Received: 2 October 2020 | Accepted: 1 March 2021

DOI: 10.1111/2041-210X.13595

RESEARCH ARTICLE

Methods in Ecology and Evolution | BRITISH ECOLOGICAL SOCIETY

Model-based ordination for species with unequal niche widths

Bert van der Veen[1,2,3] | Francis K. C. Hui[4] | Knut A. Hovstad[3,5] | Erik B. Solbu[1] | Robert B. O'Hara[2,3]

[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, Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway

[4]Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, Australia

[5]The Norwegian Biodiversity Information Centre, Trondheim, Norway

Correspondence
Bert van der Veen
Email: bert_van_der_veen@hotmail.com

Funding information
Research Council of Norway, Grant/Award Number: 272408/F40; Australian Research Council

Handling Editor: Javier Palarea-Albaladejo

Abstract

1. It is common practice for ecologists to examine species niches in the study of community composition. The response curve of a species in the fundamental niche is usually assumed to be quadratic. The centre of a quadratic curve represents a species' optimal environmental conditions, and the width its ability to tolerate deviations from the optimum.
2. Most multivariate methods assume species respond linearly to niche axes, or with a quadratic curve that is of equal width for all species. However, it is widely understood that some species have the ability to better tolerate deviations from their optimal environment (generalists) compared to other (specialist) species. Rare species often tolerate a smaller range of environments than more common species, corresponding to a narrow niche.
3. We propose a new method, for ordination and fitting Joint Species Distribution Models, based on Generalized Linear Mixed-effects Models, which relaxes the assumptions of equal tolerances.
4. By explicitly estimating species maxima, and species optima and tolerances per ecological gradient, we can better explore how species relate to each other.

KEYWORDS

joint species distribution model, model-based ordination, niche model, unconstrained quadratic ordination, unimodal response

1 | INTRODUCTION

One of the key topics addressed by community ecology is the exploration of community composition. To that end, species communities are surveyed at locations along environmental gradients. The ecological niche is then reflected in the observed distribution of a species. A species exhibits its maximum abundance, or has the highest probability of occurrence, at the optimum of the niche. The limits of a species distribution correspond to the limits of the niche, controlled by a species' tolerance to a range of environmental conditions. Different species vary in their ability to tolerate deviations from the optimum, reflecting differences in niche width, and indicating different places on the specialist–generalist spectrum.

Correspondence analysis (CA) is often used to estimate the optima of species niches with quadratic response curves. It implicitly approximates the fit of a quadratic model, which functions best under the assumptions of equally spaced optima, sites being well within the range of species optima, equal tolerances and equal or independent maxima (ter Braak, 1985). The combination of assuming equally spaced optima, equal maxima and equal tolerances gives an early niche model called the species packing model (MacArthur & Levins, 1967). The relationship of the species packing model to

CA has added to its popularity among applied ecologists (Wehrden et al., 2009).

Recent advances in the estimation of species niches have focused on performing ordination with explicit statistical models, such as Generalized Linear Latent Variable Models (GLLVMs; Warton et al. 2015). With intercepts included for row standardization, GLLVMs can fit a quadratic response curve, assuming species have equal tolerances (Hui et al., 2015; Jamil & ter Braak, 2013). When predictor variables are included, a GLLVM with quadratic response model partitions species distributions in observed (fixed effects) and latent or unobserved (random effects), similar to the partitioning of fixed and random effects in mixed-effects models when predictors are included.

The GLLVM framework is well known for its capability to fit Joint Species Distribution Models (JSDMs; Ovaskainen et al., 2017; Pollock et al., 2014; Tobler et al., 2019; Zurell et al., 2020). In the context of JSDMs, GLLVMs assume species abundances are correlated due to similarity in response to ecological gradients, modelled with predictor variables and latent variables. Latent variables can be understood as combinations of missing predictors, so that GLLVMs allow us to parsimoniously model species distributions. They are equivalent to ordination axes, representing complex ecological gradients (Halvorsen, 2012). Recently, the use of GLLVMs to perform model-based ordination has increased in popularity (Björk et al., 2018; Damgaard et al., 2020; Inoue et al., 2017; Lacoste et al., 2019; Paul, 2020). However, existing GLLVMs assume that species respond to latent variables linearly, just as all classical ordination methods do (Jamil & ter Braak, 2013). In contrast, it is widely understood that species have unequal tolerances, so that the assumption of linear responses, or at best quadratic responses with equal tolerances, is unlikely to hold in practice.

In this paper, our goal was to overcome the assumptions of equal tolerances, by formulating a GLLVM where species are allowed to respond to the latent variables in a quadratic fashion. To our knowledge, there has been no attempt to implement such a GLLVM until now. The quadratic term allows to fully estimate species niches, so that species optima and tolerances per latent variable and species maxima can all be estimated explicitly. Explicitly estimating the combination of these three parameters gives unique insight into reasons for species rarity, whether it is due to low abundance or probability of occurrence (maxima), a high degree of habitat specialization (tolerance) or due to unsuitable observed environmental conditions (optima). Due to the model-based nature of the proposed ordination method, it is possible to calculate confidence intervals for each set of parameters, providing unparalleled benefits for inference when using ordination. Additionally, assuming a quadratic response model allows to implement the concept of gradient length, as in Detrended Correspondence Analysis (DCA; Hill and Gauch, 1980), which is a measure of beta diversity commonly used by ecologists.

In contrast to classical ordination methods, GLLVMs model the latent variables as unobserved, treating them as random rather than fixed (Walker & Jackson, 2011), which consequently have to be integrated over in the likelihood. Here, we develop a variational approximation (VA) implementation after Hui et al. (2017) and Niku et al. (2019), to perform calculations quickly and efficiently. In addition to presenting the GLLVM with quadratic response model, we perform simulations to evaluate the accuracy of the VA implementation, and the capability of the GLLVM with quadratic response model to retrieve the true species-specific parameters and latent variables. We use two real-world datasets to demonstrate the use and interpretation of the proposed GLLVM with quadratic responses: (a) a small dataset of hunting spiders in a Dutch dune ecosystem (van der Aart & Smeek-Enserink, 1974), and (b) a larger dataset of Swiss alpine plant species on a strong elevation gradient (D'Amen et al., 2018).

2 | MODEL FORMULATION

The ecological niche for each species $j = 1...p$ is described here by a quadratic function involving three parameters: the optimum u_{jq} for latent variable $q = 1...d$ stored in the vector $\boldsymbol{u}_j = \{u_{j1}...u_{jq}\}$, the tolerance t_{jq} for latent variable q stored in the vector $\boldsymbol{t}_j = \{t_{j1}...t_{jq}\}$ and a species' overall maximum c_j. Optima $\boldsymbol{u}_j$ are the locations on the ecological gradients where a species exhibits its highest abundance or probability of occurrence (the maximum c_j). The tolerances $\boldsymbol{t}_j$ are a measure of the width or breadth of the niche, indicating if a species is a generalist or specialist on each ecological gradient.

Consider an $n \times p$ matrix of observations, where y_{ij} denotes the response of species j at sites $i = 1...n$. Then, we assume that conditional on a vector $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ of d latent variables where $d \ll p$, the responses y_{ij} at site i are independent observations from a distribution whose mean, denoted here as $\mathrm{E}(y_{ij}|\mathbf{z}_i)$, is modelled as:

$$\begin{aligned} g\left\{\mathrm{E}\left(y_{ij}|\mathbf{z}_i\right)\right\} &= c_j - \sum_{q=1}^{d}\left\{\frac{\left(z_{iq}-u_{jq}\right)^2}{2t_{jq}^2}\right\} \\ &= c_j - \sum_{q=1}^{d}\left(\frac{u_{jq}^2}{2t_{jq}^2} - \frac{z_{iq}^2}{2t_{jq}^2} + \frac{z_{iq}u_{jq}}{t_{jq}^2}\right), \end{aligned} \tag{1}$$

where $g\{\cdot\}$ is a known link function (e.g. the log link when responses are assumed to be Poisson, negative-binomial or gamma distributed, the probit link when the responses are assumed to be Bernoulli or ordinal distributed and the identity link for responses that are assumed to be Gaussian distributed).

