JYU DISSERTATIONS 192

Jenni Niku

On modeling multivariate abundance data with generalized linear latent variable models

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF MATHEMATICS
AND SCIENCE

JYU DISSERTATIONS 192

Jenni Niku

On modeling multivariate abundance data with generalized linear latent variable models

Esitetään Jyväskylän yliopiston matemaattis-luonnontieteellisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Gamma-salissa
helmikuun 15. päivänä 2020 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Mathematics and Science of the University of Jyväskylä,
in building Agora, hall Gamma, on February 15, 2019 at 12 o'clock noon.

JYVÄSKYLÄN YLIOPISTO
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2020

Editors
Sara Taskinen
Department of Mathematics and Statistics, University of Jyväskylä
Timo Hautala
Open Science Centre, University of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-8062-7

ISBN 978-951-39-8062-7 (PDF)
URN:ISBN:978-951-39-8062-7
ISSN 2489-9003

ABSTRACT

Niku, Jenni
On modeling multivariate abundance data with generalized linear latent variable models
Jyväskylä: University of Jyväskylä, 2020, 54 p.(+included articles)
(JYU Dissertations
ISSN 2489-9003; 191)
ISBN 978-951-39-8062-7 (PDF)

The multivariate abundance data consist typically of multiple, correlated species encountered at a set of sites, together with records of additional covariates. When analysing such data, model-based approaches have been shown to outperform classical algorithmic-based dimension reduction methods. In this thesis we consider generalized linear latent variable models, which offer a general framework for the analysis of multivariate abundance data. In order to make the models more attractive among practitioners, new computationally efficient algorithms for the parameter estimation are developed by applying closed form approximation methods, the variational approximation method and the Laplace approximation method, for the marginal likelihood and by utilizing automatic differentiation tools when implementing the algorithms. The accuracy and computational efficiency of the methods are investigated and compared to existing methods through extensive simulation studies. The developed algorithms and additional tools implemented for model diagnosis, visualization and statistical inference are collected in `R` package `gllvm`. Several examples are provided to illustrate the use of the generalized linear latent variable models in ordination and when studying the between-species correlations and the effects of environmental variables, trait variables and their interactions on ecological communities.

Keywords: Community analysis, ecological data, fourth-corner models, generalized linear models, joint modeling, Laplace approximation, latent variables, multivariate analysis, ordination, species interactions, variational approximation

Author

Jenni Niku
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Supervisor

Docent Sara Taskinen
Department of Mathematics and Statistics
University of Jyväskylä
Finland

Reviewers

Associate Professor Silvia Cagnone
Department of Statistical Sciences
University of Bologna
Italy

Professor Philip M. Dixon
Department of Statistics
Iowa State University
USA

Opponent

Professor Jouni Kuha
Department of Methodology and Department of Statistics
London School of Economics
United Kingdom

TIIVISTELMÄ

Moniulotteinen runsausdata koostuu tyypillisesti useilta paikoilta tehdyistä eläin- tai kasvilajien havainnoista. Tällaiset aineistot ovat yleisiä ekologiassa, kun tutkitaan eläin-, kasvi- tai eliöyhteisöjä, niiden vuorovaikutusta keskenään tai vuorovaikutusta suhteessa ympäristöön. Perinteisesti moniulotteista runsausdataa analysoidaan käyttäen algoritmeihin perustuvia menetelmiä, kuten pääkoordinaattianalyysia, korrespondenssianalyysia ja ei-metristä moniulotteista skaalausta. Menetelmien tavoitteena on tiivistää aineiston pääpiirteet muutamaan muuttujaan, jotka on helppo esittää visuaalisesti johtopäätösten tekemiseksi. Algoritmisten menetelmien heikkoutena on se, että tulosten luotettavuutta on vaikea arvioida.

Tilastollisten ja laskennallisesti tehokkaiden menetelmien kehityttyä, malliperusteiset menetelmät ovat kasvattaneet suosiotaan moniulotteisien runsausdatojen analysoinnissa. Malliperusteiset menetelmät mahdollistavat aineiston rakenteiden, kuten lajien välisten korrelaatioiden sekä ympäristömuuttujien ja lajipiirteiden vaikutusten, tarkan mallintamisen. Aineistolle tyypilliset ominaisuudet voidaan ottaa huomioon esimerkiksi tilastollisten jakaumien avulla. Lisäksi mallipohjaiset menetelmät tarjoavat työkaluja tilastolliseen päättelyyn ja mallinvalintaan. Näiden ominaisuuksien seurauksena malliperusteiset menetelmät antavat luotettavampia tuloksia kuin algoritmeihin perustuvat menetelmät.

Tässä väitöskirjassa tutkitaan yleistettyjen lineaaristen latenttimuuttujamallien käyttöä moniulotteisen runsausdatan analysoinnissa. Yleistettyjen lineaaristen latenttimuuttujamallien sovittaminen on laskennallisesti erittäin raskasta, kun runsausdatojen lajimäärät ovat kovin suuria. Siksi tässä työssä kehitetään laskennallisesti tehokkaita algoritmeja mallin parametrien estimoimiseksi. Laskennallinen tehokkuus saavutetaan hyödyntämällä suljetun muodon approksimaatioita marginaaliselle uskottavuusfunktiolle sekä käyttämällä automaattisia differentiointityökaluja algoritmien implementoinnissa. Laskennallista tehokkuutta ja tarkkuutta tutkitaan simulointikokeiden avulla. Menetelmien soveltuvuutta ordinaatiomenetelmänä, lajien välisten korrelaatioiden mittaamisessa, ympäristömuuttujien, lajipiirteiden ja niiden välisten interaktioiden vaikutusten tutkimisessa ja testaamisessa havainnollistetaan useiden esimerkkien avulla. Mallin sovittamiseen kehitetyt algoritmit sekä työkaluja mallien diagnostiikkaan, testaukseen ja visualisointiin on koottu `R` pakettiin `gllvm`.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor, Docent Sara Taskinen, for encouraging me to begin my graduate studies. I am very thankful for your guidance during these years and the time you invested in this project.

I am very grateful to Professor David Warton for collaboration and guidance throughout my PhD studies as well as inviting me to visit the University of New South Wales. I would also like to thank Dr. Francis Hui for the collaboration and helping me to overcome obstacles I encountered in my research. My thanks also go to Wesley Brooks and Riki Herliansyah for the co-operation in developing the computational parts of the gllvm software and for the collaboration in the second article. I wish to thank Emeritus Professor Antti Penttinen for helping me to finalize my dissertation by reading my work and by giving valuable comments on it.

My sincere thanks go to Dr. Manoj Kumar and Docent Riitta Nissinen for providing us the plant-microbial data, and for Dr. Emmanuela Daza Secco, Docent Jari Haimi and Docent Kristian Meissner for allowing us to use the testate amoebae data in our analyses.

I would like to express my gratitude for Associate Professor Silvia Cagnone and Professor Philip Dixon for agreeing to work as pre-examiners, and for Professor Jouni Kuha for serving as the opponent.

I am very grateful for the Jenny and Antti Wihuri Foundation and the Finnish Cultural Foundation for the financial support during my PhD studies. I wish to thank the Department of Mathematics and Statistics at the University of Jyväskylä for financial support and for providing the working facilities during my studies. My thanks also go to all the colleagues at the department for creating a warm and supportive atmosphere to work at.

Finally, I would like to warmly thank my parents and siblings for their support throughout my life.

Jyväskylä, January 2020

Jenni Niku

CONTENTS

LIST OF INCLUDED ARTICLES

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

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

PIII Niku, J., Hui, F. K., Taskinen, S. and Warton, D. I. Analysing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Submitted*, 2020.

PIV Niku, J., Hui, F. K., Taskinen, S. and Warton, D. I. gllvm - Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution, 10:2173–2182*, 2019.

Author's contributions

The author of this thesis has actively contributed to the research of the joint articles PI, PII, PIII and PIV. In particular, the author performed analyses, simulations and visualizations for all of the included articles and was responsible for implementing the algorithms used in the articles and in the `R` package `gllvm` introduced in article PIV. The author derived theoretical results for articles PI, PII and PIII, drafed all articles and contributed in reviewing and editing them. The original ideas for articles and research questions were formulated together with the co-authors.

1 INTRODUCTION

Multivariate abundance data are often analysed in ecological community studies. Such data typically consist of records of a large number of interacting species at a set of sites, accompanied with records of predictive covariates regarding the environmental variables or species related traits. Interest may often be in studying if the species communities differ between sites, studying between species correlations, hypothesis testing of environmental or trait effects, or making predictions for abundances.

Traditionally multivariate abundance data have been studied using classical algorithm-based methods, which often focus on producing ordination of the data, that is, visualizing multivariate abundance data in small dimensional form in order to make interpretations on the data structures. Recent methodological and technological developments, however, have allowed one to specify statistical models for the multivariate abundance data in order to capture the between species correlations and to make proper inference on the effect of predictors on communities, for instance. For model-based approaches to multivariate abundance data see, for example, Ovaskainen et al. (2010); Walker and Jackson (2011); Jamil et al. (2012); Vanhatalo et al. (2012); Brown et al. (2014); Hui et al. (2015); Thorson et al. (2015); Clark et al. (2018). With the ability to properly take into account data properties, such as mean-variance relationship in responses and correlation structures in the data, joint models have shown to outperform classical algorithmic-based dimension reduction methods by producing more reliable results when applied, for example, in ordination (Hui et al., 2015).

In this thesis we consider the analysis of multivariate abundance data using generalized linear latent variable models (GLLVMs Moustaki and Knott, 2000). The GLLVMs are based on the multivariate generalized linear models, which are extended by including latent variables to the model in order to take into account correlation structures in data. If multivariate abundance data are very high-dimensional, computational challenges are often encountered when fitting GLLVMs. In order to overcome those challenges, we develop computationally efficient estimation methods and algorithms for the model fitting by using two closed form approximations for the marginal likelihood, the Laplace approxima-

tion and the variational approximation, and making use of automatic differentation tools. The performance of the developed estimation methods and algorithms are compared to each other and to existing model fitting methods for GLLVMs using simulation studies. In addition to the computational developments, we study and illustrate the use of GLLVMs in ordination, studying between species correlations, and hypothesis testing of the environmental and trait interactions by analysing several typical multivariate abundance datasets. We also develop easy-to-use, freely available `R` package `gllvm` in order to offer computationally efficient methods for fitting GLLVMs for multivariate abundance data. Tools for inference, model checking and visualization are also provided in the `R` package.

The outline of this thesis is as follows. In Chapter 2 we describe multivariate abundance data in ecology and provide an overview on the classical methods for analysing them. In Chapter 3 we review the model-based approaches and formulate joint models for multivariate data focusing on generalized linear latent variable models. The maximum likelihood estimation and inference of the latent variable models as well as algorithms and software for model fitting are reviewed in Chapter 4. In Chapter 5 we compare the performance of the developed algorithms to several existing `R` packages for fitting GLLVMs, and show that the developed methods are computational efficient and do not loose in accuracy when compared to the most accurate competitors. In Chapter 6 we illustrate the methods via case study.

2 MULTIVARIATE ABUNDANCE DATA

Multivariate abundance data are widely used in ecological community studies and consist of records of multiple response variables measured at a set of observational units. The data are typically organized as a matrix, where rows correspond to observational units, that is, different sites or sampling units, and columns correspond to response variables such as records of interacting species. The records of responses measuring the abundance can be, for example, counts of the observed species, information about the presence of the species as a binary data or, when object of interest is not individually countable, biomass of the organic matter. As an example, below we consider the subset of the testate amoebae dataset of size 270 x 50 (Daza Secco et al., 2016), which consists of counts of $m = 50$ testate amoebae species recorded at 270 sampling sites at Finnish peatlands. This dataset will be considered in the example abalysis in Chapter 6 and in the simulation studies in article PII.

Cenacu	Cencas	Ceneco	Cenpla	Cycarc	Triarc	Trimin	Arccat
0	31	0	0	6	26	19	3
0	63	0	0	0	10	0	13
6	4	16	0	5	29	14	54
0	18	0	0	0	17	11	14
8	0	0	0	0	2	0	129

Another examples of multivariate abundance data considered in this thesis include a microbial community data consisting of counts of 985 bacteria species measured from 56 soil samples collected from high altitude soils in Europe (Nissinen et al., 2012), which were analysed in article PI. An example of biomass data, which was also considered in article PI, is the abundances of 18 coral reef species at 19 sites in Indonesia (Warwick et al., 1990), measured as a length of a ten metre transect which intersected with the species.

The testate amoebae data described above is low-dimensional having only 50 species measured at 270 sites. We thus have $m < n$. This is, however, often not the case as high-dimensionality is characteristic for multivariate abundance data. The data may include measurements from hundreds or thousands

of species and the number of species may exceed the number of sites. Such data are often encountered, when modern tools, such as metabarcoding, are used for species identification. The microbial community data in article PI serves as an example. Other typical features for abundance data are a large number of zeros and overdispersion as the species often tend to be found with large numbers or not at all. When looking at testate amoebae data, species named Cenpla seems to be pretty rare whereas Arccat is higly abundant. In addition, species are correlated due the biotic interactions, phylogeny, behavioral and biological traits and environmental conditions (Araújo and Luoto, 2007; Morales-Castilla et al., 2015). These features, which need to be taken into account when analysing multivariate abundance data, posses challenges for the analysis.

In ecological community studies one might, for example, want to study if species compositions differ in a set of sites (Björk et al., 2018; Daza Secco et al., 2018, article PI), or between species interactions (Royan et al., 2016; Inoue et al., 2017). Interactions between species can often be explained bybiological and environmental conditions. Such conditions can be defined by a large number of influencing variables which are not necessarily observable. For example, some species tend to occur at same sites as they prefer similar environmental conditions and therefore those conditions can explain the similarities and differences in species compositions at different sites. Moreover, behavioral or biological traits related to species can also mediate the effect of the environmental conditions and partly explain the observed species compositions. Such relationships in addition to the relationships between species may often be the goal of the study, and therefore data often include environmental covariates related to sites and trait covariates related to species. As an example, the effect of natural environmental or experimental conditions to the ecological communities were studied in Lammel et al. (2018) and in Daza Secco et al. (2018). In article PI we studied if environmental covariates pH value and available phosphorus affected the species compositions of microbial communities in soils. In Ribera et al. (2001) it was studied whether the species trait variables affect the species communities or if they interact with the effect of the environmental variables. Predicting species abundances may also be one of the main interests in community studies (Buisson et al., 2008).

2.1 Overview of classical methods

Classically multivariate abundance data have been analysed using algorithm-based methods, which are often based on the analysis of the association matrices, that is, matrices of pairwise distances or dissimilarity measures between sites or between species. By the dissimilarity measure we mean a value of a function measuring a dissimilarity level between two objects. Distance is a metric dissimilarity measure. Such algorithm-based methods usually focus on producing an ordination of the data, a low dimensional expression of the data which can be used to visualize the main patterns between sites in terms of their species composition.

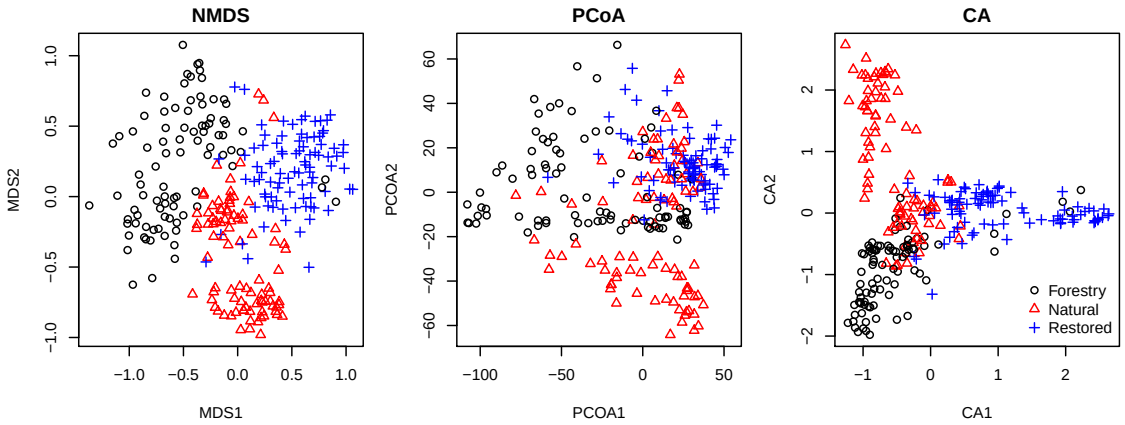

FIGURE 1 Ordination plots for the testate amoebaedata based on non-metric multidimensional scaling (NMDS) with Bray-Curtis distance, principal coordinate analysis (PCoA) and correspondence analysis (CA). The colors and symbols represent three types of peatlands in terms of the land use: peatlands used for forestry, natural peatlands and restored peatlands.

Among the most well-known ordination methods are non-metric multidimensional scaling (nMDS, Kruskal, 1964a,b), principal coordinate analysis (PCoA, Gower, 1966) and correspondence analysis (CA, Hill, 1974; Hill and Gauch Jr, 1980). For a review of classical methods for analysing multivariate abundance data, see Legendre and Legendre (2012).

As classical ordination methods are based on the association matrix, the results of the ordination depend heavily on the choice of the dissimilarity measure. In ecology, the choice of the dissimilarity measure is usually based on the previous studies (Faith et al., 1987) and therefore the reasoning of the choice relies on past empirical performance rather than the data at hand. In Warton et al. (2012) and in Warton and Hui (2017), it was shown that the performance of ordination methods depends strongly on data properties, in particular, the mean-variance relationship, and if such properties are not properly accounted for by the choice of a dissimilarity measure, the results might potentially be misleading.

Consider as an example ordination plots for the testate amoebae dataset (Daza Secco et al., 2016) based on nMDS, PCoA and CA in Figure 1. The first two methods are based on the Bray-Curtis and euclidean distances, respectively, and the third method applies a Chi-square transformation to the data matrix before using singular value decomposition to produce an ordination. The obtained ordination results are very different. The testate amoebae species were measured at three types of peatlands in terms of the land use: peatlands used for forestry, natural peatlands and restored peatlands. In the ordination plot based on the nMDS method the points are clustered according to the land use very clearly, while in the ordination plots based on PCoA and CA the samples from different types of peatlands are more mixed. Unfortunately we have no tools to evaluate which of the results describes the data best.

In order to explain ecological stuctures such as between species interactions, an analysis of the effects of explanatory variables related to observational

units or response variables is usually done using indirect or direct comparison approaches. In indirect comparisons, the effects of environmental variables are compared to the ordination axes of the abundance matrix instead of the actual response variables using correlations or regression (Legendre and Legendre, 2012, Chapter 10). Such analysis is also known as indirect gradient analysis (Jongman et al., 1987). In direct comparisons, multiple data tables, which usually are a matrix of abundances and a matrix of environmental covariates related to sites, are simultaneously analysed in order to produce an ordination with the effects of explanatory variables already taken into account. The direct comparison approach is also known as canonical analysis or direct gradient analysis (Legendre and Legendre, 2012, Chapter 11). The most well-known methods for such analyses include redundancy analysis (RDA, Legendre and Anderson, 1999) and canonical correspondence analysis (CCA, ter Braak, 1986).

In addition to the direct and indirect gradient analysis, the interest may be in understanding how behavioral and biological traits of species mediate the effect of the environmental conditions on species, and further in testing whether those associations are significant. In literature, this problem is known as the fourth-corner problem (Legendre et al., 1997). The most well-known method for studying such relationships between environmental and trait variables is ordination based method RLQ proposed by Dolédec et al. (1996). In the RLQ method, matrices of environmental covariates ($\boldsymbol{R}$), species abundance data ($\boldsymbol{L}$) and species trait covariates ($\boldsymbol{Q}$) are used to produce a matrix for describing associations between environmental and trait variables. After that the association matrix is used to produce a pair of ordinations using a singular value decomposition in order to make interpretations of the effects of the environmental and trait variables on species. Another widely used approach has been hypothesis testing via permutation test, see Legendre et al. (1997), Dray and Legendre (2008) and ter Braak et al. (2012). These methods evaluate a significance of associations between environmental and trait covariates by permutation tests.

3 JOINT MODELS FOR MULTIVARIATE ABUNDANCE DATA

Methodological and technological developments conducted over the past few decades allow us to specify statistical models for high-dimensional data, jointly across many responses. Joint models are extensions of generalized linear models (GLM, McCullagh and Nelder, 1989) for multivariate data that account for the correlation structure inherent in data simultaneously with the effect of predictors on responses. The joint models have shown to be a powerful approach for analysing multivariate abundance data as they can be used to answer a wide variety of ecological questions, such as studying between species correlations (Ovaskainen et al., 2010; Pollock et al., 2014), ordination (Walker and Jackson, 2011; Hui et al., 2015), making inferences about the effect of predictors based on confidence intervals and hypothesis testing (Jamil et al., 2012; Lammel et al., 2018), and making predictions for abundances (Ovaskainen et al., 2016c; Schliep et al., 2018).

As a model-based approach, joint models provide tools for evaluating the suitability of the model for the data at hand. In particular, residual analysis can be used to check if the chosen model and distribution are able to capture the mean-variance relationship of the data. Tools for statistical inference, prediction and model selection, such as information criterias, are also readily available. On the contrary, the classical algorithm-based approaches often lack such tools (see for example Warton et al., 2012). The model-based approaches have shown to outperform the classical algorithm-based methods in the analysis of community data, for example in ordination (Hui et al., 2015; Hui, 2017) and clustering (Hui, 2017).

Consider next the abundanes (*e.g.* counts, presence-absences, biomass) of m responses (species) recorded at n observational units (sites) which we denote by y_{ij}, $i = 1, \dots, n$, $j = 1, \dots, m$. For each observational unit i, a vector of k environmental variables, $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ik})'$, may also be recorded. In multivariate GLMs, the mean response, $\mu_{ij} = E(y_{ij})$, is regressed against a vector of covariates,

that is,

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

where $g(\cdot)$ is a known link function, β_{0j} are species-specific intercepts and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jk})'$ are species-specific regression coefficients related to the covariates.

Independency of the observational units (sites) can often be assumed by the design, but such independency of the species responses within a unit cannot be assumed due the correlation between species caused by ecological reasons as discussed in Chapter 2. Generalized linear models do not take into account any correlation between response variables that is not explained by covariates, and ignoring or misspecification of the correlation between species may result in inflated Type I errors (ter Braak et al., 2017), biased estimates and too narrow confidence intervals (Warton et al., 2015, 2016). Moreover, this can lead to false conclusions when assessing the significance of predictors in the model. Our simulation studies in article PI showed that ignoring between species correlation may cause biased estimates for the effect of the environmental variables. In article PIII we observed inflated Type I errors when the correlation structure was misspecified. In the context of multivariate GLM, the between species correlation has been earlier taken into account by using resampling techniques, such as bootstrapping (Warton, 2011; Brown et al., 2014) or permutations (ter Braak et al., 2017) in hypothesis testing. Notice however that the resampling techniques may often be very time consuming and perform worse than proper joint modeling approaches as shown in our simulation studies in article PIII. In the next sections, we consider different approaches for modeling the between species correlations within the joint models.

3.1 Generalized linear mixed models

At first, we consider models that include environmental covariates only. In order to capture the correlation structures in responses, inclusion of random effects in the model is necessary (Bolker et al., 2009). Correlation in responses can be handled in different ways, and one simple way is to include a common univariate site-specific random effect in the model (Jamil et al., 2013) in order to incorporate correlation between species. Such generalized linear mixed model (GLMM, Breslow and Clayton, 1993) for the mean response can be written as

$$g(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + u_i, \tag{2}$$

where the univariate random intercept u_i is assumed to be normally distributed with zero mean and variance σ^2. Notice however, that inclusion of a common random intercept for responses creates only a constant positive covariance between species which is certainly not a valid assumption for ecological data.

A more complicated way to create correlation structure between species is to do it straightforwardly by including site- and species-specific random effect,

u_{ij}, in model (1). This approach has been quite popular in joint modeling and has been considered in Ovaskainen et al. (2010); Clark et al. (2014) and Pollock et al. (2014), for example, and reviewed by Warton et al. (2015). Multivariate generalized linear mixed model for the mean response can be defined as,

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + u_{ij}. \tag{3}$$

Such random effects can be considered as multivariate random effects related to sites, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{im})$, which are assumed to be multivariate normal distributed with zero mean vector and unstructured covariance matrix $\boldsymbol{\Sigma}$. If the focus is on modeling relative abundance or compositional effects, the fixed site effects α_i can be included in the model for site total abundance standardization. The correlation between responses is controlled with a unstructured covariance matrix $\boldsymbol{\Sigma}$. The model offers a flexible framework for accounting for any correlation structure for small dimensional data, but it becomes computationally burdensome when the number of responses is large due to quadratically (with m) increasing number of parameters, $m(m+1)/2$, in the covariance matrix.

3.2 Generalized linear latent variable models

A more advanced approach to capture the correlation in responses is to include one or several latent variables with corresponding factor loadings in the model. This provides a flexible method for modeling any residual correlation between species not accounted for by the covariates. In generalized linear latent variable models (GLLVMs, Moustaki and Knott, 2000), the mean response μ_{ij} is regressed against a vector of $d \ll m$ latent variables, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})$, along with the k-vector of covariates $\boldsymbol{x}_i$, if available. The model can thus be written as

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \tag{4}$$

where β_{0j} and $\boldsymbol{\beta}_j$ are as in (1). The site effects α_i can be treated as fixed, as in (3), or random with a zero mean and a variance σ^2. The latter is more advisable as we noticed in article PI that the estimates of fixed site effects tend to be biased. Parameters $\boldsymbol{\gamma}_j$ are species-specific loadings for the latent variables. The latent variables are assumed to be independent across sites and standard normally distributed, $\boldsymbol{u}_i \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$, so that the zero mean and unit variance assumption fix the locations and scales of the latent variables. In order to ensure that the model is identifiable, for $m > 1$ the upper triangular of the loading matrix, $\boldsymbol{\Gamma} = [\gamma_1 \ldots \gamma_m]'$, needs to be set to zero and the diagonal elements are set positive to avoid rotational invariance (Huber et al., 2004).

In model (4) the latent variable term $\boldsymbol{u}_i'\boldsymbol{\gamma}_j$ captures any residual correlation across species not accounted for by the observed covariates $\boldsymbol{x}_i$. Further, the residual covariance matrix storing information on species co-occurrence can be calculated as $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, being of rank d (Warton et al., 2015). However, it should

be noted that this definition is on the scale of linear predictors, $\eta_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$, so it does not take into account the distribution-specific aspects such as overdispersion in the negative binomial distribution. With some correction terms such properties can be accounted for (article PIV and Ovaskainen et al., 2016a). As compared to the multivariate GLMM as defined in (3), a key feature of the GLLVM (4) is that the number of parameters, $md - d(d-1)/2$, characterizing the residual correlation, grows linearly along the number of responses m imposing a factor analytic structure for the covariance matrix.

In articles PI, PII and PIV we focus on GLLVMs as defined in (4). Article PI considers an analysis of overdispersed count and biomass data in ecology using GLLVMs. Computational aspects were considered in PII and the software for fitting the GLLVMs in PIV, respectively. In the literature, some form of the generalized linear latent variable models have been considered in several occasions. For example, item response models (IRT, see eg. Bartholomew et al., 2011), which are defined as GLLVMs for binomial and ordinal responses, were considered in Bock and Aitkin (1981). The GLLVMs in the context of ecological studies, as defined in (4), were considered by Skrondal and Rabe-Hesketh (2004); Warton et al. (2015) and Ovaskainen et al. (2017), for instance. In addition, the GLLVMs have also been extended to the case of more complex correlation structures such as temporally and/or spatially varying correlation structures for example in Thorson et al. (2015); Ovaskainen et al. (2016a); Thorson et al. (2016, 2017) and Tikhonov et al. (2019a).

3.3 Fourth-corner models

Let us next consider solutions to the fourth-corner problem using joint models. Assume that q trait covariates, $\boldsymbol{t}_j = (t_{i1}, \ldots, t_{iq})'$, in addition to the abundances y_{ij} and environmental covariates $\boldsymbol{x}_i$ are also recorded. The trait covariates can be used to explain interspecific variation in the environmental response as follows. A simple model-based approach to the fourth-corner problem was proposed by Brown et al. (2014) based on the multivariate generalized linear model. In their approach the associations between environmental and trait variables are studied by regressing the mean response μ_{ij} against the interaction between environmental and trait covariates, along with the intercept, environmental and trait variables. This, so called "fourth-corner model", for the mean responses μ_{ij} can be written as

$$g(\mu_{ij}) = \beta_0 + \boldsymbol{x}_i'\boldsymbol{\beta}_e + \boldsymbol{t}_j'\boldsymbol{\beta}_t + vec(\boldsymbol{B}_{te})'(\boldsymbol{x}_i \otimes \boldsymbol{t}_j), \tag{5}$$

where β_0 is a common intercept, $\boldsymbol{\beta}_e$ and $\boldsymbol{\beta}_t$ are vectors of the main effects for environmental covariates and trait covariates, respectively, and the $k \times q$ matrix $\boldsymbol{B}_{te}$ consists of environmental-trait interaction terms, also known as the fourth-corner coefficients. The fourth-corner coefficients explain how the species-specific functional or biological traits mediate the effect of the environmental variables. If the primary purpose is in the study of the mediating effect of the traits on the

environmental covariates across species, the main effects of the traits $\boldsymbol{\beta}_t$ can be absorbed by the species-specific intercept terms β_{0j}.

The model (5) does not take into account the between species correlations or interspecific variation in responses with respect to environmental variables that are not explained by observed trait variables. Ignoring or misspecifying the correlation may yield to biased estimates and inflated Type I errors when testing for the significance of the fourth-corner term (ter Braak, 2019; Miller et al., 2019). As a solution to the problem of inflated Type I errors, resampling techniques are proposed in several papers to complement the method when the significance of the interaction term between environmental and trait variables has been tested. For instance, Brown et al. (2014) proposed a method that bootstraps the residuals, Shipley (2010) proposed permutations for species traits and ter Braak et al. (2017) considered permutations for species and sites. ter Braak (2019) studied bootstrap and permutation methods when applied to the fourth-corner model (5) accompanied with species-specific random effects. However, as it was shown in ter Braak et al. (2017) resampling based solutions do not guarantee valid Type I errors for the bootstrapping method of Brown et al. (2014) and some permutation designs. In addition, such methods are often very time consuming and do not fix the problem of biased estimates (ter Braak, 2019) caused by the misspecified correlation in the model.

Similarly to the models that use only environmental variables as explanators to the communities, correlation structures can be incorporated in the fourth-corner models by including latent variables or random effects. Several different forms of the fourth-corner models have been proposed in the literature. Pollock et al. (2012) and Jamil and ter Braak (2013) included species-specific random intercepts and species-specific random slopes for environmental variables. Later, Warton et al. (2015) included latent variables to capture the correlation between species that is not explained by the covariates.

In article PIII, we considered an extension of the model (5), which was already used by Warton et al. (2015) in their example. A general expression of the fourth-corner latent variable model is given as

$$\begin{aligned} g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \text{ where } (r_i, \boldsymbol{u}_i')' \sim N_{d+1}(\mathbf{0}, \boldsymbol{C}_\sigma) \\ \boldsymbol{\beta}_j &= \boldsymbol{\beta}_e + \boldsymbol{B}_{te}\boldsymbol{t}_j + \boldsymbol{b}_j, \text{ where } \boldsymbol{b}_j \sim N_k(\mathbf{0}, \boldsymbol{G}). \end{aligned} \tag{6}$$

Here the parameters β_{0j} and r_i are species-specific intercepts and site-specific random intercepts, respectively. The main effects for environmental covariates, $\boldsymbol{\beta}_e$, and the fourth-corner interactions, $\boldsymbol{B}_{te}$, are as before. Similarly to the GLLVM defined in (4), we include the latent variables to capture the residual correlation between responses not accounted for by the observed covariates $\boldsymbol{x}_i$ and trait variables $\boldsymbol{t}_j$. In addition, the k-vector $\boldsymbol{b}_j = (b_{j1}, \ldots, b_{jk})'$ includes species-specific random effects for the environmental variables, which are assumed to follow a multivariate normal distribution with a zero mean vector and an unstructured $k \times k$ covariance matrix $\boldsymbol{G}$. The effect of the predictors is then a combination of the fixed effects $\boldsymbol{\beta}_e$ common for all species, the interaction terms with species traits

$\boldsymbol{B}_{te}$ to define how the traits mediate the effect of the environmental variables, and the species-specific random effects $\boldsymbol{b}_j$ to capture the interspecific variation that is not explained by the trait covariates. We include the correlation term between random site effects and latent variables, that is $corr(r_i, u_{il}) = \rho_l$, and we denote $Cov((r_i, \boldsymbol{u}_i')') = \boldsymbol{C}_\sigma$. If the main effect of the traits on responses is an object of interest, species-specific random intercepts β_{0j} can be replaced by $\beta_0 + \boldsymbol{t}_j'\boldsymbol{\beta}_t$ as in the model (5).

The model (6) above is an unifying framework that encompasses all the models of Brown et al. (2014); Pollock et al. (2012) and Jamil and ter Braak (2013). The model (6) can be reduced to model (5) used in Brown et al. (2014) by setting all variances for the random effects and latent variables to zero. The model of Pollock et al. (2012) is obtained by setting all elements in $\boldsymbol{C}_\sigma$ to zero, with an exception that in Pollock et al. (2012) intercepts β_{0j} were treated as random. Jamil and ter Braak (2013) used the same model with random site effects.

In article PIII, we compared the performance of the model (6) and the model of Jamil and ter Braak (2013) when testing significance of the fourth-corner terms using the likelihood ratio test. We noticed that the model (6) gave better Type I errors when species were correlated and an interspecific variation, not explained by traits, was inherent in data. In addition, the likelihood ratio test for the model (6) produced higher power for the test than permutation based method proposed by ter Braak et al. (2017).

4 ESTIMATION AND INFERENCE

The multivariate abundance data in ecology are often high-dimensional and may consist of abundances of hundreds or even thousands of species. This may cause computational challenges in model fitting as the number of parameters in joint models is large. As an example, in our case study in article PI which considered microbial community data with $m = 985$ species and $n = 56$ sites along with three environmental variables, the number of parameters in a generalized linear latent variable model was 6951. The computational efficiency in parameter estimation is thus an important requirement for fitting joint models for multivariate data.

In this thesis, one of the major goals was to develop computationally fast algorithms for fitting GLLVMs. In the next sections we consider the maximum likelihood estimation of the latent variable models, and provide an overview for fast methods and algorithms for the estimation and inference.

4.1 Maximum likelihood estimation of latent variable models

Consider the model estimation via maximum likelihood estimation for the generalized linear latent variable model as defined in (4), and assume for simplicity that the site effects are fixed. Write $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)'$, $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})'$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1' \ldots \boldsymbol{\beta}_m')'$, and let $\boldsymbol{\Phi}$ include all nuisance parameters, *e.g.* dispersion parameters in negative binomial distribution. Let then $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \boldsymbol{\beta}, \mathrm{vec}(\boldsymbol{\Gamma}), \boldsymbol{\Phi})$ include all model parameters as a vector. We also collect all latent variables into a vector $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_n')'$. The observational units are assumed to be independent, and the responses, $y_{i1}, \ldots, y_{im}$, at site i are assumed to be independent conditionally on latent variables, $\boldsymbol{u}_i$. By denoting the conditional probability density function for the response with $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$ and the distribution of the latent variable vector with $f(\boldsymbol{u}_i)$, that is, the density function of the multivariate normal

distribution, the complete likelihood for the GLLVM can be written as

$$L(\boldsymbol{\Psi}; \boldsymbol{u}) = \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) \right) f(\boldsymbol{u}_i) = f(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\Psi}) f(\boldsymbol{u}), \tag{7}$$

where $f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$ and $f(\boldsymbol{u}) = \prod_{i=1}^{n} f(\boldsymbol{u}_i)$ are the joint distributions of the response variables and the latent variables, respectively. Estimates for the model parameters $\boldsymbol{\Psi}$ can be obtained by maximizing the likelihood (7). However, as the likelihood function depends on the distribution of the unknown latent variables, the maximization cannot be done straightforwardly. In general, this estimation problem has attracted much attention in the statistical literature, and below we provide an overview of methods often applied for fitting GLLVMs.

