Variational Approximations for Generalized Linear Latent Variable Models

Francis K.C. Hui*[1], David I. Warton[2,3], John T. Ormerod[4,5], Viivi Haapaniemi[6], and Sara Taskinen[6]

[1]Mathematical Sciences Institute, Australian National University, Canberra, Australia

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

[3]Evolution & Ecology Research Centre, The University of New South Wales, Sydney, Australia

[4]School of Mathematics and Statistics, The University of Sydney, Sydney, Australia

[5]ARC Centre of Excellence for Mathematical & Statistical Frontiers

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

*Francis Hui, Mathematical Sciences Institute, Australian National University, Canberra, ACT, 0200, Australia. *email:* fhui28@gmail.com

Abstract

Generalized Linear Latent Variable Models (GLLVMs) are a powerful class of models for understanding the relationships among multiple, correlated responses. Estimation however presents a major challenge, as the marginal likelihood does not possess a closed form for non-normal responses. We propose a variational approximation (VA) method for estimating GLLVMs. For the common cases of binary, ordinal, and overdispersed count data, we derive fully closed form approximations to the marginal log-likelihood function in each case. Compared to other methods such as the expectation-maximization algorithm, estimation using VA is fast and straightforward to implement. Predictions of the latent variables and associated uncertainty estimates are also obtained as part of the estimation process. Simulations show that VA estimation performs similar to or better than some currently available methods, both at predicting the latent variables and estimating their corresponding coefficients. They also show that VA estimation offers dramatic reductions in computation time particularly if the number of correlated responses is large relative to the number of observational units. We apply the variational approach to two datasets, estimating GLLVMs to understanding the patterns of variation in youth gratitude and for constructing ordination plots in bird abundance data. `R` code for performing VA estimation of GLLVMs is available online.

Keywords: Factor analysis, Item response theory, Latent Trait, Multivariate analysis, Ordination, Variational approximation.

1 Introduction

In many areas of applied science, data on multiple, correlated responses are often collected, with one of the primary aims being to understand the latent variables driving these correlations. For instance, in psychometrics, subjects are given a series of questions that all relate to some latent trait/s such as gratitude. Another example is in ecology, where the abundances of many, interacting species are collected at each site, and ordination is commonly applied to visualize patterns between sites on a latent species composition space (**??**). Generalized linear latent variable models

(GLLVMs, **?**) offer a general framework for analyzing multiple, correlated responses. This is done by extending the basic generalized linear model to incorporate one or more latent variables. Specific cases of GLLVMs include factor analysis where all the responses are normally distributed, and item response theory models where the responses are binary or ordinal.

Estimating GLLVMs presents a major challenge since the marginal likelihood function, which involves integrating over the latent variables, does not posses a closed form when the responses are non-normal. In this paper, we focus on maximum likelihood estimation of GLLVMs, for which several methods have been proposed. These include Laplace's approximation (**??**), numerical integration methods such as adaptive quadrature (**?**), and the expectation-maximization (EM) algorithm or some variant of it (**??**); see **?** for a thorough review of estimation methods for GLLVMs. Many of these methods however remain computationally burdensome to use, especially the case when the number of correlated responses is large and more than one latent variable is considered.

In this article, we propose a variational approximation (VA) approach for estimating GLLVMs. A comprehensive summary of the VA approach can be found in **?**, but briefly, VA belongs to a rich class of approximations for converting a difficult optimization problem to a simpler one, whose roots begin in quantum mechanics (**?**) and were subsequently taken up in computer science to fit graphical models (**?**). With regards to statistical estimation, one attractive way of thinking about variational approximations, as discussed in Section 3, is as a means of obtaining a more tractable (potentially closed form) yet optimal approximation to an intractable likelihood (optimal in the sense of minimizing the Kullback-Leibler divergence). Over the past decade, variational methods have become increasingly popular for approximating posterior distributions in Bayesian modeling (e.g. **?**). By contrast, their use in maximum likelihood estimation for dealing with intractable likelihoods has received little attention. **?** proposed a Gaussian VA approach to maximum likelihood estimation of generalized linear mixed models, while **?** demonstrated attractive asymptotic properties of using a Gaussian VA method for Poisson mixed models. Variational EM algorithms have also been proposed specifically for random effects item response theory models (**?**) and factor analysis (**?**), but none so far have considered the broader GLLVM framework.

Motivated by examples in psychometrics and ecology we proposed a VA approach to estimating GLLVMs, with a focus on common cases of binary, ordinal, and overdispersed count data. In each case, we derive optimal forms for the variational distributions and a closed form for the VA log-likelihood. Estimation of GLLVMs is then straightforward, involving iterative updates of the model and variational parameters which can be performed using standard optimization routines such as iterative reweighted least squares. Predictions of the latent variables, their standard errors, as well as uncertainty estimates are also obtained as part of the estimation process. Simulations show that the VA approach performs similar to or better than some of the currently available methods, both in predicting the latent variables and estimating the parameters of the model, with potentially substantial reductions in computation time. We apply the proposed VA method to datasets in psychometrics and ecology, demonstrating in both examples how GLLVMs offer a model-based framework to understanding the major patterns of variation behind the correlated data on a latent space.

2 Generalized Linear Latent Variable Models

Let $\boldsymbol{y} = (\boldsymbol{y}_1 \ldots \boldsymbol{y}_n)^T$ denote an $n \times m$ response matrix, where rows $i = 1, \ldots, n$ are the observational units, and columns $j = 1, \ldots, m$ are correlated responses. A vector of p covariates, $\boldsymbol{x}_i$, may also be recorded for each observation. For a GLLVM, conditional on a vector of $d \ll m$ underlying latent variables, $\boldsymbol{u}_i$ and parameter vector $\boldsymbol{\Psi}$ (defined shortly), the responses y_{ij} are assumed to come from the exponential family of distributions, $f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp\left[\{y_{ij}\theta_{ij} - b(\theta_{ij})\}/\phi_j + c(y_{ij}, \phi_j)\right]$, where $b(\cdot)$ and $c(\cdot)$ are known functions, θ_{ij} are canonical parameters, and ϕ_j is the dispersion parameter. For simplicity, we assume all responses come from the same distribution, although the developments below can be extended to handle mixed response types through column dependent functions $b_j(\cdot)$ and $c_j(\cdot)$. The mean response, denoted as μ_{ij}, is regressed against $\boldsymbol{u}_i$, along with the p covariates if appropriate via,

