Model-based ordination

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

What should I cover in this workshop

- Model-based ordination
- Rotation (and post-hoc rotation)
- Variation explained
- Adding row-effects (and what role it plays)
- Conditioning
- Compositional data (or rather, different data types)
- Double-zero problem

Questions so far?

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

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 sometimes we lack measurements of the environment
- Thus cannot test anything

Figure 1: NIBIO

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.

In contrast to traditional ordination methods, we have a more process-based view (sampling process and ecological process)

Gradients

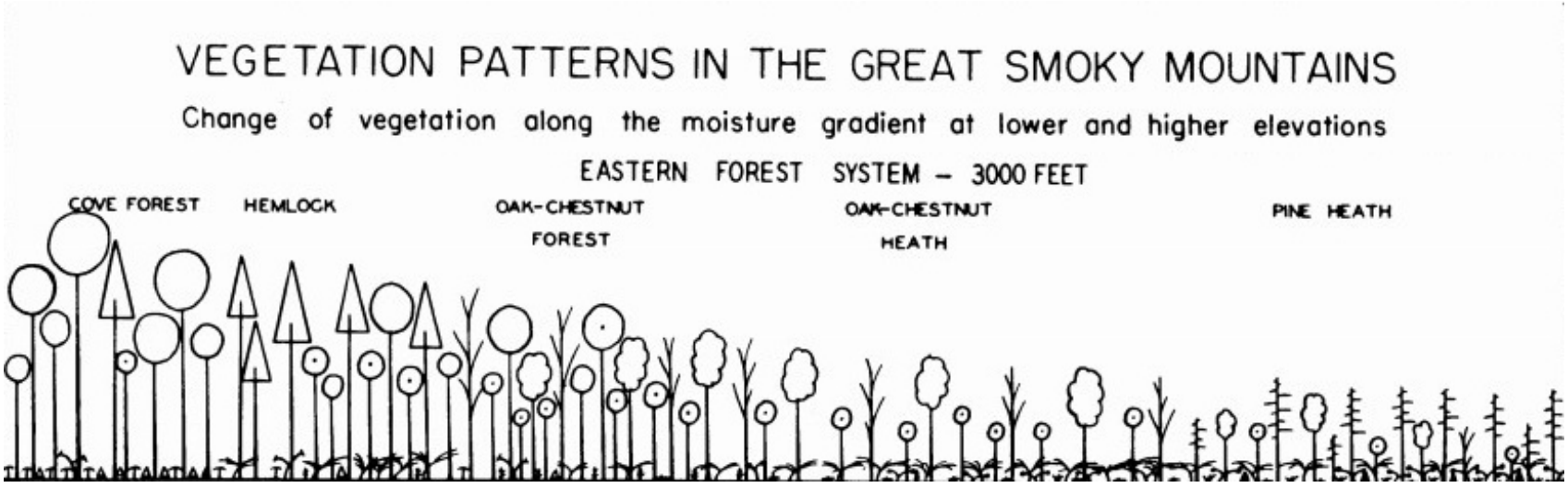


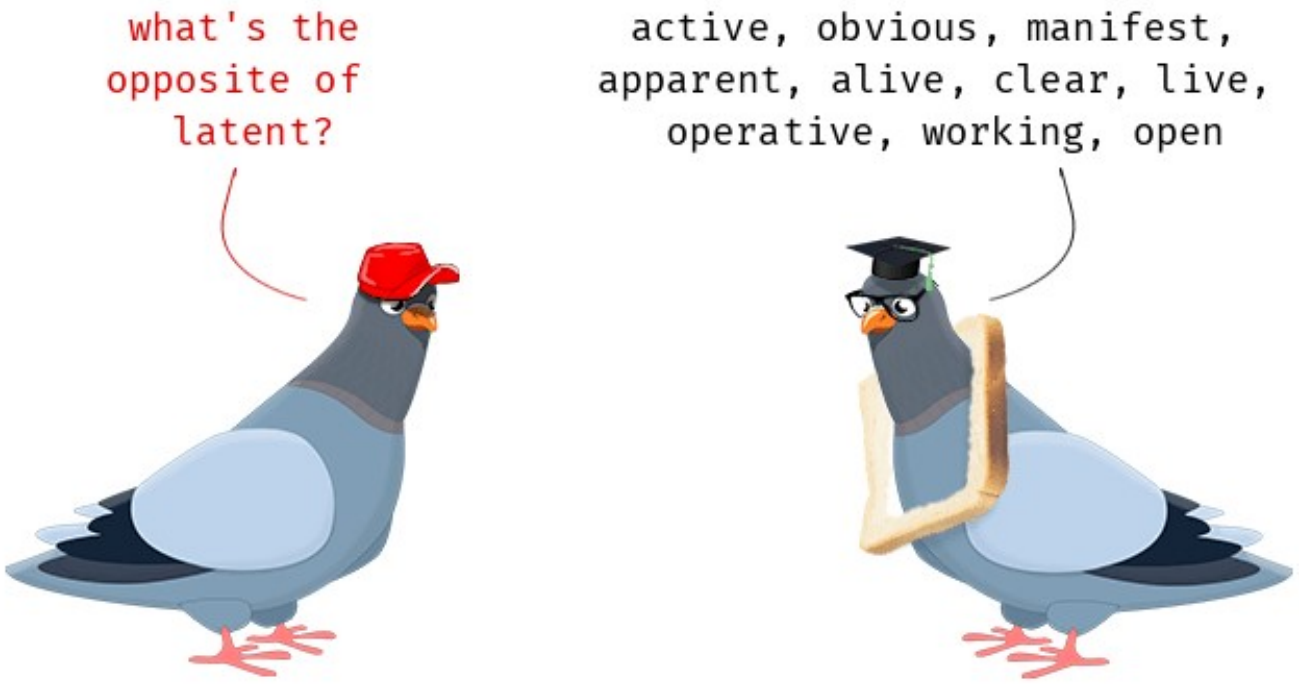
Figure 2: Whittaker 1956

There are different types of gradient, for example:

- Environmental gradient
- Complex ecological gradient
- Coenoclines

Ordination axes

"Ordination axis" has become synonymous to "latent variable"



In essence: an unobserved gradient

Ecological gradients

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

Ecological gradients

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

Which is synonymous to saying "we can probably get a way with fitting a JSDM using only a few dimensions"

Ordination as latent variable model

Many ordination methods are 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)

Ordination as latent variable model

Many ordination methods are 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:

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

This makes GLLVMs a framework for many types of ordination, with foundations in existing methods.

Ordination as latent variable model

Many ordination methods are thought of as implementing a latent variable model

The main issue? We do not know how approximate it is! We cannot validate!

They approximately implement:

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

This makes GLLVMs a framework for many types of ordination, with foundations in existing methods.

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)

Main benefits of these methods

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

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/

Validation

A "bad" looking ordination plot has often been used as indicator that the ordination method does not do well.

- PCA: horseshoe effect
- CA: arch effect
- DCA: tongue effect (and very heuristic)
- PCoA: similar to PCA
- NMDS: no species effects, no variation explained, no hypothesis testing, I can go on

Small eigenvalues: also bad.

Model-based ordination

Suggested to use Generalized Linear Latent Variable Models for unconstrained ordination

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

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

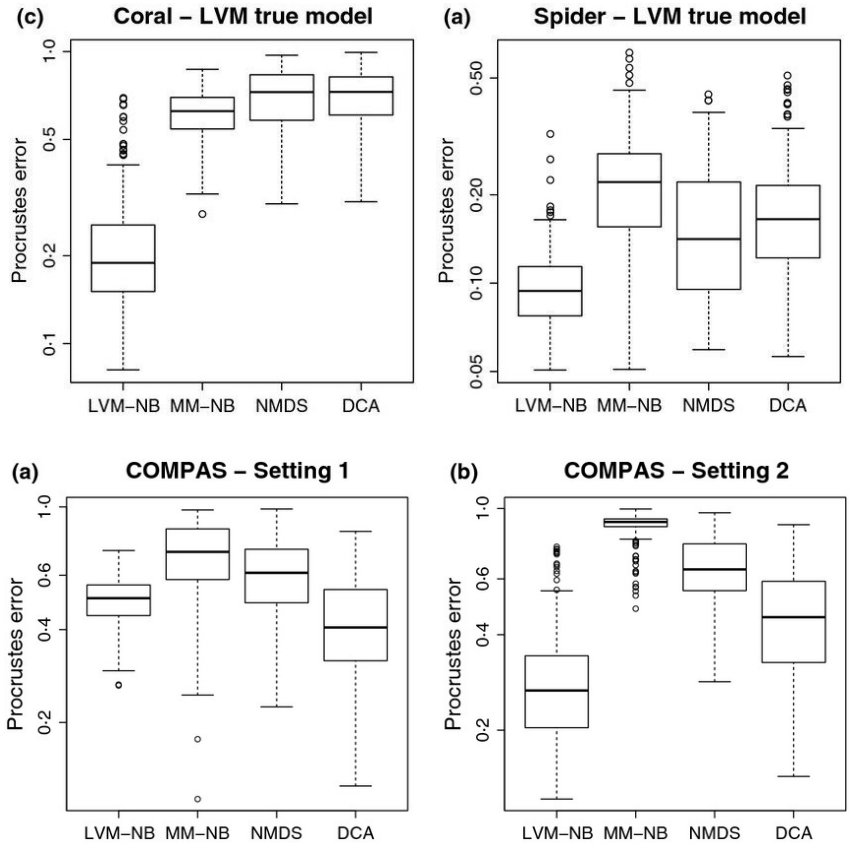
Model-based approaches to unconstrained ordination

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

Building on a long history of using latent variables in ecology (e.g., ter Braak 1985)

Unconstrained ordination



Figures from **Hui et al. 2015**

Generalised Linear Latent Variable Model

Unlike in the JSDM, model-based ordination has a focus on the latent variables. The model is:

$$\eta_{ij} = \beta_{0j} + \mathbf{u}_i^\top \boldsymbol{\gamma}_j \tag{2}$$

but now, we have a much stronger focus on the lower dimensions. And the ordination axis can be treated as fixed or random effect (but usually random).