One of the most well-known solutions for maximizing the complete likelihood function in incomplete data problems such as situations including unknown latent or random variables or missing data is to apply the Expectation-Maximization algorithm (EM algorithm, Dempster et al., 1977). The EM algorithm or some variant of it was applied to latent variable models in Sammel et al. (1997) and Hui et al. (2015), and earlier also in Bock and Aitkin (1981) for item response models. Although the EM algorithm is easy to implement, a major downside is the computational inefficiency as the algorithm usually converges very slowly.

Another popular approach for estimating models with latent or random variables is Bayesian Markov Chain Monte Carlo (MCMC, Metropolis et al., 1953; Hastings, 1970) sampling based on the complete likelihood function. Bayesian approach has been very popular in estimating hierarchical models in community level modeling in ecology (see for example Cressie et al., 2009; Ovaskainen et al., 2016b). In case of GLLVMs, Bayesian MCMC estimation is used in Ovaskainen et al. (2016a) and Hui et al. (2017). The advantage of the approach is the ability to apply it to very complex models, that is, in cases where many other methods are unfeasible. Based on the simulated values any characteristics of the posterior distribution can also be studied. Notice however that the method is computationally very intensive and the convergence might be hard to monitor (Gelman and Rubin, 1996) especially when posterior distributions for a large number of parameters need to be simultaneously estimated.

Computationally the most efficient methods for fitting models with latent variables are those that approximate the marginal likelihood in a closed form. Because the latent variables are treated like random variables, they can be integrated out of the likelihood and the inference can be based on the marginal likelihood only. By integrating over the latent variables $\boldsymbol{u}$ in (7) we obtain the marginal log-likelihood function for the GLLVM

$$\log L(\boldsymbol{\Psi}) = \log \int_{R^d} f(\boldsymbol{y}|\boldsymbol{u}; \boldsymbol{\Psi}) f(\boldsymbol{u}) d(\boldsymbol{u}). \tag{8}$$

In case of non-normal data the d-dimensional integration over latent variables cannot be solved analytically, and the marginal log-likelihood does not have a closed form solution. Several closed form approximation methods for such an integral have therefore been proposed in the literature in the context of random effect models. These include numerical integration using Gauss-Hermite quadrature (Bock and Lieberman, 1970; Bock and Aitkin, 1981; Butler and Moffitt, 1982) and adaptive Gauss-Hermite quadrature (Naylor and Smith, 1982; Liu and Pierce, 1994), Laplace approximation (Tierney and Kadane, 1986), and more recently, variational approximation method (Ormerod and Wand, 2010, 2012). In the next sections we review some closed form approximations used to fit GLLVMs in the literature.

4.2 Gauss-Hermite quadrature

The Gauss-Hermite quadrature is a numerical integration method that approximates an integral of the form $\int f(u)\exp(-u^2)du$ using a weighted sum of evaluations of the function $f(\cdot)$ at a set of locations a_r, $r = 1,\dots,R$, for the variable u. Since the function $\exp(-u^2)$ is proportional to the normal density, the approach can be applied for approximating integrals of normally distributed latent variables. Let us next consider the Gauss-Hermite quadrature for a simple generalized linear latent variable model, $g(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\gamma_j$. As the latent variables are assumed to be uncorrelated and jointly normally distributed, the marginal log-likelihood can be written as

$$\begin{aligned} l(\boldsymbol{\Psi}) &= \sum_{i=1}^{n} \log \int \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\Psi}) f(\boldsymbol{u}_i) d\boldsymbol{u}_i \\ &= \sum_{i=1}^{n} \log \int \cdots \int \prod_{j=1}^{m} f(y_{ij}|u_{i1},\dots,u_{id};\boldsymbol{\Psi}) f(u_{i1}) \cdots f(u_{id}) du_{i1} \cdots du_{id}, \end{aligned}$$

where $f(u_{il})$ is a density of the univariate standard normal distribution. Changing variables from u_{il} to $v_{il} = u_{il}/\sqrt{2}$ and applying the quadrature rule, the marginal likelihood obtains an approximation

$$l(\boldsymbol{\Psi}) \approx \sum_{i=1}^{n} \log \sum_{r_d=1}^{R} p_{r_d} \cdots \sum_{r_1=1}^{R} p_{r_1} \prod_{j=1}^{m} f(y_{ij}|a_{r_1},\dots,a_{r_d};\boldsymbol{\Psi}),$$

where R is the number of quadrature points, weights p_r are given by

$$p_r = \frac{2^{R-1}R!}{R^2[H_{R-1}(a_{ir})]^2}$$

and a_r are the roots of the Hermite polynomial $H_R(v) = (-1)^n e^{v^2}\frac{d^R}{dv^R}e^{-v^2}$. These points construct a rectangular grid for the latent variables where the function is

evaluated at. The original idea was proposed by Bock and Lieberman (1970) and the method was used within an EM algorithm for approximating integrals based on probit item response models in Bock and Aitkin (1981) and for regression models with a random intercept in Butler and Moffitt (1982). Later, in Moustaki (1996) and Moustaki and Knott (2000), GLLVMs for mixtures of binary and normal responses were fitted using Gauss-Hermite quadrature.

The function to be integrated, the product $\prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i;\boldsymbol{\Psi})$, has often a high peak and the quadrature points assess poorly the integral (Lesaffre and Spiessens, 2001). In order to obtain a sufficiently accurate approximation to the marginal log-likelihood, the Gauss-Hermite quadrature often needs a lot of quadrature points and therefore the method becomes computationally burdensome. Better performance can be provided using the extension of the Gaussian quadrature method, called adaptive Gauss-Hermite quadrature, wich was first used in the Bayesian inference by Naylor and Smith (1982). The adaptive quadrature method applies the importance sampling technique for choosing the quadrature points more carefully in order to cover better the high density area of the function to be integrated (Pinheiro and Bates, 1995). Fewer quadrature points are thus needed. A normal density $\varphi(u_{il}, \nu_{il}, \tau_{il}^2)$ approximates the posterior density of u_{il} and is used as the importance sampling density. Here ν_{il} and τ_{il}^2 are the posterior mean and the variance. The log-likelihood function $l(\boldsymbol{\Psi})$ can then be written as

$$l(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \int \cdots \int \left\{ \frac{\prod_{j=1}^{m} f(y_{ij}|u_{i1},\ldots,u_{id};\boldsymbol{\Psi}) f(u_{i1}) \cdots f(u_{id})}{\varphi(u_{i1},\nu_{i1},\tau_{i1}^2) \cdots \varphi(u_{id},\nu_{id},\tau_{id}^2)} \right\} \times \varphi(u_{i1},\nu_{i1},\tau_{i1}^2) \cdots \varphi(u_{id},\nu_{id},\tau_{id}^2) du_{i1} \cdots du_{id}.$$

By changing variables from u_{il} to $v_{il} = (u_{il} - \nu_{il})/\tau_{il}$ and applying the quadrature rule, integral obtains an approximation

$$l(\boldsymbol{\Psi}) \approx \sum_{i=1}^{n} \log \sum_{r_d=1}^{R} w_{ir_d} \cdots \sum_{r_1=1}^{R} w_{ir_1} \prod_{j=1}^{m} f(y_{ij}|a_{ir_1},\ldots,a_{ir_d};\boldsymbol{\Psi}) =: \tilde{l}(\boldsymbol{\Psi}),$$

where $a_{ir_l} = \nu_{il} + \tau_{il} a_r$ and weights w_{ir_l} are given by

$$w_{ir_l} = \sqrt{2\pi}\tau_{il} \exp(a_r^2/2) \varphi(\nu_{il} + \tau_{il} a_r) p_r.$$

For comparison of the Gauss-Hermite quadrature and adaptive quadrature methods, see Pinheiro and Bates (1995) and Rabe-Hesketh et al. (2005). Even if the method does not require as many quadrature points as the Gauss-Hermite quadrature, it still becomes computationally impractical when the number of latent variables d is moderate, even when $d > 2$. Rabe-Hesketh et al. (2002) applied the adaptive quadrature for the GLLVMs with normal, binomial, gamma, and Poisson distributed responses. Notice that the method of Rabe-Hesketh et al. (2002) for latent variable models is only available in the proprietary software `STATA`.

Asymptotic properties of the adaptive quadrature maximum likelihood es-

timators for GLLVMs were studied in Bianconcini (2014). They showed that the estimators are asymptotically consistent with rate $O_p(\max(n^{-1/2}, m^{-(R/3+1)}))$ and asymptotically normally distributed. The covariance matrix for the estimated parameters can be approximated using an inverse of the observed information matrix.

4.3 Laplace approximation

The Laplace approximation (Tierney and Kadane, 1986) is the most common and well-known method that provides a closed form approximation for the marginal likelihood. The problem of calulating the high-dimensional integral is transformed to an optimization problem which is easier to solve. The Laplace approximation method is a special case of adaptive Gauss-Hermite quadrature with only one quadrature point (Liu and Pierce, 1994), and the major advantage of the Laplace approximation is that it provides, for any distribution, response type and link function, a fully closed form approximation of the likelihood which can be maximized efficiently even in case of very complex models applied to high-dimensional data.

Consider next the Laplace approximation of the marginal log-likelihood function (8). Assuming that the responses y_{ij} come from the exponential family of distributions with mean $\mu_{ij} = E(y_{ij})$, the conditional probability distribution can be written as $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp\{y_{ij}a_j(\mu_{ij}) - b_j(\mu_{ij}) + c_j(y_{ij})\}$, where $a_j(\cdot)$, $b_j(\cdot)$ and $c_j(\cdot)$ are known functions, (see for example Dobson, 2008, Chapter 3). The marginal log-likelihood function (8) can then be written as

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

By applying the Laplace approximation method to the integral above, the marginal log-likelihood function obtains an approximation

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

where

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

and $\hat{\boldsymbol{u}}_i$ maximizes a function

$$Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = \sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) - \boldsymbol{u}_i'\boldsymbol{u}_i/2$$

with respect to $\boldsymbol{u}_i$ (Huber et al., 2004).The Laplace approximated log-likelihood function $\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)$ is treated as a new objective function instead of the actual likelihood, and estimated parameters $\hat{\boldsymbol{\Psi}}$ and predictions for the latent variables $\hat{\boldsymbol{u}}_i$ are obtained by maximizing alternately the likelihood $\tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)$ with respect to $\hat{\boldsymbol{\Psi}}$ and $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i)$ with respect to $\boldsymbol{u}_i$ using a quasi-Newton method, for example.

The asymptotic error of the Laplace approximation is known to be of order $O(m^{-1})$, (Tierney and Kadane, 1986). Therefore for high-dimensional abundance data the method provides a good approximation. In particular for GLLVMs, Huber et al. (2004) has shown that the estimates based on the Laplace approximated log-likelihood are consistent with the rate $O(m^{-1})$ and asymptotically normally distributed. The asymptotic standard errors for $\hat{\boldsymbol{\Psi}}$ are obtained via the observed information matrix,

$$\tilde{I}(\hat{\boldsymbol{\Psi}}) = -\left\{ \frac{\partial^2 \tilde{l}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)}{\partial(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)\partial(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i)'} \right\} \Bigg|_{\boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}},$$

as the asymptotic covariance matrix can be approximated by an inverse of $\tilde{I}(\hat{\boldsymbol{\Psi}})$.

The asymptotic prediction errors for $\hat{\boldsymbol{u}}_i$ are easily obtained in a similar fashion as those for $\hat{\boldsymbol{\Psi}}$, as $\widehat{Cov}(\hat{\boldsymbol{u}}_i - \boldsymbol{u}_i | \boldsymbol{y}_i) = \boldsymbol{\Gamma}(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{u}}_i)^{-1}$. However, in order to take into account the uncertainty in the parameter estimation, some adjustments for the prediction errors are needed. In the literature, the parameter uncertainty is accounted either by adjusting the prediction errors with correction terms, *e.g.* in Kackar and Harville (1984) and Booth and Hobert (1998), or by bootstrapping, see Flores-Agreda and Cantoni (2019). We apply the solution proposed by Booth and Hobert (1998); an approximative conditional mean squared errors of predictions (CMSEP). In case of the Laplace approximation method that is

$$\begin{aligned} &CMSEP(\hat{\boldsymbol{u}}_i; \boldsymbol{\Psi}, \boldsymbol{y}_i) = \boldsymbol{\Gamma}(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{u}}_i)^{-1} + \\ &\left(\frac{\partial^2 Q(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \right)^{-1} \frac{\partial^2 Q(\boldsymbol{\Psi}, \boldsymbol{u}_i)}{\partial \boldsymbol{u}_i' \partial \boldsymbol{\Psi}} \tilde{I}(\hat{\boldsymbol{\Psi}})^{-1} \frac{\partial^2 Q(\boldsymbol{\Psi}, \boldsymbol{u}_i)}{\partial \boldsymbol{u}_i \partial \boldsymbol{\Psi}'} \left(\frac{\partial^2 Q(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{u}_i' \partial \boldsymbol{u}_i} \right)^{-1} \Bigg|_{\boldsymbol{u}_i=\hat{\boldsymbol{u}}_i, \boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}. \end{aligned} \tag{9}$$

Such prediction errors can then be used for example in constructing prediction regions around ordination points.

The Laplace approximation of the marginal likelihood function in the general exponential family case for GLLVMs was first provided by Huber et al. (2004). We applied the method for the overdispersed count and biomass data in article PI assuming negative binomial, zero inflated Poisson or Tweedie distributed responses. The `R` implementations for these, as well as for gaussian and for binary data, are given in package `gllvm` (article PIV). In order to improve the accuracy, the Laplace method was extended in Bianconcini and Cagnone (2012), where a fully exponential Laplace approximation method (Tierney et al., 1989) for fitting GLLVMs was proposed in the context of general exponential family.

While being a fast method, Laplace approximation method loses in accuracy as compared to its competitors, at least in connection with discrete data and small datasets. This was shown in Joe (2008), where they compared the accuracy

of the Laplace approximation method to the adaptive Gauss-Hermite quadrature method in cases of discrete response mixed models. In the article PII the Laplace approximation method was compared to the variational approximation method in case of GLLVMs for count and binary data, and it was noticed that the variational approximation method gives more accurate estimates.

4.4 Variational approximation

Another method to approximate the marginal likelihood is the variational approximation method. The main idea in the variational approximation method is to find a closed form approximation to the integral by constructing a lower bound which has a closed form expression. By maximizing the lower bound, the distance between the approximation and the actual integral is then minimized. Originally the method was developed in machine learning research to approximate probability densities (see for example Jordan et al., 1999; Wainwright and Jordan, 2008), but it has also been used in Bayesian data analysis for approximating posterior densities (Teschendorff et al., 2005; Bishop et al., 2006; Blei et al., 2017) in order to reduce the computation times in case of high-dimensional data. In the recent years, the suitability of the variational approximation method in approximating complex marginal likelihoods in maximum likelihood estimation has also gained interest. The method has been applied in the estimation of mixed models in Ormerod and Wand (2010, 2012); Jeon et al. (2017) and Hui et al. (2019).

In the context of the maximum likelihood estimation, the variational approximation method produces a strict lower bound to approximate the marginal log-likelihood function. The estimation and inference is then based on that new objective function. In order to produce a lower bound with a closed form expression, the posterior distribution of the latent variables, $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$, is approximated by a simpler distribution, which is called a variational distribution with a density $q(\boldsymbol{u}|\boldsymbol{\xi})$ and variational parameters $\boldsymbol{\xi}$. Now, consider the marginal log-likelihood function in (8). By using the Jensen's inequality and the concavity of the logarithm function, the variational approximation approach constructs a lower bound,

$$\begin{aligned}\log \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) d\boldsymbol{u} &= \log \int \left\{ \frac{f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) q(\boldsymbol{u}|\boldsymbol{\xi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d\boldsymbol{u} \\ &\geq \int \log \left\{ \frac{f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} q(\boldsymbol{u}|\boldsymbol{\xi}) d\boldsymbol{u} =: \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}),\end{aligned}$$

which is called a variational log-likelihood. We can see that maximizing the variational likelihood is equivalent to minimizing the Kullback-Leibler distance between the true posterior, $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$, and the proposed variational density $q(\boldsymbol{u}|\boldsymbol{\xi})$.

The variational approximation method was first applied to GLLVMs by Hui et al. (2017) in the case of binary, ordinal and overdispersed count data. Follow-

ing Hui et al. (2017), we use the optimal choice for the variational densities of the latent variables and choose independent normal distributions, that is, we denote $q(\boldsymbol{u}_i|\boldsymbol{\xi}_i)$ as a density of $N_d(\boldsymbol{a}_i, \boldsymbol{A}_i)$, where $\boldsymbol{\xi} = (\boldsymbol{a}_i', vec(\boldsymbol{A}_i))'$, $\boldsymbol{a}_i$ are variational mean vectors and $\boldsymbol{A}_i$ are unstructured covariance matrices. For the GLLVM as defined in (4), the variational log-likelihood is then given by

$$\begin{aligned}\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = &\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_q(b(\eta_{ij}))}{\phi_j} + c(y_{ij}, \phi_j)\right\} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - \mathrm{tr}(\boldsymbol{A}_i) - \boldsymbol{a}_i'\boldsymbol{C}_\sigma^{-1}\boldsymbol{a}_i\right),\end{aligned} \tag{10}$$

where $\tilde{\eta}_{ij} = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{a}_i'\boldsymbol{\gamma}_j$. With $E_q(\cdot)$ we denote the expectation with respect to $q(\boldsymbol{u})$. All quantities which are constant with respect to the parameters have been omitted. Depending on the expectation $E_q(b(\eta_{ij}))$, we may obtain a closed form approximation for the marginal log-likelihood (8).

The variational approximate maximum likelihood estimators, $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{\xi}}$, for the model parameters and the variational parameters can be obtained by maximizing the variational log-likelihood $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ with respect to both the model parameters $\boldsymbol{\Psi}$ and variational parameters $\boldsymbol{\xi}$. In addition, as the $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ is treated as a log-likelihood we obtain the asymptotic covariance matrix for the parameters based on the observed information matrix

$$\underline{I}(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\xi}}) = -\left\{\frac{\partial^2\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial(\boldsymbol{\Psi}, \boldsymbol{\xi})\partial(\boldsymbol{\Psi}, \boldsymbol{\xi})'}\right\}\Bigg|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}, \boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}. \tag{11}$$

The corresponding block of the inverse of the information matrix provides approximate standard errors for the maximum likelihood estimates $\hat{\boldsymbol{\Psi}}$.

As noted in Ormerod and Wand (2012) and Hui et al. (2017), equivalency between maximizing the log-likelihood with respect to the variational parameters $\boldsymbol{\xi}$ and minimizing the Kullback-Leibler distance between the posterior and the variational density combined with the normality assumption on $q(\boldsymbol{u}_i|\boldsymbol{\xi}_i)$, imply that the estimated mean vectors $\hat{\boldsymbol{a}}_i$ provide predictions for the latent variables, as $\hat{\boldsymbol{a}}_i$ is both the variational version of the empirical Bayes and maximum a-posteriori predictor of the latent variable. In addition, a matrix $\hat{\boldsymbol{A}}_i$ provides an estimate of the posterior covariance matrix. The estimated covariance matrices can also be used to obtain prediction errors for latent variables. However, it should be noted that they do not take into account the uncertainty of the estimated parameters. Similarly to the prediction variances of the latent variables in the Laplace approximation, we can use CMSEP to approximate the prediction covariance for

the variational predictions $\hat{\boldsymbol{a}}_i$ as

$$CMSEP(\hat{\boldsymbol{a}}_i; \boldsymbol{\Psi}, \boldsymbol{y}_i) = \tag{12}$$
$$\hat{\boldsymbol{A}}_i^{-1} + \left(\frac{\partial^2 \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{a}_i' \partial \boldsymbol{a}_i} \right)^{-1} \frac{\partial^2 \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{a}_i' \partial \boldsymbol{\Psi}} \underline{I}(\hat{\boldsymbol{\Psi}})^{-1} \frac{\partial^2 \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{a}_i \partial \boldsymbol{\Psi}'} \left(\frac{\partial^2 \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})}{\partial \boldsymbol{a}_i' \partial \boldsymbol{a}_i} \right)^{-1} \Bigg|_{\boldsymbol{\xi}=\hat{\boldsymbol{\xi}}, \boldsymbol{\Psi}=\hat{\boldsymbol{\Psi}}}.$$

The variational approximation method provides very fast estimation method for GLLVMs, as described in article PII. The algorithm is implemented in the `R` package `gllvm` (article PIV). One of the few drawbacks is that the method is rather case-specific offering only a closed form approximation when specific combinations of response distributions and link functions are used. For example, in the case of the exponential family of distributions in (10), the closed form expression is obtained only if the expectation term $E_q(b(\eta_{ij}))$ can be solved analytically.

Theoretical properties of the variational approximation estimators have been studied only in few specific cases. Hall et al. (2011a) studied the theoretical properties in the case of the univariate random effect Poisson model and proved that the variational approximation maximum likelihood estimators are asymptotically consistent with rate $O_p(n^{-1/2} + m^{-1})$. Hall et al. (2011b) continued the research and proved the asymptotical normality of the estimators. Similarly, consistency and asymptotic normality of the estimators were studied by Wang and Titterington (2006) in case of Gaussian mixture models and in Bickel et al. (2013) and Celisse et al. (2012) in the case of stochastic block models. Westling and McCormick (2019) studied asymptotic properties of the variational maximum likelihood estimators based on a broad class of models and derived the asymptotic covariance matrix of the gaussian variational approximation estimator. The asymptotic properties of the variational approximation estimators have also been studied in the context of the variational Bayes approach in Wang and Blei (2019). Blei et al. (2017) give a review of the current state of the study on theoretical properties within the variational inference. For the GLLVMs, particularly, Hui et al. (2017) provided a heuristic proof of the consistency of the variational maximum likelihood estimators.

4.5 Implementation and maximization using automatic differentiation

One of the main goals of this thesis was to improve the computational efficiency when fitting generalized linear latent variable models. While the closed form approximation methods such as the Laplace approximation and the variational approximation provide efficient estimation methods, the computational efficiency can still be improved with technically advanced algorithms. We developed model fitting algorithms by utilizing automatic differentiation tools, implemented in the `R` package `TMB` (Template Model Builder, Kristensen et al., 2016). The automatic differentiation is a technology that automatically calculates the derivatives for the

functions that are specified via a programming language. The derivative for any differentiable function can be automatically calculated by applying the chain rule repeatedly to the elementary operations as all mathematical functions are based on a sequence of elementary arithmetic operations and functions. The TMB package offers a general framework for implementing complex random effect models and is inspired by a `C++` language extension AD Model Builder (Fournier et al., 2011), which provide tools for automatic differentiation techniques in statistical optimization. In particular, TMB package employs the `C++` library `CppAD` to construct gradient functions for the log-likelihood function for the model to be optimized. These functions can then be called from `R` and can be straightforwardly optimized using gradient based optimization methods.

In article PII we explain our algorithms in detail and conduct an extensive simulation studies in order to compare the new algorithms based on the variational approximation method and the Laplace approximation method implemented using `TMB` to the plain `R` implementations used in article PI and in Hui et al. (2017). The results showed that especially for the variational approximation method computation times improved significantly when TMB was used in optimization. In addition, the variational approximation method also provided the most accurate estimates for the parameters.

4.6 Software for fitting latent variable models

There exists a wide variety of software for fitting latent variable models using different estimation methods. One of the earliest ones can be found in `STATA` (Rabe-Hesketh et al., 2002; Skrondal and Rabe-Hesketh, 2004), where implementation is conducted using the adaptive Gauss-Hermite quadrature or the Laplace approximation method. The software also provides a lot of additional functionality for inference and visualization. A drawback is that the software is not freely available and therefore is not included in the comparisons presented in Chapter 5.

Some form of the EM algorithm is used in several softwares for fitting latent variable models. Computationally relatively fast algorithm for fitting some simple latent variable models is included in the `R` package `ltm` (Rizopoulos, 2006) which uses the hybrid EM algorithm implementation which utilizes adaptive quadrature for approximating integrals. The method first uses a number of EM iterations and switches then to quasi-Newton iterations. Unfortunately, `ltm` is designed for item response theory, that is, the methods are implemented only for binomial and ordinal responses. Another drawback is that only one or two latent variables can be included in the models. The EM algorithm is also used in the `R` package `mirt` (Chalmers, 2012) for item response models using the algorithm of Bock and Aitkin (1981) where integrals are approximated using Gauss-Hermite quadrature. In addition to the Gauss-Hermite quadrature EM algorithm, models are implemented using a Metropolis-Hastings Robbins-Monro (Cai, 2010)

algorithm. Drawback is that this package does not provide ready-to-use tools for ordination or studying between species correlations, for instance. The latent variable models can also be fitted using Monte Carlo EM algorithm with the `R` package `mistnet` (Harris, 2015) for binomial, ordinal and count data. However, additional tools for inference are missing.

The bayesian approach for fitting GLLVMs is available in `R` packages `boral` (Hui, 2016, 2018) and `Hmsc` (Tikhonov et al., 2019b). The latter is also available in `MATLAB`. Both of the packages provide framework for fitting latent variable models for the most common types of responses and additional tools for visualization and inference. As `boral` and `HMSC` are based on the Bayesian MCMC, model fitting is computationally burdensome.

One important contribution of this thesis is the developed `R` package `gllvm` (Niku et al., 2017) which is reviewed in the article PIV. The `gllvm` package provides computationally efficient model fitting algorithms for GLLVMs which are based on the variational and Laplace approximation methods. The methods are implemented for the most common types of responses, including count, binary, ordinal, biomass and continuous data, and for the models (4) and (6) as well as those models that can be reduced from these two. In addition, a wide variety of tools for inference, visualization and ordination are provided. Based on the simulation studies presented in the next chapter, the package provides the fastest algorithm for fitting GLLVMs as compared to the freely available competitors and it is also among the most accurate ones. The functionality of the `gllvm` is illustrated in Chapter 6.

5 COMPARISON OF ESTIMATION METHODS

In this thesis, we developed efficient estimation methods for the analysis of multivariate abundance data. In this chapter we perform a simulation study in order to compare our algorithms in the `R` package `gllvm` with other available `R` algorithms listed in Section 4.6.

We perform two simulation studies to compare accuracies and computation times of the algorithms for fitting GLLVMs for binary and count data. Data are generated using a simple mean model

$$g(\mu_{ij}) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \tag{13}$$

where intercepts β_{0j} are generated from the standard normal distribution and loadings $\boldsymbol{\gamma}_j$ and latent variables $\boldsymbol{u}_i$ are generated independently from the bivariate normal distribution with zero mean vector and covariance matrix $\boldsymbol{I}_2$, that is, two dimensional identity matrix.

In case of binary data simulation, we use a probit link and generate data from the Bernoulli distribution. We compare eight freely available methods that can be used to fit the model above, either with logit or probit link. Such methods are the variational approximation method (using probit link) and the Laplace approximation method (logit and probit) in the `gllvm` package, Bayesian MCMC methods in the packages `boral` and `Hmsc` (probit), a hybrid algorithm of EM algorithm and adaptive quadrature in the `R` package `ltm` (logit) and the Gaussian quadrature EM and Metropolis-Hastings Robbins-Monro (MHRM) algorithms in the `mirt` package (logit). The packages' default options for iterations were used if the convergence was obtained. That was the case for `gllvm` and `boral`. The `Hmsc` package did not have a default option for MCMC samples, so we used the same number of samples as in `boral`, that is, a burn-in at 10 000, total number of samples 40 000 and thinning at 30. The algorithms in the `mirt` and `ltm` packages often did not converge under the default options, so we increased the maximum number of EM iteration steps to 3000 and the maximum number of iterations for the MHRM algorithm to 5000. In the `ltm` package we used 50 EM steps and a maximum number of quasi-Newton iterations for the adaptive quadrature was

set to 500. We recorded for each algorithm the median computation times over 1000 runs. It need to be noted that the computation times are not necessarily comparable due to different convergence criteria. In the Bayesian the convergence of MCMC chains is checked afterwards and the marginal likelihood based approaches use relative convergence of a value of the likelihood.

The computation times for the binary data are presented in Table 1. We can see that the variational approximation method is the fastest method for all data sizes. The closest competitor when it comes to computational times is provided by the `ltm` before the Laplace approximation method in `gllvm` package or the MHRM algorithm in the `mirt` package. Computation times for the variational approximation method are around 35 to 300 times shorter than for the EM algorithm in the `mirt` package and 80 to 240 times shorter than for the Bayesian MCMC algorithms in the `boral` package and in the `Hmsc` package. In simulations with $n = 40$, the EM algorithm and the MHRM algorithm in `mirt` often had convergence problems and reached the maximum number of iterations without convergence.

In addition to the computation times, we also compare accuracies of the predictors for the latent variables and the parameter estimates for the latent variable loadings using scaled mean procrustes errors, similarly to the simulation setups used in Hui et al. (2017) and in article PII. Biases of the estimated intercepts are also calculated. Accuracies of the estimates given different algorithms are given in Table 2. Differences between the scaled mean procrustes errors of the predicted latent variables are very small, excluding the errors given by Bayesian approach in `Hmsc` results for the $n = m = 40$ case. In all of the cases the variational approximation method in the `gllvm` and Bayesian approach in the `boral` package provide more accurate estimates for the latent variable loadings than any other algorithm. When $n = 100$ and $m = 40$, the loadings given by `Hmsc` are almost equally accurate as compared to those given by `gllvm` and `boral`. The biases of estimates $\hat{\beta}_{0j}$ are small when the estimation is done using the variational approximation method in `gllvm` package or the Bayesian approaches in `boral` or `Hmsc`. Biases for the intercepts of the logit models are exluded from the comparisons because datasets were generated using the probit link function.

In the count data simulation, we use the same model as before with log-link and generate data from the negative binomial distribution with variance function $V(\mu_{ij}) = \mu_{ij} + \phi_j \mu_{ij}^2$, where $\phi_j = 1$, $j = 1, \ldots, m$. The `R` packages that can fit a negative binomial GLLVM for count data are `gllvm`, `boral` and `Hmsc`. Scaled mean procrustes errors of the latent variables predictions and latent variable loadings as well as biases of the estimated intercepts and dispersion parameters are also calculated.

The median computation times are presented in Table 3 and mean procrustes errors and biases are listed in Table 4. We can see that again the variational approximation method is the fastest method in all considered cases and median computation times are around 100 times shorter than when using `boral` and around 35 to 50 times shorter than when using `Hmsc`. The Bayesian MCMC algorithm in the `boral` package is the most accurate method for small sample

TABLE 1 Median computation times (in seconds) when fitting GLLVM for binary data using eight different algorithms.

algorithm	$n = 40$, $m = 40$	$n = 100$, $m = 40$	$n = 40$, $m = 100$
gllvm-VA	2.0	5.3	5.1
gllvm-LA (probit)	22.4	31.9	115.2
gllvm-LA (logit)	12.4	21.1	30.8
ltm	5.7	10.6	17.7
mirt-MHRM	42.5	43.5	101.3
mirt-EM	483.3	178.9	1523.8
boral	172.2	528.1	588.6
Hmsc	496.0	540.1	742.2

TABLE 2 Results for scaled mean Procrustes errors of predicted latent variables (scaled with n and number of latent variables) and estimated latent variable loadings (scaled with m and number of latent variables), and biases of species intercept estimates based on eight estimation algorithms. The true model was a Bernoulli GLLVM with probit link function.

	algorithm	LVs	Loadings	Bias
$n = 40, m = 40$	gllvm-VA	0.117	0.126	0.012
	gllvm-LA (probit)	0.128	0.670	-0.382
	boral	0.119	0.141	-0.001
	Hmsc	0.422	0.499	0.021
	gllvm-LA (logit)	0.131	0.659	
	ltm	0.140	0.681	
	mirt-MHRM	0.123	0.449	
	mirt-EM	0.121	0.525	
$n = 100, m = 40$	gllvm-VA	0.112	0.083	0.003
	gllvm-LA (probit)	0.116	0.284	0.061
	boral	0.114	0.080	0.012
	Hmsc	0.119	0.089	0.001
	gllvm-LA (logit)	0.117	0.471	
	ltm	0.121	0.538	
	mirt-MHRM	0.121	0.182	
	mirt-EM	0.120	0.190	
$n = 40, m = 100$	gllvm-VA	0.048	0.153	0.008
	gllvm-LA (probit)	0.054	0.785	0.168
	boral	0.054	0.137	0.007
	Hmsc	0.071	0.222	0.018
	gllvm-LA (logit)	0.057	0.754	
	ltm	0.073	0.804	
	mirt-MHRM	0.060	0.576	
	mirt-EM	0.059	0.679	

TABLE 3 Median computation times (in seconds) when fitting GLLVM for count data using four different algorithms.

algorithm	$n = 40$, $m = 40$	$n = 100$, $m = 40$	$n = 40$, $m = 100$
gllvm-VA	4.3	16.1	14.4
gllvm-LA	28.9	70.5	126.6
boral	537.4	1964.2	1935.6
Hmsc	431.4	562.9	789.1

sizes $n = 40$, if we look at latent variables, their loading and species-specific intercepts. The smallest biases of disperion parameters were obtained using the variational approximation method in `gllvm`. In case of larger datasets, `gllvm` and `boral` give almost equally accurate results. The `Hmsc` package provided the poorest results in all considered cases. This is probably due the identifiability restriction for the upper triangular of the loading matrix that is made in the `boral` and in the `gllvm` packages, but not in the `Hmsc` package.

In all simulations presented here, the variational approximation method was faster than the other methods included in comparisons. In addition, the method was equally or almost equally accurate when compared to the Bayesian MCMC algorithm in the `boral` package. The Bayesian MCMC algorithm gives highly accurate results but suffers from computational complexity which reflects to computation times.

TABLE 4 Results for scaled mean Procrustes errors of predicted latent variables (scaled with n and number of latent variables) and estimated latent variable loadings (scaled with m and number of latent variables), and biases of species intercept estimates and dispersion parameters based on four estimation algorithms. The true model was a negative binomial GLLVM with log link function.

	algorithm	LVs	Loadings	Bias $\hat{\beta}_{0j}$	Bias $\hat{\phi}_j$
$n = 40, m = 40$	gllvm-VA	0.110	0.100	0.089	-0.043
	gllvm-LA	0.118	0.137	0.173	0.228
	boral	0.086	0.095	0.083	0.259
	Hmsc	0.366	0.303	0.653	0.533
$n = 100, m = 40$	gllvm-VA	0.064	0.028	0.006	-0.054
	gllvm-LA	0.066	0.030	0.057	0.082
	boral	0.066	0.034	0.046	0.093
	Hmsc	0.074	0.048	0.545	0.292
$n = 40, m = 100$	gllvm-VA	0.053	0.108	0.129	0.128
	gllvm-LA	0.056	0.118	0.163	0.221
	boral	0.032	0.093	0.021	0.248
	Hmsc	0.311	0.462	0.671	0.541

6 APPLICATION

In this Chapter we present an illustrative example of the analysis of multivariate abundance data using generalized linear latent variable models. The models are fitted and the visualizations are executed using the agorithm based on variational approximation method implemented in `R` package `gllvm`. Consider the testate amoebae data (Daza Secco et al., 2016) that we shortly introduced in Chapter 2. The data consist of counts of the 50 testate amoebae species measured at three types of Finnish peatlands; two natural peatlands, two forested peatlands and two restored peatlands. Several samples were collected at each peatlands so that the data consist of a total of 270 sampling units. In addition, two environmental covariates temperature and pH value were recorded from each sample.

The interest is now in studying the following research questions: firstly, we study if amoebae species communities differ in terms of land use (natural, forestry, restored) in order to find out if testate amoebae communities could give valuable information about the ecological state of each peatland and, in particular, the success of restoration. Secondly, we try to find some indicator species for different types of peatlands. Finally, we study if the environmental variables (temperature and water pH) affect the species communities.

