Methods in Ecology and Evolution

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

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

*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

Running Header - gllvm R package

Word count: 3500 words

Summary

1. There has been rapid development in tools for multivariate analysis based on fully specified statistical models or "joint models". One approach attracting a lot of attention is generalized linear latent variable models (GLLVMs). However, software for fitting these models is typically slow and not practical for large datsets.
2. The R package gllvm offers relatively fast methods to fit GLLVMs via maximum likelihood, along with tools for model checking, visualization and inference.
3. The main advantage of the package over other implementations is speed *e.g.* being two orders of magnitude faster, and capable of handling thousands of response variables. These advances come from using variational approximations to simplify the likelihood expression to be maximised, automatic differentiation software for model-fitting (via the TMB package), and careful choice of initial values for parameters.
4. Examples are used to illustrate the main features and functionality of the package, such as constrained or unconstrained ordination, including functional traits in "fourth corner" models, and (if the number of environmental coefficients is not large) make inferences about environmental associations.

Keywords: High-dimensional data, joint modelling, multivariate analysis, ordination, species interactions

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

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

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

Software for fitting GLLVMs in ecology is currently quite slow computationally and not practical for large datasets. In particular, packages in the freely available software `R` 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 jth species at the ith 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 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})^\top$. A GLLVM regresses the mean abundance μ_{ij} against environmental variables and a vector of $d \ll m$ latent variables, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})^\top$:

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

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 $\boldsymbol{u}_i$ 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 $\boldsymbol{x}_i$). We assume these latent variables are independent across sites and standard normally distributed. The parameters β_{0j} are species specific intercepts, while α_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 α_i can be included for site total

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

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

If q trait covariates $\boldsymbol{t}_j = (t_{i1}, \dots, t_{iq})^\top$ are also recorded, we can use them to help explain inter-specific variation in environmental response. This leads to an extension of the so-called "fourth corner model" (Jamil and ter Braak, 2013; Brown et al., 2014) where multivariate abundance is regressed against a function of traits and environment, and the environment-trait interactions represents the fourth corner association between traits and environment. The associated fourth corner GLLVM then has mean model:

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

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

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

Estimation

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

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

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

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

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

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

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

Using the R package gllvm

The R package gllvm provides a flexible implementation for fitting GLLVMs to multivariate data. The 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, data = NULL, formula = NULL,
  num.lv = 2, family, method = "VA", row.eff = FALSE, offset = NULL,
  Power = 1.5, starting.val = "res", ...)
```

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

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

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

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

Model-based ordination

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

