

**This is a self-archived version of an original article. This version may differ from the original in pagination and typographic details.**

**Author(s):** Niku, Jenni; Hui, Francis K.C.; Taskinen, Sara; Warton, David I.

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

**Year:** 2019

**Version:** Accepted version (Final draft)

**Copyright:** © 2019 The Authors. Methods in Ecology and Evolution © 2019 British Ecological !

**Rights:** In Copyright

**Rights url:** <http://rightsstatements.org/page/InC/1.0/?language=en>

**Please cite the original version:**

Niku, J., Hui, F. K., Taskinen, S., & Warton, D. I. (2019). gllvm : Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10(12), 2173-2182. <https://doi.org/10.1111/2041-210X.13303>

# Methods in Ecology and Evolution

## 1 `gllvm` – Fast analysis of multivariate abundance data 2 with generalized linear latent variable models in R

Jenni Niku\*, Francis K. C. Hui†, Sara Taskinen\*, David I. Warton‡

3 \*Department of Mathematics and Statistics, University of Jyväskylä, Finland

†Research School of Finance, Actuarial Studies & Statistics, Australian National University, Australia

‡School of Mathematics and Statistics and Evolution & Ecology Research Centre, UNSW Sydney,  
Australia

4 **Running Header - `gllvm` R package**

5 **Word count:** 3500 words

### 6 **Summary**

- 7 1. There has been rapid development in tools for multivariate analysis based  
8 on fully specified statistical models or “joint models”. One approach  
9 attracting a lot of attention is generalized linear latent variable models  
10 (GLLVMs). However, software for fitting these models is typically slow  
11 and not practical for large datasets.
- 12 2. The R package `gllvm` offers relatively fast methods to fit GLLVMs via  
13 maximum likelihood, along with tools for model checking, visualization  
14 and inference.
- 15 3. The main advantage of the package over other implementations is speed  
16 *e.g.* being two orders of magnitude faster, and capable of handling thou-  
17 sands of response variables. These advances come from using variational  
18 approximations to simplify the likelihood expression to be maximised, au-  
19 tomatic differentiation software for model-fitting (via the TMB package),  
20 and careful choice of initial values for parameters.
- 21 4. Examples are used to illustrate the main features and functionality of  
22 the package, such as constrained or unconstrained ordination, including  
23 functional traits in “fourth corner” models, and (if the number of envi-  
24 ronmental coefficients is not large) make inferences about environmental  
25 associations.

26 *Keywords:* High-dimensional data, joint modelling, multivariate analysis, or-  
27 dination, species interactions

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:

10.1111/2041-210X.13303

This article is protected by copyright. All rights reserved.

28 Multivariate abundance data, consisting of observations of multiple interacting species  
29 (or other taxonomic group) from a set of samples, are often collected in ecological studies  
30 to characterise a community or assemblage of organisms. The term “abundance” is taken  
31 here to mean counts, presence-absence records, biomass data or any other measure of  
32 the extent to which a species may be present at a site. Common ecological questions  
33 that such data are used to answer include whether a set of sites are similar in terms of  
34 their species composition (Bjork et al., 2018), finding between species interactions and  
35 visualization of correlation patterns across species (Royan et al., 2016), hypothesis testing  
36 of environmental effects (Lammel et al., 2018), and making predictions for abundances  
37 (Buisson et al., 2008).

38 In recent years, there has been a growing movement towards the specification of statistical  
39 models for multivariate analysis in ecology (Ovaskainen et al., 2010; Warton et al., 2015;  
40 Ovaskainen et al., 2017). Of particular interest are methods that use random effects to  
41 incorporate between species correlation in models predicting species abundance as a func-  
42 tion of environmental variables, often termed joint species distribution models (Pollock  
43 et al., 2014). One exciting possibility offered by these methods is the potential to tease  
44 apart some of the causes of species co-occurrence – joint response to known environmental  
45 gradients versus other sources, *e.g.* biotic interaction.

46 A key approach for statistical modelling of multivariate abundance data is the generalized  
47 linear latent variable model (GLLVM, Skron dal and Rabe-Hesketh, 2004). A GLLVM  
48 extends the basic generalized linear model to multivariate data using a factor analytic  
49 approach, *i.e.* incorporating a small number of latent variables for each site accompanied  
50 by species specific factor loadings to model correlations between responses. These latent  
51 variables have a natural interpretation as ordination axes, but with additional capacity,  
52 *e.g.* predicting new values, controlling for known environmental variables, using standard  
53 model selection tools to choose number of ordination axes (Hui et al., 2015). One of the  
54 main advantages of GLLVMs is that they can handle situations where there are many  
55 species, because the number of parameters in the covariance model scales linearly with  
56 the number of responses (Warton et al., 2015). This is a key technical challenge – often  
57 there are more species being sampled than sites, *e.g.* microbial data often has thousands  
58 of taxa (Niku et al., 2017; Kumar et al., 2017).

59 Software for fitting GLLVMs in ecology is currently quite slow computationally and not  
60 practical for large datasets. In particular, packages in the freely available software R  
61 have been developed, *e.g.* the `boral` (Hui, 2016) and `HMSC` packages (Tikhonov et al.,

2019), but using Bayesian MCMC for estimation, which is relatively slow and not practical for large microbial datasets. More technical advances provide the opportunity to reduce computation times on some problems from hours to minutes or minutes to seconds, using variational (Hui et al., 2017) or Laplace (Niku et al., 2017) approximations to likelihoods, especially via automated differentiation software such as Template Model Builder (Kristensen et al., 2016).

This paper presents the R package `gllvm` (Niku et al., 2019a), which has been developed for rapid fitting of GLLVMs to multivariate abundance data. The package offers a framework for model-based ordination, as well as allowing us to study the effect of environmental covariates or environmental-trait interactions on responses simultaneously with the analysis of correlation patterns across species. The package also contains tools for statistical inference, model selection and visualization. While other R packages have similar functionality (Tikhonov et al., 2019; Hui, 2016), the key point of distinction is that `gllvm` fits models much faster than its immediate competitors (*e.g.* see Table 3) and is capable of modelling larger datasets. Version 1.1.7 of the `gllvm` package is currently available on the Comprehensive R Archive Network (CRAN).