In order to produce an ordination plot which reveals if species communities differ in different peatlands, we fitted a generalized linear latent variable model with two latent variables and without covariates to the data. Akaike information criterion was used to select the most suitable distribution for the responses. The criterias for the fitted GLLVMs for Poisson distributed and negative binomial distributed responses were 66405 and 33244, respectively. Also the residual analysis supported the choice of the negative binomial distributed responses over Poisson distribution. See Figure 2 for the Dunn-Smyth residuals (Dunn and Smyth, 1996) for the Poisson and the negative binomial GLLVM fit.

The main patterns in species communities can be seen in the ordination of sample sites based on the negative binomial GLLVM in Figure 3(a), where the two predicted latent variables are used as coordinates for the ordination points. Different symbols and colors refer to different uses of land, and we can see that the samples are clustered according to the land use. In addition to ordination

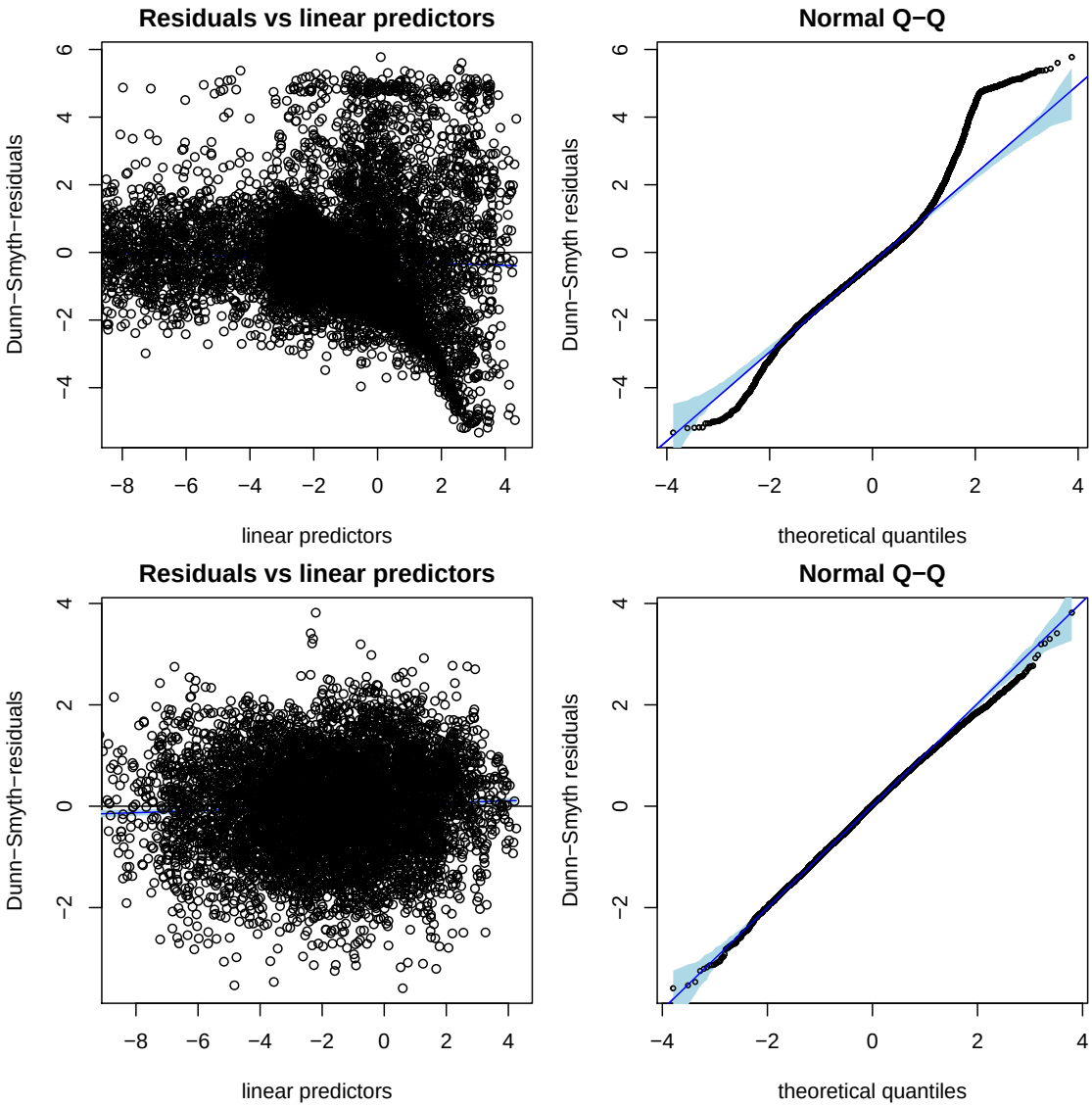

FIGURE 2 Dunn-Smyth residual plots for the Poisson (top) and the negative binomial GLLVM (bottom) fitted for the testate amoebae data. Residuals are plotted against linear predictors (left) and the normal quantile-quantile plot is plotted with simulated point-wise 95% confidence interval envelope (right).

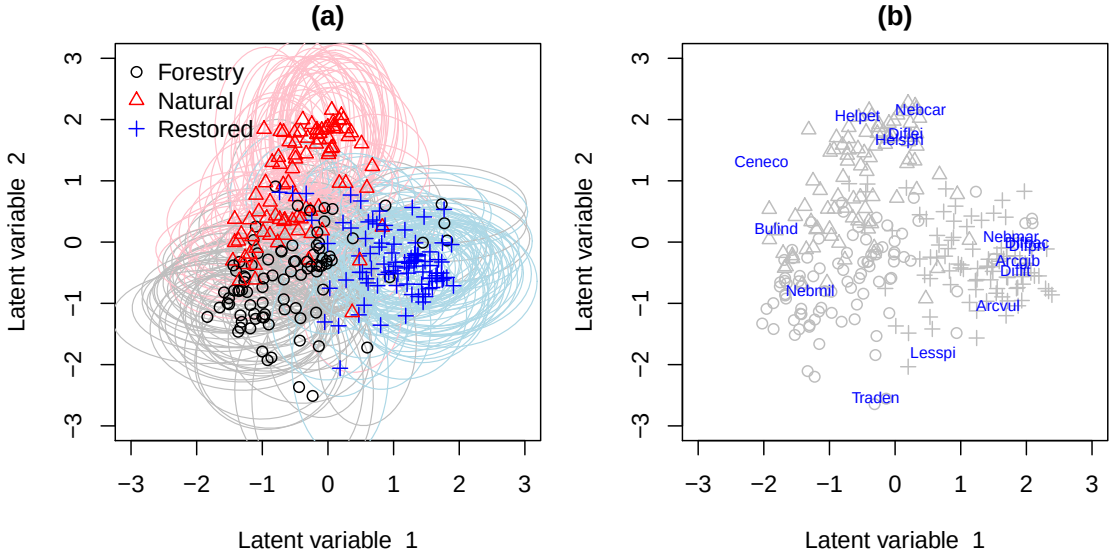

FIGURE 3 Ordination plot with 95% prediction regions (left) and biplot (right) based on the negative binomial GLLVM fitted for the testate amoebae data. In biplot 15 species with the largest facrtor loadings are printed on top of the ordination.

points, 95% prediction regions computed using prediction errors in equation 12 are added in the figure.Even if some prediction regions partly overlap the ordination strongly indicates that the testate amoebae communities differ in terms of the land use.

The biplot with 15 indicator species with the largest factor loadings is plotted in Figure 3(b). In biplot species-specific factor loadings are plotted in the same plot with the latent variables. The figure shows few typical species present at each type of peatlands. For instance, a group of six species, located at the right, are typical for restored peatlands. Having a closer look at the abundance data, one of those six species (the species named *Difbac*) is observed in 84 of the 90 samples collected at restored peatlands while it was present only in 4 forest peatland samples and 14 natural peatland samples, for instance. Similar differences in occurrences can also be found for the other five species in that group. The residual correlations based on the latent factor loadings plotted in Figure 4 can be used to find groups of positively or negatively correlated species.

In Figures 7(a) and 7(b) the ordination points are colored according to the environmental variables pH value and temperature, respectively. In Figure 1(a) we can see a light gradient in the pH values of sample sites. In Figure 1(b) the effect of the temperature is not very clear even though a group of natural sites has higher temperatures than the others. In order to formally study the effect of environmental covariates we add them to the negative binomial GLLVM. The point estimates and the approximate 95% confidence intervals for the species-specific covariate coefficients are plotted in Figure 6. The majority of the confidence intervals of the coefficients of the pH value and about half of the confidence intervals for the coefficients of the temperature do not contain the zero value indicating

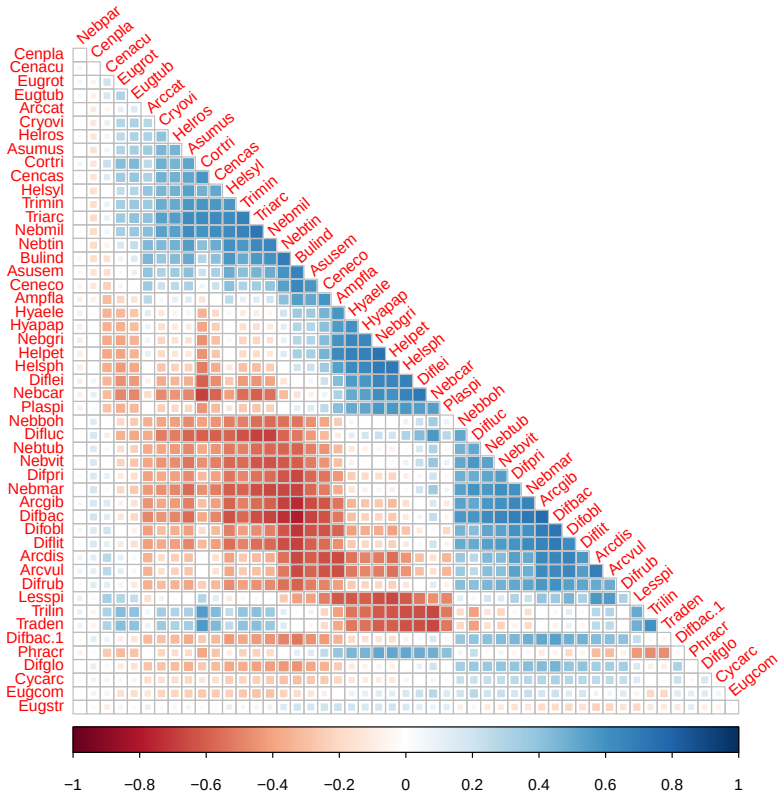

FIGURE 4 The residual correlation matrix based on latent factor loadings for the negative binomial GLLVM fitted to the testate amoebae data.

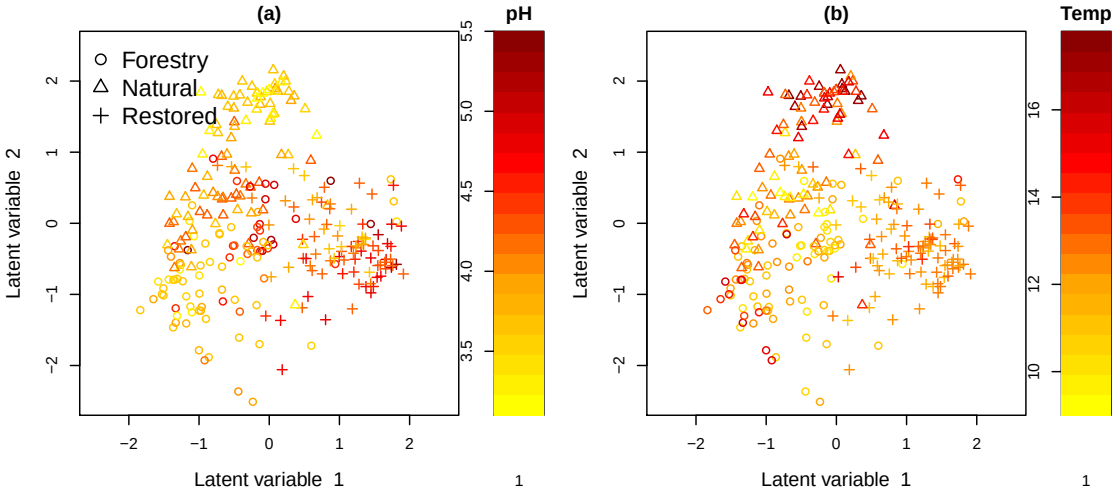

FIGURE 5 Ordination plot based on the negative binomial GLLVM fitted to the testate amoebae data. Here the points representing sample sites are colored according to the covariate values of (a) pH and (b) temperature.

that covariates have substantial effect for the species compositions. We quantified the variance explained by the environmental covariates by comparing traces of the residual covariance matrices of nested models, the model with environmental covariates and the null model without covariates. This measure is a type of pseudo-R^2 considered, for example, in Nakagawa and Schielzeth (2013). In our case, the pseudo-R^2 value indicates that the environmental variables explain 16% of the total variation. The ordination plot based on the GLLVM with the environmental covariates in Figure 7 indicates that even after the effect of the pH value and the temperature is accounted for, the effect of the land use is substantial factor driving the testate amoebae species communities.

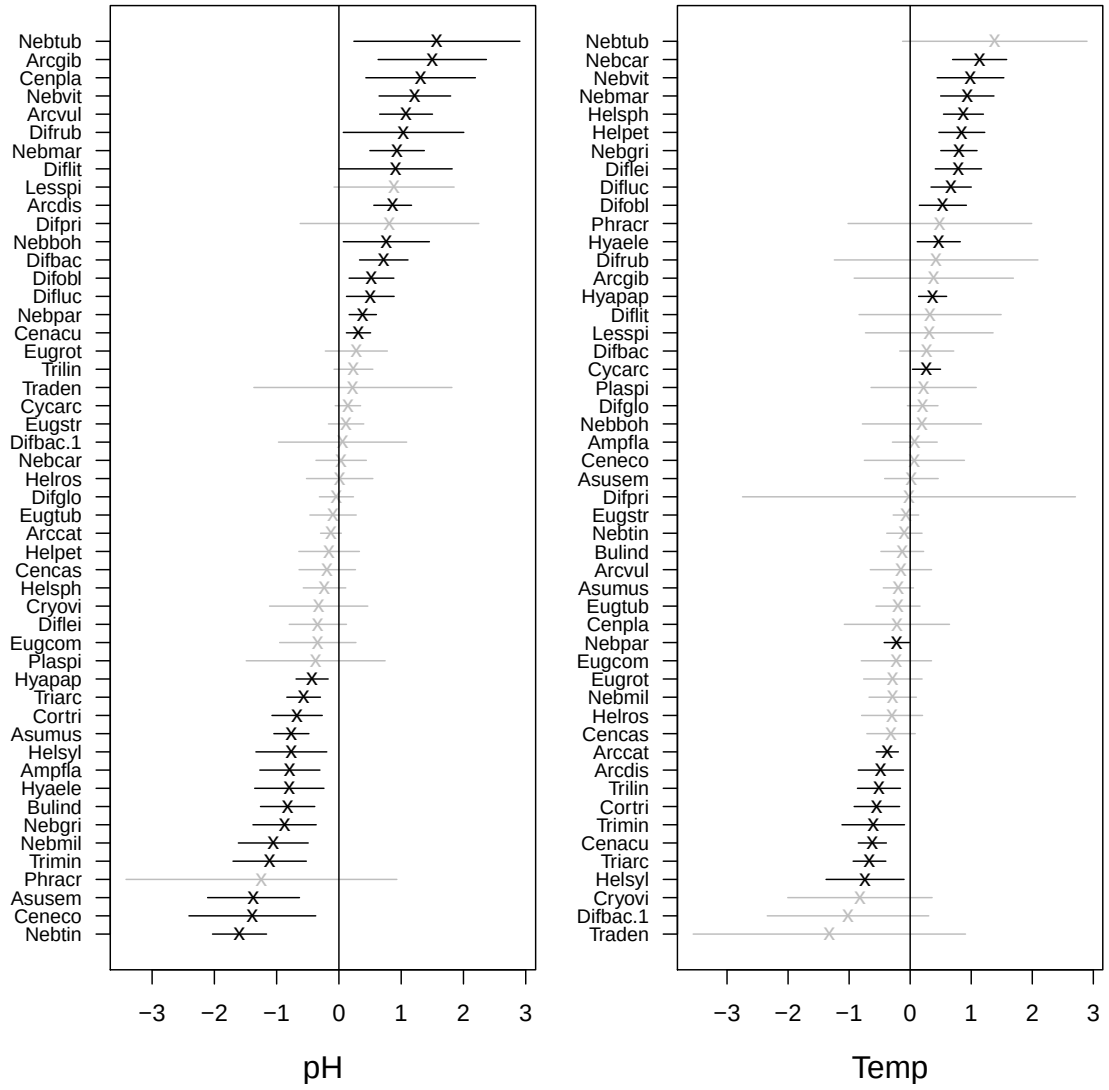

FIGURE 6 Point estimates (ticks) for coefficients of the environmental variables and their 95% confidence intervals (lines) for the negative binomial GLLVM fitted to the testate amoebae data. Lines colored in black denote intervals which do not contain zero.

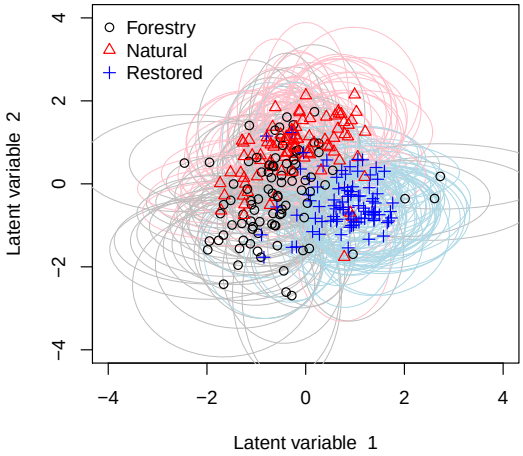

FIGURE 7 Ordination plot with 95% prediction regions based on the negative binomial GLLVM with environmental covariates.

7 SUMMARY OF ORIGINAL PUBLICATIONS

The aim of this thesis was to illustrate how multivariate abundance data can be analysed using generalized linear models and to develop computationally efficient estimation methods for fitting such models. The `R` package `gllvm` was developed in order to offer the estimation algorithms as well as the additional analysis and visualization tools for free use.

In article PI, the generalized linear latent variable models were applied to overdispersed count data and non-negative continuous data. Models were fitted by applying the Laplace approximation method for negative binomial, zero inflated Poisson and Tweedie distributed responses. Simulation studies were conducted using overdispersed counts and biomass data in order to investigate the properties of the estimated parameters. For overdispersed counts, results were shown to be similar to the estimates based on the variational approximation method. In case of biomass data, the estimates were more biased if the between species correlations were ignored. The developed methods were illustrated in ordination and in making inference on environmental variables in two case studies.

In article PII, we developed computationally efficient estimation algorithms for fitting GLLVMs based on the variational approximation method and the Laplace approximation method. By utilizing automatic differentation techniques available in the `R` package `TMB` as well as the computational effort of the `C++` language, the computational speed of the algorithms were significantly improved as compared to the plain `R` implementations. In addition, we developed a new method for choosing starting values for the parameters and latent variables in order to avoid a convergence to local maxima. Performances of new estimation algorithms were evaluated using extensive simulation studies, which indicated that the variational approximation method may potentially provide more accurate estimates for the parameters than the Laplace approximation method. The developed algorithm for the variational approximation method shortened computation times significantly.

In article PIII, we studied the performance of the GLLVMs in testing significance of the environmental-trait interactions. Results showed that the fourth-corner latent variable models were able to take into account interspecific varia-

tion in responses not explained by the observed covariates as well as between species correlations by producing valid Type I errors when significances of the fourth-corner interaction terms were tested using likelihood ratio test. In addition, powers of the likelihood ratio test used with GLLVMs were higher when compared to some existing model-based approaches and permutation tests.

In article PIV we introduced the `R` package `gllvm` for the analysis of multivariate data using GLLVMs. The package `gllvm` provides fast functions for fitting GLLVMs using ether the variational approximation method or the Laplace approximation method for the common response types used in ecology, including counts, binary, ordinal, non-negative continuous and normal distributed responses. In article PIV we illustrate the model fitting tools as well as tools for model diagnostics, visualization and inference.

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. (2014). Asymptotic properties of adaptive maximum likelihood estimators in latent variable models. *Bernoulli*, 20(3):1507–1531.

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.

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41:1922–1943.

Bishop, C. M. et al. (2006). *Pattern recognition and machine learning*. Springer, New-York.

Björk, 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(12):2714–2724.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877.

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46:443–459.

Bock, R. D. and Lieberman, M. (1970). Fitting a response model forn dichotomously scored items. *Psychometrika*, 35(2):179–197.

Bolker, B., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., and White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology and Evolution*, 24:127–135.

Booth, J. G. and Hobert, J. P. (1998). Standard errors of prediction in generalized linear mixed models. *Journal of the American Statistical Association*, 93(441):262–272.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88:9–25.

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(10):2232–2248.

Butler, J. S. and Moffitt, R. (1982). A computationally efficient quadrature procedure for the one-factor multinomial probit model. *Econometrica*, 50:761–764.

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, 75:33–57.

Celisse, A., Daudin, J.-J., and Pierre, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Chalmers, R. P. (2012). `mirt`: A multidimensional item response theory package for the `R` environment. *Journal of Statistical Software*, 48:1–29.

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). More than the sum of the parts: forest climate response from joint species distribution models. *Ecological Applications*, 24(5):990–999.

Clark, N. J., Kaiser, M. S., and Dixon, P. M. (2018). A spatially correlated auto-regressive model for count data. *arXiv:1805.08323v1*.

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.

Daza Secco, E., Haapalehto, T., Haimi, J., Meissner, K., and Tahvanainen, T. (2016). Do testate amoebae communities recover in concordance with vegetation after restoration of drained peatlands? *Mires and Peat*, 18:1–14.

Daza Secco, E., Haimi, J., Högmander, H., Taskinen, S., Niku, J., and Meissner, K. (2018). Testate amoebae community analysis as a tool to assess biological impacts of peatland use. *Wetlands Ecology and Management*, 26:597–611.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38.

Dobson, A. (2008). *An introduction to Generalized Linear Models*. Chapman& Hall/CRC.

Dolédec, S., Chessel, D., ter Braak, C. J. F., and Champely, S. (1996). Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, 3(2):143–166.

Dray, S. and Legendre, P. (2008). Testing the species traits - environment relationships: The fourth-corner problem revisited. *Ecology*, 89(12):3400–3412.

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

Faith, D. P., Minchin, P. R., and Belbin, L. (1987). Compositional dissimilarity as a robust measure of ecological distance. *Vegetatio*, 69(1-3):57–68.

Flores-Agreda, D. and Cantoni, E. (2019). Bootstrap estimation of uncertainty in prediction for generalized linear mixed models. *Computational Statistics & Data Analysis*, 130:1 – 17.

Fournier, D., Skaug, H., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M., Nielsen, A., and Sibert, J. (2011). Ad model builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, 27(2):233–249.

Gelman, A. and Rubin, D. B. (1996). Markov chain monte carlo methods in biostatistics. *Statistical Methods in Medical Research*, 5(4):339–355.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338.

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.

Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution*, 6:465–473.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109.

Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Applied statistics*, 23:340–354.

Hill, M. O. and Gauch Jr, H. (1980). Detrended correspondence analysis: an improved ordination technique. *Vegetatio*, 42:47–58.

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. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Computational Statistics & Data Analysis*, 105:1–10.

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. (2018). *`boral`: Bayesian Ordination and Regression AnaLysis*. R package version 1.6.1.

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(1):35–43.

Hui, F. K. C., You, C., Shang, H. L., and Müller, S. (2019). Semiparametric regression using variational approximations. *Journal of the American Statistical Association, doi:10.1080/01621459.2018.1518235*.

Inoue, K., Stoeckl, K., and Geist, J. (2017). Joint species models reveal the effects of environment on community assemblage of freshwater mussels and fishes in european rivers. *Diversity and Distributions*, 23(3):284–296.

Jamil, T., Opdekamp, W., van Diggelen, R., and ter Braak, C. J. (2012). Trait-environment relationships and tiered forward model selection in linear mixed models. *International Journal of Ecology*. Article 947103.

Jamil, T., Ozinga, W. A., Kleyer, M., and ter Braak, C. J. (2013). Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, 24:988–1000.

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

Jeon, M., Rijmen, F., and Rabe-Hesketh, S. (2017). A variational maximization–maximization algorithm for generalized linear mixed models with crossed random effects. *Psychometrika*, 82(3):693–716.

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

Jongman, R. H., Ter Braak, C. J. F., and Van Tongeren, O. F. R. (1987). *Data analysis in community and landscape ecology*. Pudoc: Wageningen.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kackar, R. N. and Harville, D. A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79(388):853–862.

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.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29:115–129.

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(1):6–106.

Legendre, P. and Anderson, M. J. (1999). Distance-based redundancy analysis: Testing multispecies responses in multifactorial ecological experiments. *Ecological Monographs*, 69:1–24.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L. (1997). Relating behavior to habitat: Solutions to the fourth-corner problem. *Ecology*, 78(2):547–562.

Legendre, P. and Legendre, L. (2012). *Numerical Ecology*. Elsevier, Amsterdam.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(3):325–335.

Liu, Q. and Pierce, D. A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81:624–629.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman & Hall: London.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miller, J. E. D., Damschen, E. I., and Ives, A. R. (2019). Functional traits and community composition: A comparison among community-weighted means, weighted correlations, and multilevel models. *Methods in Ecology and Evolution*, 10(3):415–425.

Morales-Castilla, I., Matias, M. G., Gravel, D., and Araújo, M. B. (2015). Inferring biotic interactions from proxies. *Trends in ecology and 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.

Naylor, J. C. and Smith, A. F. M. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):214–225.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2017). *gllvm: Generalized Linear Latent Variable Models*. R package version 1.1.2.

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.

Ormerod, J. and Wand, M. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, 21:2–17.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64(2):140–153.

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 University Press: Oxford.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, 91(9):2514–2521.

Ovaskainen, O., Roy, D. B., Fox, R., and Anderson, B. J. (2016c). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution*, 7(4):428–436.

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(5):561–576.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, 4(1):12–35.

Pollock, L. J., Morris, W. K., and Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8):716–725.

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(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.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, 128(2):301 – 323.

Ribera, I., Dolédec, S., Downie, I. S., and Foster, G. N. (2001). Effect of land disturbance and stress on species traits of ground beetle assemblages. *Ecology*, 82(4):1112–1129.

Rizopoulos, D. (2006). `ltm`: An `R` package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17:1–25.

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(8):733–742.

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 B (Statistical Methodology)*, 59:667–678.

Schliep, E. M., Lany, N. K., Zarnetske, P. L., Schaeffer, R. N., Orians, C. M., Orwig, D. A., and Preisser, E. L. (2018). Joint species distribution modelling for spatio-temporal occurrence and ordinal abundance data. *Global Ecology and Biogeography*, 27(1):142–155.

Shipley, B. (2010). Inferential permutation tests for maximum entropy models in ecology. *Ecology*, 91(9):2794–2805.

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. (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67(5):1167–1179.

ter Braak, C. J., Peres-Neto, P., and Dray, S. (2017). A critical issue in model-based inference for studying trait-based community assembly and a solution. *PeerJ*, 5:e2885.

ter Braak, C. J. F. (2019). New robust weighted averaging- and model-based methods for assessing trait-environment relationships. *Methods in Ecology and Evolution*, 10:1962–1971.

ter Braak, C. J. F., Cormont, A., and Dray, S. (2012). Improved testing of species traits-environment relationships in the fourth-corner problem. *Ecology*, 93(7):1525–1526.

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., and Caldas, C. (2005). A variational Bayesian mixture modelling framework for cluster analysis of gene-expression data. *Bioinformatics*, 21(13):3025–3033.

Thorson, J. T., Fonner, R., Haltuch, M. A., Kotaro Ono, K., and Winker, H. (2017). Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. *Canadian Journal of Fisheries and Aquatic Sciences*, 74:1794–1807.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., and Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25:1144–1158.

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. (2015). Spatial factor analysis: A new tool for estimating joint species distributions and correlations in species range. *Methods in Ecology and Evolution*, 6(6):627–637.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, 84(407):710–716.

Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D., and Ovaskainen, O. (2019a). Computationally efficient joint species distribution modeling of big spatial data. *Ecology, doi:10.1002/ecy.2929*.

Tikhonov, G., Opedal, Ø., Abrego, N., Lehikoinen, A., and Ovaskainen, O. (2019b). Joint species distribution modelling with hmsc-r. *bioRxiv: 603217*.

Vanhatalo, J., Veneranta, L., and Hudd, R. (2012). Species distribution modeling with gaussian processes: A case study with the youngest stages of sea spawning whitefish (coregonus lavaretus l. s.l.) larvae. *Ecological Modelling*, 228:49 – 58.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305.

Walker, S. C. and Jackson, D. A. (2011). Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, 81(4):635–663.

Wang, B. and Titterington, D. M. (2006). Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. *Biometrics*, 67:116–123.

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 and 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.

Warton, D. I. and Hui, F. K. C. (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.

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.

Westling, T. and McCormick, T. H. (2019). Beyond prediction: A framework for inference with variational approximations in mixture models. *Journal of Computational and Graphical Statistics, doi: 10.1080/10618600.2019.1609977*.

ORIGINAL PAPERS

PI

GENERALIZED LINEAR LATENT VARIABLE MODELS FOR MULTIVARIATE COUNT AND BIOMASS DATA IN ECOLOGY

by

Niku, J., Warton, D.I., Hui, F.K.C., and Taskinen, S. 2017

Journal of Agricultural, Biological, and Environmental Statistics, 22:498–522

Supplementary materials for this article are available at 10.1007/s13253-017-0304-7.

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

Jenni NIKU, David I. WARTON, Francis K. C. HUI, and Sara TASKINEN

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.

Supplementary materials accompanying this paper appear on-line.

Key Words: Biomass; Laplace approximation; Ordination; Overdispersed count; Species interactions.

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-

Jenni Niku (✉) and Sara Taskinen Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland (E-mail: *jenni.m.e.niku@jyu.fi*). David I. Warton School of Mathematics and Statistics and Evolution and Ecology Research Centre, The University of New South Wales, Sydney, Australia. David I. Warton School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia. Francis K. C. Hui Mathematical Sciences Institute, The Australian National University, Canberra, Australia.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 22, Number 4, Pages 498–522
DOI: 10.1007/s13253-017-0304-7

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 consist 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 Sect. 5.1 in detail. This example, which is by no means an extreme case, exhibits 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 extends the basic generalized linear model framework by including one or more latent variables, with corresponding factor loadings, as a parsimonious method of modelling 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 analysing 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 modelling (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 laboratory-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 Sect. 2, we formulate the generalized linear latent variable model framework and response distributions of interest for multivariate abundance data. In Sect. 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, Sect. 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 $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_m)'$ 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 $\boldsymbol{\Sigma}_{\text{res}} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, 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

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

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 nonzero counts (Warton 2005). But if the ecological process governing most species absences is

believed to be independent of the mechanism driving the nonzero 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

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

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 $\text{Var}(y_{ij}) = E(y_{ij})(1+p_j\mu_{ij})$. 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 $\text{Var}(y_{ij}) = \phi_j\mu_{ij}^{\nu}$, 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 analysing 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 Eq. (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 Sect. 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 "Appendix A." Here we focus on response types and distributions discussed in Sect. 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*

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

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*

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

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*

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

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 Sect. 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 (Supplementary Material)".

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 h 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.

	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

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 $m = 1000$.

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 Fig. 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 Fig. 1, both model-based ordination methods strongly outperform nMDS. The results based on Laplace approximation method and variational approximation method are almost equally good.

Finally, regarding computation time, the proposed Laplace approximation method averaged 13.2, 12.1, 159.4 and 609.3 s, respectively, to estimate the parameters and their standard

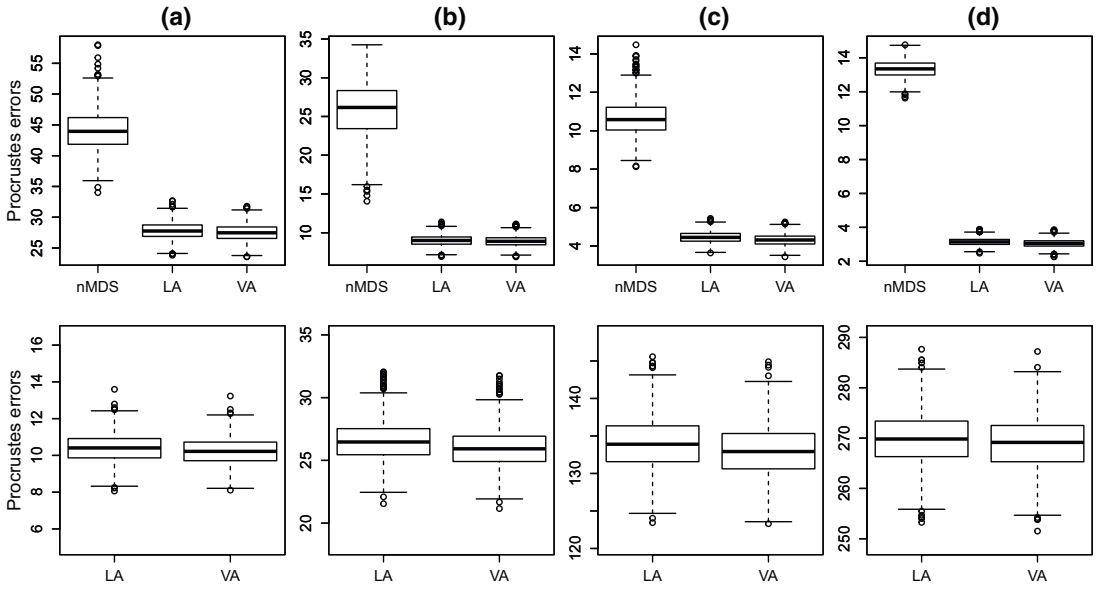

Figure 1. Comparative boxplots of Procrustes errors between true and estimated ordination points (*first row*) and true and estimated parameters $\hat{\boldsymbol{\gamma}}_j$ (*second row*). Ordination points (and parameters $\hat{\boldsymbol{\gamma}}_j$ when applicable) are obtained from non-metric 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 $m = 1000$.

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 s, 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: (a) $n = 100$ and $m = 50$, (b) $n = 50$ and $m = 100$ and (c) $n = 50$ 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

Table 2. Average biases and root mean squared errors (MSEs) of Tweedie GLLVM and Tweedie GLM estimates based on Laplace approximation method.

	GLLVM		GLM	
	Bias	RMSE	Bias	RMSE
(a)				
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
(b)				
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
(c)				
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

The true models were Tweedie GLLVMs with (a) $n = 100$ and $m = 50$, (b) $n = 50$ and $m = 100$ and (c) $n = 50$ and $m = 200$.

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 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 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, Fig. 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

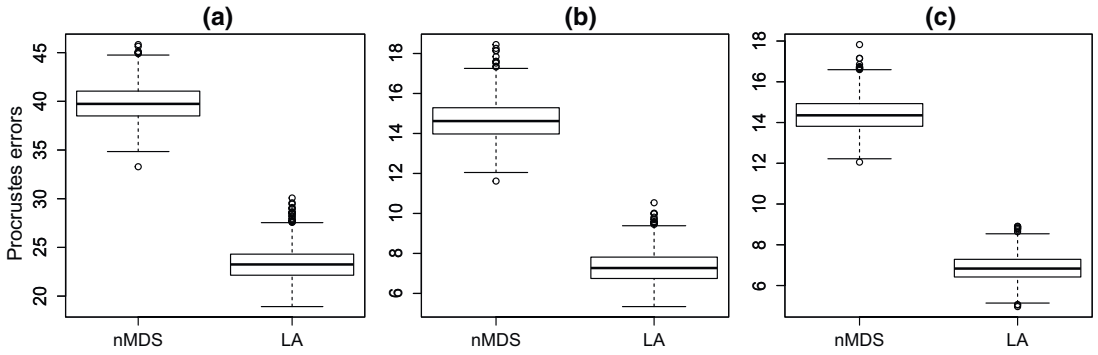

Figure 2. Comparative boxplots of Procrustes errors between true and estimated ordination points. Ordination points are obtained from non-metric multidimensional scaling (nMDS) and Tweedie GLLVM fitted using Laplace approximation method (LA). The true model in each plot was Tweedie GLLVM with **a** $n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

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

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.

