

On fast maximum likelihood -based estimation of joint species distribution models

Sara Taskinen

University of Jyväskylä, Finland



Joint work with

Jenni Niku (JyU), Pekka Korhonen (JyU), Francis Hui (ANU),
David Warton (UNSW), John Ormerod (USyd), Bert van der Veen (NTNU)

Outline

Motivation

Joint models as extensions of GLMs

Model fitting by maximum likelihood

R package `gllvm`

Example

Summary and future work

Motivation

- ▶ Joint species distribution models have gained considerable popularity in recent years in many fields of applied science.
- ▶ A prime example is modeling multivariate abundance data in community ecology, where **model-based approaches** allow us to specify a joint statistical model for abundance across many taxa¹.
- ▶ One way to formalize the models is to use **generalized linear latent variable modeling (GLLVM)** framework.
- ▶ One of the main features of GLLVMs is their capacity to handle a variety of responses types, such as (overdispersed) counts, binomial and (semi-)continuous responses, and proportions data.
- ▶ Thanks to recent advances in computational methods, we have nowadays lots of computationally scalable tools for fast and efficient fitting of GLLVMs.

¹Warton et al. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology & Evolution*, **30**, 766–779

Multivariate generalized linear models (GLM)

- ▶ The models we use are extensions of **multivariate generalized linear models** that can be used to model impacts of q environmental (site-specific) covariates on abundances, y_{ij} , $i = 1, \dots, n$, $j = 1, \dots, m$.
- ▶ Let $g(\cdot)$ be a known link function and $\mathbf{x}_i = (x_{i1}, \dots, x_{iq})'$ be a vector of environmental covariates. In multivariate GLM (stacked SDM) the mean response, denoted by $\mu_{ij} = \mathbb{E}[y_{ij}|\mathbf{x}_i]$ is assumed to be

$$g(\mu_{ij}) = \beta_{0j} + \mathbf{x}_i' \boldsymbol{\beta}_j,$$

where β_{0j} and $\boldsymbol{\beta}_j$ are (fixed) species-specific intercept and environmental effect corresponding to j th species.

- ▶ See R package `mvabund`² for tools for model fitting, inference, visualization, etc.

²Wang et al. (2012). `mvabund` – an R package for model-based analysis of multivariate abundance data. *MEE*, 3, 471-474.

Generalized linear mixed models (GLMM)

- ▶ A **joint model** for abundance requires the inclusion of random effects to capture the correlation in abundance across taxa. One way to incorporate correlation is to introduce it directly via a multivariate random effect applied to each sample.
- ▶ In **GLMM** the mean response $\mu_{ij} = \mathbb{E}[y_{ij}|\mathbf{u}_i, \mathbf{x}_i]$ is assumed to be

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \mathbf{x}_i' \boldsymbol{\beta}_j + \mathbf{u}_{ij},$$

where α_i is (optional) sample-specific intercept (fixed or random), β_{0j} and $\boldsymbol{\beta}_j$ are as before, and $\mathbf{u}_i = (u_{i1}, \dots, u_{im})' \sim \mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma})$.

- ▶ Note that when the variance-covariance matrix $\boldsymbol{\Sigma}$ controlling the correlation between taxa is assumed to be unstructured, the model fitting is problematic when $n \ll m$.

Generalized linear latent variable models (GLLVM)

- ▶ A flexible way to incorporate correlation is to regress the mean response μ_{ij} against a vector of $p \ll m$ unknown latent variables, $\mathbf{u}_i = (u_{i1}, \dots, u_{ip})'$, along with covariates.
- ▶ This forms a multivariate generalized linear latent variable model³

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \mathbf{x}_i' \boldsymbol{\beta}_j + \mathbf{u}_i' \boldsymbol{\lambda}_j,$$

where $\mathbf{u}_i \sim \mathcal{N}_p(\mathbf{0}, \mathcal{I})$ and $\boldsymbol{\lambda}_j = (\lambda_{j1}, \dots, \lambda_{jp})'$ denotes a set of species-specific loadings which quantify the relationship between the mean response and the latent variables.

³Moustaki and Knott (2000). Generalized latent trait models. *Psychometrika*, **65**, 391–411.