## Generalized linear latent variable models

A multivariate abundance dataset can be defined by a matrix of abundances, with  $n$  rows (usually sites) and  $m$  columns of responses (usually species). Denote the abundance of the  $j$ th species at the  $i$ th site as  $y_{ij}$ . A set of  $k$  environmental variables, or experimental treatments, may also be recorded at each site and stored in the vector  $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})^\top$ . A GLLVM regresses the mean abundance  $\mu_{ij}$  against environmental variables and a vector of  $d \ll m$  latent variables,  $\mathbf{u}_i = (u_{i1}, \dots, u_{id})^\top$ :

$$g(\mu_{ij}) = \eta_{ij} = \alpha_i + \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + \mathbf{u}_i^\top \boldsymbol{\gamma}_j, \quad (1)$$

where  $\boldsymbol{\beta}_j$  and  $\boldsymbol{\gamma}_j$  are vectors of species specific coefficients related to the covariates and latent variables, respectively. The latent variables  $\mathbf{u}_i$  can be thought of as unmeasured environmental variables, or as ordination scores, capturing the main axes of covariation of abundance (after controlling for observed predictors  $\mathbf{x}_i$ ). We assume these latent variables are independent across sites and standard normally distributed. The parameters  $\beta_{0j}$  are species specific intercepts, while  $\alpha_i$  are optional site effects which can be chosen as either fixed or random effects ( $\alpha_i \sim N(0, \sigma^2)$ ). The row effects  $\alpha_i$  can be included for site total

92 abundance standardization, that is, all other terms in the model can then be subsequently  
 93 interpreted as modelling *relative abundance* or compositional effects (Hui et al., 2015). To  
 94 ensure that the above model is identifiable, for  $m > 1$  the upper triangular of the loading  
 95 matrix  $\mathbf{\Gamma} = [\boldsymbol{\gamma}_1 \dots \boldsymbol{\gamma}_m]'$  needs to be set to zero and the diagonal elements positive to avoid  
 96 rotational invariance; see Hui et al. (2015) and Niku et al. (2017) for further information.

97 The residual covariance matrix, storing information on species co-occurrence that is not  
 98 explained by environmental variables, can be calculated as  $\boldsymbol{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^\top$ . This is the correct  
 99 form of correlation when the responses are Poisson distributed. In the case of negative bi-  
 100 nomial distribution with dispersion parameters  $\boldsymbol{\Phi} = (\phi_1, \dots, \phi_m)^\top$ , we adjust the diagonal  
 101 elements by adding the term  $\log(\phi_j + 1)$ , which corresponds to the variance explained by  
 102 the NB distribution. Analogously, for the binomial probit model the residual covariance  
 103 is  $\boldsymbol{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}^\top + \mathbf{I}_m$  (Ovaskainen et al., 2016).

104 If  $q$  trait covariates  $\mathbf{t}_j = (t_{j1}, \dots, t_{jq})^\top$  are also recorded, we can use them to help explain  
 105 inter-specific variation in environmental response. This leads to an extension of the so-  
 106 called “fourth corner model” (Jamil and ter Braak, 2013; Brown et al., 2014) where  
 107 multivariate abundance is regressed against a function of traits and environment, and the  
 108 environment-trait interactions represents the fourth corner association between traits and  
 109 environment. The associated fourth corner GLLVM then has mean model:

$$g(\mu_{ij}) = \eta_{ij} = \alpha_i + \beta_{0j} + \mathbf{x}_i^\top \boldsymbol{\beta}_e + (\mathbf{t}_j \otimes \mathbf{x}_i)^\top \boldsymbol{\beta}_I + \mathbf{u}_i^\top \boldsymbol{\gamma}_j, \quad (2)$$

110 where  $\boldsymbol{\beta}_e$  is a vector of main effects for environmental covariates, and  $\boldsymbol{\beta}_I$  are the fourth  
 111 corner coefficients. A main effect for traits was not included, because main effects on  
 112 abundance across species are absorbed by the intercept term  $\beta_{0j}$ . This model assumes  
 113 that all inter-specific variation in response to covariates is mediated by species, which  
 114 reduces the number of parameters related to covariates from  $mk$  in equation (1) to  $k(q+1)$   
 115 in (2).

116 In both GLLVM formulations above, a key feature is that the number of parameters  
 117 characterizing the residual correlation  $\mathbf{\Gamma}\mathbf{\Gamma}^\top$  grows linearly with the number of responses  $m$ .  
 118 This contrasts to the quadratic rate of growth when an unstructured residual covariance  
 119 matrix were assumed across responses (Pollock et al., 2014). Thus the term  $\mathbf{u}_i^\top \boldsymbol{\gamma}_j$  is able  
 120 to model residual correlation across response variables even when the number of species  
 121 is relatively large.

## 122 Estimation

123 A difficulty fitting the GLLVM is that the  $\mathbf{u}_i$ 's are unobserved and we must integrate over  
124 their possible values. Specifically, the log-likelihood function we wish to maximise has the  
125 form

$$l(\Psi) = \sum_{i=1}^n \log(f(y_{ij}, \Psi)) = \sum_{i=1}^n \log \left( \int_{\mathbb{R}^d} \prod_{j=1}^m f(y_{ij} | \mathbf{u}_i; \Psi) f(\mathbf{u}_i) d\mathbf{u}_i \right), \quad (3)$$

126 where  $\psi$  includes all model parameters. In this expression we have assumed abundances  
127 are independent across sites and any correlation across responses is captured by the latent  
128 variables  $\mathbf{u}_i$ . Thus conditional on  $\mathbf{u}_i$ , the  $y_{ij}$  are independent of each other within sites.

129 In the literature, several solutions have been proposed to the problem of integration  
130 (3), most notably adaptive quadrature (Rabe-Hesketh et al., 2002), the Monte-Carlo  
131 applications of the expectation maximization (EM) algorithm (Hui et al., 2015), and  
132 Bayesian MCMC (Tikhonov et al., 2019; Hui, 2016). For large datasets and multiple  
133 latent variables these methods are, however, time-consuming.

134 The `gllvm` package overcomes these computational problems using three key innovations:

- 135 • Maximising an approximation to the log-likelihood that is (almost completely)  
136 closed form. We provide two ways to do this – using Gaussian variational ap-  
137 proximations (VA, Hui et al., 2017) for overdispersed counts, binary and ordinal  
138 responses, or using Laplace approximations (LA, Niku et al., 2017) for other ex-  
139ponential family distributions when a fully closed form variational approximation  
140 cannot be obtained *e.g.* biomass data can be modelled by the Tweedie distribution.
- 141 • Parameter estimation makes use of automatic differentiation software in `C++` to ac-  
142 celerate computation times, via the interface provided by the R package `TMB` (Kris-  
143tensen et al., 2016).
- 144 • Careful choice of starting values. In particular, we use a factor analysis on Dunn-  
145 Smyth residuals (Niku et al., 2019b) to obtain starting values close to the anticipated  
146 solution, optionally, with jittering to check the sensitivity of the approach.

147 The end result is a package that provides more stable solutions, and is orders of magnitude  
148 faster than current competitors.

