Introduction to advanced community analysis

Bert van der Veen

Department of Mathematical Sciences, NTNU

Philosophy

- Adjust the model, not the data
- If you're not sure how to adjust the model, you just need to figure out
- Unlike "classically" where the data is adjusted to make the method fit

Classical approach

Classically:

1) Decide on a distance
2) Apply a multivariate analysis
3) Make a plot
4) Do a hypothesis test

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

Welcome! 😄

Intro

See github for all material

Sessions from 14:00 to 20:00 (Tuesday to Friday). Sessions will consist of a mix of lectures, in-class discussion, and practical exercises / case studies over Slack and Zoom.

- Tuesday: Properties of community data, VGLMs and VGLMMs
- Wednesday: Model checking, hierarchical environmental responses, JSDM
- Thursday: Model-based ordination, Ordination with covariates, Unimodal responses
- Friday: Other packages, Beyond `gllvm`, article reanalysis, analysis of own data

How we will do it

Lectures of about 45-60 minutes
Practicals of about 45-60 minutes: datasets and R

- Practical "tasks" that get more complex
- Short, live, demonstrations

Friday

1. Other R-packages for GLLVMs/JSDMs
2. A brief look at Hierarchical ordination
3. Article reanalysis
4. Own data analysis/questions/suggestions

How does that sound?

What I hope you take away

1. The `gllvm` R-package is great!
2. Performing multivariate analysis well is hard work
3. Model-based ordination methods do many things better (data properties, diagnostics)
4. One Method To Rule Them All

Logistics

All material on github

Please make sure you've downloaded data and updated R/packages

R-packages

- gllvm
- glmmTMB
- mvabund
- HMSC
- sjSDM
- ecopCopula
- vegan
- boral

and some spatial packages

- terra
- sf
- maptiles

Some resources on classical ordination

- David Zeneley's website
- Michael Palmer's website
- Numerical ecology
- Numerical ecology with R
- Data analysis in Community and Landscape ecology
- Analysis of ecological communities

Resources on model-based ordination

- Some of my other workshop repositories
- gllvm vignette website
- Oxford libraries article
- Warton 2022
- Fahrmeir and Tutz 2001
- Ovaskainen and Abrego
- Bartholomew et al. 2011
- Skrondal and Rabe-Hesketh 2004
- Zuur and Ieno (2025)

Some recommended reading

- Halvorsen (2012)
- Wang et al. (2012)
- Warton et al. (2012)
- Clark et al. (2014)
- Warton et al. (2015)
- Warton et al. (2015)
- Hui et al. (2015)
- Pollock et al. (2015)
- ter Braak and Smilauer (2015)
- Hui et al. (2017)
- Niku et al. (2017)
- Ovaskainen et al. (2017)
- Roberts (2017)
- Warton et al. (2017)
- Niku et al. (2019)
- Niku et al. (2019)
- Roberts (2019)
- Paul (2020)
- Zurell et al. (2020)
- van der Veen et al. (2021)
- Blanchet et al. (2022)
- van der Veen (2022)
- van der Veen et al. (2023)
- Korhonen et al. (2024)
- [van der Veen et al. (2024)][https://arxiv.org/abs/2408.05333]
- Tang et al. (2025)

Resources that cover all kinds of ordination methods

(none)

Motivation

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

Figure 1: Warton et al. 2015

Motivation

- We need formal, probabilistic, models for community ecology
- That incorporate data properties, rather than transforming our way out of them
- It makes for better/clearer inference
- For nicer for teaching
- Access to tools for testing assumptions
- Overall more flexibility

Why GLLVMs?

GLLVMs are a formal statistical, fully probabilistic, tool for multivariate analysis.

1) To step up your multivariate analysis
2) Maybe you want to incorporate random effects
3) Negates the need for distances