So how is this different from JSDM

1. Ordination and JSDM use the same statistical framework (GLLVMs)
2. The models take a different angle (associations versus latent variables)
3. JSDM **can** be an LVM, ordination **is** an LVM

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 macroecological
- 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 macroecological
- Is the **the sampling process** considered?

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

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 have too sparse data to estimate species effects**
3. We want to make pretty pictures

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

GLLVMs vs. classical ordination: main differences

1) GLLVMs have a real model
2) GLLVMs incorporate distributions, not distances
3) There are no eigenvalues (but there is variance)
4) Number of dimensions are set a-priori as in NMDS
5) Latent variables are found by "best fit"
6) You might not get the same solution every time
7) Forget about permutation testing
8) We do not care much about rotation
9) $\longrightarrow$

Classifying 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

Unconstrained ordination

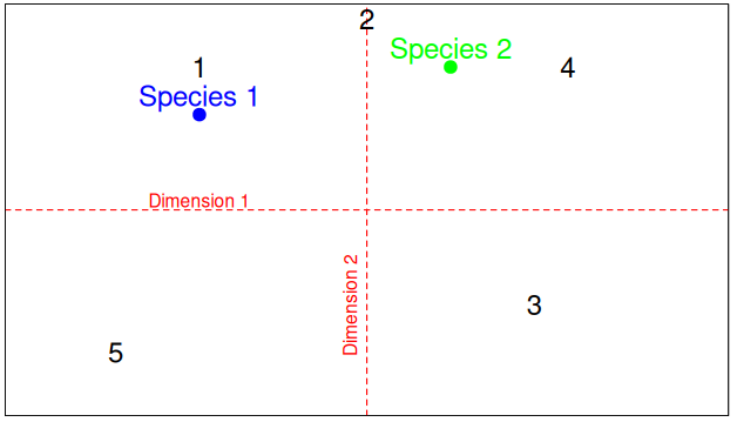
Used to:

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

To infer environmental conditions from species relationships

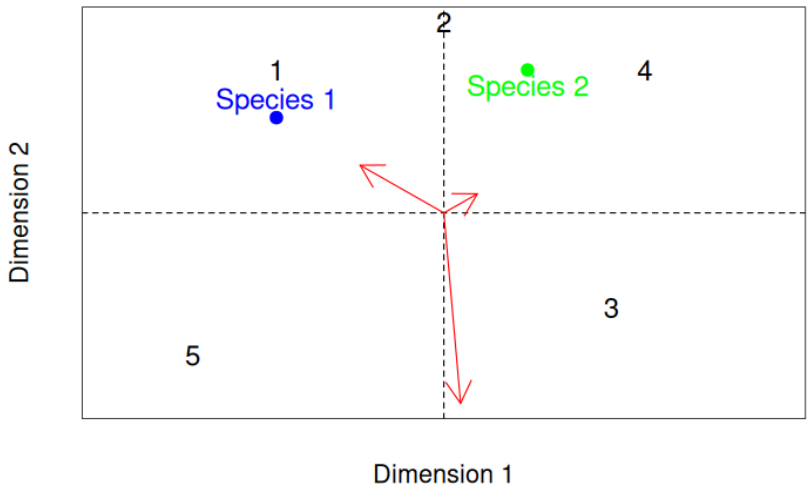
Reading an ordination plot

- Ordination plots capture the main patterns in the data
- Two coordinates close together are similar
- Close sites have similar community composition
- Close species have similar niches
- Distance in the ordination plot is analogous to correlation of JSDM
- We usually assume the environment drives patterns in an ordination



Ordination plot

- We interpret sites relative to sites, and species to species (so not usually species to sites)
- Coordinates are interpreted relative to the axes (LVs)
- In constrained ordination, arrows represent the axes-environment association
- Long arrows have a (relative) stronger effect
- The angle of arrows to the axes represent the association (orthogonal with no association)
- So, covariates help interpret the ordination



Rotation and orthogonality

Ordination methods are defined by their rotation (except NMDS?), here:

- Latent variables are orthogonal **a-priori**
- The latent variables are **not** maximum variance-rotated
- **A-posteriori** the latent variables are not orthogonal

Rotation and orthogonality

Ordination methods are defined by their rotation (except NMDS?), here:

- Latent variables are orthogonal **a-priori**
- The latent variables are **not** maximum variance-rotated
- **A-posteriori** the latent variables are not orthogonal

We can rotate them afterwards in whatever manner we want (e.g., with the `GPArotation` package) .

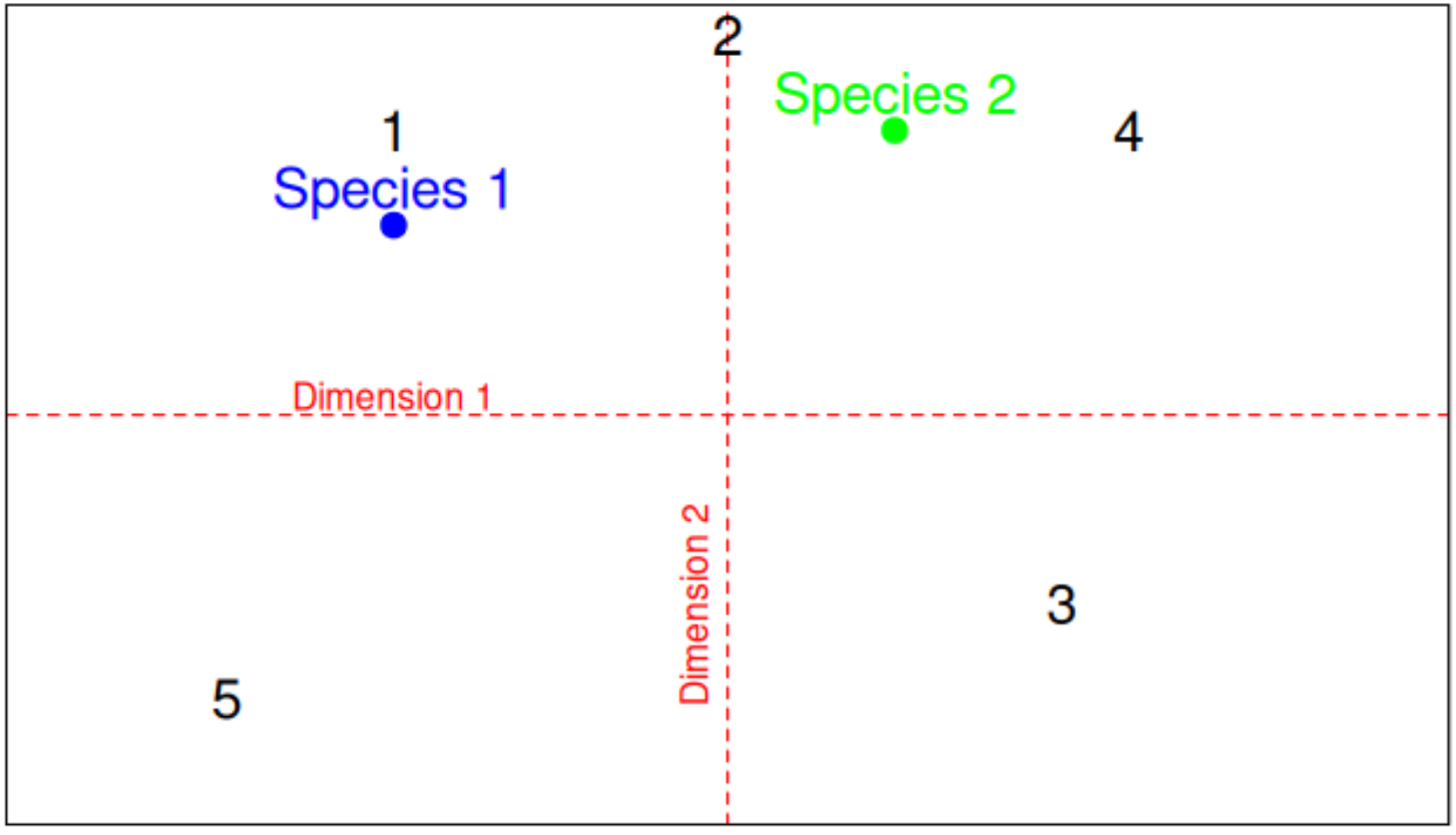
Rotation and orthogonality

Ordination methods are defined by their rotation (except NMDS?), here:

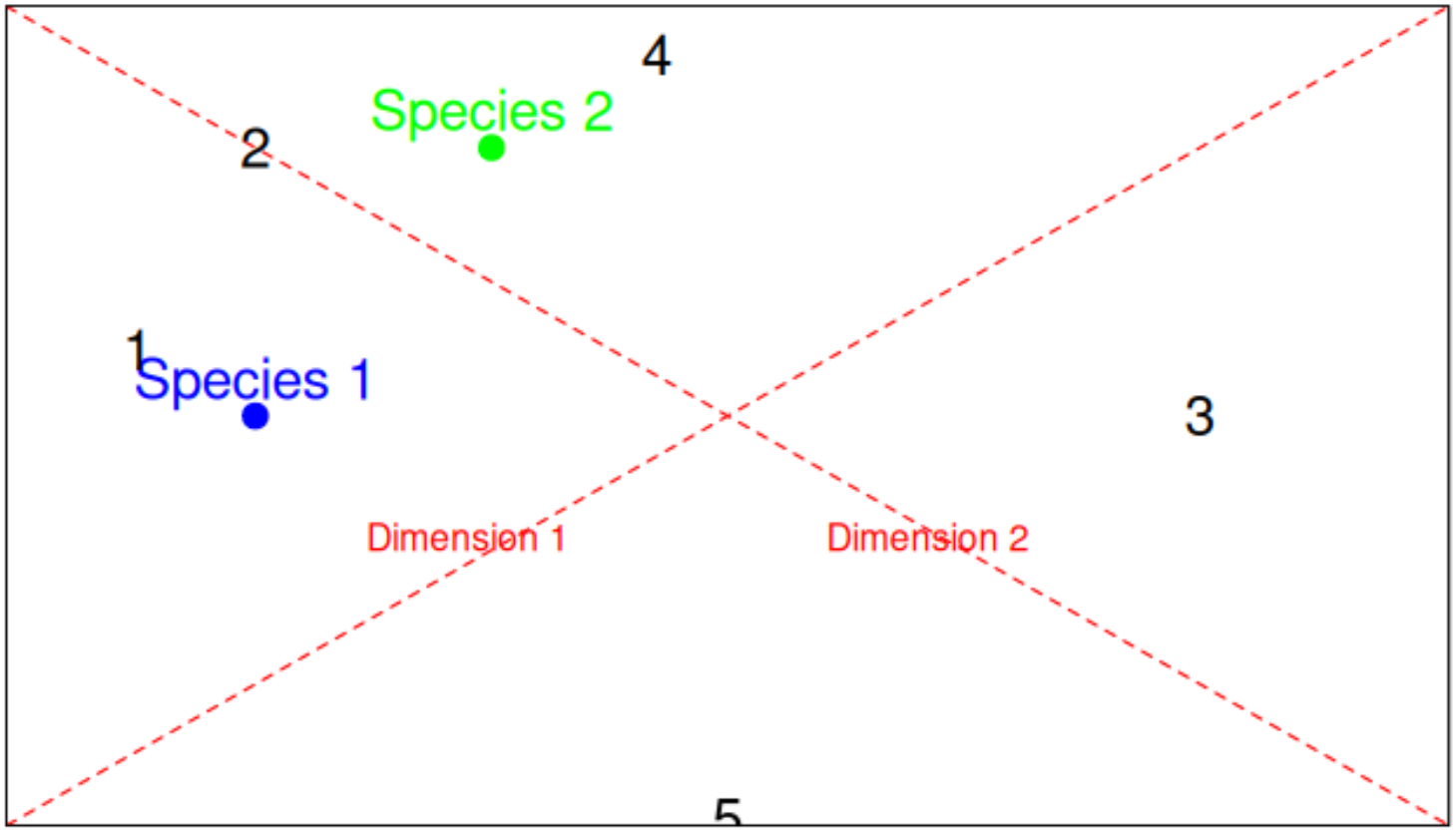
- Latent variables are orthogonal **a-priori**
- The latent variables are **not** maximum variance-rotated
- **A-posteriori** the latent variables are not orthogonal

We can rotate them afterwards in whatever manner we want (e.g., with the `GPArotation` package) . The model doesn't care.

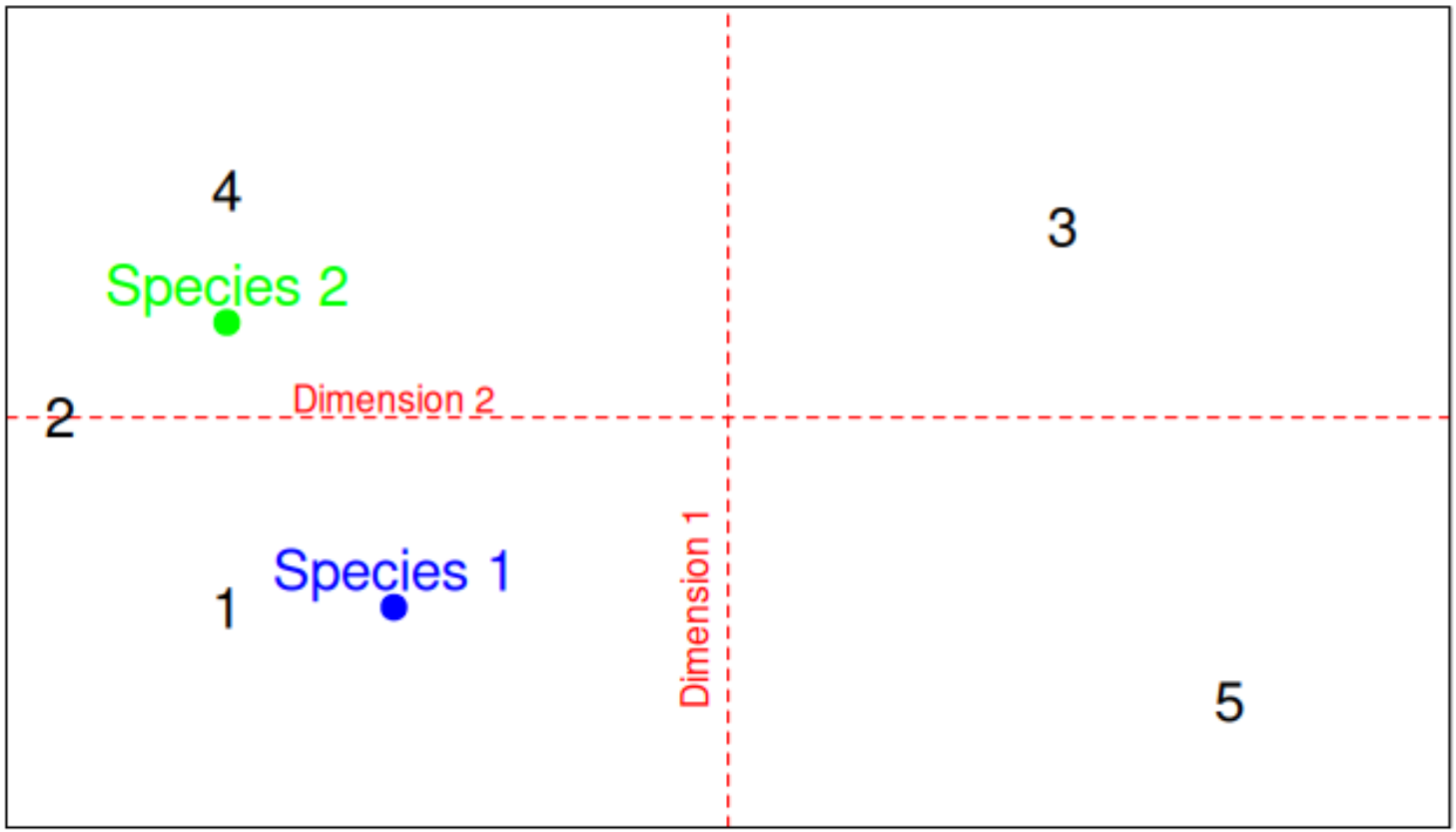
Rotation of ordination



Rotation of ordination



Rotation of ordination



Our inference of site and species (dis)similarity remains the same.

Example: Dutch Dune data

Achimill	Agrostol	Airaprae	Alopgeni	Anthodor	Bellpere	Bromhord	Chenal
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	

- A classic dataset, originally by Jongman et al. (1995)
- Ordinal classes for 30 plant species at 20 sites
- 5 covariates; A1, Moisture (5 groups), Management (4 groups), Use (3

The ordinal model: cumulative probit

More commonly, the categories are ordered.

$$\text{pr}(y_{ij} \leq k) = \Phi(\tau_{jk} - \eta_{ij}) \tag{3}$$

and

$$\text{pr}(y_{ij} = k) = \Phi(\tau_{jk} - \eta_{ij}) - \Phi(\tau_{jk-1} - \eta_{ij}) \tag{4}$$

- τ_{jk} are cut-off parameters that induce ordering
- τ_{jk} requires having at least one observation in every category
- If we have missing classes, we can re-order and skip some
- Alternatively, we can assume $\tau_{jk} = \tau_k$; same cut-offs for all species
- Controlled with `zeta.struc = common` but defaults to `species`

The effects η_{ij} are the same for all categories

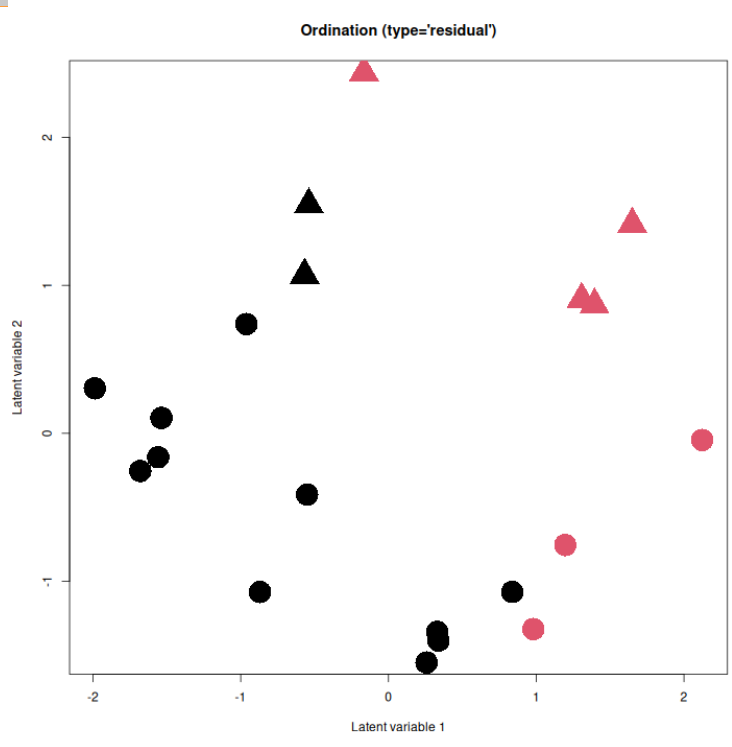
Example: unconstrained ordination

```
model1 <- gllvm(Y, num.lv = 2, family = "ordinal")
```

```
## Warning in gllvm.iter(y = y, X = X, xr = xr, lv.X = lv.X.design, for
## formula, : Can't fit ordinal model if there are species with missing
## Setting 'zeta.struc = `common`'.
```

