The New England Journal of Statistics in Data Science Volume 1, 102–109 (2023)
DOI: https://doi.org/10.51387/22-NEJSDS11

Some Noteworthy Issues in Joint Species Distribution Modeling for Plant Data

Alan E. GELFAND

Abstract

Joint species distribution modeling is attracting increasing attention in the literature these days, recognizing the fact that single species modeling fails to take into account expected dependence/interaction between species. This short paper offers discussion that attempts to illuminate five noteworthy technical issues associated with such modeling in the context of plant data. In this setting, the joint species distribution work in the literature considers several types of species data collection. For convenience of discussion, we focus on joint modeling of presence/absence data. For such data, the primary modeling strategy has been through introduction of latent multivariate normal random variables.

These issues address the following: (i) how the observed presence/absence data is linked to the latent normal variables as well as the resulting implications with regard to modeling the data sites as independent or spatially dependent, (ii) the incompatibility of point referenced and areal referenced presence/absence data in spatial modeling of species distribution, (iii) the effect of modeling species independently/marginally rather than jointly within site, with regard to assessing species distribution, (iv) the interpretation of species dependence under the use of latent multivariate normal specification, and (v) the interpretation of clustering of species associated with specific joint species distribution modeling specifications.

It is hoped that, by attempting to clarify these issues, ecological modelers and quantitative ecologists will be able to better appreciate some subtleties that are implicit in this growing collection of modeling ideas. In this regard, this paper can serve as a useful companion piece to the recent survey/comparison article by [33] in *Methods in Ecology and Evolution*.

keywords and phrases: Dirichlet process, Gaussian process, Latent factor analysis, Latent variables, Model-based clustering, Odds ratios, Spatial dependence, Species richness.

1. INTRODUCTION

Recently, in the context of plants, there has been a flood of publication on joint species distribution modeling (JSDM) in the literature [22, 29, 19, 9]. A useful comparison of such modeling has been presented in [33]. Such effort reflects the realization that observation of a community at a site anticipates dependence between the species present at that site. That is, so-called stacked species distribution modeling [13, 6], modeling the species marginally but looking at the results jointly, need not perform well. For example, with presence/absence data, such modeling tends to overestimate probability of presence for each species at a site, hence the number of presences at a site [9]. Below, we will elaborate this issue further.

Joint species distribution modeling has been developed for presence/absence data, for count (abundance) data, and for composition data [9]. Here, for simplicity in discussing the challenges of interest, we focus solely on presence/absence data. We have primary concern with the setting consisting of a large number S of species and a large number n of sites. Site $i, i = 1, 2, \ldots, n$ provides an $S \times 1$, vector, $\mathbf{Y}_i$ with entries 1 (presence) or 0 (absence). The sites may be viewed, hence modeled, as independent or spatially dependent, as appropriate. The joint species distribution modeling challenge is the need to model the set of 2^S probabilities associated with the set of possible realizations of $\mathbf{Y}_i$. Direct modeling of these probabilities is clearly infeasible even for relatively small S while we imagine S of order 10^2 or even 10^3. The common solution that has been adopted in the literature is to introduce latent variables, $\mathbf{Z}_i$ which drive the responses $\mathbf{Y}_i$. The $\mathbf{Z}_i$ are modeled as multivariate normal vectors which enables tractable model specification though still computationally demanding model fitting. There is increasing literature on this demanding model fitting when n and S are large [27] and, further, when we introduce spatial dependence [26]. However, we do not consider the computational challenge here. Rather, our focus is on issues associated with model specification.

This takes us to the specific contribution of this note. We attempt to illuminate five consequential technical issues associated with joint species distribution modeling, presented in the context of presence-absence data. We address the following: (i) how the observed presence/absence data is linked to the latent normal variables as well as the resulting implications with regard to modeling the data sites as independent or spatially dependent, (ii) the incompatibility of point referenced and areal referenced presence/absence data in spatial modeling of species distribution, (iii) the effect of modeling species independently rather than jointly within site, with regard to assessing species distribution, (iv) the interpretation of species dependence under the use of latent multivariate normal specification, and (v) the interpretation of clustering of species associated with specific joint species distribution modeling specifications.

Species distribution modeling for animals offers a much more difficult challenge due to animal movement and the scale of animal range. There is a very active literature on animal movement, almost all of it at the single species level; there is certainly nothing at the order $O(10^2)$ species which we find in the plant literature. The proposed modeling is in a very different spirit from that for plant data, both dynamic and at larger spatial scale, (see, e.g., [15]) and we do not pursue it further here.

So, the format for the paper is simple. We devote a section to each of the five foregoing issues and then present a brief concluding section.

2. ISSUE (I): LINKING BINARY RESPONSES TO LATENT NORMAL VARIABLES

This issue concerns how the observed presence/absence data is linked to the latent normal variables as well as the resulting implications with regard to modeling the data sites as independent or spatially dependent. Starting with the nonspatial case, let Y_{ij} denote the response of species j at site i and let Z_{ij} denote the associated Gaussian variable. We adopt notation in the spirit of [20], letting L_{ij}^F and L_{ij}^R denote the fixed and random effects contributions which are included additively in the modeling of Z_{ij}. More will be said about the forms of these L's below. However, as the definitions suggest, we will view L_{ij}^F as a nonrandom component in the specification for Z_{ij} (though it will have parameters – regression coefficients – in it and, in a Bayesian hierarchical modeling framework, these coefficients would be viewed and modeled as random). We will view $\mathbf{L}_i^R$ as an $S\times 1$ multivariate normal random variable with mean $\mathbf{0}$ and dependence structure given by an $S \times S$ correlation matrix, H. Then, marginally, $L_{ij}^R \sim N(0,1)$.