Multivariate analysis

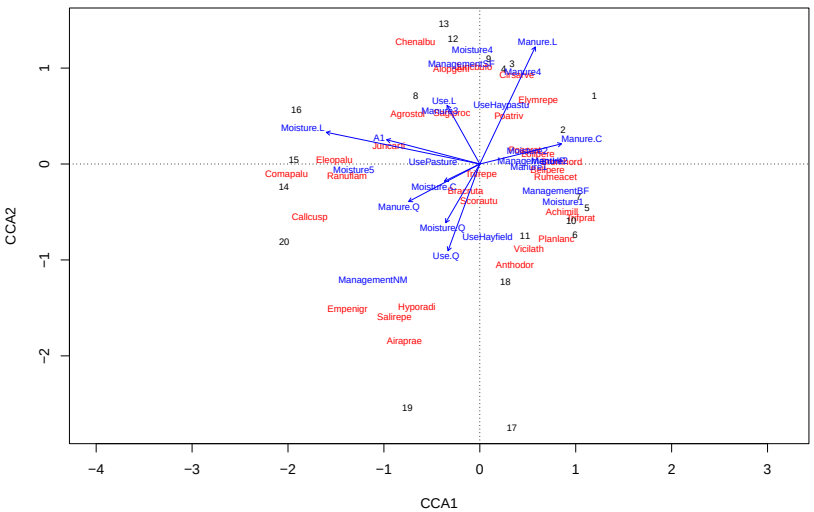


Figure 2: CCA of dune data

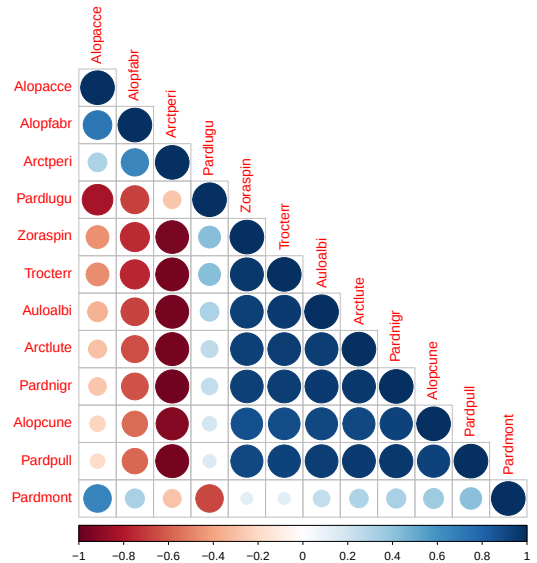


Figure 3: JSDM of spider data

Use of ordination

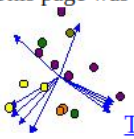
- One of the most used methods of multivariate analysis
  - but definitely not the only one
- Arranges species and samples in a low-dimensional space
  - or any column and row quantity really
- Summarizes data
- Find underlying structures (gradients/LVs)

A brief history of ordination

Milestones in Ordination: a timeline

- 1901 - Pearson develops PCA as a regression technique
- 1927 - Spearman applies factor analysis to psychology
- 1930 - Ramensky uses an informal ordination technique and the term "Ordnung" in ecology
- 1933 - Hotelling develops PCA for understanding the correlation matrix
- 1950 - Curtis and McIntosh employ the "continuum index" approach
- 1952 - Williams uses Correspondence Analysis
- 1954 - Goodall uses the term "ordination" for PCA
- 1957 - Bray-Curtis (Polar) ordination
- 1964 - Kruskal develops NMDS
- 1970's - Whittaker develops theoretical foundations of gradient analysis
- 1973 - Hill revives Correspondence Analysis
- 1976 - Canonical Correlation introduced to ecology
- 1977 - Fasham, Prentice use NMDS
- 1979 - DCA introduced by Hill and Gauch
- 1982 - Gauch's "Multivariate Analysis in Community Ecology"
- 1986 - CCA introduced by ter Braak
- 1986 - Fuzzy set ordination introduced by Roberts
- 1988 - ter Braak and Prentice's "Theory of Gradient Analysis"

This page was created and is maintained by Michael Palmer.



