Statistica Sinica **29** (2019), 1127-1154
doi:https://doi.org/10.5705/ss.202017.0482

SPATIAL JOINT SPECIES DISTRIBUTION MODELING USING DIRICHLET PROCESSES

Shinichiro Shirota[1], Alan E. Gelfand[2] and Sudipto Banerjee[1]

[1] *University of California, Los Angeles and* [2] *Duke University*

Abstract: Species distribution models usually attempt to explain the presence–absence or abundance of a species at a site in terms of the environmental features (so-called abiotic features) present at the site. Historically, such models have considered species individually. However, it is well established that species interact to influence the presence–absence and abundance (envisioned as biotic factors). As a result, recently joint species distribution models with various types of responses, such as presence–absence, continuous, and ordinal data have attracted a significant amount of interest. Such models incorporate the dependence between species' responses as a proxy for interaction. We address the accommodation of such modeling in the context of a large number of species (e.g., order 10^2) across sites numbering in the order of 10^2 or 10^3 when, in practice, only a few species are found at any observed site. To do so, we adopt a dimension-reduction approach. The novelty of our approach is that we add spatial dependence. That is, we consider a collection of sites over a relatively small spatial region. As such, we anticipate that the species distribution at a given site will be similar to that at a nearby site. Specifically, we handle dimension reduction using Dirichlet processes, which enables the clustering of species, and add spatial dependence across sites using Gaussian processes. We use simulated data and a plant communities data set for the Cape Floristic Region (CFR) of South Africa to demonstrate our approach. The latter consists of presence–absence measurements for 639 tree species at 662 locations. These two examples demonstrate the improved predictive performance of our method using the aforementioned specification.

Key words and phrases: Dimension reduction; Gaussian processes; high-dimensional covariance matrix; spatial factor model; species dependence.

1. Introduction

Understanding the distribution and abundance of species is a primary goal of ecological research. In this regard, species distribution models (SDMs) are used to investigate the regressors that affect the presence–absence and abundance

of species. They are also used to examine the prevalence of a species, predict biodiversity and richness, quantify species turnover, and assess the response to climate change (Midgley et al. (2002); Guisan and Thuiller (2005); Gelfand et al. (2006); Iverson et al. (2008); Botkin et al. (2007); McMahon et al. (2011); Thuiller et al. (2011)). These models are used to infer a species range in either a geographic space or a 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)). A further key objective is interpolation, to predict species response at locations that have not been sampled.

SDMs are most commonly fitted to presence–absence data (binary) or abundance data (counts, ordinal classfications, or proportions). Occasionally, continuous responses are used, for example, in biomass research (Dormann et al. (2012)). Predictions of species over space can be accommodated using a spatially explicit specification (Gelfand et al. (2005, 2006); Latimer et al. (2006)).

Historically, SDMs have considered species individually (Thuiller (2003); Latimer et al. (2006); Elith and Leathwick (2009); Chakraborty et al. (2011)). To make predictions at a community scale, independent models for individual species are aggregated or stacked (Calabrese et al. (2014)). However, it is well established that species interact to influence presence–absence and abundance. As a result, individual-level models tend to predict too many species per location (Guisan and Rahbek (2011)), as well as providing other misleading findings (see Clark et al. (2014), for examples). Modeling species individually does not allow underlying joint relationships to be captured (Clark et al. (2011); Ovaskainen and Soininen (2011)). That is, the problem can be viewed as the omission of the residual dependence between species.

Joint species distribution models (JSDMs) that incorporate species dependence include applications to presence–absence research (Pollock et al. (2014); Ovaskainen, Hottola and Siitonen (2010); Ovaskainen and Soininen (2011)), continuous or discrete abundance (Latimer et al. (2009); Thorson et al. (2015)), abundance with a large number of zeros (Clark et al. (2014)), and discrete, ordinal, and compositional data (Clark et al. (2017)). JSDMs jointly characterize the presence and/or abundance of multiple species at a set of locations, partitioning the drivers into two components. The first is associated with environmental suitability, and the second accounts for species dependence through the residuals, that is, adjusted for the environment. Such models incorporate the dependence between species' responses as a proxy for a formal specification of interaction.

JSDMs enhance our understanding of the distributions of species; however,

their applicability has been limited owing to computational challenges when there is a large number of species. To appreciate the potential challenge with presence–absence (binary) responses and S species, we construct 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 of the order of hundreds to thousands, rendering a contingency table analysis unfeasible. Therefore, there is a need for strategies that fit joint models in a computationally tractable manner.

To deal with these data challenges, we adopt dimension-reduction techniques, working within the Bayesian factor model setting (West (2003); Lopes and West (2004)). For instance, in the spatial case, Ren and Banerjee (2013) introduce spatial dependence into the factors using Gaussian predictive process models (Banerjee et al. (2008)). Taylor-Rodríguez et al. (2017) also consider dimension reduction within a factor modeling framework. They generate each row of the factor loading matrix from Dirichlet process realizations to enable common labels, that is, clustering across the species. They assume independent factors because their plot locations are not close to each other. Their focus is to jointly explain species' presence at plots, rather than predict the distribution at new locations. We add spatial dependence to the explanatory model to enable joint predictions at arbitrary locations over the study region.

In this regard, Thorson et al. (2015) implement a spatial factor analysis for species distribution, where they fix the factor loading matrix. Ovaskainen et al. (2016) implement the multiplicative Gamma shrinkage prior proposed by Bhattacharya and Dunson (2011) for the factor loading matrix, and introduce spatial dependence into the factors. This work is the most comparable to our approach in the sense that both are specified through hierarchical models. However, our specification directly models species dependence at the first (data) stage, whereas Ovaskainen et al. (2016) incorporate dependence in the second (probabilities) stage. We clarify this below. Furthermore, our approach enables the data to inform us about the clustering among species.

We formulate our model in the context of a large number of species (e.g., order 10^2) across a large number of sites (e.g., order 10^2 or 10^3) when, in practice, only a few species are found at any observed site. The novelty of our approach is the addition of spatial dependence. That is, we have a collection of sites over a relatively small spatial region. Thus, we anticipate that the species distribution at a given site will be similar to that at a nearby site. As noted above, we

adopt a dimension-reduction approach, following the model proposed by Taylor-Rodríguez et al. (2017). Specifically, we handle the dimension reduction using Dirichlet processes, which enables the joint labeling for species (i.e., clustering), to which we add spatial dependence using Gaussian processes (GPs).

We use both simulated data and a plant communities data set for the Cape Floristic Region (CFR) of South Africa to demonstrate our approach. The simulation study serves as a proof of concept for both continuous and binary response data. The CFR data set consists of presence–absence measurements for 639 tree species at 662 locations. These two examples demonstrate improved predictive performance of our method using the aforementioned specification.

The organization of the remainded of this paper is as follows. Section 2 introduces our motivating data and modeling strategy, that is, spatial joint species distribution models with Dirichlet processes. Section 3 provides the adaptation to binary responses, along with a discussion on identifying parameters specifically for probit models. In Section 4, we develop Bayesian inference for our model and our model comparison strategy. In Section 5, we investigate the proposed models using simulation studies for continuous and binary responses, and in Section 6 we analyze the presence–absence data from the CFR. Finally, Section 7 concludes the paper and discusses potential future work.

2. Spatial Factor Modeling with Dirichlet Processes

2.1. A motivating data example

Our data are extracted from a large database studying the distribution of plants in the Cape Floristic Region (CFR) of South Africa (Takhtajan (1986)). The CFR is one of the six floral kingdoms in the world and is located in the southwestern part of South Africa. Alhough geographically relatively small, it is extremely diverse (9,000+ species) and highly endemic (70% occur only in the CFR (Rebelo (2001)). There are more than 40,000 sites with recorded sampling within the CFR. The database from which our data set was extracted consists of more than 1,400 plots with more than 2,800 species, spanning six regions. Our data are from one of these regions and exhibit high spatial clustering, with $n = 662$ plots and $S = 639$ species. The responses are binary: presence–absence for each species' and plot (location).

The left panel of Figure 1 shows the 662 locations in the CFR data, and the right panel shows the distribution of nine selected species: 1) *Aridaria noctiflora* (ArNo); 2) *Asparagus capensis* (AsCa); 3) *Chrysocoma ciliata* (ChCi); 4)

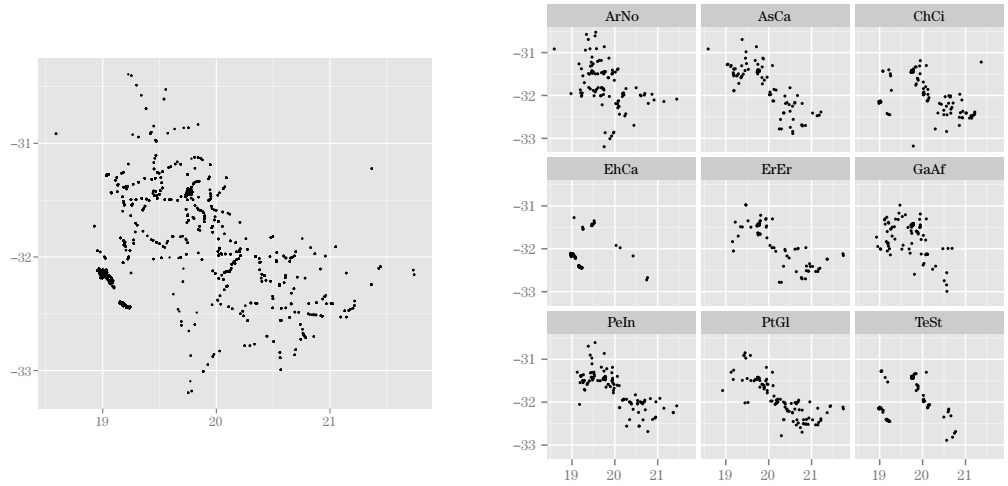

Figure 1. 662 locations in the CFR (left), and the distribution of the presence of nine selected species.

