Model-based ordination

Bert van der Veen

Department of Mathematical Sciences, NTNU

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

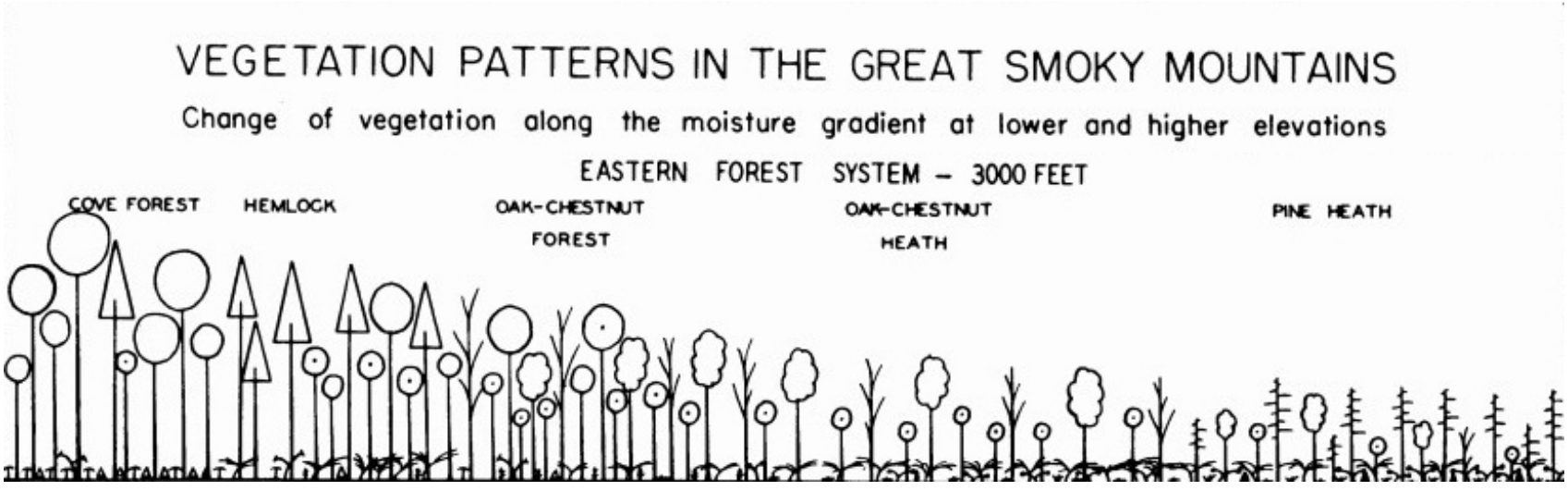
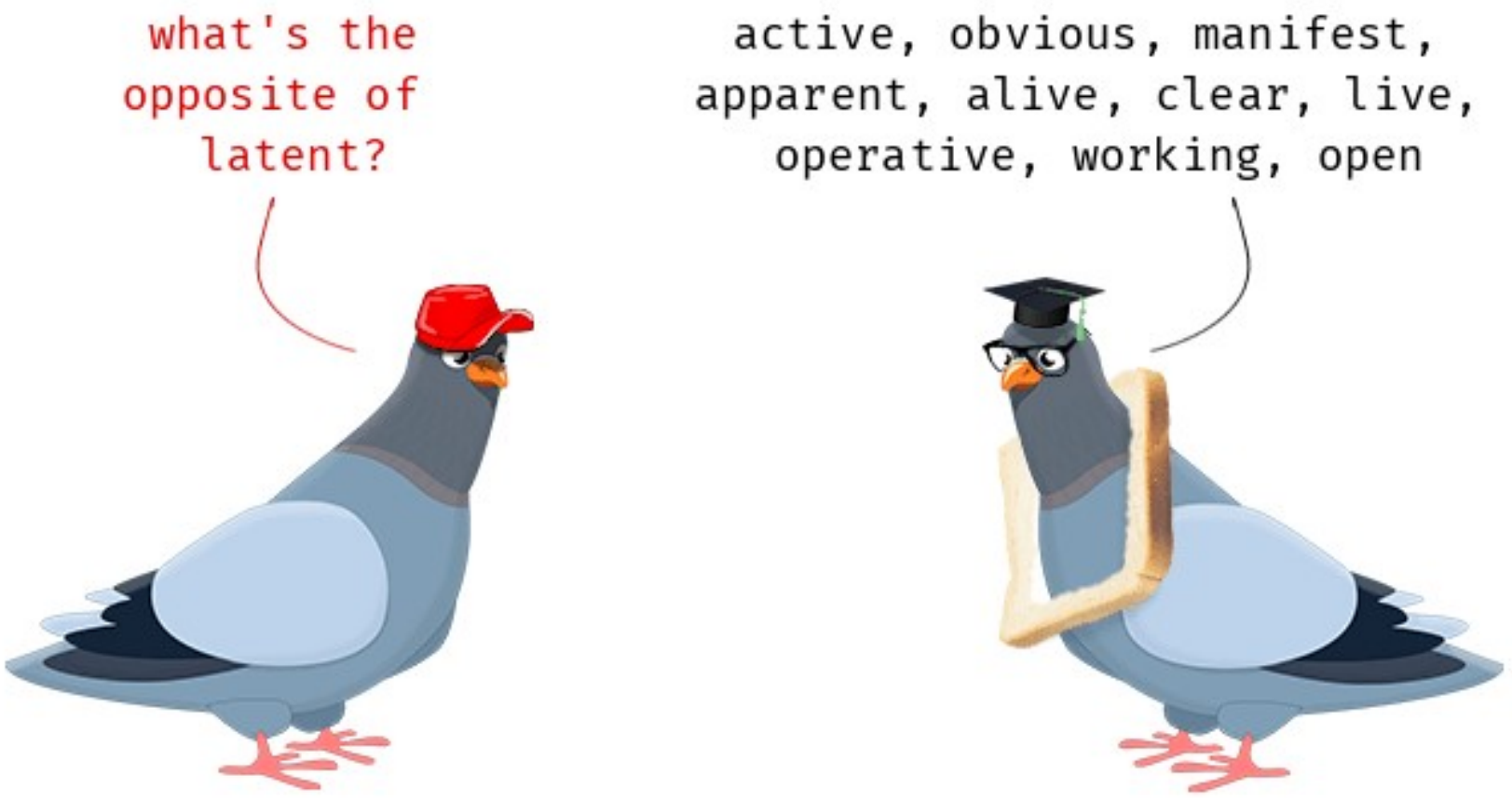