For a closer comparison to the GLLVM with linear response model (Hui et al., 2015), we formulate the GLLVM with quadratic species response curves in terms of matrix notation:

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

with a species-specific intercept β_{0j} that accounts for species mean abundances, and a vector of coefficients per species for the linear term $\boldsymbol{\gamma}_j$. We can see that a third term is added here to the existing structure of a GLLVM with linear species responses, which models tolerances per species and latent variable. Specifically, we introduce a diagonal matrix $\boldsymbol{D}_j$ of positive-only quadratic coefficients, with each diagonal

element being the quadratic effect for latent variable q and species j. The sign constraint ensures that species exhibit concave quadratic curves only. The proposed model could instead be used to estimate species minima rather than maxima, though we did not do that here as clear ecological foundations for such a model are lacking.

Let D_{jqq} denote the diagonal elements of $\boldsymbol{D}_j$ for latent variable q. Then we are able to derive the following connections between the parameters in Equations (1) and (2): $\beta_{0j} = c_j - \frac{1}{2}\sum_{q=1}^{d} u_{jq}^2/t_{jq}^2$, $\gamma_{jq} = u_{jq}/t_{jq}^2$, and $D_{jqq} = 1/\left(2t_{jq}^2\right)$. Similarly, for the formulation in Equation (2), the parameters in Equation (1) can be retrieved: $c_j = \beta_{0j} + \frac{1}{4}\sum_{q=1}^{d}\gamma_{jq}^2/D_{jqq}$, $u_{jq} = \gamma_{jq}/\left(2D_{jqq}\right)$, and $t_{jq} = 1/\sqrt{2D_{jqq}}$.

Additionally, row intercepts or predictors can be included as in Hui et al. (2017), or species traits as in Niku et al. (2019), though we have chosen to omit those terms here and focus on the case of unconstrained ordination.

Four special cases of the GLLVM with quadratic response model, as formulated in Equation (2), are worth discussing: (a) $\boldsymbol{D}_j = D$, that is, common tolerances for species, (b) $\boldsymbol{D}_j = D_{11}\mathbf{I}_d$ where $\mathbf{I}_d$ is a $d \times d$ identity matrix, that is, equal tolerances for species and latent variables, (c) when $\boldsymbol{D}_j = 0$ for a subset of the p species and (d) when $\boldsymbol{D}_j = 0$ for all p species. The first case assumes tolerances to be the same across species, but not latent variables. This species-common tolerances model might prove useful in practice, as it requires fewer observations per species than when estimating quadratic coefficients for all species, but still explicitly includes quadratic species responses in contrast to the simpler GLLVM with linear responses. In the second case, the quadratic term is not species or latent variable specific, so that it is equivalent to the GLLVM with linear species responses and random row intercepts as presented in Hui et al. (2015), which assumes tolerances to be the same for all species and latent variables. In the third case, some species respond to the latent variable linearly, while others exhibit quadratic responses. The fourth case is the most basic GLLVM with linear responses, which is the current standard in many software packages for JSDMs and model-based ordination, for example, boral (Hui, 2016), HMSC-R (Tikhonov et al., 2021) and gllvm (Niku et al., 2020).

3 | MODEL INTERPRETATION

In this section, we derive and discuss various tools that are commonly used in the application of JSDMs and ordination, such as calculating residual correlations, partitioning or decomposing residual variance, calculating gradient length and visualizing the ordination, and demonstrate how they can be adapted to the proposed GLLVM with quadratic response model.

3.1 | Residual covariance matrix

One aspect of GLLVMs is known for is modelling species residual correlations (Blanchet et al., 2020; Zurell et al., 2018), calculated from the residual covariance matrix. To facilitate calculation of the residual covariance matrix, we can reparameterize all GLLVMs as a multivariate mixed-effects model with a residual term:

$$g\left\{\mathrm{E}\left(y_{ij}|\boldsymbol{z}_i\right)\right\} = \beta_{0j} + \epsilon_{ij}. \tag{3}$$

Here, ϵ_{ij} accounts for any residual information that is not accounted for by fixed effects in the model, such as predictors or intercepts (Warton et al., 2015). Assuming the latent variables are independent for all sites, the elements of the residual covariance matrix are given by:

$$\Sigma_{jk} = \mathrm{cov}\left(\epsilon_{ij}, \epsilon_{kl}\right), \quad \forall i,k = 1\ldots n, j,l = 1\ldots p.$$

For a length p vector $\boldsymbol{\epsilon}_i$, existing JSDM implementations (e.g. Pichler & Hartig, 2020; Pollock et al., 2014) assume $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, that is, the residual term follows a multivariate normal distribution. For the GLLVM with linear species responses, it is straightforward to show that with $\epsilon_{ij} = \boldsymbol{z}_i^{\top}\boldsymbol{\gamma}_j$, then $\boldsymbol{\epsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \Gamma\Gamma^{\top}\right)$, where $\boldsymbol{\Gamma}$ is a $p \times d$ matrix of species linear coefficients for the latent variables $\boldsymbol{\gamma}_j$. In essence, GLLVMs preform a low rank approximation to the covariance matrix of a residual term. The rank of this residual covariance matrix is equal to the number of estimated latent variables d in the model for the GLLVM with linear species responses.

Turning to the GLLVM with quadratic response model, where $\epsilon_{ij} = \boldsymbol{z}_i^{\top}\boldsymbol{\gamma}_j - \boldsymbol{z}_i^{\top}\boldsymbol{D}_j\boldsymbol{z}_i$, the elements of the residual covariance matrix are:

$$\Sigma_{\mathrm{quad},jk} = \sum_{q=1}^{d}\left(\gamma_{jq}\gamma_{kq} + 2D_{jqq}D_{kqq}\right), \tag{4}$$

for which a proof is given in Appendix S1. This can be rewritten in terms of the species optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$:

$$\Sigma_{\mathrm{quad},jk} = \sum_{q=1}^{d}\left\{\left(t_{jq}^2 t_{kq}^2\right)^{-1}\left(0.5 + u_{jq}u_{kq}\right)\right\}. \tag{5}$$

Equations (4) and (5) additionally serve to demonstrate how to partition and decompose the residual variance of the GLLVM with quadratic response model, for example, per latent variable, for the linear and quadratic term separately, or both. Variance partitioning is commonly used in the application of ordination methods, for example, to determine fit (Øland, 1999), or to explore causes of residual variance (Borcard et al., 1992; Øland & Eilertsen, 1994). Predictor variables can be included in the model to account for the residual variance otherwise accounted for by the latent variables. The residual variance can be used to identify indicator species, that is, those species that best represent an ecological gradient, or to calculate a measure of R^2 (Nakagawa & Schielzeth, 2013).

Under the assumption of latent variables with zero mean, the linear and quadratic terms in the model are independent. As such, the rank of the residual covariance matrix is double that of a GLLVM with linear species responses and the same number of latent variables, $2d$. The additional quadratic term thus allows us to account for more

residual correlations between species, with fewer latent variables. This corresponds to the ecological notion that species often respond to few major complex ecological gradients (Halvorsen, 2012). From this, we see that when the number of latent variables in a GLLVM with quadratic response model exceeds ½p, there are more parameters included than in a JSDM with an unstructured residual covariance matrix. However, this is not an issue here, since for ordination purposes we are only interested in cases where there are much fewer latent variables d than species p.

3.2 | Gradient length

The length of an ecological gradient is of great interest to ecologists in the use of ordination, because it is a measure of beta diversity (Oksanen & Tonteri, 1995). Longer gradients indicate higher diversity, as spacing between sites in latent space is potentially larger. In the past, it has been emphasized that short gradients are better analysed using linear ordination methods, and longer with unimodal methods (ter Braak & Prentice, 1988). However, the GLLVM with quadratic response model allows species to exhibit both linear and unimodal responses, and so it is appropriate for both, removing the need to switch between ordination methods as a consequence of (the lack of) unimodal species responses. Regardless, gradient length could be used to decide between response models instead of, for example, information criteria. To determine gradient length from the proposed GLLVM with quadratic response model, we rescale the latent variables $\boldsymbol{z}_i$ with a diagonal covariance matrix $\boldsymbol{G}$ of size $d \times d$, to calculate ecological gradients $\tilde{\boldsymbol{z}}_i$. The measure of gradient length calculated here can be interpreted in the same manner as the gradient length provided by DCA (Hill & Gauch, 1980).

