Analysing multivariate ecological data with Generalized Linear Latent Variable Models

Bert van der Veen
Jenni Niku
Sam Perrin

CENTRE FOR BIODIVERSITY DYNAMICS
JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ
NTNU
NIBIO

Exercise material and package installation

https://github.com/BertvanderVeen/IRSAE2021GLLVMworkshop

Questions

In zoom chat or "raise" your hand

On twitter: #**GLLVMs**, @**vdVeenB** or @**J__Niku** or @**samperrinNTNU**

On github: https://github.com/BertvanderVeen/IRSAE2021GLLVMworkshop/discussions

Welcome! 😄

	Who	Affiliation	Expertise
	Bert van der Veen PhD candidate	Norwegian insitute of Bioeconomy research & Norwegian university of Science and Technology	- 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	- 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

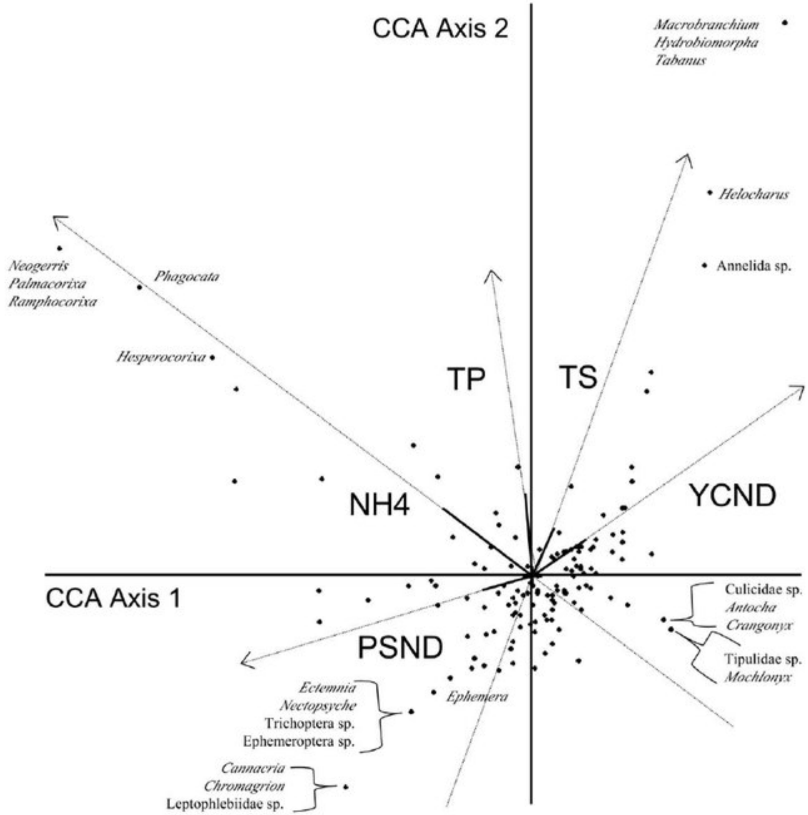
Finish

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

