Model-based ordination with constrained latent variables

Bert van der Veen[123] Francis K.C. Hui[4] Knut A. Hovstad[53]
Robert B. O'Hara[23]

[1]Department of Landscape and Biodiversity, Norwegian Institute of Bioeconomy research, Trondheim, Norway
[2]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway
[3]Centre of Biodiversity Dynamics, Norwegian University of Science and Technology, Trondheim, Norway
[4]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia
[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Summary

1. In community ecology, unconstrained ordination can be used to predict latent variables from a multivariate dataset, which generated the observed species composition.
2. Latent variables can be understood as ecological gradients, which are represented as a function of measured predictors in constrained ordination, so that ecologists can better relate species composition to the environment while reducing dimensionality of the predictors and the response data.
3. However, existing constrained ordination methods do not explicitly account for information provided by species responses, so that they have the potential to misrepresent community structure if not all predictors are measured.
4. We propose a new method for model-based ordination with constrained latent variables in the Generalized Linear Latent Variable Model framework, which incorporates both measured predictors and residual covariation to optimally represent ecological gradients. Simulations of unconstrained and constrained ordination show that the proposed method outperforms CCA and RDA.

keywords: model-based constrained ordination, unimodal response, R-squared, joint species distribution model, reduced rank regression.

Introduction

Unconstrained ordination methods help ecologists to analyse multivariate data of species communities when measurements of the environment are missing. In ordination, species and sites are arranged by their (dis)similarity, so that in unconstrained ordination similarity in environmental conditions at sites can be inferred from species composition. For example, when species preferring wet or dry circumstances are placed at opposite sides of an ordination axis, this axis will often be interpreted to represent a gradient in soil moisture. This approach of inferring the environment of species relationships can be used to generate new hypotheses (Økland 1996). However, unconstrained ordination by design does not facilitate more exact inference of species relationships and environmental conditions at sites.

When environmental conditions are measured, e.g. such as soil moisture or mean temperature, multivariate Generalized Linear Models (MGLM, Wang *et al.* 2012) can be used to provide a more thorough understanding of species-environment relationships. However, as multivariate regression methods relate the response of each species to the predictors, the number of parameters increases rapidly with the number of species and with the number of predictors. In such instances, constrained ordination (also referred to as direct gradient analysis, ter Braak & Prentice 1988) has often been used to analyse community composition data instead. Constrained ordination assumes that an underlying complex ecological gradient can be represented as a linear combination of measured predictor variables, so that the number of parameters related to the predictors scales with the number of complex ecological gradients, and not with the number of species. Constrained ordination describes a class of methods, with two notable ones being Canonical Correspondence Analysis (CCA, ter Braak 1986) and Redundancy Analysis (RDA, Rao 1964), which (also) allow researchers to arrange sites and species by their (dis)similarity. The number of ecological gradients is often considerably less than the number of predictor variables and species (Halvorsen 2012), so that constrained ordination leads to a more feasible and potentially more insightful approach for the analysis of datasets on ecological communities with a large number of predictors and species.

The practical appeal of constrained ordination is immediately apparent in the analysis of species distributions, where bioclimatic predictor variables are often used to represent a species niche (Booth 2018). In such cases, and especially when the response data is sparse, constrained ordination can be used to reduce the number of parameters relative to standard multivariate regression (Yee & Hastie 2003). Since every added predictor variable provides more flexibility in defining the ecological gradient in constrained ordination, with a large number of predictor variables, constrained and unconstrained ordination coincide in their arrangement of sites and species (Jongman *et al.* 1995; McCune 1997; ter Braak & Šmilauer 2015).

Thus, both unconstrained and constrained ordination have their roles in the analysis of ecological com-

munities. All variation in a community can be explored with unconstrained ordination, whereas the variation due to the predictors can be explored with constrained ordination (Økland 1996; ter Braak & Šmilauer 2015). On the other hand, in situations where only a few (relevant) predictor variables are measured, i.e. some important predictors remain unmeasured, constrained ordination has the potential to misrepresent community structure as any variation not explained by the measured predictors is not accounted for in the method (Økland 1996). In turn, this motivates an approach which incorporates both: 1) modelling species responses in a reduced rank form as existing constrained ordination approaches do, and 2) a means of accounting for residual variation not accounted for by the measured predictors, as standard unconstrained ordination approaches do.

Various model-based alternatives to classical constrained ordination methods have been developed in recent years, such as those made available in the `R`-package `VGAM` (Yee & Hastie 2003; Yee 2014), and those in the `R`-package `RCIM` (Hawinkel *et al.* 2019). One of the most well-known is Reduced Rank Regression (RRR, Anderson 1951), which is a model-based approach to constrained ordination that allows users to handle a range of discrete data types, and incorporate both the linear and quadratic responses in the model (Yee 2004). However, similar to the classical constrained ordination methods CCA and RDA, RRR is a purely fixed-effects model that allows for incorporating a residual error through the specification of a response distribution, but not for error and/or residual covariation between species that is associated to an ecological gradient. This means that RRR requires the assumption that the ecological gradient can be perfectly represented by predictor variables. However, in practice it can often be unclear which predictors make up an ecological gradient, so that important predictors may remain unmeasured. As such, there is great potential for residual variation, invalidating the assumption of a perfect fit for the ecological gradient.

In this article we propose a new method for model-based ordination with constrained latent variables, which we believe has the potential to fully utilize the information provided by the measured predictors variables and species responses. In the model-based approach for constrained ordination propose here, a latent variable can be understood as a complex ecological gradient, consisting of both measured and unmeasured components, in contrast to unconstrained ordination where the latent variable always not measured. As such, the proposed model simplifies to an unconstrained ordination when no predictors are measured, or to RRR when there is no residual information left to account for after including predictor variables. The proposed approach builds on the existing framework of Generalized Linear Latent Variable models (GLLVMs, Warton *et al.* 2015), which have seen various developments in unconstrained ordination during recent years (Hui *et al.* 2015; Hui 2016, 2017; Niku *et al.* 2019; Hoegh & Roberts 2020; Damgaard *et al.* 2020; van der Veen *et al.* 2021; Zeng *et al.* 2021). However, GLLVMs still lack an implementation when it comes to applications in constrained ordination. Here, we extend the GLLVM framework for model-based ordination

to the constrained case. Performing constrained ordination in the GLLVM framework allows us to relax the assumption of a perfect fit of predictors to the ecological gradient, so that the latent variables are both a function of the predictors as in the constrained case, and include residual variation provided by the response data as in the unconstrained case. In doing so, this approach allows the latent variables to better represent ecological gradients.

Through a series of simulations based on multivariate normal, presence-absence and count data, we demonstrate that even in the presence of many predictor variables, estimating species responses using RRR can perform just as well if not better than in multivariate regression (e.g., Wang *et al.* 2012) while using fewer parameters. These simulations provide a basis for the evaluation of the dimensionality of community structure and species responses. We additionally compare our proposed GLLVM approach with two popular constrained ordination methods, CCA and RDA, assessing their capability to retrieve the true ecological gradients and species responses in the presence and absence of residual variation and fixed-effects. We show that in the presence and absence of residual variation the proposed GLLVM with constrained latent variables performs similar to, if not better than, CCA and RDA in retrieving the ecological gradients and species responses. Further, when the predictors are unrelated to the ecological gradients, e.g., when the wrong predictors have been measured, the proposes GLLVM with constrained latent variables outperforms CCA and RDA.

Finally, we use two real datasets from species communities to demonstrate use of the proposed GLLVM with constrained latent variables: a dataset of alpine plants on an elevation gradient in Switzerland (D'Amen *et al.* 2017), and a dataset of vascular plants in semi-natural grasslands collected in Norway. An easy-to-use software implementation for model-based ordination with constrained latent variables is available on CRAN in the `gllvm` `R`-package.