The ordination of sites based on negative binomial GLLVM is plotted in Fig. 3a. 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 Fig. 5 in "Appendix B"). In addition, the ordination points are (also) labelled according to the sampling location (Kilpisjärvi, Ny-Ålesund and Innsbruck), and it is clear that the sites differed in terms of species composition. In Fig. 3b, a biplot based on generalized linear

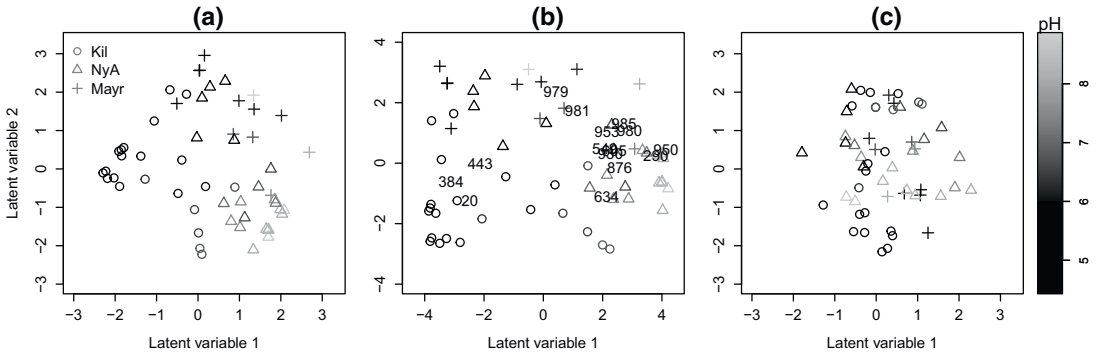

Figure 3. **a** The ordination plot of $n = 56$ sites based on generalized linear latent variable model without any covariates assuming negative binomial distributed responses. **b** The biplot, where 15 species with the largest factor loadings (in terms of distance from the origin) are printed on top 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 (**a**) to emphasize the reduction in variation. The sites in ordination plots are coloured according to their pH values and labelled according to the sampling site.

latent variable model is given. Here indices of the 15 species with largest factor loadings are added in the (rotated) ordination plot in Fig. 3a. 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 Fig. 6 in "Appendix B," and indicate that pH value was the main covariate affecting the species composition. The corresponding ordination plots are given in Fig. 7 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 Fig. 3c shows that there is no visible pattern in sampling location anymore. As the figure uses the same scale as plots in Fig. 3a, 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, 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 Fig. 8

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 Fig. 4a. 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 maximized 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 Fig. 4b exhibits a clear location difference between coral compositions

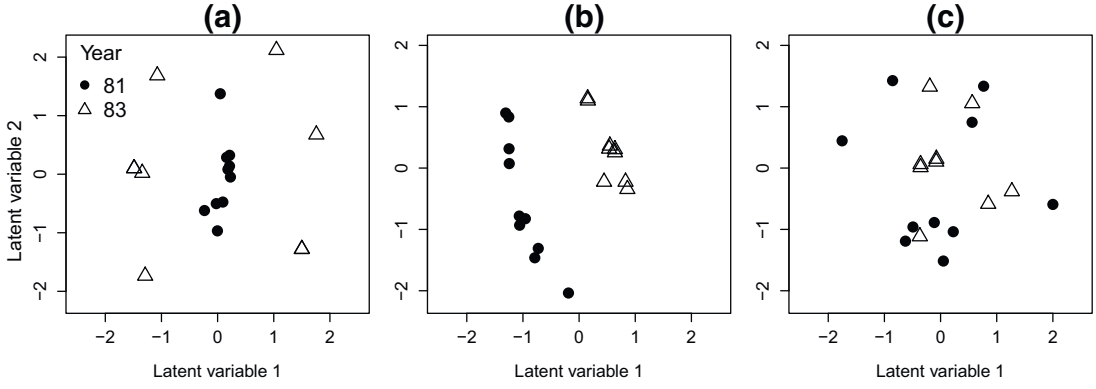

Figure 4. The ordination plots of $n = 19$ sites based on **a** non-metric multidimensional scaling, **b** Tweedie GLLVM without site effect and **c** Tweedie GLLVM with site effects. The sites in ordination plots are labelled according to the year the data was collected.

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 Fig. 4c indicate that the species compositions did not change between the two sampling times. In Fig. 9 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 modelling 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 modelling 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.

[Received January 2017. Accepted August 2017. Published Online August 2017.]

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

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

where

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

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

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

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

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

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

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

where

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

$$= \sum_{j=1}^{m} \Bigg[\exp(\hat{\eta}_{ij}) I_{(y_{ij}>0)} - \Bigg(\frac{(1-p_j)\hat{A}_{ij} \exp(\hat{\eta}_{ij})(\exp(\hat{\eta}_{ij})-1)}{p_j + (1-p_j)\hat{A}_{ij}} - \frac{(1-p_j)^2 \hat{A}_{ij}^2 \exp(2\hat{\eta}_{ij})}{(p_j + (1-p_j)\hat{A}_{ij})^2} \Bigg) I_{(y_{ij}=0)} \Bigg] \boldsymbol{\gamma}_j \boldsymbol{\gamma}_j' + \boldsymbol{I}_d,$$

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 $Q(\boldsymbol{\Psi}, \boldsymbol{u}_i) = (1/m)\left(\sum_{j=1}^{m} \log f(y_{ij}|\boldsymbol{u}_i; \boldsymbol{\Psi}) - \boldsymbol{u}_i' \boldsymbol{u}_i / 2\right)$.

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

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

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

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

where

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

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

See Figs. 5, 6, 7, 8 and 9.

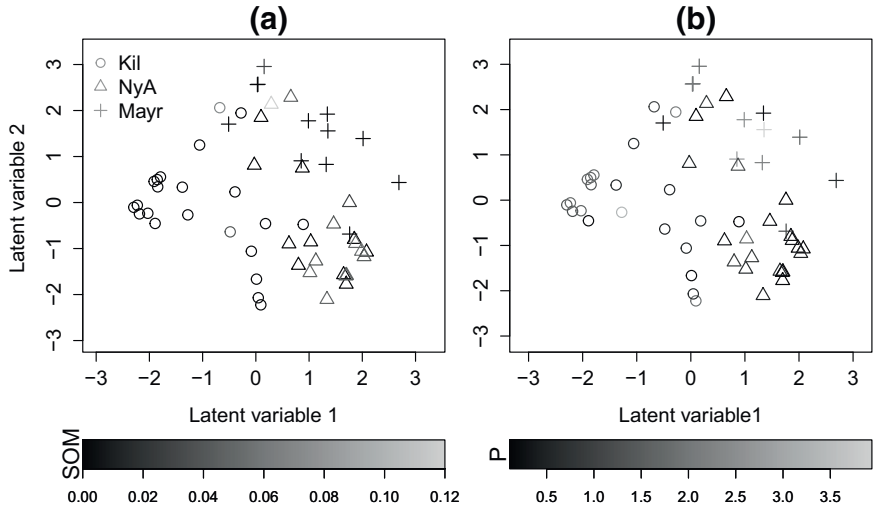

Figure 5. The ordination of $n = 56$ sites based on generalized linear latent variable model without any covariates assuming negative binomial distributed responses. The sites in ordination are coloured according to their **a** soil organic matter (SOM) values and **b** phosphorous (P) values, and labelled according to the sampling site.

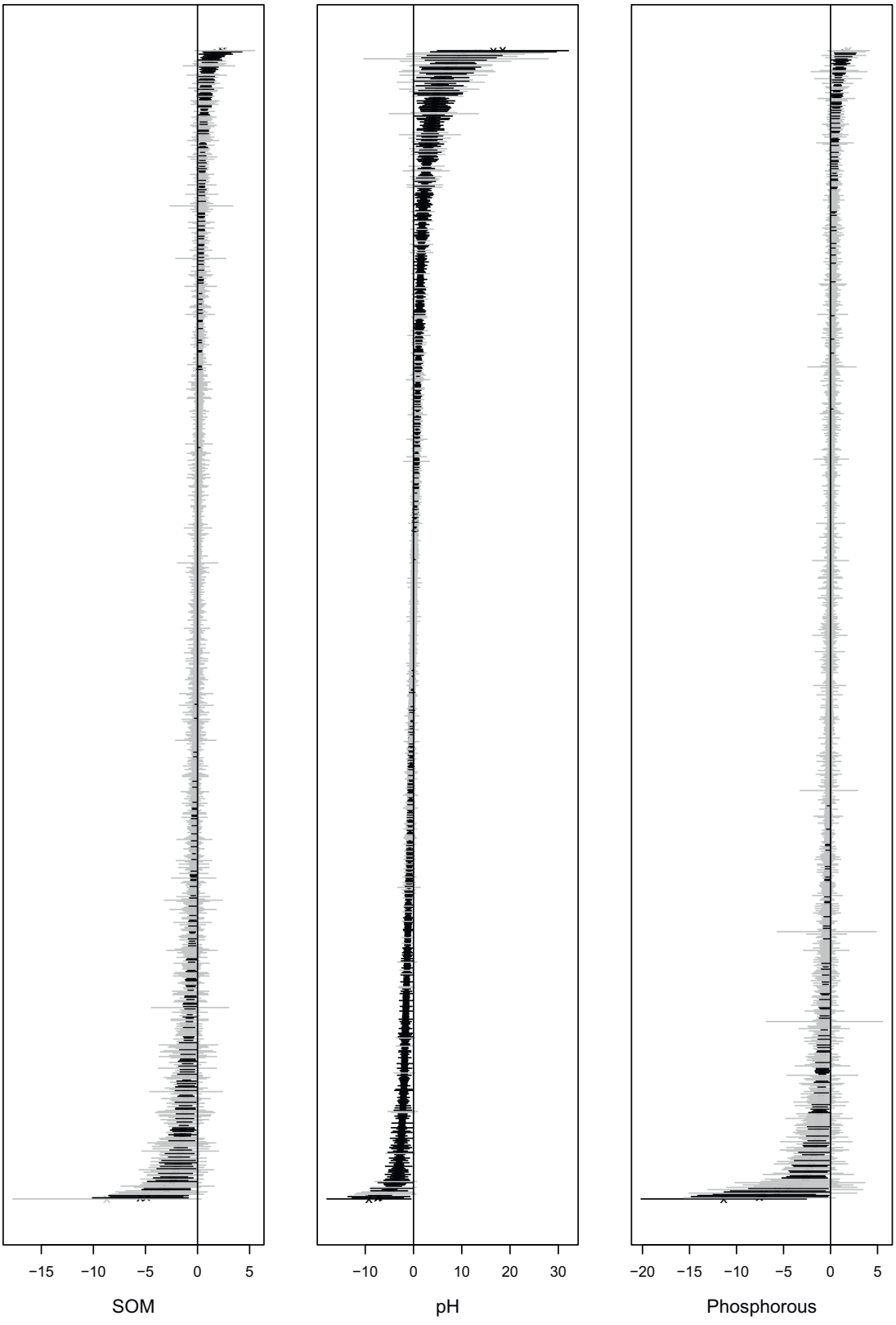

Figure 6. Ranked point estimates with 95% confidence intervals for the three environmental variables based on negative binomial GLLVM. Grey confidence intervals include the zero value.

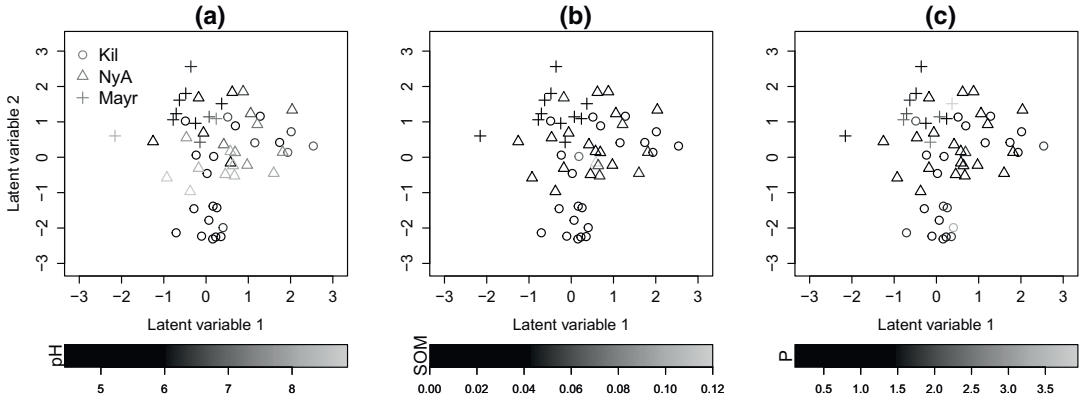

Figure 7. The ordination of $n = 56$ sites based on generalized linear latent variable model with pH, soil organic matter and phosphorous as covariates, and assuming negative binomial distributed responses. The sites in ordination are coloured according to their **a** pH values, **b** soil organic matter (SOM) values and **c** phosphorous (P) values, and labelled 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.

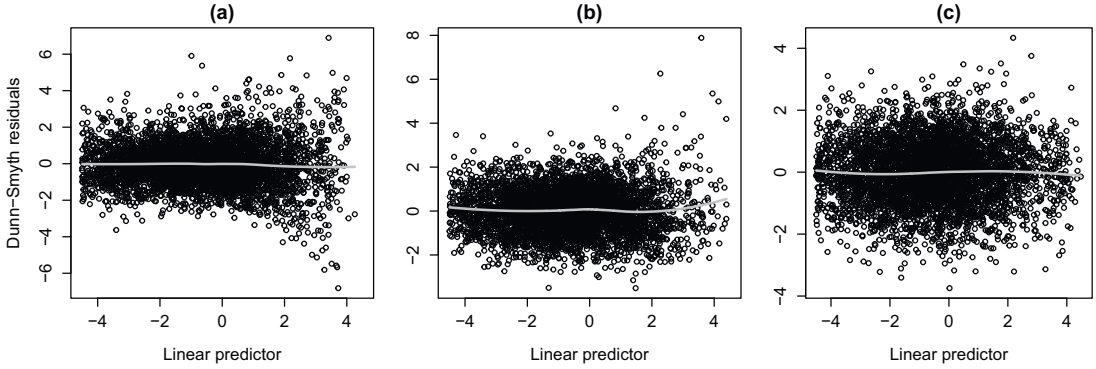

Figure 8. Dunn–Smyth residuals against linear predictors for the **a** Poisson, **b** zero-inflated Poisson and **c** negative binomial GLLVM models with pH, soil organic matter, phosphorous and categorical site as covariates. Lowess curves are included in the plots.

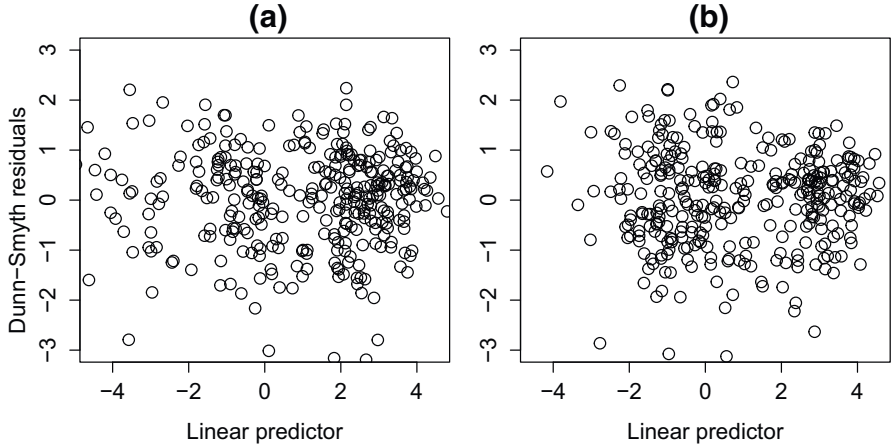

Figure 9. Dunn–Smyth residuals against linear predictors for the Tweedie models **a** without site effect and **b** with site effect.

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.

——. (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.

PII

EFFICIENT ESTIMATION OF GENERALIZED LINEAR LATENT VARIABLE MODELS

by

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., and Warton, D.I.
2019

PLOS ONE , 14(5):1–20

PLOS ONE

RESEARCH ARTICLE

Efficient estimation of generalized linear latent variable models

Jenni Niku[1]*, **Wesley Brooks**[2☯], **Riki Herliansyah**[3☯], **Francis K. C. Hui**[4☯], **Sara Taskinen**[1☯], **David I. Warton**[2,5☯]

1 Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland, **2** School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia, **3** Department of Mathematics, Kalimantan Institute of Technology, Kalimantan, Indonesia, **4** Research School of Finance, Actuarial Studies & Statistics, Australian National University, Canberra, Australia, **5** Evolution & Ecology Research Centre, The University of New South Wales, Sydney, Australia

☯ These authors contributed equally to this work.
* jenni.m.e.niku@jyu.fi

Citation: Niku J, Brooks W, Herliansyah R, Hui FKC, Taskinen S, Warton DI (2019) Efficient estimation of generalized linear latent variable models. PLoS ONE 14(5): e0216129. https://doi.org/10.1371/journal.pone.0216129

Editor: Jin Li, Geoscience Australia, AUSTRALIA

Received: November 28, 2018

Accepted: April 15, 2019

Published: May 1, 2019

Data Availability Statement: All relevant data to the first simulation setup (amoebae data) are in the Supporting Information files within the manuscript, see S2 File. The second data, Indonesian birds, are available in the supplementary material of Daniel F. R. Cleary, Timothy J. B. Boyle, Titiek Setyawati, Celina D. Anggraeni, E. Emiel Van Loon, and Steph B. J. Menken. 2007. Bird species and traits associated with logged and unlogged forest in Borneo. Ecological Applications 17:1184–1197, see Ecological Archives A017-043-A2.

Funding: J. Niku was supported by the Jenny and Antti Wihuri Foundation. S. Taskinen was

Abstract

Generalized linear latent variable models (GLLVM) are popular tools for modeling multivariate, correlated responses. Such data are often encountered, for instance, in ecological studies, where presence-absences, counts, or biomass of interacting species are collected from a set of sites. Until very recently, the main challenge in fitting GLLVMs has been the lack of computationally efficient estimation methods. For likelihood based estimation, several closed form approximations for the marginal likelihood of GLLVMs have been proposed, but their efficient implementations have been lacking in the literature. To fill this gap, we show in this paper how to obtain computationally convenient estimation algorithms based on a combination of either the Laplace approximation method or variational approximation method, and automatic optimization techniques implemented in R software. An extensive set of simulation studies is used to assess the performances of different methods, from which it is shown that the variational approximation method used in conjunction with automatic optimization offers a powerful tool for estimation.

1 Introduction

High-dimensional multivariate abundance data, which consist of records (e.g. species counts, presence-absence records, and biomass) of a large number of interacting species at a set of units or sites, are routinely collected in ecological studies. When analyzing multivariate abundance data, the interest is often in visualization of correlation patterns across species, hypothesis testing of environmental effects, and making predictions for abundances. Classical methods for analysing such data, including algorithmic-based approaches such as non-metric multidimensional scaling (nMDS) and correspondence analysis (CA), are based on distance matrices computed on some pre-specified dissimilarity measure [1]. As such, they often make wrong assumptions for key properties of the data at hand (e.g. mean-variance relationship), which can potentially lead to misleading inferential results [2, 3].

supported by CRoNoS COST Action IC1408. F.K.C. Hui and D.I. Warton were funded by Australia Research Council Discovery Project grants (DP180100836 and DP180103543, respectively). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

An alternative approach that has gained considerable attention over the past several years is generalized linear latent variable models (GLLVMs, [4]). GLLVMs start with the basic generalized linear model (GLM, [5]), classically used to model the impact of environmental covariates on abundance of one species, and extend it by incorporating latent variables to model between response correlation in a parsimonious manner. As the model makes explicit assumptions concerning the response distribution, the mean-variance relationship can be correctly modeled and verified using (for instance) residual analysis and model selection approaches. In the context of multivariate abundance data, GLLVMs were first proposed by [6] for presence-absence data, and [7] in a more general framework for model-based unconstrained ordination. By adding covariates to the model, it can be used as a model-based approach to correspondence analysis [8]. More recently, there has been an explosion in research on various extensions of GLLVMs for joint analyses of multivariate abundance data, see [9–12] among many others.

One of the main and long standing challenges with using GLLVMs is the lack of computationally efficient estimation methods. The need for fast and efficient estimation methods evolves from the fact that modern data collection tools such as metabarcoding often result in very large and high-dimensional datasets (for a recent review, see [13]), and current methods are unable to fit GLLVMs for such data in reasonable amount of time. Specifically, many of the standard methods proposed in the literature for fitting GLLVMs have a major drawback as being either computationally very intensive with high-dimensional data e.g. the Expectation Maximization algorithm [7, 14] and Bayesian Markov Chain Monte Carlo estimation [11, 15], or are computationally impractical with a larger number of latent variables, such as Gauss-Hermite quadrature [16–18]. In recent years, a number of approaches have been proposed in the literature to overcome such issues, with two of the more prominent ones being the variational approximation method to approximate the likelihood in the case of binary, ordinal and overdispersed count data [19], and the Laplace approximation method for responses from the exponential family of distributions [20], which has recently been adapted specifically for overdispersed count and biomass data in ecology [21]; Note that the Laplace approximation can be considered as a special case of adaptive quadrature with only one quadrature point. Both estimation methods provide a closed form approximation to the marginal log-likelihood that can then be maximized efficiently.

In this paper, we propose a framework for faster fitting of GLLVMs using either Laplace approximation method or the variational approximation method. Our method utilizes the `R` package `TMB` (Template Model Builder, [22]), which offers a general tool for implementing complex random effect models through simple `C++` templates. `TMB` is inspired by AD Model Builder [23], which is a `C++` language extension for solving optimization problems using automatic differentiation [24]. With growing popularity, `TMB` has been used to estimate complex non-linear models, e.g. for fitting mixed-effect models [25] and non-Gaussian state space models [26]. The algorithms we propose in this article for efficient estimation of GLLVMs have been recently implemented in the `R` package `gllvm` [27].

Another major contribution we make is to provide a new method for obtaining starting values for parameter estimation of GLLVMs. This is especially important for GLLVMs given their complex mean and latent variable structures may cause the observed likelihood to be multimodal (as discussed in [28]), and good starting values are therefore critical in order to guarantee fast convergence and to avoid local maxima. Our proposed method is based around fitting univariate GLMs to each species in order to obtain starting values for fixed parameters, and then applying a factor analysis to the Dunn-Smyth residuals [29] from the fitted GLMs as the basis for constructing starting values for the loadings and latent variables. We performed

an extensive series of simulation studies to compare the performances of estimation algorithms with and without TMB, and to compare various methods for constructing starting values. The simulation studies showed that in most cases, the variational approximation method utilizing TMB outperformed the other estimation algorithms: computation times were clearly faster than those of the other methods, the empirical mean biases and mean squared errors of the parameter estimates were smaller, and coverage probabilities of Wald-type confidence intervals were closer to their nominal level. Our simulations also show that the proposed approach for choosing starting values outperformed more standard methods such as random starting values in terms of consistency of reaching the global maximum of the likelihood, regardless of the data at hand.

The paper is organized as follows. In Section 2, we formulate a generalized linear latent variable model suitable for joint modeling of abundance data, and review the most recently proposed approximation methods. In Section 3, we explain how the estimation can be performed using TMB and introduce different methods for obtaining starting values for estimation. In section 4, we study the performances of our methods using several simulation studies. Section 5 concludes the paper.

2 Generalized linear latent variable models

Consider a sample of observations consisting of responses for m species collected at n sites, such that y_{ij} denotes the response for species $j = 1, \ldots, m$ at site $i = 1, \ldots, n$. A generalized linear latent variable model (GLLVM) regresses the mean response, denoted here as μ_{ij}, against a vector of $d \ll m$ latent variables, $\mathbf{u}_i = (u_{i1}, \ldots, u_{id})'$, along with the vector of covariates $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik})'$. 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 $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are vectors of species specific coefficients related to the covariates and latent variables, respectively. It is the term $\boldsymbol{u}_i'\boldsymbol{\gamma}_j$ which captures the residual correlation across species not accounted for by the observed covariates x_i. Moreover, a key advantage of this type of model is that it is capable of flexibly handling correlation across response variables in a parsimonious manner, with the number of parameters characterizing the correlation structure growing linearly in the number of responses m. This allows GLLVMs to be feasibly fitted to datasets with relatively large m, as often arises in practice [8].

We assume that the latent variables follow a multivariate standard normal distribution, $\boldsymbol{u}_i \sim N_d(\mathbf{0}, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes a $d \times d$ identity matrix. The assumption of zero mean and unit variance is made in order to fix the locations and scales of latent variables. We also set all the upper triangular elements of $m \times d$ matrix $\Gamma = (\boldsymbol{\gamma}_1 \cdot \cdot \boldsymbol{\gamma}_m)'$ to be zero, that is, $\gamma_{ij} = 0$ for $j > i$, and constrain its diagonal elements, γ_{ii}, to be positive in order to avoid rotation invariance and to ensure parameter identifiability.

For the GLLVM defined in Eq (1), where the α_i's are assumed to be random row effects (reflecting a nested sampling design, say), denote $\boldsymbol{u}_i^* = (\alpha_i, \boldsymbol{u}_i')'$ and $\boldsymbol{\gamma}_j^* = (1, \boldsymbol{\gamma}_j')'$ and write the model as $g(\mu_{ij}) = \eta_{ij} = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i^{*\prime}\boldsymbol{\gamma}_j^*$. Since the latent variables and random intercepts are assumed to be independent, then $\boldsymbol{u}_i^*$ follows a multivariate normal distribution with mean zero and block diagonal covariance matrix, $\boldsymbol{C}_{\sigma^2} = \text{bdiag}(\sigma^2, \boldsymbol{I}_d)$, where bdiag$(\cdot)$ is the block diagonal operator. Write the probability density function of $N(\mathbf{0}, \boldsymbol{C}_{\sigma^2})$ as $f(\boldsymbol{u}_i^*; \sigma^2)$. To complete the formulation, we assume that conditional on the latent variables $\boldsymbol{u}_i^*$ and parameter vector $\boldsymbol{\Psi}$, the responses are independent observations from the exponential family of distributions with

probability density function,

$$f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp\left\{\frac{y_{ij}a(\eta_{ij}) - b(\eta_{ij})}{\phi_j} + c(y_{ij};\phi_j)\right\}, \tag{2}$$

where $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$ are known functions and ϕ_j is a species specific dispersion parameter. Let $\boldsymbol{\Psi} = (\boldsymbol{\beta}_0', vec(\boldsymbol{B})', vec(\boldsymbol{\Gamma})', \Phi', \sigma^2)$ denote the full vector of parameters in the GLLVM, where $\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0m})'$, $\boldsymbol{B} = (\beta_1 \ldots \beta_m)'$, $\Gamma = (\gamma_1 \ldots \gamma_m)'$, and Φ includes all other nuisance parameters e.g. $\Phi = (\phi_1, \ldots, \phi_m)'$. With the conditional independence of the responses given the latent variables, we then obtain $f(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \boldsymbol{\Psi}) = \prod_{j=1}^m f(y_{ij}|\boldsymbol{u}_i^*; \boldsymbol{\Psi}) f(\boldsymbol{u}_i^*; \sigma^2)$ as the joint distribution. By integrating over latent variables $\boldsymbol{u}_i^*$ then, we obtain the following marginal log-likelihood function for a GLLVM,

$$l(\boldsymbol{\Psi}) = \sum_{i=1}^n \log\left(f(y_{ij}, \boldsymbol{\Psi})\right) = \sum_{i=1}^n \log\left(\int_{\mathbb{R}^{d+1}} \prod_{j=1}^m f(y_{ij}|\boldsymbol{u}_i^*; \boldsymbol{\Psi}) f(\boldsymbol{u}_i^*; \sigma^2) d\boldsymbol{u}_i^*\right). \tag{3}$$

For non-normal responses the above log-likelihood cannot be solved analytically. To overcome the integral in Eq (3), we consider in the following section closed-form approximations for the likelihood.

2.1 Approximations to the marginal likelihood of GLLVMs

Computationally, the most efficient likelihood based approaches for estimating GLLVMs are methods which approximate the marginal likelihood in a closed form. Of these, the most common and well known is the Laplace approximation method, which has been used extensively in the statistical literature to approximate marginal likelihood functions that cannot be resolved analytically [30]. The Laplace approximation can be easily applied to a marginal likelihood $l(\Psi) = \sum_{i=1}^n \log \int f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \Psi) f(\boldsymbol{u}_i^*)\, d\boldsymbol{u}_i^*$ with latent variables $\boldsymbol{u}_i^*$. By denoting $Q(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \Psi) = \log\{f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \Psi) f(\boldsymbol{u}_i^*)\}/m$, the likelihood can be written as $l(\Psi) = \sum_{i=1}^n \log \int \exp(mQ(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \Psi))\, d\boldsymbol{u}_i^*$. Assuming further that $\hat{\boldsymbol{u}}_i^*$ maximizes $Q(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \Psi)$, the Laplace approximation method applies a second order Taylor expansion for $Q(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \Psi)$ around the maximum $\hat{\boldsymbol{u}}_i^*$, and thus allows the integral to be performed in a tractable manner (it resembles the normalization constant for a multivariate normal distribution). For GLLVMs, the Laplace approximation was first proposed in [20], and extended by [21] to handle important distributions arising in ecology such as the negative binomial, Poisson, zero inflated Poisson and Tweedie distributed responses. For a model as defined in Eq (1) with random row effects and responses y_{ij} coming from the exponential family of distributions with mean μ_{ij} as defined in (2), the Laplace approximation of the marginal log-likelihood function can be written as follows:

$$\begin{aligned}\tilde{l}(\boldsymbol{\Psi}) \quad = \sum_{i=1}^n \Bigg(& -\frac{1}{2}\log\det\{\mathbf{G}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i^*)\} + \sum_{j=1}^m \left\{\frac{y_{ij}\, a(\hat{\eta}_{ij}) - b(\hat{\eta}_{ij})}{\phi_j} + c(y_{ij};\phi_j)\right\} \\ & -\frac{1}{2}\hat{\boldsymbol{u}}_i^{*\prime}\mathbf{C}_{\sigma^2}^{-1}\hat{\boldsymbol{u}}_i^* - \frac{1}{2}\log\det(\mathbf{C}_{\sigma^2})\Bigg),\end{aligned}$$

where

$$\mathbf{G}(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i^*) = \sum_{j=1}^m \frac{\partial^2\{-y_{ij}\, a(\eta_{ij}) + b(\eta_{ij})\}}{\partial \boldsymbol{u}_i^{*\prime}\partial \boldsymbol{u}_i^*}\bigg|_{\boldsymbol{u}_i^* = \hat{\boldsymbol{u}}_i^*} + \mathbf{C}_{\sigma^2},$$

$\hat{\eta}_{ij} = \beta_{0j} + x_i'\boldsymbol{\beta}_j + \hat{\boldsymbol{u}}_i^{*\prime}\boldsymbol{\gamma}_j^*$, $\boldsymbol{C}_{\sigma^2} = bdiag(\sigma^2, \boldsymbol{I}_d)$, $\hat{\boldsymbol{u}}_i^* = (\alpha_i, \boldsymbol{u}_i')'$ and $\hat{\boldsymbol{u}}_i^*$ maximizes

$$Q(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \boldsymbol{\Psi}) = \frac{1}{m}\left(\sum_{j=1}^{m}\left\{\frac{y_{ij}a(\eta_{ij}) - b(\eta_{ij})}{\phi_j} + c(y_{ij};\phi_j)\right\} - \frac{1}{2}\boldsymbol{u}_i^{*\prime}\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{u}_i^* - \frac{1}{2}\log\det(\boldsymbol{C}_{\sigma^2})\right)$$

with respect to $\boldsymbol{u}_i^*$. All quantities that are constant with respect to the parameters have been omitted. Some further simplification of this expression is possible when the model is defined using a canonical link function [21].

When using Laplace approximations, the estimation is performed by maximizing $\tilde{l}(\boldsymbol{\Psi})$ with respect to $\boldsymbol{\Psi}$, and $Q(\boldsymbol{y}_i, \boldsymbol{u}_i^*, \Psi)$ with respect to $\boldsymbol{u}_i^*$. The estimates $\hat{\boldsymbol{u}}_i^*$ are then used as predictions of the latent variables. Furthermore, asymptotic standard errors for $\hat{\boldsymbol{\Psi}}$ and $\hat{\boldsymbol{u}}_i^*$ are computed as the negative Hessian matrix obtained as part of the estimation process. These may form the basis for performing statistical inference for the model parameters and evaluate prediction errors for the latent variables, both of which will be examined empirically in the simulation studies in Section 4.

Another method which allows us to derive a closed form approximation for the marginal likelihood is the variational approximation method. The idea of variational approximations originates from machine learning research, where it is often used to approximate probability densities [31]. More recently, the method has gained considerable traction in Bayesian data analysis for efficiently approximating posterior densities [32, 33]. The variational approximation method is also applicable in likelihood based contexts for approximating an intractable marginal likelihood [34], although it is less frequently used in this context. Furthermore, the large sample properties of estimates and inference obtained using the variational approximation method are not thoroughly studied and remain a topic of future research [33].

The main idea behind likelihood based variational approximations is to approximate the posterior distribution of the random effects i.e., $f(\boldsymbol{u}_i^*|\boldsymbol{y}_i, \boldsymbol{\Psi})$ by a simpler distribution in order to get a closed form (or almost closed-form) expression for the marginal log-likelihood. This so called variational likelihood is a strict lower bound to the marginal log-likelihood, and is then treated as the new objective function on which to base estimation and inference. In practice, for a marginal log-likelihood function $l(\Psi) = \sum_{i=1}^{n}\log\int f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \Psi)f(\boldsymbol{u}_i^*)\,d\boldsymbol{u}_i^*$, the variational approximation approach make use of Jensen's inequality to construct this lower bound,

$$\begin{aligned}\sum_{i=1}^{n}\log\int f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \boldsymbol{\Psi})f(\boldsymbol{u}_i^*)d\boldsymbol{u}_i^* &= \sum_{i=1}^{n}\log\int\left\{\frac{f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \boldsymbol{\Psi})f(\boldsymbol{u}_i^*)q(\boldsymbol{u}_i^*|\boldsymbol{\xi})}{q(\boldsymbol{u}_i^*|\boldsymbol{\xi})}\right\}d\boldsymbol{u}_i^* \\ &\geq \sum_{i=1}^{n}\int\log\left\{\frac{f(\boldsymbol{y}_i|\boldsymbol{u}_i^*, \boldsymbol{\Psi})f(\boldsymbol{u}_i^*)}{q(\boldsymbol{u}_i^*|\boldsymbol{\xi})}\right\}q(\boldsymbol{u}_i^*|\boldsymbol{\xi})d\boldsymbol{u}_i^*,\end{aligned}$$

for some variational density $q(\boldsymbol{u}_i^*|\xi)$ with variational parameters $\boldsymbol{\xi}$. Critically, the logarithm can be brought inside the integral, thereby making integration easier for the exponential family of distributions. By maximizing the variational log-likelihood with respect to both the model parameters Ψ and variational parameters $\boldsymbol{\xi}$, we see that maximizing the variational likelihood is equivalent to minimizing the Kullback-Leibler divergence between the true posterior, $f(\boldsymbol{u}_i^*|\boldsymbol{y}_i, \Psi)$, and the proposed variational density $q(\boldsymbol{u}_i^*|\xi)$.