Example: making an ordination plot

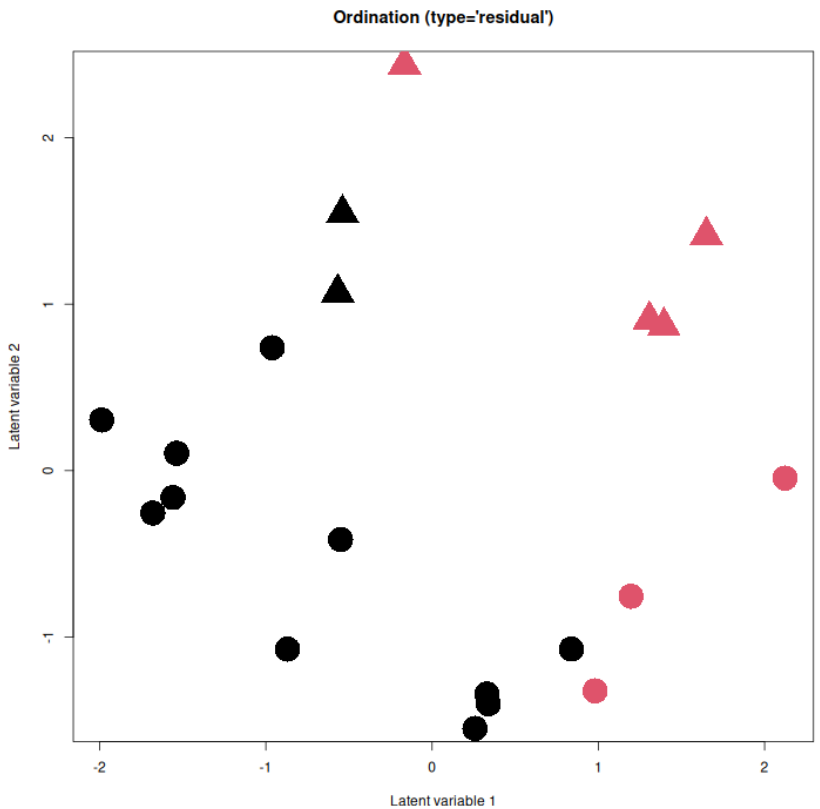
```
gllvm::ordiplot(model1, symbols = TRUE,
                s.colors = model.matrix(~0+., dune.env)[,5]+1,
                pch = model.matrix(~0+., dune.env)[,7]+16, s.cex = 4)
```



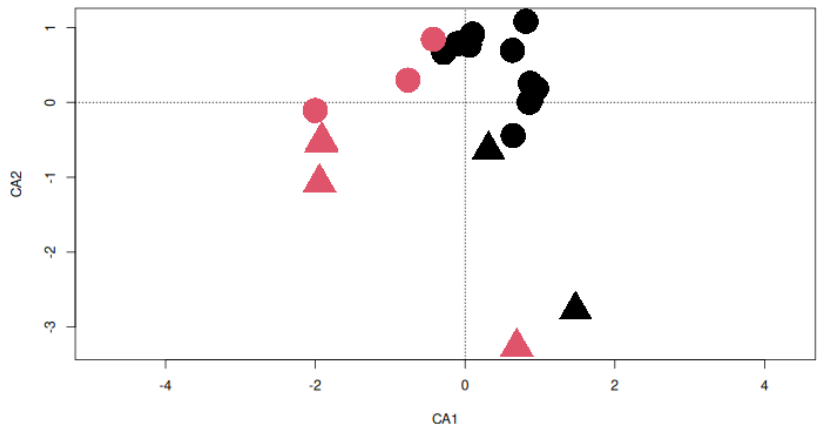
- Color by Moisture 5
- Shape by Management NM

Example: comparing ordinations

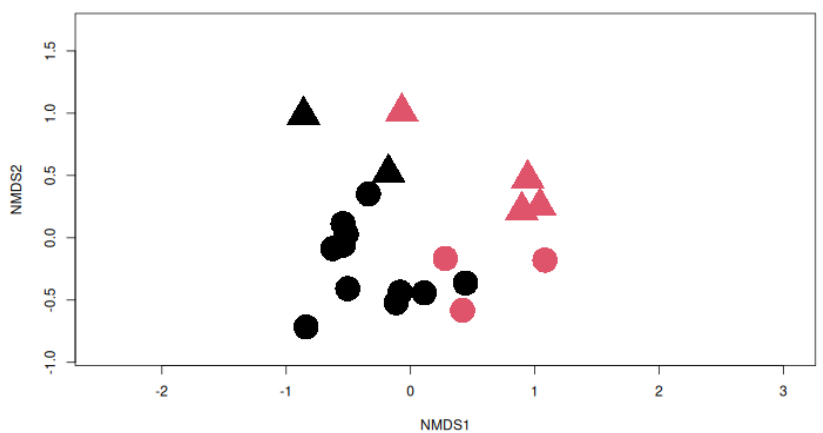
GLLVM



CA



NMDS



Example: comparing ordinations (2)

```
GLLVMscores <- getLV(model1)
CAscores <- vegan::scores(ca)
NMDSscores <- vegan::scores(nmds)
vegan::procrustes(GLLVMscores, CAscores, symmetric = TRUE)
```

```
##
## Call:
## vegan::procrustes(X = GLLVMscores, Y = CAscores, symmetric = TRUE)
##
## Procrustes sum of squares:
## 0.1961
```

```
vegan::procrustes(GLLVMscores, NMDSscores, symmetric = TRUE)
```

```
##
## Call:
## vegan::procrustes(X = GLLVMscores, Y = NMDSscores, symmetric = TRUE)
##
## Procrustes sum of squares:
## 0.1745
```

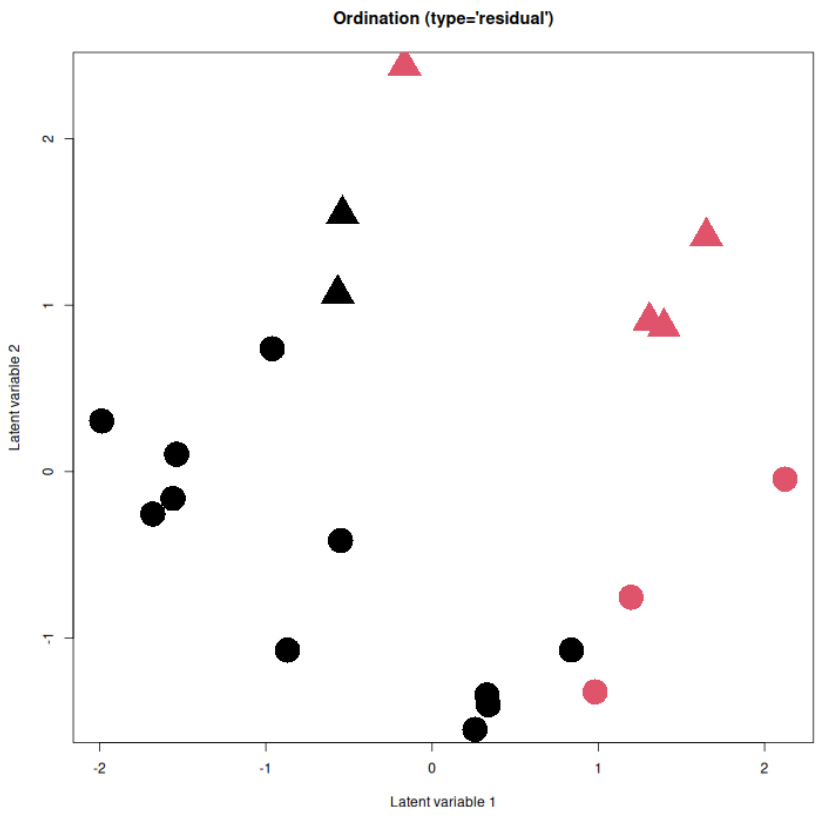
Species-specific cut-offs

```
model2 <- gllvm(apply(Y, 2, function(x)as.numeric(as.factor(x))), num.lv = 2, family = "ordinal", n.init = 3)
```

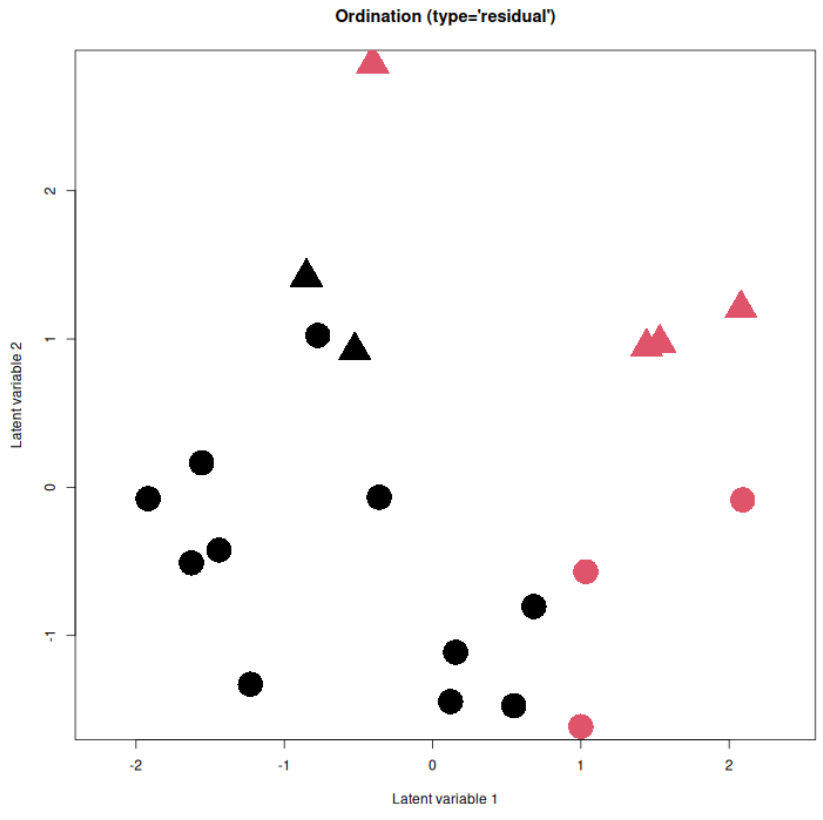
Can also explicitly be controlled with `zeta.struc = "common"` and `zeta.struc = "species"`

Example: Comparing ordinations (3)

Species-common cut-offs



Species-specific cut-offs



Example: Comparing ordinations (4)

```
GLLVMscores2 <- getLV(model2)
vegan::procrustes(GLLVMscores, GLLVMscores2, symmetric = TRUE)
```

```
##
## Call:
## vegan::procrustes(X = GLLVMscores, Y = GLLVMscores2, symmetric = TRUE)
##
## Procrustes sum of squares:
## 0.03458
```

```
AIC(model1, model2)
```

```
##          df      AIC
## model1   97 1265.992
## model2 150 1157.304
```

Note: $1263 - (150 - 97 * 2) = 1157$

Tools

What tools do we have for drawing conclusions from unconstrained ordinations?

