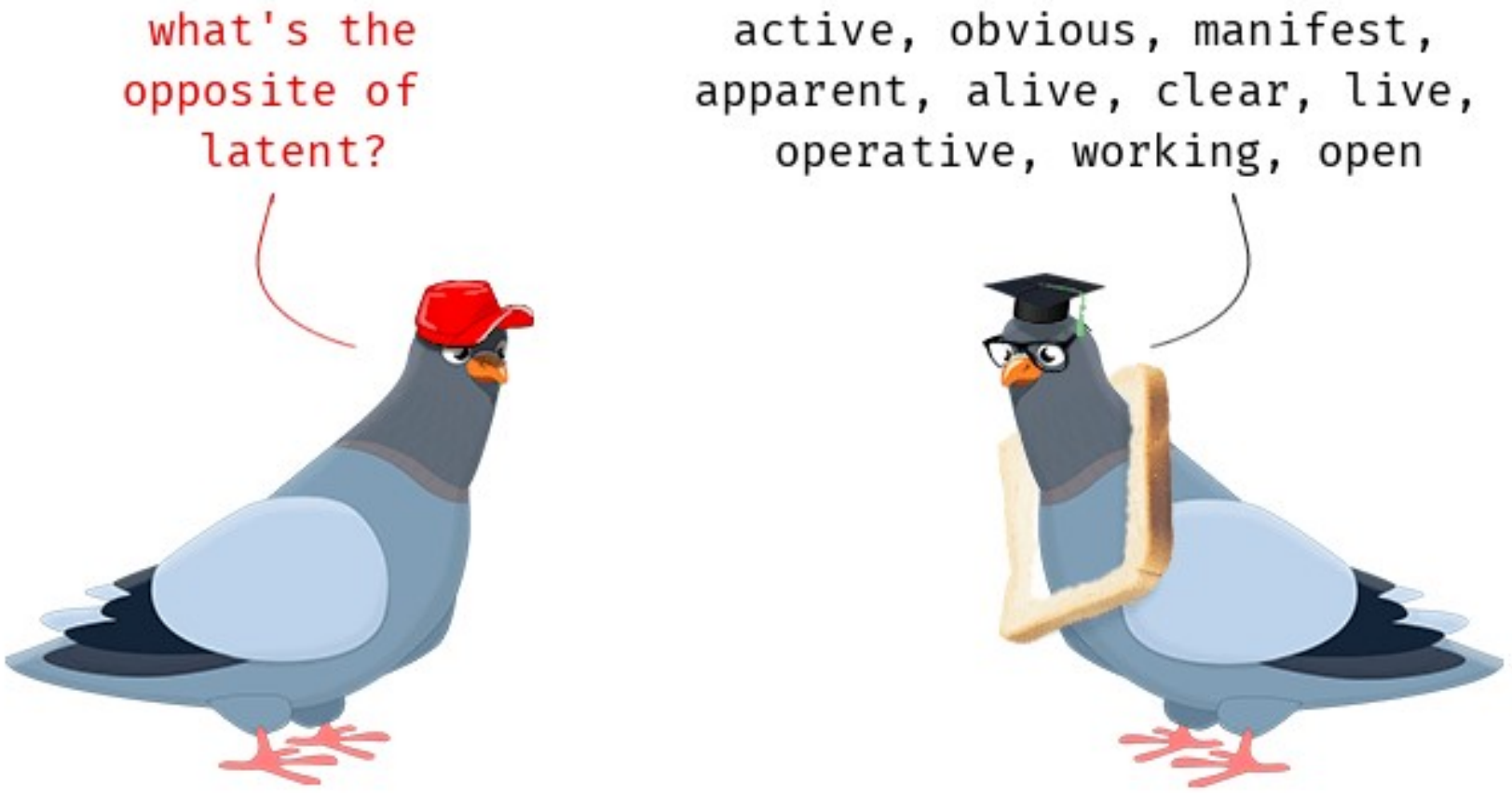


Figure 1: Whittaker 1956

- environmental gradient
- ecological gradient
- other gradients

Latent variables

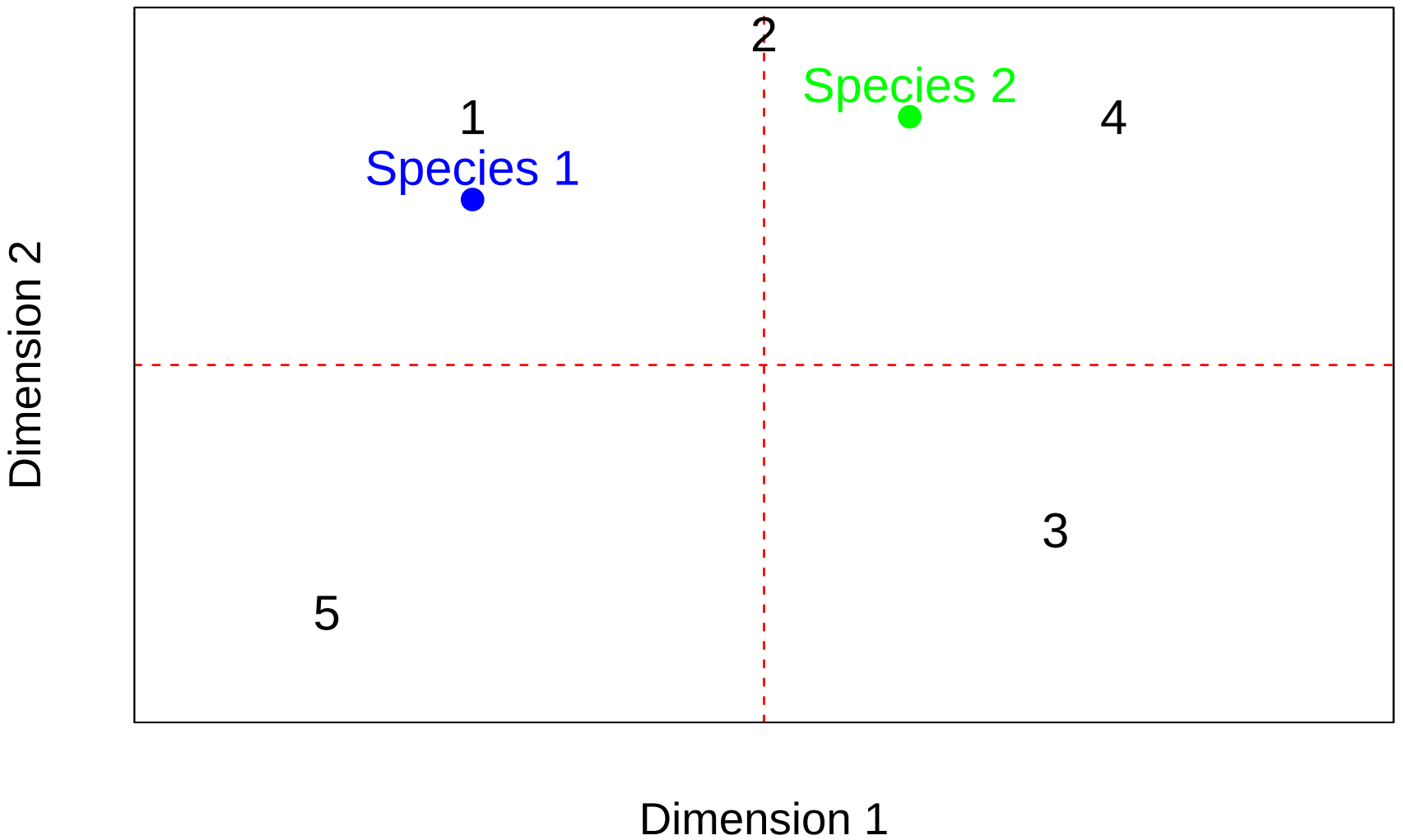


In essence: an unobserved gradient

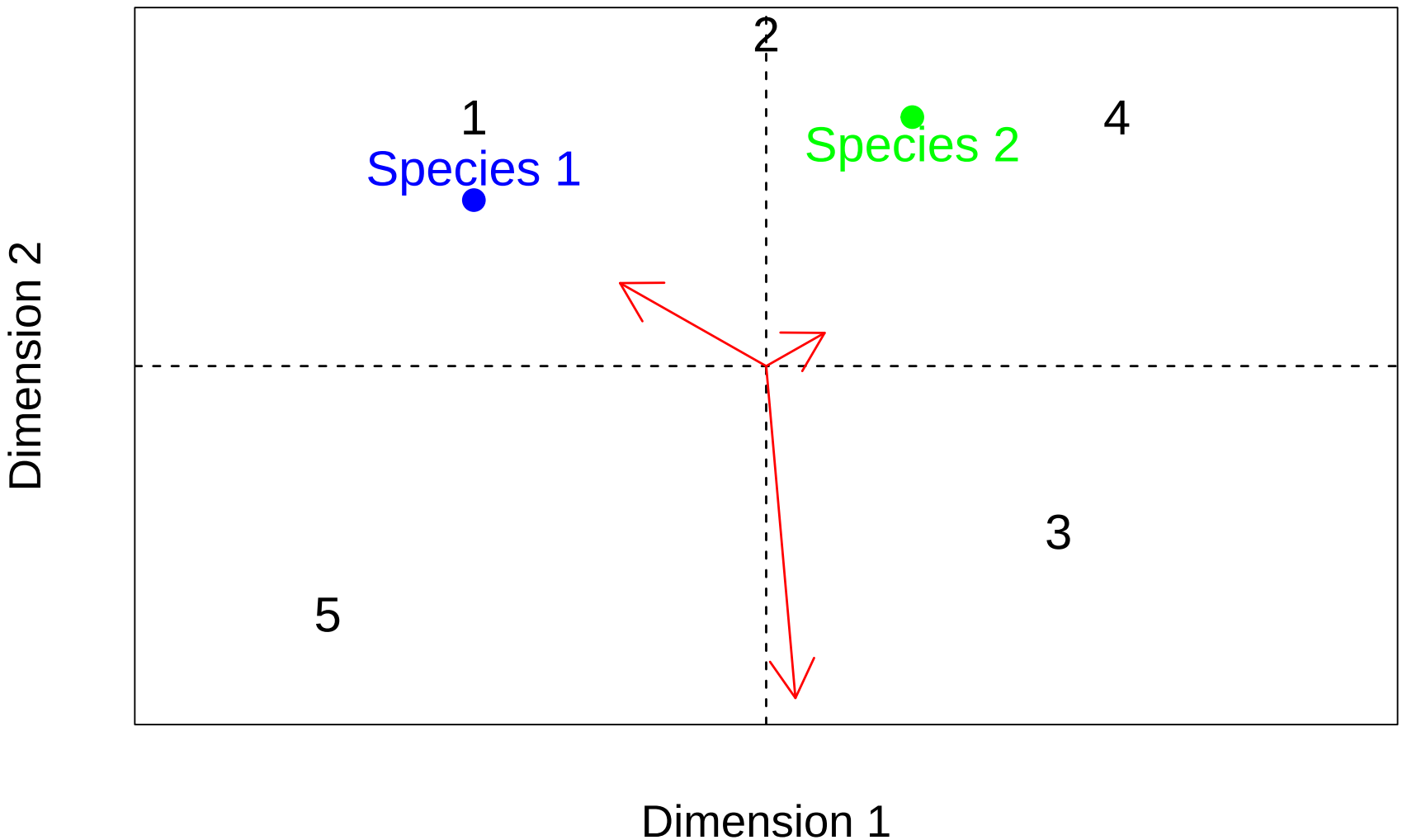
Ecological gradients

"Few major complex ecological gradients normally account for most of the variation in species composition." (Halvorsen, 2012)

Ordination plot



Ordination plot



Ordination plot

We use it to:

- Get a quick overview of patterns in the data
- Describe species **co-occurrence** patterns

When to use ordination

Mostly when we want to do dimension reduction. But also when:

1. We want to determine latent variables
   - Especially when we have not measured the environment
2. We do not have enough data to estimate species effects
3. We want to make pretty pictures

Classification of ordination

There are many ways to group ordination methods

- Indirect or direct
- Linear or unimodal
- Unconstrained or constrained
- Simple-method or distance-based

Gradient analysis

Indirect gradient analysis: patterns in species composition that may be due to environment, but without studying environmental variables
Direct gradient analysis: estimate how species are affected by environmental variables

Both are used to analyze patterns in ecological communities

Linear or Unimodal

Ordination (where does the concept come from) - arranging samples - three models according to classical methods; linear, unimodal, neither (see slides gavin)

Unconstrained or constrained