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

where $g(\cdot)$ is a known link function, $b'(\theta_{ij}) = \mu_{ij}$, β_{0j} is a column-specific intercept, and $\boldsymbol{\lambda}_j$ and $\boldsymbol{\beta}_j$ are coefficients related to the latent variables and covariates respectively. The above model allows for the case where all responses have the same regression coefficients, $\boldsymbol{\beta}_1 = \ldots = \boldsymbol{\beta}_m = \boldsymbol{\beta}$, although we keep the developments more general. Also, a row effect, τ_i, may be included in (1), e.g., to standardize for site total abundance with multivariate abundance data, ensuring that the ordination is in terms of species composition. Let $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_d)^T$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1 \ldots \boldsymbol{\beta}_p)^T$ denote the $m \times d$ and $m \times p$ matrices of regression coefficients corresponding to the latent variables and covariates respectively. Finally, let $\boldsymbol{\Psi} = \{\tau_1, \ldots, \tau_n, \beta_{01}, \ldots, \beta_{0m}, \phi_1, \ldots, \phi_m, \text{vec}(\boldsymbol{\lambda}), \text{vec}(\boldsymbol{\beta})\}$ denote all the parameters in the model.

We assume that the latent variables are drawn from independent, standard normal distributions, $\boldsymbol{u}_i \sim N_d(\boldsymbol{0}, \boldsymbol{I}_d)$ where $\boldsymbol{I}_d$ denotes a $d \times d$ identity matrix. The use of a zero mean and unit variance act as identifiability constraints to avoid location and scale invariance. We also impose constraints on the latent variable coefficient matrix to avoid rotation invariance. Specifically, we set all the upper triangular elements of $\boldsymbol{\lambda}$ to zero, and constrain its diagonal elements to be positive. Note that the assumption of independent latent variables is commonly applied (e.g. **?**), and is made without loss of generality, i.e., the independence assumption does not constrain the capacity to model the correlations between the columns of $\boldsymbol{y}$, and the model as formulated still covers the set of all rank-d covariance matrices.

3 Variational Approximation for GLLVMs

Conditional on the latent variables, the responses for each observational unit are assumed to be independent in a GLLVM, $f(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\Psi}) = \prod_{j=1}^m f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi})$. The marginal log-likelihood is then obtained by integrating over $\boldsymbol{u}_i$,

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log\{f(\boldsymbol{y}_i, \boldsymbol{\Psi})\} = \sum_{i=1}^{n} \log\left(\int \prod_{j=1}^{m} f(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\Psi}) f(\boldsymbol{u}_i)\, d\boldsymbol{u}_i\right), \tag{2}$$

where $f(\boldsymbol{u}_i)$ is a multivariate, standard normal distribution, as discussed in Section 2. As reviewed in Section 1, numerous methods have been proposed for performing the integration in (2), although many are computationally burdensome to implement. To overcome this, we propose applying a variational approximation to obtain a closed form approximation to $\ell(\boldsymbol{\Psi})$. For a generic marginal log-likelihood function $\ell(\boldsymbol{\Psi}) = \log \int f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u})\, d\boldsymbol{u}$, a commonly applied VA approach utilizes Jensen's inequality to construct a lower bound,

$$\log\left\{\int \frac{f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u}) q(\boldsymbol{u}|\boldsymbol{\xi})}{q(\boldsymbol{u}|\boldsymbol{\xi})}\right\} d\boldsymbol{u} \geq \int \log\left\{\frac{f(\boldsymbol{y}|\boldsymbol{u}, \boldsymbol{\Psi}) f(\boldsymbol{u})}{q(\boldsymbol{u}|\boldsymbol{\xi})}\right\} q(\boldsymbol{u}|\boldsymbol{\xi}) d\boldsymbol{u} \equiv \underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}), \tag{3}$$

for some variational density $q(\boldsymbol{u}|\boldsymbol{\xi})$ with parameters $\boldsymbol{\xi}$. The VA log-likelihood $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ can thus be interpreted as the Kullback-Leibler distance between $q(\boldsymbol{u}|\boldsymbol{\xi})$ and the joint likelihood $f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})$. Evidently, this is minimized by choosing the posterior distribution $q(\boldsymbol{u}|\boldsymbol{\xi}) \equiv f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$, but in order to obtain a tractable form for $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$, we choose a parametric form for $q(\boldsymbol{u}|\boldsymbol{\xi})$. Specifically, we use independent normal VA distributions for the latent variables, such that for $i = 1, \ldots, n$, we have $q(\boldsymbol{u}_i) \equiv N_d(\boldsymbol{a}_i, \boldsymbol{A}_i)$ such that $\boldsymbol{\xi}_i = \{\boldsymbol{a}_i, \text{vech}(\boldsymbol{A}_i)\}$, where $\boldsymbol{A}_i$ is an unstructured covariance matrix (although in our simulations in Section 5, we consider both unstructured and diagonal forms for $\boldsymbol{A}_i$). In Appendix **??**, we show that, in the family of multivariate normal distributions, the choice of independent VA distributions is indeed the optimal one.

With independent normal VA distributions for $\boldsymbol{u}_i$, we obtain the following result.

Lemma 1. *For the GLLVM as defined in (1), the VA log-likelihood is given by*

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{\frac{y_{ij}\tilde{\eta}_{ij} - E_q\{b(\theta_{ij})\}}{\phi_j} + c(y_{ij}, \phi_j)\right\} + \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - tr(\boldsymbol{A}_i) - \boldsymbol{a}_i^T\boldsymbol{a}_i\right),$$

where $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{a}_i^T\boldsymbol{\lambda}_j$, and all quantities constant with respect to the parameters *have been omitted.*

Estimation of the GLLVM is performed by maximizing the VA log-likelihood simultaneously over the variational parameters $\boldsymbol{\xi}$ and model parameters $\boldsymbol{\Psi}$. Note however that there remains an expectation term, $E_q\{b(\theta_{ij})\}$, which is not guaranteed to have a closed form. In **?**, this was dealt with using adaptive Gauss-Hermite quadrature. By contrast, in the next section, we show that *fully* explicit forms for $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ can be derived for some common cases of GLLVMs through a reparameterization of the models. Three responses types are of particular relevance to this article: 1) Bernoulli responses, 2) overdispersed counts, and 3) ordinal data, and in each case we obtain a closed form VA log-likelihood.