- Visualizations
- Prediction
- Variation explained

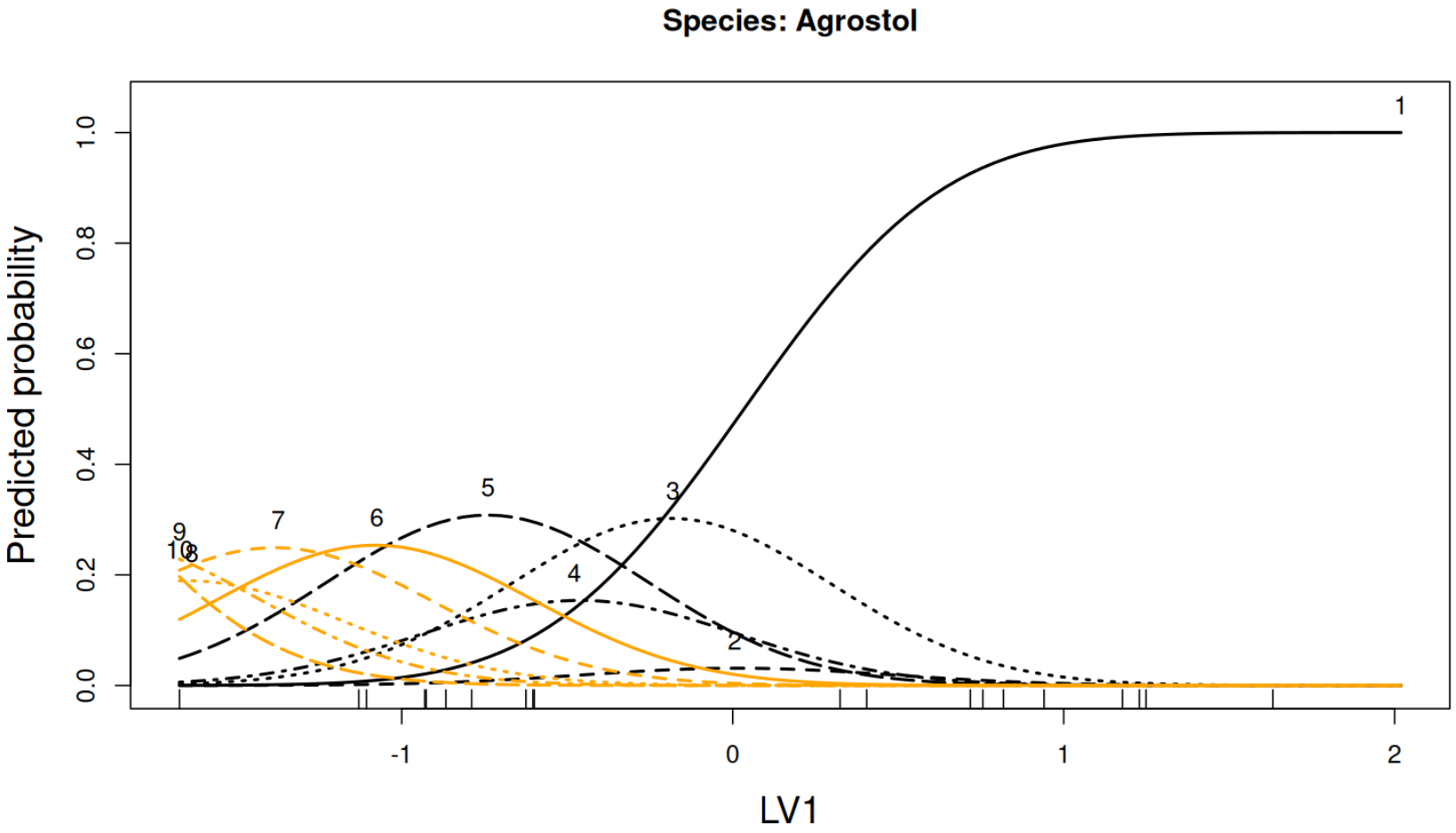
Prediction

We take our model to find out what happens under conditions that we have not observed. Here, with the latent variable.

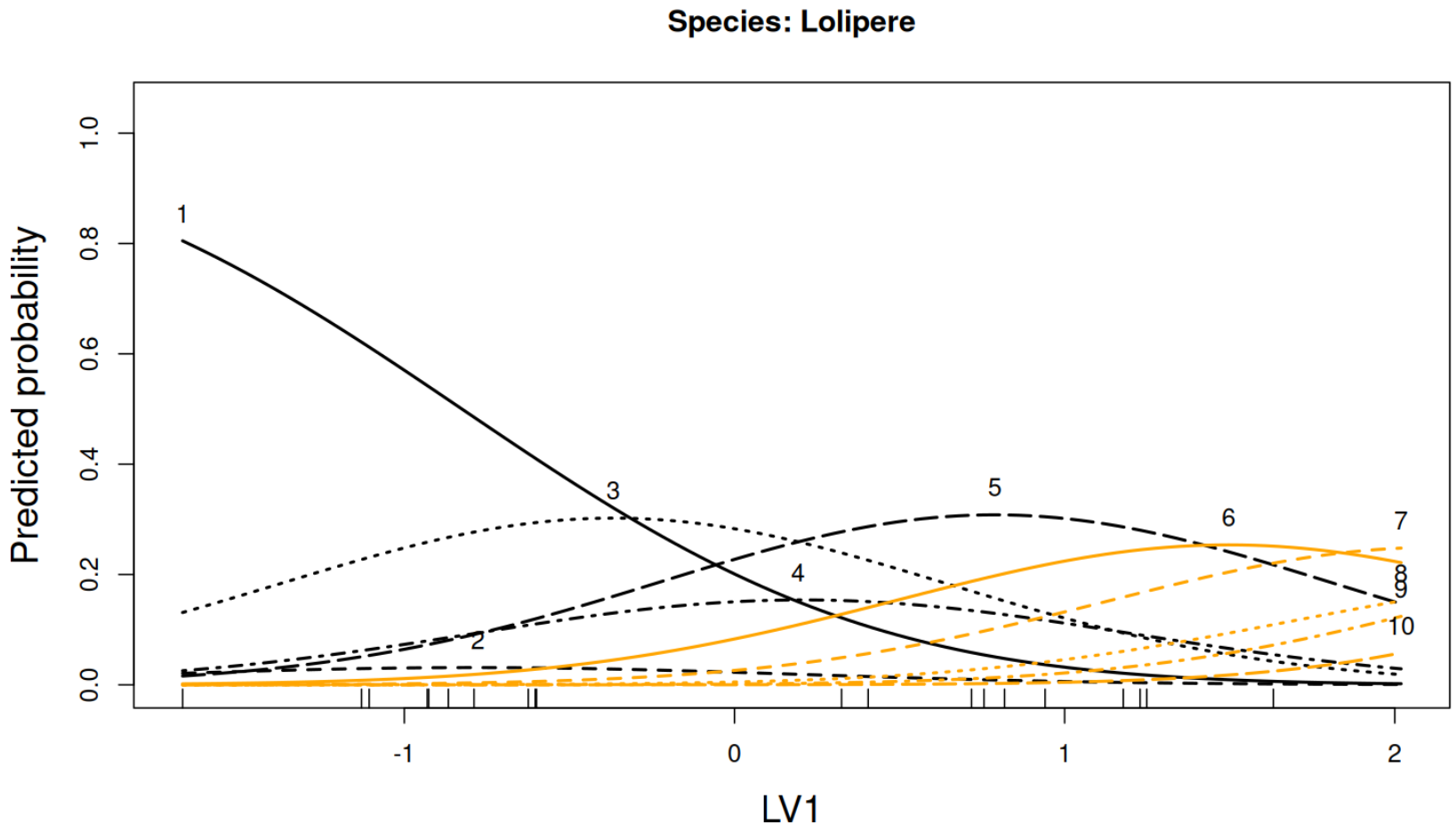
Example: prediction

```
lvs <- getLV(model1)
lv1new <- seq(from = min(lvs[,1]), to = max(lvs[,2]), length.out = 100)
preds <- predict(model1, newLV=data.frame(LV1 = lv1new, LV2 = 0),
                 type = "response")
```