Figure 1: Whittaker 1956

- environmental gradient
- ecological gradient
- other gradients

Latent variables

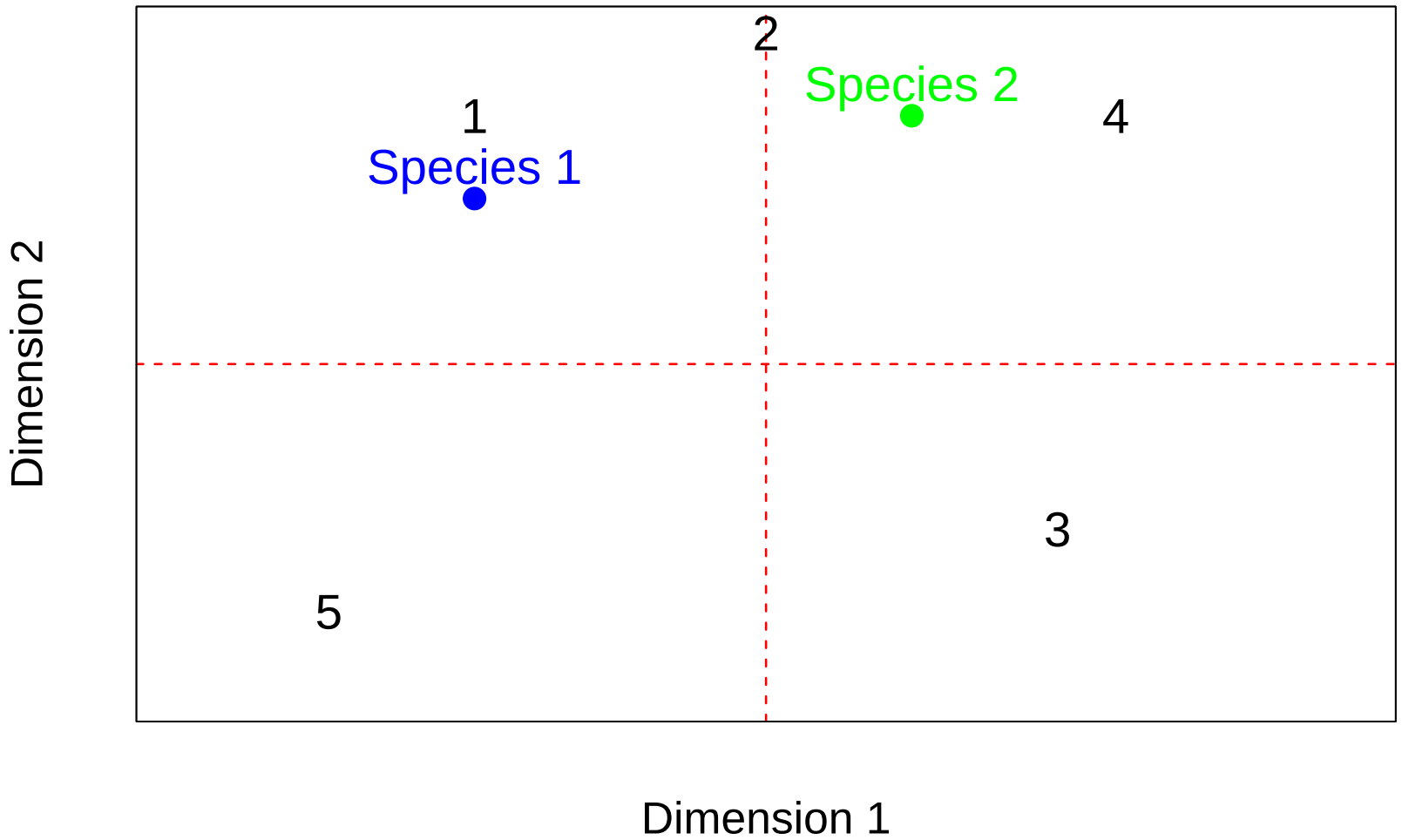


In essence: an unobserved gradient

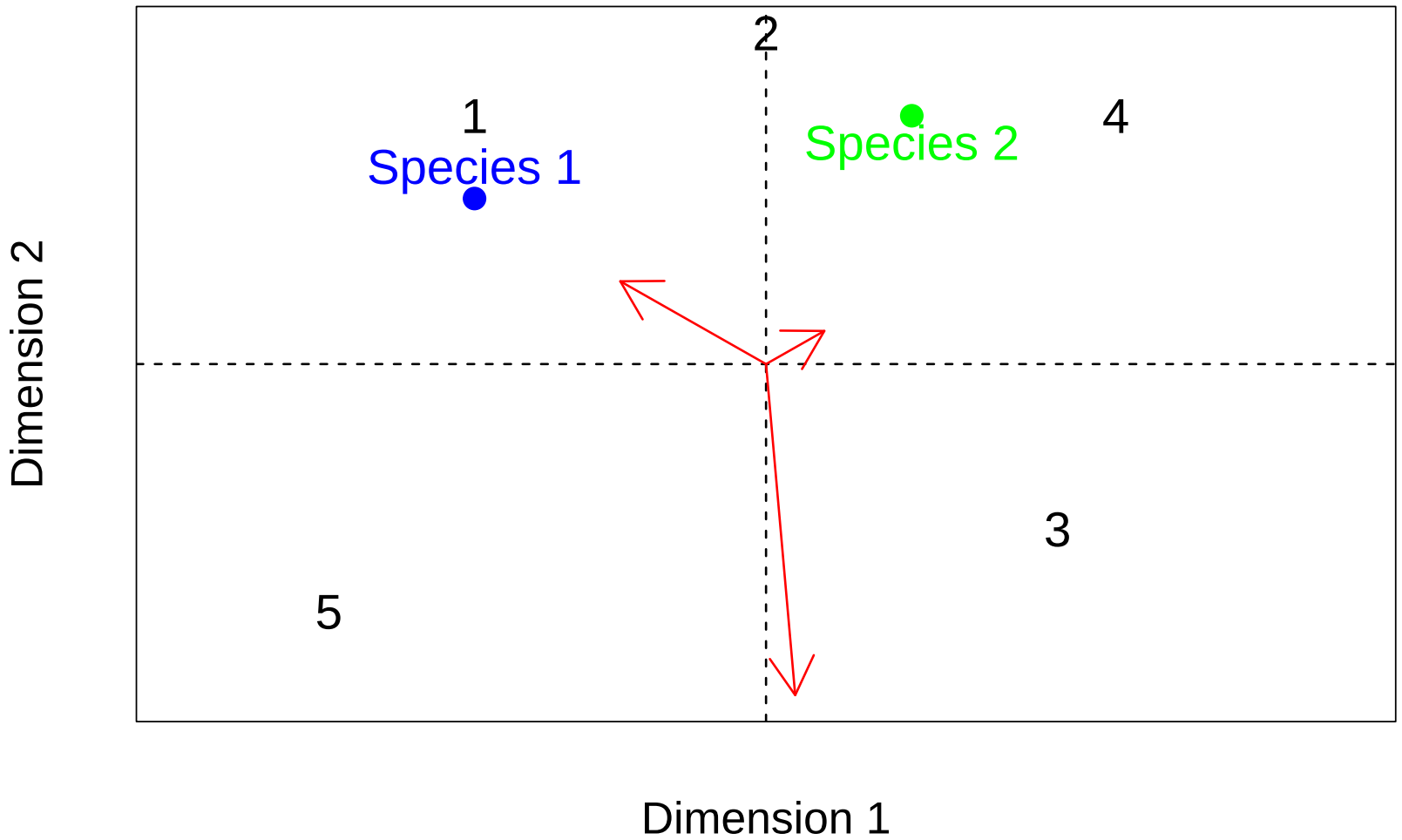
Ecological gradients

"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

Ordination plot

We use it to:

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

Species that are far away: negative association
Species that are closeby: positive association

When to use ordination

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

1. We want to estimate 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

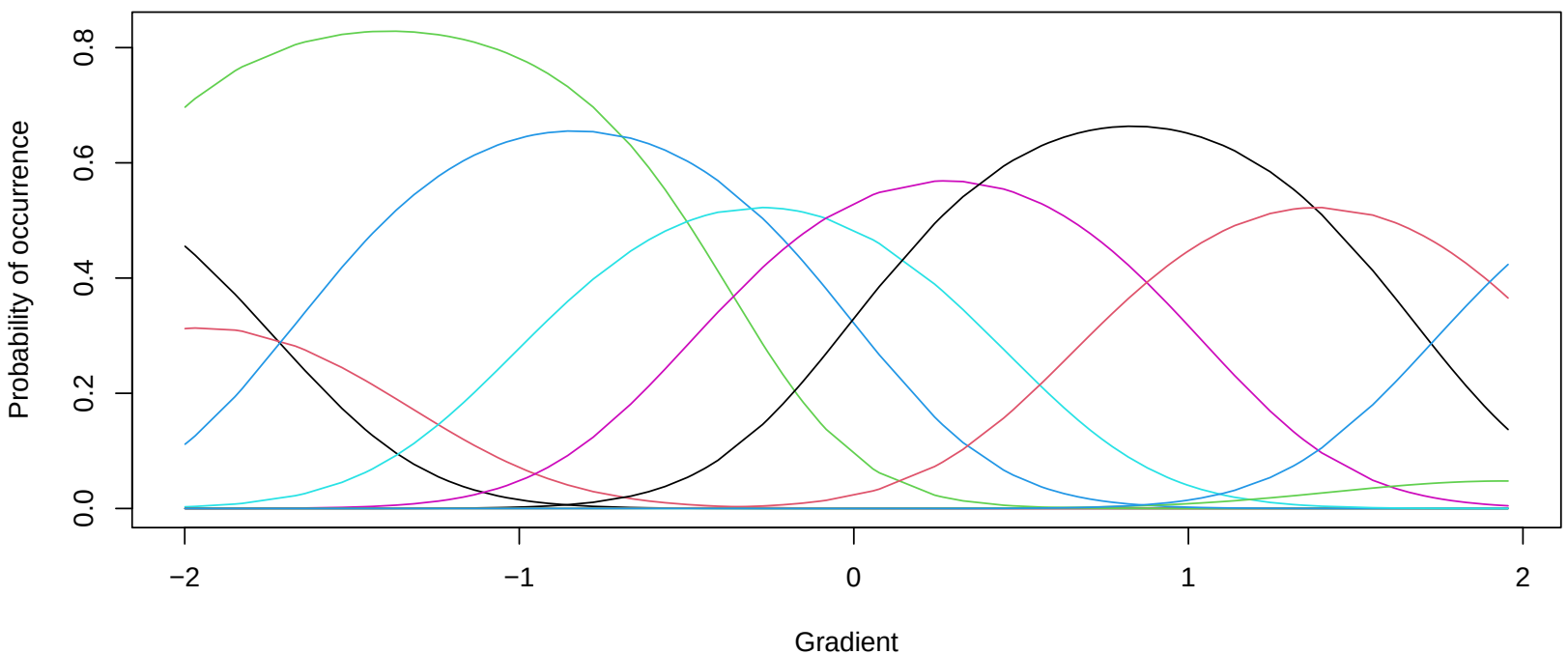
Indirect gradient analysis: patterns in species composition that may be due to environment, but without studying environmental variables
Direct gradient analysis: estimate how species are affected by environmental variables

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



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)

They approximately implement:

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

This makes for better comparison with GLLVMs

Classical ordination

Traditionally the go-to method for multivariate analysis

From p columns to $d << m$ dimensions

- Principal Component Analysis (PCA; Pearson 1901)
- Factor Analysis (FA; Spearman 1904)
- Correspondence Analysis (CA; Hirschfeld 1935)
- Non-metric Multidimensional Scaling (NMDS; Kruskal 1964a,b)
- Principal Coordinate Analysis (PCoA; Gower 1967)
- Detrended Correspondence Analysis (DCA; Hill and Gauch 1980)

Problems with classical methods

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Forum | Open Access

The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017)

David I. Warton, Francis K. C. Hui

First published: 26 July 2017 | https://doi.org/10.1111/2041-210X.12843 | Citations: 16

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Free Access

Distance-based multivariate analyses confound location and dispersion effects

David I. Warton, Stephen T. Wright, Yi Wang

First published: 06 June 2011 | https://doi.org/10.1111/j.2041-210X.2011.00127.x | Citations: 627

Correspondence site: http://www.respond2articles.com/MEE/

From JSDM to latent variable model

JSDMs focus on species, but is an ordination:

$$g\{\mathbb{E}(\mathbf{y}_i|\boldsymbol{\epsilon}_i)\} = \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_i, \qquad \text{where } \boldsymbol{\epsilon}_i = \mathbf{z}_i^\top \boldsymbol{\Gamma}^\top \tag{2}$$

- $\mathbf{z}_i$ are site scores; coordinates in the ordination
- $\boldsymbol{\Gamma}$ is a matrix of species loadings; it forms the associations

From JSDM to latent variable model

JSDMs focus on species, but is an ordination:

$$g\{\mathbb{E}(\mathbf{y}_i|\boldsymbol{\epsilon}_i)\} = \boldsymbol{\beta}_0 + \boldsymbol{\epsilon}_i, \qquad \text{where } \boldsymbol{\epsilon}_i = \mathbf{z}_i^\top \boldsymbol{\Gamma}^\top \tag{2}$$

- $\mathbf{z}_i$ are site scores; coordinates in the ordination
- $\boldsymbol{\Gamma}$ is a matrix of species loadings; it forms the associations

1. Ordination and JSDM use the same statistical framework
2. The models take a different angle
3. JSDM **can** be an LVM, ordination **is** an LVM

A new approach!..or is it?

