Bayesian Analysis (2017) **12**, Number 4, pp. 939–967

Joint Species Distribution Modeling: Dimension Reduction Using Dirichlet Processes

Daniel Taylor-Rodríguez*, Kimberly Kaufeld†, Erin M. Schliep‡, James S. Clark§, and Alan E. Gelfand¶

Abstract. Species distribution models are used to evaluate the variables that affect the distribution and abundance of species and to predict biodiversity. Historically, such models have been fitted to each species independently. While independent models can provide useful information regarding distribution and abundance, they ignore the fact that, after accounting for environmental covariates, residual interspecies dependence persists. With stacking of individual models, misleading behaviors, may arise. In particular, individual models often imply too many species per location.

Recently developed joint species distribution models have application to presence–absence, continuous or discrete abundance, abundance with large numbers of zeros, and discrete, ordinal, and compositional data. Here, we deal with the challenge of joint modeling for a large number of species. To appreciate the challenge in the simplest way, with just presence/absence (binary) response and say, S species, we have an S-way contingency table with 2^S cell probabilities. Even if S is as small as 100 this is an enormous table, infeasible to work with without some structure to reduce dimension.

We develop a computationally feasible approach to accommodate a large number of species (say order 10^3) that allows us to: 1) assess the dependence structure across species; 2) identify clusters of species that have similar dependence patterns; and 3) jointly predict species distributions. To do so, we build hierarchical models capturing dependence between species at the first or "data" stage rather than at a second or "mean" stage. We employ the Dirichlet process for clustering in a novel way to reduce dimension in the joint covariance structure. This last step makes computation tractable.

We use Forest Inventory Analysis (FIA) data in the eastern region of the United States to demonstrate our method. It consists of presence–absence measurements for 112 tree species, observed east of the Mississippi. As a proof of concept for our dimension reduction approach, we also include simulations using continuous and binary data.

*Postdoctoral Associate, SAMSI/Duke University, Research Triangle Park, NC 27709, taylor-rodriguez@samsi.info

†Postdoctoral Researcher, SAMSI/North Carolina State University, Research Triangle Park, NC 27709, kkaufeld@samsi.info

‡Assistant Professor, Department of Statistics, University of Missouri, Columbia, MO 65211, schliepe@missouri.edu

§Professor, Nicholas School of the Environment, Department of Statistical Science, Duke University, Durham, NC 27708, jimclark@duke.edu

¶Professor, Department of Statistical Science, Duke University, Durham, NC 27708, alan@stat.duke.edu

 DOI: 10.1214/16-BA1031

Keywords: abundance, hierarchical model, latent variables, Markov chain Monte Carlo, presence–absence.

1 Introduction

Understanding the processes that determine the distribution and abundance of species is a goal of ecological research. Species distribution models are used to evaluate the variables that affect the distribution and abundance of species and to predict biodiversity, including responses to climate change (Midgley et al., 2002; Guisan and Thuiller, 2005; Gelfand et al., 2006; Iverson et al., 2008; Botkin et al., 2007; Fischlin et al., 2007; McMahon et al., 2011; Thuiller et al., 2011). These models are used to infer a species range either in geographic space or in climate space (Midgley et al., 2002), to identify and manage conservation areas (Austin and Meyers, 1996), and to provide evidence of competition among species (Leathwick, 2002). The core objective is interpolation, to predict species response at plots that have not been sampled.

Species distribution models (SDMs) are most commonly fitted to presence/absence data (binary) or abundance data (counts, ordinal classifications, or proportions). Occasionally, continuous responses are used such as biomass (Dormann et al., 2012). Prediction of species over space can be accommodated using a spatially explicit specification (Gelfand et al., 2005, 2006; Latimer et al., 2006).

Within a Bayesian framework, SDMs can be fitted using hierarchical models. Hierarchical models provide a flexible way to include information regarding the distribution of a species as well as the uncertainty (Gelfand et al., 2006). The stages in a hierarchical model can describe latent processes that are ecologically important. These models enable separation of the measurement and biological process models (MacKenzie and Royle, 2005; Latimer et al., 2006; Gelfand et al., 2005).

Customarily, SDMs are fitted independently across a collection of species (Thuiller, 2003; Latimer et al., 2006; Elith and Leathwick, 2009; Chakraborty et al., 2011). To make predictions at the community scale, independent models for individual species are aggregated or stacked (Calabrese et al., 2014). However, collectively, the independent models tend to predict too many species per location (Guisan and Rahbek, 2011), as well as other misleading results (see Clark et al., 2014, for some examples). At least one source of the problem is the omission of the residual dependence between species. Modeling species individually does not allow underlying joint relationships to be exploited (Clark et al., 2011; Ovaskainen and Soininen, 2011).

Joint species distribution models (JSDMs) that incorporate species dependence now include applications to presence–absence (Pollock et al., 2014; Ovaskainen et al., 2010; Ovaskainen and Soininen, 2011), continuous or discrete abundance (Latimer et al., 2009; Thorson et al., 2015), abundance with large number of zeros (Clark et al., 2014) and, recently, discrete, ordinal, and compositional data (Clark et al., 2016). JSDMs jointly characterize the presence and/or abundance of multiple species at a set of locations, partitioning the drivers into two components, one associated with environmental suitability, the other accounting for species dependence through the *residuals*, i.e., adjusted for the environment.

Joint distributions are meaningful at all scales, but interpretation should properly reflect scale. Clark et al. (2014) discuss why the covariance matrix does not quantify 'species interactions'. For example, competition is a mutually negative interaction. However, the strongest competitors will typically have a positive correlation – tendency to co-occur is the cause, not the consequence, of competition. Likewise, predation, disease, and parasitism are not quantified by the covariance matrix. Both are asymmetric, whereas, a covariance matrix cannot be.

Although the covariance matrix cannot be interpreted as the species interactions themselves, it does depend on them. The joint distribution is critical in models of forest composition, because the canopy is 'full'. So, regardless of scale, a species cannot increase unless others decline. Accounting for co-dependence is critical for accurate prediction. Although the nature of the covariance matrix changes with aggregation (Simpson's Paradox), the need to accommodate it applies across scales (Clark et al., 2014).

These JSDMs are also specified through hierarchical models, introducing latent multivariate normal structure, capturing dependence through the associated covariance matrices. We envision vectors of data $\mathbf{Y}_i$ collected at plot i with the jth entry in the vector being the response of the jth species at plot i. We assume the $\mathbf{Y}_i$ are independent across plots, i.e., that the plots are sufficiently dispersed that a change in composition at one location does not perceptibly influence composition at another. We focus on dependence between species, because, after accounting for the environmental effects at each plot through the mean effect, residual interspecies dependence persists. We introduce this dependence at the *first* stage, viewing $\mathbf{Y}_i$ as a componentwise function $Y_{ij} = g(V_{ij})$, where $\mathbf{V}_i$ is a latent multivariate normal. For example, with biomass, $Y_{ij} = V_{ij}$ if $V_{ij} > 0$ and $Y_{ij} = 0$ if $V_{ij} \leq 0$. This is a customary Tobit specification (Cameron and Trivedi, 2005). With binary response (presence/absence), $Y_{ij} = 1$ if and only if $V_{ij} > 0$, i.e., $g(V_{ij}) = 1$ for $V_{ij} > 0$. Note that we are not modeling dependence between the (random) probabilities of presence, e.g., between $P(Y_{ij} = 1)$ and $P(Y_{ij'} = 1)$ (Ovaskainen et al., 2015). This would move the dependence to the *second* modeling stage, i.e., in the *mean* specification, introducing a probit link. We believe that, with regard to the process, modeling the dependence at the data level is more informative. Similar remarks follow for any of the responses above; we can capture them through suitable latent multivariate normals (see Clark et al., 2016).

JSDMs enhance understanding of the distribution of species, but their applicability has been limited due to computational challenges when there is a large number of species. To appreciate the challenge with even the simplest presence/absence (binary) response and S species, we have an S-way contingency table with 2^S cell probabilities at any given site. With observational data collection over space and time, as in large ecological databases, the number of species is on the order of hundreds to thousands, rendering contingency table analysis infeasible. There is need for strategies to fit joint models in a computationally tractable manner.

To deal with these large datasets it is necessary to consider data dimension reduction techniques. Common approaches include principal component analysis (Artemiou and Li, 2009, 2013), partial least squares (Naik and Tsai, 2000), and sliced inverse regression (Li, 1991). Machine learning approaches include clustering techniques such as the

Chinese Restaurant process (Blei et al., 2010), Indian buffet process (Ghahramani and Griffiths, 2005), and latent Dirichlet allocation (Blei et al., 2003).

Here, we propose a Bayesian nonparametric approach, working with Dirichlet processes, to capture dependence among species while reducing the *effective dimensionality* of the problem. We use a stick-breaking representation based upon MacEachern (1994) which defines a dependent Dirichlet process (see also Dunson and Park (2008); Chung and Dunson (2011)). The attractiveness of stick-breaking constructions is that they are computationally efficient and easy to implement. In our work we employ the Dirichlet process in a novel way to reduce dimension in the dependence structure among species. In joint species modeling, a much different Bayesian nonparametric approach by Arbel et al. (2015) models species dependence across plots through a Gaussian process and a location dependent covariance structure. However, this approach assumes independence among species.

The primary contribution of this article is to develop a computationally feasible approach for a large number of species (say order 10^3), that allows us to: (i) assess the residual dependence structure among species, (ii) identify clusters of species that exhibit similar dependence patterns, and (iii) jointly predict species behavior. Again, the proposed hierarchical model captures dependence between species at the first or "data" stage rather than at a second or "mean" stage.

Dependence is introduced in the first-stage level through plot level latent multivariate normal variables whose dimension is the number of species. Dependence is introduced using low dimensional random effects described by a low dimensional covariance matrix, which is then made nonsingular through diagonal dominance. Dimension reduction in the form of clustering of the random effects is done with a Dirichlet process. Here, we employ the Dirichlet process for clustering in a novel way. Rather than the customary clustering of the replications (in our case, this would be plots), we cluster the components of the response (here, species). Instead of replacing Gaussian random effects with Dirichlet process random effects, we use the Dirichlet process to reduce dimension in the joint covariance structure. This last step makes computation tractable and is highly scalable. Altogether, we have a form of latent factor analysis but our implementation differs from that of Bhattacharya and Dunson (2011).

We use the Forest Inventory Analysis (FIA) data in the eastern United States to demonstrate our method. It consists of roughly 100 tree species considered on 1200 hectare sized plots. As a proof of concept for our dimension reduction approach, we also include a simulation using both continuous and binary response data.

Other potential settings for our Dirichlet process dimension reduction approach include: (i) gene expression data, anticipating dependence between some of the genes, where replicates would be expression data for individuals and (ii) stock price data for stocks associated with an index, anticipating dependence between some of the stocks. Replication here would be across time and could be accommodated through independent increments in a dynamic model.

This article is structured as follows. We describe the illustrative dataset in Section 2. In Section 3, we introduce our dimension reduction approach for a multivariate collection

of species, including details about the clustering step, an adaptation to binary data, and other necessary considerations. Section 5 discusses prediction under our approach. In Section 6 we provide simulation examples. Section 7 presents the Forest Inventory Analysis data example. In Section 8 we conclude with a summary and future work.

2 Motivating Example: Forest Inventory Analysis Data

A motivating challenge is joint modeling of multiple tree species in the Forest Inventory and Analysis (FIA) Data. It is a collection of annual inventory plots (2003–2009) in 31 states in the eastern US (Woudenberg et al., 2010, FIADB version 4.0.) that includes species, size, health of trees, tree growth, mortality, and removal by harvest. It is a systematic inventory of all US forests (USDA) that is documented online in a national resource report (Smith et al., 2009). In our analysis we use presence/absence of tree species from over 20,000 FIA monitored plots from which there are measurements of roughly 1 million trees spanning more than 200 tree species.