Should we model Y_{ij} as a *function* of Z_{ij}, i.e., $Y_{ij} = I(Z_{ij} > 0)$ where I is the indicator function [as in, e.g., 22, 9] or should we model Y_{ij} using a *conditional* distribution, $[Y_{ij}|Z_{ij}]$ [as in 19]? Does it matter? We now clarify that the answer is NO if we view the sites as independent but YES if we view the sites as spatially dependent.

Under the *functional* relationship between Y_{ij} and Z_{ij}, since $Y_{ij} = I(Z_{ij} > 0)$, we have $P(Y_{ij} = 1) = P(Z_{ij} > 0)$. Consider the following two specifications for Z_{ij}:

(i) $Z_{ij} = L_{ij}^F + L_{ij}^R$
(ii) $Z_{ij} = L_{ij}^F + L_{ij}^R + \epsilon_{ij}$

where, in (ii), the ϵ_{ij} are pure error terms, i.e., independent and identically distributed normal random variables with mean 0 and variance 1. Then, with Φ denoting the standard normal cumulative distribution function, under (i), given L_{ij}^F,

$$P(Y_{ij} = 1) = P(Z_{ij} > 0) = \Phi(L_{ij}^F). \tag{2.1}$$

Under (ii), given L_{ij}^F and L_{ij}^R,

$$P(Y_{ij} = 1) = P(Z_{ij} > 0) = \Phi(L_{ij}^F + L_{ij}^R). \tag{2.2}$$

Working with model (i) for Z_{ij}, we have dependence at the first stage specification. The Z_{ij} are dependent, $\text{corr}(Z_{ij}, Z_{ij'}) = \text{corr}(L_{ij}^R, L_{ij'}^R)$, and therefore, so are the Y_{ij}. Indeed, this direct dependence approach is advocated in [9]. The concern here is that now the probability of presence has no random effects in it; simply, $P(Y_{ij} = 1) = \Phi(L_{ij}^F)$ is entirely driven by covariates. We have a basic probit regression for every species. Moreover, how do we usefully interpret a correlation between normal random variables with regard to the association between the binary variables, Y_{ij} and $Y_{ij'}$? We take up this question under challenge (iv) in Section 5 below.

If we adopt (ii) above, we model $P(Y_{ij} = 1)$ to include both fixed and random effects. It is clear that dependence is introduced through the specification for L_{ij}^R. Under this specification, dependence between species is captured in the probability of presence, the so-called second stage of a hierarchical model, as we see from (2.2). In fact, it is the correlation between $\Phi^{-1}(P(Y_{ij} = 1))$ and $\Phi^{-1}(P(Y_{ij'} = 1))$. In different words, $P(Y_{ij} = 1)$ and $P(Y_{ij'} = 1)$ are dependent but the events $Y_{ij} = 1$ and $Y_{ij'}$ are conditionally independent given these probabilities.

Furthermore, stochastic dependence between probability of presence replaces explicit modeling of interaction [9]. This raises the question of what the resulting correlation means. In this regard, it is associated with *residuals* as (ii) reveals, i.e., adjusted for the *mean*, L_{ij}^F. Moreover, at any site, we will find only a small subset of the S species present. That is, $\mathbf{Y}_i$ will be predominantly comprised of 0's. Nonetheless, we create pairwise associations for all pairs of species. So, it is evident that these associations have little to do with the actual realization of $\mathbf{Y}_i$ at site i.

Furthermore, is a positive association suggestive of encouraging co-occurrence or of a potential substitution effect, i.e., a particular species is present but another, say similar one, could equally well have been successful there? This leads to discussion presented in, e.g., [35] and [20] regarding the global species pool (all existing species), the regional species pool (those able to colonize an area), and the local species pool (those found at the finest scale considered). Recent discussion clarifying that species co-occurrences from JSDMs are not able to be interpreted directly as species interactions appears in [3] and in [5]. However, further ecological elaboration of species interaction/dependence is beyond our interest here. In the sequel, under the foregoing modeling, we view pairwise correlation/dependence between species as a surrogate for species interaction.

Now, turn to the *conditional* specification, again, $[Y_{ij}|Z_{ij}]$. Under a probit link function, $P(Y_{ij}=1)=\Phi(Z_{ij})$. So, under (i), we obtain $\Phi(L_{ij}^F+L_{ij}^R)$, the form in [20]. We can conclude that using (ii) under the functional specification or (i) under the conditional specification produces the same probability of presence. Therefore, if our goal is merely to obtain $\Phi(L_{ij}^F+L_{ij}^R)$ as the probability of presence, we can achieve this under either specification. However, if the functional specification is used, we must adopt (ii) above for the Z's. This distinction seems muddled in, e.g., [33].

Next, suppose we bring in space and spatial dependence. For plant data, we assume that the spatial scale of the study region of interest is large enough so that we can view plots as geo-coded locations. Hence, we modify notation by attaching location $\mathbf{s}_i$ to site i and writing $Y_{ij}\equiv Y_j(\mathbf{s}_i)$. Now we conceptualize a presence/absence variable, $Y_j(\mathbf{s})$ for species j at every location, $\mathbf{s}$, in the study region, say D, and, in fact, a realization of a presence/absence (binary) surface for species j, $\{Y_j(\mathbf{s}):\mathbf{s}\in D\}$. This surface is observed at $\{\mathbf{s}_i, i=1,2,\ldots,n\}$. With regard to the Z's, now we have:

(i) $Z_j(\mathbf{s})=L_j^F(\mathbf{s})+L_j^R(\mathbf{s})$ or
(ii) $Z_j(\mathbf{s})=L_j^F(\mathbf{s})+L_j^R(\mathbf{s})+\epsilon_j(\mathbf{s})$.

Here, $\epsilon_j(\mathbf{s})$ is pure error, so called white noise. That is, at each $\mathbf{s}$, we have an associated independent normal error random variable.