```
> fitp <- gllvm(y, family = poisson())
> fitp
Call:
gllvm(y = y, family = poisson())
family:
[1] "poisson"
...
AIC:  4501.263
AICc:  4178.553
BIC:  4672.209
> fit_ord <- gllvm(y, family = "negative.binomial")
> fit_ord
Call:
gllvm(y = y, family = "negative.binomial")
family:
[1] "negative.binomial"
...
AIC:  4116.173
AICc:  3717.188
BIC:  4344.568
```

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

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.

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, $||\boldsymbol{\gamma}_j||$), 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.

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

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 diagonsing model fit and performing model selection.

Model with environmental variables

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.

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

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.

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

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

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

Studying co-occurrence patterns

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

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

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

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

Incorporating functional traits into "fourth corner" models

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

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

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

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

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

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

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

Summary

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

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

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

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

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

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

Acknowledgments

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

Supporting Information

Appendix 1: R code for examples

Appendix 2: Analysis of high-dimensional microbial data

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

Authors contributions

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

Data accessibility

The ant dataset used in our examples is publicly available from the `R` package `mvabund` (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 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 well as all code used in this paper and supplementary materials is publicly available in the `R` package `gllvm` (Niku et al., 2019a) in the CRAN: `https://cran.r-project.org/web/packages/gllvm/`.

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.

Lammel, D.R., Barth, G., Ovaskainen, O., Cruz, L.M., Zanatta, J.A., Ryo, M., de Souza, E.M., and Pedrosa, F.O. (2018). Direct and indirect effects of a pH gradient bring insights into the mechanisms driving prokaryotic community structures. *Microbiome*, 6:6–106.

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., and Warton, D.I. (2019a). *gllvm: Generalized Linear Latent Variable Models*. R package version 1.1.7. https://cran.r-project.org/web/packages/gllvm/

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., and Warton, D.I. (2019b). Efficient estimation of generalized linear latent variable models. *PLoS One*, 14(5):1–20.

Niku, J., Warton, D.I., Hui, F.K.C., and Taskinen, S. (2017). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 22:498–522.

