Received: 19 February 2020 | Revised: 6 April 2021 | Accepted: 9 April 2021

DOI: 10.1002/env.2683

SPECIAL ISSUE PAPER

WILEY

Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models

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

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

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

[3]School of Mathematics and Statistics and Evolution and Ecology Research Centre, The University of New South Wales, Sydney, New South Wales, Australia

Correspondence
Jenni Niku, Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland.
Email: jenni.m.e.niku@jyu.fi

Funding information
Australian Research Council; Koneen Säätiö; Maj ja Tor Nesslingin Säätiö; Suomen Kulttuurirahasto

Abstract

In ecological community studies it is often of interest to study the effect of species related trait variables on abundances or presence-absences. Specifically, the interest may lay in the interactions between environmental and trait variables. An increasingly popular approach for studying such interactions is to use the so-called fourth-corner model, which explicitly posits a regression model where the mean response of each species is a function of interactions between covariate and trait predictors (among other terms). On the other hand, many of the fourth-corner models currently applied in the literature are too simplistic to properly account for variation in environmental and trait response and any residual covariation between species. To overcome this problem, we propose a fourth-corner latent variable model which combines the following three features: latent variables to capture the correlation between species, fourth-corner terms to account for environment-trait interactions, and species-specific random slopes for modeling excess heterogeneity between species in their environmental response. We perform an extensive numerical study comparing a variety of fourth-corner models available in the literature which account for the aforementioned sources of variation to varying degrees. Simulation results demonstrate that the proposed fourth-corner latent variable models performed well when testing for the fourth-corner (interaction) coefficients, across both Type I error and power. By comparison, some models that do not full account for all relevant sources of variation suffer from inflated Type I error leading to potentially misleading inference. The proposed method is illustrated by an example on ground beetle data.

KEYWORDS

community analysis, fourth-corner problem, generalized linear mixed model, joint species distribution model, multivariate abundance data, variational approximation

1 | INTRODUCTION

One of the main aims of statistical analyses in community ecology is to understand how species differ in their responses to the environment, and why. Specifically, if trait information on each species is measured, it is possible to study how

these traits mediate the effect of environmental conditions on species responses. In ecology, this problem of studying associations between environmental and trait variables using species abundance data is often known as the fourth-corner problem (Legendre, Galzin, & Harmelin-Vivien, 1997). Specifically, given three matrices defining the environmental data ($\boldsymbol{R}$), species abundances ($\boldsymbol{L}$), and species traits ($\boldsymbol{Q}$), we can use these to infer how the environmental variables and species traits are jointly related to species abundance. Most classical approaches to solving the fourth-corner problem use a generalized singular value decomposition applied to an environment-trait association matrix constructed using $\boldsymbol{R}$, $\boldsymbol{L}$, and $\boldsymbol{Q}$, thus leading to a pair of ordinations for making interpretations of the associations (Dolédec, Chessel, ter Braak, & Champely, 1996). Legendre et al. (1997) further introduced a hypothesis testing approach based on permutation testing to assess which associations between environmental and trait variables are significant. Classical methods were further developed in Dray and Legendre (2008), ter Braak, Cormont, and Dray (2012), and Dray et al. (2014). The strength of the interaction in these approaches is quantified by the fourth-corner correlation, the square of which is a score test statistic for trait-environment interaction in a Poisson log-linear model (ter Braak, 2017). One such classical method which has recently gained particular attention is double constrained correspondence analysis (Peng, ter Braak, Rico, & Van den Brink, 2021; ter Braak, Šmilauer, & Dray, 2018).

In the past decade, several model-based approaches have arisen in the literature for solving the fourth-corner problem, with a notable advantage being that with standardized environmental and trait variables, they also give a concrete measure of the effect size through the interpretation of relevant coefficients in the mean model. We now give an overview of these. Denote the abundances (counts, presence-absences, biomass) of m responses (species) recorded at n samples (sites) by $y_{ij}, i=1, \ldots, n, j=1, \ldots, m$. For each site i, a vector of k environmental variables, $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ik})'$, and for each species j, a vector of q trait variables $\boldsymbol{t}_j = (t_{j1}, \ldots, t_{jq})'$ are also measured. The more general form of the fourth-corner model for the mean responses, μ_{ij}, can then be formulated as

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

where $g(\cdot)$ is a known link function, β_{0j} are species-specific intercepts, the k-vector $\boldsymbol{\beta}_e$ includes main effects for environmental covariates, and the $k \times q$ matrix $\boldsymbol{B}_{et}$ consists of environmental-trait interaction terms (also known as the fourth-corner coefficients). In addition, the quantity r_i denotes random site effects which we include as means of row standardization to account for differences in species total abundance across sites. Jamil, Ozinga, Kleyer, and ter Braak (2013) and Jamil and ter Braak (2013) noted that the random site effects r_i can also accommodate quadratic response to the environment. The k-vector $\boldsymbol{b}_j$ denotes species-specific random effects for environmental variables. The precise models considered so far in the literature differ in the way the random effects are included in model (1). For instance, in the generalized linear model (GLM) approach by Brown et al. (2014), random site effects r_i and random slopes $\boldsymbol{b}_j$ were ignored, and Warton, Shipley, and Hastie (2015) proposed inference on $\boldsymbol{B}_{et}$ based on bootstrapping the set of n vector residuals across the sites. Warton, Shipley, and Hastie (2015) also showed that the method proposed in Brown et al. (2014) is a generalization of a maximum entropy approach (community assembly via trait selection, CATS) proposed by Shipley, Vile, and Garnier (2006). Easier interpretation, model selection and inference methods for CATS-regression are thus readily available. Pollock, Morris, and Vesk (2012) proposed a generalized linear mixed modeling (GLMM) approach for solving the fourth-corner problem by including species specific random intercepts, β_{0j}, and random slopes for environmental variables $\boldsymbol{e}_i$ in the model. The model was later extended by Jamil and ter Braak (2013) with the inclusion of the random site effects r_i. Most recently, ter Braak (2019) proposed to further include a site-dependent random slope for trait variables $\boldsymbol{t}_j$.

A potentially major drawback with many of the model-based approaches listed above is that they do not model any potential residual correlation between the responses. Such residual covariation could arise from a variety of sources, including but not limited to similarity in response to shared but unobserved predictors, and biotic interactions between species. More importantly, if not accounted for, this can lead to potentially invalid inference on a variety of aspects in the model (Warton et al., 2015). As mentioned above, Warton, Shipley, and Hastie (2015) attempted to circumvent this issue by resampling sites. Pollock et al. (2012) and Jamil and ter Braak (2013) took into account the randomness at the individual species level. However, modeling residual correlation across species by a single random site intercept is often too simplistic in practice (Warton, Foster, Deáth, Stoklosa, & Dunstan, 2015). In the context of testing environmental-trait interactions, the problem of ignoring residual interspecific variation to the environment (not explained by traits) was studied by ter Braak, Peres-Neto, and Dray (2017) in detail. Specifically, ter Braak et al. (2017) compared four different resampling strategies in the GLM framework and noted that resampling (bootstrapping or permuting) either sites or species tended to yield Type I error rates that were too large when testing for the fourth-corner coefficients. The $p_{\max}$ permutation test (ter Braak et al., 2012), where two separate resampling tests (site-level and species-level) are performed

and the significance is assessed by the largest of the two p-values, was shown to perform best when the data were generated according to a simple GLMM model. However, the $p_{\max}$ test also produced inflated Type I errors when simulating from models where observed trait and environmental variables interact with latent trait and environmental variables.