Model-formulation

For a multivariate dataset y_{ij} consisting of observations recorded for species $j = 1 \dots p$ and sites $i = 1 \dots n$, the proposed GLLVM with constrained latent variables is defined by the following mean model. Let $g(\cdot)$ generically denote a link function that connects the mean of the assumed response distribution (e.g., the Bernoulli distribution for presence-absence data or the negative binomial for overdispersed counts) to a linear predictor η_{ij}, with a vector $\boldsymbol{x}_{lv,i}$ of $k = 1 \dots K$ measured predictor variables e.g., solar radiation or available cover. Then we formulate the model with $q = 1 \dots d$ constrained latent variables as:

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j, \tag{1}$$

where β_{0j} is an intercept for each species j, and $\boldsymbol{B}$ is a $K \times d$ matrix of slopes per predictor and latent variable. A constrained latent variable can be understood as a complex ecological gradient, of which some components are (un)measured. As such, the vector $\boldsymbol{\gamma}_j$ includes relative species responses to both the measured component of the constrained latent variables $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ and their residual variation $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$, where I_d is a $d \times d$ identity matrix. Note, the predictors $\boldsymbol{x}_{lv,i}$ exclude an intercept term, for reasons of parameter identifiability, due to the presence of species-specific intercepts β_{0j}. More importantly, the model in equation (1) can instead be formulated in terms of a latent variable $\boldsymbol{z}_i = \boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i$, similar to the model form in unconstrained ordination methods (see e.g., Hui *et al.* 2015; van der Veen *et al.* 2021):

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{z}_i)\} = \beta_{0j} + \boldsymbol{z}_i^\top \boldsymbol{\gamma}_j, \tag{2}$$

where $\boldsymbol{z}_i \sim \mathcal{N}(\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$. The models in equation (1) and equation (2) can be straightforwardly extended with row-intercepts to model community composition instead, though we have chosen to omit that term here for ease of presentation.

Without the residual term for the latent variables, equation (1) is an ordinary reduced rank regression (RRR, Anderson 1951), similar to RR-VGLMs implemented in the `R`-package `VGAM` (Yee & Hastie 2003; Yee 2004), or classical constrained ordination methods (ter Braak & Šmilauer 2015). Then, the latent variables are represented only by the fixed-effects term $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ which represents a constrained ordination axis, and serves to reduce the dimensionality of the number of predictors K. Compared to standard multivariate regression, RRR can serve to reduce the number of parameters, as the number of parameters can be especially difficult to estimate for large K and small n. The matrix of coefficients for species $j = 1 \ldots p$ from a multivariate regression $\boldsymbol{\beta}_j$ can be reconstructed from an RRR as $\boldsymbol{\beta}_j = \boldsymbol{B}\boldsymbol{\gamma}_j$. The number of parameters in the model is then $p + d(p + K) - (d + d^2)/2$ for rank d, which can often be a more realistic assumption for ecological community data that tend to be sparse on information. Note that in some cases, the number of parameters in the reduced rank model can exceed the number of parameters used to model species responses in full rank (e.g., when $d = K$), though most commonly we assume $d < K \ll p$ so that rank-reduction is achieved.

When measurements of the environment are missing entirely, or when the predictors are unrelated to the ordination and have slopes close to zero, in essence when $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i} = \boldsymbol{0}$, then the model in equation (1) simplifies to an unconstrained ordination. In the method for model-based ordination with constrained latent variables proposed here, the term $\boldsymbol{\epsilon}_i$ is used to account for the discrepancy between the true latent variable, and the latent variable that can be predicted using the predictor variables alone. In summary, the model proposed here performs simultaneous constrained and unconstrained ordination when predictor variables are

included.

The model in equation (1) can be extended to include additional (separate) predictors, resulting in a partial constrained ordination similar to ter Braak (1988):

$$g\{\mathrm{E}(y_{ij}|\boldsymbol{x}_i, \boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i)\} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\kappa}_j + \boldsymbol{x}_{lv,i}^\top \boldsymbol{B}\boldsymbol{\gamma}_j + \boldsymbol{\epsilon}_i^\top \boldsymbol{\gamma}_j, \tag{3}$$

where $\boldsymbol{\kappa}_j$ are species coefficients for the predictors $\boldsymbol{x}_i$, and where we additionally assume that $\boldsymbol{x}_i$ and $\boldsymbol{x}_{lv,i}$ do not include the same predictor variables for reasons of parameter identifiability. Here, the effect of $\boldsymbol{x}_i$ is excluded from the constrained ordination, and are included so that the resulting ordination is interpreted conditionally on the predictors $\boldsymbol{x}_i$ and species slopes $\boldsymbol{\kappa}_j$.

The models presented so far assume that species respond linearly to the latent variables. However, it is widely acknowledged that species respond to the environment unimodally (see e.g., ter Braak 1987). van der Veen *et al.* (2021) recently presented a method for model-based ordination with quadratic responses, as a means to model species-specific environmental tolerances. All models presented here can be extended in a similar fashion, e.g. with quadratic response to the ecological gradients, which we further elaborate on in Appendix S1.

The model in equation (1) is unidentifiable without additional constraints, due to the freely varying scale parameters $\boldsymbol{\sigma}$ for the latent variables. Consider a matrix $\boldsymbol{\Gamma}$ that includes all species scores $\boldsymbol{\gamma}_j$ as row vectors, for which we fix the upper triangular entries to zero for reasons of parameter identifiability, as is usual for GLLVMs (Hui *et al.* 2015). In standard formulation of GLLVMs, due to scale invariance, the latent variables are assumed to have unit variance. Then, the species slopes $\boldsymbol{\gamma}_j$ additionally serve to determine the scale of the ordination. However, in model-based ordination with constrained latent variables, the species scores are shared for two terms, so that without extra constraints they regulate the relative scale of the second and third term in equation (1). In cases where either the fixed-effects term or the residual term is zero, this requires the model to compensate by increasing the magnitude of the species scores, and e.g., decreasing the magnitude of $\boldsymbol{B}$ if the true fixed-effects term is non-zero but the residual term is not. Therefore, we additionally choose to fix one parameter per latent variable to facilitate including freely varying scale parameters for the latent variables. Here, we choose the diagonal entries of $\boldsymbol{\Gamma}$, such that in $\boldsymbol{\Gamma}$ there are only $(p-d)d + d(d-1)/2$ parameters to estimate. This choice of the diagonal elements is arbitrary, and different elements could be chosen instead. However, the current choice is guided by the magnitude of the different parameters in the model, as now $\boldsymbol{B}$ determines the scale of the first (fixed-effects) term, so that it is (close to) zero when the predictors have no effect on the ordination. Similarly, the vector of residual standard deviations $\boldsymbol{\sigma}$ then determines the scale of the residual term, so that it is zero when there is no residual necessary in the

ordination (i.e. when the predictors perfectly represent the latent variable, as in RRR).

Consequently, the vector of standard deviations for the residual of the latent variable $\boldsymbol{\sigma}$ can additionally be used to determine when latent variables are nearly redundant, or additionally for a measure of (residual) gradient length, or to develop a method of regularization in GLLVMs. In comparison, van der Veen *et al.* (2021) considered the scale of the latent variables relative to the median tolerance of species curves being one. However, this has no meaning in models with the linear responses, and is difficult to implement in practice. Finally, it is important to note that this choice of identifiability constraint does not diminish the overall flexibility of the model, but merely clarifies the interpretation of the parameters: in essence, the latent variables are stretched or contracted so that certain species scores equal one.

Parameter estimation and model fitting

Since the proposed method is a type of GLLVM, we are required to choose an appropriate distribution to model the species observations (and their associated mean-variance relationship, see e.g., Warton & Hui 2017). For example, a Poisson or negative binomial distribution with log-link function for counts, a Binomial distribution with probit link-function for presence-absence data, or alternatively a Tweedie distribution with log-link function for biomass data. Similar to other GLLVMs proposed for model-based ordination in the literature, the residual error terms $\boldsymbol{\epsilon}_i$ are assumed to be normally distributed random variables, which thus needs to be integrated over. Consequently, the marginal log-likelihood of the proposed GLLVM with constrained latent variables as in (1) is written as:

$$\mathcal{L}(\Theta) = \sum_{i=1}^{n} \log\left\{ \int \prod_{j=1}^{p} f\left(y_{ij} | \boldsymbol{x}_i, \boldsymbol{\epsilon}_i, \Theta \right) h\left(\boldsymbol{\epsilon}_i \right) d\boldsymbol{\epsilon}_i \right\}, \tag{4}$$

where $f(y_{ij}|\boldsymbol{x}_{lv,i}, \boldsymbol{\epsilon}_i, \Theta)$ is the distribution of the responses conditional on the predictors $\boldsymbol{x}_{lv,i}$, the constrained residual error term $\boldsymbol{\epsilon}_i$, and a vector Θ. The vector Θ includes all parameters and an estimate of $\mathrm{h}(\boldsymbol{\epsilon}_i|y_{ij})$ with related variational parameters if applicable. The residual error terms are assumed to follow a multivariate normal distribution $h(\boldsymbol{\epsilon}_i) = \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 \mathrm{I}_d)$. The integration can be performed with methods previously developed for estimation and inference in GLLVMs, such as the Laplace approximation (Niku *et al.* 2017) or Variational Approximations (VA, Hui *et al.* 2017; van der Veen *et al.* 2021) and using Template Model Builder (Kristensen *et al.* 2016). Below, the models in the simulation studies and examples are fitted using VA.

