Fitting Generalized Linear Latent Variable Models using the method of Extended Variational Approximation

Master's thesis in Statistics

30th November 2020
Pekka Korhonen
Department of Mathematics and Statistics
University of Jyväskylä

University of Jyväskylä
Department of Mathematics and Statistics
Korhonen, Pekka: *Fitting Generalized Linear Latent Variable Models using the method of Extended Variational Approximation*
Master's thesis in Statistics, 41 pages
30th November 2020

Abstract

Generalized Linear Latent Variable Models (GLLVM), a family of statistical models developed on recent years, has gained a lot of attraction in applications, in particular in the field of community ecology. Ecologists are often concerned with the relationships between two or more species across a multiple test sites. Such situations naturally lead to the collection of *multivariate abundance data* and call for appropiate statistical methods to analyze such data. GLLVMs offer a model-based approach for such analyses that is also flexible in the terms of the type of abundance response at question, i.e., species count, presence/absence, biomass, and such. As their namesake implies, GLLVMs generally assume the presence of some unobserved, latent variables as predictors. These latent variables are useful, for example in the modelling of the between-species correlation, but they also introduce some computational challenges into the model fitting.

In its general form, the GLLVM marginal likelihood involves an integral over the aforementioned latent variables. Under the standard assumptions this integral cannot be solved analytically, when dealing with other than normally distributed response variables. Thus some form of numerical approximation technique is often needed. This thesis starts by introducing a *variational approximation* (VA) approach for fitting GLLVMs, which has shown to be an attractive choice in terms of both the computational efficiency and estimation precision. From there we introduce a recently proposed method of *extended variational approximation* (EVA), which extends upon the standard VA approach by allowing a wider set of response distributions and link functions to be used in modelling. Then the comparative performance of these two approaches and a popular alternative, *Laplace approximation* (LA), is addressed in simulation studies. Additionally, an example study concerning the use of EVA in ordination of plant cover data is conducted. Lastly we discuss some ideas for further development regarding the EVA approach.

The VA and LA approaches to estimation of GLLVMs are readily available in the R package gllvm, which has been used in this thesis. An implementation of the EVA approach for a few types of common response distributions was developed as a part of this thesis in R and C++ using the package TMB.

Keywords: generalized linear latent variable models, variational inference, abundance data, simulation, ordination

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Korhonen, Pekka: *Laajennettu variaatiomenetelmä yleistettyjen lineaaristen latenttimuuttujamallien sovituksessa*
Tilastotieteen Pro gradu -tutkielma, 41 sivua
30. marraskuuta 2020

Tiivistelmä

Yhteisöekologian alalla tutkijat ovat usein kiinnostuneita yhden tai useamman kasvi- tai eläinlajin välisistä esiintyvyyssuhteista eri mittauspaikoilla tai ekosysteemeissä. Tämänkaltaiset tutkimuskysymykset johtavat luonnostaan moniulotteisen runsausdatan keräämiseen. Kasvi- tai eläinlajin ekologista runsautta tietyssä ekosysteemissä voidaan kuvata esimerkiksi suoraan lajiyksilöiden lukumääränä tai binäärisenä esiintyvyysindikaattorina. Runsausvasteen tyyppi on otettava huomioon tilastollista mallia sovittaessa. *Yleistetyt lineaariset latenttimuuttujamallit* tarjoavat joustavan tavan mallintaa moniulotteista runsautta olettamalla yhden tai useamman latentin muuttujan olemassaolon. Latentit muuttujat ovat luonteeltaan satunnaisia ja havaitsemattomia. Niiden voidaan tulkita kuvaavan esimerkiksi havaitsematta jääneitä ympäristötekijöitä. Latentit muuttujat ovat hyödyllisiä, sillä niiden avulla voidaan mallintaa eri lajien välistä korrelaatiorakennetta. Latenttimuuttujamallien sovittaminen ei kuitenkaan ole erityisen suoraviivaista latenttien muuttujien havaitsemattomuudesta johtuen.

Latenttimuuttujamallia vastaava marginaalinen uskottavuusfunktio sisältää integraalin, jolla ei yleisessä tapauksessa ole analyyttistä ratkaisua. Mallin sovituksessa joudutaan tämän vuoksi käyttämään jotakin approksimatiivista menetelmää. Eräs varteenotettava vaihtoehto on niin sanottu *variaatiomenetelmä*, joka esitellään tämän tutkielman alussa. Menetelmän etuna on sekä estimointitarkkuus että laskennallinen tehokkuus. Variaatiomenetelmän selvänä heikkoutena on sen huono yleistyvyys, sillä se ei suoraan sovellu käytettäväksi kaikkien tavanomaisten vastejakaumalinkkifunktio -parien yhteydessä. Tämän vuoksi tässä tutkielmassa esitetään nyt *laajennettuksi variaatiomenetelmäksi* nimetty menetelmä. Esitettyä laajennosta verrataan sekä tavanomaiseen variaatiomenetelmään että Laplace-approksimaatioon perustuvaan kilpailevaan menetelmään aineistopohjaisten simulointikokeiden avulla. Lisäksi esitellään laajennetun variaatiomenetelmän käyttöä suoaineistolle tehtävässä ordinaatiossa. Suoaineisto on peräisin Jyväskylän yliopiston Bio- ja ympäristötieteen laitokselta. Laajennettu variaatiomenetelmä implementoitiin ohjelmointikieliä R ja C++ käyttäen muutaman tyypillisimmän latenttumuuttujamallin tapauksessa.

Avainsanoja: yleistetty lineaarinen latenttimuuttujamalli, variaatiopäättely, runsausdata, simulointi, ordinaatio

Contents

1 Introduction

In some research fields applying statistical methods, there is often a need to process and analyse multidimensional data. Several response variables of interest may be measured per observational unit, and a key research question might be about determining the structure of the relationships between the different response variables through the observation units. Typical example of this can be found from the field of community ecology, where researchers collect and use *ecological abundance data* to gain knowledge about several interacting plant or animal species across different test sites or ecosystems. Such data can naturally be represented as $n \times m$ matrix $\boldsymbol{Y}$, where rows correspond to n observational units or sites, and columns correspond to m species. Thus, an element y_{ij} of $\boldsymbol{Y}$ is the *abundance* of species $j = 1, \dots, m$ measured at site $i = 1, \dots, n$. Commonly, abundance is measured as the count of the individual units at a particular site. In this thesis however, we are going to use a bit more loose definition, allowing for the inclusion of binary 'presence/absence' data, and continuous proportion data. In general, the term abundance can be understood as any kind of appropriate measure for representation of species in a given ecosystem. The type of abundance responses needs to be accounted for when using a model based approach to analyse multivariate abundance data.

Traditionally, the analysis of multivariate abundances has been conducted using algorithmic *ordination methods*, such as *non-metric multidimensional scaling* (nMDS, Kruskal, 1964). In algorithmic ordination, a multidimensional data is reduced down to (typically) two primary axes of variation, according to some predetermined way of measuring the dissimilarities between sites. The data might need to be transformed before applying the dimensionality reduction technique. The main disadvantages of these types of methods is the lack of proper diagnostic tools when compared to model based approaches. This means that the appropriateness of the chosen dissimilarity measure or the transformation method might be difficult to assess.

Statistical modelling of multivariate abundances is a fairly recent field of development, made possible by the increase and availability in computational power. When the inter-species interactions are of interest, a *joint species distribution model* (Pollock et al., 2014) is needed. One flexible class of statistical models suited for the task are the *generalized linear mixed models* (GLMM, Jiang, 2007). In its essence, a GLMM is an extension of generalized linear model (GLM), containing random effects in addition to the fixed effects present in a standard GLM. In the purpose of joint modelling of species abundances, the random effects can be assumed to be specific to site, independent and distributed according to multivariate normal distribution, $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{im}^{\intercal}) \overset{i.i.d}{\sim} \boldsymbol{\mathcal{N}}(\boldsymbol{0}, \boldsymbol{\Sigma})$. The covariance matrix $\boldsymbol{\Sigma}$ now represents the inter-species interactions across the different test sites. Typically, no additional assumptions regarding the structure of $\boldsymbol{\Sigma}$ are made. This is often troublesome, as the amount of parameters to be estimated in the $m \times m$ matrix Σ increases quadratically with the amount of species, quickly leading into computational problems. Thus, a call for a more efficient modelling approach is reasonable. An attractive answer to this call is the class of statistical models called the *generalized linear latent variable models* (GLLVM, Skrondal and Rabe-Hesketh, 2004).

In GLLVM, the random effect z_{ij} present in GLMM is assumed to actually be a dot product of two vectors, one being the vector of the so-called *latent variable values* or *scores* $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^{\intercal}$, and the other being the vector of *latent variable loadings* $\boldsymbol{\lambda}_j = (\lambda_{j1}, \ldots, \lambda_{jp})^{\intercal}$, where p denotes the assumed number of latent variables. As the subscripts suggest, the scores are assumed to be specific to sites, while the loadings are assumed to be specific to species. This type of factorization coupled with the additional modelling assumptions described in more detail in Section 2.1 lead to a potentially far more computationally efficient way to make inferences about the inter-species correlation structure, than what the standard GLMMs can provide. Additionally, GLLVMs can be used to do model based ordination as is, simply by choosing $p = 2$.

Even though potentially more efficient than the mixed models in theory, in practice the actual estimation of latent variable models remains a difficult task, caused by the assumed existence of latent random variables. To carry out inferences about the model parameters based on maximum likelihood estimation, one needs to integrate over the latent variable space. The integral in question lacks an analytical solution in most cases and thus a numerical approximation scheme is often needed. Here, one often has to make a choice between computational efficiency, estimation accuracy and general applicability. In this thesis we extend upon one such numerical method called the *variational approximation* (VA, Ormerod and Wand, 2010). According to Warton et al. (2015) the use of VA in the fitting of GLLVM provides a promising balance between accuracy and efficiency, when compared to alternatives. What the VA approach is lacking however, is the general applicability. For example in the case of Bernoulli distributed binary abundance data, i.e. presence/absence responses, one must resort to using the probit link function, as the method is not applicable for the canonical logit link function. Thus to address this issue, we introduce an approximation method called the *extended variational approximation* (EVA), in Section 2.3. Specific derivations of the approximate log-likelihoods associated with EVA are also provided, in Section 3, for some common types of latent variable models. In Section 4, the performance of the proposed approach is addressed in three distinct simulation studies, also involving the standard VA method and one traditional alternative, the method of *Laplace approximation* (LA, Tierney and Kadane, 1986). In addition, an example case study showcasing the use of EVA in ordination of presence/absence bryophyte data is conducted in Section 5. The data was received from the Department of Biological and Environmental Science at University of Jyväskylä. Lastly, the possible developments for the future regarding the method of EVA are discussed in Section 6.

2 About GLLVMs and approximative inference

2.1 Generalized linear latent variable models

Generalized linear latent variable models (GLLVM) are, according to Huber et al. (2004), an extension to the popular and well-known framework of generalized linear models (GLM). As in the case of GLMs, the mean μ_{ij} of an exponential family type response variable y_{ij} is regressed against a linear predictor of the explanatory variables via an appropriate choice of link function. However, while GLMs assume that all of the explanatory variables are of observable nature, GLLVMs allow for the existence of some latent, unobservable covariates. Huber et al. (2004) claim that for example in the field of psychology these latent variables can be seen as an unobservable measure of a person's intelligence or some other trait that can only be tested for indirectly, such as anxiety or welfare. In the field of community ecology, latent variables can be thought of as ordination axes describing different test sites by their species abundance or composition, as noted by Hui et al. (2015). In this thesis, we are going to follow the formulation of GLLVMs in the context and terminology of analysing ecological abundance data.

Let μ_{ij} denote the mean response for *species* $j = 1, \ldots, m$ at *site* $i = 1 \ldots, n$ and let $\boldsymbol{x_i}$ be the vector of the observed environmental covariates specific to the site i. Then, a GLLVM in its standard form assumes that

$$g(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j + \boldsymbol{u}_i^\intercal \boldsymbol{\lambda}_j, \quad i = 1, \ldots, n, \ j = 1, \ldots, m, \tag{1}$$

for a known and well-behaved link function g. Here the term β_{0j} and the vector $\boldsymbol{\beta_j}$ stand for the species specific intercept and regression coefficients (related to the environmental covariates $\boldsymbol{x}_i$) as in the case of regular GLMs. Now, the vector $\boldsymbol{u}_i$ represents the values of p latent variables at site i. Then, the vector $\boldsymbol{\lambda_j}$ of species specific loadings quantifies the relationship between the species' response and the latent variables. Finally, α_i is an optional site specific parameter that can be treated either as a fixed or random effect. Hui et al. (2015) argue, that the choice of $p = 1$ or 2 latent variables is often suitable for ordination. The authors also remark that the choice of whether to include the site effects α_i or not in the model (1) corresponds to the choice of whether to plot an ordination of species abundance or composition; by including the terms α_i, the ordination is performed in the basis of the species composition, as the site effect α_i and the species specific intercept β_{0j} together act to standardize the coordinates of the ordination found in the vector $\boldsymbol{u}_i$.