## 149 Using the R package `gllvm`

150 The R package `gllvm` provides a flexible implementation for fitting GLLVMs to multivari-  
151 ate data. The main function of the `gllvm` package is `gllvm()`, which can be used to fit  
152 GLLVMs for multivariate data with the most important arguments listed in the following:

```
153 gllvm(y = NULL, X = NULL, TR = NULL, data = NULL, formula = NULL,  
154       num.lv = 2, family, method = "VA", row.eff = FALSE, offset = NULL,  
155       Power = 1.5, starting.val = "res", ...)
```

156 Data input can be specified using the “wide format” matrices via `y`, `X` and `TR` arguments,  
157 or using the long format via `data` argument, and `formula` is used for model specification  
158 (which defaults to including linear terms for all variables from `X` and `TR`, and all interac-  
159 tions between variables in `X` and variables in `TR`). The number of latent variables can be  
160 defined using the argument `num.lv`, with zero latent variables corresponding to a simple  
161 multi-response GLM that does not account for correlation across responses (Wang et al.,  
162 2012). The response distribution can be chosen using the argument `family`, and mod-  
163 els can be fitted using either the VA (`method = "VA"`, default) or with the LA (`method`  
164 `= "LA"`) method. The currently available distributions, link functions and methods for  
165 different response types are listed in Table 1.

166 Other important arguments in the `gllvm` call are `row.eff` for defining the type of row  
167 effects (`none`, `fixed` or `random`), `offset` for potential inclusion of offsets, `Power` for defining  
168 the power parameter of the Tweedie distribution (Niku et al., 2017) and `starting.val`  
169 for judicious choice of starting values for the latent variables (Niku et al., 2019b). For an  
170 overview of the available functions in `gllvm`, see Table 2.

171 Below, we demonstrate the main features of the `gllvm` package by example. In the  
172 examples we consider the `antTraits` data, which is available in the R package `mvabund`  
173 (Wang et al., 2012) and consists of counts of 41 ant species measured at 30 sites across  
174 south-east Australia, along with records of five environmental variables and five trait  
175 variables for each species. The package and the data can be loaded as follows.

```
176 > library(gllvm)  
177 > data(antTraits)  
178 > y <- as.matrix(antTraits$abund); X <- scale(as.matrix(antTraits$env))  
179 > TR <- antTraits$traits
```

## 180 Model-based ordination

181 GLLVMs can be used as a model-based approach to unconstrained ordination by including  
182 (*e.g.*) two latent variables in the model but no predictors (Walker and Jackson, 2011; Hui  
183 et al., 2015). The corresponding ordination plot then provides a graphical representation  
184 of which sites are similar in terms of their species composition. Such a model can be fitted  
185 to the `antTraits` data using the function `gllvm()` as below. We will consider two count  
186 distributions for the data – the Poisson and negative binomial (NB).

```
187 > fitp <- gllvm(y, family = poisson())
188 > fitp
189 Call:
190 gllvm(y = y, family = poisson())
191 family:
192 [1] "poisson"
193 ...
194 AIC: 4501.263
195 AICc: 4178.553
196 BIC: 4672.209
197 > fit_ord <- gllvm(y, family = "negative.binomial")
198 > fit_ord
199 Call:
200 gllvm(y = y, family = "negative.binomial")
201 family:
202 [1] "negative.binomial"
203 ...
204 AIC: 4116.173
205 AICc: 3717.188
206 BIC: 4344.568
```

207 The default printout includes information criteria, which all suggest that the NB distribu-  
208 tion is a better choice than the Poisson distribution for modelling the response. Residual  
209 plots for diagnosing model fit in Figure 1 can be obtained using the `plot()` function. Two  
210 plots for both models are of Dunn-Smyth residuals, which are randomized quantile based  
211 residuals designed for discrete data (Dunn and Smyth, 1996), plotted against linear pre-  
212 dictors, and a normal quantile-quantile plot with a simulated point-wise 95% confidence



213 interval envelope. The residual diagnostics for the Poisson model shows some overdispersion in residuals, in particular, a telltale fan-shape in the plot of residuals against fitted values. These issues are largely resolved in the NB model. Note that the latent variables in the model provide some capacity to account for overdispersion, so overdispersed counts do not always require us to move beyond the Poisson distribution, although there is clear evidence of such a need in this example.

219 Once an appropriate model has been established for the data, we can construct an ordination as a scatter plot of the predicted latent variables via the `ordiplot()` function. The species with the largest factor loadings (largest norms,  $\|\gamma_j\|$ ), and hence most strongly associated with ordination scores, can be added using the logical argument `biplot`, leading to a biplot for finding indicator species corresponding to specific sites. The `ind.spp` argument defines the number of species to be plotted.

```
225 > ordiplot(fit_ord, biplot = TRUE, ind.spp = 15,  
226 +       xlim = c(-3, 3), ylim = c(-2, 1.6))
```

227 The above command creates the biplot as shown in Figure 2 based on the GLLVM fitted to the `antTraits` data. We can see one large cluster of sites on the top with many indicator species, and few smaller clusters with only few indicator species *e.g.* sites 12–15. In Appendix 3 we apply classical algorithm-based ordination methods to the ant data and compare the results. While the results between GLLVMs and the algorithmic-based methods are quite similar, GLLVMs offer the advantage of standard tools for diagnosing model fit and performing model selection.

## 234 Model with environmental variables

235 Environmental variables can be included in the model, whether to study their effects on assemblages, or to study patterns of species co-occurrence after controlling for environmental variables.

```
238 > fit_env <- gllvm(y, X, family = "negative.binomial", num.lv = 3,  
239 +       formula = ~ Bare.ground + Shrub.cover + Volume.lying.CWD)
```

240 A model with three latent variables was chosen based on the AICc value, and residual analysis indicates that a NB distribution offered the most suitable mean-variance relationship for the responses.

243 The estimated coefficients for predictors and their confidence intervals can be plotted  
244 using the `coefplot()` function, in order to study the nature of effects of environmental  
245 variables on species.

```
246 > coefplot(fit_env, cex.ylab = 0.7, mar = c(4, 9, 2, 1),  
247 +   xlim.list = list(NULL, NULL, c(-4, 4)))
```

248 The resulting plot is given in Figure 3. Note that with a log link used, a unit change  
249 covariate  $l$  equates to a multiplicative change of  $\exp(\hat{\beta}_{jl})$  in the predicted mean  $\hat{\mu}_{ij}$  for  
250 species  $j$ . Most of the 95% confidence intervals include zero, indicating that the majority  
251 of the species do not exhibit evidence of a strong association between environment and  
252 species abundance. This may be due to a lack of information in the data, as much as  
253 being due to a lack of environmental association after accounting for potential residual  
254 species covariation.