Example: prediction

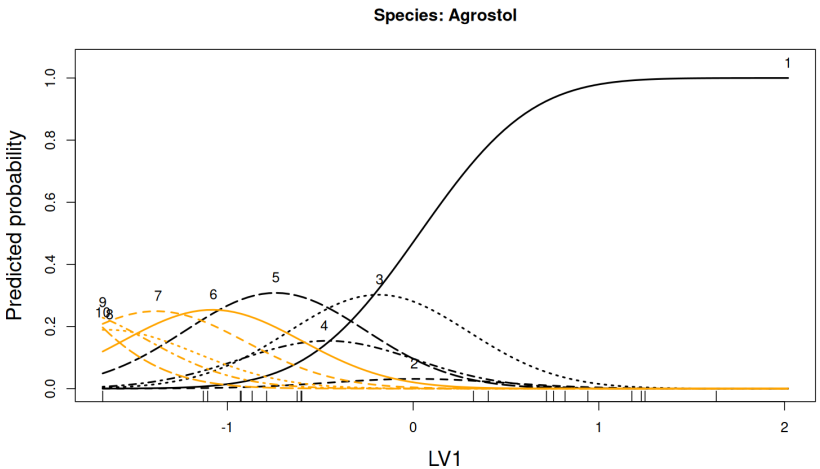


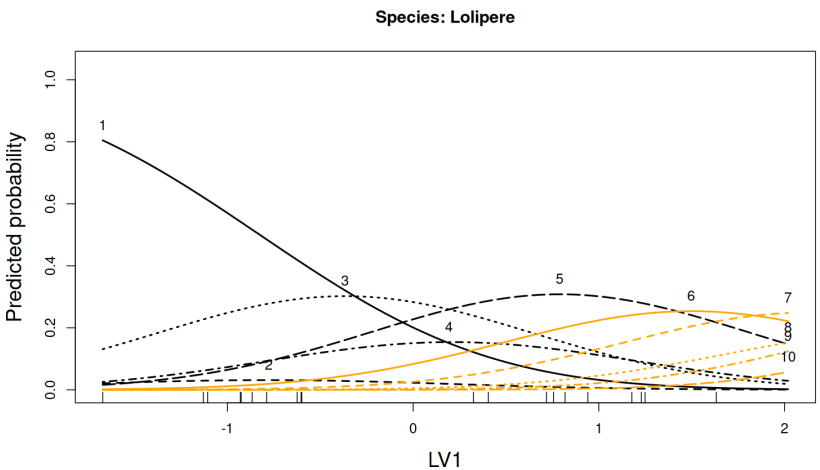
Example: prediction



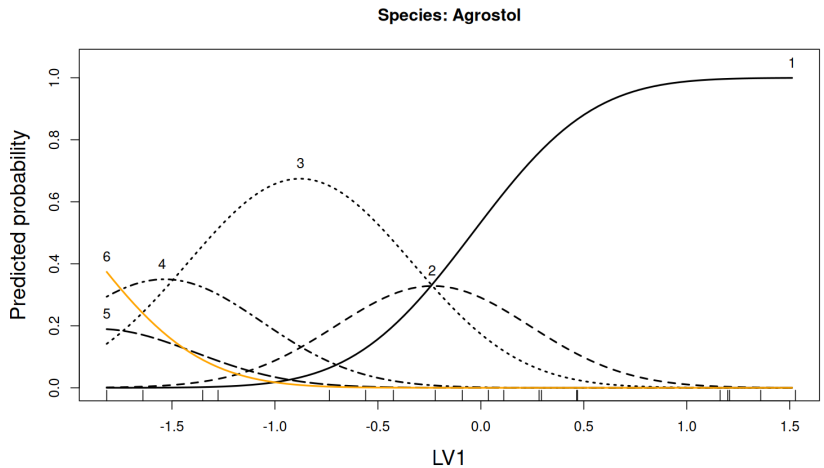
Example: prediction (2)

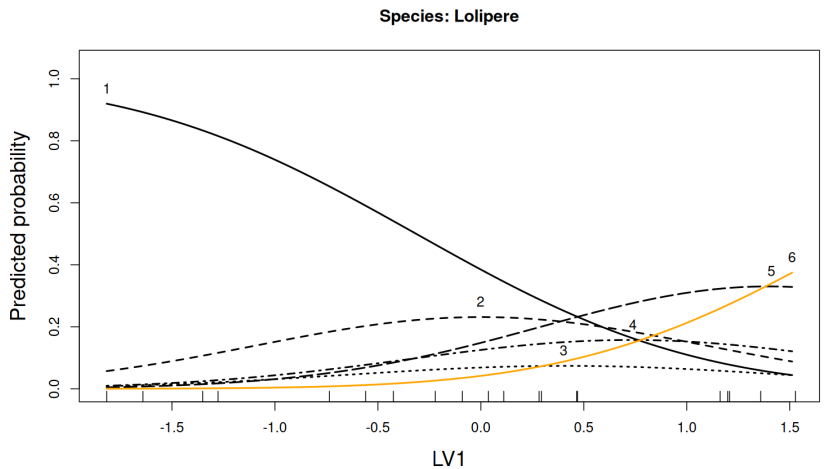
Species-common cut-offs





Species-specific cut-offs





Variation explained

In classical ordination, the eigenvalues tells us the variation explained by each dimension.
Latent variable models are not that straightforward.

- We do not estimate all axes
- We have no variation explained on the response scale
- We can do model selection (in a way similar to "stress" in NMDS)
- Or get a relative measure of variation explained

Variation explained

In classical ordination, the eigenvalues tells us the variation explained by each dimension.
Latent variable models are not that straightforward.

- We do not estimate all axes
- We have no variation explained on the response scale
- We can do model selection (in a way similar to "stress" in NMDS)
- Or get a relative measure of variation explained

Try not to confuse variation explained with ecological importance

(Relative) variation explained

$$\begin{aligned} \boldsymbol{\eta}_i &= \beta_{0j} + \sum_{q=1}^{d} z_{iq}\boldsymbol{\gamma}_q, \qquad \text{where } \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &= \beta_{0j} + \boldsymbol{\epsilon}_i, \qquad \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sum_{q=1}^{d} \boldsymbol{\gamma}_q) \end{aligned} \tag{5}$$

(Relative) variation explained

$$\begin{aligned} \boldsymbol{\eta}_i &= \beta_{0j} + \sum_{q=1}^{d} z_{iq}\gamma_q, \qquad \text{where } \mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \\ &= \beta_{0j} + \boldsymbol{\epsilon}_i, \qquad \text{where } \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \sum_{q=1}^{d} \boldsymbol{\gamma}_q) \end{aligned} \tag{5}$$

```
var.q = getResidualCov(model1)$var.q
var.q;var.q/sum(var.q)
```

```
##       LV1       LV2
## 42.58724 30.67649
```

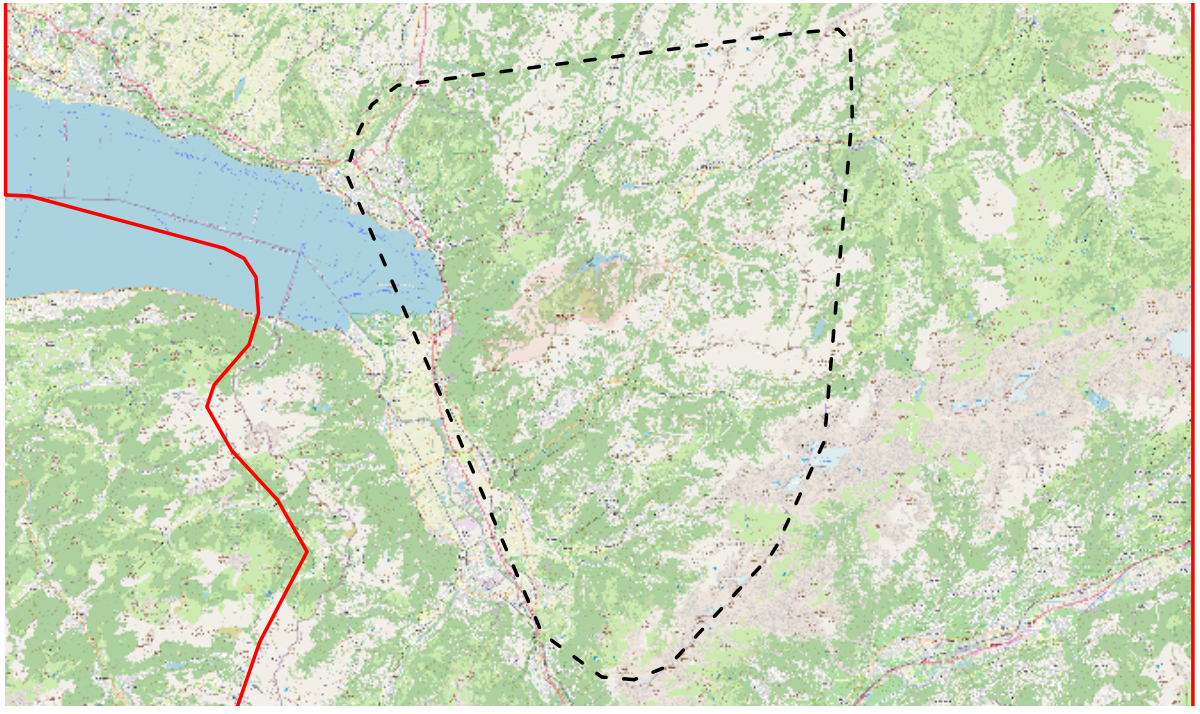
```
##       LV1       LV2
## 0.5812868 0.4187132
```

What does this mean for importance of dimensions?

- Importance does not equal variation explained
- The first dimension may or may not be most relevant
- Dimensions with less variation may be equally important to represent the community
- We measure "importance" differently (e.g., by fit via `AIC` or `BIC`)

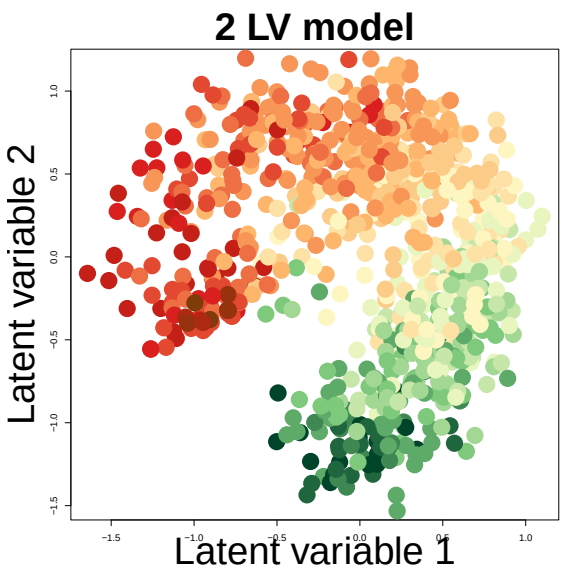
Example: alpine plants in Switzerland

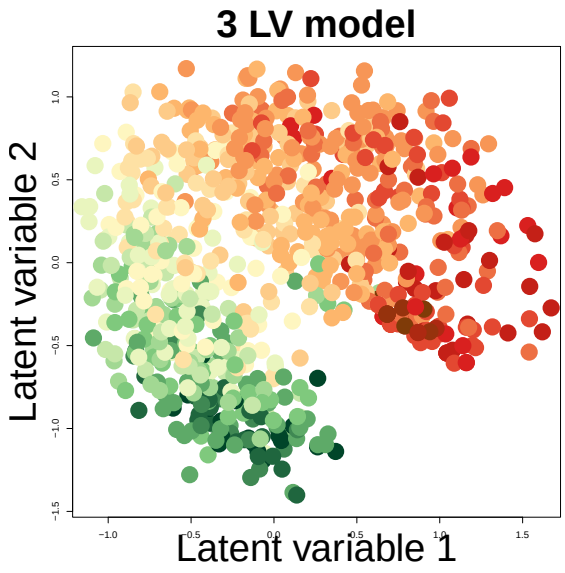
- Data by D'amen et al. (2017)
- Occurrence of 175 species at 840 $4m^2$ plots
- Sampled on an elevation gradient