Ehrharta calycina (EhCa); 5) *Eriocephalus ericoides* (ErEr); 6) *Galenia africana* (GaAf); 7) *Pentzia incana* (PeIn); 8) *Pteronia glomerata* (PtGl); and 9) *Tenaxia stricta* (TeSt). These species are selected because they are observed at more than 100 locations (plots). Some species reveal strong spatial clustering (e.g., EhCa and TeSt).

The total number of binary responses is $n \times S = 662 \times 639 = 423,018$. The overall number of presences is 6,980, or 1.65% of the total number of binary responses. This emphasizes that, although we have many species in our data set, only a few are present on any given plot. Among the $S = 639$ species, 351 are observed in at most five locations. We discard these species and retain $S = 288$ species across the 662 locations for model fitting.

2.2. Our model

Let $\mathcal{D} \subset \mathbb{R}^2$ be a bounded study region, $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ be a set of plot locations, where $\boldsymbol{s}_i \in \mathcal{D}$ for $i = 1, \ldots, n$, and $\boldsymbol{U}_i := \boldsymbol{U}(\boldsymbol{s}_i) \in \mathbb{R}^S$ be an $S \times 1$ latent vector of continuous variables at location $\boldsymbol{s}_i$. Under independence for the locations, the model for $\boldsymbol{U}_i$ is specified as

$$\boldsymbol{U}_i = \mathbf{B}\boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \overset{iid}{\sim} \mathcal{N}_S(\mathbf{0}, \boldsymbol{\Sigma}), \quad \text{for} \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{B}$ is an $S \times p$ coefficient matrix, $\boldsymbol{x}_i$ is a $p \times 1$ covariate vector at location $\boldsymbol{s}_i$, and $\boldsymbol{\Sigma}$ is an $S \times S$ covariance matrix for species. This model has $\mathcal{O}(S^2)$

parameters: $S(S+1)/2$ parameters from $\mathbf{\Sigma}$, and pS parameters from $\mathbf{B}$. For example, for $S = 300$ species and $p = 3$ covariates, the model contains 46,050 parameters.

Taylor-Rodríguez et al. (2017) propose a dimension-reduction approximation to $\mathbf{\Sigma}$ that allows the number of parameters to grow linearly in S. They approximate $\mathbf{\Sigma}$ with $\mathbf{\Sigma}^* = \mathbf{\Lambda}\mathbf{\Lambda}^T + \sigma_\epsilon^2 \mathbf{I}_S$, and replace the above model with

$$\boldsymbol{U}_i = \mathbf{B}\boldsymbol{x}_i + \mathbf{\Lambda}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i, \quad \boldsymbol{\epsilon}_i \sim \mathcal{N}_S(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_S), \quad \text{for} \quad i = 1, \ldots, n, \tag{2.2}$$

where the random vectors $\boldsymbol{w}_i$ are independent and identically distributed (i.i.d). with $\boldsymbol{w}_i \sim \mathcal{N}_r(\mathbf{0}, \mathbf{I}_r)$ and $\mathbf{\Lambda}$ is an $S \times r$ matrix with $r \ll S$. Now, $\mathbf{\Sigma}^*$ has only $Sr+1$ parameters, and the estimation problem of $\mathcal{O}(S^2)$ parameters is reduced to that of $\mathcal{O}(S)$ parameters. We refer to this specification as the dimension-reduced nonspatial model.

Although $\mathbf{\Lambda}\mathbf{\Lambda}^T$ has rank r, including the nugget variance, $\sigma_\epsilon^2\mathbf{I}$, ensures that $\mathbf{\Sigma}^*$ is nonsingular. Taylor-Rodríguez et al. (2017) further propose sampling the rows of $\mathbf{\Lambda}$ from a Dirichlet process mixture (DPM) using a stick-breaking representation (Sethuraman (1994)). This representation is attractive within a Gibbs sampling setting (see, e.g., Escobar (1994); Escobar and West (1995); MacEachern (1994); Bush and MacEachern (1996); Neal (2000)) owing to a Pólya urn scheme representation that enables a straightforward simulation from the needed full conditional distributions.

Under the stick-breaking construction, we say the random distribution, G, follows a DP with base measure H and precision parameter α, $G \sim DP(\alpha H)$, if $G(\cdot) = \sum_{l=1}^{\infty} p_l \delta_{\theta_l}(\cdot)$. Here, $p_1 = \xi_1$, $p_l = \xi_l \prod_{h=1}^{l-1}(1-\xi_h)$ $(h \geq 2)$, with i.i.d. $\xi_l \sim \text{Beta}(1, \alpha)$, 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, this approach yields ties when realizations are drawn; the Pólya urn scheme representation draws from an atomic distribution with point masses at the already seen values, with the remaining mass on H. Thus, the DP enables us to perform model clustering. We use this feature to allow some rows of $\mathbf{\Lambda}$ to be common, which corresponds to clustering the species in terms of their residual dependence behavior, as we clarify below.

According to (2.2), the $\boldsymbol{U}_i$ are conditionally independent, given $\boldsymbol{B}$ and $\mathbf{\Lambda}$; that is, the $\boldsymbol{w}_i$ are independent across locations. However, because the plot locations in our data set are relatively close together, we introduce spatial dependence into $\boldsymbol{w}_i$, which enables us to improve the prediction for new plot locations in the study region.

To provide the hierarchical formulation for this model, let $\mathbf{Z} = [\boldsymbol{Z}_1 : \cdots :$

$\boldsymbol{Z}_N]^T$ (with $\boldsymbol{Z}_j \sim H$) denote an $N \times r$ matrix the rows of which make up all potential atoms. In this setup, we need a vector of grouping labels $\boldsymbol{k} = (k_1, \ldots, k_S)$ $(1 \le k_l \le N)$ such that the l-th row of $\boldsymbol{\Lambda}$ is equal to $\boldsymbol{Z}_{k_l}$. Note that $\boldsymbol{\Lambda}$ can be represented by $\boldsymbol{\Lambda} = \mathbf{Q}(\boldsymbol{k})\mathbf{Z}$, where $\mathbf{Q}(\boldsymbol{k}) = [\boldsymbol{e}_{k_1} : \cdots : \boldsymbol{e}_{k_S}]^T$ is $S \times N$, with $\boldsymbol{e}_{k_l}$ denoting the N-dimensional vector with a one in position k_l, and zeros elsewhere. Letting $\mathbf{W} = [\boldsymbol{w}_1 : \cdots : \boldsymbol{w}_n]^T$ be the $n \times r$ spatial factor matrix, our approximate model is

$$
\begin{aligned}
\boldsymbol{U}_i|\boldsymbol{k}, \mathbf{Z}, \boldsymbol{w}_i, \mathbf{B}, \sigma_\epsilon^2 &\sim \mathcal{N}_S(\mathbf{B}\boldsymbol{x}_i + \mathbf{Q}(\boldsymbol{k})\mathbf{Z}\boldsymbol{w}_i, \sigma_\epsilon^2 \mathbf{I}_S), \quad \text{for} \quad i = 1, \ldots, n, \\
\mathbf{W}^{(h)} &\sim \mathcal{N}_n(\mathbf{0}, \mathbf{C}_\phi), \quad \text{for} \quad h = 1, \ldots, r, \\
k_l|\boldsymbol{p} &\sim \sum_{j=1}^{N} p_j \delta_j(k_l), \quad \text{for} \quad l = 1, \ldots, S, \\
\boldsymbol{Z}_j|\mathbf{D}_{\boldsymbol{Z}} &\sim \mathcal{N}_r(\mathbf{0}, \mathbf{D}_{\boldsymbol{Z}}), \quad \text{for} \quad j = 1, \ldots, N, \\
Z_{1,h} &> 0, \quad \text{for} \quad h = 1, \ldots, r, \\
\boldsymbol{p} &\sim \mathcal{GD}_N(\boldsymbol{a}, \boldsymbol{b}), \\
\mathbf{D}_{\boldsymbol{Z}} &\sim \mathcal{IW}\left(2 + r - 1, 4\mathrm{diag}\left(\frac{1}{\eta_1}, \ldots, \frac{1}{\eta_r}\right)\right), \\
\eta_h &\sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{10^4}\right), \quad \text{for} \quad h = 1, \ldots, r,
\end{aligned}
\tag{2.3}
$$

where $\mathcal{GD}_N$ is an N-dimensional generalized Dirichlet distribution. In addition, $\mathbf{W}^{(h)} = (w_1^{(h)}, \ldots, w_n^{(h)})^T$ is the h-th column of $\mathbf{W}$ ($n \times 1$ vector) and is distributed as an n-variate normal vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}_\phi = [\exp(-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|)]_{i,i'=1,\ldots,n}$, that is, it is a realization of a GP with an exponential covariance function at the sites in $\mathcal{S}$. We refer to the above modeling specification as the dimension-reduced spatial model. Again, Taylor-Rodríguez et al. (2017) consider the entries in $\mathbf{W}^{(h)}$ to be independent across i (i.e., across sites), whereas we introduce spatial dependence across i through a GP for each column of $\mathbf{W}$. Furthermore, we restrict $k_1 = 1$ and all components of $\boldsymbol{Z}_1 = (Z_{1,1}, \ldots, Z_{1,r})^T$ to be positive in order to identify the covariance structure, as discussed in Ren and Banerjee (2013). We provide more detail in Section 3.1.