Fourth-corner GLLVM

- ▶ If r trait covariates $\mathbf{t}_j = (t_{j1}, \dots, t_{jr})'$ are also recorded, we can use them to explain interspecific variation in environmental response.
- ▶ This leads to an extension of the so-called **fourth-corner model**⁴.
- ▶ The **fourth-corner GLLVM** then has a mean model

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \mathbf{x}_i' \boldsymbol{\beta}_e + (\mathbf{t}_j \otimes \mathbf{x}_i)' \boldsymbol{\beta}_l + \mathbf{u}_i' \boldsymbol{\lambda}_j,$$

where $\boldsymbol{\beta}_e$ is a vector of main effects for environmental covariates, and $\boldsymbol{\beta}_l$ is the fourth-corner coefficient.

⁴Jamil and ter Braak (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, **1**, e95.

Generalized linear latent variable models (GLLVM)

- ▶ The term $\mathbf{u}_i' \boldsymbol{\lambda}_j$ accounts for any residual correlation not accounted for by the covariates. Latent variables also represent missing predictors.
- ▶ On the linear predictor scale, the $m \times m$ covariance matrix is $\boldsymbol{\Sigma} = \mathbf{A}\mathbf{A}'$, that is, GLLVMs are **reduced-rank** versions of GLMMs with general covariance structure.
- ▶ The number of latent variables (p) controls model complexity. One can use model selection tools (AIC, BIC, etc.) to guide the choice.
- ▶ If $p = 2$, latent variables and loadings can be used to build model-based ordination plots and biplots⁵.

⁵Hui et al. (2015). Model-based approaches to unconstrained ordination. *MEE*, **6**, 399–411.

Model fitting by maximum likelihood

- ▶ Collect now all model parameters and latent variables into vectors Ψ and $\mathbf{u} = (\mathbf{u}'_1, \dots, \mathbf{u}'_n)'$, respectively.
- ▶ We estimate the model parameters using the maximum likelihood method, that is, we find such Ψ which maximizes the marginal likelihood

$$\log L(\Psi; \mathbf{u}) = \log \left(\int f(\mathbf{y}|\mathbf{u}, \Psi) f(\mathbf{u}) d\mathbf{u} \right).$$

Benefits of model-based approaches and the use of ML

- ▶ One can explicitly account for key statistical properties of the data (e.g. mean-variance relationship) by choosing proper response distribution,
- ▶ use residual analysis tools for model checking,
- ▶ use model selection tools (AIC, BIC, etc.) to choose the most appropriate model for data at hand,
- ▶ use the standard tools developed for statistical inference (large-sample theory for ML estimates),
- ▶ enables fast model-fitting.

Laplace approximation (LA)

- ▶ As p -dimensional integral in $\log L$ cannot be solved analytically, numerical approximation methods are needed.
- ▶ Computationally most feasible maximum likelihood approaches for fitting GLLVMs are those that approximate the marginal likelihood in a closed form.
- ▶ A classical approach is the Laplace approximation (LA) method.⁶
- ▶ LA is easy to implement and it can handle any response distribution and link function combination.
- ▶ Unfortunately the method performs poorly with highly discrete responses (binary, ordinal, etc.).

⁶Niku et al. (2017a), Generalized linear latent variable models for multivariate count and biomass data in ecology, *JABES*, **22**, 498–522.

Variational approximation (VA)

- ▶ One recent, attractive choice in likelihood-based estimation is the **variational approximation method**⁷.
- ▶ The basic idea of VA is to develop so-called **variational lower bound** to marginal log-likelihood function.
- ▶ Plug-in so-called **variational distributions** $q(\cdot)$ of the latent variables to $\log L$ and apply Jensen's inequality to construct a lower bound,

$$\begin{aligned}\log L(\boldsymbol{\Psi}) &= \log \int \left\{ \frac{f(\mathbf{y}|\mathbf{u}, \boldsymbol{\Psi})f(\mathbf{u})q(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u} \\ &\geq \int q(\mathbf{u}) \log \left\{ \frac{f(\mathbf{y}|\mathbf{u}, \boldsymbol{\Psi})f(\mathbf{u})}{q(\mathbf{u})} \right\} d\mathbf{u} \triangleq \ell_{\text{VA}}(\boldsymbol{\Psi}|q).\end{aligned}$$

⁷Ormerod and Wand (2012). Explaining variational approximations. *JCGS*, **21**, 2–17

Variational approximation (VA)