Finally, we propose that the estimator of $\boldsymbol{\Psi}$ based on maximizing Lemma 1 is estimation consistent (as in **?**). That is, let $(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\xi}})$ denote the maximizer of $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$. Then as $n \to \infty$ and $m \to \infty$, we have $\hat{\boldsymbol{\Psi}} \xrightarrow{p} \boldsymbol{\Psi}_0$ where $\boldsymbol{\Psi}_0$ denotes the true parameter point and $\hat{\boldsymbol{\Psi}}$ is the VA estimator. A heuristic proof of this is provided in Appendix **??**. Logically, consistency of the estimators depends critically on the accuracy of the VA log-likelihood approximation to the true marginal likelihood (**?**). In brief, a central limit theorem based argument shows that the posterior distribution $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$ is asymptotically normally distributed as $m \to \infty$, and therefore with $q(\boldsymbol{u}|\boldsymbol{\xi})$ chosen as a normal distribution then the VA log-likelihood is expected to converge to the true likelihood, i.e., the lower bound in (3) gets sharper as $m \to \infty$.

3.1 Bernoulli Responses

When the responses are binary, we assume a Bernoulli distribution and use the probit link function. Equivalently, we introduce an auxiliary variable, z_{ij}, which is normally distributed with mean η_{ij} and unit variance, and set $y_{ij} = 1$ if $z_{ij} \geq 0$ and $y_{ij} = 0$ otherwise. We thus have $f(y_{ij}|z_{ij}, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \mathrm{I}(z_{ij} \geq 0)^{y_{ij}} \mathrm{I}(z_{ij} < 0)^{1-y_{ij}}$ where $z_{ij} \sim N(\eta_{ij}, 1)$, where $\mathrm{I}(\cdot)$ denotes the indicator function. Under this parameterization, the marginal log-likelihood requires integrating over both $\boldsymbol{u}_i$ and z_{ij}, that is, $\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \Big(\int \int \prod_{j=1}^{m} f(y_{ij}|z_{ij}, \boldsymbol{u}_i, \boldsymbol{\Psi}) f(z_{ij}) f(\boldsymbol{u}_i) \, dz_{ij} d\boldsymbol{u}_i \Big)$. However, the key advantage with introducing the auxiliary variable is that it leads to a closed form for $\underline{\ell}(\mathbf{\Psi}; q)$. To show this, we first choose a VA distribution $q(z_{ij})$, which we assume to be independent of $q(\boldsymbol{u}_i)$. The following guides this choice.

Lemma 2. *The optimal choice of $q(z_{ij})$, in the sense of maximizing the lower bound $\underline{\ell}(\mathbf{\Psi}, \boldsymbol{\xi})$, is a truncated normal distribution with location parameter $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{a}_i^T \boldsymbol{\lambda}_j$, scale parameter 1, and limits $(-\infty, 0)$ if $y_{ij} = 0$, and $(0, \infty)$ if $y_{ij} = 1$.*

All proofs may be found in Appendix **??**. Combining the above result with our choice of $q(\boldsymbol{u}_i)$ as a normal distribution leads to the result below.

Theorem 1. *The VA log-likelihood for the Bernoulli GLLVM with probit link is given by the following expression*

$$\underline{\ell}(\mathbf{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[y_{ij} \log\{\Phi(\tilde{\eta}_{ij})\} + (1 - y_{ij}) \log\{1 - \Phi(\tilde{\eta}_{ij})\} \right] - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j + \frac{1}{2} \sum_{i=1}^{n} \left(\log\det(\boldsymbol{A}_i) - tr(\boldsymbol{A}_i) - \boldsymbol{a}_i^T \boldsymbol{a}_i \right),$$

where $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{a}_i^T \boldsymbol{\lambda}_j$ and all other quantities that are constant with respect to the parameters have been omitted.

Note the first summation in Theorem 1 is independent of $\boldsymbol{A}_i$, meaning the estimates of $\boldsymbol{A}_i$ are the same for all observations. Maximizing $\underline{\ell}(\mathbf{\Psi}, \boldsymbol{\xi})$ in Theorem 1 is straightforward, since the VA log-likelihood involves only separate summands over i and j, and can be performed, for instance, by iterating the following steps until convergence:

1. For $j = 1, \ldots, m$, update $(\beta_{0j}, \boldsymbol{\beta}_j)$ by fitting a probit Generalized Linear Model (GLM) with $\boldsymbol{x}_i$ as covariates and $\tau_i + \boldsymbol{a}_i^T \boldsymbol{\lambda}_j$ entered as an offset.

2. For $j = 1, \ldots, m$, update $\boldsymbol{\lambda}_j$ by fitting a penalized probit GLM, where $\boldsymbol{a}_i$ are treated as covariates, $\tau_i + \beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j$ is entered as an offset, and the ridge penalty $(1/2) \sum_{i=1}^{n} \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j$ is used. The GLM fitting process must also account for constraints on $\boldsymbol{\lambda}_j$.

3. For $i = 1, \ldots, n$, update τ_i and $\boldsymbol{a}_i$ by fitting a penalized probit GLM, where $\boldsymbol{\lambda}_j$ are treated as covariates, $\beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j$ is entered as an offset, and the ridge penalty $\boldsymbol{a}_i^T\boldsymbol{a}_i$ is used. Then a closed form update can be used for $\boldsymbol{A}_i$, specifically, $\boldsymbol{A}_i = \left(\boldsymbol{I}_d + \sum_{j=1}^{m}\boldsymbol{\lambda}_j\boldsymbol{\lambda}_j^T\right)^{-1}$.

Note that rather than updating the column or row specific parameters separately, we could instead apply optimization routines to update all parameters at once, i.e. update all $\{\beta_{01}, \ldots, \beta_{0m}, \text{vec}(\boldsymbol{\lambda}), \text{vec}(\boldsymbol{\beta})\}$, then update all $(\tau_1, \ldots, \tau_n, \boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$, and then $\boldsymbol{A}_i$. Finally, we point out that had we used the logit link instead, then by Lemma 1 the resulting VA log-likelihood would involve a term $E_q[\log\{1+\exp(\eta_{ij})\}]$, and therefore would involve numerical integration to calculate and optimize. By contrast, using a probit link and thus Lemma 2 offers a fully closed form VA log-likelihood.

