Clustering species with residual covariance matrix in Joint Species Distribution models

Daria Bystrova, Giovanni Poggiato, Billur Bektaş, Julyan Arbel, James Clark, Alessandra Guglielmi, Wilfried Thuiller

frontiers

Clustering species with residual covariance matrix in Joint Species Distribution models

Daria Bystrova [1,2,*], Giovanni Poggiato [1], Billur Bektaş [1], Julyan Arbel [2], James S. Clark [3,4,5], Alessandra Guglielmi [6] and Wilfried Thuiller [1]

[1] *Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LECA, F-38000 Grenoble, France*
[2] *Univ. of Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, F-38000 Grenoble, France*
[3] *Univ. of Grenoble Alpes, INRAE, LESSEM, F-38000 Grenoble, France*
[4] *Nicholas School of the Environment, Duke University, Durham, North Carolina, 27708 USA*
[5] *Department of Statistical Science, Duke University, Durham, North Carolina, 27708 USA*
[6] *Dipartimento di Matematica, Politecnico di Milano, Milan, Italy*

Correspondence*:
Corresponding Author
daria.bystrova@inria.fr

ABSTRACT

Modelling species distributions over space and time is one of the major research topics in both ecology and conservation biology. Joint Species Distribution models (JSDMs) have recently been introduced as a tool to better model community data, by inferring a residual covariance matrix between species, after accounting for species' response to the environment. However, these models are computationally demanding, even when latent factors, a common tool for dimension reduction, are used. To address this issue, Taylor-Rodriguez et al. (2017) proposed to use a Dirichlet process, a Bayesian nonparametric prior, to further reduce model dimension by clustering species in the residual covariance matrix. Here, we built on this approach to include a prior knowledge on the potential number of clusters, and instead used a Pitman–Yor process to address some critical limitations of the Dirichlet process. We therefore propose a framework that includes prior knowledge in the residual covariance matrix, providing a tool to analyze clusters of species that share the same residual associations with respect to other species. We applied our methodology to a case study of plant communities in a protected area of the French Alps (the Bauges Regional Park), and demonstrated that our extensions improve dimension reduction and reveal additional information from the residual covariance matrix, notably showing how the estimated clusters are compatible with plant traits, endorsing their importance in shaping communities.

Keywords: biodiversity modelling, dimension reduction, joint species distribution model, latent factors, Bayesian nonparametrics, plant communities

1 INTRODUCTION

Understanding and predicting the distribution of species across space and time is one of the central questions in ecology [50]. As such, species distribution models (SDMs) are essential tools to investigate how species respond to environment [19, 13, 20]. The main principle is to relate *individual* species observations to a set of environmental predictors. The estimated relationship between species and the environment allows to infer the environmental niche of the species and then to predict its distribution for new environmental conditions, either in space or time, or in both [19, 29, 20]. While SDMs could be used to study species assemblages [a technique commonly called stacked SDM (sSDM), see 15, 5], they were meant to model and predict the distribution of individual species. To model species assemblages, recent statistical advances yield to Joint Species Distribution Models (JSDMs) [43, 9, 37, 52], which are multivariate extensions of generalized linear regression models (GLM) [other approaches can be found in 21, 51]. In JSDMs, the regression coefficients are related to the response of species to the environment, as in SDMs, while the correlation among the residuals describe the pairwise-species dependencies not explained by the environment.

Since JSDMs were created to deal with community data, they are gaining popularity with the ever-increasing developments of novel methods for community assessment, such as environmental DNA (eDNA) metabarcoding [47]. However, their application to large datasets still faces strong limitations such as computational costs and the interpretation of the residual covariance matrix. Indeed, JSDMs are computationally expensive because the number of estimated parameters in the residual correlation matrix grows quadratically with the number of species. There are several approaches to address this problem. For JSDMs that are based on the multivariate probit model, computational reduction can be achieved by efficient parallel sampling [6, 39] to fit a full covariance matrix in a frequentist framework. Another common solution relies on dimension reduction through latent variable models (LVM) [52], where the effective dimension of the model is reduced by a low-rank approximation of the residual covariance matrix [52, 35, 48, 24]. While the approximation with low rank values could capture the residual associations in the covariance matrix for a large number of species and significantly improve convergence and computational time [52], their wide applications to large dataset is still prohibited [39].

A growing number of species is not only a problem from a computational viewpoint but it also makes the interpretation of the residual correlations challenging. For example, for 100 species, 4 950 pairwise residual correlations are estimated, which represent species associations patterns that are not explained by the environmental covariates and can depend on many factors: model misspecification, missing covariates, and less likely, biotic interactions [42]. Moreover, the symmetric constraint of any covariance matrix impedes to detect any asymmetric dependence between species (e.g. hierarchical competition, predator-prey, [12, 42]). Therefore, the complexity of the pattern increases with the dimension of the problem and blurs the interpretation of the residual covariance matrix inferred by JSDMs.

To improve such an interpretability of the residual covariance matrix recent works proposed to reduce the number of non-zero residual correlations between species. This is usually done by applying sparsity inducing regularization (e.g. L_1, elastic net) to the correlation matrix [e.g elastic-net 39] or its inverse, the precision matrix [8]. However, latent factor JSDMs usually fail to produce sparse matrices [39]. We believe that providing additional assumption on the structure of the residual covariance matrix could be a promising avenue. For instance, we might consider block-wise structure of the covariance matrix, such that residual associations would vary between the blocks, instead of individual observations [30]. In the JSDM case, it means that we can consider groups of species that share the same association patterns with respect to the other ones. In this case, the model would capture the main associations between (and within) groups of species instead of the species level ones.

Incorporating expert-based knowledge about this block structure of the covariance matrix would further improve this model. Interestingly, most JSDMs are implemented within a Bayesian framework, that naturally allows the incorporation of a prior knowledge, but few ecological studies have actually exploited this feature [1]. Choosing the prior knowledge that we want to give to the residual covariance matrix is tricky, but feasible. For instance, in a species-rich foodweb, there are a fair amount of species that share the same preys and predators with others, forming what is usually called trophic groups [38]. If they are known, or inferred with a specific approach like a stochastic block model [25], the number of trophic groups can be used as a prior to reduce the residual correlation matrix. In a similar way, plant functional groups have been designed to group species that share the same traits, respond the same way to environment, and interact the same way with species from other groups [2]. We believe that the prior knowledge on the number of groups of interest (e.g. trophic or functional) can be used as a prior for the block structure of the residual covariance matrix, which could help to reduce the dimension, and the same time, might help the interpretability of the residual covariance matrix.

Recently, Taylor-Rodriguez et al. [48] proposed a dimension reduction method that combines a latent variable approach with an additional clustering of the variance-covariance matrix using a Dirichlet process prior. That allows to further reduce the effective dimension and improve the computational efficiency of the model, but in the proposed model, clustering was mainly a tool for dimension reduction, without focusing on further cluster interpretation. That paper also did not address prior information that could be used with the Dirichlet process to inform the number of desired species groups.