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

CrossMark

Model-based thinking for community ecology

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

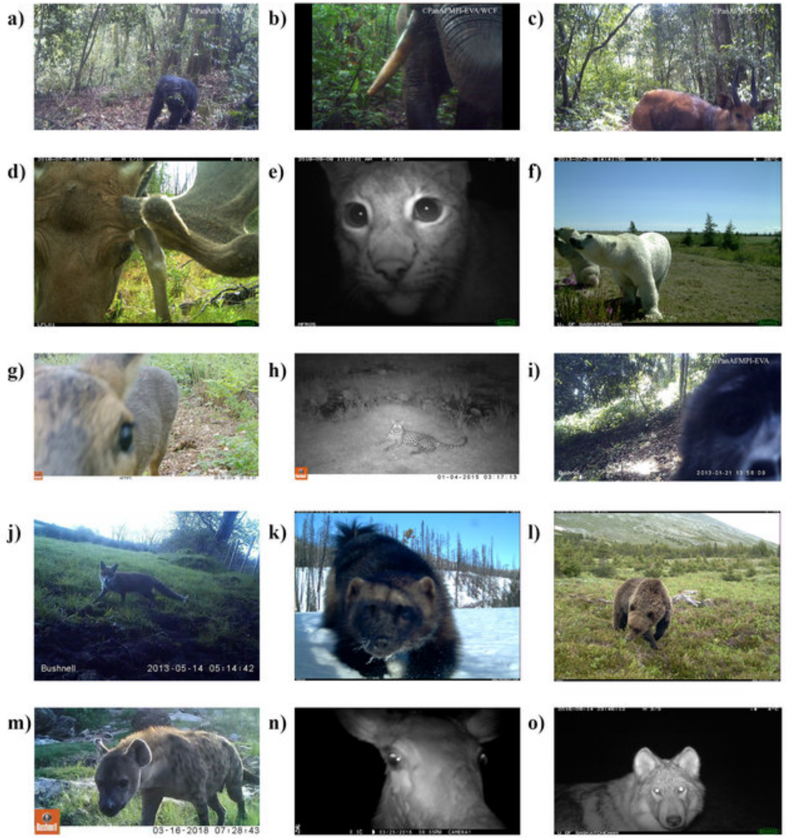
Received: 29 January 2014 / Accepted: 30 May 2014 / Published online: 19 November 2014

Gathering data

We go out, register species at multiple sites

© Geir-Harald Strand / NIBIO

e.g., camera trap data



Caravaggi et al. 2020

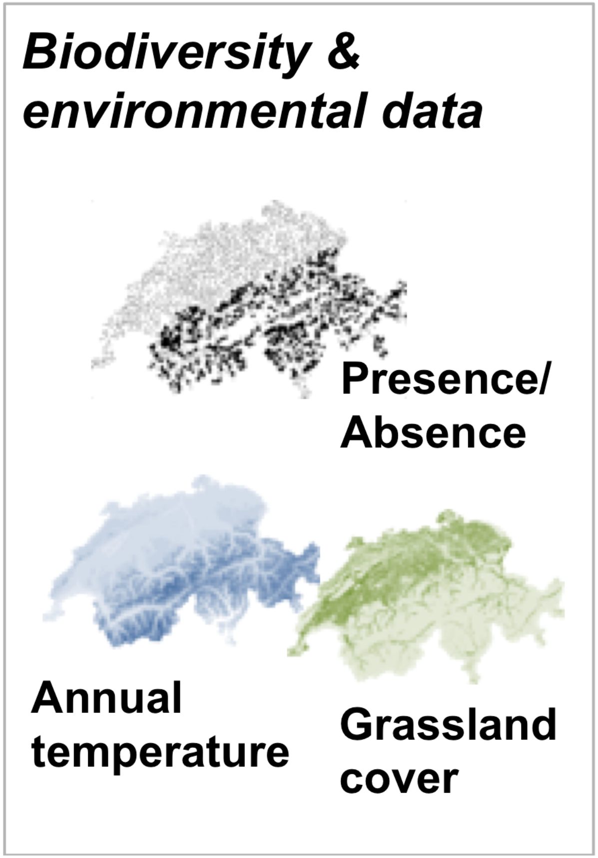
"Multivariate"

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

	Species 1	Species 2	Species 3	Species 4	Species 5
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

"Multivariable"

Multiple **predictors** that represent the environment



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

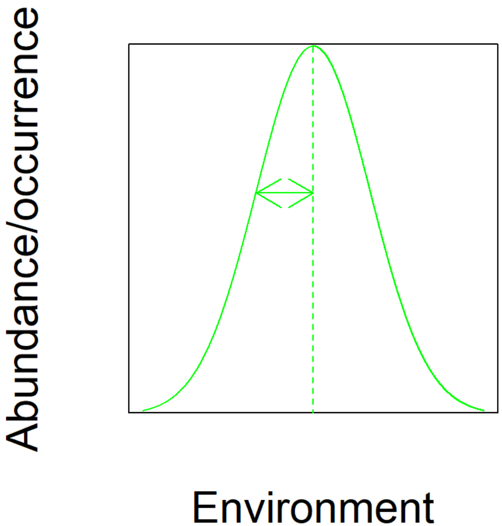
To clarify

- 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?