## 255 Studying co-occurrence patterns

256 Latent variables induce correlation across response variables, and so provide a means of  
257 estimating correlation patterns across species, and the extent to which they can be ex-  
258 plained by environmental variables. As explained previously, information on correlation is  
259 stored in the factor loadings, and the `getResidualCor()` function can be used to estimate  
260 the correlation matrix of the linear predictor across species. This can be visualised using  
261 the `corrplot` package:

```
262 > cr <- getResidualCor(fit_env)  
263 > library("corrplot"); library("gclus");  
264 > corrplot(cr[order.single(cr), order.single(cr)], diag = FALSE, type =  
265 + "lower", method = "square", tl.cex = 0.8, tl.srt = 45, tl.col = "red")
```

266 Regions coloured in dark blue on Figure 4 indicate clusters of species that are positively  
267 correlated with each other, after controlling for covariation in species explained by the  
268 environmental terms in `fit_env`. There are also two regions coloured in red, indicating  
269 negative correlation between pairs of species. The effect of the environmental variables  
270 on the between species correlations can be seen by comparing the correlation matrix in  
271 Figure 4 to the correlation matrix given by the model without environmental variables,  
272 see example in Appendix 1, where the correlation patterns are considerably different from

273 one another. Correlations can also be visualized in a residual biplot (Appendix 1). The  
274 traces of residual covariances obtained via the `getResidualCov()` function can be used  
275 to quantify the amount of variation in the data explained by environmental variables  
276 (Warton et al., 2015), see Appendix 1.

## 277 Incorporating functional traits into “fourth corner” models

278 In the previous section, environmental associations were studied by fitting separate terms  
279 for each species, without attempting to explain why different species respond differently  
280 to the environment. Adding functional traits to the model offers the potential to explain  
281 why species differ in environmental response. The fourth corner model in equation (2)  
282 can be fitted by using the argument `TR` to include traits, and the argument `formula` is  
283 used to specify the model.

```
284 > fit_4th <- gllvm(y, X, TR, family = "negative.binomial", num.lv = 3,  
285 +   formula = y ~ (Bare.ground + Shrub.cover + Volume.lying.CWD) +  
286 +   (Bare.ground + Shrub.cover + Volume.lying.CWD) :  
287 +   (Pilosity + Polymorphism + Webers.length))
```

288 As previously, coefficients can be plotted using the function `coefplot()`. The environmen-  
289 tal-trait interaction terms, also known as the fourth corner terms, can also be visualized  
290 using the function `levelplot()` from the package `lattice`, see Appendix 1 for example  
291 code. The resulting plots in Figure 5 indicate that interactions of the trait variable  
292 `Polymorphism` with `Bare.ground` and `Webers.length` with `Volume.lying.CWD` have the  
293 strongest effects on ant abundances. Notice that `Pilosity` and `Polymorphism` are factors  
294 and `gllvm()` recognises this.

295 By using a maximum likelihood framework, `gllvm` offers likelihood-based machinery for  
296 model-based inference. A particular example is likelihood ratio testing via the `anova()`  
297 function when comparing nested models. In Figure 5, for example, all the trait-envi-  
298 ronment interactions appear to be relatively small and most of the confidence intervals of  
299 the coefficients include zero values. But to formally test whether these traits vary en-  
300 vironment, in the below code we fitted a second model without traits and performed a  
301 likelihood ratio test. Notice that in order to separate the next model from the one which  
302 has species specific coefficients for environmental variables, we include `TR` matrix to the  
303 function call.

```

304 > fit_4th2 <- gllvm(y, X, TR, family = "negative.binomial", num.lv = 3,
305 +   formula = y ~ (Bare.ground + Shrub.cover + Volume.lying.CWD))
306 > anova(fit_4th, fit_4th2)
307 Model 1 : y ~ (Bare.ground + Shrub.cover + Volume.lying.CWD)
308 Model 2 : y ~ (Bare.ground + Shrub.cover + Volume.lying.CWD) +
309 (Bare.ground + Shrub.cover + Volume.lying.CWD) : (Pilosity + Polymorphism
310 + Webers.length)
311   Resid.Df      D Df.diff P.value
312 1      1025  0.00000      0
313 2      1007 18.90272     18 0.397837

```

314 Based on the output from applying the `anova()` function, the  $p$ -value suggests that the  
315 simpler model where traits were not included is more appropriate i.e., there is no strong  
316 evidence of traits mediating the environmental response of species.

317 The validity of any model-based inference procedure relies on the assumptions of its  
318 underlying model. Note that the above test is based on `fit_4th`, a model that made the  
319 strong assumption that all interspecific variation in environmental response is captured  
320 by the trait in the model. Tests based on such models can have inflated false positive  
321 rates when this assumption is violated, as can be shown using simulations with missing  
322 trait predictors (ter Braak, 2019). We are working on an extension of our model, using a  
323 random slope across species, to capture variation in environmental response not captured  
324 by the trait model. Tests based on such a model can be expected to have much-improved  
325 robustness to missing predictors in the trait model.

## 326 Summary

327 In this paper, we introduced the R package `gllvm` for the analysis of multivariate abun-  
328 dance data using GLLVMs. The package caters for the types of response variables most  
329 commonly seen in ecology, including presence-absence data, overdispersed counts, biomass  
330 and ordinal data. The main point of difference between `gllvm` and other packages for fit-  
331 ting GLLVMs (Tikhonov et al., 2019; Hui, 2016) is that our algorithm is much faster for  
332 model-fitting, and thus capable of handling much larger datasets. Computational effi-  
333 ciency was achieved by avoiding MC approaches to estimation, and instead making use of  
334 recent innovations for maximum likelihood estimation as discussed in *Estimation*. Table 3  
335 illustrates this by comparing the computation time of `gllvm` to `boral` with default set-

336 tings (40 000 total iterations, warm-up at 10 000, thinning at 30), for the three example  
337 models of this paper. Computation times were over 140 times shorter when using `gllvm`,  
338 analysing the data in seconds rather than minutes. Note that this example dataset was  
339 relatively small, and differences in computation time become practically meaningful for  
340 larger datasets. For example, for the metagenomic dataset of Niku et al. (2017), with 56  
341 rows and 985 responses, `gllvm` fitted a two latent variable model without predictors in  
342 15 minutes, while `boral` (under default settings) took 10 hours, without achieving con-  
343 vergence. Even larger datasets again can be handled by `gllvm`, for which analysis is  
344 otherwise infeasible with currently available packages.

345 A second point of difference between `gllvm` and competing packages is that it uses a  
346 maximum likelihood framework, and thus can employ likelihood-based tools for inference.  
347 Familiar generic R functions like `AIC`, `BIC` and `anova` can be applied to `gllvm` objects,  
348 although as previously we emphasise that `anova` results will only be reliable when testing  
349 hypotheses concerning a relatively small number of parameters. To compare, packages  
350 that fit GLLVMs under a Bayesian framework would return full posterior distributions for  
351 both parameters and latent variables (Tikhonov et al., 2019; Hui, 2016), while our likeli-  
352 hood based framework returns approximate confidence intervals for parameters, assuming  
353 estimators are normally distributed. On the other hand, performing Bayesian hypothesis  
354 testing presents a bigger challenge compared to using likelihood based hypothesis testing  
355 as the `gllvm` package implements.