Here, we build on this recent work to propose a novel framework that allows for a clustering of residual associations that makes use of prior information. In doing so, we addressed the following questions: (1) Can prior knowledge, combined with dimension reduction on the structure of the residual covariance matrix, improve model inference in JSDM? (2) Can estimated clusters be interpreted in ways that help us understand species communities?

In the following, we first describe the model and our extension that improves clustering properties by incorporating prior knowledge on the number of species that share residual associations. We then introduce Pitman–Yor process, a more flexible Bayesian nonparametric prior, which is less sensitive to miscalibration than the Dirichlet process. We hypothesized that species within the same cluster have similar functional strategies. As a show case, we investigated this hypothesis within the scope of the case study of Bauges National Park.

2 THE FRAMEWORK

2.1 Statistical model

We provide a formal description of our model, which is an extension of the model in [48] developed to reduce the dimensionality of the inference in JSDMs, in the particular framework of Generalized Joint Attribute Modelling (GJAM) [9]. GJAM allows to model many types of species observations (presence-absence, counts, biomass and others) altogther. We present the model and its application for presence-absence data, but since our approach is an extension of GJAM, it is valid for most responses.

To study species distributions, we model a response variable $\boldsymbol{y}_i$ with respect to a set of p environmental covariates $\boldsymbol{x}_i = \{x_{i\ell}\}_{\ell=1}^{p}$, at each site $i = 1, \dots, n$, where $\ell = 1, \dots, p$ represents the ℓ-th environmental covariate. The response variable $\boldsymbol{Y}_i \in \{0, 1\}$ is a vector where each element y_{ij} contains the observation for species $j = 1, \dots, S$ at site i. JSDMs model the response variable using what is commonly called the multivariate probit model [7], where for species j at site i the probability of presence is modelled through a latent normal variable z_{ij} as $\Pr(y_{ij} = 1) = \Pr(z_{ij} > 0)$. In dimension reduction approach suggested by

[48] $\boldsymbol{z}_i$ is modelled as:

$$\boldsymbol{z}_i = \boldsymbol{B}\boldsymbol{x}_i + \boldsymbol{\Lambda}\boldsymbol{w}_i + \boldsymbol{\epsilon}_i, \qquad \boldsymbol{\epsilon}_i \overset{\text{iid}}{\sim} N_S(0_S, \sigma_\epsilon^2 I_S), \tag{1}$$

where $\boldsymbol{B}$ is the $S \times p$ matrix of regression coefficients and $\boldsymbol{x}$ is the $p \times n$ matrix of measured covariates and $\boldsymbol{w}_i \overset{\text{iid}}{\sim} N(0, \boldsymbol{I}_r)$ are the latent factors, $\boldsymbol{\Lambda}$ is the $S \times r$ matrix of factor loadings. The number of factors r is supposed to be comparably smaller than S ($r \ll S$). Here, latent normal variable $\boldsymbol{z}_i$ has residual covariance matrix $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \boldsymbol{I}_S$, which have dimension $S \times r + 1$ and is less than $S(S+1)/2$ in general case [see details 48]. In this model, the number of latent factors r is fixed, and can be chosen to maximize the goodness-of-fit or some informative criteria like DIC, BIC [16] or using cross-validation. While choosing the number of latent factors r, it is important to verify that the matrix $\boldsymbol{\Lambda}$ has a full column rank and the model is well-identified [48, 18].

If the residual covariance matrix $\boldsymbol{\Sigma}^*$ represents the co-occurrence pattern not explained by the environment, latent factor models can provide further insights in this residual correlation. Indeed, the low-rank matrix $\boldsymbol{\Lambda}$ would represent the main axes of variation of the residual co-occurrence pattern. Moreover, latent factors $\boldsymbol{w}_i$ could represent missing environmental predictors at site i, and rows of matrix $\boldsymbol{\Lambda}$ ($\boldsymbol{\lambda}_j$) encode the response of species j to these missing predictors. Therefore, latent factors can highlight both the main axis of co-variation and a common (or opposite) response to unmeasured covariates [see Chapter 7 of 34].

A further dimension reduction proposed by [48], that finds common rows in $\boldsymbol{\Lambda}$, is described in the next section.

2.2 Clustering in the residual covariance matrix

Latent factors allow to model a 'tall and skinny' $S \times r$ matrix $\boldsymbol{\Lambda}$ instead of a 'tall and wide' $S \times S$ matrix $\boldsymbol{\Sigma}$. Further dimension reduction proposed in [48] is based on the reduction of this 'tall and skinny' $\boldsymbol{\Lambda}$ matrix to a 'short and skinny' one. To do so, the authors find common rows in $\boldsymbol{\Lambda}$, exploiting the clustering properties of the *Dirichlet process* (DP), a prior distribution used routinely in Bayesian nonparametric statistics. By finding common rows in $\boldsymbol{\Lambda}$, the DP creates clusters (i.e. groups) of species that share the same values of $\boldsymbol{\Lambda}_j$. Therefore, only the $N < S$ unique values of the rows of $\boldsymbol{\Lambda}^*$ are estimated, that will then be repeated for species in the same clusters to obtain $\boldsymbol{\Lambda}$ and then $\boldsymbol{\Sigma}^*$. As a consequence, the model no more estimates the 'tall and skinny' $\boldsymbol{\Lambda}$, but only the 'short and skinny' $\boldsymbol{\Lambda}^*$ matrix. Species in the same cluster will have the same value of the corresponding rows of $\boldsymbol{\Lambda}$, and consequently these species will also have the same rows and columns in $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \boldsymbol{I}_S$. In other words, species in the same cluster are similar in their residual covariance matrix. Therefore, we will say hereafter that we *cluster species depending on their associations with respect to other species*. This approach allows to reduce dimension of the model and infer groups of species with the same residual correlation structure

In the model proposed by [48], clustering was only a tool for dimension reduction, and the paper did not focus on the further interpretation of the clusters that we just discussed. However, the clustering resulting from the Dirichlet process prior depends on the prior specification [11] for which we offer two extensions. In the first extension, we provide a flexible method to incorporate prior information on the number of clusters that allows clusters to better represent the underlying data. For the second extension, we introduce another Bayesian nonparametric prior called the Pitman–Yor process, which overcomes some limitations of the Dirichlet process and is more suitable for ecological applications.

2.2.1 Dirichlet process formulations: **DP** and **DP**$_c$

We describe here the original Dirichlet process formulation proposed by [48], as well as an extension of it which allows the introduction of a prior knowledge in a flexible way, respectively denoted as **DP** and **DP**$_c$ (calibrated Dirichlet process model).

The Dirichlet process is used in Bayesian statistics as a prior distribution over distributions. In other words, sampling from the Dirichlet process provides a distribution that has the important feature of being discrete, thus clustering samples naturally [14]. This process is parameterized by the base distribution H, which is the mean of the Dirichlet process and the concentration parameter α that regulates how the distribution drawn from the Dirichlet process is concentrated around its mean (base distribution) (see the formal definition in Appendix section S.4).

