Received: 25 March 2024 | Accepted: 26 September 2024

DOI: 10.1111/2041-210X.14437

Methods in Ecology and Evolution

RESEARCH ARTICLE

A comparison of joint species distribution models for percent cover data

Pekka Korhonen[1] | Francis K. C. Hui[2] | Jenni Niku[3,4] | Sara Taskinen[1] | Bert van der Veen[5]

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

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

[3]Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

[4]Department of Biological and Environmental Science, University of Jyväskylä, Jyväskylä, Finland

[5]Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway

Correspondence
Pekka Korhonen
Email: pekka.o.korhonen@jyu.fi

Funding information
Australian Research Council, Grant/Award Number: DP230101908; HiTEc COST Action, Grant/Award Number: CA21163; Jenny ja Antti Wihurin Rahasto, Grant/Award Number: 00220161; Koneen Säätiö, Grant/Award Number: 201903741; Research Council of Finland, Grant/Award Number: 453691

Handling Editor: Jiangshan Lai

Abstract

1. Joint species distribution models (JSDMs) have gained considerable traction among ecologists over the past decade, due to their capacity to answer a wide range of questions at both the species- and the community-level. The family of generalised linear latent variable models in particular has proven popular for building JSDMs, being able to handle many response types including presence-absence data, biomass, overdispersed and/or zero-inflated counts.
2. We extend latent variable models to handle percent cover response variables, with vegetation, sessile invertebrate and macroalgal cover data representing the prime examples of such data arising in community ecology.
3. Sparsity is a commonly encountered challenge with percent cover data. Responses are typically recorded as percentages covered per plot, though some species may be completely absent or present, that is, have 0% or 100% cover, respectively, rendering the use of beta distribution inadequate.
4. We propose two JSDMs suitable for percent cover data, namely a hurdle beta model and an ordered beta model. We compare the two proposed approaches to a beta distribution for shifted responses, transformed presence-absence data and an ordinal model for percent cover classes. Results demonstrate the hurdle beta JSDM was generally the most accurate at retrieving the latent variables and predicting ecological percent cover data.

KEYWORDS
beta regression, community-level modelling, latent variable model, ordination, percent cover data, zero-inflation

1 | INTRODUCTION

Measurements of percent cover are typical in many ecological studies of plant communities, macroalgae, or sessile animals. By their nature, for example, limited seed dispersal, tendency for clumping and lack of self-locomotion, the notion of 'individual' may not always be meaningful or easy to determine regarding such organisms. In such cases, it is often sensible to use the percentage covered (of a given study area) by species as its measure of abundance, rather than counts or simple presence/absence.

For instance, data on percent cover at a given study site are typically collected by taking measurements on multiple plots or along line transects. These measurement can vary a lot in manner: percent cover may be determined visually by practitioners, possibly through aggregation of standardised subplots, or through pin-point methods, that is, by placing a given amount of pins randomly across the study area and recording the proportion of 'hits' for each species part of the study (Damgaard et al., 2020). Instead of representing cover as a percentage, cover data may also (classically) be represented as ordered classes, for example, using the Braun-Blanquet (1932) or Daubenmire (1959) scale. Finally, with improving technologies, automated percentage cover data collection procedures based on high-resolution images (say) are expected to increase in the future. For a comprehensive review of methods to measure vegetation cover data, we refer the reader to Damgaard and Irvine (2019). In this article, we focus on the analysis of percent cover data where species are allowed to overlap each other, that is, the sum covered by all species in a plot can exceed 100%.

For percent cover data, regression assuming a beta distribution offers a natural starting choice for modelling. For cover class data, a reasonable default would be the cumulative logit model (also known as proportional odds model, McCullagh, 1980). On the other hand, proper analysis of cover data is often hindered by high percentages of observations recorded to be zero that is, the responses are sparse. This causes issues particularly with models for continuous data, such as the beta or Dirichlet regression, which are unable to accommodate zero responses altogether. If the amount of zeros is relatively low, or if instead of being structural the zeros are the result of inadequate sampling, one can replace them with some small values or via imputation. However, when the zeros are structural, that is related to the actual underlying ecological process in question, hurdle models should instead be considered. Distinguishing between structural and sampling zeros is a hard problem overall, and, especially with percent cover data, is heavily influenced by the method of data collection. For instance, the pin-point method, although more objective in general than visual assessment, commonly underestimates or misses completely the covers of small or rare species (Bråkenhielm & Qinghong, 1995). We refer the reader to Blasco-Moreno et al. (2019) and references therein for a more detailed discussion about the differences between structural zeros and sampling zeros, noting that to properly diagnose between structural or sampling zeros, it is often crucial to consult an expert knowledgeable regarding the types of species and/or environments in question. As a related topic, one might also consider multispecies occupancy models, which extend ideas from mark-recapture methods into modelling of ecological communities, to account for imperfect detection of species (e.g. Rota et al., 2016; Tobler et al., 2019; Warton et al., 2016).

In this paper, we investigate the analysis of percent cover data in the context of joint species distribution modelling (JSDM; Hui et al., 2023; Ovaskainen & Abrego, 2020; Pollock et al., 2014; Warton et al., 2015). JSDMs are a powerful approach to analyse various types of community composition data, providing researchers with a general framework to draw inferences about, for example, co-occurrence patterns between different species, community covariation and its attribution to environmental filtering versus possible biotic interactions, and model-based ordinations on species assemblages. There is a suite of statistical software available for fitting (various flavours of) JSDMs, for example, the R-packages boral (Hui, 2016), Hmsc (Tikhonov, Opedal, et al., 2020), gllvm (Niku, Hui, et al., 2019) and VAST (Thorson, 2019). All of these adopt generalised linear latent variable models (GLLVMs; Skrondal & Rabe-Hesketh, 2004) as the basis for fitting JSDMs, using either Bayesian Markov chain Monte Carlo (MCMC) or approximate likelihood-based methods for estimation and inference. This article focuses on the latter, particularly following in the vein of the gllvm R-package (Niku, Hui, et al., 2019) which combines variational approximations (see Korhonen et al., 2023, and references therein) with automatic differentiation techniques from the TMB R-package (Kristensen et al., 2016) to facilitate computationally efficient and scalable estimation. The comparative strengths of the gllvm package include fast estimation and ease of use, with the interface and the diagnostic tools made available having been designed for practitioners who are used to working with the (generalised) linear modelling environment in base R, that is, the functions lm() and glm(). By contrast, MCMC-based estimation, such as that implemented in the boral and Hmsc packages, tends to be very slow for JSDMs and convergence of the Markov chains can be difficult to diagnose for users not familiar with Bayesian methodology. On the other hand, both Hmsc and VAST are currently more flexible in incorporating complex dependency structures in the modelling such as spatio-temporal correlations.

Although a flexible framework by design, research and readily available implementations of JSDMs/GLLVMs for percent cover data specifically have been relatively lacking. Exceptions to this are the works of Damgaard et al. (2020), who proposed a method for model-based ordination of pin-point cover data utilising a re-parameterised Dirichlet-multinomial distribution. This is appropriate for data that, instead of percentages, includes the counts of the 'hits' of the pins. As such, their model is unable to account for structural zeros. More recently, Kettunen et al. (2023) introduce a similar type of model with a more general structure, letting some subsets of species to be in direct competition for space, meaning they cannot overlap one another, while simultaneously allowing it for others. In this article, we consider JSDMs for percent cover setting with a focus on model-based ordination.

For visualising community composition data, ordination plots display observational units according to their scores on a small set of latent axes, such that units closer to each other in the ordination can be deemed to be more similar in species composition or relative abundance (van der Veen et al., 2021; Warton et al., 2015). Ordination plots have long been used by ecologists to analyse cover data, and particularly Braun-Blanquet cover class data, as seen, for example, in Islebe and Velázquez (1994), Cilliers and Bredenkamp (2000) and Härdtle et al. (2005). While

traditional ordination methods are algorithmic and based on distance measures (e.g. non-metric multidimensional scaling or NMDS; Kruskal, 1964a, 1964b), model-based approaches to ordination have surged in popularity over the past decade due their capacity to (also) incorporate environmental covariates, complex dependence structures, and species' traits and phylogeny information (e.g. Popovic et al., 2022; van der Veen et al., 2021, 2023). Compared to some earlier approaches to modelling cover data (see, e.g. Herpigny & Gosselin, 2015, and the included references), on top of offering effect and uncertainty quantification, model-based ordination is able to bridge the gap between the familiar (ordination) and more sophisticated tools (statistical models) for practising ecologists.

The remainder of this article is structured as follows. We begin by introducing GLLVMs as a method for analysing multivariate percent cover data with exact zeros, and possibly exact ones. Afterward, we briefly review existing approaches for modelling multivariate percent cover data, starting with a GLLVM assuming a beta distribution in combination with a common transformation used for exact zeros and ones (Smithson & Verkuilen, 2006). We then propose a new three-part hurdle model extension of the beta GLLVM, extending the work of Ospina and Ferrari (2012); Liu and Kong (2015) to the case of multiple correlated responses. We also review the cumulative logit GLLVM for ordinal (cover class), responses and propose a model inspired by the recent work of Kubinec (2023) on the ordered beta distribution. Note, that the methods proposed in this article are aimed towards handling structural zeros, rather than zeros resulting from insufficient sampling or detection. We perform a series of numerical comparisons between the various JSDMs using both simulated artificial cover data, and by making predictions based on real-world cover data. For the former, we are particularly interested in the different models' ability to accurately recover the latent variables under model misspecification and increasing rate of zeros. Correct retrieval of latent variable scores is an important aspect of a model's performance, due to their part in the resulting model-based ordination. The comparison also includes the popular algorithmic alternative NMDS for distance-based ordination. For predictive comparisons with real-world cover data, we split the datasets into training and test sets, and use various metrics to assess point prediction and classification performance. Special consideration is paid to the effect of the sparsity (or equivalently, recorded prevalence) of the species on predictive performance. We conclude the article with some general remarks and discussion on the results and ideas for future studies.