First, for a species-common tolerances model, we note that the quadratic term in Equation (2), that is, $\boldsymbol{z}_i^\top \boldsymbol{D}\boldsymbol{z}_i$, can instead be written as $\sum_{q=1}^{d} z_{iq}^2 D_{qq}$, so that $\tilde{z}_{iq} = z_{iq}\sqrt{D_{qq}}$, and $\tilde{\boldsymbol{z}}_i \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{G} = 2\boldsymbol{D}$. Then, the length per ecological gradient is approximately $4G_{qq}^{\frac{1}{2}}$ (i.e. the approximate width of a normal distribution).

Second, for the species-specific tolerances model, we note that one of the uses of gradient length in the past has been to rescale the latent variables so that an ordination diagram can be understood in terms of compositional turnover (Hill & Gauch, 1980). This requires the mean species tolerances to be one (as is the case for the species-common tolerances model, under the rescaling suggested above), so that the covariance matrix of the ecological gradient in the species-specific tolerances model is $G_{qq} = \frac{1}{2p}\sum_{j=1}^{p} D_{jqq}$ and the matrix of quadratic coefficients $\boldsymbol{D}_j$ is scaled by the inverse of the covariance matrix of the ecological gradient, $\boldsymbol{G}^{-1}$. However, we choose to use the median of the species tolerances t_{jq} instead, as it more accurately represents gradient length with both linear and quadratic responses of species in the model. In general, the proposed quadratic model allows further exploration of measures of gradient length by, for example, using the mean tolerance of species with clear quadratic responses, rather than the median of all tolerances.

3.3 | Ordination diagram

Usually, results from an ordination are inspected visually, by jointly plotting site and species scores. For a GLLVM with linear responses, this can be done by constructing a biplot (Gabriel, 1971). Biplots perform a linear approximation of a matrix, and thus are expected to perform poorly when species exhibit quadratic responses: biplots will create an arch when the residual variance of the linear term is smaller than the residual variance of the quadratic term. When the linear and quadratic terms are independent, as is the case here (see above), a biplot can visualize them separately.

Instead, we propose that species optima and tolerances can be plotted directly, so that species niches are visualized in a two-dimensional latent (ecological) space from a top-down perspective. However, since species are allowed to exhibit linear responses in the quadratic response model, optima and tolerances can be very large. If plotting both directly, this will lead to species with large optima and wide niches dominating the plot. The first issue can be prevented by only visualizing species optima that are close to, or within, the range of the estimated site scores, and by using arrows to indicate the location of the remaining optima (similarly as in Gabriel, 1971). The widths of the niches can be represented as ellipses using the precision of estimated species tolerances, to provide an impression of species co-occurrence patterns. The precision, calculated as the inverse of the squared species tolerances $1/t_{jq}^2$, can be interpreted as 'narrowness' of the ecological niche (i.e. a small precision corresponds to a wide niche). Then, a larger ellipse corresponds to a larger residual variance of the quadratic term of a latent variable, drawing emphasis to potential indicator species.

Additionally, information on sites, such as the predicted locations and prediction regions, can be added (Hui et al., 2017). Information for the sites can be used to infer the distance of sites to the species optima (i.e. the suitability of sites for species), or to the edges of species niches (see the hunting spiders example below).

Finally, based on the discussion in the two subsections above, there are two ways of scaling the ordination diagram: (a) by the residual variance per latent variable, or (b) by using the mean or median tolerance. In the first scaling, the diagram is scaled to draw attention to the latent variable that explains most variance in the model. However, the second scaling has a more ecological intuitive interpretation; if the tolerances are assumed to be common for species, the second scaling produces an ordination diagram in units of compositional turnover (Gauch, 1982). When the linear term in the model does not explain a larger proportion of the total residual variance per latent variable relative to the quadratic term, these scalings produce similar results.

4 | MODEL ESTIMATION

We propose to use VAs (Hui et al., 2017) for estimation and inference for the GLLVM with quadratic response model. Broadly speaking, VA is a general technique used to provide a closed-form

approximation to the marginal log-likelihood of a model with random effects or latent variables, when an analytical solution is not available. Computationally, VA can be orders of magnitude faster than MCMC, numerical integration or even the Laplace approximation (Niku et al., 2019), and without loss of accuracy (Hui et al., 2017). However, the calculation of the VA log-likelihood needs to be derived on a case-by-case basis. In contrast, the Laplace approximation can be applied automatically in many cases (Kristensen et al., 2016), although it is not possible to apply that here for the GLLVM with quadratic response model (K. Kristensen, pers. comm., 8 March 2019).

The marginal log-likelihood of a GLLVM is given by:

$$\mathscr{L}(\Theta) = \sum_{i=1}^{n} \log \left\{ \int_{-\infty}^{\infty} \prod_{j=1}^{p} f\left(y_{ij} | \boldsymbol{z}_i, \Theta\right) h(\boldsymbol{z}_i) d\boldsymbol{z}_i \right\}, \quad (6)$$

where $f(y_{ij}|\boldsymbol{z}_i, \Theta)$ is the distribution of the species responses given the latent variables. As mentioned previously, and as per Hui et al. (2015), we assume the distribution of the latent variables $h(\boldsymbol{z}_i)$ to be multivariate standard normal, that is, $h(\boldsymbol{z}_i) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The vector Θ includes all parameters in the model $\Theta = \left\{\beta_{01} \ldots \beta_{0j}, \gamma_{11} \ldots \gamma_{jq}, D_{111} \ldots D_{jqq}\right\}^{\top}$.

In VA, we construct a lower bound to Equation (6), by assuming that the posterior distribution of the latent variables can be approximated by a closed-form distribution, for example, a multivariate normal distribution (this is also referred to as the variational distribution). We then treat this lower bound as our new objective function, on which we base estimation and inference of the model parameters, as well as predictions of the latent variables. More details on the motivation and background of variational approximations are available in the study by Ormerod and Wand (2010, 2012). Hui et al. (2017) showed that, for GLLVMs with linear responses, the optimal variational distribution is multivariate normal $\boldsymbol{z}_i \sim \mathcal{N}\left(\boldsymbol{a}_i, \mathbf{A}_i\right)$, with mean $\boldsymbol{a}_i$ and covariance matrix $\mathbf{A}_i$, so we will adopt this choice here as well. While we do not anticipate a multivariate normal distribution to be the optimal variational distribution for a GLLVM with quadratic response model, we nevertheless choose to follow the same assumption to facilitate computational efficiency and a closed form for the resulting VA log-likelihood. The means of the variational distribution $\boldsymbol{a}_i$ can be understood as predicted locations of sites, that is, site scores in an ordination. The covariance matrices of the variational distributions $\mathbf{A}_i$ provide the necessary information to construct prediction regions.

In Appendix S2 we provide derivations for the log-likelihood of common response types in community ecology, such as count data (Poisson, a Poisson–Gamma derivation of the negative-binomial distribution for overdispersed counts and both assuming a log-link function), binary data and ordinal data (both with probit-link function), as well as positive continuous data (gamma, with log-link function) and continuous data (Gaussian, with an identity-link function). Additionally, some information on calculating approximate confidence intervals for (functions of) the parameters is included in Appendix S2. Recommendations on stabilizing the fitting of GLLVMs with a quadratic response model are included in Appendix S3.

5 | SIMULATION STUDY

To assess how well the proposed model retrieves the true latent variables $\boldsymbol{z}_i$, optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$, we performed simulations for six response distributions; (1) Gaussian, (2) gamma, (3) Poisson, (4) negative-binomial, (5) Bernoulli and (6) ordinal. The R code used for the simulations is provided in Appendix S4. For each of the distributions, we simulated 1,000 datasets with different numbers of sites and species. A consequence of restricting the quadratic response model to concave shapes only is that it often simulates a large number of negative values (on the link scale, generally more so than the GLLVM with linear species responses), providing a challenge in testing its accuracy, especially for small datasets.

