Variance partitioning-based priors for species distribution models

Luisa Ferrari[1], Massimo Ventrucci[1]

[1] University of Bologna, Italy

E-mail for correspondence: luisa.ferrari5@unibo.it

Abstract: Species distribution models for community ecology data are usually quite complex because of the need to account for many abiotic factors, with potentially non-linear effects, as well as residual spatio-temporal correlation, which capture abiotic phenomena. The use of variance partioning-based priors recently emerged in the literature could be an effective and intuitive strategy to deal with the high flexibility often required in this field. In this work, we discuss how to extend this new class of priors to species distribution models containing spatial and temporal smooth effects.

Keywords: Bayesian species distribution models; Intuitive priors; IGMRF.

1 Introduction

Surveys in the field of community ecology collect large datasets on the abundance of different species at certain locations and time points. Multiple factors are believed to influence abundance patterns. Species distribution models (SDM) are often expressed as generalized linear mixed models (GLMM) with fixed effects for the *abiotic* factors and random effects capturing the residual spatio-temporal correlation, reflecting the so-called *biotic* phenomena (e.g. predator-prey abundance cycles, species' spatial segregation, symbiotic or competitive relationships). Further complexity arises in a joint SDM framework, where several approaches to model between-species correlation structures have been proposed including latent variable models (Tikhonov et al. (2020)) and spatio-temporal basis functions (Hui et al. (2023)). All these approaches involve richly-parametrized GLMMs that require regularization to avoid overfitting.
Regardless of the chosen approach, it is often the case that ecologists have prior insight into the relative importance of each factor in explaining the response. As a consequence, a Bayesian approach would be particularly beneficial in these applications to impose a regularization based on prior information. We argue that thinking in terms of quantities like proportions of variance due to the individual model components is more intuitive than considering the original variance

This paper was presented at the 38th International Workshop on Statistical Modelling (IWSM), Durham University, UK, 14–19 July 2024.

parameters. This can be achieved using *variance partitioning* (VP)-based priors (Franco-Villoria et al. (2022), Fuglstad et al. (2020)) which make use of a reparametrization of the variance parameters of a mixed model into a total variance and a simplex vector containing the proportional contributions to the total variance from each model component.
The common advantage of VP priors consists in the fact that it is much easier to introduce prior information in the model using these new parameters. One can easily implement very different types of prior knowledge on the variance contributions of the different model components, based on what is known about the case study at hand. As an example, assuming a Uniform distribution on the simplex would reflect ignorance a priori about the relative importance of each term, while a Dirichlet inducing sparsity on the proportions of variance would provide a suitable solution to perform variable selection in sparse linear regression. Furthermore, a hierarchical decomposition of the total variance through subsequent splits can be chosen to favour shrinkage towards simpler model structures (Franco-Villoria et al. (2022), Fuglstad et al. (2020)).
The VP-based priors proposed so far only deal with specific effects, e.g. stationary or linear effects. Challenges arise in their extension to complex models, such as SDMs which often contain smooth effects of continuous covariates as well as Intrinsic Gaussian Markov random fields (IGMRFs) for spatial and time effects. The goal of this paper is to develop a unified VP framework applicable to more complex settings, such as SDMs.

2 Proposal

Consider the following SDM in which the linear predictor of a generic abundance response can be written as an additive model of P linear effects for the $X_1, \dots, X_P$ covariates, a smooth effect over spatial coordinates (S_1, S_2), and another smooth effect for time T. The smooth effects are both expressed using a finite-dimensional basis, $\mathbf{B}_S(\cdot)$ and $\mathbf{B}_T(\cdot)$, and a corresponding set of coefficients, $\mathbf{u}$ and $\mathbf{v}$ respectively:

$$\eta = \mu + \sum_{p=1}^{P} X_p \beta_p + \mathbf{B}_S(S_1, S_2)^{\mathrm{T}} \mathbf{u} + \mathbf{B}_T(T)^{\mathrm{T}} \mathbf{v}. \tag{1}$$

A latent Gaussian model is assumed on all coefficient sets, i.e. they are specified as Normally distributed with 0 mean and fixed precision matrix conditional upon a single scale parameter: $\beta_p|\sigma_p^2 \sim N(0, \sigma_p^2) \quad p = 1, \dots, P$, $\mathbf{u}|\sigma_S^2 \sim N(\mathbf{0}, \sigma_S^2 \mathbf{Q}_S^{-1})$, $\mathbf{v}|\sigma_T^2 \sim N(\mathbf{0}, \sigma_T^2 \mathbf{Q}_T^{-1})$. The VP parameters can then be defined as:

$$V = \sum_{p=1}^{P} \sigma_p^2 + \sigma_S^2 + \sigma_T^2 \qquad \boldsymbol{\omega} = \left[\frac{\sigma_1^2}{V}, \dots, \frac{\sigma_P^2}{V}, \frac{\sigma_S^2}{V}, \frac{\sigma_T^2}{V} \right] \tag{2}$$

The great advantage of VP-based priors comes from the possibility of assigning priors directly on the total variance in the linear predictor (i.e. V) and the set of proportions of variance due to each effect (i.e. $\boldsymbol{\omega}$). However, it is not guaranteed that these intuitive interpretations actually match the VP parameters in (2). This only occurs if all model components in (1) are processes on a comparable, *standardized* scale so that the elements of $\boldsymbol{\omega}$ actually represent the corresponding variance contributions.