Initial values

Both with and without residual error term, the algorithm used to fit the models presented here is sensitive to the initial values. In this article, we adapt the approach used in the `gllvm` `R`-package to overcome this, and obtain reasonable starting values. Specifically, for ordinary RRR, we followed a similar procedure to that described by Files *et al.* (2019), where we generate starting values for $\boldsymbol{B}$ and $\boldsymbol{\gamma}_j$ by first fitting a multivariate linear model with predictor variables to the Dunn-Smyth residuals (Dunn & Smyth 1996) of an intercept-only MGLM. We then performed a QR-decomposition on the matrix of regression coefficients to obtain the starting values for $\boldsymbol{B}$ and $\boldsymbol{\gamma}_j$. For constrained latent variables, we performed a factor analysis on the Dunn-Smyth residuals of an intercept-only MGLM, and then regress the estimated factor scores to receive initial values for the predictor slopes $\boldsymbol{B}$. The residuals of the regression provided initial values for $\boldsymbol{\epsilon}_i$, and the loadings from the factor analysis were taken as the initial values for $\boldsymbol{\gamma}_j$.

Inference

In this section we present various tools for inference and prediction for the proposed method of model-based ordination with constrained latent variables.

Constrained ordination diagram

Separate ordination diagrams can be constructed for both terms in equation (1) to explore species relationships for the predictors and for residual variation, or an ordination diagram can be constructed including both terms to present species co-occurrence patterns within a single plot. In GLLVMs with unconstrained latent variables, we can obtain predictions of the residuals $\boldsymbol{\epsilon}_i$ using e.g., the means of variational distributions (Hui *et al.* 2017) or the maximum a-posteriori prediction from the Laplace approximation (Niku *et al.* 2017). Note however these latent variables are assumed to fully consist of residual information as discussed previously. In contrast, with model-based ordination using constrained latent variables, we instead consider using the predicted site scores $\boldsymbol{z}_i$ for ordination, which can be constructed based on the predicted $\boldsymbol{\epsilon}_i$'s along with the estimated value for $\boldsymbol{B}$. We consider these site scores similar to the weighted average (WA) scores (Palmer 1993; McCune 1997) provided by classical constrained ordination methods such as CCA or RDA, since the residual term $\boldsymbol{\epsilon}_i$, accounts for variation in the response not explained by the predictor variables. As such, $\boldsymbol{B}^\top \boldsymbol{x}_{lv,i}$ are linear predictor (LC) scores, that do not include additional information on the latent variable provided by the response data (Palmer 1993). Note RRR always only includes LC scores, which are not generally recommended for inference by community ecologists (McCune 1997).

A constrained ordination diagram of latent variables including both the effects of the predictor variables, and $\boldsymbol{\epsilon}_i$ will, in many instances, provide a similar ordination as when latent variables are assumed to be unconstrained. On the other hand, constrained ordination allows the predictor effects to be represented in an ordination diagram, in the form of arrows based on the rows of $\boldsymbol{B}$. The length of the arrow is proportional to the magnitude of the parameter estimate, so that the predictor with the largest estimate is presented as the longest arrow, although note that we correct the arrow length using the standard deviation of each predictor (see e.g., Figure 3 below). Statistical uncertainty of the slope estimates for the predictors can be further represented using the colour of the arrows, for example by colouring the arrow less intensely for predictor slope estimates for which the corresponding confidence interval includes zero for at least one of the ordination dimensions.

In an ordination diagram, the predicted site scores are plotted to represent (dis)similarity between sites in an ordination. Furthermore, Niku (2020) constructed prediction regions using the Conditional Mean Squared Error of Predictions (CMSEPs, Booth & Hobert 1998) to represent the statistical uncertainty of the site scores in an ordination diagram. To fully and properly convey confidence in the dissimilarity of sites, we adopt the same approach, but adapt the calculation for the case of constrained ordination (see Appendix S2 for details of the calculation). These prediction intervals can be used to provide a larger degree of certainty in the dissimilarity of sites, and tend to be larger than in unconstrained ordination, as they represent both the uncertainty of the fixed-effects and of the residual error term.

Model selection

As the number of predictors increases, and flexibility is added in the modelling of the site scores and species responses, the standard deviations of the residual term $\boldsymbol{\sigma}$ are likely to get smaller. Determining the optimal number of latent variables and the most suitable predictor variables for a constrained ordination is thus an important problem for our method, although it can be a challenging exercise as the number of potential models may be quite large. In the `vegan` `R`-package various tools are available to find the combination of predictor variables that optimally represents the latent variable, such as stepwise selection using permutation P-values or an adjusted $R^2_{\boldsymbol{B}}$ (Oksanen *et al.* 2020). As a model-based approach, we can leverage conventional methods such as hypothesis testing, information criteria (Burnham & Anderson 2002), residual diagnostics (Hartig 2021) among others for assessing the optimal number of latent variables and predictors, predictions, as well as assessing other model assumptions such as the distribution of the responses. For example, the importance of predictors in a model-based ordination with constrained latent variables can be assessed with use of a Wald-statistic and associated p-values, or with confidence intervals. We illustrate an example of

determining predictor importance later on in our applications of two real datasets.

Similarly, the question of whether to perform constrained or unconstrained ordination, and whether LC scores or the model-based equivalent of WA scores are more suitable for representing the ecological community, can be solved using model-selection tools such as information criteria.

Predictor importance

Similarly to other GLLVMs, the residual covariance matrix associated with the latent variables can be calculated (see Appendix S3) to examine species associations and determine the residual variation in the response, beyond that due to the measured predictors. By fitting a second unconstrained model, the variation explained by the predictors in the response can also be determined, similar to the approach presented by Warton *et al.* (2015) based on relative differences in the trace of the residual covariance matrix. Here however, with model-based ordination using constrained latent variables, we focus on determining the importance of predictors in explaining the latent variables. Since the latent variables are by definition unmeasured, calculating importance of the predictors through e.g. a partial $R^2_{\boldsymbol{B}}$ as in ordinary linear regression, is not directly possible. As such, to assess the importance of predictors in explaining the latent variables, we adopt an approach similar to that presented by Edwards *et al.* (2008), which also avoids having to fit a second model (which can be computationally intensive for a large number of sites n and species p). Specifically, Edwards *et al.* (2008) developed a measure of $R^2_{\boldsymbol{B}}$ for linear mixed-effects models based on the fit of a single model, which Jaeger *et al.* (2017) extended to the generalized linear mixed-effects model and implemented in the `r2glmm` `R`-package, using a multivariate Wald-statistic for the testing of fixed-effects. Thus, the proportion of (generalized) variance explained by all predictors for all latent variables is:

$$R^2_{\boldsymbol{B}} = \frac{\omega \mathrm{vec}(\hat{\boldsymbol{B}})^\top (\hat{\boldsymbol{\Sigma}})^{-1} \mathrm{vec}(\hat{\boldsymbol{B}})(dK)^{-1}}{1 + \omega \mathrm{vec}(\hat{\boldsymbol{B}})^\top (\hat{\boldsymbol{\Sigma}})^{-1} \mathrm{vec}(\hat{\boldsymbol{B}})(dK)^{-1}}, \tag{5}$$

where $\mathrm{vec}(\hat{\boldsymbol{B}})$ is a vectorized version of the matrix of estimated predictor slopes $\hat{\boldsymbol{B}}$ with corresponding estimated covariance matrix $\hat{\boldsymbol{\Sigma}}$, and ω is a ratio based on the residual degrees of freedom: $\frac{dK}{dn-dK}$, for the total number of sites n, the number of constrained latent variables d, and the number of predictors K. Note this $R^2_{\boldsymbol{B}}$ can also be calculated on a per predictor variable basis (with numerator degrees of freedom d), or per latent variable and predictor (with unit numerator degrees of freedom), to retrieve a semi-partial $R^2_{\boldsymbol{B}}$. This semi-partial $R^2_{\boldsymbol{B}}$ is interpreted as the capability of that predictor to explain the (generalized) residual variation unaccounted for after accounting for all other predictors in the full model (Edwards *et al.* 2008).