For prior specifications, we assume $\sigma_\epsilon^2 \sim \mathcal{IG}(a/2, b/2)$ and $\mathbf{B}_l \sim \mathcal{N}(\mathbf{0}, c\mathbf{I}_p)$ for $l = 1, \ldots, S$ where $\mathbf{B}_l$ is l-th row of $\mathbf{B}$. In practice, we suggest using a weakly informative prior specification, for example, $a = 2$ or 3, $b \le 0.1$, and $c = 100$. We assume a uniform prior for ϕ, $\phi \sim \mathcal{U}[\phi_{\min}, \phi_{\max}]$, with $\phi_{\max} = -\log(0.01)/d_{\min}$ and $\phi_{\min} = -\log(0.05)/d_{\max}$, where $d_{\max}$ and $d_{\min}$ are the minimum and maximum observed intersite distances, respectively, across all locations, following

Wang and Wall (2003). In our data sets, $d_{\max} = 3.292$ and $d_{\min}$ is set to a very small number, but within the limits of machine precision to avoid overflow. Thus, the induced *effective range* d_0, that is, the distance at which spatial correlation is negligible (falls below 0.05), is about the same as the maximum observed intersite distance (see, e.g., (Banerjee, Carlin and Gelfand (2014))).

Next, we offer a few clarifying remarks about the roles of Λ and $\mathbf{w}_h$.

Remark 1. The initial specification in (2.2) is a nonspatial nondimension-reduced model. The only model comparisons we make are between the dimension-reduced nonspatial and spatial models because both of these models have the same approximation form for the covariance, $\mathbf{\Sigma}^* = \mathbf{\Lambda}\mathbf{\Lambda}^T + \sigma_\epsilon^2 \mathbf{I}_S$. In this regard, we argue that $\mathbf{\Lambda}$ should not be location dependent. Furthermore, $\mathbf{\Lambda}\mathbf{\Lambda}^T$ is a feature of the taxonomy and, thus, should not be spatially varying.

Remark 2. We can clarify the interpretation of the clustering resulting from modeling the rows of $\mathbf{\Lambda}$ through a Dirichlet process. If we cluster the rows of $\mathbf{\Lambda}$, then we do not cluster the species by their means because each species gets its own vector of regression coefficients from $\mathbf{B}$. Instead, the residual covariance structure is clustered. If row $\mathbf{\Lambda}_l = \mathbf{\Lambda}_{l'}$, then the row entries for $U_i^{(l)}$ and $U_i^{(l')}$ in $\mathbf{\Sigma}^*$ are identical. That is, when species are clustered at an iteration of the Markov chain Monte Carlo (MCMC) fitting, they have the same dependence structure as those of all other species.

Therefore, posterior clustering is interpreted for a pair of species having a similar dependence to that of all of the other species, adjusted for the regressors. This may make a useful ecological interpretation of the clustering difficult. Alternatively, because attempting to formally model species interactions is challenging, we instead view the modeling of the residual dependence as a proxy. Then, we might attach an interpretation of similar dependence with other species as a similar interaction with other species.

Remark 3. With regard to modeling the spatial dependence structure, in principle, each species might have its own spatial range/decay parameter. However, under the dimension reduction, we can include at most $r \ll S$ decay parameters. Thus, an issue is whether incorporating a common decay parameter for the latent GPs, (i.e., a separable model) will sacrifice much compared with employing r decay parameters when r is, say, 3 to 5. The implications for the species-level spatial dependence behavior are expected to be negligible. Moreover, with r decay parameters ordered (as, e.g., in Ren and Banerjee (2013)) to obtain well-behaved MCMC, the chain may not move well over this constrained space for the

parameters. Lastly, if we have an $S \times 1$ binary vector at each location, we would not expect the data to carry much information about a set of r decay parameters.

2.3. Interpretation

Here, we provide a technical elaboration of the foregoing remarks. Given $\boldsymbol{w}_i$, the conditional expectations for the l-th and l'-th rows of $\boldsymbol{U}_i$ are

$$E[U_i^{(l)}|\boldsymbol{w}_i] = \mathbf{B}_l\boldsymbol{x}_i + \boldsymbol{\Lambda}_l\boldsymbol{w}_i \quad E[U_i^{(l')}|\boldsymbol{w}_i] = \mathbf{B}_{l'}\boldsymbol{x}_i + \boldsymbol{\Lambda}_{l'}\boldsymbol{w}_i. \tag{2.4}$$

We see that the random effect provides an additional component in the mean explanation. This is usually interpreted as capturing the effects of unmeasured/unobserved predictors at location $\boldsymbol{s}_i$. Thus, $\boldsymbol{\Lambda}_l\boldsymbol{w}_i$ and $\boldsymbol{\Lambda}_{l'}\boldsymbol{w}_i$ inform us about the residual variance adjusted for the fixed effects in the model. In addition, we can study two features associated with the pair $\boldsymbol{\Lambda}_l\boldsymbol{w}_i$ and $\boldsymbol{\Lambda}_{l'}\boldsymbol{w}_i$. The first is the covariance between them, which specifies the (l, l')-th entry in $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$. The second is the expected distance between them, $E(\|\boldsymbol{\Lambda}_l\boldsymbol{w}_i - \boldsymbol{\Lambda}_{l'}\boldsymbol{w}_i\|^2) = (\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})^T$.

If $(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})^T$ is small, this implies that we have multiple ties for the two species in their row selection in $\boldsymbol{\Lambda}$. Therefore, the residual random effects are similar for the two species, providing a similar residual adjustment. This is not related to their mean contribution. However, more importantly, this means that the pair have a similar dependence structure to that of all remaining species. Evidently, when the l-th and l'-th rows of $\boldsymbol{\Lambda}$ share the same cluster, $(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})^T = \mathbf{O}$ (the matrix of zeros). More generally, the labels do not change much across iterations in the model fitting (see below). Thus, $(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})(\boldsymbol{\Lambda}_l - \boldsymbol{\Lambda}_{l'})^T$ takes a discrete set of values for many pairs.

A different perspective makes the spatial random effects orthogonal to the fixed effects (e.g., Hodges and Reich (2010); Hughes and Haran (2013); Hanks et al. (2015)). Let $\mathbf{X} = [\boldsymbol{x}_1 : \cdots : \boldsymbol{x}_n]^T$ and $\mathbf{U} = [\boldsymbol{U}_1 : \cdots : \boldsymbol{U}_n]^T$, and let $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ be the projection matrix associated with $M(\mathbf{X})$, the column space of $\mathbf{X}$. Then, we can write

$$E[\mathbf{U}|\mathbf{W}] = \mathbf{X}\mathbf{B}^T + \mathbf{P}\mathbf{W}\boldsymbol{\Lambda}^T + (\mathbf{I}_n - \mathbf{P})\mathbf{W}\boldsymbol{\Lambda}^T. \tag{2.5}$$

Thus, we can rewrite this conditional mean as

$$E[\mathbf{U}|\mathbf{W}] = \mathbf{X}\mathbf{B}^{*T} + \mathbf{W}^*\boldsymbol{\Lambda}^T, \tag{2.6}$$

where $\mathbf{B}^{*T} = \mathbf{B}^T + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\boldsymbol{\Lambda}^T$ and $\mathbf{W}^* = (\mathbf{I}_n - \mathbf{P})\mathbf{W}$. This approach deals with *spatial confounding* which describes multicollinearity among spatial covariates $\mathbf{X}$ and spatial random effects $\mathbf{W}$. Paciorek (2010) demonstrated that

this confounding can lead to bias in estimation, especially when the spatial random effects $\mathbf{W}$ are spatially smooth and have a large effective range of spatial autocorrelation. Hanks et al. (2015) consider spatial confounding in a geostatistical (continuous spatial support) setting. They demonstrate that the orthogonalization above provides computational benefits, but that its resulting Bayesian credible intervals can be inappropriately narrow under model misspecification.

In conclusion, confounding is only a problem while interpretating rather than predicting the coefficient matrix, $\mathbf{B}$. In particular, in our application below, Figures 7 and 8 reveal the difference in estimation between $\mathbf{B}$ and $\mathbf{B}^*$. We anticipate that the ecological reader will consider the regressors and the role they play when random effects are introduced,that is how much confounding there is in the data and model.

3. Adaptation to Binary Response, (i.e., Presence–Absence Data)

For binary presence–absence response data, a logit or probit model specification is often assumed. To work with binary responses, we adapt the data-augmentation algorithm proposed by Chib and Greenberg (1998) for a multivariate probit regression, which improves the mixing of the MCMC algorithm. Taylor-Rodríguez et al. (2017) consider the probit model specification,

$$Y_i^{(l)} = \begin{cases} 1 & U_i^{(l)} > 0 \\ 0 & U_i^{(l)} \leq 0 \end{cases}, \quad \text{for} \quad l = 1, \ldots, S, \quad i = 1, \ldots, n, \tag{3.1}$$

where $U_i^{(l)}$ is an auxiliary variable. We model $U_i^{(l)}$ as presented in Section 2.2. The form in (3.1) implies that we sample the latent $U_i^{(l)}$ from a truncated normal distribution within the MCMC iteration.

Note that we specify that $Y_i^{(l)} = g(U_i^{(l)}) = \mathbf{I}(U_i^{(l)} > 0)$. The latent Us are part of the first stage model specification, that is, $Y_i^{(l)}$ is a function of $U_i^{(l)}$. The latent process driving the binary responses is specified at the data stage. This contrasts with specifying a conditional distribution, $[Y_i^{(l)}|U_i^{(l)}]$ (e.g., $P(Y_i^{(l)} = 1) = p(U_i^{(l)})$), where $p(\cdot)$ would be a regression in $U_i^{(l)}$ (e.g., $\Phi(\alpha_0 + \alpha_1 U_i^{(l)})$). This moves the Us to a second-stage model specification and yields a probit regression.

To clarify, the former states that $Y_i^{(l)}$ arises deterministically from the $U_i^{(l)}$ surface. The latter states that we have a Bernoulli trial with a probit link function at each i. It is not clear whether the former is better than the latter. It may be preferred because we are directly modeling the dependence, joint and spatial,

between $U_i^{(l)}$ and $U_{i'}^{(l')}$, and hence between $Y_i^{(l)}$ and $Y_{i'}^{(l')}$, rather than deferring the dependence to the second stage, that is, to the presence–absence surface with conditionally independent Bernoulli trials at each location, given the surface. This is the distinction between our approach and that of Ovaskainen et al. (2016).

