

Analysing multivariate ecological data with Generalized Linear Latent Variable Models

Bert van der Veen

Jenni Niku

Sam Perrin

Robert Brian O'Hara



Exercise material and package installation

<https://github.com/BertvanderVeen/BES2020GLLMworkshop>

Questions

In zoom chat to **Bob** or

- 🐦 On twitter: #**GLLVMs**, @**vdVeenB** or @**J__Niku** or @**samperrinNTNU** or @**BobOHara**
- 🗨️ On github: <https://github.com/BertvanderVeen/BES2020GLLMworkshop/discussions>

Welcome! 😊

Who



Bert van der Veen
PhD candidate

Affiliation

Norwegian insititute of
Bioeconomy research &
Norwegian univeristy 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 candidate

Norwegian university of
Science and Technology

- Fresh water ecology
- Invasion ecology
- Species distribution modeling



Robert Brian O'Hara
Professor

Norwegian University of
Science and Technology

- Statistical ecology
- Species distribution modeling
- Data integration

Program day 2

Topic

- Species associations
- Species distribution modeling
- Generalized Linear Latent Variable models

Duration

20 minutes

Who



Break

5 minutes

- **gllvm** R-package (Niku et al. 2019) 20 minutes
 - How to: species distributions with GLLVMs
-
- Exercise: break out 10 minutes



Break

5 minutes

- Some ecology 20 minutes
 - Exercise: break out 15 minutes
-
- Discuss results exercise



Gathering data

We go out, register species at multiple sites



© Geir-Harald Strand / NIBIO

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

| | 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**
- E.g. measurements of the environment

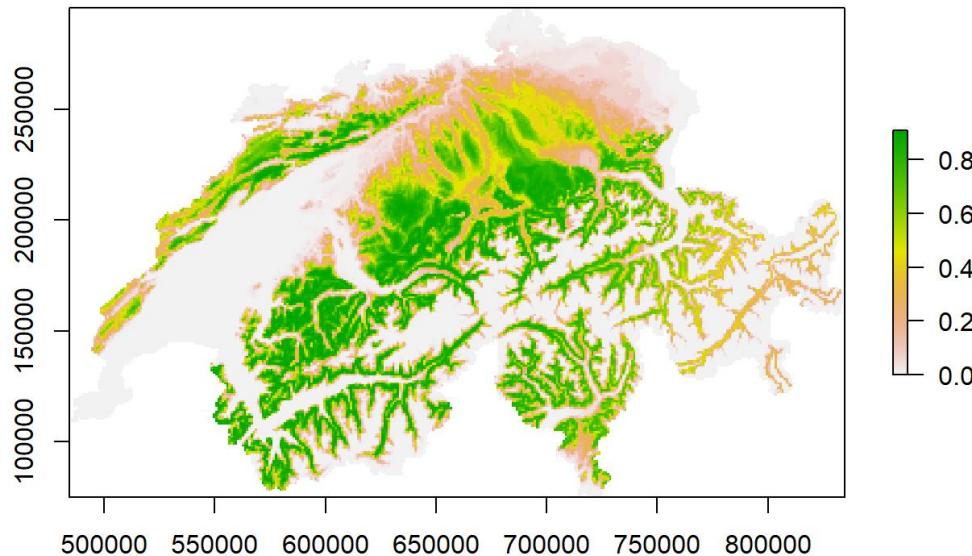
| | Predictor 1 | Predictor 2 | Predictor 3 | Predictor 4 | Predictor 5 | Predictor 6 |
|--------|-------------|-------------|-------------|-------------|-------------|-------------|
| Site 1 | 2.3321 | 3.0445 | 0.0000 | 3.0445 | 4.4543 | 3.9120 |
| Site 2 | 3.0493 | 3.2581 | 1.7918 | 1.0986 | 4.5643 | 1.6094 |
| Site 3 | 2.5572 | 3.5835 | 0.0000 | 2.3979 | 4.6052 | 3.6889 |
| Site 4 | 2.6741 | 4.5109 | 0.0000 | 2.3979 | 4.6151 | 2.9957 |
| Site 5 | 3.0155 | 2.3979 | 0.0000 | 0.0000 | 4.6151 | 2.3026 |
| Site 6 | 3.3810 | 3.4340 | 3.4340 | 2.3979 | 3.4340 | 0.6931 |

