

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 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 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 1

Topic

- Ecological gradients
- Ordination
- Generalized Linear Latent Variable models

Duration

20 minutes

Who



Break

5 minutes

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

20 minutes



- Exercise: break out

10 minutes

Break

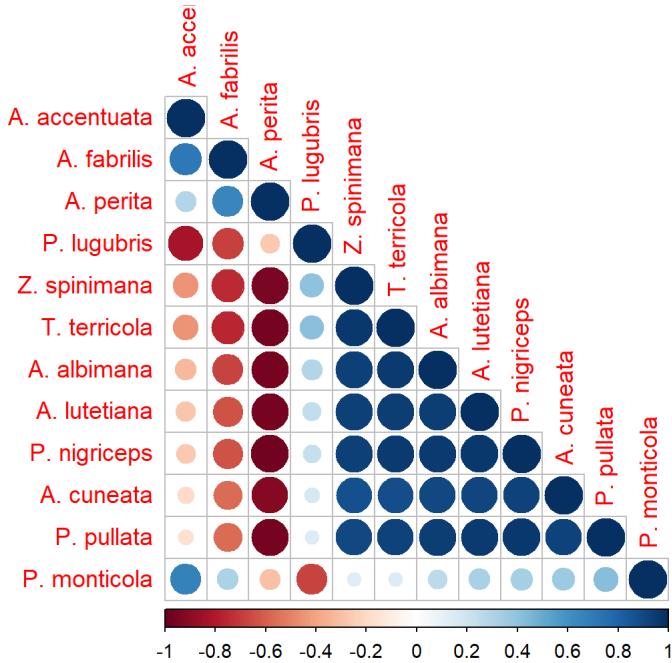
5 minutes

- Some ecology
- Exercise: break out
- Discuss results exercise

20 minutes



Program day 2



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, 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)

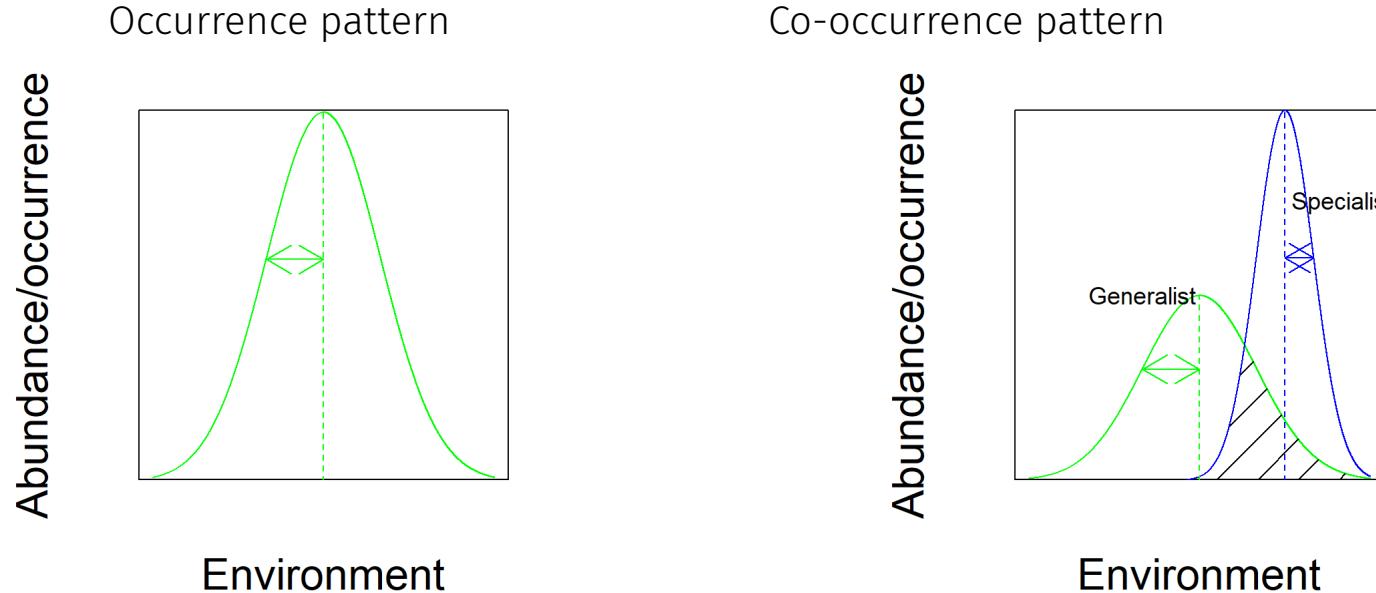
Why analyse multivariate data?