The variational approximation method was applied to the estimation of GLLVMs by [19] and it was shown that it is optimal in some sense to choose, as variational densities $q(\cdot)$, independent normal distributions for the latent variables for each observational unit. Following on

from this, for our GLLVM model in Eq (1) with random row effects we choose $q(\boldsymbol{u}_i^*|\xi_{\boldsymbol{u}_i^*}) = N_{d+1}(\boldsymbol{a}_i, \boldsymbol{A}_i)$ for $i = 1, \ldots, n$, where $\xi_{\boldsymbol{u}_i^*} = (\boldsymbol{a}_i, \text{vec}(\boldsymbol{A}_i)')'$, $\boldsymbol{A}_i = bdiag(A_{\alpha_i}, \boldsymbol{A}_{\boldsymbol{u}_i})$ and $\boldsymbol{A}_{\boldsymbol{u}_i}$ is an unstructured $d \times d$ covariance matrix. For responses coming from the exponential family of distributions with the canonical link function, this leads to the variational approximation of the GLLVM log-likelihood as follows:

$$\ell(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_{q^*}\{b(\eta_{ij})\}}{\phi_j} + c(y_{ij}, \phi_j)\right\} + \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - \text{tr}(\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{A}_i) - \boldsymbol{a}_i'\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{a}_i - \log\det(\boldsymbol{C}_{\sigma^2})\right),$$

where $\tilde{\eta}_{ij} = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{a}_i'\boldsymbol{\gamma}_j^*$, $\boldsymbol{C}_{\sigma^2} = bdiag(\sigma^2, \boldsymbol{I}_d)$ and a_i and $\boldsymbol{A}_i$ are the mean and the covariance matrix of a variational density, respectively. All quantities constant with respect to the parameters have been omitted. Notice the lower bound includes the expectation term $E_{q^*}\{b(\eta_{ij})\}$, which is not guaranteed to have a closed form for any distribution form the exponential family. Through reparameterization of the GLLVM, fully explicit forms for $\underline{\ell}(\boldsymbol{\Psi}, \xi)$ can be derived for some common occurring responses in multivariate abundance data, such as binary, ordinal and overdispersed count responses [19].

One attractive feature of likelihood based variational approximations is that the estimated means of the variational distributions, $\hat{\boldsymbol{a}}_i$, $i = 1, \ldots, n$, provide a natural predictor for the latent variables $\boldsymbol{u}_i^*$, while the estimated covariance matrices $\hat{\boldsymbol{A}}_i$ along with the assumed variational density $q(\boldsymbol{u}_i^*|\xi)$ can be used as the basis for constructing prediction intervals [34]. Both quantities are obtained directly from the maximization procedure. Furthermore, asymptotic standard errors for the model parameters can be obtained by using the block inverse matrix of the negative Hessian of $\underline{\ell}(\boldsymbol{\Psi}, \xi)$, (see also [35]).

3 Implementation

Two advances are made in this paper, which enable faster, more reliable fitting of GLLVMs than previous implementations of Laplace or variational approximations. First, we write software to make use of automatic differentiation software in the `TMB` package [22]. Secondly, we make strategic choices for the starting values of the parameters in the GLLVM, in order to improve speed and stability of the estimation algorithms. Our simulations presented later demonstrate that these changes improve speed by an order of magnitude, as well as improving reliability by increasing the accuracy of the estimates.

3.1 Implementation with `TMB`

The closed form approximate marginal log-likelihoods proposed in the previous section are often maximized using some gradient-based optimization algorithms. This presents a computational challenge as it means that the gradient functions need to be calculated for each response distribution and specific model separately. To overcome this, we use Template Model Builder (`TMB`) for fitting GLLVMs. `TMB` is a general `R` package for fitting non-linear mixed effects and latent variable models based on AD Model Builder, which is a `C++` language extension for solving statistical optimization problems using automatic differentiation techniques [23]. To perform optimization using `TMB` in general, the complete log-likelihood for the model of interest is written in `C++`, from which `TMB` employs the `C++` library 'CppAD' to efficiently construct functions for calculating the associate gradient and Hessian. These functions written can then be called from `R`, and can be straightforwardly passed into gradient based optimization methods

such as `optim()` or `nlminb()`. After optimization, the Hessian matrix is obtained as a side product and can be used to calculate standard errors for parameters. Note however initialization of the model and the choice of starting values must be done in `R`.

For models involving random effects, `TMB` uses the Laplace approximation method. As a result, we can straightforwardly adapt it for maximizing the Laplace approximation of the GLLVM log-likelihood in Section 2.1 based on the following steps:

1. Write the complete log-likelihood for the responses and latent variables in `C++` using the `TMB` model template and compile it.
2. Set initial values for the model parameters and the latent variables in `R`; see Section 3.2.
3. Create the `TMB` object using `TMB::MakeADFun()` with data, initial values and the objective function as input, specifying the names of the parameters to be integrated out of the likelihood using argument `random` in `TMB::MakeADFun()`. The Laplace approximation method will then be automatically applied to the complete likelihood, and gradient and Hessian functions for the marginal log-likelihood will be constructed.
4. Optimize the objective function using `optim()` or `nlminb()` in `R`.
5. Calculate the Hessian matrix in `R` using `optimHess()`, from which the standard errors for the model parameters as well as prediction errors for the latent variables can be obtained.

Notice that the initialization in Step 2 is crucial for the model fitting as poor initial values may yield to convergence problems. We return to the selection of starting values in Section 3.2.

Since `TMB` allows maximization of any likelihood function, it can also be used to optimize the variational approximation to the marginal log-likelihood for GLLVMs. In this case, we can treat the variational parameters ξ as additional model parameters and maximize the variational approximation to the log-likelihood based on the following steps:

1. Write the variational approximation lower bound for the log-likelihood in `C++` using `TMB` model template and compile it.
2. Set initial values for the model parameters and the variational parameters in `R`; see Section 3.2.
3. Create the `TMB` object using `TMB::MakeADFun()` with data, initial values and the objective function as input. The gradient and Hessian for the variational approximated log-likelihood will then be automatically calculated using `TMB::MakeADFun()`.
4. Optimize the objective function using `optim()` or `nlminb()` in `R`.
5. Calculate the Hessian matrix in `R` using `optimHess()`, from which standard errors for the model parameters as well as prediction errors for the latent variables may be obtained by applying block inversion for the negative Hessian matrix.

Finally, for all the implementations we considered, we parameterized any dispersion parameters and variance components in terms of their log transformed values in to avoid boundary issues in estimation and inference i.e. $\log(\sigma)$, $\log(\phi)$, and so on.

3.2 Starting values

With GLLVMs and models involving a large number of latent random effects, the importance of selecting the initial values of model parameters is particularly important. When the

observed likelihood function is multimodal, maximization algorithms can often end up in local maxima if the initial values for parameters are not sufficiently close enough to the global maximum. A widely used strategy to work around this issue is to use several random starting values and to pick up the solution with highest log-likelihood value. In case of complex models and large datasets however, the use of several random starting values may however be too time consuming.

We propose a new data driven method for constructing initial values for parameters in a GLLVM. In this approach, we first fit a GLM, $g(E(y_{ij})) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j$, to each response variable (species), from which the obtained estimates of β_{0j} and β_j are used as starting values for the fixed parameters in the GLLVM. Starting values for latent variables u_i and their loadings γ_j are then constructed by applying factor analysis to the Dunn-Smyth residuals [29] from the fitted GLMs. Furthermore, the matrices of starting values for the latent variables and the loadings obtained via factor analysis are rotated so that the upper triangle of the loading matrix is zero, so as to adhere to the parameter identifiability constructed below Eq (1). As starting values for the random row effects, we use a vector of zeros. The key idea underlying this approach to constructing starting values lies in the Dunn-Smyth residuals, which are defined for the observation y_{ij} as

$$r_{ij} = \Phi^{-1}(z_{ij}F_{ij}(y_{ij}) + (1 - z_{ij})F_{ij}^{-}(y_{ij})), \tag{6}$$

where Φ and F_{ij} are the cumulative distribution functions of the standard normal distibution and the response variable, respectively, F_{ij}^{-} is the limit as F_{ij} is approached from the negative side, and z_{ij} is a random variable generated from the standard uniform distribution. Dunn-Smyth residuals have the attractive property that if model assumptions are correct, then the residuals are exactly normally distributed. The normality of the residuals motivates us to use the classical factor analysis on the residuals from the fitted GLMs, in particular, because they contain information regarding the residual correlation across species not accounted for by the observed covariates. For the remainder of this article, we will refer to this method for constructing starting values as `res`.

An extension to the above method is `resX`, where the starting values are obtained in a similar fashion as in `res`, with the crucial difference being that `resX` uses X sets of starting values for the latent variables. These are obtained by "jittering" starting values by adding random variation from a normal distribution to the latent variables obtained using `res`. In our simulation studies we use a jitter variance of 0.2^2 and $X = 3$ sets of starting values (we will thus refer to this approach as `res3` in Section 4). With X sets of starting values, which only differ in the latent variables (the starting values for the $\boldsymbol{B}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Phi}$ remain the same), the estimation procedure then proceeds as we would with random starting values. That is, a GLLVM is fitted using those X different sets of starting values, and the fit with the highest log-likelihood value is then considered the best fitting GLLVM for that dataset.

In the simulation studies in the following section, we will compare `res` and `res3` to two alternative and common methods for constructing starting values: 1) a method referred to as `zero`, where we use zero initial values for all parameters; 2) a method referred to as `random`, where we simulate initial values for latent variables from a multivariate standard normal distribution, while (as previously) a GLM is fitted to each response variable against environmental variables and latent variables to get starting values for fixed parameters and loadings. Note that the difference between `random` and `res`/`res3` is that the latter makes use of the residual information from the multivariate GLM to directly construct the starting values for the latent variables and loadings, while the former simulates these randomly.

4 Simulation studies

We performed a series of simulation studies to compare the performance of different model fitting algorithms with and without automatic differentiation using `TMB`, using either the Laplace approximation or variational approximation, and with different starting value strategies (`res`, `res3`, `zero`, `random`). For fitting algorithms without automatic differentiation, we implemented both the Laplace and variational approximations in plain `R` code by manually defining their respective approximate likelihoods and their gradient functions. Details of the simulation design are given below.

4.1 Simulation designs

We considered GLLVMs with multivariate count and binary data, and based our simulation studies on two real datasets: the first dataset consists of abundances of testate amoebae in Finnish peatlands [36], and the second dataset consists of abundances of bird species in Indonesia [37].

The first simulation setup was based on the testate amoebae data [36], which consist of counts of $m = 48$ testate amoebae species measured from $n = 263$ sampling sites across six peatlands in southern and central Finland. Two environmental variables, water pH and water temperature, were also recorded at each sampling site. We conducted simulation studies based on the original count data as well as based on binary data obtained by converting counts to presence-absences. As mean models, we used $\log(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for counts and $\Phi^{-1}(\mu_{ij}) = \beta_{0j} + \boldsymbol{x}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for presence-absences, where x_i includes the values for the two covariates recorded at site i, and u_i includes two latent variables. Notice that with two-dimensional latent variables, GLLVMs can be used as a model-based ordination method as described in [7]. The parameters for the true model used to simulate multivariate abundance data were obtained by fitting a negative binomial (Bernoulli) GLLVM to the real data, consisting of counts (presence-absences) of observed amoebae species. To study the effect of sample size on performance, we constructed nested subsets of size $n = 50$, 120, 190 and 260 randomly sampling from the sites and used parameters of the fitted model, which corresponded the sites in subsets, to generate datasets of the desired sizes. We generated $K = 500$ datasets for each value of n, and for each dataset we fitted GLLVMs using the four starting value strategies and both approximation methods with and without automatic differentiation.

The second simulation setup was based on Indonesian bird data [37], which consists of counts of $m = 177$ bird species measured from $n = 37$ sites in Central Kalimantan, Indonesia. We conducted a simulation study for the original count data as well as for the binary data obtained by converting counts to presence-absences. We used $\log(\mu_{ij}) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for counts and $\Phi^{-1}(\mu_{ij}) = \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for presence-absence data, with parameters for the true model based on a negative binomial GLLVM fitted to the count data and a Bernoulli GLLVM fitted to the binary data. In this simulation study, we varied the number of species, that is, we used four different numbers of randomly selected species, $m = 30$, 60, 100 and 140. As in the previous setup, the parameters for the true model were obtained by fitting a negative binomial (Bernoulli) GLLVM to the data in the case of counts (presence-absences), and the parameters that corresponded the species in each subset were used obtain a dataset of the desired size. For each value of m, we generated $K = 500$ datasets, and for each dataset we fitted GLLVMs using four different starting value strategies and both approximation methods with and without automatic differentiation.

In addition to the above two simulation setups, we included another design based on the Indonesian birds data, where we added a random row effect to the simulation model.

Specifically, the true mean models were given by $\log(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for counts and $\Phi^{-1}(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{u}_i'\boldsymbol{\gamma}_j$ for presence-absence data, where α_i is a random effects assumed to follow a normal distribution with zero mean and variance 0.25. We fitted these models with random row effects using `TMB` only. The reason for this is that the plain `R` implementations of [21] do not cater for random row effects, and other simulations had already demonstrated that the `TMB` implementation is more computationally efficient.

Note that the first simulation setup, based on a dataset with a large sample size, varied n, while the second simulation setup, based on a dataset with a species rich community (large m), varied m. Hence we looked at the effects of varying each of sample size and of number of responses, but do so one simulation at a time. These simulations were computationally intensive, with a total running time across all simulations of 5 weeks on a Intel Xeon E7-8837 (2.67GHz) processor with 25 CPUs.

4.2 Overdispersed counts

We being by presenting the results from negative binomial GLLVM under the first simulation design, and compared variational approximation and Laplace approximation methods implemented with and without `TMB`, using the starting value method `res`; see Section 4.4 for the reason behind this choice of starting value approach. Fig 1 plots the median computation times, and demonstrates that the variational approximation method implemented using `TMB` was substantially faster than the other estimation methods. The `TMB` implementation of the Laplace approximation method was also faster than the plain `R` implementation for the smallest sample size.

The results in Table 1 suggest that the advantages in computation time did not come at the cost of estimation and inferential accuracy. In fact, the average biases across all species and root mean squared errors tended to be smaller for the variational approximation method compared to the Laplace approximation method. With very small n, the differences between the two approximation methods were particularly noticeable. For both methods, the estimates for log-dispersion parameters were comparably biased when the sample size was very small. When the sample size increased, the variational approximation method in particular performed better, with differences between the two variational approximation implementations becoming very small. For the Laplace approximation method, although the differences in average biases were small, the differences in coverage probabilities and mean confidence interval widths were comparably larger than its variational counterpart. Furthermore, the implementation which did not use `TMB` tended to provide overly narrow confidence intervals for almost all parameters.

In order to evaluate the performance of the estimated latent variable loadings, $\hat{\gamma}_j$, and predicted latent variables, $\hat{\boldsymbol{u}}_i$, we list in Table 2 the mean Procrustes errors between the estimated and the true values ([28], Chapter 8.4). These are scaled according to the sample size and number of species to make comparisons easier. Results indicated that for small n, compared to the Laplace approximation method, the variational approximation method produced smaller Procrustes errors for both latent variables and loadings. As expected, the difference between Procrustes errors based on different methods decreased when n increased.

In addition to the results presented in Tables 1 and 2, we also evaluated the acurracy of competing models by adapting the variation explained based on cross-validation (denoted here as VE), as proposed in [38, 39], for our text with simulated binary and count data. Specifically, for each simulation setup we compared the predictive performance of the correponding GLLVM to the null model i.e, a model including only an species-specific intercept only, using

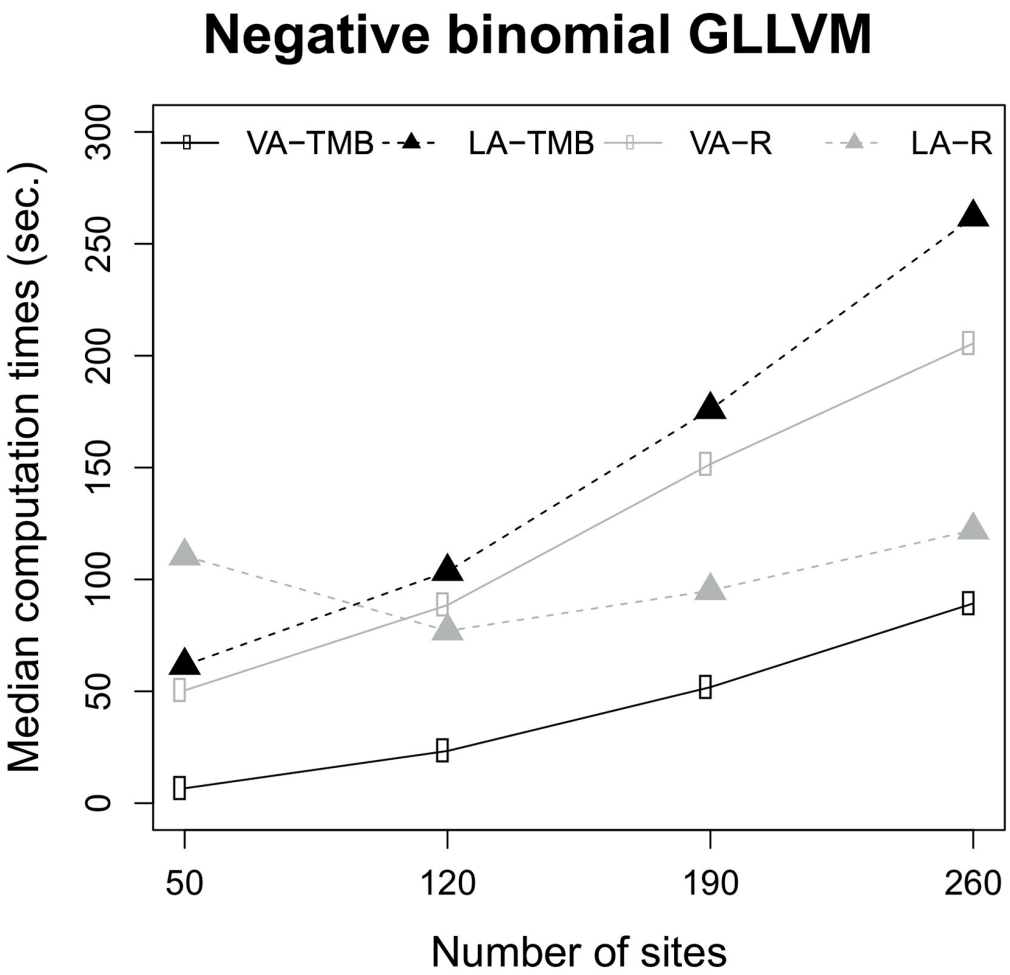

Fig 1. Median computation times for negative binomial GLLVMs. Times for the plain R (gray) and the TMB implementations (black) for the variational approximation (VA, solid line) method and the Laplace approximation (LA, dashed line) method for a negative binomial GLLVM with two covariates and two latent variables. The simulation setup was based on testate amoebae data.

https://doi.org/10.1371/journal.pone.0216129.g001

the formula

$$VE_k = 1 - \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} |\hat{\mu}_{ij}^{(k)} - \mu_{ij}|}{\sum_{i=1}^{n}\sum_{j=1}^{m} |\hat{\mu}_{ij,null}^{(k)} - \mu_{ij}|},$$

where for the kth simulated dataset with $k = 1, \ldots, 500$, the quantities $\hat{\mu}_{ij}^{(k)}$ and $\hat{\mu}_{ij,null}^{(k)} = g^{-1}(\hat{\beta}_{0j})$ denote the predicted means from the fitted GLLVM and from a null model, respectively. The true means, which were used to generate the training datasets, are denoted by μ_{ij}. Because we are using simulated data and therefore can generate multiple training datasets, as opposed to a real application where we only have the one realized dataset, then there is less motivation to use cross-validation when calculating VE i.e, the natural variation across folds can be well

Table 1. Average biases, root mean squared errors (RMSE), coverage probabilities of 95% confidence intervals and mean confidence interval widths (CI) for negative binomial GLLVM estimates based on the plain R and the TMB implementations for the variational approximation and the Laplace approximation methods. The true model parameters were obtained by fitting a negative binomial GLLVM with two environmental covariates for the testate amoebae data with counts of $m = 48$ species recorded at $n = 50$, 120, 190 and 260 sites. Parameter β_0 refers to the species specific intercepts, β_{pH} and β_{temp} to the coefficients of water pH and water temperature and log ϕ to the log transformed dispersion parameters.

n		VA-TMB				LA-TMB			
		Bias	RMSE	Cover	CI	Bias	RMSE	Cover	CI
50	β_0	-0.32	0.85	0.94	3.09	-0.92	2.24	0.93	5.14
	β_{pH}	-0.03	0.63	0.95	2.44	0.01	0.90	0.95	2.94
	β_{temp}	0.02	0.73	0.93	2.76	-0.05	0.97	0.93	3.31
	log ϕ	-0.38	0.67	0.92	2.35	-2.80	5.12	0.95	76.72
120	β_0	-0.05	0.49	0.94	1.78	-0.33	0.99	0.95	2.53
	β_{pH}	-0.04	0.40	0.95	1.55	-0.01	0.46	0.95	1.67
	β_{temp}	0.02	0.37	0.96	1.48	0.00	0.46	0.96	1.65
	log ϕ	-0.06	0.36	0.94	1.48	-0.59	1.57	0.95	5.13
190	β_0	0.03	0.40	0.92	1.36	-0.19	0.62	0.96	1.80
	β_{pH}	-0.04	0.32	0.95	1.20	-0.01	0.34	0.95	1.27
	β_{temp}	0.01	0.30	0.97	1.24	0.00	0.36	0.96	1.34
	log ϕ	0.02	0.30	0.93	1.16	-0.24	0.62	0.95	1.81
260	β_0	0.07	0.36	0.91	1.15	-0.13	0.46	0.96	1.46
	β_{pH}	-0.04	0.27	0.96	1.05	-0.02	0.29	0.96	1.10
	β_{temp}	0.01	0.25	0.97	1.05	0.01	0.29	0.97	1.11
	log ϕ	0.06	0.28	0.91	0.99	-0.15	0.36	0.95	1.24
		VA-R				LA-R			
50	β_0	-0.31	0.85	0.95	3.15	-0.94	2.34	0.84	4.60
	β_{pH}	-0.03	0.63	0.95	2.48	-0.00	0.86	0.72	2.18
	β_{temp}	0.02	0.73	0.94	2.80	-0.05	0.98	0.67	2.19
	log ϕ	-0.38	0.67	0.93	2.42	-1.44	2.39	0.51	3.27
120	β_0	-0.05	0.49	0.95	1.79	-0.31	0.97	0.89	2.17
	β_{pH}	-0.04	0.40	0.95	1.56	-0.02	0.48	0.79	1.54
	β_{temp}	0.02	0.37	0.96	1.49	0.00	0.46	0.81	1.61
	log ϕ	-0.06	0.36	0.95	1.49	-0.40	0.86	0.56	0.85
190	β_0	0.03	0.40	0.92	1.37	-0.18	0.60	0.91	1.55
	β_{pH}	-0.04	0.32	0.95	1.20	-0.02	0.39	0.77	1.22
	β_{temp}	0.01	0.30	0.97	1.24	-0.00	0.39	0.79	1.30
	log ϕ	0.02	0.30	0.93	1.17	-0.21	0.48	0.58	0.63
260	β_0	0.07	0.36	0.91	1.15	-0.12	0.45	0.89	1.26
	β_{pH}	-0.04	0.27	0.96	1.05	-0.03	0.39	0.71	1.04
	β_{temp}	0.01	0.25	0.97	1.05	0.01	0.34	0.77	1.11
	log ϕ	0.06	0.28	0.91	0.99	-0.13	0.35	0.59	0.53

https://doi.org/10.1371/journal.pone.0216129.t001

accounted by the natural variation across simulated datasets. Also, note because we are working with discrete data, then we choose to calculate VE based on the predicted mean scale μ_{ij} rather than on the response scale. The median VE values for negative binomial GLLVMs fitted to counts simulated based on amoebae dataset are listed in Table 3. The results indicate that the predictive accuracy improves as the number of sites increases. The accuracy is slightly higher when the variational approximation method is used. Further, when $n > 50$, the Laplace approximation method using the R implementation gives clearly lower VE values than the method using the TMB implementation.

Table 2. Scaled mean Procrustes errors of predicted latent variables and estimated latent variable loadings for negative binomial GLLVM estimates based on the plain R and the TMB implementations for the variational approximation and the Laplace approximation methods. The true model parameters were obtained by fitting a negative binomial GLLVM for the testate amoebae data with counts of $m = 48$ species recorded at $n = 50$, 120, 190 and 260 sites.

n	VA-TMB		LA-TMB		VA-R		LA-R	
	LVs	Loadings	LVs	Loadings	LVs	Loadings	LVs	Loadings
50	0.256	0.346	0.296	0.497	0.256	0.347	0.328	0.489
120	0.198	0.198	0.208	0.296	0.198	0.198	0.219	0.276
190	0.185	0.147	0.189	0.213	0.185	0.148	0.213	0.195
260	0.177	0.118	0.179	0.150	0.177	0.119	0.216	0.135

https://doi.org/10.1371/journal.pone.0216129.t002

The simulation results based on the negative binomial GLLVMs fitted for Indonesian bird data, with and without random row effect are given in S2 Appendix. Broadly speaking, they returned similar conclusions to those reported above. However, for both methods the log standard deviations of the random row effects were highly biased when the number of species was $m = 30$ but accuracy improved substantially with larger m. In addition, the predictive accuracy improves when the number of species increases.

4.3 Binary responses

Below we use the second simulation design to compare the performance of both approximation methods implemented with and without `TMB` for GLLVMs with binary responses. As previously, starting values obtained via the `res` method.

Fig 2 presents the computation times of various methods used to fit GLLVMs to binary responses. Similar to the simulation involving overdispersed counts, the variational approximation method implemented using `TMB` was substantially faster than all the other methods for all considered cases. It was also interesting to note that the median computation times for the Laplace approximation method implemented using `TMB` scaled very poorly with increasing n.

Table 4 lists the average biases, root mean squared errors, 95% coverage probabilities and mean confidence interval widths for estimates of the GLLVM without random row effects from different estimation methods. As in the case of overdispersed counts, the number of species did not have much effect on the estimates of species specific intercepts, β_0. The variational approximation method performed better overall in each of the considered cases, producing less biased estimates, smaller root mean squared errors and coverage probabilities closer to the nominal coverage level of 95%. By contrast, the estimates based on the Laplace approximation were severely biased, especially when the sample size was small. When m increased, the biases became smaller for both methods and the coverage probabilities approached to the nominal 95% level when the Laplace approximation were used. Results for the scaled mean Procrustes errors in Table 5 showed that errors were tended to be smaller when the variational

Table 3. Median VE values of negative binomial GLLVMs for 500 simulated datasets using the plain R and the TMB implementations for the variational approximation and the Laplace approximation methods. The datasets were based on a negative binomial GLLVM fitted for the testate amoebae data with counts of $m = 48$ species recorded at $n = 50$, 120, 190 and 260 sites.

n	VA-TMB	LA-TMB	VA-R	LA-R
50	0.27	0.19	0.27	0.21
120	0.48	0.43	0.42	0.29
190	0.53	0.50	0.53	0.35
260	0.56	0.54	0.56	0.39

https://doi.org/10.1371/journal.pone.0216129.t003

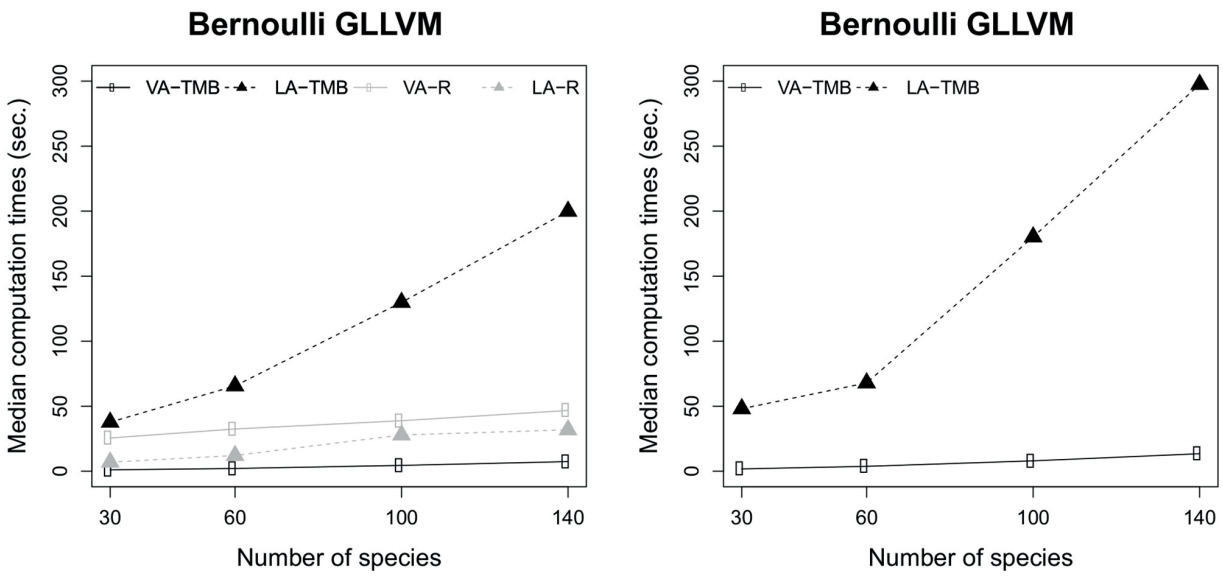

Fig 2. Median computation times for Bernoulli GLLVMs. Times for the plain `R` (gray) and the `TMB` implementations (black) for the variational approximation (VA, solid line) method and the Laplace approximation (LA, dashed line) method for a Bernoulli GLLVM with two latent variables. The left plot is for the model without row effects and right one with random row effects. The simulation setup was based on the Indonesian birds data.

https://doi.org/10.1371/journal.pone.0216129.g002

approximation method was used in estimation compared to the Laplace approximation method. As in the simulation settings with overdispersed counts, the mean Procrustes errors for latent variables predictions decreased with an increasing number of species m.

Variation explained was computed for Bernoulli GLLVMs as in Section 4.2, and the median VE values are listed in Table 6. Based on the results, differences in predictive accuracies improve with increasing m. The variance explained is substantially lower for the Laplace approximation method compared to the variational approximation method when number of species is small, but equally good for large m.

Supporting information S2 Appendix reports results for simulations based on the Indonesian bird dataset with a random row effect, and for simulations based on the testate amoebae

Table 4. Average biases, root mean squared errors (RMSEs), coverage probabilities of 95% confidence intervals and mean confidence intervals widths (CI) for GLLVM estimates based on the plain `R` and the `TMB` implementations for the variational approximation and the Laplace approximation methods. The true model parameters were obtained by fitting a Bernoulli GLLVM with probit link function for the Indonesian birds data with presence-absences of m = 30, 60, 100 and 140 species recorded at n = 37 sites.

m		VA-`TMB`				LA-`TMB`			
		Bias	RMSE	Cover	CI	Bias	RMSE	Cover	CI
30	β_0	0.05	0.29	0.93	1.27	-4.43	18.24	0.73	5.22
60	β_0	-0.03	0.30	0.98	1.55	-0.22	7.77	0.89	5.23
100	β_0	-0.03	0.35	0.96	1.55	-0.05	5.37	0.92	3.19
140	β_0	-0.03	0.39	0.96	1.57	-0.04	1.04	0.92	2.07
		VA-R				LA-R			
30	β_0	0.05	0.29	0.93	1.27	-0.01	0.46	0.81	1.31
60	β_0	-0.03	0.30	0.98	1.54	-0.14	0.67	0.83	1.57
100	β_0	-0.03	0.35	0.96	1.55	-0.12	0.95	0.84	1.69
140	β_0	-0.03	0.39	0.96	1.56	-0.10	0.94	0.83	1.49

https://doi.org/10.1371/journal.pone.0216129.t004

Table 5. Scaled mean Procrustes errors of predicted latent variables and estimated latent variable loadings for GLLVM estimates based on the plain R and the TMB implementations for the variational approximation and the Laplace approximation methods. Values are scaled with the number of sites and number of species for comparisons. The true model parameters were obtained by fitting a Bernoulli GLLVM with probit link function for the Indonesian birds data with presence-absences of $m = 30, 60, 100$ and 140 species recorded at $n = 37$ sites.

m	VA-TMB		LA-TMB		VA-R		LA-R	
	LVs	Loadings	LVs	Loadings	LVs	Loadings	LVs	Loadings
30	0.556	0.122	0.615	0.140	0.556	0.122	0.615	0.173
60	0.185	0.098	0.204	0.160	0.185	0.098	0.204	0.141
100	0.129	0.095	0.144	0.130	0.129	0.095	0.144	0.139
140	0.098	0.091	0.109	0.121	0.098	0.091	0.109	0.126

https://doi.org/10.1371/journal.pone.0216129.t005

data when converted to presence-absence data. Results were broadly similar to those reported for β_0 in Table 4, with the variational approximation leading to more accurate and precise estimates, while the Laplace approximation method tended to produce severely biased estimates particularly at small sample sizes. For both approximation methods, the log standard deviations of the random row effects were biased when the number of species m was small.

4.4 Starting value comparisons

To study the sensitivity of model fitting results to starting values, we compared the performances of four starting value selection strategies explained in section 3.2. As a global performance measure, we used the log-likelihood values obtained from `res3` as a reference level, and compared differences between this and the three other methods (`res`, `zero`, `random`).

Boxplots of the differences in log-likelihood values are given in Fig 3 for negative binomial GLLVMs fitted for the Testate amoebae data with $n = 260$ sites and $m = 48$ species, and for Bernoulli GLLVMs fitted for the Indonesian bird data with $n = 37$ sites and $m = 140$ species. When the TMB implementation of the variational approximation method was used the differences between the log-likelihood values based on `res3` and the other three methods were relatively small. The biggest differences were seen when the Laplace approximation method and the variational approximation method were implemented without `TMB` and applied to binary data. The full results with simulated datasets of different sizes may be found in S3 Appendix. In all of the considered cases, `res3` and `res` were consistently among the best starting values strategies giving the highest log-likelihood values, while the performances of `zero` and `random` depended strongly on the simulation setup.

In addition to the differences in log-likelihood values illustrated in Fig 3 for Bernoulli GLLVMs and in S3 Appendix for negative binomial GLLVMs, we also list for binary responses of the Indonesian bird data the average biases, root mean squared errors, 95% coverage probabilities and mean confidence interval widths for species specific intercept estimates as well as

Table 6. Median VE values of Bernoulli GLLVMs for 500 simulated datasets using the plain R and the TMB implementations for the variational approximation and the Laplace approximation methods. The datasets were based on a Bernoulli GLLVM with probit link function fitted for the Indonesian birds data with presence-absences of $m = 30, 60, 100$ and 140 species recorded at $n = 37$ sites.

m	VA-TMB	LA-TMB	VA-R	LA-R
30	0.23	0.08	0.24	0.08
60	0.30	0.28	0.30	0.26
100	0.34	0.30	0.31	0.31
140	0.36	0.35	0.36	0.36

https://doi.org/10.1371/journal.pone.0216129.t006

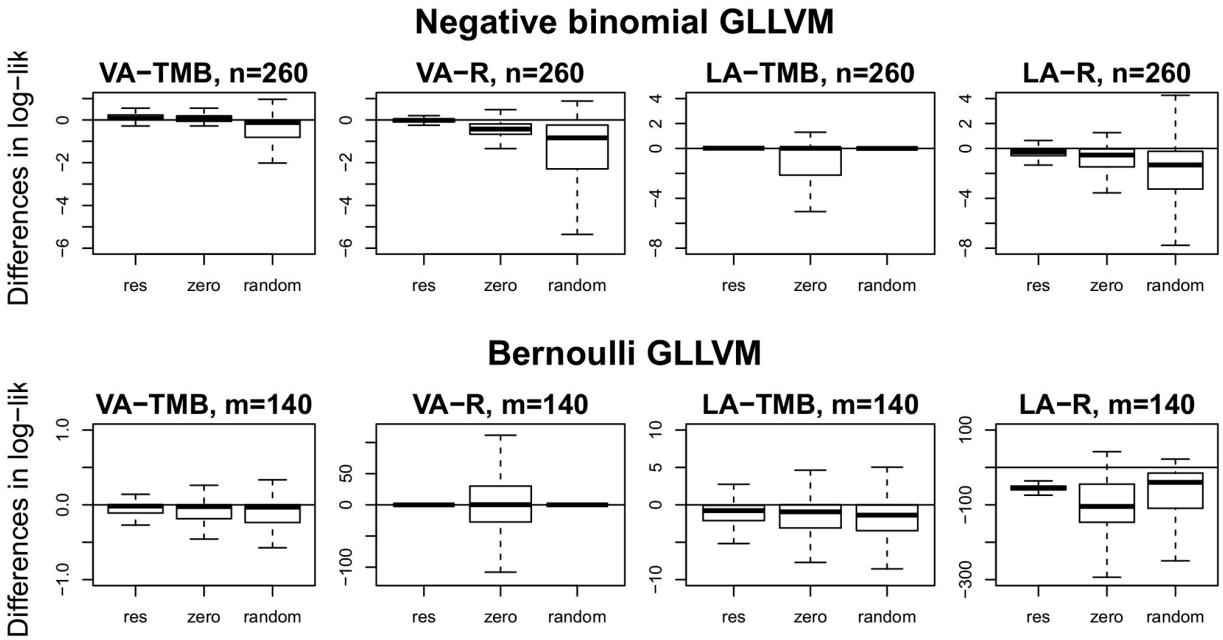

Fig 3. Differences in log-likelihood value when strategies `res`, `zero` and `random` are compared to `res3`. The true models were based on negative binomial GLLVM fitted for the Testate amoebae data with $n = 260$ sites and Bernoulli GLLVM fitted for the Indonesian bird data with $m = 140$ species. A negative value means that performance of the corresponding starting value strategy is worse than that of `res3`. Notice that columns have different scales.

https://doi.org/10.1371/journal.pone.0216129.g003

scaled mean Procrustes errors of predicted latent variables and estimated latent variable loadings for all methods included in comparisons in S3 Appendix.

