Generalized Linear Latent Variable Models

Bert van der Veen

Department of Mathematical Sciences, NTNU

Outline

- GLLVMs background
- Difference to classical methods
- `gllvm` R-package

Questions so far?

Model-based thinking for community ecology

Warton et al. 2015

- Classical methods ignore properties of the ecological process
- They are purely algorithmic (and severely outdated)
- There are few links to theory

Model-based analysis

- Accounts for properties of the ecological process
- Is flexible
- Has clear connections with testable hypotheses
- Are computationally intensive 🥴
- Provides diagnostic tools

Some other (approximate) latent variable models

- Principal Component Analysis (Pearson 1901)
- Factor analysis (Spearman 1904)
- Redundancy Analysis (Rao 1964)
- Principal Coordinate Analysis (Gower 1966)
- Correspondence Analysis (Benzecri 1973)
- Detrended Correspondence Analysis (Hill and Gauch 1980)
- Canonical Correspondence Analysis (ter Braak 1986)
- Non-metric Multidimensional Scaling (Kruskal 1964)
- t-SNE (van der Maaten 2008)
- UMAP (McInnes and Healy 2018)

Some other (approximate) latent variable models

- Principal Component Analysis (Pearson 1901)
- Factor analysis (Spearman 1904)
- Redundancy Analysis (Rao 1964)
- Principal Coordinate Analysis (Gower 1966)
- Correspondence Analysis (Benzecri 1973)
- Detrended Correspondence Analysis (Hill and Gauch 1980)
- Canonical Correspondence Analysis (ter Braak 1986)
- Non-metric Multidimensional Scaling (Kruskal 1964)
- t-SNE (van der Maaten 2008)
- UMAP (McInnes and Healy 2018)

GLLVM replaces all of these, and does it better.

YOU GET AN UPGRADE! AND YOU GET AN UPGRADE!
EVERYBODY GETS AN UPGRADE!
makeameme.org

Contemporary multivariate methods

- Multivariate GLMs (Wang et al. 2012)
- Vector GAMs (Yee et al. 1996,2003,2010,2015)
- Joint SDMs (Pollock et al. 2014 , Clark et al. 2014)
- Row-column interaction models (Hawinkel et al. 2019)
- Generalized Linear Latent Variable Models (Skrondal and Rabe-Hesketh 2004, Hui et al. 2015, Warton et al. 2015, Ovaskainen et al. 2017)

Contemporary multivariate methods

- Multivariate GLMs (Wang et al. 2012)
- Vector GAMs (Yee et al. 1996,2003,2010,2015)
- Joint SDMs (Pollock et al. 2014 , Clark et al. 2014)
- Row-column interaction models (Hawinkel et al. 2019)
- Generalized Linear Latent Variable Models (Skrondal and Rabe-Hesketh 2004, Hui et al. 2015, Warton et al. 2015, Ovaskainen et al. 2017)

and many more (e.g., clustering)

Goals of ordination

- To order sites and species along gradients
- To reduce dimensions (because there are many)
  - for unconstrained ordination: species
  - for constrained ordination: species and predictors
- generally: to facilitate inference on (sparse) multivariate data

Generalized Linear Latent Variable Model (GLLVM)

- A framework for model-based multivariate analysis
- That does dimension reduction (i.e., ordination)
- There is no distance measure
- You do need to specify:
    1. A distribution
    2. A link function
    3. The model its structure
- Latent variables are found by *best fit* (and the first is not always the most important)
- We (can) treat them as random effect when appropriate

Response distribution

$$y_{ij} \sim f\left\{g^{-1}\left(\eta_{ij}\right), \phi_j\right\} \tag{1}$$

Response distribution

$$y_{ij} \sim f\left\{g^{-1}\left(\eta_{ij}\right), \phi_j\right\} \tag{1}$$

1. Community data

Response distribution

$$y_{ij} \sim f\left\{ g^{-1}\left(\eta_{ij}\right), \phi_j \right\} \tag{1}$$

2. Response distribution

Response distribution

$$y_{ij} \sim f\left\{g^{-1}\left(\eta_{ij}\right), \phi_j\right\} \tag{1}$$

3. (inverse) Link function

Response distribution

$$\boxed{y_{ij}} \sim \boxed{f}\left\{\boxed{g^{-1}}\left(\boxed{\eta_{ij}}\right), \boxed{\phi_j}\right\} \tag{1}$$