The FIA has a sampling protocol that is applied consistently throughout the country. This protocol consists of a quasi-systematic design that covers all ownerships across the United States, and these sampling efforts yield a national sampling intensity of one plot for every 2,438 hectares (Bechtold and Patterson, 2005). Each FIA plot is made up of four circular subplots with a radius of 7.32 meters (m) each, and the centroids for the subplots are 35.58 m apart. The aggregate area of the four subplots is 673.34 m^2 or ≈ 0.067 hectares (ha).

FIA plots are too small to adequately summarize local distribution and abundance. With a plot area equal to 25 m on a side, each supports only a handful of large trees, and most species occur sporadically. Predicting the composition on a plot this small is neither feasible, nor desirable. Instead we focus on the 1-ha scale, an area large enough to accurately represent forest structure; we aggregate plots to model at this scale. Given that FIA plots are located relatively far, nearest neighbors in geographic space need not represent similar environments. As such, geographic aggregation of FIA data can potentially group ridgetops with bottomlands. To avoid this issue, we combine FIA plots to the hectare scale by grouping them according to their covariate values. This translates the problem from a joint species distribution analysis in geographic space to one in covariate space. Prediction at this scale is more relevant to questions in ecology and climate change. The resulting joint distribution describes the tendency of species to respond together to the environment at the (meaningful) ha-scale, beyond what is captured in the mean structure of the model. To avoid ambiguity when discussing this application in what follows, we refer to the plots aggregated in covariate space simply as “plots”, and to the original ones as “FIA plots”.

To construct hectare-size plots in covariate space we follow the procedure described in Schliep et al. (2016). The authors used a k-means clustering algorithm based on stand age, temperature, and precipitation. Through this method, groups of 16 FIA plots (to make up a 1.07 ha sized plots) are identified, each of which minimizes the distance between the observations in the group and the cluster centroid, providing joint presence/absence of species at the one hectare scale (in covariate space).

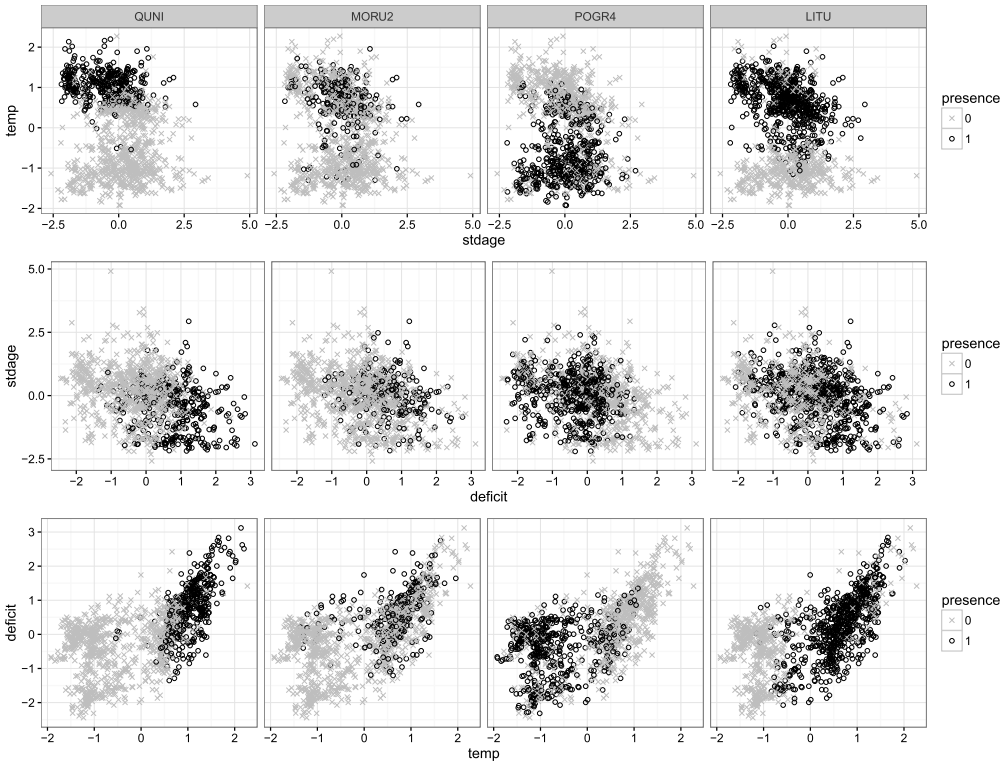


Figure 1: Presence/absence status of four randomly selected species at hectare sized plots in covariate space. Each *location* represents the centroid of the 16 combined FIA plots.

Species that were either too abundant (present in all but 50 plots or fewer) or too sparse (in fewer than 50) were removed such that inference on the regression parameters can be made for the species retained in the model. Data used here consist of presence/absence for $S = 112$ species. We use standard climate predictors, temperature and hydrothermal deficit, where the latter is the number of cumulative degree hours for months with a negative water balance. The values of these covariates for the plots are the means of the 16 FIA plots. Because the clustering is done in covariate space, the covariate values within the 16 plots are similar. After grouping, we have approximately 1200 hectare sized plots with presence/absence data. Figure 1 displays presence/absence status at the plots for four randomly selected species mapped onto the clustering covariate space.

3 A Joint Species Distribution Model Using the Dirichlet Process

In this section, we formalize the joint species distribution model. We begin with the model for the latent process $\mathbf{V}_i$. We then connect it to presence/absence data. As noted

in the introduction, we do not specify Dirichlet process models for the plots, i.e., for the $\mathbf{V}_i$ across i; we do not replace normal models for errors or random effects with Dirichlet process models. Rather, we use the Dirichlet process to provide a dimension reduction in specifying the dependence structure.

Suppose there are $j = 1, \cdots, S$ species at $i = 1, \cdots, n$ plots. For plot i, an S-dimensional vector, $\mathbf{V}_i$, is modeled as

$$\mathbf{V}_i = \mathbf{B}\mathbf{x}_i + \boldsymbol{\epsilon}_i \text{ with } \boldsymbol{\epsilon}_i \sim \mathrm{N}_S(\mathbf{0}_S, \Sigma), \tag{1}$$

where $B_{S\times p} = \begin{pmatrix} \boldsymbol{\beta}_1' \\ \vdots \\ \boldsymbol{\beta}_S' \end{pmatrix}$ is the coefficient matrix.

This model has $O(S^2)$ parameters, $\binom{S}{2} + S$ from Σ and, with p predictors in $\mathbf{x}_i$, including an intercept, there are pS parameters in $\mathbf{B}$. For example, for $S = 100$ species and $p = 3$ predictors, the model contains 5,350 parameters. We propose a dimension reduction approximation to Σ that allows the number of parameters to grow linearly in S.

We approximate Σ with $\Sigma^\star = \mathbf{A}\mathbf{A}' + \sigma_\varepsilon^2 \mathrm{I}$ and replace (1) with:

$$\mathbf{V}_i = \mathbf{B}\mathbf{x}_i + \mathbf{A}\mathbf{w}_i + \boldsymbol{\varepsilon}_i, \tag{2}$$

where the random vectors $\mathbf{w}_i$ are i.i.d. with $\mathbf{w}_i \sim \mathrm{N}_r(\mathbf{0}_r, \mathrm{I}_r)$ and $\boldsymbol{\varepsilon}_i \sim \mathrm{N}_S(\mathbf{0}_S, \sigma_\varepsilon^2 \mathrm{I})$, and $\mathbf{A}$ is an $S \times r$ matrix with $r < S$.

In fact, we think of $r << S$ and $\mathbf{A}$ as 'tall and skinny.' $\Sigma^\star$ has $Sr + 1$ unknowns clarifying the conversion of an $O(S^2)$ problem to an $O(S)$ problem. Evidently, the rank of $\mathbf{A}\mathbf{A}'$ is only r but adding $\sigma_\varepsilon^2 I$ yields diagonal dominance and a nonsingular matrix.[1] We require i.i.d. $\mathbf{w}_i$ in order to ensure that the $\mathbf{V}_i$ are conditionally independent given $\mathbf{A}$ and σ^2 as they are in (1) given Σ. A further reduction in the number of parameters can be attained by finding common rows in $\mathbf{A}$; a method to achieve this is described in the following section.

3.1 Clustering the Rows of A with a Dirichlet Process

Briefly, the Dirichlet process (DP) provides stochastic models for random distributions. It is widely used in the Bayesian nonparametric literature as a prior for distributions rather than using customary parametric distributions adopting priors for the parameters (Escobar and West, 1995; Ishwaran and James, 2001; Papaspiliopoulos and Roberts, 2008). It finds attractive application in hierarchical modeling to provide random distributions for random effects.

A constructive and useful representation of the DP is the stick-breaking formulation of Sethuraman (1994). It is ideally suited for implementation within a Gibbs sampling setting (see Escobar, 1994; Escobar and West, 1995; MacEachern, 1994; Bush and MacEachern, 1996; Neal, 2000) due to a Pólya urn scheme representation which

[1] One can imagine adding a more general diagonal matrix to AA' but, in our experience, $\sigma_\varepsilon^2 I$ performs well enough.

provides easy sampling from full conditional distributions. Under a stick breaking construction, we say the random distribution, G, follows a DP with base measure H and precision parameter α, $G \sim \mathrm{DP}(\alpha H)$, if

$$G(\cdot) \quad = \quad \sum_{l=1}^{\infty} p_l \delta_{\theta_l}(\cdot), \tag{3}$$

where $p_1 = \xi_1$, $p_l = \xi_l \prod_{h=1}^{l-1}(1-\xi_h)$ $(h \geq 2)$ with i.i.d. $\xi_l \sim \mathrm{Beta}(1,\alpha)$, and $\delta_{\theta_l}(\cdot)$ is the Dirac delta function at θ_l where $\theta_l \sim H$. Because it is almost surely a discrete distribution, it enables ties; hence, it enables clusters. We make use of this feature to allow some rows of $\mathbf{A}$ to be common, which corresponds to clustering species in their dependence behavior.

As shown in Ishwaran and Zarepour (2000), the sum in (3) can be approximated with a truncated version $\sum_{l=1}^{N} p_l \delta_{\theta_l}(\cdot)$, where $\xi_l \sim \mathrm{Beta}(\frac{\alpha}{N}, \frac{(N-l)}{N}\alpha)$ $(l < N)$, and $\xi_N = 1$ is needed to ensure $\sum_{l=1}^{N} p_l = 1$. The weights defined in this fashion come from a generalized Dirichlet distribution $\mathcal{GD}(\mathbf{a},\mathbf{b})$, which is conjugate with multinomial sampling.

We refer to this finite approximation as $\mathrm{DP}_N(\alpha H)$, and it is the version that we employ in developing our clustering strategy. The $\mathrm{DP}_N(\alpha H)$ formulation facilitates sampling with the blocked Gibbs sampler of Ishwaran and James (2001), which avoids marginalizing out the base measure, instead allowing the prior to be involved in the Gibbs sampling scheme.