Overall, these findings suggest that `res` and `res3` were the best strategies for choosing starting values. All methods `res`, `zero` and `random` have been implemented as different options (with the same names) in the `R` package `gllvm` with `res` as the default.

5 Discussion

In this article, we studied two closed form approximations (the Laplace approximation and variational approximation) for the marginal log-likelihood of a generalized linear latent variable model. We showed how the closed form approximations can be implemented efficiently using automatic optimization techniques implemented in `R` with the help of the package `TMB`. In addition, a new method for choosing the starting values for our estimation algorithms was proposed. The performances of the two approximation methods and different starting values strategies were compared using several simulation studies for overdispersed count and binary data, which are often encountered in biological and ecological studies. Results indicated that for both response types the variational approximation implementations tended to outperform the Laplace approximation implementations, both in terms of computation speed and estimation and inferential accuracy. These findings are congruent with the results of Hui *et al.* [7], where the performance of the variational approximation method was compared to the Laplace approximation method and the MCEM algorithm for count and binary data, and also to Gauss-Hermite Quadrature in the case of binary data. However, more comprehensive comparisons between the variational approximation method and other estimation methods, eg. the Gauss-Hermite Quadrature, would be useful and interesting in the future.

The Laplace approximation method implemented without automatic optimization showed the poorest performance in all of the considered cases. The differences between the `TMB` and `R` implementations, especially with the Laplace approximation, are most likely due to the differences in the optimization algorithms. In the `R` implementation we used a block-coordinate optimization in which we cycled between iterative updates of one of regression coefficients, latent variables and nuisance parameters, until convergence. We postulate that this led to a less targeted exploration of the parameter space with an increased chance of getting trapped in a local maximum. In the case of binary data, the variational approximation implementations performed substantially better than their Laplace approximation counterparts. This supports earlier findings that the Laplace approximation method often performs poorly with highly discrete responses [40].

All simulation studies further showed that we can obtain more accurate predictions of the latent variables by increasing the number of species, m. For the Laplace method this is explained by the asymptotic error, which is known to be of order $O(m^{-1})$ [41]. Although not proven here, we conjecture that for the variational approximation method, the asymptotic error is $O(m^{-1})$; see also the heuristic proof of consistency in [19]. However, more accurate estimates for model parameters can be obtained only by increasing the sample size, n.

Another way to of obtaining more accurate estimates and inferential for the parameters in a GLLVM is by introducing structure that allows us to borrow strength across species (response) in order to estimate regression and/or loading parameters. Not only does this decrease the number of parameters in the model, it also means that these new parameters are a function of n and m, and thus accuracy of their estimation and inference should improve when either the number of sites and/or species increases. An examples is using functional traits in order to mediate the species environment relationships (sometimes called a "fourth corner model", [42]): the resulting fourth corner coefficients parameters are then common to all species and estimation should improve as both a function of n and m both. Fourth corner models with latent variables can also be fitted using the `R` package `gllvm`, which implements both the Laplace and variational approximation methods.

Comparison of computation times clearly indicate that the `TMB` implementation of the variational approximation method is much faster than that both implementations of the Laplace approximation, with the difference becoming greater when the data are higher-dimensional. There are a number of reasons for this: first, we specified the variational approximation of the likelihood directly in `C++`, while for a Laplace approximation we only specified the integrand, and asked the `TMB` package to use automatic differentiation to calculate a Laplace approximation. This automation of the Laplace approximation offers considerable flexibility, and makes it relatively easy to fit some quite complex models, because the joint likelihood in the integrand is usually relatively easy to derive. However, it seems that not specifying a fully closed form (approximated) marginal log-likelihood comes at a computational cost. Another reason for a difference in computational time is that all variational parameters are handled like fixed parameters, which makes estimation faster than dealing with random effects. The other possible reason for more rapid growth in computation time for the Laplace approximation method, when m increases, comes from the complexity of the approximation itself, where there is a term $\log\det\{G(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i^*)\}$, where $G(\boldsymbol{\Psi}, \hat{\boldsymbol{u}}_i^*)$ has dimension m, and so computing its determinant has a complexity that grows at a rate $O(m^3)$.

Overall, our findings suggest present a strong case for the use of the variational approximation method as a primary method for performing likelihood based estimation and inference in GLLVMs. Because it is relatively accurate and very quick, variational approximation on `TMB` provides a platform for upscaling analyses to large datasets. To date we have used the software to fit a dataset of size 174×985 in 61 minutes. In future work, we plan to generalize GLLVMs,

as well as the `gllvm` package, so that it can handle spatial and or temporal correlation inherent in the data, as well as offer some data-driven forms of order and variable selection (see for example [43]).

Supporting information

S1 Appendix. Proof of the variational approximation of the likelihood of GLLVMs.
(PDF)

S2 Appendix. Additional simulation results. Results of the negative binomial GLLVM simulation for the Indonesian birds data and the Bernoulli GLLVM simulation for the testate amoebae data.
(PDF)

S3 Appendix. Full results for the starting value comparisons.
(PDF)

S1 File. R code for simulations.
(R)

S2 File. Amoebae data.
(ZIP)

Acknowledgments

JN was supported by the Jenny and Antti Wihuri Foundation. ST was supported by CRoNoS COST Action IC1408. FKCH and DIW were funded by Australia Research Council Discovery Project grants (DP180100836 and DP180103543, respectively).

Author Contributions

Conceptualization: Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton.

Data curation: Jenni Niku.

Formal analysis: Jenni Niku.

Investigation: Jenni Niku.

Methodology: Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton.

Software: Jenni Niku, Wesley Brooks, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton.

Validation: Jenni Niku.

Visualization: Jenni Niku.

Writing – original draft: Jenni Niku, Riki Herliansyah.

Writing – review & editing: Jenni Niku, Riki Herliansyah, Francis K. C. Hui, Sara Taskinen, David I. Warton.

References

1. Legendre P, Legendre L. Numerical ecology (3rd edition). vol. 24. Elsevier; 2012.

2. Warton DI, Wright ST, Wang Y. Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution. 2012; 3:89–101. https://doi.org/10.1111/j.2041-210X.2011.00127.x
3. Warton DI, Hui FKC. The central role of mean-variance relationships in the analysis of multivariate abundance data: a response to Roberts (2017). Methods in Ecology and Evolution. 2017; 8:1408–1414. https://doi.org/10.1111/2041-210X.12843
4. Skrondal A, Rabe-Hesketh S. Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models. Boca Raton: Chapman & Hall; 2004.
5. McCullagh P, Nelder JA. Generalized Linear Models. Chapman & Hall; 1989.
6. Walker SC, Jackson DA. Random-effects ordination: describing and predicting multivariate correlations and co-occurrences. Ecological Monographs. 2011; 81(4):635–663. https://doi.org/10.1890/11-0886.1
7. Hui FKC, Taskinen S, Pledger S, Foster SD, Warton DI. Model-based approaches to unconstrained ordination. Methods in Ecology and Evolution. 2015; 6:399–411. https://doi.org/10.1111/2041-210X.12236
8. Warton DI, Blanchet FG, O'Hara R, Ovaskainen O, Taskinen S, Walker SC, et al. So many variables: Joint modeling in community ecology. Trends in Ecology and Evolution. 2015; 30:766–779. https://doi.org/10.1016/j.tree.2015.09.007 PMID: 26519235
9. Ovaskainen O, Abrego N, Halme P, Dunson D. Using latent variable models to identify large networks of species-to-species associations at different spatial scales. Methods in Ecology and Evolution. 2016; 7:549–555. https://doi.org/10.1111/2041-210X.12501
10. Thorson JT, Ianelli JN, Larsen EA, Ries L, Scheuerell MD, Szuwalski C, et al. Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. Global Ecology and Biogeography. 2016; 25:1144–1158. https://doi.org/10.1111/geb.12464
11. Ovaskainen O, Tikhonov G, Norberg A, Guillaume Blanchet F, Duan L, Dunson D, et al. How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters. 2017; 20:561–576. https://doi.org/10.1111/ele.12757 PMID: 28317296
12. Tikhonov G, Abrego N, Dunson D, Ovaskainen O. Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution. 2017; 8:443–452. https://doi.org/10.1111/2041-210X.12723
13. Bálint M, Bahram M, Eren AM, Faust K, Fuhrman JA, Lindahl B, et al. Millions of reads, thousands of taxa: microbial community structure and associations analyzed via marker genes. FEMS Microbiology Reviews. 2016; 40(5):686–700. https://doi.org/10.1093/femsre/fuw017 PMID: 27358393
14. Sammel MD, Ryan LM, Legler JM. Latent variable models for mixed discrete and continuous outcomes. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1997; 59:667–678. https://doi.org/10.1111/1467-9868.00090
15. Blanchet FG. HMSC: Hierachical modelling of species community; 2014. Available from: http://CRAN.R-project.org/package=HMSC.
16. Moustaki I. A latent trait and a latent class model for mixed observed variables. British Journal of Mathematical and Statistical Psychology. 1996; 49:313–334. https://doi.org/10.1111/j.2044-8317.1996.tb01091.x
17. Moustaki I, Knott M. Generalized latent trait models. Psychometrika. 2000; 65:391–411. https://doi.org/10.1007/BF02296153
18. Cagnone S, Moustaki I, Vasdekis V. Latent variable models for multivariate longitudinal ordinal responses. British Journal of Mathematical and Statistical Psychology. 2009; 62(2):401–415. https://doi.org/10.1348/000711008X320134 PMID: 18625083
19. Hui FKC, Warton DI, Ormerod JT, Haapaniemi V, Taskinen S. Variational approximations for generalized linear latent variable models. Journal of Computational and Graphical Statistics. 2017; 26:35–43. https://doi.org/10.1080/10618600.2016.1164708
20. Huber P, Ronchetti E, Victoria-Feser M. Estimation of generalized linear latent variable models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004; 66:893–908. https://doi.org/10.1111/j.1467-9868.2004.05627.x
21. Niku J, Warton DI, Hui FKC, Taskinen S. Generalized linear latent variable models for multivariate count and biomass data in ecology. Journal of Agricultural, Biological, and Environmental Statistics. 2017; 22:498–522. https://doi.org/10.1007/s13253-017-0304-7
22. Kristensen K, Nielsen A, Berg CW, Skaug H, Bell BM. TMB: Automatic Differentiation and Laplace Approximation. Journal of Statistical Software. 2016; 70(5):1–21. https://doi.org/10.18637/jss.v070.i05
23. Fournier D, Skaug H, Ancheta J, Ianelli J, Magnusson A, Maunder M, et al. AD Model Builder: using Automatic Differentiation for Statistical Inference of Highly Parameterized Complex Nonlinear Models.

Optimization Methods and Software. 2011; 27(2):233–249. https://doi.org/10.1080/10556788.2011.597854

24. Griewank A, Walther A. Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation. Society for Industrial and Applied Mathematics (SIAM). 2008;.
25. Thorson JT, Fonner R, Haltuch MA, Kotaro Ono K, Winker H. Accounting for spatiotemporal variation and fisher targeting when estimating abundance from multispecies fishery data. Canadian Journal of Fisheries and Aquatic Sciences. 2017; 74:1794–1807. https://doi.org/10.1139/cjfas-2015-0598
26. Albertsen CM, Whoriskey K, Yurkowski D, Nielsen A, Flemming JM. Fast fitting of non-Gaussian state-space models to animal movement data via Template Model Builder. Ecological Society of America. 2015; 96(10):2598–2604.
27. Niku J, Brooks W, Herliansyah R, Hui FKC, Taskinen S, Warton DI. gllvm: R package version 0.1.0. 2017;.
28. Bartholomew DJ, Knott M, Moustaki I. Latent variable models and factor analysis: A unified approach. Wiley: New York; 2011.
29. Dunn PK, Smyth GK. Randomized quantile residuals. Journal of Computational and Graphical Statistics. 1996; 5:236–244. https://doi.org/10.1080/10618600.1996.10474708
30. Wolfinger R. Laplace's approximation for nonlinear mixed models. Biometrika. 1993; 80:791–795. https://doi.org/10.1093/biomet/80.4.791
31. Wainwright MJ, Jordan MI. Graphical Models, Exponential Families, and Variational Inference. Foundations and Trends in Machine Learning. 2008; 1:1–305. https://doi.org/10.1561/2200000001
32. Bishop CM. Pattern recognition and machine learning. Springer; 2006.
33. Blei DM, Kucukelbir A, McAuliffe JD. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association. 2017; 112:859–877. https://doi.org/10.1080/01621459.2017.1285773
34. Ormerod JT, Wand MP. Gaussian variational approximate inference for generalized linear mixed models. Journal of Computational and Graphical Statistics. 2012; 21:2–17. https://doi.org/10.1198/jcgs.2011.09118
35. Westling T, McCormick TH. Beyond prediction: A framework for inference with variational approximations in mixture models. arXiv preprint arXiv:151008151v4. 2017;.
36. Daza Secco E, Haapalehto T, Haimi J, Meissner K, Tahvanainen T. Do testate amoebae communities recover in concordance with vegetation after restoration of drained peatlands? Mires and Peat. 2016; 18:1–14.
37. Cleary DFR, Genner MJ, Boyle TJB, Setyawati T, Angraeti CD, Menken SBJ. Associations of bird species richness and community composition with local and landscape-scale environmental factors in Borneo. Landscape Ecology. 2005; 20:989–1001. https://doi.org/10.1007/s10980-005-7754-y
38. Li J. Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained. Environmental Modelling & Software. 2016; 80:1–8. https://doi.org/10.1016/j.envsoft.2016.02.004.
39. Li J. Assessing the accuracy of predictive models for numerical data: Not r nor r2, why not? Then what? PLOS ONE. 2017; 12(8):1–16. https://doi.org/10.1371/journal.pone.0183250
40. Joe H. Accuracy of Laplace approximation for discrete response mixed models. Computational Statistics & Data Analysis. 2008; 52:5066–5074. https://doi.org/10.1016/j.csda.2008.05.002
41. Tierney L, Kadane JB. Accurate Approximations for Posterior Moments and Marginal Densities. Journal of the American Statistical Association. 1986; 81(393):82–86. https://doi.org/10.1080/01621459.1986.10478240
42. Brown AM, Warton DI, Andrew NR, Binns M, Cassis G, Gibb H. The fourth-corner solution—using predictive models to understand how species traits interact with the environment. Methods in Ecology and Evolution. 2014; 5:344–352. https://doi.org/10.1111/2041-210X.12163
43. Hui FKC, Tanaka E, Warton DI. Order selection and sparsity in latent variable models via the ordered factor LASSO. Biometrics. 2018;In press. https://doi.org/10.1111/biom.12888 PMID: 29750847

PIII

ANALYSING ENVIRONMENTAL-TRAIT INTERACTIONS IN ECOLOGICAL COMMUNITIES WITH FOURTH-CORNER LATENT VARIABLE MODELS

by

Niku, J., Hui, F. K., Taskinen, S. and Warton, D. I 2020

Submitted

Analysing Environmental-Trait Interactions in Ecological Communities with Fourth-Corner Latent Variable Models

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

[1]Department of Mathematics and Statistics, University of Jyväskylä, Finland
[2]Research School of Finance, Actuarial Studies & Statistics, Australian National University, Australia
[3]School of Mathematics and Statistics and Evolution & Ecology Research Centre, The University of New South Wales, Sydney, Australia

Abstract

In ecological community studies it is often of interest to study the effect of species related trait variables on abundances or presence-absences. Specifically, the interest may lay in the interactions between environmental and trait variables. An increasingly popular approach for studying such interactions is the so-called fourth-corner model, which explicitly posits a regression model where the mean response of each species is a function of interactions between covariate and trait predictors (among other terms). On the other hand, the array of fourth-corner models currently applied in the literature do not necessarily account for interspecific variation in the environmental response, nor for any residual covariation between species. To overcome this problem, in this article we propose a fourth-corner latent variable model which combines the following three features: latent variables to capture the correlation between species, fourth-corner terms to account for environment-trait interactions, and species-specific random slopes for modelling excess heterogeneity between species in their environmental response. Simulation studies demonstrate that the proposed method outperformed competitors when testing for the fourth-corner (interaction) coefficients, across Type I error and power simulations. The method is illustrated by an example on ground beetle data.

Keywords: Community analysis, fourth-corner problem, generalized linear mixed model, joint species distribution model, multivariate abundance data, variational approximation.

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

1 Introduction

One of the main aims of statistical analyses in community ecology is to understand how species differ in their responses to the environment, and why. Specifically, if trait information on each species is measured, it is possible to study how these traits mediate the effect of environmental conditions on species responses. In ecology, this problem of studying associations between environmental and trait variables using species abundance data is often known as the fourth-corner problem (Legendre et al., 1997). Specifically, given three matrices defining the environmental data ($\boldsymbol{R}$), species abundances ($\boldsymbol{L}$), and species traits ($\boldsymbol{Q}$), we can use these to infer how the environmental variables and species traits are jointly related to species abundance. Most classical approaches to solving the fourth-corner problem use a generalized singular value decomposition applied to a environment-trait association matrix constructed using $\boldsymbol{R}$, $\boldsymbol{L}$ and $\boldsymbol{Q}$, thus leading to a pair of ordinations for making interpretations of the associations (Dolédec et al., 1996). Legendre et al. (1997) further introduced a hypothesis testing approach based on permutation testing to assess which associations between environmental and trait variables are significant. Classical methods were further developed in Dray and Legendre (2008), ter Braak et al. (2012) and Dray et al. (2014). The Permutation tests are often used to make conclusions on which environmental and trait variables are associated with each other. However, no information on the strength of the interactions is obtained.

In the past decade, several model-based approaches have arisen in the literature for solving the fourth-corner problem, with a notable advantage being that they also give a concrete measure the effect size through the interpretation of relevant coefficients in the mean model. We now give an overview of these. Denote the abundances (counts, presence-absences, biomass) of m responses (species) recorded at n samples (sites) by y_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$. For each site i, a vector of k environmental variables, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ik})^\top$, and for each species j, a vector of q trait variables $\boldsymbol{t}_j = (t_{j1}, \ldots, t_{jq})^\top$ are also measured. The more general form of the fourth-corner model for the mean responses, μ_{ij}, can then be formulated as

$$g(\mu_{ij}) = r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + \text{vec}(\boldsymbol{B}_{et})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i), \tag{1}$$

where $g(\cdot)$ is a known link function, β_{0j} are species-specific intercepts, the k-vector $\boldsymbol{\beta}_e$ includes main effects for environmental covariates, and the $k \times q$ matrix $\boldsymbol{B}_{et}$ consists of environmental-trait interaction terms (also known as the fourth-corner coefficients). Also, r_i denotes random site effects which we include as means of row standardisation, while the k-vector $\boldsymbol{b}_j$ denotes species-specific random effects for environmental variables. The precise models considered so far in the literature differ in the way the random effects are included in model (1). For instance, in the generalized linear model (GLM) approach by Brown et al. (2014), random site effects

r_i and random slopes $\boldsymbol{b}_j$ were ignored, and Warton et al. (2015b) proposed inference on $\boldsymbol{B}_{et}$ based on bootstrapping the set of n vector residuals across the sites. Warton et al. (2015b) also showed that the method proposed in Brown et al. (2014) is a generalisation of a maximum entropy approach (community assembly via trait selection, CATS) proposed by Shipley et al. (2006). Easier interpretation, model selection and inference methods for CATS-regression are thus readily available. Pollock et al. (2012) proposed a generalized linear mixed modeling (GLMM) approach for solving the fourth-corner problem by including random site effects r_i and random slopes for environmental variables $\boldsymbol{e}_i$ in the model. The model was later extended by Jamil and ter Braak (2013) by treating the species-specific intercepts, β_{0j}, also as random. Most recently, ter Braak (2019) proposed to further include a random slope for trait variables $\boldsymbol{t}_j$.

A major drawback with all of the above model-based approaches listed above is that they do not model residual correlation between responses arising from direct biotic interactions between species, missing predictors, among a host of possible reasons. As mentioned above, Warton et al. (2015b) attempted to circumvent this issue by resampling sites. Pollock et al. (2012) and Jamil and ter Braak (2013) took into account the randomness at the individual species level, but did not account for residual correlation across species. In the context of testing environmental-trait interactions, the problem of ignoring residual interspecific variation to the environment (not explained by traits) was studied by ter Braak et al. (2017) in detail. Specificially, ter Braak et al. (2017) compared four different resampling strategies in the GLM framework and noted that resampling (bootstrapping or permuting) either sites or species tended to yield Type I error rates there were too small and an associated loss in power when testing for the fourth-corner coefficients. The p_{max} permutation test (ter Braak et al., 2012), where two separate resampling tests (site-level and species-level) are performed and the significance is assessed by the largest of the two p-values, was shown to perform best when the data were generated according to a simple GLMM model. However, the p_{max} test also produced inflated Type I errors when simulating from models where observed trait and environmental variables interact with latent trait and environmental variables. More recently, ter Braak (2019) compared different model-based testing approaches (likelihood-ratio test, parametric bootstrap test and permutation-based p_{max} test) for testing fourth-corner interaction terms, with tests based on a GLMM, where random slopes for trait variables $\boldsymbol{t}_j$ were also included, were shown to outperform other GLM and GLMM based tests for interaction.

An alternative approach to resampling-based procedures for testing the environmental-trait interactions is to use a generalised linear latent variable model (GLLVM) to explicitly model the between species correlation using a factor analytical approach, and then employ (say) standard asymptotic likelihood ratio tests. The past five years has seen an explosion in the use of latent variable models for community level modeling; see Warton et al. (2015a), Warton et al. (2016), Ovaskainen et al. (2017), Bjork et al. (2018) Niku et al. (2017b), among many others.

The fourth-corner latent variable model which we consider in this article, and which was first considered by Warton et al. (2015a), is based on extending the fourth-corner GLM of Brown et al. (2014) by including site-specific random row intercepts to account for the variation between sites, and species-specific random slopes for environmental variables for capturing the interspecific variation in responses not explained by the traits. In addition, latent variables with corresponding loadings are included to capture any residual correlation between species which is not explained by observed environmental and trait variables. While similar models have also been developed in the Bayesian context by Hui (2016) and Tikhonov et al. (2019), the performances of such models for assessing environment-trait interactions (in terms of producing valid inference) have not been studied before, let alone compared to existing procedures (including those reviewed above) in the literature.

In this paper, we propose a fast and efficient maximum likelihood based estimation algorithm for the fourth-corner latent variable model, and apply it to the environment-trait interaction testing problem. Specifically, when testing the fourth-corner coefficients, we employ a simple likelihood ratio testing approach. Importantly, this is made possible by including the necessary terms in the mean structure to ensure that all relevant sources of heterogeneity and residual correlation are accounted, thereby leading to valid statistical inference. The performance of the proposed interaction test is compared with tests based on GLMMs (Pollock et al., 2012; Jamil and ter Braak, 2013; ter Braak et al., 2017) through simulation, as we investigate Type I error rates under the null hypothesis and power under varying alternative hypotheses.

The paper is organized as follows, in Section 2 we define our fourth-corner latent variable model and discuss the associated estimation and inferential procedures based on fast variational approximations (Hui et al., 2017; Niku et al., 2019a). In Section 3, we perform simulation studies for comparing Type I errors and powers of different tests for the interaction term. Finally, in Section 4 we illustrate our method by applying it to ground beetle data (Ribera et al., 2001).

2 Model definition and estimation

Using the notation previously introduced, the

fourth-corner latent variable model with random site effects (intercepts) and random slopes is defined by the following mean regression model,

$$g(\mu_{ij}) = \eta_{ij} = r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i) + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \tag{2}$$

or equivalently formulated in a hierarchical fashion,

$$g(\mu_{ij}) = \eta_{ij} = r_i + \beta_{0j} + \boldsymbol{e}_i'\boldsymbol{\beta}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \text{ where } (r_i, \boldsymbol{u}_i')' \sim N_{d+1}(\boldsymbol{0}, \boldsymbol{C}_\sigma) \quad (3)$$
$$\boldsymbol{\beta}_j = \boldsymbol{\beta}_e + \boldsymbol{B}_{te}\boldsymbol{t}_j + \boldsymbol{b}_j, \text{ where } \boldsymbol{b}_j \sim N_k(\boldsymbol{0}, \boldsymbol{G}).$$

As in model (1), we let β_{0j} denote the species-specific intercepts, k-vector $\boldsymbol{\beta}_e$ denote the main effects for the environmental covariates, and $k \times q$ matrix $\boldsymbol{B}_{te}$ denote the environmental-trait interaction matrix on which testing will be performed. The random site intercepts, r_i, are assumed to follow a normal distribution with zero mean and variance σ^2, $r_i \sim N(0, \sigma^2)$. Notice that if the site effects are treated as fixed, then the main effects for environmental covariates, $\boldsymbol{\beta}_e$, can be omitted. The vector $\boldsymbol{b}_j$ includes k species-specific random effects for environmental variables, which are assumed to follow a multivariate normal distribution with zero mean vector and unstructured $k \times k$ covariance matrix $\boldsymbol{G}$, $\boldsymbol{b}_j \sim N_k(\boldsymbol{0}, \boldsymbol{G})$. If random slope parameters are included in the model, then the effect of predictors is a combination of the fixed effects, $\boldsymbol{\beta}_e$, which are common to all species, the interaction terms with species traits, $\boldsymbol{B}_{te}$, which define how traits mediate the effect of environmental variables, and the random effects for species, $\boldsymbol{b}_j$, which capture the interspecific variation not explained by traits. Finally, the d-vector $\boldsymbol{\gamma}_j$ includes species-specific factor loadings for d-variate ($d \ll m$) latent variables, $\boldsymbol{u}_i$, which are assumed to follow a multivariate standard normal distribution, $\boldsymbol{u}_i \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$, where $\boldsymbol{I}_d$ denotes a $d \times d$ identity matrix. The zero mean and unit variance fix the locations and scales of latent variables and ensure parameter identifiability (Huber et al., 2004). In turn, the term $\boldsymbol{u}_i'\boldsymbol{\gamma}_j$ captures the residual correlation between species not accounted for by the observed covariates $\boldsymbol{e}_i$ and trait variables $\boldsymbol{t}_j$. Covariance matrix $Cov((r_i, \boldsymbol{u}_i')') = \boldsymbol{C}_\sigma$ is formed so that we include the correlation term between site effects and latent variables, $corr(r_i, u_{il}) = \rho_l$. We denote the matrix of loadings $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1 \cdots \boldsymbol{\gamma}_m)'$, and set all the upper triangular elements of $m \times d$ matrix $\boldsymbol{\Gamma}$ to be zero and constrain its diagonal elements to be positive in order to avoid rotation invariance and (again) ensure parameter identifiability (Huber et al., 2004). Note that this constraint on the loading matrix does not reduce the flexibility of the model; indeed, the residual between species covariance matrix (given the environmental and trait predictors) is straightforwardly seen to be $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}'$, from which we see that the residual covariance is modelled parsimoniously via rank-reduction.

Model (3) serves as a unifying framework that encompasses models proposed previously in Pollock et al. (2012), Jamil and ter Braak (2013) and Brown et al. (2014). If we set all variances of random effects, r_i and $\boldsymbol{b}_j$, and latent variables $\boldsymbol{u}_i$ in model (3) to zero, the model reduces to the fourth-corner GLM of Brown et al. (2014). If we set the covariance matrix of random row effects and latent variables, $\boldsymbol{C}_\sigma$, to zero, we get a similar model as in Pollock et al. (2012), with an exception that Pollock et al. (2012) treated species-specific intercepts, β_{0j}, as random. Jamil

and ter Braak (2013) extended the model proposed in Pollock et al. (2012) by adding random site effects, r_i, in the model.

Let $\boldsymbol{\Psi} = \{\boldsymbol{\beta}_0', \boldsymbol{\beta}_e', vec(\boldsymbol{B}_{te})', vec(\boldsymbol{\Gamma})', \boldsymbol{\Phi}', vec(\boldsymbol{C}_\sigma), vec(\boldsymbol{G})\}$ denote the full vector of parameters in the fourth-corner latent variable model, where $\boldsymbol{\beta}_0 = \{\beta_{01}, \ldots, \beta_{0m}\}'$ is the vector of all species-specific intercepts, $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_m)'$ includes all other nuisance parameters, e.g., dispersion parameters of the negative binomial or the Tweedie distribution as in Niku et al. (2017b). Furthermore, we denote $\boldsymbol{r} = (r_1, \ldots, r_n)'$, $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_m')'$ and $\boldsymbol{u} = (\boldsymbol{u}_1', \ldots, \boldsymbol{u}_n')'$ as the full vector of site intercepts, species-specific random effects, and latent variables, respectively. Conditional on the latent variables and parameter vector $\boldsymbol{\Psi}$, the responses are assumed to be independently distributed and we obtain the joint distribution $f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u}; \boldsymbol{\Psi}) = \prod_{i=1}^{n}\prod_{j=1}^{m} f(y_{ij}|r_i, \boldsymbol{b}_j, \boldsymbol{u}_i; \boldsymbol{\Psi})$. By integrating over random effects $\boldsymbol{r}$ and $\boldsymbol{b}$ and latent variables $\boldsymbol{u}$ then, we obtain the following marginal log-likelihood function for the fourth-corner latent variable model,

$$l(\boldsymbol{\Psi}) = \log\left\{\int f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u}; \boldsymbol{\Psi}) f(\boldsymbol{r}, \boldsymbol{u}; \boldsymbol{C}_\sigma) f(\boldsymbol{b}; \boldsymbol{G}) d(\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u})\right\}. \tag{4}$$

For multivariate abundance data, the response distribution $f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u}; \boldsymbol{\Psi})$ is not assumed to be a multivariate normal distribution (since the responses are usually discrete with a strong non-constant mean-variance relationship). Consequently, the integration over latent variables and random effects does not have a closed form. To overcome this issue then, a common and computationally efficient approach is to approximate the integral using approaches such as the Laplace (Niku et al., 2017b) or variational (Hui et al., 2017) approximation, which subsequently provide either a closed or nearly closed form approximation to the marginal log-likelihood (4). In Niku et al. (2019a) it was shown that computationally convenient estimation algorithms for GLLVMs can be obtained by combining the Laplace or variational approximation methods with automatic optimization techniques implemented in R software, for computationally efficient estimation.

In this paper, we adopt the variational approximation approach for approximating the marginal log-likelihood in (4). As part of using the variational approximation method, we need to define so-called variational distributions for the random effects $\boldsymbol{a}$ and $\boldsymbol{b}$, and the latent variables $\boldsymbol{u}$, which effectively act as the approximate posterior distributions for these latent quantities. For ease of computation, while also being a sensible choice in an asymptotic sense (Hui et al., 2017; Blei et al., 2017), we propose to use independent normal distributions. Specifically, for $i = 1, \ldots, n$, we set the variational density of the random effects $q(r_i, \boldsymbol{u}_i)$ as independent multivariate normal distributions $N_{d+1}(\boldsymbol{a}_i, \boldsymbol{A}_i)$, while for response $j = 1, \ldots, m$ we set the variational density of the random effects $q(\boldsymbol{b}_j)$ as independent multivariate normal distributions $N_k(\boldsymbol{a}_{bj}, \boldsymbol{A}_{bj})$. Here, $\boldsymbol{a}_i$ and $\boldsymbol{a}_{nj}$ denote mean vectors of length $(d+1)$ and k respectively, while $\boldsymbol{A}_{bj}$ and $\boldsymbol{A}_i$ are assumed to be positive definite and unstructured covariance matrices of dimension $(d+1) \times (d+1)$

and $k \times k$, respectively. Following these assumptions, and assuming that y_{ij} comes from the exponential family of distributions with mean $\mu_{ij} = E(y_{ij})$, such that $f(y_{ij}|r_i, \boldsymbol{b}_j, \boldsymbol{u}_i; \boldsymbol{\Psi}) = \exp\{(y_{ij}\eta_{ij} + b(\eta_{ij}))/\phi_j + c(y_{ij}, \phi_j)\}$, where $b(\cdot)$ and $c(\cdot)$ are known functions, then the resulting variational log-likelihood function is given by

$$\begin{aligned}\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = &\sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_q\{b(\eta_{ij})\}}{\phi_j} + c(y_{ij}, \phi_j)\right\} \\ &+ \frac{1}{2}\sum_{i=1}^{n}\left\{\log\det(\boldsymbol{A}_i) - \text{tr}(\boldsymbol{C}_\sigma^{-1}\boldsymbol{A}_i) - \boldsymbol{a}_i'\boldsymbol{C}_\sigma^{-1}\boldsymbol{a}_i - \log\det(\boldsymbol{C}_\sigma)\right\} \\ &+ \frac{1}{2}\sum_{j=1}^{m}\left\{\log\det(\boldsymbol{A}_{bj}) - \text{tr}(\boldsymbol{G}^{-1}\boldsymbol{A}_{bj}) - \boldsymbol{a}_{bj}'\boldsymbol{G}^{-1}\boldsymbol{a}_{bj} - \log\det(\boldsymbol{G})\right\},\end{aligned}$$

where $\tilde{\eta}_{ij} = \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + a_{bj}) + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i) + \boldsymbol{a}_i'(1, \boldsymbol{\gamma}_j')'$, and all quantities constant with respect to the parameters have been omitted. Notice that above $E_q(\cdot)$ denotes the expectation with respect to $q(\boldsymbol{b})q(\boldsymbol{r}, \boldsymbol{u})$, which does not necessarily have a closed form. In Hui et al. (2017) it was shown that by reparametrizing GLLVMs, fully closed form variational log-likelihoods can be obtained in case of binary, ordinal and count data. A proof for the above formula is provided in Appendix B.

By treating the variational log-likelihood function as a new objective function, we can then fit and perform inference on the fourth-corner latent variable model. For instance, maximization of $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ with respective to both model $\boldsymbol{\Psi}$ and variational $\boldsymbol{\xi}$ parameters produces relevant estimates, with the latter acting also as predictions for the latent variables and random effects. Specifically, the variational distributions $q(r_i, \boldsymbol{u}_i)$ and $q(\boldsymbol{b}_j)$ serve as approximate posterior distributions for all latent quantities, which can be used for ordination. The asymptotic standard errors for model parameters can be computed using the the observed information matrix (negative Hessian) as described in Hui et al. (2017). This allows us to construct confidence intervals as well as to conduct Wald tests for the model parameters. Likelihood ratio tests are also readily available and will be applied in the next section for testing the fourth-corner interaction terms. All the inferential methods listed above are implemented in the `R` package `gllvm` (Niku et al., 2017a). The package uses Template Model Builder (`TMB`, Kristensen et al., 2016) for automatic differentation of the log-likelihood function to enable efficient parameter estimation. For further details of the implementation, we refer to Niku et al. (2019b).

