

# Analysing multivariate ecological data with Generalized Linear Latent Variable Models

Bert van der Veen  
Jenni Niku  
Sam Perrin



# Exercise material and package installation

<https://github.com/BertvanderVeen/IRSAE2021GLLMworkshop>

## Questions

In chat or

- 🐦 On twitter: #**GLLMs**, @**vdVeenB** or @**J\_\_Niku** or @**BigDrFishBoi**
- 🗣 On github: <https://github.com/BertvanderVeen/IRSAE2021GLLMworkshop/discussions>

or **join zoom!**

## Installation

```
install.packages("gllvm")
```

Ask in chat if you have issues

# Welcome! 😊



## Who

Bert van der Veen  
PhD candidate

## Affiliation

Norwegian institute of  
Bioeconomy research &  
Norwegian university of  
Science and Technology

## Expertise

- Statistical ecology
- Ordination
- Species distribution modeling



Jenni Niku  
Postdoc

University of Jyväskylä

- Statistical ecology
- Species distribution modeling



Sam Perrin  
PhD

Norwegian university of  
Science and Technology  
Ducky AS

- Fresh water ecology
- Invasion ecology
- Species distribution modeling

# Program

## Topic

- Basics of Multivariate analysis
- Latent variables
- Generalized Linear Latent Variable models

## Who



*Questions / Break*

- **gllvm** R-package (Niku et al. 2019)
- How to: model-based ordination with GLLVMs



*Questions / Break*

- Some ecology with species associations



*Finish*

- Model-based quadratic ordination
- Model-based constrained ordination
- Wrap-up



# What are these GLLVMs?

## Generalized Linear Latent Variable Models

- Model-based multivariate analysis
- Mixed-effects model
- Joint Species Distribution Model Ovaskainen et al. 2017
- Unconstrained ordination Hui et al. 2015
- Constrained ordination
- "Factor analytic approach"

*But in general, for the analysis of species (co-)occurrence patterns.*

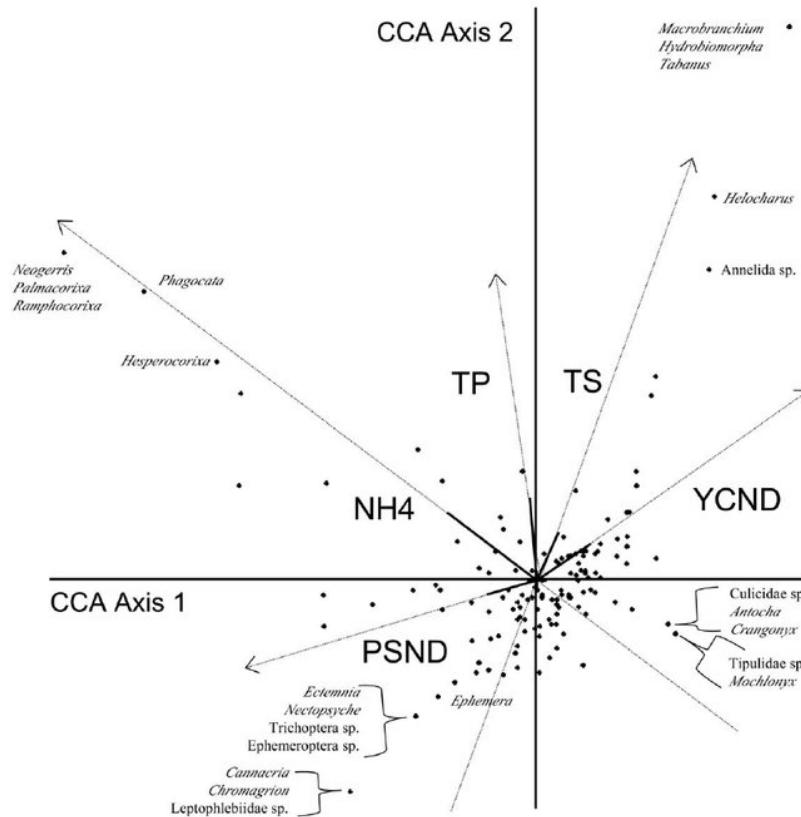
- **GLLVMs are flexible**



# Classical multivariate analysis

E.g., PCA, CA, DCA, NMDS

With **eigenvectors** and using **distance metrics**



Maul et al. 2004

# Classical multivariate analysis

Here we go model-based!