To summarize, a high semi-partial $R^2_{\boldsymbol{B}}$ indicates importance of a predictor in explaining the latent variables. We demonstrate use of the semi-partial $R^2_{\boldsymbol{B}}$ in the real data examples below.

Simulation studies

We performed two separate simulation studies for the proposed GLLVM with constrained latent variables: 1) we simulated from a MGLM (i.e. with full rank species responses) using bioclimatic predictor variables, and compared the accuracy of RRR (i.e. model-based ordination with constrained latent variables and $\boldsymbol{\epsilon}_i = \mathbf{0}$) and a MGLM to retrieve the true species responses, and 2) we simulated unconstrained and constrained ordinations to compare the capability of the proposed GLLVM to retrieve the true latent variables $\boldsymbol{z}_i = \boldsymbol{B}^\top \boldsymbol{x}_{lv,i} + \boldsymbol{\epsilon}_i$ and species loadings $\boldsymbol{\gamma}_j$, in comparison to the WA scores from CCA and RDA. Since it is more difficult to accurately predict the latent variables when the number of species is small, and since more sites provide more information to estimate species responses, in both simulations we included few species but more sites. `R`-code for the simulation studies is included in Appendix S4.

In our first simulation study, to get a realistic collinearity structure in the predictor variables, we simulated 1000 random points across the European Union, Switzerland and Norway, using the `sp` `R`-package (Pebesma & Bivand 2005), at which we retrieved 19 bioclimatic variables using the `raster` `R`-package (Hijmans 2020). Afterwards, we simulated $K = 19$ bioclimatic predictor variables at $n = 100$ new sites, from a multivariate normal distribution with a zero mean vector and the covariance matrix set equal to the sample covariance matrix of the 1000 random points. We then standardized the predictor variables to have mean zero and variance one. We removed one predictor that was almost fully collinear with another predictor, as it resulted in numerical issues, so that the final number of predictors was $K = 18$. We then used the extended hunting spider dataset with $n = 100$ sites and $p = 12$ species from van der Aart & Smeek-Enserink (1975), to fit a MGLM with a Poisson distribution and log-link for the count responses and including the 18 predictor variables above, and subsequently used the estimated species-specific slopes from this model as the true slopes in our simulation. Additionally, we simulated species-specific intercepts from the uniform distribution Uniform$(-1, 1)$ corresponding to species of low abundance or occurrence. We simulated 1000 datasets assuming either Gaussian responses, Poisson counts or Bernoulli presence-absence responses. The variance associated with Gaussian responses was assumed to be one. For each simulated dataset, we fitted a MGLM along with RRR with 2-8 latent variables in equation (1). The rank of the matrix of species responses can maximally be $\min(p, K)$, so that it was 12 here. The MGLMs included the same number of parameters as the true model, namely 228 parameters, whereas the fitted RRR models included 69 parameters for $d = 2$, 96 for $d = 3$, 122 for $d = 4$, and 147 for $d = 5$, 171 for $d = 6$, 194 for $d = 7$, and 216 for $d = 8$. We did not include a rank 9 model in our example, as it would have exceeded the number of parameters in the MGLM (237 parameters versus 228 parameters). Finally, we calculated the symmetric Procrustes error between the true and and estimated matrix of species responses for each simulated dataset, as to compare accuracy of

the methods in retrieving species responses, using the `vegan` R-package (Oksanen *et al.* 2020).

The results of the first simulation study are summarized in Figure 1, which show that as the number of latent variables in the RRR increased, the accuracy of the estimated species responses improved. At rank 5, the accuracy was similar to that of the MGLM, yet the number of parameters included in the model was about 65% of that of the true model. In general, the estimated species responses were accurately estimated across all distributions, though for Bernoulli responses the error was generally higher than for Gaussian or Poisson responses.

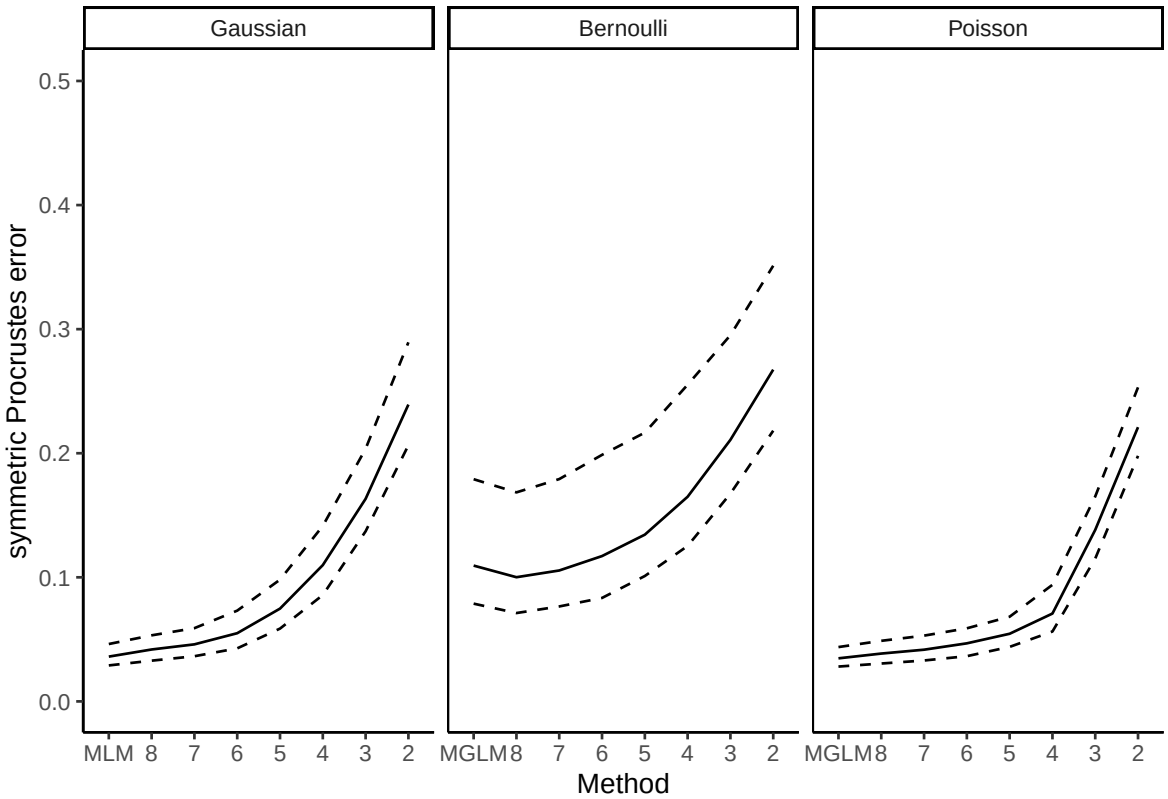

Figure 1: Simulation results for the 1000 multivariate GLMs and various reduced rank regressions fitted to Gaussian, Bernoulli and Poisson response datasets simulated from a multivariate GLM, with $n = 100$ sites and $p = 12$ species. The true model contained 18 bioclimatic predictor variables with a realistic degree of collinearity. The symmetric Procrustes error of the estimated and true species responses is shown on the y-axis. Numbers 2-8 indicate reduced rank models with that rank. The solid line represents the median symmetric Procrustes error, and dashed lines the first and third quartiles.

For the second simulation study, we considered datasets with $n = 100$ sites and with $p = 30$ species. We simulated three forms of the model in equation (1): with 1) non-zero predictor slopes $\boldsymbol{B}$ and residuals $\boldsymbol{\epsilon}_i$, 2) non-zero predictor slopes without the residuals, i.e. RRR, and 3) with $\boldsymbol{B} = \boldsymbol{0}$ and non-zero residuals, i.e. an unconstrained ordination.