- Community ecology has been doing it for a hundred years
- e.g. Forbes (1907) or Goodall (1954)
- Walker and Jackson (2011): Random-effects ordination!
- Hui et al. (2015): Model-based unconstrained ordination

BIOMETRICS 41, 859–873
December 1985

Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Model

Cajo J. F. ter Braak
TNO Institute of Mathematics, Information Processing and Statistics,
P. O. Box 100, 6700 AC Wageningen, The Netherlands

ECOLOGICAL MONOGRAPHS
ECOLOGICAL SOCIETY OF AMERICA

Article

Random-effects ordination: describing and predicting multivariate correlations and co-occurrences

Steven C. Walker ✉, Donald A. Jackson

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Special Feature: New Opportunities at the Interface Between Ecology and Statistics | Free Access

Model-based approaches to unconstrained ordination

Francis K.C. Hui ✉, Sara Taskinen, Shirley Pledger, Scott D. Foster, David I. Warton

First published:23 July 2014 | https://doi.org/10.1111/2041-210X.12236 | Citations: 57

Generalized Linear Latent Variable Model

- Still a mixed-effects model (so can incorporate random effects)
- d latent variables, usually treated as random-effect
- Produces ordination
  - No varimax rotation
  - Much more flexible
  - Clear link with the original data, no distances
- We can use tools from ordination and regression

JSDM vs. ordination

JSDMs build more heavily on SDMs than on traditional multivariate analysis Ordination methods have been criticised for being too descriptive rather than predictive nature

Ovaskainen and Abrego 2021

JSDM vs. ordination

JSDMs build more heavily on SDMs than on traditional multivariate analysis Ordination methods have been criticised for being too descriptive rather than predictive nature

Ovaskainen and Abrego 2021

ordination did it first Walker and Jackson 2011

JSDM vs. ordination

The differences is in how we think of the model:

- Do we formulate on the basis of latent variables or associations
- Do we look at patterns in the ordination, or patterns on a map?
- Do we believe the "axes" have meaning, or not?
- The scale at which we operate: local or marcoecological
- Is the **the sampling process** considered?

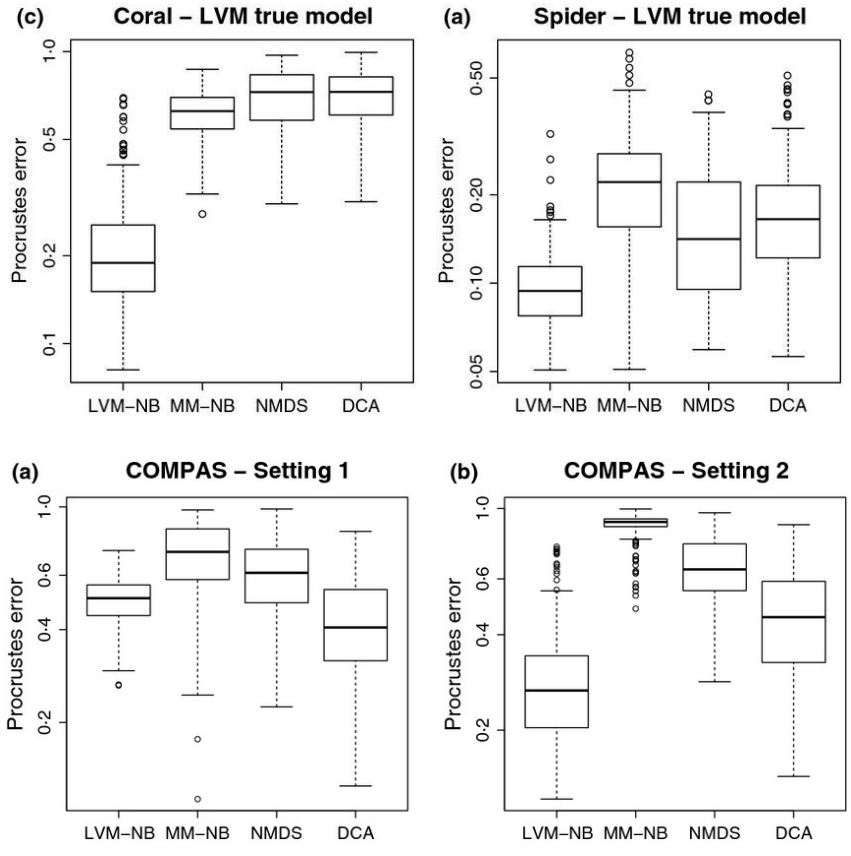
JSDM vs. ordination

The differences is in how we think of the model:

- Do we formulate on the basis of latent variables or associations
- Do we look at patterns in the ordination, or patterns on a map?
- Do we believe the "axes" have meaning, or not?
- The scale at which we operate: local or marcoecological
- Is the **the sampling process** considered?

Both of the angles have a lot to teach us about community ecology

Unconstrained ordination



Figures from **Hui et al. 2015**

Example 1

Wadden sea data Dewenter et al. (2023)

- Abundance (counts) or Biomass of macrozoobenthos
- Covariates
- Transects at islands (Norderney, Spiekeroog, Wangerooge)

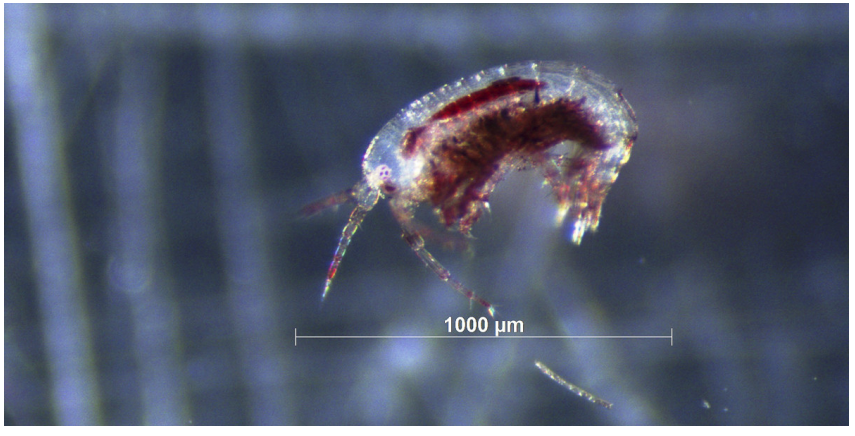


Figure 2: nioz.nl

```
Ya <- read.csv("../data/waddenY.csv")[,-c(1:2)];
Ya <- Ya[, colSums(ifelse(Ya==0,0,1))>2]
X <- read.csv("../data/waddenX.csv")
X[,unlist(lapply(X,is.numeric))] <- scale(X[,unlist(lapply(X,is.numeric))])
```

