Beyond vanilla GLLVMs

Bert van der Veen

Department of Mathematical Sciences, NTNU

Questions so far?

Some topics covered in this workshop

- Community data, Vector GLM and, GLMMs (day 1)
- Residual diagnostics and model comparison, fourth-corner approach and phylogenetic random effects, JSDM (day 2)
- All forms of ordination (day 3)
- Other R packages and beyond GLLVMs (day 4)

Topics not covered in this workshop

- Joint models with imperfect detection (Tobler et al. 2019, Doser et al. 2023, Hogg et al. 2021)
- Mixture models for clustering species (Hui et al. 2015, Hui 2017)
- Large scale spatio-temporal analysis
- Other methods for analysing community ecological data
- Ordination of excess zeros

Correlated LVs

At present, it is possible to incorporate spatially or temporally correlated LVs in GLLVMs.

$$\mathbf{u}_q \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}) \tag{1}$$

As we know: assuming independence messes up our results. Space/time is no different than species' correlation.

Received: 5 January 2024 | Accepted: 20 August 2024

DOI: 10.1111/2041-210X.14420

Methods in Ecology and Evolution

RESEARCH ARTICLE

Spatial confounding in joint species distribution models

Francis K. C. Hui[1] | Quan Vu[1] | Mevin B. Hooten[2]

Correlated LVs

At present, it is possible to incorporate spatially or temporally correlated LVs in GLLVMs.

$$\mathbf{u}_q \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{1}$$

Biased covariate effects and the ordination

As we know: assuming independence messes up our results.

Space/time is no different than species' correlation.

Received: 5 January 2024 | Accepted: 20 August 2024
DOI: 10.1111/2041-210X.14420

Methods in Ecology and Evolution

RESEARCH ARTICLE

Spatial confounding in joint species distribution models

Francis K. C. Hui[1] | Quan Vu[1] | Mevin B. Hooten[2]

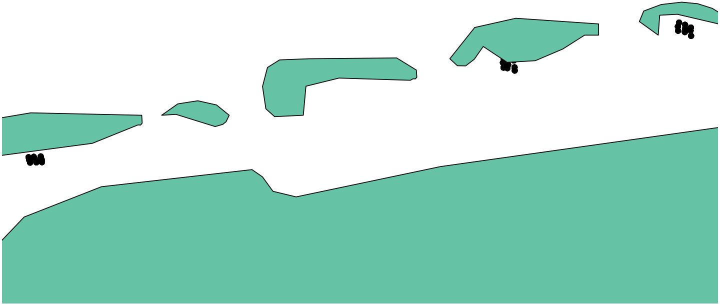
Spatial latent variable models

The goal: sites in close vicinity are plased close together in the ordination (incorporate extra information)

- Difficult to fit quickly; we have a site by site covariance matrix
- With (for example) 1000 sites, we need to invert a very big matrix
- Tricks need to be applied to avoid that; similar issue to the phylogenetic model
- This is why (e.g.,) INLA (Rue et al., 2009) was developed

`gllvm`, like `glmmTMB`, does not yet handle this in an efficient fashion

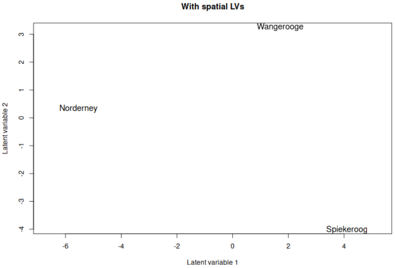
Example: Wadden data

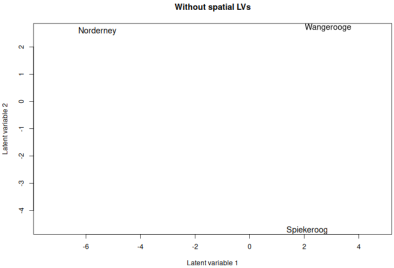


Example: fit correlated LVs

```
model1 <- gllvm(Y, studyDesign = X[,"island",drop=FALSE],
                num.lv = 2, lvCor = ~corExp(1|island),
                distLV = aggregate(coords, FUN = mean, by =list(X$island)),
                family = "tweedie", Power = NULL,
                maxit = 10e4, sd.errors = FALSE)
# Compare to IID LVs
model2 <- gllvm(Y, studyDesign = X[,"island",drop=FALSE],
                num.lv = 2, lvCor = ~(1|island), family = "tweedie",
                sd.errors = FALSE, Power = NULL)
```

Example: ordination plots





Spatial field parameters: 5.2510121×10^{-4}, 8.2825986×10^{-4}

Phylogenetic latent variable models

To bring phylogeny in the ordination (or other species information) the ideas are the same, but requires the loadings to be random effects.