To construct the true model, we first simulated $K = 5$ predictor variables following a multivariate

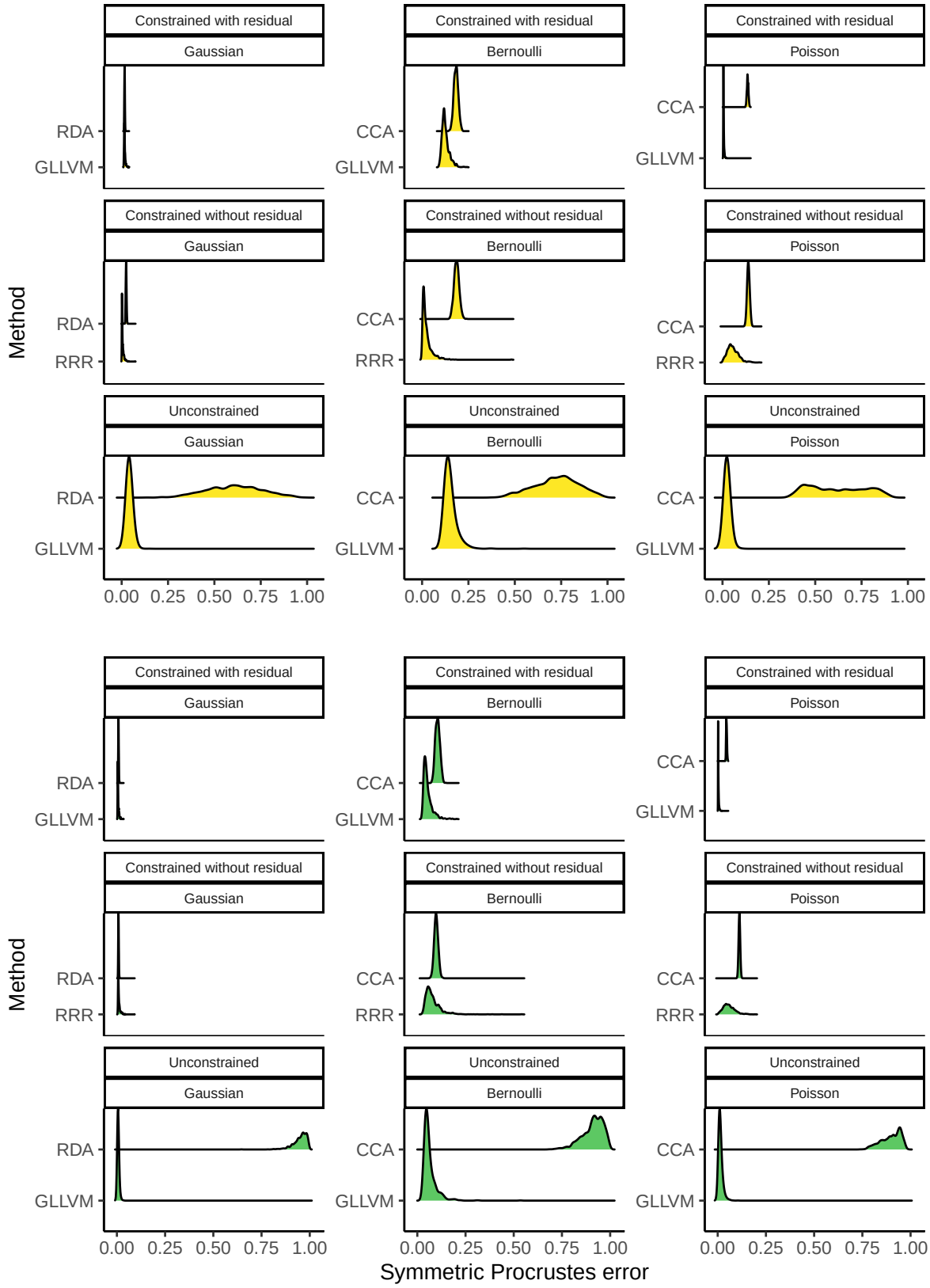

Figure 2: Results for ordination methods fitted to 1000 simulated datasets with Gaussian, Bernoulli, and Poisson response datasets. Simulations labled 'unconstrained' followed the same true model as simulations labled 'constrained', but instead slopes for the predictor variables were fixed to zero i.e., $\boldsymbol{B} = \mathbf{0}$. For simulations labled 'constrained without residual', the true model was the same as that of 'constrained', but without the residual variation i.e., $\boldsymbol{\epsilon}_i = \mathbf{0}$. The procrustes error of the latent variables $\boldsymbol{z}_i$ (yellow) and species scores $\boldsymbol{\gamma}_j$ are presented.

standard normal distribution. Next, we generated the true slope coefficients $\boldsymbol{B}$ by applying a factor analysis to the simulated predictor variables, with two dimensions. We simulated the true intercepts β_{0j} from Uniform$(-1, 1)$, and species coefficients for the latent variables $\boldsymbol{\gamma}_j$ independently from Uniform$(-2, 2)$. Finally, we simulated the constrained residual error $\boldsymbol{\epsilon}_i$ by first sampling from a bivariate standard normal distribution, after which we regressed the sampled realization against the simulated predictor variables, and used the residual from the regression as the residual error in the true model. This ensures that the true residual error $\boldsymbol{\epsilon}_i$ was independent of the simulated predictor variables by construction. We simulated 1000 datasets each from the Gaussian, Bernoulli, and Poisson distributions. The variance associated with Gaussian responses was again assumed to be one. To each dataset, we fitted a GLLVM with two constrained latent variables while CCA was fitted to the datasets with Bernoulli and Poisson responses, and RDA to the datasets with Gaussian distributed responses. Both classical ordination methods were fitted using the `vegan` `R`-package, which we also used for the calculation of a symmetric Procrustes error between the simulated latent variables and the latent variables retrieved from the proposed GLLVM, CCA, and RDA, and the same for the species scores (Peres-Neto & Jackson 2001; Oksanen *et al.* 2020). The results are summarized in Figure 2.

In general, for the proposed GLLVM with constrained latent variables, with and without residual (i.e. RRR), but also without fixed-effects (i.e. unconstrained ordination), we consistently and with little variability managed to retrieve the true latent variables $\boldsymbol{z}_i$ and species loadings $\boldsymbol{\gamma}_j$. For constrained ordination and for Gaussian distributed responses, RDA compared to GLLVMs and RRR performed similarly. However, when RDA was fitted to datasets where the predictor variables had no relation to the latent variables, it was unable to retrieve the latent variables or species loadings. In all cases, GLLVMs and RRR performed better than CCA. Similarly to RDA, CCA was not able to retrieve the latent variables or the species loadings if the predictor variables had no relation to the latent variables.

Worked examples

We demonstrate applications for the proposed GLLVM with constrained latent variables on two ecological datasets: 1) a dataset of Swiss alpine plants (D'Amen *et al.* 2018), and 2) a dataset of vascular plants collected in Levanger, Norway.

Swiss alpine plants

The first example includes a presence-absence dataset of alpine plants, collected in the western Swiss Alps. The dataset was collected on a strong elevation gradient, including sites in both lowland and alpine envi-

ronments (D'Amen *et al.* 2018). In total the dataset includes $n = 791$ plots and $p = 175$ plant species, after excluding plots with fewer than two observations and species with less than three presence observations (D'Amen *et al.* 2018). Six predictor variables were included in the study: degree days above zero, slope, moisture index, total solar radiation over the year, topography, and elevation. All predictors were standardized to have mean zero and variance one.

Fitting a range of models while testing for the optimal number of dimensions and predictors would be computationally burdensome and time consuming, so for demonstration purposes we here fit a model with $d = 2$ latent variables and assuming a Bernoulli distribution for the responses, using all predictors, and with one quadratic coefficient per latent variable. Previously, van der Veen *et al.* (2021) found that using a quadratic response model lead to better predictions of the ecological gradient for this dataset, so we adopt that approach here. We then base our inference of predictor importance on the confidence interval of predictor slopes i.e., a Wald-statistic with accompanying p-values, and the approach for $R^2_{\boldsymbol{B}}$ presented above.