To clarify

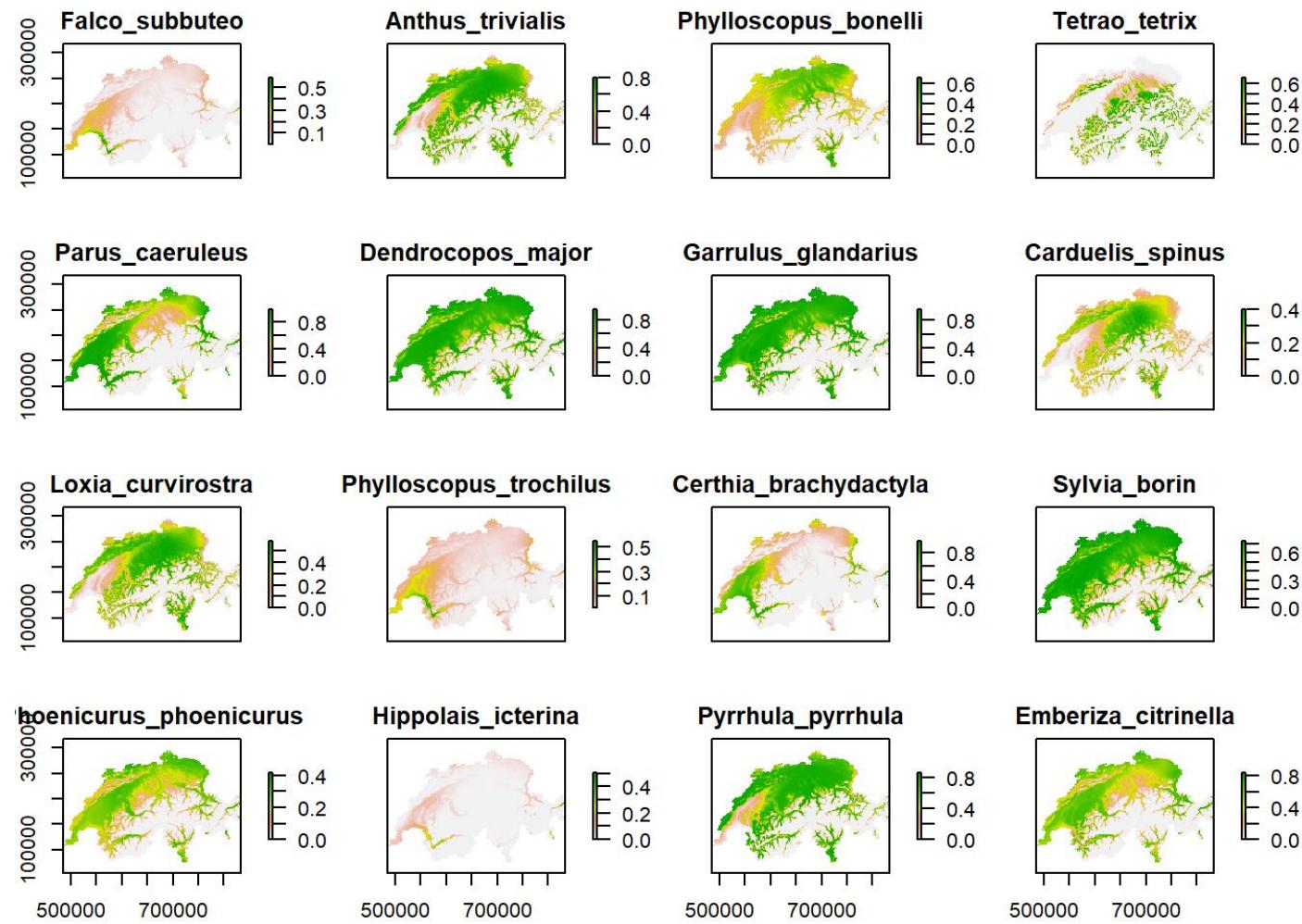
- Both data and model can be univariate or multivariate
- Multivariate data can be analysed with both multivariate and univariate models (SDM, ordination)
- Multivariable data can be used in multivariate or univariate analysis
 - Generally the same for all responses
 - (But, note that the model can of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

Why analyse species distributions?