Background

1. CWM + RDA *Doledec et al. (1996)*
2. Double constrained ordination *Lebreton et al. (1988), ter Braak et al. (2018)*
3. Fourth corner (LV) Models *Brown et al. (2014), Ovaskainen at al. (2017), Niku et al. (2021)*
4. Model-based double constrained ordination (Yee's `VGAM`, ter Braak and van Rossum's `douconca`)

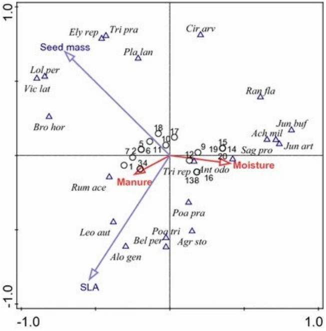


Figure 2: Quadriplot
ter Braak et al. (2018)

Reduced-rank modeling

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{E})\right\} = \mathbf{1}\boldsymbol{\beta}^\top + \mathbf{E}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2}$$

This is a GLMM

Reduced-rank modeling

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{E})\right\} = \mathbf{1}\beta^\top + \mathbf{E}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \tag{2}$$

This is an unconstrained ordination

Reduced-rank modeling

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{E})\right\} = \mathbf{1}\beta^\top + \mathbf{E}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top \tag{2}$$

This is an unconstrained ordination

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{X})\right\} = \mathbf{1}\beta^\top + \mathbf{XB} \tag{3}$$

This is a GLM

Reduced-rank modeling

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{E})\right\} = \mathbf{1}\beta^\top + \mathbf{E}, \qquad \boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boxed{\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top}) \tag{2}$$

This is an unconstrained ordination

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{X})\right\} = \mathbf{1}\beta^\top + \boxed{\mathbf{X}\mathbf{B}_{lv}\boldsymbol{\Gamma}^\top} \tag{3}$$

This is a constrained ordination

Reduced-rank modeling

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{E})\right\} = \mathbf{1}\boldsymbol{\beta}^\top + \mathbf{U}\boldsymbol{\Gamma}^\top, \qquad \mathbf{U} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

This is an unconstrained ordination

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{X})\right\} = \mathbf{1}\boldsymbol{\beta}^\top + \mathbf{U}\boldsymbol{\Gamma}^\top, \qquad \mathbf{U} = \mathbf{X}\mathbf{B} \tag{3}$$

This is a constrained ordination

4th corner reduced-rank

What if we plug the model from yesterday:

$$\eta_{ij} = \beta_{0j} + \mathbf{x}_i^\top(\boldsymbol{\beta}_x + \mathbf{b}_j) + \mathbf{tr}_j^\top \mathbf{B}_{xtr}\mathbf{x}_i \tag{4}$$

into the reduced rank model.

Reduced-rank 4th corner

$$\eta_{ij} = \beta_{0j} + ... + \mathbf{u}_i^\top \boldsymbol{\gamma}_j, \qquad \text{where } \boldsymbol{\gamma}_j = \boldsymbol{\varepsilon}_j + \mathbf{B}_{x,tr}\mathbf{tr}_j \quad (5)$$

if you recall from yesterday (where I wrote $\mathbf{b}_j$ instead of $\boldsymbol{\varepsilon}_j$), the covariate effects were random. So now, we get random loadings in the ordination.

Hierarchical ordination (O'Hara & van der Veen, 2022)

This is the idea behind "hierarchical ordination".

$$\underbrace{g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{U},\boldsymbol{\Gamma})\right\} = \mathbf{1}\boldsymbol{\beta}^\top}_{\text{Intercept}} + \underbrace{\mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top}_{\text{Ordination}} \tag{6}$$

All effects go into the ordination

Pros: fewer parameters, easy visualization

Cons: generally more difficult to fit/estimate

Hierarchical ordination (O'Hara & van der Veen, 2022)

This is the idea behind "hierarchical ordination".

$$g\left\{\mathbb{E}(\mathbf{Y}|\mathbf{U},\boldsymbol{\Gamma})\right\} = \mathbf{1}\boldsymbol{\beta}^\top + \mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top \tag{6}$$

Intercept

Ordination

This becomes an ordination with a true varimax rotation; the axes are orthogonal and ordered, not sensitive to species ordering.

