Latent factor models: a tool for dimension reduction in joint species distribution models

Daria Bystrova, Giovanni Poggiato, Julyan Arbel, Wilfried Thuiller

Chapter 1

Latent factor models: a tool for dimension reduction in joint species distribution models

1.1. Introduction

Understanding how species are distributed in space has been one of the main goals of Ecology [Von Humbolt, GUI 05]. In particular, investigating which factors drive species distributions within communities, across regions or along environmental gradients can improve our understanding of fundamental ecological processes underlying such patterns, as well as our ability to anticipate future biodiversity changes [GUI 17; THU 13]. When building models to explain and predict the distribution of organisms, we necessarily need to ask the same questions as the early biogeographers. It is now clear that three main conditions need to be met for a species to occupy a site and maintain populations [see Figure 1.1, PUL 00; LOR 04; SOB 07] :

– the species has to physically reach the site, i.e. to access the region [BAR 11];

– the abiotic environmental conditions (i.e. temperature, precipitation...) must be physiologically suitable for the species ;

– the biotic environment (interactions with other species) must be suitable for the species.

The first condition is a matter of species **dispersal** capacity from those areas previously occupied by the species. It includes the biogeographic history of the species, and thus all factors limiting its distribution from the place where it first originated,

Chapter written by Daria BYSTROVA, Giovanni POGGIATO, Julyan ARBEL and Wilfried THUILLER.

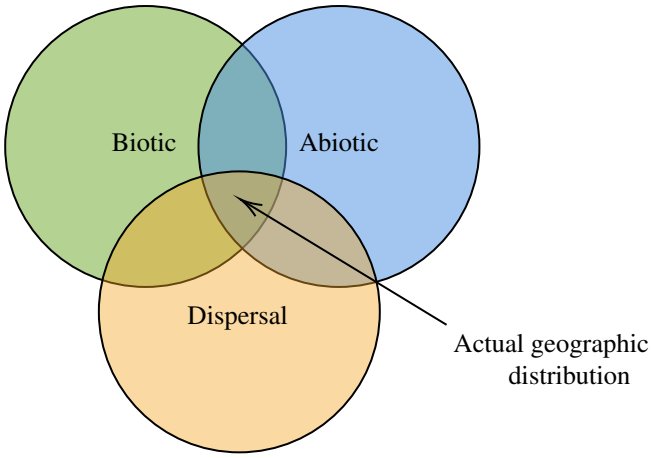

Figure 1.1 – The three factors that determine the actual distribution of a species [SOB 05].

such as barriers to migration, biotic and abiotic dispersal vectors, rare long distance dispersal, etc.

The second condition is the matter of abiotic **habitat suitability** for the target species, which means that the combination of abiotic environmental variables at the site - often referred to as environmental suitability - are within the range of environmental conditions that the species requires to grow and maintain viable populations. These suitable environmental conditions are what ecologists call the *environmental niche* [HUT 57].

The third condition concerns **biotic interactions**, i.e. interactions with other organisms, either positive (commensalism, mutualism) or negative (competition, predation), which themselves are influenced by the environment through their influence on all organisms in the local community.

From a statistical point of view, the most common tools to model how species are distributed in space are species distribution models (SDMs, [GUI 05]). There are a variety of SDMs that differ in their underlying statistical algorithms and flexibility [GUI 05; MER 14; GUI 17], but they all relate the presence or abundance, and sometimes the absence, of a species to a set of environmental variables and project this relationship in space and/or time. While SDMs have proven to be very useful and reliable in many different areas and fields [see YAT 18; GUI 17, for reviews], they also have well-known limitations and assumptions that run counter to ecological niche theory [GUI 00] and that may question the robustness of their predictions. A first major criticism of SDMs is that they model species independently of each other, making

the assumption that species respond individualistically to the environment. As a consequence, SDMs can only capture the implicit combined effect of both abiotic and biotic environments. Despite these limitations, researchers have also used SDMs to predict species communities in space and time. In that case, single species predictions are simply stacked together (e.g. stacked SDMs, [GUI 05]) by summing either the species' probabilities of occurrence [CAL 14] or the binary-transformed predictions [GUI 11]. In the end, going from single species predictions to species communities commonly relies on a two-step procedure without any consideration of error propagation and without a joint-estimation of all model parameters.

With the increasing availability of community data (thanks to new sampling techniques like environmental DNA (eDNA) metabarcoding [TAB 12]), researchers now aim to model community as whole, and not as the stacked response of species [CLA 14]. The species are then modelled together, giving birth to Joint Species Distribution Models (JSDMs) [POL 14; OVA 17; CLA 16]. These models estimate the relationship of each species with respect to environmental covariates through a regression, like SDMs, and additionally infer a correlation matrix among species from the regression residuals. This correlation matrix reflects species co-occurrence patterns not explained by the environmental predictors and may arise from model mis-specifications, missing covariates or, importantly species interactions. Since the number of parameters in the residual correlation matrix scale quadratically with the number of species, these methods are computationally challenging. Latent factor models, that provide a low-rank approximation of this matrix, have naturally raised as a computationally efficient solution for JSDMs [WAR 15]. In this chapter we present latent factor models in the context of JSDMs, emphasising their usefulness in community ecology. We apply latent factor models to plant species along 18 elevation gradients in the French Alps, belonging to the long-term observatory ORCHAMP (www.orchamp.osug.fr).

Within this book, two other chapters also develop methodologies for JSDMs: Chiquet et al. (Log-normal Poisson model) and Mortier et al. (Supervised component generalized linear regression and extensions). Chiquet et al. chapter focuses on the multivariate Poisson log-normal (PLN) model with abundance data, while ours essentially covers presence-absence data. Inference for this PLN model is done in a classical (non Bayesian) setting with a variational approximation, while we follow a Bayesian approach and use a Markov chain Monte Carlo algorithm to sample from the posterior distribution, thus offering posterior credible intervals. Mortier et al. chapter has a slightly different focus on how to combine predictors into components in order to lead to optimal learning. A classical (non Bayesian) approach is used, and the case study tackles abundance data.

1.2. Joint species distribution models

To study species distribution we relate a response variable $\mathbf{Y}_n$ to a set of p environmental covariates $\mathbf{X}_n = (X_{n\ell})_{\ell=1}^{p}$, at each site $n = 1, \ldots, N$. $\mathbf{Y}_n \in \mathbb{R}^S$ is a vector where each element Y_{ns} contains the observation for species $s = 1, \ldots, S$ at site n. Most JSDMs are based an extension of Generalised Linear Models (GLMs), where they assume that the response variable is distributed as F, whose mean is given by a regression term and a residual multivariate random effect: For species s at site n, this writes:

$$Y_{ns} \sim F(\mu_{ns}, \phi_s), \tag{1.1}$$

$$g(\mu_{ns}) = \beta_{0s} + t(\boldsymbol{\beta}_s)\mathbf{X}_n + e_{ns}, \tag{1.2}$$

$$\mathbf{e}_n \overset{\text{iid}}{\sim} \mathcal{N}_S(0, \boldsymbol{\Sigma}), \tag{1.3}$$

where F is the assumed distribution for the data with mean μ and dispersion parameter ϕ_s (that is usually not accounted for when modelling presence-absence data), and $t(\beta)$ denotes the transposition of β. Function g is called the *link function*. The vectors β_{0s} and $\boldsymbol{\beta}_s$ represent the intercept and regression coefficients for species s, that describe the relationship between each species and the environmental covariates. Thanks to these coefficients, we can therefore define the suitable environmental conditions for each species, the *environmental niche*. Note here that the environmental covariates could also integrate the abundance or presence-absence of species (REF). Residual correlations among species are captured by $\boldsymbol{\Sigma}$, a symmetric and positive-definite variance-covariance matrix (that has the constrain to be a correlation matrix for presence-absence data). The elements of $\boldsymbol{\Sigma}$ reflect species co-occurrence patterns that are not explained by the environmental predictors, and can arise from noise in the data, model mi-specification, missing predictors, and species associations.