3.1. Identifiability issues

We aim to learn the dependence structure between species using $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \mathbf{I}_S$, and to extract the clustering behavior for the rows of $\boldsymbol{\Lambda}$. However, it is well known that, with random $\mathbf{W}$, the entries in $\boldsymbol{\Lambda}$ and σ_ϵ^2 are not identified. Thus, we briefly review the identifiability problems in factor and probit models. The identifiability problems for each of these specifications are mutually connected.

First, consider the loading matrices and factors under dimension-reduction. For posterior inference, we identify $\boldsymbol{\Lambda}\boldsymbol{w}$, but not $\boldsymbol{\Lambda}$ and $\boldsymbol{w}$. Some restriction on the factor loading matrices is required (Geweke and Singleton (1980); Lopes and West (2004)). A widely used approach is to fix certain elements of $\boldsymbol{\Lambda}$, usually to zero, such as restricting $\boldsymbol{\Lambda}$ to be upper or lower triangular matrices with strictly positive diagonal elements (Geweke and Zhou (1996)). This restriction enables a direct interpretation of the latent factors and loading matrices.

Alternatively, Ren and Banerjee (2013) discuss the difference related to identifiability according to whether the elements in the factors across locations ($\mathbf{W}^{(h)}$ for $h = 1, \ldots, r$) are independent or are spatially structured across locations. In the former case, the dependence structure is invariant to any orthogonal transformation of $\boldsymbol{\Lambda}$. We can have an infinite number of equivalent matrices of factor loadings. However, in the second case, they argue that only two types of linear transformations, namely reflections and permutations, lead to nonidentifiability. To avoid these types of nonidentifiability, Ren and Banerjee (2013) put a positivity restriction on the elements of the first row of $\boldsymbol{\Lambda}$. This is available for our modeling as well, but does not impose constant constraints on $\boldsymbol{\Lambda}$. Therefore, the elements of $\boldsymbol{\Lambda}$ and $\boldsymbol{w}$ still cannot be identified. However, the restrictions suggested by Ren and Banerjee (2013) enable us to identify the covariance structure of the latent process (i.e., Cov[vec($\mathbf{U}$)]), which is one of our goals.

4. Bayesian Inference

4.1. Model fitting

The full joint likelihood is

$$\mathcal{L} \propto (\sigma_\epsilon^2)^{(nS/2+1)} \prod_{i=1}^{n} \exp\left(-\frac{1}{2\sigma_\epsilon^2}\|\boldsymbol{U}_i - \mathbf{B}\boldsymbol{x}_i - \mathbf{Q}(\boldsymbol{k})\mathbf{Z}\boldsymbol{w}_i\|^2\right)$$
$$\times |\mathbf{C}_\phi|^{-1/2} \prod_{h=1}^{r} \exp\left(-\frac{1}{2}\mathbf{W}^{(h)T}\mathbf{C}_\phi^{-1}\mathbf{W}^{(h)}\right)\mathcal{IG}\left(\sigma_\epsilon^2\middle|\frac{a}{2},\frac{b}{2}\right)\prod_{l=1}^{S}\mathcal{N}(\mathbf{B}_l|\mathbf{0},c\mathbf{I}_p)$$
$$\times |\mathbf{D}_{\boldsymbol{Z}}|^{-1/2} \prod_{j=1}^{N} \exp\left(-\frac{1}{2}\boldsymbol{Z}_j^T\mathbf{D}_{\boldsymbol{Z}}^{-1}\boldsymbol{Z}_j\right) \times \prod_{l=1}^{S}\sum_{j=1}^{N} p_j\delta_j(k_l)\pi(\boldsymbol{p}|\mathbf{0},\boldsymbol{\alpha})$$
$$\times \mathcal{IW}\left(\mathbf{D}_{\boldsymbol{Z}}\middle|2+r-1, 4\mathrm{diag}\left(\frac{1}{\eta_1},\ldots,\frac{1}{\eta_r}\right)\right)\prod_{h=1}^{r}\mathcal{IG}\left(\eta_h\middle|\frac{1}{2},\frac{1}{10^4}\right)\mathcal{U}(\phi|\phi_{\min},\phi_{\max}). \tag{4.1}$$

Our sampling algorithm is similar to that of Taylor-Rodríguez et al. (2017), except for the sampling of $\mathbf{W}$ and ϕ. In our case, the elements of $\mathbf{W}$ are spatially correlated, but Gibbs sampling is still available. We describe the full sampling steps, including sampling of $\mathbf{W}$ and ϕ, in the Appendix.

4.2. Model comparison

Our model comparison focuses on the improvement of predictive performance at held-out locations. We implement out-of-sample predictive performance checks with respect to held-out samples of entire plots, rather than holding out samples of species within plots. This is in accord with our spatial modeling objective, which is to improve the predictive performance for held-out locations.

For the continuous-response case, the predictive performance is assessed by calculating the Euclidean distances between the true values and the conditional predictions, predicting $100p\%$ of the plots, conditional on the remaining $100(1-p)\%$ plots. We denote the number of plots of test data by m and the out-of-sample response matrix (test data) by $\mathbf{U}_{pred} = (\boldsymbol{U}_{1,pred},\ldots,\boldsymbol{U}_{m,pred})$ at locations $\mathcal{S}_{pred} = \{\boldsymbol{s}_{i_1},\ldots,\boldsymbol{s}_{i_m}\}$.

The criterion used to assess the predictive ability of the algorithm is the predictive mean squared error (PMSE), given by

$$\mathrm{PMSE} = \frac{1}{Sn_p}\sum_{i=1}^{m}(\boldsymbol{U}_{i,pred} - \hat{\boldsymbol{U}}_{i,pred})^T(\boldsymbol{U}_{i,pred} - \hat{\boldsymbol{U}}_{i,pred}), \tag{4.2}$$

where $\hat{\boldsymbol{U}}_{i,pred}$ is the posterior mean estimate of $\boldsymbol{U}_{i,pred}$.

For binary responses, we use the Tjur R^2 coefficient of determination (Tjur (2009)), which compares the estimated probabilities of the presence between the observed ones and the observed zeros. For species j, this quantity is given by

$TR_j = (\hat{\pi}_j(1) - \hat{\pi}_j(0))$, where $\hat{\pi}_j(1)$ and $\hat{\pi}_j(0)$ are the average probabilities of the presence for the observed ones and zeros, respectively, of the j-th species across the locations. The larger the TR_j, the better the discrimination is. We calculate an average TR measure across species, that is, $TR = (1/S)\sum_{j=1}^{S} TR_j$.

5. A Simulation Study

5.1. Continuous responses

We investigate the parameter recovery of our proposed model for continuous responses. We use the same locations ($n = 662$) and covariate information as in the CFR data. As covariate information, we include the following: (1) elevation, (2) mean annual precipitation, and (3) mean annual temperature. These values are standardized. The setting for the simulated data is

$$\begin{aligned} &q = 5, \quad p = 3, \quad S = 300, \quad K_{true} = 10, \quad \sigma_\epsilon^2 = 1, \\ &\boldsymbol{U}_i \sim \mathcal{N}(\tilde{\mathbf{B}}\boldsymbol{x}_i + \mathbf{Q}_{true}(\boldsymbol{k})\mathbf{Z}_{true}\boldsymbol{w}_i, \sigma_\epsilon^2 \mathbf{I}_S), \quad i = 1, \ldots, n, \\ &\tilde{\mathbf{B}}_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \quad l = 1, \ldots, S, \\ &\mathbf{W}^{(h)} \sim \mathcal{N}(\mathbf{0}, \mathbf{C}_\phi), \quad h = 1, \ldots, q, \\ &\mathbf{Z}_{true} = (\boldsymbol{Z}_{1,true}, \ldots, \boldsymbol{Z}_{K_{true},true})^T. \end{aligned} \tag{5.1}$$

Here, q denotes the fixed number of factors under the simulation. $\mathbf{W}^{(h)}$ is the h-th column of $\mathbf{W}$, an n-variate normal vector with mean $\mathbf{0}$ and covariance matrix $\mathbf{C}_\phi = [\exp(-\phi\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\|)]_{i,i'=1,\ldots,n}$. Here, we set $\phi = 2$. The label k_l is uniformly sampled from K_{true} labels for $l = 1, \ldots, S$. $\mathbf{Q}_{true}(\boldsymbol{k})$ and $\mathbf{Z}_{true}$ are $S \times K_{true}$ and $K_{true} \times q$ matrices, respectively. Each component of $\boldsymbol{Z}_{k,true}$ is uniformly selected from $\{-1, -0.5, 0, 0.5, 1\}$; for example, a realization might be $\boldsymbol{Z}_{k,true} = (0.5, -0.5, 0, 0, 1)^T$, such that $\boldsymbol{Z}_{k,true} \neq \boldsymbol{Z}_{k',true}$ for $k < k' = 1, \ldots, K_{true}$, and we set $\boldsymbol{Z}_{1,true} = 0.5\mathbf{1}_q$. We forced $\boldsymbol{Z}_{k,true}$ to be quite different from each other in order to facilitate the recovery of the number of clusters, especially for the binary case. We set $\boldsymbol{Z}_{1,true} = 0.5\mathbf{1}_q$ to keep all components of $\boldsymbol{Z}_{1,true}$ positive in order to meet the identifiability condition discussed in Section 3.1.