We denote by $\boldsymbol{\lambda}_j,\ j = 1, \dots, S$ the rows of the matrix $\boldsymbol{\Lambda}$ in (1). The original **DP** model uses the Dirichlet process as follows:

$$\boldsymbol{\lambda}_j \mid G \overset{\text{iid}}{\sim} G, \quad j = 1, \dots, S, \tag{2}$$

$$G \sim \text{DP}(\alpha H), \tag{3}$$

where G is the probability distribution drawn from the Dirichlet process prior. [48] chose the r-dimensional normal distribution as the base distribution H, and used a fixed concentration parameter α. By the properties of the Dirichlet process, the distribution G is almost surely discrete, so that there will be repeated values in the sampled rows $\boldsymbol{\lambda}_j$ (i.e., there is non-zero probability that two rows collide). The unique values of $\boldsymbol{\lambda}_j$ form the $N \leq S$ matrix $\boldsymbol{\Lambda}^*$. We hereafter call a 'cluster' (group) the subset of species whose $\boldsymbol{\lambda}_j$ coincide.

The main advantage of the Dirichlet process (and other more flexible Bayesian nonparametric priors) is that it does not pre-specify the exact number of clusters. Dirichlet process prior induce prior distribution on the number of clusters. We may fix features of the induced prior number of clusters through the concentration parameter α, that regulates the clustering properties of the Dirichlet process (see details in Appendix Section S.4).

The Dirichlet process is the most widespread nonparametric prior due to its computational ease, but it has several limitations. The main one is precisely that clustering properties are regulated by only one parameter, α. As pointed out in [11], this concentration parameter has a strong effect on the posterior distribution of the number of clusters. Indeed, the prior distribution on the number of clusters of a Dirichlet process prior is highly peaked. As a consequence one would require a high sample size to counterbalance such a strong prior weight, resulting in a low posterior probability to have a number of clusters far from the prior mean.

To overcome this limitation, we added a hierarchical layer for the α parameter to let the model choose values for α, thus providing flexibility to the posterior number of clusters. We chose a Gamma distribution as a hyperprior for α, so that $\alpha \sim \text{Ga}(\nu_1, \nu_2)$, where ν_1, ν_2 are hyperparameters. We implemented a within-Gibbs Metropolis–Hasting step (see for details Appendix Section S.6), to sample from the posterior distribution of this parameter [45]. As in [48] we use the Dirichlet multinomial process for approximating the Dirichlet process [31]. We hereafter refer to this model as **DP**$_c$ (calibrated Dirichlet Process). By conveniently choosing the hyperparameters of the Gamma distribution, we can calibrate the expected value of the prior distribution on the number of clusters induced by the DP. Indeed, the expected number of clusters for the Dirichlet multinomial process is [26, Example 2]:

$$\mathbb{E}[K_{n,\alpha,N}] = N - \frac{(N-1)(\alpha + 1 - \frac{\alpha}{N})_{n-1}}{(\alpha+1)_{n-1}}, \tag{4}$$

where $(x)_n = x(x+1)\dots(x+n-1)$ denotes the increasing factorial coefficient, for any real number x and integer n. By further sampling from parameter α and from $\text{Ga}(\nu_1, \nu_2)$ and using (4) we can ultimately determine the values of the hyperparameters (ν_1, ν_2) that guarantee that the prior expected number of clusters K_S matches our prior knowledge on the number of clusters K^*, i.e $\mathbb{E}[K_S] = K^*$ (see details in Appendix Section S.5.1). In our case, an ecologically-driven prior knowledge is used to specify the prior belief on the number of clusters K^*. The ground truth on the value of K^* is hard to be known, but due to the larger prior variance provided by the hierarchical modelling of α the prior is not fully informative, and allows the inclusion of an eventually uncertain prior knowledge too. A sensitivity analysis has to be carried to confirm such a choice.

As a side note, notice that in the model proposed by [48] one may suitably select the parameter N in order to fix the prior mean on the number of groups, but this could lead to an extremely peaky prior distribution (Figure S4), which may not result in a flexible model. While providing more flexibility to the clustering properties of the model, the Dirichlet Process is still limited by its dependence on a single parameter α. We therefore propose another extension of the model, by introducing the Pitman–Yor process.

2.2.2 Pitman–Yor process formulation: $\mathbf{PY_c}$

The Pitman–Yor (PY) process is a flexible generalization of the Dirichlet process (see the full description in Appendix Section S.4). Indeed the Pitman–Yor process is characterized by the base measure H, the concentration α and, importantly, by a discount parameter $\sigma \in [0, 1)$. When $\sigma = 0$, the Pitman–Yor process is anything but the Dirichlet Process. The parameter σ influences the variance of the prior number of clusters, and a high value of σ leads to high variance for the distribution of the prior number of groups. As a consequence, the posterior distribution is less constrained by the prior, and the resulting clustering is more flexible. Denote by K_S the prior number of clusters for S samples. Another property of Pitman–Yor process is that the number K_S grows more rapidly with the number of species S than for the Dirichlet process [40]. For the Pitman–Yor process the number K_S follows a power-law, i.e $\mathbb{E}[K_S]$ grows as S^σ when $S \to \infty$, while for the Dirichlet process it grows logarithmically as $log(S)$ when $S \to \infty$ [40]. Moreover, the cluster size distribution also shows power-law under Pitman–Yor process [41]. For many real applications, this power-law property is a more natural assumption than in the Dirichlet process, where we generally have a small number of clusters with a high number of observations, and a large number of clusters with only a few observations.

We therefore considered a Pitman–Yor process as a prior for the rows of $\mathbf{\Lambda}$, similarly to the **DP** and $\mathbf{DP_c}$ models.

$$\boldsymbol{\lambda}_j \mid G \overset{\text{iid}}{\sim} G, \quad j = 1, \dots, S, \tag{5}$$

$$G \sim \text{PY}(\alpha, \sigma, H), \tag{6}$$

where H is the base measure as in (3), α is the concentration parameter, and σ is the discount parameter. In our model we used the finite-dimensional Pitman–Yor multinomial process proposed by [27], which approximates the Pitman–Yor process and allows tractable computation (more details in Appendix Section S.4).

We assumed parameters α and σ as fixed following [27], and that the Pitman–Yor multinomial process is flexible enough and does not require a prior distribution on hyperparameters. We use the prior distribution on the number of clusters for the Pitman–Yor multinomial process to compute the prior expected number of clusters $\mathbb{E}[K_S]$ and variance $\mathbb{V}[K_S]$ of this prior distribution. We can set $\mathbb{E}[K_S] = K^*$ and specify the variance $\mathbb{V}[K_S]$ to reflect the desired level of uncertainty and then solve the system of equations numerically.

However, the solution could be computationally challenging for some values of parameters. In addition, certain pairs of expectation and standard deviation are not easily attainable (see Figure 6 in [4]). Here, we firstly fix parameter σ, choosing the distribution variability, and then find the parameter α, such that $\mathbb{E}[K_S] = K^*$ (see details in Appendix Section S.5.2). We refer to this model as $\mathbf{PY}_\mathrm{c}$ (calibrated Pitman–Yor process model).