Suppose $L_j^F(\mathbf{s})$ is a surface which is continuous except for a set of measure 0 over D. What this means here is that typically, environmental regressors are available at areal scales making $L_j^F(\mathbf{s})$ continuous over D except for the boundaries between areas. The total area of these boundaries is 0 relative to the area of D. Further, suppose $L_j^R(\mathbf{s})$ is a realization of a Gaussian process [2] which produces mean square continuous realizations.[1] Then, under (i), $Z_j(\mathbf{s})$ is a continuous surface except for a set of measure 0 while under (ii) $Z_j(\mathbf{s})$ is everywhere discontinuous because the pure error $\epsilon_j(\mathbf{s})$ surface is.

Again, consider the functional specification, now $Y_j(\mathbf{s})=I(Z_j(\mathbf{s})>0)$ (which is referred to as a clipped Gaussian field in the literature [e.g., 10]), and the conditional specification, now $[Y_j(\mathbf{s})|Z_j(\mathbf{s})]$. Suppose, we work with (ii) yielding a probability of presence surface, $P(Y_j(\mathbf{s})=1)=\Phi(L_j^F(\mathbf{s})+L_j^R(\mathbf{s}))$. Then, following the previous paragraph, for species j, the probability of presence surface is a.e. continuous over D. However, under the conditional specification, each $Y_j(\mathbf{s})$ is drawn as a conditionally independent Bernoulli variable given its probability of presence. Hence, the realized presence/absence surface, $\{Y_j(\mathbf{s}):\mathbf{s}\in D\}$ here is everywhere discontinuous. This seems unsatisfying; the realized presence/absence surface should manifest *local* smoothness, local subregions where it is 0, local subregions where it is 1.

Back to the functional specification, under (ii), since $Z_j(\mathbf{s})$ is everywhere discontinuous, we can not obtain local continuity for the $Y_j(\mathbf{s})$ surface. However, under (i), if the $Z_j(\mathbf{s})$ surface is continuous, with the functional specification, we can obtain local continuity for the $Y_j(\mathbf{s})$ surface. The point here is that, with spatial modeling, if we value local smoothness in the realized presence/absence surface, if we think that such smoothness more appropriately captures real world behavior of process realizations, then we should work with the functional specification since this smoothness can never be achieved with the conditional specification.

However, to work with the functional specification under (i), we encounter a technical problem. Suppose we define $L_{ij}^F=\boldsymbol{X}^T(\boldsymbol{s}_i)\boldsymbol{\beta}_j$ and $L_{ij}^R=w_j(\mathbf{s}_i)$. The problem concerns the difference between the probability of presence surface under (i) vs. under (ii). Because of the spatial dependence imposed on the presence/absence surface under (i), the realized presence surface, $\Phi(\boldsymbol{X}^T(\boldsymbol{s})\boldsymbol{\beta}_j)$ has to "agree" with the observed presences and absences. Under (ii), smoothness is imposed on the probability of presence surface, i.e., $\Phi(\boldsymbol{X}^T(\boldsymbol{s})\boldsymbol{\beta}_j+w_j(\mathbf{s}))$ but not on the realized presence/absence surface. With the latter, we can observe a presence that has small probability of occurring or an absence that has a small probability of occurring. As a result, the probability of presence surface does not have to work as hard to fit the data. Under the functional model, the GP has to react strongly to observed presences and absences. Under the conditional modeling, it has to react less so. Therefore, when fitting the functional model, the $w_j(\boldsymbol{s})$ surface becomes spiky in the neighborhood of a presence in order to explain well the observed presence. The flexibility of the GP produces a posterior which is too sensitive to the data.

A potential solution is to replace $\epsilon_j(\mathbf{s})$ with $v_j(\mathbf{s})$, a second spatial Gaussian process, exchanging the discontinuity everywhere of the former with the spatial continuity of the latter. That is, still using the functional form, $Y_j(\boldsymbol{s}_i)=1,0$ according to $Z(\boldsymbol{s}_i)\geq 0, <0$, we have two GP's in specifying $Z_j(\boldsymbol{s})$, i.e., $Z_j(\boldsymbol{s})=\boldsymbol{X}^T(\boldsymbol{s})\boldsymbol{\beta}_j+w_j(\boldsymbol{s})+v_j(\boldsymbol{s})$. Here, $w_j(\boldsymbol{s})$ has a larger range, a smaller decay parameter while $v_j(\boldsymbol{s})$ has a smaller range with a larger decay parameter. That is, the w process seeks to capture the spatial dependence in the process while the v process only serves as a device to introduce smoothness.

3. ISSUE (II): INCOMPATIBILITY OF POINT-REFERENCED AND AREAL UNIT PRESENCE/ABSENCE DATA

This issue concerns the incompatibility of point referenced and areal referenced presence/absence data in spatial modeling of species distribution. To do so requires explicit discussion regarding what an observed presence means along

[1] A sufficient condition is that the correlation function of the Gaussian process be continuous at $\mathbf{0}$.

with the associated implications. The problem is whether presence/absence is viewed as an event at point level or at areal level. Is it a Bernoulli trial at say location $\mathbf{s}$ or is it the event that the number of individuals of a species in a set, say A, is ≥ 1?

If we model presence/absence at point level, then $Y(\mathbf{s}) = 1$ is the result of a Bernoulli trial at location $\mathbf{s}$. However, under point level modeling, what does $Y(A)$ mean? A coherent probabilistic definition specifies it as a block average, i.e., a realization of $Y(A)$ is $Y(A) = \int_A 1(Y(\mathbf{s}) = 1)d\mathbf{s}/|A|$ (where $|A|$ is the area of A). It is the proportion of the $Y(\mathbf{s})$ in A that equal 1; it is not a Bernoulli trial and $P(Y(A) = 1) = 0$ since the probability that almost every Bernoulli trial in A results in a 1 equals 0. We can calculate $E(Y(A)) = \int_A p(\mathbf{s})d\mathbf{s}/|A|$ with $p(\mathbf{s})$ specified as in the previous section. That is, $E(Y(A))$ becomes the average probability of presence over A. It is the probability that, at a randomly selected location in A, the species is present. If $p(\mathbf{s})$ is constant over A then $E(Y(A))$ is this constant probability. It is interpreted at point level; it is the probability of presence at any site in A.