Hierarchical ordination

The "Hierarchical" in ordination

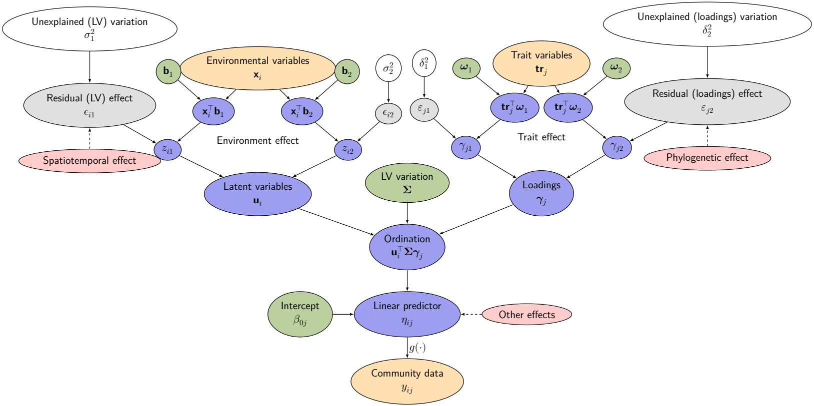
Concurrent ordination:

$$\mathbf{U} = \boxed{\mathbf{X}_{lv}\mathbf{B}_{lv}} + \boxed{\mathbf{E}} \tag{7}$$

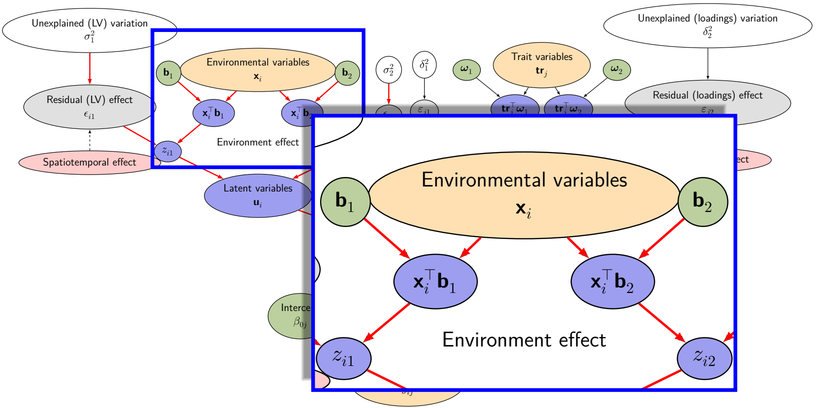
1. Covariate effects
2. LV-level errors

van der Veen et al. (2023)

