Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology

Jenni Niku*[1], David I. Warton[2,3], Francis K.C. Hui[4], and Sara Taskinen[1]

[1]Department of Mathematics and Statistics, University of Jyväskylä, Finland

[2]School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, Sydney, Australia

[3]School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia

[4]Mathematical Sciences Institute, The Australian National University, Australia

Abstract

In this paper we consider generalized linear latent variable models that can handle overdispersed counts and continuous but non-negative data. Such data are common in ecological studies when modelling multivariate abundances or biomass. By extending the standard generalized linear modelling framework to include latent variables, we can account for any covariation between species not accounted for by the predictors, notably species interactions and correlations driven by missing covariates. We show how estimation and inference for the considered models can be performed efficiently using the Laplace approximation method, and use simulations to study the finite-sample properties of the resulting estimates. In the overdispersed count data case, the Laplace approximated estimates perform similarly to the estimates based on variational approximation method, which is another method that provides a closed form approximation of the likelihood. In the biomass data case, we show that ignoring the correlation between taxa affects the regression estimates unfavourably. To illustrate how our methods can be used in unconstrained ordination and in making inference on environmental variables, we apply them to two ecological datasets: abundances of bacterial species in three arctic locations in Europe and abundances of coral reef species in Indonesia.

Keywords: biomass; Laplace approximation; ordination; overdispersed count; species interactions

*Corresponding author: Jenni Niku, *email:* jenni.m.e.niku@jyu.fi

1 Introduction

In many studies in community ecology, multivariate abundance data are often collected, comprising the records of a large number of interacting species at a set of observational units or sites. Such data are characterized by two main features. First, the data are high-dimensional in that the number of species, many of which may interact, is often close to or exceeding the number of sites. Second the data almost always are not or cannot be suitably transformed to be normally distributed. Instead, the most common types of responses recorded include presence-absence records, overdispersed species counts, biomass (non-negative, continuous data often with large number of zeros, representing the total mass of a species found at a site), and heavily discretized percent cover data.

As a motivating example, we consider data on diversity of plant-associated bacteria (Nissinen et al., 2012). The data consists of counts of 1276 interacting bacteria species measured from different habitats (bulk soil) in 56 sites across three locations. The study design is explained in Section 5.1 in detail. This example, which is by no means an extreme case, exhibit both of the above characteristics, with the number of species approximately 23 times that of the number of sites, and the counts being highly overdispersed with nearly half of the species present at ten or fewer sites.

Multivariate abundance data are often collected to answer a number of key questions concerning the species community. In our motivating dataset for instance, Nissinen et al. (2012) were interested in performing an ordination to visualize whether sites are similar in terms of their species composition, which could be helpful in planning future sampling designs as well as identifying the drivers of microbial community composition such as soil physiochemical properties. They were also interested in conducting multivariate inference on the associations between climate zone, environment and soil microflora on microbial communities associated with plant or with particular plant species. Such analyses have important implications to help in interpreting drivers of biological associations (bacteria-plant) as well as abiotic factors (Männistö et al., 2007; Chu et al., 2010). A model-based analysis of such data poses some major challenges not just due to the high-dimensionality and non-normality of the data, as previously discussed, but also because of the (potentially) complex between species interactions. Analogous to longitudinal data, while the observational units (sites) are often independent by design, we cannot assume that species within a unit are independent: species responses are likely to be correlated due to a host of ecological reasons, such as biotic interactions, phylogeny and missing covariates (Araújo and Luoto, 2007; Morales-Castilla et al., 2015). Ignoring the correlation between species responses may result in inflated Type I errors and too narrow confidence intervals when assessing the significance of one or more predictors in the model, and too narrow prediction intervals when extrapolating key community quantities such as species richness into new sites and/or under various

climate scenarios (Warton et al., 2015, 2016).

Over the past few years, the above challenges have spurred a variety of work into model-based joint analysis of multivariate abundance data. One promising approach, as reviewed by Warton et al. (2015), is generalized linear latent variable models (GLLVMs, Moustaki and Knott, 2000). This rich class of models extend the basic generalized linear model framework by including one or more latent variables, with corresponding factor loadings, as a parsimonious method of modeling any residual correlation between species not accounted by the covariates. Warton et al. (2015) showed how GLLVMs overcome the challenges discussed above to offer a viable approach for analyzing multivariate abundance data. Specifically, by using a factor analytic type approach based on rank reduction to model the high-dimensional between species covariance matrix, GLLVMs offer a viable method of constructing model-based (residual) ordination and biplots, as well as conducting multivariate inference such as hypothesis testing of environmental and/or treatment effects, environment-by-trait interactions, and how species interactions vary at different spatial and temporal scales; see Letten et al. (2015) and Ovaskainen et al. (2016a) for recent applications of GLLVMs to multivariate abundance data.

While a promising approach, one of the major and outstanding challenges with using GLLVMs is computationally efficient estimation and inference. Since the responses are not normally distributed, the marginal likelihood, which involves integrating out the unknown latent variables, does not possess a closed form. This problem in general has attracted much attention in the statistical literature, and below we review several of the well-known methods proposed to overcome this issue. In Moustaki (1996) and Moustaki and Knott (2000), GLLVMs for mixtures of binary and normal responses were fitted using Gauss-Hermite quadrature. This was expanded upon by Rabe-Hesketh et al. (2002), who proposed adaptive Gauss-Hermite quadrature to fit GLLVMs, allowing for normal, binomial, gamma, and Poisson distributed responses. While quadrature in general works well for simple latent variable models, the method scales poorly with the number of latent variables, and becomes computationally impractical if the number of latent variables is moderate e.g., exceeds two. Another drawback is that the method of Rabe-Hesketh et al. (2002) is only available in the proprietary software STATA. More recently, Hui et al. (2016) proposed a fast variational approximation method to approximate the likelihood in the case of binary, ordinal and overdispersed count data. While quick, the method is rather case specific, offering only a closed approximation for specific combinations of response distributions and link functions. Furthermore, little is known about the theoretical properties of variational approximations as a framework e.g., the convergence rate and asymptotic normality of Gaussian variational approximation estimates has been derived in only specific cases such as Poisson mixed models with a random intercept (Hall et al., 2011a,b).

The most well-known approach for estimating GLLVMs is to apply an Expectation Maximization (EM)

algorithm or some variant of it, as in Sammel et al. (1997) and Hui et al. (2015). In the ecology literature however, with the growing popularity in hierarchical approaches to community level modeling (Cressie et al., 2009; Ovaskainen et al., 2016b), most of the applications of GLLVMs have instead employed Bayesian Markov Chain Monte Carlo estimation based on the complete likelihood function (Blanchet, 2014; Ovaskainen et al., 2016a; Hui, 2016). A major downside of both Markov Chain Monte Carlo and the EM algorithm estimation though is that they are computationally very intensive: the E-step in the EM algorithm (still) does not possess a closed form, and so some form of Monte-Carlo integration is still necessary.

Computational efficiency is a key requirement of methods of parameter estimation, given the sizes of datasets now encountered in practice in ecology. While historically most multivariate abundance datasets had a few hundred variables, modern lab-based sampling and classification techniques, such as metabarcoding in Yu et al. (2012) commonly result in datasets exceeding a thousand response variables, as in our microbial application. As such, the most feasible maximum likelihood approaches for fitting GLLVMs in the foreseeable future are those that approximate the marginal likelihood as a closed form, in particular, a variational approximation (where applicable), or as in this paper, a Laplace approximation.

In this paper, we propose estimating and performing inference with GLLVMs using the Laplace approximation for overdispersed count and biomass data, motivated by multivariate abundance data in ecology. Although the Laplace method is a special case of adaptive Gauss-Hermite quadrature with only one quadrature point, one of the major advantages of the Laplace approximation is that it provides a general but fully closed form approximation of the likelihood, which can be maximized efficiently even for very complex models applied to high-dimensional data such as overdispersed species counts in our motivating example. This article is not the first to propose the Laplace approximation for GLLVMs, but the key innovation is our extension particularly to handle overdispersed counts and biomass data in ecology. Huber et al. (2004) previously provided a Laplace approximation of the likelihood function in the general exponential family case, with mixtures of binomial and normal responses serving as examples. This was extended by Bianconcini and Cagnone (2012), who proposed a fully exponential Laplace approximation method for fitting GLLVMs. They also treated the general exponential family case, but focused on ordinal data in simulation studies. This article differs from these previous works though in that we are motivated specifically by multivariate abundance data in ecology, and provides the first Laplace approximated likelihood forms for response distributions appropriate for overdispersed count and biomass data. More precisely, we derive forms in the case of negative binomial or zero-inflated Poisson distributions for overdispersed counts and the Tweedie distribution for biomass data. To our knowledge, the Laplace approximation method has not been formally considered for any of these distributions so far. Notice that the two other important response types in ecology, that is, presence-absence records and heavily discretized percent cover data, can be handled with the

tools provided by Huber et al. (2004) for binary responses and Bianconcini and Cagnone (2012) for ordinal responses, respectively.

The paper is organized as follows. In Section 2, we formulate the generalized linear latent variable model framework and response distributions of interest for multivariate abundance data. In Section 3, Laplace approximations of the likelihood functions are derived, and estimation and inference based on these are discussed. Section 4 provides a simulation study to compare the performance of Laplace approximation estimates to variational approximation estimates in the case of overdispersed count data. In the case of biomass data, we empirically illustrate the detrimental effect of ignoring the correlation inherent in the responses on parameter estimates. Finally, Section 5 applies the proposed Laplace approximated GLLVMs to the microbial community data (Nissinen et al., 2012) and coral community data (Warwick et al., 1990), in both cases demonstrating how common aspects of inference such as ordination can be performed within a model-based framework via the Laplace approximation.

2 GLLVMs for Multivariate Abundance Data

Let $\boldsymbol{Y}$ denote a $n \times m$ response matrix, where rows $i = 1, \ldots, n$ are observational units (sites) and columns $j = 1, \ldots, m$ consist of m-variate correlated responses (species). For each site $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$, a k-vector of environmental covariates, denoted here as $\boldsymbol{x}_i$, may also be recorded.