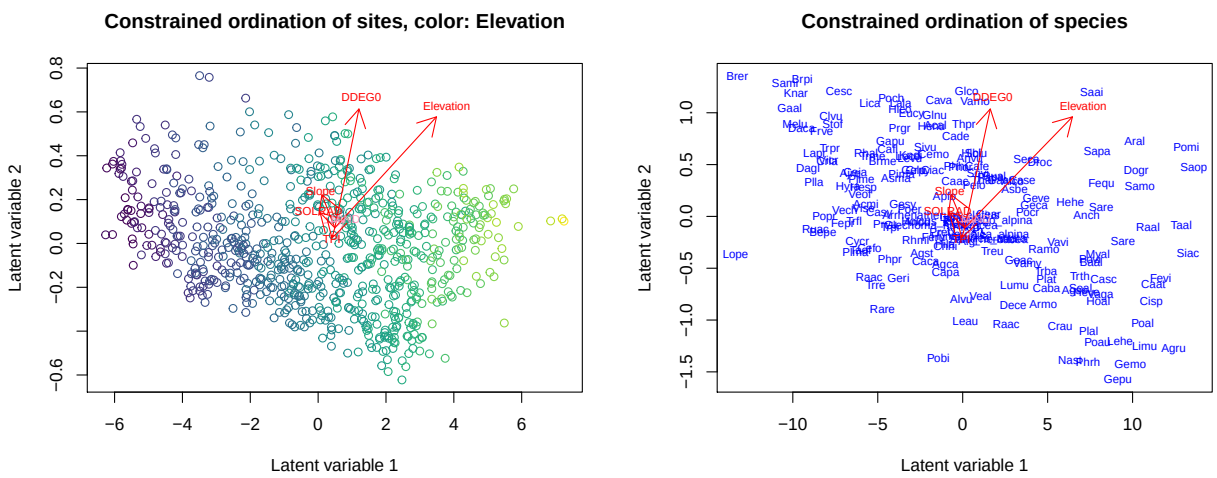

Figure 3: Constrained ordination diagram for the Swiss alpine plants data. Darker colors in the left plot plot indicate low elevation, whereas lighter colors indicate sites at higher elevation. Arrows represent predictor slopes for each latent variable, with arrow length being proportional to the magnitude of the slope estimate. Arrows shown in pink represent slope estimates of which the confidence intervals included zero for one of the dimensions. The latent variables have been rotated to principal direction, so that the first latent variable explains most variation. In the right plot, the abbreviated species names represent optima from the quadratic response model, which are drawn as arrows if they are too far removed from the latent variable. Detailed results, and a list of species names, are included in Appendix S5: Table S1, Table S2.

The results from the constrained ordination are summarized in Appendix S5: Table S1. Similar to van der Veen *et al.* (2021), elevation and degree days above zero were the two predictors most related to the two predicted latent variables. For the first latent variable, the slope for the elevation predictor was four times

as large as the slope for degree days above zero. For the second latent variable, the magnitude of the slopes for elevation and degree days above zero was similar. The results are visually presented in Figure 3.

Based on the fitted GLLVM with constrained latent variables containing all six predictor variables, the residual standard deviations were 0.00 (95% confidence interval: -0.02, 0.02) and 0.34 (0.32, 0.36), indicating that the first latent variable could be fully represented by the predictors. The $R^2_{\boldsymbol{B}}$ for the linear regressions of the latent variables was 0.34. The semi-partial $R^2_{\boldsymbol{B}}$ for the predictors was 0.08 (degree days above zero), 0.14 (slope), 0.02 (moisture index), 0.06 (solar radiation), 0.04 (topography), and 0.15 (elevation), indicating that elevation and slope were most important for representing the latent variables. Our results overall were similar to those for CCA, where the largest correlation of a predictor with the first axis was for elevation, and for the second axis slope (see also Appendix S5: Figure S1).

Semi-natural grasslands in Norway

The second example contains observations of vascular plants collected at Levanger, Norway in 2008. In total $n = 132$ plots of $1m^2$ were pseudo-randomly positioned across five "zones". For each zone, coordinates were sampled randomly, and plants were recorded if the coordinate was located in a semi-natural grassland. We excluded sites for which there were no soil measurements available, so that the final dataset included $n = 116$ sites. In total, the dataset included observations of $p = 132$ vascular plants, but we chose to exclude species with fewer than 3 observations, so that the final dataset included observations of $p = 64$ vascular plant species. Some of the grasslands were grazed by sheep and cattle, whereas others were abandoned and had no management. The study area was approximately five kilometres long and four kilometres wide, and located along a ridge with an east-west direction. The collected data is in the form of percentage cover per species in a plot, so that the total coverage of all plants in a plot can exceed 100 percent. In total, the dataset included 36 different predictor variables of various water-soluble soil nutrients such as pH, phosphorus, potassium, calcium, and organic matter content for the first 0 - 10 centimetre of soil and additionally for the next 10 - 20 centimetre layer. Though the soil variables for 0 - 10- and 10 - 20 centimetres soil depth to some degree are collinear, they represent different aspects of the edaphic site conditions; at 0 - 10 centimetre deep the soil properties are more affected by current conditions such as management and present vegetation. The results from soil samples at 10 - 20 centimetres depth were included to represent the mineral content of deeper soil layers less influenced by present management and vegetation. Below, soil variables for the first 0 - 10 centimetres are indicated by a one, and for the 10 - 20 centimetres layer with a two. Various other predictor variables were also recorded, such as whether a plot was grazed or abandoned, the cover of trees, shrubs, litter, and the height of the different layers, slope, and aspect. For demonstration purposes, we focus

on the soil property variables, and on the effects of grazing, so that the final number of predictors in the constrained ordination was $K = 10$.

We fitted a model with two constrained latent variables, assuming a Tweedie distribution with power parameter 1.3, as that seemed to provide the best fit as determined from re-fitting and examining of residual plots. A Tweedie distribution has the potential to predict percentages larger than 100 here, and as such it might be more realistic to fit the model using a beta distribution. In contrast, the Tweedie will provides a more flexible mean-variance relationship than the beta distribution. Additionally, the data here include zeros, which are not possible to include with a beta distribution. To demonstrate the possibility of additional predictors, as per equation (3), we included whether a plot was grazed or not as an additional fixed effect. Thus, the ordination will be conditional on the effect of grazing, so that by colouring sites based on this predictor, it should then become clear that the effect has been accounted for outside of the ordination. More details from the constrained model for the Levanger grasslands dataset, including the effect of grazing on individual plant species, and a table of estimates and standard errors for the ordination, are included in Appendix S5: Table S3, Figure S3.

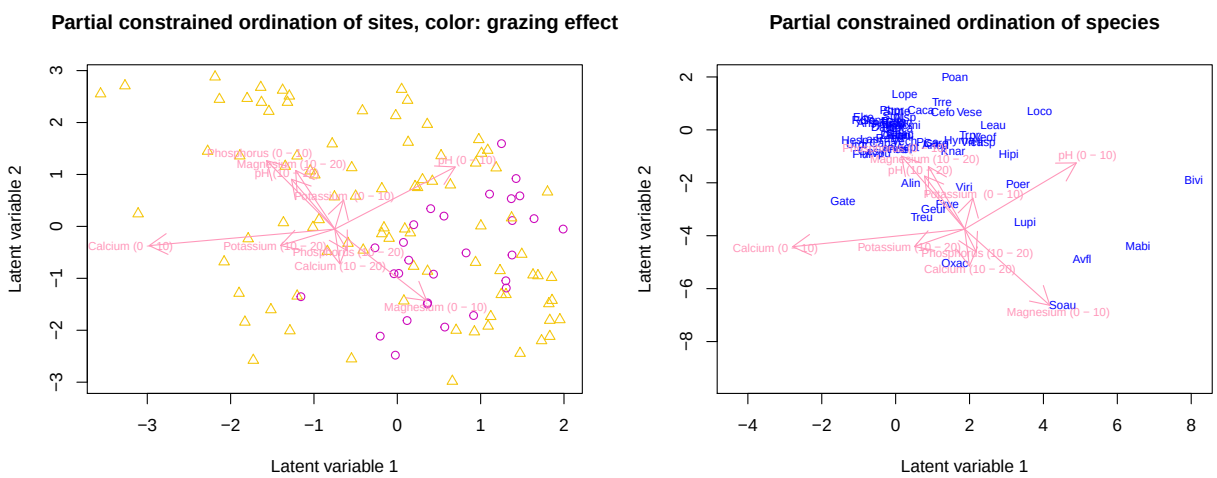

Figure 4: Constrained ordination of sites (left) and species (right) from the Levanger grasslands dataset. The constrained ordination is conditional on the effect of grazing, so that the effect excluded from the ordination diagram. To emphasize this, sites have been marked by their classification: yellow triangles are the reference plots, whereas purple circles indicate grazed plots. By conditioning on the effect of grazing, this constrained ordination now solely focuses on arranging sites (and potentially species) based on soil properties. Estimates and standard errors of the effects are provided in Appendix S5: Table S3, though none of the effects were statistically significant for both dimensions. A list of species names is included in Appendix S5: Table S4.