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. and Wagner, H. (2018). vegan: Community Ecology Package. R package version 2.5-3. https://CRAN.R-project.org/package=vegan

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution*, 7:549–555.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91:2514–2521.

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, 20:561–576.

Pollock, L.J., Tingley, R., Morris, W.K., Golding, N., O'Hara, R.B., Parris, K.M., Vesk, P.A., and McCarthy, M.A. (2014). Understanding co-occurrence by modelling species simultaneously with a joint species distribution model (JSDM). *Methods in Ecology and Evolution*, 5:397–406.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, 2:1–21.

Royan, A., Reynolds, S.J., Hannah, D.M., Prudhomme, C., Noble, D.G., and Sadler, J.P. (2016). Shared environmental responses drive co-occurrence patterns in river bird communities. *Ecography*, 39:733–742.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Chapman & Hall, Boca Raton.

ter Braak, C.J.F. (2019) New robust weighted averaging- and model-based methods for assessing trait-environment relationships. *Methods in Ecology and Evolution*, In press.

Tikhonov, G., Opedal, Ø., Abrego, N., Lehikoinen, A., and Ovaskainen, O. (2019). Joint species distribution modelling with `HMSC-R`. *bioRxiv* preprint, https://doi.org/10.1101/603217.

Walker, S.C. and Jackson, D.A. (2011). Random-effects ordination: Describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, 81:635–663.

Wang, Y., Naumann, U., Wright, S.T., and Warton, D.I. (2012). mvabund - an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, 3:471–474.

Warton, D.I., Blanchet, F.G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S.C., and Hui, F.K.C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology and Evolution*, 30:766–779.