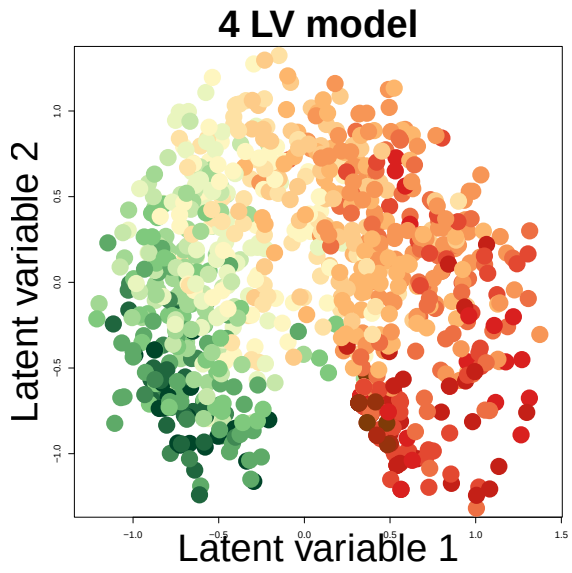


Example: fit unconstrained ordinations

```
model3 <- gllvm(Y, num.lv = 2, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
model4 <- gllvm(Y, num.lv = 3, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
model5 <- gllvm(Y, num.lv = 4, family = "binomial", sd.errors = FALSE, diag.iter = 0, optim.method = "L-BFGS-B
```







- After rotation all three show the same (elevation) pattern
- Before rotation, the 3LV model exhibits elevational patterns on LV 1 and 3

Example: re-rotating 3 LV model

```
vegan::procrustes(getLV(model3), getLV(model4)[, 1:2], symmetric = TRUE)$ss
```

```
## [1] 0.4804651
```

```
rot <- GPArotation::Varimax(getLoadings(model4))$Th
vegan::procrustes((getLV(model4)%*%rot)[,1:2], getLV(model3), symmetric = TRUE)$ss
```

```
## [1] 0.2683858
```

```
cbind(cor(getLV(model4), X$ELEV), cor(getLV(model4)%*%rot, X$ELEV))
```

```
##            [,1]       [,2]
## LV1  0.81904771 0.79264055
## LV2 -0.09767524 0.27823543
## LV3  0.16211116 0.01917828
```

Example: variation

```
##          LV1       LV2      LV3      LV4
## 1   86.91255 148.13208       NA       NA
## 2 165.39895  54.06302 81.12508       NA
## 3 134.67558 112.75264 43.10753 33.92187
```

The dominant gradient occurs on different LVs, but does explain most variation here.

Example: model selection

```
IC <- cbind(AIC(model3, model4, model5)[,2], AICc(model3, model4, model5), BIC(model3, model4, model5)[,2])
```

```
##            AIC     AICc      BIC
## 2 LVs 76767.82 76771.57 81954.47
## 3 LVs 72776.62 72783.27 79675.66
## 4 LVs 72471.26 72481.61 81072.79
```

```
c("2 LV" = goodnessOfFit(model3$y, object = model3)$RMSE,
"3 LV" = goodnessOfFit(model4$y, object = model4)$RMSE,
"4 LV" = goodnessOfFit(model5$y, object = model5)$RMSE)
```

```
##      2 LV      3 LV      4 LV
## 0.2625611 0.2454750 0.2372903
```

Example: model selection

```
IC <- cbind(AIC(model3, model4, model5)[,2], AICc(model3, model4, model5), BIC(model3, model4, model5)[,2])
```

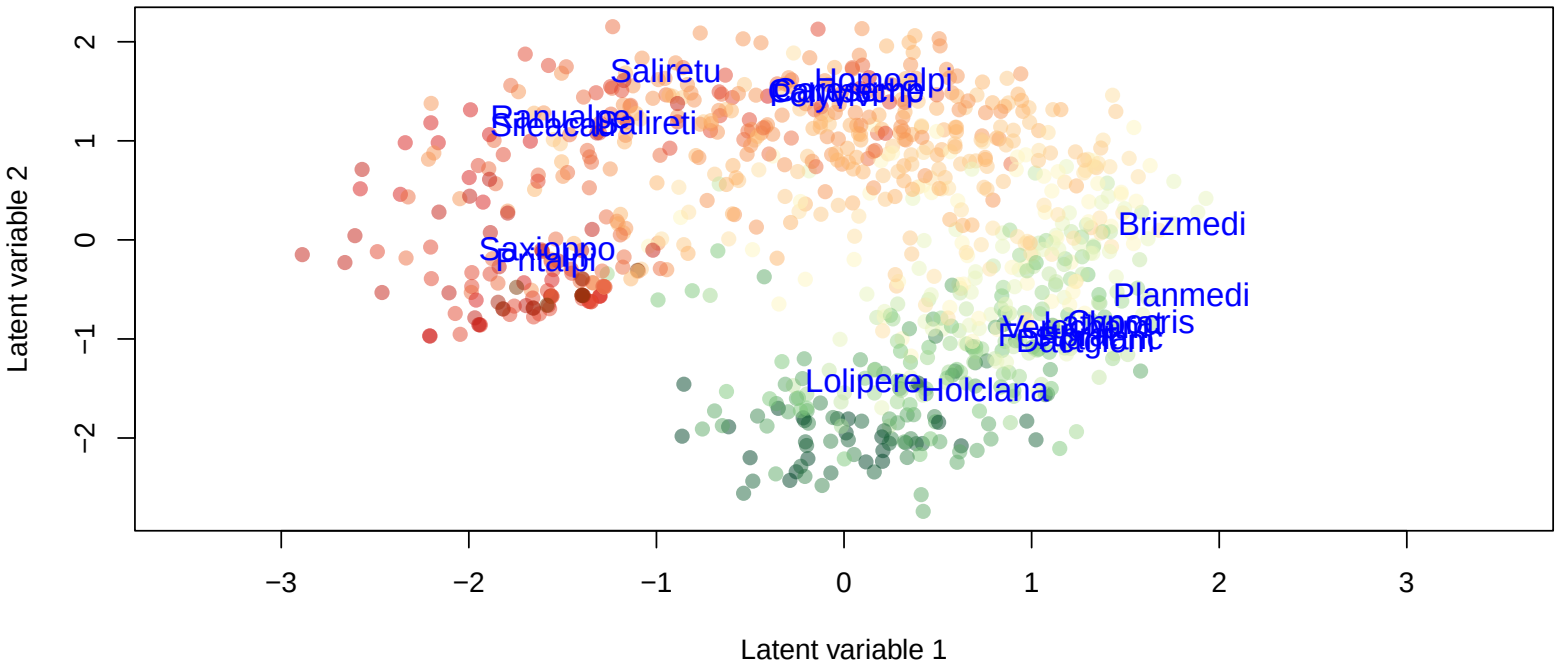
```
##            AIC     AICc      BIC
## 2 LVs 76767.82 76771.57 81954.47
## 3 LVs 72776.62 72783.27 79675.66
## 4 LVs 72471.26 72481.61 81072.79
```

```
c("2 LV" = goodnessOfFit(model3$y, object = model3)$RMSE,
"3 LV" = goodnessOfFit(model4$y, object = model4)$RMSE,
"4 LV" = goodnessOfFit(model5$y, object = model5)$RMSE)
```

```
##      2 LV      3 LV      4 LV
## 0.2625611 0.2454750 0.2372903
```

Predictive performance might not be best from the model with 2 dimensions, but it will usually capture the dominant gradients correctly.