In GLLVMs, the mean response $\mu_{ij} = E(y_{ij})$ is regressed against a vector of $d \ll m$ latent variables, denoted as $\boldsymbol{u}_i$, along with the vector of k covariates if available. That is,

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

where $g(\cdot)$ is a known link function, and α_i are β_{0j} denote row effects and species-specific intercepts respectively. While optional, row and column effects may be included to account for differences in site and species total abundance. For example, a row effect is included to ensure that the latent variables quantify differences in species composition only, as opposed to species abundance (a combination of composition and site total abundance; see Hui et al., 2015, for more details). The vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ denote species-specific regression coefficients and loadings, that is, coefficients related to the covariates and latent variables, respectively.

In model (1), the term $\boldsymbol{u}_i'\boldsymbol{\gamma}_j$ captures any residual correlation across species not accounted for by the observed covariates $\boldsymbol{x}_i$. We assume that the latent variables are drawn from independent, standard normal distributions, $\boldsymbol{u}_i \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes a $d \times d$ identity matrix. The purpose of the zero mean and unit variance assumption is to fix the locations and scales of the latent variables (see Chapter 5, Skrondal

and Rabe-Hesketh, 2004). Also, to avoid rotation invariance and ensure parameter identifiability, we set all the upper triangular elements of $m \times d$ matrix $\mathbf{\Gamma} = (\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_m)'$ to zero, and constrain its diagonal elements to be positive (Huber et al., 2004). It is important to emphasize that these constraints do not limit the flexibility of the GLLVM to model between species correlation: there are no restrictions on the form of the residual covariance matrix induced by (1), namely $\mathbf{\Sigma}_{\text{res}} = \mathbf{\Gamma}\mathbf{\Gamma}'$, aside from it being of reduced rank d.

We now study specific cases of GLLVMs of key relevance to multivariate abundance data in ecology, namely, overdispersed species counts and biomass (a continuous, non-negative value typically obtained as total mass of a species at a site).

2.1 Species Counts

Species counts are often overdispersed due to their clustered nature i.e., species tend to be found in large numbers or not at all. A standard approach is to assume a negative binomial distribution for the response, $y_{ij} \sim \text{NegBin}(\mu_{ij}, \phi_j)$, where ϕ_j is a species-specific dispersion parameter, and choose $g(\cdot)$ to be the log link function. The probability density function is given by

$$f(y_{ij}|\boldsymbol{u}_i, \mathbf{\Psi}) = \frac{\Gamma(y_{ij} + 1/\phi_j)}{y_{ij}!\Gamma(1/\phi_j)} \left(\frac{\mu_{ij}}{1/\phi_j + \mu_{ij}}\right)^{y_{ij}} \left(\frac{1}{1 + \mu_{ij}\phi_j}\right)^{1/\phi_j}, \tag{2}$$

such that $E(y_{ij}) = \mu_{ij}$ and the quadratic mean-variance relationship $V(\mu_{ij}) = \mu_{ij} + \mu_{ij}^2\phi_j$. When $\phi_j \to 0$, the response variable approaches the Poisson distribution.

The negative binomial distribution is often appropriate when the zeros (species absences) in the data can be explained via the same environmental filtering mechanism as the non-zero counts (Warton, 2005). But if the ecological process governing most species absences is believed to be independent of the mechanism driving the non-zero counts, then a more appropriate and common choice is a zero-inflated Poisson (ZIP) model (Welsh et al., 1996; Martin et al., 2005). A ZIP model assumes that responses are either structural zeros obtained with probability p or Poisson distributed count values obtained with probability $1 - p$. If $y_{ij} \sim ZIP(p_j, \mu_{ij})$, the probability distribution function is

$$f(y_{ij}|\boldsymbol{u}_i, \mathbf{\Psi}) = \begin{cases} p_j + (1 - p_j)\exp(-\mu_{ij}), & \text{if } y_{ij} = 0, \\ (1 - p_j)\exp(-\mu_{ij})\mu_{ij}^{y_{ij}}/y_{ij}!, & \text{if } y_{ij} > 0. \end{cases}, \tag{3}$$

where μ_{ij} is modelled as in (1) with log link function. Here we assume the probability of extra zeros is modelled for each species separately and without reference to the covariates. Under the ZIP model, $E(y_{ij}) = \mu_{ij}(1 - p_j)$ and $Var(y_{ij}) = E(y_{ij})(1 + p_j\mu_{ij})$. When $p_j = 0$, the ZIP model reduces to the Poisson model. Finally, notice the negative binomial distribution could also be extended to account for

extra zeros (e.g., Welsh et al., 1996). Zero-inflated negative binomial models however can often fit poorly to overdispersed count data and can suffer from convergence problems (Warton, 2005; Rodrigues-Motta et al., 2013), and so we do not pursue such a model in this article.

2.2 Biomass Data

For biomass data, which take continuous but non-negative values, an often appropriate assumption is the Tweedie distribution (Jorgensen, 1997). For a comprehensive discussion on Tweedie models and their suitability for biomass data, see Foster and Bravington (2013). If y_{ij} follows a Tweedie distribution, then $E(y_{ij}) = \mu_{ij}$ and $Var(y_{ij}) = \phi_j \mu_{ij}^{\nu}$, where ϕ_j is a species-specific dispersion parameter and ν is a power parameter controlling the shape of the distribution. The mean-variance relationship is thus explicitly defined by Taylor's power law (Taylor, 1961), which empirically arises under a range of ecological processes (Kendal, 2004).

The Tweedie distribution does not possess an explicit analytic form, but the density function can be evaluated numerically. For a typical power parameter value, $1 < \nu < 2$, a Tweedie random variable follows a compound Poisson distribution, and the probability distribution function can be written as

$$f(y_{ij}; \boldsymbol{u}_i, \boldsymbol{\Psi}) = \begin{cases} \exp\left(-\frac{\mu_{ij}^{2-\nu}}{\phi_j(2-\nu)}\right), & y = 0 \\ W(y_{ij}, \phi_j, \nu) \exp\left\{\left(\frac{y_{ij}\mu_{ij}^{1-\nu}}{1-\nu} - \frac{\mu_{ij}^{2-\nu}}{2-\nu}\right)/\phi_j\right\}/y_{ij}, & y > 0 \end{cases}, \tag{4}$$

where $W(y_{ij}, \phi_j, \nu) = \sum_{k=1}^{\infty} W_k$, and

$$W_k = \frac{y_{ij}^{-k\alpha}(\nu-1)^{\alpha k}}{\phi_j^{k(1-\alpha)}(2-\nu)^k k! \Gamma(-k\alpha)}$$

with $\alpha = (2-\nu)/(1-\nu)$. The function $W(y_{ij}, \phi_j, \nu)$ can be evaluated numerically using the method described in Dunn and Smyth (2005). Foster and Bravington (2013) and Dunstan et al. (2013) noted that a Tweedie distribution is equivalent to the distribution obtained by summing a Poisson number of gamma random variables. Such a parametrization makes it particularly suitable for example in analyzing marine data, e.g., the total weight of a fish species at a site can be considered as the sum of the individual fish weights, where the number of fish caught is given by a Poisson random variable and the weight of each fish follows a gamma distribution.

3 The Laplace approximation for GLLVMs

Consider again a $n \times m$ matrix, $\boldsymbol{Y}$, of observed responses and GLLVMs as defined in equation (1). Write $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})'$, $\boldsymbol{B} = (\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_m)'$ and $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 \ldots \boldsymbol{\gamma}_m)'$, and collect all the model parameters as a vector $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \text{vec}(\boldsymbol{B}), \text{vec}(\boldsymbol{\Gamma}), \boldsymbol{\Phi})$, where without loss of generality $\boldsymbol{\Phi}$ is used to denote any nuisance parameters depending on the assumed distribution, i.e., $\phi_1, \ldots, \phi_m$ for the negative binomial and Tweedie distributions and $p_1, \ldots, p_m$ for the ZIP distribution. Here vec$(\cdot)$ is the vectorizing operator, which stacks the columns of a matrix in a column vector. Conditionally on latent variables $\boldsymbol{u}_i$, the responses $y_{i1}, \ldots, y_{im}$ at site i are assumed to be independent, such that $f(\boldsymbol{y}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \prod_{j=1}^m f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})h(\boldsymbol{u}_i)$, where $h(\boldsymbol{u}_i) = N_d(\boldsymbol{0}, \boldsymbol{I}_d)$. The marginal distribution of $\boldsymbol{y}_i$ is obtained by integrating over the distribution of $\boldsymbol{u}_i$, leading to the log-likelihood function

$$l(\boldsymbol{\Psi}) = \sum_{i=1}^n \log\{f(\boldsymbol{y}_i, \boldsymbol{\Psi})\} = \sum_{i=1}^n \log\left(\int \prod_{j=1}^m f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})h(\boldsymbol{u}_i)\, d\boldsymbol{u}_i\right). \tag{5}$$

For the distributions discussed in Section 2, as well as for non-normally distributed responses in general, the marginal likelihood in (5) involves a d-dimensional integral, which cannot be solved analytically. We propose to overcome this by applying a Laplace approximation to $l(\boldsymbol{\Psi})$. The Laplace approximation for the log-likelihood in the case of the general exponential family is given in Huber et al. (2004), and is reviewed in the Appendix A. Here we focus on response types and distributions discussed in Section 2, which are frequently collected in ecology.

Consider first the negative binomial distribution which, for fixed dispersion parameters ϕ_j, is a member of the exponential family. Thus a Laplace approximation for the log-likelihood function can be derived directly from the general result of Huber et al. (2004).

Theorem 1. *The Laplace approximation $\tilde{l}$ of the log-likelihood function in negative binomial GLLVM in (2) is given by*

$$\begin{aligned}\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{i=1}^n \Bigg(&- \frac{1}{2}\log\det\{\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)\} + \sum_{j=1}^m \Bigg\{ y_{ij}\hat{\eta}_{ij} - \left(y_{ij} + \frac{1}{\phi_j}\right)\log\{1 + \phi_j \exp(\hat{\eta}_{ij})\} \\ &+ y_{ij}\log(\phi_j) + \log\Gamma\left(y_{ij} + \frac{1}{\phi_j}\right) - \log(y_{ij}!) - \log\Gamma\left(\frac{1}{\phi_j}\right)\Bigg\} - \frac{\hat{\boldsymbol{u}}_i'\hat{\boldsymbol{u}}_i}{2}\Bigg),\end{aligned}$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{j=1}^m \frac{(\phi_j y_{ij} + 1)\exp(\hat{\eta}_{ij})}{\{1 + \phi_j \exp(\hat{\eta}_{ij})\}^2}\boldsymbol{\gamma}_j\boldsymbol{\gamma}_j' + \boldsymbol{I}_d,$$

with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i{}'\boldsymbol{\gamma}_j$, *and* $\hat{\boldsymbol{u}}_i$ *is the maximum of*

$$Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = \sum_{j=1}^{m} \left\{ y_{ij}\eta_{ij} + y_{ij}\log(\phi_j) - \left(y_{ij} + \frac{1}{\phi_j}\right)\log\left\{1 + \phi_j \exp(\eta_{ij})\right\} + \log\Gamma\left(y_{ij} + \frac{1}{\phi_j}\right) - \log(y_{ij}!) - \log\Gamma\left(\frac{1}{\phi_j}\right)\right\} - \frac{\boldsymbol{u}_i'\boldsymbol{u}_i}{2}.$$

If the dispersion parameters ϕ_j are unknown as is usually the case, they can be estimated jointly with the other model parameters by maximizing $\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)$.