a) Unconstrained ordination: explore main drivers of variation (i.e., indirect gradient analysis)
b) Constrained ordination: filter variation due to covariates (i.e., gradient analysis)
c) (Concurrent ordination)

Unconstrained is mostly descriptive, constrained can also be used for hypothesis testing.
Both can be understood as estimating latent variables.
Note: method is different from the figure, but usually referred to with same/similar names

Ordination as latent variable models

Some ordination methods can be thought of as implementing a latent variable model

- ter Braak (1985)
- Jongman et al. (1995)
- van der Veen et al. (2022, section 3 chapter 1)

I will write these on the slides, usually they look like:

$$y_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\theta}_j \tag{1}$$

This makes for better comparison with GLLVMs tomorrow.

Unconstrained ordination

Sometimes simply referred to as “ordination”

Unconstrained ordination methods

- Principal Component Analysis (Pearson, 1901)
- Correspondence Analysis (Hirschfield, 1935)
- Detrended Correspondence Analysis (Hill and Gauch, 1980)
- Principal Coordinates Analsysis (Gower, 1966)
- Non-metric Multidimensional Scaling (Kruskal, 1964)

Unconstrained ordination

Goal: to explore co-occurrence patterns
Problem: data forms high-dimensional space

- Why do species co-occur?
  - Similar environmental preferences
  - Similar history in the environment
  - Might results in *Interactions*
- But we lack measurements of the environment
- Thus cannot test anything