356 The GLLVM framework is distinct from methods historically used for ordination in ecol-  
357 ogy, such as non-metric multi-dimensional scaling (nMDS, as in `vegan`, Oksanen et al.,  
358 2018) and duality diagrams (as in `ade4`, Dray and Dufour, 2007). A key point of distinc-  
359 tion is that a GLLVM specifies a statistical model for the data intended to capture key  
360 data properties. In particular, multivariate abundance data typically have a strong mean-  
361 variance relationship, which if not accounted for, often introduces artifacts into analyses  
362 (Warton et al., 2012; Warton and Hui, 2017). Specifying a statistical model that aims to  
363 capture this mean-variance relationship, and using diagnostic tools to check its adequacy  
364 (Figure 1), can avoid this issue.

365 In the future, we plan to broaden the scope of the `gllvm` package to handle spatial and  
366 temporal correlations that often characterise observational multivariate abundance data,  
367 by allowing the latent variables to be structured rather than assuming independence  
368 across observational units. We will also extend the fourth corner models by including  
369 species specific random slopes for the predictors, to account for interspecific variation

370 in environmental response that is *not* explained by traits. The code repository for the  
371 package can be found from github, see <https://github.com/JenniNiku/gllvm> .

## 372 **Acknowledgments**

373 The work of JN was supported by the Wihuri Foundation. The work of ST was supported  
374 by the CRoNoS COST Action IC1408. The work of FKCH and DIW was supported by  
375 Australia Research Council Discovery Project grants (DP180100836 and DP150100823,  
376 respectively), FKCH was also supported by an ANU cross disciplinary grant.

## 377 **Supporting Information**

378 Appendix 1: R code for examples

379 Appendix 2: Analysis of high-dimensional microbial data

380 Appendix 3: Comparing model-based and algorithm-based ordination methods

## 381 **Authors contributions**

382 JN, FKCH, ST and DIW conceived the ideas and designed methodology; JN was main  
383 responsible for implementing the application; All authors contributed to the writing,  
384 reviewing and editing of the draft and gave final approval for publication.

## 385 **Data accessibility**

386 The ant dataset used in our examples is publicly available from the R package `mvabund`  
387 (Wang et al., 2012) in the Comprehensive R Archive Network: <https://cran.r-project.org/web/packages/mvabund/>. The microbial data (Kumar et al., 2017) is published in  
388 European Nucleotide Archive under the project number PRJEB17695, <https://www.ebi.ac.uk/ena/data/view/PRJEB17695>. A subset of this data used in Appendix 2, as  
389 well as all code used in this paper and supplementary materials is publicly available in  
390 the R package `gllvm` (Niku et al., 2019a) in the CRAN: <https://cran.r-project.org/web/packages/gllvm/>.  
391  
392  
393

## References

- Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley.
- Bjork, J.R., Hui, F.K.C., O'Hara, R.B., and Montoya, J.M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, 27:2714–2724.
- Brown, A.M., Warton, D.I., Andrew, N.R., Binns, M., Cassis, G., and Gibb, H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, 5:344–352.
- Buisson, L., Thuiller, W., Lek, S., Lim, P., and Grenouillet, G. (2008). Climate change hastens the turnover of stream fish assemblages. *Global Change Biology*, 14:2232–2248.
- Dray, S. and Dufour, A.B. (2007). The `ade4` Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 22:1–20.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244.
- Hui, F.K.C. (2016). `boral` – bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution*, 7:744–750.
- Hui, F. K.C., Taskinen, S., Pledger, S., Foster, S.D., and Warton, D.I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, 6:399–411.
- Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V., and Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26:35–43.
- Jamil, T. and ter Braak, C.J. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, 1:e95.
- Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H., and Bell, B.M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, 70:1–21.
- Kumar, M., Brader, G., Sessitsch, A., Mäki, A., van Elsas, J.D., and Nissinen, R. (2017) Plants Assemble Species Specific Bacterial Communities from Common Core Taxa in Three Arcto-Alpine Climate Zones. *Frontiers in Microbiology*, 8:12.



- 423 Lammel, D.R., Barth, G., Ovaskainen, O., Cruz, L.M., Zanatta, J.A., Ryo, M., de Souza,  
424 E.M., and Pedrosa, F.O. (2018). Direct and indirect effects of a pH gradient bring  
425 insights into the mechanisms driving prokaryotic community structures. *Microbiome*,  
426 6:6–106.
- 427 Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., and Warton, D.I.  
428 (2019a). *gllvm: Generalized Linear Latent Variable Models*. R package version 1.1.7.  
429 <https://cran.r-project.org/web/packages/gllvm/>
- 430 Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., and Warton, D.I.  
431 (2019b). Efficient estimation of generalized linear latent variable models. *PLoS One*,  
432 14(5):1–20.
- 433 Niku, J., Warton, D.I., Hui, F.K.C., and Taskinen, S. (2017). Generalized linear la-  
434 tent variable models for multivariate count and biomass data in ecology. *Journal of*  
435 *Agricultural, Biological, and Environmental Statistics*, 22:498–522.
- 436 Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D.,  
437 Minchin, P.R., O’Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E.  
438 and Wagner, H. (2018). *vegan: Community Ecology Package*. R package version 2.5-3.  
439 <https://CRAN.R-project.org/package=vegan>
- 440 Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016). Using latent variable  
441 models to identify large networks of species-to-species associations at different spatial  
442 scales. *Methods in Ecology and Evolution*, 7:549–555.
- 443 Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence  
444 by multivariate logistic regression generates new hypotheses on fungal interactions.  
445 *Ecology*, 91:2514–2521.
- 446 Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson,  
447 D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? A  
448 conceptual framework and its implementation as models and software. *Ecology Letters*,  
449 20:561–576.
- 450 Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O’Hara, R.B., Parris, K.M., Vesk,  
451 P.A., and McCarthy, M.A. (2014). Understanding co-occurrence by modelling species  
452 simultaneously with a joint species distribution model (JSDM). *Methods in Ecology*  
453 *and Evolution*, 5:397–406.