Plant Ecol (2015) 216:669–682  
DOI 10.1007/s11258-014-0366-3



## Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·  
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

Received: 29 January 2014/Accepted: 30 May 2014/Published online: 19 November 2014  
© Springer Science+Business Media Dordrecht 2014

# Gathering data

We go out, register species at multiple sites



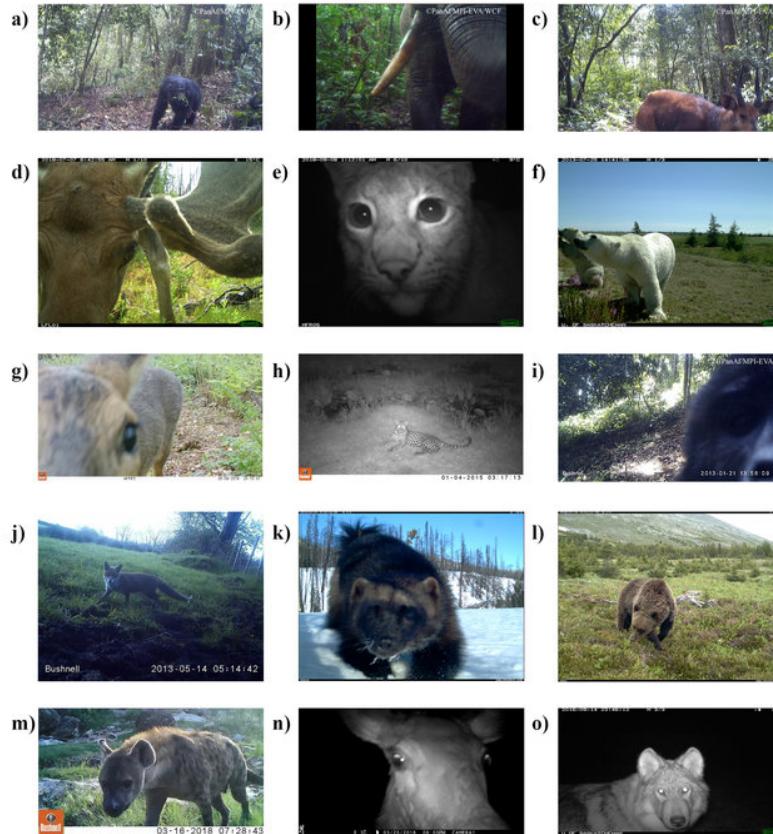
© Geir-Harald Strand / NIBIO

# "Multivariate"

- What does multivariate mean?
- Multivariate: multiple **responses**
- E.g. counts of species at sites

	<b>Species 1</b>	<b>Species 2</b>	<b>Species 3</b>	<b>Species 4</b>	<b>Species 5</b>
Site 1	25	10	0	0	0
Site 2	0	2	0	0	0
Site 3	15	20	2	2	0
Site 4	2	6	0	1	0
Site 5	1	20	0	2	0