To specify the hierarchical formulation for this model, let $\mathbf{Z} = (Z_j')_{j=1}^{N}$ (with $Z_j \overset{iid}{\sim} H$) denote the $N \times r$ matrix whose rows make up all potential atoms (i.e., vector values that the rows $\mathbf{a}_l'$ of $\mathbf{A}$ may take). In this setup, we need a vector of grouping labels $\mathbf{k} = (k_1, k_2, \ldots, k_S)$ $(1 \leq k_l \leq N)$ such that $\mathbf{a}_l = Z_{k_l}$. Note that $\mathbf{A}$ can be represented by $\mathbf{A} = Q(\mathbf{k})\mathbf{Z}$, where $Q(\mathbf{k})_{S\times N} = (\mathbf{e}_{k_1}, \mathbf{e}_{k_2}, \ldots, \mathbf{e}_{k_S})'$, and $\mathbf{e}_{k_l}$ is the N-dimensional vector with a 1 in position k_l and 0's elsewhere. Using this notation, our approximate model is given by

$$\begin{aligned}
\mathbf{V}_i|\mathbf{k},\mathbf{Z},\mathbf{w}_i,\mathbf{B},\sigma_\varepsilon^2 &\sim \mathrm{N}_S(\mathbf{B}\mathbf{x}_i + Q(\mathbf{k})\mathbf{Z}\mathbf{w}_i, \sigma_\varepsilon^2 \mathrm{I}_S), \quad \text{for} \quad i = 1,\ldots,n,\\
[\mathbf{B},\sigma_\varepsilon^2] &\propto \frac{1}{\sigma_\varepsilon^2},\\
\mathbf{w}_i &\sim \mathrm{N}_r(\mathbf{0},\mathrm{I}_r),\\
k_l|\boldsymbol{p} &\overset{iid}{\sim} \sum_{j=1}^{N} p_j \delta_j(k_l), \quad \text{for} \quad l = 1,\ldots,S,\\
Z_j|\mathbf{D}_\mathbf{z} &\overset{iid}{\sim} \mathrm{N}_r(\mathbf{0},\mathbf{D}_\mathbf{z}), \quad \text{for} \quad j = 1,\ldots,N,\\
\boldsymbol{p} &\sim \mathcal{GD}_N(\mathbf{a}_\alpha,\mathbf{b}_\alpha),\\
\mathbf{D}_\mathbf{z} &\sim \mathcal{IW}(2+r-1, 4\,\mathrm{diag}(1/\eta_1,\ldots,1/\eta_r)),\\
\eta_h &\sim \mathcal{IG}(1/2, 1/10^4), \quad \text{for} \quad h = 1,\ldots,r,
\end{aligned} \tag{4}$$

where $\mathrm{N}_S(\cdot,\cdot)$ is the S-dimensional multivariate normal distribution. Also, $\mathcal{GD}_N(\mathbf{a}_\alpha,\mathbf{b}_\alpha)$ corresponds to the N-dimensional Generalized Dirichlet process with $\mathbf{a}_\alpha = (\frac{\alpha}{N},\ldots,\frac{\alpha}{N})$ and $\mathbf{b}_\alpha = (\frac{\alpha(N-1)}{N}, \frac{\alpha(N-2)}{N}, \ldots, \frac{\alpha}{N})$, and $\delta_j(\cdot)$ represents a point mass at j.

As suggested in Ishwaran and Zarepour (2000), even for a very large number of species (S), a moderate level of truncation should suffice to approximate a DP(αH). In both our simulations and case study we set $N = \min\{150, S\}$. The specification for the prior of $\mathbf{D_z}$ follows the noninformative strategy to sample covariance matrices described in Huang and Wand (2013). Further details on implementation of the sampling algorithm for the hierarchical setup in (4) are provided in Supplementary Appendix A (Taylor-Rodríguez et al., 2016). Finally, we note that this approach for generating a reduced rank matrix $\mathbf{A}$, differs considerably from the approach in Bhattacharya and Dunson (2011). There, entries in $\mathbf{A}$ arise from inverse gamma scaled normals.

3.2 Adaptation for Binary Response (Presence/Absence)

In ecological surveys, by far the most common response is the set of presence/absence indicators for the set of species recorded at each plot. The FIA data described in Section 2 is an example of this. To work with binary responses, we resort to an adaptation of the data-augmentation algorithm proposed by Chib (1998) for multivariate probit regression, which improves the mixing of the MCMC algorithm. This modification is known in the literature as the parameter-expansion data-augmentation (PX-DA) algorithm (Liu and Wu, 1999; Lawrence et al., 2008; Schliep and Hoeting, 2015). It allows us to use the machinery proposed for the continuous case as a latent model. The PX-DA strategy was also considered by (Clark et al., 2016) to model non-continuous responses. In either case, approximation is needed to handle a large collection of species. The augmentation consists of introducing multivariate normal latent random variables, which are used to obtain a full conditional posterior density where the entire covariance matrix can be sampled. The sampled covariance is then re-scaled as a correlation matrix to accommodate the identifiability constraints imposed by the probit link. In our case, the approach is further modified to accommodate the dimension reduction step as described below.

Again, let $\Sigma^\star = \mathbf{A}\mathbf{A}' + \sigma^2_\varepsilon \mathrm{I}$ for some $S \times r$ matrix $\mathbf{A}$, and denote by $\mathbf{R} = D^{-1/2}\Sigma^\star D^{-1/2}$, where D is the diagonal matrix containing $\mathrm{diag}(\Sigma^\star)$. We can use data augmentation with the binary likelihood, assuming that, for plot i, $\mathbf{V} \sim \mathrm{N}_{n\times S}(\mathbf{X}\mathbf{B}', \mathbf{R}, \mathrm{I}_n)$, where $\mathbf{X}$ is the $n \times p$ matrix of predictors and $\mathbf{B}$ is the $S \times p$ matrix of regression coefficients. Assume that the matrix of binary responses is given by $\mathbf{Y} = [\mathbf{Y}_1, \ldots, \mathbf{Y}_n]'$, where $\mathbf{Y}'_i = (Y_{i1}, \ldots, Y_{iS})$ for $i = 1, \ldots, n$. Recall that we connect $\mathbf{Y}_i$ with $\mathbf{V}_i$ through $Y_{ij} = I(V_{ij} > 0)$ so that the contribution to the likelihood for species j in plot i is $I^{Y_{ij}}_{\{V_{ij}>0\}} I^{1-Y_{ij}}_{\{V_{ij}\le 0\}}$. With $\mathbf{y}_i = (y_{i1} \ldots, y_{iS})$ the binary vector of observed presences indicators at plot i, we have

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i) = \int_{\Gamma(y_{iS})} \cdots \int_{\Gamma(y_{i1})} (2\pi)^{-S/2} |\mathbf{R}|^{-1/2} \times \exp\left\{-\frac{1}{2}(\mathbf{V}_i - \mathbf{B}\mathbf{x}_i)' \mathbf{R}^{-1} (\mathbf{V}_i - \mathbf{B}\mathbf{x}_i)\right\} d\mathbf{V}_i,$$

where $\mathbf{V}_i$ is the ith row of $\mathbf{V}$, and $\Gamma(y_{ij})$ is $(-\infty, 0]$ if $y_{ij} = 0$ and $(0, \infty)$ if $y_{ij} = 1$.

Now, let $\mathbf{V}^\star_i = D^{1/2}\mathbf{V}_i$ and note that $\mathbf{V}^\star_i \sim \mathrm{N}_S(\mathbf{B}^\star \mathbf{x}_i, \Sigma^\star)$, where $\mathbf{B}^\star = D^{1/2}\mathbf{B}$. This change of variable doesn't affect the probabilities for $\mathbf{Y}_i$ (Lawrence et al., 2008). Hence,

$$\Pr(\mathbf{Y}_i = \mathbf{y}_i) = \int_{\Gamma(y_{iS})} \cdots \int_{\Gamma(y_{i1})} (2\pi)^{-S/2} \left|\Sigma^\star\right|^{-1/2} \times \exp\left\{-\frac{1}{2}\left(\mathbf{V}_i^\star - \mathbf{B}^\star \mathbf{x}_i\right)' \Sigma^{\star\,-1} \left(\mathbf{V}_i^\star - \mathbf{B}^\star \mathbf{x}_i\right)\right\} d\mathbf{v}_i^\star,$$

which in turn implies the expanded likelihood given by

$$\mathcal{L}(\mathbf{B}^\star, \Sigma^\star, \mathbf{V}^\star | \mathbf{Y}) = |\Sigma|^{-n/2} \left(\prod_{i=1}^{n} \exp\left\{-\frac{1}{2}\left(\mathbf{V}_i^\star - \mathbf{B}^\star \mathbf{x}_i\right)' \Sigma^{\star\,-1} \left(\mathbf{V}_i^\star - \mathbf{B}^\star \mathbf{x}_i\right)\right\} \times \prod_{j=1}^{S} \mathrm{I}_{\left\{V_{ij}^\star > 0\right\}}^{\,y_{ij}} \mathrm{I}_{\left\{V_{ij}^\star \leq 0\right\}}^{\,1-y_{ij}} \right) [\mathbf{B}^\star] \left[\Sigma^\star\right]. \tag{5}$$

As in the continuous response case, we can introduce an $n \times r$ matrix of standard normal random variables $\mathbf{W}$, such that $\mathbf{V}^\star | \mathbf{B}^\star, \mathbf{A}, \mathbf{W}, \sigma_\varepsilon^2 \sim \mathrm{N}_{n \times S}(\mathbf{X}\,\mathbf{B}^{\star\,\prime} + \mathbf{W}\mathbf{A}', \sigma_\varepsilon^2 \mathrm{I}_S, \mathrm{I}_n)$, where $\mathrm{N}_{a \times b}(M, V_{col}, V_{row})$ represents the $a \times b$-dimensional matrix normal distribution with mean M, and column and row covariance matrix V_{col} and V_{row}, respectively.

Hence, the expanded likelihood can now be expressed as

$$\mathcal{L}_{PA}(\mathbf{B}^\star, \mathbf{W}, \mathbf{A}, \sigma_\varepsilon^2, \mathbf{V}^\star | \mathbf{Y}) \propto (\sigma_\varepsilon^2)^{-nS/2} \left(\prod_{i=1}^{n} \exp\left\{-\frac{1}{2\sigma_\varepsilon^2} \|\mathbf{V}_i^\star - \mathbf{B}^\star \mathbf{x}_i - \mathbf{A}\mathbf{w}_i\|^2 - \frac{1}{2}\|\mathbf{w}_i\|^2\right\} \times \prod_{j=1}^{S} \mathrm{I}_{\left\{V_{ij}^\star > 0\right\}}^{\,y_{ij}} \mathrm{I}_{\left\{V_{ij}^\star \leq 0\right\}}^{\,1-y_{ij}} \right) [\mathbf{B}^\star] [\mathbf{A}] \left[\sigma_\varepsilon^2\right]. \tag{6}$$

With the expanded likelihood in (6), the sampling algorithm becomes

1. Sample $\mathbf{V}_i^\star \sim \mathrm{tr.N}_S(\mathbf{B}^\star \mathbf{x}_i + \mathbf{A}\mathbf{w}_i, \sigma_\varepsilon^2 \mathrm{I}_S; \Gamma(\mathbf{y}_i))$. This proposal density is convenient as it corresponds to drawing from univariate truncated normal random variables.
2. Draw $\mathbf{A} = Q(\mathbf{k})\mathbf{Z}$ and σ_ε^2 as in the continuous case (see in Supplementary Appendix A sampling steps for $\mathbf{Z}, \mathbf{k}, \boldsymbol{p}$ and σ_ε^2), where the full conditional densities for $\mathbf{A}$ and σ_ε^2 depend on $\mathbf{V}^\star, \mathbf{B}^\star$, and $\mathbf{W}$.
3. Assuming a flat prior on $\mathbf{B}^\star$, the full conditional posterior distribution for $\mathbf{B}^\star$ is given by $\mathrm{N}_{S \times p}((\mathbf{V}^\star - \mathbf{W}\mathbf{A}')'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}, \sigma_\varepsilon^2(\mathbf{X}'\mathbf{X})^{-1}, \mathrm{I}_S)$.
4. Finally, obtain the variables on the correlation scale using the transformations $\mathbf{V} = D^{-1/2}\mathbf{V}^\star$, $\mathbf{B} = D^{-1/2}\mathbf{B}^\star$ and $\mathbf{R} = D^{-1/2}(\mathbf{A}\mathbf{A}' + \sigma_\varepsilon^2 \mathrm{I})D^{-1/2}$.

3.3 Non-Negative Response (Continuous Abundance)

In many species distribution surveys, the biomass of living organisms is recorded as an indication of species abundance. This type of response is continuous, but takes on 0

values whenever the species is absent from a plot. Functionally, we have a Tobit model defined by

$$Y_{ij} = \begin{cases} V_{ij} & \text{if } V_{ij} > 0 \\ 0 & \text{if } V_{ij} \leq 0 \end{cases}, \tag{7}$$

where Y_{ij} corresponds to the random variable that measures continuous abundance for species j at plot i, and V_{ij} is an element of the S-dimensional vector $\mathbf{V}_i$ as specified in (2).