Study design

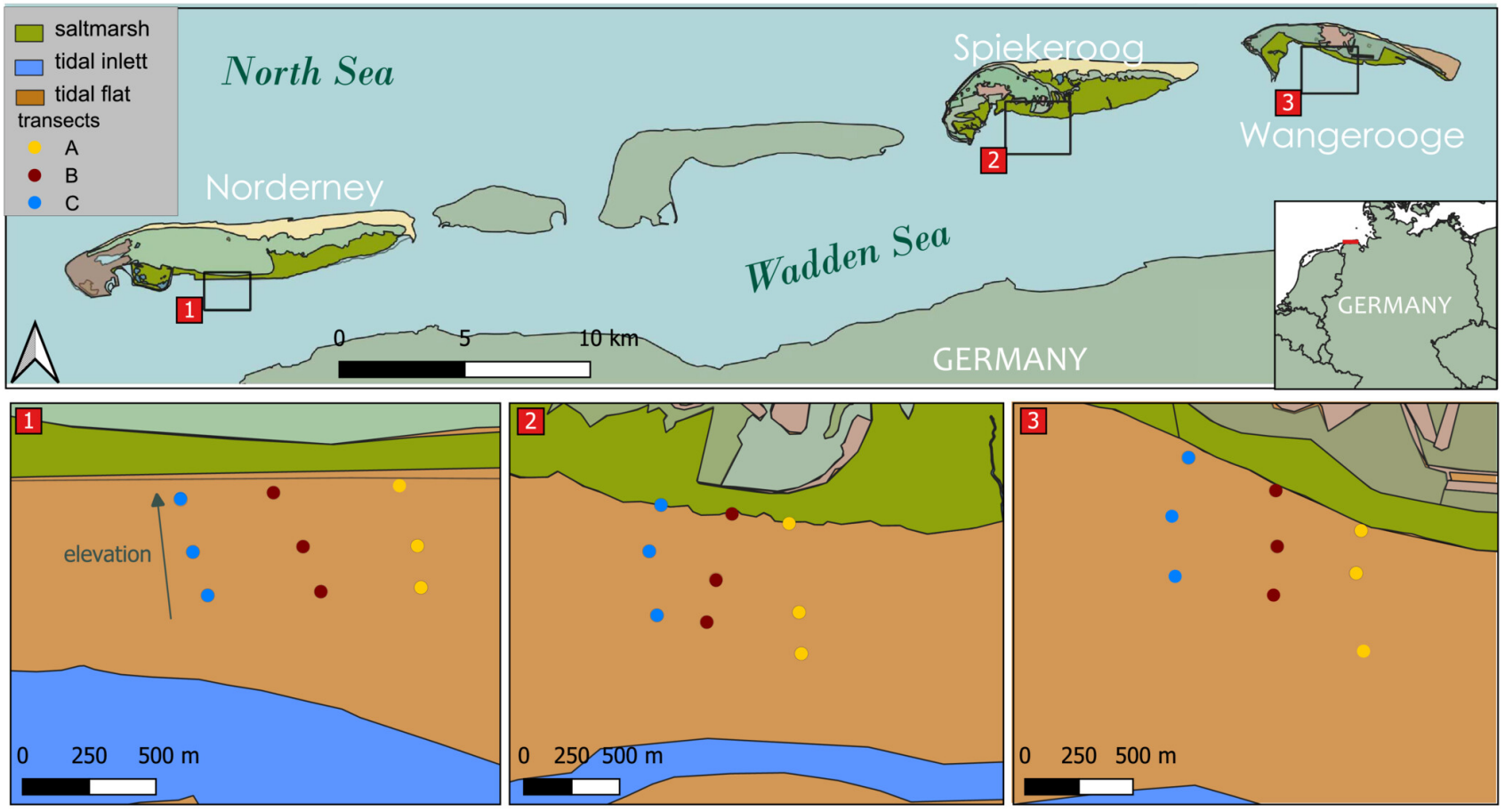
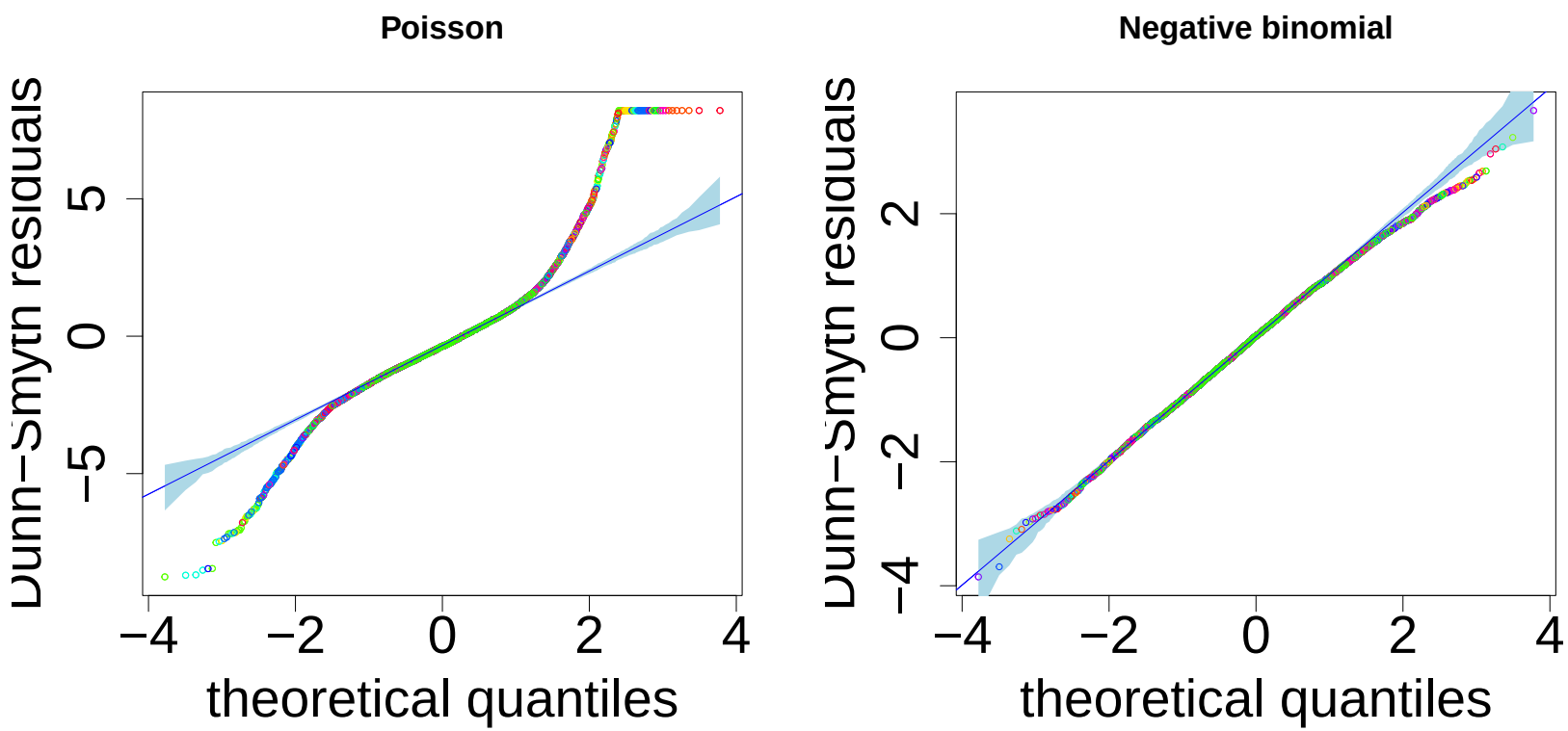


Figure 3: Dewenter et al. Fig 1.

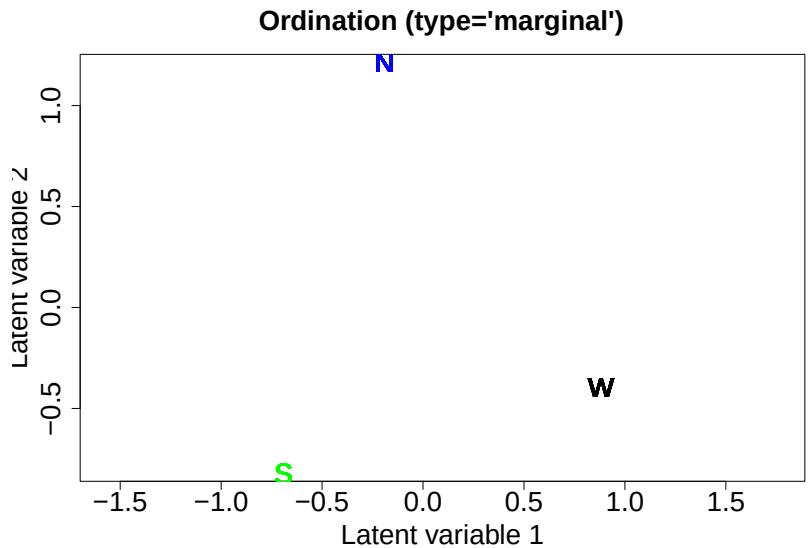
Unconstrained ordination

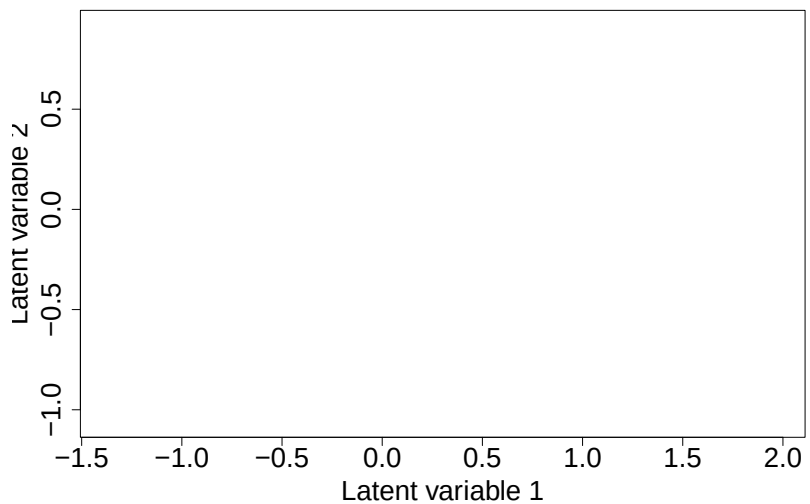
```
plot(uord1, which = 2, caption = "Poisson", cex = 2, cex.lab = 2, cex.axis
plot(uord2, which = 2, caption = "Negative binomial", cex = 2, cex.lab = 2
```



Group-level unconstrained ordination

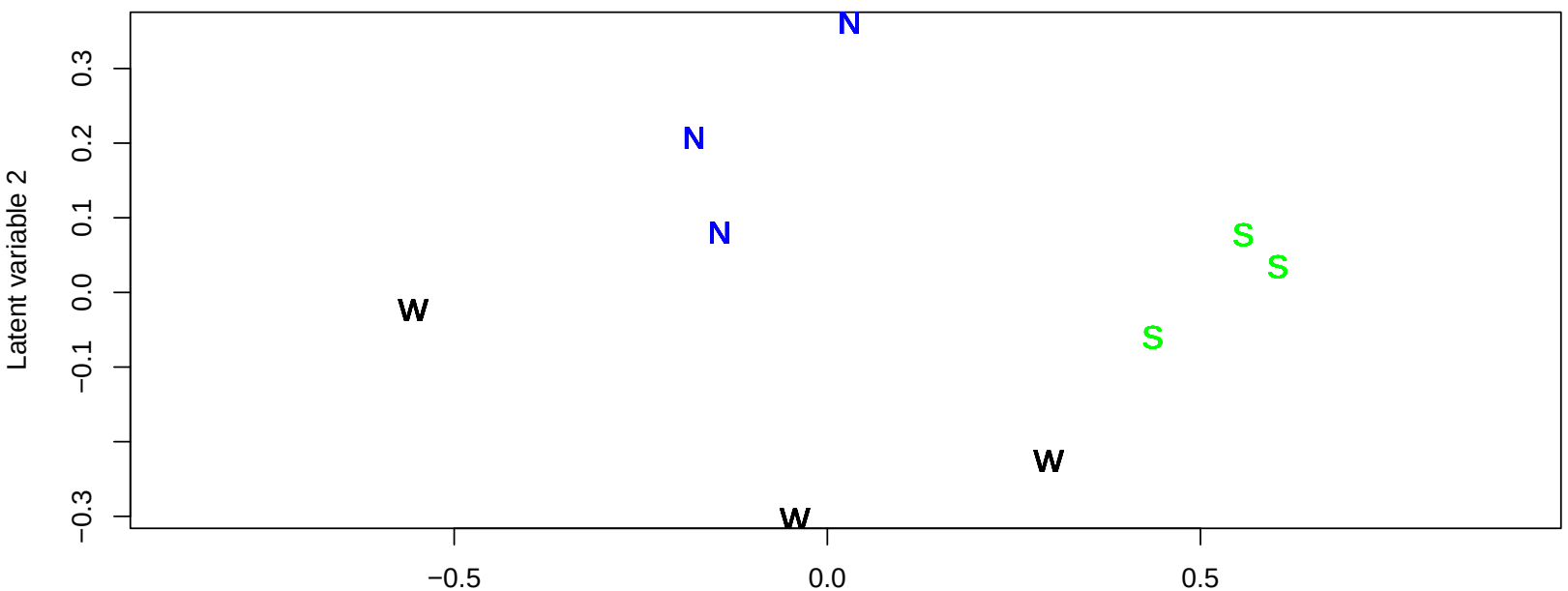
```
uord3 <- gllvm::gllvm(Ya, X = X,
         num.RR = 2, lv.formula = ~diag(1|island), randomB = "LV",
         family = "negative.binomial", disp.formula = rep(1,ncol(Ya)))
uord4 <- gllvm::gllvm(Ya, X = X,
         num.RR = 2, lv.formula = ~diag(1|island/transect), randomB = "LV"
         family = "negative.binomial", disp.formula = rep(1,ncol(Ya)), sta
```