2 | GLLVMs FOR PERCENT COVER DATA

Cover data in ecology typically comprise records for the proportion of a plot that species, for example, plants or other sessile organisms, occupy. Denote the coverage of species j in sample i as Y_{ij} for $i = 1, \dots, n, j = 1, \dots, m$, where Y_{ij} belongs in the closed-interval $[0, 1]$. Statistical modelling of proportion data that includes zeros and/or ones is challenging in general, though a variety of models have been proposed in the literature for univariate responses, that is, single species; hurdle beta model that allow for exactly zeros *or* one as responses was introduced in Ospina and Ferrari (2012), and ordered beta model was proposed recently by Kubinec (2023). Here we propose a number of extensions for these regression models to the setting of joint species distribution modelling, using GLLVMs, for multivariate percent cover responses.

In GLLVMs, we consider regressing the mean of each response $\mu_{ij} = E(Y_{ij})$ against a vector of $d \ll m$ latent variables, $\boldsymbol{u}_i = (u_{i1}, \dots, u_{id})^\top$, along with the q-vector of covariates $\boldsymbol{x}_i = (x_{i1}, \dots, x_{iq})^\top$ as follows

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

where $g(\cdot)$ is a known link function, the vectors $\boldsymbol{\beta}_j = (\beta_{j1}, \dots, \beta_{jq})^\top$ and $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jd})^\top$ denote species-specific regression coefficients and loadings, respectively, β_{0j} denote species-specific intercepts, and α_i denote (optional) row effects. The latent variables $\boldsymbol{u}_i$ can be considered as ordination scores that capture the correlation across species after accounting for observed covariates $\boldsymbol{x}_i$. These are typically assumed to follow a d-dimensional standard normal distribution, $\boldsymbol{u}_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$.

As mentioned above, the latent variables $\boldsymbol{u}_i$ can be used for ordination, and most commonly when $d = 2$ (van der Veen et al., 2023). Ordination methods construct a low-dimensional presentation of the high-dimensional matrix of responses. As an example, Figure 1 presents a model-based unconstrained ordination based on two new forms of GLLVMs we propose for percent cover data, and fitted to a vascular plant dataset from 151 peatland sites across Finland; see the study on prediction capabilities later on for further details. The vegetation cover data is extremely sparse with a large proportion of covers recorded to be exactly zero. When the sites are coloured according to peatland type, we see clear clusters forming accordingly: this showcases the ability of the (our proposed) GLLVMs for multivariate percent cover data to account for unobserved characteristics of the data when performing model-based ordination.

Yet another important application of JSDMs to ecological data is to be able to inspect species associations. With GLLVMs, this can be done through the residual covariance matrix $\Longrightarrow = \mathcal{C}\mathcal{C}^\top$, where $\mathcal{C}$ is the $m \times d$ matrix with the loadings γ_j set as columns. Pairs of species with high (negative) covariance are strongly (negatively) associated with each other among the environments present in the study (see for instance Astarloa et al., 2019; Pollock et al., 2014). For a general review on (other) applications of GLLVMs in ecology, including applications and potential inference tools, available computational tools, and similarities to mixed models, we refer to Warton et al. (2015) and Ovaskainen and Abrego (2020). For a practically oriented introduction to using GLLVMs on ecological data, we can also recommend the original software articles relating to the popular R packages as a great starting point, for example, Hui (2016) on boral; Niku, Hui, et al. (2019) on gllvm; or Tikhonov, Opedal, et al. (2020) on Hmsc.

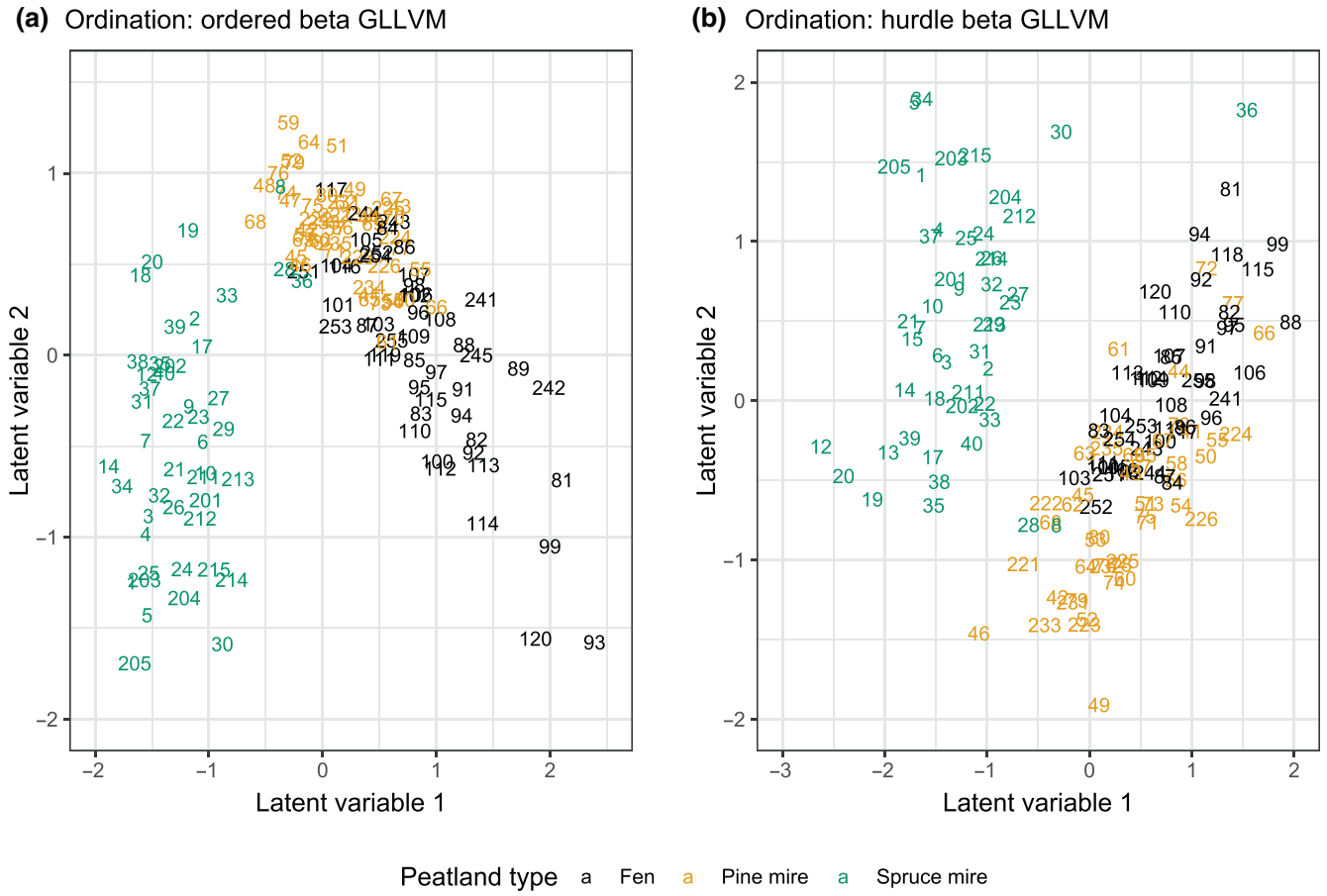

FIGURE 1 Model-based unconstrained ordination plots based on (a) ordered beta, and (b) hurdle beta generalised linear latent variable models (GLLVMs) fitted to the vascular plant cover dataset. Sites are coloured according to their peatland type. These clear clusters in the latent variable scores would dissipate if the peatland type was included in the GLLVM as a covariate.

3 | MODELS

3.1 | Beta GLLVM

A popular, starting approach for modelling ecological percent cover data is to use a beta distribution defined for a bounded continuous interval. That is, assume $Y_{ij} \in (0, 1)$, meaning it can take any value between but can not exactly be equal to zero or one. Then we assume a beta distribution $Y_{ij} \sim \text{Beta}(\mu_{ij}, \phi_j)$ with mean μ_{ij} and species-specific precision parameter $\phi_j > 0$. The probability density function of Y_{ij} is given by

$$f_{\text{beta}}(Y_{ij}; \mu_{ij}, \phi_j) = \frac{\Gamma(\phi_j)}{\Gamma(\mu_{ij}\phi_j)\Gamma(\phi_j(1-\mu_{ij}))} Y_{ij}^{\mu_{ij}\phi_j - 1}(1-Y_{ij})^{\phi_j(1-\mu_{ij})-1}. \quad (2)$$

For a GLLVM, we can relate μ_{ij} to the covariates and latent variables using Equation (1), where $g(\mu) = \log(\mu/(1-\mu))$ is most commonly set to a logit link function.