3.2 Overdispersed Counts

For count data, a standard option is to assume a Poisson distribution with log link function. In such a case, the VA log-likelihood for a Poisson GLLVM is given by the following

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\left\{y_{ij}\tilde{\eta}_{ij} - \exp\left(\tilde{\eta}_{ij} + \frac{1}{2}\boldsymbol{\lambda}_j^T\boldsymbol{A}_i\boldsymbol{\lambda}_j\right)\right\} + \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - \text{tr}(\boldsymbol{A}_i) - \boldsymbol{a}_i^T\boldsymbol{a}_i\right),$$

where $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{a}_i^T\boldsymbol{\lambda}_j$, and all quantities constant with respect to the parameters are omitted. The proof of the above is similar to the derivation of the VA log-likelihood for the Poisson mixed model in **?**, and is omitted here. In many settings however, count data are overdispersed. A prime example of this is multivariate abundance data in ecology, where many species tend to be found in large numbers or not at all. To handle this, one could assume a negative binomial distribution with quadratic mean-variance relationship, $\text{Var}(y_{ij}) = \mu_{ij} + \mu_{ij}^2/\phi_j$, where ϕ_j is the response-specific overdispersion parameter. From Lemma 1 however, it can be shown this results in the expectation term $E_q[\log\{1 + \phi_j\exp(\eta_{ij})\}]$, which requires numerical methods to deal with. To overcome this, we propose using a Poisson-Gamma random effects model instead, $f(y_{ij}|\nu_{ij}, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \exp(-\nu_{ij})(\nu_{ij})^{y_{ij}}/y_{ij}!$, where $\nu_{ij} \sim \text{Gamma}(\phi_j, \phi_j/\mu_{ij})$, and $\log(\mu_{ij}) =$

η_{ij}. The parameterization produces the same quadratic mean-variance relationship as the negative binomial distribution. However, it can be shown that the optimal VA distribution for ν_{ij} is a Gamma distribution with shape $(y_{ij} + \phi_j)$ and rate $\{1 + \phi_j \exp(-\tau_i - \beta_{0j} - \boldsymbol{x}_i^T \boldsymbol{\beta}_j - \boldsymbol{a}_i^T \boldsymbol{\lambda}_j + \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j / 2)\}$. Combining this result with choice of $q(\boldsymbol{u}_i)$ leads to the following fully closed form.

Theorem 2. *The VA log-likelihood for Poisson-Gamma GLLVM with log link is given by the following expression*

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(y_{ij} \left(\tilde{\eta}_{ij} - \frac{1}{2} \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j \right) - (y_{ij} + \phi_j) \log \left\{ \phi_j + \exp \left(\tilde{\eta}_{ij} - \frac{1}{2} \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j \right) \right\} \right.$$
$$\left. + \log \Gamma(y_{ij} + \phi_j) - \frac{\phi_j}{2} \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j \right) + n\{\phi_j \log(\phi_j) - \log \Gamma(\phi_j)\}$$
$$+ \frac{1}{2} \sum_{i=1}^{n} \left(\log \det(\boldsymbol{A}_i) - tr(\boldsymbol{A}_i) - \boldsymbol{a}_i^T \boldsymbol{a}_i \right),$$

where $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{a}_i^T \boldsymbol{\lambda}_j$, $\Gamma(\cdot)$ *is the Gamma function, and all other quantities that are constant with respect to the parameters have been omitted.*

To update the VA log-likelihood above, we can iterate the following steps until convergence:

1. For $j = 1, \ldots, m$, update $(\beta_{0j}, \boldsymbol{\beta}_j, \phi_j)$ by fitting a negative binomial GLM, with $\boldsymbol{x}_i$ as covariates and $\tau_i + \boldsymbol{a}_i^T \boldsymbol{\lambda}_j - (1/2) \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j$ entered as an offset.

2. For $j = 1, \ldots, m$, update $\boldsymbol{\lambda}_j$ using a optimization routine such as the Quasi-Newton method.

3. For $i = 1, \ldots, n$, update τ_i and $\boldsymbol{a}_i$ by fitting a penalized negative binomial GLM, where $\boldsymbol{\lambda}_j$ are treated as covariates, $\beta_{0j} + \boldsymbol{x}_i^T \boldsymbol{\beta}_j - (1/2) \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j$ is entered as an offset, and the ridge penalty $\boldsymbol{a}_i^T \boldsymbol{a}_i$ is used. Then a fixed–point algorithm can be used to update $\boldsymbol{A}_i$, specifically, using the formula $\boldsymbol{A}_i = \left(\boldsymbol{I}_d + \sum_{j=1}^{m} \boldsymbol{\lambda}_j \boldsymbol{\lambda}_j^T W_{ij} \right)^{-1}$, where $W_{ij} = \phi_j (y_{ij} + \phi_j) / (\phi_j + \exp\left(\tilde{\eta}_{ij} - (1/2) \boldsymbol{\lambda}_j^T \boldsymbol{A}_i \boldsymbol{\lambda}_j \right)$.

3.3 Ordinal Data

Ordinal responses can be handled by extending the Bernoulli GLLVM in Section 3.1 to use cumulative probit regression. Suppose y_{ij} can take one of K_j possible levels, $\{1, 2, \ldots, K_j\}$. Then for each $i = 1, \ldots, n; j = 1, \ldots, p$, we define the vector $(y^*_{ij1}, \ldots, y^*_{ijK_j})$ where $y^*_{ijk} = 1$ if $y_{ij} = k$ and zero otherwise. Next, we introduce an auxiliary variable z_{ij} that is normally distributed with mean η_{ij} and unit variance, and define a vector of cutoffs $\zeta_{j0} < \zeta_{j1} < \ldots < \zeta_{jK_j}$ for each response column, with $\zeta_{j0} = -\infty$ and $\zeta_{jK_j} = +\infty$, such that $y^*_{ijk} = 1$ (equivalently, $y_{ij} = l$) if $\zeta_{j(k-1)} < z_{ij} < \zeta_{jk}$. Under this parameterization, the conditional likelihood of the responses follows a multinomial distribution, $f(y_{ij}|z_{ij}, \boldsymbol{u}_i, \boldsymbol{\Psi}) = \prod_{k=1}^{K_j} \mathrm{I}(\zeta_{j(k-1)} < z_{ij} < \zeta_{jk})^{y^*_{ijk}} \quad \text{where} \quad z_{ij} \sim N(\eta_{ij}, 1)$. With both the cutoffs and the intercept β_{0j} included, the model is unidentifiable due to location invariance. We thus set $\zeta_{j1} = 0$, and freely estimate the remaining cutoffs $\zeta_{j2} < \ldots < \zeta_{j(K_j-1)}$. Setting $\zeta_{j1} = 0$ and keeping the intercept in the model ensures that in the case of $K_j = 2$, the parameterizations of the ordinal and Bernoulli GLLVMs are equivalent. The following guides the choice of $q(z_{ij})$.