- 454 Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized  
455 linear mixed models using adaptive quadrature. *Stata Journal*, 2:1–21.
- 456 Royan, A., Reynolds, S.J., Hannah, D.M., Prudhomme, C., Noble, D.G., and Sadler,  
457 J.P. (2016). Shared environmental responses drive co-occurrence patterns in river bird  
458 communities. *Ecography*, 39:733–742.
- 459 Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multi-*  
460 *level, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton.
- 461 ter Braak, C.J.F. (2019) New robust weighted averaging- and model-based methods for  
462 assessing trait-environment relationships. *Methods in Ecology and Evolution*, In press.
- 463 Tikhonov, G., Opedal, Ø., Abrego, N., Lehikoinen, A., and Ovaskainen, O.  
464 (2019). Joint species distribution modelling with HMSC-R. *bioRxiv* preprint,  
465 <https://doi.org/10.1101/603217>.
- 466 Walker, S.C. and Jackson, D.A. (2011). Random-effects ordination: Describing and pre-  
467 dicting multivariate correlations and co-occurrences. *Ecological Monographs*, 81:635–  
468 663.
- 469 Wang, Y., Naumann, U., Wright, S.T., and Warton, D.I. (2012). mvabund - an R package  
470 for model-based analysis of multivariate abundance data. *Methods in Ecology and*  
471 *Evolution*, 3:471–474.
- 472 Warton, D.I., Blanchet, F.G., O’Hara, R., Ovaskainen, O., Taskinen, S., Walker, S.C.,  
473 and Hui, F.K.C. (2015). So many variables: Joint modeling in community ecology.  
474 *Trends in Ecology and Evolution*, 30:766–779.
- 475 Warton, D. I. and Hui, F. K. (2017). The central role of mean-variance relationships in  
476 the analysis of multivariate abundance data: a response to Roberts (2017). *Methods in*  
477 *Ecology and Evolution*, 8:1408–1414.
- 478 Warton, D.I., Wright, S.T. and Wang, Y. (2012). Distance-based multivariate analyses  
479 confound location and dispersion effects. *Methods in Ecology and Evolution*, 3:89–101.

Table 1: Overview of available distributions with the mean,  $E(y_{ij})$ , and mean-variance,  $V(\mu_{ij})$ , functions, estimation methods and link functions for various response types in `gllvm`.

Response	Distribution	Method	Link	Description
Counts	Poisson	VA/LA	log	$E(y_{ij}) = \mu_{ij}, V(\mu_{ij}) = \mu_{ij}$
	NB	VA/LA	log	$E(y_{ij}) = \mu_{ij}, V(\mu_{ij}) = \mu_{ij} + \phi_j \mu_{ij}^2$ , where $\phi_j > 0$ is a dispersion parameter
	ZIP	LA	log	$E(y_{ij}) = (1 - p_j)\mu_{ij}, P(y_{ij} = 0) = p_j$ , $V(\mu_{ij}) = \mu_{ij}(1 - p_j)(1 + \mu_{ij}p_j)$
Binary	Bernoulli	VA/LA	probit	$E(y_{ij}) = \mu_{ij}, V(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$
		LA	logit	
Biomass	Tweedie	LA	log	$E(y_{ij}) = \mu_{ij}, V(\mu_{ij}) = \phi_j \mu_{ij}^\nu$ , where $1 < \nu < 2$ is a power parameter and $\phi_j > 0$ is a dispersion parameter
Ordinal	Multinomial	VA	probit	Cumulative probit model
Normal	Gaussian	VA/LA	identity	$E(y_{ij}) = \mu_{ij}, V(y_{ij}) = \phi_j^2$

Table 2: Overview of functions available in `gllvm`.

Function	Description
<code>gllvm()</code>	Fits a generalized linear latent variable model
<code>anova.gllvm()</code>	Analysis of deviance for ‘ <code>gllvm</code> ’ objects
<code>coefplot.gllvm()</code>	Plots covariate coefficients and confidence intervals
<code>logLik.gllvm()</code>	Log-likelihood of an object of class ‘ <code>gllvm</code> ’
<code>residuals.gllvm()</code>	Dunn-Smyth residuals for ‘ <code>gllvm</code> ’ model
<code>summary.gllvm()</code>	Summarizing ‘ <code>gllvm</code> ’ model fits
<code>ordiplot.gllvm()</code>	Plots latent variables from a ‘ <code>gllvm</code> ’ model
<code>plot.gllvm()</code>	Plots diagnostics for a ‘ <code>gllvm</code> ’ object
<code>confint.gllvm()</code>	Confidence intervals for ‘ <code>gllvm</code> ’ model parameters
<code>predict.gllvm()</code>	Obtains predictions from a ‘ <code>gllvm</code> ’ model
<code>getResidualCov.gllvm()</code>	Calculates residual covariance matrix for a ‘ <code>gllvm</code> ’ fit
<code>getResidualCor.gllvm()</code>	Calculates residual correlations for a ‘ <code>gllvm</code> ’ fit
<code>getPredictErr.gllvm()</code>	Prediction errors for predicted latent variables
<code>simulate.gllvm()</code>	Generate new data based on a ‘ <code>gllvm</code> ’ fit

Table 3: Computation times in seconds (on a Intel Core i7-3770 (3.4GHz)) to fit the example GLLVM objects of this paper using `gllvm` and `boral` (with default settings) using. The `gllvm` reduces computation times from minutes to seconds for each example.

	<code>fit_ord</code>	<code>fit_env</code>	<code>fit_4th</code>
<code>gllvm</code>	4.0	10.0	10.3
<code>boral</code>	595.4	1483.6	1529.9