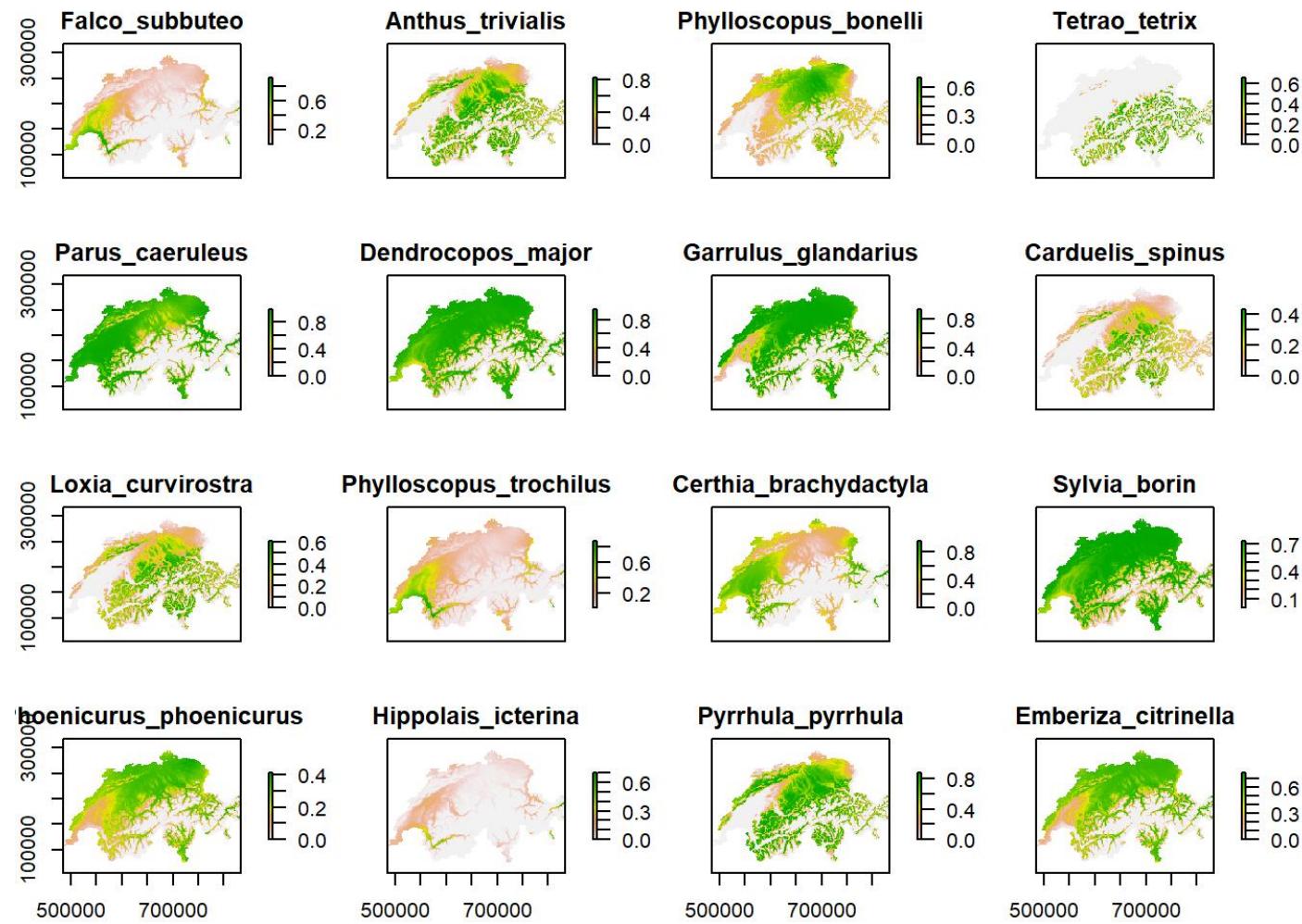
- For example:
 - Where does my species occur?
 - Where can my species occur?
 - How might that change in the future?
 - Informs us of the state of a species



But then for more species



And potentially predict into the future

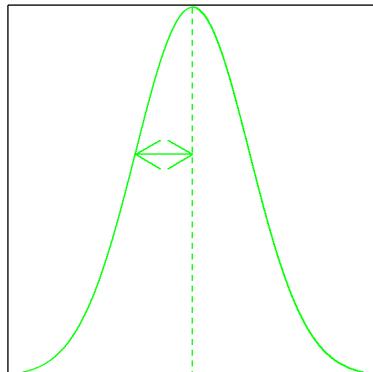


Species associations

- Our data contains information on where **multiple species** occur at the same time!
 - Similar environmental preferences
 - Similar history in the environment
 - Might result in **Interactions**
- Multiple species form a **community**