2.3 Clustering analysis

We summarize the posterior distribution of the clusters to obtain a clustering (i.e. partition of species) for each model **DP**, $\mathbf{DP}_\mathrm{c}$ and $\mathbf{PY}_\mathrm{c}$ (the procedure is described in Appendix Section S.7). Notice that there is a difference between the posterior expected number of clusters and the number of clusters of the estimated clustering, obtained by the algorithm described in Appendix Section S.7 for posterior inference on partition space. The former describes the distribution of the number of clusters in Markov chain Monte Carlo (MCMC) samples [45], while the latter represents the number of clusters in the single partition that best represents the posterior distribution of the clusters in the MCMC samples. Generally, even if one has certain prior knowledge on the number of groups, it is possible that there is no information on the cluster composition. In our case study, we have a prior expected number of clusters and we also have a cluster composition. For this reason, we could also assess the composition of obtained cluster in-depth. To do so, we measured the distance between the clusters obtained by the models and the clusters we used as a prior belief. We used the adjusted Rand index (ARI) [23], which is the corrected for chance proportion of the number of agreements (species clustered similarly) in all possible pairs of species divided by the total number of all possible pairs. This value is between 0 and 1, where 1 corresponds to exactly the same cluster composition. Additionally, we checked how the choice of the prior number of groups affects the posterior distribution (sensitivity to the prior, Appendix Section S.8.3).

3 CASE STUDY

3.1 Study site and species information

We illustrated our methods with data on plant species in Bauges Natural Regional Park (France) available from the Alpine Botanical Conservatory (CBNA) and previously analysed by [49]. We included as covariates the first two principal components of the environmental variables presented by [49], including a quadratic term (using orthogonal polynomials to reduce correlation among the covariates). We considered presence-absence for the 111 most abundant species (present at least in 2% of sites) across the 1 139 plots. For details on the data processing steps, see Appendix Section S.1. We considered the 16 plant functional groups (PFGs) that were built in [49]. These PFGs have been obtained through hierarchical clustering, in order to build groups of species that have a similar functional role: they have a similar tolerance of abiotic conditions, dispersal abilities, resistance to disturbance (grazing and mowing), response to competition for light (whether they germinate and grow under specific light conditions), competitive effects (estimated by the height of the species) and demographic characteristics (life-form, longevity, age of maturity). We refer to [49] for a complete description of PFGs and how they were classified. The number of these groups were used to specify the $\mathbf{DP}_\mathrm{c}$ and $\mathbf{PY}_\mathrm{c}$ priors, by fixing $\mathbb{E}[K_S] = 16$. Note that the number of groups, but not their composition, was used for prior specification.

3.2 Implementation and specification of the models

We applied our $\mathbf{DP}_\mathrm{c}$ and $\mathbf{PY}_\mathrm{c}$ models together with the original **DP** model in dimension reduction with the default settings on the Bauges plant data. We fitted the models using Bayesian inference via MCMC using a Gibbs sampling scheme. For the original **DP** model, we used the R package GJAM [48]. We implemented the $\mathbf{DP}_\mathrm{c}$ and $\mathbf{PY}_\mathrm{c}$ models in R by extending the original GJAM R package. In particular, we implemented

an additional adaptive Metropolis–Hasting step (for $\mathbf{DP}_c$) and the multi-step algorithm proposed by [27] to sample from $\mathbf{PY}_c$ (see details in Appendix Section S.6). Our code (an extension of the GJAM package) can be found at the first author's Github repository (https://github.com/dbystrova/GJAM_clust). The prior on the number of clusters was set using the number of plant functional groups (PFGs) as described above (see Appendix Section S.5 for the calibration method and for the importance of such step). For the sake of comparison, we used the same default non informative priors suggested by [9] for all other parameters of the three models. Convergence was assessed through the calculation of Gelman–Rubin diagnostics [17] or visual inspection of the trace plots.

For the dimension reduction regime in GJAM model, the number of latent factors in the first step of dimension reduction needed to be specified. The number of latent factors was chosen using the deviance information criterion (DIC) [46] (see details in Appendix Section S.2). Model fit was evaluated at the species level. Prediction performances were not the main objective of the paper, as we do not expect the residual covariance matrix to impact predictive mean values [32, 39, 53]. However, we did check that the model fitted well the data both on training and test set. The dataset was randomly partitioned into a training and a testing dataset, using $70\%/30\%$ ratio. We fitted models on the training dataset and then predicted species occurrences on the testing data, comparing the predicted and the actual occurrences, similarly to [32, 53] (cross-validation is not a doable task due to the computational costs of the models). For each species, we measured the predictive performances by calculating the area under the receiver operating characteristic curve (AUC) on both training (AUC_{in}) and testing datasets (AUC_{out}).

3.3 Ecological representation of the clusters

We hypothesized that species within the same clusters might have similar functional strategies as measured by distance in trait space. We considered the following traits: Landolt nutrient indicator, Landolt light indicator, height (in the logarithmic scale), specific leaf area (SLA), leaf dry matter content (LDMC), leaf carbon concentration (LCC), and leaf nitrogen concentration (LNC) [3]. All traits presented here were available for at least 70 % of the species. For a more intuitive understanding, we assigned traits with a similar role in the community assembly process [2] into four categories: competitive effect (height, SLA, LDMC, LCC, LNC), tolerance to abiotic and biotic conditions (Landolt nutrient indicator, Landolt light indicator), interaction via light resources (height, SLA, Landolt light indicator) and interaction via soil resources (LNC, Landolt nutrient indicator). Specifically, we calculated the following species-specific ratio for each species, each category of traits and each clustering method (including the PFGs) to measure whether species within the same cluster share a similar range of functional traits:

$$\text{Species grouped-trait ratio} = \frac{\text{mean(distance to other species)}_{\text{within cluster}}}{\text{mean(distance to other species)}_{\text{all species}}}$$

In accordance with our hypothesis, we expected these distributions of species grouped-trait ratios to be close to zero, however not exactly zero, as exact zero would indicate the singleton clusters. This specifically indicates that the species were closer to within-cluster species than to species in the other clusters, thus fitted clusters could represent similar functional strategies. However, in our interpretation, we also penalized the clustering method when the number of singleton clusters increase as they do not serve the aim for clustering functionally similar species.

3.4 Results

3.4.1 Prediction evaluation

Table S2 in Appendix Section S.8.1 provides the predictive performances (both in-sample and out-of-sample) of the models, that are all very similar across models. The data are well explained (mean value of AUC_{in} is around 0.755), and the performances do not drop on the test dataset (AUC_{out} around 0.745).

3.4.2 Clustering properties of the models

The posterior distribution of the number of clusters of the **DP** model with a mean value of 35 was substantially lower than the prior mean of 56 (Table S1 in Appendix Section S.5.3, Figure 1). Larger variances for the $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models reduced prior weight and thus posterior distributions remained closer to their prior mean of 16, yielding a posterior mean near 20 (Figure 1).