As the beta distribution cannot account for zeros (or ones), a common solution is to apply a transformation that shifts the responses slightly away from the bounds. One popular transformation come from Smithson and Verkuilen (2006), who suggested replacing Y_{ij} by $Y^*_{ij} = (Y_{ij}(N-1) + 0.5)/N$ and then modelling Y^*_{ij} using Equation (2). Although such a beta GLLVM on transformed responses is simple to fit and produces credible results when the number of recorded zeros and ones is small, an obvious drawback is that the exact zero and one responses can carry important information which may be lost in the process of the transformation. Put another way, it may be ecologically more reasonable to model the zeros and ones separately rather than through a single continuous distribution. From a statistical perspective, particularly with lots of recorded zeros/ones, the responses are pushed up against the boundary and applying beta GLLVMs on transformed responses does not actually address this issue (see O'Hara & Kotze, 2010; Warton, 2018, on the analogous issue of using log transformations for count data).

3.2 | Hurdle beta GLLVM

A less heuristic approach to account for recorded zeros and ones in percent cover data is to explicitly model their respective probabilities of arising, and doing so separately from values in between zero and one. For instance, in the broader context of modelling

semi-continuous responses where exact zeros can arise, for example, biomass, we can consider a model with two distinct parts:

$$\mathbb{P}(Y=y)=\begin{cases}\rho, & \text{if } y=0,\\ (1-\rho)\cdot p(y), & \text{if } y>0,\end{cases}$$

where ρ controls the probability of a zero occurring and $p(y)$ denotes some generic distribution that can only generate positive values. The above can be extended to more than two parts, and falls under a class of models generally called the *hurdle models* (Cragg, 1971).

As an aside, note another commonly used approach to dealing with excess amount of zeros in ecological modelling is zero-inflated regression (e.g. Martin et al., 2005; Wenger & Freeman, 2008). Zero-inflation is most often associated with models for count data and refers to a sort of mixture model consisting of the base model together with a process that generates additional zeros.

We propose a new hurdle beta GLLVM that accommodates both recorded zeros and ones in multivariate percent cover data. The proposed approach builds on the ideas of Ospina and Ferrari (2012); Liu and Kong (2015) and Kubinec (2023) and is comprised of the following three-part beta-distribution-based approach to modelling Y_{ij}:

$$P\left(Y_{ij};\mu_{ij},\mu_{ij}^0,\mu_{ij}^1,\phi_j\right)=\begin{cases}\mu_{ij}^0, & \text{if } Y_{ij}=0,\\ \left(1-\mu_{ij}^0\right)\mu_{ij}^1, & \text{if } Y_{i,j}=1,\\ \left(1-\mu_{ij}^0\right)\left(1-\mu_{ij}^1\right)\cdot f_{\text{beta}}\left(Y_{ij};\mu_{ij},\phi_j\right), & \text{if } Y_{ij}\in(0,1).\end{cases} \quad (3)$$

Analogous to Equation (1), we use a link function to connect the parameters μ_{ij}, μ_{ij}^0 and μ_{ij}^1 to covariates and latent variables. That is, for the zero- and one-parts, respectively, we use $g\left(\mu_{ij}^0\right)=\eta_{ij}^0=\beta_{0j}^0+\mathbf{x}_i^\top\beta_j^0+\mathbf{u}_i^\top\gamma_j^0$, and $g\left(\mu_{ij}^1\right)=\eta_{ij}^1=\beta_{0j}^1+\mathbf{x}_i^\top\beta_j^1+\mathbf{u}_i^\top\gamma_j^1$, and $g(\cdot)$ is set to the logit link function. The quantity $f_{\text{beta}}(Y;\mu,\phi)$ is defined as per Equation (2).

It is not hard to see the hurdle beta GLLVM works by simultaneously modelling the absences (zeros), full coverage (ones), and percent covers between this. The three linear predictors η_{ij}^0,η_{ij}^1 and η_{ij} need not contain the same set of environmental covariates $\mathbf{x}_i$ (or even the same latent variables $\mathbf{u}_i$). For example, if the data contain a moderate to high rate of zeros but only relatively few ones, then we may choose to use a simple structure for η_{ij}^1 including the intercept and latent variables only. Also, expert knowledge may inform that a certain covariate is known to have an effect on percent covers of all species present, but does not influence the actual likelihood of presence. Then it might be sensible to leave such as covariate out of η_{ij}^0 while keeping it in η_{ij}. This degree of flexibility is a key strength of the hurdle beta GLLVM. For ease of presentation though, and for focusing on the case of producing a single model-based ordination, we have set up the model such that the same $\mathbf{x}_i$'s and $\mathbf{u}_i$'s occur in all three linear predictors.

3.3 | Cumulative logit GLLVM

The cumulative logit regression or proportional odds model (McCullagh, 1980) is a popular approach for analysing ordered categorical, that is, ordinal, responses. In community ecology, it finds its use in analysing cover class datasets where, instead of cover percentages or counts of specimens, the responses instead comprise labels indicating the class each species is sorted into. Two well-known examples of vegetation cover classification systems are given by Braun-Blanquet (1932) and Daubenmire (1959).

For each species j, assume a classification scheme consisting of levels labelled from 1 to K_j, ordered from low (or zero) coverage to high (or full) coverage. Then a cumulative logit GLLVM is characterised by the following distribution for the responses:

$$P\left(Y_{ij};\eta_{ij},c_1^{(j)},\ldots,c_{K_j-1}^{(j)}\right)=\begin{cases}\rho_{ij}^1, & \text{if } Y_{ij}=1,\\ \rho_{ij}^k-\rho_{ij}^{k-1}, & \text{if } 2\le Y_{ij}\le K_j-1,\\ 1-\rho_{ij}^{K_j-1}, & \text{if } Y_{ij}=K_j.,\end{cases} \quad (4)$$

where $g\left(\rho_{ij}^k\right)=c_k^{(j)}-\eta_{ij}$ with $g(\cdot)$ set to the logit link, $\eta_{ij}=\mathbf{x}_i^\top\boldsymbol{\beta}_j+\mathbf{u}_i^\top\boldsymbol{\gamma}_j$ where the species-specific intercept is omitted for reasons of parameter identifiability, and $c_1^{(j)}<c_2^{(j)}<\cdots<c_{K_j-1}^{(j)}$ are a set of cutoff parameters specific to the species $j=1,\ldots,m$. To ensure the model can be fitted, for every species $j=1,\ldots,m$ the data must include at least one observation in each of the corresponding K_j levels. Note for interpretation, it may make more sense to use a common set of cutoffs for all species that is one set of parameters $c_1<\cdots<c_{K-1}$ and reintroduce β_{0j}, for example, if a common cover class system is used. On the other hand, while using species-common cutoffs may (also) ease fitting of the cumulative logit GLLVM, it offers less flexibility than allowing species-specific cutoffs as in Equation (4); it is our recommendation for practitioners to carefully consider the classification systems employed for each species during the data collection process before deciding which particular version of the model to adopt.

3.4 | Ordered beta GLLVM

We propose an extension of the cumulative logit GLLVM, called the ordered beta GLLVM to handle percent cover data that includes records of exact zeros and ones. The proposed approach is based on the idea of an ordered beta distribution by Kubinec (2023), which formulates a conditional cumulative logit process for responses belonging in one of the classes $\{\{0\},(0,1),\{1\}\}$ that is, zeros, between zero and one, or ones. Conditional on being in the second class, the percent cover is then represented by a beta distribution. In doing so, the ordered beta GLLVM greatly reduces the total amount of parameters to be estimated compared with the hurdle beta GLLVM in Equation (3). This may prove advantageous in situations where the latter tends to overfit or there is not enough information in the multivariate percent cover data to adequately model the probability of zeros and ones separately.

We now formulate the ordered beta GLLVM in more detail. For species $j=1,\ldots,m$, let z_{ij} denote an underlying continuous variable, and define two cutoff parameters $\zeta_{j0}<\zeta_{j1}$ such that $Y_{ij}=0$ occurs when $z_{ij}<\zeta_{j0}$, $Y_{ij}=1$ occurs when $z_{ij}>\zeta_{j1}$, and $Y_{ij}\in(0,1)$ occurs

when $\zeta_{j0} < z_{ij} < \zeta_{j1}$. Conditional on $Y_{ij} \in (0, 1)$, the response variable follows a beta distribution Y_{ij} as per Equation (2). By assuming z_{ij} follows a logistic distribution, then marginalising over z_{ij} we obtain the following distribution for the percent cover responses that characterises the ordered beta GLLVM,

$$P(Y_{ij}; \eta_{ij}, \phi_j) = \begin{cases} \rho_{ij}^0, & \text{if } Y_{ij} = 0, \\ (\rho^1 - \rho^0) \cdot f_{\text{beta}}(Y_{ij}; \mu_{ij}, \phi_j), & \text{if } Y_{ij} \in (0, 1), \\ 1 - \rho_{ij}^1, & \text{if } Y_{ij} = 1, \end{cases} \quad (5)$$

where $g\left(\rho_{ij}^k\right) = \zeta_k^{(j)} - \eta_{ij}$ and $g(\cdot)$ the logit link as in cumulative logit GLLVM, and $\eta_{ij} = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_j + \boldsymbol{u}_i^\top \boldsymbol{\gamma}_j$ The ordered beta GLLVM looks somewhat similar to the case of a cumulative logit GLLVM in Equation (4) with three classes, except the middle class is coupled with a standard beta regression model as in Equation (2). Here, only one linear predictor is needed to model all parts of the data, and connects all three possible 'states' of the response (either it is recorded as exactly zero, or between zero and one, or exactly one). One can further interpret the model, and the corresponding underlying ecological process, through the notion of censoring: imagine that the underlying continuous variable (modelled here as a function of η_{ij}) captures both the relevant ecological factors for the likelihood of presence, and for the cover of species j at site i. Then, the cutoff parameter ζ_{j0} serves as a threshold that η_{ij} needs to surpass in order for the odds to be in favour of species j populating site i. If η_{ij} exceeds ζ_{j1}, the odds favour species j to cover the whole of site i.