The latent variables $\boldsymbol{u}_i$ in (1) are assumed to be independent and follow a standard multivariate normal distribution. GLLVMs also assume that the responses $y_{i1}, \dots, y_{im}$ at site i are conditionally independent given the latent variables $\boldsymbol{u}_i$. This essentially means, that the correlation structure between responses is completely accounted by the latent variables. Additionally, the vectors of site-wide responses $\boldsymbol{y}_1, \dots, \boldsymbol{y}_n$ are assumed to be independent from each other. Now, use $\boldsymbol{\Psi}$ to denote the vector of all model parameters, that is $\boldsymbol{\Psi} = (\boldsymbol{\alpha}, \boldsymbol{\beta}_0, \text{vec}(\boldsymbol{\beta}), \text{vec}(\boldsymbol{\Lambda}))$, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\intercal$, $\boldsymbol{\beta}_0 = (\beta_{01}, \dots, \beta_{0m})^\intercal$ and $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ are matrices containing the vectors $\boldsymbol{\beta}_j$ and $\boldsymbol{\lambda}_j$ as columns, respectively. Also let $\boldsymbol{u} = (\boldsymbol{u}_1^\intercal, \dots, \boldsymbol{u}_n^\intercal)^\intercal$ denote a vector of the latent variables. Then the full likelihood function for a GLLVM is

$$L(\boldsymbol{\Psi}; \boldsymbol{u}) = \prod_{i=1}^{n} \left(\prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) \right) f(\boldsymbol{u}_i), \tag{2}$$

where $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$ is the conditional p.d.f. for the response y_{ij} and $f(\boldsymbol{u}_i)$ is the standard (multivariate) normal p.d.f. for the latent variable $\boldsymbol{u}_i$. We can denote the joint distribution of the latent variables as $f(\boldsymbol{u}) = \prod_{i=1}^{n} f(\boldsymbol{u}_i)$ and the joint distribution of the responses as $f(\boldsymbol{y}|\boldsymbol{u}, \Psi) = \prod_{i=1}^{n} \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$, thus simplifying (2) into

$$L(\boldsymbol{\Psi}; \boldsymbol{u}) = f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}). \tag{3}$$

After finding the specific complete likelihood function $L(\boldsymbol{\Psi}; \boldsymbol{u})$ as in (2) and (3), a reasonable next step would be to apply the standard maximum likelihood estimation techniques to get estimates for the model parameters $\boldsymbol{\Psi}$. However, due to their unobservant nature, the latent variables cause some issues to this, and the usual methods are typically of no use. One approach to dealing with these issues is to use the marginal likelihood function

$$L(\boldsymbol{\Psi}) = \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) d(\boldsymbol{u}) \tag{4}$$

in the estimation process, instead of the full likelihood in (3). This is the approach we are going to focus on in this thesis. Do note however, that methods operating more directly on the complete likelihood exist, including for example the *Expectation-Maximization algorithm* (EM algorithm, Dempster et al., 1977) or *Markov Chain Monte Carlo sampling* (MCMC, Metropolis et al., 1953). Both of these methods are quite general and widely used across the different fields of applied and computational statistics.

2.2 Variational approximation

While looking at the marginal GLLVM likelihood in (4), or rather, the marginal log-likelihood

$$\ell(\boldsymbol{\Psi}) = \log \left(\int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) d(\boldsymbol{u}) \right), \tag{5}$$

one will quickly run in to additional problems, as the integral in (4) or (5) does not possess an analytical solution in the case of non-normal responses and thus a some form of numerical approximation technique is needed, as noted by Huber et al. (2004). One such technique is the use of *variational approximations* (VA), a general class of methods for approximative inference first popularized in statistical literature by Ormerod and Wand (2012, 2010) in the contexts of Bayesian inference and generalized linear mixed models (GLMM) – a family of statistical models closely related to GLLVMs. The basic idea of variational approximation is to replace a complex optimization problem, possibly involving intractable integrals – such as in (5) – by an easier, approximative one. For GLLVMs (or GLLMs) this can be done by first introducing the so-called *variational lower bound*

$$\begin{aligned}
\ell(\boldsymbol{\Psi}) &= \log \left(\int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) d(\boldsymbol{u}) \right) \\
&= \int q(\boldsymbol{u}) \log \left\{ \frac{f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u})} \right\} d(\boldsymbol{u}) \\
&\quad + \int q(\boldsymbol{u}) \log \left\{ \frac{q(\boldsymbol{u})}{f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})} \right\} d(\boldsymbol{u}) \\
&\geq \int q(\boldsymbol{u}) \log \left\{ \frac{f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u})} \right\} d(\boldsymbol{u}) \triangleq \underline{\ell}(\boldsymbol{\Psi}|q),
\end{aligned} \tag{6}$$

where $q(\boldsymbol{u})$ denotes the density of a *variational distribution* of the latent variables $\boldsymbol{u}$. For a more detailed derivation and explanation of (6), refer to Ormerod and Wand (2010, p.142, 152). Note that the term $\underline{\ell}(\boldsymbol{\Psi}|q)$ in (6) actually coincides with the negative *Kullback-Leibler divergence* (Kullback and Leibler, 1951) between $q(\boldsymbol{u})$ and $f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})$.

For an arbitary variational density q, the lower bound in (6) does not offer much in terms of tractability. Thus, a assumption of belonging to some parametric family of distributions $\{q(\boldsymbol{u}|\boldsymbol{\xi}) : \boldsymbol{\xi} \in \boldsymbol{\Xi}\}$ is often imposed on q. We will further assume this parametric family to be the family of products of multivariate normal densities, that is

$$q(\boldsymbol{u}) = q(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = q_1(\boldsymbol{u}_1) \cdots q_n(\boldsymbol{u}_n),$$

where each $q_i(\boldsymbol{u}_i)$ is the density of multivariate normal distribution governed by the parameters mean $\boldsymbol{a}_i$ and a diagonal covariance matrix $\boldsymbol{A}_i$. Taking this development into consideration, the lower bound in (6) now takes the form of

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q) = \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log \left\{ \frac{f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d(\boldsymbol{u}). \tag{7}$$

Now, by solving the maximization problem

$$(\widehat{\underline{\boldsymbol{\Psi}}}, \widehat{\underline{\boldsymbol{\xi}}}) = \underset{\boldsymbol{\Psi}, \boldsymbol{\xi}}{\operatorname{argmax}}\, \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q),$$

one finds the variational approximation $\widehat{\underline{\boldsymbol{\Psi}}}$ of the maximum likelihood estimate for the model parameter vector $\boldsymbol{\Psi}$. Then standard error estimates for the model parameters can be found from the approximate Fisher information matrix corresponding to the variational log-likelihood $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q)$. The estimates $\widehat{\boldsymbol{a}}_i$ of the variational means $\boldsymbol{a}_i$ corresponding to $\widehat{\underline{\boldsymbol{\xi}}}$ can be used as the variational approximations for the predictive scores of the latent variable effects.

According to Ormerod and Wand (2010), the key advantage of the variational approximation is it's computational efficiency when compared to common alternatives, for example the use of Markov Chain Monte Carlo or Laplace's approximation. When applicable, (7) gives a closed form expression which can then be easily maximized using a standard, readily available optimization software, such as the function optim in R. The authors do note however, that the approximation accuracy of variational approximations is generally more limited than that of MCMC based methods.

Using (7) to derive the exact formulas of variational log-likelihoods for common response types (count, overdispersed count, presence/absence, biomass and such) is not in the scope of this thesis. Instead, a more detailed treatment will be reserved for an approach we call the *extended variational approximation* (EVA), upon which we are going to focus on next. For a more refined overview of the standard variational approximation approach to GLLVMs laid out above, refer to Hui et al. (2017).

2.3 Extended variational approximation

A major disadvantage of the standard VA approach laid out above is the fact that the further derivation of the variational log-likelihood from (7) depends heavily on the assumed distribution of the responses and the link function g in (1). In fact, a tractable, closed form expression might not be available even for some popular response-link combinations, such as in the case of logistic Bernoulli GLLVM, for example. In the case of Bernoulli distributed (i.e. presence/absence) responses, one has to resort to using the probit link function, even if logit or complementary log-log (cloglog) would provide a better description of the data at hand. In order to overcome this issue and to broaden the applicability of VA, we present an approach to GLLVM fitting we now simply call the *extended variational approximation* (EVA). Do note, that a very similar method going by the name *delta method variational inference* has been presented by Wang and Blei (2013).

Derivation of the EVA approach starts by expanding the $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q)$ in (7) in a following way:

$$\begin{aligned}\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q) &= \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log \left\{ \frac{f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d(\boldsymbol{u}) \\ &= \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log \left\{ \frac{f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d(\boldsymbol{u}) \\ &= \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) d(\boldsymbol{u}) + \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log \left\{ \frac{f(\boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d(\boldsymbol{u}). \qquad (8)\end{aligned}$$

Next up, the log-density of the responses, $\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi})$, is approximated by it's quadratic Taylor expansion with respect to the latent variables $\boldsymbol{u}$. The center of the expansion is taken to be the mean of the variational distribution, that is $\boldsymbol{a}$. This approximation gives us that

$$\begin{aligned}\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) \approx{}& \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) + (\boldsymbol{u} - \boldsymbol{a})^\intercal \, \nabla_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) \\ &+ \frac{1}{2} (\boldsymbol{u} - \boldsymbol{a})^\intercal \, \nabla^2_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) \, (\boldsymbol{u} - \boldsymbol{a}). \qquad (9)\end{aligned}$$

Now, by substituting $\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi})$ in (8) by its Taylor expansion (9), we get the

EVA-log-likelihood

$$
\begin{aligned}
\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q) &= \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) d(\boldsymbol{u}) + \int q(\boldsymbol{u}|\boldsymbol{\xi}) \log \left\{ \frac{f(\boldsymbol{u}|\boldsymbol{\Psi})}{q(\boldsymbol{u}|\boldsymbol{\xi})} \right\} d(\boldsymbol{u}) \\
&\approx \int q(\boldsymbol{u}|\boldsymbol{\xi}) \Big\{ \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) + (\boldsymbol{u} - \boldsymbol{a})^\intercal \nabla_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) \\
&\qquad + \frac{1}{2} (\boldsymbol{u} - \boldsymbol{a})^\intercal \nabla^2_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi})\, (\boldsymbol{u} - \boldsymbol{a}) \Big\} d(\boldsymbol{u}) \\
&\quad + \int q(\boldsymbol{u}|\boldsymbol{\xi}) \big\{ \log f(\boldsymbol{u}|\boldsymbol{\Psi}) - \log q(\boldsymbol{u}|\boldsymbol{\xi}) \big\} d(\boldsymbol{u}) \\
&= \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) + \frac{1}{2} \operatorname{Tr}(\nabla^2_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) \boldsymbol{A}) \\
&\quad - D_{\mathrm{KL}}\big(q(\boldsymbol{u}|\boldsymbol{\xi}) \,||\, f(\boldsymbol{u}|\boldsymbol{\Psi}) \big) \triangleq \ell_{\mathrm{EVA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}|q), \qquad (10)
\end{aligned}
$$

where $D_{\mathrm{KL}}\big(q(\boldsymbol{u}|\boldsymbol{\xi}) \,||\, f(\boldsymbol{u}|\boldsymbol{\Psi}) \big)$ is the Kullback-Leibler divergence from the distribution of the latent variables, $f(\boldsymbol{u}|\boldsymbol{\Psi})$, to the variational distribution, $q(\boldsymbol{u}|\boldsymbol{\xi})$. In the derivation of (10) we used the result that for quadratic forms $\boldsymbol{X}^\intercal \boldsymbol{A} \boldsymbol{X}$, where $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{m}$ and $\operatorname{Var}(\boldsymbol{X}) = \boldsymbol{S}$, it holds that $\mathbb{E}[\boldsymbol{X}^\intercal \boldsymbol{A} \boldsymbol{X}] = \operatorname{Tr}(\boldsymbol{A}\boldsymbol{S}) + \boldsymbol{m}^\intercal \boldsymbol{A} \boldsymbol{m}$ (Mathai and Provost, 1992, p.50), together with the fact that $\mathbb{E}_{q(\boldsymbol{u}|\boldsymbol{\xi})}[\boldsymbol{u}] = \boldsymbol{a}$. Additionally, as both of the densities $f(\boldsymbol{u}|\boldsymbol{\Psi})$ and $q(\boldsymbol{u}|\boldsymbol{\xi})$ were assumed to be multivariate normal, we can expand (10) further by using the well-known formula (Duchi, 2007, p.13) for KL-divergence between two p-dimensional multivariate normal distributions, leading us to

$$
\begin{aligned}
\ell_{\mathrm{EVA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) &+ \frac{1}{2} \operatorname{Tr}(\nabla^2_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi}) \boldsymbol{A}) \\
&- \frac{1}{2} \big\{ \operatorname{Tr}(\boldsymbol{A}) + \boldsymbol{a}^\intercal \boldsymbol{a} - p - \log\det(\boldsymbol{A}) \big\}. \qquad (11)
\end{aligned}
$$

Now by solving the maximization problem

$$
(\widehat{\boldsymbol{\Psi}}_{\mathrm{EVA}}, \widehat{\boldsymbol{\xi}}_{\mathrm{EVA}}) = \underset{\boldsymbol{\Psi}, \boldsymbol{\xi}}{\operatorname{argmax}}\, \ell_{\mathrm{EVA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}),
$$

one attains the extended variational approximation $\widehat{\boldsymbol{\Psi}}_{\mathrm{EVA}}$ for the maximum likelihood estimate $\widehat{\boldsymbol{\Psi}}$ of the model parameters.