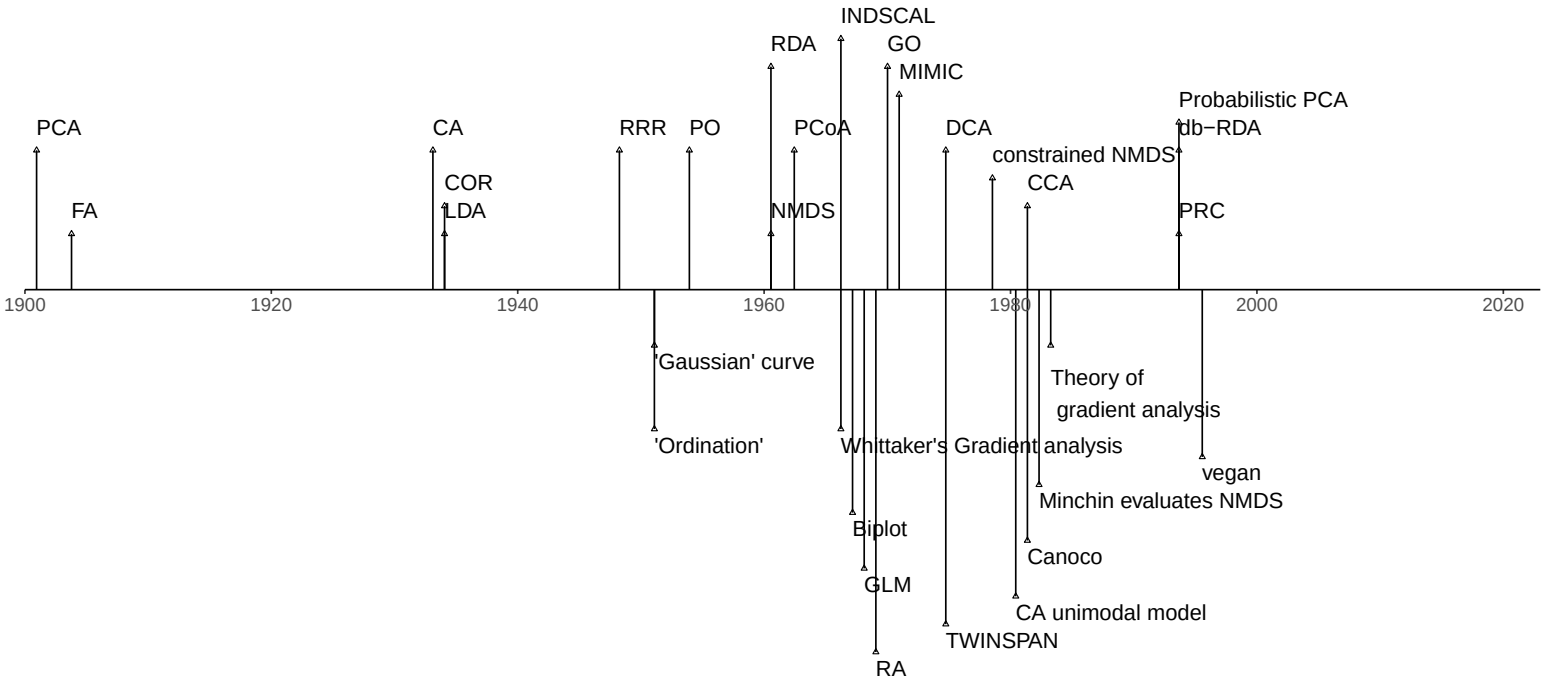
To the ordination web page

Historical Perspective

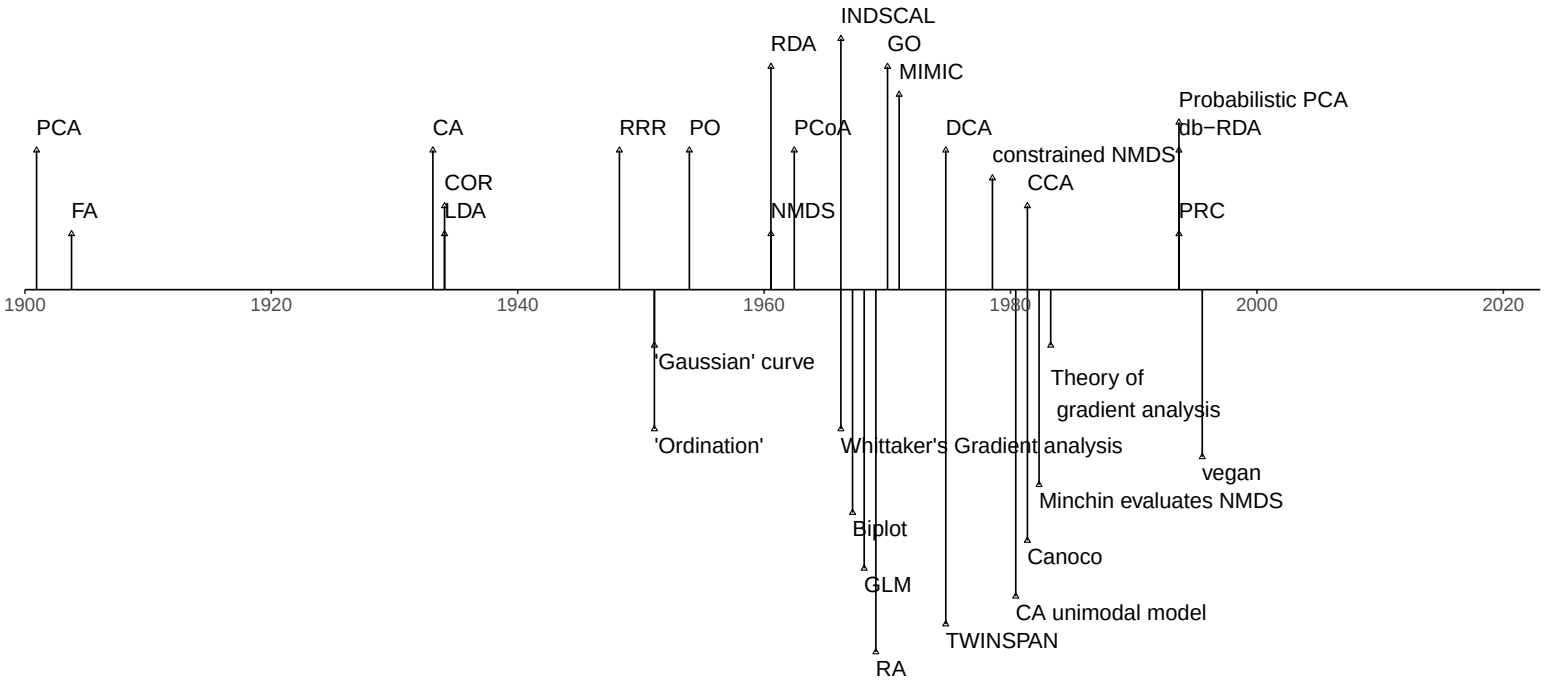
1901	-	Pearson develops PCA as a regression technique.
1927	-	Spearman applies factor analysis to psychology.
1930	-	Ramensky uses an informal ordination technique and introduces the term 'ordnung' into ecology.
1954	-	D.W. Goodall introduces PCA into ecology and proposes the term 'ordination'.
1970	-	R.H. Whittaker develops theoretical foundations of gradient analysis, especially unimodal species responses and turnover along environmental gradients.
1971	-	K.R. Gabriel develops biplot graphical display.
1973	-	M.O. Hill re-invents correspondence analysis and introduces CA (as 'reciprocal averaging') into ecology.
1986	-	Cajo ter Braak invents canonical correspondence analysis (CCA) and released CANOCO software.
1988	-	Cajo ter Braak and Colin Prentice's "A theory of gradient analysis" (Advances in Ecological Research 18; 271-317) that unifies indirect and direct gradient analysis and highlights the importance of underlying species response models.
1998, 2002	-	Cajo ter Braak and Petr Šmilauer CANOCO 4 & 4.5 software and manual.