Now, suppose we consider the locations of all individuals in a study region as a random point pattern. Then, if $N(A)$ is the number of individuals in set A, $P(\text{presence } \mathit{in} \ A) = P(N(A) \geq 1)$. Here, assuming say, a nonhomogeneous Poisson process (NHPP) or, more generally a log Gaussian Cox process (LGCP) with intensity $\lambda(\mathbf{s})$ (see Illian et al. (2008) for a full discussion of NHPPs and LGCPs), $N(A) \sim Po(\lambda(A))$ where $\lambda(A) = \int_A \lambda(s)ds$. Then, taking the areal unit definition of a presence in A, we seek $P(Y(A) = 1) = P(N(A) \geq 1) = 1 - e^{-\lambda(A)}$. Viewing the data as a collection of observed presences imagines the data as presence-only; there are no absences [32, 7]. This conceptualization enables the foregoing definition of presence/absence. However, the probability of a presence is only defined given A and, evidently, will depend on the size/scale of A. As a result, it is unclear how to specify a meaningful probability of presence surface. Perhaps the best option would be a *gridded* surface for some choice of A? Furthermore, the definition of probability of presence as "one or more" observations of the species in A yields local distortion to any such surface; $N(A) = 1$ or $N(A) = 11$ are treated the same with regard to probability of presence in A [1].

The two foregoing definitions associated with presence/absence are incompatible and the fundamental difference between them seems to have been missed in the literature (though see [11]). The conceptualization for the first choice is that we go to fixed "point" locations and see what is there; we are not sampling a point pattern. We model a surface over a domain D which captures the probability of presence at every location in D. The conceptualization for the second is that we identify an area of interest D and, theoretically, we census it completely for all of the occurrences of the point pattern (though in practice we never have the sampling effort to a study region completely). We model an intensity which, using the definition above, provides a probability of presence for a given A. The intensity surface can be normalized to a density surface under which the probability of an event at a "dimensionless" point is 0. That is, this density has nothing to do with modeling a Bernoulli trial at a point by specifying a probability of presence at the point, hence a probability of presence surface.

Furthermore, if presented with a collection of plots and observed presence/absence for those plots, one would not model the data as a point pattern. No point pattern was observed; there is no way to model an intensity. We would treat the plots as points in space and use a version of the foregoing presence/absence regression models. To reconcile the differences above it may be useful to think more carefully about what the distribution of a species looks like within a specified region, D and the associated implications. See [11] for further discussion in this regard.

4. ISSUE (III): THE EFFECT OF DEPENDENCE VS. INDEPENDENCE AT SITE LEVEL

With regard to assessing species distribution, this issue concerns the effect of modeling species independently rather than jointly within site. For example, with presence/absence data, stacking may tend to overestimate probability of presence for each species at a site. Hence the number of presences, the richness, at a site [13, 9] may be overestimated. This can be potentially more problematic when a large number of species are examined.

Specifically, [13] offer the following criticisms of stacked species distribution modeling: (i) without adding a dispersal filter (e.g., seed dispersal pathways) it may incorrectly predict species in areas that appear environmentally suitable but that are outside their colonizable or historical range; (ii) it does not consider any constraints based on the carrying capacity of the local environment which determine the maximum number of species that may co-occur; and (iii) it does not explicitly consider any rules based on biotic interactions that control species co-occurrences and can exclude species from a community. As a result, it is anticipated that too many species can be predicted to occur in a geographical unit by stacked species distribution models.

We offer a stochastic perspective through formalization of species richness. Species richness records the number of distinct species present at a site and is commonly used to characterize species distributions at sites. With the foregoing notation, the observed richness at site $\boldsymbol{s}$ is $\texttt{Rich}(\boldsymbol{s}) = \sum_{j=1}^{S} 1_j(\boldsymbol{s})$. To be clear, $1_j(\boldsymbol{s})$ is the indicator of whether species j is present at location $\boldsymbol{s}$. Further, $\{1_j(\boldsymbol{s}), j = 1, 2, \ldots, S\}$ does not constitute a multinomial trial but, rather, a set of dependent Bernoulli trials. Whether we model species independently using stacked species distribution models or dependently using JSDM's, $E(\texttt{Rich}(\boldsymbol{s})) = \sum_{j=1}^{S} E(1_j(\boldsymbol{s})) = \sum_{j=1}^{S} P(Y_j(\boldsymbol{s}) = 1) = \sum_{j=1}^{S} p_j(\boldsymbol{s})$ with forms for $p_j(\boldsymbol{s})$ supplied above. Though the forms are the same, these expectations need not agree since, following the argument of the previous paragraph, the $p_j(\boldsymbol{s})$ are expected to be different under an independence model vs. a JSDM. Probabilistically, because the joint model considers the data for all of the species at a site while the individual models consider the data only for the individual species at the site, unconstrained by the overall presence/absence at the site, intuitively, we might anticipate the latter expectations to be larger, suggesting prediction of higher richness using a stacked species distribution model.