The choice of the distribution F and the link function g depends on the response variable $\mathbf{Y}_n$ to be modelled. JSDMs typically model presence/absence, counts, biomass and many others due to the heterogeneity of ecological data and of the sampling campaigns. For presence-absence data, most models assume a Bernoulli distribution and a probit link function [see GLM MCC 89]. However, this is quite common to remplace the probit link function by a latent variable parameterisation [CHI 98] to make the model computationally more efficient. Since species community data may contain observations of species documented in multifarious ways (e.g. presence-absence and counts), several JSDMs have been implemented to address this challenge [OVA 17; CLA 17].

Interestingly, many JSDMs can model the regression coefficients hierarchically:

$$\boldsymbol{\beta}_s \overset{\text{iid}}{\sim} \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{V}). \tag{1.4}$$

This allows for a share of information across species on their response to the environment, so that the estimation of the niche of rarely observed species could 'borrow strength' from those of common species assuming that they do not behave fundamentally differently [OVA 11]. Moreover, it is possible to account for functional traits and/or phylogeny by including them in $\boldsymbol{\mu}$ and/or $\boldsymbol{V}$ [see e.g. OVA 20, Chapter 6 for detailed description].

1.3. Dimension reduction with latent factors

The model described above suffers from the "curse of dimensionality", since the covariance matrix scales quadratically with the number of species. Indeed, the number of free parameters of the covariance matrix when modelling S species is $S(S+1)/2$. For example, for $S = 100$ species, the number of parameters of the covariance matrix is $5\,000+$. Nowadays, dealing with large datasets that contain observational data over space and time, the number of modelled species can easily exceed several thousands, making inference challenging and endless computational times. Hence, there is a need for dimension reduction approaches in JSDMs.

To address this challenge, several authors proposed a low-rank approximation of the covariance matrix of JSDMs, through the use of latent factors [WAR 15]. Starting from the original model (1.2), we assume a factorized representation of the residual random effects $\mathbf{e}_{ns}$, as a product of factor loadings and latent factors:

$$\mathbf{e}_{ns} = t(\mathbf{T}_s)\mathbf{Z}_n \quad \text{where} \quad \mathbf{Z}_n \overset{\text{iid}}{\sim} \mathcal{N}_K(0, \mathbf{I}_K). \tag{1.5}$$

The vector $\mathbf{T}_s \in \mathbb{R}^K$ is called the *factor loading* of species s; the collection of $t(\mathbf{T}_s)$, $s = 1, \ldots, S$, constitutes the rows of the so-called *factor loadings* matrix $\mathbf{T}$ (of dimension $S \times K$). The Gaussian random vectors $\mathbf{Z}_n \in \mathbb{R}^K$ are called *latent factors*. Crucially, note that under this factorised representation of the residual random effect, the residual covariance becomes now: $\boldsymbol{\Sigma} = \mathbf{T}t(\mathbf{T})$. By taking the number of latent factors $K \ll S$, the parameters to be modelled are drastically reduced.

Latent factors $\mathbf{Z}_n$ can be seen as a set of unmeasured covariates at site n, and the factor loadings $\mathbf{T}_s$ as the response of species s to these unmeasured covariates. A common (or opposite) response to these unmeasured covariates introduces a positive (or negative) correlation between species.

A critical feature of this dimension reduction is to appropriately select the number of latent factors. On one hand, we need $K \ll S$ to reduce model complexity. On the other hand, we have to provide to the model the flexibility (that increases with a higher K) that is necessary to fully capture the required information from the data.

The number of factors controls the complexity of the model. The challenge is to find the appropriate number of factors such that the model is simple and tractable,

yet appropriately capturing the covariance structure. Interestingly, this question arises also in most multivariate analyses where an optimal number of components has to be chosen. There are several approaches to address this issue in a Bayesian framework. One way is to initially fix K and then run a model selection with a range of K values. This is typically done by using information-theoretic criteria such as the deviance information criterion (DIC) [SPI 02] or the Watanabe–Akaike information criteria (WAIC) [WAT 10].

1.4. Inference

These models could be fitted either in the maximum likelihood framework or in the Bayesian one. The key difference between the two approaches is that maximum likelihood methods consider the model parameters as fixed (but unknown) quantities, while in the Bayesian approach they are considered as random [ELL 04]. As a consequence, the Bayesian framework allows to introduce a prior information on the parameters, that might come from expert knowledge or previous studies. Bayesian methods also differ in the quantification of uncertainty: while maximum likelihood methods usually provide point parameter estimates and confidence interval, the Bayesian approach can provide the full distribution of the estimated parameters (the so-called posterior distribution).

Bayesian inference is particularly suitable in Ecology due to its flexibility and computational tractability when dealing with highly complex models. Indeed, modelling Nature is challenging due to the complexity and stochasticity of its underlying processes. This motivates the use of the Bayesian framework to analyse ecological data [CLA 06]. Introducing prior information in Bayesian models allows to incorporate various historical/external information and expert opinion for improving the models. Additionally, parameter estimations in these complex models are uncertain, and the Bayesian approaches are particularly suited for dealing with such an uncertainty.

As shortly mentioned above, a Bayesian framework implies to select suitable priors for model parameters: $\boldsymbol{\beta}_s, \boldsymbol{\Sigma}$. Incorporating prior information in the model could improve parameter estimates, but if priors are specified incorrectly, they could potentially bias the model, especially when only few observations are available. In practice, it is quite often difficult to specify correctly prior distributions reflecting prior knowledge. In this chapter we present the case of more widespread or non-informative priors, but informative choices are also possible [CLA 16; BYS 20]. The prior distribution for regression coefficients is usually a multivariate normal and an inverse Wishart for the covariance matrix, and all hyperpriors are chosen to be vague.

1.5. Ecological interpretation of latent factors

We described latent factors from the mathematical point of view, but what do they imply in term of ecological hypotheses and interpretation? In model (1.3), we described the residuals $\mathbf{e}_i$ for site i as a Gaussian vector whose covariance matrix $\mathbf{\Sigma}$ was unconstrained. This correlation reflects species co-occurrence patterns that are not explained by the environmental predictors, and may arise from model misspecifications, missing covariates or species associations. We can also leverage on the non-independence between species to improve the co-occurrence and conditional predictions (see Section 1.6). Latent factors not only allow to reduce the dimension of the model and to deal with a larger number of species, but they also yield crucial ecological insights.