- Inspired by Michael Palmer's and John Birk's
- But in need of a little update

The rich history of multivariate analysis

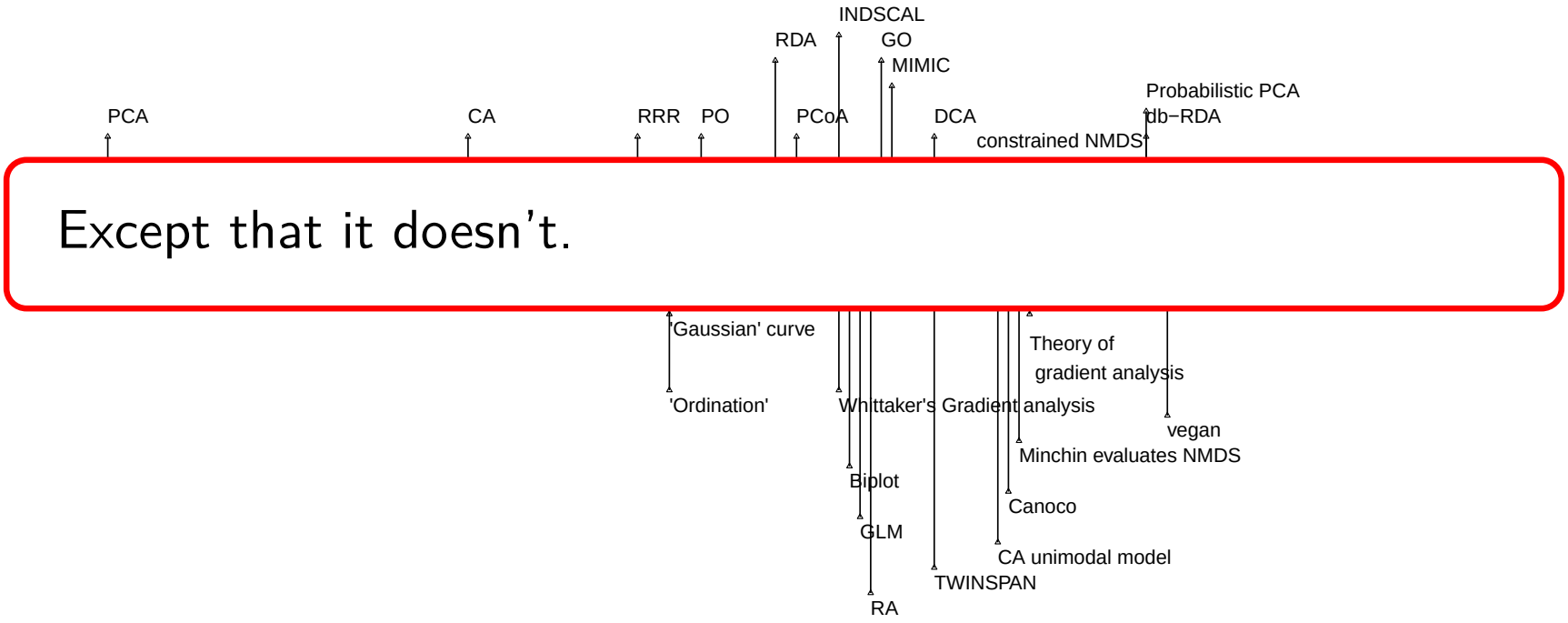


The rich history of multivariate analysis



Seems like everything stops at the turn of the millennium?

The rich history of multivariate analysis



Except that it doesn't.

Seems like everything stops at the turn of the millennium?

Contemporary multivariate analysis

- Vector GLM(M)s
- Joint Species Distribution Models
- Model-based ordination

All three of these can be fitted with the `gllvm` R-package.

Vector Generalised Linear (Mixed effects) Models

We can fit specialised mixed-effects models to community data:

$$g\{\mathbb{E}(y_{ij}|\boldsymbol{\lambda}_i)\} = \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{z}_i^\top \boldsymbol{\lambda}_j, \qquad \boldsymbol{\lambda}_j \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (1)$$

- The time that we were computationally limited is long past
- We can analyse multivariate data like any other
- Leverage information across all species in our data to learn about community processes