3 Simulation studies

Three simulation studies were conducted to evaluate the ability of the proposed fourth-corner latent variable model to account for unobserved random variation in multivariate count data and to compare the method to the p_{max} test of ter Braak et al. (2017). In the first simulation setup, we study the Type I errors of the likelihood ratio test for testing the null hypothesis H_0 : $\boldsymbol{B}_{te} = \mathbf{0}$ based on the fourth-corner latent variable model in (3), for a situation where interspecific variation and correlation between species is inherent in data. In the second setting, we examined the power of the proposed test, that is, the empirical probability of finding the significant interaction between environmental and trait variables, under varying alternative hypotheses. For comparison, we consider four variants of model (3), consisting of two GLMM type models and two GLLVM models with $d = 1$ and $d = 2$ latent variables. These are denoted as follows

$$
\begin{aligned}
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'\boldsymbol{\beta}_e + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i), && \text{glmm(r)} \\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i), && \text{glmm(r + e)} \\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'\boldsymbol{\beta}_e + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i) + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, && \text{gllvm(d lv + r)} \\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i) + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, && \text{gllvm(d lv + r + e)}
\end{aligned}
$$

Here we assume that d is known, but in practice, model selection tools, such as AIC and BIC, can be used to guide the selection. Data will be simulated under $d = 2$, so results for $d = 1$ give some insight into the performance of GLLVM when the covariance structure has been misspecified. As mentioned above, the likelihood ratio tests based on above models were also compared to the p_{max} test (ter Braak et al., 2017). This approach was applied to log-likelihood ratio tests from model glmm(r) fitted under a Poisson GLM (the Poisson being used for computational efficiency ter Braak et al., 2017), and involves taking the largest of the two P-values formed by permuting either rows or columns of predictors. Additionally, we include in some of our comparisons a variant of the p_{max} permutation test considered in ter Braak (2019), where it is applied to generalized linear mixed models with random site effects and random slopes.

In the first simulation setup, we compared the Type I errors based on the likelihood ratio tests to those of the p_{max} test. We generated datasets according to the negative binomial distribution using two sample sizes and dimensions: (a) $m = 40$ and $n = 70$, and (b) $m = 70$ and $n = 40$. As a mean model we used model (3) with one trait and one environmental variable but with species intercepts written as $\beta_{0j} + t_j\beta_t$, with β_{0j} generated independently from the uniform distribution $U(-1, 1)$, and $\beta_t = 0.3$. The value for the variance of the random row effects was $\sigma^2 = 0.3$ and also $\beta_e = 0.3$. The fourth-corner coefficient B_{te} was set to zero in order to assess Type I error. The species-specific dispersion parameters were all set to $\phi_j = 0.5$.

In order to create unobserved correlation structure between species, we generated two-dimensional

latent variables, $\boldsymbol{u}_i = (u_{i1}, u_{i2})'$, from the bivariate standard normal distribution, and simulated the values of the associated loadings $\boldsymbol{\gamma}_j$ independently from the standard normal distribution. Finally, we generated the random slopes b_j from a normal distribution with mean zero and variance from the range $\sigma_b^2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. That is, variances of random slopes b_j that causes interspecific variation not explained by the covariates, were increased from 0 to 1. To recap, the latent variables can also be interpreted to include latent environmental covariates and their loadings as effects of latent traits on latent environmental variables, while the random slopes generate unexplained random variation on species that is not explained by the observed traits, and can therefore be interpreted as latent traits. The latent variables and their loadings, random effects and covariates were re-generated for each simulated datasets.

Given the above set up, we simulated 1000 datasets from the negative binomial distribution. For each dataset, we then calculated p-values based on likelihood ratio tests from the four models listed above, as well as the p_{max} test applied to both GLMs and GLMMs, for assessing the null hypothesis $H_0 : B_{te} = 0$. The resulting Type I errors are presented in Figure 1. Results indicate that the fourth-corner latent variable models, gllvm($d\,lv + r + e$) with $d = 1$ and $d = 2$, provided empirical Type I errors that were reasonably close to the nominal significance level of 5% for all values of σ_b^2. In contrast, the p_{max} test applied to GLMs tended to be too conservative and produced Type I errors below the nominal level especially for small values of σ_b^2. In fact, the p_{max} test applied to GLMMs produced Type I errors of exactly zero for all values of σ_b^2 tests and was therefore excluded from Figure 1. The Type I errors for likelihood ratio tests based on models that do not include species-specific random slopes were severely inflated especially for large values of σ_b^2, while the likelihood ratio test based on the mixed model with random intercept and slope, gllmm($r + e$), performed similarly to the tests based on latent variable models.

In the second simulation setup, we introduced correlation between the residual correlation term and the observed trait by setting $corr(\gamma_{j2}, t_j) = 0.5$, where γ_{j2} are the latent variable loadings for the latent variables u_{i2}. In practice, loadings γ_{j2} and traits t_j were generated from a bivariate normal distribution with unit variances and 0.5 correlation. This can be interpreted as a situation in which the effect of the observed trait differs between sites and is not fully explained by the observed environmental variables. The methods that provided inflated Type I errors in the previous setting, namely GLLVM and GLMM with random intercepts, were excluded from the comparison. Type I errors presented in Figure 2 show that the fourth-corner latent variable model with $d = 2$ typically maintains close to nominal Type I error, although rising to almost 0.1 in one case. The p_{max} test applied to GLMs provided Type I errors close to the nominal level as well, while it was overly conservative in the first setting. The p_{max} test applied to GLMMs performed similarly to the first simulation, with zero Type I error in most cases. The likelihood ratio test based on the fourth-corner latent variable model with $d = 1$ and the mixed model

with random intercept and slope both produced inflated type I errors for small values of σ_b^2. These results suggest that the inclusion of latent variables is necessary in order to capture the additional source of residual between species correlation. In Appendix A.1 and A.2 we present some additional simulation studies which show that the model-based test based on the GLMM with random site effects and slopes performed worse than the test based on fourth-corner latent variable model. Overall, these simulation results demonstrate the robust performance of the fourth-corner latent variable model due to its capability to capture all the relevant sources of covariation.

In the third simulation setup, we compared the power of the various testing procedures. The methods that provided inflated Type I errors in the first simulation study were excluded from the comparison, meaning only four methods were included for comparison. We again generated 1000 datasets using the similar setup as in the first simulation study with $n = 70$ and $m = 40$, but varied the interaction term B_{te} such that $B_{te} \in \{0, 0.1, 0.2, 0.3, 0.4\}$. As variances for random slope effects, we considered $\sigma_b^2 \in \{0, 0.4, 0.8\}$. The power simulation for the setup with $n = 40$ and $m = 70$ was excluded as results were similar compared to the previous one. The resulting empirical powers of the p_{max} test and three different likelihood ratio tests are plotted in Figure 3. In all cases, the likelihood ratio tests based on the fourth-corner latent variable models provide higher probabilities for detecting the significant interaction between environmental and trait variables as compared to the p_{max} test. This was not surprising given the p_{max} test is, by construction, conservative since it involves performing two permutation tests and then choosing the less conservative of the two. Indeed, this conservatism is reflected in the Type I error results seen in Figure 1. The likelihood ratio test applied to gllmm(r+e) performed well but was slightly less powerful than the tests based on fourth-corner latent variable models when the value for σ_b^2 was small.

4 Case study

We applied the proposed fourth-corner latent variable model to a dataset consisting of counts of $m = 68$ ground beetle species recorded at $n = 87$ sites across Scotland (Ribera et al., 2001). The original data also included 17 environmental variables recorded at each site and 20 trait variables for each species. Ribera et al. (2001) studied whether the morphology and life traits of ground beetle species can be related to the environmental variability of the habitats. For illustration purposes, we consider using a subset of $k = 4$ environmental variables: land use management intensity score (Management), percentage moisture content (Moist), elevation, and soil pH value, along with four species trait covariates: total length (LTL) and pronotum height (LPH), overwintering (OVE, with two levels: 1 = only adults; 2 = adults and larvae or only larvae), and breeding season (BRE, with three levels: 1 = spring; 2 =, summer; 3 =,

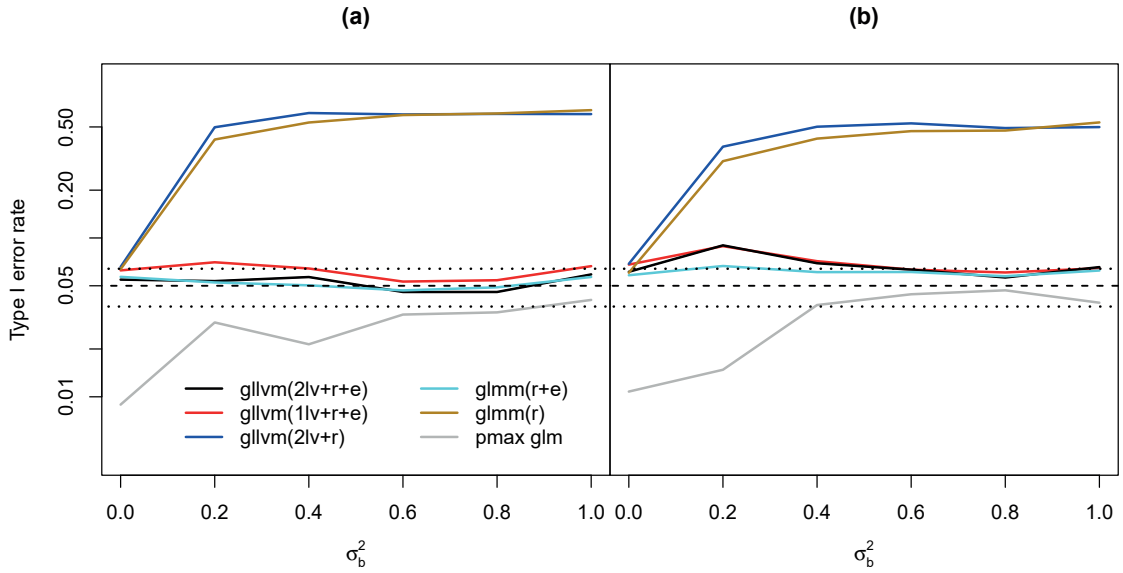

Figure 1: Type I error rates for likelihood ratio tests based on GLMM with random intercepts (glmm(r)), GLMM with random intercepts and slopes (glmm($r + e$)), GLLVM with random intercepts (gllvm($2lv + r$)), GLLVM with random intercepts and slopes and d latent variables (gllvm($d\,lv + r + e$))), and the p_{max} test applied to GLMs. The p_{max} test applied to GLMMs produced Type I errors of exactly zero for all values of σ_b^2 and was therefore excluded from the plot. On the left, data size is (a) $n = 70$ sites and $m = 40$ species, on the right (b) $n = 40$ sites and $m = 70$ species. The variance of the random slope effects, σ_b^2 is plotted on x-axis. Dashed line is the nominal level 0.05 and dotted lines around nominal level correspond values for sample proportions which are not significantly different from 0.05.

autumn or winter). This set of environmental and trait variables were among the most important covariates affecting the ground beetle communities based on the analysis of Ribera et al. (2001). All quantitative covariates were centered and scaled to have variance one before the analysis, while dummy variables were set up for OVE and BRE, meaning there were a total of $q = 5$ predictors in the vector of traits $\boldsymbol{t}_j$.

We first tested if the interactions between environmental and trait covariates were significant using likelihood ratio tests based on the fourth-corner latent variable model with one and two latent variables, GLMMs with random row and slope parameters included, and the p_{max} test. Table 1 lists the AIC values for different models as well as p-values given by three likelihood ratio tests. The GLLVM with random row effects and random slopes and two latent variables had the lowest value of AIC, suggesting that both latent variables and species-specific random effects were needed to model additional sources of (co)variation, while the p-values for all three tests were less than 0.001 providing clear evidence of an interactions between the considered environmental and trait variables. By contrast, the p_{max} test with 999 permutations gives a p-value of 0.143 when testing for the fourth-corner interaction term. The result is thus consistent with the simulation study results showing the conservativeness of the p_{max} test.

The estimated coefficients for the environmental covariates and interaction terms based on the

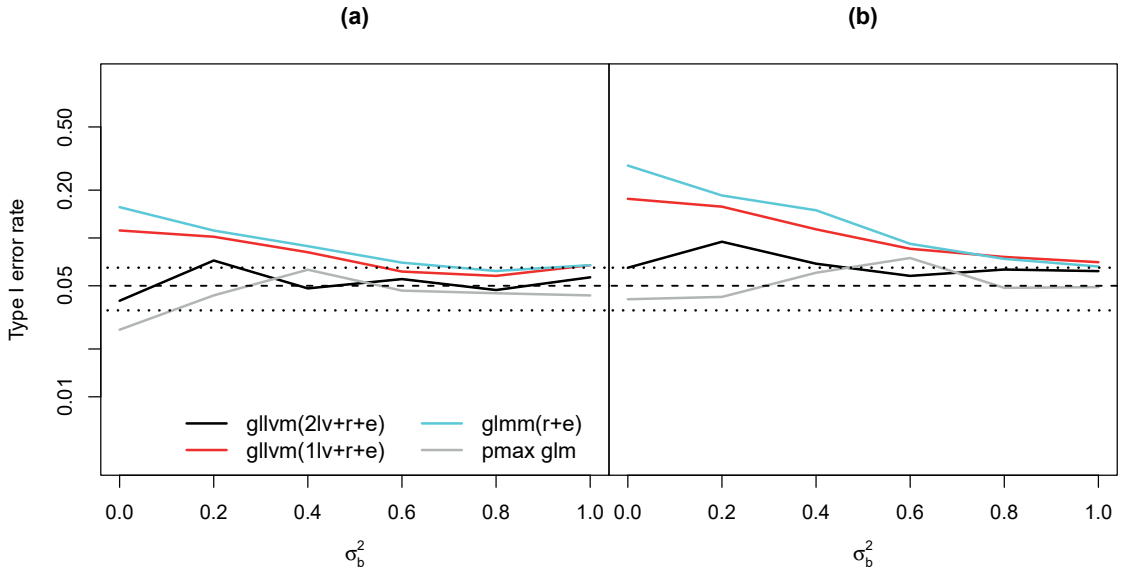

Figure 2: Type I error rates for likelihood ratio tests based on GLMM with random intercepts and slopes (glmm($r+e$)), GLLVM with random intercepts, slopes and d latent variables (gllvm($d\,lv+r+e$))), and the p_{max} test applied to GLMs. The p_{max} test applied to GLMMs produced Type I errors of exactly zero for almost all values of σ_b^2, except $\sigma_b^2 = 0$, and was therefore excluded from the plot. On the left, data size is (a) $n = 70$ sites and $m = 40$ species, on the right (b) $n = 40$ sites and $m = 70$ species. The variance of the random slope effects, σ_b^2, is plotted on the x-axis. Dashed line is the nominal level 0.05 and dotted lines around nominal level correspond to values for sample proportions which are not significantly different from 0.05.

Table 1: The values of AIC for the two fourth-corner latent variable models, and the GLMM model fitted to the ground beetle dataset. Also shown are the p-values for the corresponding likelihood ratio test of the fourth-corner interaction terms.

	glmm(r+e)	gllvm(1lv+r+e)	gllvm(2lv+r+e)
AIC	18706	18496	**18424**
p value	<0.001	<0.001	<0.001

GLLVM with two latent variables are plotted in Figure 4. The strongest negative interactions were between management intensity and breeding season, as well as between management intensity and pronetum height. In other words, high management intensity was found to have a large negative effect on species that breed during the summer and have larger body size. The strongest positive effects occurred in interactions between elevation and breeding season and between management intensity and total length. That is,, species having breeding season in summer succeeded better in high altitude environments as compared to species which breed during other seasons. Finally, predictions for species-specific random slopes for the environmental covariates and their associated 95% uncertainty intervals are plotted in Figure 5; recall from Section 2 these were generated based on the normal variational distributions and the estimated values of $\boldsymbol{a}_{bj}$ and $\boldsymbol{A}_{bj}$. From this, we can see that the interspecific variation in responses, which is not explained by the traits, is highest for the effect of the moisture content and management

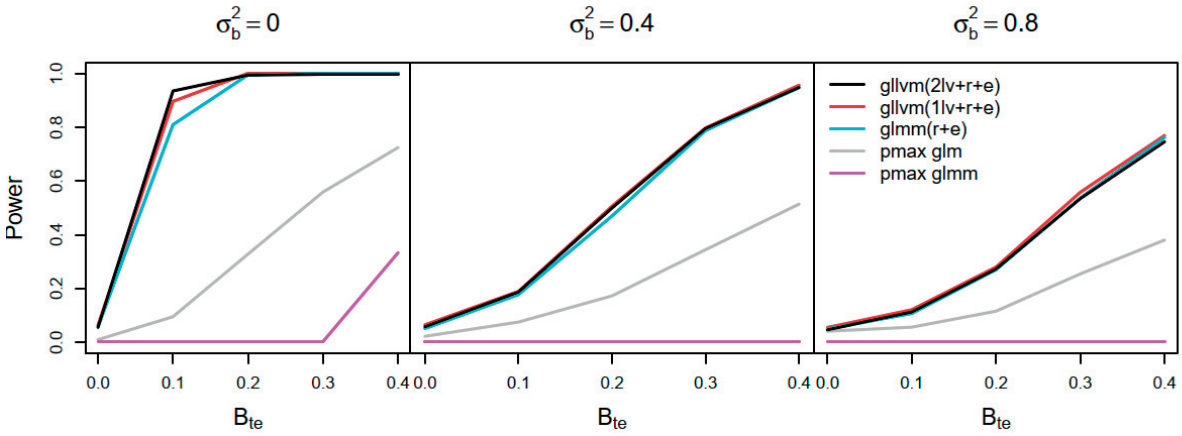

Figure 3: Power as a function of the effect size B_{te} for the likelihood ratio tests based on GLMM with random intercepts and slopes (glmm$(r + e)$), GLLVM with random intercepts, slopes and d latent variables (gllvm$(d\,lv + r + e)$), and the p_{max} test applied to GLMs and GLMMs.

intensity and lowest for the effect of the elevation.

5 Discussion

In this article, we have proposed a fourth-corner latent variable model that accounts for two key sources of error in current implementations of fourth-corner model, namely the failure of traits to capture all interspecific variation (species-specific error), and the failure to account for the residual correlation between species (site-specific error) not explained by the environmental and trait variables. With a model-based approach, we are able to account for both sources of additional variation through the inclusion of additional species-specific random slopes, and site-specific latent variables. The approach is shown to be an extension of the recently introduced model-based approaches in Pollock et al. (2012), Jamil and ter Braak (2013) and Brown et al. (2014). We adopted an efficient estimation and inference approach base on variational approximations, and compared its finite sample performance to classical competitors for assessing the importance of fourth-corner interaction terms. Results showed that the proposed approach maintains close to nominal Type I error levels when testing for the fourth-corner coefficients, while power can be substantially better than other resampling-based procedures. Moreover, not accounting for the additional species-specific variation not explained by traits led to inflated Type I errors. We also found that even when the correlation model was not correct, that is, a fourth-corner GLLVM with one latent variable was used when two were needed, the approach continued to perform reasonably well, and tended to do better than alternatives. In Appendix A.1, we present one setting where the fourth-corner latent variable model (along with all the other approaches) can fail: namely when there are missing predictors which are correlated with the observed ones. Such methods can fail here because it leads to confounding, thus biased esti-

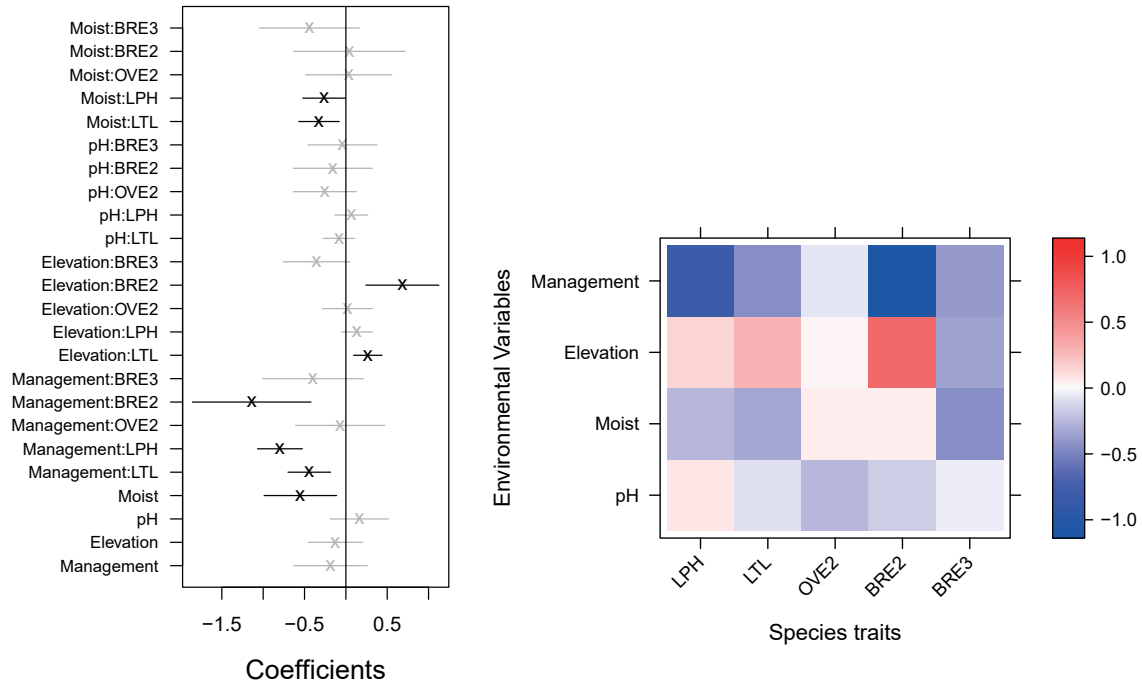

Figure 4: Point estimates and associated 95% confidence intervals for coefficients (left), along with a level plot (right) for fourth-corner interaction terms from a fourth-corner latent variable model with two latent variables fitted to the ground beetle data. The confidence intervals that do not contain zero are in black while those that do contain zero are in grey and faded.

mation and uncertainty quantification for the associated regression coefficients (see for instance Paciorek, 2010, on the related issue of confounding).

Model (1) included a random effect to capture species-specific variation in environmental response, not captured by traits. Because species tend to respond to the environment in complex and sophisticated ways, and because our data collection process rarely captures all these reasons, it seems a sensible working assumption to always expect such species-specific variation. Simulations in ter Braak et al. (2017), and those in this paper, emphasise the importance of including such a term. This paper additionally shows that it is important to capture residual correlation in abundance across species, which can be achieved using latent variables. In future research, we will examine other data-driven approaches to selecting the number of latent variables (Hui et al., 2018), as well as extensions to other incorporate other sources of variation such as spatio-temporal correlations (e.g., adapting the work of Thorson et al., 2016), and imperfect detection (Warton et al., 2016; Tobler et al., 2019).

References

Bjork, J. R., Hui, F. K., 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.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112:859–877.

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.

Dolédec, S., Chessel, D., ter Braak, C. J. F., and Champely, S. (1996). Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, 3(2):143–166.

Dray, S., Choler, P., Dolédec, S., Peres-Neto, P. R., Thuiller, W., Pavoine, S., and ter Braak, C. J. F. (2014). Combining the fourth-corner and the rlq methods for assessing trait responses to environmental variation. *Ecology*, 95(1):14–21.

Dray, S. and Legendre, P. (2008). Testing the species traits - environment relationships: The fourth - corner problem revisited. *Ecology*, 89(12):3400–3412.

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

Hui, F. K., Tanaka, E., and Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor lasso. *Biometrics*, 74:1311–1319.

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., 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(1):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(5):1–21.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L. (1997). Relating behavior to habitat: Solutions to thefourth-corner problem. *Ecology*, 78(2):547–562.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2017a). *gllvm: Generalized Linear Latent Variable Models*. R package version 1.2.

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019a). Efficient estimation of generalized linear latent variable models. *PLOS ONE*, 14(5):e0216129.

Niku, J., Hui, F. K. C., Taskinen, S., and Warton, D. I. (2019b). gllvm – fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10:2173–2182.

Niku, J., Warton, D. I., Hui, F. K. C., and Taskinen, S. (2017b). Generalized linear latent variable models for multivariate count and biomass data in ecology. *Journal of Agricultural, Biological, and Environmental Statistics*, 22(4):498–522.

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(5):561–576.

Paciorek, C. J. (2010). The importance of scale for spatial-confounding bias and precision of spatial regression estimators. *Statistical Science*, 25:107–125.

Pollock, L. J., Morris, W. K., and Vesk, P. A. (2012). The role of functional traits in species distributions revealed through a hierarchical model. *Ecography*, 35(8):716–725.

Ribera, I., Dolédec, S., Downie, I. S., and Foster, G. N. (2001). Effect of land disturbance and stress on species traits of ground beetle assemblages. *Ecology*, 82(4):1112–1129.

Shipley, B., Vile, D., and Garnier, r. (2006). From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science*, 314(5800):812–814.

ter Braak, C. J., Peres-Neto, P., and Dray, S. (2017). A critical issue in model-based inference for studying trait-based community assembly and a solution. *PeerJ*, 5:e2885.

ter Braak, C. J. F. (2019). New robust weighted averaging- and model-based methods for assessing trait - environment relationships. *Methods in Ecology and Evolution*, 10:1962–1971.

ter Braak, C. J. F., Cormont, A., and Dray, S. (2012). Improved testing of species traits - environment relationships in the fourth - corner problem. *Ecology*, 93(7):1525–1526.

Thorson, J. T., Ianelli, J. N., Larsen, E. A., Ries, L., Scheuerell, M. D., Szuwalski, C., and Zipkin, E. F. (2016). Joint dynamic species distribution models: a tool for community ordination and spatio-temporal monitoring. *Global Ecology and Biogeography*, 25:1144–1158.

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

Tobler, M. W., Kéry, M., Hui, F. K., Guillera-Arroita, G., Knaus, P., and Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, 100(8):e02754.

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(10):737–738.

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

Warton, D. I., Shipley, B., and Hastie, T. (2015b). Cats regression – a model-based approach to studying trait-based community assembly. *Methods in Ecology and Evolution*, 6(4):389–398.

A Additional simulations

A.1 Additional simulations with random parameters

We compared the methods used in Section 3 by mimicking the simulation setup of ter Braak et al. (2017), and the methods used in Section 3 were included in the comparisons. We generated 1000 datasets with 40 species and 40 sites from the negative binomial distribution with mean model

$$g(\mu_{ij}) = \mu_0 + R_i + C_j + b_{te}t_j e_i + b_{ze}z_j e_i + b_{tx}t_j x_i + b^*_{zx}z^*_j x^*_i + \epsilon_{ij}, \tag{5}$$

and variance $V(\mu_{ij}) = \mu_{ij} + \mu^2_{ij}$. Here intercept equals $\mu_0 = log(30)$.Row effects were generated as $R_i = a_0 e_i + a_1 e_i^2 + \epsilon_{ri}$, with $\epsilon_{ri} \sim N(0, 0.01)$, and column effects similarly by $C_j = c_0 t_j + c_1 t_j^2 + \epsilon_{tj}$, with $\epsilon_{tj} \sim N(0, 0.01)$. Observed environmental variable e_i and trait t_j were generated from standard normal distribution $N(0, 1)$. Independent latent environmental variables x_i and x^*_i and traits z_j and z^*_j were also generated from $N(0, 1)$. Parameters b_{te}, b_{ze}, b_{tx} and b^*_{zx} are effects for associations. Term $b^*_{zx}z^*_j x^*_i$ represents here the correlation structure among species and sites and can be interpreted similarly to the latent variable term $\boldsymbol{u}'_i\boldsymbol{\gamma}_j$ in fourth corner latent variable model, with is only one latent variable. Error terms ϵ_{ij} were generated from normal distribution, $\epsilon_{ij} \sim N(0, 0.2)$. We test the null hypothesis $H_0 : b_{te} = 0$ and calculate Type I error rates for random trait case, where $b_{te} = 0$, $b_{ze} \in \{0, 0.2, 0.4, 0.6, 0.8\}$, $b_{tx} = 0$, and random trait and random environmental variable case, where $b_{te} = 0$, $b_{ze} = b_{tx} \in \{0, 0.2, 0.4, 0.6, 0.8\}$. We set $a_0 = 0.05$, $a_1 = -0.1$, $c_0 = 0.05$, $c_1 = -0.1$ and $b^*_{zx} = 0.2$.

Based on the results in Figure 6(a) and 6(b), the likelihood ratio test based on the GLLVMs with one and two latent variables, random slopes and random row effects provided Type I errors

close the nominal level 0.05 in all considered cases excluding the case $b_{ze} = b_{tx} = 0.2$ where the Type I errors exceeded significantly the nominal level 0.05. Such peak is seen in all conducted simulation setups with a small sample size. The likelihood ratio test based on the GLMM with random slopes and random row effects produced close to valid Type I errors for the random trait case (Figure 6(a)) but inflated Type I errors for the random trait and random env case (Figure 6(b)). The p_{max} test applied for GLM worked quite well for the random trait case, but produced slightly inflated Type I errors for the random trait and random environmental variable case when effect sizes for b_{ze} and b_{tx} were larger than 0.4. The likelihood ratio tests based on GLLVM and GLMM which did not include random slopes producedtoo large Type I errors.

In Figure 7 the Type I errors were calculated using the same mean model as above, except observed environmental variables e_i and latent environmental variables x_i as well as observed traits t_j and latent traits z_j were generated so that they were correlated, that is, $corr(e_i, x_i) = 0.3$ and $corr(t_j, z_j) = 0.3$. Such correlations lead to a confounding effect and the results show that if this is the case all methods produced too inflated Type I errors.

A.2 Simulations for the ground beetle data

Simulations based on subsets of the ground beetle data considered in Section 4 were conducted by using the mean model

$$\log(\mu_{ij}) = r_i + \beta_{0j} + t_j\beta_t + e_i(\beta_e + b_j) + B_{te}t_je_i + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \tag{6}$$

and variance $V(\mu_{ij}) = \mu_{ij} + \phi_j\mu_{ij}^2$. As the observed environmental variable e_i we used covariate moisture content and for the observed trait t_j we used LTL from the data. Values for the parameters β_{0j}, β_t, β_e and γ_j were based on negative binomial GLLVM with two latent variables fitted for the ground beetle data. Values for latent variables $\boldsymbol{u}_i$ were based on predicted latent variables and values for site effects r_i were based on predicted random site effects. Type I errors for the hypothesis $H_0 : B_{te} = 0$ were calculated based on 800 generated datasets when a variance of the random slopes b_j was varied with a range, $\sigma_b^2 \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$.

Results are provided in Figure 8 and are quite similar to the results showed in Section 3 for the GLLVMs and the p_{max} test. Type I errors for the GLMM with random site effects and random slopes are notably inflated.

B Proof for the variational approximation of the likelihood of the

Assume that the responses come from the exponential family of distributions with density $f(y_{ij}|\boldsymbol{r}_i, \boldsymbol{u}_i, \boldsymbol{b}_j; \boldsymbol{\Psi}) = \exp\{(y_{ij}\eta_{ij} - b(\eta_{ij}))/\phi_j + c(y_{ij}, \phi_j)\}$. The variational approximation for the marginal log-likelihood can then be obtained as follows

$$
\begin{aligned}
\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) &= \int \log\left\{\frac{f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}; \boldsymbol{\Psi}) f(\boldsymbol{r}, \boldsymbol{u}; \boldsymbol{C}_\sigma) f(\boldsymbol{b}; \boldsymbol{G})}{q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b})}\right\} q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b}) d(\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}), \\
&= \int (\log f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}; \boldsymbol{\Psi}) + \log f(\boldsymbol{r}, \boldsymbol{u}; \boldsymbol{C}_\sigma) + \log f(\boldsymbol{b}; \boldsymbol{G}) - \log q(\boldsymbol{r}, \boldsymbol{u}) - \log q(\boldsymbol{b})) \\
&\quad \times q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b}) d(\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}), \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} E_q\{\log f(\boldsymbol{y}_{ij}|(r_i, \boldsymbol{u}_i, \boldsymbol{b}_j), \boldsymbol{\Psi})\} + \sum_{i=1}^{n} E_q\{\log f((r_i, \boldsymbol{u}_i; C_\sigma))\} \\
&\quad + \sum_{j=1}^{m} E_q\{\log f(\boldsymbol{b}_j; \boldsymbol{G})\} + \sum_{i=1}^{n} E_q\{-\log q(r_i, \boldsymbol{u}_i|\boldsymbol{\xi})\} + \sum_{j=1}^{m} E_q\{-\log q(\boldsymbol{b}_j|\boldsymbol{\xi})\},
\end{aligned}
$$

where E_q is expectation with respect to variational density $q(\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}) = q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b})$. Expectation $E_q\{-\log q(r_i, \boldsymbol{u}_i)\}$ is the definition to the entropy of $q(r_i, \boldsymbol{u}_i)$ which equals to $\log\det(2\pi e \boldsymbol{A}_i)/2$ and similarly $E_q\{-\log q(\boldsymbol{b}_j)\} = \log\det(2\pi e \boldsymbol{A}_{bj})/2$. When we omit all quantities constant with respect to the parameters, the above equals to

$$
\begin{aligned}
\underline{\ell}(\boldsymbol{\Psi},\boldsymbol{\xi}) =& \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_{q^*}\{b(\eta_{ij})\}}{\phi_j} + c(y_{ij},\phi_j)\right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\left\{\log\det \boldsymbol{A}_i - E_q\left\{(r_i,\boldsymbol{u}_i')\boldsymbol{C}_\sigma^{-1}(r_i,\boldsymbol{u}_i')' + \log\det(\boldsymbol{C}_\sigma)\right\}\right\} \\
&+ \frac{1}{2}\sum_{j=1}^{m}\left\{\log\det \boldsymbol{A}_{bj} - E_q\left\{\boldsymbol{b}_j{}'\boldsymbol{G}^{-1}\boldsymbol{b}_j + \log\det(\boldsymbol{G})\right\}\right\} \\
=& \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_q\{b(\eta_{ij})\}}{\phi_j} + c(y_{ij},\phi_j)\right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - \mathrm{tr}(\boldsymbol{C}_{\sigma^2}^{-\frac{1}{2}}\boldsymbol{A}_i\boldsymbol{C}_{\sigma^2}^{-\frac{1}{2}}) - \boldsymbol{a}_i'\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{a}_i - \log\det(\boldsymbol{C}_{\sigma^2})\right) \\
&+ \frac{1}{2}\sum_{j=1}^{m}\left(\log\det(\boldsymbol{A}_{bj}) - \mathrm{tr}(\boldsymbol{G}^{-\frac{1}{2}}\boldsymbol{A}_{bj}\boldsymbol{G}^{-\frac{1}{2}}) - \boldsymbol{a}_{bj}'\boldsymbol{G}^{-1}\boldsymbol{a}_{bj} - \log\det(\boldsymbol{G})\right) \\
=& \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_q\{b(\eta_{ij})\}}{\phi_j} + c(y_{ij},\phi_j)\right\} \\
&+ \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - \mathrm{tr}(\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{A}_i) - \boldsymbol{a}_i'\boldsymbol{C}_{\sigma^2}^{-1}\boldsymbol{a}_i - \log\det(\boldsymbol{C}_{\sigma^2})\right) \\
&+ \frac{1}{2}\sum_{j=1}^{m}\left(\log\det(\boldsymbol{A}_{bj}) - \mathrm{tr}(\boldsymbol{G}^{-1}\boldsymbol{A}_{bj}) - \boldsymbol{a}_{bj}'\boldsymbol{G}^{-1}\boldsymbol{a}_{bj} - \log\det(\boldsymbol{G})\right),
\end{aligned}
$$

where $\tilde{\eta}_{ij} = \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{a}_{bj}) + vec(\boldsymbol{B}_{te})'(\boldsymbol{t}_j \otimes \boldsymbol{e}_i) + \boldsymbol{a}_i'(1, \boldsymbol{\gamma}_j')'$. The matrix $\boldsymbol{C}_\sigma^{-1/2}$ is the square root of $\boldsymbol{C}_\sigma^{-1}$ which means that $\boldsymbol{C}_\sigma^{-\frac{1}{2}}\boldsymbol{C}_\sigma^{-\frac{1}{2}} = \boldsymbol{C}_\sigma^{-1}$. This operation is possible for positive semidefinite matrices $\boldsymbol{C}_\sigma$ and $\boldsymbol{G}$. The same result holds for matrix $\boldsymbol{G}$.