We estimate the posterior distributions for the objects in $\{\tilde{\mathbf{B}}, \mathbf{Z}, \mathbf{W}, \boldsymbol{k}, \sigma_\epsilon^2, \phi\}$ using the algorithms described in appendix A. The prior specification is

$$\sigma_\epsilon^2 \sim \mathcal{IG}(2, 0.1), \quad \phi \sim \mathcal{U}[\phi_{\min}, \phi_{\max}], \quad \tilde{\mathbf{B}}_l \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}_p), \quad \text{for} \quad l = 1, \ldots, S, \tag{5.2}$$

where $\phi_{\min} = 0.909$ and $\phi_{\max} = 46,052$. We implement dimension reduction, selecting $r = 5$ and $N = 150$ ($> K_{true}$ and $< S$). We run the MCMC, discarding

Table 1. Estimation results for continuous responses.

	True	Mean	Stdev	95% Int
ϕ	2	2.095	0.226	[1.600, 2.585]
σ_ϵ^2	1	1.000	0.003	[0.993, 1.006]

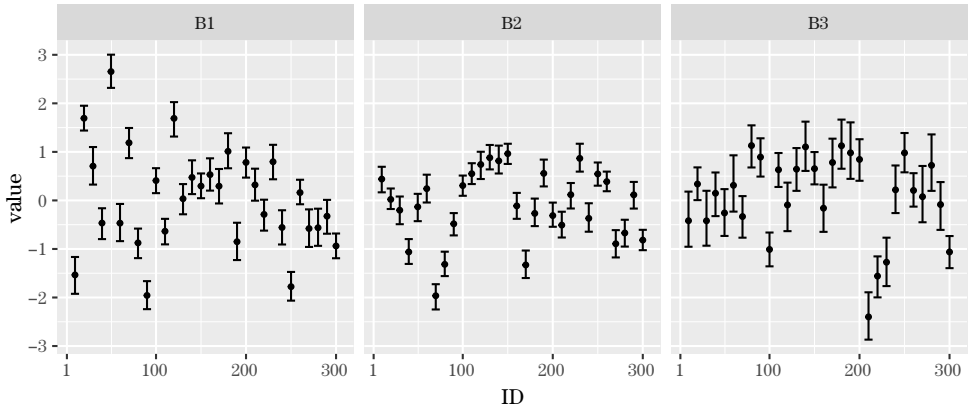

Figure 2. Estimated 95% CIs of $\tilde{\mathbf{B}}$ with continuous responses for 30 selected species. Black dots denote the true values.

the first 20,000 samples as a burn-in period, preserving the subsequent 20,000 samples as posterior samples.

Table 1 provides the estimation results for our model fitting. Both the decay parameter ϕ and the nugget variance σ_ϵ^2 are well recovered.

Figure 2 shows the 95% credible intervals (CIs) for $\tilde{\mathbf{B}}$ for 30 selected species (chosen every 10 species) by our model. With $\tilde{\mathbf{B}}$ identified in the case of continuous responses, the true parameter values are well recovered for both cases. Figure 3 reveals the sampled $\boldsymbol{k}$ of our spatial model for all species with a maximum posterior probability. Indeed, in this simulation study, the $\boldsymbol{k}$s for both models are completely recovered. In other words, the number of components of $\boldsymbol{k}$ is 10 ($= K_{true}$) with posterior probability one for both the independence and the spatial models. The sampled $\boldsymbol{k}$s for both models are also the same as the simulated $\boldsymbol{k}$ with posterior probability one.

In addition, we compare the true covariance $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \mathbf{I}_S$ with the estimated covariance $\hat{\boldsymbol{\Sigma}}^* = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \hat{\sigma}_\epsilon^2 \mathbf{I}_S$, where $\hat{\boldsymbol{\Lambda}}$ and $\hat{\sigma}_\epsilon^2$ are the posterior means of $\boldsymbol{\Lambda}$ and σ_ϵ^2 under the spatial and independent models, respectively. This comparison is motivated by the possibility that, with dependence in the spatial factors, the estimated covariance structure might be distorted by assuming independent

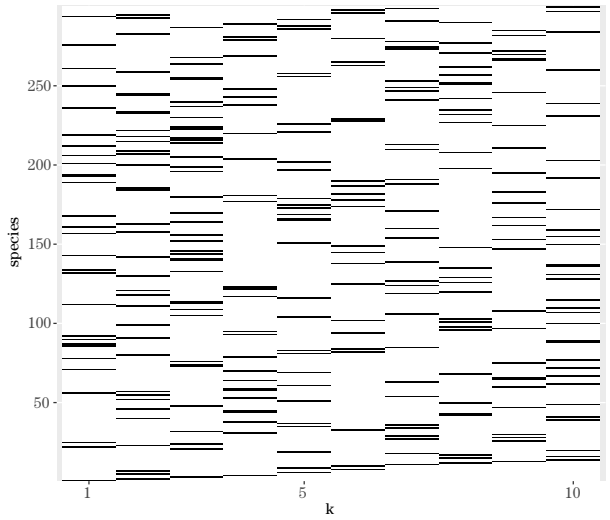

Figure 3. For continuous responses, the 0-1 map (0: white, 1: black) of the sampled $\boldsymbol{k}$ for the spatial model with a maximum posterior probability. Each species has only one label.

factors. We calculate the Frobenius norm (i.e., $\|\mathbf{A}\|_F = \sqrt{\sum_{l=1}^{S}\sum_{l'=1}^{S}|a_{ll'}|^2}$), for the difference $\boldsymbol{\Sigma}^* - \hat{\boldsymbol{\Sigma}}^*$. The values are 161.8 for the independent model and 31.13 for the spatial model. Hence, when factors have spatial dependence, the independence model appears to provide a less precise estimation of $\hat{\boldsymbol{\Sigma}}^*$.

Finally, we investigate the predictive performance of our spatial model. As discussed in Section 4.2, the predictive performance is assessed by calculating the Euclidean distances between the true values and the conditional predictions, predicting 20% of the plots, conditional on observing the remaining 80% of the plots. The estimated PMSE for our spatial model is 1.144 and that for the independence model is 2.069. Thus, the spatial model reveals an approximately 45% improvement over the independent model.

5.2. Binary responses

In addition to the continuous case, we investigate the parameter recovery and estimated covariance structure for binary responses. In the binary case, all parameter settings are the same as those in the continuous case, except for the observed response,

$$Y_i^{(l)} = \begin{cases} 1, & U_i^{(l)} > 0 \\ 0, & U_i^{(l)} \le 0 \end{cases}, \quad i = 1, \ldots, n, \quad l = 1, \ldots, S. \tag{5.3}$$

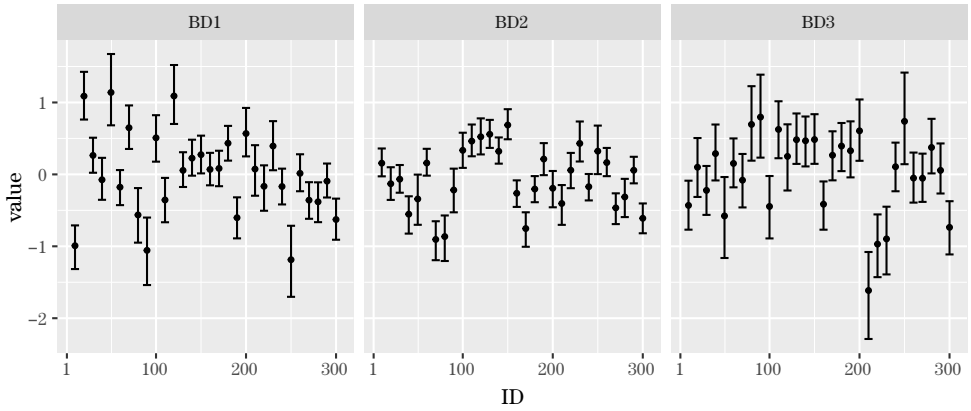

Figure 4. Estimated 95% CIs of $\mathbf{D}_{\boldsymbol{\Sigma}}^{-1/2}\tilde{\mathbf{B}}$ with binary response for 30 selected species. Black dots denote the true values.

We sample $\mathbf{U}$ as auxiliary responses within MCMC iterations. Again, we discard the first 20,000 samples as a burn-in period and preserve the subsequent 20,000 samples as posterior samples. The same prior specification is assumed for ϕ and $\tilde{\mathbf{B}}$ and we fix $\sigma_\epsilon^2 = 1$. The posterior mean of ϕ is 1.687 (95% CI [1.237, 2.422]) so the true value is well recovered.

For the binary case, $\tilde{\mathbf{B}}$ is not identifiable. Taylor-Rodríguez et al. (2017) estimate $\mathbf{B}$ using a scaled correlation matrix, $\mathbf{R} = \mathbf{D}_{\boldsymbol{\Sigma}^*}^{-1/2}\boldsymbol{\Sigma}^*\mathbf{D}_{\boldsymbol{\Sigma}^*}^{-1/2}$, that is, $\mathbf{B} = \mathbf{D}_{\boldsymbol{\Sigma}^*}^{-1/2}\tilde{\mathbf{B}}$, following the discussion in Lawrence et al. (2008). We adopt this choice as well, because applying the change of variables $(\tilde{\mathbf{B}}, \boldsymbol{\Sigma}^*)$ to $(\mathbf{B}, \mathbf{R})$ does not affect the probabilities for $\boldsymbol{Y}_i$, but identifies $\mathbf{B}$ as unaffected by the change of the scale matrix, $\mathbf{D}_{\boldsymbol{\Sigma}^*}$. Figure 4 shows the 95% CIs for $\mathbf{D}_{\boldsymbol{\Sigma}}^{-1/2}\tilde{\mathbf{B}}$ for 30 selected species (chosen every 10 species) under our model. The true parameter values are well recovered.

Figure 5 shows the 0-1 map of the sampled $\boldsymbol{k}$ for the spatial model with a maximum posterior probability. As in the continuous case, $\boldsymbol{k}$ is completely recovered. That is, the estimated number of clusters is 10 with posterior probability one, and $\boldsymbol{k}$ is the same as the true $\boldsymbol{k}$ with posterior probability one after a sufficiently long burn-in period.