Thus the posterior cluster estimate from the **DP** model estimated more clusters than did the $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models (18, and 20 respectively), which were closer to the number of PFG groups (16); see point estimates in Figure 1. Figure 3 provides the ARI as similarity measure between clusters estimated by each model and the PFGs. The posterior cluster estimates for all models are distant from the PFGs as the value of the ARI measure between each of the clusters and PFGs is close to zero. The **DP** is however more distant from PFGs than $\mathbf{DP}_c$ and $\mathbf{PY}_c$. The $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models yielded cluster estimates similar to one another (Figure 3). Pairwise similarities with a random partition in the Appendix (see Figure S5 in Appendix Section S.8.2) show that PFGs are closer to the estimated clusters than a random partition.

We have tested sensitivity to prior specification for the $\mathbf{PY}_c$ and $\mathbf{DP}_c$ models, specifying prior at lower (8) and larger (56) values (Figure 2). $\mathbf{PY}_c$ model which has a larger variance for the prior distribution of the number of clusters, appeared to be less sensitive to prior specification than $\mathbf{DP}_c$.

3.4.2.1 Clusters interpretation

The clusters estimated by $\mathbf{DP}_c$ and $\mathbf{PY}_c$ represent functional strategies (Figure 4), particularly for traits related to tolerance. The resulting clustering of the **DP** model contains many singleton clusters (i.e. clusters with one species), which have zero grouped-trait ratio, thus imply lower overall grouped-trait ratio for **DP** model (Figure 4). Figures S7 - S9 in Appendix show the residual covariance matrix inferred by the models.

4 DISCUSSION

Understanding what are the main environmental drivers of species distributions and biodiversity is one of the main goad of ecology. This task requires to consider a large number of species with as a consequent high computational cost of the models employed, whose feasibility depends on dimension reduction and the inclusion of an expert-based prior knowledge. Here, we presented an extension of the dimension reduction approach for joint species distribution models proposed by [48], by incorporating prior knowledge and by providing a more flexible clustering method. While reducing the dimension of the model, we provide a tool to create groups of species that share the same associations with respect to other species. For studies where a specific residual covariance structure is desirable our approach brings new flexibility to JSDMs. Our application shows a case where residual covariance is structured in agreement with functional traits, suggesting that these traits determine presence-absence beyond what is explained by the mean structure of the model.

4.1 Clustering properties of the models

The results of our case study confirm the importance of carefully choosing the prior in the **DP** model proposed by [48]. The **DP** specifies greater prior weight on the number of clusters than can be desirable in some applications. For this application where we specified a prior mean that was far from the posterior (i.e. using the default settings of [48]), the posterior distribution of the number of clusters of the **DP** model

moved far from the prior, but without the full flexibility we offer here (Figure 1). Large prior variances on the $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models make them less informative. In this application where we specified a prior mean close to the posterior, we found prior-posterior agreement.
By tuning the parameter N, the **DP** model would have also recovered the desired number of clusters. However due its very peaky prior distribution (i.e. strong prior weight), it has less flexibility to move far from the prior when sample size is limited (sample size in our case, number of plots is $n = 1\,139$, number of species is $S = 111$)). For this reasons we did not test for the ability of this model in the case study. Finally, the fact that the posterior distribution of the number of clusters of $\mathbf{PY}_c$ for different prior choices stay close, confirms that the prior on the number of clusters for the models ($\mathbb{E}[K_S] = 16$) is well chosen.

4.2 The importance of prior elicitation

Including prior knowledge is an appealing feature of Bayesian statistics, which is however often unused, or worse, misused [1]. Expert-based prior knowledge on species interactions has always been available, and it is now getting more and more accessible [28]. While co-occurrence networks should not be interpreted as interaction networks, we claim that this prior knowledge can help to separate the effect of the environment from the one of biotic interactions, to improve inference of interaction networks, but also to account for biotic interactions in predictive distribution models. In our case, prior knowledge does not concern particular species-specific interactions, but informs the model on the number of groups of species that share the same associations with other species. Although these associations should not be confounded with interactions, we suggest that our model $\mathbf{DP}_c$ is a first attempt to include prior information when building co-occurrence networks. Since time-series contain much more informations on biotic interactions than snapshot data, we could further extend our model to cluster the autoregressive coefficients of dynamic JSDMs [36, 10], in order to truly include a prior knowledge on the structure of the interaction network.

4.3 Clustering species in JSDMs framework

Thanks to new sampling techniques [e.g. eDNA, 47], community data are becoming more and more available. Learning the structure of a co-occurrence network from data with a large number of species is particularly demanding, since for a given number of nodes S, there exists 2^S possible networks. Moreover, even in case a correct inference is possible, it is not an easy task to visualize, and then summarize and analyze a large network. Clustering species allows to zoom out from the species level, focusing on a broader scale, easier to model, visualize and describe [33]. Indeed, our model both reduce the dimension of the problem and enable a better understanding of the ecosystem under study, showing how these clusters are strongly linked to functional traits. We emphasize that our method is conceptually different from applying a clustering method (e.g. hierarchical clustering) on the inferred residual correlation matrix, because in that case species in the same cluster do not exactly share the same residual associations and the dimensions of the model would not be reduced. Finally, we notice that since we cluster the residual correlation matrix, we do not filter out indirect associations [22, 44]. To do so, our model should be extended to cluster the residual partial correlation matrix (i.e. the inverse of the residual correlation matrix) to truly represent a co-occurrence network, and not a network of marginal correlations (that represent both direct and indirect associations).

4.4 The role of functional traits to shape community assemblages

With our case study, we show how the proposed clustering methodology could facilitate the description and provide better insights of the residual covariance matrix. Firstly, while we acknowledge that such a residual covariance matrix should not be interpreted as a species interaction network, we believe that we can still attribute a certain ecological meaning to the residual associations between species. Indeed, species within the same inferred clusters share similar competitive abilities, similar tolerance level to abiotic and biotic conditions and interact in the same way even when we consider ecological processes at

different levels such as interactions for light and soil resources (Figure 4). Moreover, species within the same clusters tend to be positively correlated (Appendix Figures S8, S9 in the Appendix S.9). For example, with both clustering methods ($\mathbf{DP}_c$ and $\mathbf{PY}_c$), we observed that *Sorbus aria* (i.e. mountainous tree) and *Hieracium murorum* form a cluster. The latter being a mountainous understory herbaceous species it needs the shade provided by the former: therefore, the two are positively correlated in the residual covariance matrix. Another example is given by five species (*Lonicera xylosteum, Corylus avellana, Mercurialis perennis, Hedera helix, Fraxinus excelsior*) from different life forms that are always grouped together with both clustering methods ($\mathbf{DP}_c$ and $\mathbf{PY}_c$) for that they are all found mainly at forest edges and the herbaceous understory species, *Mercurialis perennis*, benefiting from the shade of the trees. In addition, the same cluster is negatively correlated with the big cluster no. 5 (built with $\mathbf{PY}_c$ method, Appendix Figure S9) that is mainly grouping lowland to subalpine species that are shade-intolerant but can tolerate nutrient poor soils. In sum, the groups of species that we build represent those species that tend to co-occur together more than expected by the lens of observed environmental variables. Notice that this might also be an indication that species within the same clusters also happen to be in similar habitats, suggesting missing environmental variables. Notably, the fact that species within the same clusters tent to show similar values of the Landolt nutrient indicator suggests that soil properties might explain some of the residual correlations (Appendix Figure S3).