- ▶ To obtain a tractable form for $\ell_{\text{VA}}(\boldsymbol{\Psi}|q)$, we choose a parametric form for q . Specifically, we employ a mean-field approximation and set $q(\mathbf{u}|\boldsymbol{\xi}) = \prod_{i=1}^n q_i(\mathbf{u}_i|\boldsymbol{\xi}_i)$, where $q_i(\mathbf{u}_i|\boldsymbol{\xi}_i) = \mathcal{N}_p(\mathbf{a}_i, \mathbf{A}_i)$ yielding

$$\begin{aligned}\ell_{\text{VA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q) &= \int q(\mathbf{u}|\boldsymbol{\xi}) \log f(\mathbf{y}|\mathbf{u}, \boldsymbol{\Psi}) d\mathbf{u} \\ &\quad + \frac{1}{2} \sum_{i=1}^n \{ \log \det(\mathbf{A}_i) - \mathbf{a}_i' \mathbf{a}_i - \text{tr}(\mathbf{A}_i) \}.\end{aligned}$$

- ▶ VA has proven to be accurate and computationally efficient in cases where $\ell_{\text{VA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q)$ can be attained in a closed form⁸.
- ▶ Sadly this is not always the case. A prime example is the case of Bernoulli distributed responses with logit link function.

⁸Hui et al. (2017). Variational approximations for generalized linear latent variable models. *JCGS*, **26**, 35-43

Extended variational approximation (EVA)

- Motivated by the method of Laplace approximation, we approximate $\log f(\mathbf{y}|\mathbf{u}, \Psi)$ by its second-order Taylor expansion w.r.t. the latent variables \mathbf{u} , centered around \mathbf{a} , that is,⁹

$$\begin{aligned}\log f(\mathbf{y}|\mathbf{u}, \Psi) &\approx \log f(\mathbf{y}|\mathbf{a}, \Psi) + (\mathbf{u} - \mathbf{a})' \nabla_{\mathbf{u}} \log f(\mathbf{y}|\mathbf{u}, \Psi) \Big|_{\mathbf{u}=\mathbf{a}} \\ &\quad + \frac{1}{2} (\mathbf{u} - \mathbf{a})' \nabla_{\mathbf{u}}^2 \log f(\mathbf{y}|\mathbf{u}, \Psi) \Big|_{\mathbf{u}=\mathbf{a}} (\mathbf{u} - \mathbf{a}).\end{aligned}$$

- This leads to a closed-form approximation of the variational lower bound for GLLVMs with **any type of response and link function combination**, as now

$$\begin{aligned}\int q(\mathbf{u}) \log f(\mathbf{y}|\mathbf{u}, \Psi) d\mathbf{u} \\ \approx \log f(\mathbf{y}|\mathbf{a}, \Psi) + \frac{1}{2} \text{tr}(\nabla_{\mathbf{u}}^2 \log f(\mathbf{y}|\mathbf{u}, \Psi) \Big|_{\mathbf{u}=\mathbf{a}} \mathbf{A}).\end{aligned}$$

⁹Korhonen et al. (2022). Fast and universal estimation of latent variable models using extended variational approximations. [arXiv:2107.02627](https://arxiv.org/abs/2107.02627)

Estimation and inference

- ▶ GLLVM model fitting is performed by maximizing $\ell_{(\mathbf{E})\text{VA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q)$ simultaneously over $\boldsymbol{\Psi}$ and $\boldsymbol{\xi}$.
- ▶ The estimated variational distributions $\hat{q}_i(\mathbf{u}_i) = N_p(\hat{\mathbf{a}}_i, \hat{\mathbf{A}}_i)$ provide an approximation of $f(\mathbf{u}|\mathbf{y}, \boldsymbol{\Psi})$, that is, $\hat{\mathbf{a}}_i$ provide appropriate approximations to best predictors of \mathbf{u}_i (BP), and $\hat{\mathbf{A}}_i$ can be used to measure their variability.
- ▶ For the analysis of model parameters, the approximate asymptotic standard errors may be obtained using

$$\mathbf{I}(\hat{\boldsymbol{\Psi}}_{(\mathbf{E})\text{VA}}, \hat{\boldsymbol{\xi}}_{(\mathbf{E})\text{VA}}) = - \frac{\partial^2 \ell_{(\mathbf{E})\text{VA}}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial(\boldsymbol{\Psi}, \boldsymbol{\xi}) \partial(\boldsymbol{\Psi}, \boldsymbol{\xi})^T} \bigg|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}_{(\mathbf{E})\text{VA}}, \boldsymbol{\xi}=\hat{\boldsymbol{\xi}}_{(\mathbf{E})\text{VA}}}.$$