First, to study the accuracy of the VA approximation, we simulated datasets of p = 20–100 species in increments of 10, while keeping the number of sites constant at n = 100. Hui et al. (2017) argued that the VA log-likelihood is expected to converge to the true likelihood as $p \to \infty$, as with many species the posterior for the site scores is likely to be approximately normal due to the central limit theory. This will allow us to study the finite sample properties of the VA approximation for the proposed model. Second, to explore the sample size required to accurately estimate the species-specific parameters, for example, species optima $\boldsymbol{u}_j$ and tolerances $\boldsymbol{t}_j$, we simulated datasets of n = 20–100 sites in increments of 10, while keeping the number of species constant at p = 100.

As a true model, we considered a GLLVM with quadratic response model and d = 2 latent variables, which was constructed as follows. The latent variables were simulated following a multivariate standard normal distribution, that is, $\boldsymbol{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Second, the species maxima c_j were simulated as Uniform(2,6), as this was approximately the range of species maxima in the best fitting model for the hunting spider dataset below. Next, the true optima u_{jq} were simulated within the range of the realized latent variables (approximately between −2 and 2) following a uniform distribution. Lastly, species tolerances were simulated as Uniform(0.2,1), corresponding to species niches ranging from narrow to the full width of the latent variable. Resulting species-specific intercepts β_{0j} from Equation (2) approximately ranged between −15 and 20, but tended to be more positive than negative, with a median of 2.6. For the Gaussian, negative-binomial and gamma distributions, the dispersion parameter for all species was set equal to 1. For the ordinal distribution, we assumed six classes with the true cut-offs being 0, 1, 2, 3, 4, 5, meaning that species were most often absent (category 0), while they were rarely very abundant (category 5). When fitting a model to each simulated dataset, we assumed the number of latent variables was known prior to fitting (i.e. we did not select the number of latent variables).

We measured performance of the GLLVM with quadratic response model by the prediction of the latent variables $\boldsymbol{z}_i$ and the species optima $\boldsymbol{u}_j$. The species optima are a function of both the

linear and quadratic coefficients and should provide a good overall measure of performance for retrieving the true species-specific parameters, in addition to being of specific interest to ecologists. We measured discrepancy to the true parameter values using the Procrustes error (Peres-Neto & Jackson, 2001). For this, we excluded the optimum of the first species on the second latent variable as this was fixed to zero for reasons of parameter identifiability (Hui et al., 2015). Since the GLLVM with quadratic response model allows species to exhibit linear responses, which have optima tending to infinity, we also chose to remove all optima larger than 10 and smaller than −10, that is, for those species that clearly lacked a sufficiently strong quadratic signal in the simulated datasets. Including these optima would result in a biased view of the accuracy of the optima that can be estimated by the model. For clarity and transparency, we additionally present the number of optima removed for each of the datasets, to further provide an impression of the data requirements of the proposed model.

For all of the models fitted to Gaussian and gamma response datasets, typically none or only a few optima were excluded, meaning that the median number excluded was zero. In general, and not surprisingly, more optima were excluded for models fitted to datasets where n/p was small and for discrete distributions. For example, when $n = 20$ sites and $p = 100$ species, so that the true model included a total 200 species optima, the median number of optima excluded for datasets with Poisson responses was 4 (2–5, first and third quartiles), for datasets with negative-binomial responses this was 7 (5–10), for datasets with Bernoulli responses this was 44 (40–47) and for datasets with ordinal responses this was 20 (17–24). In contrast, for datasets where n/p was large, considerably fewer optima were excluded across all response types. For example, when $n = 100$ and $p = 100$, and for Poisson responses, the median of excluded optima was 1 (1–3), for negative-binomial response datasets this was 6 (5–7), while for Bernoulli response datasets the median number of optima excluded was with a median of 29 (27–32) still large and for ordinal response datasets this was 13 (11–15).

The symmetric Procrustes error per distribution and for the different sized datasets is presented in Figure 1. As expected, the GLLVM with quadratic responses was more accurate for datasets with larger p and larger n. For all distributions, the latent variables

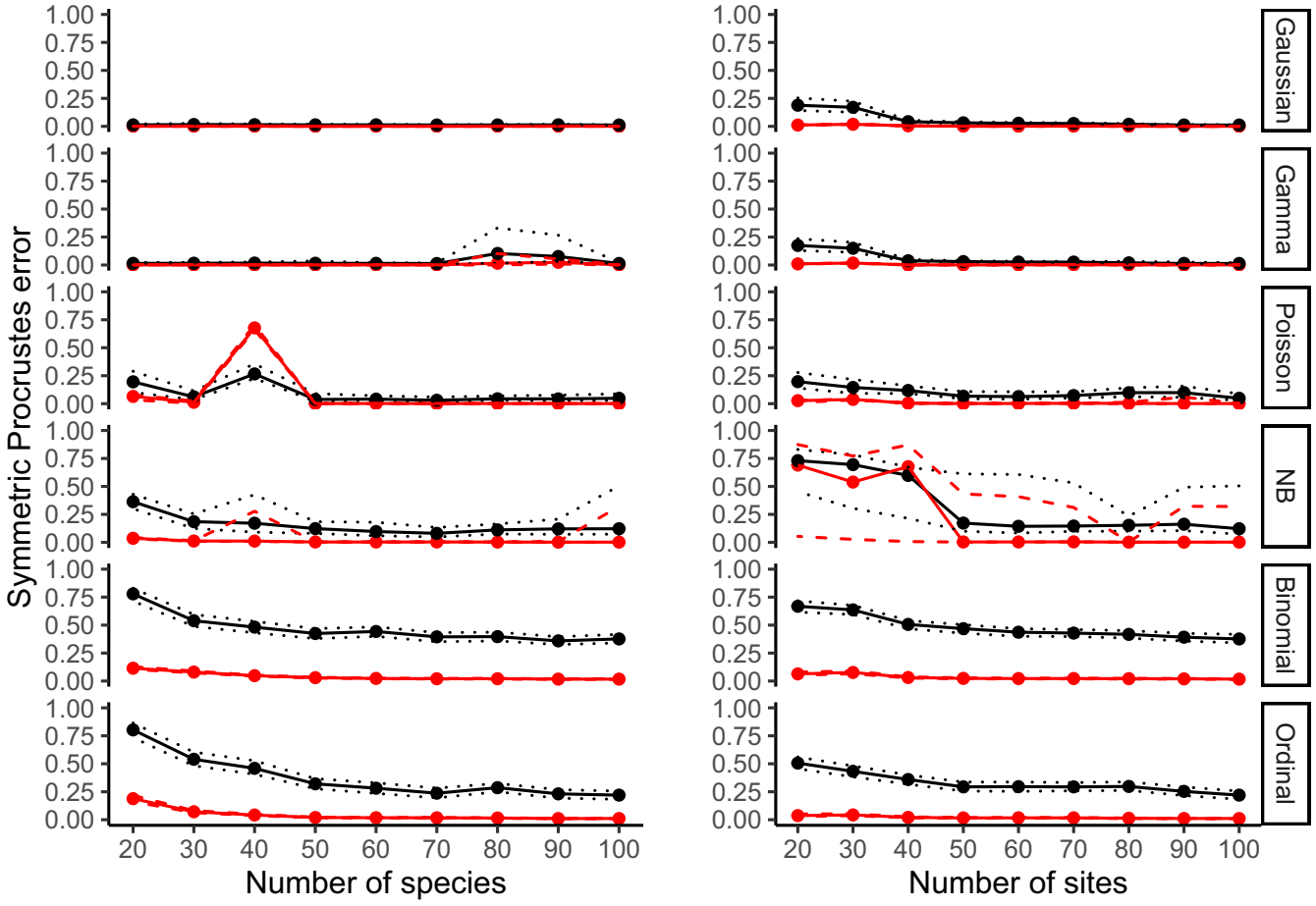

FIGURE 1 Simulation results for the 1,000 GLLVMs fitted to each dataset and response distribution, with the symmetric Procrustes error calculated based on optima that could be estimated (optima outside the range (−10,10) were excluded). The left column shows simulations where the number of sites was kept constant at $n = 100$, and analogous for the right column with $p = 100$. The figure includes the median Procrustes Error for species optima (black) and latent variables (red), with the first and third quartiles represented as dotted (optima) and dashed (latent variables) lines