Next, for a ZIP model, the Laplace approximation of the log-likelihood function is given as follows. Note that this is not part of the exponential family and so we cannot directly use results from Huber et al. (2004).

Theorem 2. *The Laplace approximation* $\tilde{l}$ *of the log-likelihood function for the zero-inflated Poisson GLLVM in (3) is given by*

$$\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{i=1}^{n} \left(-\frac{1}{2}\log\det\left\{\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)\right\} + \sum_{j=1}^{m} \Big(\log(p_j + (1-p_j)\hat{A}_{ij}) I_{(y_{ij}=0)} + \left\{\log(1-p_j) - \exp(\hat{\eta}_{ij}) + y_{ij}\hat{\eta}_{ij} - \log(y_{ij}!)\right\} I_{(y_{ij}>0)} \Big) - \frac{\hat{\boldsymbol{u}}_i'\hat{\boldsymbol{u}}_i}{2} \right),$$

where $A_{ij} = \exp\{-\exp(\eta_{ij})\}$,

$$\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{j=1}^{m} \left(\exp(\hat{\eta}_{ij}) I_{(y_{ij}>0)} - \left(\frac{(1-p_j)\hat{A}_{ij}\exp(\hat{\eta}_{ij})(\exp(\hat{\eta}_{ij}) - 1)}{p_j + (1-p_j)\hat{A}_{ij}} - \frac{(1-p_j)^2\hat{A}_{ij}^2\exp(2\hat{\eta}_{ij})}{(p_j + (1-p_j)\hat{A}_{ij})^2} \right) I_{(y_{ij}=0)} \right) \boldsymbol{\gamma}_j\boldsymbol{\gamma}_j' + \boldsymbol{I}_d,$$

with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i{}'\boldsymbol{\gamma}_j$ *and* $\hat{A}_{ij} = \exp\{-\exp(\hat{\eta}_{ij})\}$, *and* $\hat{\boldsymbol{u}}_i$ *is the maximum of*

$$Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = \sum_{j=1}^{m} \left(\log(p_j + (1-p_j)A_{ij}) I_{(y_{ij}=0)} + \left\{\log(1-p_j) - \exp(\eta_{ij}) + y_{ij}\eta_{ij} - \log(y_{ij}!)\right\} I_{(y_{ij}>0)} \right) - \frac{\boldsymbol{u}_i'\boldsymbol{u}_i}{2}.$$

Finally for the Tweedie distribution, we have the following result.

Theorem 3. *A Laplace approximation* $\tilde{l}$ *of the log-likelihood function in Tweedie GLLVM in (4),*

is given by

$$\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log\det\left\{\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)\right\} + \sum_{j=1}^{m} \left[\left\{ \log \hat{W}(y_{ij}, \phi_j, \nu) - \log(y_{ij}) \right\} I_{(y_{ij}=0)} \right.\right.$$
$$\left.\left. + \frac{1}{\phi_j} \left(\frac{y_{ij} \exp\{(1-\nu)\hat{\eta}_{ij}\}}{1-\nu} - \frac{\exp\{(2-\nu)\hat{\eta}_{ij}\}}{2-\nu} \right) \right] - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right),$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{j=1}^{m} \frac{1}{\phi_j} \left[(2-\nu) \exp\{(2-\nu)\hat{\eta}_{ij}\} - y_{ij}(1-\nu) \exp\{(1-\nu)\hat{\eta}_{ij}\} \right] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j' + \boldsymbol{I}_d$$

with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i' \boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i{}' \boldsymbol{\gamma}_j$*, and* $\hat{\boldsymbol{u}}_i$ *is the maximum of*

$$Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = \sum_{j=1}^{m} \left[\left\{ \log \hat{W}(y_{ij}, \phi_j, \nu) - \log(y_{ij}) \right\} I_{(y_{ij}=0)} + \frac{1}{\phi_j} \left(\frac{y_{ij} \exp\{(1-\nu)\eta_{ij}\}}{1-\nu} - \frac{\exp\{(2-\nu)\eta_{ij}\}}{2-\nu} \right) \right]$$
$$- \frac{\boldsymbol{u}_i' \boldsymbol{u}_i}{2}.$$

Note a common power parameter ν is used for all species. This is done mainly for reasons of stability, as there is typically very little information within each species to estimate the power parameter, and previous studies have shown that most species tend to have very similar values of ν (Dunstan et al., 2013).

3.1 Estimation and Inference

In all of the cases above, the Laplace approximated likelihood has a fully closed form, and therefore parameter estimates, $\hat{\boldsymbol{\Psi}}$, and predictions of the latent variables, $\hat{\boldsymbol{u}}_i$ for the GLLVM are easily obtained by using standard quasi-Newton optimization routines available in `R` and alternately maximizing $\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)$ and $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i)$ until convergence. For this, we have developed an `R` package `gllvm`, which is now available on GitHub and implements the framework proposed in this paper among other functionalities.

For Laplace's method, the asymptotic error is of order $O(m^{-1})$, where m is the number of species. The method is therefore well suited and provides a good approximation for high-dimensional abundance data where m/n is often close to or exceeds one. As discussed in Huber et al. (2004) the Laplace approximated estimates solve the M-estimation equations, thus their consistency and asymptotic normality follow under general assumptions (Chapters 6.2-6.3, Huber and Ronchetti, 2009). Furthermore, the asymptotic standard errors for $\hat{\boldsymbol{\Psi}}$ are easy to compute as the observed information matrix (negative Hessian) is obtained as part of the estimation process. This allows us to construct confidence intervals as well as conduct Wald tests for the model parameters. Likelihood ratio tests are also readily available, although with the small sample

sizes as well as the fact that removing a covariate from the model actually removes m coefficients, their use requires careful consideration. In our examples, we use instead the corrected Akaike information criterion for variable selection, although this is by no means the only information criterion one could employ.

Regarding ordination, similar to Hui et al. (2015) we can construct an ordination plot using predicted latent variables from the fitted GLLVM. The asymptotic standard errors for $\hat{\boldsymbol{u}}_i$ are easily obtained in a similar fashion as those for $\hat{\boldsymbol{\Psi}}$, and can be used for example in constructing prediction regions around ordination points. In particular if $d = 2$, then $\hat{\boldsymbol{u}}_i$ is a pair of coordinates representing the position of the site i in a latent two-dimensional indirect gradient space. Furthermore, the coefficients $\boldsymbol{\gamma}_j$ quantify how each species response relates to the latent variables. Therefore, we can construct a model-based biplot, where the site ordinations give an indication of how species composition differs across sites, while plotting the species loadings identify the indicator species characterizing the sites.

In Section 5, we illustrate how the model-based inference discussed above using GLLVMs can be applied, using two ecological datasets.

4 Simulation studies

To evaluate the finite-sample properties of estimates obtained using the Laplace approximation method, we performed two simulation studies on overdispersed count and biomass data. Details on the simulation setups as well as example `R` code are given in Appendix C.

4.1 Overdispersed counts

In the overdispersed count data case, we compared the Laplace approximation estimates to those given by variational approximation method (Hui et al., 2016). To our knowledge, this is the only other maximum likelihood based method currently available which can handle negative binomial GLLVMs in a computationally feasible manner. In Hui et al. (2015, 2016), MCMC based methods and the EM algorithm were used in estimation and inference respectively, but we found these methods to be computationally so intensive that they could not be included for comparison. For instance, in our initial testing with the simulation setup (d) below, MCMC based method took approximately 12 hours to fit the negative binomial GLLVM.

The simulation setup was as follows. We simulated $K = 1500$ datasets according to the negative binomial model using four different sample sizes and dimensions: (a) $n = 100$ and $m = 50$, (b) $n = 50$ and $m = 100$, (c) $n = 50$ and $m = 500$ and (d) $n = 50$ and $m = 1000$. Note that especially response matrices with $m \gg n$ typically arise with multivariate abundance data in ecology. As a mean model, we used $\log(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$, meaning no covariates were included in the model. The true latent variables, $\boldsymbol{u}_i$, were

generated from the mixture of bivariate normal distributions all having covariance matrices $0.5I_2$, means $(-1, 1)$, $(2, 1.5)$ and $(0.5, -1.5)$, and proportions 0.4, 0.3 and 0.3, respectively. The sites thus exhibit a clustering on a latent variable space. The population parameters $\boldsymbol{\gamma}_j$ were generated so that all the elements in both columns were generated independently from a uniform distribution $U(-2, 2)$. The population row parameters α_i and species-specific parameters β_{0j} were generated from a uniform distribution $U(-1, 1)$, and the dispersion parameters were set to $\phi_j = 1$ for all species j.

Table 1: Average biases, root mean squared errors (RMSEs), coverage probabilities of 95% confidence intervals and mean CI widths for GLLVM estimates based on Laplace approximation and variational approximation methods. The true models were negative binomial GLLVMs with (a) $n = 100$ and $m = 50$, (b) $n = 50$ and $m = 100$, (c) $n = 50$ and $m = 500$ and (d) $n = 50$ and $m = 1000$.