Implementation using TMB

- ▶ We use TMB (Template Model Builder) for fast estimation of model parameters¹⁰.
- ▶ TMB is an R package for fitting using automatic differentiation (AD) in optimization.
- ▶ TMB can calculate first and second order derivatives of the likelihood function written in C++. The likelihood function can be called from R and optimized using `optim()`.
- ▶ Standard errors for estimated parameters will be obtained by producing Hessian matrix with `optimHess()` in R.
- ▶ The full implementation of LA, VA and EVA via TMB are available in the R package `gllvm`¹¹.

¹⁰Kristensen et al. (2015). TMB: Automatic differentiation and Laplace approximation, *JSS*, **70**, 1–21.

¹¹Niku et al. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R, *MEE*, **10**, 2173–2182

Reponse distributions available in R package gl1vm

Response	Distribution	Link	Method
Binary	Bernoulli	logit	EVA/LA
		probit	EVA/VA/LA
Counts	Poisson	log	VA/LA
	Negative binomial	log	EVA/VA/LA
	ZIP	log	LA
Plant cover	Beta	probit	EVA/LA
Biomass	Tweedie	log	EVA/LA
Ordinal	Multinomial	probit	VA
Normal	Gaussian	Identity	VA/LA
Non-negative continuous	Gamma	log	VA/LA

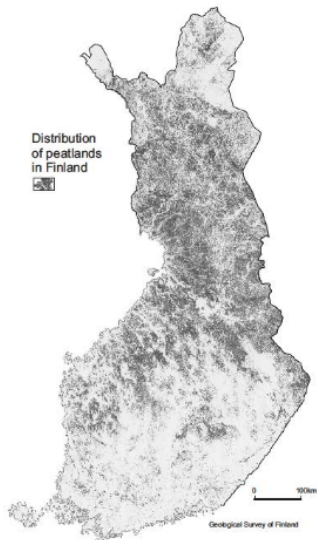
Data input

Main function of the `gllvm` package is `gllvm()`, which can be used to fit GLLVMs for multivariate data with the most important arguments listed in the following:

```
gllvm(y = NULL, X = NULL, TR = NULL, family, num.lv = NULL,  
formula = NULL, method = "VA", row.eff = FALSE, n.init = 1,...)
```




- `y`: matrix of abundances
- `X`: matrix or data.frame of environmental variables
- `TR`: matrix or data.frame of trait variables
- `family`: distribution for responses
- `num.lv`: number of latent variables
- `method`: approximation used ("VA", "EVA" or "LA")
- `row.eff`: type of row effects
- `n.init`: number of random starting points for latent variables

Example: Finnish peatland study



Example: Finnish peatland study

- ▶ Finnish environment institute was looking for new tools for ecological monitoring of peatlands.
- ▶ Former studies indicated that amoeba species might be useful when assessing biological impacts of peatland use.
- ▶ In Daza Secco et al. (2018)¹² we studied if amoeba species communities differ in terms of land use (natural, forestry, restored), and how environmental covariates affect species abundance.

¹²Daza Secco et al. (2018). Testate amoebae community analysis as a tool to assess biological impacts of peatland use, *Wetlands Ecology and Management*, **26**, 597-611.   

Data

- ▶ Three types of study sites located in the boreal zone of Central and Western Finland:
 - ▶ two **natural** peatlands
 - ▶ two peatlands that are used for **forestry**
 - ▶ two **restored** peatlands
- ▶ 15 sampling plots from each study site. Altogether $n = 270$ moss samples were taken.
- ▶ Amoeba species were identified and counted. Altogether $m = 51$ species were detected.
- ▶ Environmental covariates: water pH, temperature and water table depth).