An alternative approach to resampling-based procedures for testing the environmental-trait interactions is to try to construct a model that explicitly takes into account all the relevant sources of variation, including between species correlation and interspecific variation, through the inclusion of one or more random effects (or variations thereof.) In doing so, we can then employ, say, likelihood-based methods for estimation and inference. Recently, ter Braak (2019) used such as approach by introducing a GLMM based fourth-corner multilevel model (we refer to it by GLMM3), which differs from equation 1 primarily by adding site-specific random slopes for trait variables, $\mathbf{t}_j'\mathbf{u}_i$ where the $\mathbf{u}_i$ are independent multivariate normal random effects with variance-covariance matrix $\mathbf{\Sigma}_u$. ter Braak (2019) compared different model-based testing approaches (likelihood-ratio test, parametric bootstrap test and permutation-based $p_{\max}$ test) for testing the fourth-corner interaction term, and the likelihood ratio tests based on the GLMM3 model was shown to outperform other GLM and GLMM based tests for interaction. However, a potential issue with this approach is that it assumes that all between-species correlation can be captured by the measured traits. Specifically, adding $\mathbf{t}_j'\mathbf{u}_i$ to the linear predictor is equivalent to introducing a multivariate normal random intercept ϵ_{ij} with variance-covariance matrix proportional to $\mathbf{T}\mathbf{\Sigma}_u\mathbf{T}'$, where $\mathbf{T}$ is a matrix whose jth row is $\mathbf{t}_j$. The question is whether such a term is sufficient to capture between species correlation.

We consider here another potential model, namely, a generalized linear latent variable model (GLLVM), that explicitly accounts for between species correlations using a factor analytical approach. The past 5 years has seen an explosion in the use of latent variable models for community level modeling; see Warton, Blanchet, et al. (2015), Warton et al. (2016), Ovaskainen et al. (2017), Bjork, Hui, O'Hara, and Montoya (2018) Niku, Warton, Hui, and Taskinen (2017), among many others. The fourth-corner latent variable model, which we consider in this article, builds on the models proposed previously in Jamil and ter Braak (2013) and Warton, Blanchet, et al. (2015), while also extending the fourth-corner GLM of Brown et al. (2014) by including site-specific random row intercepts to account for the variation between sites, and species-specific random slopes for environmental variables for capturing the interspecific variation in responses not explained by the traits. In addition, latent variables with corresponding loadings are included to capture any residual correlation between species which is not explained by observed environmental and trait variables. While similar models have also been developed in the Bayesian context by Hui (2016) and Tikhonov, Opedal, Abrego, Lehikoinen, and Ovaskainen (2019), the performances of such models for assessing environment-trait interactions (in terms of producing valid inference) have not been studied before, let alone compared with existing procedures (including those reviewed above) in the literature.

To fit such models, we extend a fast and efficient maximum likelihood-based estimation algorithm, presented in (Niku et al., 2019; Niku, Hui, Taskinen, & Warton, 2019), for the fourth-corner latent variable model, and apply it to the environment-trait interaction testing problem. Specifically, when testing the fourth-corner coefficients, we employ a simple likelihood ratio testing approach. Importantly, this is made possible by including the necessary terms in the mean structure to ensure that all relevant sources of heterogeneity and residual correlation are accounted for. The performance of the proposed interaction test is compared with tests based on GLMMs (Jamil & ter Braak, 2013; Pollock et al., 2012; ter Braak, 2019; ter Braak et al., 2017), as well as with the $p_{\max}$ permutation test (ter Braak et al., 2012) through simulation studies, as we investigate both Type I error rates under the null hypothesis and powers of the various tests under several scenarios.

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

2 | MODEL DEFINITION AND ESTIMATION

Using the notation previously introduced, we propose a fourth-corner latent variable model with random site effects (intercepts) and random slopes, as defined by the following mean regression model,

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

or equivalently formulated in a hierarchical fashion,

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

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

Model (3) serves as a unifying framework that encompasses models proposed previously in Pollock et al. (2012), Jamil and ter Braak (2013), and Brown et al. (2014). If we set all variances of random effects, r_i and $\boldsymbol{b}_j$, and latent variables $\boldsymbol{u}_i$ in model (3) to zero, the model reduces to the fourth-corner GLM of Brown et al. (2014). If we set the covariance matrix of random row effects and latent variables, $\boldsymbol{\Sigma}_u$, to zero, we get a similar model as in Pollock et al. (2012), with an exception that Pollock et al. (2012) treated species-specific intercepts, β_{0j}, as random. Jamil and ter Braak (2013) extended the model proposed in Pollock et al. (2012) by adding random site effects, r_i, while ter Braak (2019) added site-specific random trait effects in the model. The latter is closely related to the method proposed here, in that it also models residual species covariance matrix $\boldsymbol{\Sigma}$ via rank reduction. The key distinction however is that the method of ter Braak (2019) method assumes $\boldsymbol{\Sigma}$ is a quadratic function of measured traits $\mathbf{t}_j$, while our latent variable approach imposes no restrictions on the residual covariance matrix, beyond restricting its rank.

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

$$l(\boldsymbol{\Psi}) = \log\left\{\int f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u}; \boldsymbol{\Psi}) f(\boldsymbol{r}, \boldsymbol{u}; \boldsymbol{\Sigma}_u) f(\boldsymbol{b}; \boldsymbol{\Sigma}_b) d(\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u})\right\}. \quad (4)$$

For multivariate abundance data, the response distribution $f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{b}, \boldsymbol{u}; \boldsymbol{\Psi})$ is not assumed to be a multivariate normal distribution (since the responses are usually discrete with a strong nonconstant mean-variance relationship). Consequently, the integration over latent variables and random effects does not have a closed form. To overcome this issue then,

a common and computationally efficient approach is to approximate the integral using approaches such as the Laplace (Niku, Warton, et al., 2017) or variational (Hui et al., 2017) approximation, which subsequently provide either a closed or nearly closed form approximation to the marginal log-likelihood (4). In Niku, Brooks, et al. (2019) it was shown that computationally convenient estimation algorithms for GLLVMs can be obtained by combining the Laplace or variational approximation methods with automatic optimization techniques implemented in R software, for computationally efficient estimation. Either the variational approximation or the Laplace approximation could be used to fit the proposed model. In this article, we choose to use the variational approximation in the simulations due to its computational efficiency, and the Laplace approximation in the application to allow the calculation of a common information criteria across methods. Applying the Laplace approximation to the chosen model is relatively straightforward, and goes similarly to the applications in Niku, Warton, et al. (2017) and Niku, Brooks, et al. (2019), and so we focus more closely here on the details of the variational approximation.

Specifically, we adopt the variational approximation approach for approximating the marginal log-likelihood in (4). Some theory has been developed to support the quality of variational approximations in general, for example, in (Hall, Pham, Wand, & Wang, 2011; Hui et al., 2017), and Wang and Blei (2019). However, theoretical study for the approximation of the specific type of model presented above has not been done as yet, and we view this as a subject of future research. As part of using the variational approximation method, we need to define so-called variational distributions for the random effects $\boldsymbol{r}$ and $\boldsymbol{b}$, and the latent variables $\boldsymbol{u}$, which effectively act as the approximate posterior distributions for these latent quantities. For ease of computation, while also being a sensible choice in an asymptotic sense (Blei, Kucukelbir, & McAuliffe, 2017; Hui et al., 2017), we propose to use independent normal distributions. Specifically, for $i = 1, \ldots, n$, we set the variational density of the random effects $q(r_i, \boldsymbol{u}_i)$ as independent multivariate normal distributions $N_{d+1}(\boldsymbol{a}_i, \boldsymbol{A}_i)$, while for response $j = 1, \ldots, m$ we set the variational density of the random effects $q(\boldsymbol{b}_j)$ as independent multivariate normal distributions $N_k(\boldsymbol{a}_{bj}, \boldsymbol{A}_{bj})$. Here, $\boldsymbol{a}_i$ and $\boldsymbol{a}_{bj}$ denote mean vectors of length $(d+1)$ and k, respectively, while $\boldsymbol{A}_i$ and $\boldsymbol{A}_{bj}$ are assumed to be positive definite and unstructured covariance matrices of dimension $(d+1)\times(d+1)$ and $k \times k$, respectively. Following these assumptions, and assuming that y_{ij} comes from the exponential family of distributions with mean $\mu_{ij} = E(y_{ij})$, such that $f(y_{ij}|r_i, \boldsymbol{b}_j, \boldsymbol{u}_i; \boldsymbol{\Psi}) = \exp\{(y_{ij}\eta_{ij} + b(\eta_{ij}))/\phi_j + c(y_{ij}, \phi_j)\}$, where $b(\cdot)$ and $c(\cdot)$ are known functions, then the resulting variational log-likelihood function is given by