Figure 2: NIBIO

Unconstrained ordination

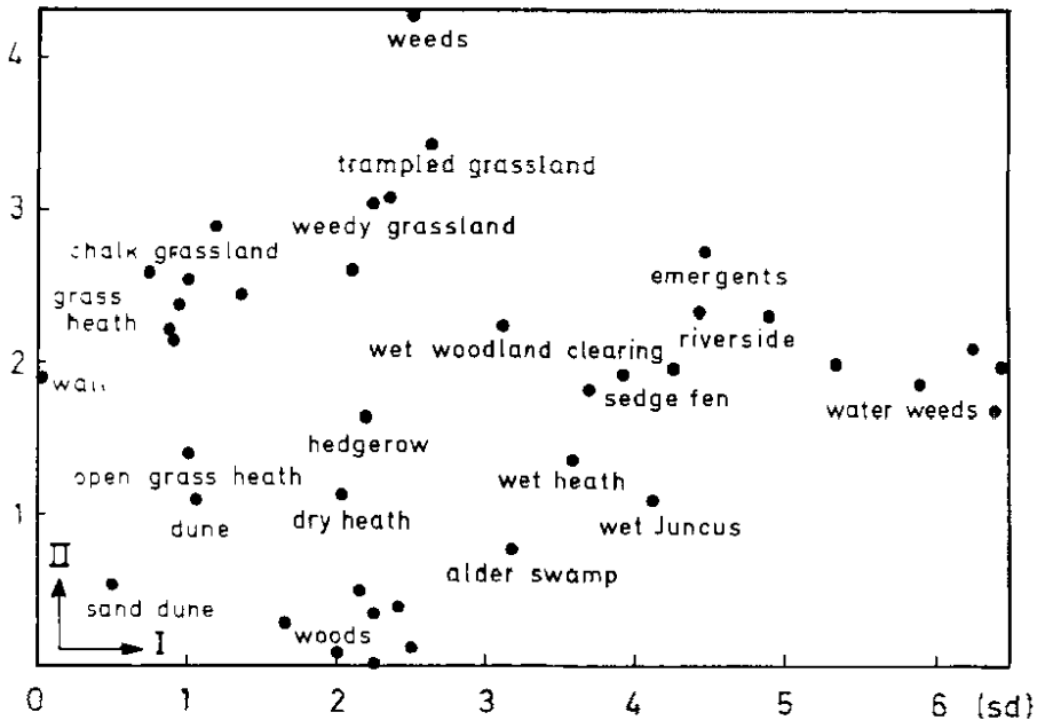


Fig. 8. DCA ordination of a vegetation survey of southeast England (data of H.J.B. Birks, 876 species in 3270 relevés, clustered into 40 composite samples). The first axis goes from dry to wet conditions, and the second axis from woodland to weed communities.

Figure 3: Hill and Gauch 1980

The ecological process

Ecological gradient theory informs us about the process

- Type of response curve
- Measured and/or unmeasured components
- Spatial and/or temporal components
- Functional traits or Phylogeny
- Et cetera.

Method attempts to capture underlying process

Unconstrained ordination

Used to:

- Visualize patterns in data
- Draw 2D plots
- Generate hypotheses
- Explore drivers of community composition

To infer environmental conditions from species relationships

Dutch Dune data

Achimill	Agrostol	Airaprae	Alopgeni	Anthodor	Bellpere	Bromhord	Ch
1	0	0	0	0	0	0	
3	0	0	2	0	3	4	
0	4	0	7	0	2	0	
0	8	0	2	0	2	3	
2	0	0	0	4	2	2	
2	0	0	0	3	0	0	
2	0	0	0	2	0	2	
0	4	0	5	0	0	0	
0	3	0	3	0	0	0	
4	0	0	0	4	2	4	

- Another classical dataset, originally by Jongman et al. (1995)
- Ordinal classes for 30 plant species at 20 sites

Principal Component Analysis

PCA was developed by Pearson (1901) and implements:

$$\overline{\mathbf{Y}} = \mathbf{UDV}^\top \tag{2}$$

which is the same as the model:

$$y_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\theta}_j \tag{3}$$

- Intercepts: $\beta_{0j} = \frac{1}{n}\sum_{i=1}^{n} y_j$
- Site scores: $\mathbf{z}_i^\top = \mathbf{u}_i^\top$
- Species scores: $\boldsymbol{\theta}_j = \mathbf{D}\mathbf{v}_j$ (scaling can instead be swept into the site scores)

So: PCA fits assumes linear responses of species to the ordination axis

PCA

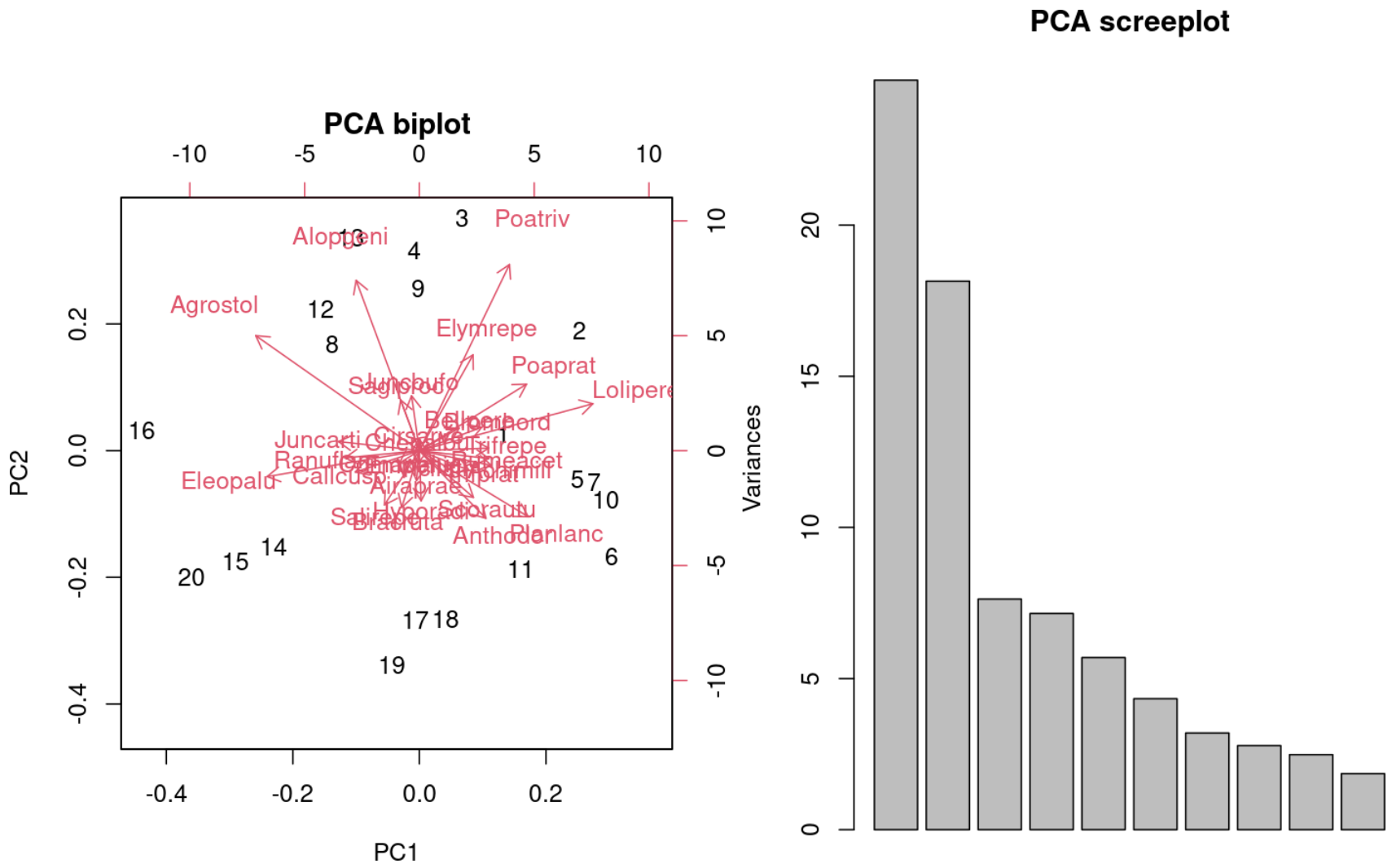
- PCA draws a line through the direction of most spread
- Then draws another one orthogonal to that
- Continues until we have as many axes as species
- Usually we only take a few axes
- Troubles with horseshoe effect

PCA of dune data

```
par(mfrow=c(1,2))
data(dune, package = "vegan") #load data
Y <- dune
PCA <- prcomp(Y)

biplot(PCA, main = "PCA biplot");plot(PCA, main = "PCA screeplot")
```

PCA of dune data



PCA of dune data

- Not generally recommended for community data
- Only for "short" gradients due to linearity assumption
- Mostly useful for Gaussian-distributed responses