Using the connection between Y_{ij} and V_{ij} provided by (7), the likelihood for continuous abundance data is given by

$$\mathcal{L}_{CA}(\mathbf{B}, \mathbf{W}, \mathbf{A}, \sigma_\varepsilon^2, \mathbf{V}|\mathbf{Y}) \propto \left(\sigma_\varepsilon^2\right)^{-\frac{nS}{2}} \left(\prod_{i=1}^{n} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \|\mathbf{V}_i - \mathbf{B}\mathbf{x}_i - \mathbf{A}\mathbf{w}_i\|^2 - \frac{1}{2}\|\mathbf{w}_i\|^2 \right\} \times \prod_{j=1}^{S} \left(\mathrm{I}_{\{V_{ij} \leq 0\}}\right)^{\mathrm{I}_{\{Y_{ij}=0\}}} \left(\mathrm{I}_{\{V_{ij}=Y_{ij}\}}\right)^{\mathrm{I}_{\{Y_{ij}>0\}}} \right) \times [\mathbf{B}]\,[\mathbf{A}]\left[\sigma_\varepsilon^2\right]. \tag{8}$$

This formulation provides for direct Gibbs sampling, with full conditionals for the latent variables given by

$$V_{ij} \sim \begin{cases} \text{tr.N}\left(\mathbf{x}_i'\boldsymbol{\beta}_j + \mathbf{w}_i'\mathbf{a}_j, \sigma_\varepsilon^2; (-\infty, 0]\right)) & \text{if } Y_{ij} \leq 0 \\ \mathrm{I}_{\{Y_{ij}\}} & \text{otherwise} \end{cases},$$

where $\boldsymbol{\beta}_j$ and $\mathbf{a}_j$ correspond to the jth row of $\mathbf{B}$ and $\mathbf{A}$, respectively. Sampling for $\mathbf{B}$, $\mathbf{A}$ and σ_ε^2 is performed in the same way as in the continuous case (see Supplementary Appendix A).

4 Prediction

While explanation through the covariates is usually the objective in species distribution modeling, prediction also plays a key role. Prediction can be envisioned for unobserved plots, for instance, predicting probabilities of presence–absence, or predicting the biomass for the set of species at unobserved plots. With independent plots, this becomes prediction of a latent multivariate vector using the fitted regression. Given that species are correlated within plots, if the response for some of the species is available at a new plot (e.g., presence/absence), one may take advantage of this fact to further inform prediction of the responses at the new plot for the unobserved species, conditionally on values for the observed ones. This is useful if only information on subset of species is available for some new plots and prediction on the remaining species is of interest. Alternatively, this feature can be exploited in formulating hypothetical experiments. If we assume certain species are present at a site, which species are likely to join them; if we assume certain species are absent at a site, which species are likely to replace them.

As mentioned earlier, individual level models when aggregated tend to overestimate richness – the number of distinct species at a plot (Clark et al., 2016; Calabrese et al.,

2014). Under joint modeling the data we are fitting are vectors, each of 0's and 1's, such that each vector has an implicit observed richness. As a result, when the joint model is fitted, these observations inform about the number of distinct species at a plot. We will expect to see substitution among species that are equally suited to the available environment at the new plot. However, we will not expect the overall number of species to be overestimated.

We briefly describe how such prediction is implemented for the latent variables and then add the modifications for binary and non-negative data. First we derive the posterior predictive distribution. Recall that our latent fitting model is $\mathbf{V}_i = \mathbf{B}\mathbf{x}_i + \mathbf{A}\mathbf{w}_i + \boldsymbol{\varepsilon}_i$, where $\mathbf{A} = Q(\mathbf{k})\mathbf{Z}$ depends on the label vector $\mathbf{k} = (k_1, \ldots, k_S)$ through $Q(\mathbf{k})$. Again, $Q(\mathbf{k})$ is the $S \times N$ matrix that assigns the rows of $\mathbf{Z}$ to the rows of $\mathbf{A}$ according to the clustering specified by the vector of groupings $\mathbf{k}$. Also recall that the $N \times r$ matrix $\mathbf{Z}$ is such that its rows, denoted by Z_j, are iid with prior distribution $\mathrm{N}_r(0, \mathbf{D}_\mathbf{z})$, for $j = 1, \ldots, N$. Finally, we have $\mathbf{w}_i \overset{iid}{\sim} \mathrm{N}_q(0, \mathrm{I}_r)$, $\boldsymbol{\varepsilon}_i \overset{iid}{\sim} \mathrm{N}_q(0, \sigma_\varepsilon^2 \mathrm{I}_S)$ for $i = 1, \ldots, n$, and the labels $k_l | \boldsymbol{p} \overset{iid}{\sim} \sum_{j=1}^N p_j \delta_j(k_l)$ for $l = 1, \ldots, S$, with the p_j's drawn from a generalized Dirichlet process.

We sort the columns of $\mathbf{V}$ to have in the first S_p columns the response for species that will be predicted at the new plots. The remaining $S - S_p$ correspond to those species specified as observed at the new plots. The response matrix at the n_o new locations is given by $\mathbf{V}^o = [\mathbf{V}^o_{pred}\ \mathbf{V}^o_{obs}]$. Next, denoting $\boldsymbol{\theta} = (\mathbf{B}, \mathbf{Z}, \mathbf{k}, \boldsymbol{p}, \sigma_\varepsilon^2)$ and letting $\mathbf{X}^o$ be the matrix of environmental features at the new plots, the posterior predictive distribution after integrating out the $\mathbf{w}_i$'s, is

$$\Pr(\mathbf{V}^o_{pred} | \mathbf{V}, \mathbf{V}^o_{obs}) \quad = \quad \int_{\boldsymbol{\Theta}} p(\mathbf{V}^o_{pred} | \boldsymbol{\theta}, \mathbf{V}, \mathbf{V}^o_{obs}) p(\boldsymbol{\theta} | \mathbf{V})\, \mathrm{d}\boldsymbol{\theta}. \tag{9}$$

Now, letting $\boldsymbol{\mu}_i = \mathbf{B}\mathbf{x}_i$ and $\Sigma^\star = Q(\mathbf{k})\mathbf{Z}\mathbf{Z}'Q(\mathbf{k})' + \sigma_\varepsilon^2 \mathrm{I}_S$, we have

$$p(\boldsymbol{\theta} | \mathbf{V}) \quad \propto \quad \left(\prod_{i=1}^n \phi_S\left(\mathbf{V}_i \middle| \boldsymbol{\mu}_i, \Sigma^\star\right) \right) \pi_\theta(\boldsymbol{\theta}),$$

and

$$p(\mathbf{V}^o_{pred} | \boldsymbol{\theta}, \mathbf{V}, \mathbf{V}^o_{obs}) \quad = \quad \prod_{h=1}^{n_o} \phi_{S_p}\left(\mathbf{V}^o_{h,pred} \middle| \boldsymbol{\mu}_{h,pred|obs}, \Sigma_{pred|obs}\right),$$

where $\phi_k(\cdot | \boldsymbol{\mu}, \Sigma)$ denotes the k-dimensional normal density with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix Σ, and

$$\boldsymbol{\mu}_{h,pred|obs} = \boldsymbol{\mu}_{h,pred} + \Sigma_{1,2}\Sigma_{2,2}^{-1}\left(\mathbf{V}^o_{h,obs} - \boldsymbol{\mu}_{h,obs}\right),$$

with $\boldsymbol{\mu}_{h,pred} = \mathbf{B}_{pred}\mathbf{x}^o_h$, $\boldsymbol{\mu}_{h,obs} = \mathbf{B}_{obs}\mathbf{x}^o_h$, and

$$\Sigma_{pred|obs} = \Sigma_{1,1} - \Sigma_{1,2}\Sigma_{2,2}^{-1}\Sigma'_{1,2}, \text{ with } \Sigma^* = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma'_{1,2} & \Sigma_{2,2} \end{pmatrix}.$$

To sample the predicted responses, instead of integrating out the parameters, we can obtain samples from $\Pr(\mathbf{V}^o_{pred} | \mathbf{V}, \mathbf{V}^o_{obs})$ by using the MCMC draws of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta} | \mathbf{V})$, and then drawing $\mathbf{V}^o_{pred}$ from $p(\mathbf{V}^o_{pred} | \boldsymbol{\theta}, \mathbf{V}, \mathbf{V}^o_{obs})$.

Adaptation to the binary case is clear. The latent variables for the responses of the observed species need to be drawn to predict the latent variables for the unobserved species. Positive latent responses for the unobserved species are set to 1 (presence), and negative latent responses are set to 0 (absence). With presence/absence, $\mathbf{Y}^o$ is now a matrix of binary responses at new locations, while $\mathbf{V}^o$ represents the matrix with the associated latent variables. Using analogous notation to that from Section 3.2, we have

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \int \mathcal{L}_{PA}(\boldsymbol{\theta}, \mathbf{V}|\mathbf{Y}) d\mathbf{V},$$

and

$$p(\mathbf{Y}^o_{pred}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{Y}^o_{obs}) \propto \prod_{h=1}^{n_o} \int p(\mathbf{V}^{\star o}_{h,pred}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{V}^{\star o}_{h,obs}) \times \prod_{j=1}^{S_p} \mathrm{I}_{\{V^\star_{hj}>0\}}{}^{Y_{hj}} \mathrm{I}_{\{V^\star_{hj}\leq 0\}}{}^{1-Y_{hj}} \mathrm{d}\mathbf{V}^{\star o}_{h,pred},$$

where $V^\star_{hj}$ and Y_{hj} represent the elements of the vectors $\mathbf{V}^{\star o}_{h,pred}$ and $\mathbf{Y}^o_{h,pred}$, respectively. Here, $p(\mathbf{V}^{\star o}_{h,pred}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{V}^{\star o}_{h,obs}) = \phi_{S_p}(\mathbf{V}^{\star o}_{h,pred}|\boldsymbol{\mu}_{h,pred|obs}, \Sigma_{pred|obs})$, making the obvious modifications to $\boldsymbol{\mu}_{h,pred|obs}$ and $\Sigma_{pred|obs}$ in the definitions provided for the continuous case to accommodate $\mathbf{V}^{\star o}$ instead of $\mathbf{Y}^o$ and $\mathbf{B}^\star$ instead of $\mathbf{B}$. Similarly, prediction is done by using the MCMC draws from $\boldsymbol{\theta}$.

Prediction with continuous abundance data follows a similar strategy. For the observed species at the new plots, the latent variables at new plots are drawn from normals truncated to be negative for species that are absent, and are set equal to the response for those species that take on positive values of the response. In this case

$$p(\boldsymbol{\theta}|\mathbf{Y}) \propto \int \mathcal{L}_{CA}(\boldsymbol{\theta}, \mathbf{V}|\mathbf{Y}) d\mathbf{V}$$

and

$$p(\mathbf{Y}^o_{pred}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{Y}^o_{obs}) \propto \prod_{h=1}^{n_o} \prod_{j=1}^{S_p} g(Y_{hj}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{V}^{\star o}_{h,obs}),$$

where

$$g(Y_{hj}|\boldsymbol{\theta}, \mathbf{Y}, \mathbf{V}^{\star o}_{h,obs}) = \begin{cases} \int_{\mathbb{R}^-} \phi(V_{hj}|\mu_{(hj|obs)}, \sigma^2_{(hj|obs)}) dV_{hj} & \text{if } Y_{hj} = 0 \\ \phi(Y_{hj}|\mu_{(hj|obs)}, \sigma^2_{(hj|obs)}) & \text{otherwise} \end{cases},$$

where $\mu_{(hj|obs)} = \mathrm{E}[V_{hj}|\theta, \mathbf{V}^o_{h,obs}]$ and $\sigma^2_{(hj|obs)} = \mathrm{var}(V_{hj}|\theta, \mathbf{V}^o_{h,obs})$.