For linear effects, it is sufficient to use the standardized version of X_p for $p = 1, ..., P$. However, it is not as simple for effects defined using a generic basis matrix, e.g. the spatial and temporal effects in this model. We propose a scaling procedure inspired by the work of Sørbye and Rue (2014) on IGMRFs that guarantees that the parameter of V and $\boldsymbol{\omega}$ match their intuitive interpretation. This is achieved by scaling each of the bases in the model by the square root of a term-specific constant C defined as the variance of the corresponding process conditional on $\sigma^2 = 1$ and marginalizing over the covariates' distribution. For example, the constant for the temporal effect is defined as:

$$C_T = \int_{t \in \mathcal{T}} \mathbf{B}_T(t)^{\mathrm{T}} \mathbf{Q}_T^{-1} \mathbf{B}_T(t) \cdot \pi(t) \, \mathrm{d}t \tag{3}$$

where $\mathcal{T}$ is the support of interest for variable T and $\pi(t)$ is its probability distribution. C_S is analogously defined using a given $\mathcal{S}$ support and $\pi(s_1, s_2)$ density. This scaling procedure can be viewed as a generalization of the standardization procedure used for linear effect, as C simplifies to the variance of the corresponding covariate in this case. We argue that VP-based priors can be safely employed only after scaling each term in the model according to this procedure. An advantage of the scaling procedure lies in the possibility of immediately evaluating the variance partition structure of the model considering the posterior distribution of the $\boldsymbol{\omega}$ vector. This is possible because after scaling each entry will represent the proportional contribution of a model component to the response variability. A challenging aspect in the scaling constant definition in Equation 3 is that it requires the choice of a distribution $\pi(\cdot)$ for the corresponding covariate. While it is reasonable to assume a Uniform distribution over the spatio-temporal support, this becomes a non-trivial choice in the case in which the procedure must be applied to other types of effects, such as smooth effects of continuous covariates.

3 Application

3.1 Data

The model defined in Equation (1) is applied to the NOAA-NEFSC fall bottom trawl survey dataset, studied in Hui et al. (2023) and publicly available at: https://github.com/fhui28/CBFM. The survey contains presence/absence data for 39 fish species from $N = 5892$ different space-time locations in the North-West Atlantic region, spanning a 20-year period. Figure 1 shows the study region with the number of species found in each location. Information about 5 environmental covariates is also available: surface temperature and salinity, bottom temperature and salinity, depth. A binary variable indicating the type of vessel collecting the data at each location can be used as an additional covariate.

3.2 Model and results

The model of Equation 1 is applied to each of the 39 species from the survey to illustrate how the proposed method provides a simple and intuitive way to study the contributions of different factors on the variability of an occurrence response. A logistic model is chosen to link the linear predictor to the binary presence-absence response for each species. The five environmental covariates

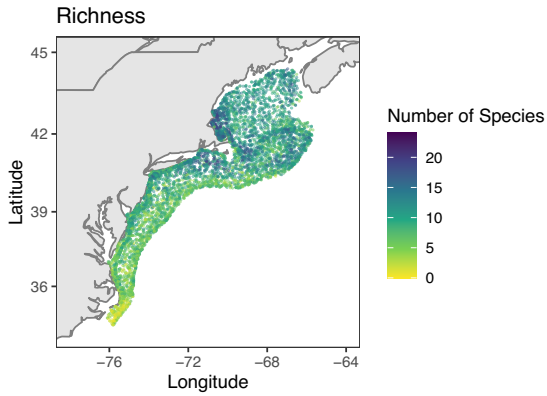

FIGURE 1. Number of different species, i.e. richness, detected in each of the locations from the survey.

and the vessel dummy are entered into the model with linear effects, following standardization. A 2-dimensional B-Spline basis with an Intrinsic CAR model (Besag et al. (1991)) precision matrix is used for the spatial effect, whose knots are equally spaced on a grid of 50x50km cells. A B-Spline with 20 basis functions is chosen for the temporal effect, with a 1^{st} order random walk prior on the coefficients. A Uniform distribution is assumed over the observed spatio-temporal support for the computation of the scaling constants C_S and C_T.