Warton, D. I. and Hui, F. K. (2017). The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017). *Methods in Ecology and Evolution*, 8:1408–1414.

Warton, D.I., Wright, S.T. and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, 3:89–101.

Tables

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

	`fit_ord`	`fit_env`	`fit_4th`
`gllvm`	4.0	10.0	10.3
`boral`	595.4	1483.6	1529.9

Figures

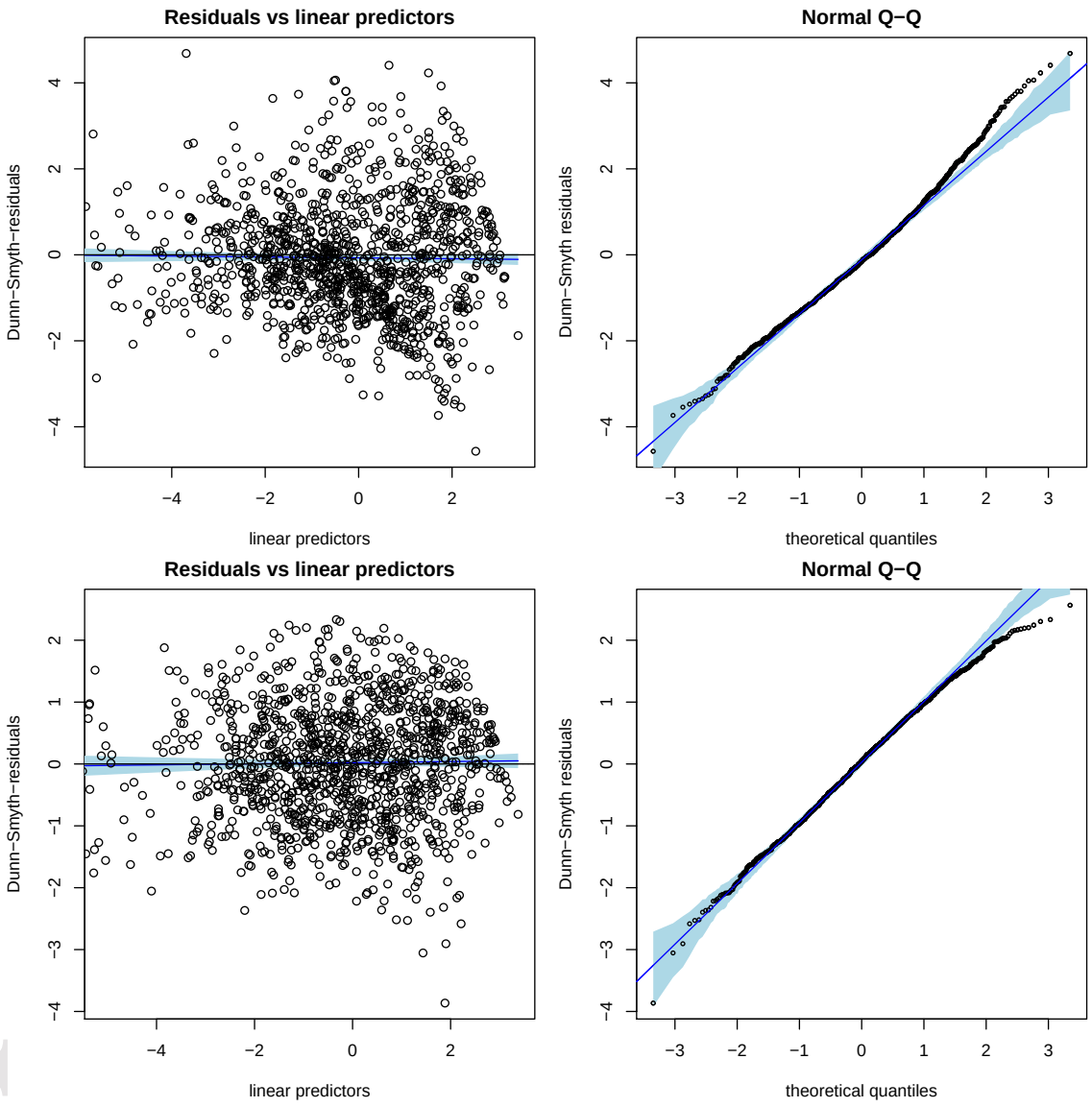

Figure 1: Residual plots for the Poisson GLLVM (top) and the NB-GLLVM (bottom) applied for model-based ordination. Specifically, Dunn-Smyth residuals are plotted against linear predictors (left), while simulated point-wise 95% confidence interval envelope are added in the normal quantile-quantile plot (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.

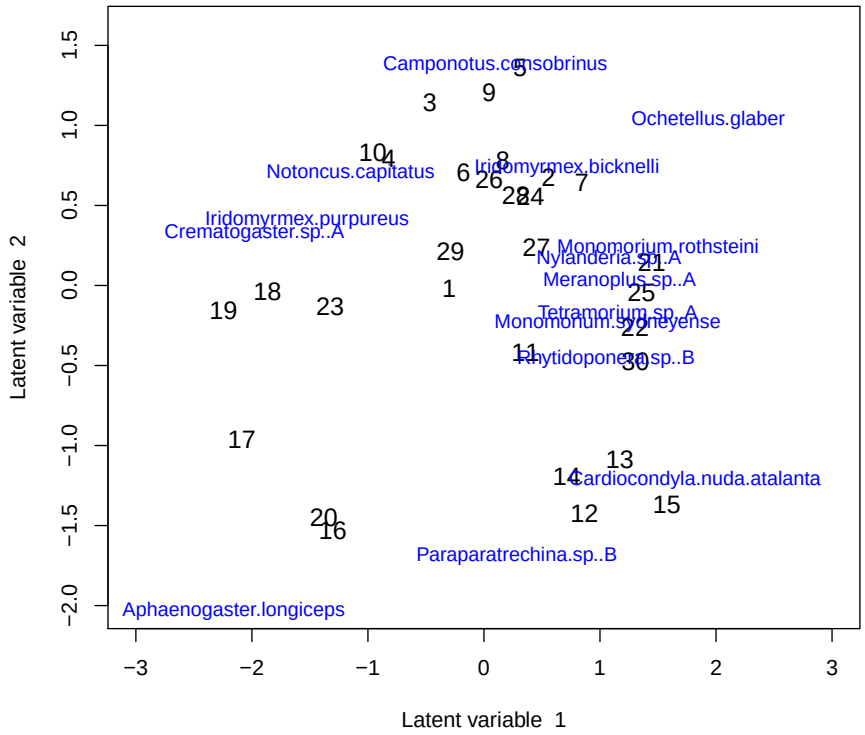

Figure 2: A biplot with 15 indicator species based on the NB-GLLVM fitted to the ant data. The numbers correspond to the site indices.

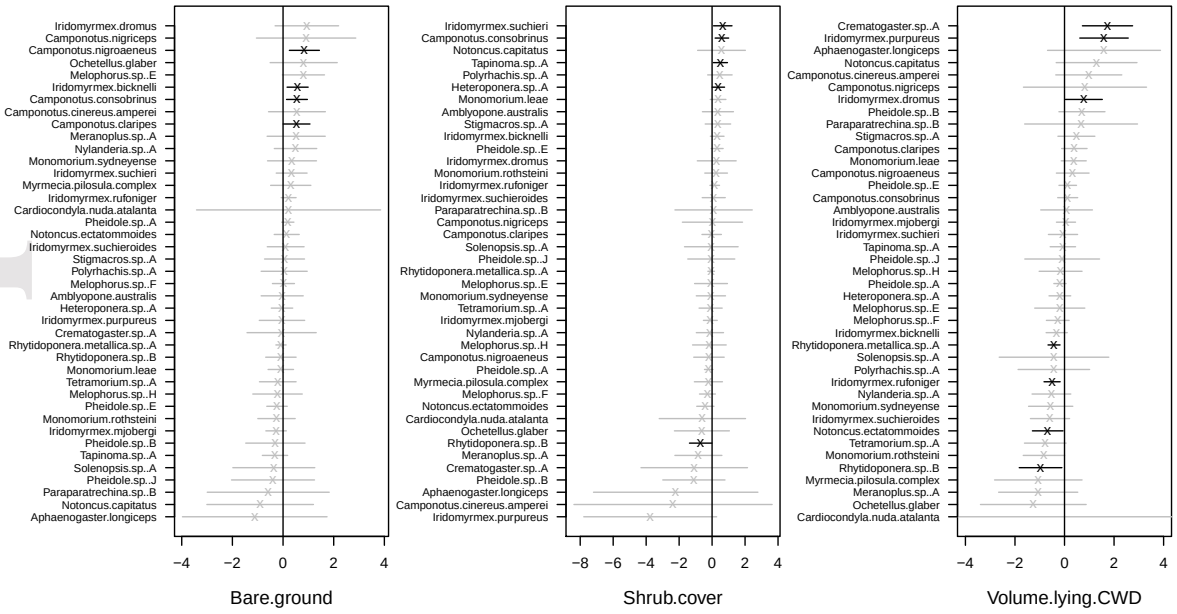

Figure 3: Plots of the point estimates (ticks) for coefficients of the environmental variables and their 95% confidence intervals (lines) for the NB-GLLVM, with those colored in grey (black) denoting intervals (not) containing zero. The x-axis of the coefficient plot of the third variable is truncated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

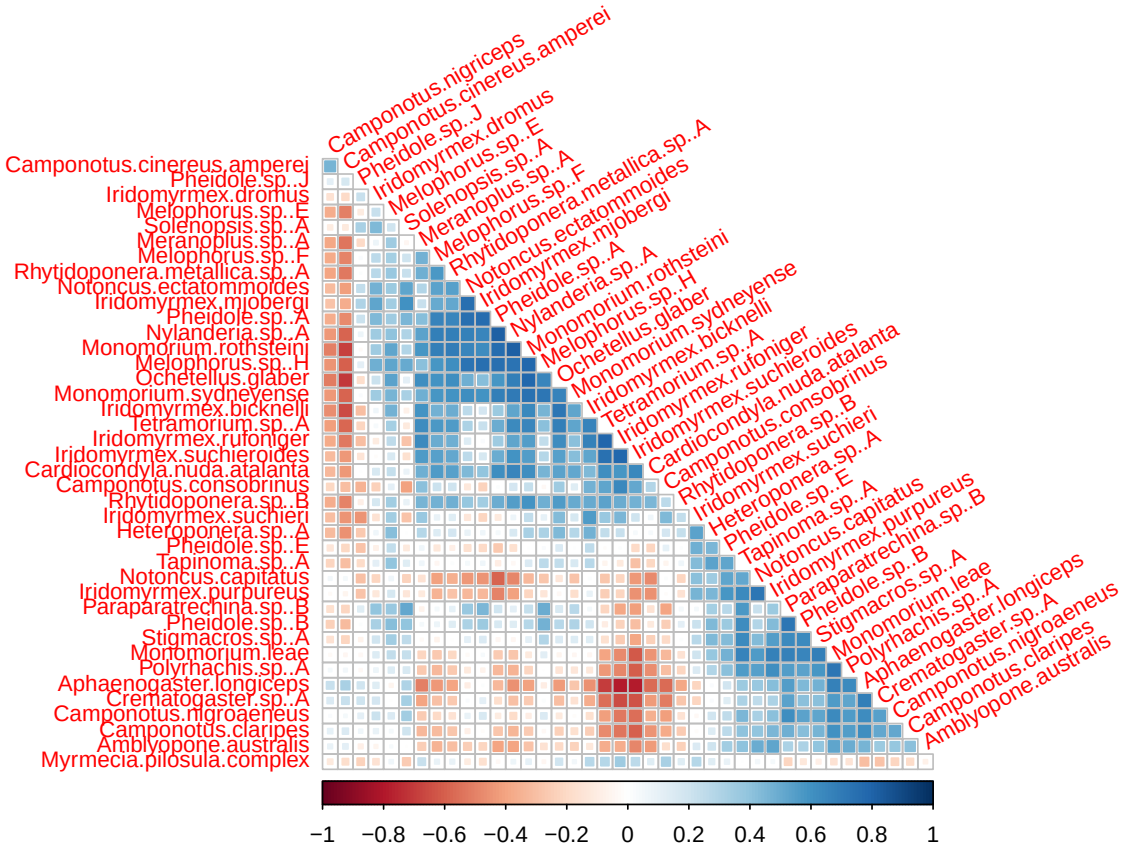

Figure 4: Residual correlation matrix based on latent factor loadings for the NB-GLLVM with environmental covariates.

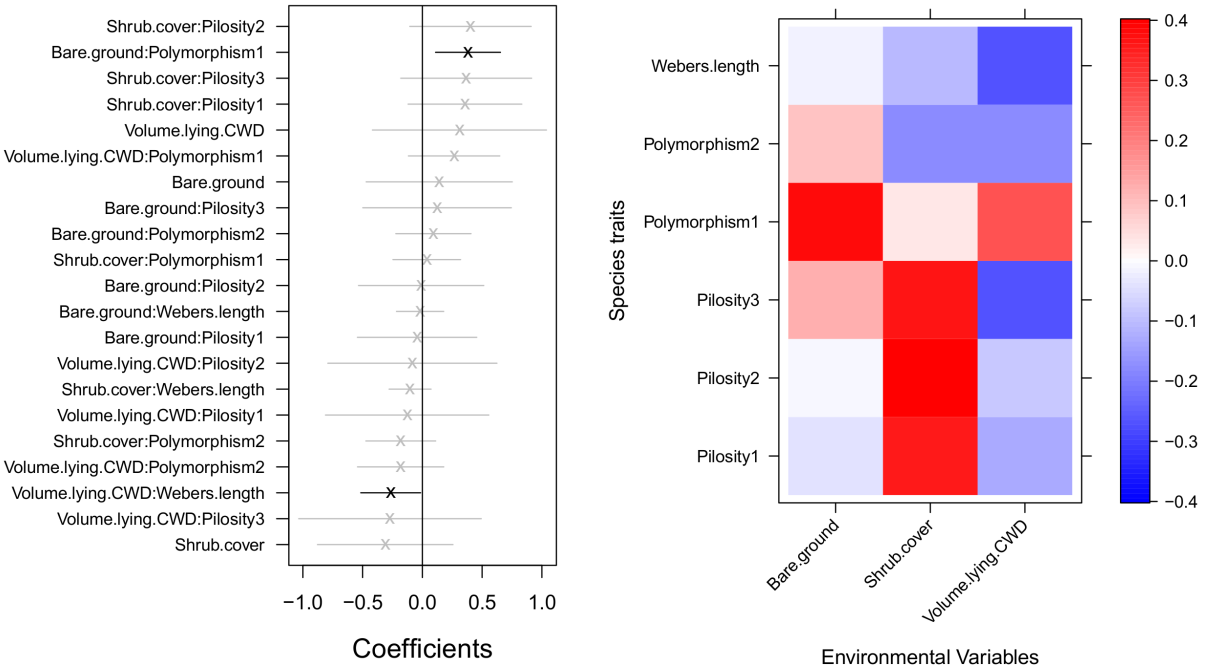

Figure 5: A plot of the estimated coefficients (ticks) and their 95% confidence intervals (lines) for all terms in the fourth corner model (left), and a level plot for the fourth corner interaction terms (right) in the NB-GLLVM. The colors offer an indication of the signs and magnitudes of the point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