Wang and Blei (2013) also present two additional options for a reasonable choice of the point around which the quadratic expansion of $\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi})$ is centered. First of these alternatives is to take the center to be the point (in the latent variable space) that maximizes $\log f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi})$. This choice leads to a distinct approximative method the authors call the *Laplace variational inference*. The other of the alternatives is to center the expansion around the mean of the variational distribution from the previous iterative step of the optimization algorithm. However, according to the authors this approach did not often lead to desirable results in terms of the model convergence.

3 EVA for some common GLLVMs

In this section we are going to derive the specific EVA log-likelihoods based on (11), for some of the response distributions and link functions commonly encountered in ecological studies concerned about joint species distribution modeling. Namely, we are going to focus on modelling count data in the case of both Poisson distributed and overdispersed responses, binary data using logit and probit links and finally, percent cover data with Beta distributed responses using logit link.

3.1 Models for count data

The concept of species abundance in ecology is perhaps most often thought of as the amount of individual units of the given species on a specific site. Thus in a way, latent variable models targeting responses that represent counts might well be the most fundamental ones for community ecology. First we are going to look at the case of Poisson GLLVM. Though needlessly simple and often misspecified because of the prevalence of overdispersion, the Poisson model provides an easy-to-follow example of using (11) for the specific model at hand. Up next we will be looking at a negative binomial model, a bit more involved alternative for the Poisson model capable of accounting for overdispersion.

3.1.1 Poisson GLLVM

Assume now that the response variables follow a Poisson distribution, that is, $y_{ij} \sim \text{Poisson}(\mu_{ij})$, where μ_{ij} is such that

$$\log(\mu_{ij}) = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j + \boldsymbol{u}_i^\intercal \boldsymbol{\lambda}_j = \eta_{ij}, \quad i = 1, \dots, n, \ j = 1, \dots, m, \tag{12}$$

and thus $\mu_{ij} = \exp(\eta_{ij})$. Then, the likelihood and the log-likelihood of the response y_{ij} are given as

$$f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp\big(y_{ij}\log(\mu_{ij}) - \mu_{ij}\big)$$

and

$$\log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) = y_{ij}\eta_{ij} - \exp(\eta_{ij}), \tag{13}$$

where constant terms have been omitted. Now, by looking at (11), one sees that the real model specific part of the EVA log-likelihood in need of additional derivation is the term $\nabla^2_{\boldsymbol{u}} \log f(\boldsymbol{y}|\boldsymbol{a}, \boldsymbol{\Psi})$. Now by differentiating $\log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})$ in (13) twice with respect to $\boldsymbol{u}_i$, we get that

$$\frac{\partial \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i} = \big[y_{ij} - \exp(\eta_{ij})\big]\, \boldsymbol{\lambda}_j^\intercal$$

and

$$\frac{\partial^2 \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_i^\intercal} = -\exp(\eta_{ij})\, \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j^\intercal.$$

Next, observe that

$$\mathrm{Tr}\left(\nabla^2_{\boldsymbol{u}_i} \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{a}_i, \boldsymbol{\Psi}) \boldsymbol{A}_i\right) = \mathrm{Tr}\left(-\exp(\eta_{ij}) \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j^\intercal \boldsymbol{A}_i\right) = -\exp(\eta_{ij})\, \boldsymbol{\lambda}_j^\intercal \boldsymbol{A}_i \boldsymbol{\lambda}_j.$$

Finally, by (11), this leads to the following form of the EVA log-likelihood for Poisson GLLVM:

$$\begin{aligned}\ell_{\mathrm{LVA}}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = &\sum_{i=1}^{n} \sum_{j=1}^{m} \left\{ y_{ij} \tilde{\eta}_{ij} - \exp(\tilde{\eta}_{ij}) - \frac{1}{2} \exp(\tilde{\eta}_{ij})\, \boldsymbol{\lambda}_j^\intercal \boldsymbol{A}_i \boldsymbol{\lambda}_j \right\} \\ &- \frac{1}{2} \sum_{i=1}^{n} \{\mathrm{Tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^\intercal \boldsymbol{a}_i - \log\det(\boldsymbol{A}_i)\},\end{aligned}$$

where $\tilde{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j + \boldsymbol{a}_i^\intercal \boldsymbol{\lambda}_j$, that is, the linear predictor η_{ij} evaluated at the variational mean $\boldsymbol{a}_i$.

3.1.2 Negative binomial GLLVM

Let y_{ij} follow the negative binomial distribution governed by mean μ_{ij} and variance $\mu_{ij} + \mu_{ij}^2/\phi_j$, where ϕ_j is a species specific dispersion parameter. Using logarithmic link function yields the same model equation as in (12). The likelihood and the log-likelihood functions given the response y_{ij} are

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

and

$$\log f(y_{ij}|\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{\Psi}) = \log\Gamma(y_{ij}+\phi_j^{-1}) - \log\Gamma(\phi_j^{-1}) + \phi_j^{-1}\log(\phi_j^{-1}) \\ + y_{ij}\eta_{ij} - (y_{ij}+\phi_j^{-1})\log(\mu_{ij}+\phi_j^{-1}),$$

where Γ is the Gamma function and constant terms w.r.t. model parameters and latent variables have been omitted from the log-likelihood. Differentiation w.r.t. $\boldsymbol{u}_i$ gives us

$$\frac{\partial \log f(y_{ij}|\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{\Psi})}{\partial \boldsymbol{u}_i} = \left[y_{ij} - \frac{\phi_j^{-1}+y_{ij}}{\phi_j^{-1}+\mu_{ij}}\mu_{ij} \right]\boldsymbol{\lambda}_j^\intercal,$$

and

$$\frac{\partial^2 \log f(y_{ij}|\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{\Psi})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_i^\intercal} = \left[\frac{\phi_j^{-1}+y_{ij}}{(\phi_j^{-1}+\mu_{ij})^2}\mu_{ij}^2 - \frac{\phi_j^{-1}+y_{ij}}{\phi_j^{-1}+\mu_{ij}}\mu_{ij} \right]\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^\intercal.$$

Putting it all together using (11) we arrive at the following EVA log-likelihood

$$\ell_{\text{LVA}}(\boldsymbol{\Psi},\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\Bigg\{ \log\Gamma(y_{ij}+\phi_j^{-1}) - \log\Gamma(\phi_j^{-1}) + \phi_j^{-1}\log(\phi_j^{-1}) \\ + y_{ij}\tilde{\eta}_{ij} - (y_{ij}+\phi_j^{-1})\log(\tilde{\mu}_{ij}+\phi_j^{-1}) \\ + \frac{1}{2}\left[\frac{\phi_j^{-1}+y_{ij}}{(\phi_j^{-1}+\tilde{\mu}_{ij})^2}\tilde{\mu}_{ij}^2 - \frac{\phi_j^{-1}+y_{ij}}{\phi_j^{-1}+\tilde{\mu}_{ij}}\tilde{\mu}_{ij} \right]\boldsymbol{\lambda}_j^\intercal\boldsymbol{A}_i\boldsymbol{\lambda}_j \Bigg\} \\ - \frac{1}{2}\sum_{i=1}^{n}\big\{ \text{Tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^\intercal\boldsymbol{a}_i - \log\det(\boldsymbol{A}_i)\big\}, \tag{14}$$

where $\tilde{\eta}_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\intercal\boldsymbol{\beta}_j + \boldsymbol{a}_i^\intercal\boldsymbol{\lambda}_j$ and $\tilde{\mu}_{ij} = \exp(\tilde{\eta}_{ij})$.

The advantage of the negative binomial model when compared to the Poisson model laid out before is the fact that it is able to better manage overdispersion. Assumptions of the Poisson model say that $\text{Var}(y_{ij}) = \mu_{ij} = \mathbb{E}[y_{ij}]$, which is often not the case. Overdispersion is diagnosed, when the observed variance of the responses is greater than the observed mean. The additional dispersion parameters ϕ_j present in the negative binomial model and the assumption that $\text{Var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi_j$ give the model a better chance to adjust to the situation. However, having more parameters to be estimated from the data often translates to a greater computational load and possibly some decreased modeling accuracy in the cases where overdispersion is actually not present.

3.2 Models for binary data

In the context of community ecology, binary abundances $y_{ij} \in \{0,1\}$ can be thought of as indicators of either presence or absence of the species j at the test site i. As typically in statistics, the responses y_{ij} are now assumed to follow a Bernoulli distribution with probability or mean $\mu_{ij} = g^{-1}(\eta_{ij})$, which gives us

$$f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \mu_{ij}^{y_{ij}}(1-\mu_{ij})^{1-y_{ij}}$$

and

$$\log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) = y_{ij}\log(\mu_{ij}) + (1-y_{ij})\log(1-\mu_{ij}). \tag{15}$$

The steps of the derivation to follow differ by the choice of the link function g used. Here we are going to look at two cases in particular, the one where g is taken to be the logit function and the other in which it is taken to be the probit function.

3.2.1 Bernoulli logit GLLVM

Assume the relationship between the probability μ_{ij} and the linear predictor η_{ij} to be governed by the logit link function. Then

$$\operatorname{logit}(\mu_{ij}) = \log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} = \alpha_i + \beta_{0j} + \boldsymbol{x}_i^\intercal \boldsymbol{\beta}_j + \boldsymbol{u}_i^\intercal \boldsymbol{\lambda}_j$$

and

$$\mu_{ij} = \operatorname{logit}^{-1}(\eta_{ij}) = \frac{\exp(\eta_{ij})}{\exp(\eta_{ij})+1}.$$

Then (15) takes the form of

$$\begin{aligned}\log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) &= y_{ij}\log(\mu_{ij}) + (1-y_{ij})\log(1-\mu_{ij})\\ &= y_{ij}\log\left(\frac{\exp(\eta_{ij})}{\exp(\eta_{ij})+1}\right) + (1-y_{ij})\log\left(\frac{1}{\exp(\eta_{ij})+1}\right)\\ &= y_{ij}\eta_{ij} - \log\big(\exp(\eta_{ij})+1\big).\end{aligned}$$

Differentiating w.r.t. $\boldsymbol{u}_i$ gives

$$\frac{\partial \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i} = \left[y_{ij} - \frac{\exp(\eta_{ij})}{\exp(\eta_{ij}) + 1} \right] \boldsymbol{\lambda}_j^\intercal,$$

and

$$\frac{\partial^2 \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_i^\intercal} = - \left[\frac{\exp(\eta_{ij})}{\big(\exp(\eta_{ij}) + 1\big)^2} \right] \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j^\intercal,$$

which leads us to the following expression for EVA log-likelihood based on (11):

$$\begin{aligned} \ell_{EVA}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{m} & \left\{ y_{ij} \tilde{\eta}_{ij} - \log\big(\exp(\tilde{\eta}_{ij}) + 1\big) - \frac{\exp(\tilde{\eta}_{ij})}{2\big(1 + \exp(\tilde{\eta}_{ij})\big)^2} \boldsymbol{\lambda}_j^\intercal \boldsymbol{A}_i \boldsymbol{\lambda}_j \right\} \\ & - \frac{1}{2} \sum_{i=1}^{n} \{ \mathrm{Tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^\intercal \boldsymbol{a}_i - \log\det(\boldsymbol{A}_i) \}. \end{aligned} \tag{16}$$

The logistic Bernoulli GLLVM is a good example of a situation where the ordinary VA approach, as described in section 2.2, fails to provide a tractable closed form approximation of the log-likelihood function. Thus the use of EVA to fit the model is very reasonable. The method of Laplace approximation (LA, Tierney and Kadane, 1986) can however also be used in the case of logit model and thus one of the simulation studies described in Section 4.1 is concerned with comparing the performance of EVA and LA in this particular modeling scenario.