4. Linear predictor ("the model") $\longrightarrow \eta_{ij}$

Response distribution

$$y_{ij} \sim f\left\{g^{-1}\left(\eta_{ij}\right), \phi_j\right\} \tag{1}$$

5. Dispersion parameter

Latent variable distribution

$$\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

- In unconstrained ordination, we assume LVs to be multivariate standard normal in distribution
  - They absorb all residual variation, so we assume $\mathbf{z}_i = \boldsymbol{\epsilon}_i$
- This is similar to classical ordination methods (orthogonality)
- But in GLLVMs they are random effects (more appropriate)
- In constrained ordination this is different (see)

GLLVM Likelihood

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{ \int \prod_{j=1}^{m} f\left(y_{ij} \mid \mathbf{z}_i, \Theta \right) h\left(\mathbf{z}_i \right) d\,\mathbf{z}_i \right\}, \qquad (3)$$

Take away: ugly integral sign that we cannot analytically solve

The model

$$\eta_{ij} = \beta_{0j} + \ldots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \qquad (4)$$

The model

$$\eta_{ij} = \beta_{0j} + \ldots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \tag{4}$$

1. Linear predictor

The model

$$\eta_{ij} = \beta_{0j} + \dots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \tag{4}$$

2. Species intercept

The model

$$\eta_{ij} = \beta_{0j} + \ldots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \tag{4}$$

3. Other stuff

The model

$$\eta_{ij} = \beta_{0j} + \ldots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \tag{4}$$

4. Ordination

The model

$$\eta_{ij} = \beta_{0j} + \ldots + \boxed{\mathbf{z}_i^\top \boldsymbol{\gamma}_j} \tag{4}$$

5. Site scores (ordination axis/latent variables)
6. Species loadings

GLLVM advantages

Combining regression with the best of ordination

GLLVM advantages

Combining regression with the best of ordination

- Model selection
- Confidence intervals
- Diagnostic tools: residuals, derivatives, information criteria
- Adjustable model structure

GLLVM advantages

Combining regression with the best of ordination

- Model selection
- Confidence intervals
- Diagnostic tools: residuals, derivatives, information criteria
- Adjustable model structure
- Ordination for all kinds of datatypes in one framework
  - Count data, cover data, binary data, continuous data
  - Poisson, zero-inflated, negative binomial, binomial, ordinal, beta, gamma, etc.

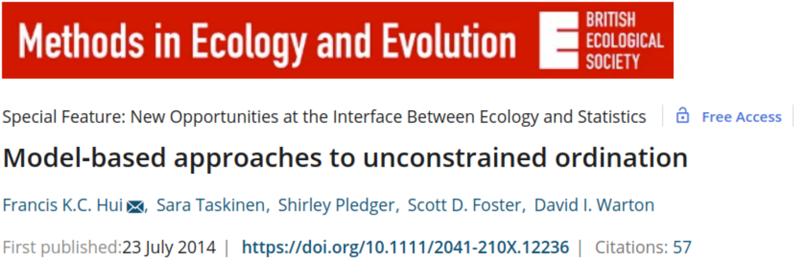
GLLVM advantages

Combining regression with the best of ordination

- Model selection
- Confidence intervals
- Diagnostic tools: residuals, derivatives, information criteria
- Adjustable model structure
- Ordination for all kinds of datatypes in one framework
  - Count data, cover data, binary data, continuous data
  - Poisson, zero-inflated, negative binomial, binomial, ordinal, beta, gamma, etc.
- Many tools from ordination too
  - Ordination plots
  - Gradient length
  - Variance partitioning (to some degree)

Model-based ordination

Suggested to use Generalized Linear Latent Variable Models for unconstrained ordination



Methods in Ecology and Evolution

BRITISH ECOLOGICAL SOCIETY

Special Feature: New Opportunities at the Interface Between Ecology and Statistics | Free Access

Model-based approaches to unconstrained ordination

Francis K.C. Hui, Sara Taskinen, Shirley Pledger, Scott D. Foster, David I. Warton

First published:23 July 2014 | https://doi.org/10.1111/2041-210X.12236 | Citations: 57

Building on a long history of using latent variables in ecology (e.g., ter Braak 1985)

GLLVMs vs. classical ordination: main differences

1) GLLVMs have a real model
2) GLLVMs incorporate distributions, not distances
3) There are no eigenvalues (but there is variance)
4) Number of dimensions are set a-priori as in NMDS
5) Latent variables are found by "best fit"
6) You might not get the same solution every time
7) Forget about permutation testing
8) We do not care much about rotation
9) $\longrightarrow$