were often better retrieved than the species optima. This is not surprising, as the species optima are a function of two parameters, particularly the inverse of the quadratic coefficients, so that a small change in the quadratic coefficients can result in a large change in the species optima. When fitted to Gaussian or gamma response datasets, the model performed best. The accuracy of the estimated species optima and latent variables was only slightly lower for datasets with Poisson responses, and was also similar for datasets with negative-binomial responses and a large number of sites. If the number of sites is small, the variation in accuracy of the latent variable and of species optima was considerably larger for datasets with negative-binomial responses. Since the quadratic response model can, even without negative-binomial distribution, simulate overdispersed counts compared to the linear response model, these results were not surprising. In many cases, negative-binomial distributed datasets contained less information than datasets with Poisson-distributed responses, which makes accurate estimation increasingly difficult. The model was not accurate for Bernoulli or ordinal response datasets with small p. Fortunately, data of ecological communities often contain many species. For small n, models fitted to datasets with Bernoulli responses were not accurate, whereas models fit to datasets with ordinal responses showed slightly better performance. This too was not surprising, as datasets with ordinal responses include more information compared to datasets with Bernoulli responses. When the number of sites and species increased above 40, the performance of the GLLVM with quadratic responses in both cases improved considerably. Regardless, especially for Bernoulli responses, the simulated datasets often included too little information for many species to accurately estimate the parameters.

6 | APPLICATIONS TO REAL DATA

We applied the proposed GLLVM with quadratic response model to two different datasets: (a) the well-known hunting spider dataset collected by van der Aart and Smeek-Enserink (1974) in Dutch dunes, available in the `mvabund R` package (Wang et al., 2012), and (b) a dataset of plants in the Swiss Alps (available in the dryad database; D'Amen et al., 2017).

6.1 | Hunting spiders

For the hunting spider dataset, van der Aart and Smeek-Enserink (1974) used pitfall traps to collect spiders over a 60-week period, resulting in a dataset of counts for each of the $n = 28$ sites and $p = 12$ species. It has been used in the testing of ordination methods before (e.g. ter Braak, 1985, 1986; Hui et al., 2015; Yee, 2004), providing the possibility for comparison here. We used the Akaike information criterion corrected for small sample sizes (AICc; Burnham and Anderson, 2002) to find the model that best fitted the hunting spider dataset. We fitted GLLVMs with $d = 1$–3 latent variables, with linear and quadratic responses, including equal, common or unequal tolerances, and fixed row intercepts, all with Poisson or negative-binomial distributions (see Appendix S5 for the details). After selecting the model structure and number of latent variables, we continued to explore different sets of initial values to find the model that maximizes the VA log-likelihood. The best model included $d = 3$ latent variables and unequal tolerances, though a model with unequal tolerances $d = 2$ latent variables and fixed row intercepts was a close second contender (difference of 2.2 in AICc; see Appendix S5). The results for the two latent variables of the final model fit, which explained most residual variation, are presented in Figure 2.

We used the residual variance to determine which latent variables explained most variation, that is, were most important to consider for inference. For the GLLVM with quadratic response model, the first and third latent variables explained most variation in the model; 31% and 58%, respectively, so we will discuss the results of these below. Overall, the GLLVM with quadratic responses explained two and a half times more residual variation than a GLLVM with linear responses and the same number of latent variables. The lengths of the ecological gradients were 5.48 (3.96–7.00, 95% confidence interval), 3.68 (2.65–4.71) and 4.77 (3.10–6.44).

ter Braak (1985) interpreted the first ordination axis of DCA as 'a composite gradient of soil moisture and openness of habitat', as determined by regressing the ordination axis on variables measuring the amount of bare sand, soil moisture and the percentage cover by mosses at sites. Yee (2004) concluded that reflection of the soil surface had the strongest relationship with the first latent variable estimated using a Vector Generalized Additive Model. Similarly, the first latent variable in the GLLVM here has a strong relation with reflection of the surface (correlation coefficient of 0.83), the percentage cover of moss (0.82) and the cover of fallen leaves (−0.75). The second latent variable was related to the cover provided by the herb layer (0.70), and the third latent variable with soil water content (0.77).

ter Braak (1985) and Yee (2004) both visualized quadratic curves of the first latent variable using variations of Poisson regression and Generalized Additive Models respectively. There are clear similarities between the height and the location of species response curves for the first latent variable, and the corresponding response curves described by ter Braak (1985) and Yee (2004). Similarly, Figure 2 here shows a similar arrangement of species as Figure 1 in ter Braak (1986).

ter Braak (1985) concluded that most species exhibited unimodal curves on the first latent variable, though the benefit of a quadratic response model was least to the species *Alopecosa fabrilis*, *Arctosa perita* and *Pardosa lugubris*. Similarly, the optimum of *Pardosa lugubris* could not be estimated by VGAM. Here, as in Yee (2004), *Pardosa lugubris* and *Trochosa terricola* were the most abundant species. On the first latent variable, only the optima of *Pardosa lugubris* and *Pardosa monticola* were located outside the range of the latent variable. On the third latent variable, only the optima of *Arctosa lutetiana* were unobserved. Similar to the conclusion by ter Braak (1985), the confidence intervals for the quadratic coefficients

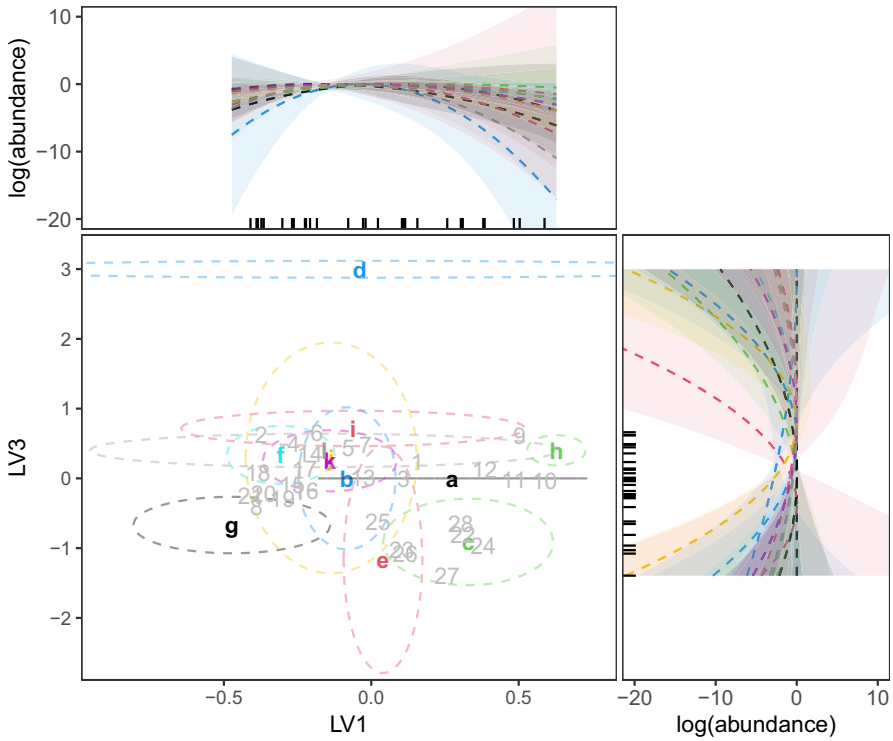

FIGURE 2 Ordination plot for the first two latent variables of the final GLLVM fit to the hunting spider dataset, scaled by the residual variances. Species optima are shown as letters, indicating the following species: a = *Alopecosa accentuata*, b = *Alopecosa cuneata*, c = *Alopecosa fabrilis*, d = *Arctosa lutetiana*, e = *Arctosa perita*, f = *Alonia albimana*, g = *Pardosa lugubris*, h = *Pardosa monticola*, i = *Pardosa nigriceps*, j = *Pardosa pullata*, k = *Trochosa terricola*, l = *Zora spinimana*. Ellipses represent the precision of the ecological niche, which can be interpreted as 'narrowness', so that large or wide ellipses represent species with narrow response curves. Species quadratic curves are included as side panels, with 95% confidence interval bands. Site locations are represented by grey numbers, though prediction regions have not been included in favour of readability