3.2.2 Bernoulli probit GLLVM

According to the probit model, the relationship between μ_{ij} and the linear predictor η_{ij} is assumed to be

$$\mathrm{probit}(\mu_{ij}) = \Phi^{-1}(\mu_{ij}) = \eta_{ij}, \tag{17}$$

where Φ is the cumulative distribution function of the standard normal distribution. Now (15) becomes

$$\begin{aligned} \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) &= y_{ij} \log(\mu_{ij}) + (1 - y_{ij}) \log(1 - \mu_{ij}) \\ &= y_{ij} \log\big(\Phi(\eta_{ij})\big) + (1 - y_{ij}) \log\big(1 - \Phi(\eta_{ij})\big). \end{aligned}$$

Then,

$$
\begin{aligned}
\frac{\partial \log f(y_{ij}|\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{\Psi})}{\partial \boldsymbol{u}_i} &= \left[y_{ij}\frac{\phi(\eta_{ij})}{\Phi(\eta_{ij})} - (1-y_{ij})\frac{\phi(\eta_{ij})}{1-\Phi(\eta_{ij})} \right] \boldsymbol{\lambda}_j^{\mathsf{T}} \\
\frac{\partial^2 \log f(y_{ij}|\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{\Psi})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_i^{\mathsf{T}}} &= \left[y_{ij}\frac{\phi'(\eta_{ij})\Phi(\eta_{ij}) - (\phi(\eta_{ij}))^2}{(\Phi(\eta_{ij}))^2} \right] \boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^{\mathsf{T}} \\
&\quad - \left[(1-y_{ij})\frac{\phi'(\eta_{ij})(1-\Phi(\eta_{ij})) + (\phi(\eta_{ij}))^2}{(1-\Phi(\eta_{ij}))^2} \right] \boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^{\mathsf{T}},
\end{aligned}
$$

where $\phi(\eta_{ij})$ denotes the p.d.f. of the standard normal distribution evaluated at η_{ij}, and $\phi'(\eta_{ij})$ denotes the first derivative of the said p.d.f. evaluated at η_{ij}. This gives us the following EVA log-likelihood (11)

$$
\begin{aligned}
\ell_{EVA}(\boldsymbol{\Psi},\boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \Bigg\{ & y_{ij}\log\big(\Phi(\eta_{ij})\big) + (1-y_{ij})\log\big(1-\Phi(\eta_{ij})\big) \\
& + \frac{1}{2}\left[y_{ij}\frac{\phi'(\eta_{ij})\Phi(\eta_{ij}) - (\phi(\eta_{ij}))^2}{(\Phi(\eta_{ij}))^2} \right] \boldsymbol{\lambda}_j^{\mathsf{T}}\boldsymbol{A}_i\boldsymbol{\lambda}_j \\
& - \frac{1}{2}\left[(1-y_{ij})\frac{\phi'(\eta_{ij})(1-\Phi(\eta_{ij})) + (\phi(\eta_{ij}))^2}{(1-\Phi(\eta_{ij}))^2} \right] \boldsymbol{\lambda}_j^{\mathsf{T}}\boldsymbol{A}_i\boldsymbol{\lambda}_j \Bigg\} \\
& - \frac{1}{2}\sum_{i=1}^{n}\big\{ \operatorname{Tr}(\boldsymbol{A}_i) + \boldsymbol{a}_i^{\mathsf{T}}\boldsymbol{a}_i - \log\det(\boldsymbol{A}_i)\big\}.
\end{aligned} \tag{18}
$$

The probit model (18), unlike the preceeding logit model (16), is a type of a GLLVM where the ordinary VA approach can be applied in the model estimation process. It remains reasonable however to assess the comparative performance of these two types of variational inference -based methods, as the results can bring valuable insight about the accuracy and the efficiency of the EVA approach as a whole. Thus two of the three distinct simulation studies conducted as a part of this thesis involve comparing VA, LA and EVA approaches against each other in the case of the Bernoulli-probit GLLVM.

3.3 Beta model for percent cover data

Assume now that the responses y_{ij} represent the percent cover of the (plant) species j at site i. That is, $100 \times y_{ij}$ tells the percentage that the species j covers of the total area observed at site i. Then, y_{ij} is continuous and constrained to the closed unit interval $[0, 1]$. If however, the endpoints can be excluded, meaning that no species can be assumed to cover all (or none) of the observable area at any site, then a multivariate Beta model can be specified, $y_{ij} \sim \text{Beta}(\mu_{ij}, \phi_j)$. That is, to assume that the responses y_{ij} follow a Beta distribution with mean μ_{ij} and variance $\mu_{ij}(1-\mu_{ij})/(1+\phi_j)$, where ϕ_j is a dispersion parameter specific to species j. Then, the responses have the following log-density:

$$\begin{aligned}\log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi}) &= \log\Gamma(\phi_j) - \log\Gamma(\mu_{ij}\phi_j) - \log\Gamma\big((1-\mu_{i,j})\phi_j\big) \\ &\quad + (\mu_{ij}\phi_j - 1)\log(y_{i,j}) + \big((1-\mu_{ij})\phi_j - 1\big)\log(1-y_{ij}). \quad (19)\end{aligned}$$

By assuming the logit link function, the relationship between μ_{ij} and η_{ij} is

$$\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = \eta_{ij} \implies \mu_{ij} = \frac{\exp(\eta_{ij})}{\exp(\eta_{ij})+1}$$

Differentation w.r.t. $\boldsymbol{u}_i$ gives

$$\begin{aligned}&\frac{\partial \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i} \\ &= \left[-\psi(\mu_{ij}\phi_j)\phi_j\mu'_{ij} + \psi\big((1-\mu_{ij})\phi_j\big)\phi_j\mu'_{ij} + \phi_j\mu'_{ij}\log\left(\frac{y_{ij}}{1-y_{ij}}\right)\right]\boldsymbol{\lambda}_j^\intercal,\end{aligned}$$

where ψ is the digamma function, $\psi(x) = \frac{d}{dx}\log\big(\Gamma(x)\big)$ and μ'_{ij}, is a shorthand notation for $\mu'(\eta_{ij}) = \frac{\exp(\eta_{ij})}{(\exp(\eta_{ij})+1)^2}$. Further differentiation gives us

$$\begin{aligned}\frac{\partial^2 \log f(y_{ij}|\boldsymbol{x}_i, \boldsymbol{u}_i, \boldsymbol{\Psi})}{\partial \boldsymbol{u}_i \partial \boldsymbol{u}_i^\intercal} &= \big[-\psi_1(\mu_{ij}\phi_j)\phi_j^2(\mu'_{ij})^2 - \phi_j\mu''_{ij}\psi(\mu_{ij}\phi_j)\big]\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^\intercal \\ &\quad + \big[-\psi_1\big((1-\mu_{ij})\phi_j\big)\phi_j(\mu'_{ij})^2 + \psi((1-\mu_{ij})\phi_j)\phi_j\mu''_{ij}\big]\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^\intercal \\ &\quad + \big[\phi_j\mu''_{ij}\big(\log(y_{ij}) - \log(1-y_{ij})\big)\big]\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^\intercal, \qquad (20)\end{aligned}$$

where ψ_1 is the trigamma function, $\psi_1(x) = \frac{d}{dx}\psi(x)$ and μ''_{ij} is a shorthand for $\mu''(\eta_{ij}) = \frac{\exp(3\eta_{ij})-\exp(\eta_{ij})}{(\exp(\eta_{ij})+1)^4}$. As usual, by plugging (19) and (20) into (11), one gets the desired approximate log-likelihood function.

3.4 Notes about computer implementations

As mentioned, the use of both the standard method of variational approximation and the method of Laplace approximation is possible when fitting GLLVMs using the package gllvm (Niku et al., 2019b). The extended variational approximation approach was implemented as part of this thesis. The implementation was done partly in R and partly in C++ using the package TMB (Kristensen et al., 2016). First, the negative EVA log-likelihoods were written in C++. Then, by using TMB, the negative log-likelihoods were compiled, resulting in R object containing both the negative log-likelihood function and it's gradient function, attained by the technique of automatic differentiation. Lastly, the negative log-likelihood and the gradient were passed to some optimization procedure readily available in R – such as the function optim – leading to the approximative maximum likelihood estimates and predictions for the latent variable values.

4 Simulation studies

In this section, we present both the simulation settings used and the results and conclusions that followed.

4.1 Simulation settings

The main purpose of the method of extended variational approximation was to broaden the class of the specific models for which a variational approximation type approach could be performed. This is clearly achieved by the method of EVA as described in section 2.3. What remains unknown however is the actual performance of EVA, in terms of both computational efficiency and estimation accuracy, when compared to orher methods of approximative inference. Poor performance in either efficiency or accuracy could render the approach useless in practice. Thus, three distinct simulation studies were conducted as part of this thesis, comparing EVA against the method of standard VA and the method of LA, both readily implemented in the package gllvm (Niku et al., 2019b). The simulation setups to be presented next essentially follow the ones used previously in Niku et al. (2019a).

4.1.1 Simulation study #1: amoeba data

The first study involves all three approaches VA, LA and EVA in the case of both count and binary responses simulated based on a data set analysed in Secco et al. (2016). The data set contains counts of 48 different amoeba species measured from 263 test sites. Additionally, the data set contained measurements from two environmental covariates, totaling to 263 observations for both. First, four sets of row indices were generated in the following way; the first index set was randomly drawn sample of 50 row indices out of the 263 sites of the full data. The second set was then created by augmenting the first one by 70 additional randomly drawn indices. The third and fourth sets further added 70 additional row indices each, thus resulting to four sets of indices corresponding to $n = 50, 120, 190$ and 260 rows of the original data. Then, a negative binomial GLLVM with logarithmic link function (defined in Section 3.1.2), was fitted to the full data, using the VA approach. Row effects α_i were not included. The parameter estimates of the full model were then used as the 'true parameter values'. Thousand sets of simulated responses were generated for each of the four index sets, using the parameter estimates from the full model corresponding to the given indices. This resulted in to thousand simulated sets of count responses for each number n of test sites $n = 50, 120, 190, 260$. The number of species was kept as a constant, $m = 48$, for all of the total of 4000 simulations.

The purpose of this study was to observe how the methods of VA, LA and EVA perform in the case of the negative binomial latent variable model as the number of test sites, n increases. Thus, all three of the methods VA, LA and EVA were used to fit a negative binomial GLLVM on the 1000 simulated sets of responses for each value of n. On each iteration, the same set of initial values was used for all methods. The initial values were attained by first fitting a multi-response negative binomial GLM on the simulated data set at question, from which the initial values for the regression and dispersion parameters were attained. Then, a standard factor analysis procedure was carried out on the GLM residuals, giving initial values for the latent variable scores and loadings. This method of initializing the model parameters is readily available in the package gllvm (Niku et al., 2019b).

The methods were compared in the terms of average bias and root mean squared error (RMSE) regarding the estimates of regression coefficients $(\beta_{0j}, \beta_{1j}, \beta_{2j})^\intercal$ and the dispersion parameters ϕ_j. RMSE of a model parameter estimate $\hat{\theta}$ is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\theta - \hat{\theta}\right)^2},$$

where θ is the corresponding true parameter value and N is the amount of simulation runs.

Total computation times for each set of 1000 runs were also recorded for each approach and compared. Additionally, the *Procrustes error* of both the latent variable predictions and the loading estimates were compared. By Peres-Neto and Jackson (2001), the Procrustes error of the latent variable predictions is defined as

$$\text{Procrustes error} = \sum_{i=1}^{n}\sum_{r=1}^{p}\left(\hat{u}_{ir} - u_{ir}\right)^2,$$

where $\hat{u}_{ir}$ is the *Procrustes-rotated* prediction of the value of the rth latent variable for site i; with u_{ir} being the corresponding true coordinate value. The Procrustes-rotation re-scales the fitted and the true latent variable scores to common size and then successively rotates the fitted scores until the sum of the squared residuals between the fitted values $\hat{u}_{ir}$ and true the values u_{ir} – the Procrustes error – is minimized. Similarly, the Procrustes error can also be calculated for the latent variable loading estimates $\hat{\lambda}_{ij}$, as was done as a part of the studies..

In addition to the case of negative binomial model, the simulation procedure as outlined above was also carried out in the case of Bernoulli probit GLLVM (17) simply by transforming the count valued responses in the original amoeba data into binary indicators of presence and absence. Meaning that responses, for which $y_{ij} > 0$ were transformed in to 1. Comparison were done using the same metrics as in the case of negative binomial model. The beta GLLVM discussed in Section 3.3 does not possess a readily available implementation in the gllvm package using either of the LA or VA approaches, and thus it was not included in the simulation studies.