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

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

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

3 | SIMULATION STUDIES

Three simulation studies were conducted to evaluate the ability of the proposed fourth-corner latent variable model to account for unobserved random variation in multivariate count data, in comparison to a variety of other fourth-corner models currently available in the literature. In the first simulation setup, we study the Type I errors of the likelihood ratio test for testing the null hypothesis $H_0 : \boldsymbol{B}_{te} = \boldsymbol{0}$ based on the fourth-corner latent variable model in (3), for a situation where interspecific variation and correlation between species is inherent in data. The Type I error simulations are repeated in the second setting with a more complex correlation structure that is related to the traits. In the third setting, we examined the power of the proposed test, that is, the empirical probability of finding the significant interaction between environmental and trait variables, under varying alternative hypotheses. For comparison, we consider four variants of model (3), consisting of two GLMM models with and without species-specific random slopes for environmental variables (GLMM2 and GLMM1, respectively) and two GLLVM models with and without random slopes and d-dimensional latent variables (GLLVM2(d) and GLLVM1(d), respectively). These are denoted as follows

$$\begin{aligned}
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'\boldsymbol{\beta}_e + \boldsymbol{e}_i'\boldsymbol{B}_{te}\boldsymbol{t}_j, && \text{GLMM1}\\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + \boldsymbol{e}_i'\boldsymbol{B}_{te}\boldsymbol{t}_j, && \text{GLMM2}\\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'\boldsymbol{\beta}_e + \boldsymbol{e}_i'\boldsymbol{B}_{te}\boldsymbol{t}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, && \text{GLLVM1}(d)\\
g(\mu_{ij}) &= r_i + \beta_{0j} + \boldsymbol{e}_i'(\boldsymbol{\beta}_e + \boldsymbol{b}_j) + \boldsymbol{e}_i'\boldsymbol{B}_{te}\boldsymbol{t}_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j. && \text{GLLVM2}(d).
\end{aligned}$$

Here, we assume that d is known, but in practice, model selection tools, such as AIC and BIC, can be used to guide the selection (Burnham & Anderson, 2002). We simulated data under $d = 2$, as described below, and so results for $d = 1$ and $d = 0$ (GLMM2) give some insight into the performance of GLLVM when the covariance structure has been misspecified. As additional method for comparison, we also included the likelihood ratio test based on the multilevel model of ter Braak (2019), which we denote as GLMM3. The mean model of GLMM3 can be defined by

$$g(\mu_{ij}) = \beta_0 + b_{0j} + e_i'(\beta_e + b_j) + u_{0i} + t_j(\beta_t + u_i) + e_i'B_{te}t_j,$$

where b_j and u_i are assumed to be independent with variances σ_b^2 and σ_u^2. A random effect u_{0i} can be seen as a site-specific error term or site-specific random intercept, similar to r_i. Finally, we also included the $p_{\max}$ permutation test of ter Braak et al. (2017) for comparison. This approach applies the log-likelihood ratio tests to a Poisson GLM (the Poisson being used for computational efficiency, ter Braak et al., 2017), and involves taking the largest of the two p-values formed by permuting either rows or columns of predictors.

In the first simulation setup, we compared the Type I errors based on likelihood ratio tests to those of the $p_{\max}$ test. We generated datasets according to the negative binomial distribution using two sample sizes and dimensions: (a) $m = 40$ and $n = 70$, and (b) $m = 70$ and $n = 40$. As a simulation model, we used

$$\log(\mu_{ij}) = r_i + \beta_{0j} + t_j\beta_t + e_i'(\beta_e + b_j) + e_i'B_{te}t_j + \boldsymbol{u}_i'\boldsymbol{\gamma}_j, \tag{5}$$

with one trait and one environmental variable generated independently from the standard normal distribution. The species intercepts β_{0j} were generated independently from the uniform distribution $U(-1, 1)$, and $\beta_t = 0.3$. The value for the variance of the random row effects was $\sigma^2 = 0.3$, while we set $\beta_e = 0.3$. The fourth-corner coefficient B_{te} was set to zero in order to assess Type I error. The species-specific dispersion parameters were all set to $\phi_j = 0.5$.

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

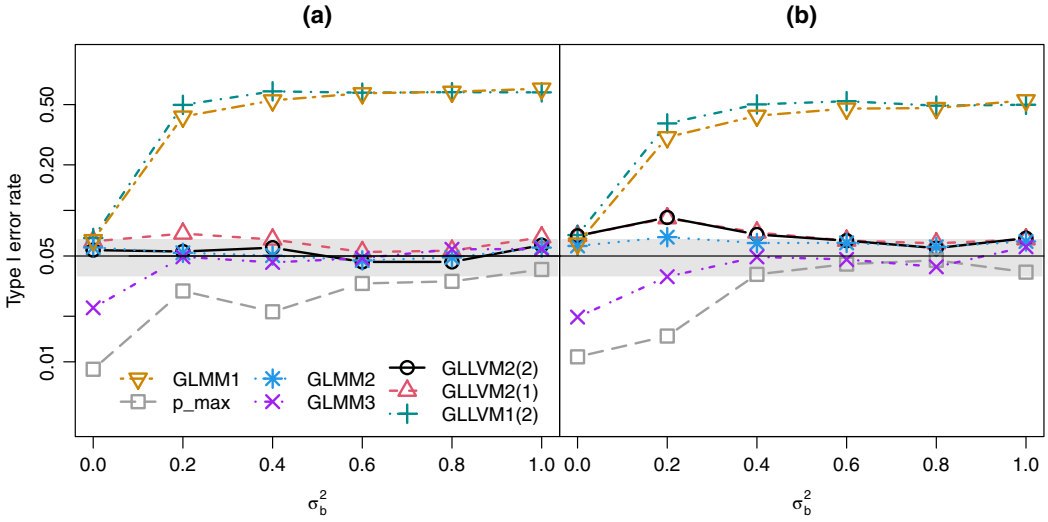

FIGURE 1 Type I error rates for likelihood ratio tests based on GLMM with random intercepts (GLMM1), GLMM with random intercepts and slopes (GLMM2), GLLVM with random intercepts (GLLVM1(2)), GLLVM with random intercepts and slopes and $d = 1$ or 2 latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3), and the $p_{\max}$ test applied to GLMs. On the left, data size is (a) $n = 70$ sites and $m = 40$ species, on the right (b) $n = 40$ sites and $m = 70$ species. The variance of the random slope effects, σ_b^2 is plotted on x-axis. A gray envelope around the nominal level 0.05 corresponds to values for sample proportions which are not significantly different from 0.05. GLLVM, generalized linear latent variable model; GLM, generalized linear model; GLMM, generalized linear mixed modeling

Given the above set up, we simulated 1000 datasets, assuming a negative binomial distribution for the response. For each dataset, we then calculated p-values based on likelihood ratio tests from the four models listed above, the likelihood ratio test based on GLMM3, as well as the $p_{\max}$ test applied to GLMs, for assessing the null hypothesis $H_0: B_{te} = 0$. The resulting Type I errors are presented in Figure 1. Results indicate that the fourth-corner latent variable models, GLLVM2(d) with $d = 1$ and $d = 2$, provided empirical Type I errors that were reasonably close to the nominal significance level of 5% for all values of σ_b^2. The Type I errors for likelihood ratio tests based on models GLLVM1(2) and GLMM1, which do not include species-specific random slopes, were severely inflated for values of σ_b^2 greater than zero, while the likelihood ratio test based on any of the considered models with such random effect, including GLMM2 as well as GLMM3, provided empirical Type I errors close to the nominal level. A possible explanation for the failure of the GLLVM1(2) could be due to model underfitting. That is, the latent variable term may not have been able to properly and simultaneously account for both the variation caused by the latent variables and variation in species caused by the random slopes (especially given the latent variables and the environmental variable were generated independently of each other). The $p_{\max}$ test applied to GLMs controls the Type I error well, although tended to produce Type I errors below the nominal level especially for small values of σ_b^2. The $p_{\max}$ is based on the sequential test procedure (Goeman & Solari, 2010), and if the assumptions hold true, the test controls Type I error rate in the sense that it is smaller than or equal to the nominal level. Therefore, it is not considered as a fault, that the error rate is smaller than the nominal level. However, Type I errors well below the nominal level may be an indication of a lower power of the test in such scenarios compared with the other methods, which will be explored in the third simulation setup.