The residual standard deviations of the constrained latent variables were 1.11 (95% confidence interval: 0.41, 1.82) and 0.78 (0.17, 1.40). The predictors poorly explained the latent variables, as indicated by the low $R^2_{\boldsymbol{B}}$ for the latent variables of 0.08, so that the semi-partial $R^2_{\boldsymbol{B}}$ of all predictors was close to zero, but

largest for calcium. For CCA, the largest correlation for the first axis was for pH (0 - 10) and Ca (10 - 20), and K (10 - 20) for the second axis. The variation that could be explained by the predictors was less than 18% of the overall variation (see also Appendix S5: Figure S2).

In the light of the above results with $R^2_{\boldsymbol{B}}$, it was not surprising that most of the estimated predictor slopes i.e., elements of $\boldsymbol{B}$, were accompanied by a large statistical uncertainty, so that no predictor estimates were significantly different from zero for both latent variables simultaneously (see Appendix S5: Table S3). For the first dimension, the effect of pH at soil depth 0 - 10 centimetres was statistically significant, as well as the effect of calcium. For the second dimension, the effects of potassium (0 - 10 cm) and magnesium (0 - 10 cm and 10 - 20 cm) were statistically significant. The grazed and abandoned sites occupy a similar space in the ordination diagram (Figure 4), as was expected when performing a partial constrained ordination.

Discussion

In this article, we present a new method for model-based ordination with constrained latent variables, or alternatively for estimating species responses in a reduced-rank form, using the GLLVM framework (Warton *et al.* 2015). Constrained ordination allows ecologists to order sites and species using measured predictors, and as such to better examine species-environment relationships (Ter Braak 1987). When there is no effect of the predictors, the model proposed here simplifies to an unconstrained ordination. Similarly, without residual term for the latent variables and with predictor variables, the model simplifies to a reduced rank regression (RRR) and is similar to the popular constrained ordination methods CCA and RDA.

In studies of species distribution modelling, the bioclimatic variables often used are collinear (Júnior & Nóbrega 2018), so that it may not always be possible to accurately estimate all parameters. In the first simulation study, performed competitively with MGLMs in estimating species responses, even while it included fewer parameters. Future research on the use and performance of RRR could attempt to further explore the dimensionality of species communities, which we here found to be lower than assumed by a MGLM. Though most ecologists consider between two and four dimensions for ordination (Halvorsen 2012), our results imply that at least five dimensions are necessary to accurately estimate species responses. However, the dimensionality of species responses might differ between communities based on e.g., the number of environmental gradients that underlay the structure of a community (see also Tobler *et al.* 2019).

In classical constrained ordination methods such as CCA and RDA, latent variables are assumed to be perfectly represented by predictor variables, so that those methods do not account for residual information unaccounted for by the predictors (ter Braak & Šmilauer 2015). In contrast, the approach presented here is capable of simultaneous unconstrained and constrained ordination though it requires having measured

predictors, unlike in model-based unconstrained ordination (Hui *et al.* 2015). In the second simulation study in this article, we showed that if the predictors are unrelated to the true latent variables, so that the true model is that of an unconstrained ordination, CCA and RDA are unable to retrieve the latent variables and species responses, in contrast to the proposed GLLVM. In reality, it can often be unclear which predictors represent the latent variables, so that accounting for additional residual information can be important. Though CCA and RDA do not account for residual information in the ordination explicitly, our results suggest that the use of WA scores sufficiently mitigates that deficiency. WA scores can be considered minimally constrained, unlike LC scores (Palmer 1993), which have shown to not be sufficiently robust for inference (McCune 1997). Accounting for residual information from species responses addresses the concern shared by community ecologists over discarding ecological gradients that the ecologist is unaware of (Palmer 1993; Økland 1996; McCune 1997; ter Braak & Šmilauer 2015).

We demonstrated how to apply model-based ordination with constrained latent variables using two example datasets, one of Swiss alpine plants (D'Amen *et al.* 2017) and another of vascular plants in Norway. In those instances, the residual variation unaccounted for by the predictor variables demonstrated the need to account for residual variation in community ecological studies using dimension reduction techniques. We assessed importance of the predictors in the constrained ordination using a semi-partial $R^2_{\boldsymbol{B}}$ (Edwards *et al.* 2008), which can be calculated for our proposed model-based constrained ordination irrespective of whether the residual term is included or not (though omitting the residual term will affect the magnitude of the $R^2_{\boldsymbol{B}}$ statistic). In the first example, elevation and slope were the most important predictors in explaining the latent variables for the Swiss alpine plants dataset according to their semi-partial $R^2_{\boldsymbol{B}}$ values, and calcium for the Levanger grasslands dataset. For the Levanger grasslands dataset, the measured soil variables included as predictors did not provide a good representation of the latent variables, as indicated by a low $R^2_{\boldsymbol{B}}$ value for the entire model and by the parameter estimates and their corresponding Wald-tests for the predictor slopes $\boldsymbol{B}$.

Model-based ordination with constrained latent variables provides a suitable alternative for the modelling of multiple species responses with or without residual term or other random-effects (such as random row-intercepts). Our proposed approach provides access to standard tools for statistical inference such as statistical uncertainties for parameters estimates, model-selection tools, p-values related to a Wald-statistic that can be used to determine significance of the effect of predictors, and more such as residual diagnostics, all of which are available as part of the `gllvm` `R`-package (Niku *et al.* 2020), including a vignette that demonstrates the use of the proposed method. To conclude, the method presented here provides an extended version for various types of multivariate analyses, including fixed-effects constrained ordination, partial constrained ordination, MGLMs, and in general has merit for the ordering of sites and species.

Acknowledgements

The authors would like to thank Synnøve Nordal Grenne, Liv S. Nilsen and Line Rosef, who together with Knut Anders Hovstad collected the Levanger grasslands dataset, which is available on Figshare to download (https://doi.org/10.6084/m9.figshare.15143937.v1). The authors would like to thank Cajo ter Braak for providing the extended hunting spiders dataset, used in the first simulation study. Jenni Niku kindly helped improving the software implementation for the `gllvm` `R`-package. B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H was supported by an Australia Research Council Discovery Fellowship (grant number DE200100435).

Authors contributions

B.V. conceived the ideas. B.V. and F.K.C.H designed the methodology. All authors contributed to the writing, reviewing and editing of the draft and gave final approval for publication.

References

Anderson, T.W. (1951). Estimating Linear Restrictions on Regression Coefficients for Multivariate Normal Distributions. *The Annals of Mathematical Statistics*, **22**, 327–351. Retrieved June 29, 2021, from https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-22/issue-3/Estimating-Linear-Restrictions-on-Regression-Coefficients-for-Multivariate-Normal-Distributions/10.1214/aoms/1177729580.full

Booth, T.H. (2018). Why understanding the pioneering and continuing contributions of BIOCLIM to species distribution modelling is important. *Austral Ecology*, **43**, 852–860. Retrieved June 29, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1111/aec.12628

Booth, J.G. & Hobert, J.P. (1998). Standard Errors of Prediction in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, **93**, 262–272. Retrieved June 15, 2021, from https://www.tandfonline.com/doi/abs/10.1080/01621459.1998.10474107

Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer-Verlag, New York. Retrieved July 5, 2021, from https://www.springer.com/gp/book/9780387953649

D'Amen, M., Mod, H.K., Gotelli, N.J. & Guisan, A. (2017). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Dryad*. Retrieved June 29, 2020, from http://datadryad.org/stash/dataset/doi:10.5061/dryad.8mv11

D'Amen, M., Mod, H.K., Gotelli, N.J. & Guisan, A. (2018). Disentangling biotic interactions, environmental filters, and dispersal limitation as drivers of species co-occurrence. *Ecography*, **41**, 1233–1244. Retrieved June 9, 2020, from https://onlinelibrary.wiley.com/doi/abs/10.1111/ecog.03148

Damgaard, C., Hansen, R.R. & Hui, F.K.C. (2020). Model-based ordination of pin-point cover data: Effect of management on dry heathland. *Ecological Informatics*, **60**, 101155. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S1574954120301059

Dunn, P.K. & Smyth, G.K. (1996). Randomized Quantile Residuals. *Journal of Computational and Graphical Statistics*, **5**, 236–244. Retrieved from http://www.jstor.org/stable/1390802

Edwards, L.J., Muller, K.E., Wolfinger, R.D., Qaqish, B.F. & Schabenberger, O. (2008). An R2 statistic for fixed effects in the linear mixed model. *Statistics in Medicine*, **27**, 6137–6157. Retrieved July 5, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3429

Files, B.T., Strelioff, M. & Bonnevie, R. (2019). *Bayesian reduced-rank regression with stan*. ARMY RESEARCH LAB ADELPHI MD Playa Vista United States.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, **35**, 1–165.