Joint Species Distribution Modeling

- First suggested by Ovaskainen et al. 2010
- Named by Pollock et al. 2014

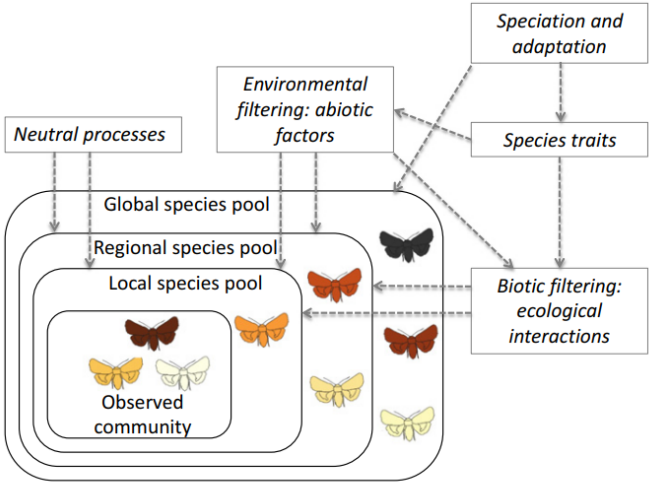


Figure 1 A conceptual diagram of the assembly processes influencing ecological communities at different spatiotemporal scales. The composition and dynamics of local, regional and global communities are influenced by the combined effects of environmental filters, biotic interactions and neutral processes. The responses of the species to these factors depend on their traits, which are ultimately shaped by evolutionary history and therefore constrained by phylogenetic relationships.

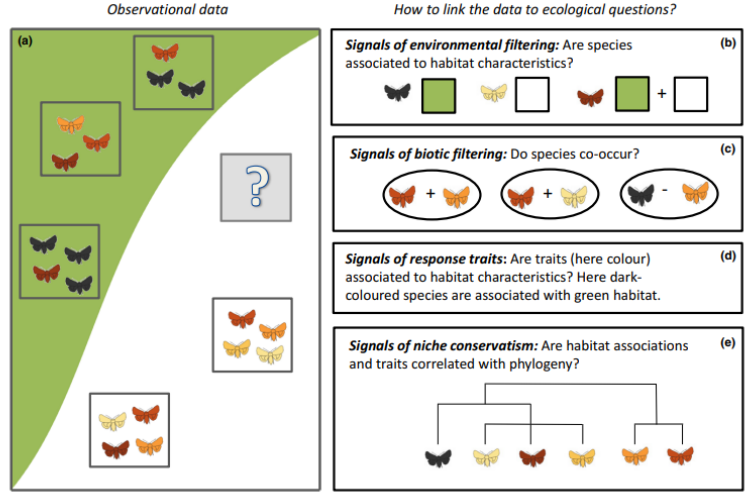


Figure 2 A conceptual illustration of some key questions in community ecology. The green and white colours represent differences in the environmental

Figure 4: Figures from Ovaskainen et al. 2017

Joint Species Distribution Modeling (2)

- A niche process generates data
- Distributions are driven by 1) environment and 2) interactions
- Statistically, interaction = correlation of species
- So we model that:

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

Can be fit using standard mixed-effects modeling software:

```
glmer(abundance~species+x:species+(0+species|sites))
```