Again, we compare the true covariance $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \mathbf{I}_S$ and the estimated covariance $\hat{\boldsymbol{\Sigma}}^* = \hat{\boldsymbol{\Lambda}}\hat{\boldsymbol{\Lambda}}^T + \mathbf{I}_S$ for the spatial and independent models. The calculated Frobenius norms are 156.1 for the independent model and 73.09 for the spatial model. The value for the spatial model is smaller than that of the independent

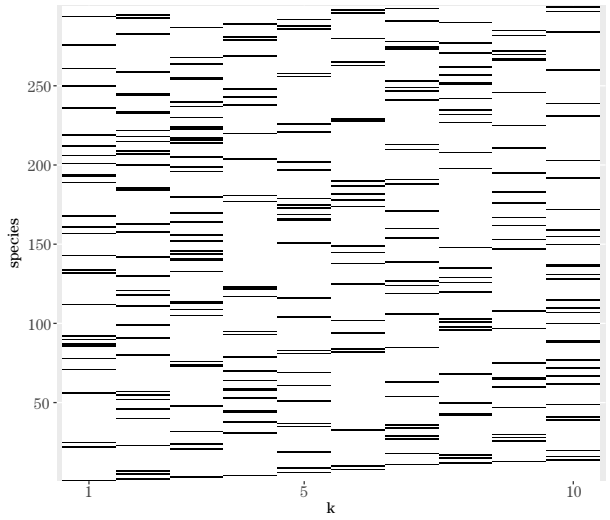

Figure 5. For binary responses, the 0-1 map (0: white, 1: black) of sampled $\boldsymbol{k}$ for the spatial model with a maximum posterior probability. Each species has only one label.

model, but larger than that of the spatial model with continuous responses. Finally, we investigate the predictive performance of our spatial model using the TR measures introduced in Section 4.2. The values are 0.5603 for the spatial model and 0.415 for the independent model; thus, the spatial model outperforms the independent model.

6. Real Data Application

From Section 2.1, the total number of binary responses is $n \times S = 662 \times 639 = 423,018$. The number of $Y_{l,i} = 1$ is 6,980, or 1.65% of all binary responses. Discarding the 351 species that are observed at at most five locations, we preserve $S = 288$ species for the model fitting. Longitude and latitude are transformed into easting and northing scales. Then, these scales are normalized by 100 km; thus, $\|\boldsymbol{s}_i - \boldsymbol{s}_{i'}\| = 1$ means the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$ is 100 km. Again, as covariate information, we include the following: (1) elevation, (2) mean annual precipitation, (3) mean annual temperature. These values are standardized.

In the analysis below, we set $r = 5$ (following Taylor-Rodríguez et al. (2017)). (We also conducted a sensitivity analysis for the choice of r; see below.) The prior specification is

$$\phi \sim \mathcal{U}[\phi_{\min}, \phi_{\max}], \quad \mathbf{B}_l \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}_p), \quad \text{for} \quad l = 1, \ldots, S, \tag{6.1}$$

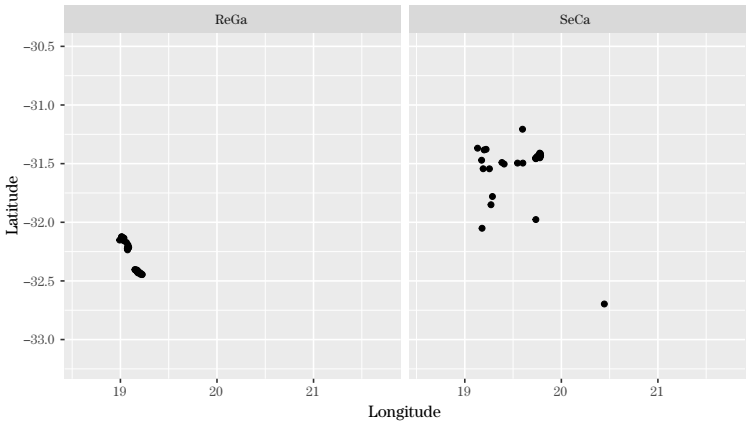

Figure 6. The distribution of ReGa (left) and SeCa (right).

where $\phi_{\min} = 0.909$ and $\phi_{\max} = 46,052$ and we fix $\sigma_\epsilon^2 = 1$. We discard the first 20,000 samples as a burn-in and preserve the subsequent 20,000 samples as posterior samples.

The estimated value of ϕ is 2.314 (95% CI [1.614, 3.589]), which reflects the spatial dependence for the factors. Among 288 species, the labels for 280 species are fixed with a posterior probability of one; that is, the same labels are selected for each of the 280 species for every posterior sample. The number of distinct labels, that is, associated with at least one species, is 22 with a posterior probability of one.

We also calculated the inefficiency factor (IF) which is the ratio of the numerical variance of the estimate from the MCMC samples relative to that from hypothetically uncorrelated samples. The IF is defined as $1 + 2\sum_{s=1}^{\infty} \rho_s$, where ρ_s is the sample autocorrelation at lag s. It suggests that the relative number of correlated draws necessary to attain the same variance of the posterior mean from the uncorrelated draws (Chib (2001)). The IFs for the parameters are $53 \sim 140$. Because we retain 20,000 samples as posterior draws, we preserve at least $20{,}000/140 \approx 142$ samples from the stationary distribution. The computational time for 40,000 iterations with five factors is 3,211 minutes.

We pick up two species, as discussed in Section 1, that share the same label arising from a large negative, and, hence, influential $\mathbf{W}\mathbf{\Lambda}^T$. The first species is *Restio gaudichaudianus* (ReGa), which shows large absolute values of $\mathbf{X}\mathbf{B}_l^T$. The second is *Senecio cardaminifolius* (SeCa), which shows small absolute values. Figure 6 shows the distribution of ReGa and SeCa. Both species show very

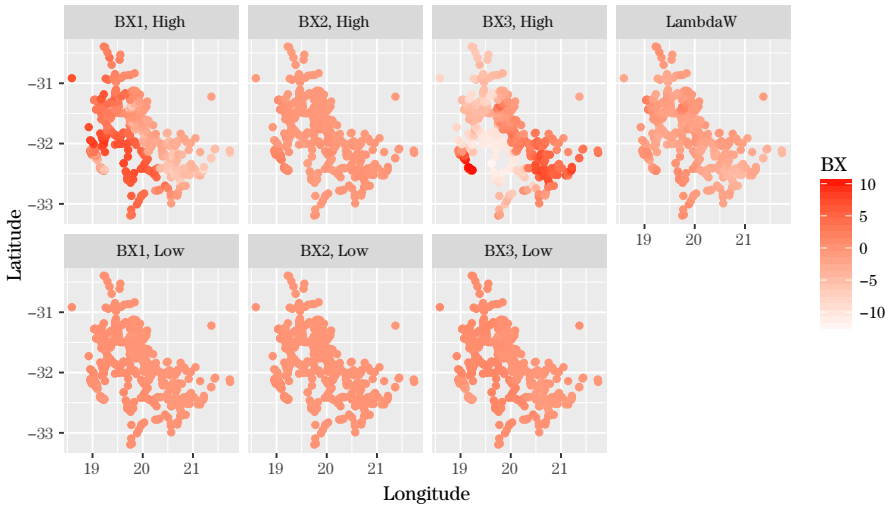

Figure 7. Estimated $\mathbf{XB}_l^T$ and $\mathbf{W}\boldsymbol{\Lambda}_l^T$ for ReGa (high, top) and Seca (low, bottom).

different distribution patterns, with ReGa concentrated in a small southwest area.

Figure 7 shows the estimation results for $\mathbf{XB}_l^T$ and $\mathbf{W}\boldsymbol{\Lambda}_l^T$. Because they share the same label, $\mathbf{W}\boldsymbol{\Lambda}_l^T$ is the same for both species. For ReGa, $\mathbf{XB}_l^T$ reveals larger variation than that of SeCa. In addition, $\mathbf{W}\boldsymbol{\Lambda}_l^T$ shows relatively negative values that exert a significant influence on the presence probability of SeCa. Figure 8 demonstrates the estimation results for the orthogonalized versions, $\mathbf{XB}_l^{*T}$ and $\mathbf{W}^*\boldsymbol{\Lambda}_l^T$, as defined in Section 2.3. Although the difference is small, the surface of $\mathbf{W}^*\boldsymbol{\Lambda}_l^T$ has larger positive values than those of $\mathbf{W}\boldsymbol{\Lambda}_l^T$. However, the figure suggests that spatial confounding effects are relatively small.

Next, we investigate the predictive performance of our model. As a sensitivity check with respect to the number of factors, Figure 9 shows the TR measure for the independence model with five factors (first boxplot) and for spatial models with different numbers of factors. The figure suggests that the spatial model with $r = 3$ factors performs best, while the spatial model with five factors is similar. Both models show better predictive performance than that of the independence model with five factors. In addition, having a greater number of factors does not improve the performance of the models.

Lastly, we compare the predictive performance between our models and the stacked "independence" model. Here, the independence model means that spatial random effects are introduced independently across species. Hence, the stacked

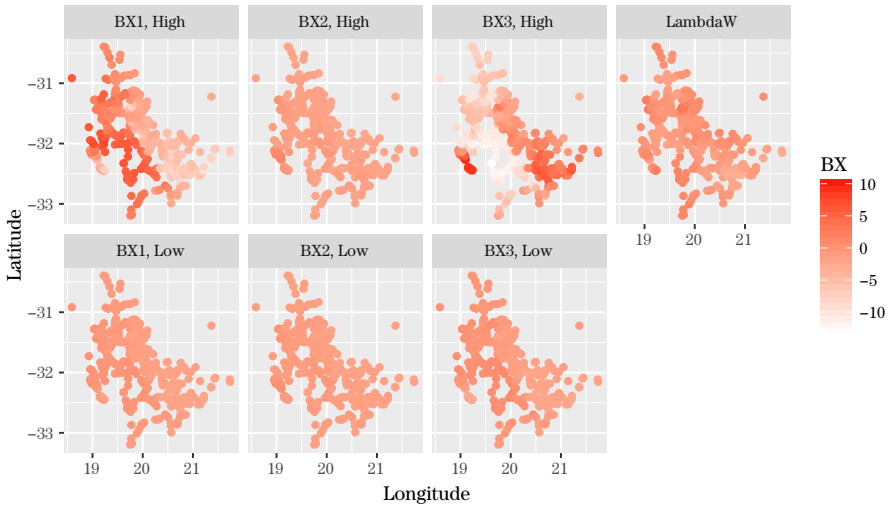

Figure 8. Estimated orthogonalized $\mathbf{XB}_l^{*T}$ and $\mathbf{W}^*\mathbf{\Lambda}_l^T$ for ReGa (high, top) and SeCa (low, bottom).