First of all, this new representation still makes it possible to infer the residual covariance matrix among taxa: as shown previously, latent factors $\mathbf{T}$ factorise the covariance matrix into $\mathbf{\Sigma} = \mathbf{T}t(\mathbf{T})$. Therefore, species that are highly correlated have similar latent loadings. How can these latent loadings be interpreted?

In the latent factor representation of JSDMs it is natural to think as the term $t(\mathbf{T}_s)\mathbf{Z}_n$ in Equation (1.5) as a random effect term of a vector of latent covariates $\mathbf{Z}_n$ and their related species-specific coefficients $\mathbf{T}_s$. These latent covariates can be seen as *missing environmental predictors* and therefore provide a means of solving the longstanding problem of missing covariates modelling. So doing, species with similar latent loadings share the same response to missing covariate and are thus expected to share similar occurrence patterns. Therefore, they are more correlated.

Latent factors can also be thought as *ordination axes*, that represent the main axes of (co)variations of abundances across taxa. By forcing the number of latent factors to $K = 2$, it is possible to visualise on a biplot both the sites ordination, thanks to the latent variables $\mathbf{Z}_n$, and the ordination of taxa, with the latent loadings $\mathbf{T}_s$. Therefore, species that have close latent loadings will be close in the low dimensional space represented by the biplot, and therefore highly correlated. By evaluating this model-based ordination before and after the inclusion of measured environmental covariates, we can understand how much the co-occurrence suggested by an unconstrained model (i.e. without environmental covariates) can be explained by a shared response to environmental covariates.

1.6. On the interpretation of JSDMs

Although JSDMs are receiving increasing attention, there has been a lack of clarification on both the ecological processes they incorporate and on their specific commonalities and advantages with respect to SDMs. Since JSDMs infer a correlation matrix from the residual, it is tempting to think these residual correlations can inform

about biotic interactions [POL 14] or even that JSDMs "account for biotic interactions in species distribution models" [WIL 19]. As highlighted by [POG 21], these tempting ideas should be avoided. JSDMs can provide additional information on species co-occurrences, but cannot separate the biotic and the abiotic effects, and their predictions on species distribution inevitably coincide with those of SDMs. However, JSDMs have the great advantage of leveraging on the residual correlation matrix to provide conditional predictions, which can be of great help in empirical studies, as we show in the case study below.

1.7. Case study

1.7.1. *Introduction of the dataset*

We present hereafter an application of latent factor models to a plant community dataset recently published by [MAR 20]. The data are being collected within ORCHAMP, a long-term observatory of mountain ecosystems aiming to observe, understand and model biodiversity and ecosystem functioning over space and time. ORCHAMP is built around multiple elevational gradients that range from about 900 m to 3000 m, and have been chosen to have a homogeneous exposure and slope along the gradient, a typical vegetation for the elevation levels (with woods dominating the lower parts and alpine meadows the higher parts), so that all the gradients as a whole are representative of the environmental and topographical variability of the French Alps. Between 2016 and 2018, at least five sampling plots were installed along 18 gradient, with an average of 200m elevation difference, for a total of 99 plots (Figure 1.2). Here, we study the response of plant species to climate, the physico-chemistry properties and the microbial activities of the soil. We applied latent factor models to a selection of 44 plant species, whose occurrences were recorded in at least 20 sites over the 99 sites, together with climatic variables, soil physico-chemical properties and exoenzymatic activities. Latent factor models are particularly suitable to study the response of plant communities for the reasons described above. We aim to understand which species share the same response to the environment, and how eventual changes of climate and soil could affect these plants. Moreover, we are interested in the inference of the residuals correlation among species, the correlation matrix $\mathbf{\Sigma}$ that is given by $\mathbf{\Lambda}t(\mathbf{\Lambda})$. Thanks to the latent factor representation, we will be able not only to infer the residual associations among species, but also to represent species and sites on ordination axes, after filtering from the environment.

Using this dataset, [MAR 20] highlighted how Growing Degree Days, (GDD, the annual sum of average daily degrees above zero), the total potential exoenzymatic activity (total EEA, the sum of all measured exoenzyme activities), soil pH and the ratio between soil carbon and nitrogen (soil C/N) determine the distributions of the44 plant species. We therefore choose to include these four variables as the covariates of

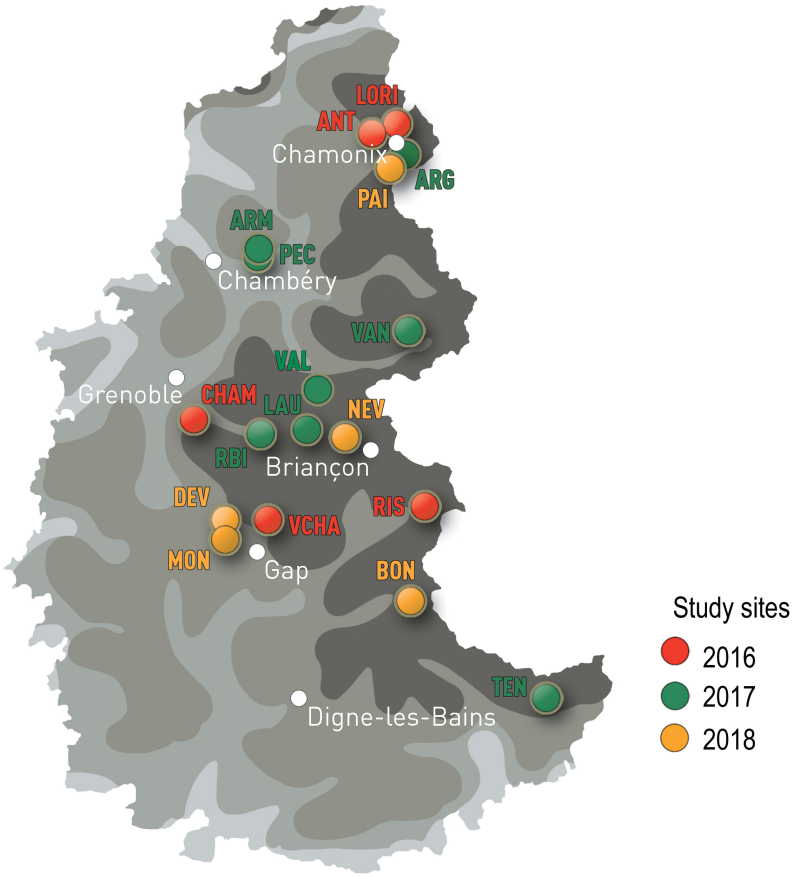

Figure 1.2 – Localisation and names of the 18 gradients of ORCHAMP.

the latent factor model. In line with [MAR 20] we considered the square of the GDD, due to the unimodal response of species to this variable.