By connecting the likelihood of presence and the magnitude of cover through a single linear predictor η_{ij}, the ordered beta GLLVM thus arguably presents a more 'continuous' data-generating process, in contrast to the hurdle beta GLLVM which effectively separates the three types of classes $\{\{0\}, (0, 1), \{1\}\}$ into distinct parts. Put another way, the hurdle beta model reflects an assumption that the presence and the cover of a species may be driven by almost completely independent ecological processes, which may not be realistic for the data at hand. Indeed, Kubinec (2023) argues that with the hurdle beta model, it is possible for a set of environmental covariates to simultaneously have an increasing effect on both the probability of observing zeros and the probability of observing ones, which is usually not expected ecologically. On the other hand, in situations where the data *does* carry enough information about the distinct the ecological processes generating the zeros, ones or percent cover continuous responses, the hurdle beta GLLVM is expected to perform similar to or better than its ordered beta counterpart.

4 | MODEL FITTING AND VALIDATION

When working with non-continuous responses and a non-identity link function/s, GLLVMs can be estimated using approximate likelihood-based methods, with the need for approximation arising since the latent variables cannot be integrated out analytically. One class of approaches that has garnered a lot of attention recently is variational approximations for GLLVMs, which have shown to be an attractive choice over alternatives such as Laplace approximation or quadrature rules (Korhonen et al., 2023); see also the recent quasi-likelihood approach of Kidzinski et al. (2022). For the particular models studied in this article, we use the class of extended variational approximations (EVA) described in Korhonen et al. (2023), which allows efficient approximate likelihood-based fitting and inference even when tractable, closed-form expressions cannot be immediately obtained. We also coded every model in a similar manner, utilising TMB (Kristensen et al., 2016) for its fast implementation of automatic differentiation, so as to ensure any comparisons we make are not confounded by differences in estimation and inferential methods.

Turning to inference briefly, model selection with GLLVMs can be performed using information criteria, such as Akaike information criterion (AIC, e.g. Burnham & Anderson, 2002). AIC can be (also) used, for example, for determining a suitable number d for the dimension of the latent variables in the model, although other model selection approaches such as regularisation are possible here (Hui et al., 2018) With smaller total response matrix sizes nm, it is advisable to use BIC (Bayesian information criterion) instead; otherwise, over-parameterised models could get chosen almost unilaterally; see also Tredennick et al. (2021) for a general work on the topic of model selection in ecology. Finally, assessing the suitability of a GLLVM fit and associated assumptions for the multivariate percent cover data at hand can be conducted in a manner similar to that of ordinary linear models, that is, by visual inspection of residual plots. Outside of normal responses, randomised quantile or Dunn-Smyth residuals (Dunn & Smyth, 1996) are commonly used with GLLVMs. Predictive accuracy of an estimated GLLVM can be assessed by comparing new set of responses that is, hold-out data to the corresponding values predicted by the model; see Kidzinski et al. (2022) and the later section on comparing predictions for examples of this.

5 | NUMERICAL STUDY

5.1 | Simulation design

With a focus on model-based ordination, we used two simulation setups to compare the Procrustes error between the predicted and true latent variables under increasing degrees of sparsity for multivariate percent cover data. Briefly, the Procrustes error can be thought of as the mean squared error of two matrices after accounting for differences in scale and rotation (Oksanen et al., 2018). We compared latent variables obtained from three different beta-distribution-based GLLVMs (shifted, ordered, hurdle), the cumulative logit model GLLVM on ordinally classified responses, and a Bernoulli logit GLLVM on presence-absence transformed responses (i.e. the percent cover was converted to zeros and ones). We also included non-metric multidimensional scaling (NMDS) with either the Bray-Curtis or Jaccard dissimilarity measures, as an algorithmic distance-based alternative to ordination.

The first and second simulation setups used the ordered beta GLLVM and the hurdle beta GLLVM, respectively, as the true data-generating processes. In both of these cases, we simulated 1500 multivariate percent cover datasets of [0,1]-responses with $n = 180$ (units) and $m = 240$ (species). We considered the mean proportion of zero observations across the m species to be varying as $p = 10\%, 20\%, \ldots, 90\%$, while the proportion of ones was kept constant at 5%. When fitting the cumulative logit GLLVM, we assumed common cutoff parameters for all species and converted the simulated percent cover responses to class numbers in accordance with the Daubenmire system. For simplicity, both simulation setups featured no predictor variables, with the species-specific intercepts β_{0j} drawn from a uniform distribution $\mathcal{U}(-1, 1)$, the elements of the loading vectors $\boldsymbol{\gamma}_j$ drawn independently from $\mathcal{U}(-2, 2)$, and the latent variables $\mathbf{u}_i$ drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I}_2)$. When simulating from the hurdle beta GLLVM, the additional loadings $\boldsymbol{\gamma}_j^0$ and $\boldsymbol{\gamma}_j^1$ were also drawn from $\mathcal{U}(-2, 2)$, while the cutoff parameters $\zeta_k^{(j)}$ and additional intercepts $\left(\beta_{0j}^0, \beta_{0j}^1\right)$ in the true ordered and hurdle beta GLLVMs, respectively, were chosen to best fulfil the desired proportions of zeros and ones as discussed above.

5.2 | Results of numerical study

Results for the Procrustes error are shown in Figure 2, noting that NMDS with Jaccard dissimilary metric was omitted due to it performing almost identically to the Bray–Curtis dissimilarity. Overall, NMDS consistently performed the worse at recovering the true latent variables values. Of the model-based approaches, the simple beta GLLVM with transformed responses struggled the most especially when the mean proportion of zeros was quite high and the data-generating model was hurdle beta GLLVM (Figure 2b). Unsurprisingly, the performance of the Bernoulli model shows improvement as the simulated responses start to resemble presence-absence data more, that is, as the sparsity level increases. Note the adverse effects of presence-absence transformation of cover responses in terms of information loss have been studied, for example, in Van der Maarel (1979), and we would only advocate for the Bernoulli model in scenarios where the data is extremely sparse and the models based on modifications of the beta distribution run into convergence issues even with simplified model structures.

Generally, the hurdle beta GLLVM performed best across both scenarios, which was to be expected given it is more flexible than the ordered beta GLLVM and was the true model in the second simulation setup. The second setup also highlights the important differences between the ordered beta and hurdle models when it comes to the assumptions made about data generation; when the zero part of the data is generated through a process distinct from the continuous part, around the halfway mark that is, 50%–60% zeros, the ordered beta starts to deviate more and more from the 'baseline' model (hurdle), falling behind the Bernoulli and the cumulative logit models towards the end.

Finally, the cumulative logit GLLVM performed only slightly worse than its ordered and hurdle beta counterparts across the board in the first setup (Figure 2a) but was noticeably worse than these two models in the second scenario when the data was dense, that is, the mean proportion of zero responses was low (Figure 2b). Such a result however can be attributed more to the behaviour of the ordered and hurdle beta GLLVMs themselves: when the multivariate percent cover data was relatively dense, these two models outperform the cumulative logit GLLVM since converting such data to cover classes results in a non-negligible loss of information in the responses. However as the data became more sparse (and discrete) the performance of the ordered and hurdle beta GLLVMs starts to deteriorate while cumulative logit GLLVM continues to perform similarly (since in this case converting to cover classes does not lose as much information).

6 | APPLICATION AND COMPARISON OF PREDICTIVE PERFORMANCES

6.1 | Data and setup

We compared the out-of-sample predictive capabilities of the three beta-distribution-based GLLVMs on four real multivariate percent cover data sets, utilising hold-out set splits. The first two datsets originate from the Santa Barbara Coastal Long Term Ecological Research (SBC LTER; Reed & Miller, 2023) site, where cover percentages of more than 150 sessile invertebrates and macroalgae taxa were recorded between years 2000 and 2020, along 40 m × 2 m permanent transects at 11 kelp forest sites across coastal Southern California, including two sites situated on islands. The number of transects differed among sites, varying between two and eight with a total of 44, while the two datasets were defined separately for algae and invertebrates. About half of the taxa corresponding to extremely rare species were removed prior to analysis. When fitting the various GLLVMs, we set the latent variables to be at the transect level, leading to a total of $n = 44$ two-dimensional latent variables scores to estimate. Each model was fitted to data recorded from 2000 to 2017, with years 2018–2020 held out to assess predictive performance. The two datasets also included two key environmental covariates: rockiness of the seabed, and the number of stripes of giant kelp. These were used together with year as covariates in $\boldsymbol{x}_i$.