Correspondence analysis

CA was developed by Hirschfield (1935) and implements:

$$\mathbf{Y} = \text{diag}(\sqrt{\sum_{j=1}^{m} y_{ij}})\mathbf{UD}\text{diag}(\sum_{i=1}^{n}\sum_{j=1}^{m} y_{ij})\mathbf{V}^{\top}\text{diag}(\sqrt{\sum_{i=1}^{n} y_{ij}}) \quad (4)$$

which is the same as the model:

$$y_{ij} = \mathbf{z}_i^{\top}\boldsymbol{\theta}_j \quad (5)$$

- First dimension captures expected frequencies
- Site scores: $\mathbf{z}_i^{\top} = \mathbf{u}_i^{\top}\sqrt{\sum_{j=1}^{m} y_{ij}}$
- Species scores: $\boldsymbol{\theta}_j = \mathbf{D}\mathbf{v}_j\sqrt{\sum_{j=1}^{n} y_{ij}}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij}$ (scaling can instead be swept into the site scores)

CA as unimodal model

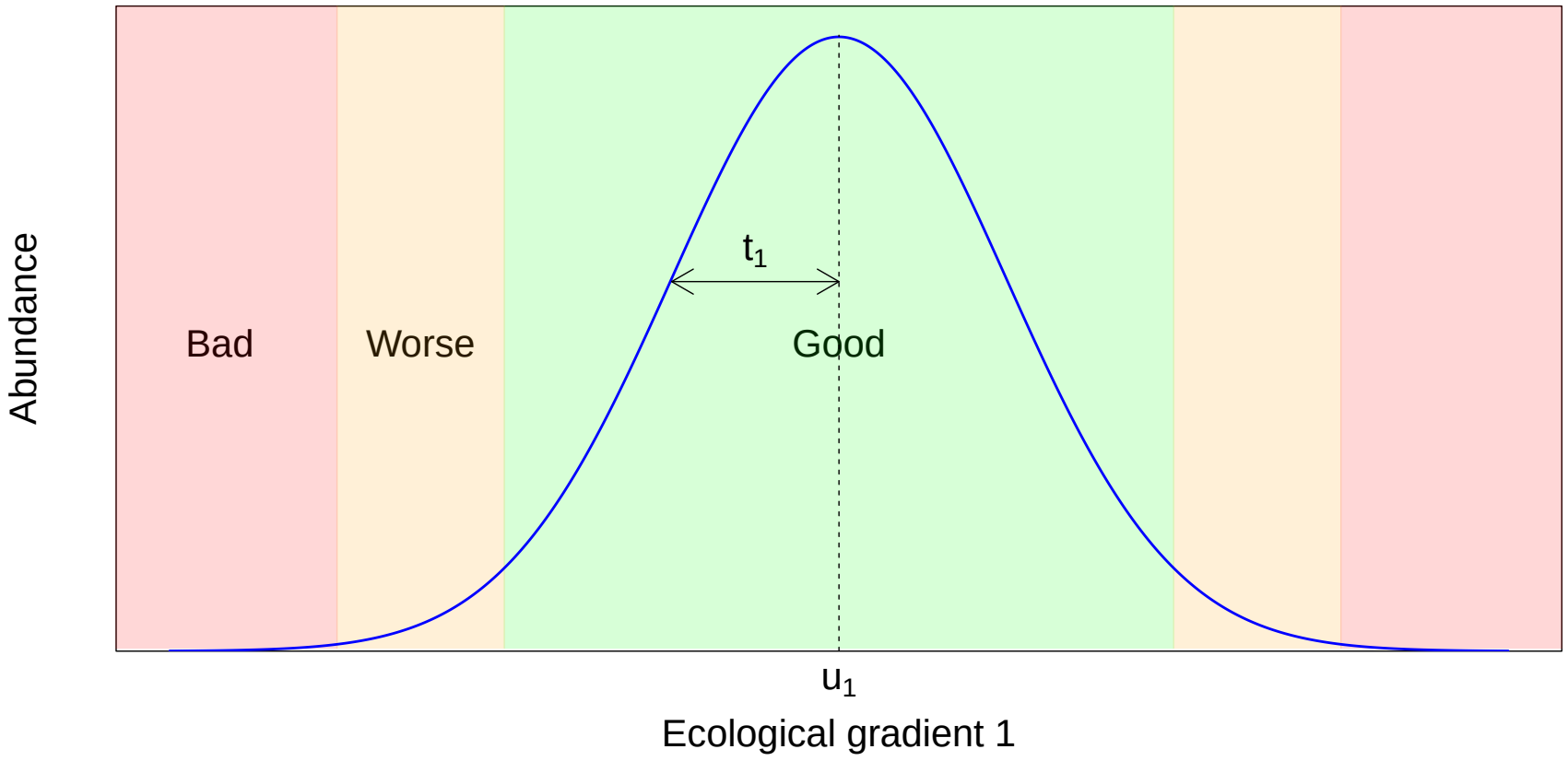
ter Braak (1985) showed that the CA solution approximates the MLEs of the model:

$$\begin{aligned} \mathrm{g}^{-1}(y_{ij}) &= \alpha_i + \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\theta}_j \\ &= c_j - \frac{1}{2}\frac{(z_i - u_j)^2}{t^2} \end{aligned} \tag{6}$$

- c_j the maximum
- u_j the optimum
- t the tolerance

Also referred to as "species packing" (MacArthur 1970)