Quadratic unconstrained ordination

```
uord5 <- gllvm::gllvm(Ya, X = X, quadratic = TRUE,
         num.RR = 2, lv.formula = ~diag(1|island/transect), randomB = "LV"
         family = "negative.binomial", disp.formula = rep(1,ncol(Ya)), sta
```



Residual ordination

Any effect not "in" the ordination (but in the model) "filters" the ordination (i.e., conditioning)

- We might want to account for pseudo-replication (transect as in the example)
- Exclude the effect of some "confounder"
- When we do not have enough data to estimate all effects outside of the ordination, but some

Residual ordination: the model

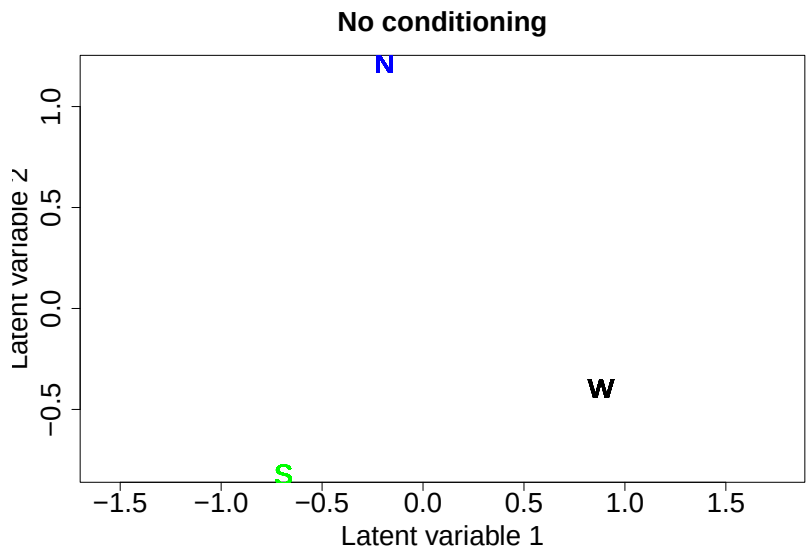
$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \quad (3)$$

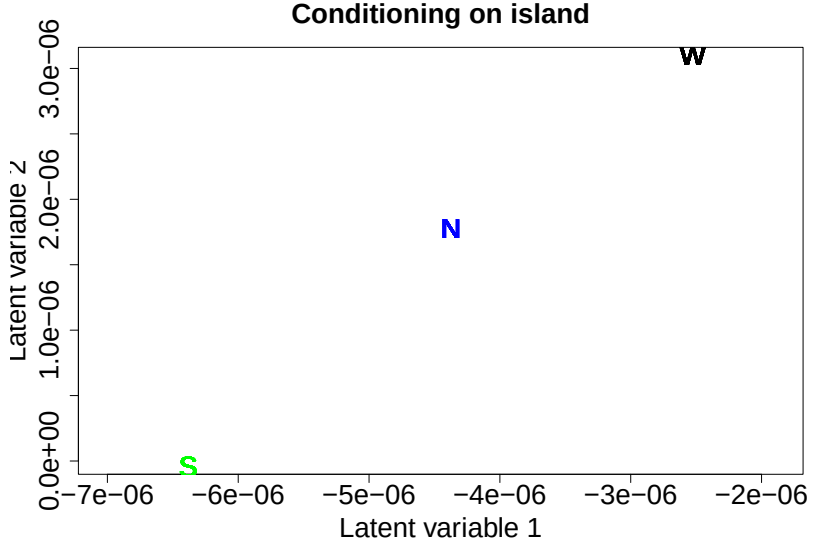
Covariates with species-specific coefficients ("conditioning")

- No longer an unconstrained ordination: covariates are involved
- For binary data it is a JSDM (more later)
- We can also use it to adjust the ordination (take an effect out)
- We estimate species-specific effect $\boldsymbol{\beta}_j$ so need a good amount of data

Example with Wadden data: conditioning

```
gllvm::ordiplot(uord3, symbols=TRUE, pch=X$island, arrow.scale=0, s.cex =
```





Residual ordination

In the same way, we can decide to "pull" any effect in or out of the ordination.

Constrained

Goal: to determine if (how) environment affects community composition

Problem: many possible drivers (if not, multivariate GLM would do the trick)

- Why are sites different?
- Why do species co-occur (or not)?
- Which components of the environment are most important for the community?

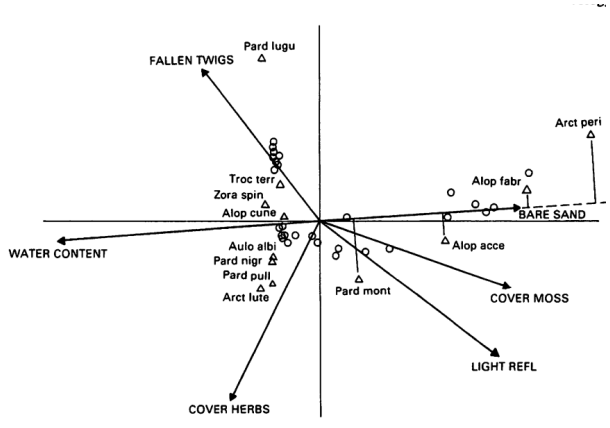


FIG. 1. The distribution of 12 species of hunting spiders caught in pitfall traps in a Dutch dune area.

Figure 4: ter Braak 1986

Constrained ordination

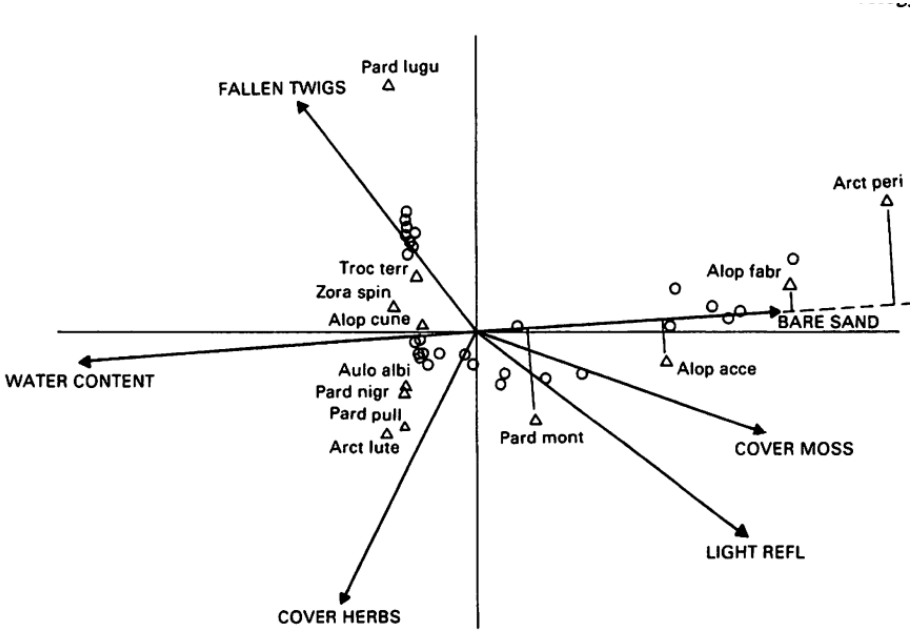


FIG. 1. The distribution of 12 species of hunting spiders caught in pitfall traps in a Dutch dune area.

Figure 5: ter Braak 1986

Now three quantities: so we call this a **triplot**. The arrows show the association of covariates to the axes.

Methods for constrained ordination

- Redundancy Analysis (Rao 1964)
- Canonical Correspondence Analysis (ter Braak 1986)
- RR-GLMs (Yee et al. 1996,2003,2010,2015)
- Row-column interaction models (Hawinkel et al. 2019)
- GLLVMs (van der Veen et al. 2023)

Canonical Correspondence Analysis

- Although RDA was developed much earlier, CCA has been the leading constrained ordination method
- ter Braak (1986) developed CCA as a combination of ordination and regression
- Each axis is restricted (constrained) by covariate information
- CCA approximates Gaussian Ordination (i.e., to the unimodal model, Johnson and Altman, 1999)