1.7.2. *R package used*

To analyse the dataset, we used the R package Hmsc [TIK 19; TIK 20]. This package makes inference on the parameters of the models by sampling from the posterior distribution through Markov chain Monte Carlo (MCMC) sampling. Hmsc implements the latent factors methodology of [BHA 11], where the number of latent factors

is automatically chosen via shrinkage. Although we shall not describe all the features of this package, let us mention the interesting feature that it allows hierarchical modelling of the regression coefficients, and allows both functional traits and phylogeny to be included. This feat enables the user to study the dependence between functional traits and the environment, and to quantify the importance of phylogeny on species distribution. Moreover, it allows an explicit spatial and temporal dependence between sites to be included, improving the performance of the model. Here, however, we do not include any of these features to strictly describe the application of latent factor models. [1]

1.7.3. *Implementation and convergence diagnosis*

We run two MCMC chains of 1500 samples each, with 500 burn-in iterations and no thinning. These models are usually computationally demanding, and the computations for this model notably took around 3 hours. Figure 1.3 shows that all the models clearly converged. The effective sampling size (nESS) of the chains is very high for most parameters, and the potential scale reduction factor (psrf) was always close to one (the description of these measures can be found in [GEL 04]). Thanks to the Bayesian framework, the full posterior distribution of the parameters was available and could then be used to compute a point estimate (posterior mean) and credible intervals (through posterior quantiles) for all parameters.

1.7.4. *Results and discussion*

We evaluated the predictive performance of the model both in in-sample prediction and in cross-validation (due to the high computational costs, we performed a 2-fold cross-validation only). We evaluated the model on these tasks by calculating, for each species, the True Skill Statistic (TSS), which has the advantage to account both for the model sensitivity (i.e. proportion of observed presences predicted as presences) and specificity (i.e. the proportion of observed absences predicted as absences, [ALL 06]). TSS can vary from -1 to 1, where $+1$ indicates perfect fit and values of zero or less indicate a performance no better or worse than random [ALL 06]). Since the TSS requires a threshold to transform species' probability of presence into binary presence-absence data, we selected the threshold that maximises the TSS values. We also evaluated the Root Mean Square Error (RMSE) of each species. In general, the model has good abilities to fit the data (mean in-sample TSS is equal to 0.63, Figure 1.4), but a scarcer ability to generalize on new data (in cross-validation the mean TSS drops by 0.5 and RMSE increases by 0.25). Model performances vary across species, with some species that were poorly modelled (three species had a cross-validation TSS

1. The R code can be found at https://oliviergimenez.github.io/code_livre_variables_cachees/bystrova.html.

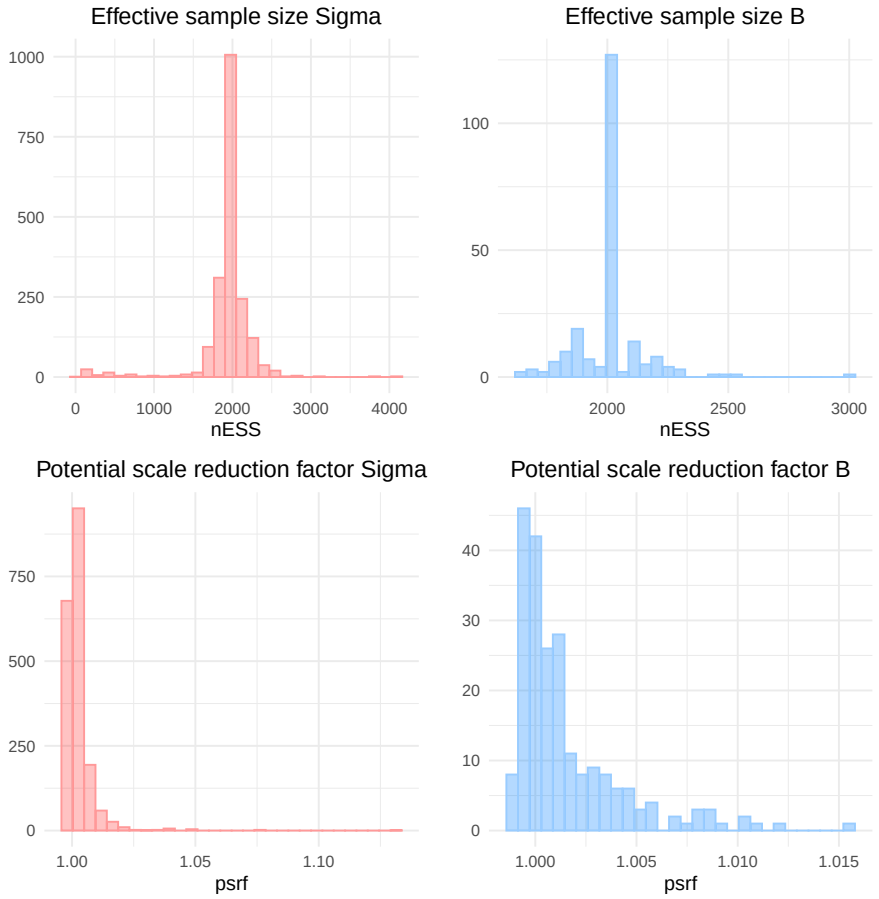

Figure 1.3 – Effective sample size (top panel) and potential scale reduction factor (bottom panel) for the correlation matrix $\boldsymbol{\Sigma}$ (Sigma, left panel) and the regression coefficient $\boldsymbol{\beta}$ (B, right panel).

score equal to 0) and others whose distribution was very explained (cross-validation TSS over 0.3).

The regression coefficients tell us how species respond to the environment, and in this example, their heterogeneity shows how plant species have different responses to the environment (Figure 1.5). In general, climate (represented by GDD) has a significant effect on the distribution of a few number of species only. Instead, soil properties had a higher explanatory power. Many species notably show a trade-off along the gradients of soil characteristics: species that have a positive response to soil

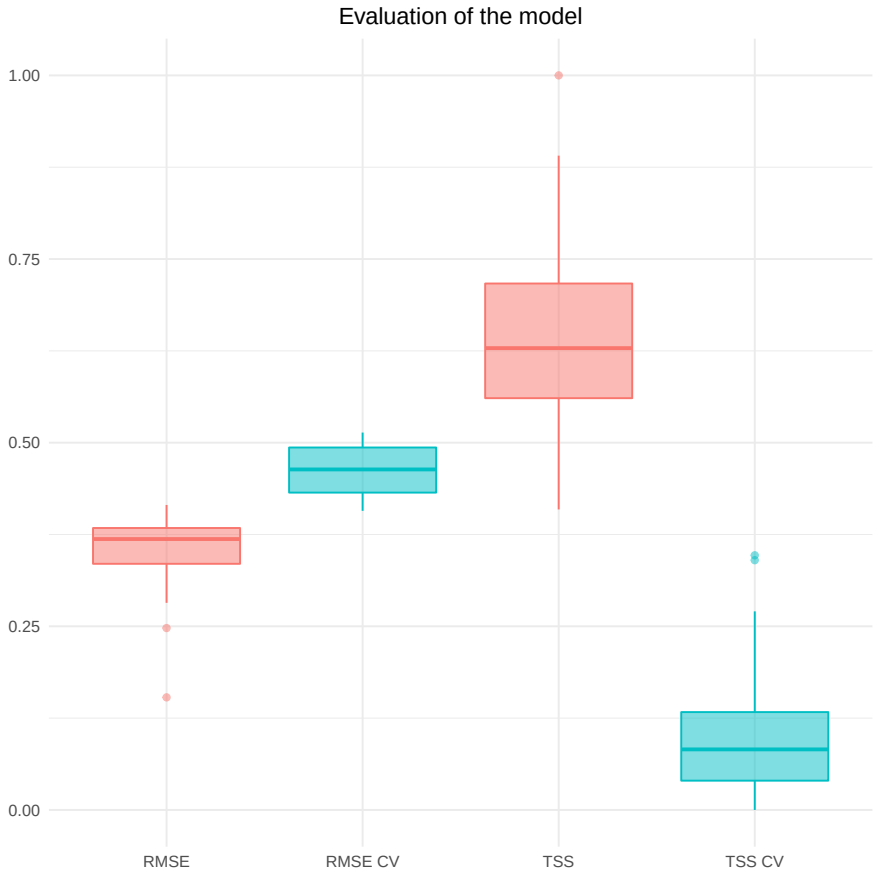

Figure 1.4 – Distribution of TSS and RMSE score across species for in-sample prediction (red) and 2-fold cross-validation (blue).