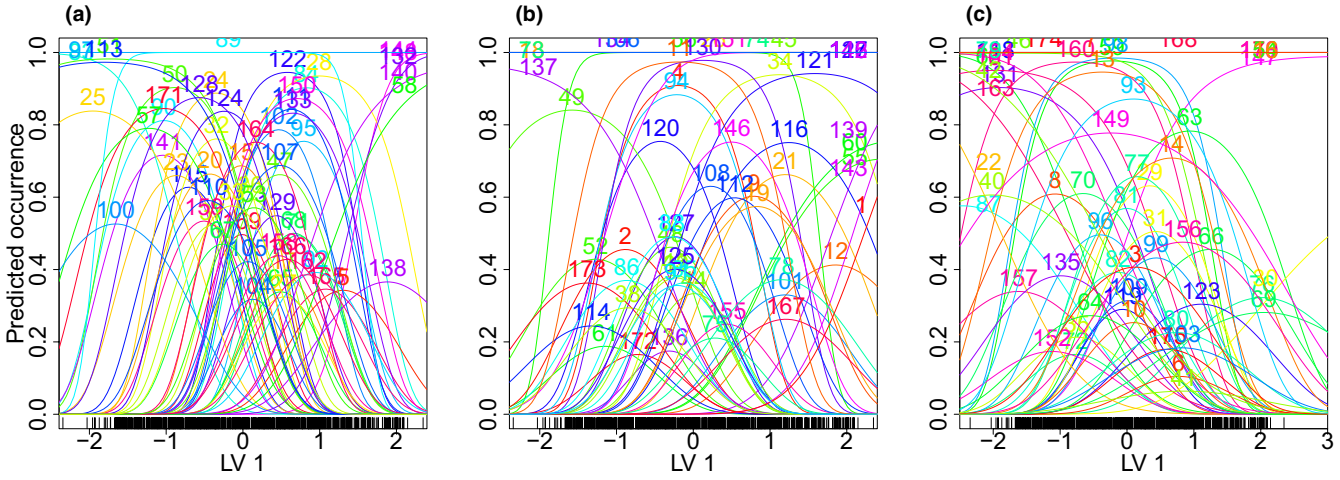

FIGURE 3 One-dimensional figures for the GLLVM fit to the Swiss alpine plants dataset. Each plot includes approximately one third of the species in the dataset, which have been sorted based on their variation explained, so that the first plot includes species explaining most of the variation. Plot (a) represents 63% of the residual variation, plot (b) represents 27% of the residual variation, and plot (c) represents 10% of the residual variation for the first latent variable. The rug plot at the bottom indicates predicted locations of the sites. The numbers correspond with the species names in Figure 4

of *Pardosa lugubris* and *Arctosa perita* included zero on all latent variables, in addition to *Arctosa lutetiana*. From all species on all latent variables, *Arctosa lutetiana* had the smallest tolerance (0.33, on the first latent variable).

6.2 | Swiss alpine plants

In the second application, n = 912 plots of 4 m^2 each were used to record binary data on p = 175 plant species. More species were recorded, but in the original study of this dataset species with less than 22 presences were excluded (D'Amen et al., 2018). Though fitting the model with these species would not have presented any computational issues, their estimates could not necessarily be expected to be accurate. Plots were located on a strong elevation gradient ranging from 375 m to 3,210 m a.s.l. (D'Amen et al., 2018). To improve computation time, we excluded 72 plots without any presences, and 103 plots with less than six presences, so that the final dataset included n = 737 plots.

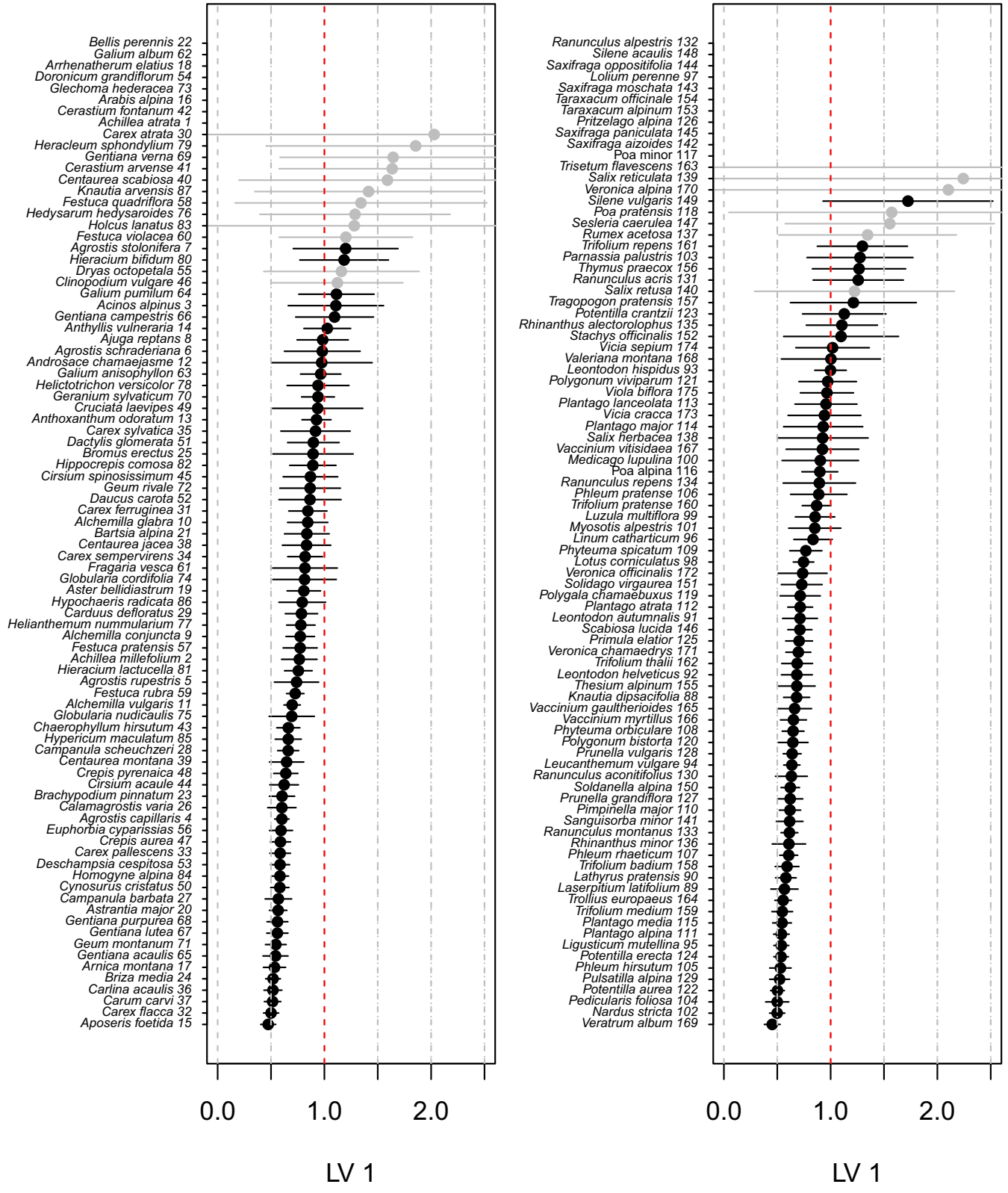

FIGURE 4 Species tolerances and approximate 95% confidence intervals derived using the Delta method, of the first latent variable from the model with unequal tolerances, fitted to the Swiss plants dataset. When tolerances cross 1 (indicated with a red dashed line), species have partially unobserved niches (regardless of the location of their optima). The panels show the first and second half of species in the dataset, respectively, ordered by the size of their tolerances. Species of which the confidence interval for the quadratic coefficients crosses 0 are shown in grey. Species at the top of the plot, seemingly without tolerances, exhibit near linear responses, so that their tolerances are very large. Grey dashed lines are added at increments of 0.5 as visual aid

Instead of selecting the optimal number of latent variables, we directly fitted the proposed GLLVM with quadratic response model to the data, using a Bernoulli distribution and with $d = 2$ latent variables, for the purpose of constructing an ordination diagram. We tested different sets of initial values and retained the model that had the highest log-likelihood.