Abundance/occurrence

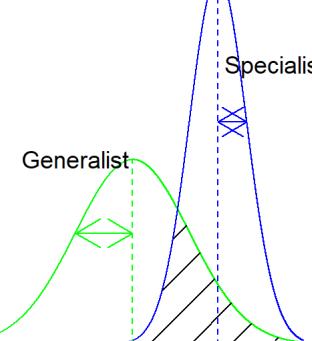
Occurrence pattern



Environment

Co-occurrence pattern

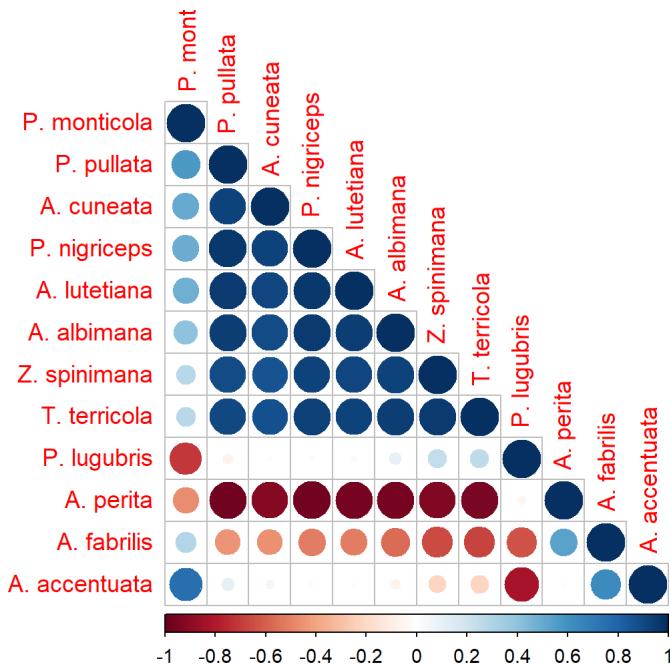
Abundance/occurrence



Environment

Co-occurrence

- Occurrence data can tell us about where multiple species occur
- How species are "associated"
- Statistically: correlation
- So we are looking for patterns in our data



Analysing co-occurrence patterns

- Community ecology has been doing it for a hundred years
- e.g. Forbes (1907) or Goodall (1954)
- Ordination: Principal Component Analysis (PCA), Correspondence analysis (CA), NMDS
- Niche overlap
- Some discussion in (2020): Co-occurrence is not evidence of ecological interactions"

Joint modeling

- Account for correlation between taxa
- "borrow" information from other species
- Estimate species associations
- Concept: fit a single model for all species
 - Faster
 - Less tedious
 - Explicitly model species co-occurrence
 - Etc.

Patterns in our data

Can be **observed** (covariates) or **latent**

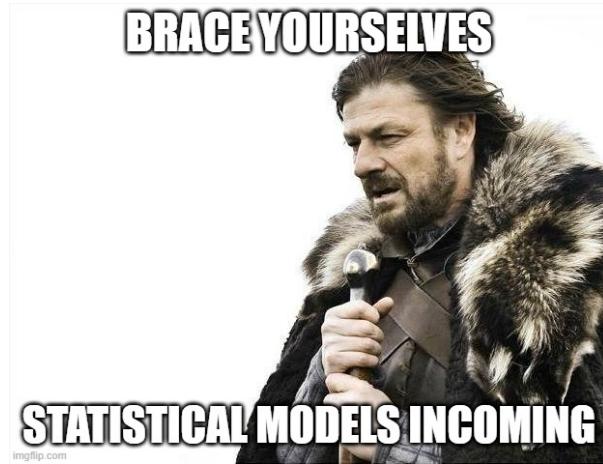
what's the
opposite of
latent?



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



Questions so far?



Specifying a multivariate statistical model

- β_{0j} intercept per species
- X_{ik} site-specific covariates
- β_{0j} species-specific coefficients

$$g(\mathbf{E}(y_{ij})) = \beta_{0j} + \mathbf{X}_i^\top \boldsymbol{\beta}_j \quad (2)$$

- Stacked SDM or `glm()` function
- For example: **mvabund** Wang et al. 2012
- No random-effects
- Hypothesis testing for multivariate data

A Multivariate Mixed-effects model

- Add residual for $i = 1 \dots n$ sites and $j = 1 \dots p$ species
- Structure Σ by species
- "Joint species distribution model" Pollock et al. 2014
- Can fit using standard mixed-effects modeling software.

$$g(\mathbb{E}(y_{ij})) = \beta_{0j} + \mathbf{X}_i^\top \boldsymbol{\beta}_j + \epsilon_{ij}, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \Sigma) \quad (3)$$

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 species associations with latent variables
- So JSMD = ordination?
- "Model-based approaches to unconstrained ordination" Hui et al. 2015

All the benefits from regression and ordination!