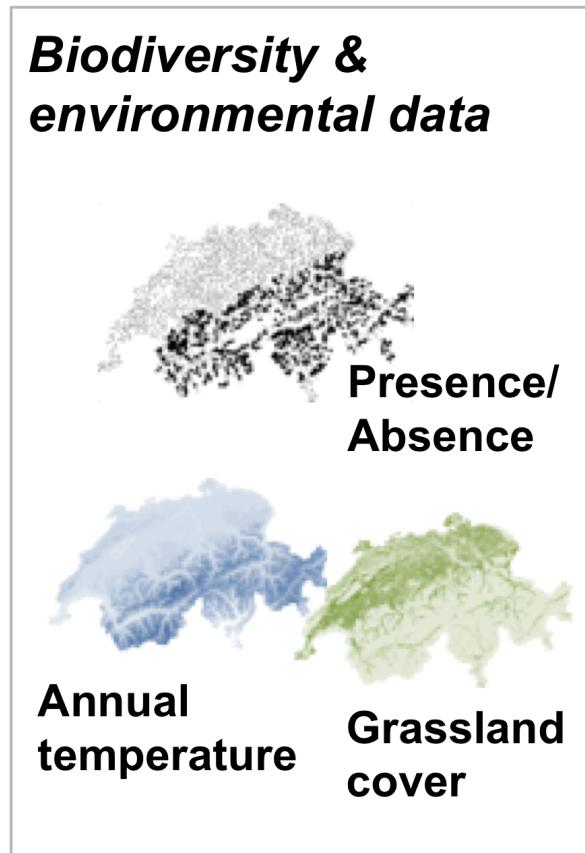
# e.g., camera trap data



Caravaggi et al. 2020

# "Multivariable"

Multiple **predictors** that represent the environment



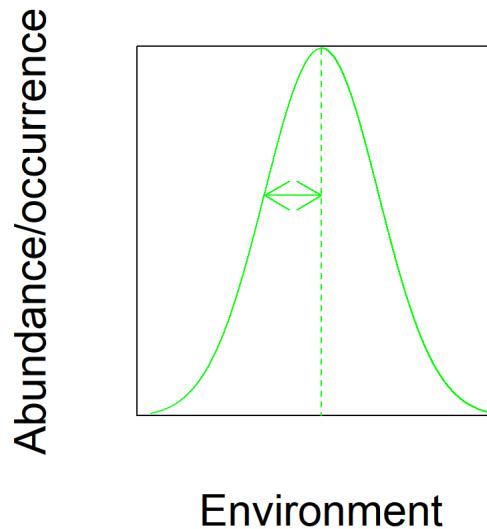
<https://damariszurell.github.io/SDM-Intro/>

# To summarise

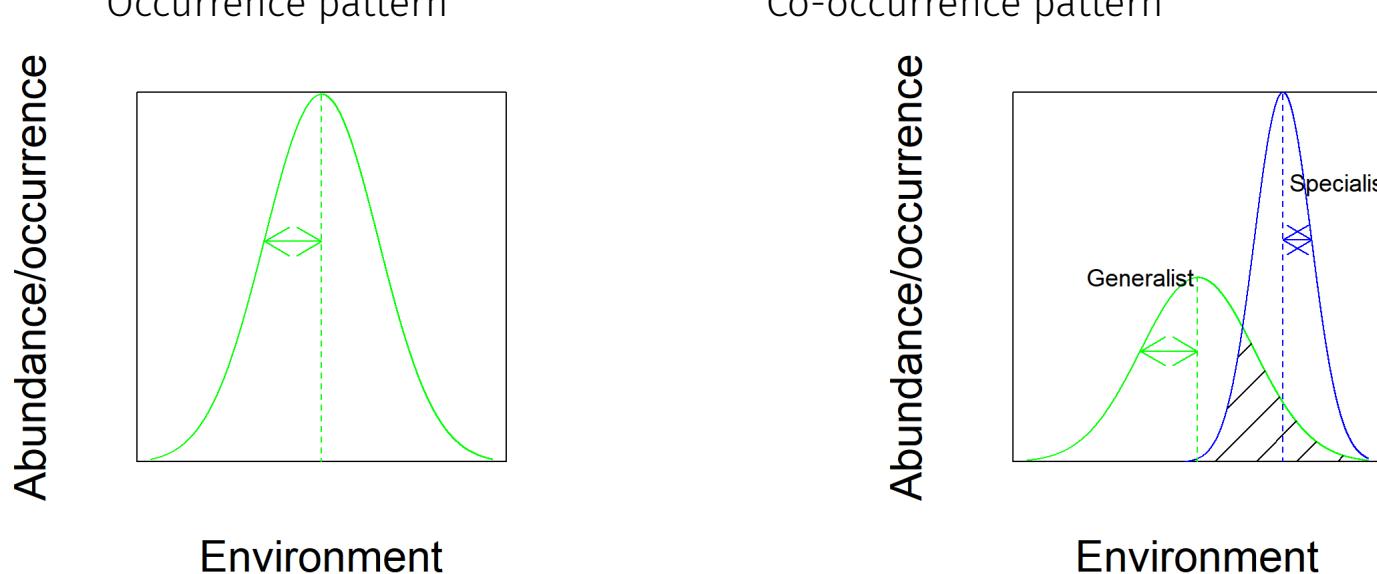
- Both data and method can be univariate or multivariate
- Multivariate data can be analysed with both multivariate and univariate methods (SDM, CA)
- Multivariable data can be used in multivariate or univariate analysis
  - Generally the same for all responses
  - (But, note that the model can of course set terms to zero)