The second pair of datasets come from a Finnish longitudinal study on effects of peatland restoration, comprising over 250 species of vascular plants and mosses collected on 151 sites across Finland between years 2006 and 2022 (Elo et al., 2016). Within each site, the percent cover of each species was measured on 10 placed plots of size 1m^2. Utilising a balanced study design, the sites varied evenly on a number of key environmental factors: type (fen, pine mire and spruce mire), treatment level (drained, pristine and restored), and productivity level (high vs. low). Similar to the two SBC LTER datasets, the two datasets were defined separately for

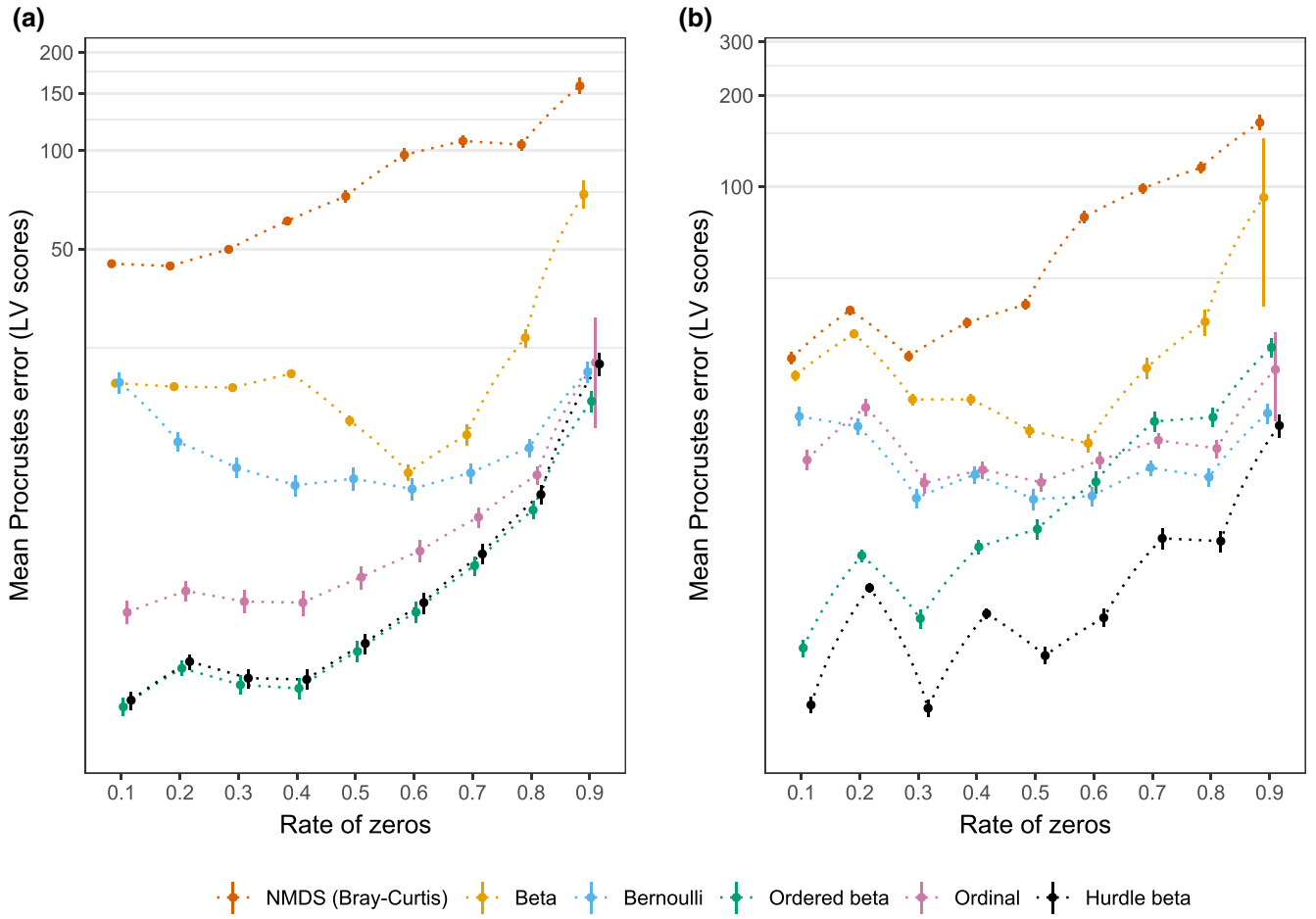

FIGURE 2 Means and standard deviations of the Procrustes errors between the predicted and the true latent variable scores, where multivariate percent cover data were generated using the: (a) ordered beta generalised linear latent variable model (GLLVM), and (b) hurdle beta GLLVM. A trimming factor of 5% was used to remove effects of the most extreme values resulting from extremely volatile fits. The points are slightly jittered to avoid visual overlap.

vascular plant and moss species, and in both we subset to only species for which there existed at least one observation per each environmental factor level. The latent variables were estimated at the site level, resulting in a total of $n = 151$ two-dimensional ordination scores to estimate. We estimated GLLVMs using data from 2006 to 2021 and held out data from the final year (2022) to assess predictive performance. The same datasets were analysed recently in Elo et al. (2024), where the effect of peatland restoration on individual species occurrences was modelled in a Bayesian setting using an alternative hierarchical JSDM but assuming a hurdle normal model on log-transformed and normalised cover percentages. Our two proposed models, which directly target percent cover responses without need for data transformation and standardisation that can potentially adversely affect the mean–variance relationship, offers a valuable addition to this application.

We assessed performances of the fitted GLLVMs using four metrics: mean absolute error of prediction (MAEP) and root mean square error of prediction (RMSE), which were calculated for the beta, hurdle beta, and ordered GLLVMs and for each species individually, and area under the receiver operating characteristic curve (AUC) and Tjur's pseudo-R^2 (Tjur, 2009) for classifying presences and absences. The latter two measures assess capacity to discriminate between zeros and ones and were calculated for binary, ordinal, ordered beta and hurdle beta GLLVMs. With the ordinal GLLVM, we assumed species-common cutoff parameters and used seven classes with the following bounds: $\{0\}, (0,0.02], (0.02,0.05], (0.05,0.25], (0.25,0.5], (0.5,0.7]$ and $(0.7,1]$. We note the selection of classification scale is expected to have an impact on performance of the ordinal model, and our choice here is based on exploratory data analysis. Future research could examine how sensitive prediction, interpretation, and statistical inference for ordinal models are in general to the scale of classification.

As all of the data sets either entirely lacked, or had very small amount, of records exactly equal to one (full cover of a measurement area), then for the hurdle beta GLLVM in Equation (3) we only included the part for modelling zeros and values strictly between zero and one. Furthermore, for the ordered beta GLLVM in Equation (5), we used species-common upper cutoff parameters.

6.2 | Results of prediction comparisons

Figures 3 and 4 plot the MAEP and RMSE, respectively, as a function of total species prevalence, p, across each of the four datasets. Across both figures, the beta GLLVM tended to perform worse than the two zero-accommodating approaches when predicting for the rarer species, which is to be expected given the capacity of the hurdle and ordered betas GLLVMs to systemically handle the recorded zeros. The differences between these three models diminished when predicting for the more commonly occurring species. Note the actual metrics are small in magnitude for the rarer species overall; this is generally not surprising given the more discrete nature responses in such cases. Between the hurdle and ordered beta GLLVMs, there were no noticeable differences in performance.

Figure 5 presents the values of AUC and Tjur R^2 across the four datasets when plotted against the recorded group mean prevalence. For each dataset, the recorded group mean prevalence was obtained by computing the proportion of non-zero observations in the complete dataset for each species, clustering species based on these recorded prevalences into a small number of groups, and then calculating the mean prevalence of each group. The values of AUC and Tjur R^2 were then calculated correspondingly for each group. Note since both metrics are determined based on estimated probabilities of presence, then only models capable of producing these estimates were included for comparison, that is, the beta GLLVM is omitted. Overall, the hurdle beta GLLVM model followed the Bernoulli logit model closely in its ability to best discriminate between zero and non-zero responses across all levels of prevalence. The ordered beta GLLVM performed similarly to the hurdle and Bernoulli GLLVMs in terms of AUC, although its performance was closer to the ordinal/cumulative logit GLLVM, which performed worst overall, when it came to Tjur R^2.

7 | DISCUSSION

In this paper, we compared different joint species distribution modelling approaches for multivariate percent cover data. Such data are encountered, for example, in studies where percent cover of several plants or other sessile organisms is measured across multiple sites, as per the four real datasets we investigated. The typical ecological questions arising from such studies include whether sites exhibit similar species composition, how environmental factors influence community composition, and whether we can predict vegetation or macroalgal cover at new sites and/or over time. This article focused on comparing methods in terms of community composition analysis