In the second simulation setup, we introduced correlation between the residual correlation term and the observed trait by setting $\text{corr}(\gamma_{j2}, t_j) = 0.5$, where γ_{j2} is the loading corresponding to the second latent variable u_{i2}. In practice, loadings γ_{j2} and traits t_j were generated from a bivariate normal distribution with unit variances and 0.5 correlation. This can be interpreted as a situation in which the effect of the observed trait differs between sites and is not fully explained by the observed environmental variables. This feature is highlighted in the additional simulation setup in Appendix B. The methods that provided inflated Type I errors in the previous setting, namely, GLLVM and GLMM without random slopes, were excluded from the comparison. Therefore, only five methods were compared. Type I errors presented in Figure 2 show that the likelihood ratio test based on the fourth-corner latent variable model with $d = 1$ (GLLVM2(1)) and the mixed model with random intercept and slope (GLMM2) both produced inflated Type I errors, especially for small values of σ_b^2. The results of the additional simulations (see Figure B1 in Appendix B) support the conclusion that

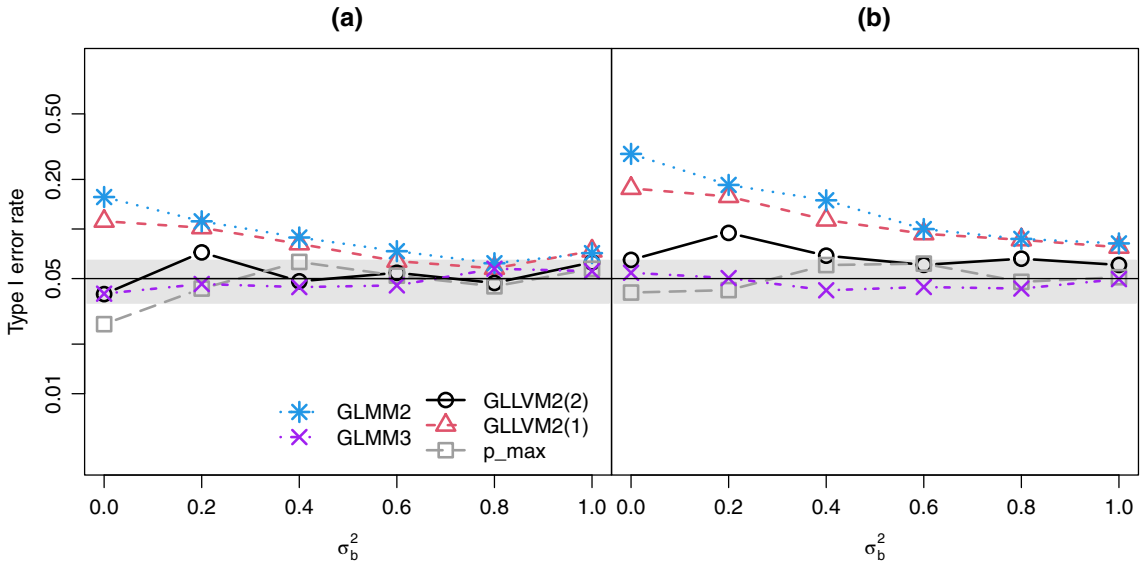

FIGURE 2 Type I error rates for likelihood ratio tests based on GLMM with random intercepts and slopes (GLMM2), GLLVM with random intercepts, slopes and $d = 1, 2$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3), and the $p_{\max}$ test applied to GLMs. On the left, data size is (a) $n = 70$ sites and $m = 40$ species, on the right (b) $n = 40$ sites and $m = 70$ species. The variance of the random slope effects, σ_b^2, is plotted on the x-axis. A gray envelope around the nominal level 0.05 corresponds to values for sample proportions which are not significantly different from 0.05. GLLVM, generalized linear latent variable model; GLM, generalized linear model; GLMM, generalized linear mixed modeling

the GLMM2 in particular is incapable of capturing the relevant sources of variation. The fourth-corner latent variable model with $d = 2$ typically maintained close to nominal Type I error, although rising to almost 0.1 in one case with a small sample size $n = 40$ and variance $\sigma_b^2 = 0.2$. Similar results can also be seen in Figure B1(b) in the Appendix B, where the simulation model was chosen to mimic a situation in which the effect of the observed trait differed between sites. Under such a model, the results then showed moderate inflation for GLLVM2(2) in one case. The $p_{\max}$ test applied to GLMs provided Type I errors close to the nominal level in Figure 2, but also exhibited slight inflation in the additional simulation in Appendix (see Figure B1(b)). The most consistent results were provided by the GLMM3 model, both in Figures 2 and B1(b), with Type I errors very close to the nominal level. Recall that the GLMM3 model differed from other GLMM implementations by having a random slope for traits, which uses traits to try and approximate the residual correlation structure across species. Overall, these results indicate that misspecification of the correlation structure of the responses can lead to invalid results in cases where correlation structure is not independent of the predictors.

In the third simulation setup, we compared the power of the various testing procedures. The methods that provided inflated Type I errors in the first simulation study were excluded from the comparison, meaning only five methods were included for comparison. We again generated 1000 datasets using the similar setup as in the first simulation study with $n = 70$ and $m = 40$, but varied the interaction term B_{te} such that $B_{te} \in \{0, 0.1, 0.2, 0.3, 0.4\}$. As variances for random slope effects, we considered $\sigma_b^2 \in \{0, 0.4, 0.8\}$. The power simulation for the setup with $n = 40$ and $m = 70$ was excluded as results were similar compared with the previous one. The resulting empirical powers of the $p_{\max}$ test and four different likelihood ratio tests are plotted in Figure 3. In all cases, the likelihood ratio tests based on any of the four models provide higher probabilities for detecting a true nonzero interaction between environmental and trait variables as compared with the $p_{\max}$ test. This was not surprising given the $p_{\max}$ test is, by construction, conservative since it involves performing two permutation tests and then choosing the more conservative of the two. Indeed, this conservatism was reflected in the Type I error results seen in Figure 1. Likelihood ratio tests applied to GLMM2 or GLMM3 performed well, but were slightly less powerful than the tests based on the fourth-corner latent variable models when the value for σ_b^2 was small. This can be seen especially when $B_{te} = 0.1$ in the case of $\sigma_b^2 = 0$. The difference between GLLVM with the correct number of latent variables and GLMM3 became slightly more clear when we used a more complex residual correlation structure defined by four latent variables (see Figure B3 in Appendix B for additional results), for which the residual covariance structure could not be well approximated by measured traits $\mathbf{t}_j$.