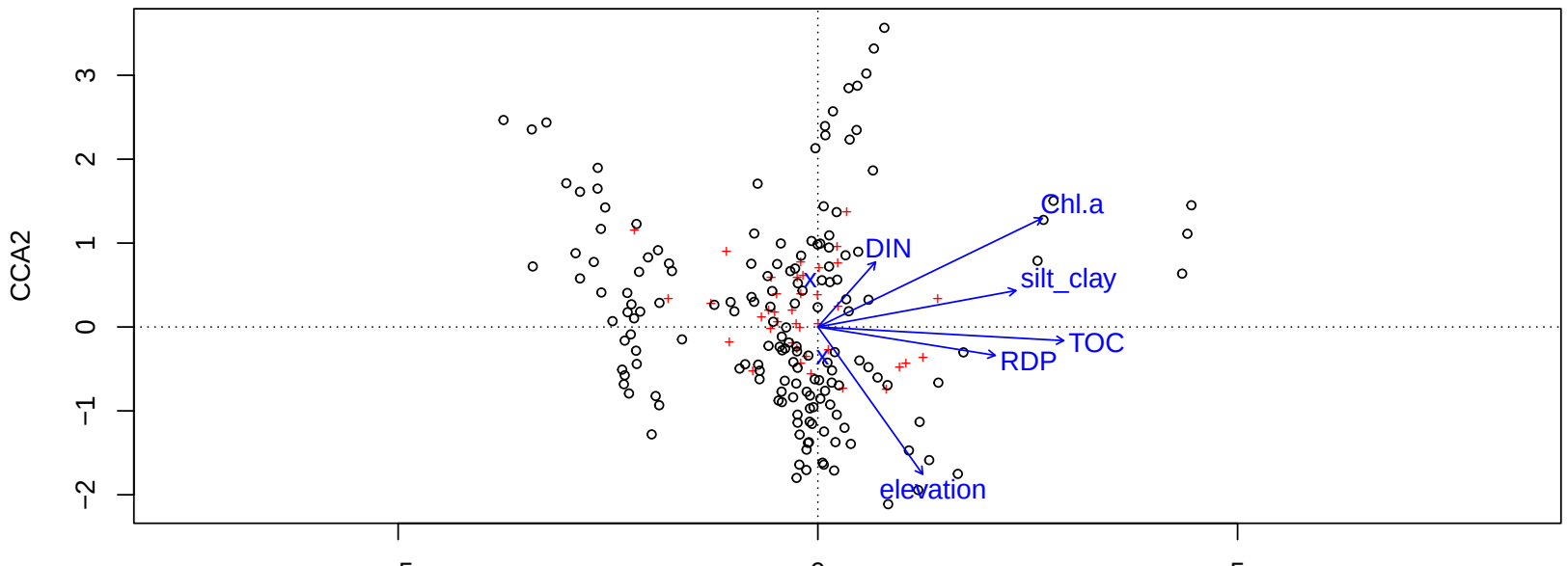
Canonical Correspondence Analysis: arrows

The covariate coefficients **B** are referred to as **canonical** coefficients.

- `vegan` does not use these for plotting
- Instead it uses sample correlation coefficients as recommended by ter Braak (1986)
- The canonical coefficients can be "unstable" due to multicollinearity
- In `gllvm`, we do use **B** (more details later)

Canonical Correspondence Analysis example

```
cca <- vegan::cca(Ya~Condition(transect)+Condition(island)+
                    elevation+TOC+DIN+RDP+Chl.a+silt_clay+season, X)
vegan::ordiplot(cca)
```



Different scores

The CCA algorithm gives raise to two sets of site scores:

1) Linear combination (LC) scores (similar to model-based constrained ordination)
2) Weighted average (WA) scores (similar to model-based concurrent ordination)

WA scores are usually recommended for plotting (Palmer, 1993)

Different scores

The CCA algorithm gives raise to two sets of site scores:

1) Linear combination (LC) scores (similar to model-based constrained ordination)
2) Weighted average (WA) scores (similar to model-based concurrent ordination)

WA scores are usually recommended for plotting (Palmer, 1993)

I disagree

Constrained ordination

In R e.g.

For constrained ordination:

- `vegan` - classical methods
- `VGAM` - cool algorithm, faster than `gllvm`, but not so easy to use (and no random effects)
- `gllvm` - easy to use

Constrained ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{4}$$

So far, we have assumed that the latent variable is estimated by species information alone: $\mathbf{z}_i = \boldsymbol{\epsilon}_i$

Constrained ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{4}$$

So far, we have assumed that the latent variable is estimated by species information alone: $\mathbf{z}_i = \boldsymbol{\epsilon}_i$

Constrained ordination instead assumes that we also have covariates in the ordination: $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i$

Constrained ordination: the model

Plugging in $\mathbf{z}_i = \boldsymbol{\epsilon}_i$ we get:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top \mathbf{B} \boldsymbol{\gamma}_j \tag{5}$$

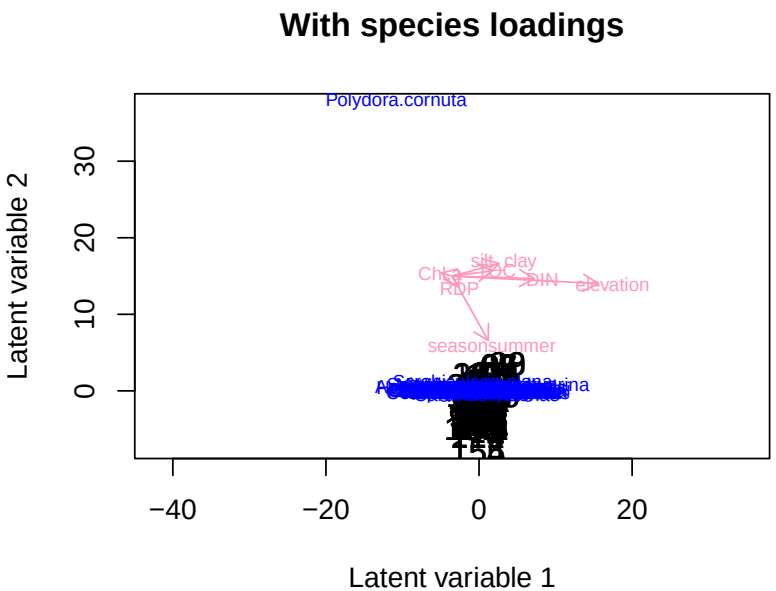
From this we see that $\boldsymbol{\beta}_j \overset{d}{\approx} \mathbf{B}\boldsymbol{\gamma}_j$

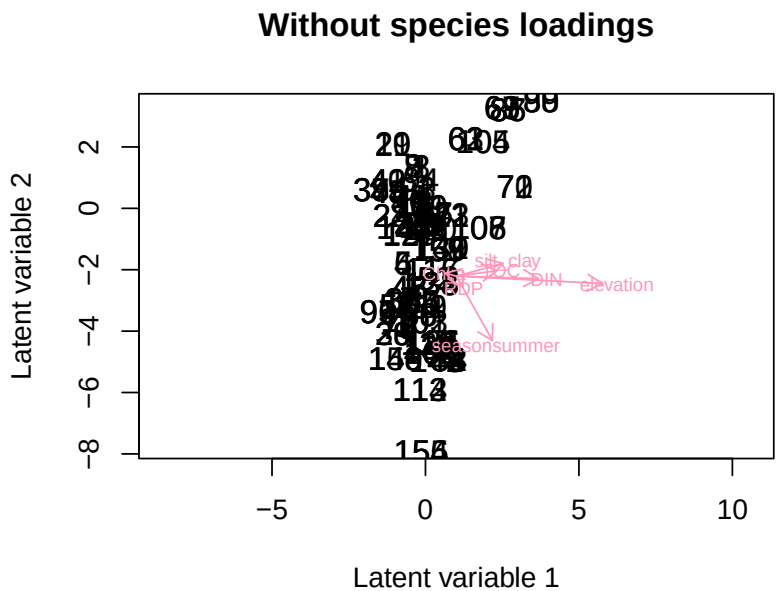
- These are the (reduced rank) approximated species-specific covariate coefficients
- We can extract these, and inspect them with statistical uncertainty
- So we use information across the whole community, to estimate species-specific responses

Constrained ordination with `gllvm`

Example with Wadden data

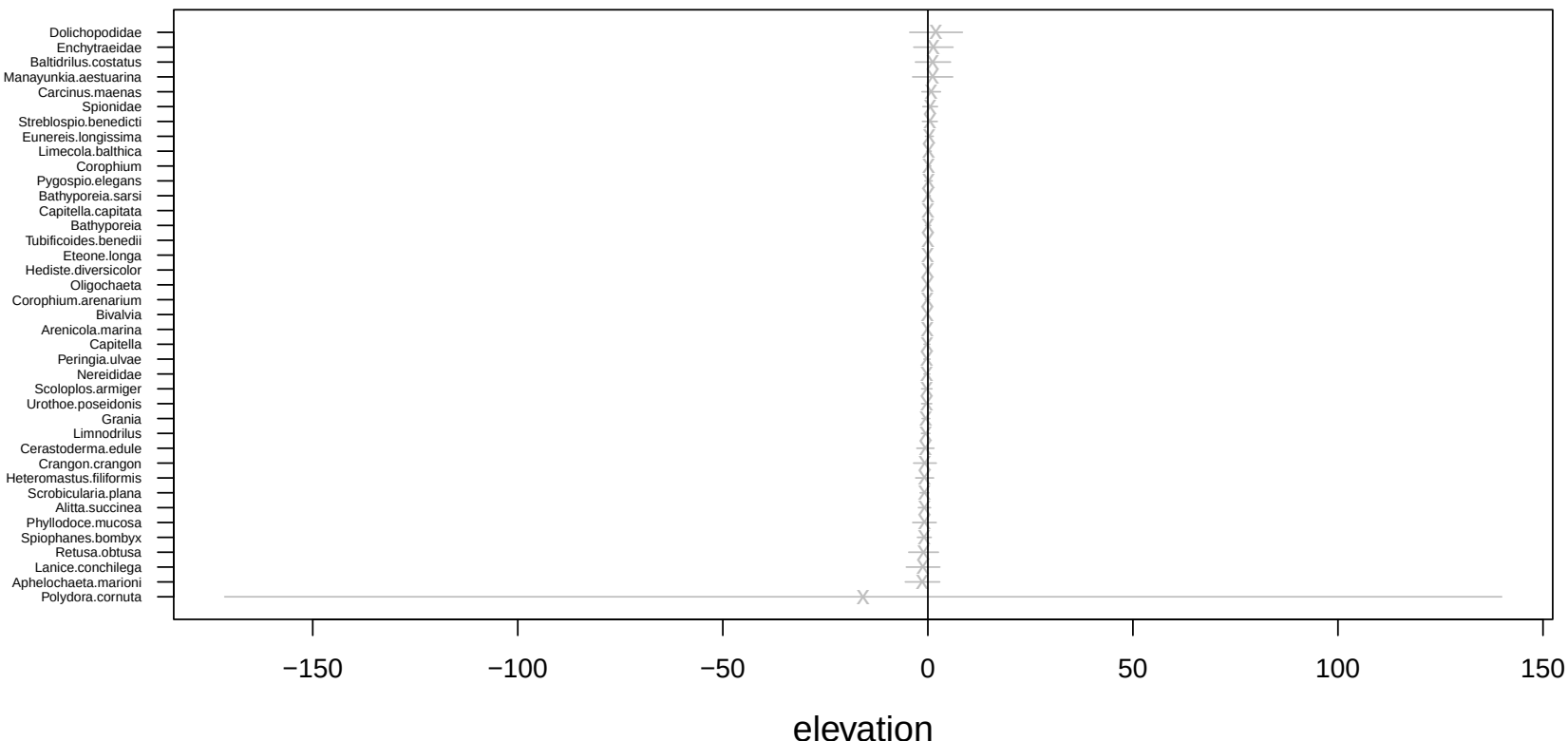
```
cord <- gllvm::gllvm(y = Ya, X, num.RR = 2, family = "negative.binomial",
        lv.formula = ~elevation+TOC+DIN+RDP+Chl.a+silt_clay+season,
        row.eff = ~(1|island/transect), studyDesign = X[, c("island","tran
        disp.formula = rep(1,ncol(Ya)))
```