Unimodal responses



Multiple unimodal responses

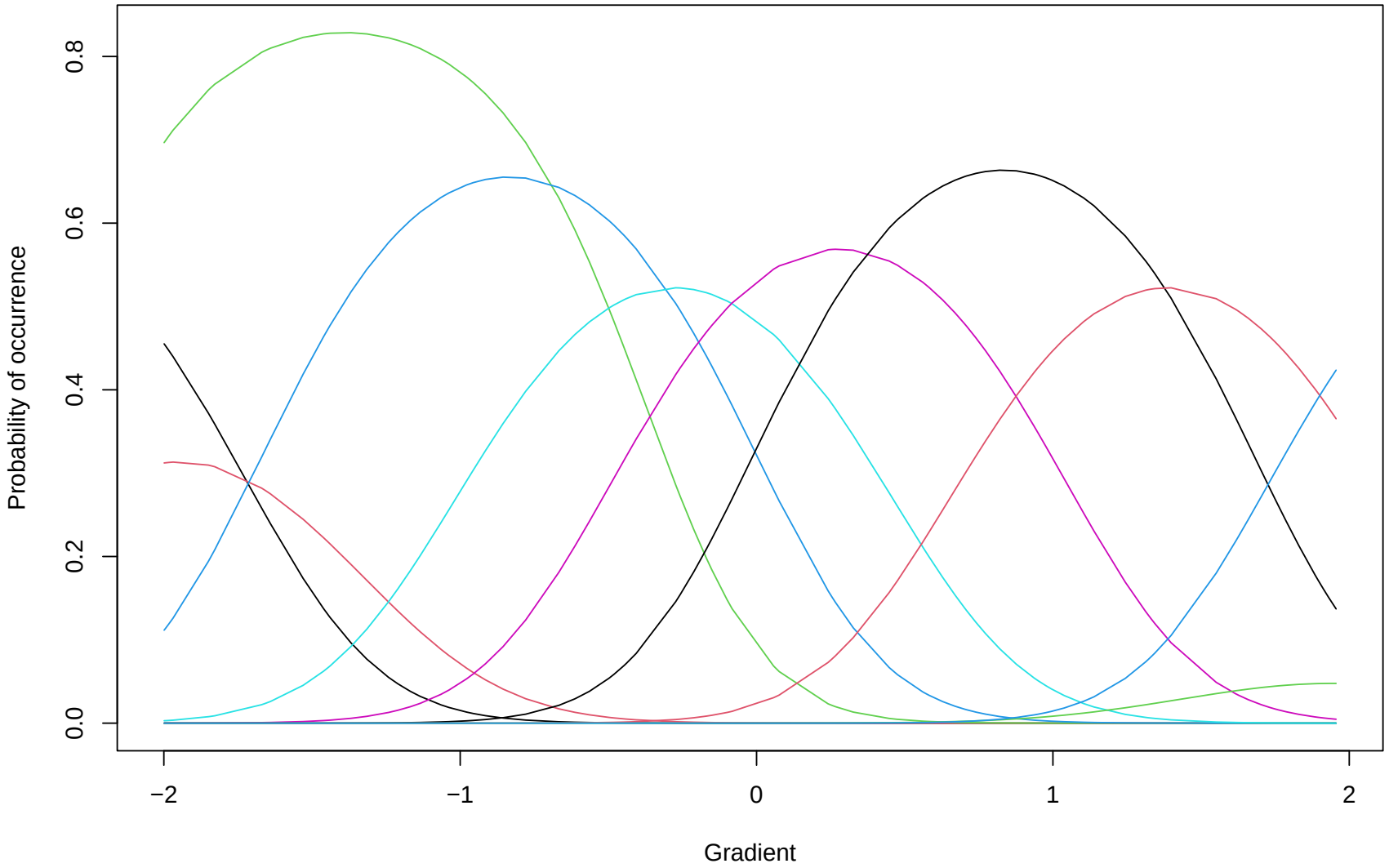


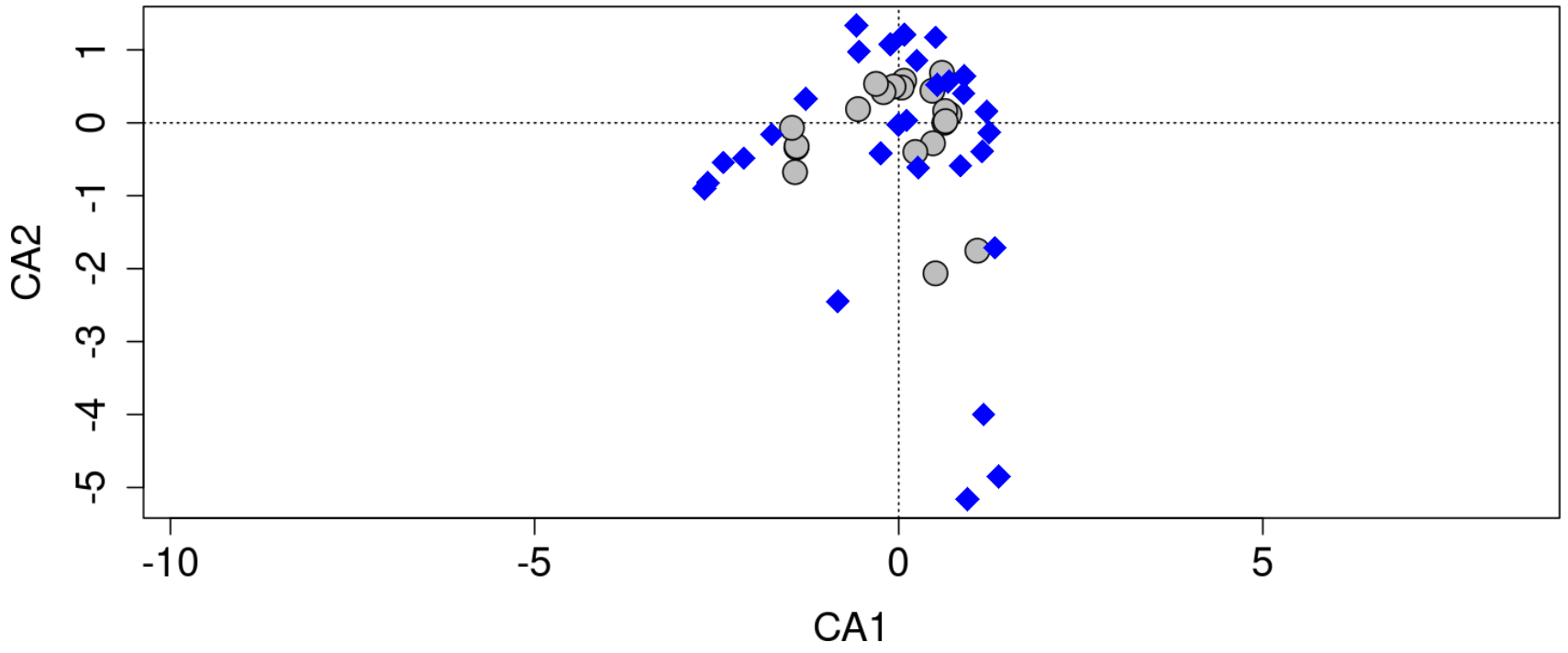
Figure 1: Example of unimodal species responses (Stevens 2021)

CA of dune data

```
CA <- cca(Y)
plot(CA, type="n", scaling="sites", cex.lab=1.5, cex.axis = 1.5)
points(CA, scaling = "sites", cex = 2, bg = "grey", pch = 21)
points(CA, "species", col="blue", cex=2, scaling="sites", pch = 18)
```

Some R magic for easier reading

CA of dune data



- Grey points are sites
- Blue names represent species optima

Detrended Correspondence Analysis

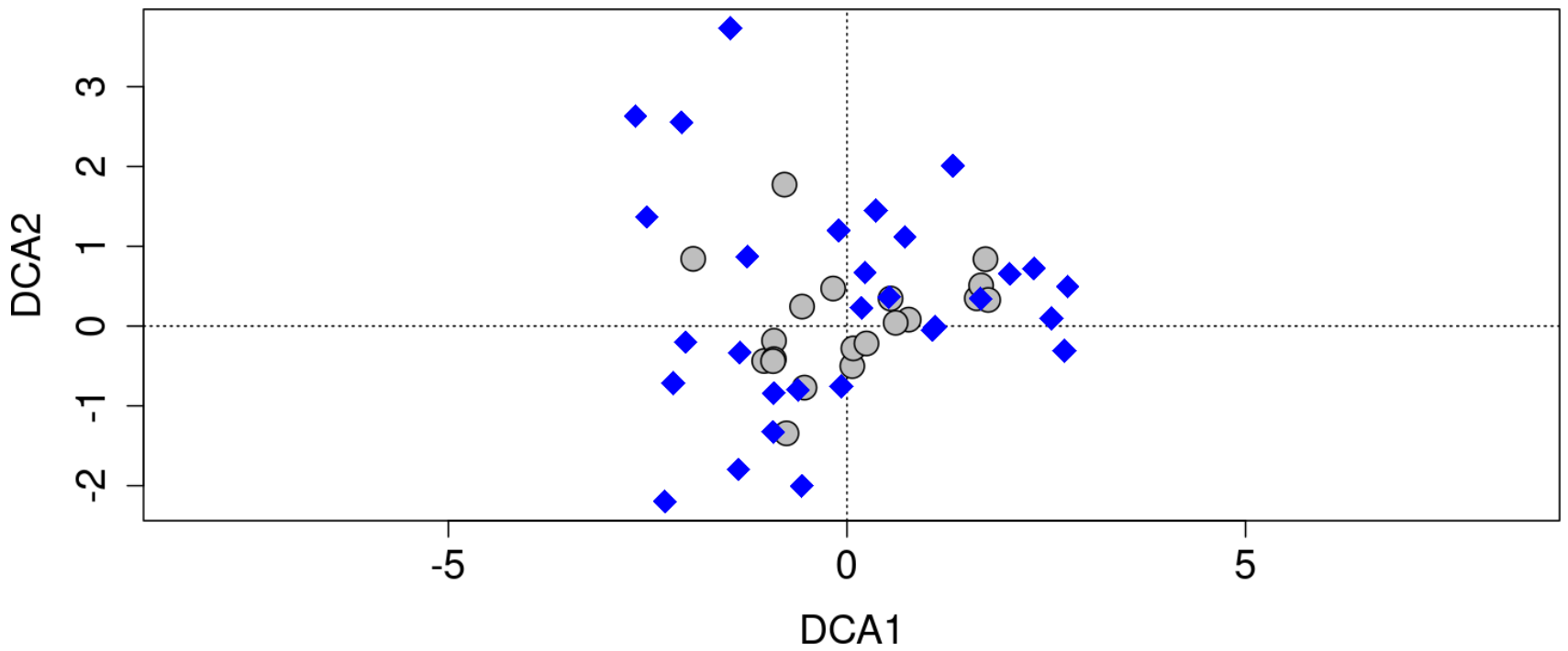
Developed by Hill and Gauch (1980). "Improves" CA in a few ways:

- Detrends (removing relationships of axes)
- Non-linearly rescales the axes
- I.e., removing "Arch" and "Edge" effects
- No statistical basis for these actions, so difficult to connect to a model.
- Relation to original data becomes fuzzy
- Often introduces its own "tongue" effect
- Sensitive to its settings
- Can sometimes remove also ecological effects

DCA of dune data

```
DCA <- decorana(Y)
plot(DCA, type="n", cex.lab=1.5, cex.axis = 1.5)
points(DCA, cex = 2, bg = "grey", pch = 21)
points(DCA, "species", col="blue", cex=2, scaling="sites", pch = 18)
```

DCA of dune data



DCA of dune data

```
summary(DCA, display = "none")
```

```
##
## Call:
## decorana(veg = Y)
##
## Detrended correspondence analysis with 26 segments.
## Rescaling of axes with 4 iterations.
## Total inertia (scaled Chi-square): 2.1153
##
##                        DCA1   DCA2    DCA3    DCA4
## Eigenvalues          0.5117 0.3036 0.12125 0.14267
## Additive Eigenvalues 0.5117 0.2985 0.12242 0.12984
## Decorana values      0.5360 0.2869 0.08136 0.04814
## Axis lengths         3.7004 3.1166 1.30055 1.47888
```

Most popular statistic of DCA is its axis lengths

DCA axis length

DCA axis length (in "SD" units) is a statistic often used to decide if a dataset exhibit unimodal responses.

Long axis: turnorver is high
Short axis length: turnover is low

- Species curves span over about 4SD
- Sites further than 4SD apart have few species in common

Distances and transformations

Ordination is a world based on distances and data transformations

Attitude: the right distance measures solves all your issues

- PCA: euclidean distance
- (D)CA: χ^2 distance

Here "distance" has a particular meaning (see workshop by Gavin Simpson)
Distance of sites; represent (dis)similarity in multidimensional space
Downside: no species effects

Idea: change distance measure to accommodate data properties

Principal Coordinate Analysis

Pretty much PCA of a distance matrix (so also a similar implied model)

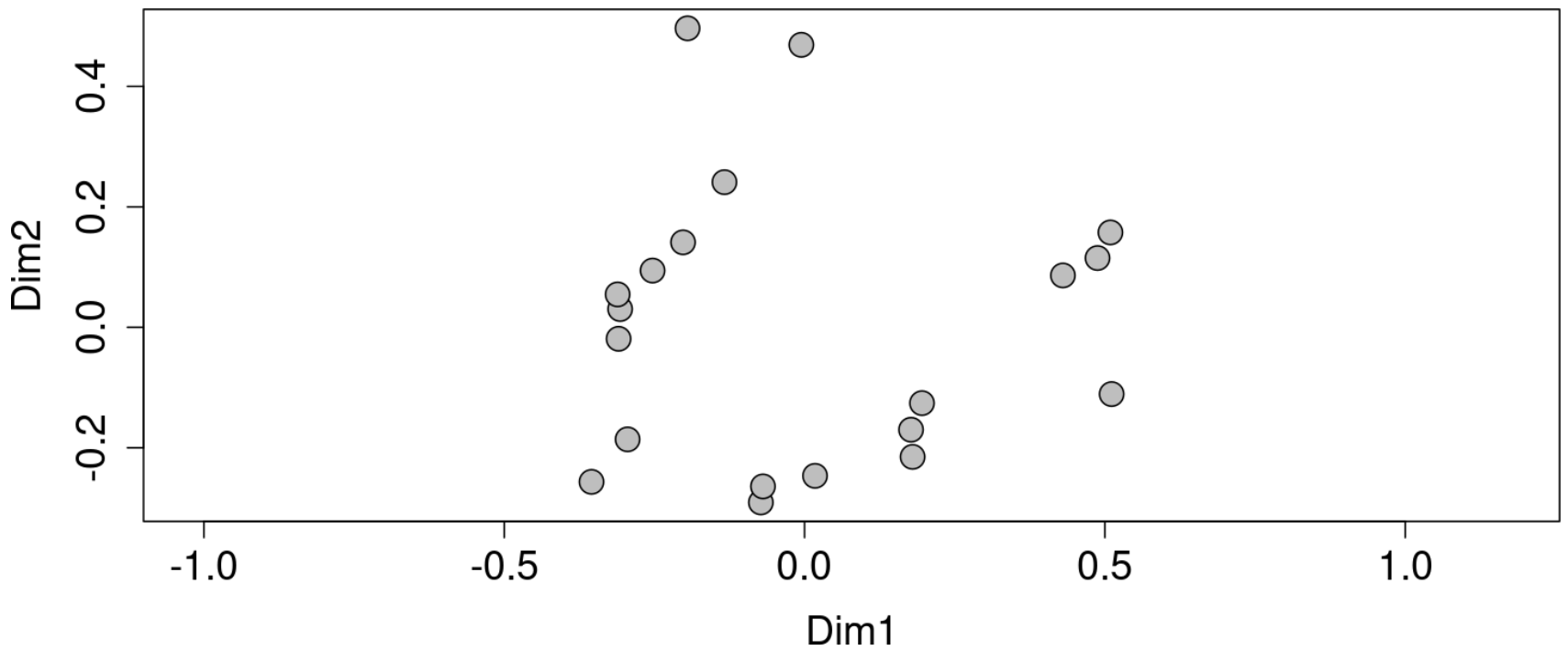
1) Calculate distance matrix and element-wise square it
2) Row-center
3) Column-center
4) Apply eigendecomposition with some scaling ($-\frac{1}{2}$)

Sometimes, eigenvalues can be negative

PCoA of dune data

```
PCoA <- cmdscale(vegdist(Y, method = "bray"))
ordiplot(PCoA, type = "n", cex.lab = 1.5, cex.axis = 1.5)
points(PCoA, pch = 21, bg = "grey", cex = 2)
```

PCoA of Dune data



NMDS

Unlike PCoA, NMDS is an iterative method that does not rely on an eigendecomposition.