Moreover, we believe that these results will also have practical advantages. The PFG building framework [2] allows to group species according to their functional strategies in the aim of reducing the botanical complexity in dynamic vegetation models. As shown here, our models provide clusters that could represent similarity in tolerance to abiotic and biotic conditions and their competitive ability at least as much as PFGs. Hence, the obtained clusters in our case can be considered as a valuable alternative to the PFG building framework, that requires the availability of many functional traits for most species.

Despite these improvements and advantages, we also acknowledge some possible pitfalls. Notably, missing covariates have always the potential to drive the patterns in the residual covariance matrix. The fact that our clusters performed well in representing the traits related to tolerance to abiotic conditions might be an indication of such a problem. Among these traits, Landolt nutrient indicator represents soil nutritive requirements of plants and was quite well represented by the clusters (Appendix Figure S3). Having in mind that we were not able to include soil data among the covariates for this case study due to data availability, it is possible that the residual co-occurrence patterns are also driven by the soil properties. Another way forward in the framework would also be including habitat and soil information as covariates to further investigate if we can retrieve different patterns in the residual covariance matrix that are more directly related to biotic interactions.

5 CONCLUSION

We propose a statistical framework that allows an additional but ecologically meaningful dimension reduction of joint species distribution models and includes prior knowledge in the residual covariance matrix, providing a tool to infer clusters of species that share the same residual associations with respect to other species. The case study shows that the obtained clusters of species are ecologically meaningful, and correlated with functional traits. Therefore, our model can also be seen as an alternative way to build functional groups without having to measure all necessary traits.

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

AUTHOR CONTRIBUTIONS

WT, JA, DB, and GP conceived the idea. AG, JA, DB, and GP developed statistical methods. DB and GP implemented the methods in R. DB, GP and BB analyzed the data for the case study. DB, GP, BB, JA, JC, WT wrote the first draft of the manuscript and all authors contributed substantially to the revisions.

FUNDING

DB, GP & WT were supported by the GAMBAS project funded by the French National Research Agency (ANR-18-CE02-0025). JA was supported by the Grenoble Alpes Data Institute, funded by the French National Research Agency (ANR-15-IDEX-02). JSC and WT also acknowledged support from the Programme d'Investissement d'Avenir under project FORBIC (18-MPGA-0004). This work also received funding from the ERA-Net BiodivERsA - Belmont Forum, with the national funder Agence National pour la Recherche (FutureWeb: ANR-18-EBI4–0009) to WT and the National Science Foundation (NSF grant No 1854976) to JSC. This work has been partially supported by MIAI @ Grenoble Alpes, (ANR-19-P3IA-0003).

DATA AVAILABILITY STATEMENT

The datasets analyzed for this study can be found at the first author's Github repository (https://github.com/dbystrova/GJAM_clust).

REFERENCES

[1] Banner, K. M., Irvine, K. M., and Rodhouse, T. (2020). The Use of Bayesian Priors in Ecology: The Good, The Bad, and The Not Great. *Methods in Ecology and Evolution, forthcoming*

[2] Boulangeat, I., Philippe, P., Abdulhak, S., Douzet, R., Garraud, L., Lavergne, S., et al. (2012). Improving plant functional groups for dynamic models of biodiversity: at the crossroads between functional and community ecology. *Global Change Biology* 18, 3464–3475

[3] Brun, P., Zimmermann, N. E., Graham, C. H., Lavergne, S., Pellissier, L., Münkemüller, T., et al. (2019). The productivity-biodiversity relationship varies across diversity dimensions. *Nature Communications* 10, 1–11

[4] Bystrova, D., Arbel, J., King, G. K. K., and Deslandes, F. (2021). Approximating the clusters' prior distribution in Bayesian nonparametric models. In *Third Symposium on Advances in Approximate Bayesian Inference*

[5] Calabrese, J. M., Certain, G., Kraan, C., and Dormann, C. F. (2014). Stacking species distribution models and adjusting bias by linking them to macroecological models. *Global Ecology and Biogeography* 23, 99–112

[6] Chen, D., Xue, Y., and Gomes, C. (2018). End-to-end learning for the deep multivariate probit model (Stockholmsmässan, Stockholm Sweden: PMLR), vol. 80 of *Proceedings of Machine Learning Research*, 932–941

[7] Chib, S. and Greenberg, E. (1998). Analysis of Multivariate Probit Models. *Biometrika* 85, 347–361. doi:10.1093/biomet/85.2.347

[8] Chiquet, J., Robin, S., and Mariadassou, M. (2019). Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning* (PMLR), 1162–1171

[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, 34–56. doi:10.1002/ecm.1241

[10] Clark, J. S., Scher, C. L., and Swift, M. (2020). The emergent interactions that govern biodiversity change. *Proceedings of the National Academy of Sciences* 117, 17074–17083. doi:10.1073/pnas.2003852117

[11] De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37, 212–229

[12] Dormann, C. F., Bobrowski, M., Dehling, D. M., Harris, D. J., Hartig, F., Lischke, H., et al. (2018). Biotic interactions in species distribution modelling: 10 questions to guide interpretation and avoid false conclusions. *Global Ecology and Biogeography* 27, 1004–1016. doi:10.1111/geb.12759

[13] Elith, J. and Leathwick, J. R. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annual review of ecology, evolution, and systematics* 40, 677–697

[14] Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1, 209–230

[15] Ferrier, S. and Guisan, A. (2006). Spatial modelling of biodiversity at the community level. *Journal of Applied Ecology* 43, 393–404

[16] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis* (Chapman and Hall/CRC)

[17] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science* 7, 457–472

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

[19] Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology Letters* 8, 993–1009. doi:10.1111/j.1461-0248.2005.00792.x

[20] Guisan, A., Thuiller, W., and Zimmermann, N. E. (2017). *Habitat suitability and distribution models: with applications in R* (Cambridge University Press)

[21] Harris, D. J. (2015). Generating realistic assemblages with a joint species distribution model. *Methods in Ecology and Evolution* 6, 465–473

[22] Harris, D. J. (2016). Inferring species interactions from co-occurrence data with Markov networks. *Ecology* 97, 3308–3314. doi:10.1002/ecy.1605

[23] Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification* 2, 193–218

[24] Hui, F. K. (2016). boral–Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods in Ecology and Evolution* 7, 744–750