Example with Wadden data

```
gllvm::coefplot(cord, which.Xcoef="elevation")
```



Constrained ordination

- Species effects can be retrieved for any covariate
- Extreme results occur, usually due to insufficient data
- GLLVMs picks up on extreme clustering -very- well

Constrained ordination

The first implementation of CO that can be combined with random effects

- Random site effects (outside ordination)
- Random canonical coefficients (more in a few slides)

Common misconception

Post-hoc relating unconstraied ordination axes to environmental covariates is **not** equivalent to a constrained ordination

Also it is bad practice: please do not do it. Instead **adjust your model**.

Hybrid ordination

- Incorporate both constrained and unconstrained ordination
- But without explicit connection
- Default in `vegan` and you can also do it in `gllvm` (use both `num.RR` and `num.lv`)

Concurrent

- In practice, constrained and unconstrained ordination are often combined into an analysis
- Variation not due to the environment is discared, while potentially of large importance
- *Concurrent ordination* is a new type of ordination method that combines unconstrained and constrained ordination

Concurrent ordination

Concurrent:'existing or happening at the same time' (Oxford's dictionary)

Concurrent ordination

Concurrent:'existing or happening at the same time' (Oxford's dictionary)

1. Suggested in van der Veen et al. (2023)
2. Performs both unconstrained and constrained ordination **simultaneously**

3. Ordination axes have **measured** and **unmeasured** components
4. Covariates *inform* rather than *constrain*
5. Separates out drivers of community composition

Concurrent ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

Concurrent ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

1. $\mathbf{z}_i = \boldsymbol{\epsilon}_i$, **unconstrained**

Concurrent ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

1. $\mathbf{z}_i = \boldsymbol{\epsilon}_i$, **unconstrained**
2. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i$, **constrained**

Concurrent ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

1. $\mathbf{z}_i = \boldsymbol{\epsilon}_i$, **unconstrained**
2. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i$, **constrained**
3. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\epsilon}_i$, **concurrent**

Concurrent ordination: the model

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

1. $\mathbf{z}_i = \boldsymbol{\epsilon}_i$, **unconstrained**
2. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i$, **constrained**
3. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\epsilon}_i$, **concurrent**

Often unconstrained and concurrent ordinations are similar

Concurrent ordination: site scores

$$\eta_{ij} = \beta_{0j} + \mathbf{z}_i^\top \boldsymbol{\gamma}_j \tag{6}$$

The model is flexible, $\mathbf{z}_i$ can be all kinds of things.

1. $\mathbf{z}_i = \boldsymbol{\epsilon}_i$, residual
2. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i$, marginal
3. $\mathbf{z}_i = \mathbf{B}^\top \mathbf{x}_i + \boldsymbol{\epsilon}_i$, conditional

Often unconstrained and concurrent ordinations are similar

Concurrent ordination: the model

Essentially a linear mixed-effects model of the latent variable $\mathbf{z}_i$
(hiearchcally)

Concurrent ordination with `gllvm`

Example with Wadden data

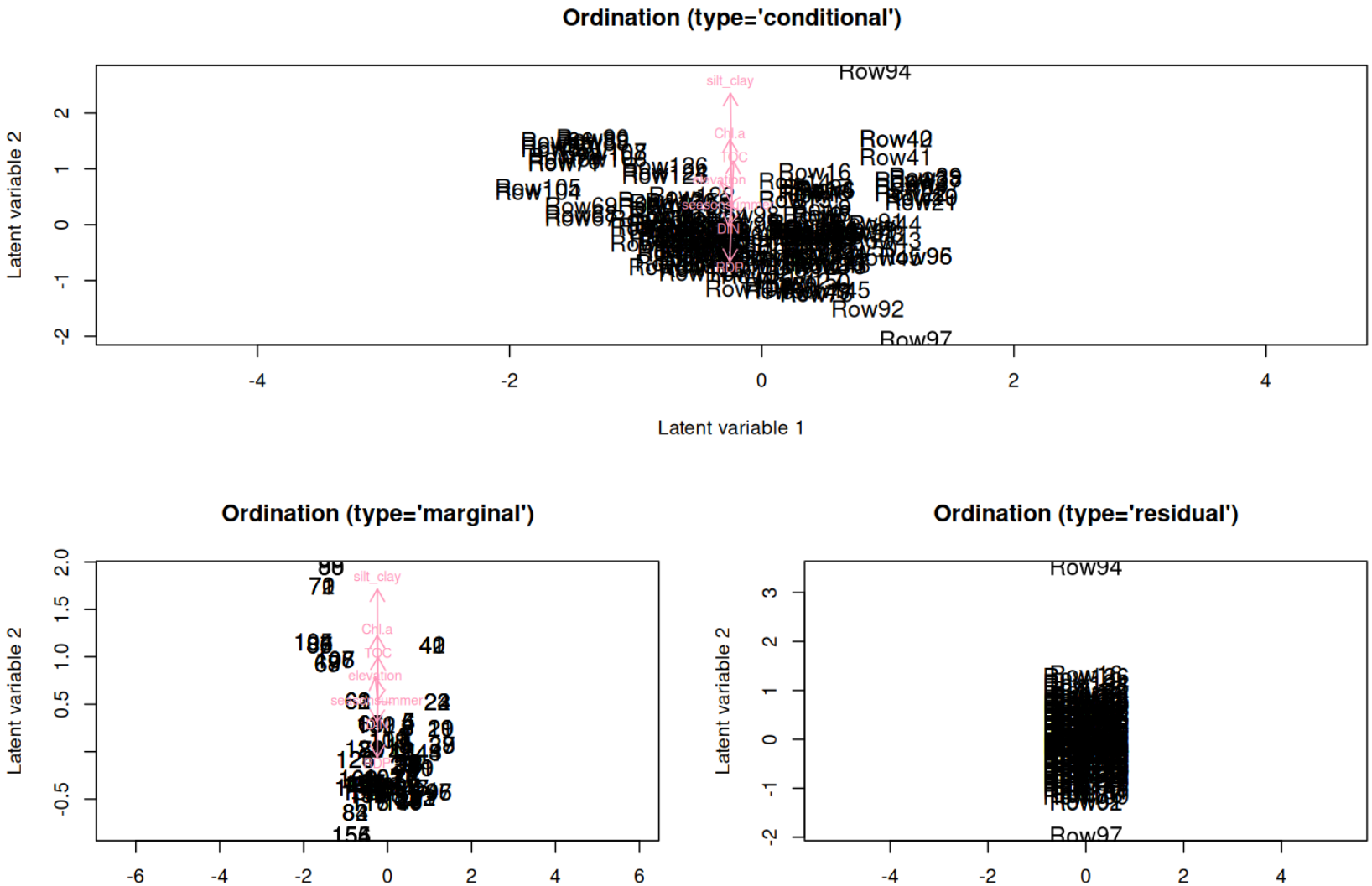
```
cnord <- gllvm::gllvm(y = Ya, X, num.lv.c = 2, family = "negative.binomial
         lv.formula = ~elevation+TOC+DIN+RDP+Chl.a+silt_clay+season,
         row.eff = ~(1|island/transect), studyDesign = X[, c("island","tra
         disp.formula = rep(1,ncol(Ya)), n.init = 3)
coef(cnord, parm="Cancoef")
```

```
##                          CLV1          CLV2
## elevation    -0.0010555096  4.890698e-03
## TOC           0.0001835056  9.143895e-03
## DIN          -0.0003019491 -4.429945e-03
## RDP          -0.0001773025 -1.177149e-02
## Chl.a        -0.0001715056  1.368980e-02
## silt_clay    -0.0001218514  2.333842e-02
## seasonsummer -0.0001556384 -3.185406e-05
```

Example with Wadden data

```
layout(matrix(c(1,1,2,3), 2, 2, byrow=TRUE))
gllvm::ordiplot(cnord, type= "conditional", rotate = FALSE)
gllvm::ordiplot(cnord, type = "marginal", rotate = FALSE)
gllvm::ordiplot(cnord, type = "residual", rotate = FALSE)
```

Example with Wadden data



Random canonical coefficients

We can treat the canonical coefficients as random with `randomB`

- This is usually faster
- Treats the "bouncing beta" problem
- Models correlation between species due to environment

1) LV: canonical coefficients of the same ordination axis come from the same distribution

- Shrinkage over LVs

2) P: canonical coefficients of the same covariate come from the same distribution

- Shrinkage over covariates

3) single: all come from the same distribution

Random canonical coefficients

We can also incorporate **correlation** of the effects. This is a model where we have:

- Associations due to environment `getEnvironCor`
- Correlation between environment effects (requires `lv.formula` lme4-style)
- Which means we can examine the reduced-rank approximated random effects ‘RandomCoefPlot“