- Developed by Kruskal (1964)
- Popularized by Minchin (1987)
- NMDS minimizes a measure called "stress" (lower is better)
- Invariant to rotation, but usually post-hoc receives a rotation

Vegetatio 69: 89–107, 1987

An evaluation of the relative robustness of techniques for ecological ordination

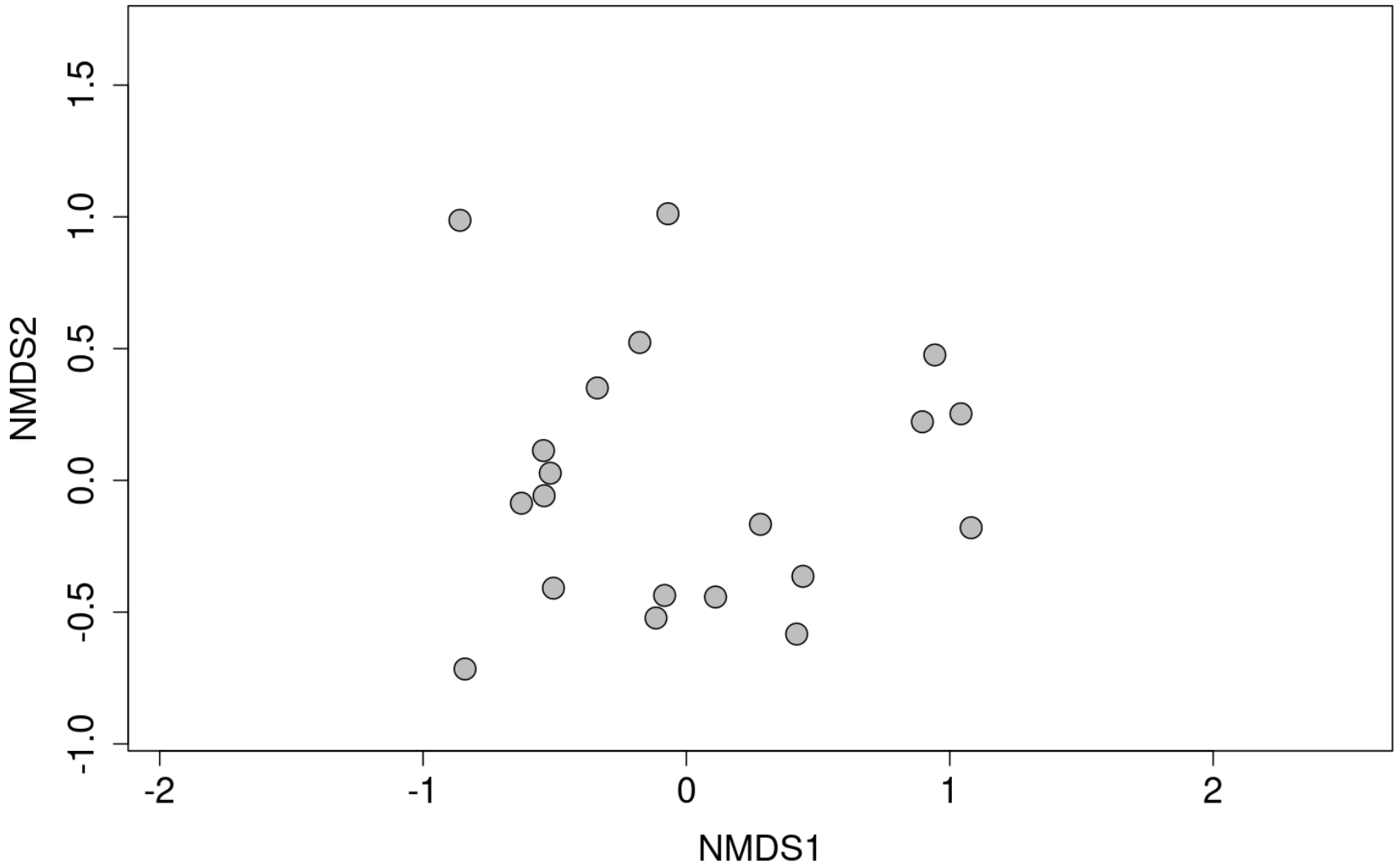
Peter R. Minchin*
CSIRO Division of Water and Land Resources, G.P.O. Box 1666, Canberra, 2601, Australia

NMDS of Dune data

```
NMDS <- metaMDS(Y)
```

```
## Run 0 stress 0.1192678
## Run 1 stress 0.1192678
## ... Procrustes: rmse 3.031091e-05  max resid 9.235996e-05
## ... Similar to previous best
## Run 2 stress 0.1808911
## Run 3 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 0.02026942  max resid 0.06495824
## Run 4 stress 0.1183186
## ... New best solution
## ... Procrustes: rmse 1.071382e-05  max resid 3.378876e-05
## ... Similar to previous best
## Run 5 stress 0.1808911
## Run 6 stress 0.1192678
## Run 7 stress 0.1192678
```

NMDS of Dune data



NMDS implied model

- NMDS just makes "pretty" pictures
- It is not a statistical method, so harder to connect to a model
- But see Brady (1985)

Constrained ordination methods

- RDA
- (D)CCA
- ~~Distance-based~~
  - Note: Superimposition of vectors on NMDS is not constrained ordination
  - Constrained NMDS exists but it is never used

Unconstrained ordination is fun, but what if you want more? Constrained ordination relates ordination axes to environmental variables.

Constrained ordination filters variation due to covariates, and throws the rest away

Dune data covariates

```
data(dune.env, package = "vegan")
X = dune.env
knitr::kable(head(X, 10), format="latex", booktabs = T)
```

A1	Moisture	Management	Use	Manure
2.8	1	SF	Haypastu	4
3.5	1	BF	Haypastu	2
4.3	2	SF	Haypastu	4
4.2	2	SF	Haypastu	4
6.3	1	HF	Hayfield	2
4.3	1	HF	Haypastu	2
2.8	1	HF	Pasture	3
4.2	5	HF	Pasture	3
3.7	4	HF	Hayfield	1
3.3	2	BF	Hayfield	1

Redundancy Analysis

PCA but with covariates:

1) Estimate β in $\overline{\mathbf{Y}} = \overline{X}\beta$
2) Apply PCA to the fitted values

So we have the model:

$$y_{ij} = \beta_{0j} + \mathbf{x}_i^\top \mathbf{B}\boldsymbol{\gamma}_j + \epsilon_{ij} \tag{7}$$