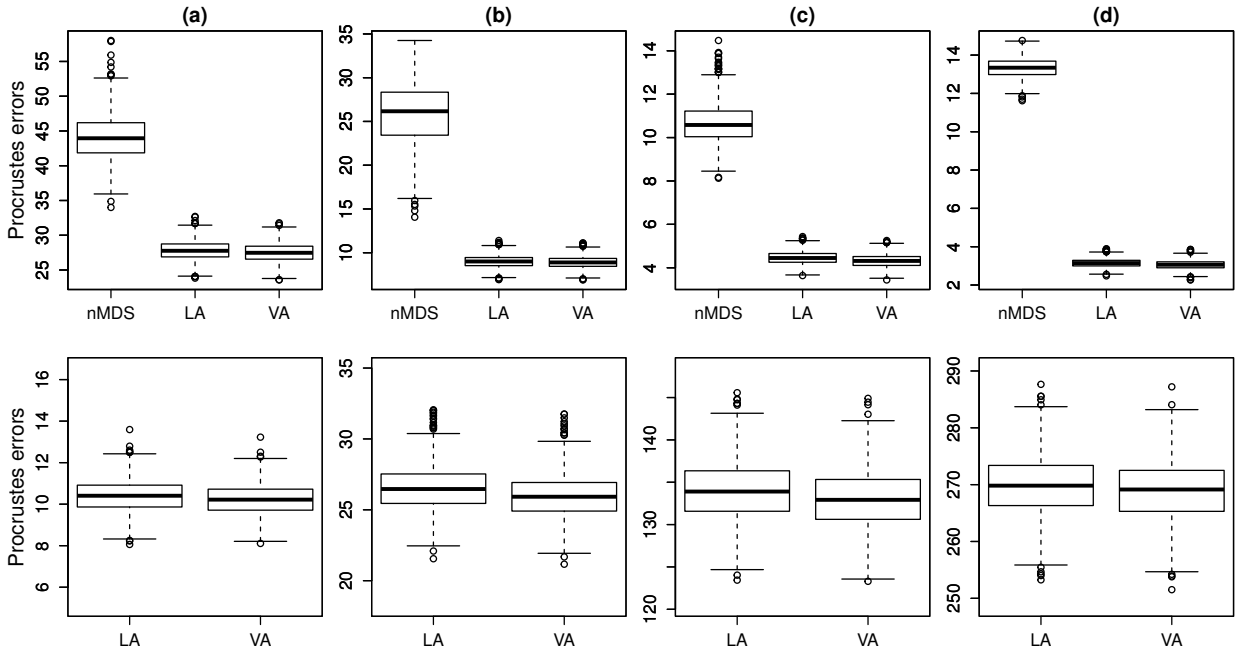
		Laplace				Variational			
		Bias	RMSE	Coverage	CI width	Bias	RMSE	Coverage	CI width
(a)	β_0	0.07	0.23	0.86	0.68	0.07	0.20	0.96	0.92
	α	-0.11	0.51	0.72	1.14	-0.11	0.39	0.85	1.14
	ϕ	-0.08	0.32	0.96	1.26	-0.03	0.30	0.96	1.16
(b)	β_0	0.10	0.30	0.88	0.97	0.10	0.30	0.95	1.27
	α	-0.19	0.30	0.95	1.21	-0.19	0.29	0.95	1.08
	ϕ	-0.12	0.42	0.97	2.30	-0.09	0.40	0.99	2.33
(c)	β_0	0.13	0.32	0.87	1.02	0.13	0.32	0.93	1.29
	α	-0.22	0.28	0.95	1.12	-0.22	0.27	0.96	1.12
	ϕ	-0.10	0.41	0.98	2.30	-0.10	0.40	0.98	2.31
(d)	β_0	0.15	0.32	0.84	0.99	0.15	0.33	0.86	1.15
	α	-0.24	0.45	0.74	1.11	-0.25	0.41	0.60	1.11
	ϕ	-0.10	0.41	0.98	2.30	-0.10	0.40	0.98	2.31

Table 1 lists the average biases, root mean squared errors, coverage probabilities of 95% confidence intervals and mean confidence interval widths for estimates of α_i, β_{0j} and ϕ_j, when the Laplace and variational approximation methods were used to fit the models assuming negative binomial distributed responses. Results indicate that both methods performed similarly, with slight but noticeable biases especially for the row parameter α_i when $n \ll m$. In some cases the coverage probabilities were a lot smaller or higher than the designated level 0.95. Notice that instead of using here large-sample theory, more accurate intervals could have be obtained using, for instance, resampling based methods. This approach was however not considered due to large computational burden, and we reserve this for avenue for future empirical research.

To evaluate the performance of estimated $\boldsymbol{\gamma}_j$ and predicted latent variables, $\boldsymbol{u}_i$, the mean Procrustes errors between the estimated and true parameter values were computed (Bartholomew et al., 2011, Chapter 8.4). The Procrustes error can be thought of as the mean squared error of two matrices after accounting for differences in rotation and scale. The boxplots of Procrustes errors based on Laplace approximation method and variational approximation method are given in Figure 1. To compare the performances of model based

ordination methods to a classical algorithmic based ordination method, non-metric multidimensional scaling (nMDS), the mean Procrustes errors between the true latent variables and nMDS ordination points were also computed. As seen in Figure 1, both model based ordination methods strongly outperform nMDS. The results based on Laplace approximation method and variational approximation method are almost equally good.

Figure 1: Comparative boxplots of Procrustes errors between true and estimated ordination points (first row) and true and estimated parameters $\hat{\boldsymbol{\gamma}}_j$ (second row). Ordination points (and parameters $\hat{\boldsymbol{\gamma}}_j$ when applicable) are obtained from non-metric multidimensional scaling (nMDS) and negative binomial GLLVM fitted using Laplace approximation method (LA) and variational approximation method (VA). The true model in each plot was negative binomial GLLVM with (a) $n = 100$ and $m = 50$, (b) $n = 50$ and $m = 100$, (c) $n = 50$ and $m = 500$ and (d) $n = 50$ and $m = 1000$.



Finally, regarding computation time, the proposed Laplace approximation method averaged 13.2, 12.1, 159.4 and 609.3 seconds respectively to estimate the parameters and their standard errors using models in simulation settings (a) to (d) above. This was a substantial gain on the corresponding mean computation times for variational approximation method, which averaged 56.4, 54.9, 233.4 and 650.9 seconds, respectively. The main reason for differences in computation times is that for these setups, the variational approximation needs to estimate $5n$ variational parameters (corresponding to the mean and covariance parameters in the variational distribution) on top of the model parameters.

4.2 Biomass data

In the case of biomass data, we used simulations to study the effect of ignoring the correlation between taxa on regression estimates. We used only Laplace approximation method to fit the models, as there are currently no alternative maximum likelihood based methods available for fitting GLLVMs to biomass data.

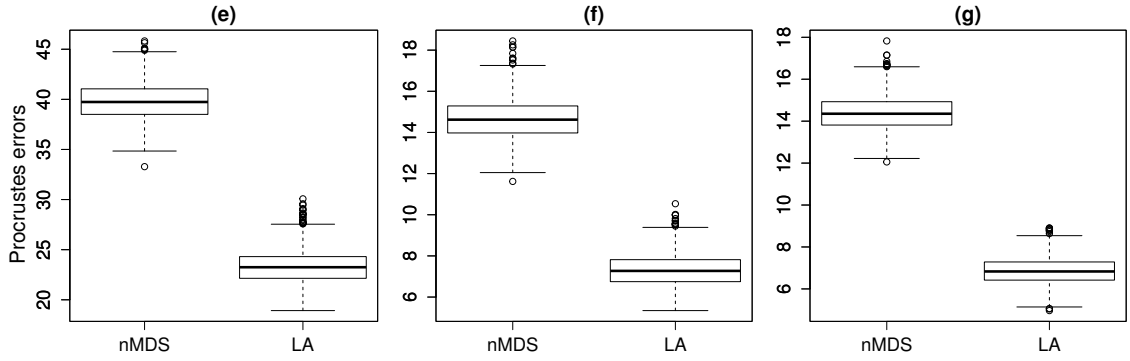
The simulation setup differed slightly from the one used previously for overdispersed counts. Specifically, we simulated $K = 1500$ datasets according to the Tweedie model with fixed power parameter $\nu = 1.6$ using three different sample sizes with dimensions: (e) $n = 100$ and $m = 50$, (f) $n = 50$ and $m = 100$ and (g) $n = 50$ and $m = 200$. As a mean model, we used $\log(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$, with two covariates included in the model. The true latent variables for the GLLVM, $\boldsymbol{u}_i$, were generated from a three component mixture of bivariate normal distributions all having covariance matrices $0.5I_2$, with differing means $(-1, 1)$, $(1.5, 1.5)$ and $(0.5, -1.5)$, and proportions 0.4, 0.3 and 0.3 respectively. The first covariate x_{i1} was generated from the standard normal distribution and the second covariate x_{i2} from the exponential distribution with rate $\lambda = 1$. Finally, as per the overdispersed count simulation, we constructed $\boldsymbol{\gamma}_j$ such that all elements in both columns were obtained from the uniform distribution $U(-2, 2)$, while the species-specific covariate coefficients $\boldsymbol{\beta}_j$ and intercept parameters β_{0j} were chosen from the uniform distribution $U(-1, 1)$. The dispersion parameters were set to $\phi_j = 1$ for all species j.

Table 2: Average biases and root mean squared errors (MSEs) of Tweedie GLLVM and Tweedie GLM estimates based on Laplace approximation method. The true models were Tweedie GLLVMs with (e) $n = 100$ and $m = 50$, (f) $n = 50$ and $m = 100$ and (g) $n = 50$ and $m = 200$.

		GLLVM		GLM	
		Bias	RMSE	Bias	RMSE
(e)	β_0	0.06	0.31	1.15	1.37
	β_1	0.03	0.16	-0.09	0.18
	β_2	-0.08	0.32	0.02	0.21
	ϕ	-0.03	0.12	2.06	2.71
(f)	β_0	-0.02	0.25	0.97	1.17
	β_1	0.00	0.17	-0.20	0.32
	β_2	-0.03	0.23	0.06	0.34
	ϕ	-0.07	0.18	1.79	2.44
(g)	β_0	-0.02	0.27	0.94	1.12
	β_1	-0.00	0.17	-0.18	0.32
	β_2	-0.03	0.25	0.06	0.34
	ϕ	-0.07	0.18	1.63	2.10

Table 2 lists the average biases and mean squared errors for regression estimates based on a Tweedie GLLVM compared to a Tweedie generalized linear model (GLM). The latter does not include any latent variables to account for residual correlation between species i.e., it assumes the species are independent after

Figure 2: Comparative boxplots of Procrustes errors between true and estimated ordination points. Ordination points are obtained from non-metric multidimensional scaling (nMDS) and Tweedie GLLVM fitted using Laplace approximation method (LA). The true model in each plot was Tweedie GLLVM with (e) $n = 100$ and $m = 50$, (f) $n = 50$ and $m = 100$ and (g) $n = 50$ and $m = 200$.



accounting for correlations due to the observed predictors $\boldsymbol{x}_i$. In all of the considered setups ignoring the correlation yields biased estimates with high variability, particularly for the species specific intercepts and overdispersion parameters. Additionally, Figure 2 displays the boxplots of Procrustes errors between true and predicted latent variables, as well as those between the true latent variables and ordination points given by nMDS. Again, the model based approach of GLLVM yields substantially better ordination results.