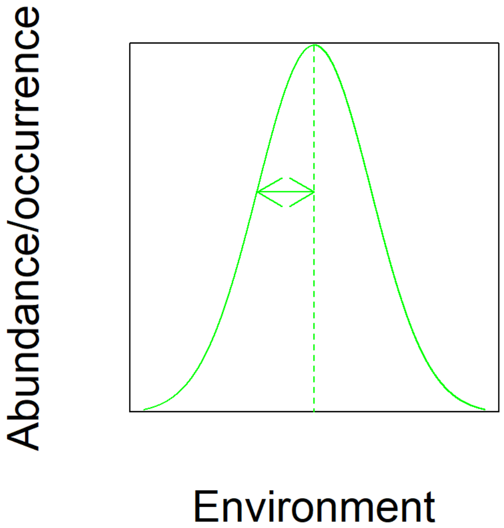
- Interest in **co**-occurrence patterns
  - In contrast to only **occurrence** patterns (a species distribution)
- Why do species co-occur?
  - Similar environmental preferences
  - Similar history in the environment
  - Might result in **Interactions**
- Multiple species form a **community**

Occurrence pattern

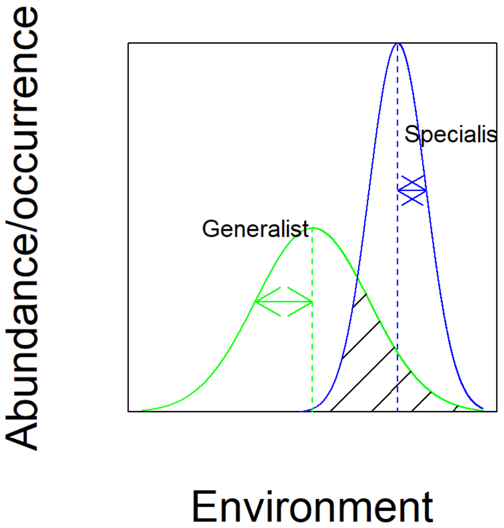


Why analyse multivariate data?