[25] Lee, C. and Wilkinson, D. J. (2019). A review of stochastic block models and extensions for graph clustering. *Applied Network Science* 4, 122

[26] Lijoi, A., Prünster, I., and Rigon, T. (2020). Finite-dimensional discrete random structures and Bayesian clustering. *Carlo Alberto Notebooks*

[27] Lijoi, A., Prünster, I., and Rigon, T. (2020). The Pitman–Yor multinomial process for mixture modeling. *Biometrika* 107, 891–906

[28] Maiorano, L., Montemaggiori, A., Ficetola, G. F., O'Connor, L., and Thuiller, W. (2020). TETRA-EU 1.0: A species-level trophic metaweb of European tetrapods. *Global Ecology and Biogeography* doi:10.1111/geb.13138

[29] Merow, C., Smith, M. J., Edwards Jr, T. C., Guisan, A., McMahon, S. M., Normand, S., et al. (2014). What do we gain from simplicity versus complexity in species distribution models? *Ecography* 37, 1267–1281. doi:10.1111/ecog.00845

[30] Moscone, F., Tosetti, E., and Vinciotti, V. (2017). Sparse estimation of huge networks with a block-wise structure. *The Econometrics Journal* 20, S61–S85

[31] Muliere, P. and Secchi, P. (2003). Weak convergence of a Dirichlet-multinomial process. *Georgian Mathematical Journal* 10, 319–324

[32] Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., et al. (2019). A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels. *Ecological Monographs* 89, 834– 848. doi:10.1002/ecm.1370

[33] Ohlmann, M., Miele, V., Dray, S., Chalmandrier, L., O'connor, L., and Thuiller, W. (2019). Diversity indices for ecological networks: a unifying framework using hill numbers. *Ecology Letters* 22, 737–747

[34] Ovaskainen, O. and Abrego, N. (2020). *Joint Species Distribution Modelling: With Applications in R* (Cambridge University Press)

[35] Ovaskainen, O., Abrego, N., Halme, P., and Dunson, D. (2016). Using latent variable models to identify large networks of species-to-species associations at different spatial scales. *Methods in Ecology and Evolution* 7, 549–555. doi:10.1111/2041-210X.12501

[36] Ovaskainen, O., Tikhonov, G., Dunson, D., Grøtan, V., Engen, S., Saether, B.-E., et al. (2017). How are species interactions structured in species-rich communities? A new method for analysing time-series data. *Proceedings of the Royal Society B: Biological Sciences* 284. doi:10.1098/rspb.2017.0768

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

[38] O'Connor, L. M., Pollock, L. J., Braga, J., Ficetola, G. F., Maiorano, L., Martinez-Almoyna, C., et al. (2020). Unveiling the food webs of tetrapods across europe through the prism of the eltonian niche. *Journal of Biogeography* 47, 181–192

[39] Pichler, M. and Hartig, F. (2020). A new method for faster and more accurate inference of species associations from novel community data. *arXiv preprint arXiv:2003.05331*

[40] Pitman, J. (2006). *Combinatorial Stochastic Processes: Ecole d'Eté de Probabilités de Saint-Flour XXXII-2002* (Springer)

[41] Pitman, J. and Yor, M. (1997). The two-parameter poisson-dirichlet distribution derived from a stable subordinator. *The Annals of Probability* , 855–900

[42] Poggiato, G., Münkemüller, T., Bystrova, D., Arbel, J., Clark, J., and Thuiller, W. (2021). On the interpretations of joint modelling in community ecology. *Trends in Ecology and Evolution, in press*

[43] Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., et al. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model. *Methods in Ecology and Evolution* 5, 397–406. doi:10.1111/2041-210X.12180

[44] Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K., and Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution* 10, 1571–1583

[45] Robert, C. P. and Casella, G. (2004). *Monte Carlo statistical methods* (Springer-Verlag, New York)

[46] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series b (statistical methodology)* 64, 583–639

[47] Taberlet, P., Coissac, E., Hajibabaei, M., and Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology* 21, 1789–1793

[48] 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, 939–967

[49] Thuiller, W., Guéguen, M., Bison, M., Duparc, A., Garel, M., Loison, A., et al. (2018). Combining point-process and landscape vegetation models to predict large herbivore distributions in space and time—A case study of Rupicapra rupicapra. *Diversity and Distributions* 24, 352–362

[50] Thuiller, W., Münkemüller, T., Lavergne, S., Mouillot, D., Mouquet, N., Schiffers, K., et al. (2013). A road map for integrating eco-evolutionary processes into biodiversity models. *Ecology letters* 16, 94–105

[51] Vanhatalo, J., Hartmann, M., and Veneranta, L. (2020). Additive multivariate gaussian processes for joint species distribution modeling with heterogeneous data. *Bayesian Anal.* 15, 415–447. doi:10.1214/19-BA1158

[52] Warton, D. I., Blanchet, F. G., O'Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., et al. (2015). So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* 30, 766–779

[53] Wilkinson, D. P., Golding, N., Guillera-Arroita, G., Tingley, R., and McCarthy, M. A. (2019). A comparison of joint species distribution models for presence–absence data. *Methods in Ecology and Evolution* 10, 198–211