5 Examples

5.1 Microbial Community Data

We applied Laplace approximated GLLVMs on the bacterial species data discussed in Nissinen et al. (2012). Altogether eight different sampling sites were selected from three locations. Three of the sites were in Kilpisjärvi, Finland, three in Ny-Ålesund, Svalbard, Norway, and two in Mayrhofen, Austria. From each sampling site, several soil samples were taken and their bacterial species were recorded. The data consist of $m = 1276$ bacterial species counts measured from $n = 56$ sites. The sites can be considered as independent from each other since bacterial communities are known to be very location specific. As many of the species were observed only in few sites, we decided to exclude such rare species and considered only species present at five of more sites. This reduced the number of species to $m = 985$. In addition to bacteria counts, three continuous environmental variables (pH, available phosphorous and soil organic matter) were measured from each soil sample.

In order to study whether the effect of environmental variables is seen in an unconstrained ordination plot, we first considered a generalized linear latent variable model with two latent variables and no predictors,

and constructed an ordination plot based on the predicted latent variables. Due to small sample size, the corrected Akaike information criterion, AIC_c, was used for selecting which count distribution was most appropriate for the data (Burnham and Anderson, 2002). The values for AIC_c (scaled by n and subtracted by 1942) based on the Poisson, negative binomial and ZIP models are given in the first column of Table 3, with results indicating that the negative binomial model fitted the data best. The ZIP model outperformed the model assuming Poisson counts.

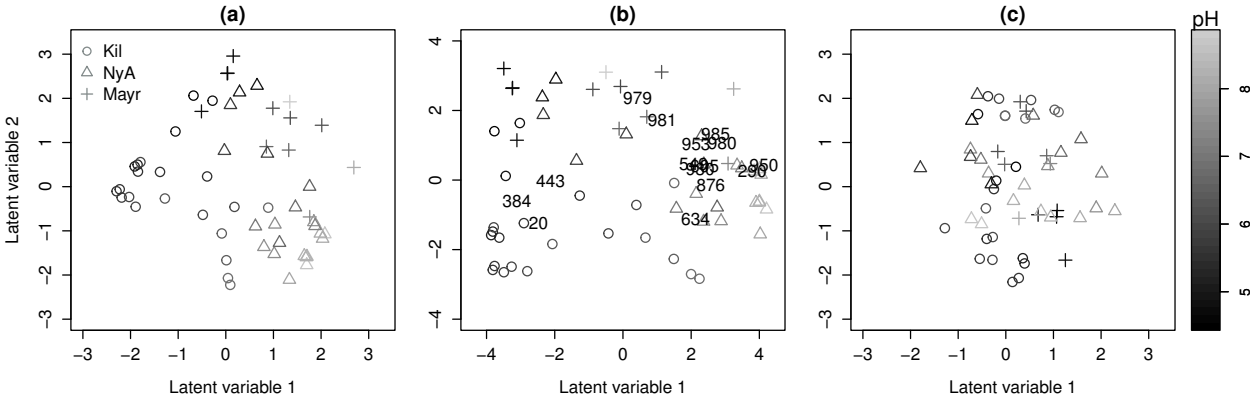
Table 3: Values of AIC_c (scaled by n and subtracted by 1942) for Poisson, negative binomial (NB) and ZIP GLLVMs (1) without covariate, (2) with pH as a covariate, (3) with pH, soil organic matter and phosphorous as covariates, (4) with pH included along with a site effect and (5) with all three soil covariates included along with a site effect.

	(1)	(2)	(3)	(4)	(5)
Poisson	771	674	463	395	244
NB	178	150	86	59	0
ZIP	630	547	377	311	189

The ordination of sites based on negative binomial GLLVM is plotted in Figure 3(a). The sites are coloured according to their pH values. A very clear gradient in the pH values of sites is observed, while there was less evidence of such a pattern with the two other soil variables (see Figure B1 in Appendix B). In addition, the ordination points are (also) labeled according to the sampling location (Kilpisjärvi, Ny-Ålesund and Innsbruck), and it is clear that the sites differed in terms of species composition. In Figure 3(b), a biplot based on generalized linear latent variable model is given. Here indices of the 15 species with largest factor loadings are added in the (rotated) ordination plot in Figure 3(a). The biplot suggests a small set of indicator species which prefer sites with low pH values and a larger set of indicator species for high pH sites.

In order to study whether the environmental variables alone are capable of explaining the variation in species composition across sites, we included them as explanatory variables in the GLLVM. Points estimates with 95% confidence intervals are plotted in Figure B2 in Appendix B, and indicate that pH value was the main covariate affecting the species composition. The corresponding ordination plots are given in Figure B3 in Appendix B, and they indicate that even though the effect of environmental variables on ordination vanishes, the ordination still exhibits a sampling location effect. Several Kilpisjärvi sites in particular seem to be different from the others. To account for this, we further added the sampling location as a categorical covariate into the model. The resulting ordination plot in Figure 3(c) shows that there is no visible pattern in sampling location anymore. As the figure uses the same scale as plots in Figure 3(a), it is clear that a lot of covariation in ordination is explained by the covariates included in the model. When comparing nested models, in particular, the model with environmental covariates to the null model, and the model with all covariates to the model with environmental covariates, the deviances are 5144.6 and 4830.1, respectively,

Figure 3: (a) The ordination plot of $n = 56$ sites based on generalized linear latent variable model without any covariates assuming negative binomial distributed responses. (b) The biplot, where 15 species with the largest factor loadings (in terms of distance from the origin) are printed on top of the (rotated) site ordination to illustrate indicator species for sites with low and high pH values. (c) The ordination plot based on generalized linear latent variable model with environmental variables and sampling location as covariates. The plot (c) uses the same scale as Figure (a) to emphasize the reduction in variation. The sites in ordination plots are coloured according to their pH values and labeled according to the sampling site.



suggesting that about 6% of the total covariation is due to environmental covariates based on the marginal log-likelihood. Notice that changes in log-likelihood are not the only approach to quantifying variance explained, and other methods like extensions of pseudo R^2 are possible (see for instance recent work by Nakagawa and Schielzeth, 2013, for the case of generalized linear mixed models). Notice also that the corrected AIC_c picks the model with these covariates i.e., the negative binomial GLLVM with all three covariates and sampling location, as the best model (Table 3).

Finally, as a diagnostic tool, we plotted Dunn-Smyth residuals (Dunn and Smyth, 1996) against linear predictors for Poisson, zero-inflated Poisson and negative binomial GLLVM models with pH, soil organic matter, phosphorous and site as covariates. The plots in Figure B4 in Appendix B show residuals for 100 randomly selected species to make any patterns in the plots more apparent. Specifically, the plot for the Poisson model displays a fan-shaped pattern, which means that the model is not capable of capturing the overdispersion in the data, while the plot for the ZIP model displays skew with a lowess curve showing a positive trend in residuals. By contrast, the Dunn-Smyth residuals given by negative binomial GLLVMs are uniformly distributed around zero indicating an appropriate fit to the data.

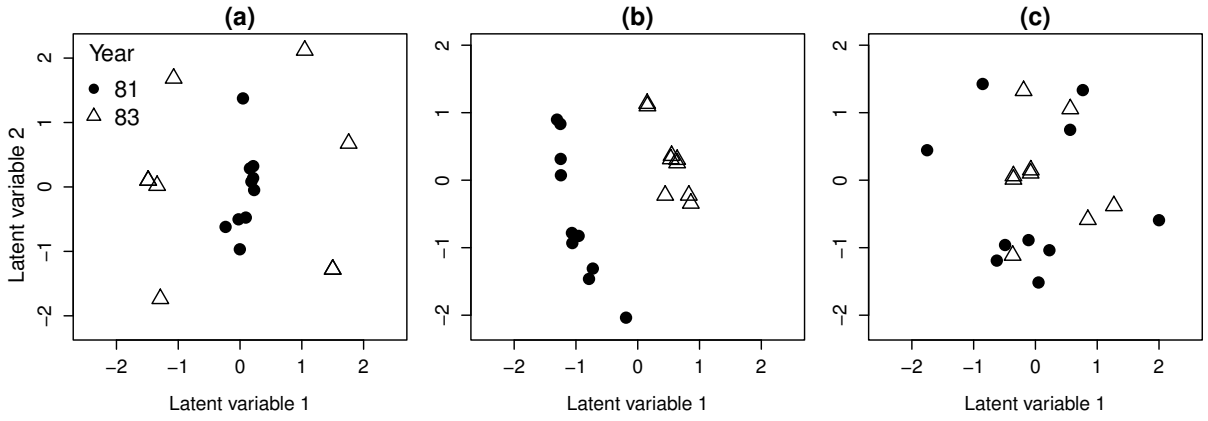
5.2 Coral data

As the second example, we consider abundances of coral reef species collected in Tikus island, Indonesia (Warwick et al., 1990). The abundance of each reef species was measured as the length (in centimetres)

of a ten metre transect which intersected with the species. The data were collected during 1981-1988, but in this example we only consider measurements taken in 1981 and in 1983. The reason for this is that there was an El Niño event in 1982-1983 causing a tenfold decrease in site total abundance between the two sampling times. The aim is to study whether this event had any effect on the community structure, beyond the effect on total abundance. We consider species with more than four presences over the two years. Also one record for a site in 1983 that contained no presences was removed. The final data set thus contains $n = 19$ sites and $m = 18$ species.