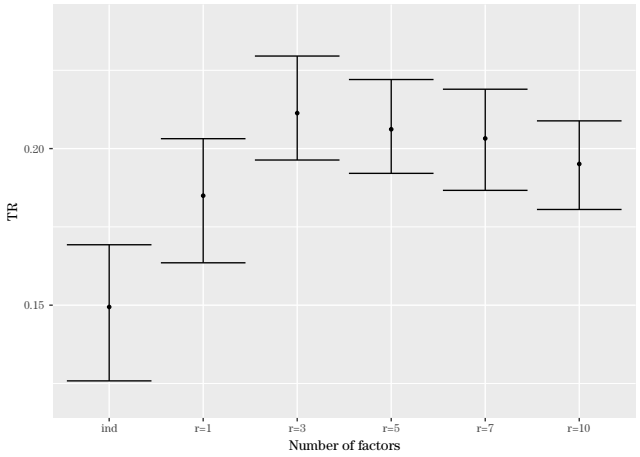

Figure 9. TR measure for each number of factors.

independence model incorporates spatial dependence, but not dependence among species. We calculate the conditional TR measure, denoted by $TR_{k|Y^{(l)}=1}$ and $TR_{k|Y^{(l)}=0}$ if we condition on species l being present or absent, respectively, as investigated in Taylor-Rodríguez et al. (2017). We illustrate this conditional TR measure at 134 held-out locations by conditioning on the presence–absence state of *Aridaria noctiflora* (ArNo) and obtain the posterior probability of the presence of *Pteronia glomerata* (PtGl). These species share the same label, with poste-

Table 2. Tjur R for PtGl conditional on ArNo at 134 held-out locations.

		PtGl		$TR_{PtGl\|ArNo}$	
		0	1	Independent	Joint
ArNo	0	$n_{00} = 100$	$n_{01} = 12$	0.2263	0.2523
	1	$n_{10} = 17$	$n_{11} = 5$	0.2574	0.2874

rior probability one. Furthermore, the posterior mean correlation between the two species is 0.4011, which is relatively high. We calculate $TR_{PtGl|Y_{ArNo}=1}$ and $TR_{PtGl|Y_{ArNo}=0}$ under both the joint model with $r = 5$ and the stacked independence model (Table 2). The joint model shows better validation performance.

7. Summary and Future Work

We have proposed a spatial joint species distribution model with a Dirichlet process dimension reduction for the factor loading matrix. The former enables dependence across spatial locations, and the latter enables dependence across species. We show that introducing spatial dependence into the factors improves the out-of-sample predictive performance over the study region under both continuous and binary species responses using both simulated and real data.

In future work, we will extend our model to handle more challenging responses. For instance, we often observe a compositional data response vector, that is, a response that lies on a simplex in R^S, but that allows for point masses at zeros. Another challenge is the case of a large number of spatial locations, for instance, at continental scales, resulting in perhaps $n \approx 10^6$. In this case, we will explore recently developed sparse GP approximations such as the nearest neighbor Gaussian processes (NNGP, Datta et al. (2016)) or the multiresolution Gaussian processes (MGP, Katzfuss (2017)). Another direction is a more detailed investigation of the effects of additional decay parameters with regard to the covariance matrices of the spatial factors. Ren and Banerjee (2013) allow different decay parameters for spatial factor models, ϕ_h, for $h = 1, \ldots, r$ using the Gaussian predictive process approximation by Banerjee et al. (2008). Without some approximation of the GPs, inferences with different decay parameters require computing matrix factorizations r times when sampling ϕ_h for $h = 1, \ldots, r$. This is computationally demanding, even when the number of locations is moderate. The NNGP or MGP may be useful in such situations.

Acknowledgment

The computational results are obtained using Ox version 7.1 (Doornik (2007)). The work of the first and third authors was supported, in part, by federal grants NSF/DMS 1513654, NSF/IIS 1562303, and NIH/NIEHS 1R01ES027027. The authors thank Matthew Aiello-Lammens and John A. Silander, Jr. for providing the Cape Floristic Region data as well as for motivation and useful conversations about the problem.

Appendix A. Details of Model Fitting

Sampling B

Let $\boldsymbol{x}_i$ be a $p \times 1$ location dependent covariate vector, which is assumed common for the $l = 1, \ldots, S$ species. For $\mathbf{B}_l$, we have $\mathbf{B}_l \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{B}_l}, \boldsymbol{\Sigma}_{\mathbf{B}})$ where

$$\boldsymbol{\mu}_{\mathbf{B}_l} = \boldsymbol{\Sigma}_{\mathbf{B}} \frac{1}{\sigma_\epsilon^2} \mathbf{X}^T (\mathbf{U}^{(l)} - \mathbf{W}(\mathbf{Z}^T \mathbf{Q}(\boldsymbol{k})^T)^{(l)}), \quad \boldsymbol{\Sigma}_{\mathbf{B}} = \left(\frac{\mathbf{X}^T \mathbf{X}}{\sigma_\epsilon^2} + \frac{1}{c} \mathbf{I}_S \right)^{-1} \tag{A.1}$$

with $\mathbf{U}^{(l)}$ is the l-th column of matrix $\mathbf{U}$ and $(\mathbf{Z}^T \mathbf{Q}(\boldsymbol{k})^T)^{(l)}$ the l-th column of matrix $\mathbf{Z}^T \mathbf{Q}(\boldsymbol{k})^T$.

Sampling Z

Sampling $\mathbf{Z}$ employs almost the same algorithm as in Taylor-Rodríguez et al. (2017). In our case, the first row of $\boldsymbol{\Lambda}$ is positive, we set $\boldsymbol{Z}_1$ as the first row of $\boldsymbol{\Lambda}$. For $j = 1$,

- let $S_1 = \{l = 1, \ldots, S, \text{s.t.} k_l = 1\}$ and let $|S_1|$ denote the cardinality of S_1. Using these definitions the full conditional distribution for $\boldsymbol{Z}_1$ is given by $\boldsymbol{Z}_1 \sim \mathcal{TN}_r(\boldsymbol{\mu}_{\boldsymbol{Z}_1}, \boldsymbol{\Sigma}_{\boldsymbol{Z}_1})$ where $\mathcal{TN}_r$ is multivariate truncated normal distribution defined on $(0, \infty)^r$ and

$$\boldsymbol{\mu}_{\boldsymbol{Z}_1} = \boldsymbol{\Sigma}_{\boldsymbol{Z}_1} \mathbf{W}^T \frac{1}{\sigma_\epsilon^2} \sum_{l \in S_1} (\mathbf{U}^{(l)} - \mathbf{X}\mathbf{B}_l^T), \quad \boldsymbol{\Sigma}_{\boldsymbol{Z}_1} = \left(\frac{|S_1|}{\sigma_\epsilon^2} \mathbf{W}^T \mathbf{W} + \mathbf{D}_{\boldsymbol{Z}}^{-1} \right)^{-1}. \tag{A.2}$$

The full conditional for other rows of $\mathbf{Z}$ depends on whether or not the row considered was chosen to be at least one row from $\boldsymbol{\Lambda}$, For $j = 2, \ldots, N$

1. If $j \notin \boldsymbol{k}$, sample $\boldsymbol{Z}_j \sim \mathcal{N}_r(\mathbf{0}, \mathbf{D}_{\boldsymbol{Z}})$.
2. Otherwise, let $S_j = \{l = 1, \ldots, S, \text{s.t.} k_l = j\}$ and let $|S_j|$ denote the cardinality of S_j. Using these definitions the full conditional distribution for $\boldsymbol{Z}_j$

is given by $\boldsymbol{Z}_j \sim \mathcal{N}_r(\boldsymbol{\mu}_{\boldsymbol{Z}_j}, \boldsymbol{\Sigma}_{\boldsymbol{Z}_j})$ where

$$\boldsymbol{\mu}_{\boldsymbol{Z}_j} = \boldsymbol{\Sigma}_{\boldsymbol{Z}_j}\mathbf{W}^T\frac{1}{\sigma_\epsilon^2}\sum_{l\in S_j}(\mathbf{U}^{(l)} - \mathbf{X}\mathbf{B}_l^T), \quad \boldsymbol{\Sigma}_{\boldsymbol{Z}_j} = \left(\frac{|S_j|}{\sigma_\epsilon^2}\mathbf{W}^T\mathbf{W} + \mathbf{D}_{\boldsymbol{Z}}^{-1}\right)^{-1} \tag{A.3}$$

with $\mathbf{B}_l$ the l-th row of matrix $\mathbf{B}$.

Sampling W

Sampling $\mathbf{W}$ requires the matrix factorization for n-dimensional covariance matrices. For $h = 1, \ldots, r$,

$$[\mathbf{W}^{(h)}|\cdot] \propto \prod_{i=1}^{n}\exp\left(-\frac{1}{2\sigma_\epsilon^2}\|\boldsymbol{U}_i - \mathbf{B}\boldsymbol{x}_i - \mathbf{Q}(\boldsymbol{k})\mathbf{Z}\boldsymbol{w}_i\|^2\right) \times \exp\left(-\frac{1}{2}\mathbf{W}^{(h)T}\mathbf{C}_\phi^{-1}\mathbf{W}^{(h)}\right). \tag{A.4}$$

Although Gibbs sampling is available, $\mathcal{O}(n^3)$ computational time is required.