In this case study, the VP-based prior approach is used to reflect the assumption that not all effects are likely to affect the abundance of each species, but rather a few (species-specific) factors are assumed to be responsible for most of the variability. This assumption can be introduced through the choice of a symmetric Dirichlet prior on the vector of proportions: $\boldsymbol{\omega} \sim \text{Dir}(0.5)$. The marginal prior induced on each of the $\boldsymbol{\omega}$ elements is represented as a solid black line in the left panel of Figure 2; as we can see, this prior assigns most probability mass near 0 indicating that $\omega_j = 0$ (no effect) is favoured a priori. The prior specification is completed by a vague prior on the intercept μ and a Jeffreys on V.

The models are fitted using the R-INLA software. Thanks to scaling, the posterior distribution of $\boldsymbol{\omega}$ can directly answers questions about variance partitioning without further transformations. The left panel of Figure 2 shows the marginal posterior distributions of the proportions of variance $\boldsymbol{\omega}$ entries for a single species (*Weakfish*). The plot shows how the prior choice helps in the identification of the most important factors affecting occurrence as most factors are shrunk towards 0. The right panel shows the posterior median of $\boldsymbol{\omega}$ for six different species. Along with conclusions about individual species, this plot can help assess the variance partitioning for the community as a whole: for example, the spatial component appears to be a relevant term for all the species in this subset.

4 Discussion

This work proposes a new way to analyze SDMs that can incorporate prior knowledge about the relative importance of different factors affecting species abundance

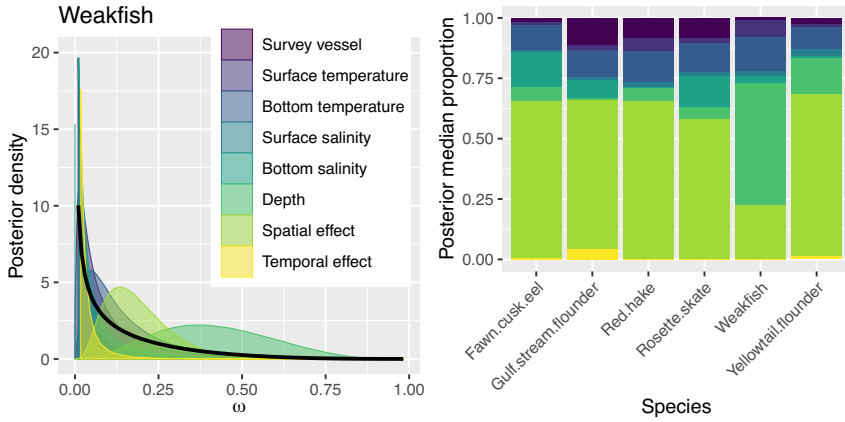

FIGURE 2. Left panel: comparison between the prior distribution on each ω_j (solid black line) and their posterior density for the *Weakfish* species. Right panel: posterior median of each ω_j for six different species.

and give immediate and intuitive posterior outputs about variance partitioning. The class of models of Equation 1 represents just an illustration of a larger theoretical framework developed to correctly apply VP-based priors to a broader class of SDMs, which can include for example smooth effects of abiotic factors, among others. Future challenges include exploring the application of VP-based priors in the context of joint species distribution models.

Acknowledgments: This work was funded by the European Union under the NextGeneration EU Programme within the Plan "PNRR - Missione 4 "Istruzione e Ricerca" - Componente C2 Investimento 1.1 "Fondo per il Programma Nazionale di Ricerca e Progetti di Rilevante Interesse Nazionale (PRIN)" by the Italian Ministry of University and Research (MUR), Project title: "METAbarcoding for METAcommunities: towards a genetic approach to community ecology (META2)", Project code: 2022PA3BS2 (CUP E53D23007580006), MUR D.D. financing decree n. 1015 of 07/07/2023.

References

Besag, J., York, J. and Mollié, A. (1991).Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1 – 20.

Franco-Villoria, M., Ventrucci, M. and Rue, H. (2021). Variance partitioning in spatio-temporal disease mapping models. *Statistical Methods in Medical Research*, **31**, 1566 — 1578.

Fuglstad, G.A., Hem, I.G., Knight, A., Rue, H. and Riebler, A. (2020). Intuitive joint priors for variance parameters. *Bayesian Analysis*, **15**, 1109 – 1137.

Hui, F. K., Warton, D. I., Foster, S. D. and Haak, C. R . (2023). Spatiotemporal joint species distribution modelling: A basis function approach. *Methods*

in Ecology and Evolution, **58**, 2150 – 2164.

Sørbye, S. H. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, **8**, 39 – 51.

Tikhonov, G., Opedal, Ø. H., Abrego, N., Lehikoinen, A., de Jonge, M. M., Oksanen, J. and Ovaskainen, O. (2020). Joint species distribution modelling with the R-package Hmsc. *Methods in ecology and evolution*, **11**, 442 – 447.