Problem: Number of parameters grows quadratically

“Fun” ecological inference

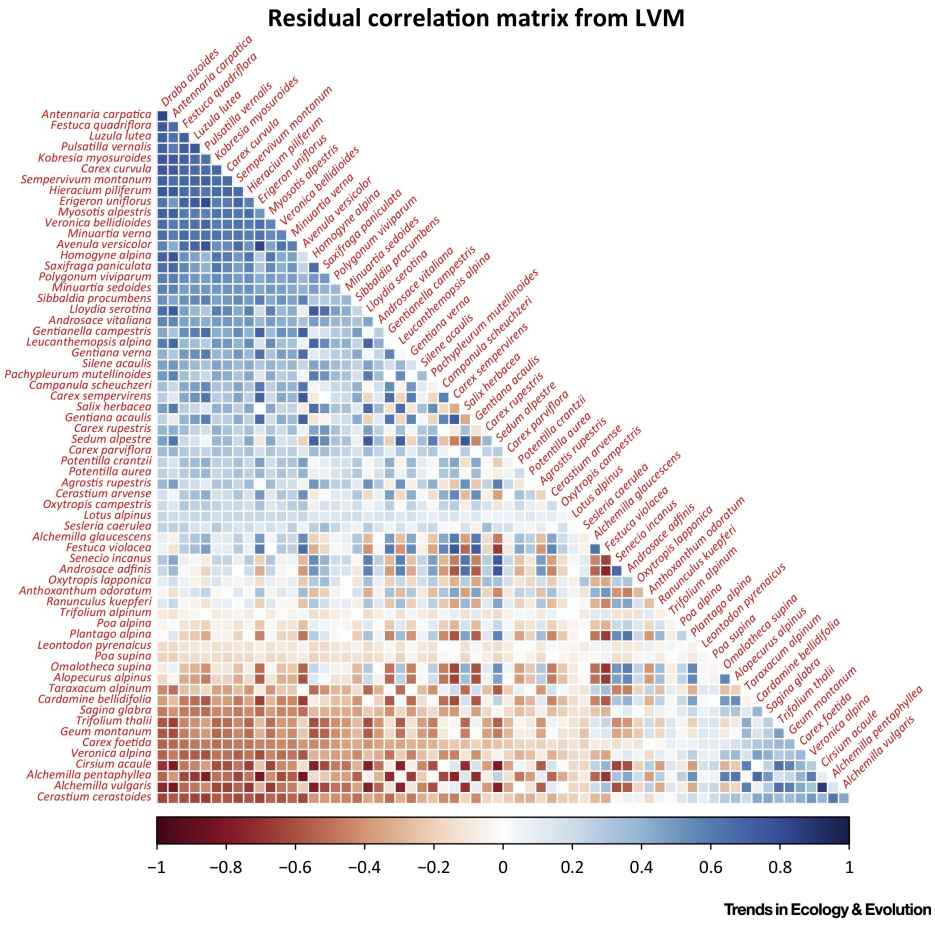


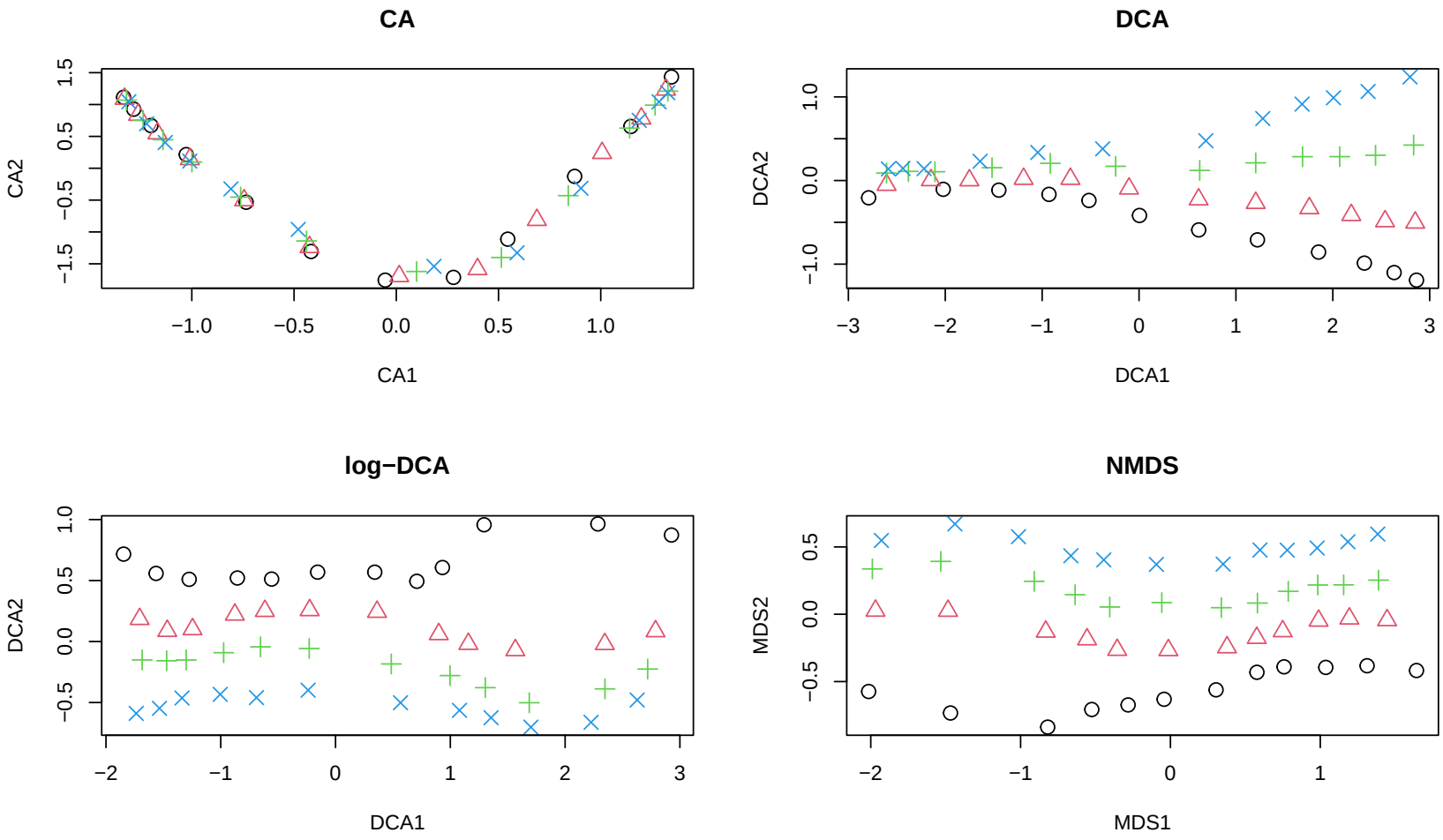
Figure 5: Warton et al. 2015 Alpine plants