Data matrix

	Name	LUse	Site	pH	temp	Cenacu	Cencas	Ceneco	Cenpla	Cycarc	Triarc	Trimin
1	asuo	na	1	4.6	10.3	3324	906	0	6043	1209	0	1511
2	asuo	na	1	4.2	9.3	7148	1787	0	1489	2383	0	8935
3	asuo	na	1	4.0	10.9	14671	386	0	5405	1158	0	1158
4	asuo	na	1	3.9	10.0	4009	1002	0	1403	1403	0	11025
5	asuo	na	1	3.9	9.2	1638	1489	0	2829	1042	0	2085
6	asuo	na	1	3.7	9.5	3475	204	0	0	0	0	6541
7	asuo	na	1	3.9	9.2	727	4242	0	1333	1576	0	364
8	asuo	na	1	4.0	10.3	6515	0	0	5212	1303	0	1303
9	asuo	na	1	3.9	9.4	4765	298	0	596	596	0	1191
10	asuo	na	1	3.6	9.0	2743	1097	0	823	686	0	1372
11	asuo	na	1	3.7	8.8	1002	200	0	401	200	0	13832
12	asuo	na	1	3.7	9.1	1714	286	0	571	1714	0	23989
13	asuo	na	1	3.5	9.3	1180	0	0	131	262	0	5900
14	asuo	na	1	3.7	9.1	3552	0	0	0	618	0	2780
15	asuo	na	1	3.8	10.1	2934	1235	0	927	309	0	10810

- To visualize the main trends between sampling sites in terms of their species composition we fitted a GLLVM model with $p = 2$ latent variables, that is,

$$\log(\mu_{ij}) = \alpha_{p(i)} + \beta_{0j} + \mathbf{u}_i' \boldsymbol{\lambda}_j,$$

where $\alpha_{p(i)}$ is now a plot-level random effect.

- To overcome overdispersion in responses, negative binomial distribution was used.
- Model can be fitted using `gllvm` package with

```
fit <- gllvm(y = Y, family = "negative binomial", num.lv = 2,  
            row.eff = ~(1 | Site),...)
```

- ▶ Predicted latent variables can plotted on a standard scatterplot to look for patterns between sites (model-based unconstrained ordination plot).

```
ordiplot(fit)
```

- ▶ If estimated factor loadings are added in the ordination plot, we can find indicator species for sites (model-based biplot).

```
ordiplot(fit, biplot = TRUE)
```

- ▶ For constrained ordination, see van der Veen et al., (2021).¹³

¹³van der Veen et al. (2021). Model-based ordination with constrained latent variables, bioRxiv.

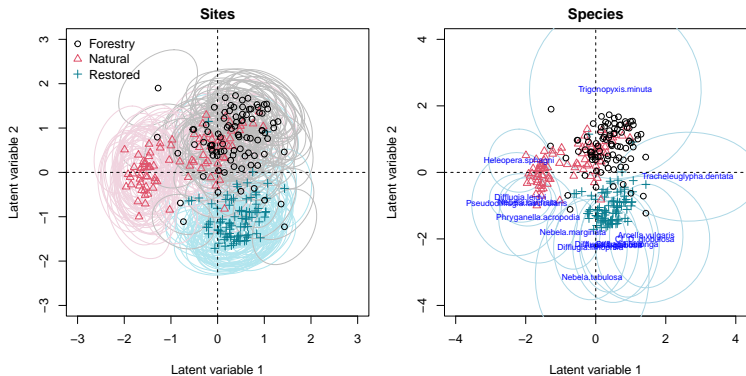


Figure: Model-based unconstrained ordination of the sites in the peatland data, along with 95% CMSEP-based prediction regions (left) and model-based biplot showing 15 indicator species for sites (right).

- To find out how environmental variables affect the abundance, we fitted NB-GLLVM including water pH, temperature, water table depth and land use type (as a factor with dummy variables) as covariates

$$\log(\mu_{ij}) = \alpha_{p(i)} + \beta_{0j} + \mathbf{x}'_i \boldsymbol{\beta}_j + \mathbf{u}'_i \boldsymbol{\lambda}_j,$$

- This can be done using:

```
fitX <- gllvm(y = Y, X = X, family = "negative binomial",  
             num.lv = 2, row.eff = ~(1 | Site),...)
```

- Information on correlation stored in the factor loadings can be used to estimate the correlation matrix of the linear predictor across species:

```
getResidualCor(fitX)
```

- The amount of covariation within and between species that is explained by the covariates can be quantified by calculating the relative change in the trace of the estimated residual covariance matrix $\hat{\Sigma} = \hat{\Gamma}\hat{\Gamma}'$.
- According to NB-GLLVM, water pH, temperature and water table depth together explain 31.8% of the covariation in the model. When the land usage was also included as covariate, 47.0% of covariation was explained.
- Plots of the estimated regression coefficients and corresponding 95% Wald intervals for covariates can be plotted using

```
coefplot(fitX, ...)
```