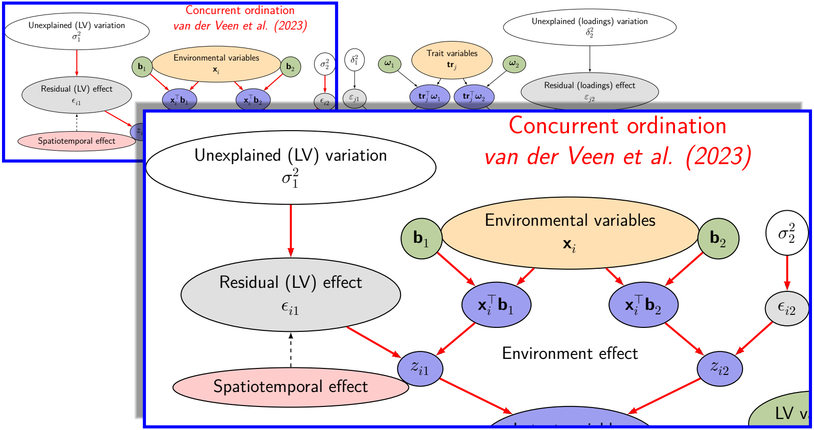
Hierarchical ordination: the framework



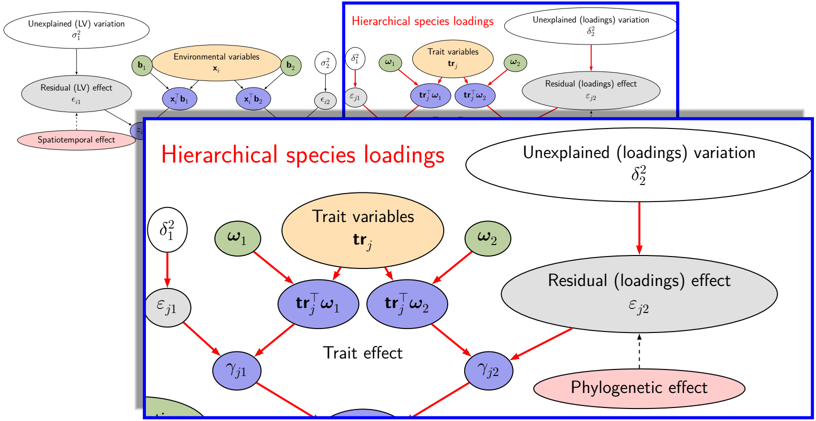
Hierarchical ordination: the framework



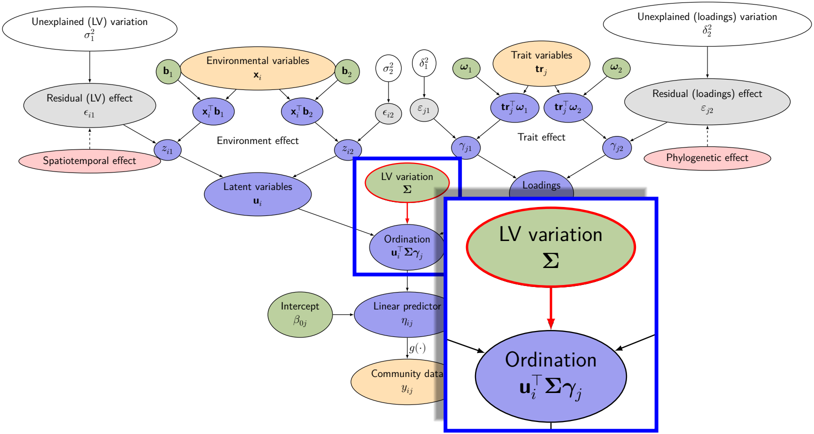
Hierarchical ordination: the framework



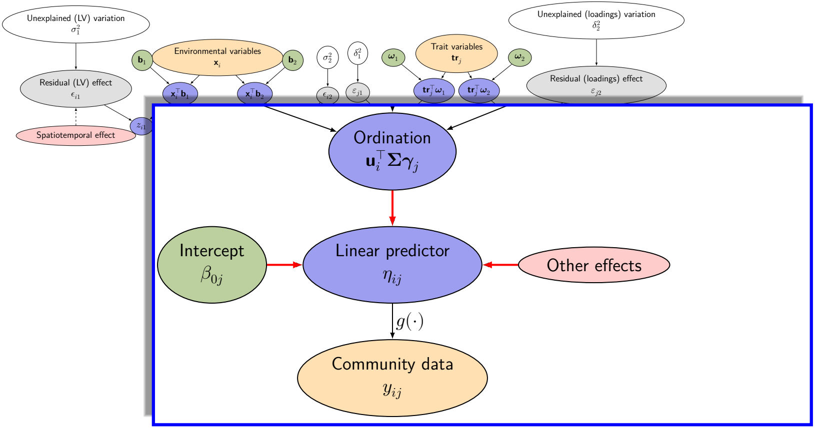
Hierarchical ordination: the framework



Hierarchical ordination: the framework



Hierarchical ordination: the framework



HO $\rightarrow$ 4th corner

$$\mathbf{U}\,\boldsymbol{\Sigma}\,\boldsymbol{\Gamma}^\top =$$

$$\underbrace{\mathbf{XB}\,\boldsymbol{\Sigma}\,\mathcal{E}^\top}_{\text{Main effects}} + \overbrace{\mathbf{XB}\,\boldsymbol{\Sigma}\,\boldsymbol{\Omega}^\top\mathbf{TR}^\top}^{\text{4th corner terms}} + \underbrace{\mathbf{E}\,\boldsymbol{\Sigma}\,(\mathbf{TR}\boldsymbol{\Omega} + \mathcal{E})^\top}_{\text{Other stuff}} \tag{8}$$