Turning to the second moments, we expect to incorrectly estimate uncertainty in richness when the indicator variables in the sum are not independent. That is, $\text{Var}(\texttt{Rich}(\boldsymbol{s})) = \text{Var}(\sum_{j=1}^{S} 1_j(\boldsymbol{s}))$ should reflect the chance of joint presence or absence. Formally, $\text{Var}(\texttt{Rich}(\boldsymbol{s})) = \text{Var}(\sum_{j=1}^{S} 1_j(\boldsymbol{s})) = \sum_{j=1}^{S} \text{Var}(1_j(\boldsymbol{s})) + 2\sum_{j<j'} \text{Cov}(1_j(\boldsymbol{s}), 1_{j'}(\boldsymbol{s}))$. However, in obvious notation, $\text{Cov}(1_j(\boldsymbol{s}), 1_{j'}(\boldsymbol{s})) = p_{(j,j')}(\boldsymbol{s}) - p_j(\boldsymbol{s})p_{j'}(\boldsymbol{s})$. Departure from independence will affect this term and there how departure from independence can affect the variance in observed richness. Finally, the above is all in the context of a single site so the same conclusions apply whether we are building a spatial or a nonspatial specification.

As a last thought here, perhaps the most direct way to demonstrate the benefit of the joint modeling is with conditional prediction. This strategy does not depend upon whether or not the model fitting was done spatially. At a site, suppose we attempt to predict presence/absence for a species, given we know the presence/absence state of some other species at that site. The conditional prediction probabilities will be suitably adjusted given this information. The model for the species under stacking will ignore this information. See [34] in this regard.

5. ISSUE (IV): INTERPRETATION OF DEPENDENCE UNDER LATENT MULTIVARIATE NORMAL DISTRIBUTIONS

This issue concerns the interpretation of species dependence under the use of latent multivariate normal specification. As pointed out in Section 2, the pairwise associations arising under the latent multivariate normal have little to do with the actual realization of Y_i at site i. Perhaps more importantly, the pairwise correlations between species arising under the normal model provide little understanding of the nature of/strength of dependence between species. For a pair of species, envisioning a 2×2 table for presence/absence, the odds ratio provides a useful tool for learning about species dependence with regard to presence/absence. Specifically, a positive log odds ratio captures sympatry, i.e., encouraging joint occurrence or joint absence. A negative log odds ratio captures allopatry, i.e., discouragement of co-occurrence. As an aside, since independence modeling underlies stacked species distribution models, such models will not be able to capture sympatric or allopatric behavior for pairs of species. [12] provide a full discussion of the role of odds ratios in interpretation of species dependence in JSDMs. Here, we extract a few thoughts.

For the JSDMs above, again, dependence across species is captured through the pairwise correlation between species in the latent bivariate normal distribution. We do not model the 2×2 table of probabilities, $p_{a,b}^{(j,j')}, a, b = 0, 1$ associated with species j and j' directly but, rather, we model the parameters in the latent multivariate normal distribution and, as a result, each of these probabilities is a function of these parameters.

However, there is no direct connection between say $\rho^{(j,j')}$, the correlation in the latent multivariate normal between species j and species j', and the odds ratio associated with the induced 2×2 table of joint probabilities for the species pair, $j, j')$ at site i. Specifically, suppose the latent bivariate normal distribution for $\begin{pmatrix} Z_{ij} \\ Z_{ij'} \end{pmatrix}$ has mean $\begin{pmatrix} \mu_i^{(j)} \\ \mu_i^{(j')} \end{pmatrix}$ and correlation matrix $\begin{pmatrix} 1 & \rho^{(j,j')} \\ \rho^{(j,j')} & 1 \end{pmatrix}$. Then, the odds ratio for species j and j' at site i,

$$\begin{aligned} \theta_i^{(j,j')} &= \frac{p_{i,00}^{(j,j')} p_{i,11}^{(j,j')}}{p_{i,10}^{(j,j')} p_{i,01}^{(j,j')}} \\ &= \frac{P(Z_{ij} < 0, Z_{ij'} < 0)P(Z_{ij} \geq 0, Z_{ij'} \geq 0)}{P(Z_{ij} \geq 0, Z_{ij'} < 0)P(Z_{ij} < 0, Z_{ij'} \geq 0)}. \end{aligned} \tag{5.1}$$

The expressions for the double integrals in (5.1) show that each probability is a function of $\mu_i^{(j)}$, $\mu_i^{(j')}$, and $\rho^{(j,j')}$. [12] prove that $\theta_i^{(j,j')}$ is non-decreasing in $\rho^{(j,j')}$ for fixed $\mu_i^{(j)}$ and $\mu_i^{(j')}$. However, in the presence of $\mu_i^{(j)}$, $\mu_i^{(j')}$, the latent correlations do not determine the strength/magnitude of the odds ratios.

Specifically, this result should be applied to $W_{ij} = Z_{ij} - \mu_i^{(j)}$ where say, $\mu_i^{(j)} = \mathbf{X}_i^T \boldsymbol{\beta}_j$ and $W_{ij'} = Z_{ij'} - \mu_i^{(j')}$ where again, $\mu_i^{(j')} = \mathbf{X}_i^T \boldsymbol{\beta}_{j'}$. As a result, $P(Z_{ij} < 0, Z_{ij'} < 0) = P(W_{ij} < c_{ij}, W_{ij'} < c_{ij'})$, where $c_{ij} = -\mathbf{X}_i^T \boldsymbol{\beta}_j$ and $c_{ij'} = -\mathbf{X}_i^T \boldsymbol{\beta}_{j'}$, is non-decreasing in $\rho^{(j,j')}$ for any $\mathbf{X}_i$, $\boldsymbol{\beta}_j$, and $\boldsymbol{\beta}_{j'}$ and therefore so is the associated odds ratio, $\theta(\mathbf{X}_i, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'})$. As a result, for a given $\rho^{(j,j')}$, we can see the response of $\theta(\mathbf{X}_i, \boldsymbol{\beta}_j, \boldsymbol{\beta}_{j'})$ to changes in $\mathbf{X}_i$ for given coefficient vectors; we can understand how the odds ratio varies across environmental niches. In different words, JSDMs *disentangle* the role of the environment from the role of biotic interactions in the model specification. With these models, odds ratios provide a measure of association that *unifies* the effects of the biotic and abiotic conditions while enabling assessment of the effect of each on the association.