Warwick et al. (1990) applied non-metric multidimensional scaling on this data and concluded that stress due to El Niño event increases variability in coral communities; see also Figure 4(a). Later Hui et al. (2015) applied GLLVM based ordination methods to the corresponding, converted presence-absence data, and showed that there was in fact no evidence of a difference in dispersion across the two sampling times. We now repeat their analyses using a GLLVM assuming Tweedie distributed responses. The power parameter ν was estimated using a profile likelihood approach, testing several different parameter values and selecting the one ($\nu = 1.1$) which maximised the profile likelihood. At first, the generalized linear latent variable model without site effects was fitted to produce an ordination of species abundance, i.e., including effects on total abundance as well as on relative abundance. The ordination plot in Figure 4(b) exhibits a clear location difference between coral compositions in 1981 and 1983, reflecting the El Niño event. Secondly, a GLLVM with site effects was fitted in order to study ordinations of species composition. The results in Figure 4(c) indicate that the species compositions did not change between the two sampling times.

Figure 4: The ordination plots of $n = 19$ sites based on (a) non-metric multidimensional scaling (b) Tweedie GLLVM without site effect and (c) Tweedie GLLVM with site effects. The sites in ordination plots are labeled according to the year the data was collected.



In Figure B5 in Appendix B the residual plots are given for the GLLVM models (b) and (c).

6 Discussion

In this paper we illustrated how generalized linear latent variable models can be used to model multivariate abundance data and biomass data, that is, data common in ecological studies. When modeling multivariate abundance data (overdispersed counts), we assumed negative binomial or zero-inflated Poisson models for responses. For biomass data (continuous but non-negative data) the Tweedie distributed responses were assumed. Notice however that these distributions just serve as examples and the method can be tailored to handle any response distribution.

Although the generalized linear latent variable models are straightforward to derive, the major challenge is the lack of computationally efficient estimation tools. In this paper, we used the Laplace approximation method for the estimation and inference. The general form for the Laplace approximation in case of exponential family is given in Huber et al. (2004), and we have extended this to the zero-inflated Poisson, negative binomial and Tweedie distributions cases, which involve additional nuisance parameters. Other case-by-case extensions may sometimes be required, e.g. to handle ordinal data, and one could argue that a disadvantage of the Laplace method is the need for case-by-case derivation of estimation algorithms. In such case, automated differentiation offers a way forward in this regard, e.g. the Template Model Builder software (Kristensen et al., 2016) can potentially simplify estimation procedures, as it requires specification of the complete likelihood only, and implementation is based on `C++` code. More importantly however, such general software nevertheless employs the same Laplace approximation considered in this article as the basis for estimation and inference in GLLVMs.

Simulation studies indicated that such estimation method performs well when modeling overdispersed counts and continuous, non-negative data. However, as shown in Joe (2008) the Laplace approximation can become less adequate when the conditional distributions of the responses are highly discrete. In such settings, such as for binary and ordinal responses, we may consider other approximations method e.g. the variational approximation approach as in Hui et al. (2016). All these choices are available in `R` package `gllvm`, which is associated with this article. In our two examples we illustrated how generalized linear latent variable models can be applied to produce ordination plots as well as to make inferences on environmental covariates on species communities.

The generalized latent variable model considered in this paper can be generalized in several ways. If q trait covariates $\boldsymbol{t}_j$ are also recorded and one wishes to study the environmental-trait interaction, a simple way to do it is via model $g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_e + \text{vec}'(\boldsymbol{B}_I')(\boldsymbol{t}_j \otimes \boldsymbol{x}_i) + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$. Here $\boldsymbol{\beta}_e$ is now a main effect for the environment, common for all species, and $\boldsymbol{B}_I'$ is an interaction matrix, which tells us how well traits explain variation in the environmental response. Notice that, as compared to (1), the above model

includes far less parameters to be estimated and tested. In ecology, the model (without latent variables) is known as a fourth corner model (Brown et al., 2014). Another way to reduce the number of parameters is to introduce random effects into the model. For instance, using a random rather than fixed site effect might be beneficial as, based on our simulation studies, the fixed site estimates seem to be slightly biased in the case of the latter. We will consider the fourth corner latent variable model and random effect models in our future studies.

Acknowledgements

We thank the Associate Editor and the referees for their helpful comments. We also thank Dr Manoj Kumar and Dr Riitta Nissinen for providing us the plant-microbial diversity data. JN and ST were supported by the Academy of Finland grants 251965 and 283323.

A Proofs

A.1 Laplace approximations for the general exponential family

Assume that the responses y_{ij} come from the exponential family of distributions with mean $\mu_{ij} = E(y_{ij})$, and write $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp\{y_{ij}a_j(\mu_{ij}) - b_j(\mu_{ij}) + c_j(y_{ij})\}$, where $a_j(\cdot)$, $b_j(\cdot)$ and $c_j(\cdot)$ are known functions, and $\boldsymbol{\Psi}$ includes all model parameters. The log-likelihood function (5) for parameter vector $\boldsymbol{\Psi}$ now equals

$$l(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \int \left[\prod_{j=1}^{m} \exp\left\{ y_{ij}\, a_j(\mu_{ij}) - b_j(\mu_{ij}) + c_j(y_{ij}) \right\} \right] \times (2\pi)^{-\frac{d}{2}} \exp\left(-\frac{1}{2} \boldsymbol{u}_i' \boldsymbol{u}_i \right) d\boldsymbol{u}_i,$$

and the Laplace approximation of the log-likelihood function is

$$\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log\det\{\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)\} + \sum_{j=1}^{m} \{y_{ij}\, a_j(\mu_{ij}) - b_j(\mu_{ij}) + c_j(y_{ij})\} - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right),$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{j=1}^{m} \frac{\partial^2 \{-y_{ij}\, a_j(\mu_{ij}) + b_j(\mu_{ij})\}}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \bigg|_{\boldsymbol{u}_i = \hat{\boldsymbol{u}}_i} + \boldsymbol{I}_d,$$

and $\hat{\boldsymbol{u}}_i$ is the maximum of $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = (1/m) \left(\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) - \boldsymbol{u}_i' \boldsymbol{u}_i / 2 \right)$ with respect to $\boldsymbol{u}_i$. The result has been proven in Huber et al. (2004).

A.2 Poisson responses

Species counts can be modelled as Poisson distributed responses, $y_{ij} \sim Poisson(\mu_{ij})$, and log link function. Then $a_j(\mu_{ij}) = \log(\mu_{ij}), b_j(\mu_{ij}) = \mu_{ij}$, and $c_j(y_{ij}) = -\log(y_{ij}!)$. Then the following Laplace approximation $\tilde{l}$ for the log-likelihood function is obtained

$$\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{i=1}^{n} \left(-\frac{1}{2} \log \det \left(\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) \right) + \sum_{j=1}^{m} \left[y_{ij} \hat{\eta}_{ij} - \exp(\hat{\eta}_{ij}) - \log(y_{ij}!) \right] - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right),$$

where $\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = \sum_{j=1}^{m} \exp(\hat{\eta}_{ij}) \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j' + \boldsymbol{I}_d$, with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i' \boldsymbol{\beta}_j + \hat{\boldsymbol{u}_i}' \boldsymbol{\gamma}_j$, and $\hat{\boldsymbol{u}}_i$ is the maximum of

$$Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = \frac{1}{m} \left[\sum_{j=1}^{m} \left[y_{ij} \eta_{ij} - \exp(\eta_{ij}) - \log(y_{ij}!) \right] - \frac{\boldsymbol{u}_i' \boldsymbol{u}_i}{2} - \frac{d}{2} \log(2\pi) \right].$$

A.3 Proof of Theorem 2

Assume that the responses y_{ij} come from the zero-inflated Poisson distribution with mean $E(y_{ij}) = (1-p_j)\mu_{ij}$ and density of the form (3). The log-likelihood function (5) then equals

$$\begin{aligned} l(\boldsymbol{\Psi}) = & \sum_{i=1}^{n} \log \Bigg(\int \prod_{j=1}^{m} \exp \big(\log \left[p_j + (1-p_j) \exp\{-\exp(\eta_{ij})\} \right] I_{(y_{ij}=0)} \\ & + \left\{ \log(1-p_j) - \exp(\eta_{ij}) + y_{ij}\eta_{ij} - \log(y_{ij}!) \right\} I_{(y_{ij}>0)} \big) \Bigg) \\ & \times (2\pi)^{-\frac{d}{2}} \exp \left(-\frac{1}{2} \boldsymbol{u}_i' \boldsymbol{u}_i \right) d\boldsymbol{u}_i. \end{aligned}$$

Hence, the Laplace approximation of the log-likelihood function is

$$\begin{aligned} \tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) = & \sum_{i=1}^{n} \left(-\frac{1}{2} \log \det \left\{ \boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) \right\} + \sum_{j=1}^{m} \log f(y_{ij} | \hat{\boldsymbol{u}}_i; \boldsymbol{\Psi}) - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right) \\ = & \sum_{i=1}^{n} \Bigg(-\frac{1}{2} \log \det \left\{ \boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) \right\} + \sum_{j=1}^{m} \Big(\log \left(p_j + (1-p_j) \hat{A}_{ij} \right) I_{(y_{ij}=0)} \\ & + \left\{ \log(1-p_j) - \exp(\hat{\eta}_{ij}) + y_{ij}\hat{\eta}_{ij} - \log(y_{ij}!) \right\} I_{(y_{ij}>0)} \Big) - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \Bigg), \end{aligned}$$

where

$$
\begin{aligned}
\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) &= \frac{\partial^2}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \left[-\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) + \frac{\boldsymbol{u}_i' \boldsymbol{u}_i}{2} \right] \Bigg|_{\boldsymbol{u}_i = \hat{\boldsymbol{u}}_i} \\
&= \sum_{j=1}^{m} \frac{\partial^2 \left\{ \exp(\eta_{ij}) I_{(y_{ij}>0)} - \log(p_j + (1-p_j) A_{ij}) I_{(y_{ij}=0)} \right\}}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \Bigg|_{\boldsymbol{u}_i = \hat{\boldsymbol{u}}_i} + \boldsymbol{I}_d \\
&= \sum_{j=1}^{m} \left[\exp(\hat{\eta}_{ij}) I_{(y_{ij}>0)} - \left(\frac{(1-p_j) \hat{A}_{ij} \exp(\hat{\eta}_{ij}) (\exp(\hat{\eta}_{ij}) - 1)}{p_j + (1-p_j) \hat{A}_{ij}} \right. \right. \\
&\quad \left. \left. - \frac{(1-p_j)^2 \hat{A}_{ij}^2 \exp(2\hat{\eta}_{ij})}{(p_j + (1-p_j) \hat{A}_{ij})^2} \right) I_{(y_{ij}=0)} \right] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j' + \boldsymbol{I}_d,
\end{aligned}
$$

with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i' \boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i{}' \boldsymbol{\gamma}_j$ and $\hat{A}_{ij} = \exp\{-\exp(\hat{\eta}_{ij})\}$, and $\hat{\boldsymbol{u}}_i$ is the maximum of $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = (1/m) \left(\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) - \boldsymbol{u}_i' \boldsymbol{u}_i / 2 \right)$.