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

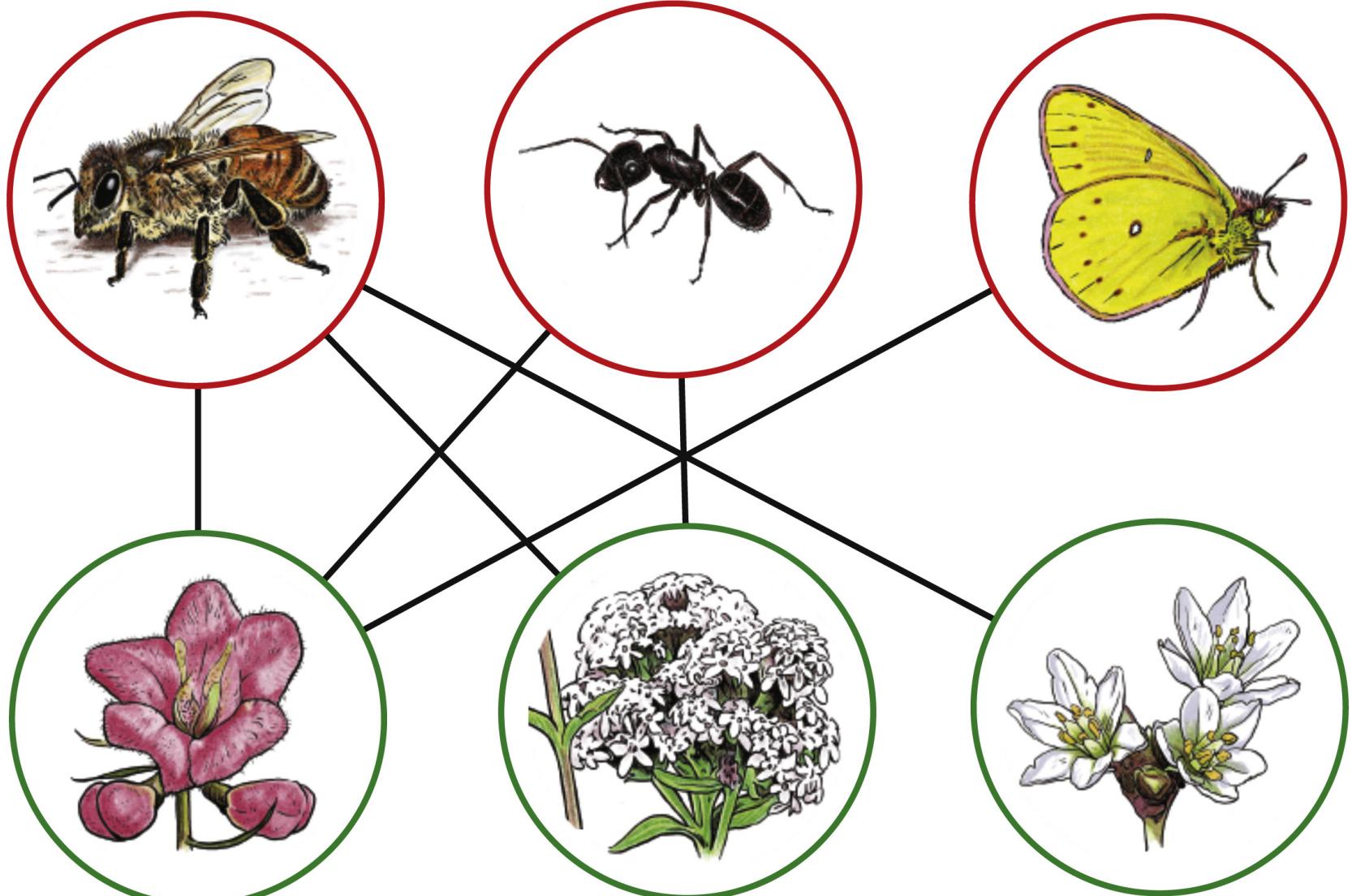
Occurrence pattern

rank

Why analyse multivariate data?



But then for more species

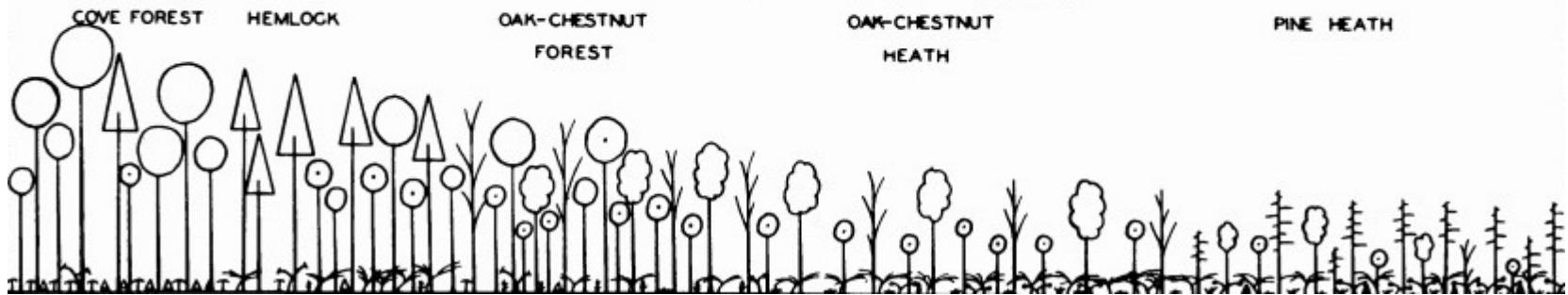


Turnover along a gradient

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)

Ecological gradient analysis

"Gradient analysis is a research approach for study of spatial patterns of species." (**Whittaker, 1967**)