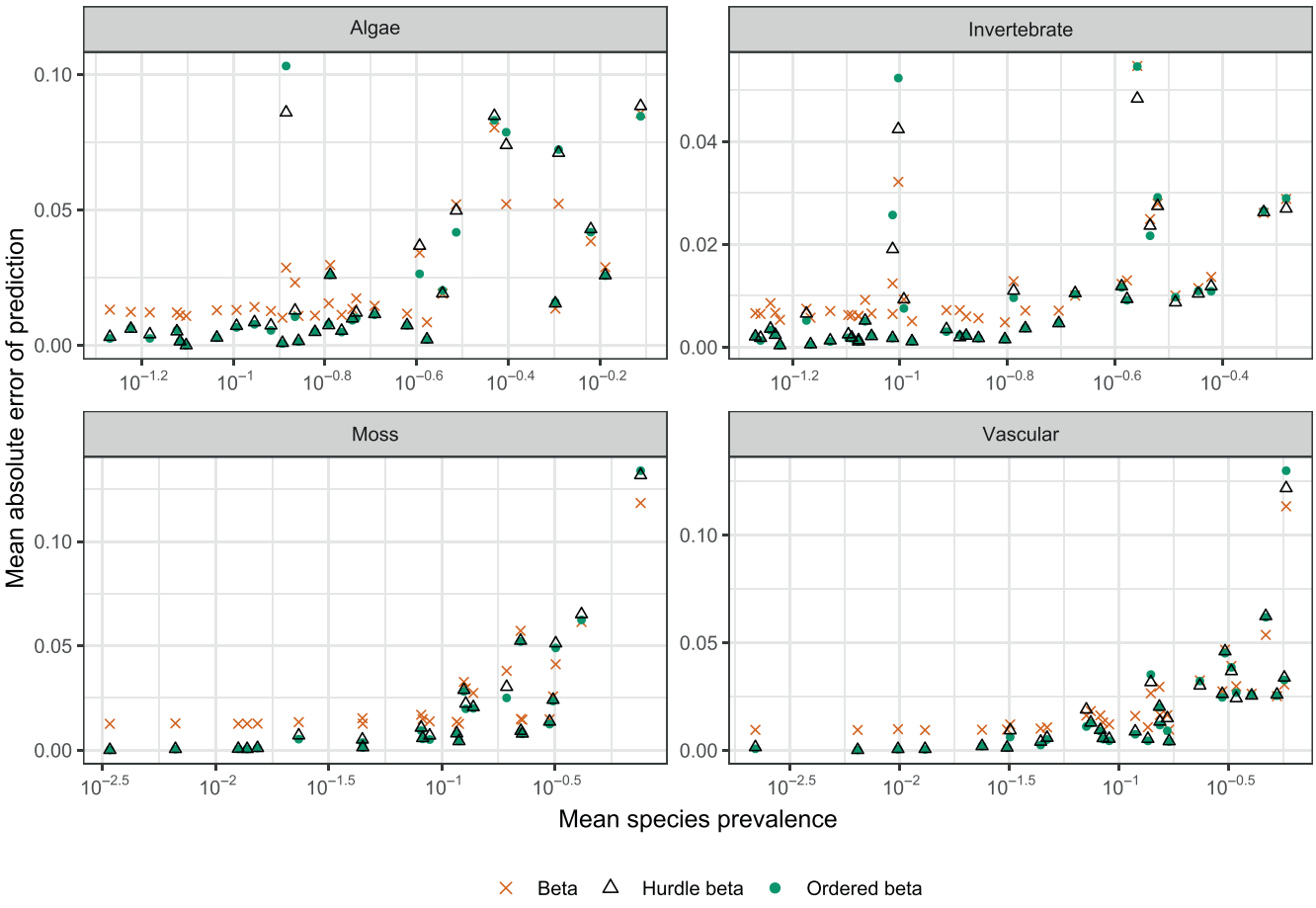

FIGURE 3 Mean absolute error of prediction as a function of mean species prevalence for beta, hurdle beta, and ordered beta generalised linear latent variable models, across the four real multivariate percent cover datasets.

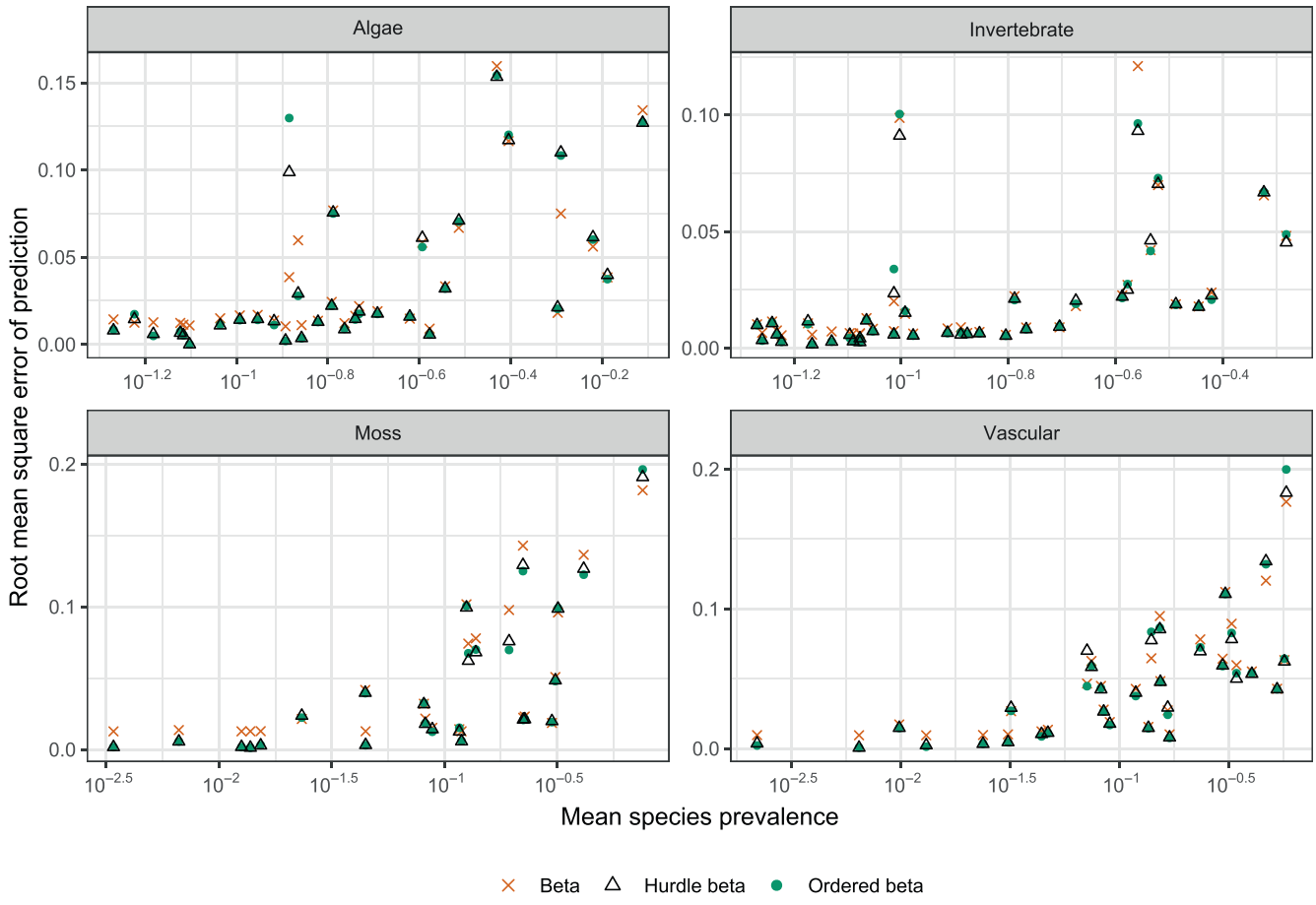

FIGURE 4 Root mean square error of prediction as a function of mean species prevalence for beta, hurdle beta, and ordered beta generalised linear latent variable models, across the four real multivariate percent cover datasets.

through model-based ordination, and in making out-of-sample predictions. We explored and extended several of GLLVMs to handle percent cover data with exact zeros and ones, most notably a three-part hurdle beta GLLVM, a cumulative logit GLLVM for cover class responses, and an ordered beta GLLVM, and discussed potential differences in the underlying ecological interpretations and assumptions of the proposed beta-based models. Regression models based on beta distribution are known to be able to accommodate a wider variety of shapes and scales for values in $(0, 1)$ than a typical normal model (Cribari-Neto & Zeileis, 2010), and this added flexibility could help in providing new and better insights about ecosystem management in studies such as Elo et al. (2024). Furthermore, the proposed models were presented in a way that facilitates adoption into other approaches to joint species distribution frameworks and software, for example, sparse JSDMs (Pichler & Hartig, 2021), copula models (Popovic et al., 2022), or community-level basis function models (Hui et al., 2023). Many of these recent approaches place emphasis on efficient computation, for example, the sJSDM package (Pichler & Hartig, 2021) utilises PyTorch (Paszke et al., 2019) to facilitate GPU-based estimation (optional) of JSDMs, which in the seminal paper proved to be even faster than gllvm. However it is important to highlight that this is rapidly moving landscape, for example, gllvm now has parallel computations enabled for many parts (van der Veen & O'Hara, 2024), while Hmsc has also recently been upgraded to utilise GPU-accelerated sampling (Rahman et al., 2024). Extensive and up-to-date comparisons of the various available approaches to JSDMs (e.g. along the lines of Norberg et al., 2019; Wilkinson et al., 2019), and how they fare in analysing all kinds of ecological abundance response including percent cover data, is a fertile area of future investigation.

As ecological percent cover data are often sparse with a large number of recorded zeros along with potentially a number of full coverages (recorded ones), it is of primary interest to study how methods performed when observed species prevalence decreased. Failing to account for excess zeros, whether due to insufficient sampling design or the ecological process itself, in the analysis of abundance data can lead to biased estimates and erroneous conclusions; see Blasco-Moreno et al. (2019) among others regarding this issue in the context of count data modelling. When comparing ordinations using simulation studies, we found the hurdle beta GLLVM exhibited the best overall performance and were fairly robust under increasing rate of zeros, while the classical beta GLLVM on transformed responses performed poorly. Both the hurdle beta and the ordered beta GLLVMs performed relatively well when it came to prediction and classification, and better than the cumulative logit and the classical beta GLLVMs.