# Why analyse multivariate data?

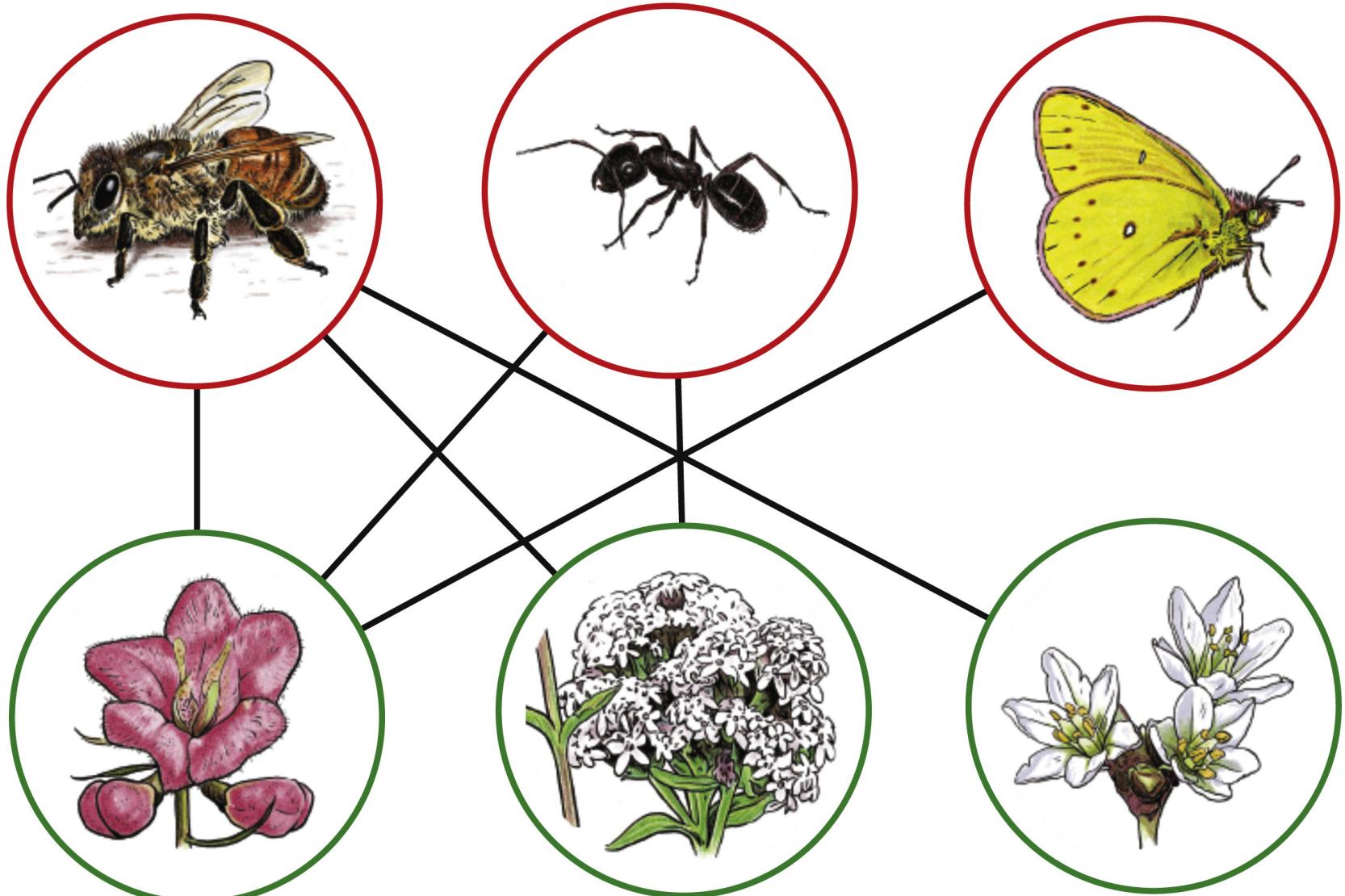
- Interest in **co**-occurrence patterns
  - In contrast to only **occurrence** patterns (a species distribution)
- Why do species co-occur?
  - Similar environmental preferences
  - Similar history in the environment
  - Might result in **Interactions**
- Multiple species form a **community**



# Why analyse multivariate data?



# But then for more species



# Some examples

- Macroinvertebrate monitoring Paul 2020
- Effects of habitat on fish abundance Brown et al. 2017
- Associations of wood-decaying fungi Maynard et al. 2018
- Beta-diversity patterns in tree species Draper et al. 2019
- Effects of management on heathland Damgaard et al. 2020



state.nj.us

# Joint modeling with latent variables

- Accounts for correlation between taxa
- "Borrows" information from other species for estimation
- Provides species associations
- Concept: fit a single model for all species
- I.e., a "Joint Species Distribution Model"
  - Faster
  - Less tedious
  - Explicitly model species co-occurrence
  - Etc.