Our sites describe the environment. Multiple 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
- Can be represented as a single factor, covariate, predictor, latent variable, ordination axis

Gradients can be **observed** or **latent**

what's the
opposite of
latent?



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



Ecological gradients

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

In essence:

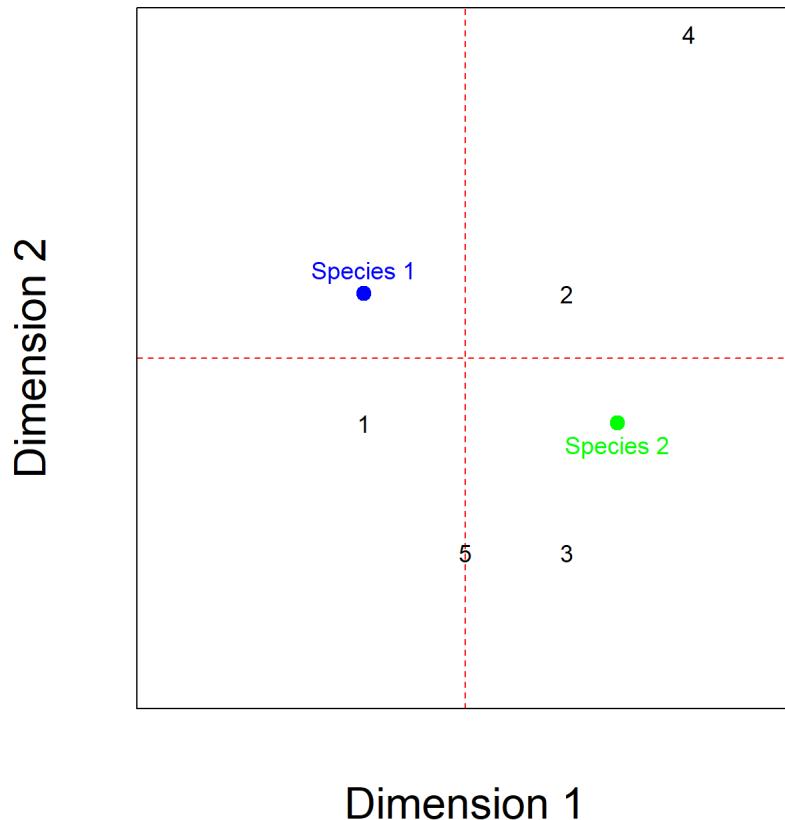
Community structure is generally low-dimensional.