4.1.2 Simulation study #2: bird data

The second simulation setting mirrored the first one using a bird data set analysed in Cleary et al. (2005) as it's basis. Here we fix the amount of sites $n = 37$, and study the effect of the amount of the species m on the performance of the three approaches, when fitting a negative binomial or binary probit GLLVM. Thus, four sets of column indices were generated in the same manner as the sets of row indices in the first study, leading to thousand sets of simulated responses for each of the four increasing column index sets, with $m = 30, 60, 100, 140$. The original data set had observations from 177 species, of which 29 were considered to be rare enough (observed at most on 2 of the 37 test sites) to not be involved in the studies, reducing the total amount of species to 148. Unlike in the case of the amoeba data, the simulations carried out with the bird data did not involve any environmental covariates. The metrics of comparisons were the same as in the first study: average bias and RMSE for the species specific intercepts β_{0j} and dispersions ϕ_j (in the case of the negative binomial model), and average Procrustes error for the latent variable scores u_{ir} and loadings λ_{ij}. The same procedure of model parameter initialization was used as in the first simulation study.

4.1.3 Simulation study #3: logit model

The aim of the third simulation study was to compare EVA to LA in a modelling situation, where the standard VA could not be applied. A logit latent variable model for Bernoulli distributed responses, as discussed in Section 3.2.1, represents one such situation. Both the amoeba and bird data were used, similarly as in the cases of probit models in the two studies discussed above, leading to eight sets of

1000 simulation runs for both methods. The true model, on which the simulations were based on, was fitted using the LA approach for both the amoeba and the bird data. The metrics of comparison were the same as in the previous studies, that is, average computation time, bias and RMSE of the regression parameters and the Procrustes errors of both the latent variable scores and loadings.

4.2 Results

In this section, we present the results of the simulation studies described previously in section 4.1.

4.2.1 Simulation study #1

The first simulation study was concerned with the comparative performance of the EVA method, when the amount of samples, or sites, n increases. The four values used for the sample size were $n_1 = 50$, $n_2 = 120$, $n_3 = 190$ and $n_4 = 260$. The amount of species was kept as constant, $m = 48$. The study was conducted for both a negative binomial count data GLLVM and a binary data probit GLLVM.

4.2.1.1 Negative binomial model. The proportion of negative and finite log-likelihoods, denoted by ℓ_0, was quite low for the method LA in the case of the two of the largest sample sizes n_3 and n_4, meaning that the method of LA was quite prone to converging on to a local maximum. The average computational times, biases, RMSEs and Procrustes errors reported in Table 1 were calculated using only the results from model fits on which the log-likelihood attained a proper value. This decision was made so that the instability of the specific implementation would not influence the results on which the actual methods were compared, by too large of a margin, as tge low ℓ_0 of the method LA was suspected to be related to some issue of sensitivity on initial values. Additionally, a trimming factor of 2% was used when calculating the average bias and RMSE, meaning that the the most extreme 2% of the values were omitted, reducing the effect of possible outliers.

In terms of computational efficiency, the methods of VA and EVA seemed to perform quite evenly, while the average computation time $\hat{\mathrm{t}}$ was much higher for LA. The ranges for computation time (in seconds) were $(1.41, 9.67)$, $(1.62, 5.94)$ and $(11.26, 75.42)$ for EVA, VA and LA, respectively, in the case of the smallest sample size n_1, and $(6.37, 49.55)$, $(8.75, 44.96)$ and $(68.75, 226.18)$ and for the largest sample size n_4, when only the 'good fits' were considered. The mean computation times are reported in Table 1.

In terms of the regression parameter estimation accuracy, VA was the winner in all cases, producing the lowest average biases and RMSEs, as can be seen from Table 1. When measuring by the Procrustes error of the latent variable scores, all three methods performed quite equally. On average, both VA and LA were a bit more accurate than EVA in the estimation of the latent variable loadings. Each method managed to improve in accuracy on all cases as the sample size n increased.

Overall, the method VA can be considered as the best performing approach both in terms of speed and accuracy. In terms of estimation accuracy, the method LA takes the second place, with EVA coming in quite close. However, as EVA attained a much higher proportion of proper fits ℓ_0 and held on closely to the good computational speed of VA, the method EVA can be reasonably considered to have been the second best approach overall.

Table 1: Summary of the amoeba data based simulation study involving a negative binomial GLLVM. The column ℓ_0 marks the proportion of 'proper fits', indicated by finite and negative log-likelihood value. The computation time $\hat{t}$, biases, RMSEs and Procrustes errors are reported as averages calculated using only the proper fits. In addition, a trimming factor of 2% was used when calculating the average biases and RMSEs. Sample sizes were $n_1 = 50$, $n_2 = 120$, $n_3 = 190$ and $n_4 = 260$. Amount of species was kept constant, $m = 48$.

				Bias				RMSE				PE	
		$\hat{t}$	ℓ_0	β_0	β_1	β_2	ϕ	β_0	β_1	β_2	ϕ	u	λ
n_1	EVA	3.50	0.996	-1.404	0.016	-0.150	-4.745	2.320	0.982	0.987	5.518	0.266	0.508
	VA	3.46	1.000	-0.401	-0.073	-0.079	-3.651	0.834	0.658	0.669	4.559	0.230	0.316
	LA	23.21	0.964	-1.307	0.010	-0.139	-4.735	2.137	0.927	0.946	5.511	0.269	0.487
n_2	EVA	8.92	1.000	-0.583	0.014	-0.047	-2.952	1.072	0.493	0.476	4.097	0.194	0.328
	VA	7.86	1.000	-0.113	-0.024	-0.026	-1.630	0.458	0.380	0.368	3.410	0.186	0.186
	LA	52.64	0.949	-0.545	0.013	-0.045	-2.954	0.990	0.476	0.463	4.115	0.198	0.303
n_3	EVA	14.72	1.000	-0.359	-0.011	-0.022	-2.171	0.722	0.340	0.350	3.440	0.187	0.239
	VA	14.45	1.000	-0.032	-0.033	-0.010	-0.814	0.347	0.296	0.283	2.978	0.185	0.139
	LA	91.10	0.399	-0.327	-0.013	-0.013	-2.129	0.655	0.332	0.339	3.460	0.189	0.212
n_4	EVA	17.95	1.000	-0.258	-0.016	-0.009	-1.681	0.552	0.293	0.287	2.997	0.177	0.183
	VA	19.93	1.000	0.012	-0.037	-0.002	-0.267	0.291	0.260	0.243	2.729	0.176	0.109
	LA	134.46	0.215	-0.257	-0.018	-0.009	-1.775	0.543	0.287	0.283	3.075	0.179	0.171

4.2.1.2 Bernoulli probit model. Contrary to the case of the negative binomial model, all fits acquired from the amoeba data based binary response probit model simulation study had a finite and non-negative log-likelihood value. Thus the column ℓ_0 was dropped from the summary in Table 2. The quantities in the summary were reported as averages, with a 2% trimming factor used in the calculation of bias and RMSE.

As can be seen from Table 2, the method EVA was, on average, the most computationally efficient among the three approaches. The method LA was again vastly slower than either VA or EVA in all cases, though interestingly the average computation time of LA generally decreased as the sample size n increased.

The method VA was again the approach producing the most accurate results, on average. The method LA struggled with bad accuracy in the case of the smallest sample size $n_1 = 50$, particularly with the estimation of the intercept parameter β_0 and the latent variable loadings λ. However, the accuracy of all methods generally improved as the sample size increased, and the results provided by the methods EVA and LA were quite indistinguishable in the cases of two of the largest sample sizes.

Based on these results, the method VA should be considered as the default choice when choosing to fit a Bernoulli probit GLLVM on a binary presence/absence abundance data, unless a strict computational efficiency is considered as only the decisive factor, in which case the use of EVA approach could perhaps be reasoned. If for some reason only EVA or LA are considered and the sample size n is on the smaller size, then EVA could be the better choice.

Table 2: Summary of the amoeba data based simulation study involving a binary response probit GLLVM. The computation time $\hat{t}$, biases, RMSEs and Procrustes errors are reported as averages. In addition, a trimming factor of 2% was used when calculating the average biases and RMSEs. Sample sizes were $n_1 = 50$, $n_2 = 120$, $n_3 = 190$ and $n_4 = 260$. Amount of species was kept constant, $m = 48$.

			Bias			RMSE			PE	
		$\hat{t}$	β_0	β_1	β_2	β_0	β_1	β_2	u	λ
n_1	EVA	0.74	-0.312	-0.056	-0.043	0.719	0.465	0.418	0.353	0.489
	VA	1.86	-0.276	-0.052	-0.053	1.063	0.374	0.354	0.269	0.285
	LA	93.86	-18.457	0.943	0.820	42.267	16.119	12.244	0.319	0.870
n_2	EVA	2.04	-0.235	0.020	0.010	0.546	0.297	0.253	0.276	0.417
	VA	3.69	-0.027	-0.004	-0.013	0.224	0.197	0.180	0.242	0.146
	LA	104.73	-2.834	0.521	0.172	8.174	2.576	1.862	0.264	0.793
n_3	EVA	5.25	-0.187	0.020	0.019	0.449	0.210	0.199	0.272	0.345
	VA	8.51	-0.004	-0.001	0.001	0.161	0.146	0.142	0.250	0.111
	LA	87.95	-0.691	0.078	0.035	2.235	0.664	0.495	0.269	0.545
n_4	EVA	11.41	-0.122	0.002	0.007	0.331	0.172	0.153	0.247	0.247
	VA	15.39	0.016	-0.014	-0.007	0.133	0.127	0.119	0.233	0.082
	LA	83.75	-0.247	0.013	0.009	0.889	0.281	0.234	0.245	0.313

4.2.2 Simulation study #2

The second simulation study was based on bird data set, where $m \gg n$. That is, the amount of species was larger than the amount of sites. The aim of the study was then to assess how the performance of EVA compares to that of standard VA and LA when the amount of sites is kept as constant and the amount of species is increased gradually. The four values used for the amount of species were $m_1 = 40$, $m_2 = 60$, $m_3 = 100$ and $m_4 = 140$. The amount of test sites was $n = 37$ in all cases. The study setting was conducted for both a negative binomial and a Bernoulli probit GLLVM.

4.2.2.1 Negative binomial model. As in the case of the amoeba data based study, the negative binomial model had a tendency to converge in to an improper solution when using the methods of LA or EVA. For EVA, the occurrence of these incidents was quite minor. The proportion of proper fits, ℓ_0, can be seen from Table 3. The other quantities in the summary are reported based on only the negative and finite fits. A trimming factor of 2% was again used when calculating the average biases and RMSEs.

As can be seen from Table 3, the average computation times followed the already established trend, with LA being clearly the slowest method among the three, while EVA and VA attained fairly even speeds. EVA attained a bit lower average computation time than VA in the cases of two of the smallest values for amount of species, m_1 and m_2, with the situtation being reversed in the case of the largest value m_4.

In terms of both average bias and RMSE, the standard VA was clearly the most accurate approach, with LA placing second by a small margin over EVA. Generally, the estimates got slightly more imprecise, as the amount of species increased. When, measured by the Procrustes error of latent variable scores, the three methods performed almost indistinguishably. The differences were more drastic in the Procrustes error of latent variable loadings, with VA achieving the lowest error for all values of m, and LA coming in at second place with slightly lower errors than EVA.

The method VA can again be considered as a good default choice for the given situation, based on these results. Between the methods of EVA and LA, the choice comes down to whether to prioritise speed and reliability (higher ℓ_0) over a slightly more accurate estimation, or vice versa.

Table 3: Summary table of the bird data based simulation study involving a negative binomial GLLVM. The column ℓ_0 marks the proportion of 'proper fits', indicated by finite and negative log-likelihood value. The computation time $\hat{t}$, biases, RMSEs and Procrustes errors are reported as averages calculated using only the proper fits. In addition, a trimming factor of 2% was used when calculating the average biases and RMSEs. The amount of sites was kept as constant, $n = 37$, while the amounts of species varied, $m_1 = 40$, $m_2 = 60$, $m_3 = 100$ and $m_4 = 140$.

				Bias		RMSE		PE	
		$\hat{t}$	ℓ_0	β_0	ϕ	β_0	ϕ	u	λ
m_1	EVA	1.00	0.981	-0.301	-0.710	0.721	1.413	0.519	0.567
	VA	1.61	1.000	0.072	0.196	0.378	1.360	0.493	0.558
	LA	8.31	0.871	-0.264	-0.694	0.654	1.410	0.513	0.559
m_2	EVA	3.29	0.991	-0.304	-0.598	0.693	1.244	0.238	0.469
	VA	3.99	1.000	-0.080	-0.140	0.384	1.100	0.214	0.388
	LA	25.16	0.916	-0.293	-0.599	0.663	1.240	0.234	0.457
m_3	EVA	6.04	0.999	-0.347	-0.642	0.784	1.308	0.122	0.459
	VA	6.01	1.000	-0.142	-0.344	0.458	1.144	0.113	0.351
	LA	47.55	0.855	-0.334	-0.642	0.751	1.304	0.122	0.449
m_4	EVA	9.90	1.000	-0.370	-0.656	0.810	1.279	0.086	0.478
	VA	7.56	1.000	-0.188	-0.434	0.497	1.133	0.081	0.349
	LA	74.62	0.869	-0.366	-0.655	0.799	1.280	0.086	0.465

4.2.2.2 Bernoulli probit model. A summary of the results from the bird data experiment regarding a Bernoulli probit GLLVM can be seen from Table 4. EVA was on average the fastest method among the three across all values of m. At worst, LA produced average computation times higher by two orders of magnitude, when compared to both VA and EVA. In terms of the average bias and RMSE regarding the intercept parameter β_0, the results produced by LA again fell short compared to that of EVA or VA. The methods EVA and VA managed to be quite competitive, with VA being more accurate on all but the largest value of m. As m increased, the RMSE increased for VA, and decreased for EVA. Similar development can also be seen for bias.