Lemma 3. *The optimal choice of $q(z_{ij})$, in the sense of maximizing the lower bound $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$, is a truncated normal distribution with mean $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{a}_i^T\boldsymbol{\lambda}_j$, variance 1, and limits $(\zeta_{j(k-1)}, \zeta_{jk})$ if $y^*_{ijk} = 1$.*

The above is a straightforward extension of Lemma 2. We therefore have the following result.

Theorem 3. *The VA log-likelihood for ordinal GLLVM using cumulative probit regression is given by the following expression*

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{K_j} y^*_{ijl}\left[\log\left\{\Phi(\zeta_{jk} - \tilde{\eta}_{ij}) - \Phi(\zeta_{j(k-1)} - \tilde{\eta}_{ij})\right\}\right] - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\boldsymbol{\lambda}_j^T\boldsymbol{A}_i\boldsymbol{\lambda}_j$$
$$+ \frac{1}{2}\sum_{i=1}^{n}\left(\log\det(\boldsymbol{A}_i) - tr(\boldsymbol{A}_i) - \boldsymbol{a}_i^T\boldsymbol{a}_i\right),$$

where $\tilde{\eta}_{ij} = \tau_i + \beta_{0j} + \boldsymbol{x}_i^T\boldsymbol{\beta}_j + \boldsymbol{a}_i^T\boldsymbol{\lambda}_j$, $\zeta_{j0} = -\infty$ and $\zeta_{jK_j} = +\infty$, $\zeta_{j1} = 0$, and all other quantities that are constant with respect to the parameters have been omitted.

Maximizing the VA log-likelihood in Theorem 3 follows the same approach as the iterative steps provided for the binary response case at the end of Section 3.1, with the only difference between that instead of probit GLMs, we fit cumulative probit regression models in steps one and two instead. Note that cumulative probit regression models will also provide estimates of the cutoffs ζ_{jk}, or alternatively, a Quasi-Newton optimization routine can be used to update the cutoffs as an additional step.

4 Inference and Prediction

After fitting the GLLVM, we are often interested in interpretation and analysis of the model parameters $\boldsymbol{\Psi}$, as well prediction and ordination of the latent variables $\boldsymbol{u}_i$. For the former, we can treat $\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi})$ as a log-likelihood function, with $(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\xi}})$ as the maximum likelihood estimates (MLEs), and base inference around this. For instance, approximate asymptotic standard errors may be obtained based on the observed information matrix evaluated at the MLEs, given by

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

Note $\boldsymbol{I}(\hat{\boldsymbol{\Psi}}, \hat{\boldsymbol{\xi}})$ consists of three blocks corresponding to the negative Hessian matrices with respect to $\hat{\boldsymbol{\Psi}}$, $\hat{\boldsymbol{\xi}}$, as well as their cross derivatives. The Hessian matrix with respect to $\hat{\boldsymbol{\xi}}$ exhibits a block diagonal structure due to the independence of $\boldsymbol{u}_i$ with respect to the VA distribution. If row effects τ_i are not included, then the Hessian matrix with respect to $\hat{\boldsymbol{\Psi}}$ also exhibits a block diagonal structure. In summary, the three blocks can be calculated in $O(\max(m, n))$ operations, after which blockwise inversion can be used to obtain the covariance matrix. Confidence intervals and approximate Wald tests for the model parameters $\hat{\boldsymbol{\Psi}}$ can then be implemented.

For ordination, the two most common methods of constructing predictions for the latent variables are empirical Bayes and maximum a-posteriori, which correspond respectively to the mean and mode of the posterior distribution $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$. For estimation methods such as numerical integration, constructing these predictions and estimates of their uncertainty require additional computation after the GLLVM is fitted. In the Gaussian VA framework however, maximizing with respect to $\boldsymbol{\xi}$ is equivalent to minimizing the Kullback-Leibler distance between $q(\boldsymbol{u}|\boldsymbol{\xi})$ and $f(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$. Therefore with the normality assumption on $q(\boldsymbol{u}|\boldsymbol{\xi})$, it follows that for the cluster i, the vector $\hat{\boldsymbol{a}}_i$ is both the variational versions of the empirical Bayes and maximum a-posteriori predictors of the latent variables and $\hat{\boldsymbol{A}}_i$ provides an estimate of the posterior covariance matrix. Importantly, both $\hat{\boldsymbol{a}}_i$ and $\hat{\boldsymbol{A}}_i$ are obtained directly from the estimation algorithm, as was seen in Section 3. In summary, the Gaussian VA approach quite naturally lends itself to the problem of predicting latent variables and constructing ordination plots, with $\hat{\boldsymbol{a}}_i$ can be used as the point predictions and $\hat{\boldsymbol{A}}_i$ can be used to construct prediction regions around these points.

5 Simulation Study

We performed a simulation study to compare our proposed VA approach to several currently available methods for fitting GLLVMs. Two settings were considered: the first simulated binary response datasets resembling those in item response theory, while the second setting simulated datasets resembling overdispersed species counts in ecology. In both settings, we assessed performance based on computation time, and the difference between the true and estimated parameter values/latent variables as calculated using the symmetric Procrustes error (see Chapter 8.4, **?**). The Procrustes error is commonly used as a method of comparing different methods of ordination, and can be thought of as the mean squared error of two matrices after accounting for differences in rotation and scale. It is an appropriate method of evaluating performance in this simulation, given we are interested in an overall measure of how well the latent variables and parameters from the fitted model matched those of the true model, while accounting for potential differences in scaling and rotation that have no bearing on a model's performance given their arbitrariness. We calculated the Procrustes error via the `procrustes` function in the `R` package `vegan` (**?**).