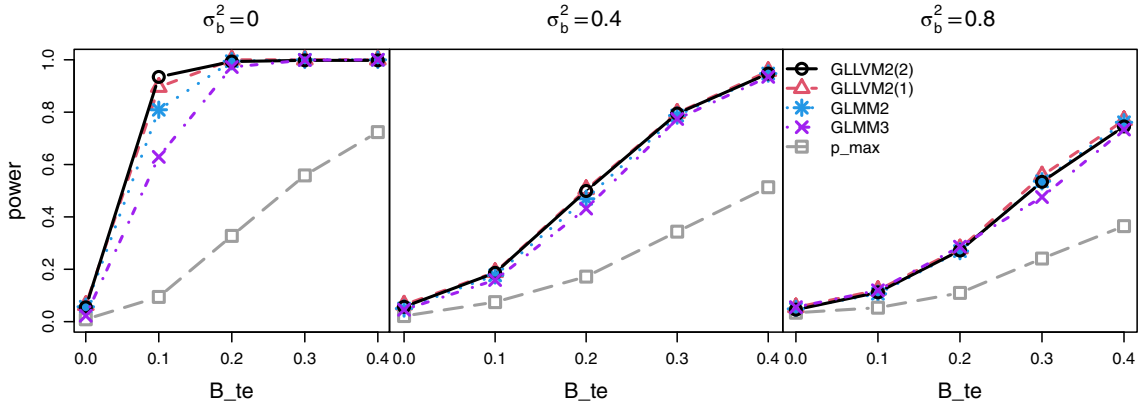

FIGURE 3 Power as a function of the effect size B_{te} for the likelihood ratio tests based on GLMM with random intercepts and slopes (GLMM2), GLLVM with random intercepts, slopes and $d = 1, 2$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3) and the $p_{\max}$ test applied to GLMs. GLLVM, generalized linear latent variable model; GLM, generalized linear model; GLMM, generalized linear mixed modeling

TABLE 1 The values of AIC for the two fourth-corner latent variable models, and the GLMM model fitted to the ground beetle dataset

	GLLVM2(1)	GLLVM2(2)	GLMM2	GLMM3	$p_{\max}$
AIC	18,294	**18,077**	19,098	18,913	NA
p-value	<0.001	<0.001	<0.001	<0.001	0.143

Note: Also shown are the p-values for the corresponding likelihood ratio test of the fourth-corner interaction terms, Bold values indicates lowest AIC value.
Abbreviations: GLLVM, generalized linear latent variable model; GLMM, generalized linear mixed modeling.

4 | CASE STUDY

We applied the proposed fourth-corner latent variable model to a dataset consisting of counts of $m = 68$ ground beetle species recorded at $n = 87$ sites across Scotland (Ribera et al., 2001). The original data also included 17 environmental variables recorded at each site and 20 trait variables for each species. Ribera et al. (2001) studied whether the morphology and life traits of ground beetle species can be related to the environmental variability of the habitats. For illustration purposes, we consider using a subset of $k = 4$ environmental variables: land use management intensity score (Management), percentage moisture content (Moist), elevation, and soil pH value, along with four species trait covariates: total length and pronotum height, overwintering (OVE, with two levels: 1= only adults; 2 = adults and larvae or only larvae), and breeding season (BRE, with three levels: 1 = spring; 2 = summer; 3 = autumn or winter). This set of environmental and trait variables were among the most important covariates affecting the ground beetle communities based on the analysis of Ribera et al. (2001). All quantitative covariates were centered and scaled to have variance one before the analysis, while dummy variables were set up for OVE and BRE, meaning there were a total of $q = 5$ predictors in the vector of traits $\boldsymbol{t}_j$.

We first tested if the interactions between environmental and trait covariates were significant using likelihood ratio tests based on the fourth-corner latent variable model with one and two latent variables, GLMMs with random row and slope parameters included, and the $p_{\max}$ test. Table 1 lists the AIC values for various models, as well as p-values given by four likelihood ratio tests and the $p_{\max}$. The GLLVM with random row effects and random slopes and two latent variables had the lowest value of AIC, suggesting that both latent variables and species-specific random effects were needed to model additional sources of (co)variation, while the p-values for all LR tests were less than 0.001 providing clear evidence of an interactions between the considered environmental and trait variables. By contrast, the $p_{\max}$ test with 999 permutations gives a p-value of 0.143 when testing for the fourth-corner interaction term. The result is thus consistent with the simulation study results showing the conservativeness of the $p_{\max}$ test. To ensure AIC values were comparable, all models were fitted using TMB (via `glmmTMB` or `gllvm`) using a Laplace approximation. Note that using a variational approximation method as detailed in Section 2 would have given similar p-values and conclusions, and so the choice of the approximation method itself is not critical. Rather, when comparing the AICs, because GLMM3 was designed for and

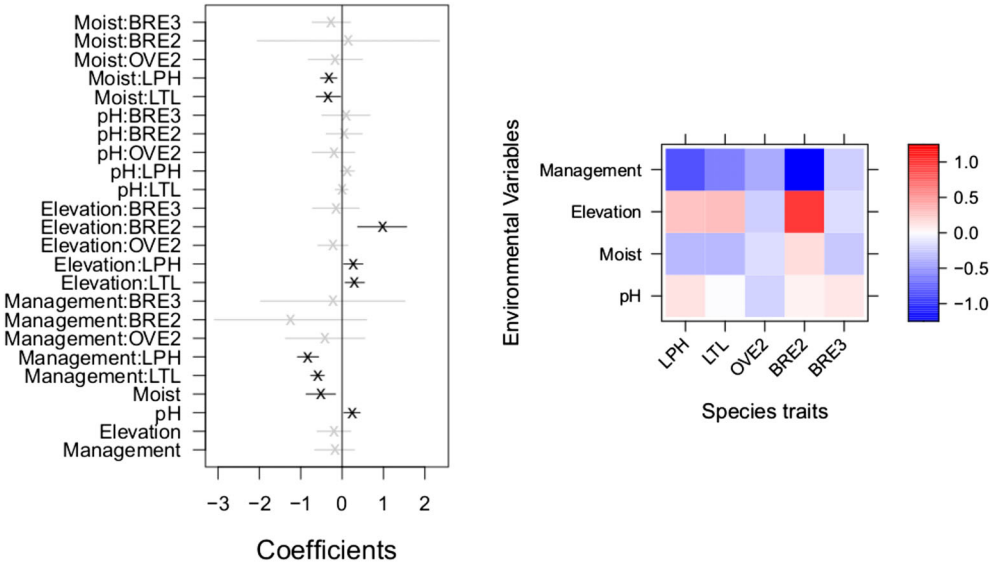

FIGURE 4 Point estimates and associated 95% confidence intervals for coefficients (left), along with a level plot (right) for fourth-corner interaction terms from a fourth-corner latent variable model with two latent variables fitted to the ground beetle data. The confidence intervals that do not contain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

fitted using the `glmmTMB` package which employs the Laplace approximation, then to facilitate comparison we also fitted our proposed method GLLVM2 (via the `gllvm` package) using a Laplace approximation.

The estimated coefficients for the environmental covariates and interaction terms based on the GLLVM with two latent variables are plotted in Figure 4. The strongest negative interactions were between management intensity and total length, as well as between management intensity and pronetum height. In other words, high management intensity was found to have a large negative effect on species that have larger body size. The strongest positive effects occurred in interactions between elevation and breeding season. That is, species having breeding season in summer succeeded better in high altitude environments as compared with species which breed during other seasons. Finally, predictions for species-specific random slopes for the environmental covariates and their associated 95% uncertainty intervals are plotted in Figure 5; the uncertainty intervals were constructed based on the conditional mean squared error of prediction (Booth & Hobert, 1998). From this, we can see that the interspecific variation in responses, which is not explained by the traits, is highest for the effect of the moisture content and management intensity and low or nearly nonexistent for the effect of the elevation and the pH value. Finally, we note that adjustments of the confidence intervals to account for multiple comparison was not done here, and we leave this as an avenue for future research.