C/N, often have a negative one to soil pH and/or total EEA and vice-versa (Figure 1.5). These results are consistent with [MAR 20], where the authors, that also considered functional traits (but no residual correlation), showed how such a behaviour reflects the functional trade-offs between conservatives and exploitative species. Exploitative plants are advantaged in nutrient-rich places with mild climate, while conservative species succeed in places were the soil condition are harsh thanks to their adaptation that allow them to survive in stressful situations. As a concluding remark, note that some species do not respond significantly to any of the environmental covariates, and these are the same species for which the TSS and RMSE scores are particularly poor. By analysing the residual correlation matrix, we can understand species co-occurrence

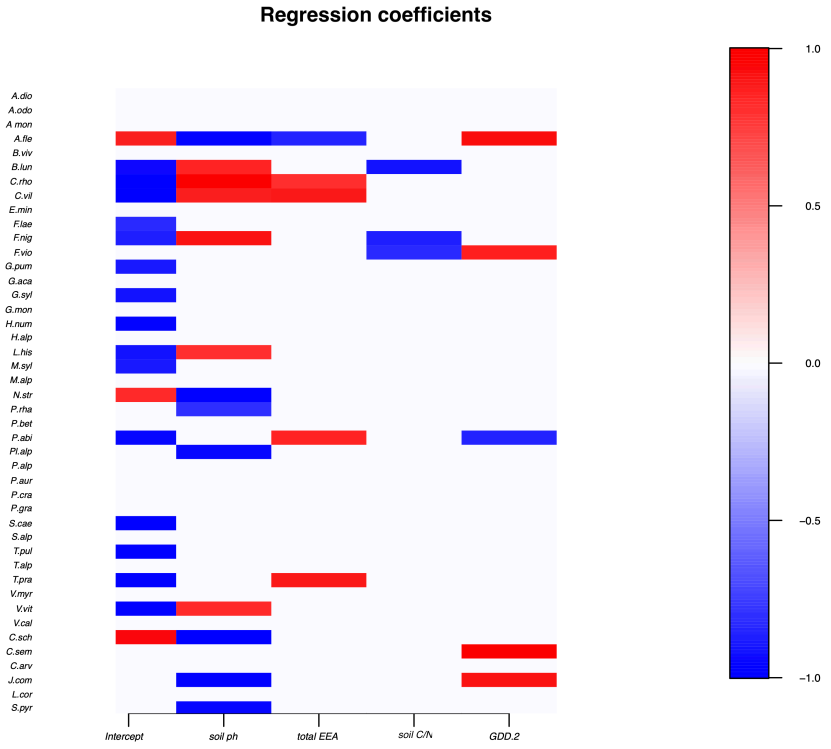

Figure 1.5 – Posterior support values for species regression coefficients. Red if the bounds of the 90% credible interval are both positive, white if the credible interval overlaps 0 and blue if both bounds are negative.

patterns that are not described by the environment, and provide insights about the phenomena that generate them. In the residual correlation matrix of this case study, the species are interestingly divided into two groups. Most plants tend to be positively correlated with species belonging to the same group, but negatively correlated with those of the other group (Figure 1.6). One group (that contains most of the species) is characterised by herbaceous plants that characterise alpine meadows (e.g. *Festuca violacea*, *Sesleria caerula*, *Carex curvula* and *Gentiana acaulis*), while the other one contains trees (e.g. *Picea Abies*), shade-preferring shrubs (e.g. *Vaccinium myrtillus*

and *Vaccinium vitis-ideae*) and herbaceous species that are found in forests and humid habitats (e.g. *Melampyrum sylvaticum* and *Chaerophyllum villarsii*).

This residual correlation matrix highlights ecological phenomena that are well recognised. In fact, along elevational gradients, trees are limited by climatic conditions that prevent their survival above certain altitudes. As a consequence, herbaceous plants that need a high amount of light are excluded from the forests and are only found in open habitats, whereas other herbaceous plants (and shade-friendly shrubs) need the shade provided by the trees, and are therefore found in closed and/or humid habitats. The residual co-occurrence matrix not only endorses the biotic phenomena we described above, but also suggests to include habitat as an additional covariate, that might explain some of this residual co-occurrence patterns and improve model predictions.

Thanks to the latent factor representation, we can try to better understand where these correlations come from. A natural way of doing so is via *ordination*, as explained in Section 1.5. We project the species in the space of the first two latent loadings (the first two columns of $\mathbf{T}$) and the sites in the first two latent factors thanks to a biplot (Figure 1.7). With such representation, we can think of latent factors as missing covariates, and represent sites depending on these missing covariates. Species loadings are therefore the response of species to such missing covariates. If two species are close on the biplot and far from the origin, they respond in the same way to these missing covariates, and are thus more correlated. For example, we see that *Picea abies*, *Melampyrum sylvaticum* and *Vaccinium myrtillus* tend to respond differently from the other species to these missing variables, and in fact, as said above, they are negatively correlated with them.

The type of habitat of the sites is one of the environmental predictors that we haven't included in the study, and that could potentially impact species distribution, and interestingly some of the species highlighted in the previous biplot (Figure 1.7) tend to prefer close rather than open habitats compared to other species. We therefore marked each site as forest (indexed by 1) or grassland (0), re-run the model including this additional covariate and analyse its updated ordination plot (Figure 1.8). The above mentioned species, that tend to behave as outliers when the habitat information was not include, are now closer to the rest of species. The species pool now tend to be more evenly distributed in the ordination space, even if some trends are still remarkable. Notably, we can still see a gradient, with sub-alpine species in the upper-right corner of the biplot and alpine species in the bottom-left one. If this can this be still partially due to some unmeasured environmental variable, this might also be due to the influence of species on each others, with *Picea Abies* that provides the shade for other species.