481 **Figures**

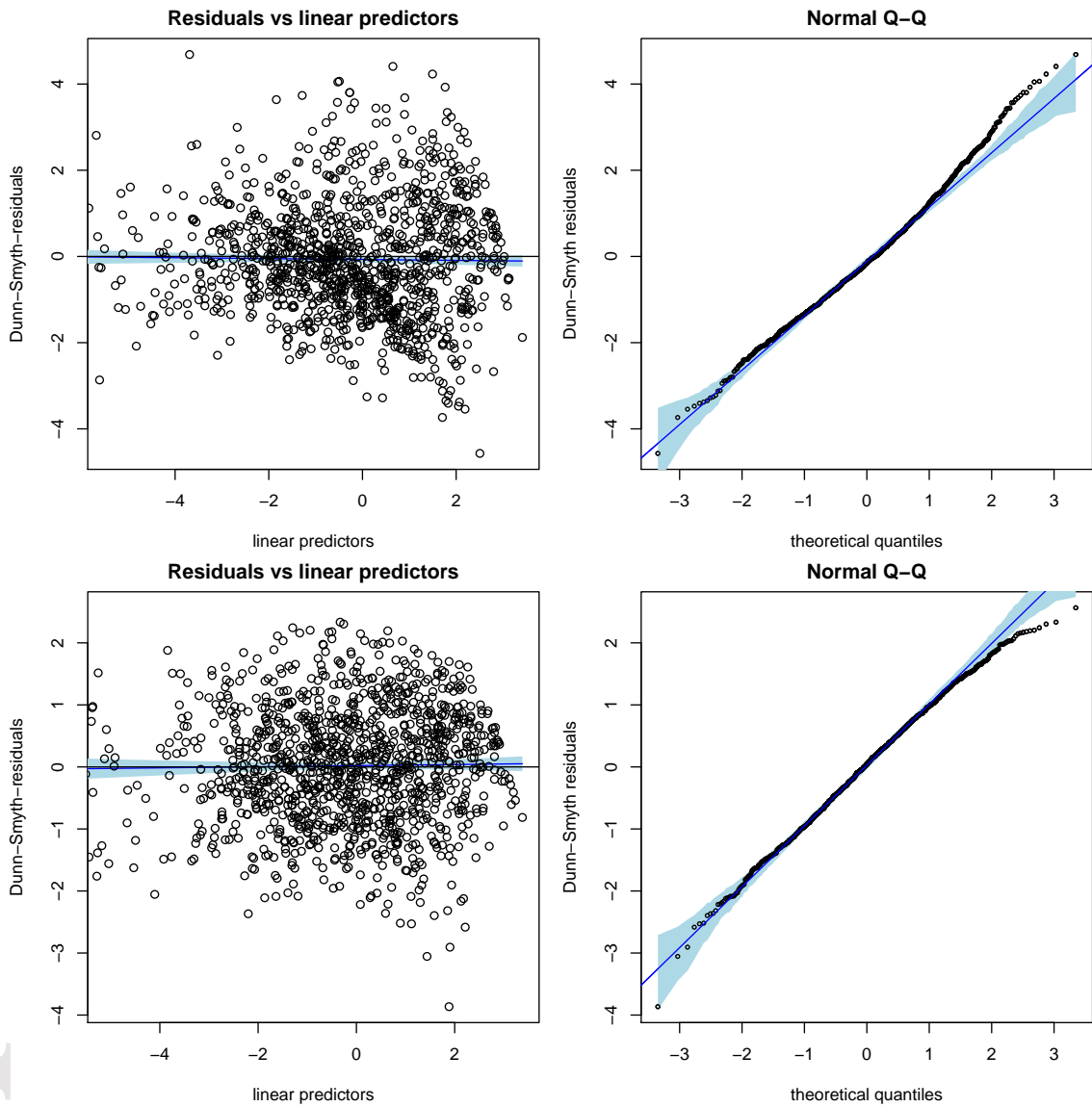


Figure 1: Residual plots for the Poisson GLLVM (top) and the NB-GLLVM (bottom) applied for model-based ordination. Specifically, Dunn-Smyth residuals are plotted against linear predictors (left), while simulated point-wise 95% confidence interval envelope are added in the normal quantile-quantile plot (right). The fan shape and unusually large residuals for the Poisson GLLVM suggest data are slightly overdispersed compared to the Poisson distribution. The lack of pattern and smaller residuals for the NB-GLLVM, suggests a better model fit to the data.

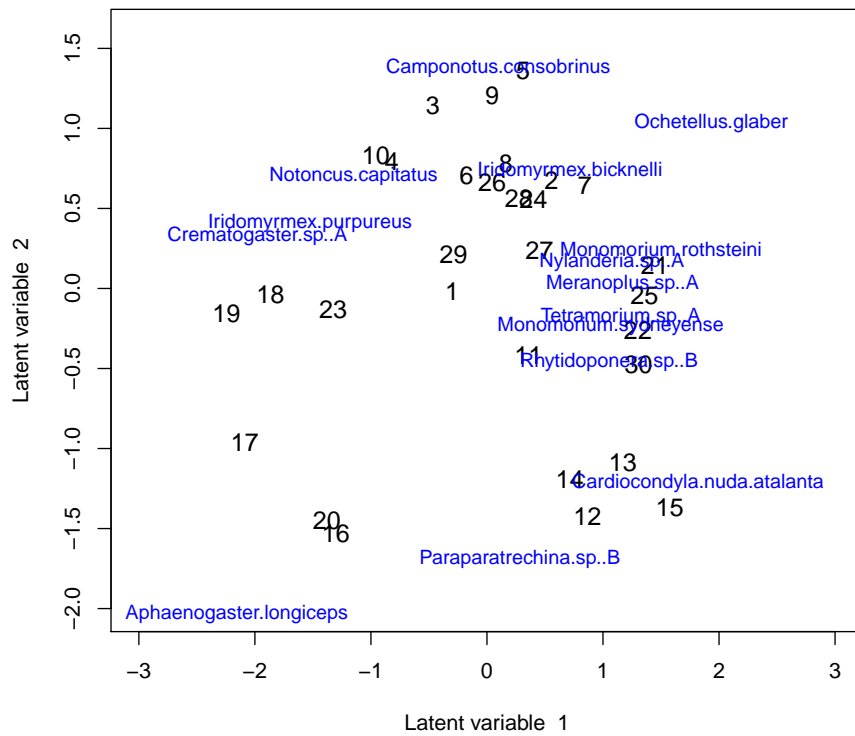


Figure 2: A biplot with 15 indicator species based on the NB-GLLVM fitted to the ant data. The numbers correspond to the site indices.

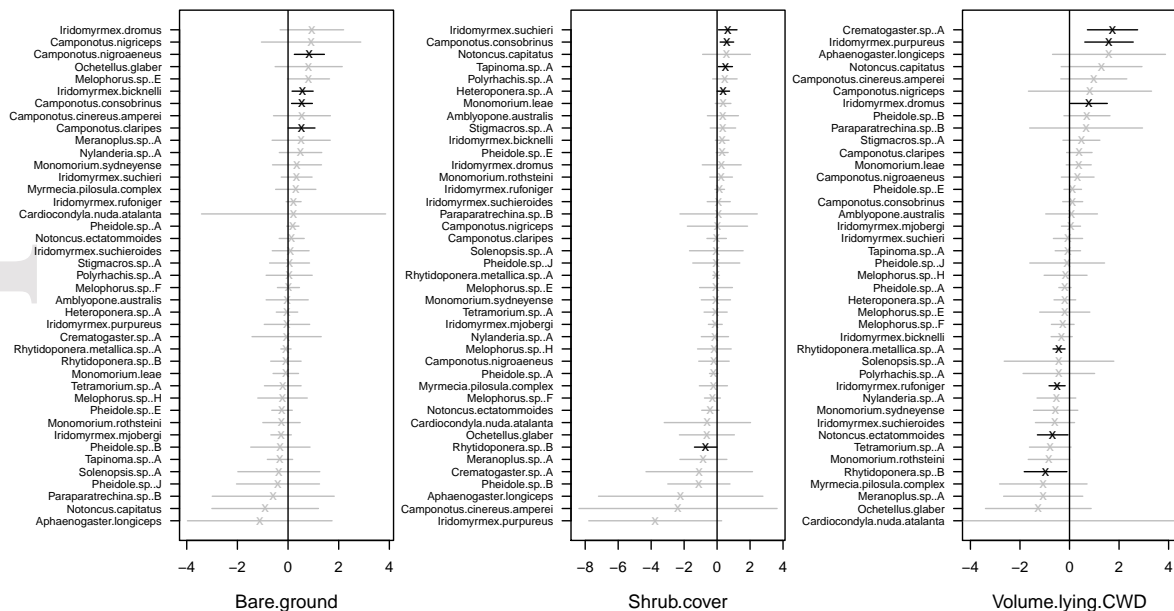


Figure 3: Plots of the point estimates (ticks) for coefficients of the environmental variables and their 95% confidence intervals (lines) for the NB-GLLVM, with those colored in grey (black) denoting intervals (not) containing zero. The x-axis of the coefficient plot of the third variable is truncated due to very wide confidence interval for one of the coefficients.