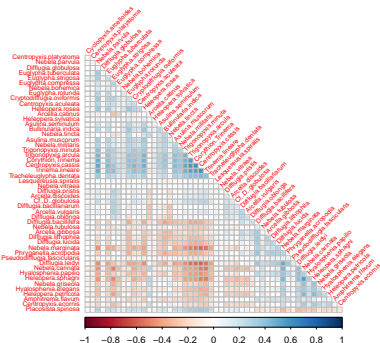
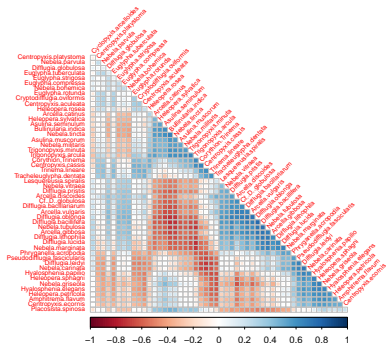


Figure: Residual correlation matrix based on latent factor loadings for the NB-GLLM without (left) and with (right) environmental covariates.

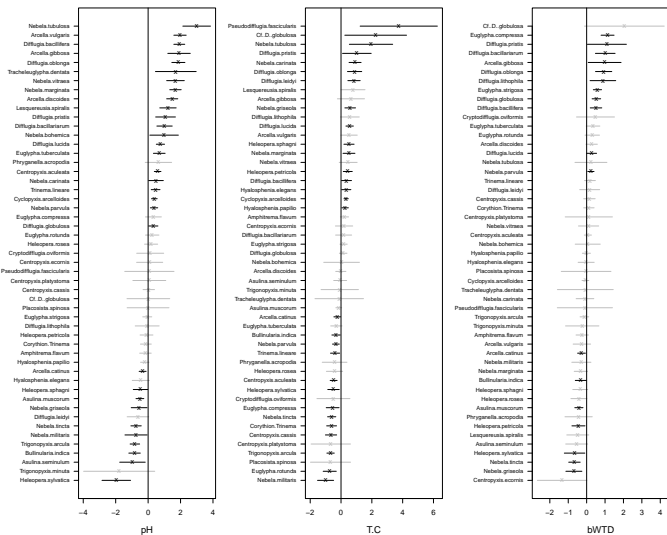


Figure: Coefficient plots containing the point estimates and 95% Wald confidence intervals for the effect of water pH, temperature and water table depth.

- For diagnosing model fit one can plot randomized quantile-based residuals designed for discrete data (Dunn and Smyth, 1996) against linear predictors:

```
plot(fitX)
```

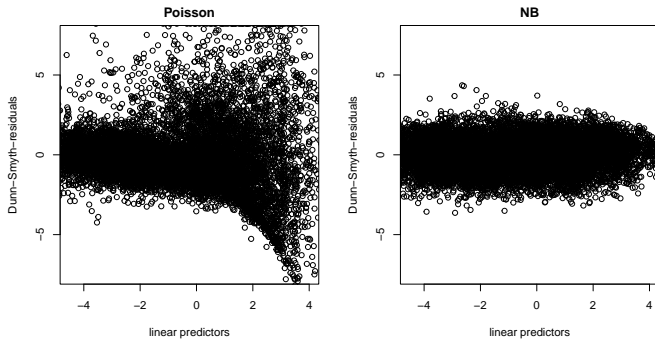


Figure: Residual plots for the Poisson GLLVM (left) and the NB-GLLVM (right).

Summary

- ▶ GLLVMs offer a flexible framework for modeling multivariate abundance data.
- ▶ The R package `gllvm` offers relatively fast methods to fit GLLVMs via maximum likelihood, along with tools for model checking, visualization and inference.
- ▶ Methods implemented in package are applicable for the most common types of responses in ecological studies: presence-absence, overdispersed counts, biomass and percent cover data.

Future work

- ▶ Implementation of mixed-response GLLVMs.
- ▶ GLLVMs involving temporally and/or spatially dependent latent variables.
- ▶ GLLVMs for compositional data.
- ▶ Computational solutions for high-dimensional data settings (parallel computation techniques, mini-batching, etc.). Regularization.
- ▶ Large sample properties of estimators.

For more details...

<https://jenniniku.github.io/gllvm/>