Example: inference

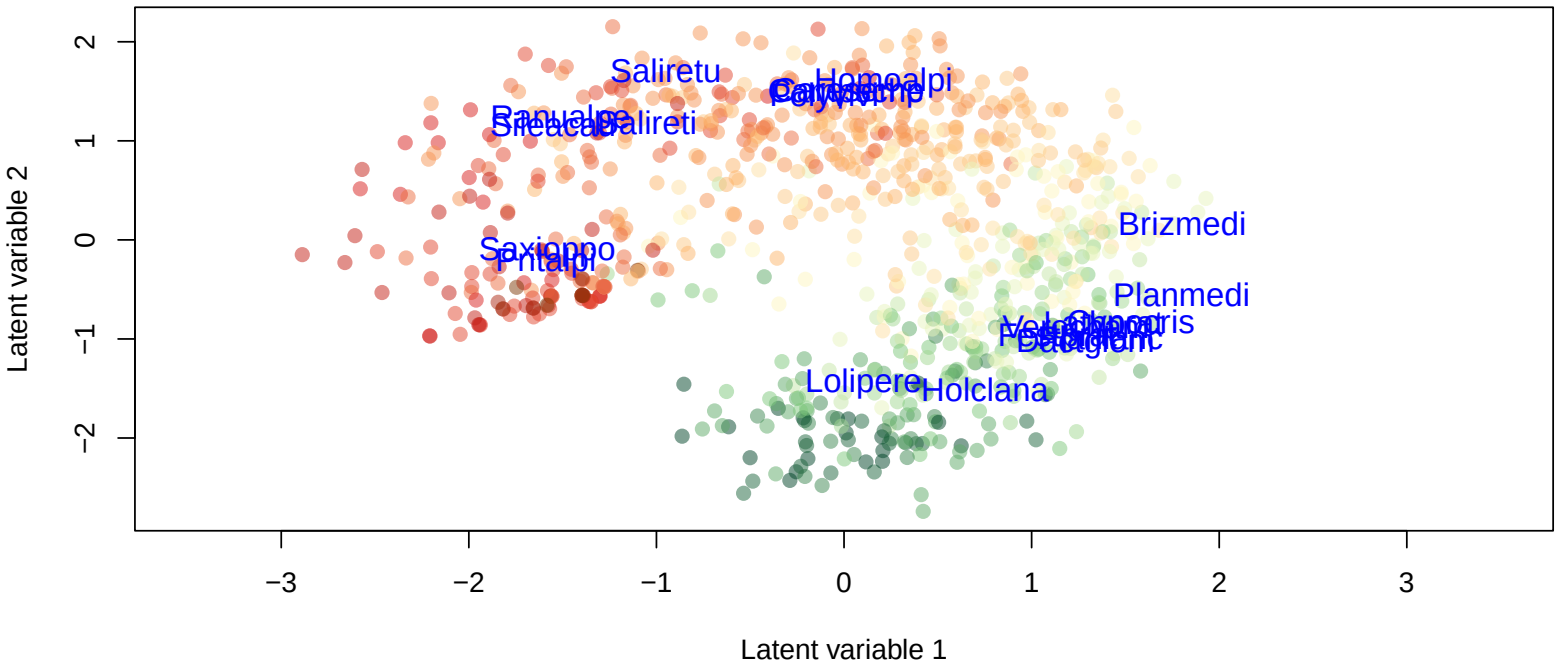


Pritzelago alpine

Salix retusa

Lolium perenne

Example: inference

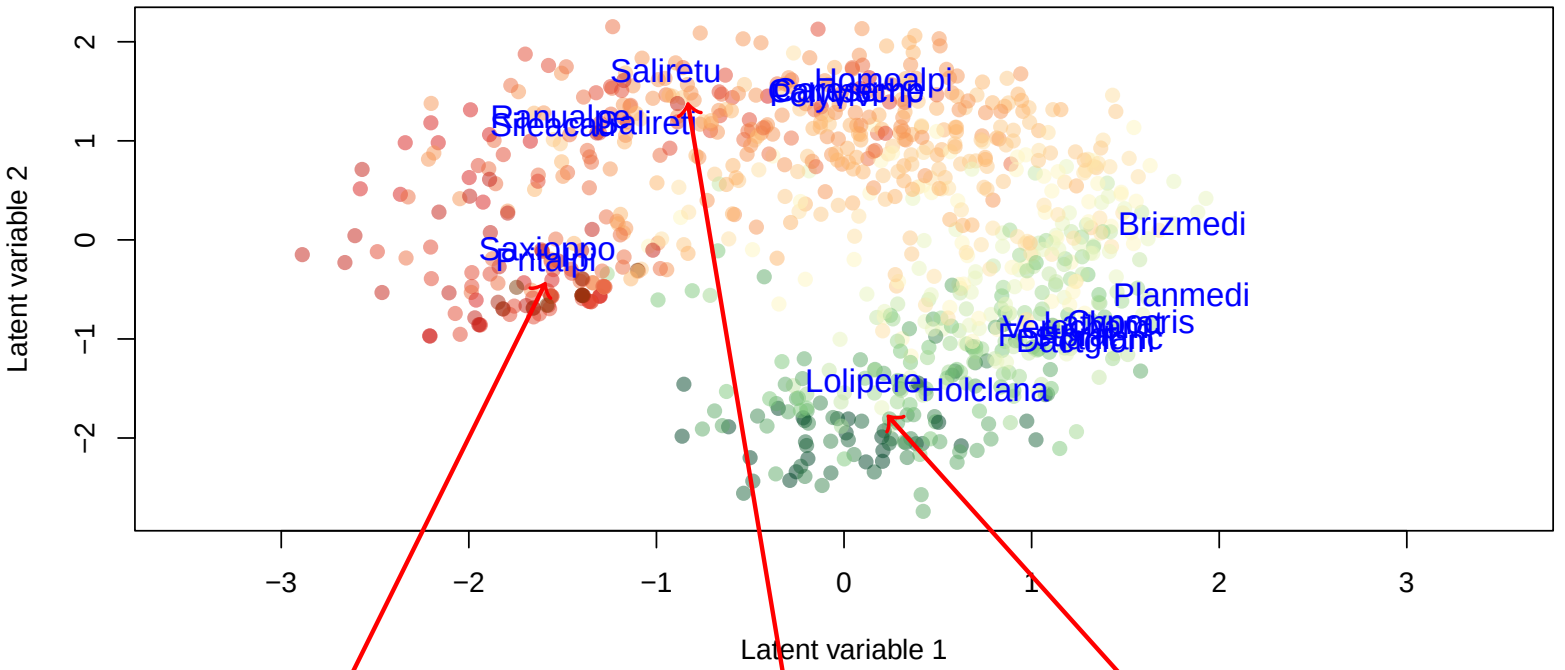


Pritzelago alpine

Salix retusa

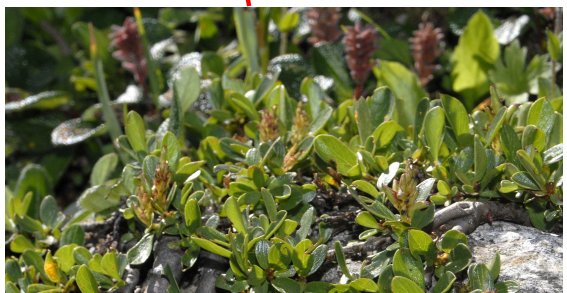
Lolium perenne

Example: inference



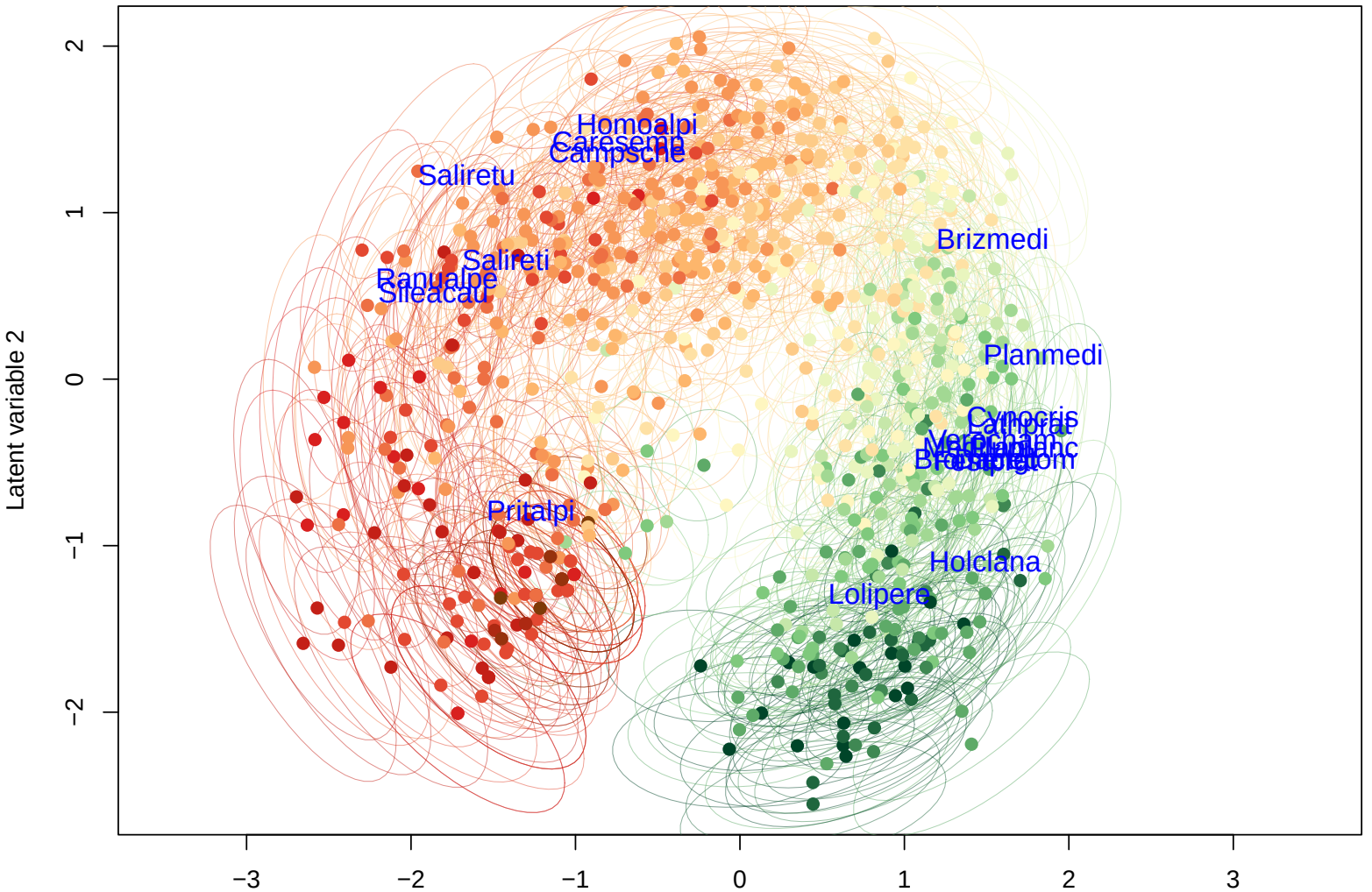
Pritzelago alpine

Salix retusa



Lolium perenne

Example: prediction regions



Conclusion

- Ordination is used to reduce parameters in a complex model
- Ecologically, to explore co-occurrence patterns via a low-dimensional space
- Unconstrained ordination only arranges sites and species based on the community data
- Without information on the environment, we use species' known preferences for inference
- Model-based ordination leverages from regression and ordination tools
  - Biplots
  - Uncertainties
  - Residual diagnostics
  - Information criteria
  - A flexible model
  - Prediction