The first latent variable explained 75% of the overall residual variation in the model, of which 50% was accounted for by the linear term. The length of the first ecological gradient was 4.79 (3.94–5.64, 95% confidence interval), and the length of the second ecological gradient 3.66 (2.86–4.45). Since the first latent variable explained considerably more residual variation than the second, we here focus our inference on that alone for illustration purposes. The species response curves for the first latent variable are visualized in Figure 3a–c. To improve readability, species are numbered by their location in the dataset, for which the corresponding names are included in Figure 4, which also shows species tolerances for the first latent variable, with approximate 95% confidence intervals.

The original dataset additionally included multiple predictor variables, measuring the growing degree-days above zero, a moisture index, total solar radiation over the year, slope, topography and elevation (van der Veen et al., 2021). In an attempt to identify the ecological gradient represented by the first latent variable, we post hoc calculated correlation coefficients between the predictors and the first latent variable. From all predictor variables, elevation was most correlated with the first latent variable (a correlation coefficient of 0.93), though this was collinear with growing degree-days above zero and the moisture index. We additionally fitted two unconstrained GLLVMs with linear species responses and with two latent variables, one of which included a random row intercept, and again calculated a correlation coefficient between the latent variables and elevation. Jamil and ter Braak (2013) showed that a mixed-effects model with random row intercept can account for the squared term of the latent variable. Here, the random row intercept was indeed related to the square of the first latent variable (correlation coefficient of −0.82). The GLLVM with linear species responses but without a row intercept estimated the ecological gradient less successfully (highest correlation coefficient with the elevation predictor of −0.71), than when a row intercept was included (highest correlation coefficient of 0.92). To test more explicitly for the effect of elevation, we additionally fitted a GLLVM with quadratic latent responses and elevation included as a predictor (both the linear and quadratic term, but without sign constraints, though most species exhibited concave curves), and with two latent variables. Including the predictor variable reduced the residual variance to 36% of that in the unconstrained model. The results presented here are from the unconstrained model, though the effect of elevation is presented in Appendix S5, Figure S1.

Of the $p = 175$ species included in the model, 36 had optima that were unobserved, of which 20 were larger than 10 or smaller than −10. The environmental tolerances from species of which the confidence interval for the quadratic coefficients on the first latent variable did not include zero ranged from 0.45 (*Veratrum album*) to 1.72 (*Silene vulgaris*) with a median tolerance of 0.73 and a standard deviation of 0.22. We examined groups of plants at the extremes of the gradient, that is, plants that had optima of minus two or smaller, and plants with optima of two or larger, to further investigate whether the estimated latent variable from the GLLVM with quadratic responses represented an elevation gradient. This approach allowed us to distinguish two groups of plants, the first indicative of lowlands (see Figure 3). In contrast, plant species included on the opposite side of the latent variable were clearly indicative of alpine conditions. Here, we focus our inference on the alpine plants, as those are likely to be most affected by climate change (Walther et al., 2005). All species with optima larger than 2 had confidence intervals for the quadratic coefficients that included zero. Three alpine species had optima located between 1.5 and 2: *Androsace chamaejasme* (1.85, −0.10 to 3.81), *Polygonum viviparum* (1.59, 0.77–2.40) and *Salix herbacea* (1.88, −0.04 to 3.80). Of these three species, *Salix herbacea* had the lowest maximum: −0.34. All three species had a wide response curve on the first latent variable, with tolerances near 1.

Figure 4 clearly shows some species that have smaller tolerances, thus more specialized species are present in the dataset. Six species had a tolerance of 0.50 or smaller: *Aposeris foetida*, *Carex flacca*, *Nardus stricta*, *Pedicularis foliosa*, *Potentilla aurea* and *Veratrum album*.

7 | DISCUSSION

In this article, we extended the GLLVM approach of Hui et al. (2015), to estimate the niches of species with quadratic responses to unobserved ecological gradients. We fitted and performed inference for the GLLVM with quadratic response model by extending the VA approach from Hui et al. (2017). The relation between latent variable models (i.e. unobserved ecological gradients) and ecological niches has been well-described for classical ordination methods (ter Braak & Prentice, 1988; Jongman et al., 1995), yet a method (either classical or model-based) to perform unconstrained (residual) ordination without limiting assumptions for species tolerances has not been available to date.

The similarity in responses of species to unobserved environments can be assessed by examining optima and tolerances, for example, visually using an ordination diagrams, to identify overlap in species distributions, or alternatively by examining a matrix of residual correlations between species. Determining if species exhibit fully quadratic curves in response to ecological gradients, whether tolerances are the same for all species per ecological gradient, or if the equal tolerances assumption is suited for a dataset, comes down to a problem of model selection for GLLVMs. To that end, future research can further investigate approaches such as regularization (e.g. possibly extending the approach of Hui et al., 2018), hypothesis testing or the use of confidence intervals of the quadratic coefficients. Similar to DCA, the GLLVM with quadratic response model provides estimates of gradient length. Here, gradient length is calculated from the quadratic coefficients, which are estimated via a variational approximation approach to maximizing the marginal likelihood function.

For datasets with 50 species and 50 sites or more, the GLLVM with separate quadratic responses for all species accurately retrieved ecological gradients and species-specific parameters, though for continuous responses or counts it was possible to accurately estimate parameters with fewer species or sites. In general, when fitting the GLLVM with quadratic response model to binary or ordinal responses, more information is required than for other data types (similarly as reported in Yee, 2004). However, this is conditional on the information content in a dataset, and the number of required sites and species here should only be considered as a rough rule of thumb. For observed environmental variables, ter Braak and Looman (1986) reported from simulations on estimates of species optima by weighted averaging that, 'with 10–13 presences, the variances of species optima are appreciable'. In our simulations, even with the number of sites fixed at $n = 100$, 24% of species had 13 or fewer presences, indicating difficulty in achieving a sufficient information content in presence–absence datasets to accurately estimate species optima.

We studied the response curves of species to ecological gradients for hunting spiders in a Dutch dune ecosystem (van der Aart & Smeek-Enserink, 1974), and for Swiss alpine plants (D'Amen et al., 2017), using the GLLVM with quadratic response model. Various specialist species can be identified in both datasets, as species with small tolerances on one or multiple latent variables. Specialist species are more likely to be affected by future changes in the environment, and as such their identification is of critical importance to community ecology, to better focus recommendations for conservation efforts.

Modelling rare species is often difficult in community ecology as few ordination methods have the capability to explicitly do so. The quadratic response model has great potential for community ecology, as it can simultaneously accommodate common (large tolerances and maximum i.e. a wide and high niche) and rare species (small tolerances and maximum i.e. a narrow and low niche). The quadratic response model naturally predicts species with unobserved optima, narrow niches and small maxima to have the fewest observations. Since the quadratic response model includes two species-specific parameters per latent variable, and thus requires more information in the data for accurate estimation of parameters than when assuming linear species responses, it potentially requires a large dataset to include sufficient information on rare species and accurately estimate the corresponding parameters. However, the example in this paper using the dataset of counts for hunting spiders (van der Aart & Smeek-Enserink, 1974) suggests that a GLLVM with quadratic response model can be feasible to fit even to small datasets. Regardless, assuming quadratic coefficients to be the same for all species per latent variable might be more realistic for many ecological datasets, while still providing the benefit of an explicit quadratic response model, with all the benefits it provides—calculating species optima, tolerances, maxima, gradient length and their corresponding statistical uncertainties. An additional advantage of a GLLVM-type approach is the ability to use information from both common and rare species to improve estimation of ecological gradients. Even if optima of species with too few observations cannot be accurately estimated, species preferences can be identified based on the ecological gradient, in relation to the response curve of more common species, and based on the direction of the maximum (slope). Without penalization or borrowing information for estimation from more abundant species though, the (quadratic) coefficients for species with few observations are not necessarily expected to be accurate.

An easy-to-use implementation of the quadratic response model with GLLVMs is available in the `gllvm` R package (Niku et al., 2020).