Model-based ordination

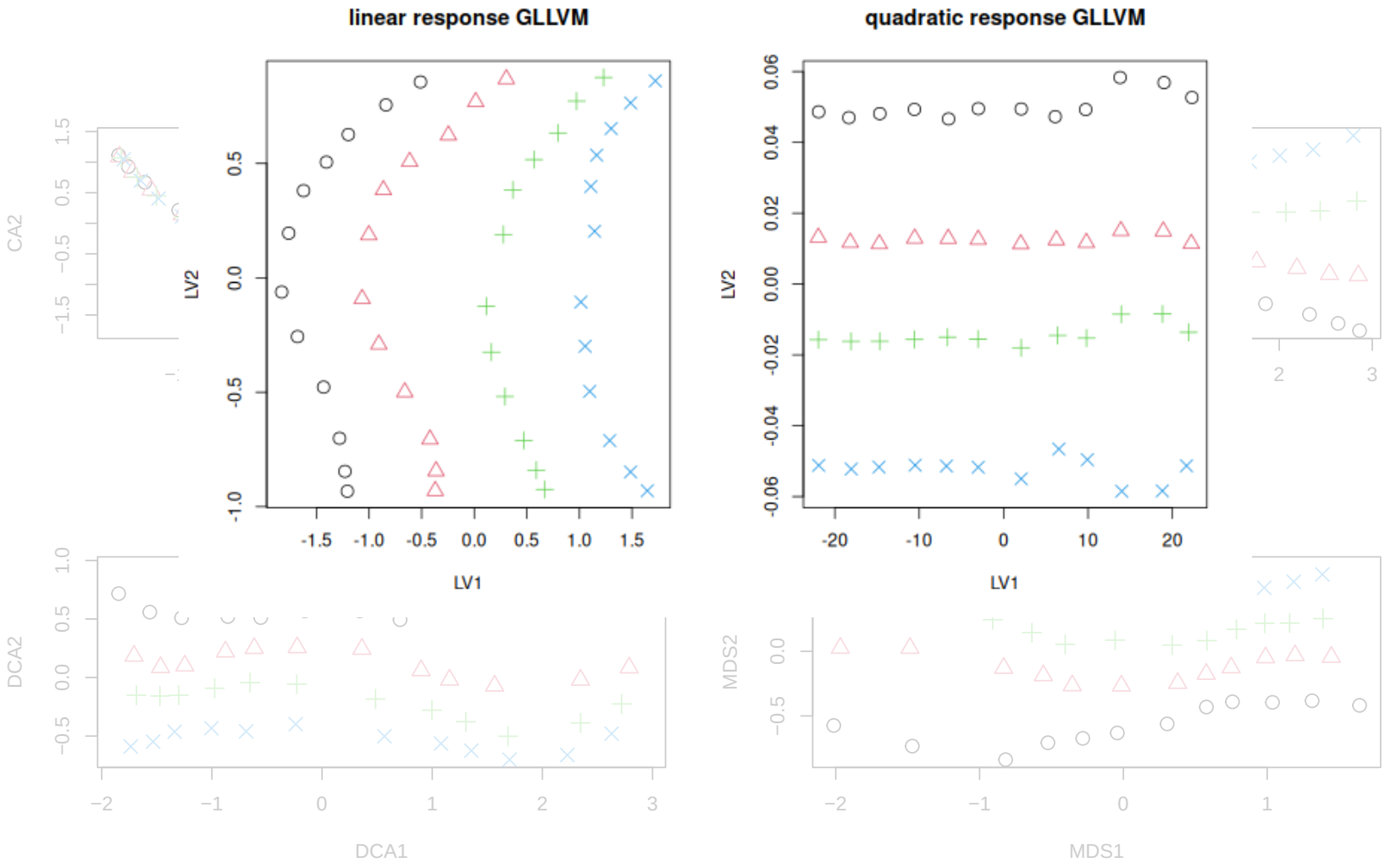
- Generalised Linear Latent Variable Model
- Adds "factor analytic" structure to Σ
- $\epsilon_{ij} = \mathbf{u}_i^\top \boldsymbol{\gamma}_j$
- i.e. $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top)$
- Faster and fewer parameters:
- Number of parameter doesn't grow so fast

$$\boldsymbol{\Sigma} = \begin{bmatrix} \gamma_{11} & 0 & 0 \\ \gamma_{12} & \gamma_{22} & 0 \\ \vdots & \ddots & \vdots \\ \gamma_{1j} & \cdots & \gamma_{dj} \end{bmatrix} \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1j} \\ 0 & \gamma_{22} & \ddots & \vdots \\ 0 & 0 & \cdots & \gamma_{dj} \end{bmatrix} \tag{3}$$

The Minchin dataset



The Minchin dataset



Why model-based

- Flexibility
- Accommodate properties of data and sampling processes
- Now straightforward available and fast
- More advanced ecological inference and prediction

We can do all kinds of fun things

- Residual diagnostics
- Species-specific effects from constrained ordination
- Random-effects
- Flexible models
- Etc.

Challenges

- Complex model
  - products of random-effects
  - identifiability
- No analytical solution (need approximate methods)
- Computationally intensive (need fast fitting)
- Non-convex objective function (need robust optimization)
- Usability needs a lot more improvements

The gllvm R-package

Many people have put a lot of work into development of the methods presented here

The gllvm R-package

- Fast
- Easy to use
- Many different model structures
- (Un)constrained ordination with random-effects
- Tools for ordination (biplot) and regression (model selection, statistical uncertainties)
- (very) Active support via github (Jenni Niku, me ☺️)

Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K.C. Hui, Pekka Korhonen, Sara Taskinen, Bert van der Veen and David I. Warton (2025). gllvm: Generalized Linear Latent Variable Models. R package version 2.0.2

gllvm

Received: 7 May 2019 | Accepted: 5 September 2019

DOI: 10.1111/2041-210X.13303

APPLICATION

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

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

Jenni Niku[1] | Francis K. C. Hui[2] | Sara Taskinen[1] | David I. Warton[3]