Fitting unconstrained ordination

In R e.g.,

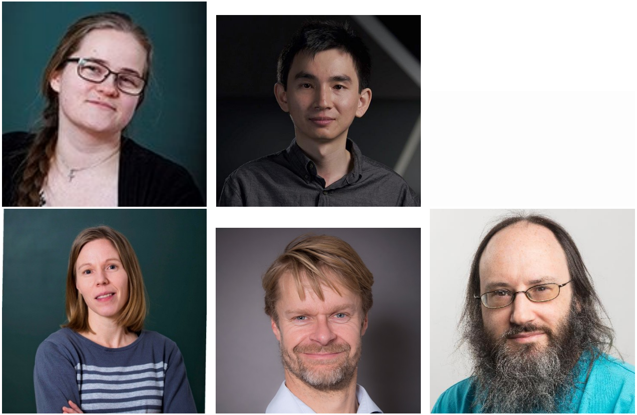
- `HMSC` - Bayesian (slow) with a lot of tools
- `Boral` - Bayesian (slow) easy to use
- `ecoCopula` - even (super) faster (but marginal interpretation)
- `glmmTMB` - very easy to use, can include many other random-effects
- `gllvm` - fast and easy to use

Fitting ordination with predictors

In R e.g.,

- VGAM - robust algorithm, quick, many distributions
- RCIM - flexible response models
- gllvm - straightforward interface, also random effects

The `gllvm` R-package



Jenni Niku (JYU), Francis Hui (ANU), Pekka Korhonen (JYU), Sara Taskinen (JYU), David Warton (UNSW), Bob O'Hara (NTNU)

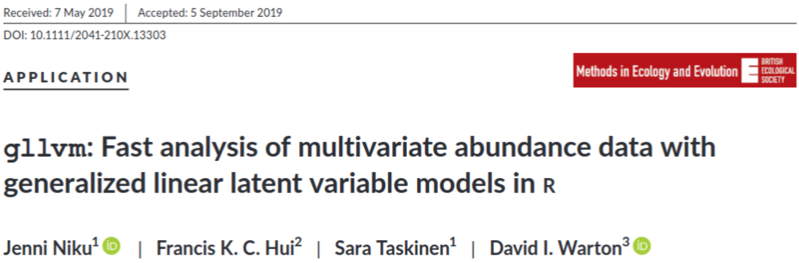
Many people have put a lot of work into development of the methods presented

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 (2023). gllvm: Generalized Linear Latent Variable Models. R package version 1.4.3.

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

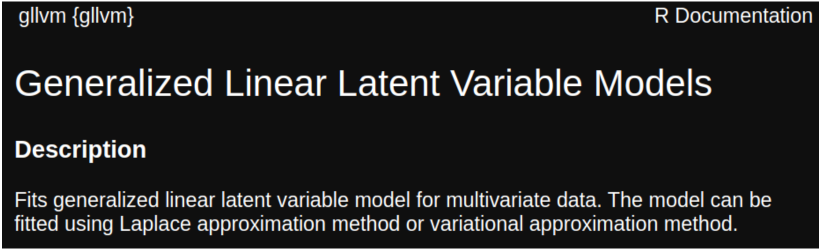
$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{ \int \prod_{j=1}^{m} f\left(y_{ij} | \mathbf{z}_i, \Theta \right) h\left(\mathbf{z}_i \right) d\mathbf{z}_i \right\}, \qquad (5)$$

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` (often takes long)
- `method` (LA, VA, EVA)

gllvm(.) arguments continued

- studyDesign
- dist
- colMat (new)
- colMat.rho.struct (new)
- corWithin
- quadratic (unimodal response model)
- row.eff
- offset
- randomB (for constrained/concurrent)
- randomX (for 4th corner)
- 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()` (new)
- `predictLVs()`
- `getPredictErr()`
- `getResidualCor()` and `getResidualCov()`
- `getEnvironmentalCor()` (new) and `getEnvironmentalCov()` (new)
- `optima()` and `tolerances()`
- `simulate()`

new: in development version, not on CRAN yet

Defaults

- without `lv.formula`: 2 unconstrained LVs
- with `lv.formula`: 0 unconstrained LVs
- `method = "VA"`
- `sd.errors = TRUE`
- `optimizer = "optim"` with `optim.method  = "BFGS"`
- `row.eff = FALSE`
- `Power = 1.1`
- `seed = NULL`

We will look at how/why to change these these in the workshop.