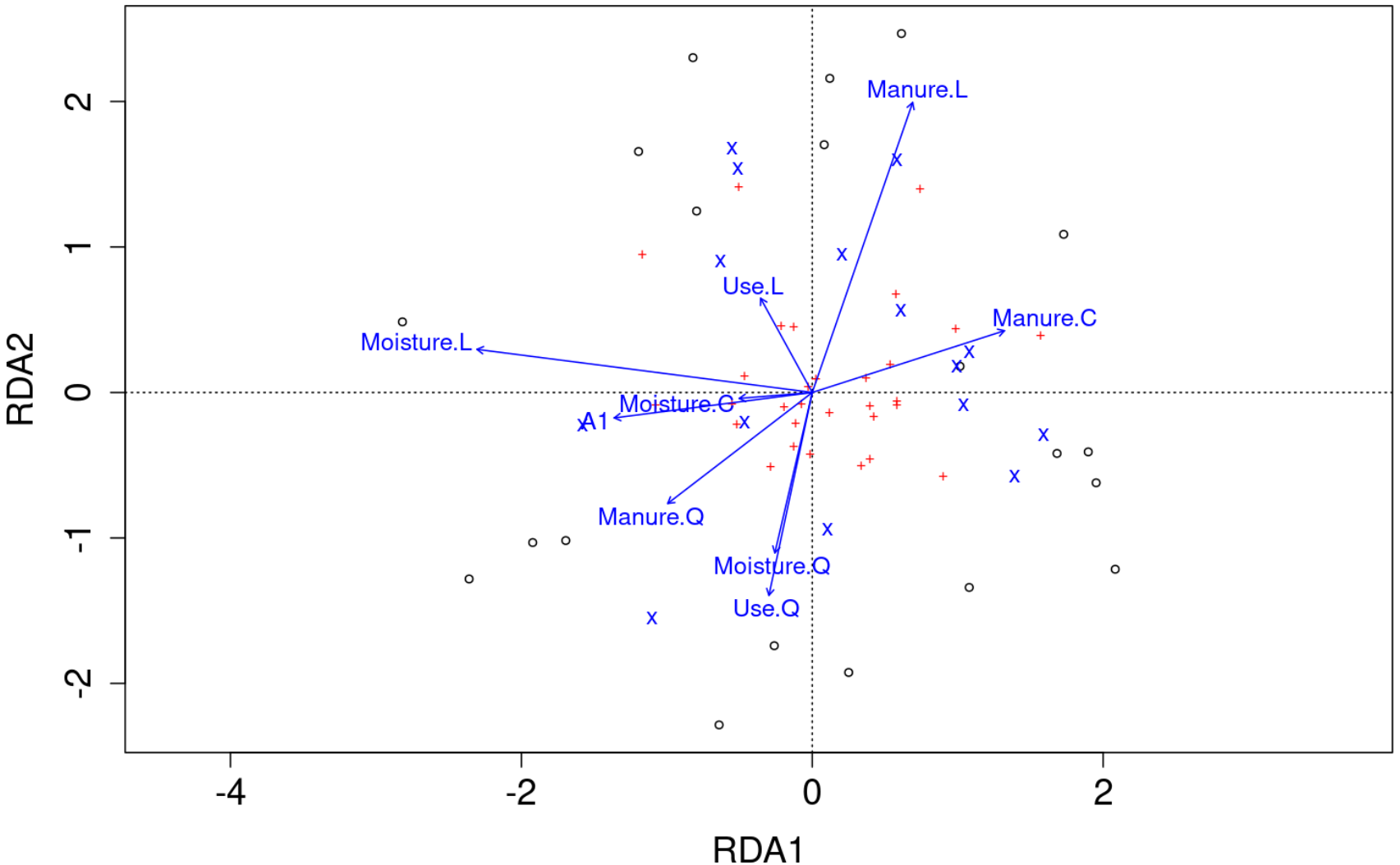
- $\mathbf{B} = \boldsymbol{\beta}\mathbf{V}\mathbf{D}^{-1}$ are the canonical coefficients
- $\boldsymbol{\gamma}_j$ are given by $\mathbf{v}_j$

So RDA explicitly assumes normality, linearity, homoscedasticity, etc.

RDA of Dune data

```
RDA <- rda(Y~A1+Moisture+Management+Use+Manure,data=X)
ordiplot(RDA, cex=3, cex.lab = 1.5, cex.axis = 1.5)
```

RDA of Dune data



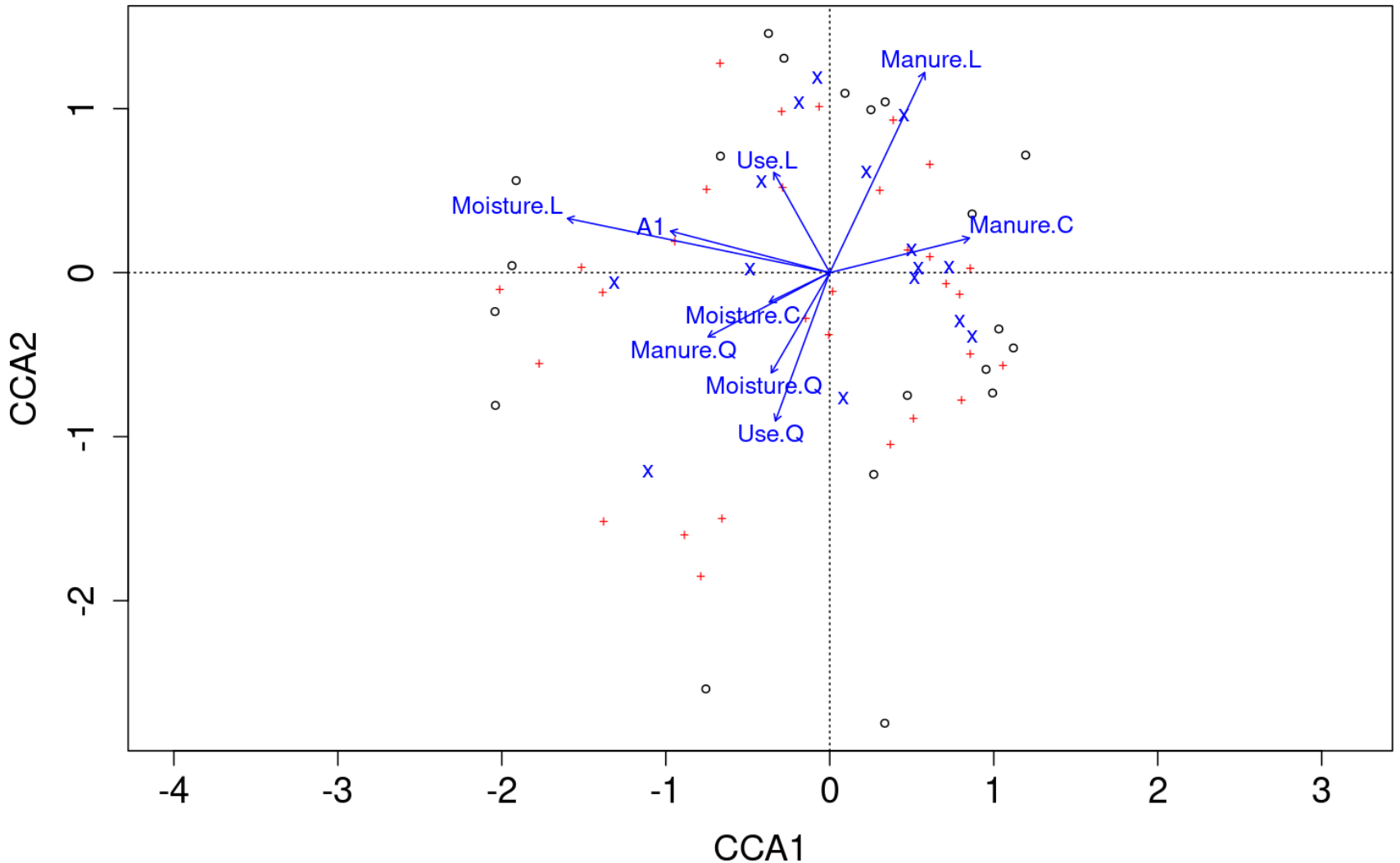
Canonical Correspondence Analysis

- Same model as CA, but constrained
- Widely acknowledge to be an excellent method
- With weighted average scores (not LC)

CCA of Dune data

```
CCA <- cca(Y~A1+Moisture+Management+Use+Manure,data=X)
ordiplot(CCA, cex=3, cex.lab = 1.5, cex.axis = 1.5)
```

CCA of Dune data



Site scores

Weighted Average scores: should be preferred, less sensitive to noise

Linear Combination scores

Ecology, 74(8), 1993, pp. 2215–2230

PUTTING THINGS IN EVEN BETTER ORDER:
THE ADVANTAGES OF CANONICAL CORRESPONDENCE ANALYSIS[1]

MICHAEL W. PALMER
Department of Botany, Oklahoma State University, Stillwater, Oklahoma 74078 USA

Main benefits of these methods

1) Easy to use
2) Loads of resources
3) Isuess, artefacts, usecases are all well known
4) Permutation testing is readily available
5) Variance partitioning is straightforward

Summary classical methods

- Methods either perform indirect or direct gradient analysis
- Are assumed to have some kind of response model (linear, unimodal, sometimes neither)
- All of these method have an implicit distance measure
- Discrete data: (D)C(C)A, or some distance-based method
- Continuous data and linear responses: PCA or RDA