5 | DISCUSSION

In this article, we have proposed a fourth-corner latent variable model that accounts for two key sources of error in current implementations of fourth-corner model, namely, the failure of traits to capture all interspecific variation (species-specific error), and the failure to account for the residual correlation between species (site-specific error) not explained by the environmental and trait variables. With a model-based approach, we are able to account for both sources of additional variation through the inclusion of additional species-specific random slopes, and site-specific latent variables. The approach is shown to be an extension of the recently introduced model-based approaches in Pollock et al. (2012), Jamil and ter Braak (2013), and Brown et al. (2014). We adopted an efficient estimation and inference approach based on variational approximations, and compared its finite sample performance to classical competitors for assessing the importance of fourth-corner interaction terms. Results showed that the proposed approach (GLLVM2) maintains close to nominal Type I error levels when testing for the fourth-corner coefficients, while power can be substantially better than resampling-based procedures. Importantly, models which fail to account for additional species-specific variation not due to traits, such as GLMM1 and GLLVM1, produced inflated Type I errors and hence misleading inference.

While GLMM3 and the proposed latent variable method (GLLVM) tended to maintain close to nominal Type I error, each method strayed from nominal levels in some simulations—GLMM3 occasionally being too conservative and GLLVM occasionally being too liberal (see Figure 1, also Figure B1). The former is the more desirable situation, although we do not think that the general behavior is due to a structural weakness of either method. While Type I error control tended to be better for GLMM3, this will not always be the case, as shown in an additional simulation with quadratic site-specific

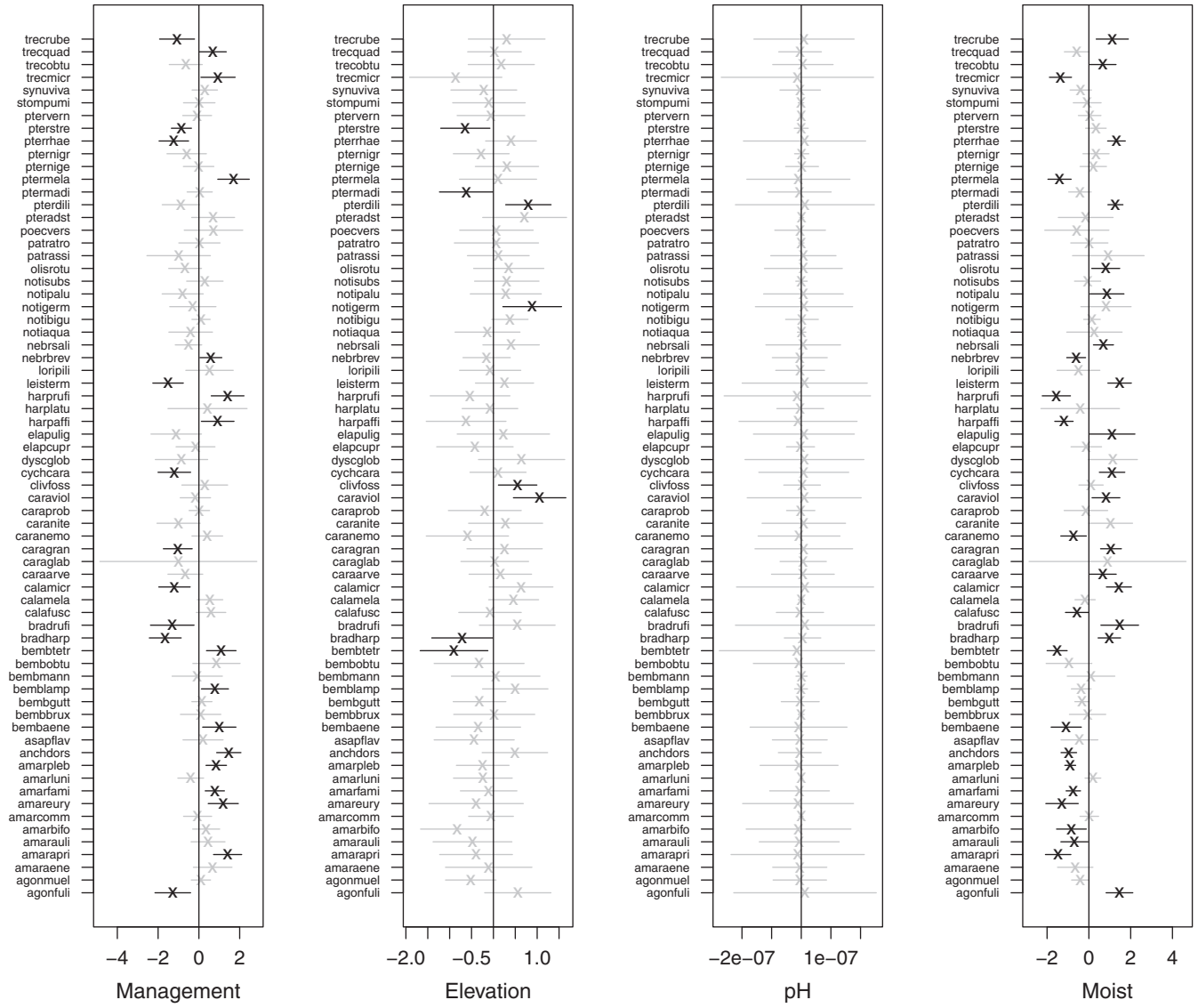

FIGURE 5 Point predictions for species-specific random slopes and associated 95% uncertainty intervals from a fourth-corner latent variable model with two latent variables fitted to the ground beetle data

effects of traits (see Figure B4 in Appendix B). In Appendix B Figure B2, we present an important scenario in which any regression model fails and so also all fourth-corner models will fail: namely when there are missing predictors that are correlated with the observed predictors. Such methods can fail here because it leads to confounding, thus biased estimation and uncertainty quantification for the associated regression coefficients (see, for instance, Paciorek, 2010, on the related issue of confounding).

Model (1) included a random effect to capture species-specific variation in environmental response, not captured by traits. Because species tend to respond to the environment in complex and sophisticated ways, and because our data collection process rarely captures all these reasons, it seems a sensible working assumption to always expect such species-specific variation. Simulations in ter Braak et al. (2017), and those in this article, emphasize the importance of including such a term. This article additionally shows that it is important to capture residual correlation in abundance across species, which can be achieved using latent variables, as in model (3). In future research, we will examine other data-driven approaches to selecting the number of latent variables (Hui, Tanaka, & Warton, 2018), as well as extensions to incorporate other sources of variation such as phylogenic (Ovaskainen et al., 2017) or spatio-temporal correlations (e.g., adapting the work of Ren & Banerjee, 2013; Thorson et al., 2016; Taylor-Rodriguez et al., 2019; Tikhonov et al., 2020), and imperfect detection (Tobler et al., 2019; Warton et al., 2016). Adjusting the inference for multiple comparisons when a moderate number of fourth-corner interactions terms are present is also a topic of future investigation.

ACKNOWLEDGMENTS

The authors would like to thank the associate editor and the reviewers for their invaluable feedback and suggestions that have greatly improved this article. The work of J. Niku was supported by the Maj ja Tor Nessling foundation and the Finnish Cultural Foundation and the work of S. Taskinen was supported by the Kone foundation. F.K.C. Hui was supported by two Australian Research Council Discovery grants.

ORCID

Jenni Niku https://orcid.org/0000-0002-7992-2598

REFERENCES

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

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

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

Brown, A. M., Warton, D. I., Andrew, N. R., Binns, M., Cassis, G., & Gibb, H. (2014). The fourth-corner solution - using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution*, *5*, 344–352.

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

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

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

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

Goeman, J. J., & Solari, A. (2010). The sequential rejection principle of familywise error control. *The Annals of Statistics*, *38*(6), 3782–3810.

Hall, P., Pham, T., Wand, M. P., & Wang, S. S. J. (2011). Asymptotic normality and valid inference for gaussian variational approximation. *The Annals of Statistics*, *39*(5), 2502–2532.

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

Hui, F. K. 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.

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*(1), 35–43.

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

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

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and laplace approximation. *Journal of Statistical Software*, *70*(5), 1–21.

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

Niku, J., Brooks, W., Herliansyah, R., Hui, F. K. C., Taskinen, S., & Warton, D. I. (2017). *gllvm: Generalized linear latent variable models*. R package version 1.2.3.