## Key references

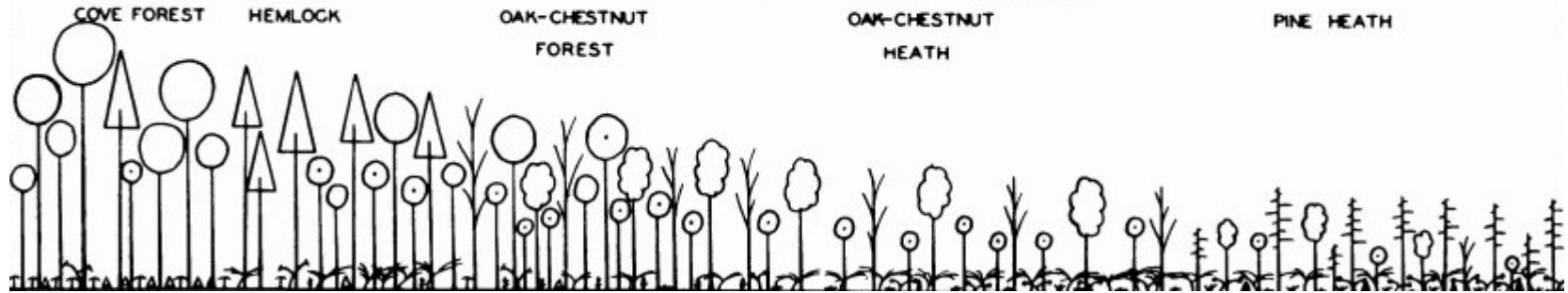
- Warton et al. 2015: "So many variables: Joint modeling in community ecology"
- Blanchet et al. 2020: "Co-occurrence is not evidence of ecological interactions"
- Poggiato et al. 2021: "On the interpretations of joint modeling in community ecology"

# Latent variables?

## VEGETATION PATTERNS IN THE GREAT SMOKY MOUNTAINS

Change of vegetation along the moisture gradient at lower and higher elevations

EASTERN FOREST SYSTEM - 3000 FEET



Whittaker 1967

Variables can be **observed** or **latent**

what's the  
opposite of  
latent?



active, obvious, manifest,  
apparent, alive, clear, live,  
operative, working, open



*If not measured, it is latent, like a random-effect.*

# Ecological gradient analysis

"Gradient analysis is a research approach for the study of spatial patterns of species." Whittaker 1967

Our sites describe the environment. Multiple environmental gradients can form a **complex** gradient.

	Predictor 1	Predictor 2	Predictor 3	Predictor 4	Predictor 5
Site 1	2.3321	3.0445	0.0000	3.0445	4.4543
Site 2	3.0493	3.2581	1.7918	1.0986	4.5643
Site 3	2.5572	3.5835	0.0000	2.3979	4.6052
Site 4	2.6741	4.5109	0.0000	2.3979	4.6151
Site 5	3.0155	2.3979	0.0000	0.0000	4.6151

# Ecological gradients

1) Ecological gradient: gradual change in the environment

- e.g. temperature

2) Complex gradient: change in several ecological gradients

- e.g., soil moisture and acidity on an elevation gradient
- e.g., a gradual urban to rural change in the landscape
- Can be represented as a single factor, covariate, predictor, latent variable, ordination axis

*So a latent variable is an ecological gradient or ordination axis representing one, or multiple, missing predictors.*

# Latent variables

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

In essence:

Community structure is generally low-dimensional.