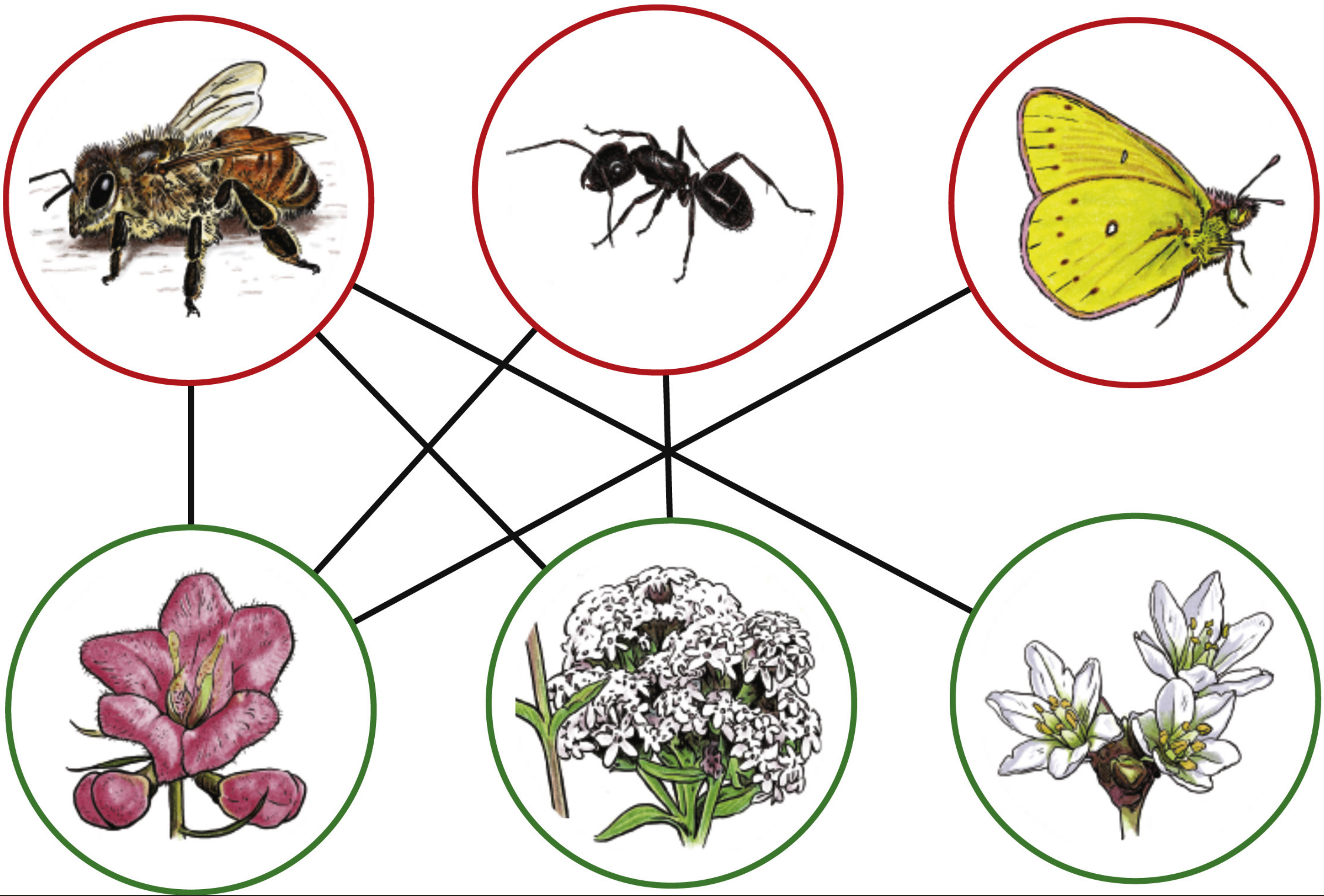
Occurrence pattern



Co-occurrence pattern



But then for more species



Joint modeling with latent variables

- Accounts for correlation between taxa
- "Borrows" information from other species for estimation

- Provides species associations

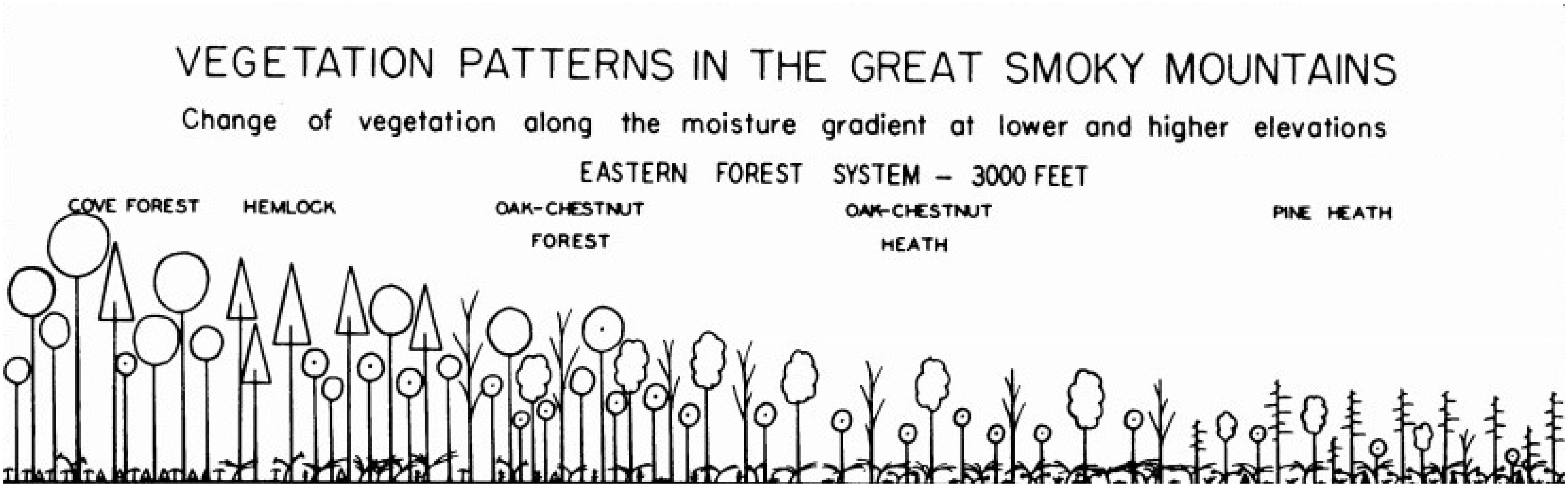
- Concept: fit a single model for all species