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*(5), e0216129.

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.

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

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., 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*(5), 561–576.

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

Peng, F.-J., ter Braak, C. J., Rico, A., & Van den Brink, P. J. (2021). Double constrained ordination for assessing biological trait responses to multiple stressors: A case study with benthic macroinvertebrate communities. *Science of the Total Environment*, *754*, 142171.

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

Ren, Q., & Banerjee, S. (2013). Hierarchical factor models for large spatially misaligned data: A low-rank predictive process approach. *Biometrics*, *69*(1), 19–30.

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

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

Taylor-Rodriguez, D., Finley, A., Datta, A., Babcock, C., Andersen, H., Cook, B., Morton, D., & Banerjee, S. (2019). Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. *Statistica Sinica*, *29*(3), 1155–1180.

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

ter Braak, C. J. F. (2017). Fourth-corner correlation is a score test statistic in a log-linear trait-environment model that is useful in permutation testing. *Environmental and Ecological Statistics*, *24*, 219–242.

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

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

ter Braak, C. J. F., Šmilauer, P., & Dray, S. (2018). Algorithms and biplots for double constrained correspondence analysis. *Environmental and Ecological Statistics*, *25*, 171–197.

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

Tikhonov, G., Duan, L., Abrego, N., Newell, G., White, M., Dunson, D., & Ovaskainen, O. (2020). Computationally efficient joint species distribution modeling of big spatial data. *Ecology*, *101*(2).

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

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

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

Warton, D., Foster, S., Deáth, G., Stoklosa, J., & Dunstan, P. (2015). Model-based thinking for community ecology. *Plant Ecology*, *216*, 669–682.

Warton, D. I., Blanchet, F. G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2016). Extending joint models in community ecology: A response to beissinger et al. *Trends in Ecology & Evolution*, *31*(10), 737–738.

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

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

SUPPORTING INFORMATION

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

How to cite this article: Niku J, Hui FKC, Taskinen S, Warton DI. Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics*. 2021;e2683. https://doi.org/10.1002/env.2683

APPENDIX A. PROOF FOR THE VARIATIONAL APPROXIMATION OF THE LIKELIHOOD FUNCTION

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

$$\underline{\ell}(\boldsymbol{\Psi}, \boldsymbol{\xi}) = \int \log\left\{\frac{f(\boldsymbol{y}|\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}; \boldsymbol{\Psi}) f(\boldsymbol{r}, \boldsymbol{u}; \boldsymbol{\Sigma}_u) f(\boldsymbol{b}; \boldsymbol{\Sigma}_b)}{q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b})}\right\} q(\boldsymbol{r}, \boldsymbol{u}) q(\boldsymbol{b}) d(\boldsymbol{r}, \boldsymbol{u}, \boldsymbol{b}),$$

$$
\begin{aligned}
&= \int (\log f(\boldsymbol{y}|\boldsymbol{r},\boldsymbol{u},\boldsymbol{b};\boldsymbol{\Psi}) + \log f(\boldsymbol{r},\boldsymbol{u};\boldsymbol{\Sigma}_u) + \log f(\boldsymbol{b};\boldsymbol{\Sigma}_b) - \log q(\boldsymbol{r},\boldsymbol{u}) - \log q(\boldsymbol{b})) \\
&\quad \times q(\boldsymbol{r},\boldsymbol{u})q(\boldsymbol{b})d(\boldsymbol{r},\boldsymbol{u},\boldsymbol{b}), \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} E_q\{\log f(\boldsymbol{y}_{ij}|(r_i,\boldsymbol{u}_i,\boldsymbol{b}_j),\boldsymbol{\Psi})\} + \sum_{i=1}^{n} E_q\{\log f((r_i,\boldsymbol{u}_i;\boldsymbol{\Sigma}_u))\} \\
&\quad + \sum_{j=1}^{m} E_q\{\log f(\boldsymbol{b}_j;\boldsymbol{\Sigma}_b)\} + \sum_{i=1}^{n} E_q\{-\log q(r_i,\boldsymbol{u}_i|\boldsymbol{\xi})\} + \sum_{j=1}^{m} E_q\{-\log q(\boldsymbol{b}_j|\boldsymbol{\xi})\},
\end{aligned}
$$

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

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

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

APPENDIX B. ADDITIONAL SIMULATIONS

B.1 Additional simulations with random parameters

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

$$
g(\mu_{ij}) = \mu_0 + R_i + C_j + b_{te}t_je_i + b_{ze}z_je_i + b_{tx}t_jx_i + b_{zx}^*z_j^*x_i^* + \epsilon_{ij}, \tag{B1}
$$

and variance $V(\mu_{ij}) = \mu_{ij} + \mu_{ij}^2$. Here intercept equals $\mu_0 = \log(30)$. Row effects were generated as $R_i = a_0e_i + a_1e_i^2 + \epsilon_{ri}$, with $\epsilon_{ri} \sim N(0, 0.01)$, and column effects similarly by $C_j = c_0t_j + c_1t_j^2 + \epsilon_{tj}$, with $\epsilon_{tj} \sim N(0, 0.01)$. Observed environmental

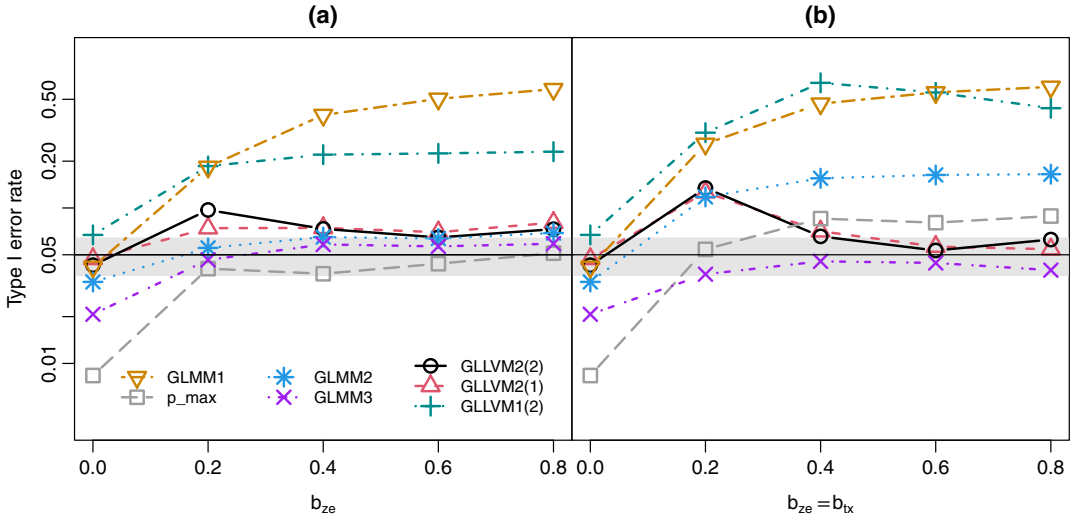

FIGURE B1 Type I error rates obtained using simulation setup described in Appendix B.1 for likelihood ratio tests based on GLMM with random intercepts (GLMM1), GLMM with random intercepts and slopes (GLMM2), GLLVM with random intercepts (GLLVM1(2)), and GLLVM with random intercepts, slopes, and $d = 1, 2$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3) and the $p_{\max}$ test. Generated datasets consisted of $n = 40$ sites and $m = 40$ species. A gray envelope around the nominal level 0.05 corresponds to values for sample proportions which are not significantly different from 0.05. GLLVM, generalized linear latent variable model; GLMM, generalized linear mixed modeling

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

Based on the results in Figure B1(a,b), the likelihood ratio test based on the GLLVMs with one and two latent variables, random slopes and random row effects provided Type I errors close the nominal level 0.05 in all considered cases excluding the case $b_{ze} = b_{tx} = 0.2$ where the Type I errors exceeded significantly the nominal level 0.05. Such peak is seen in all conducted simulation setups with a small sample size. The likelihood ratio test based on the GLMM with random slopes and random row effects produced close to valid Type I errors for the random trait case (Figure B1(a)) but inflated Type I errors for the random trait and random env case (Figure B1(b)). The $p_{\max}$ test applied for GLM worked quite well for the random trait case, but produced slightly inflated Type I errors for the random trait and random environmental variable case when effect sizes for b_{ze} and b_{tx} were larger than 0.4. The likelihood ratio tests based on GLLVM and GLMM which did not include random slopes produced too large Type I errors.