6. ISSUE (V): INTERPRETATION OF CLUSTERING OF SPECIES

This issue concerns interpretation for joint species distribution modeling specifications which impose clustering of species. When S is large, it is natural to attempt to cluster the species, here seeking data-driven clustering rather than say taxonomic or morphological clustering. Further, we seek model-based clustering rather than ad hoc clustering. With independent sites, such clustering has been proposed by [27]. Can we attach useful interpretation to the resulting clustering? Suppose we include spatial dependence between sites and again seek model-based clustering. An approach for such clustering has been proposed by [26]. Again, can we attach useful interpretation to the resulting clustering?

Continuing our notation for L^F_{ij} and L^R_{ij} above, we have $L^F_{ij} = \mathbf{X}_i^T \boldsymbol{\beta}_j$ where $\mathbf{X}_i$ denotes the vector of environmental covariates associated with site i and $\boldsymbol{\beta}_j$ is a species-specific coefficient vector. Collecting to a vector for site i, we can write $\mathbf{L}^F_i = \mathbf{B}\mathbf{X}_i$ where $\mathbf{B}$ is an $S \times p$ matrix whose jth row provides the regression coefficients for species j. Similarly, collect the L^R_{ij} to a vector $\mathbf{L}^R_i$ which is to be modeled as an $S \times 1$ vector of random effects. Under independence of sites, these vectors are independent and identically distributed as multivariate normals, say $\mathbf{L}^R_i \sim MVN(\mathbf{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\Sigma}$ is an $S \times S$ covariance matrix.

In working with binary responses, to be able to identify the coefficients in $\mathbf{B}$, we need to work with a correlation matrix rather than a covariance matrix. So, in model fitting we would set $\mathbf{R} = \mathbf{D}^{-1/2}\boldsymbol{\Sigma}\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is the diagonal matrix consisting of the diagonal elements of $\boldsymbol{\Sigma}$. As an aside, with regard to model fitting, this enables adaptation of the data-augmentation algorithm proposed by [8] for multivariate probit regression and is known in the literature as the parameter-expansion data-augmentation (PX-DA) algorithm [18, 17].

$\boldsymbol{\Sigma}$ has $S(S+1)/2$ distinct entries and with S on the order of 10^2 or 10^3 as above, it becomes infeasible to infer about $\boldsymbol{\Sigma}$. The solution that has been adopted in the fully model-based JSDM literature is to employ a dimension reduction in the form of a so-called latent factor analysis [4, 23]. We write L^R_{ij} in the form $\boldsymbol{\lambda}_j^T \boldsymbol{\eta}_i$ where each of these vectors is $r \times 1$ with $r << S$. (In applications typically r is at least 3 but at most 10.) As a result, we can write $\mathbf{L}^R_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i$ where $\boldsymbol{\Lambda}$ is $S \times r$. We envision r latent factors where the entries in $\boldsymbol{\eta}_i$ are r independent $N(0,1)$ variables and the rows of $\boldsymbol{\Lambda}$ provide the so-called factor loadings for the collection of species. The induced covariance matrix for $\mathbf{L}^R_i$ is $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$, creating the dependence structure between the species, that is $(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T)_{jj'}$ is the covariance between species j and j'. Since $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T$ is of rank r, not full rank, a diagonal matrix $\mathbf{V}$ with positive entries $V_{jj} = \sigma^2_j$ is added, yielding the diagonally dominant matrix, $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \mathbf{V}$ as the full rank approximation to $\boldsymbol{\Sigma}$. This corresponds to adding pure error to the model for the Z's, that is, to adopting model (ii) in Section 2 for the Z_{ij}'s.

We now have rS unknowns in $\boldsymbol{\Lambda}$ with S more in the $\mathbf{V}$ matrix.[2] So, the number of unknowns is now order S rather than order S^2, achieving the desired dimension reduction. It is well known that the $\boldsymbol{\Lambda}$ matrix is not identified. Various strategies have been proposed in the literature to deal with this issue; [see e.g., 23]. However, [27] introduce model based clustering in the specification of $\boldsymbol{\Lambda}$ to address the identifiability problem. It is implemented through the stick-breaking representation of the Dirichlet process [25] which provides specification of random discrete distributions and, therefore, results in a tie when two observations take on the same discrete value. To be clear, this approach enables ties in the $\boldsymbol{\lambda}_j$ vectors. It means that the S rows in $\boldsymbol{\Lambda}$ will not all be unique. Specifically, in a Markov chain Monte Carlo model fitting implementation, at each iteration the Dirichlet process yields a random number $(< S)$ of unique choices for the rows of $\boldsymbol{\Lambda}$. As a result, ties are created in the random effects structure and, for each iteration, the number of distinct rows in $\boldsymbol{\Lambda}$ is the number of clusters for the species associated with that iteration.

So, the clustering resulting from modeling the rows of $\boldsymbol{\Lambda}$ through a Dirichlet process is not clustering the species by their means since each species gets its own vector of regression coefficients from $\mathbf{B}$. Rather, it is clustering on the residual covariance structure. If $\boldsymbol{\lambda}_j = \boldsymbol{\lambda}_{j'}$, then the row entries for $\mathbf{Z}_j$ and $\mathbf{Z}_{j'}$ in $\boldsymbol{\Sigma}^*$ are identical. In other words, species j and j' have the same dependence structure with all other species, adjusted for the regressors. If pairwise residual dependence is viewed as a surrogate for pairwise species interaction, then, we might view common dependence with other species as a surrogate for common interaction with other species.