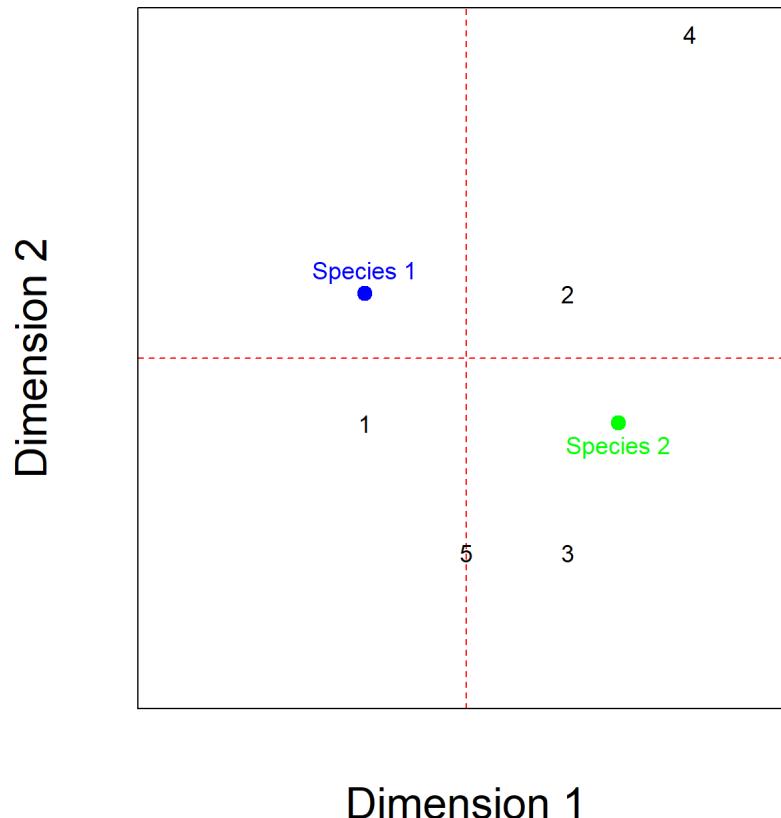
e.g.

- Procrustus analysis
- Biplots
- Model-selection
- Residual diagnostics
- Appropriate mean-variance relationships
- Hypothesis testing
- etc.



Ordination: visual inspection

- Most common tool is the biplot Gabriel 1971
- Distance between species indicates dissimilarity
- Distance between sites indicates dissimilarity
- What is a latent variable?



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
- Can be represented as a single factor, covariate, predictor, latent variable, ordination axis

Ecological gradients

"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.

Generalized Linear Latent Variable Models

- GLLVM for short
- Add factor analytic structure to Σ
- $\epsilon_{ij} = \mathbf{u}_i^\top \boldsymbol{\theta}_j$
 - 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

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

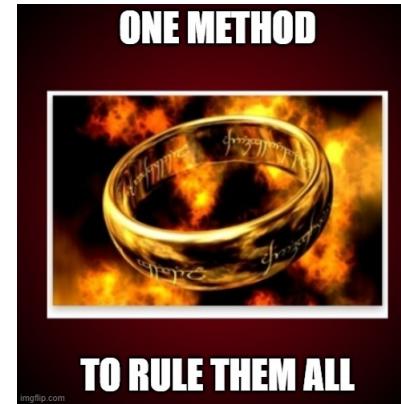
Generalized Linear Latent Variable Models

- Still a mixed-effects model
- Low-rank approach to Σ
- d latent variables treated as random-effect
- Produces ordination
 - "site scores": \mathbf{u}_i
 - "species scores" or "loadings": $\boldsymbol{\theta}_j$
 - No varimax

$$g(\text{E}(y_{ij})) = \beta_{0j} + \mathbf{X}_i^\top \boldsymbol{\beta}_j + \mathbf{u}_i^\top \boldsymbol{\theta}_j, \quad \mathbf{u}_i \sim \mathcal{N}(0, \mathbf{I}) \quad (5)$$

Compared to SDMs

- Multivariate GLM is a GLLVM without latent variables
 - (so the same for a GAM)
 - Multivariate GLMM is a GLLVM if the covariance matrix is of reduced rank
-
- **GLLVMs are a flexible approach to estimate species distributions**
 - **That can readily be extended**



More complex JSDM?

e.g. spatially dependent sites

e.g. temporal replicates

e.g. including species relatedness



APPLICATION | Open Access |

Joint species distribution modelling with the R-package HMSC

Gleb Tikhonov, Øystein H. Opedal, Nerea Abrego, Aleksi Lehikoinen, Melinda M. J. de Jonge, Jari Oksanen, Otso Ovaskainen

First published: 25 December 2019 | <https://doi.org/10.1111/2041-210X.13345> | Citations: 19