A.4 Proof of Theorem 3

Assume that the responses y_{ij} come from the Tweedie distribution with mean $E(y_{ij}) = \mu_{ij}$ and density of the form (4). The log-likelihood function (5) then equals

$$
\begin{aligned}
l(\boldsymbol{\Psi}) = &\sum_{i=1}^{n} \log \left(\int \prod_{j=1}^{m} \exp \left(-\frac{\mu_{ij}^{2-\nu}}{\phi_j (2-\nu)} \right) I_{(y_{ij}=0)} + \frac{1}{y_{ij}} \tilde{W}(y_{ij}, \phi_j, \nu) \exp \left\{ \frac{1}{\phi_j} \left(\frac{y_{ij} \mu_{ij}^{1-\nu}}{1-\nu} - \frac{\mu_{ij}^{2-\nu}}{2-\nu} \right) \right\} I_{(y_{ij}>0)} \right) \\
&\times (2\pi)^{-\frac{d}{2}} \exp \left(-\frac{1}{2} \boldsymbol{u}_i' \boldsymbol{u}_i \right) d\boldsymbol{u}_i.
\end{aligned}
$$

Hence, the Laplace approximation of the log-likelihood function is

$$
\begin{aligned}
\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) &= \sum_{i=1}^{n} \left(-\frac{1}{2} \log \det \{ \boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) \} + \sum_{j=1}^{m} \log f(y_{ij}|\hat{\boldsymbol{u}}_i; \boldsymbol{\Psi}) - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2} \log \det \{ \boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) \} + \sum_{j=1}^{m} \left[\left\{ \log \tilde{W}(y_{ij}, \phi_j, \nu) - \log(y_{ij}) \right\} I_{(y_{ij}>0)} \right. \right. \\
&\quad \left. \left. + \frac{1}{\phi_j} \left(\frac{y_{ij} \exp\{(1-\nu)\hat{\eta}_{ij}\}}{1-\nu} - \frac{\exp\{(2-\nu)\hat{\eta}_{ij}\}}{2-\nu} \right) \right] - \frac{\hat{\boldsymbol{u}}_i' \hat{\boldsymbol{u}}_i}{2} \right),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{\Gamma}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i) &= \frac{\partial^2}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \left[-\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) + \frac{\boldsymbol{u}_i' \boldsymbol{u}_i}{2} \right] \Bigg|_{\boldsymbol{u}_i = \hat{\boldsymbol{u}}_i} \\
&= \sum_{j=1}^{m} \frac{\partial^2}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \frac{1}{\phi_j} \left(-\frac{y_{ij} \exp\{(1-\nu)\eta_{ij}\}}{1-\nu} + \frac{\exp\{(2-\nu)\eta_{ij}\}}{2-\nu} \right) \Bigg|_{\boldsymbol{u}_i = \hat{\boldsymbol{u}}_i} + \boldsymbol{I}_d \\
&= \sum_{j=1}^{m} \frac{1}{\phi_j} \left[(2-\nu) \exp\{(2-\nu)\hat{\eta}_{ij}\} - y_{ij}(1-\nu) \exp\{(1-\nu)\hat{\eta}_{ij}\} \right] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j' + \boldsymbol{I}_d,
\end{aligned}
$$

with $\hat{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i' \boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i{}' \boldsymbol{\gamma}_j$ and $\hat{A}_{ij} = \exp\{-\exp(\hat{\eta}_{ij})\}$, and $\hat{\boldsymbol{u}}_i$ is the maximum of $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = (1/m) \left(\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) - \boldsymbol{u}_i' \boldsymbol{u}_i / 2 \right)$.

B Additional Application Results

Figure B1: The ordination of $n = 56$ sites based on generalized linear latent variable model without any covariates assuming negative binomial distributed responses. The sites in ordination are coloured according to their (a) soil organic matter (SOM) values and (b) phosphorous (P) values, and labelled according to the sampling site.

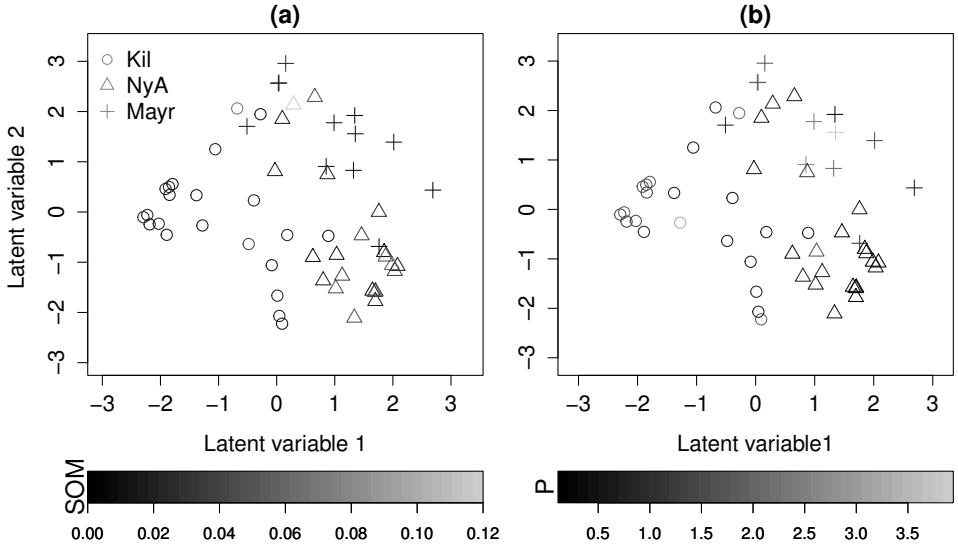


Figure B2: Ranked point estimates with 95% confidence intervals for the three environmental variables based on negative binomial GLLVM. Grey confidence intervals include the zero value.

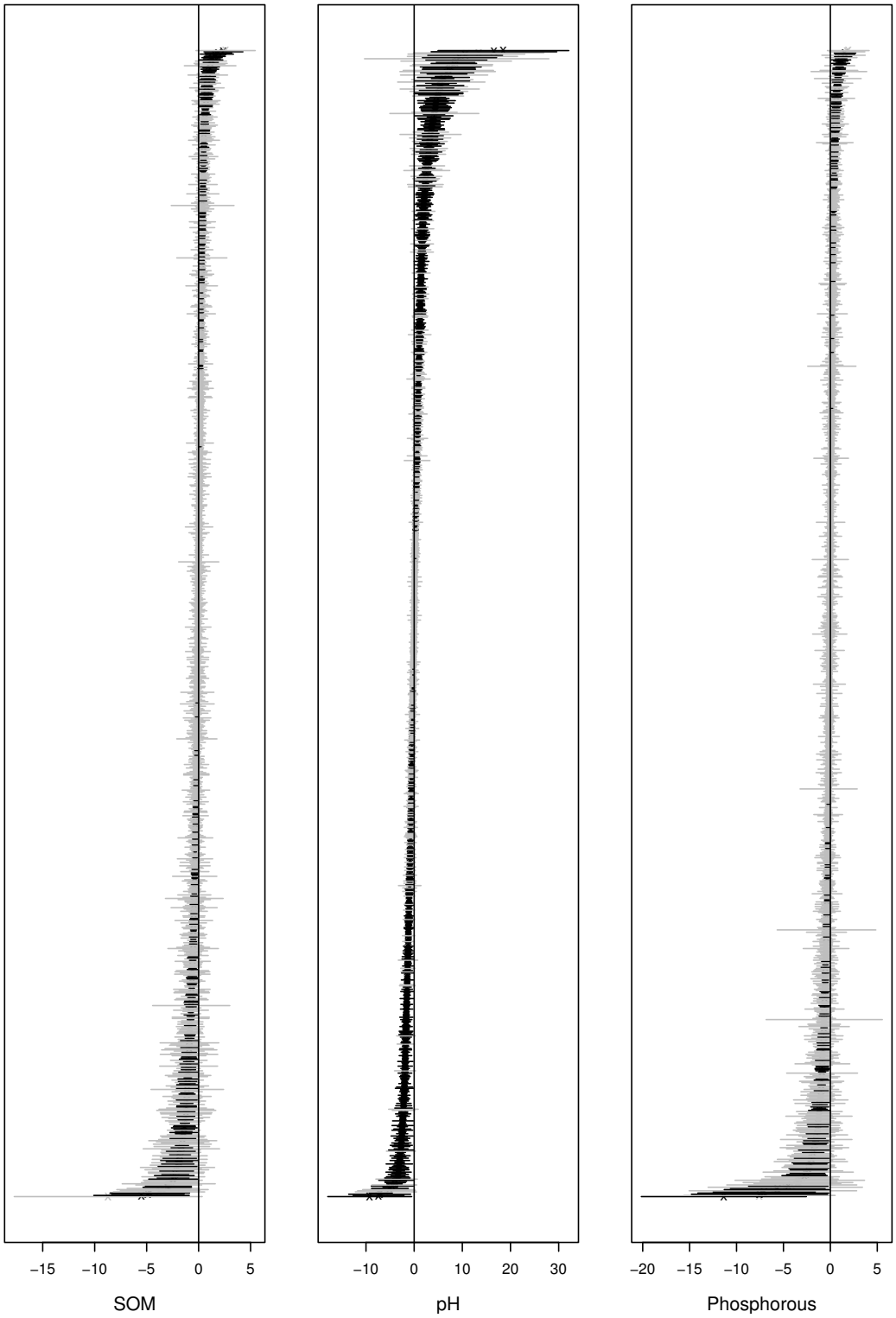


Figure B3: The ordination of $n = 56$ sites based on generalized linear latent variable model with pH, soil organic matter and phosphorous as covariates, and assuming negative binomial distributed responses. The sites in ordination are coloured according to their (a) pH values, (b) soil organic matter (SOM) values and (c) phosphorous (P) values, and labeled according to the sampling site. The effect of environmental variables vanishes, but the ordination is affected by the sampling location few Kilpisjärvi sites being different from the others what comes to species composition.