There has been some recent work to cluster the $\boldsymbol{\beta}_j$'s, i.e., to cluster the coefficient vectors across species [16, 14]. The Dirichlet process modeling that has been employed for the second order dependence structure can be adopted for the first order mean structure if such clustering is of interest. However, since no identifiability challenges are raised with regard to the $\boldsymbol{\beta}_j$'s, such clustering is not *needed* for the fitting the means.[3] In the hierarchical setting, an $S \times p$ matrix variate normal distribution is adopted for $\mathbf{B}$. In fact, it is greatly simplified to provide a vague independent normal distribution for each of the entries in $\mathbf{B}$ [see 27, for details].

Next, we bring in space and spatial dependence to the clustering problem, following the notation of Section 2. Again, we envision a presence/absence variable, $Y_j(\mathbf{s})$ for species j at every location, $\mathbf{s}$, in study region D. With regard to the Z's, again we have: (i) $Z_j(\mathbf{s}) = L^F_j(\mathbf{s}) + L^R_j(\mathbf{s})$ and (ii) $Z_j(\mathbf{s}) = L^F_j(\mathbf{s}) + L^R_j(\mathbf{s}) + \epsilon_j(\mathbf{s})$. Under either (i) or (ii), we envision a multivariate spatial process for $\mathbf{Z}(\mathbf{s})$. That is, we envision dependence within the components/species at a

[2] In practice, setting $\mathbf{V} = \sigma^2\mathbf{I}$ typically provides an adequate approximation.

[3] There is no identifiability problem for the βs because the Xs are observed unlike with the λs where the ηs are not observed.

given $\mathbf{s}$ but also, spatial dependence between $\mathbf{Z}(\mathbf{s})$ and $\mathbf{Z}(\mathbf{s}')$. We use the functional specification, $Y_j(\mathbf{s}) = I(Z_j(\mathbf{s}) > 0)$ to impart spatial dependence for the Z's to the Y's. Such specification requires an $S \times S$ cross-covariance function, say $C(\mathbf{s}, \mathbf{s}')$ which is such that $(C(\mathbf{s}, \mathbf{s}'))_{jj'} = \text{cov}(Z_j(\mathbf{s}), Z_{j'}(\mathbf{s}'))$ [2]. Under (i) or (ii) it becomes $\text{cov}(L^R_j(\mathbf{s}), L^R_{j'}(\mathbf{s}'))$.

So, $\mathbf{L}^R(\mathbf{s})$ becomes an S-dimensional Gaussian process over D where, again, we consider S to be large. Coregionalization [31, 2] is a convenient way to introduce dimension reduction here; we write $\mathbf{L}^R(\mathbf{s})$ as a linear transformation of a lower dimensional (say r) process. Analogous to the nonspatial case, we write $L^R(\mathbf{s}) = \mathbf{\Lambda}\boldsymbol{\eta}(\mathbf{s})$ where $\mathbf{\Lambda}$ is again, $s \times r$ (with $r << S$ and now $\boldsymbol{\eta}(\mathbf{s})$ is an r-dimensional Gaussian process with its own $r \times r$ cross-covariance function, say $C_{\boldsymbol{\eta}}(\mathbf{s}, \mathbf{s}')$. The induced cross-covariance function for $\mathbf{L}^R(\mathbf{s})$, hence for $\mathbf{Z}(\mathbf{s})$, is $\mathbf{\Lambda} C_{\boldsymbol{\eta}}(\mathbf{s}, \mathbf{s}')\mathbf{\Lambda}^T$.

Choices for $\boldsymbol{\eta}(\mathbf{s})$ include supplying an r-dimensional cross-covariance function but, with dependence induced between species through $\mathbf{\Lambda}$, independent components in $\boldsymbol{\eta}(\mathbf{s})$ are sufficient. In fact, as noted in [26], the components can be r independent replicates of a Gaussian process with common correlation function $\rho(\mathbf{s}, \mathbf{s}'; \theta_\eta)$. As a result, the induced cross-covariance function for $\mathbf{L}^R(\mathbf{s})$, hence for $\mathbf{Z}(\mathbf{s})$ simplifies to $\rho(\mathbf{s}, \mathbf{s}'; \theta_\eta)\mathbf{\Lambda}\mathbf{\Lambda}^T$. As far as specification for $\mathbf{\Lambda}$, with interest in clustering, the same specification for $\mathbf{\Lambda}$, as in the nonspatial case, using a Dirichlet process for the rows, can be employed.

Returning to interpretation, again we are clustering on the rows of $\mathbf{\Lambda}$; we are clustering on the residual covariance structure. If $\boldsymbol{\lambda}_j = \boldsymbol{\lambda}_{j'}$, then the row entries for $\mathbf{Z}_j$ and $\mathbf{Z}_{j'}$ in $\mathbf{\Sigma}^*$ are identical; species j and j' have the same *local* dependence structure with all other species. In addition, under the dimension-reduced cross-covariance function for $\mathbf{Z}(\mathbf{s})$ and $\mathbf{Z}(\mathbf{s}')$, $\text{cov}(Z_j(\mathbf{s}), Z_h(\mathbf{s}')) = \text{cov}(Z_{j'}(\mathbf{s}), Z_h(\mathbf{s}'))$ for all $h \neq j, j'$. The spatial modeling adds the further interpretation that decay in spatial dependence, in terms of distance, for species j with all other species is identical to that for species j' with all other species.

7. CLOSING COMMENTS

It is useful to note that the study design may add an extra level to the data, e.g., species occur on trees with trees located within sites as in [20]. Suppose then we add a subscript to the Y's, i.e., Y_{ikj} with design level k *nested* within design level i. The design need not be balanced, we can have $k = 1, 2, \ldots, n_i$. Now, the latent Z process becomes Z_{ikj}, with analogy to (i) and (ii) above. Now, we can have $L^F_{ikj} = \mathbf{X}_i^T \boldsymbol{\beta}_j + \mathbf{W}_{ik}^T \boldsymbol{\gamma}_j$, incorporating both design level i and design level k fixed effects. Similarly, we can have $L^R_{ikj} = \boldsymbol{\lambda}_{1j}^T \boldsymbol{\eta}_i + \boldsymbol{\lambda}_{2j}^T \boldsymbol{\omega}_{ik}$, incorporating design level i and design level k random effects. All of the foregoing discussion involving issues (i)–(v) above, both spatial and nonspatial, can be elaborated to this setting.