Prediction can also be used for model validation/comparison. That is, a portion of the data can be held out from the fitting and used for validation. In Section 5, using simulated data, we offer a few examples of how this can be done.

5 Simulations

In this section, we conduct simulations for both the continuous and the binary case. The simulation for the continuous case is merely intended as a proof of concept. That is, on the latent scale, we show how well we can recover the specified multivariate dependence structure. The binary case is intended to demonstrate what we can expect to recover with real presence/absence data. The performance of the algorithm is assessed in terms of out-of-sample prediction. For the continuous case, we generate data for a large number of species using two illustrative simulation mechanisms according to how the variance–covariance matrix is specified. At plot i, the continuous joint response vector (e.g., biomass) for S species is drawn from $\mathbf{y}_i \sim \mathrm{N}_S(\mathbf{0}_S, \Sigma)$. For the binary case, we generate the data using one choice of covariance structure and compare the parameter estimates and predictions with and without dimension reduction, as well as assuming independent models for each species. To fit the models without dimension reduction and compare the results with the reduced dimension approach, we consider a smaller number of species than in the continuous case.

In generating the continuous responses we consider two types of covariance matrices for 1000 species (see Figure 2 for an illustration). For each data generating mechanism, we employ $r = 5, 10, 20$ to assess the effect of the approximation. In the binary case, we only use a single covariance structure (imposed on the latent scale) and simulate data for 100 species. This is manageable to fit even without implementing a dimension reduction scheme, and allows us to compare the estimates and prediction with and without dimension reduction. With both types of data we also obtain the results from the independent stacked models. Below we describe in detail the simulation setup used in each case.

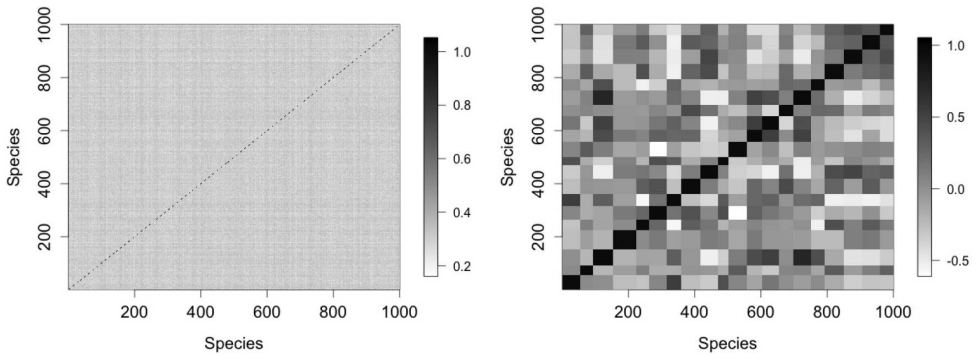
5.1 Continuous Case

The two covariance structures considered for the simulations with continuous responses are given by

1. *Expected equicorrelation*: Assumes $\Sigma = \Phi\Phi'$ where the ith row of Φ is given by $\boldsymbol{\phi}_i = \theta_0\mathbf{v}_0 + \theta_1\mathbf{v}_i$ with the $\mathbf{v}$'s i.i.d. from $\mathrm{N}_S(\mathbf{0}, \sigma_v^2\mathrm{I}_S)$. Under this alternative we have that $E[\Sigma_{ii}] = E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_i] = S\sigma_v^2(\theta_0^2 + \theta_1^2)$ and $E[\Sigma_{ij}] = E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_j] = S\sigma_v^2\theta_0^2$.
2. *Clustered covariance*: This structure assumes that $\Sigma = \mathbf{A}_{true}\mathbf{A}_{true}' + \tau^2\mathrm{I}_S$, where $\mathbf{A}_{true}$ is an $S \times q$ matrix (with $q \leq S$). Each row of $\mathbf{A}_{true}$ is assigned a label randomly from $\kappa = 1, \ldots, \mathbb{K}_{true}$, where $\mathbb{K}_{true} < S$, such that if the lth row is assigned label κ, then $\mathbf{a}_l = \mathbf{v}_\kappa$, with $\mathbf{v}_\kappa \sim \mathrm{N}_q(\mathbf{0}, \sigma_v^2 I_q)$.

We view equicorrelation as essentially a straw man; it is not what we expect in practice. The second scenario, clustering of species, is more likely in real data and is, in fact, what our DP dimension reduction approach is designed to capture. Below, we will see that, indeed, our approach does work well.

Using these two alternatives, 10 independent data sets were generated in order to examine the ability of the algorithm to recover the true covariance structure and to



(a) equicorrelated (b) clustered with $\mathbb{K}_{true} = 20$

Figure 2: Covariance structures.

perform out-of-sample prediction. Assessing out-of-sample predictive performance is done with respect to a hold out sample of species within plots as opposed to using a hold out sample of plots on all species. Predictive performance is assessed by calculating the Euclidean distances between the true values and the conditional predictions, predicting 5%, 10%, 20% and 50% of the species, conditional on the remaining 95%, 90%, 80% and 50% species, respectively. We denote the out-of-sample response matrix by $\mathbf{V}^o$. The rows of $\mathbf{V}^o$, denoted by $\mathbf{V}^o_i$, correspond to some permutation of $(\mathbf{V}^o_{i,pred}\ \mathbf{V}^o_{i,obs})$, where $\mathbf{V}^o_{i,obs}$ is the vector of responses for species that are assumed observed and $\mathbf{V}^o_{i,pred}$ are those considered for prediction.

The prediction vector, $\mathbf{V}^o_{i,pred}$ is given by $(V_{im_1}, \ldots, V_{im_{S_p}})$, with S_p denoting the number of out of sample species chosen to make prediction for, and $m_1, m_2, \ldots, m_{S_p}$ denote the set of indices chosen at random for the species to be predicted. Similarly, $\hat{\mathbf{V}}^o_{i,pred}$ denotes the vector of predicted values. Therefore, $\mathbf{V}^o_{i,obs}$ corresponds to the remaining $S - m_{S_p}$ responses. The criterion used to assess predictive ability of the algorithm is the root mean squared predictive error (RMSPE), given by

$$\text{RMSPE} = \sqrt{\frac{1}{S_p n_o}\sum_{i=1}^{n_o}(\mathbf{V}_{i,pred} - \hat{\mathbf{V}}_{i,pred})'(\mathbf{V}_{i,pred} - \hat{\mathbf{V}}_{i,pred})}.$$

To test the algorithm we built the $\mathbf{A}$ matrix in model (2) setting $r = 5, 10, 20$ columns. The parameters chosen for each of the covariance types are:

Expected equicorrelation $\theta_0 = \sqrt{\frac{0.3}{\psi}}, \theta_1 = \sqrt{\frac{0.7}{\psi}}$ and $\sigma_v^2 = \frac{\psi}{S}$, where $\psi = 2$. This choice of parameters yields $E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_i] = 1$ for the diagonal elements, and $E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_j] = 0.3$ for the off-diagonal ones.

Clustered covariance $\mathbb{K}_{true} = 80, q = 20$ and $\sigma_v^2 = 3$.

For the simulation experiments with continuous response, we assume $S = 1000$ species on $n = 3000$ plots. With every covariance structure, 10 datasets were generated at random from $\mathbf{y}_i \sim \mathrm{N}_S(\mathbf{0}_S, \Sigma)$ for $i = 1, \ldots, n$. Using each of these datasets, the algorithm was applied individually for the different values of r. The results extracted from each run of the algorithm are the Euclidean distances between the true out of sample dataset $\mathbf{V}^o_{pred}$ and their posterior predictive means $\hat{\mathbf{V}}^o_{pred}$.

In Supplementary Appendix B we included Figures 1 and 2 which display, for a single dataset, a comparison between: (i) the true and estimated covariance parameters, and (ii) the observed responses with the out-of-sample predicted values. The figures shown were obtained for $\mathbf{A}$ with $r = 5$ columns, corresponding to a single dataset generated either with the equicorrelation or the clustered covariance.

When the observed data is generated using the expected equicorrelation structure, our methodology attractively recovers two distinctly different groups of parameters, namely, those on the diagonal and those off the diagonal. This is expected, following from the difference in magnitude in their expected values, which are $E[\Sigma_{ii}] = E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_i] = S\sigma_v^2(\theta_0^2+\theta_1^2)$ and $E[\Sigma_{ij}] = E[\boldsymbol{\phi}_i'\boldsymbol{\phi}_j] = S\sigma_v^2\theta_0^2$ for the diagonal and off-diagonal elements, respectively.

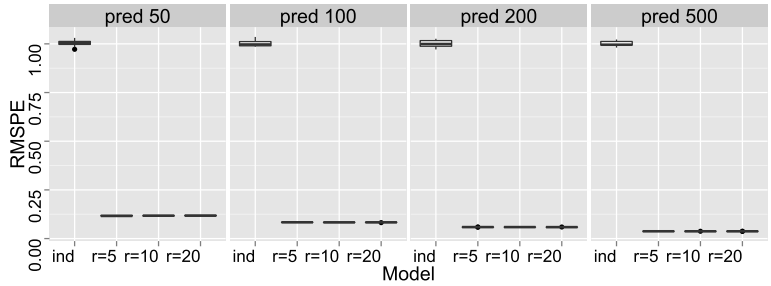
At a first glance, in terms of prediction the method appears to perform poorly under equicorrelation. However, the very nature of the data generating mechanism induces this behavior. To better understand the role that expected equicorrelation plays, consider the conditional expectation at a single plot replacing Σ by its expected (equicorrelation) covariance matrix $\Lambda = E[\Sigma] = \left(\begin{smallmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{smallmatrix}\right)$. The diagonal entries of Λ are given by $S\sigma_v^2(\theta_0^2+\theta_1^2)$ and the off-diagonal elements are $S\sigma_v^2\theta_0^2$, such that the posterior conditional predictive means are

$$E[\mathbf{V}_{pred}|\mathbf{y}_{obs}, \Lambda] \quad = \quad \frac{\theta_0^2(S-S_p)}{\theta_1^2+\theta_0^2(S-S_p)}\bar{V}_{obs}\mathbf{1}_{S_p},$$

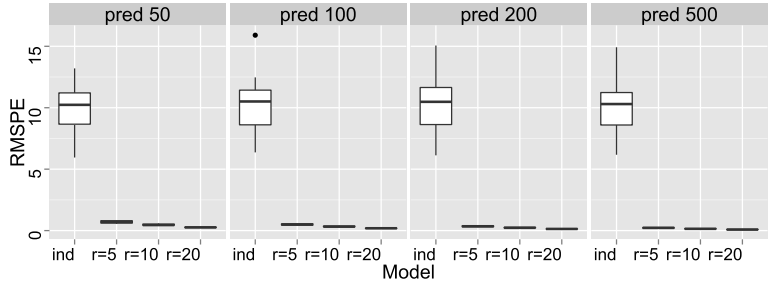
where $\boldsymbol{J}$ is an $S_p \times (S-S_p)$ matrix of ones and $\bar{V}_{obs}$ is the average of the entries in $\mathbf{V}_{obs}$. Therefore, the conditional mean for $\mathbf{y}_{pred}$ at a given plot is equal for all of its entries. So even though prediction from our method appears to render poor results, it is in fact behaving as if the true covariance matrix was equicorrelated.

With the clustered covariance structure our approach fares quite well, both recovering the true covariance and predicting new observations accurately (see Figures 1 and 2 in Supplementary Appendix B (Taylor-Rodríguez et al., 2016)).

The notable improvement in conditional prediction by accounting for the interspecific dependence is also observed in Figure 3. Under both covariance structures used to generate the true data, modeling the dependence dramatically reduces the error in out-of-sample prediction. In this figure we can also assess the effect of increasing r, the number of columns in $\mathbf{A}$. For the equicorrelated covariance, the results indicate that adding more columns yields an insubstantial improvement. With the clustered covariance structure, the RMSPE slowly decreases as r grows until $r = q = 20$, the number of columns in $\mathbf{A}_{true}$. Hence, when the true covariance structure is of the same form as that from our approximation, the error is minimized when $\mathrm{rank}(\mathbf{A}) = \mathrm{rank}(\mathbf{A}_{true})$.