We finally want to leverage on the information that we assessed in the residual correlation matrix to improve the predictions of the model. We saw how the *Picea abies*

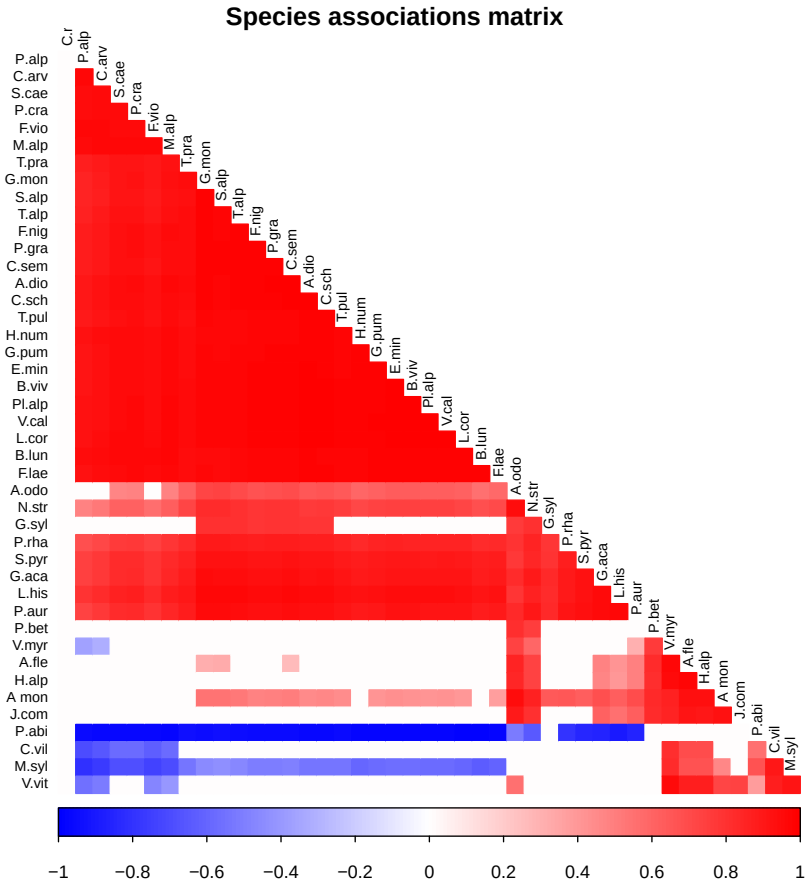

Figure 1.6 – The residual correlation matrix. Only significant values (i.e. 95% credible interval do not overlap zero) are shown.

provides the shade and moisture that allows shrub species like the *Vaccinium*. (*Vaccinium myrtillus* and *Vaccinium vitis-ideae*) and shadow-friendly herbaceous species such as *Melampyrum sylvaticum* and *Chaerophyllum villarsii*) to thrive, while at the same time preventing the survival of herbaceous species that need lots of light, such as *Festuca violacea*. To improve our ability to predict one (or more) of the species described above, we predict the probability of occurrence of species conditionally on the presence (or absence) of *Festuca violacea*, an herbaceous plants that characterises alpine grasslands. This is very similar to include *Festuca violacea* as predictor for the unobserved species. While including the other species as predictors is a doable option for communities with small number of species, it is not straightforward to do it if

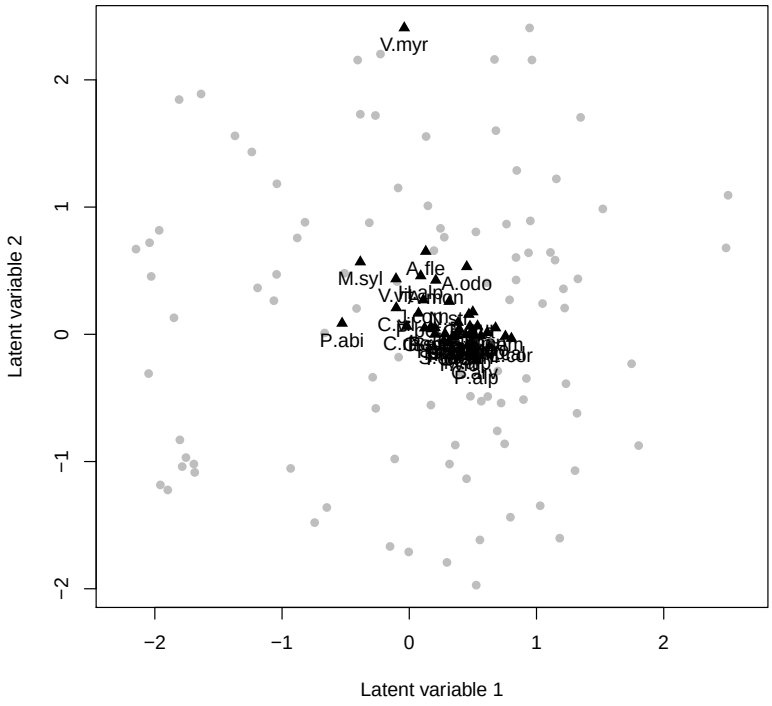

Figure 1.7 – Model based ordination analysis. The two latent variables can be seen as missing covariates, and the position of species (black triangle) on the plot the way species respond to those missing covariates. Species close in the the latent variable species are positively correlated and viceversa.

there are tens or hundreds of species. In contrast, conditional prediction can be made also for a great number of species, without the need to run the model again. When conditioning on *Festuca violacea*, the predictive power of the model improve, in particular concerning cross-validation predictions, where the mean TSS score gains 80% (from 0.1 to 0.19) with respect to the non-conditional predictions. This is particularly true for species that show a particular residual correlation (negative or positive) with *Festuca violacea*. Therefore, we focused on *Poa Alpina*, *Campanula scheuchzeri*, *Soldanella alpina*, *Viola calcarata* and *Euphrasia minima*, which, like *Festuca violacea*, characterize sub-alpine pastures and are often found together, and on the tree *Picea abies*, which as said before, takes the light that would allow the *Festuca violacea* to survive. For example, we consider an alpine meadow in the region of Devoluy (south

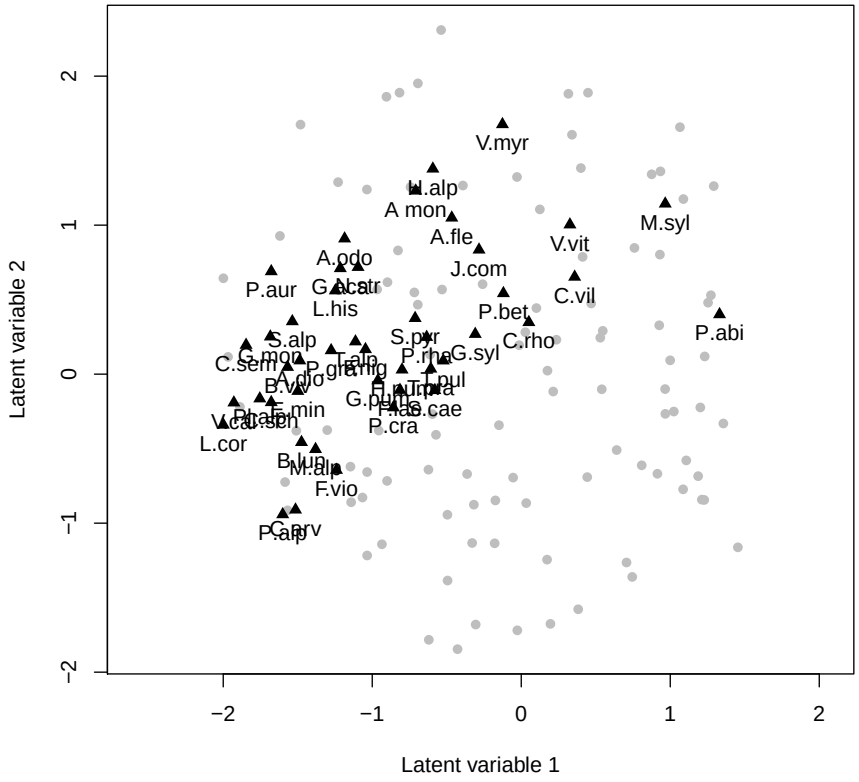

Figure 1.8 – Model based ordination analysis, as above, but when we include the habitat as an additional covariate of the model.

of France), at an altitude of 2100m, where all the above mentioned herbaceous species are present, and *Picea abies* is absent. Figure 1.9a shows how cross-validation predictions conditioned by the presence of *Festuca violacea* increase the probability of the species that are actually present, and decrease the one of the tree, which is absent. This leads more generally to a marked improvement in the cross-validation AUC (Figure 1.9b).