We anticipate that areas for future work will include further development for: (i) issues of missed detection or misclassification [30], (ii) bringing in trait information including intra-specific trait variation [24, 19], (iii) introduction of dynamics, i.e., data over time as well as over space [28], (iv) data types beyond presence/absence, e.g., abundance or composition data [9], and, perhaps most importantly, (v) faster and more efficient computation [33, 28, 21]. Evidently, there is still much life in attempting to explain joint distribution of plant life.

ACKNOWLEDGEMENTS

Thanks go to Jim Clark, Otso Ovaskainen, Erin Schliep, Shinichiro Shirota, John Silander, Daniel Taylor-Rodriguez, and Gleb Tikhonov for numerous conversations which opened up the issues taken up here and led to the attempt to bring some further clarity to them.

Accepted 21 July 2022

REFERENCES

[1] Aarts, G., Fieberg, J. and Matthiopoulos, J. (2012). Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution* **3**, 177–187.

[2] Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data.* 2nd edn. Chapman & Hall/CRC, Boca Raton, FL, USA. MR3362184

[3] Barner, A. K., Coblentz, K. E., Hacker, S. D. and Menge, B. A. (2018). Fundamental contradictions among observational and experimental estimates of non-trophic species interactions. *Ecology* **99**, 557–566.

[4] Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika* 291–306. https://doi.org/10.1093/biomet/asr013. MR2806429

[5] Blanchet, F. G., Cazelles, K. and Gravel, D. (2020). Co-occurrence is not evidence of ecological interactions. *Ecology Letters* **23**, 1050–1063.

[6] 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.

[7] 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. *JRSS-C* **60**, 757–776. https://doi.org/10.1111/j.1467-9876.2011.00769.x. MR2844854

[8] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika* **85**(2) 347–361.

[9] Clark, J. S., Nemergut, D., Seyednasrollah, B., Turner, P. J. and Zhang, S. (2017). Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data. *Ecological Monographs* **87**(1) 34–56.

[10] De Oliveira, V. (2000). Bayesian prediction of clipped Gaussian random fields. *CompStatDataAnal* **34**, 299–314.

[11] Gelfand, A. E. and Shirota, S. (2019). Preferential sampling for presence/absence data and for fusion of presence/absence data with presence–only data. *Ecological Monographs* **89**(3).

[12] Gelfand, A. E. and Shirota, S. (2021). On the role of odds ratios in joint species distribution modeling. *Environmental and Ecological Statistics* (forthcoming).

[13] 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.

[14] Hefley, T. (2020). Model selection for ecological community data using tree shrinkage priors. ArXiv preprint, 2005.14303.

[15] Hooten, M. B., Johnson, D. S., McClintock, B. T. and Morales, J. M. (2017). *Animal Movement: Statistical Models for Telemetry Data*. CRC Press, Boca Raton. MR3889901

[16] Johnson, D. S. and Sinclair, E. H. (2017). Modeling joint abundance of multiple species using Dirichlet process mixtures. *Environmetrics* **28**, e2440. https://doi.org/10.1002/env.2440. MR3634110

[17] 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. https://doi.org/10.1198/004017008000000064. MR2439877

[18] Liu, J. S. and Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association* **94**(448) 1264–1274. https://doi.org/10.2307/2669940. MR1731488

[19] 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**(4) 428–436.

[20] Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T. and Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters* **20**(5) 561–576.

[21] Pichler, M., Boreux, V., Klein, A. M., Schleuning, M. and Hartig, F. (2019). Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. *Methods in Ecology and Evolution* **11**, 281–293.

[22] 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.

[23] Ren, Q. and Banerjee, S. (2013). Hierarchical factor models for large spatially misaligned data: a low-rank predictive process approach. *Biometrics* **69**(1) 19–30. https://doi.org/10.1111/j.1541-0420.2012.01832.x. MR3058048

[24] Schliep, E. M., Gelfand, A. E., Mitchell, R. M., Aiello-Lammens, M. E. and Silander Jr, J. A. (2018). Assessing the joint behaviour of species traits as filtered by environment. *Methods in Ecology and Evolution* **9**(3) 716–727.

[25] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica sinica* 639–650. MR1309433

[26] Shirota, S., Gelfand, A. E. and Banerjee, S. (2018). Spatial Joint Species Distribution Modeling using Dirichlet Processes. *Statistica Sinica*. MR3932512

[27] Taylor-Rodriguez, 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**(4) 939–967. https://doi.org/10.1214/16-BA1031. MR3724974

[28] Tikhonov, G. (2018). Bayesian latent factor approaches for modeling ecological species communities. Technical Report, Ph.D. Thesis, Faculty of Biological and Environmental Sciences, University of Helsinki.

[29] 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**(6) 627–637.

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

[31] Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications*. 3rd edn. Springer-Verlag.

[32] Warton, D. I. and Shepherd, L. C. (2010). Poisson point process models solve the pseudo-absence problem for presence-only data in ecology. *AnnalsAppldStats* **4**, 1383–1402. https://doi.org/10.1214/10-AOAS331. MR2758333

[33] Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. and McCarthy, M. A. (2018). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution*.

[34] Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R. and McCarthy, M. A. (2021). Defining and evaluating predictions of joint species distribution models. *Methods in Ecology and Evolution* **12**, 394–404.

[35] Zobel, M. (1997). The relative of species pools in determining plant species richness: an alternative explanation of species coexistence? *Trends in Ecology & Evolution* **12**(7) 266–269.

Alan E. Gelfand. Department of Statistical Science, Duke University, USA. E-mail address: alan@duke.edu