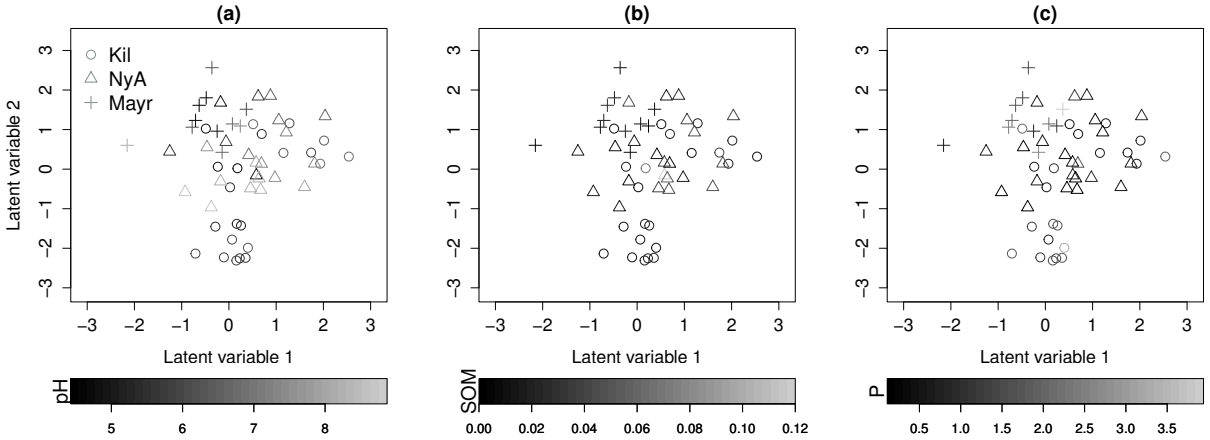


Figure B4: Dunn-Smyth residuals against linear predictors for the (a) Poisson, (b) zero inflated Poisson and (c) negative binomial GLLVM models with pH, soil organic matter, phosphorous and categorical site as covariates. Lowess curves are included in the plots.

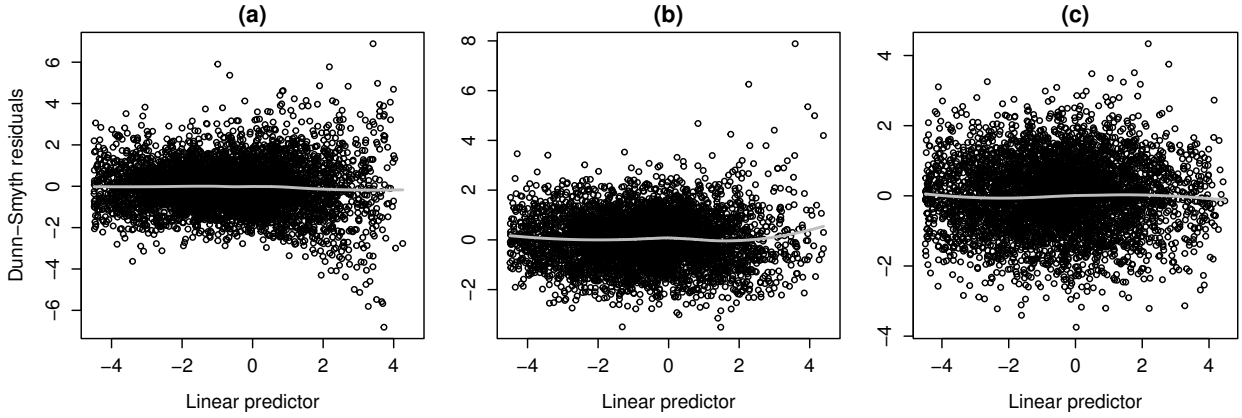
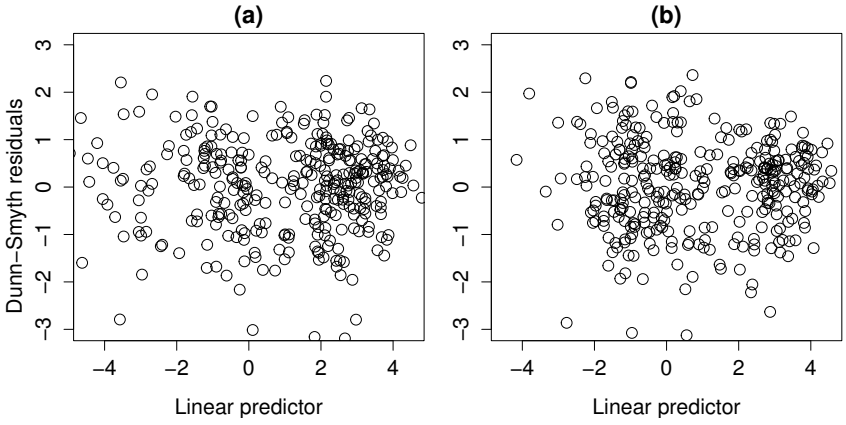


Figure B5: Dunn-Smyth residuals against linear predictors for the Tweedie models (a) without site effect and (b) with site effect.



References

Araújo, M. B. and Luoto, M. (2007). The importance of biotic interactions for modelling species distributions under climate change. *Global Ecology and Biogeography*, 16:743–753.

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Wiley: New York.

Bianconcini, S. and Cagnone, S. (2012). Estimation of generalized linear latent variable models via fully exponential Laplace approximation. *Journal of Multivariate Analysis*, 112:183–193.

Blanchet, F. (2014). *HMSC: Hierachical modelling of species community*. R package version 0.6-2.

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.

Burnham, K. and Anderson, D. (2002). *Model selection and multimodel inference: Al practical information-theoretic approach*. Springer.

Chu, H., Fierer, N., Lauber, C. L., Caporaso, J. G., Knight, R., and Grogan, P. (2010). Soil bacterial diversity in the arctic is not fundamentally different from that found in other biomes. *Environmental Microbiology*, 12:2998–3006.

Cressie, N., Calder, C. A., Clark, J. S., Hoef, J. M. V., and Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, 19(3):553–570.

Dunn, P. K. and Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5:236–244.

Dunn, P. K. and Smyth, G. K. (2005). Series evaluation of tweedie exponential dispersion model densities. *Statistics and Computing*, 15:267–280.

Dunstan, P. K., Foster, S. D., Hui, F., and Warton, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of Agricultural, Biological and Environmental Sciences*, 18:357–375.

Foster, S. D. and Bravington, M. V. (2013). A Poisson–Gamma model for analysis of ecological non-negative continuous data. *Environmental and ecological statistics*, 20:533–552.

Hall, P., Ormerod, J. T., and Wand, M. (2011a). Theory of gaussian variational approximation for a poisson mixed model. *Statistica Sinica*, 21:369–389.

Hall, P., Pham, T., Wand, M. P., Wang, S. S., et al. (2011b). Asymptotic normality and valid inference for Gaussian variational approximation. *The Annals of Statistics*, 39:2502–2532.

Huber, P. and Ronchetti, E. (2009). *Robust Statistics*. Wiley: New York.

Huber, P., Ronchetti, E., and Victoria-Feser, M. (2004). Estimation of generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66:893–908.

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., Ormerod, J., Haapaniemi, V., and Taskinen, S. (2016). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*. In press.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Computational Statistics & Data Analysis*, 5066–5074:52.

Jorgensen, B. (1997). *The Theory of Dispersion Models*. Chapman & Hall.

Kendal, W. S. (2004). Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. *Ecological Complexity*, 1(3):193–209.

Kristensen, K., Nielsen, A., Berg, C., Skaug, H., and Bell, B. (2016). Tmb: Automatic differentiation and laplace approximation. *Journal of Statistical Software, Articles*, 70(5):1–21.

Letten, A. D., Keith, D. A., Tozer, M. G., and Hui, F. K. (2015). Fine-scale hydrological niche differentiation through the lens of multi-species co-occurrence models. *Journal of Ecology*, 103:1264–1275.

Männistö, M. K., Tiirola, M., and Häggblom, M. M. (2007). Bacterial communities in arctic fjelds of finnish lapland are stable but highly ph-dependent. *FEMS Microbiology Ecology*, 59:452–465.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005). Zero tolerance ecology: improving ecological inference by modelling the source of zero observations. *Ecology letters*, 8:1235–1246.

Morales-Castilla, I., Matias, M. G., Gravel, D., and Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in ecology & evolution*, 30(6):347–356.

Moustaki, I. (1996). A latent trait and a latent class model for mixed observed variables. *British Journal of Mathematical and Statistical Psychology*, 49:313–334.

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

Nakagawa, S. and Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods In Ecology And Evolution*, 4:133–142.

Nissinen, R., Männistö, M., and van Elsas, J. (2012). Endophytic bacterial communities in three arctic plants from low arctic fell tundra are cold-adapted and host-plant specific. *FEMS Microbiology Ecology*, 82:510–522.

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016a). 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., de Knegt, H. J., and Delgado Sanchez, M. d. M. (2016b). *Quantitative Ecology and Evolutionary Biology: Integrating Models with Data*. Oxford: Oxford University Press.

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

Rodrigues-Motta, M., Pinheiro, H. P., Martins, E. G., Araujo, M. S., and dos Reis, S. F. (2013). Multivariate models for correlated count data. *Journal of Applied Statistics*, 40:1586–1596.

Sammel, M. D., Ryan, L. M., and Legler, J. M. (1997). Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59:667–678.

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

Taylor, L. R. (1961). Aggregation, variance and the mean. *Nature*, 189:732 – 735.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 16:275–289.

Warton, D. I., Blanchet, F. G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., and Hui, F. K. (2016). Extending Joint Models in Community Ecology: A Response to Beissinger et al. *Trends in Ecology & Evolution*, 31:737–738.

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.

Warwick, R., Clarke, K., and Suharsono (1990). A statistical analysis of coral community responses to the 1982–83 el niño in the thousand islands, indonesia. *Coral Reefs*, 8:171–179.

Welsh, A. H., Cunningham, R. B., Donnelly, C., and Lindenmayer, D. B. (1996). Modelling the abundance of rare species: statistical models for counts with extra zeros. *Ecological Modelling*, 88:297–308.

Yu, D. W., Ji, Y., Emerson, B. C., Wang, X., Ye, C., Yang, C., and Ding, Z. (2012). Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, 3:613–623.