- I.e., a "Joint Species Distribution Model"
  - Faster
  - Less tedious
  - Explicitly model species co-occurrence
  - Etc.

Key references

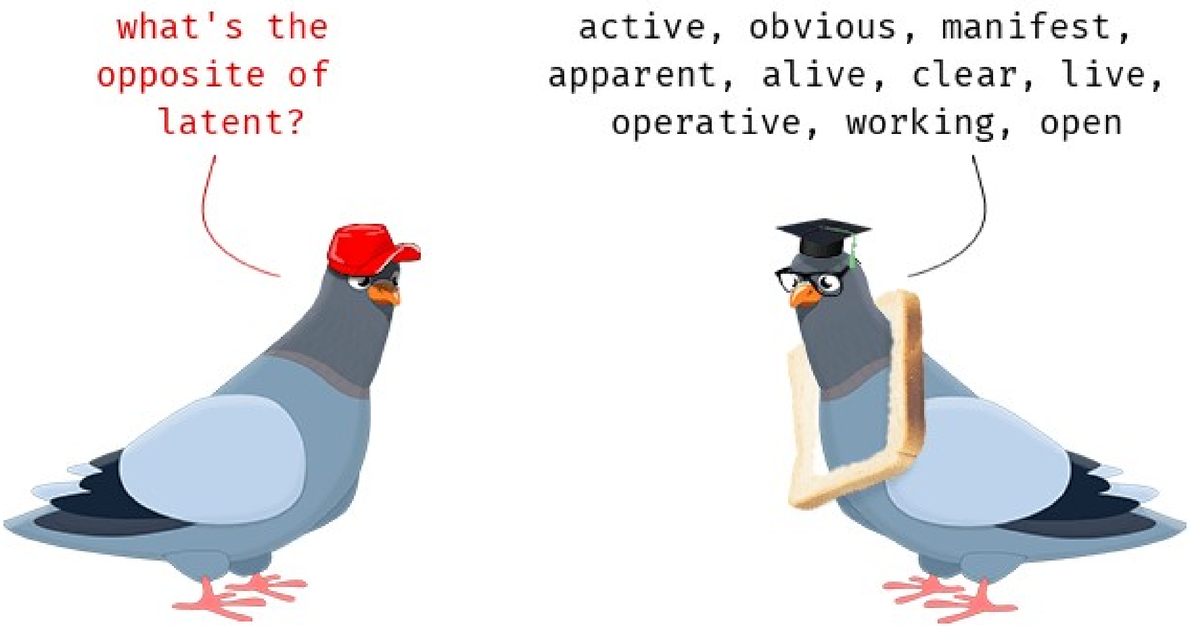
- Warton et al. 2015: "So many variabes: Joint modeling in community ecology"
- Poggiato et al. 2021: "On the interpretations of joint modeling in community ecology"
- Blanchet et al. (2020): "Co-occurrence is not evidence of ecological interactions"

Latent variables?



Whittaker 1967

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



If not measured, it is latent, like a fixed versus 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.