- Originally published in 2019 by Niku et al.
- For JSDM, unconstrained, and residual ordination
- Since then it has been considerably extended
- Models are fitted in C++ with Template Model Builder

(Kristensen et al. 2015)

Likelihood approximation

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{ \int \prod_{j=1}^{m} f\Big(y_{ij} | \mathbf{z}_i, \Theta \Big) h\Big(\mathbf{z}_i \Big) d\mathbf{z}_i \right\}, \qquad (4)$$

The package has three methods for approaching the estimation:

- Laplace approximation (LA)
- Variational approximation (VA)
- Extended variational approximation (EVA)

Sometimes we want or need to switch between these

Main function: `gllvm(.)`

gllvm {gllvm} R Documentation

Generalized Linear Latent Variable Models

Description

Fits generalized linear latent variable model for multivariate data. The model can be fitted using Laplace approximation method or variational approximation method.

This has many arguments

- `y` (community data)
- `X` (environment)
- `TR` (traits)
- `data`
- `formula`
- `family`
- `num.lv` (unconstrained ord.)
- `num.lv.c` (concurrent ord.)
- `num.RR` (constrained ord.)
- `lv.formula`
- `sd.errors` (can take long)
- `method` (LA, VA, EVA)

gllvm(.) arguments continued

- studyDesign
- dist
- colMat
- colMat.rho.struct
- corWithin
- quadratic
- row.eff
- offset
- randomB
- randomX
- beta0com
- zeta.struc (only applies to ordinal)
- link (only applies to binomial)
- Ntrials (only applies to binomial)
- Power (only applies to tweedie)
- seed (for reproducibility)
- scale.X (for 4th corner)
- return.terms
- gradient.check (convergence check)
- disp.formula (dispersion parameters)
- control
- control.va
- control.start

Distributions in gllvm

Type of data	Distribution	Method	Link
Normal	Gaussian	VA/LA	identity
Counts	Poisson	VA/LA	log
	NB	VA/LA	log
	ZIP	VA/LA	log
	ZINB	VA/LA	log
	binomial	VA/LA	probit
	binomial	LA	logit
Binary	Bernoulli	EVA VA/LA	probit logit
Ordinal	Multinomial	VA	cumulative probit
Biomass	Tweedie	EVA/LA	log
Positive continuous	Gamma	VA/LA	log
	Exponential	VA/LA	log
Percent cover	beta	LA/EVA	probit/logit
with zeros or ones	ordered beta	EVA	probit

Functions in the package

- `gllvm()`
- `logLik()`
- `summary()`
- `confint()`
- `predict()`
- `coefplot()`
- `randomCoefplot()`
- `plot()` and `residuals()`
- `se()` and `vcov()`
- `getLV()`
- `getLoadings()`
- `predictLVs()`
- `phyloPlot`
- `VP()` and `plotVP()`
- `getPredictErr()`
- `getResidualCor()` and `getResidualCov()`
- `getEnvironmentalCor()` and `getEnvironmentalCov()`
- `optima()` and `tolerances()`
- `simulate()`

Model structures

- Covariates outside of ordination: `X` and `formula` ("conditioning" or "covariate-adjusted")
- Unconstrained ordination: `num.lv` and (optional) `quadratic`
- Constrained ordination: `num.RR` and (optional) `lv.formula` or `randomB` or `quadratic`
- Concurrent ordination: `num.lv.c` and (optional) `lv.formula` or `randomB` or `quadratic`
- Fourth-corner LVM: `X` and `TR` and `formula` and (optional) `randomX` or `beta0comm`
- Random species effects: `formula` and `X` and (optional) `beta0comm`
  - Phylogenetic effects: `colMat` and `colMat.rho.struct`
- Random site effects: `row.eff` and (optional) `dist` or `studyDesign`

Some of these can be combined, not all

E.g., no traits with constrained/concurrent ordination that might be the only limitation at present

Insight into the algorithm

Little shiny app here: https://bertvdveen.shinyapps.io/shinygllvm/
Or, e.g., : `gllvm(y, family = "poisson", TMB = FALSE, plot = TRUE)`

Multiple starting values

One "quirk" about such models: they can give different solutions each time.

`gllvm` strategy (think: `lme4` + `metaMDS` combined) :

- `starting.val`: different types of (smartly) generated starting values
- `jitter.var`: add a little noise to starting values
- `optimizer`: changing it can help at times
- `n.init`: run model multiple times and pick best
- `n.init.max`: maximum number of tries before exit

Efficient estimation of generalized linear latent variable models

Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton

Published: May 1, 2019 • https://doi.org/10.1371/journal.pone.0216129

Where to go

Bugs: https://github.com/JenniNiku/gllvm/issues
Questions: https://github.com/JenniNiku/gllvm/discussions
Examples: https://jenniniku.github.io/gllvm/

Citing: `citation("gllvm")`

Outline

- ~~Introduction~~
- Brainstorm community data
- Vector GLMs
- Multispecies random effects

15-20 minute break 15:30-15:45

45 minute break 17:45-18:30


LETS GET STARTED
makeameme.org