EVA had the highest Procrustes errors of latent variable scores, with VA and LA attaining almost equal precision on two of the largest values of m. LA had clearly the most trouble among the three in estimation of the latent variable loadings, as can be seen from Table 4.

Judging by these results, VA once again gets to be considered the most reasonable default choice when choosing to fit a type of model in question. The method of LA would be very hard to recommended in any such situation. Meanwhile, it would be interesting to assess, that whether the EVA could keep on improving in accuracy over the VA approach, if the ratio of the amount of species to the amount of sites would be increased even further.

Table 4: Summary table of the bird data based simulation study involving a binary response probit GLLVM. The computation time $\hat{t}$, bias, RMSE and Procrustes errors are reported as averages calculated. In addition, a trimming factor of 2% was used when calculating the average bias and RMSE. The amount of sites was kept as constant, $n = 37$, while the amounts of species varied, $m_1 = 40$, $m_2 = 60$, $m_3 = 100$ and $m_4 = 140$.

			Bias	RMSE	PE	
		$\hat{t}$	β_0	β_0	u	λ
m_1	EVA	0.27	-0.047	0.539	0.588	0.676
	VA	1.03	0.049	0.283	0.492	0.539
	LA	37.34	-6.778	33.478	0.557	0.848
m_2	EVA	0.80	-0.049	0.482	0.334	0.538
	VA	1.51	-0.012	0.326	0.236	0.395
	LA	59.03	-2.568	46.035	0.271	0.875
m_3	EVA	1.63	-0.050	0.441	0.200	0.426
	VA	2.22	-0.037	0.375	0.140	0.333
	LA	102.05	-2.038	35.452	0.158	0.890
m_4	EVA	2.86	-0.028	0.384	0.170	0.399
	VA	3.29	-0.050	0.412	0.094	0.318
	LA	152.92	-0.821	24.403	0.106	0.891

4.2.3 Simulation study # 3

The third simulation study used both the amoeba and the bird data sets to compare EVA against LA in the estimation of a logistic GLLVM for Bernoulli distributed responses. The true model was fitted using the LA approach.

4.2.3.1 Amoeba data. As with all of the previous cases, EVA was again considerably faster method to fit the model, than LA. Table 5 contains a summary of the results of the amoeba data based simulations. Noticeably, there is a stark contrast between methods when it comes to the accuracy in terms of the regression coefficients, in favor of LA. In general, both of the methods struggled with the regression coefficients, particularly with the intercept β_0. The average biases and RMSEs produced were much higher than in any of the previous studies. This raises concerns about possible sensitivity issues relating to the choice of the method used to fit the true model on which the simulations were based on. For increased generality, in the future, the same simulation setting should be conducted using EVA as the approach in estimation of the true model.

The Procrustes errors of latent variable scores and loadings were rather more in line with previous simulation studies. LA was in general the more accurate one of the methods. Neither of the methods showed noticeable improvement in accuracy relating to the latent variables when the sample size n was increased.

Based solely on this evidence, the use of LA would be more favorable when fitting a logistic GLLVM on binary response data, where the amount of sites n exceeds the amount of species m. Both methods performed quite poorly overall, especially with the smaller values of n. Replication of the study using EVA in the true model fitting would be a sensible next step in future developments.

Table 5: Summary of the amoeba data based simulation study involving a binary response logistic GLLVM. The computation time $\hat{t}$, biases, RMSEs and Procrustes errors are reported as averages calculated. In addition, a trimming factor of 2% was used when calculating the average biases and RMSEs. Sample sizes were $n_1 = 50$, $n_2 = 120$, $n_3 = 190$ and $n_4 = 260$. Amount of species was kept constant, $m = 48$.

			Bias			RMSE			PE	
		$\hat{t}$	β_0	β_1	β_2	β_0	β_1	β_2	u	λ
n_1	EVA	3.25	-44.112	0.974	2.946	77.446	30.527	24.392	0.280	0.704
	LA	41.45	-51.493	1.588	2.227	86.491	34.308	27.043	0.282	0.697
n_2	EVA	9.90	-21.110	4.865	3.172	38.620	13.500	10.480	0.238	0.754
	LA	98.49	-19.600	4.809	2.352	37.177	12.442	9.091	0.233	0.774
n_3	EVA	20.63	-13.511	4.277	2.187	23.417	7.732	6.733	0.240	0.768
	LA	127.47	-9.273	2.552	0.967	17.703	5.460	4.135	0.235	0.727
n_4	EVA	35.89	-9.647	3.920	1.494	16.746	6.324	4.388	0.221	0.785
	LA	165.00	-4.996	1.550	0.526	10.327	3.448	2.126	0.217	0.661

4.2.3.2 Bird data. When fitting a logistic GLLVM on simulations based on the binary response bird data, EVA was much faster method than LA, on average, as can be seen from the summary in Table 6. The differences between the methods in the average bias and RMSE of the intercept β_0 were not as stark as in the case of the amoeba data. Either way, the bias and the RMSE again seemed rather bloated when compared to the results from the first and second simulation study. This might be another indicator of the need to replicate the study using EVA as baseline for the true model, and compare the results. Based only on these results, LA seems more accurate in the estimation of the intercept β_0. A similar story holds for the latent variables, as LA attained, in general, lower average Procrustes errors. Thus essentially, the choice between the methods depends on which is deemed more important in a given situation, computational efficiency (EVA), or estimation accuracy (LA).

Table 6: Summary table of the bird data based simulation study involving a binary response logistic GLLVM. The computation time $\hat{t}$, bias, RMSE and Procrustes errors are reported as averages calculated. In addition, a trimming factor of 2% was used when calculating the average bias and RMSE. The amount of sites was kept as constant, $n = 37$, while the amounts of species varied, $m_1 = 40$, $m_2 = 60$, $m_3 = 100$ and $m_4 = 140$.

			Bias	RMSE	PE	
		$\hat{t}$	β_0	β_0	u	λ
m$_1$	EVA	1.07	-6.527	55.789	0.541	0.892
	LA	22.48	-7.740	55.172	0.554	0.866
m$_2$	EVA	2.27	-3.799	36.149	0.278	0.846
	LA	18.86	-3.697	35.505	0.277	0.832
m$_3$	EVA	3.99	-3.198	28.079	0.161	0.860
	LA	24.06	-2.877	27.690	0.160	0.855
m$_4$	EVA	5.79	-2.034	20.584	0.100	0.625
	LA	30.97	-1.750	21.586	0.099	0.681

4.2.4 Conclusion

As a general verdict based on the results from the three simulation studies laid out above, the standard VA approach can be deemed as the most sensible option among the three in all of the cases where it can be applied in the fitting of a GLLVM. In general, computation speed was the sole possibly decisive factor in which the EVA managed to perform better in some situations, the differences being quite minor however. What is clear, is that the LA is definitely the slowest method of the three, by a fairly large margin.

In terms of accuracy, EVA and LA performed fairly evenly overall. Smaller values for the amount of sites n seemed to favor EVA over LA more clearly, with the differences subsiding as the sample size increased. The results from the third study leaned heavily toward LA, but the comparatively bloated errors raised a question of possible sensitivity issues of the simulation setting relating to the choice of the fitting approach used in estimation of the true model. To asses this, the third study should be repeated using EVA in place of LA in the fitting of the true model. Similarly, the sensitivity of the first two studies could be assessed by replicating the study settings using both EVA and LA, or maybe some fourth method not part of the actual comparisons, to reduce the possibility of favoritism. Alternatively, the use of completely synthetic simulation settings could be reasoned. The decision to use real data sets as a basis for the simulations was made based on the notion of it possibly leading to more realistic approximations of abundance data collected in real world applications.

5 Example: model based ordination of peatland data

In this section, we use EVA to fit a logistic Bernoulli GLLVM for a model based ordination of multivariate presence/absence vegetation data, collected across 120 Finnish peatlands of varying environmental backgrounds. A detailed description of the data collection process and the related study setting can be found in Elo et al. (2016). As said, the data set contained measurements from 120 test sites, with the response variables of interest being the percent covers of total of 86 bryophyte species. Ten 1m^2 sample plots were used on each site to conduct the measurements, leading to a 1200×86 response matrix. Of the 120 sites, half were considered to be in a pristine state, while the other half were sites disturbed by drainage. The pristine sites were considered as the control group in the original study, with the disturbed sites forming the treatment group. Additionally, the sites consisted evenly of three distinct peatland ecosystem types, those being spruce mires, pine mires and fens. Third environmental classifier considered was productivity, with 59 sites being deemed as sites of low productivity, and the rest labeled as sites of high productivity.

Empty rows, or rows with missing values were removed from the response matrix, as well as columns corresponding to species with less than five observations total. Then, the plant cover percentages were pooled site-wise, as in Elo et al. (2016), by calculating the mean abundance in the ten sample plots found on each site. These actions led to response matrix of size 119×65. Then, diverting from the original article, the mean coverages were transformed into binary indicators of presence or absence. Originally, the aim was to use the model for beta distributed responses discussed in 3.3, but the very high proportion of measured coverages of zero percent (about 84%) deemed the use of beta model to be problematic. In the future, some type of augmented beta model allowing for the inclusion of zero cover responses should be considered and implemented.

As was previously mentioned, a GLLVM with the amount of latent variables specified to be $p = 2$ leads naturally to two dimensional ordination plots, by simply plotting the scatterplot of the resulting latent variable score predictions $\hat{\boldsymbol{u}}_i = (\hat{u}_{i1}, \hat{u}_{i2})^\intercal$. First, a null model was specified, that is, a GLLVM with no environmental covariates included,

$$\mathrm{logit}(\mu_{ij}) = \beta_{0j} + \boldsymbol{u}_i^{\mathsf{T}} \boldsymbol{\lambda}_j, \quad i = 1, \ldots, 119,\ j = 1, \ldots, 65. \tag{21}$$

The ordination plot resulting from the null model (21) can be seen in Figure 1. The sites in the first plot have been colored according to ecosystem type. The sites form very clearly visible clusters according to the ecosystem type, with fens tending to top portion of the plot and spruce mires tending to bottom portion of the plot. In the Figure 2, the coloring corresponds to productivity level. The sites of high production tend heavily to the right portion of the plot, while sites of low production tend to left.

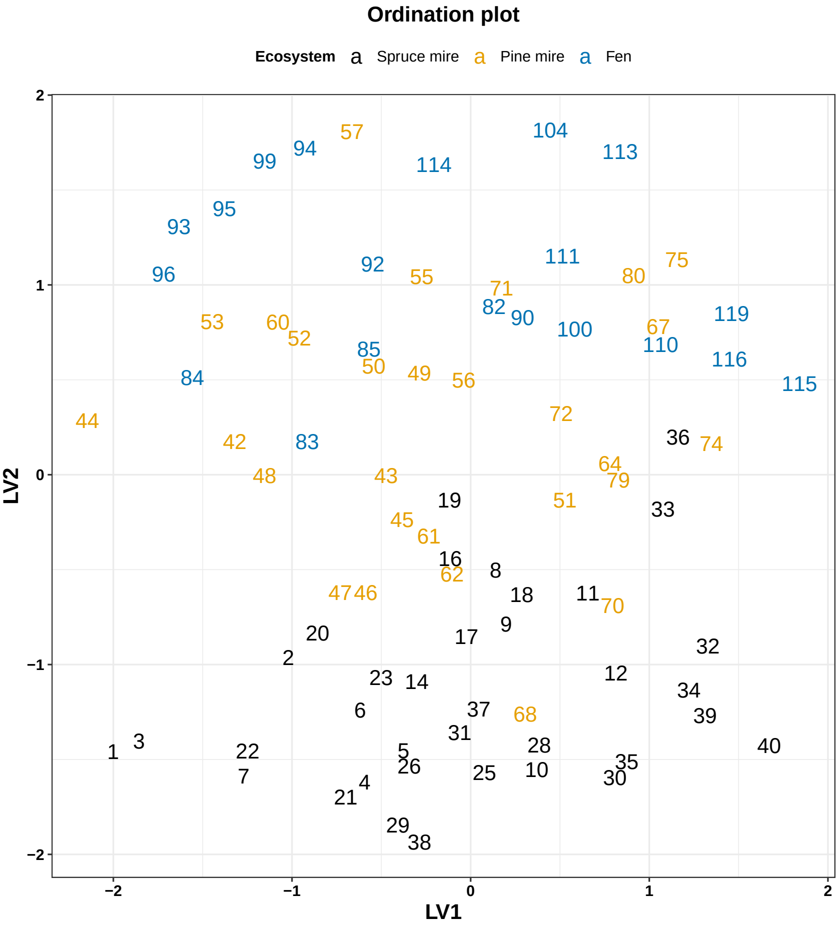

Figure 1: Ordination of the test sites present in the peatland data set, according to the null model (21), i.e. a binary logistic GLLVM without any environmental covariates. Sites are colored according to ecosystem type; sites classified as spruce mires on black, pine mires on orange and fens on blue. There is a clear clustering of the sites by the ecosystem type, with spruce mires tending to the lower portion of the plot, while the sites classified as fens tend to top.