(a) Equicorrelation



(b) Clustered

Figure 3: Out-of-sample RMSPE for continuous simulations using the (a) unstructured and (b) clustered covariances. Results compared the independence model to the dimension reduced approach (assuming $r = 5, 10, 20$) when predicting 50, 100, 200 and 500 species (out of 1000) from 10 simulated datasets with $n = 3000$.

Although not shown here, it is also reassuring that *a posteriori* the number of clusters is highly concentrated about 80, the true number of clusters $\mathbb{K}_{true}$ in $\mathbf{A}_{true}$.

5.2 Binary Case

For presence–absence data, we consider a single covariance type (imposed on the latent scale), of the form $\Sigma = (\Psi\Psi')^{-1}$, where Ψ is an $S \times S$ matrix generated from independent standard normal random variables, which yields inverse Wishart draws. In this simulation, we considered only $S = 100$ species on 1000 different plots in order to be able to compare our results with those from the estimation procedure that considers dependence among species but has no dimension reduction. The mean structure is defined by an intercept and a single predictor. Thus, the true model contains 100×2 regression coefficients and $\binom{100}{2} = 4950$ correlation parameters.

Our dimension reduction approach requires a choice of r, the number of columns to be used for the $\mathbf{A}$ matrix. Noting the similarity of this formulation with factor analysis models, some guidelines regarding this choice can be extracted from this extensive body of literature. A pervasive notion in this literature is that the relevant number of factors (i.e., columns in $\mathbf{A}$) generally ranges from small to moderate (e.g., see Lopes and West, 2004).

In this spirit, letting $\mathbb{K}$ denote the number of unique cluster labels, then $\text{rank}(\mathbf{A}) = \min\{r, \mathbb{K}\}$. Given that $\mathbf{A}$ must be of full column rank for the model to be well identified (Geweke and Singleton, 1980), then $\mathbb{K}$ (or an approximation) can act as an upper bound on r. The number of unique clusters is unknown, but in our experience, for a given r, $\mathbb{K}$ can be assessed relatively fast by monitoring the number of unique clusters drawn throughout the MCMC algorithm. If this value appears too small relative to r, then r must be reduced. Another indication that r is inadequately large is the appearance of multimodalities in the posterior densities of the elements in $\mathbf{A}$ (Aguilar and West, 2010).

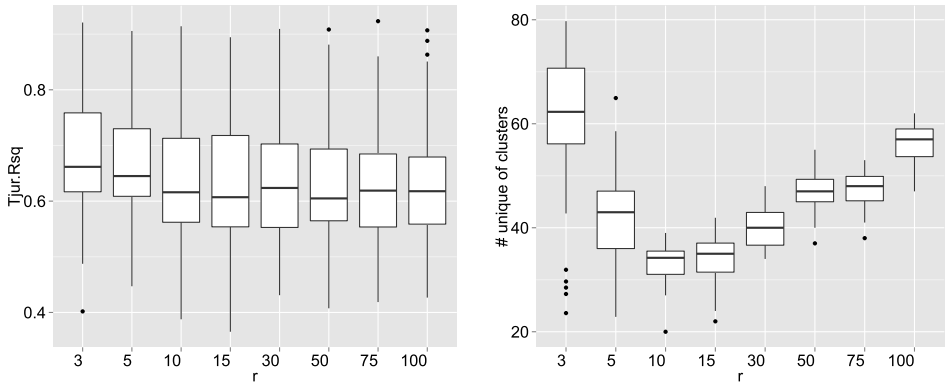
A straightforward approach for selecting r is to perform sensitivity analysis varying r, and selecting the value that optimizes some criterion. Here, for the dimension reduced approach we fit models with $r = 3, 5, 10, 15, 30, 50, 75, 100$. To identify a suitable number of columns we use the Tjur R^2 coefficient of determination (Tjur, 2009), which compares the estimated probabilities of presence between the observed ones and the observed zeros. For species j, this quantity is given by $TR_j = (\hat{\bar{\pi}}_j^{(1)} - \hat{\bar{\pi}}_j^{(0)})$, where $\hat{\bar{\pi}}_j^{(1)}$ and $\hat{\bar{\pi}}_j^{(0)}$ are the average probabilities of presence for the observed ones and zeros of the jth species, respectively. The larger the TR_j, the better the discrimination. An overall criterion is to obtain the average over all species given by $\overline{TR} = \frac{1}{S}\sum_{j=1}^{S} TR_j$, which we employ in this section. We would hope to find a peak in $\overline{TR}$ as we increase r. If there is little variation in $\overline{TR}$ over r, then we need not be concerned about the choice of r and simply choose a suitably parsimonious model.

Although the analysis that follows is conducted for a single dataset, to observe the behavior of the Tjur R^2 for the proposed simulation scenario, we generate 50 independent datasets, and for each of them calculate the posterior mean of $\overline{TR}$ for the different values of r (displayed in Figure 4). Additionally, for these 50 datasets we extracted the posterior mean for $\mathbb{K}$ to determine if identifiability issues might arise.

Here we observe that the model fit is similar for the different values of r, with $r = 3$ being slightly better than the rest. Also, for values $r \geq 50$ there might be model identifiability problems given that $\mathbb{K} \leq r$.

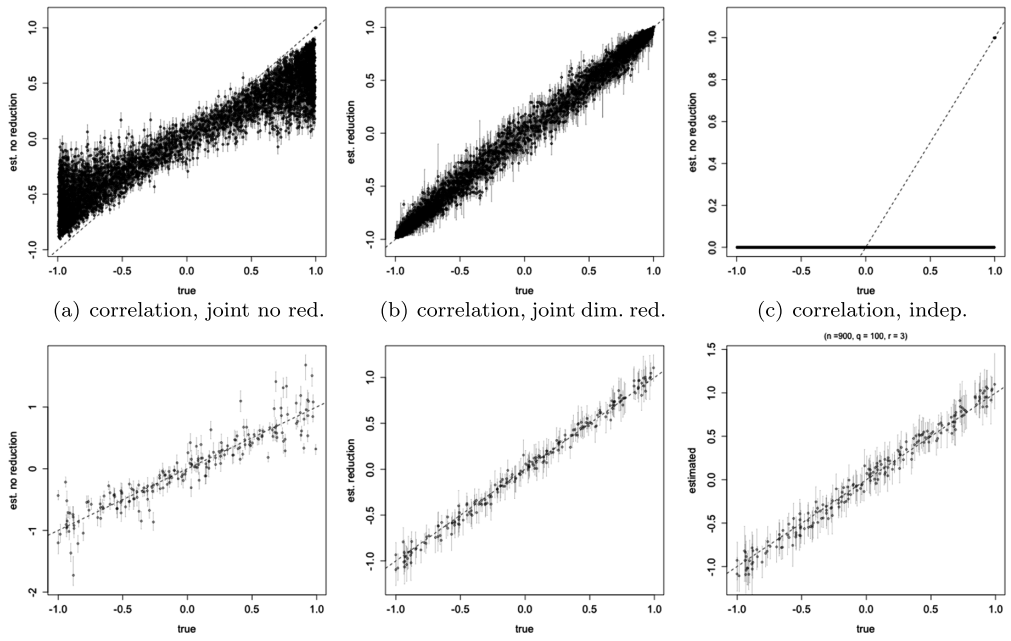
We focus on a single dataset chosen at random from the 50 that were generated, and use the results from the model with $r = 3$, given the behavior observed in Figure 4. For comparison, we also perform the estimation using (i) the joint model without reducing the dimensionality fitted with the gjam R package (Clark et al., 2016), and (ii) a model where the species are fitted independently and the results are stacked together.

In Figure 5, we compare the parameter estimates obtained from the joint model without reduction (left column), the joint model with dimension reduction (middle column), and the independence model (right column). The top row in Figure 5 compares the ability of the three methods to recover the correlation parameters by contrasting their estimates to the true values (using the means and 95% credible intervals). In this case, the joint method without dimension reduction attempts to estimate all of the $\binom{100}{2} = 4950$ parameters in the correlation matrix; however, it appears to have some difficulties recovering the parameters for the sample size considered. Conversely, the



(a) Posterior mean of $\overline{TR}$ in 50 datasets (b) Posterior mean of $\mathbb{K}$ in 50 datasets

Figure 4: (a) Goodness-of-fit measured by the posterior mean of Tjur R^2 for 50 simulated datasets with binary response for 100 species in 1000 plots having $r = 3, 5, 10, 15, 30, 50, 75, 100$. (b) Posterior mean for the unique number of clusters ($\mathbb{K}$) drawn for each of 50 datasets for the value of r considered.



(a) correlation, joint no red. (b) correlation, joint dim. red. (c) correlation, indep.

(d) reg. coef., joint no red. (e) reg. coef., joint dim. red. (f) reg. coef., indep.

Figure 5: Binary simulation results with 100 species at 1000 plots. Comparison between joint model without reduction (top row), joint model with dimension reduction (middle row), and independent model (bottom row).

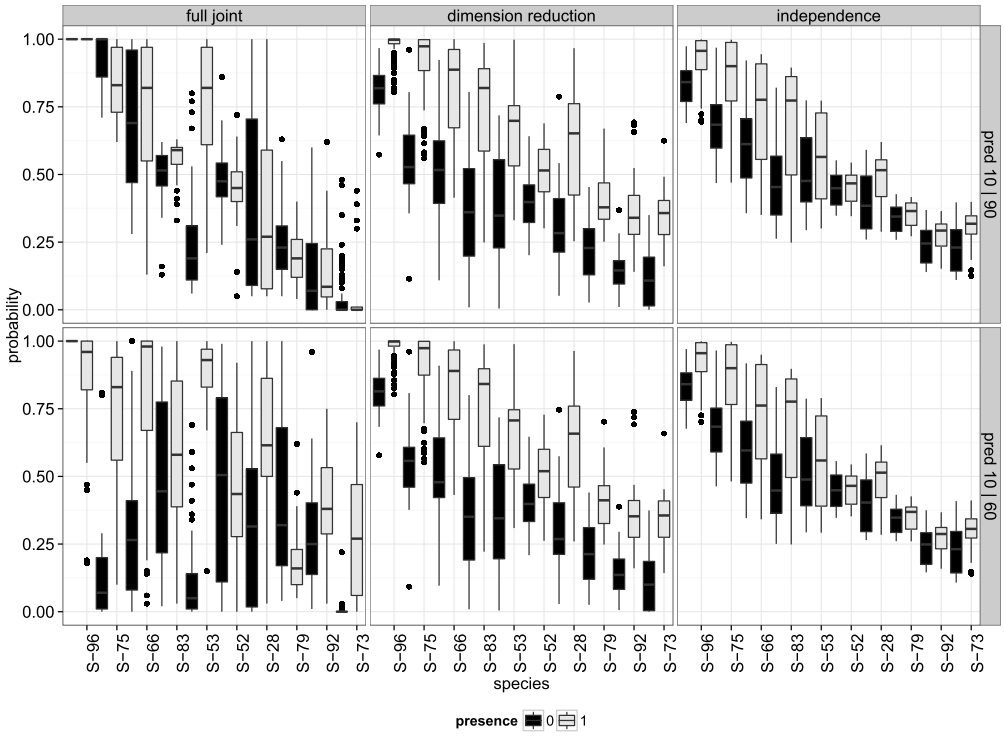


Figure 6: Binary simulation comparison of the predicted conditional probabilities of presence for 10 species given the observed presence status of 90 and 60 species, respectively.

dimension reduction approach, by only requiring 301 parameters, identifies an underlying lower dimensional structure, which, in this scenario, yields outstanding recovery of the correlation matrix. The bottom row in the figure compares the fitted regression coefficients to the true ones (means and 95% credible intervals). Here we observe that all three approaches do a reasonably good job, with slightly more accuracy using either the dimension reduced or the independence approach.