ACKNOWLEDGEMENTS

The authors thank the Spatial Ecology Group at the University of Lausanne, and Manuela D'Amen in specific, for providing the elevation data from the Swiss Alpine plants dataset. The elevation data were originally retrieved from the Swiss Federal Office of Topography. They thank Cajo ter Braak and an anonymous reviewer for helpful comments on earlier drafts of the manuscript. B.V. was supported by a scholarship from the Research Council of Norway (grant number 272408/F40). F.K.C.H. was supported by two Australian Research Council Discovery grants.

AUTHORS' CONTRIBUTIONS

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

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1111/2041-210X.13595.

DATA AVAILABILITY STATEMENT

The hunting spider dataset from the first example is available in the mvabund R package (Wang et al., 2012). The Swiss alpine plants dataset from the second example is available for download in the dryad database (D'Amen et al., 2017), with separate elevation data (van der Veen et al., 2021).

ORCID

Bert van der Veen https://orcid.org/0000-0003-2263-3880
Francis K. C. Hui https://orcid.org/0000-0003-0765-3533
Knut A. Hovstad https://orcid.org/0000-0002-7108-0787
Erik B. Solbu https://orcid.org/0000-0002-6023-3100
Robert B. O'Hara https://orcid.org/0000-0001-9737-3724

REFERENCES

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

Blanchet, F. G., Cazelles, K., & Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters*, *23*, 1050–1063. https://doi.org/10.1111/ele.13525

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*, 1045–1055. https://doi.org/10.2307/1940179

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). Springer-Verlag.

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*, https://doi.org/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. https://doi.org/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. *bioRxiv*, 2020.03.05.980060.

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika*, *58*, 453–467. https://doi.org/10.1093/biomet/58.3.453

Gauch, H. G. (1982). *Multivariate analysis in community ecology*. Cambridge University Press.

Halvorsen, R. (2012). A gradient analytic perspective on distribution modelling. *Sommerfeltia*, *35*, 1–165. https://doi.org/10.2478/v10208-011-0015-3

Hill, M. O., & Gauch, H. G. (1980). Detrended correspondence analysis: An improved ordination technique. *Vegetatio*, *42*, 47–58.

Hui, F. K. C. (2016). Boral Bayesian ordination and regression analysis of multivariate abundance data in r. *Methods in Ecology and Evolution*, *7*, 744–750.

Hui, F. K. C., Tanaka, E., & Warton, D. I. (2018). Order selection and sparsity in latent variable models via the ordered factor LASSO. *Biometrics*, *74*, 1311–1319. https://doi.org/10.1111/biom.12888

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

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

Inoue, K., Stoeckl, K., & Geist, J. (2017). Joint species models reveal the effects of environment on community assemblage of freshwater mussels and fishes in European rivers. *Diversity and Distributions*, *23*, 284–296. https://doi.org/10.1111/ddi.12520

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

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

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21.

Lacoste, É., Weise, A. M., Lavoie, M.-F., Archambault, P., & McKindsey, C. W. (2019). Changes in infaunal assemblage structure influence nutrient fluxes in sediment enriched by mussel biodeposition. *Science of the Total Environment*, *692*, 39–48. https://doi.org/10.1016/j.scitotenv.2019.07.235

MacArthur, R., & Levins, R. (1967). The limiting similarity, convergence, and divergence of coexisting species. *The American Naturalist*, *101*, 377–385. https://doi.org/10.1086/282505

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

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2019). Efficient estimation of generalized linear latent variable models. *PLoS One*, *14*, e0216129. https://doi.org/10.1371/journal.pone.0216129

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., Warton, D. I., van der Veen, B. (2020). GIlvm: Generalized linear latent variable models. https://github.com/JenniNiku/gllvm

Oksanen, J., & Tonteri, T. (1995). Rate of compositional turnover along gradients and total gradient length. *Journal of Vegetation Science*, *6*, 815–824. https://doi.org/10.2307/3236395

Øland, R. H. (1999). On the variation explained by ordination and constrained ordination axes. *Journal of Vegetation Science*, *10*, 131–136.

Øland, R. H., & Eilertsen, O. (1994). Canonical Correspondence Analysis with variation partitioning: Some comments and an application. *Journal of Vegetation Science*, *5*, 117–126.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153. https://doi.org/10.1198/tast.2010.09058

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*, 2–17. https://doi.org/10.1198/jcgs.2011.09118

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F. G., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*, 561–576. https://doi.org/10.1111/ele.12757

Paul, W. (2020). Covariate-adjusted species response curves derived from long-term macroinvertebrate monitoring data using classical and contemporary model-based ordination methods. *Ecological Informatics*, *60*, 101159. https://doi.org/10.1016/j.ecoinf.2020.101159

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. https://doi.org/10.1007/s004420100720

Pichler, M., & Hartig, F. (2020). A new method for faster and more accurate inference of species associations from novel community data. arXiv:2003.05331 [q-bio, stat]. Retrieved from http://arxiv.org/abs/2003.05331

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., & McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, *5*, 397–406.

ter Braak, C. J. F. (1985). Correspondence analysis of incidence and abundance data: Properties in terms of a unimodal response model. *Biometrics*, *41*, 859–873. https://doi.org/10.2307/2530959

ter Braak, C. J. (1986). Canonical correspondence analysis: A new eigenvector technique for multivariate direct gradient analysis. *Ecology*, *67*, 1167–1179. https://doi.org/10.2307/1938672

ter Braak, C. J. F., & Looman, C. W. N. (1986). Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, *65*, 3–11. https://doi.org/10.1007/BF00032121

ter Braak, C. J. F., & Prentice, I. C. (1988). In M. Begon, A. H. Fitter, E. D. Ford, & A. Macfadyen (Eds.), *A theory of gradient analysis*. *Advances in ecological research* (pp. 271–317). Academic Press.

Tikhonov, G., Ovaskainen, O., Oksanen, J., de Jonge, M., Opedal, O., & Dallas, T. (2021). Hmsc: Hierarchical model of species communities. https://CRAN.R-project.org/package=Hmsc

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. https://doi.org/10.1002/ecy.2754

van der Aart, P., & Smeek-Enserink, N. (1974). 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). Data from: {Model}-based ordination for species with unequal niche widths. *Dryad*, https://doi.org/10.5061/dryad.pnvx0k6m1
Walker, S. C., & Jackson, D. A. (2011). Random-effects ordination: Describing and predicting multivariate correlations and co-occurrences. *Ecological Monographs*, *81*, 635–663. https://doi.org/10.1890/11-0886.1
Walther, G.-R., Beißner, S., & Burga, C. A. (2005). Trends in the upward shift of alpine plants. *Journal of Vegetation Science*, *16*, 541–548. https://doi.org/10.1111/j.1654-1103.2005.tb02394.x
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.
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 in Ecology & Evolution*, *30*, 766–779. https://doi.org/10.1016/j.tree.2015.09.007
Wehrden, H. V., Hanspach, J., Bruelheide, H., & Wesche, K. (2009). Pluralism and diversity: Trends in the use and application of ordination methods 1990–2007. *Journal of Vegetation Science*, *20*, 695–705. https://doi.org/10.1111/j.1654-1103.2009.01063.x
Yee, T. W. (2004). A new technique for maximum-likelihood canonical Gaussian ordination. *Ecological Monographs*, *74*, 685–701. https://doi.org/10.1890/03-0078
Zurell, D., Pollock, L. J., & Thuiller, W. (2018). Do joint species distribution models reliably detect interspecific interactions from co-occurrence data in homogenous environments? *Ecography*, *41*, 1812–1819. https://doi.org/10.1111/ecog.03315
Zurell, D., Zimmermann, N. E., Gross, H., Baltensweiler, A., Sattler, T., & Wüest, R. O. (2020). Testing species assemblage predictions from stacked and joint species distribution models. *Journal of Biogeography*, *47*, 101–113. https://doi.org/10.1111/jbi.13608

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: van der Veen B, Hui FKC, Hovstad KA, Solbu EB, O'Hara RB. Model-based ordination for species with unequal niche widths. *Methods Ecol Evol*. 2021;12:1288–1300. https://doi.org/10.1111/2041-210X.13595