## **At this point you might think:**

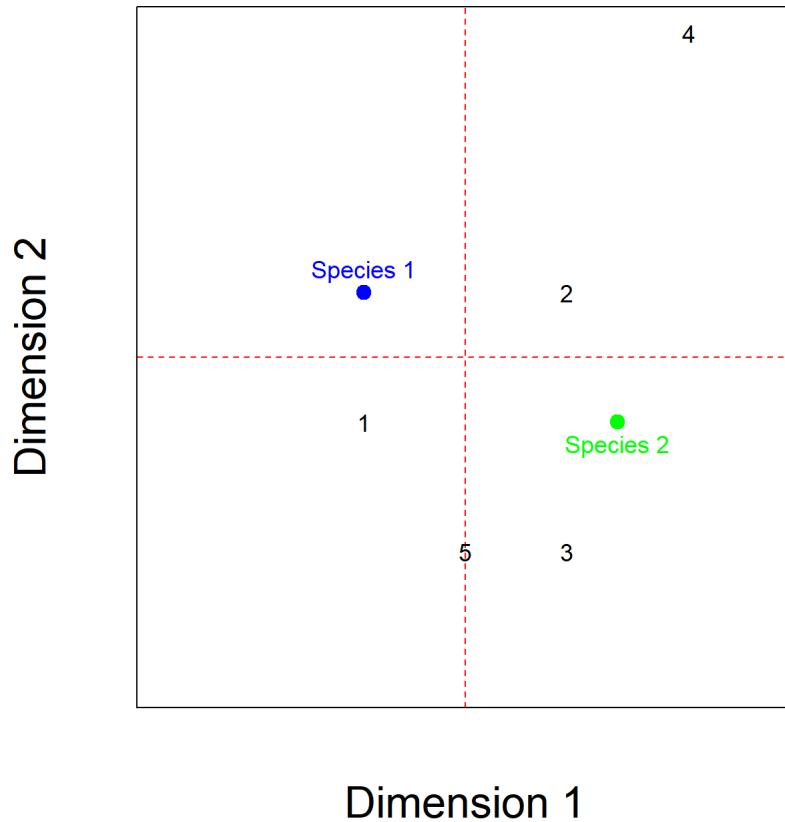
- Community ecology has been doing it for a hundred years!
- e.g. Forbes (1907)
- Ordination:
  - Principal Component Analysis Pearson 1901
  - Correspondence analysis Hirschfeld 1934
  - NMDS Kruskal 1964a,b
- Niche overlap

# Analysing multivariate data: ordination

- Termed by David Goodall in 1954: "An essay in the use of factor analysis"
- Applied factor analysis For the analysis of data on a plant community
- Reducing dimension of data
- ordering species or samples along an ecological gradient
- classically e.g.,
  - Principal Component Analysis (PCA; `prcomp()`)
  - Correspondence Analysis (CA; `cca()` in **vegan**)
  - Multidimensional scaling (PCoA; `cmdscale()`, NMDS; `metaMDS()` in **vegan**)
  - **Factor analysis:** Precursor to GLLVMs (FA; `factanal()`)
- *Treats latent variables as fixed-effects*
- So a multivariate Generalized Linear Model (kind of), i.e., a joint model

# Ordination: visual inspection

- Most common tool is the biplot Gabriel 1971
- Distance between species indicates dissimilarity
- Distance between sites indicates dissimilarity



# Classical methods have some issues..

- Ordination axis (ecological gradient) treated as fixed (parameter)
- Horseshoe or arch effect (PCA, CA)
- Difficult (near to impossible) to check any assumptions
- Mean-variance relationships Warton and Hui 2017

**In general, not very flexible.**

# Model-based thinking

Concept: apply regression concepts to multivariate analysis Warton et al. 2015



# Model-based ordination to the rescue!

- Ordination = dimension reduction
- Represent the latent complex ecological gradient
- A model like in regression
- Represent species associations with latent variables
- So JSMD = ordination? Yes! (for GLLVMs)
- "Model-based approaches to unconstrained ordination" Hui et al. 2015

**All the benefits from regression and ordination!, e.g.:**

- Procrustus analysis
- Biplots
- Model-selection
- Residual diagnostics
- Appropriate mean-variance relationships
- Hypothesis testing
- No distance metrics



# First: specify a multivariate GLM

- $\beta_{0j}$  intercept per species
- $x_{ik}$  site-specific predictors
- $\beta_j$  species-specific slopes

$$g(E(y_{ij}|\mathbf{x}_i)) = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j \quad (1)$$

- Stacked SDM (fit with `glm(..)` function)
- No random-effects

## Second: a mixed-effects model

- Add residual for  $i = 1 \dots n$  sites and  $j = 1 \dots p$  species

$$g(\mathbb{E}(y_{ij}|\boldsymbol{x}_i)) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_{ij}, \quad \boldsymbol{u}_i \sim \mathcal{N}(0, \Sigma) \quad (2)$$

- Structure  $\boldsymbol{u}_{ij}$  with  $\Sigma$  by species
- $\Sigma$  are species covariances or *associations*
- A "joint species distribution model" Pollock et al. 2014
- Can be fit using standard mixed-effects modeling software.