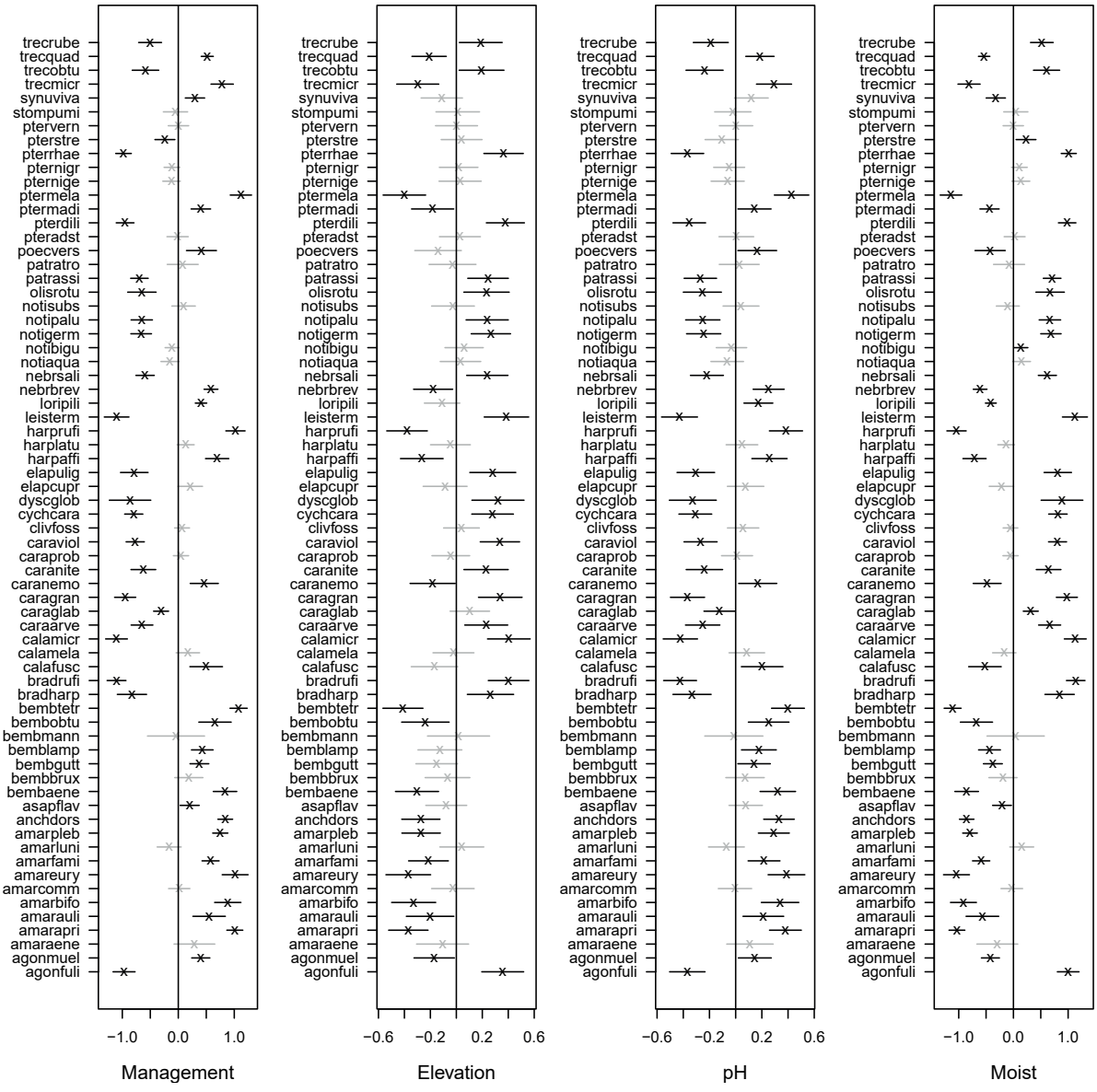

Figure 5: Point predictions for species-specific random slopes and associated 95% uncertainty intervals from a fourth-corner latent variable model with two latent variables fitted to the ground beetle data.

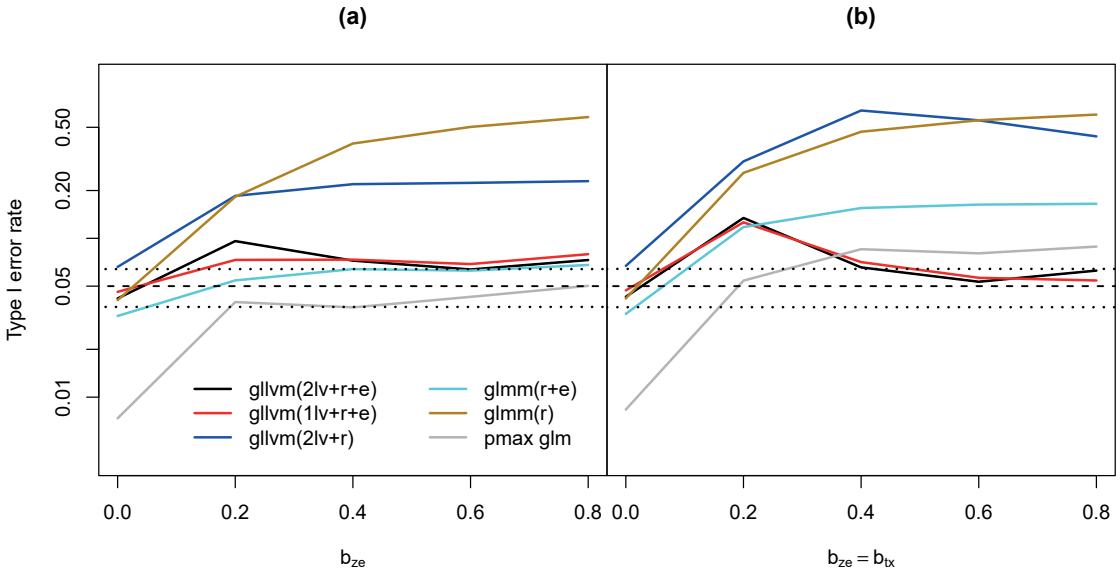

Figure 6: Type I error rates obtained using simulation setup described in Appendix A.1 for likelihood ratio tests based on GLMM with random intercepts (glmm(r)), GLMM with random intercepts and slopes (glmm($r + e$)), GLLVM with random intercepts (gllvm($2lv + r$)), and GLLVM with random intercepts, slopes, and $d = 1, 2$ latent variables (gllvm($d\,lv + r$)), and the p_{max} test. Generated datasets consisted of $n = 40$ sites and $m = 40$ species. Dashed line is the nominal level 0.05 and dotted lines around nominal level correspond values for sample proportions which are not significantly different from 0.05.

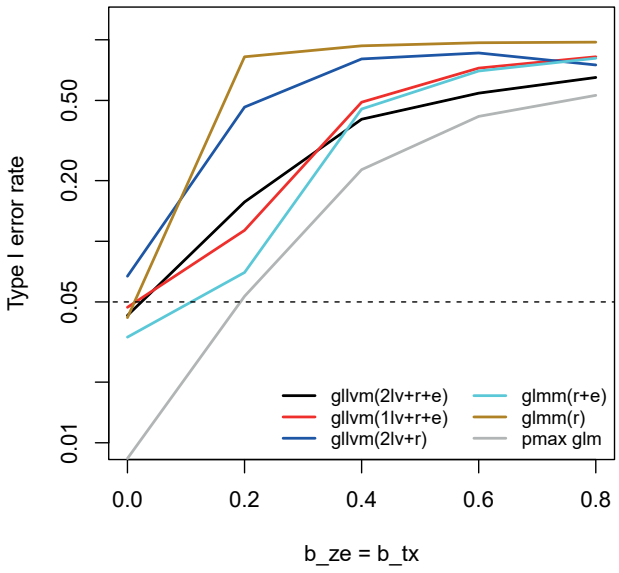

Figure 7: Type I error rates obtained using simulation setup described in Appendix A.1 for likelihood ratio tests based on GLMM with random intercepts (glmm(r)), GLMM with random intercepts and slopes (glmm($r + e$)), GLLVM with random intercepts (gllvm($2lv + r$)), and GLLVM with random intercepts, slopes, and $d = 1, 2$ latent variables (gllvm($d\,lv + r$)), and the p_{max} test. Generated datasets consisted of $n = 40$ sites and $m = 40$ species. In the mean model, we used latent environmental variables x_i and latent trait variables z_j which were correlated with the observed environmental e_i and observed trait variables t_j with correlation of 0.3.

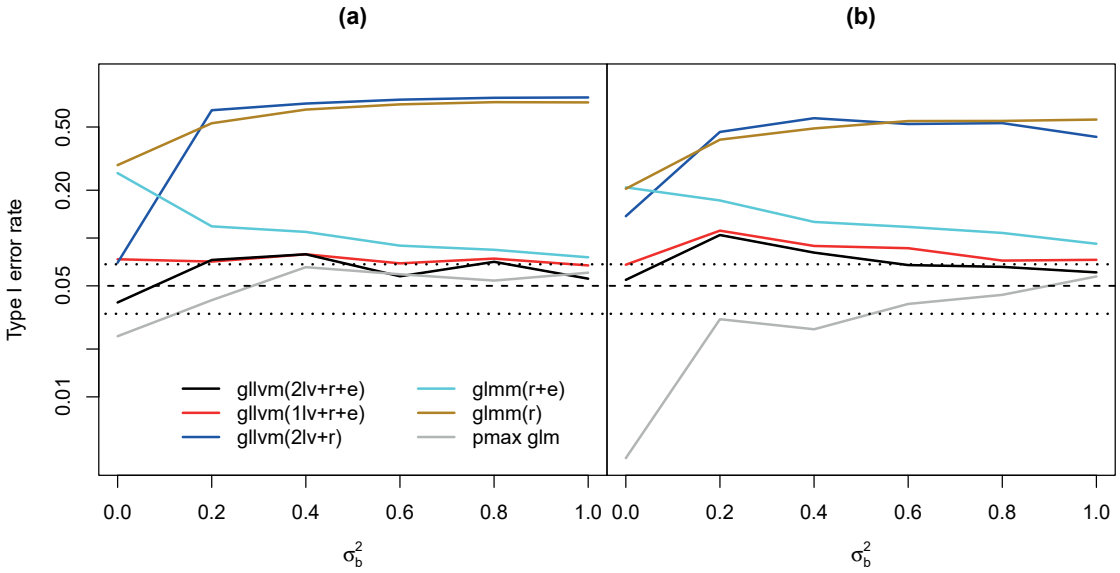

Figure 8: Type I error rates for likelihood ratio tests based on GLMM with random intercepts and slopes (glmm($r + e$)), GLLVM with random intercepts (gllvm($2lv + r$)), and GLLVM with random intercepts, slopes, and $d = 1, 2$ latent variables (gllvm($d\,lv + r$)), and the p_{max} test. The simulation setup is based on the subsets of the beetle data with (a) $n = 80$ and $m = 40$ and (b) $n = 35$ and $m = 68$.

PIV

GLLVM - FAST ANALYSIS OF MULTIVARIATE ABUNDANCE DATA WITH GENERALIZED LINEAR LATENT VARIABLE MODELS IN R

by

Niku, J., Hui, F. K., Taskinen, S. and Warton, D. I 2019

Methods in Ecology and Evolution, 10:2173–2182

Received: 7 May 2019 | Accepted: 5 September 2019

DOI: 10.1111/2041-210X.13303

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

APPLICATION

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

Jenni Niku[1] | Francis K. C. Hui[2] | Sara Taskinen[1] | David I. Warton[3]

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

[2]Research School of Finance, Actuarial Studies & Statistics, Australian National University, Canberra, Australia

[3]School of Mathematics and Statistics and Evolution & Ecology Research Centre, UNSW Sydney, Canberra, Australia

Correspondence
Jenni Niku
Email: jenni.m.e.niku@jyu.fi

Funding information
Jenny ja Antti Wihurin Rahasto; Australia Research Council Discovery Project, Grant/Award Number: DP150100823 and DP180100836; CRoNoS COST, Grant/Award Number: IC1408

Handling Editor: Sarah Goslee

Abstract

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 datasets.
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, for example, 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 maximized, 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
abundance data, generalized linear latent variable models, high-dimensional data, joint modelling, maximum likelihood, multivariate analysis, ordination, species interactions

1 | INTRODUCTION

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 characterize 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 is similar in terms of their species composition (Bjork, Hui, O'Hara, & Montoya, 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, Thuiller, Lek, Lim, & Grenouillet, 2008).

In recent years, there has been a growing movement towards the specification of statistical models for multivariate analysis in ecology (Ovaskainen, Hottola, & Siitonen, 2010; Ovaskainen et al., 2017; Warton et al., 2015). 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, for example, biotic interaction.

A key approach for statistical modelling of multivariate abundance data is the generalized linear latent variable model (GLLVM, Skrondal & Rabe-Hesketh, 2004). A GLLVM extends the basic generalized linear model to multivariate data using a factor analytic approach, that is, 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, for example, predicting new values, controlling for known environmental variables, using standard model selection tools to choose number of ordination axes (Hui, Taskinen, Pledger, Foster, & Warton, 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, for example, microbial data often have thousands of taxa (Kumar et al., 2017; Niku, Warton, Hui, & Taskinen, 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, for example, the `boral` (Hui et al., 2016 and `HMSC` packages (Tikhonov, Opedal, Abrego, Lehikoinen, & Ovaskainen, 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, Warton, Ormerod, Haapaniemi, & Taskinen, 2017) or Laplace (Niku et al., 2017) approximations to likelihoods, especially via automated differentiation software such as Template Model Builder (Kristensen, Nielsen, Berg, Skaug, & Bell, 2016).

This paper presents the R package `gllvm` (Niku et al., 2017), 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 environment–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 (Hui, 2016; Tikhonov et al., 2019), 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).

2 | 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, $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{id})^\top$:

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

where $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ are vectors of species specific coefficients related to the covariates and latent variables, respectively. The latent variables $\boldsymbol{u}_i$ 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 that 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 $\boldsymbol{\Gamma} = [\gamma_1 \ \ldots \ \gamma_m]'$ needs to be set to zero and the diagonal elements to be set positive to avoid rotational invariance; 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 $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$. This is the correct form of correlation when the responses are Poisson distributed. In the case of negative binomial distribution with dispersion parameters $\boldsymbol{\Phi} = (\phi_1, \ldots, \phi_m)^\top$, 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 $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top + \boldsymbol{I}_m$ (Ovaskainen, Abrego, Halme, & Dunson, 2016).

If q trait covariates $\boldsymbol{t}_j = (t_{i1}, \ldots, t_{iq})^\top$ are also recorded, we can use them to help explain interspecific variation in environmental response. This leads to an extension of the so-called 'fourth corner model' (Brown et al., 2014; Jamil & ter Braak, 2013) 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:

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

where $\boldsymbol{\beta}_e$ is a vector of main effects for environmental covariates, and $\boldsymbol{\beta}_l$ is the fourth corner coefficient. 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 interspecific 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 Equation 2.

In both GLLVM formulations mentioned above, a key feature is that the number of parameters characterizing the residual correlation $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top$ grows linearly with the number of responses m. This contrasts

with the quadratic rate of growth when an unstructured residual covariance matrix was 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.

3 | 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 maximize has the form

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

where $\boldsymbol{\psi}$ includes all model parameters. In this expression, we have assumed that abundances are independent across sites and any correlation across responses are 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, Skrondal, & Pickles, 2002), the Monte Carlo applications of the expectation maximization (EM) algorithm (Hui et al., 2015) and Bayesian MCMC (Hui, 2016; Tikhonov et al., 2019). For large datasets and multiple latent variables, these methods are, however, time-consuming.

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

- Maximizing the log-likelihood using (almost completely) closed form approximation. 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, for example, 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.

4 | 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, Naumann, Wright, & Warton, 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.

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	$E(y_{ij})=\mu_{ij}$, $V(y_{ij})=\phi_j^2$

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 are available in the R package mvabund (Wang et al., 2012) and consist 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
```

5 | 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 (Hui et al., 2015; Walker & Jackson, 2011). 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 given below. We will consider two count distributions for the data – the Poisson and negative binomial (NB).

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()	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 an Intel Core i7-3770 (3.4 GHz)) 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	1,483.6	1,529.9

```
> 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 & 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 show some overdispersion in residuals, in particular, a telltale fan shape in the plot of residuals against fitted values. These issues are largely

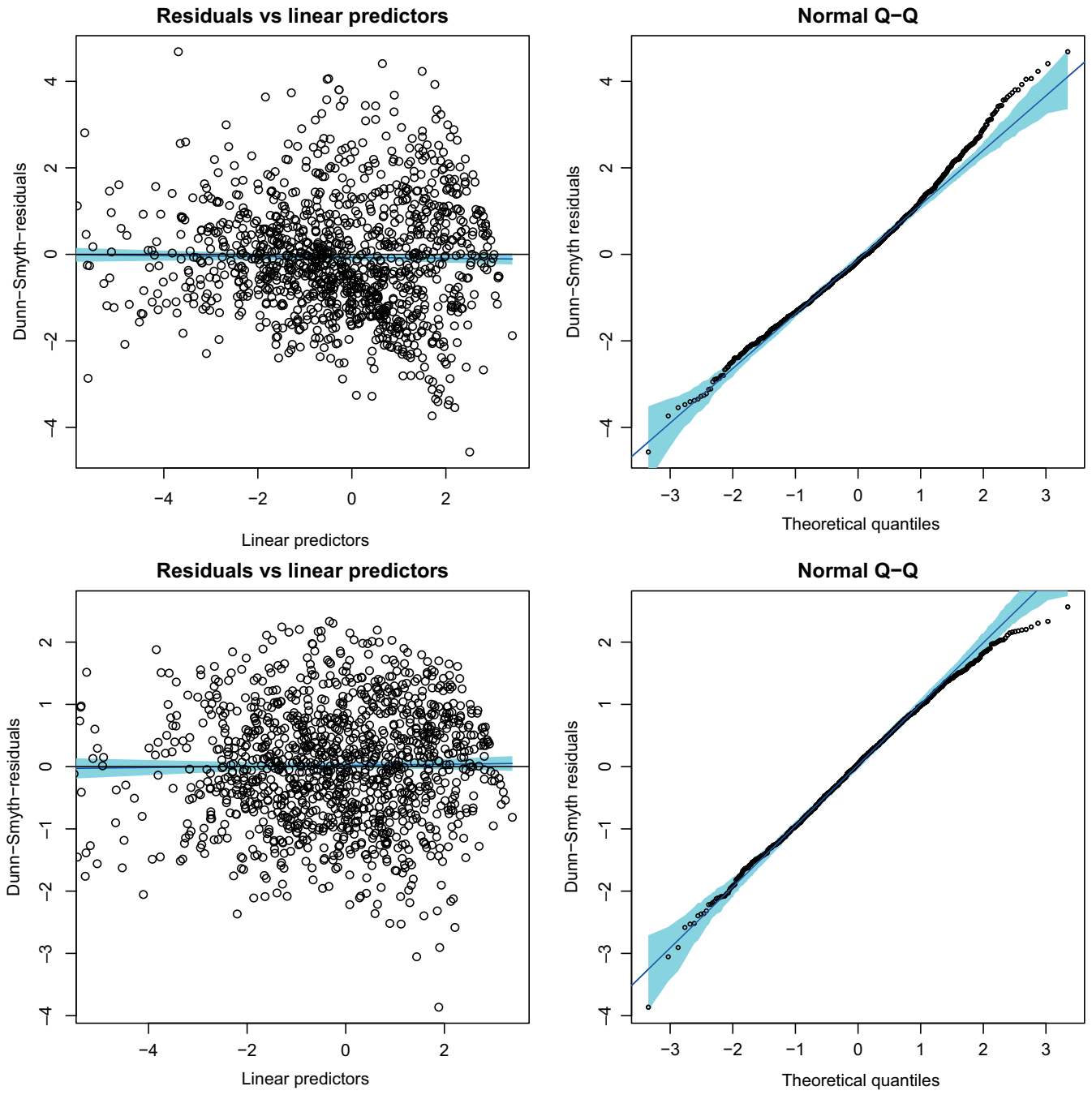

FIGURE 1 Residual plots for the Poisson GLLVM (top) and the NB-GLLVM (bottom) applied for model-based ordination. Specifically, Dunn-Smyth residuals are plotted against linear predictors (left), while simulated point-wise 95% confidence interval envelope is added in the normal quantile−quantile plot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

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, $||\gamma_j||$), 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))
```

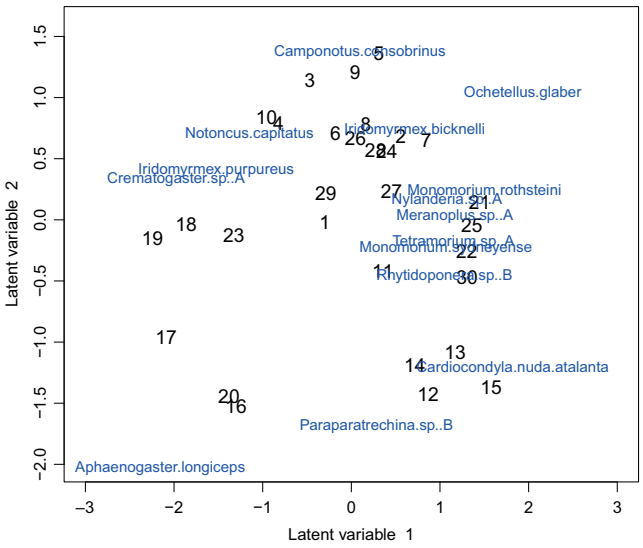

FIGURE 2 A biplot with 15 indicator species based on the NB-GLLVM fitted to the ant data. The numbers correspond to the site indices

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, for example, sites 12–15. In Appendix S3, we apply classical algorithm-based ordination methods to the ant data and compare the results. While the results between GLLVMs and the algorithm-based methods are quite similar, GLLVMs offer the advantage of standard tools for diagnosing model fit and performing model selection.

6 | 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)))
```

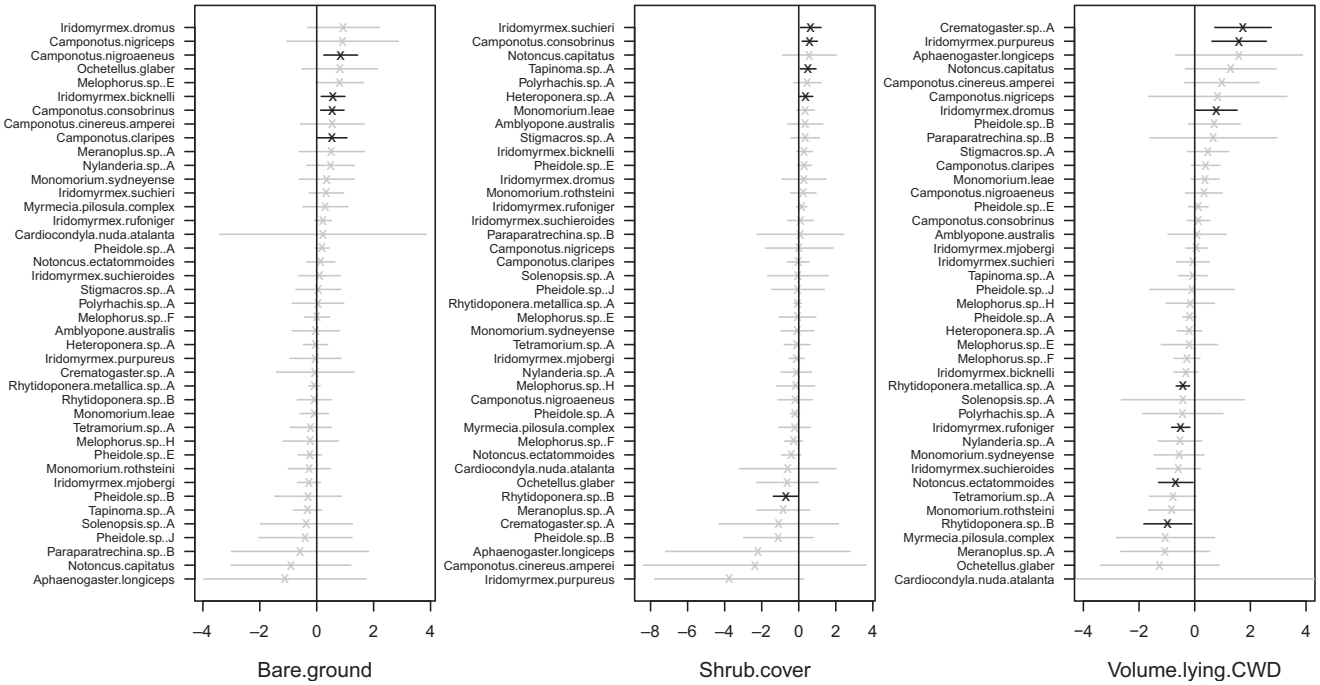

FIGURE 3 Plots of the point estimates (ticks) for coefficients of the environmental variables and their 95% confidence intervals (lines) for the NB-GLLVM, with those coloured in grey (black) denoting intervals (not) containing zero. The *x*-axis of the coefficient plot of the third variable is truncated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

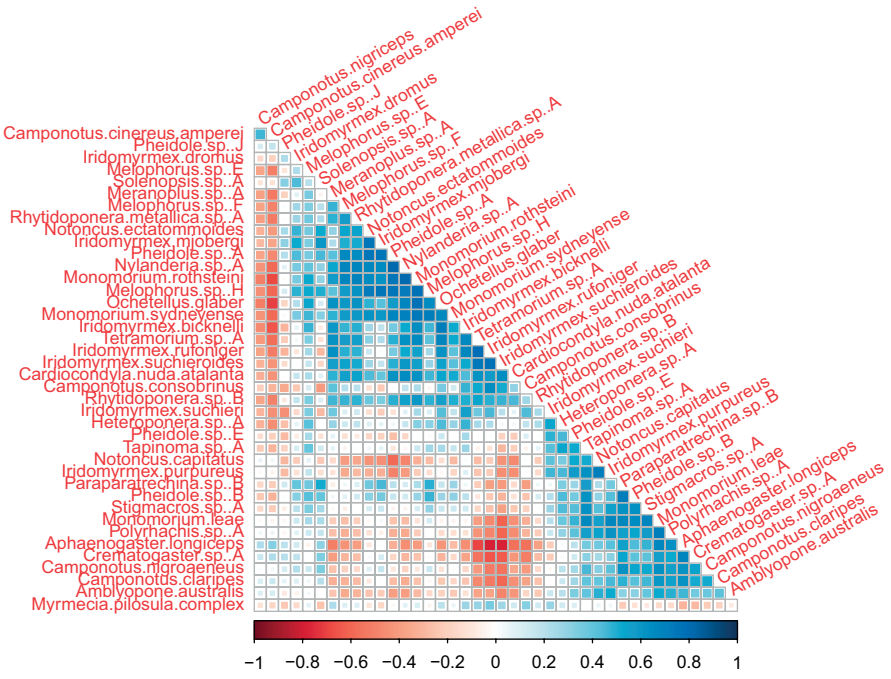

FIGURE 4 Residual correlation matrix based on latent factor loadings for the NB-GLLVM with environmental covariates

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 does 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.

7 | 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 visualized 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 S1, where the correlation patterns are considerably different from one another. Correlations can also be visualized in a residual biplot (Appendix S1). 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 S1.

8 | 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 environment–trait interaction terms, also known as the fourth corner terms, can also be visualized using the function `levelplot()` from the package `lattice`, see Appendix S1 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()` recognizes 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, that is, 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.

9 | 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 (Hui, 2016; Tikhonov et al., 2019) 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 settings (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 min, while boral (under default settings) took 10 hr, 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 emphasize 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 (Hui, 2016; Tikhonov et al., 2019), 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 & 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 artefacts into analyses (Warton & Hui,

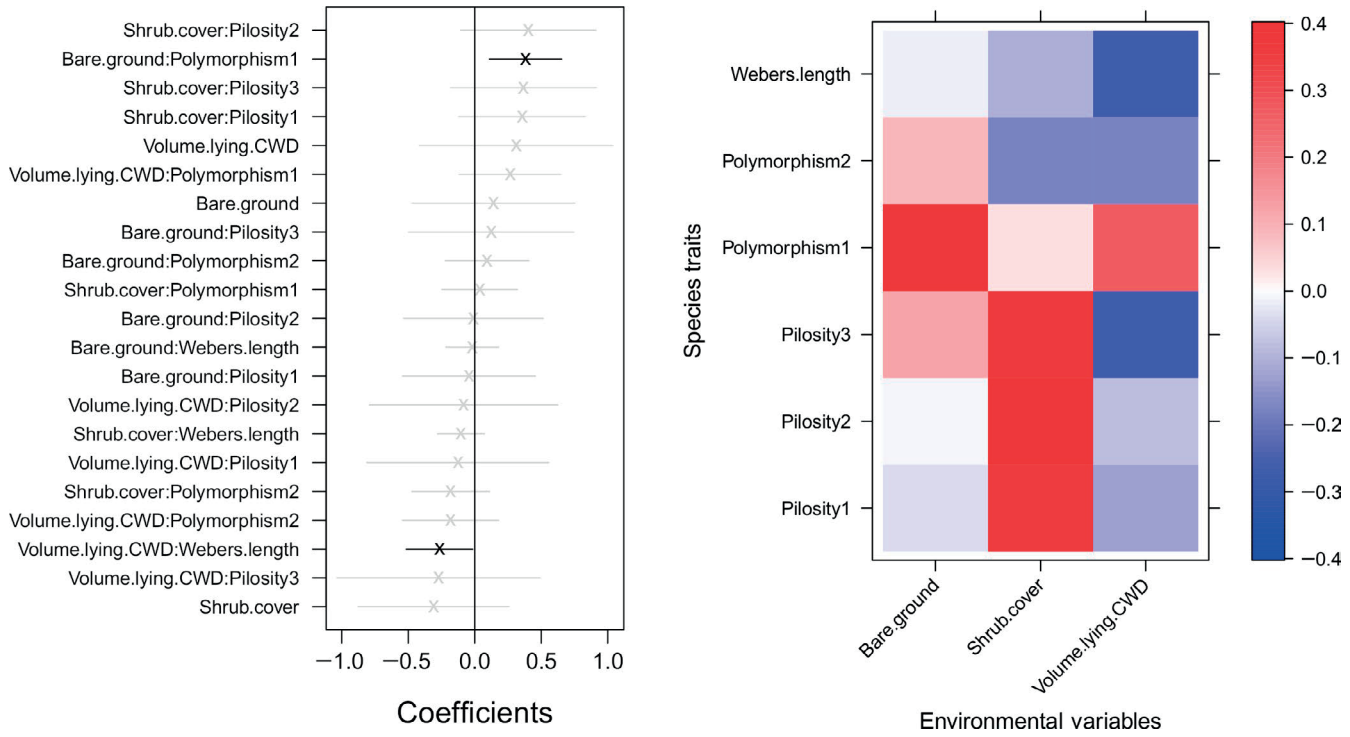

FIGURE 5 A plot of the estimated coefficients (ticks) and their 95% confidence intervals (lines) for all terms in the fourth corner model (left), and a level plot for the fourth corner interaction terms (right) in the NB-GLLVM. The colours offer an indication of the signs and magnitudes of the point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

2017; Warton, Wright, & Wang, 2012). 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 characterize 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.

ACKNOWLEDGEMENTS

The work of J.N. was supported by the Wihuri Foundation. The work of S.T. was supported by the CRoNoS COST Action IC1408. The work of F.K.C.H. and D.I.W. was supported by Australia Research Council Discovery Project grants (DP180100836 and DP150100823, respectively), F.K.C.H. was also supported by an ANU cross disciplinary grant.

AUTHORS' CONTRIBUTIONS

J.N., F.K.C.H., S.T. and D.I.W. conceived the ideas and designed methodology; J.N. was mainly responsible for implementing the application; All authors contributed to the writing, reviewing and editing of the draft and gave final approval for publication.

DATA AVAILABILITY STATEMENT

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) are published in European Nucleotide Archive under the project number PRJEB17695, https://www.ebi.ac.uk/ena/data/view/PRJEB17695. A subset of these data used in Appendix S2, 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/.

ORCID

Jenni Niku https://orcid.org/0000-0002-7992-2598
David I. Warton https://orcid.org/0000-0001-9441-6645

REFERENCES

Bjork, J. R., Hui, F. K. C., O'Hara, R. B., & Montoya, J. M. (2018). Uncovering the drivers of host-associated microbiota with joint species distribution modelling. *Molecular Ecology*, *27*, 2714–2724. https://doi.org/10.1111/mec.14718

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & 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. https://doi.org/10.1111/2041-210x.12163

Buisson, L., Thuiller, W., Lek, S., Lim, P., & Grenouillet, G. (2008). Climate change hastens the turnover of stream fish assemblages. *Global Change Biology*, *14*, 2232–2248. https://doi.org/10.1111/j.1365-2486.2008.01657.x

Dray, S., & Dufour, A. B. (2007). The ADE4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*, 1–20. https://doi.org/10.18637/jss.v022.i04

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*, 236–244. https://doi.org/10.1080/10618600.1996.10474708

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. https://doi.org/10.1111/2041-210x.12514

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*, 399–411. https://doi.org/10.1111/2041-210x.12236

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., & Taskinen, S. (2017). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, *26*, 35–43. https://doi.org/10.1080/10618600.2016.1164708

Jamil, T., & ter Braak, C. J. (2013). Generalized linear mixed models can detect unimodal species-environment relationships. *PeerJ*, *1*, e95. https://doi.org/10.7717/peerj.95

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, *70*, 1–21. https://doi.org/10.18637/jss.v070.i05

Kumar, M., Brader, G., Sessitsch, A., Mäki, A., van Elsas, J. D., & Nissinen, R. (2017). Plants assemble species specific bacterial communities from common core taxa in three arcto-alpine climate zones. *Frontiers in Microbiology*, *8*, 12. https://doi.org/10.3389/fmicb.2017.00012

Lammel, D. R., Barth, G., Ovaskainen, O., Cruz, L. M., Zanatta, J. A., Ryo, M., … 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., & 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., & Warton, D. I. (2019b). Efficient estimation of generalized linear latent variable models. *PLoS ONE*, *14*(5), 1–20. https://doi.org/10.1371/journal.pone.0216129

Niku, J., Warton, D. I., Hui, F. K. C., & 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. https://doi.org/10.1007/s13253-017-0304-7

Oksanen, J., Guillaume Blanchet, F., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., … 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., & 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. https://doi.org/10.1111/2041-210x.12501

Ovaskainen, O., Hottola, J., & Siitonen, J. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. *Ecology*, *91*, 2514–2521. https://doi.org/10.1890/10-0173.1

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., … 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. https://doi.org/10.1111/ele.12757

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., … 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. https://doi.org/10.1111/2041-210x.12180

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *Stata Journal*, *2*, 1–21. https://doi.org/10.1177/1536867x0200200101

Royan, A., Reynolds, S. J., Hannah, D. M., Prudhomme, C., Noble, D. G., & Sadler, J. P. (2016). Shared environmental responses drive co-occurrence patterns in river bird communities. *Ecography*, *39*, 733–742. https://doi.org/10.1111/ecog.01703

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman & Hall.

ter Braak, C. J. F. (2019). New robust weighted averaging- and model-based methods for assessing trait-environment relationships. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13278

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

Walker, S. C., & Jackson, D. A. (2011). Random-effects ordination: Describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, *81*, 635–663. https://doi.org/10.1890/11-0886.1

Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). MVABUND – An R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, *3*, 471–474. https://doi.org/10.1111/j.2041-210x.2012.00190.x

Warton, D. I., Blanchet, F. G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So many variables: Joint modeling in community ecology. *Trends in Ecology and Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Warton, D. I., & 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. https://doi.org/10.1111/2041-210x.12843

Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, *3*, 89–101. https://doi.org/10.1111/j.2041-210x.2011.00127.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Niku J, Hui FKC, Taskinen S, Warton DI. `gllvm`: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods Ecol Evol*. 2019;00:1–10. https://doi.org/10.1111/2041-210X.13303