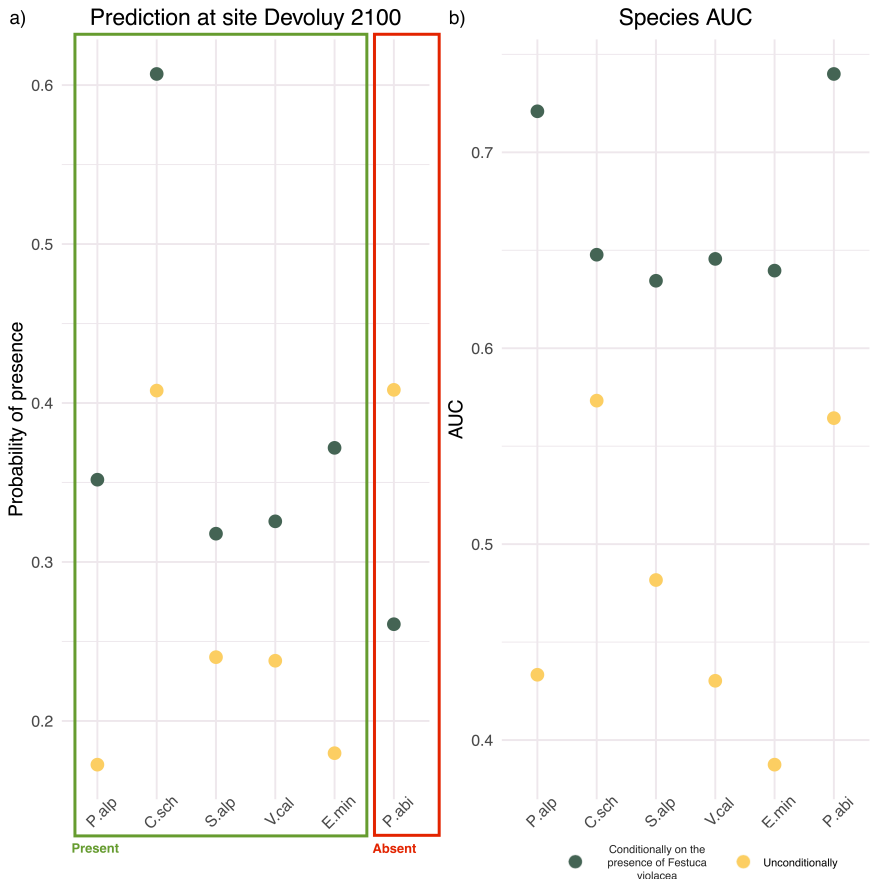

Figure 1.9 – Cross-validation predicted probability of presence (a) and cross-validation AUC (b) of Poa Alpina, Campanula scheuchzeri, Soldanella alpina, Viola calcarata, Euphrasia minima and Picea Abies conditionally on Festuca violacea (green) and unconditionally (yellow). At site Devoluy 2100 all the herbaceous species of above were present (green box) while Picea abies was absent (red box).

1.8. Conclusion

Joint species distribution models (JSDMs) have been recently proposed as an extension of species distribution models (SDMs) that infers residual correlations between species, reflecting co-occurrence patterns not explained by the environmental predictors. These models should be interpreted with care [POG 21], but they still

provide important insights on community assemblage processes. In particular, the application of latent factors to JSDMs can provide further advantages. Indeed, latent factors reduce the dimension of the residual covariance matrix, and the related computational costs that were one of the strongest limitations of early JSDMs. Moreover, by measuring the main axes of residual co-variation between species, they also allow for a residual model based ordination of species and sites. This is particularly interesting when one aims at studying species response to missing environmental variables, that is naturally measured by latent variables. Nevertheless, considering latent factors instead of a full residual covariance matrix can have some drawbacks. First, latent factor models increase their dimension with the number of sites. As a consequence, when dealing with many sites and few species, it is computationally more interesting to model a full residual covariance matrix. Moreover, it is not possible to sparsify the residual covariance matrix induced by latent factor models, a feature that has just been proposed as a solution to improve the interpretability of JSDMs [see PIC 20] in the case of the full residual covariance matrix.

JSDMs have been implemented in many R packages, each with its particular features [see WIL 19, for a review]. In our case study we chose to work with Hmsc because of its broad documentation and the large number of options it includes. Among them, it allows to take into account functional traits and phylogeny, and easily computes conditional predictions. Hmsc is a complete package, easy to start working with, but it is computationally heavy. In order to have faster results, we suggest to work with the package proposed by [PIC 20], whose features remain quite limited for now.

1.9. Bibliography

[ALL 06] ALLOUCHE O., TSOAR A., KADMON R., "Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)", *Journal of Applied Ecology*, vol. 43, num. 6, p. 1223-1232, 2006.

[BAR 11] BARVE N., BARVE V., JIMÉNEZ-VALVERDE A., LIRA-NORIEGA A., MAHER S. P., PETERSON A. T., SOBERON J., VILLALOBOS F., "The crucial role of the accessible area in ecological niche modeling and species distribution modeling", *Ecological Modelling*, vol. 222, num. 11, p. 1810 - 1819, 2011.

[BHA 11] BHATTACHARYA A., DUNSON D. B., "Sparse Bayesian infinite factor models", *Biometrika*, p. 291–306, 2011.

[BYS 20] BYSTROVA D., POGGIATO G., BEKTAS B., ARBEL J., CLARK J. S., GUGLIELMI A., THUILLER W., "Clustering species with residual covariance matrix in joint species distribution models", *Preprint*, 2020.

[CAL 14] CALABRESE J. M., CERTAIN G., KRAAN C., DORMANN C. F., "Stacking species distribution models and adjusting bias by linking them to macroecological models", *Global Ecology and Biogeography*, vol. 23, num. 1, p. 99-112, 2014.

[CHI 98] CHIB S., GREENBERG E., "Analysis of Multivariate Probit Models", *Biometrika*, vol. 85, p. 347-361, 06 1998.

[CLA 06] CLARK J. S., GELFAND A. E., *Hierarchical modelling for the environmental sciences: statistical methods and applications*, Oxford University Press on Demand, 2006.

[CLA 14] CLARK J. S., GELFAND A. E., WOODALL C. W., ZHU K., "More than the sum of the parts: forest climate response from joint species distribution models", *Ecological Applications*, vol. 24, num. 5, p. 990–999, Wiley Online Library, 2014.

[CLA 16] CLARK J. S., NEMERGUT D., SEYEDNASROLLAH B., TURNER P., ZHANG S., "Generalized joint attribute modeling for biodiversity analysis: Median-zero, multivariate, multifarious data", *Ecological Monographs*, 11 2016.