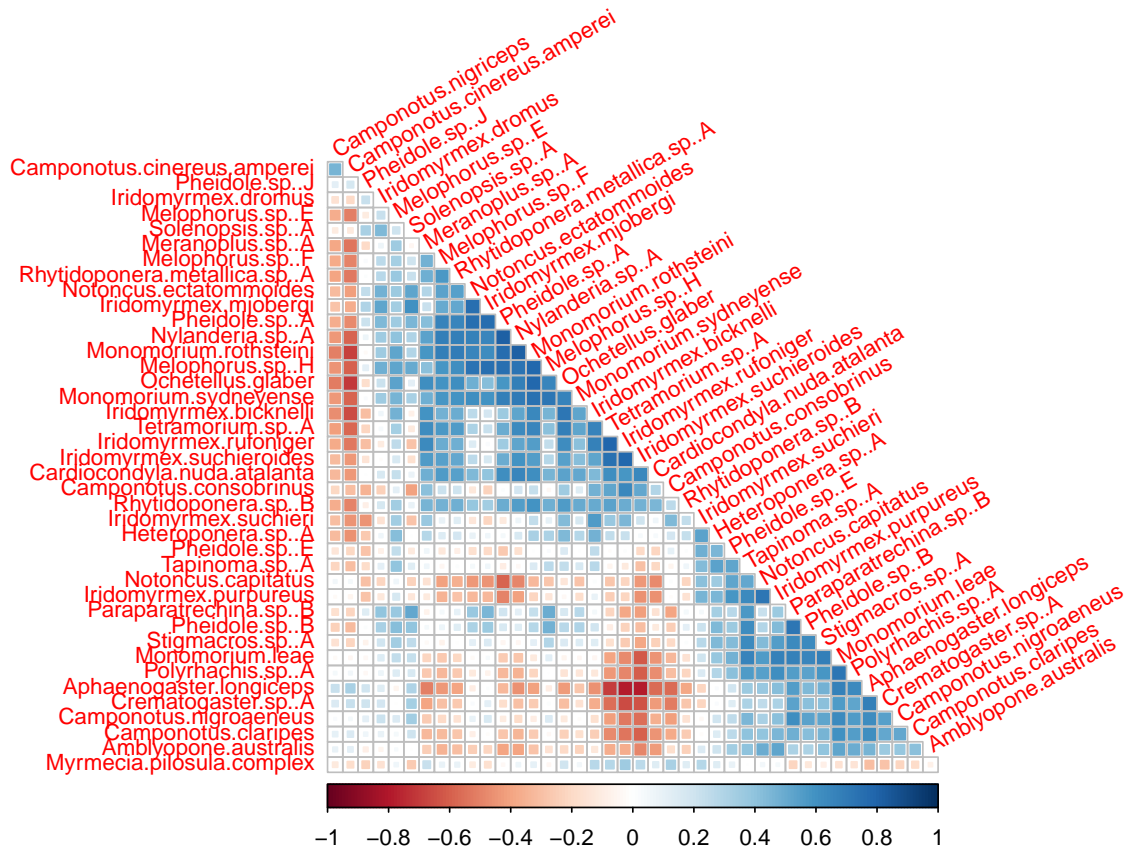


Figure 4: Residual correlation matrix based on latent factor loadings for the NB-GLLVM with environmental covariates.

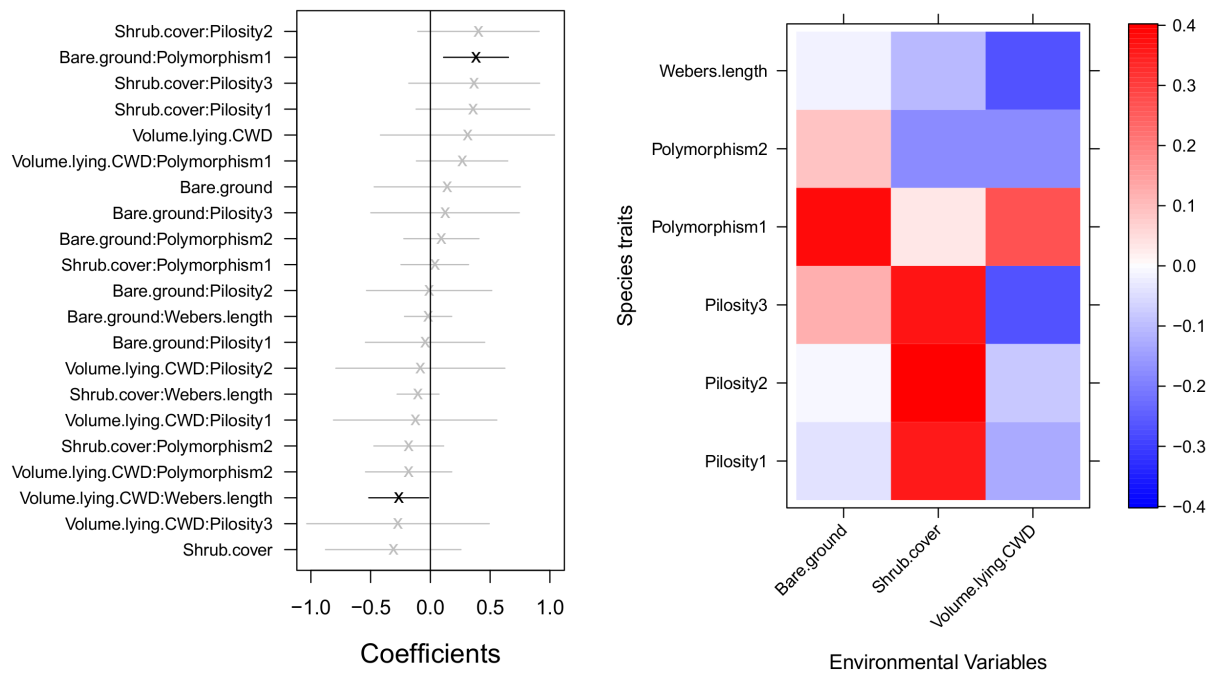


Figure 5: A plot of the estimated coefficients (ticks) and their 95% confidence intervals (lines) for all terms in the fourth corner model (left), and a level plot for the fourth corner interaction terms (right) in the NB-GLLVM. The colors offer an indication of the signs and magnitudes of the point estimates.