Analysing co-occurrence patterns

At this point you might think:

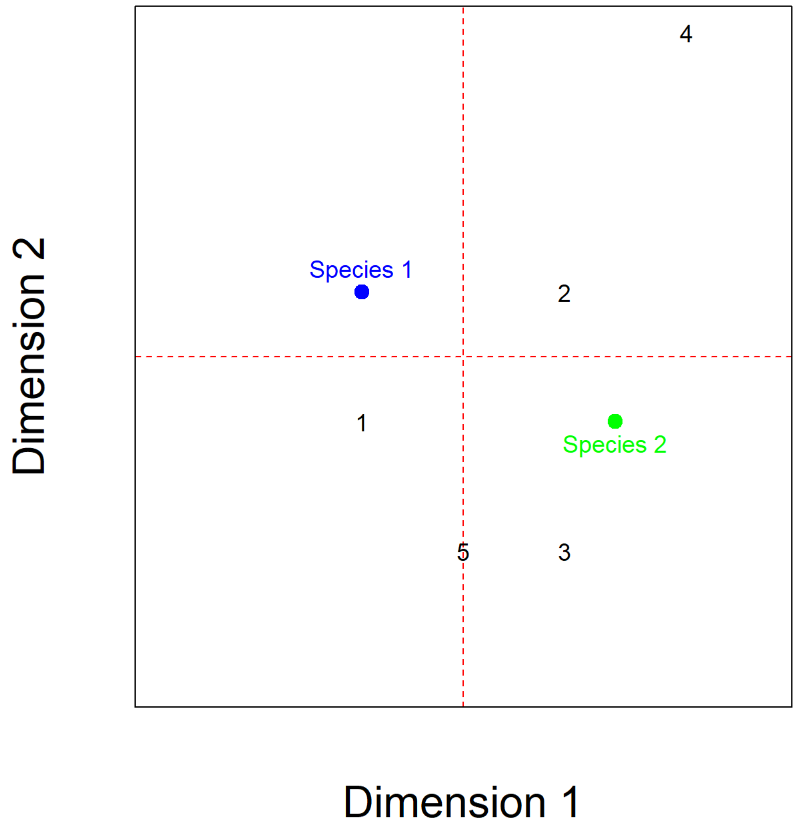
- Community ecology has been doing it for a hundred years!
- e.g. Forbes (1907)
- Ordination: Principal Component Analysis Pearson 1901, Correspondence analysis Hirschfeld 1934, NMDS (Kruskal 1964a,b)
- Niche overlap

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



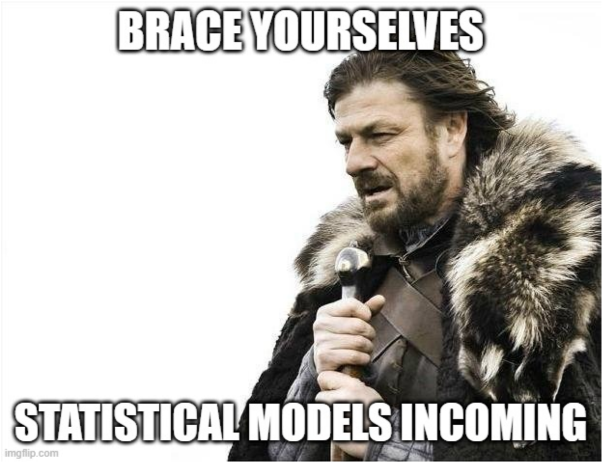
Classical methods have some issues..

- 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
    - Explicit statistical models
    - Residual diagnostics
    - Model selection
    - et cetera



Specifying a multivariate statistical model

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

$$g(\mathrm{E}(y_{ij}|\boldsymbol{x}_i)) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j \tag{2}$$

- Stacked SDM or `glm(.)` function
- Without random-effects

A Multivariate Mixed-effects model

- Add residual for $i = 1 \dots n$ sites and $j = 1 \dots p$ species
- Structure Σ by species
- Σ are species covariances or *associations*
- A "joint species distribution model" Pollock et al. 2014

$$g(\mathrm{E}(y_{ij}|\boldsymbol{x}_i)) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Sigma) \tag{3}$$

- Can be fit using standard mixed-effects modeling software.