[CLA 17] CLARK J. S., NEMERGUT D., SEYEDNASROLLAH B., TURNER P. J., ZHANG S., "Generalized joint attribute modeling for biodiversity analysis: median-zero, multivariate, multifarious data", *Ecological Monographs*, vol. 87, num. 1, p. 34-56, 2017.

[ELL 04] ELLISON A. M., "Bayesian inference in ecology", *Ecology letters*, vol. 7, num. 6, p. 509–520, Wiley Online Library, 2004.

[GEL 04] GELMAN A., CARLIN J. B., STERN H. S., RUBIN D. B., *Bayesian Data Analysis*, Chapman and Hall/CRC, 2nd ed. edition, 2004.

[GUI 00] GUISAN A., ZIMMERMANN N. E., "Predictive habitat distribution models in ecology", *Ecological Modelling*, vol. 135, num. 2, p. 147 - 186, 2000.

[GUI 05] GUISAN A., THUILLER W., "Predicting species distribution: offering more than simple habitat models", *Ecology Letters*, vol. 8, num. 9, p. 993-1009, 2005.

[GUI 11] GUISAN A., RAHBEK C., "SESAM – a new framework integrating macroecological and species distribution models for predicting spatio-temporal patterns of species assemblages", *Journal of Biogeography*, vol. 38, num. 8, p. 1433-1444, 2011.

[GUI 17] GUISAN A., THUILLER W., ZIMMERMANN N. E., *Habitat Suitability and Distribution Models: With Applications in R*, Ecology, Biodiversity and Conservation, Cambridge University Press, 2017.

[HUT 57] HUTCHINSON, "Population studies: Animal ecology and demography", *Bulletin of Mathematical Biology*, vol. 53, num. 1, p. 193 - 213, 1957.

[LOR 04] LORTIE C. J., BROOKER R. W., CHOLER P., KIKVIDZE Z., MICHALET R., PUGNAIRE F. I., CALLAWAY R. M., "Rethinking plant community theory", *Oikos*, vol. 107, num. 2, p. 433-438, 2004.

[MAR 20] MARTINEZ-ALMOYNA C., PITON G., ABDULHAK S., BOULANGEAT L., CHOLER P., DELAHAYE T., DENTANT C., FOULQUIER A., POULENARD J., NOBLE V., RENAUD J., ROME M., SAILLARD A., CONSORTIUM T. O., THUILLER W., MÜNKEMÜLLER T., "Climate, soil resources and microbial activity shape the distributions of mountain plants based on their functional traits", *Ecography*, vol. 43, num. 10, p. 1550-1559, 2020.

[MCC 89] MCCULLAGH P., NELDER J., *Generalized Linear Models, Second Edition*, Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series, Chapman & Hall, 1989.

[MER 14] MEROW C., SMITH M. J., EDWARDS JR T. C., GUISAN A., MCMAHON S. M., NORMAND S., THUILLER W., WÜEST R. O., ZIMMERMANN N. E., ELITH J., "What do we gain from simplicity versus complexity in species distribution models?", *Ecography*, vol. 37, num. 12, p. 1267-1281, 2014.

[OVA 11] OVASKAINEN O., SOININEN J., "Making more out of sparse data: hierarchical modeling of species communities", *Ecology*, vol. 92, num. 2, p. 289-295, 2011.

[OVA 17] OVASKAINEN O., TIKHONOV G., NORBERG A., GUILLAUME BLANCHET F., DUAN L., DUNSON D., ROSLIN T., ABREGO N., "How to make more out of community data? A conceptual framework and its implementation as models and software", *Ecology Letters*, vol. 20, num. 5, p. 561-576, 2017.

[OVA 20] OVASKAINEN O., ABREGO N., *Joint Species Distribution Modelling: With Applications in R*, Ecology, Biodiversity and Conservation, Cambridge University Press, 2020.

[PIC 20] PICHLER M., HARTIG F., "A new method for faster and more accurate inference of species associations from novel community data", 2020.

[POG 21] POGGIATO G., MÜNKEMÜLLER T., BYSTROVA D., ARBEL J., CLARK J., THUILLER W., "On the interpretations of joint modelling in community ecology", *Trends in Ecology and Evolution, in press*, 2021.

[POL 14] POLLOCK L. J., TINGLEY R., MORRIS W. K., GOLDING N., O'HARA R. B., PARRIS K. M., VESK P. A., MCCARTHY M. A., "Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model", *Methods in Ecology and Evolution*, 2014.

[PUL 00] PULLIAM H., "On the relationship between niche and distribution", *Ecology Letters*, vol. 3, num. 4, p. 349-361, 2000.

[SOB 05] SOBERON J., PETERSON A., "Interpretation of Models of Fundamental Ecological Niches and Species Distributional Areas", *Biodiversity informatics*, 2005.

[SOB 07] SOBERON J., "Grinnellian and Eltonian niches and geographic distributions of species", *Ecology Letters*, vol. 10, num. 12, p. 1115-1123, 2007.

[SPI 02] SPIEGELHALTER D. J., BEST N. G., CARLIN B. P., VAN DER LINDE A., "Bayesian measures of model complexity and fit", *Journal of the royal statistical society: Series b (statistical methodology)*, vol. 64, num. 4, p. 583–639, Wiley Online Library, 2002.

[TAB 12] TABERLET P., COISSAC E., HAJIBABAEI M., RIESEBERG L. H., "Environmental DNA", *Molecular ecology*, vol. 21, num. 8, p. 1789–1793, Wiley Online Library, 2012.

[THU 13] THUILLER W., MÜNKEMÜLLER T., LAVERGNE S., MOUILLOT D., MOUQUET N., SCHIFFERS K., GRAVEL D., "A road map for integrating eco-evolutionary processes into biodiversity models", *Ecology Letters*, vol. 16, num. s1, p. 94-105, 2013.

[TIK 19] TIKHONOV G., OVASKAINEN O., OKSANEN J., DE JONGE M., OPEDAL O., DALLAS T., Hmsc: Hierarchical Model of Species Communities, 2019, R package version 3.0-4.

[TIK 20] TIKHONOV G., OPEDAL Ã. H., ABREGO N., LEHIKOINEN A., DE JONGE M. M. J., OKSANEN J., OVASKAINEN O., "Joint species distribution modelling with the r-package Hmsc", *Methods in Ecology and Evolution*, vol. 11, num. 3, p. 442-447, 2020.

[WAR 15] WARTON D., D. FOSTER S., DE'ATH G., STOKLOSA J., K. DUNSTAN P., "Model-based thinking for community ecology", *Plant Ecology*, vol. 216, p. 669-682, 05 2015.

[WAT 10] WATANABE S., OPPER M., "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory", *Journal of Machine Learning Research*, vol. 11, num. 12, 2010.

[WIL 19] WILKINSON D. P., GOLDING N., GUILLERA-ARROITA G., TINGLEY R., MCCARTHY M. A., "A comparison of joint species distribution models for presence–absence data", *Methods in Ecology and Evolution*, vol. 10, num. 2, p. 198–211, Wiley Online Library, 2019.

[YAT 18] YATES K., BOUCHET P., CALEY M., K E. A. M., "Outstanding challenges in the transferability of ecological models", *Trends in Ecology & Evolution*, vol. 33, num. 10, p. 790–802, Cell Press (Elsevier), August 2018.