In Figure B2, the Type I errors were calculated using the same mean model as above, except observed environmental variables e_i and latent environmental variables x_i as well as observed traits t_j and latent traits z_j were generated so that they were correlated, that is, $\text{corr}(e_i, x_i) = 0.3$ and $\text{corr}(t_j, z_j) = 0.3$. Such correlations lead to a confounding effect and the results show that if this is the case all methods produced too inflated Type I errors.

B.2 Simulations with four latent variables

We conducted an additional simulation setup similar to the power simulation in Section 3, except that the mean model included four latent variables generated from the multivariate standard normal distribution. Here, we compared GLLVM2(d) with $d = 2$ and $d = 4$ latent variables and GLMM3. Results are shown in Figure B3.

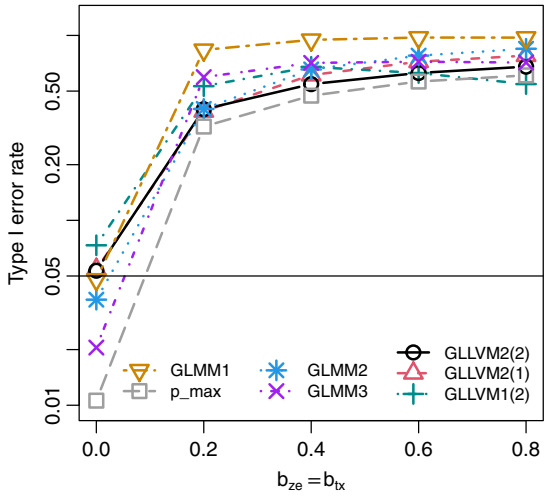

FIGURE B2 Type I error rates obtained using simulation setup described in Appendix B.1 for likelihood ratio tests based on GLMM with random intercepts (GLMM1), GLMM with random site intercepts and slopes for environmental variables (GLMM2), GLLVM with random site intercepts (GLLVM1(2)), and GLLVM with random intercepts, slopes, and $d = 1, 2$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3), and the $p_{\max}$ test. Generated datasets consisted of $n = 40$ sites and $m = 40$ species. In the mean model, we used latent environmental variables x_i and latent trait variables z_j which were correlated with the observed environmental e_i and observed trait variables t_j with correlation of 0.3. These settings lead to such confounding, in which any regression model fails and so also all fourth-corner models will fail. GLLVM, generalized linear latent variable model; GLMM, generalized linear mixed modeling

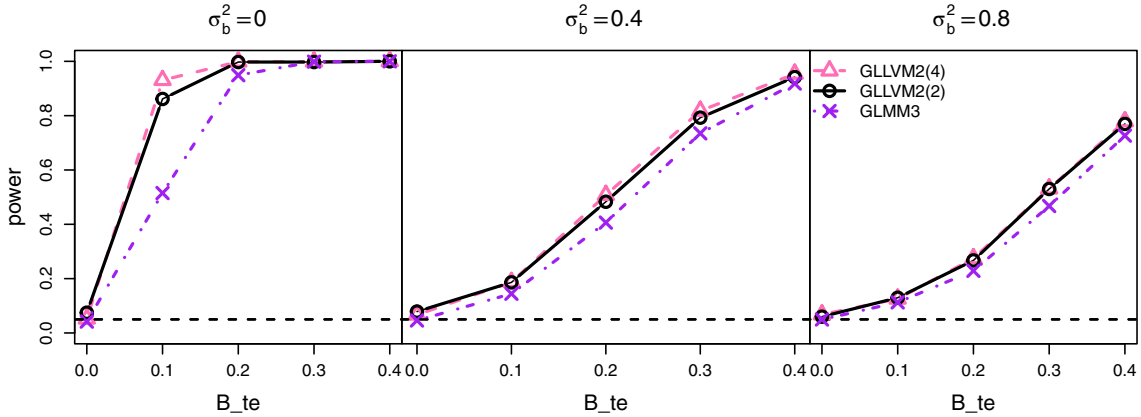

FIGURE B3 Power as a function of the effect size B_{te} for the likelihood ratio tests based on GLLVM with random intercepts, slopes and $d = 2, 4$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3). The datasets were generated based on the simulation model defined by Equation (5), with four latent variables, similarly to the first simulation setup in Section 3. GLLVM, generalized linear latent variable model; GLMM, generalized linear mixed modeling

B.3 Simulations with complex correlation structure based on quadratic traits

In this additional simulation setup, we used a more complex correlation structure for the species by defining latent variable loadings using both linear and quadratic terms for traits. As a simulation model we used

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

with one trait and one environmental variable generated independently from the standard normal distribution. Species intercepts β_{0j} were generated independently from the uniform distribution $U(-1, 1)$ and $\beta_t = 0.3$. The value for the variance of the random row effects was set to 0.3, while $\beta_e = 0.3$. The fourth-corner coefficient B_{te} was set to zero in order to assess Type I error. The species-specific dispersion parameters were all set to $\phi_j = 0.5$. The random slopes b_j were generated from a normal distribution with mean zero and variance from the range $\sigma_b^2 \in \{0, 0.4, 0.8\}$. Unobserved correlation structure between species was constructed by generating two-dimensional latent variables, $\boldsymbol{u}_i = (u_{i1}, u_{i2})'$, from the bivariate standard normal distribution. Finally, the loadings $\boldsymbol{\gamma}_j$ were generated so that $\gamma_{j1} = -2 - 0.4t_j + 0.3t_j^2 + \epsilon_{j1}$, $\epsilon_{j1} \sim N(0, 0.1)$ and $\gamma_{j2} = 0.5t_j + \epsilon_{j2}$, $\epsilon_{j1} \sim N(0, 0.7)$. We generated 1000 datasets assuming a negative binomial distribution for the response, and considered $n = 150$ units and $m = 25$ species. Here, we compared GLLVM2(d) with $d = 2$ latent variables and GLMM3. Results are presented in Figure B4. In terms of the Type I error, GLMM3 performed slightly worse than GLLVM, showing moderate inflation. This could be fixed by modifying GLMM3 to include an additional random site effect for the quadratic trait term.

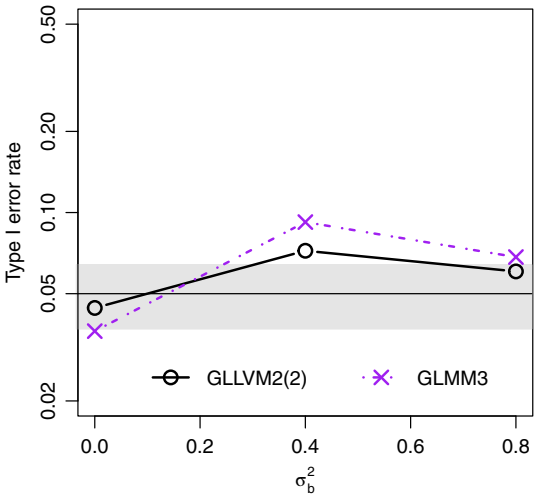

FIGURE B4 Type I error rates obtained using simulation setup described in Appendix B.3 for likelihood ratio tests based on GLLVM with random intercepts, slopes and $d = 2$ latent variables (GLLVM2(d)), GLMM with random intercepts for sites and species and random slopes for both environmental and trait variables (GLMM3). GLLVM, generalized linear latent variable model; GLMM, generalized linear mixed modeling