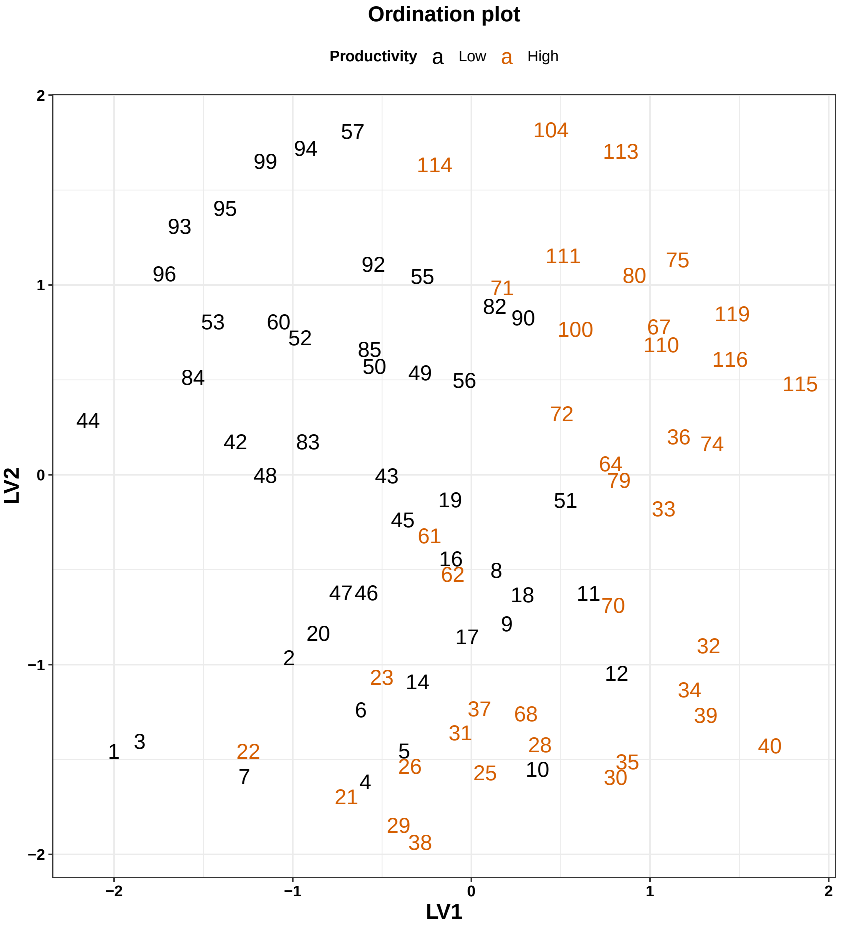

Figure 2: Ordination of the test sites present in the peatland data set, according to the null model (21), i.e. a binary logistic GLLVM without any environmental covariates. The sites are colored according to productivity level, sites of low productivity on black and sites of high productivity on red. Sites of low productivity tend heavily to left, while sites of high productivity gravitate to right.

The notes above suggest that both the productivity and ecosystem type should possibly be included in to the latent variable model as predictors, leading to model specified as

$$\operatorname{logit}(\mu_{ij}) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \beta_{3j}x_{3i} + \boldsymbol{u}_i^{\intercal}\boldsymbol{\lambda}_j, \tag{22}$$

where x_{1i} and x_{2i} are dummy variables corresponding to the categories of pine mire and fen ecosystems, respectively, and x_{3i} indicates whether the site i is classified as site of high production or not. This model resulted in to the ordination plot shown in Figure 3. When looking by ecosystem type, the sites seem to be fairly mixed when compared to Figure 1. The sites considered to be in pristine state seem to cluster heavily in to the upper portion of the plot, while the sites disturbed by drainage tend towards the lower portion of the plot.

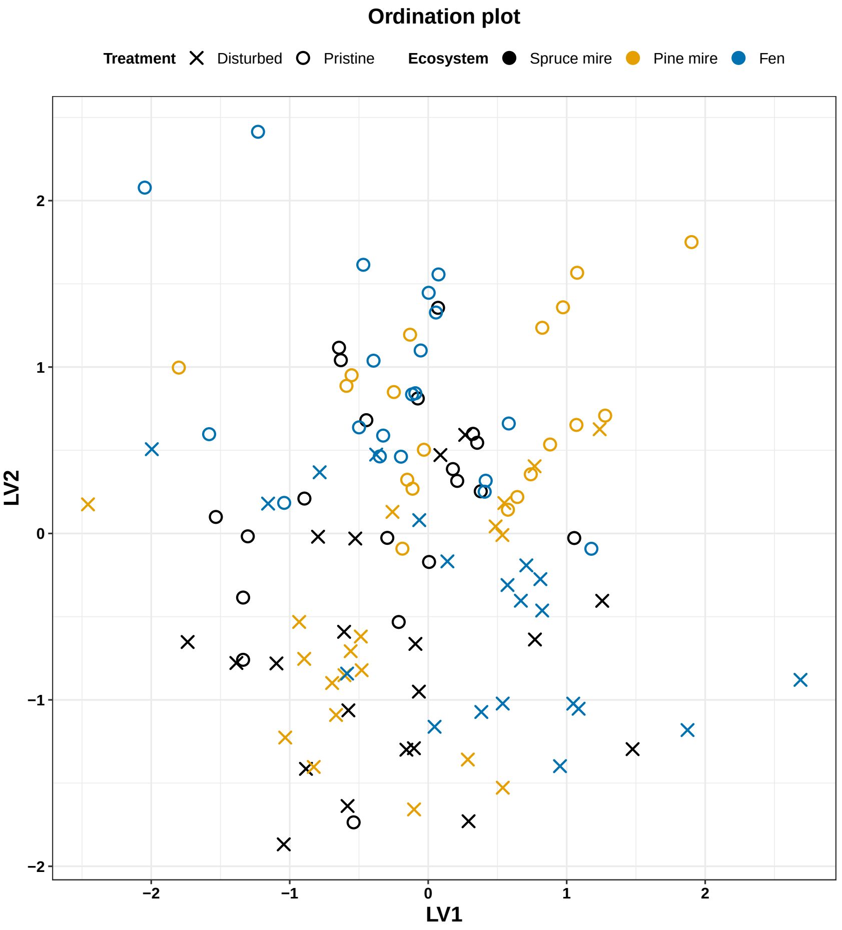

Figure 3: Ordination of the test sites according to the model (22). In terms of ecosystem type, the sites appear to be quite mixed. When looking at the sites by the treatment level, the sites classified as pristine can be seen to form a cluster on to the upper portion of the plot.

The third and final model fitted included all three of the environmental covariates present in the data set, leading to a model equation of the form.

$$\operatorname{logit}(\mu_{ij}) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \beta_{3j}x_{3i} + \beta_{4j}x_{4i} + \boldsymbol{u}_i^\intercal \boldsymbol{\lambda}_j, \tag{23}$$

where x_{1i}, x_{2i} and x_{3i} are as in the model (22) and x_{4i} is an indicator of site i being in pristine state or not. The resulting ordination plot is shown in Figure 4. Now, the effects of the treatment level seems to have been taken care of by the model. Moreover, the sites seem to fall more tightly around the origin. Sites further away from the origin tend to be sites with more species being present, than on average. For example, the sites labeled as 7, 28 and 57 had presence of 17, 18 and 15 species, respectively, of the included 65, with the average being around 10.

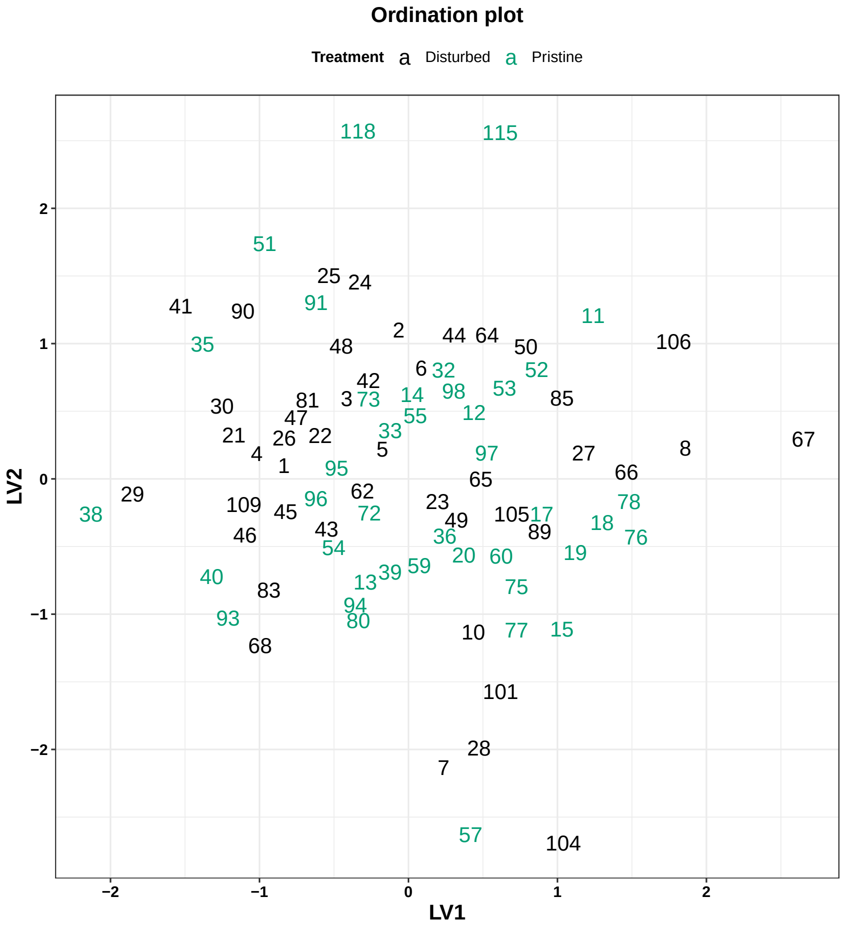

Figure 4: Ordination of the sites according to the model (23), with the ecosystem type and both treatment and productivity levels as covariates. Visibly clear clustering in terms of treatment level is no more present and the sites seem to be not as spread out as with the previous models. The sites are labelled according to the site number.

With latent variable models, the residual covariance matrix, calculated as $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\intercal$ using the loading matrix $\mathbf{\Lambda}$, carries information about the inter-species relationships after the effects of the measured environmental covariates have been controlled for. For example, trace of the matrix $\mathbf{\Sigma}$ can be thought as a measure of unexplained variation. The ratio of traces can be used as a 'pseudo-R^2 type' method to compare nested models. In this example, the trace of the residual covariance matrix resulting from the third model (23) was about 86.5% of that of the second model (22), meaning that the inclusion of treatment level as a predictor explained approximately 13.5% of the variation in the species abundances. When comparing the third model (23) to the null model (21), the ratio of the traces suggests that the three environmental covariates together managed to explain approximately 87.6% of the variation present in the abundances.

The soundness of a GLLVM fit can be assessed much in the same way as in the case of typical GLMs, by the use of residual plots. For discrete responses, the so-called *randomized quantile*, or *Dunn-Smyth residuals* (Dunn and Smyth, 1996) are often employed. The Dunn-Smyth residuals are defined as

$$r_{q,ij} = \mathbf{\Phi}^{-1}(c_{ij}), \tag{24}$$

where $c_{ij} \sim U\big(\lim_{y \uparrow y_{ij}} F(y|\hat{\mu}_{ij}, \hat{\phi}), F(y_{ij}|\hat{\mu}_{ij}, \hat{\phi})\big)$, with $F(\cdot|\hat{\mu}_{ij}, \hat{\phi})$ being the c.d.f. of the response y_{ij}. This definition gives us residuals following exact standard normal distribution, in the case of properly estimated model parameters. The Dunn-Smyth residuals can be plotted against the values of the linear predictor $\hat{\eta}_{ij}$. Normal quantile-quantile plots can also be constructed. These two types of plots, corresponding to the final model (23), can be seen in Figure 5. Neither of the residual plots show anything problematic regarding the model fit.