Let $\mathbf{Z}^{(h)}$ be h-th column vector of $\mathbf{Z}$, $\mathbf{Z}^{(-h)}$ and $\mathbf{W}^{(-h)}$ be remaining matrices after deleting $\mathbf{Z}^{(h)}$ and $\mathbf{W}^{(h)}$, respectively. The full conditional is

$$\begin{aligned}
[\mathbf{W}^{(h)}|\cdot] &\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}\left(\mathbf{U} - \mathbf{X}\mathbf{B}^T - \mathbf{W}\mathbf{Z}^T\mathbf{Q}(\boldsymbol{k})^T\right)^T\left(\mathbf{U} - \mathbf{X}\mathbf{B}^T - \mathbf{W}\mathbf{Z}^T\mathbf{Q}(\boldsymbol{k})^T\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2}\mathbf{W}^{(h)T}\mathbf{C}_\phi^{-1}\mathbf{W}^{(h)}\right) \\
&\propto \exp\left(-\frac{1}{2\sigma_\epsilon^2}\left(\mathbf{U} - \mathbf{X}\mathbf{B}^T - \mathbf{W}^{(-h)}\mathbf{Z}^{(-h)T}\mathbf{Q}(\boldsymbol{k})^T - \mathbf{W}^{(h)}\mathbf{Z}^{(h)T}\mathbf{Q}(\boldsymbol{k})^T\right)^T\right. \\
&\quad \left.\times\left(\mathbf{U} - \mathbf{X}\mathbf{B}^T - \mathbf{W}^{(-h)}\mathbf{Z}^{(-h)T}\mathbf{Q}(\boldsymbol{k})^T - \mathbf{W}^{(h)}\mathbf{Z}^{(h)T}\mathbf{Q}(\boldsymbol{k})^T\right)\right) \\
&\quad \times \exp\left(-\frac{1}{2}\mathbf{W}^{(h)T}\mathbf{C}_\phi^{-1}\mathbf{W}^{(h)}\right) \\
&= \mathcal{N}(\boldsymbol{\mu}_{w_h}, \boldsymbol{\Sigma}_{w_h}),
\end{aligned} \tag{A.5}$$

where

$$\boldsymbol{\mu}_{w_h} = \boldsymbol{\Sigma}_{w_h}\frac{1}{\sigma_\epsilon^2}\left(\mathbf{U} - \mathbf{X}\mathbf{B}^T - \mathbf{W}^{(-h)}\mathbf{Z}^{(-h)T}\mathbf{Q}(\boldsymbol{k})^T\right)\mathbf{Q}(\boldsymbol{k})\mathbf{Z}^{(h)}, \tag{A.6}$$

$$\boldsymbol{\Sigma}_{w_h} = \left(\mathbf{C}_\phi^{-1} + \frac{\|\mathbf{Z}^{(h)T}\mathbf{Q}(\boldsymbol{k})^T\|^2}{\sigma_\epsilon^2}\mathbf{I}_n\right)^{-1}. \tag{A.7}$$

Sampling ϕ

The full conditional distribution for ϕ is

$$|\mathbf{C}_\phi|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{W}^{(h)T}\mathbf{C}_\phi^{-1}\mathbf{W}^{(h)}\right)\mathbf{I}(\phi_{\min}<\phi<\phi_{\max}). \tag{A.8}$$

We implement a Metropolis-Hastings algorithm.

Sampling $\boldsymbol{k}$

For the vector of labels $\boldsymbol{k}$, the full conditional distribution is $[\boldsymbol{k}|\cdot] = \prod_{l=1}^{S}(\sum_{j=1}^{N} p_{l,j}\delta_j(k_l))$ with

$$p_{l,j} \propto p_j \times \exp\left(-\frac{1}{2\sigma_\epsilon^2}\|\mathbf{U}^{(l)} - \mathbf{X}\mathbf{B}_l^T - \mathbf{W}\boldsymbol{Z}_j\|^2\right). \tag{A.9}$$

Sampling $\boldsymbol{p}$

The full conditional distribution for $\boldsymbol{p}$, given conjugacy of the $\mathcal{GD}$ distribution with multinomial sampling, the draws of $\boldsymbol{p}$ are

$$p_1 = \xi_1, \tag{A.10}$$

$$p_j = (1-\xi_1)\ldots(1-\xi_{j-1})\xi_j, \quad \text{for} \quad j = 2,3,\ldots,N-1, \tag{A.11}$$

$$p_N = 1 - \sum_{j=1}^{N-1} p_j, \tag{A.12}$$

with $\xi_j \sim \text{Beta}(\alpha/N + \sum_{l=1}^{S} I_{(k_l=j)}, (N-1)/N\alpha + \sum_{s=j+1}^{N}\sum_{l=1}^{S} I_{(k_l=s)})$ for $j = 1,\ldots,N-1$.

Sampling σ_ϵ^2

By conjugacy of the prior for σ_ϵ^2 with the normal likelihood, the full conditional distribution is

$$\sigma_\epsilon^2 \sim \mathcal{IG}\left(\frac{nS+a}{2}, \frac{\sum_{i=1}^{n}\|\boldsymbol{U}_i - \mathbf{B}\boldsymbol{x}_i - \mathbf{Q}(\boldsymbol{k})\mathbf{Z}\boldsymbol{w}_i\|^2 + b}{2}\right). \tag{A.13}$$

Sampling $\mathbf{D}_{\boldsymbol{Z}}$

$$\mathbf{D}_{\boldsymbol{Z}} \sim \mathcal{IW}\left(\mathbf{D}_{\boldsymbol{Z}}|2+r+N-1, \mathbf{Z}^T\mathbf{Z} + 4\text{diag}\left(\frac{1}{\eta_1},\ldots,\frac{1}{\eta_r}\right)\right). \tag{A.14}$$

References

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**, 95–106.

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for*

Spatial Data, 2nd Edition. Chapman and Hall/CRC.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008). Gaussain predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 825–848.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98**, 291–306.

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

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

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**, 99–112.

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**, 757–776.

Chib, S. (2001). Markov chain Monte Carlo methods: computation and inference. In: Handbook of Econometrics. 5 (Edited by G. Elliott, C. W. J. Granger and A. Timmermann), 3569–3649. Amsterdam: North Holland Press.

Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**, 347–361.

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**, 1273–1287.

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**, 990–999.

Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. and Zhange, S. (2017). Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data. *Ecological Monographs* **87**, 34–56.

Datta, A., Banerjee, S., Finley, A. O. and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association* **111**, 800–812.

Doornik, J. (2007). *Ox: Object Oriented Matrix Programming.* Timberlake Consultants Press.

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

Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual Reviews of Ecology, Evolution, and Systematics* **40**, 677–697.

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

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

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**, 1–20.

Gelfand, A. E., Silander, J. A., Wu, S., Latimer, A., Lewis, P. O., Rebelo, A. G. and Holder, M. (2006). Explaining species distribution patterns through hierarchical modeling. *Bayesian Analysis* **1**, 41–92.

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**, 133–137.

Geweke, J. F. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The Review of Financial Studies* **9**, 557–587.

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**, 1433–1444.

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

Hanks, E. M., Schliep, E. M., Hooten, M. B. and Hoeting, J. A. (2015). Restricted spatial regression in practice: geostatistical models, confounding, and robustness under model misspecification. *Environmetrics* **26**, 243–254.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64**, 325–334.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **75**, 139–159.

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**, 390–406.

Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association* **112**, 201–214.

Latimer, A., Banerjee, S., Jr, H. S., Mosher, E. and Jr, J. S. (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**, 144–154.

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

Lawrence, E., Bingham, D., Liu, C. and Nair, V. N. (2008). Bayesian inference for multivariate ordinal data using parameter expansion. *Technometrics* **50**, 182–191.

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

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14**, 41–67.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics* **23**, 727–741.

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*

and Evolution **26**, 249–259.

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**, 445–451.

Neal, R. M. (2000). Markov chain sampling Methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

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

Ovaskainen, O., Roy, D. B., Fox, R. and Anderson, B. J. (2016). Uncovering hidden spatial structure in species communities with spatially explicit joint species distribution models. *Methods in Ecology and Evolution* **7**, 428–436.

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

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

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**, 397–406.

Rebelo, T. (2001). *SASOL Proteas: A Field Guide to the Proteas of South Africa (2nd Ed).* Fernwood Press.

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

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

Takhtajan, A. (1986). *Floristic Regions of the World.* University of California Press.

Taylor-Rodríguez, D., Kaufeld, K., Schliep, E. M., Clark, J. S. and Gelfand, A. E. (2017). Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis* **12**, 939–967.

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* **6**, 627–637.

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

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**, 531–534.

Tjur, T. (2009). Coefficients of determination in logistic regression models-A new proposal: the coefficient of discrimination. *The American Statistician* **63**, 366–372.

Wang, F. and Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics* **4**, 569–582.

West, M. (2003). Bayesian factor regression models in the large p, small n paradigm. In Bayesian Statistics 7 (Edited by J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West), 723–732. Oxford University Press.

Department of Biostatistics, University of California, Los Angeles. 650 Charles E. Young Drive South Los Angeles, CA 90095-1772, USA.

E-mail: shinichiro.shirota@gmail.com

Department of Statistics, Duke University, Durham, NC 27708-0251, USA.

E-mail: alan@duke.edu

Department of Biostatistics, University of California, Los Angeles. 650 Charles E. Young Drive South Los Angeles, CA 90095-1772, USA.

E-mail: sudipto@ucla.edu

(Received November 2017; accepted September 2018)