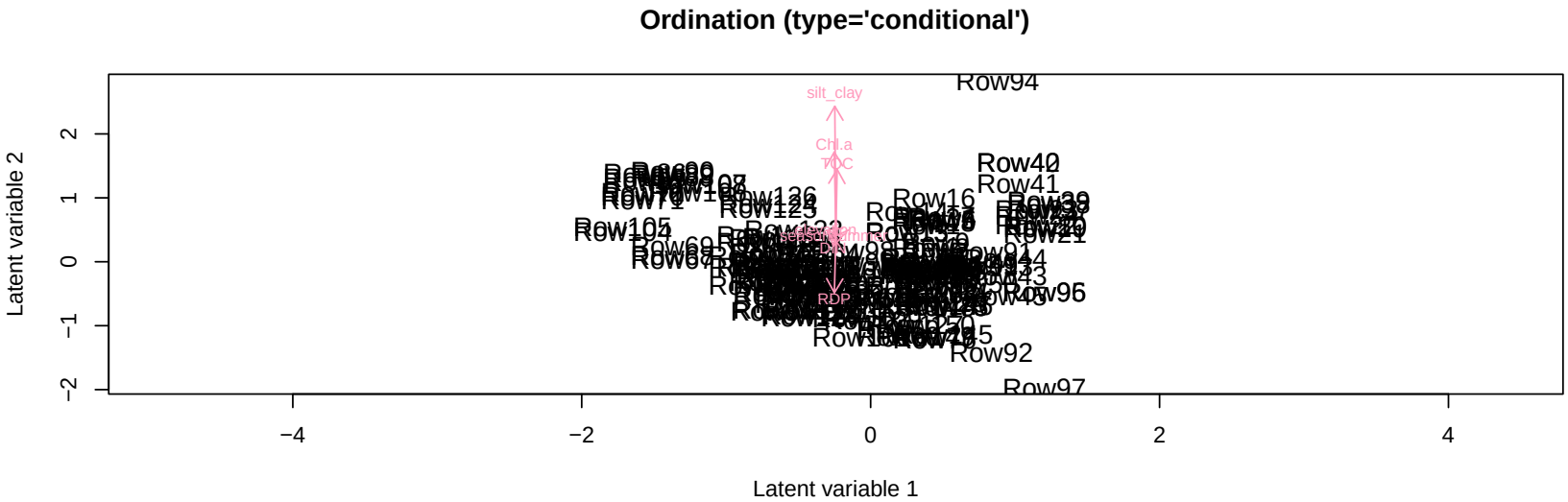
Example with Wadden data

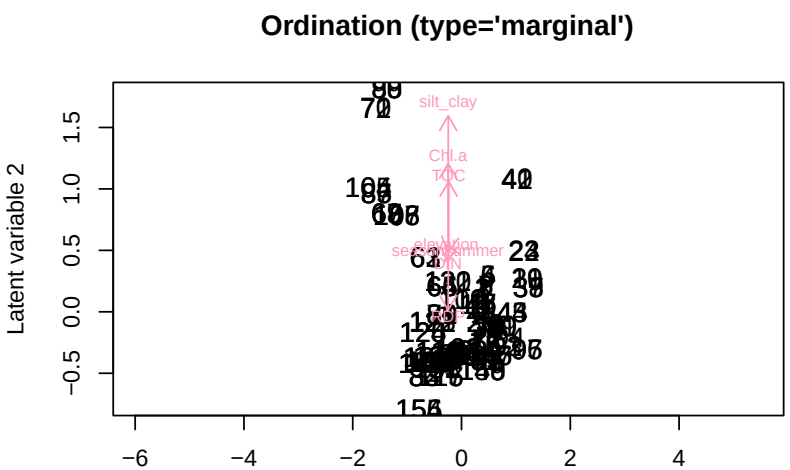
```
cnord2 <- gllvm::gllvm(y = Ya, X = X, num.lv.c = 2, randomB = "LV",
          family = "negative.binomial",
          lv.formula = ~elevation+TOC+DIN+RDP+Chl.a+silt_clay+season,
          row.eff = ~(1|island/transect), studyDesign = X[, c("island","tr
          disp.formula = rep(1,ncol(Ya)), n.init = 3)
```

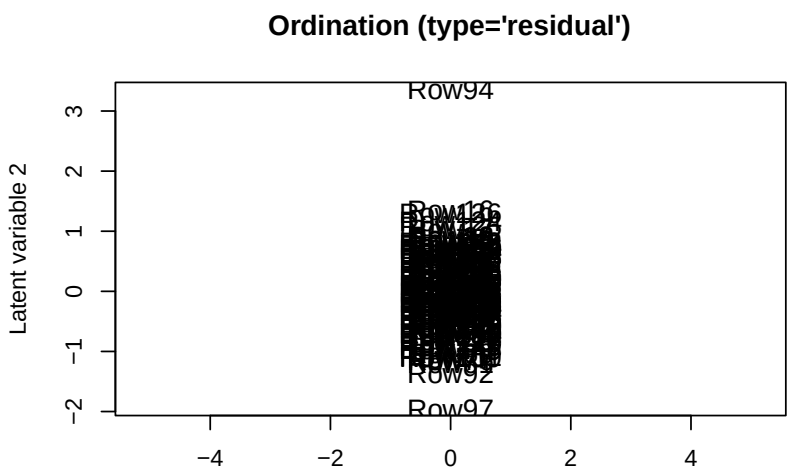
or random-effects style (correlation for effects in same brackets):

```
cnord2b <- gllvm::gllvm(y = Ya, X = X, num.lv.c = 2, randomB="LV",
           family = "negative.binomial",
           lv.formula = ~(0+elevation+TOC+DIN+RDP+Chl.a+silt_clay+season|1
           row.eff = ~(1|island/transect), studyDesign = X[, c("island","t
           disp.formula = rep(1,ncol(Ya)), n.init = 3)
```

Example with Wadden data

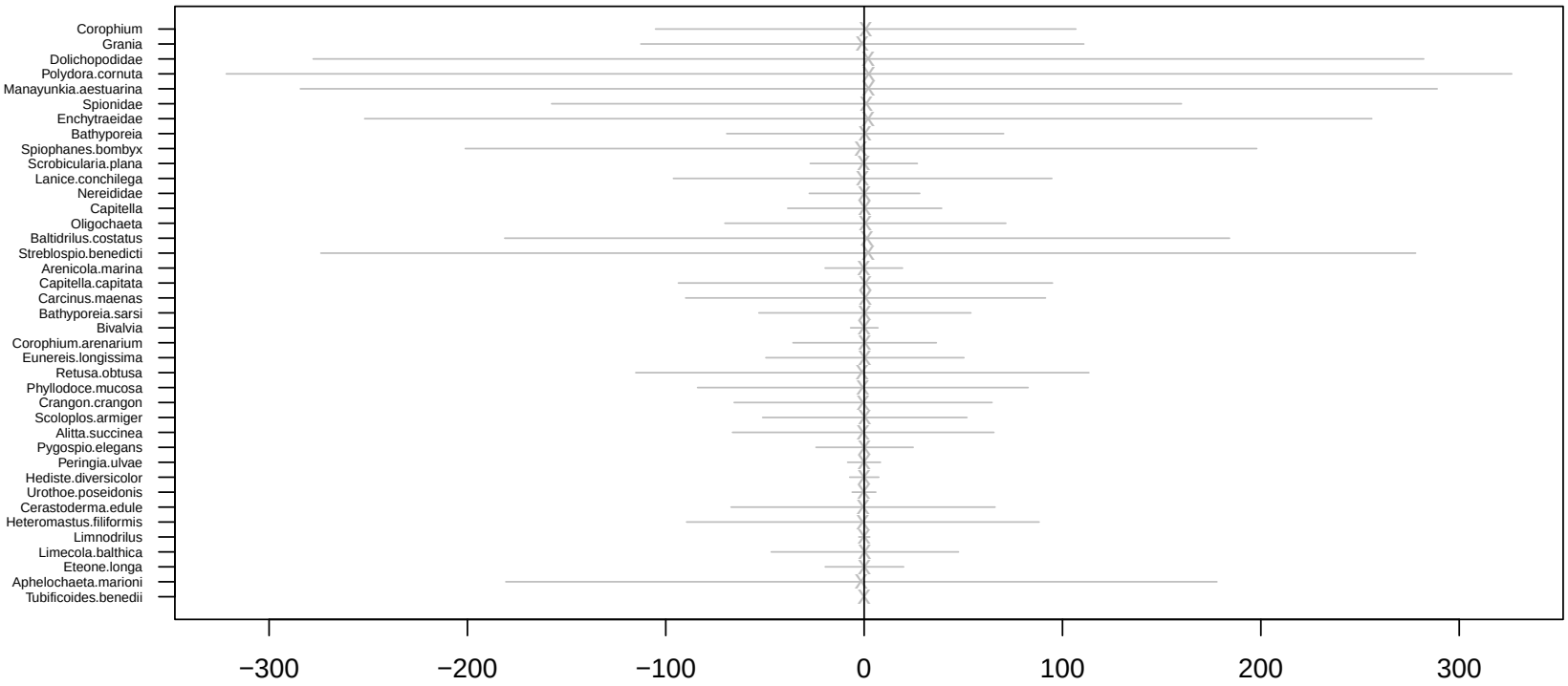






Example with Wadden data

```
gllvm::randomCoefplot(cnord2, which.Xcoef="elevation")
```



Summary

- Ordination has four flavours in GLLVM:
  - Unconstrained ordination
  - Residual ordination
  - Constrained ordination
  - Concurrent ordination (combining unconstrained and constrained)
- Random (correlated) canonical coefficients via the `randomB` argument
- Partial ordination by adding (fixed or random) effects via `formula`
- Or conditioning on species-common (fixed or random) effects via 'row.eff"

How to choose

It all depends what we want out of an ordination.

How to choose

It all depends what we want out of an ordination.

- Exploratory analysis of the community: unconstrained ordination
- Species-environment relationships:
  - residual or constrained ordination
  - concurrent or hybrid ordination
- Associations
  - Residual: unconstrained, residual or concurrent ordination
  - Environmental: constrained or concurrent ordination