In the simulation exercise considered, our method, in addition to fairing well for estimation, outperforms the joint unreduced and independent models in terms of prediction. To assess out-of-sample conditional prediction, for each of 10 species, we obtained the probability of presence at sites where the species was in fact present, and the probability of presence when the species was in fact absent. The conditioning is on the presence–absence status of 90 (top row) and 60 (bottom row) species (Figure 6), respectively, for each of the three approaches considered. The 10 species being predicted are sorted from most to least prevalent, thus the decreasing trend observed in the probabilities.

For the simulated dataset considered, the joint approach without reduction struggles when attempting to estimate this many parameters for the number of samples available.

For some species, it clearly separates the probabilities of presence at occupied and unoccupied plots, but for other species the distributions of these probabilities appear to be practically indistinguishable. This behavior seems to be affected by the number of species on which the conditioning takes place for the joint unreduced model (compare the top and bottom plots on the left column in Figure 6). With so many parameters and the available sample size, both estimation and prediction appear to breakdown without doing reduction. On the other hand, while prediction using the independence model roughly captures the trend dictated by prevalence for each species, the probabilities of presence at occupied and unoccupied plots (unaffected by conditioning when assuming independence) barely separate from each other. In contrast, our reduced dimension approach assigns visibly higher probabilities to plot/species combinations where the species was in fact present than to those plot/species combinations where the species was absent.

6 Forest Inventory Data Analysis

We apply our methodology to hectare scale plots obtained by aggregating FIA data in covariate space, since, as clarified in Section 2, FIA plots are too small to allow for meaningful fitting or interpretation. The model yields predictions in covariate space and enables the analysis of species response to changes in the environment. The aggregated data consists of presence–absence data in covariate space for 112 species on 1200 plots. Of these, 1100 are used in model fitting the model, whereas the remaining 100 were held out for out-of-sample prediction and subsequent validation. Also, we only consider temperature and deficit as predictors. In this analysis, we compared the results obtained for the joint dimension reduced approach against the independence model. We focus on comparing (i) goodness of fit, and (ii) predicted species richness.

To assess model goodness-of-fit and compare the independent approach against reduced dimension joint models for $r = 3, 5, 10, 15, 30, 50$, we obtained the median and 95% credible sets for $\overline{TR}$ (see Figure 7), as defined in Section 5.2. As expected, the joint approach with dimension reduction yields a remarkably better fit than the independent model for all values of r considered. For the choices of r considered, the median $\overline{TR}$ values were all similar. We present the analysis for $r = 5$ since this resulted in the largest median $\overline{TR}$. Before moving forward with the analysis, we look at the posterior distribution of the number of unique clusters, $\mathbb{K}$, to establish if the model is well determined. With $r = 5$, the 95% credible interval for $\mathbb{K}$ lies between 42 and 51 with a median value of 47, indicating that we should not run into indeterminacy issues by fixing the $\mathbf{A}$ matrix to have $r = 5$ columns.

Tjur's R^2 can additionally be used to assess out-of-sample predictive performance making use of the dependence between species. In particular, by considering a hold-out sample of plots (in covariate space in this application), one may obtain at each holdout plot the posterior predictive conditional expectations for species j given the presence/absence status of species l. With these, we may calculate the conditional Tjur R^2, which we denote by $TR_{j|Y_l=0}$ and $TR_{j|Y_l=1}$, if we condition on species l being absent or present, respectively. Being able to condition in this fashion can enhance the quality of the predictions when species j and l show strong dependence.

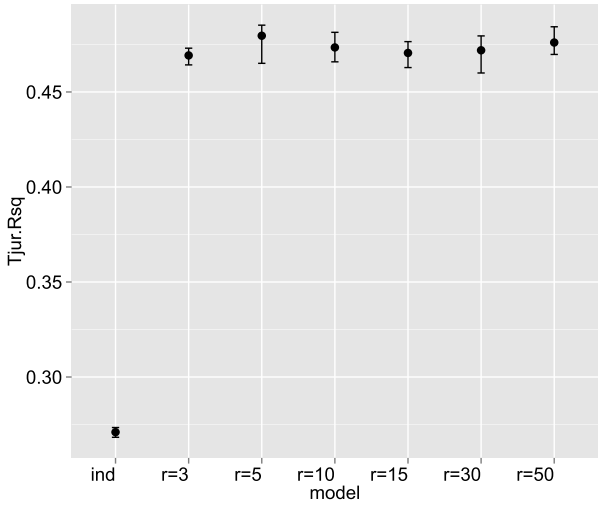


Figure 7: Goodness of fit measured by the posterior median and 95% credible sets of the Tjur's R^2 coefficient averaged over all species ($\overline{TR}$), comparing the independent model to the dimension reduced joint models with $r = 3, 5, 10, 15, 30$ and 50.

		NYSY		$TR_{\text{NYSY}\mid Y_{\text{FAGR}}}$	
		0	1	Joint ($r = 5$)	Independent
FAGR	0	$n_{00} = 22$	$n_{01} = 12$	0.7515	0.5633
	1	$n_{10} = 25$	$n_{11} = 41$	0.5085	0.1945

Table 1: Tjur R for NYSY conditional on FAGR in a holdout sample at 100 plots.

We illustrate this alternative use for Tjur's R^2 at 100 holdout plots by conditioning on the presence–absence state of *Fagus grandifolia* (FAGR) – present in 66 plots and absent in the remaining 34 – and obtain the posterior probability of presence for *Nyssa sylvatica* (NYSY) at each pseudo-plot. Using the probabilities of presence, we calculate $TR_{\text{NYSY}|Y_{\text{FAGR}}=1}$ and $TR_{\text{NYSY}|Y_{\text{FAGR}}=0}$ under both the joint model with $r = 5$ and the stacked independent models (Table 1). The posterior 95% credible interval for the correlation parameter (on the latent scale) between FAGR and NYSY lies between 0.2957 and 0.4163, which is relatively high. As such, we expect that conditioning on the presence–absence state of FAGR should improve the quality of the predictions for NYSY. As observed in Table 1, the improved predicted ability of the joint model is evident, when FAGR is either absent or present.

Using the selected model, among the 6,216 correlation parameters in $\mathbf{A}\mathbf{A}' + \sigma_\varepsilon^2 \mathrm{I}$, we found that 72.4% of the 95% credible intervals excluded 0 and were positive, 12% excluded 0 and were negative. For the 95% credible intervals of the 112 temperature

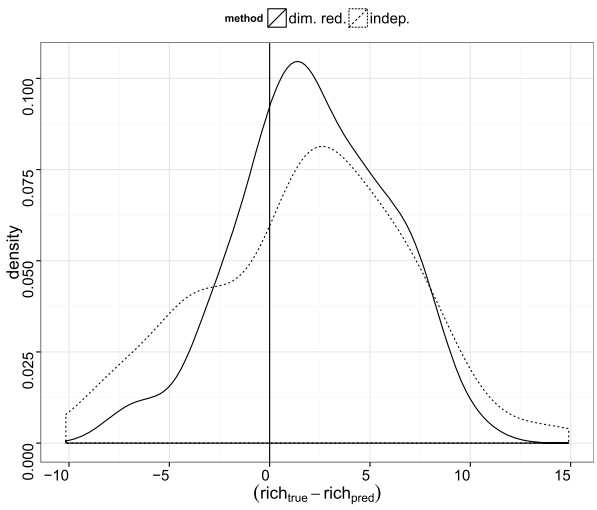


Figure 8: Posterior distribution for the mean difference between the true and predicted species richness in 100 out-of-sample plots.

coefficients, 69% were positive and excluded 0 while 26.8% were negative and excluded 0. For deficit, 25.9% of the credible sets excluded 0 and were positive while 40.2% did not contain 0 and were negative. The signs resulting for the regression coefficients of both predictors are consistent with the environment features the different species are best suited for (see Figure 4 in Supplementary Appendix C). For instance, the temperature coefficient for *Quercus laurifolia* (QULA3) is large and positive, which is consistent with the higher temperatures found throughout its natural range, i.e., southeastern and south-central US. Conversely, *Betula papyrifera* (BEPA) is found throughout northern continental US where the temperatures are on average lower and its coefficient is large and negative (see Figure 4). This predictive approach can be applied to scenarios for climate change as a joint response across all dominant species.

To validate the claim made in the literature that predicting richness from individual models overestimates the number of species (see, for example, Guisan and Rahbek, 2011; Clark et al., 2016; Calabrese et al., 2014), we considered a hold out sample of 100 plots. At each of these plots we calculated the expected richness using the MCMC parameter draws.

With the expected richness values estimated at each plot, we calculated the difference between the true richness and the predicted one, and with these derived the posterior distribution of the average differences across all plots (Figure 8). This analysis corroborates the claim made in Guisan and Rahbek (2011). When modeling the species distributions jointly, the mean differences between the true and predicted richness cluster more tightly about 0 than when using the independence model. The latter specification concentrates more mass on large negative values, confirming that it tends to overpredict richness.

7 Summary and Future Work

We have extended recent activity in joint species distribution modeling. In particular, we have noted that, with many species, perhaps hundreds or thousands, fitting explicit joint species distribution models over several plots will become computationally infeasible. We have offered a first stage latent multivariate normal modeling approach, incorporating a dimension reduction strategy to capture interspecies residual dependence, using Dirichlet processes in a novel way that scales linearly in the number of species. We have shown that this approximation enables estimation of dependence structure, prediction, and clustering of species. The versatility of the framework developed, and the ease with which it can be incorporated into strategies with other types of responses, make it a powerful tool to tackle novel problems. Among others, this approach can be used to address the challenges of microbiome data analysis (where presence–absence of hundreds or thousands of OTUs – Operational Taxonomic Units – is jointly observed), or the massive species distribution datasets currently being collected within the continental scale National Ecological Observatory Network (NEON) project.

Our approach can be applied to other data settings, particularly abundance which can be measured in various ways, e.g., counts, ordinal abundances, proportions of coverage, biomass, basal area, etc., each requiring a different first stage transformation from the latent multivariate normal. To make the methodology readily available for practitioners, it has been fully implemented in the gjam2.0 package (Clark et al., 2016) in R for binary, count, and non-negative continuous responses.

At fine spatial scales, accounting for spatial dependence may be appropriate. A straightforward alternative could be to build into the mean component some function of the spatial coordinates – in the simplest case, by including the linear terms and interaction between the coordinates. However, other more sophisticated strategies, where the dependence is modeled directly, merit exploration. In these we will have both within plot and between plot dependence arising across many species and many plots. Lastly, utilizing climate scenarios, we can explore dynamic models to attempt to forecast the evolution of joint species distributions in response to changing climate.

Supplementary Material

Appendices: Joint Species distribution modeling: dimension reduction using Dirichlet processes (DOI: 10.1214/16-BA1031SUPP; .pdf).