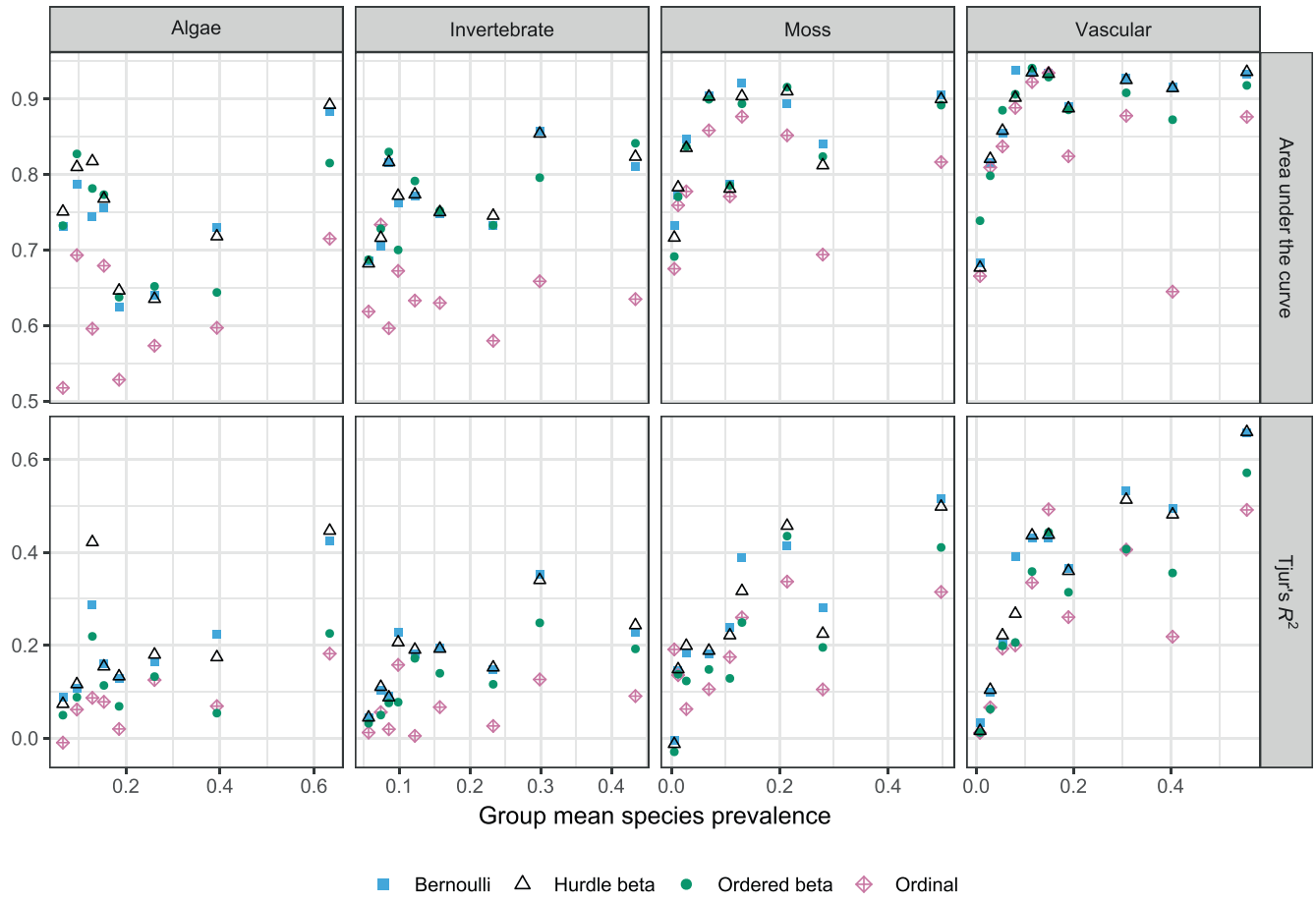

FIGURE 5 Area under the curve (top row) and Tjur's R^2 (bottom row) as a function of recorded group mean prevalence for the four real multivariate percent cover datasets. Recorded group mean prevalence was obtained by clustering species based on their recorded prevalences in the complete dataset into a small number of groups, and then calculating the mean prevalence of each group. The *y*-axes presents the corresponding metric for each group.

Even though the four datasets used to compare predictions were quite different in terms of type, size or sampling design, further studies into the robustness and performance of the proposed JSDMs under varied sampling schemes could be warranted, akin to, for example, work done in Zhang et al. (2018) for presence/absence and biomass data. Note that the issues regarding robustness, sampling irregularities and rare species are shared by pretty much all existing families of (joint) species distribution models (Norberg et al., 2019), and as such present an important topic for active research. Similarly, the effects of trying to ameliorate scenarios of extreme data sparsity by adding shrinkage penalties to the models or, for Bayesian JSDM approaches, regularising through weakly informative priors, could also be worth exploring in detail. For recent work along these lines, see Scharf and Nestler (2019) on elastic net regularised factor analysis, Kidzinski et al. (2022) on penalised quasi-likelihood estimation of GLLVMs, Chung et al. (2015) on imposing a weakly informative prior on covariance parameters of hierarchical model. Alternative model formulations could also be investigated, for example, the *k-ZIG* model proposed in Ghosh et al. (2012), which essentially applies the usual zero-inflated Poisson or negative binomial construction onto itself recursively *k* times in order to better suit scenarios with extreme excess of zeros. Link functions based on the generalised extreme value distribution, such as the special case of complementary log–log link, have been argued to fare better when analysing presence-absence data with highly disproportionate classes (Wang & Dey, 2010). Finally, whenever the study design allows, strategies leveraging auxiliary information in the manner described in Clark et al. (2014) could prove fruitful for tackling sparsity also in the percent cover data case.

Another area of future research concerns extending the GLLVM framework to other types of cover responses, such as data collected using the pin-point method or for compositional data increasingly seen in community ecology. For such data, there exists several competing frameworks, for example, classic log-ratio analysis (Aitchison, 1982), regression models based on Dirichlet or Dirichlet-multinomial distributions (Damgaard et al., 2020; Kettunen et al., 2023), and distributions directly on the composition itself (e.g. Scealy & Wood, 2023). With most of these approaches, incorporating structural zeros comes with challenges; the standard log-ratio transformations of Aitchison geometry are not defined for zero observations, and similarly the standard

Dirichlet or Dirichlet-multinomial distributions are also incapable of handling structural zeros, while implementing, for example, (multinomial-)Dirichlet hurdle model in the fashion of Equation (3) in likelihood-based setting requires may not be straightforward (see also Tang & Chen, 2018, for a Bayesian treatment of such models). Meanwhile, the pin-point method is known to be prone to miss or underestimate cover of small or rare species (Bråkenhielm & Qinghong, 1995), perhaps calling for the need to consider (multispecies) occupancy models, which utilise a sub-model and repeated visits to observation sites to control for insufficient detection (e.g. Tobler et al., 2019).

Extending the ordered beta and hurdle beta GLLVMs to handle spatially or spatio-temporally dependent latent variables could also prove fruitful; presumably, ecosystems closer to each other geographically are expected to also be more alike in the community structures. Similarly, many multivariate percent cover datasets originate from longitudinal studies, including both the SBC LTER (Reed & Miller, 2023) and the Finnish peatland (Elo et al., 2016) datasets considered in this article. With GLLVMs, spatial and/or temporal extensions could be employed by considering a more general assumption for the latent variables, for example, the covariance of the latent variables across different observational units is characterised by an autoregressive structure or coming from a Matérn covariance function. Estimating parameters for these types of more general covariance structures is typically very challenging in high-dimensional settings, and such GLLVMs would most likely require employing additional approximation techniques, for example, nearest neighbour Gaussian processes (e.g. Tikhonov, Duan, et al., 2020) or some form of a basis-function/fixed-rank kriging approach (e.g. see the recent work of Hui et al., 2023). Figuring out how to best combine such techniques with efficient GLLVM fitting algorithms, for example, variational approximations (Niku, Brooks, et al., 2019), presents an important avenue for future research.

AUTHOR CONTRIBUTIONS

All authors conceived the idea for the manuscript, Pekka Korhonen and Jenni Niku implemented the methodology, and performed all performance studies. Pekka Korhonen led the writing of the manuscript, while Francis K. C. Hui, Jenni Niku, Sara Taskinen, and Bert van der Veen contributed to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

PK was funded by the Wihuri Foundation (00220161), and PK, JN and ST were funded by the Kone Foundation (201903741). ST was funded by the Research Council of Finland (453691) and the HiTEc COST Action (CA21163). FKCH was funded by an Australian Research Council Discovery Project (DP230101908). We thank Merja Elo at University of Finland and Santtu Kareksela at Metsähallitus Parks & Wildlife Finland for providing us the Finnish peatland dataset. We also acknowledge CSC—IT Center for Science, Finland, for computational resources.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14437.

DATA AVAILABILITY STATEMENT

The kelp forest community data (Reed & Miller, 2023) are available in SBC LTER database https://doi.org/10.6073/pasta/0af1a5b0d9dde5b4e5915c0012ccf99c. The vascular plant and moss community data from Finnish peatlands (Elo et al., 2016, 2024) are available in Zenodo repository https://zenodo.org/records/10906943. The Bernoulli logit and beta GLLVMs models are available as part of the R-package gllvm (Niku, Hui, et al., 2019), while implementations for the hurdle beta, ordered beta and cumulative logit GLLVMs are available as part of the development version of the package at https://github.com/JenniNiku/gllvm, or at https://doi.org/10.5281/zenodo.13880825 (Niku et al., 2024). The gllvm GitHub repository also includes a vignette instructing how to apply the hurdle beta model on a part of the SBC LTER dataset. NMDS is implemented in the vegan R-package (Oksanen et al., 2018).

ORCID

Pekka Korhonen https://orcid.org/0000-0003-2650-3645
Francis K. C. Hui https://orcid.org/0000-0003-0765-3533
Jenni Niku https://orcid.org/0000-0002-7992-2598
Sara Taskinen https://orcid.org/0000-0001-9470-7258
Bert van der Veen https://orcid.org/0000-0003-2263-3880

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *44*, 139–160.