1. Induces site correlations
2. Induces species correlations

HO $\rightarrow$ 4th corner

$$\mathbf{U}\boldsymbol{\Sigma}\boldsymbol{\Gamma}^\top =$$

$$\underbrace{\mathbf{XB}\boldsymbol{\Sigma}\mathcal{E}^\top}_{\text{Main effects}} + \overbrace{\mathbf{XB}\boldsymbol{\Sigma}\boldsymbol{\Omega}^\top\mathbf{TR}^\top}^{\text{4th corner terms}} + \underbrace{\mathbf{E}\boldsymbol{\Sigma}(\mathbf{TR}\boldsymbol{\Omega}+\mathcal{E})^\top}_{\text{Other stuff}} \tag{8}$$

1. Induces site correlations
2. Induces species correlations

HO $\rightarrow$ 4th corner

Then we also have:

$$\boldsymbol{\beta} = \mathbf{B}\boldsymbol{\Sigma}\mathcal{E}^\top + \mathbf{B}\boldsymbol{\Sigma}\boldsymbol{\Omega}^\top\mathbf{TR}^\top \tag{9}$$

I.e., a hierarchical linear regression again of species responses to the environment.

Example: carabid beetles

- Data from *Ribera et al. (2001)*
  - Counts
  - 87 sites
  - 68 species
  - Beetles caught with pitfall traps
  - 26 environmental variables and 27 traits
  - in Scotland
    - Management, elevation, soil moisture, pH

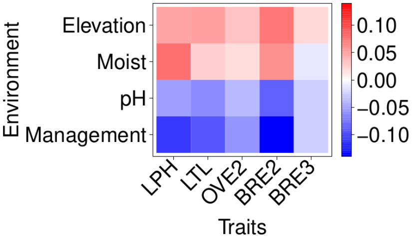
Example: carabid beetles

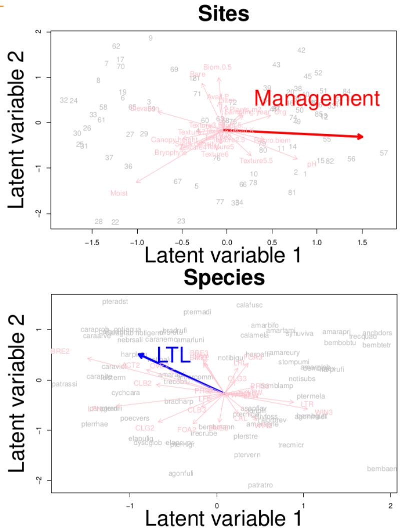
- Data from *Ribera et al. (2001)*
  - Counts
  - 87 sites
  - 68 species
  - Beetles caught with pitfall traps
  - 26 environmental variables and 27 traits
  - in Scotland
    - Management, elevation, soil

4887 → 484 parameters (for two dimensions)

Case study: results*

- Reduced-rank approximated terms (e.g., 4th corner)
- Ordination plots
- Residual correlation plots





* *fitted with Nimble* (de Valpine et al., 2017)

Next generation ordination

I think its been demonstrate, model-based multivariate analysis has many advantages

- Explicit and flexible model
- Testable assumptions
- Random effects to induce correlation
- Makes good ordinations

Sampling processes

In `gllvm` we think a lot about modeling ecological processes. However,

- Data are often not randomly sampled
- Might contain bias
- Presence-only from citizen scientists
- Mixed response types

Acommodating the sampling processes warrants further research.

Potential next steps

The big benefit of this model-based multivariate framework: we can adjust to model to accommodate our wishes. as far as technically possible

Potential next steps

The big benefit of this model-based multivariate framework: we can adjust to model to accommodate our wishes. as far as technically possible

More response types for unimodal response model?

Phylogenetic ordination

Ordination of the zero-inflated component

Spatio-temporal modeling is still a challenge.

Spatio-temporal modeling

- In space and/or time dimensions are usually large
- This slows down models considerably
- Requires additional approximations (e.g, as in INLA, CBFM, or NNGPs as in HMSC)

Wishlist `gllvm`

We continue to work on things such as:

- Improvements in usability
- Improvements of robustness (optimisation)
- Fast spatial random effects
- Incorporate space, traits and Phylogeny into ordinations

Wishlist `gllvm`

We continue to work on things such as:

- Improvements in usability
- Improvements of robustness (optimisation)
- Fast spatial random effects
- Incorporate space, traits and Phylogeny into ordinations

Do you have any wishes?

The end

One thing is clear: some very good multivariate methods have been developed in the last decade.