References

Aguilar, O. and West, M. (2010). "Bayesian Dynamic Factor Models and Portfolio Allocation." *Journal of Business & Economic Statistics*, 18(3): 338–357. papers2://publication/doi/10.1080/07350015.2000.10524875. 956

Arbel, J., King, C. K., Raymond, B., Winsley, T., and Mengersen, K. L. (2015). "Application of a Bayesian nonparametric model to derive toxicity estimates based on the

response of Antarctic microbial communities to fuel-contaminated soil." *Ecology and Evolution*, 5(13): 2633–2645. 942

Artemiou, A. and Li, B. (2009). "On principal components and regression: A statistical explanation of a natural phenomenon." *Statistica Sinica*, 19(4): 1557. MR2589197. 941

Artemiou, A. and Li, B. (2013). "Predictive power of principal components for single-index model and sufficient dimension reduction." *Journal of Multivariate Analysis*, 119: 176–184. MR3061422. doi: http://dx.doi.org/10.1016/j.jmva.2013.04.015. 941

Austin, M. and Meyers, J. (1996). "Current approaches to modelling the environmental niche of eucalypts: implication for management of forest biodiversity." *Forest Ecology and Management*, 85(1): 95–106. 940

Bechtold, W. A. and Patterson, P. L. (2005). "The enhanced forest inventory and analysis program: national sampling design and estimation procedures." Technical report, US Department of Agriculture Forest Service, Southern Research Station Asheville, North Carolina. 943

Bhattacharya, A. and Dunson, D. B. (2011). "Sparse Bayesian infinite factor models." *Biometrika*, 98(2): 291–306. MR2806429. doi: http://dx.doi.org/10.1093/biomet/asr013. 942, 947

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies." *Journal of the ACM (JACM)*, 57(2): 7. MR2606082. doi: http://dx.doi.org/10.1145/1667053.1667056. 942

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." *The Journal of Machine Learning Research*, 3: 993–1022. 942

Botkin, D. B., Saxe, H., Araujo, M. B., Betts, R., Bradshaw, R. H., Cedhagen, T., Chesson, P., Dawson, T. P., Etterson, J. R., Faith, D. P., et al. (2007). "Forecasting the effects of global warming on biodiversity." *Bioscience*, 57(3): 227–236. 940

Bush, C. A. and MacEachern, S. N. (1996). "A semiparametric Bayesian model for randomised block designs." *Biometrika*, 83(2): 275–285. 945

Calabrese, J. M., Certain, G., Kraan, C., and Dormann, C. F. (2014). "Stacking species distribution models and adjusting bias by linking them to macroecological models." *Global Ecology and Biogeography*, 23(1): 99–112. 940, 949, 961

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge University Press. 941

Chakraborty, A., Gelfand, A. E., Wilson, A. M., Latimer, A. M., and Silander, J. A. (2011). "Point pattern modelling for degraded presence-only data over large regions." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(5): 757–776. MR2844854. doi: http://dx.doi.org/10.1111/j.1467-9876.2011.00769.x. 940

Chib, S. (1998). "Analysis of multivariate probit models." *Biometrika*, 85(2): 347–361. http://biomet.oupjournals.org/cgi/doi/10.1093/biomet/85.2.347 947

Chung, Y. and Dunson, D. B. (2011). "The local Dirichlet process." *Annals of the Institute of Statistical Mathematics*, 63(1): 59–80. MR2748934. doi: http://dx.doi.org/10.1007/s10463-008-0218-9. 942

Clark, J. S., Bell, D. M., Hersh, M. H., Kwit, M. C., Moran, E., Salk, C., Stine, A., Valle, D., and Zhu, K. (2011). "Individual-scale variation, species-scale differences: inference needed to understand diversity." *Ecology Letters*, 14(12): 1273–1287. 940

Clark, J. S., Gelfand, A. E., Woodall, C. W., and Zhu, K. (2014). "More than the sum of the parts: forest climate response from joint species distribution models." *Ecological Applications*, 24(5): 990–999. 940, 941

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P., and Zhange, S. (2016). "Median-zero, multivariate, multifarious data: generalized joint attribute modeling for biodiversity analysis." *Ecological Monographs, in press*. 940, 941, 947, 949, 956, 961, 962

Dormann, C. F., Schymanski, S. J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B., et al. (2012). "Correlation and process in species distribution models: bridging a dichotomy." *Journal of Biogeography*, 39(12): 2119–2131. 940

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2): 307–323. MR2521586. doi: http://dx.doi.org/10.1093/biomet/asn012. 942

Elith, J. and Leathwick, J. R. (2009). "Species distribution models: ecological explanation and prediction across space and time." *Annual Review of Ecology, Evolution, and Systematics*, 40(1): 677. 940

Escobar, M. D. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89(425): 268–277. MR1266299. 945

Escobar, M. D. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 945

Fischlin, A., Midgley, G. F., Hughs, L., Price, J., Leemans, R., Gopal, B., Turley, C., Rounsevell, M., Dube, P., Tarazona, J., et al. (2007). "Ecosystems, their properties, goods and services." 940

Gelfand, A. E., Schmidt, A. M., Wu, S., Silander, J. A., Latimer, A., and Rebelo, A. G. (2005). "Modelling species diversity through species level hierarchical modelling." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1): 1–20. MR2134594. doi: http://dx.doi.org/10.1111/j.1467-9876.2005.00466.x. 940

Gelfand, A. E., Silander, J. A., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G., Holder, M., et al. (2006). "Explaining species distribution patterns through hierarchical modeling." *Bayesian Analysis*, 1(1): 41–92. MR2227362. doi: http://dx.doi.org/10.1214/06-BA102. 940

Geweke, J. F. and Singleton, K. J. (1980). "Interpreting the likelihood ratio statistic in factor models when sample size is small." *Journal of the American Statistical Association*, 75(369): 133–137. 956

Ghahramani, Z. and Griffiths, T. L. (2005). "Infinite latent feature models and the Indian buffet process." In *Advances in Neural Information Processing Systems*, 475–482. 942

Guisan, A. and Rahbek, C. (2011). "SESAM–a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages." *Journal of Biogeography*, 38(8): 1433–1444. 940, 961

Guisan, A. and Thuiller, W. (2005). "Predicting species distribution: offering more than simple habitat models." *Ecology Letters*, 8(9): 993–1009. 940

Huang, A. and Wand, M. P. (2013). "Simple marginally noninformative prior distributions for covariance matrices." *Bayesian Analysis*, 8(2): 439–452. MR3066948. doi: http://dx.doi.org/10.1214/13-BA815. 947

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. http://www.tandfonline.com/doi/abs/10.1198/016214501750332758. MR1952729. doi: http://dx.doi.org/10.1198/016214501750332758. 945, 946

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87(2): 371–390. MR1782485. doi: http://dx.doi.org/10.1093/biomet/87.2.371. 946, 947

Iverson, L. R., Prasad, A. M., Matthews, S. N., and Peters, M. (2008). "Estimating potential habitat for 134 eastern US tree species under six climate scenarios." *Forest Ecology and Management*, 254(3): 390–406. 940

Latimer, A., Banerjee, S., Sang Jr, H., Mosher, E., and Silander Jr, J. (2009). "Hierarchical models facilitate spatial analysis of large data sets: a case study on invasive plant species in the northeastern United States." *Ecology Letters*, 12(2): 144–154. 940

Latimer, A. M., Wu, S., Gelfand, A. E., and Silander Jr, J. A. (2006). "Building statistical models to analyze species distributions." *Ecological Applications*, 16(1): 33–50. 940

Lawrence, E., Bingham, D., Liu, C., and Nair, V. N. (2008). "Bayesian inference for multivariate ordinal data using parameter expansion." *Technometrics*, 50(2): 182–191. http://www.tandfonline.com/doi/abs/10.1198/004017008000000064. MR2439877. doi: http://dx.doi.org/10.1198/004017008000000064. 947

Leathwick, J. (2002). "Intra-generic competition among Nothofagus in New Zealand's primary indigenous forests." *Biodiversity & Conservation*, 11(12): 2177–2187. 940

Li, K.-C. (1991). "Sliced inverse regression for dimension reduction." *Journal of the American Statistical Association*, 86(414): 316–327. MR1137117. 941

Liu, J. and Wu, Y. (1999). "Parameter expansion for data augmentation." *Journal of the American Statistical Association*, 37–41. http://amstat.tandfonline.com/doi/full/10.1080/01621459.1999.10473879. MR1731488. doi: http://dx.doi.org/10.2307/2669940. 947

Lopes, H. F. and West, M. (2004). "Bayesian model assessment in factor analysis." 14: 41–67. MR2036762. 955

MacEachern, S. N. (1994). "Estimating normal means with a conjugate style Dirichlet process prior." *Communications in Statistics – Simulation and Computation*, 23(3): 727–741. MR1293996. doi: http://dx.doi.org/10.1080/03610919408813196. 942, 945

MacKenzie, D. I. and Royle, J. A. (2005). "Designing occupancy studies: general advice and allocating survey effort." *Journal of Applied Ecology*, 42(6): 1105–1114. 940

McMahon, S. M., Harrison, S. P., Armbruster, W. S., Bartlein, P. J., Beale, C. M., Edwards, M. E., Kattge, J., Midgley, G., Morin, X., and Prentice, I. C. (2011). "Improving assessment and modelling of climate change impacts on global terrestrial biodiversity." *Trends in Ecology & Evolution*, 26(5): 249–259. 940

Midgley, G., Hannah, L., Millar, D., Rutherford, M., and Powrie, L. (2002). "Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot." *Global Ecology and Biogeography*, 11(6): 445–451. 940

Naik, P. and Tsai, C.-L. (2000). "Partial least squares estimator for single-index models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4): 763–771. MR1796290. doi: http://dx.doi.org/10.1111/1467-9868.00262. 941

Neal, R. M. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: http://dx.doi.org/10.2307/1390653. 945

Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2015). "Using latent variable models to identify large networks of species-to-species associations at different spatial scales." *Methods in Ecology and Evolution*, 7: 549–555. 941

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). "Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions." *Ecology*, 91(9): 2514–2521. 940

Ovaskainen, O. and Soininen, J. (2011). "Making more out of sparse data: hierarchical modeling of species communities." *Ecology*, 92(2): 289–295. 940

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: http://dx.doi.org/10.1093/biomet/asm086. 945

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM)." *Methods in Ecology and Evolution*, 5(5): 397–406. 940

Schliep, E. M., Gelfand, A. E., Clark, J. S., and Tomasek, B. J. (2016). "Biomass prediction using density dependent diameter distribution." *Manuscript submitted for publication*. 943

Schliep, E. M. and Hoeting, J. A. (2015). "Data augmentation and parameter expansion for independent or spatially correlated ordinal data." *Computational Statistics & Data Analysis*, 90: 1–14. MR3354825. doi: http://dx.doi.org/10.1016/j.csda.2015.03.020. 947

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 945

Smith, W. B., Miles, P. D., Vissage, J. S., Pugh, S. A., et al. (2009). "Forest resources of the United States. 2007." *General Technical Report-USDA Forest Service*, WO-78, United States Department of Agriculture, Forest Service. 943

Taylor-Rodríguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S., and Gelfand, A. E. (2016). "Appendices: Joint Species distribution modeling: dimension reduction using Dirichlet processes." *Bayesian Analysis*. doi: http://dx.doi.org/10.1214/16-BA1031SUPP. 947, 954

Thorson, J. T., Scheuerell, M. D., Shelton, A. O., See, K. E., Skaug, H. J., and Kristensen, K. (2015). "Spatial factor analysis: a new tool for estimating joint species distributions and correlations in species range." *Methods in Ecology and Evolution*. 940

Thuiller, W. (2003). "BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change." *Global Change Biology*, 9(10): 1353–1362. 940

Thuiller, W., Lavergne, S., Roquet, C., Boulangeat, I., Lafourcade, B., and Araujo, M. B. (2011). "Consequences of climate change on the tree of life in Europe." *Nature*, 470(7335): 531–534. 940

Tjur, T. (2009). "Coefficients of determination in logistic regression models? A new proposal: The coefficient of discrimination." *The American Statistician*, 63(4): 366–372. MR2751755. doi: http://dx.doi.org/10.1198/tast.2009.08210. 956

Woudenberg, S. W., Conkling, B. L., O'Connell, B. M., LaPoint, E. B., Turner, J. A., Waddell, K. L., et al. (2010). "The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2." 943

Acknowledgments

The authors thank Nikolay Bliznyuk for sharing computational resources through the HyperGator at the University of Florida, Bradley Tomasek for building the FIA dataset used in the analysis, Andrew Womack for providing useful code, and Andres Felipe Barrientos for valuable discussions. The work of the first and second authors were partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. The work of the third, fourth, and fifth authors was supported in part by NSF-CDI-0940671 and NSF-EF-1137364. The first, fourth and fifth authors were also partially supported by NSF-EF-1550911.