Analysing multivariate data: ordination

- Termed by David Goodall (1954)
- Applied factor analysis to a community
- Reducing dimension of data
- ordering species or samples along an ecological gradient
- 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()`)

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.

Questions so far?



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
- X_{ik} site-specific covariates
- β_j species-specific slopes

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

- Stacked SDM or `glm(.)` function
- For observed gradients (predictors)

A Multivariate Mixed-effects model

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

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

- 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!

- Represent the latent complex ecological gradient
- A model like in regression
- "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
- No distance metrics
- Hypothesis testing
- etc.

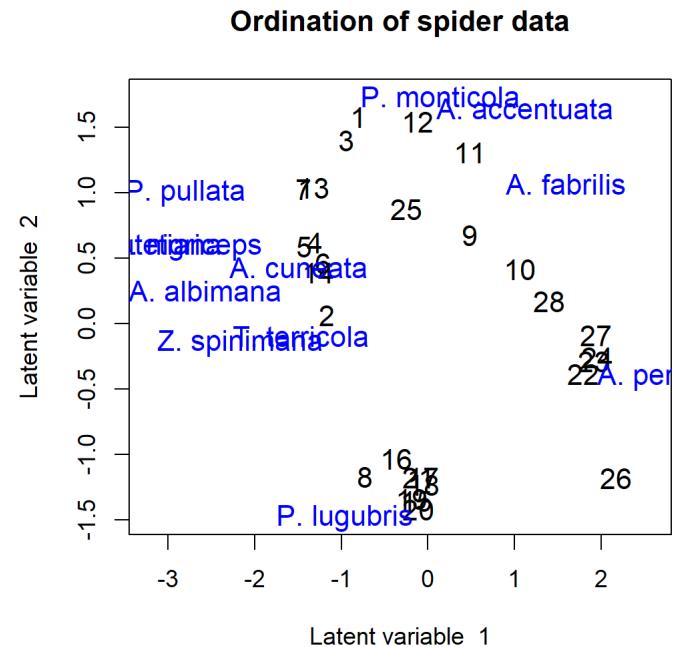
Generalized Linear Latent Variable Models

- GLLVM for short
- Add factor analytic structure to Σ
- Ordination = dimension reduction
- $\epsilon_{ij} = \mathbf{u}_i^\top \boldsymbol{\theta}_j$
 - i.e. $\epsilon_i \sim \mathcal{N}(0, \boldsymbol{\theta}_j \boldsymbol{\theta}_j^\top)$
- 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
 - d latent variables treated as random-effect
 - Produces ordination
 - "site scores" : \boldsymbol{u}_i
 - "species scores" or "loadings": $\boldsymbol{\theta}_j$
 - No varimax



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

Compared to (unconstrained) classical ordination

- Principal Component Analysis is (essentially) a GLLVM with normal distribution
- Correspondence analysis
 - approximate GLLVM (with Poisson/binomial distribution) ter Braak 1985
 - With row-effect Hui et al. 2015
- Non-metric multidimensional scaling?
 - Relation is complicated (but produces similar results in practice)

- **GLLVMs are more flexible**
- **The statistical model can be extended**



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