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

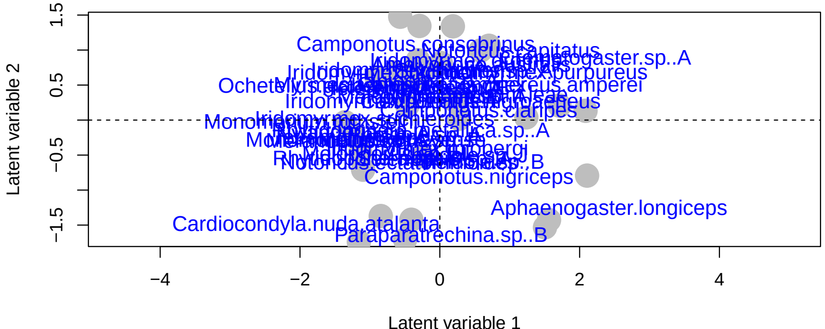
ordiplot(.)

Insight into the algorithm

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

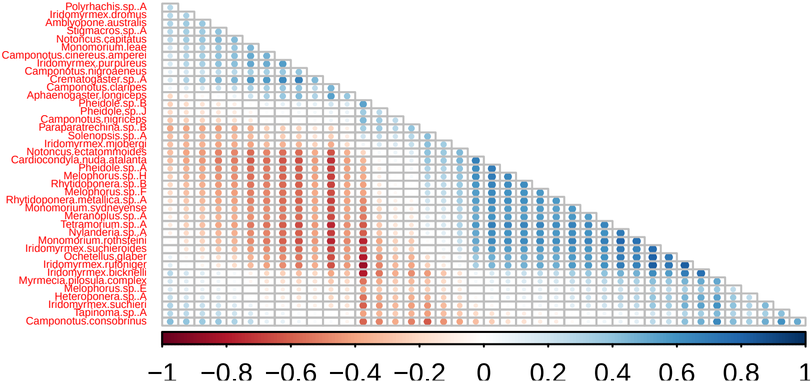
Unconstrained ordination/JSDM

```
library(gllvm)
data("antTraits"); Y <- antTraits$abund
uord <- gllvm(Y, family = "negative.binomial")
ordiplot(uord, biplot = TRUE, main = NA, symbols = TRUE,
         s.colors = "grey", cex.spp = 1.2, s.cex = 3, pch = 16)
```



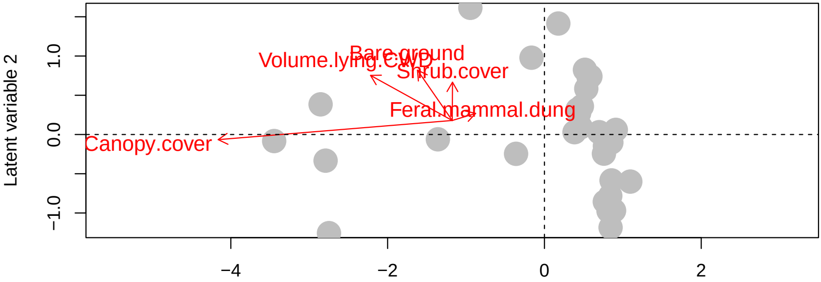
Unconstrained ordination/JSDM

```
library(corrplot)
corrplot(getResidualCor(uord), type = "lower", order = "AOE", tl.pos = "l"
         tl.cex = 0.3, diag = FALSE, win.asp = 0.5)
```



Constrained ordination

```
X = scale(antTraits$env)
coord <- gllvm(Y, X = X, num.RR = 2, family = "negative.binomial",
               randomB="LV")
ordiplot(coord, main = NA, symbols = TRUE, s.colors = "grey",
         cex.spp = 1.2, s.cex = 3, pch = 16, arrow.ci = FALSE,
         cex.env = 1.2, arrow.scale = 2, lab.dist = 0.3)
```



Multiple starting values

One “quirk” about such models: they can give different solutions each time. `gllvm` strategy:

- `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

Signs that your model has not converged

- Weird ordination plot (somtimes)
- Large gradient values (use gradient.check cautiously)
- Singular hessian
- Parameters on boundary
- Very large species loadings (usually combined with small sigma.lv)
- Small site scores
- Probably more

To-do

Many extensions are still possible, and in the pipeline

- emmeans support
- Fitting robustness
- Traits in ordination
- Spatial/temporal LVs
- Variance partitioning
- Zero-inflated modeling
- Mixed response types

But, we are few and only human

Where to go

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

To conclude

- In active development
- Parallelisation
- Suggestions welcome
- Let's dive in