5.1 Setting 1

Binary datasets were simulated from GLLVMs with $d = 2$ latent variables and assuming the probit link, considering different combinations of $n = \{50, 100, 200\}$ and $m = \{10, 40\}$. Each true model was constructed by first simulating a $n \times 2$ matrix of true latent variables, such that 50% of the values were generated from a bivariate normal distribution with mean (-2,2), 30% from a bivariate normal distribution with mean (0,-1), and the remaining 20% from a bivariate normal distribution with mean (1,1). In all three normal distributions, the covariance matrix was set to the identity matrix. This leads to a three-cluster pattern, although overall the groups are not easily distinguished (see Figure **??** in Appendix **??**). Next, a $m \times 2$ matrix of latent variable coefficients was generated, with the first column consisting of an evenly spaced ascending sequence from -2 to 2, and the second column consisting of an evenly spaced descending sequence from 1 to -1. Finally, an intercept for each item was simulated from a uniform distribution $U[-1, 1]$. For each true GLLVM, we simulated 1000 datasets.

Six methods for fitting item response models were compared: 1) the VA method in Theorem 1 and assuming a diagonal form for $\boldsymbol{A}_i$, 2) the VA method in Theorem 1 and assuming an unstructured form for $\boldsymbol{A}_i$, 3) the Laplace approximation (**?**), where we wrote our own code to compute the estimates (see supplementary material), 4) the `ltm` function in the `R` package `ltm` (**?**), which uses a hybrid algorithm combining EM and quasi-Newton optimization, with the integration performed using Gauss-Hermite quadrature and the default of 15 quadrature points, 5) the EM algorithm of **?** with the integration performed using fixed point quadrature with 21 quadrature points, and 6) The Metropolis-Hastings Robbins-Monro algorithm (MHRM, **?**). Both methods 5 and 6 are available in the `mirt` function in the `R` package `mirt` (**?**), with their respective default settings used.

Overall, the two VA methods and the Laplace approximation performed best in estimation and prediction (Table 1A). The most telling difference was at $m = 40$ and $n = 50, 100$, where the large number of items relative to the number of observations caused the hybrid, standard EM, and MHRM algorithms to suffer from instability in estimating the coefficients $\boldsymbol{\lambda}$. By contrast, assuming a normal posterior distribution for the $\boldsymbol{u}_i$'s as VA does led to significantly lower mean

Procrustes error for the $\boldsymbol{\lambda}$'s in these settings. The VA method assuming an unstructured form for $\boldsymbol{A}_i$ performed slightly better than the VA method assuming a diagonal form, although we emphasize that the differences in mean Procrustes error between these two versions were minor. Finally, while its performance was similar to the two VA approaches, the Laplace approximation tended to suffer from convergence problems, with updates between successive iterations not always producing an increase in the log-likelihood and there being a strong sensitivity to starting points. Similar convergence problems were also encountered in **?**, who compared the Laplace approximation to several extensions they proposed for estimating GLLVMs, and may be a result of the joint likelihood, i.e. the integrand in equation (2), being far from normally distributed for when the responses are binary.

Table 1: Results for (A) mean Procrustes error (latent variables $\boldsymbol{u}$/latent variable coefficients $\boldsymbol{\lambda}$), and (B) computation time in seconds for simulation Setting 1. Methods compared included the two VA methods assuming either diagonal or unstructured forms for $\boldsymbol{A}_i$, the Laplace approximation, and methods in the `ltm` and `mirt` packages. Computation time includes prediction for the latent variables and calculation of standard errors for the model parameters.

m	n	VA-diag	VA-unstruct	Laplace	`ltm`-hybrid	`mirt`-EM	`mirt`-MHRM
				A: Mean Procrustes error			
	50	0.320/0.136	0.320/0.136	0.305/0.143	0.323/0.394	0.317/0.375	0.314/0.278
10	100	0.317/0.090	0.315/0.089	0.373/0.080	0.328/0.299	0.310/0.184	0.306/0.196
	200	0.278/0.074	0.277/0.076	0.346/0.075	0.311/0.172	0.288/0.093	0.289/0.114
	50	0.145/0.131	0.140/0.116	0.153/0.119	0.213/0.472	0.136/0.400	0.144/0.242
40	100	0.168/0.077	0.161/0.069	0.170/0.072	0.156/0.313	0.160/0.215	0.161/0.197
	200	0.160/0.053	0.150/0.046	0.155/0.053	0.152/0.186	0.152/0.102	0.153/0.088
				B: Mean computation time			
	50	6.56	9.88	8.57	6.69	6.59	19.52
10	100	11.65	19.15	13.27	8.66	7.90	25.08
	200	21.80	33.61	26.71	15.30	9.02	32.07
	50	17.57	41.19	27.84	10.10	82.04	42.98
40	100	27.65	63.30	35.84	17.90	126.79	69.01
	200	61.46	126.90	72.94	29.20	188.42	83.48

With the usual caveats regarding implementation in mind, our implementation of the VA method assuming a diagonal matrix for $\boldsymbol{A}_i$ was slightly faster than the Laplace approximation, with both methods not surprisingly being substantially quicker than the VA method assuming an unstructured $\boldsymbol{A}_i$ (Table 1B). The standard EM algorithm from `mirt` was the fastest method at $m = 10$, but by far the slowest method at $m = 40$. The hybrid EM algorithm also performed strongly in computation time, although it was the worst performer in terms of estimating $\boldsymbol{\lambda}$ (Table 1A). Finally, both VA methods and the Laplace approximation scaled worse than the other methods with increasing n, a result which is not surprising given that these methods introduce an additional set of parameters for each new observation: VA explicitly introduces $(\boldsymbol{a}_i, \boldsymbol{A}_i)$ for each $i = 1, \ldots, n$, while for the Laplace approximation the posterior mode is estimated for each observation.