Astarloa, A., Louzao, M., Boyra, G., Martinez, U., Rubio, A., Irigoien, X., Hui, F. K. C., & Chust, G. (2019). Identifying main interactions in marine predator–prey networks of the Bay of Biscay. *ICES Journal of Marine Science*, *76*, 2247–2259.

Blasco-Moreno, A., Pérez-Casany, M., Puig, P., Morante, M., & Castells, E. (2019). What does a zero mean? Understanding false, random and structural zeros in ecology. *Methods in Ecology and Evolution*, *10*, 949–959.

Bråkenhielm, S., & Qinghong, L. (1995). Comparison of field methods in vegetation monitoring. *Water, Air, and Soil Pollution*, *79*, 75–87.

Braun-Blanquet, J. (1932). *Plant sociology: The study of plant communities*. McGraw-Hill.

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

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. *Journal of Educational and Behavioral Statistics*, *40*, 136–157.

Cilliers, S., & Bredenkamp, G. (2000). Vegetation of road verges on an urbanisation gradient in potchefstroom, South Africa. *Landscape and Urban Planning*, *46*, 217–239.

Clark, J. S., Gelfand, A. E., Woodall, C. W., & Zhu, K. (2014). More than the sum of the parts: Forest climate response from joint species distribution models. *Ecological Applications*, *24*, 990–999.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, *39*, 829–844.

Cribari-Neto, F., & Zeileis, A. (2010). Beta regression in r. *Journal of Statistical Software*, *34*, 1–24.

Damgaard, C., Hansen, R. R., & Hui, F. K. C. (2020). Model-based ordination of pin-point cover data: Effect of management on dry heathland. *Ecological Informatics*, *60*, 101155.

Damgaard, C. F., & Irvine, K. M. (2019). Using the beta distribution to analyse plant cover data. *Journal of Ecology*, *107*, 2747–2759.

Daubenmire, R. F. (1959). A canopy-coverage method of vegetational analysis. *Northwest Science*, *33*, 43–64.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*, 236–244.

Elo, M., Kareksela, S., Haapalehto, T., Vuori, H., Aapala, K., & Kotiaho, J. S. (2016). The mechanistic basis of changes in community assembly in relation to anthropogenic disturbance and productivity. *Ecosphere*, *7*, e01310.

Elo, M., Kareksela, S., Ovaskainen, O., Abrego, N., Niku, J., Taskinen, S., Aapala, K., & Kotiaho, J. S. (2024). A large-scale and long-term experiment to identify effectiveness of ecosystem restoration. Preprint available on *bioRxiv* at https://www.biorxiv.org/content/early/2024/04/03/2024.04.02.587693

Ghosh, S., Gelfand, A. E., Zhu, K., & Clark, J. S. (2012). The k-zig: Flexible modeling for zero-inflated counts. *Biometrics*, *68*, 878–885.

Härdtle, W., Oheimb, G. V., & Westphal, C. (2005). Relationships between the vegetation and soil conditions in beech and beech-oak forests of northern Germany. *Plant Ecology*, *177*, 113–124.

Herpigny, B., & Gosselin, F. (2015). Analyzing plant cover class data quantitatively: Customized zero-inflated cumulative beta distributions show promising results. *Ecological Informatics*, *26*, 18–26.

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., Foster, S. D., & Haak, C. R. (2023). Spatiotemporal joint species distribution modelling: A basis function approach. *Methods in Ecology and Evolution*, *14*, 2150–2164.

Islebe, G., & Velázquez, A. (1994). Affinity among mountain ranges in megamexico: A phytogeographical scenario. *Vegetatio*, *115*, 1–9.

Kettunen, J., Mehtätalo, L., Tuittila, E.-S., Korrensalo, A., & Vanhatalo, J. (2023). Joint species distribution modeling with competition for space. *Environmetrics*, *35*, e2830.

Kidzinski, L., Hui, F. K. C., Warton, D. I., & Hastie, T. J. (2022). Generalized matrix factorization: Efficient algorithms for fitting generalized linear latent variable models to large data arrays. *Journal of Machine Learning Research*, *23*, 1–29.

Korhonen, P., Hui, F. C., Niku, J., & Taskinen, S. (2023). Fast and universal estimation of latent variable models using extended variational approximations. *Statistics and Computing*, *33*, 26.

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

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*, 115–129.

Kubinec, R. (2023). Ordered beta regression: A parsimonious, well-fitting model for continuous data with lower and upper bounds. *Political Analysis*, *31*, 519–536.

Liu, F., & Kong, Y. (2015). Zoib: An R package for Bayesian inference for beta regression and zero/one inflated beta regression. *The R Journal*, *7*, 34.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*, 1235–1246.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)*, *42*, 109–142.

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.

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., van der Veen, B., Warton, D., Korhonen, P., Hui, F. K. C., Taskinen, S., & Brooks, W. (2024). gllvm (1.4.8). Zenodo. https://doi.org/10.5281/zenodo.13880825

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., … Ovaskainen, O. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs*, *89*, e01370.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*, 118–122.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., & Wagner, H. (2018). *vegan: Community ecology package*. R package version 2.5-2.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*, 1609–1623.

Ovaskainen, O., & Abrego, N. (2020). *Joint species distribution modelling: With applications in R*. Cambridge University Press.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. *32*, pp. 8024–8035). Curran Associates, Inc.

Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, *12*, 2159–2173.

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.

Popovic, G. C., Hui, F. K. C., & Warton, D. I. (2022). Fast model-based ordination with copulas. *Methods in Ecology and Evolution*, *13*, 194–202.

Rahman, A. U., Tikhonov, G., Oksanen, J., Rossi, T., & Ovaskainen, O. (2024). Accelerating joint species distribution modeling with Hmsc-HPC: A 1000× faster GPU deployment. *bioRxiv*. https://doi.org/10.1101/2024.02.13.580046

Reed, D. C., & Miller, R. J. (2023). *SBC LTER: Reef: Kelp forest community dynamics: Cover of sessile organisms, uniform point contact*. LTER Network Member Node. https://doi.org/10.6073/pasta/0af1a5b0d9dde5b4e5915c0012ccf99c

Rota, C. T., Ferreira, M. A. R., Kays, R. W., Forrester, T. D., Kalies, E. L., McShea, W. J., Parsons, A. W., & Millspaugh, J. J. (2016). A

multispecies occupancy model for two or more interacting species. *Methods in Ecology and Evolution*, *7*, 1164–1173.
Scealy, J. L., & Wood, A. T. A. (2023). Score matching for compositional distributions. *Journal of the American Statistical Association*, *118*, 1811–1823.
Scharf, F., & Nestler, S. (2019). Should regularization replace simple structure rotation in exploratory factor analysis? *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 576–590.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. CRC Press.
Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*, 54–71.
Tang, Z.-Z., & Chen, G. (2018). Zero-inflated generalized Dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, *20*, 698–713.
Thorson, J. T. (2019). Guidance for decisions using the vector autoregressive spatio-temporal (VAST) package in stock, ecosystem, habitat and climate assessments. *Fisheries Research*, *210*, 143–161.
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*, e02929.
Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M. J., Oksanen, J., & Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in Ecology and Evolution*, *11*, 442–447.
Tjur, T. (2009). Coefficients of determination in logistic regression models—A new proposal: The coefficient of discrimination. *The American Statistician*, *63*, 366–372.
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.
Tredennick, A. T., Hooker, G., Ellner, S. P., & Adler, P. B. (2021). A practical guide to selecting models for exploration, inference, and prediction in ecology. *Ecology*, *102*, e03336.
Van der Maarel, E. (1979). Transformation of cover-abundance values in phytosociology and its effects on community similarity. *Vegetatio*, *39*, 97–114.
van der Veen, B., Hui, F. K. C., Hovstad, K. A., & O'Hara, R. B. (2023). Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling. *Methods in Ecology and Evolution*, *14*, 683–695.
van der Veen, B., Hui, F. K. C., Hovstad, K. A., Solbu, E. B., & O'Hara, R. B. (2021). Model-based ordination for species with unequal niche widths. *Methods in Ecology and Evolution*, *12*, 1288–1300.
van der Veen, B., & O'Hara, R. B. (2024). Fast fitting of phylogenetic mixed effects models. *arXiv*, 2408.05333. https://doi.org/10.48550/arXiv.2408.05333
Wang, X., & Dey, D. K. (2010). Generalized extreme value regression for binary response data: An application to B2B electronic payments system adoption. *The Annals of Applied Statistics*, *4*, 2000–2023.
Warton, D. I. (2018). Why you cannot transform your way out of trouble for small counts. *Biometrics*, *74*, 362–368.
Warton, D. I., Blanchet, F. G., O'Hara, R., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2016). Extending joint models in community ecology: A response to Beissinger et al. *Trends in Ecology & Evolution*, *31*, 737–738.
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.
Wenger, S. J., & Freeman, M. C. (2008). Estimating species occurrence, abundance, and detection probability using zero-inflated distributions. *Ecology*, *89*, 2953–2959.
Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., & McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*, *10*, 198–211.
Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2018). Comparing the prediction of joint species distribution models with respect to characteristics of sampling data. *Ecography*, *41*, 1876–1887.

How to cite this article: Korhonen, P., Hui, F. K. C., Niku, J., Taskinen, S., & van der Veen, B. (2024). A comparison of joint species distribution models for percent cover data. *Methods in Ecology and Evolution*, *00*, 1–14. https://doi.org/10.1111/2041-210X.14437