Hartig, F. (2021). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models*. Retrieved from https://CRAN.R-project.org/package=DHARMa

Hawinkel, S., Kerckhof, F.-M., Bijnens, L. & Thas, O. (2019). A unified framework for unconstrained and constrained ordination of microbiome read count data. *PLOS ONE*, **14**, e0205474. Retrieved July 5, 2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0205474

Hijmans, R.J. (2020). *Raster: Geographic data analysis and modeling*. Retrieved from https://CRAN.R-project.org/package=raster

Hoegh, A. & Roberts, D.W. (2020). Evaluating and presenting uncertainty in model-based unconstrained ordination. *Ecology and Evolution*, **10**, 59–69. Retrieved July 5, 2021, from https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.5752

Hui, F.K.C. (2016). Boral – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in r. *Methods in Ecology and Evolution*, **7**, 744–750. Retrieved August 21, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12514

Hui, F.K.C. (2017). Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Computational Statistics & Data Analysis*, **105**, 1–10. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S0167947316301724

Hui, F.K.C., Taskinen, S., Pledger, S., Foster, S.D. & Warton, D.I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, **6**, 399–411. Retrieved April 24, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12236

Hui, F.K.C., Warton, D.I., Ormerod, J.T., Haapaniemi, V. & Taskinen, S. (2017). Variational Approximations for Generalized Linear Latent Variable Models. *Journal of Computational and Graphical Statistics*, **26**, 35–43. Retrieved April 24, 2020, from https://doi.org/10.1080/10618600.2016.1164708

Jaeger, B.C., Edwards, L.J., Das, K. & Sen, P.K. (2017). An R2 statistic for fixed effects in the generalized linear mixed model. *Journal of Applied Statistics*, **44**, 1086–1105. Retrieved July 7, 2021, from https://doi.org/10.1080/02664763.2016.1193725

Jongman, R., ter Braak, C. & van Tongeren, O. (Eds.). (1995). *Data analysis in community and landscape ecology*. Cambridge university press, Cambridge.

Júnior, P.D.M. & Nóbrega, C.C. (2018). Evaluating collinearity effects on species distribution models: An approach based on virtual species simulation. *PLOS ONE*, **13**, e0202403. Retrieved June 29, 2021, from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0202403

Kristensen, K., Nielsen, A., Berg, C.W., Skaug, H. & Bell, B. (2016). TMB: Automatic Differentiation and Laplace Approximation. *J. Stat. Soft.*, **70**.

McCune, B. (1997). Influence of Noisy Environmental Data on Canonical Correspondence Analysis. *Ecology*, **78**, 2617–2623. Retrieved April 22, 2021, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/0012-9658

Niku, J. (2020). On modeling multivariate abundance data with generalized linear latent variable models. *JYU dissertations*. Retrieved June 30, 2021, from https://jyx.jyu.fi/handle/123456789/67735

Niku, J., Brooks, W., Herliansyah, R., Hui, F.K.C., Taskinen, S., Warton, D.I. & van der Veen, B. (2020). *Gllvm: Generalized linear latent variable models*. Retrieved from https://CRAN.R-project.org/package=gllvm

Niku, J., Hui, F.K.C., Taskinen, S. & Warton, D.I. (2019). Gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, **10**, 2173–2182. Retrieved May 13, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13303

Niku, J., Warton, D.I., Hui, F.K.C. & Taskinen, S. (2017). Generalized Linear Latent Variable Models for Multivariate Count and Biomass Data in Ecology. *JABES*, **22**, 498–522. Retrieved April 24, 2020, from https://doi.org/10.1007/s13253-017-0304-7

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E. & Wagner, H. (2020). *Vegan: Community ecology package*. Retrieved from https://CRAN.R-project.org/package=vegan

Palmer, M.W. (1993). Putting Things in Even Better Order: The Advantages of Canonical Correspondence Analysis. *Ecology*, **74**, 2215–2230. Retrieved from http://www.jstor.org/stable/1939575

Pebesma, E.J. & Bivand, R.S. (2005). Classes and methods for spatial data in R. *R News*, **5**, 9–13. Retrieved from https://CRAN.R-project.org/doc/Rnews/

Peres-Neto, P.R. & Jackson, D.A. (2001). How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia*, **129**, 169–178. Retrieved August 8, 2020, from https://doi.org/10.1007/s004420100720

Rao, C.R. (1964). The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, **26**, 329–358. Retrieved from http://www.jstor.org/stable/25049339

ter Braak, C.J. (1986). Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis. *Ecology*, **67**, 1167–1179. Retrieved May 25, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.2307/1938672

ter Braak, C.J.F. (1987). *Unimodal models to relate species to environment*. PhD thesis thesis, Ter Braak, S.l. Retrieved July 5, 2021, from https://library.wur.nl/WebQuery/wurpubs/2436

Ter Braak, C.J.F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, **69**, 69–77. Retrieved July 11, 2021, from https://doi.org/10.1007/BF00038688

ter Braak, C. (1988). Partial canonical correspondence analysis. *Classification and related methods of data analysis: proceedings of the first conference of the International Federation of Classification Societies (IFCS), Technical University of Aachen, FRG, 29 June-1 July 1987*, 551–558.

ter Braak, C.J.F. & Prentice, I.C. (1988). A Theory of Gradient Analysis. *Advances in Ecological Research* (eds M. Begon, A.H. Fitter, E.D. Ford & A. Macfadyen), pp. 271–317. Academic Press. Retrieved July 24, 2020, from http://www.sciencedirect.com/science/article/pii/S006525040860183X

ter Braak, C.J.F. & Šmilauer, P. (2015). Topics in constrained and unconstrained ordination. *Plant Ecol*, **216**, 683–696. Retrieved June 9, 2021, from https://doi.org/10.1007/s11258-014-0356-5

Tobler, M.W., Kéry, M., Hui, F.K.C., Guillera-Arroita, G., Knaus, P. & Sattler, T. (2019). Joint species distribution models with species correlations and imperfect detection. *Ecology*, **100**, e02754. Retrieved July 12, 2021, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1002/ecy.2754

van der Aart, P. & Smeek-Enserink, N. (1975). Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Netherlands Journal of Zoology*, **25**, 1–45.

van der Veen, B., Hui, F.K.C., Hovstad, K.A., Solbu, E.B. & O'Hara, R.B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, **n/a**. Retrieved April 20, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13595

Wang, Y., Naumann, U., Wright, S.T. & Warton, D.I. (2012). Mvabund– an R package for model-based analysis of multivariate abundance data. *Methods in Ecology and Evolution*, **3**, 471–474. Retrieved August 12, 2020, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210X.2012.00190.x

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C. & Hui, F.K.C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends Ecol. Evol. (Amst.)*, **30**, 766–779.

Warton, D.I. & Hui, F.K.C. (2017). The central role of mean-variance relationships in the analysis of multivariate abundance data: A response to Roberts (2017). *Methods in Ecology and Evolution*, **8**, 1408–1414. Retrieved July 12, 2021, from https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.12843

Yee, T.W. (2004). A New Technique for Maximum-Likelihood Canonical Gaussian Ordination. *Ecological Monographs*, **74**, 685–701. Retrieved June 9, 2020, from https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/03-0078

Yee, T.W. (2014). Reduced-rank vector generalized linear models with two linear predictors. *Computational Statistics & Data Analysis*, **71**, 889–902. Retrieved July 5, 2021, from https://www.sciencedirect.com/science/article/pii/S0167947313000273

Yee, T.W. & Hastie, T.J. (2003). Reduced-rank vector generalized linear models. *Statistical Modelling*, **3**, 15–41. Retrieved June 9, 2021, from https://doi.org/10.1191/1471082X03st045oa

Zeng, Y., Zhao, H. & Wang, T. (2021). Model-Based Microbiome Data Ordination: A Variational Approximation Approach. *Journal of Computational and Graphical Statistics*, **0**, 1–13. Retrieved July 5, 2021, from https://doi.org/10.1080/10618600.2021.1882467

Økland, R.H. (1996). Are ordination and constrained ordination alternative or complementary strategies in general ecological studies? *Journal of Vegetation Science*, **7**, 289–292. Retrieved from http://www.jstor.org/stable/3236330