In addition to the simulation above, we also assessed VA estimation for a larger number of latent variables. Specifically we simulated binary datasets from GLLVMs with $d = 5$ latent variables, with a three-cluster pattern in the latent variables and coefficients generated in a similar manner to the design above. Details are presented in Appendix **??**, and again demonstrate the strong performance of the two VA methods in terms of estimation of coefficients, prediction of latent variables, and computation time.

5.2 Setting 2

We simulated overdispersed count data by modifying one of the models fitted to the birds species dataset (see Appendix **??** for the details of the example) and treating it as a true model. Specifically, we considered a GLLVM which assumed a Poisson-Gamma model, $d = 2$ latent variables, no covariates and included site effects. We then modified it to include two covariates, by generating a $n \times 2$ matrix of covariates with elements simulated from the standard normal distribution, and a corresponding $m \times 2$ matrix of regression coefficients with elements simulated from a uniform distribution $U[-2, 2]$. This modified GLLVM was then treated as the true model. Datasets were simulated with the same number of sites as in the original dataset ($n = 37$) and with a varying the numbers of species, $m = \{30, 50, 100\}$. Since the original dataset consisted of 96 species, then for the cases of $m = 30$ and 50 we took a random sample from the 96 set of species coefficients, while for the case of $m = 100$ we randomly sampled four additional species coefficients for inclusion. Note this simulation setting focused on datasets with m/n close to or exceeding 1 – such wide response matrices are a common attribute of multivariate abundance data in ecology. For each true GLLVM, we simulated 200 datasets.

We compared the following four methods of estimation: 1) the VA method in Theorem 2 and assuming a diagonal form for $\boldsymbol{A}_i$, 2) the VA method in Theorem 2 and assuming an unstructured form for $\boldsymbol{A}_i$, 3) the Laplace approximation (**?**) assuming negative binomial counts, and 2) the Monte Carlo EM (MCEM, **?**) algorithm used in **?** assuming negative binomial counts, where 2000 Monte Carlo samples were used to perform the integration involved in the E-step. Due to its long computation time (see results Table 2), we limited the maximum number of iterations for the MCEM algorithm to 100 iterations. We also considered the three estimation methods assuming Poisson counts, but not surprisingly their performances were considerably worse than assuming overdispersed data, and so their results have been omitted. More generally, we are unaware of any non-proprietary software available for fitting GLLVMs to overdispersed count data.

Table 2: Results for (A) mean Procrustes error (latent variables $\boldsymbol{u}$/latent variable coefficients $\boldsymbol{\lambda}$/covariate coefficients $\boldsymbol{\beta}$) and (B) computation time in seconds for simulation Setting 2. Methods compared included the two VA methods assuming either diagonal or unstructured forms for $\boldsymbol{A}_i$, the Laplace approximation, and the MCEM algorithm. Computation time includes prediction for the latent variables and calculation of standard errors for the model parameters.

m	VA-diag	VA-unstruct	Laplace	MCEM
		A: Mean Procrustes error		
30	0.551/0.802/0.066	0.562/0.797/0.066	0.580/0.807/0.071	0.587/0.807/0.080
50	0.394/0.815/0.070	0.408/0.820/0.070	0.403/0.823/0.073	0.450/0.828/0.074
100	0.274/0.819/0.068	0.295/0.819/0.068	0.291/0.818/0.071	0.335/0.828/0.071
		B: Mean computation time (secs.)		
30	26.53	74.35	75.56	8413.53
50	28.62	63.19	145.07	13905.12
100	53.10	102.18	362.19	26605.92

Overall, the VA method assuming a diagonal form for $\boldsymbol{A}_i$ performed best both in terms of mean Procrustes errors and computation time, followed by the VA method assuming an unstructured form for $\boldsymbol{A}_i$ and the Laplace approximation (Table 2). It should be noted though that, similar to Setting 1, the differences in mean Procrustes error between the two versions of VA were minor. The MCEM algorithm performed worst, having the highest mean Procrustes errors for both the latent variables $\boldsymbol{u}$ and for the covariate coefficients $\boldsymbol{\beta}$, while also taking significantly longer to fit the model than the approximation methods. This dramatic difference in computation time could be attributed to the fact that the M-step in MCEM estimation (effectively) involves fitting models to a dataset of nmB observations, compared to both the VA methods and the Laplace approximation that involve fitting models to a dataset with nm observations. Finally, we note that unlike setting 1, the Laplace approximation did not suffer from any convergence problems here with count response datasets. This was most likely due to the joint likelihood being relatively normally distributed compared to the more discrete, binary response setting.

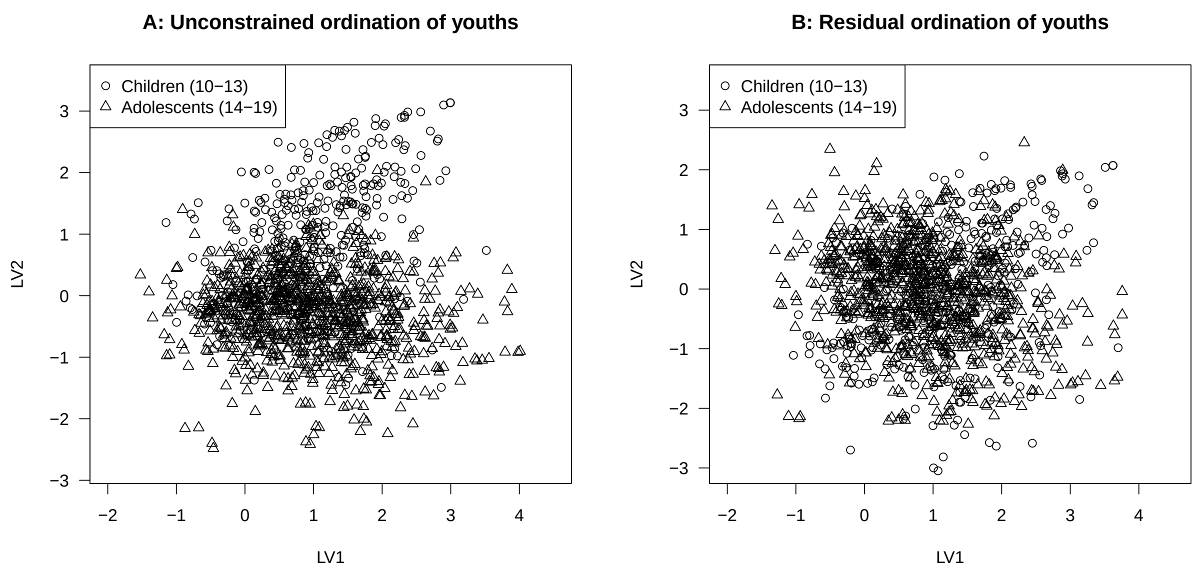
6 Application: Gratitude in Youths

We illustrate the application of the proposed VA method a cross-sectional dataset on several gratitude scales for youths. The dataset is available from the R package psychotools (**?**), and consists of ratings (ordinal responses) on $m = 25$ gratitude scales from $n = 1327$ youths. We also note that the scales have differing numbers of levels, with maximum number of levels ranging from five to nine. The age of each youth (to the nearest integer year) was also available. Details on the psychometric background of the dataset may be found in **?**.