A good way to visualize the co-occurrence structure among the species is to construct a correlation plot based on the residual covariance matrix $\mathbf{\Sigma}$. The correlation plot resulting from the model (23) can be seen in Figure 6. The red blocks of squares correspond to groups of species with highly negative correlations of abundance. These are species that tend to not occur on same sites, after the effects of the three environmental predictors have been controlled for. Meanwhile, the blue regions consist of species having positively correlated absences, i.e. species that have high chance of co-occurrence.

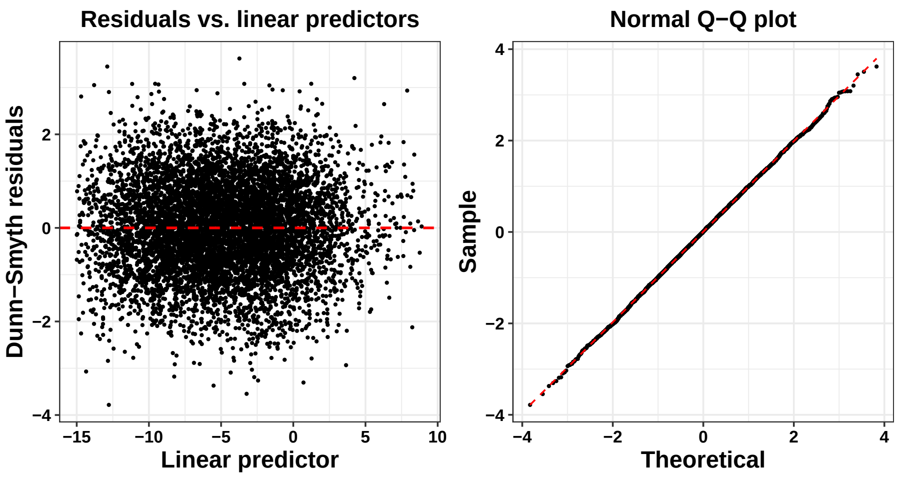

Figure 5: Two types of residual plots resulting from the final model (23). Here, the residuals used were the randomized quantile residuals presented in Dunn and Smyth (1996).

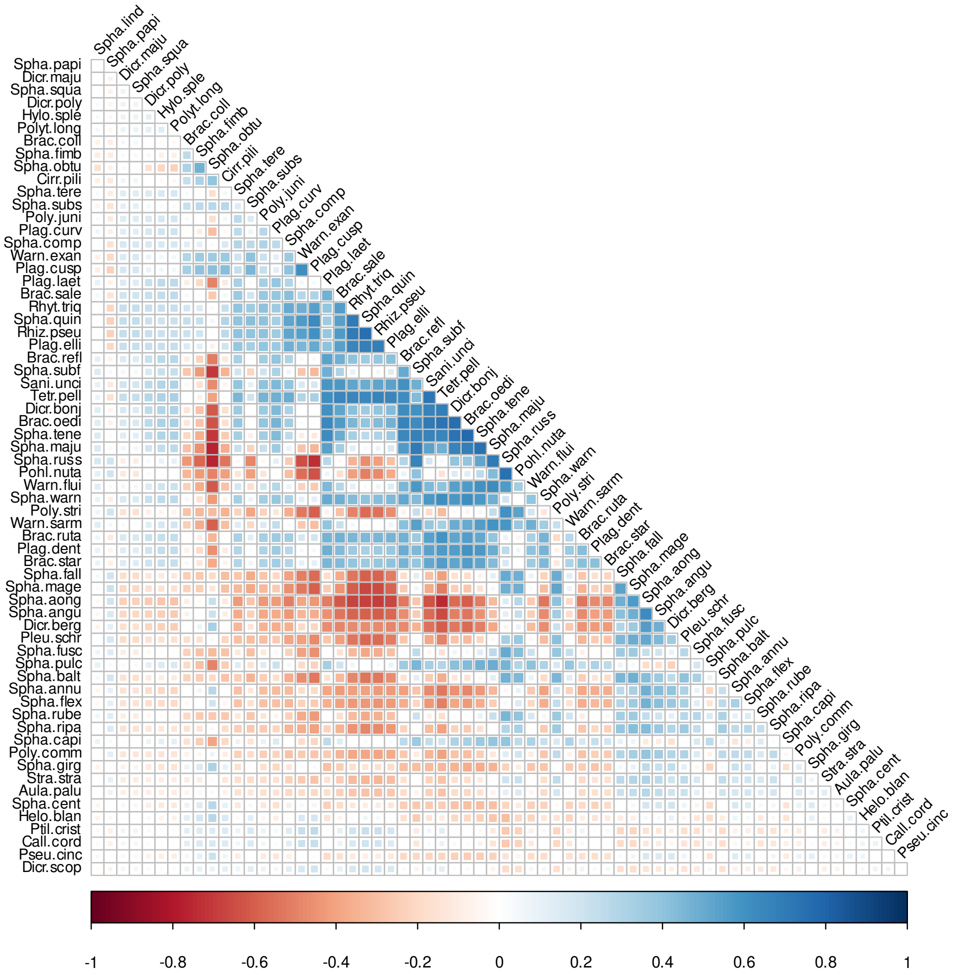

Figure 6: A correlation plot resulting from the logistic latent variable model (23) containing all three of the environmental covariates; ecosystem type and treatment (disturbance) and productivity levels. The noticeably red regions corresponds to species that have high negative correlation in abundance, i.e. species that tend to not co-exist on same sites – after the effects of the environmental covariates have been controlled for. On the opposite, the blue regions mark the species with positively correlated probabilities of presence.

6 Discussion

The method of EVA, as described in Section 2.3, managed well to broaden the amount of feasible response distributions and link functions, when compared to the standard VA (Section 2.2). In addition, in the simulation studies conducted, EVA was generally not far off from the accuracy of LA, one of the more popular alternatives. In terms of computational efficiency, EVA was vastly superior to LA in all cases studied, and slightly faster than VA in most cases. As discussed already in the Section 4.2.4, a replication of the simulation studies should be conducted, aiming to eliminate the possible effect of favoring the method chosen for fitting the true model. Additionally, completely synthetic simulation studies could also be considered.

Two distinct branches for future development regarding EVA can be seen, one concerned with improving and extending the current "basic" implementation of EVA into a more readily available form for practitioners to use, and the other concerned with extending EVA to the cases of some of the more general GLLVMs, like the so-called *fourth corner models* (Brown et al., 2014), or models with spatially or temporally correlated latent variables.

The current iteration of EVA considered rests firmly on the simple, standard definition of GLLVM as given in Section 2.1. Basic implementations concerning the specific models derived in Section 3 are already in place, yet lacking the methods required for sophisticated model diagnostics, for the method to be usable as a "black box" by researchers in applied fields. For instance, a routine for calculation of standard error estimates could be added, as well as significance tests for model parameters. Residual analysis tools, based on the Dunn-Smyth residuals (24) in the discrete cases, could be added too. The possibility of specifying random row effects α_i needs to be implemented, as the current version considers only fixed ones. Some additional models should also be considered, including for example a GLLVM for Tweedie distributed (with *power parameter* $1 < \nu < 2$) responses (Jorgensen, 1997), a popular choice of distribution for modelling biomass data, as well as models for zero-inflated Poisson and zero-one-inflated Beta distributed responses. Special attention could also be paid to improving numerical stability of the algorithms, as numerical overflow and underflow tends to happen frequently with link functions involving logarithmic transformations.

The possibility of using Taylor approximation of higher order, in place of 9 in the derivation of EVA, should be assessed. Hypothetically, higher order approximation should lead to a more accurate estimation of model parameters, at the cost of computational efficiency. Alternatively, the effects of the choice of the center of expansion could also be studied, first considering for example the method of Laplace variational inference discussed in Wang and Blei (2013).

The framework of GLLVMs can be generalized further by including additional terms in to the model equation (1), or by relaxing the assumptions of independence of the latent variables. In addition to the environmental covariates discussed in this thesis, the inclusion of species traits and the interactions between environmental and species traits could be considered, leading effectively to the topic of fourth corner models (Brown et al., 2014). The assumption of independent latent variables could be replaced with dependent latent variables, with the dependency structure perhaps being governed by some temporal or spatial (or spatio-temporal) stochastic process. For example, a assumption that the latent variable values are more similar on sites geographically closer to each other, than on sites distant from each other, sounds a plausible description for many real world situations. Generalization of the method of EVA in to these situations could be explored. The method of EVA on GLLVMs on sparse data presents one possible additional direction for future research, as well robust methods and regularization for GLLVM estimation. The bryophyte data considered in Section 5 is a good example of sparse data set, as almost all of the elements of the response matrix were zero. The extension of EVA (and GLLVMs in general) into situations of sparse data could perhaps improve the prospects of using actual percent cover response model, instead of the presence/absence model considered in this thesis.

References

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". In: *Methods in Ecology and Evolution* 5.4, pp. 344–352.

Cleary, D. F. R., Genner, M. J., Boyle, T. J. B., Setyawati, T., Angraeti, C. D., and Menken, S. B. J. (2005). "Associations of bird species richness and community composition with local and landscape-scale environmental factors in Borneo". In: *Landscape Ecology* 20.8, pp. 989–1001.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society, Series B* 39(1), pp. 1–38.

Duchi, J. (2007). "Derivations for linear algebra and optimization". In: *Berkeley, California* 3, pp. 2325–5870.

Dunn, P. K. and Smyth, G. K. (1996). "Randomized quantile residuals". In: *Journal of Computational and Graphical Statistics* 5.3, pp. 236–244.

Elo, M., Kareksela, S., Haapalehto, T., Vuori, H., Aapala, K., and Kotiaho, J. S. (2016). "The mechanistic basis of changes in community assembly in relation to anthropogenic disturbance and productivity". In: *Ecosphere* 7.4, e01310.

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

Hui, F. K. C., Taskinen, S., Pledger, S., Foster, S. D., and Warton, D. I. (2015). "Model-based approaches to unconstrained ordination". In: *Methods in Ecology and Evolution* 6(4), pp. 399–411. DOI: 10.1111/2041-210X.12236.

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., and Taskinen, S. (2017). "Variational approximations for generalized linear latent variable models". In: *Journal of Computational and Graphical Statistics* 26.1, pp. 35–43.

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications.* Springer Science & Business Media.

Jorgensen, B. (1997). *The theory of dispersion models.* CRC Press.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016). "TMB: Automatic Differentiation and Laplace Approximation". In: *Journal of Statistical Software* 70.5, pp. 1–21. DOI: 10.18637/jss.v070.i05.

Kruskal, J. B. (1964). “Nonmetric multidimensional scaling: a numerical method”. In: *Psychometrika* 29.2, pp. 115–129.

Kullback, S. and Leibler, R. A. (1951). “On Information and Sufficiency”. In: *The Annals of Mathematical Statistics* 22, pp. 79–86.

Mathai, A. M. and Provost, S. B. (1992). *Quadratic forms in random variables: theory and applications*. Dekker.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). “Equation of State Calculations by Fast Computing Machines”. In: *The Journal of Chemical Physics* 21(6), pp. 1087–1092. DOI: 10.1063/1.1699114.

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”. In: *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”. In: *Methods in Ecology and Evolution* 10.12, pp. 2173–2182.

Ormerod, J. T. and Wand, M. P. (2012). “Gaussian Variational Approximate Inference for Generalized Linear Mixed Models”. In: *Journal of Computational and Graphical Statistics* 21(1), pp. 2–17. DOI: 10.1198/jcgs.2011.09118.

— (2010). “Explaining Variational Approximations”. In: *The American Statistician* 64(2), pp. 140–153. DOI: 10.1198/tast.2010.09058.

Peres-Neto, P. R. and Jackson, D. A. (2001). “How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test”. In: *Oecologia* 129.2, pp. 169–178.

Pollock, Laura J, Tingley, Reid, Morris, William K, Golding, Nick, O’Hara, Robert B, Parris, Kirsten M, Vesk, Peter A, and McCarthy, Michael A (2014). “Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)”. In: *Methods in Ecology and Evolution* 5.5, pp. 397–406.

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

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. Crc Press.

Tierney, L. and Kadane, J. B. (1986). “Accurate approximations for posterior moments and marginal densities”. In: *Journal of the american statistical association* 81.393, pp. 82–86.

Wang, C. and Blei, D. M. (2013). “Variational inference in nonconjugate models”. In: *Journal of Machine Learning Research* 14.Apr, pp. 1005–1031.

Warton, D. I., Blanchet, F. G., O’Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C, and Hui, F. K. C. (2015). “So many variables: joint modeling in community ecology”. In: *Trends in Ecology & Evolution* 30.12, pp. 766–779.