In `lme4`:

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

- Σ has $p(p+1)/2$ parameters (which increases quadratically with # species)

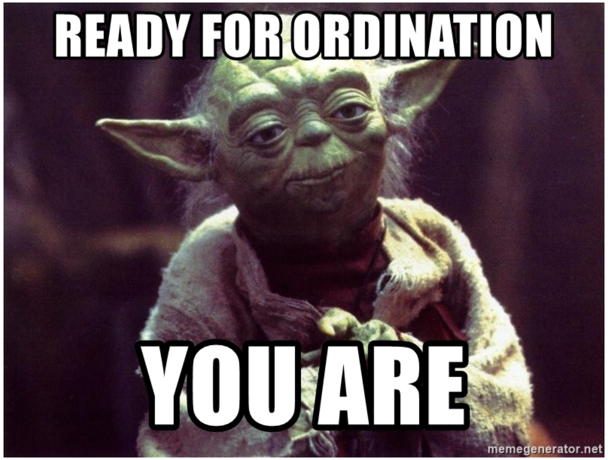
Model-based ordination to the rescue!

- Ordination = dimension reduction
- Represent the latent complex ecological gradient
- A model like in regression
- Represent species associations with latent variables
- So JSDM = 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



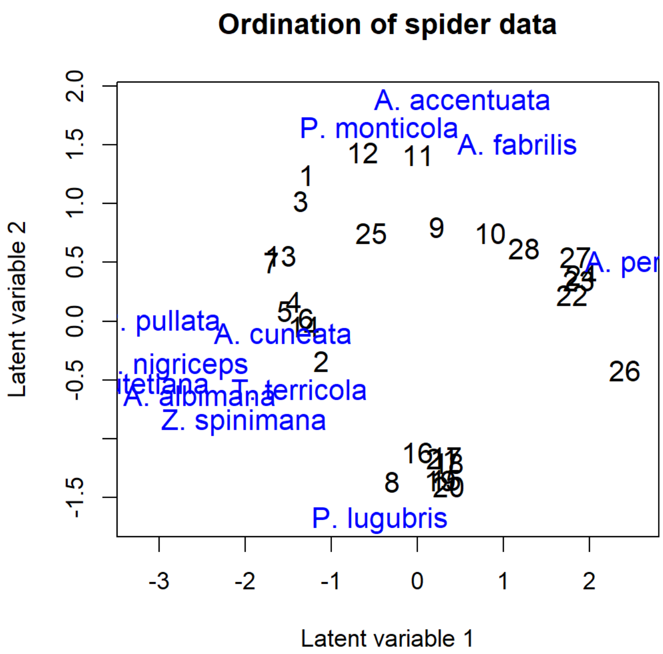
Generalized Linear Latent Variable Models

- GLLVM for short
- Add factor analytic structure to Σ
- Ordination = dimension reduction
- $\epsilon_{ij} = \boldsymbol{u}_i^\top \boldsymbol{\theta}_j$
  - i.e. $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\theta}_j \boldsymbol{\theta}_j^\top)$
- Faster and fewer parameters:
  - Number of parameter doesn't grow so fast
  - More latent variables, better estimation of Σ

$$\Sigma = \begin{bmatrix} \theta_{11} & 0 & 0 \\ \theta_{12} & \theta_{22} & 0 \\ \vdots & \ddots & \vdots \\ \theta_{1j} & \cdots & \theta_{dj} \end{bmatrix} \begin{bmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1j} \\ 0 & \theta_{22} & \ddots & \vdots \\ 0 & 0 & \cdots & \theta_{dj} \end{bmatrix} \tag{4}$$

Generalized Linear Latent Variable Models

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



$$g(\mathrm{E}(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i)) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\theta}_j, \qquad \boldsymbol{u}_i \sim \mathcal{N}(0, \boldsymbol{I}) \tag{5}$$

Compared to (unconstrained) classical ordination

So GLLVMs can be used as, e.g.,

- Joint species distribution model (with reduced-rank residual covariance matrix),
- Multivariate GLM (no random-effects),
- Unconstrained Ordination
- Constrained ordination (see last part)

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

- **GLLVMs are flexible**
- **Because the statistical model can be extended**

`gllvm` R-package

Methods in Ecology and Evolution — BRITISH ECOLOGICAL SOCIETY

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: #**GLLVMs**, @**vdVeenB** or @**J__Niku** or @**samperrinNTNU**

On github: https://github.com/BertvanderVeen/IRSAE2021GLLVMworkshop/discussions

Or on zoom.