We fitted a GLLVM assuming ordinal responses, $d = 2$ latent variables, and no covariates. We chose to use $d = 2$ latent variables in both examples for the purposes of ordination, to visualize the main patterns between youths of various ages. For the VA method, estimation was performed assuming an unstructured form for the covariance matrix $\boldsymbol{A}_i$; we also considered a diagonal form for $\boldsymbol{A}_i$, and similar results were obtained.

A scatterplot of the predicted latent gratitude scores for each youth ($\boldsymbol{a}_i$) showed a separation between children (10–13 years old) and adolescents (14–19 years old), as seen in Figure 1A. The

Figure 1: Results for the gratitude in youths dataset: (A), unconstrained ordination using a GLLVM with $d = 2$ LVs and no covariates, (B) residual ordination using the same model but with an binary predictor included to differentiate between child versus adolescent. The coordinates for each youth are represented by different symbols, as based on their age classification to child or adolescents.



elements of the estimated coefficient matrix $\boldsymbol{\lambda}$ were all greater than zero except for the second coefficient in five of the gratitude scales, which were significantly less than zero (LOSD 2 to 6; see estimates and standard errors in Table **??** of Appendix **??**). This was not surprising, given these five scales were reverse scored, i.e., a *lower* score reflected a higher sense of gratitude. More importantly though, it indicated that LOSD 2 to 6 were the most effective at differentiating between the levels of gratitude in children versus adolescents.

Given the above results, we therefore constructed a "residual ordination" plot by fitting a GLLVM with the setup as above, except a categorical predictor was now included to indicate whether the youth was a child or adolescent (10–13 versus 14–19 years old). From the resulting fit, the coefficients $\boldsymbol{\beta}$ for this covariate showed adolescents scored significantly higher for LOSD 2 to 6 as well as significantly lower for three other gratitude scales (GAC 1 to 3) compared to children (see Table **??** in Appendix **??**). Moreover, the residual ordination plot no longer presented any substantial pattern for age (Figure 1B), although the lack of any other covariates available in the dataset meant that we could verify whether the residual pattern was perhaps driven by other covariates.

Finally, to assess the goodness of fit for the $d = 2$ model, we performed Monte-Carlo cross-validation, where for each of iteration we randomly sampled 10% of the rows (youths) out to act as a test observations, with the remaining 90% constituting the training dataset. GLLVMs (with no covariates included) ranging from $d = 1$ to 5 were then fitted to each training dataset, using the VA approach, and then the predictive marginal log-likelihood of the test observations was calculated. This procedure was repeated 50 times. Results definitively showed that $d = 1$ latent variables was insufficient, while the predictive performance improved marginally as we transitioned from $d = 2$ to 5 (see Figure **??** in Appendix **??**). This suggested $d = 2$ latent variables was successful in capturing most of the correlation between the responses.

Aside from the above example, we also considered a second dataset comprising counts of bird species collected at sites across Indonesia. Results for this application are found in Appendix **??**. In particular, the design of simulation setting 2 in Section 5.2 was based off this example.

7 Discussion

In this article, we have proposed a variational approximation method for estimating GLLVMs, deriving fully closed form approximations to the log-likelihood for the common cases of binary, ordinal, and overdispersed count data. Estimation is straightforward to implement compared to other methods such as numerical quadrature. The VA approach also returns predictions of the latent variables and uncertainty estimates as part of the estimation procedure. Simulations showed that the VA approach performs similar to or better than some of popular methods used for fitting GLLVMs, with potentially significant reductions in computation time. The `R` code for performing VA estimation of GLLVMs is available in the supplementary material of this article, and in future work we plan to integrate (even faster versions of) these functions into the `mvabund` package (**?**). In this simulations, the VA method performed especially well in settings where m/n is non-negligible. Such data are common in ecology, and thus the VA approach shows a lot of promise for fast fitting of community-level models (such of those of **??**) that also account for inter-species correlation. Since species tend to respond to the environment in rather complex ways however, the VA approach considered in this paper would need to be extended to handle flexible methods of modeling the linear response, e.g. replacing $\boldsymbol{x}_i^T\boldsymbol{\beta}_j$ and $\boldsymbol{u}_i^T\boldsymbol{\lambda}_j$ in (1) with smoothing terms. Many applications of item response theory models assume a discrete instead of continuous distribution for the latent variables, and extending the VA approach to such cases would prove useful not only for psychometrics data, but may also have strong potential in collaborative filtering and latent class models where the datasets are often very high-dimensional (e.g., **??**). Finally, we only offered a heuristic argument for the estimation consistency of the VA estimators for GLLVMs, and substantial research remains to be done to broaden the results of **?** and **?** to show that variational approximations in general produces estimators that are consistent and asymptotically normal, and what these rates of convergence are.

Acknowledgements

FKCH supported by the Australian Research Council discovery project grant DP140101259. ST was supported by the Academy of Finland grant 251965. DIW was supported by an Australian Research Council Future Fellowship (FT120100501).

Supplementary Material

Appendices: Appendix A contains proofs for all theorems and lemmas. Appendix B contains additional simulation results. Appendix C contains additional results for the applications. Appendix D contains the additional application to the birds species count dataset.

R code: The R code for estimating GLLVMs using the VA method and the Laplace approximation, performing simulation Setting 1 and Example 2, and a "readme" file describing each of the files, are contained in a zip file (ms-VAGLLVM.zip).