In `lme4`:

```
glmer(abundance~species+x:species+
      (0+species|sites), family="poisson", data=data)
```

- Too many parameters in  $\Sigma$
- Number of parameters increases rapidly
- Often not feasible to fit

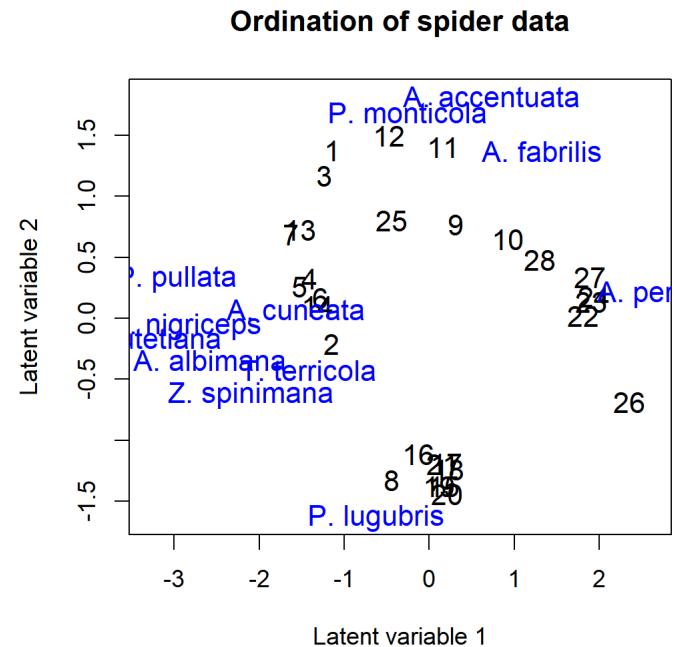
# Reduced rank $\Sigma$

- Add factor analytic structure to  $\Sigma$
- Dimension reduction = ordination
- $u_{ij} = \epsilon_i^\top \theta_j$ 
  - i.e.  $u_i \sim \mathcal{N}(0, \theta_j \theta_j^\top)$
- Faster and fewer parameters:
  - Number of parameter doesn't grow so fast
  - More latent variables, better estimation of  $\Sigma$
- So we are estimating residual covariances!

$$\Sigma = \begin{bmatrix} \theta_{11} & 0 & 0 \\ \theta_{12} & \theta_{22} & 0 \\ \vdots & \ddots & \vdots \\ \theta_{1j} & \cdots & \theta_{dj} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1j} \\ 0 & \theta_{22} & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{dj} \end{bmatrix} \quad (3)$$

# Generalized Linear Latent Variable Models

- Still a mixed-effects model
- $d$  latent variables treated as random-effect
- Produces ordination
  - "site scores" :  $\epsilon_i$
  - "species scores" or "loadings":  $\theta_j$
- Species and sites far apart are dissimilar
- E.g., because species prefer different environments



$$g(\mathbb{E}(y_{ij} | \mathbf{x}_i, \epsilon_i)) = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \epsilon_i^\top \boldsymbol{\theta}_j, \quad \epsilon_i \sim \mathcal{N}(0, \mathbf{I}) \quad (4)$$

# gllvm R-package

## Methods in Ecology and Evolution



APPLICATION

Free Access

**gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R**

Jenni Niku✉, Francis K. C. Hui, Sara Taskinen, David I. Warton

First published: 21 September 2019 | <https://doi.org/10.1111/2041-210X.13303> | Citations: 4

# Break / Questions

🐦 On twitter: #**GLLMs**, @**vdVeenB** or @**J\_\_Niku** or @**BigDrFishBoi**

🐙 On github: <https://github.com/BertvanderVeen/IRSAE2021GLLMworkshop/discussions>

Or in the chat.