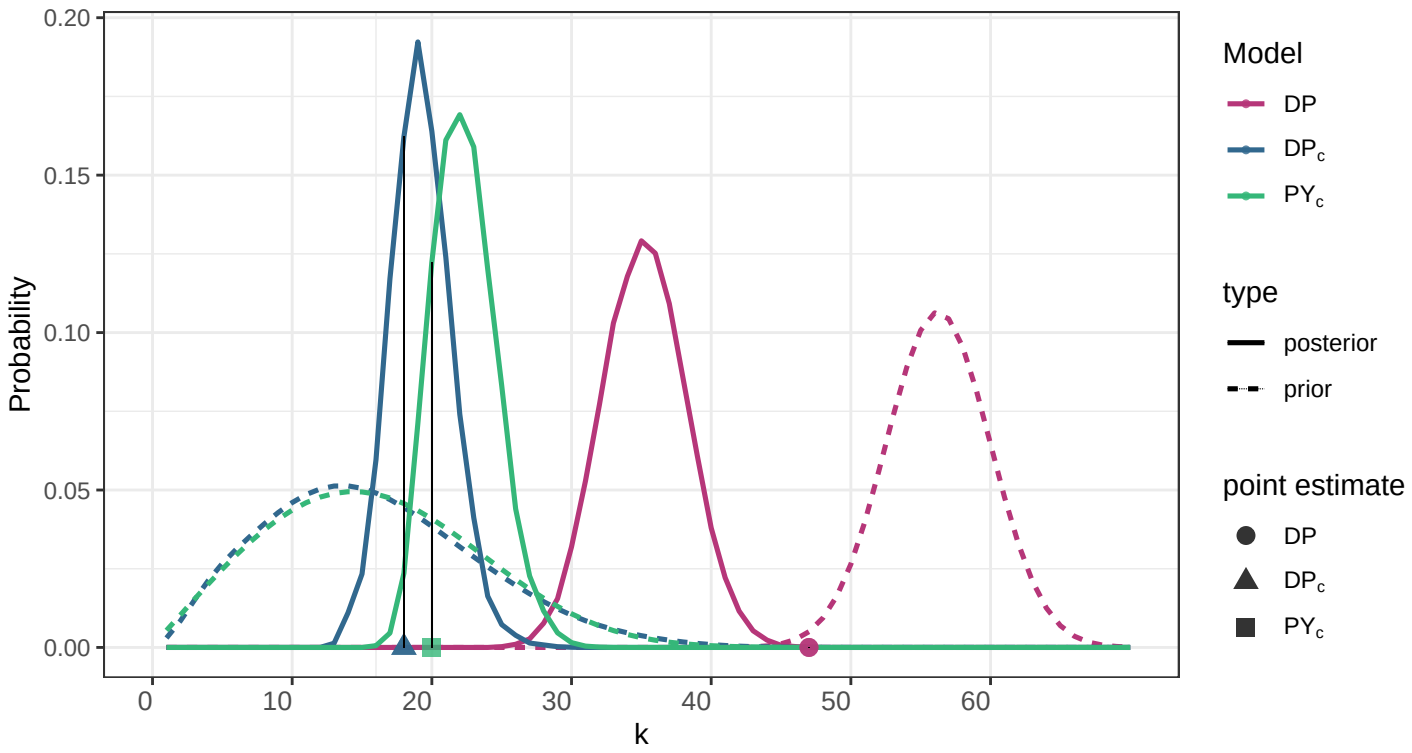

Figure 1. Prior distribution and posterior estimation of the number of clusters corresponding to **DP**, **DP**$_c$, **PY**$_c$ models. For **DP**$_c$ and **PY**$_c$ models, prior distribution is specified such that $\mathbb{E}[K_S]$ matches the prior ecological knowledge (in our case it is the number of PFGs). Posterior estimation for all the models is represented by the posterior distribution of the number of clusters (solid lines) and the number of clusters of the posterior cluster estimate (points on x-axis). Refer to the clustering estimation procedure described in Appendix Section S.7 for a pointer on why the size of the cluster estimate can be distant from the bulk of the posterior distribution.

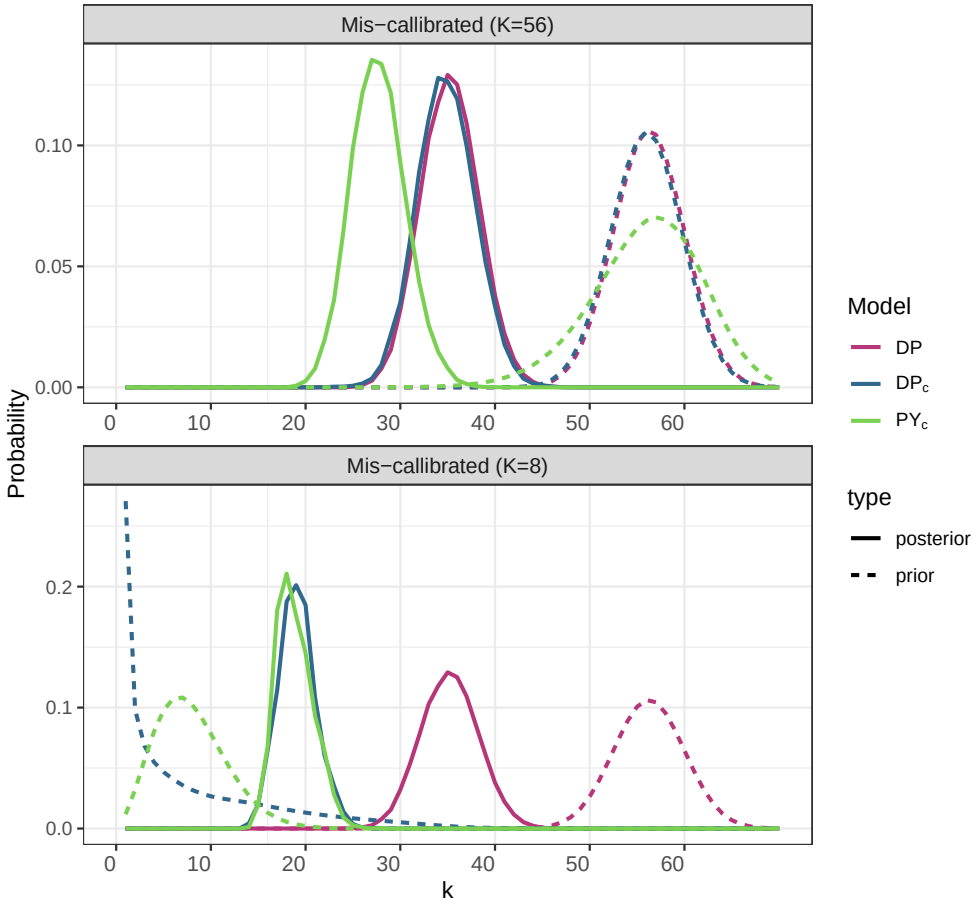

Figure 2. Prior and posterior distribution of the number of clusters for the models $\mathbf{DP_c}$, $\mathbf{PY_c}$ corresponding to the different prior specification of the number of clusters, where prior expected number of clusters $\mathbb{E}[K_S]$ take values in $\{8, 56\}$.

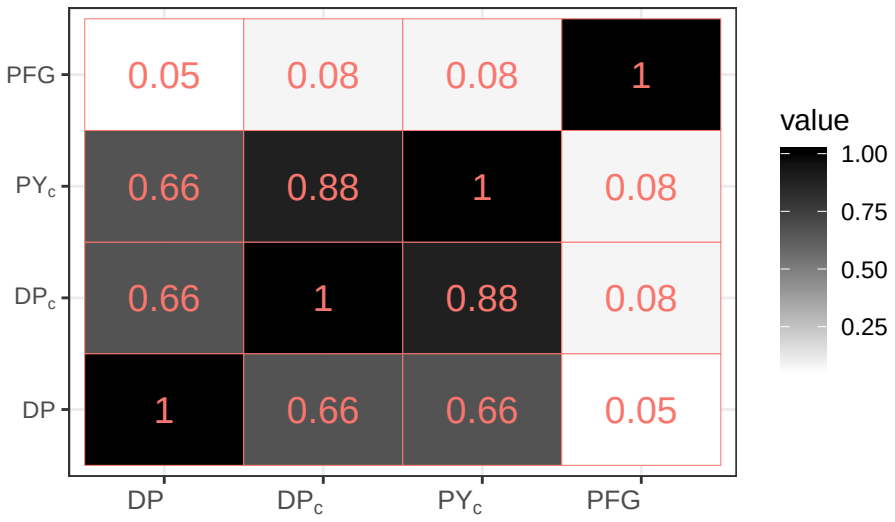

Figure 3. Pairwise ARI similarities between the PFGs and the clusters estimated by the models (**DP**, **DP**$_c$, **PY**$_c$)

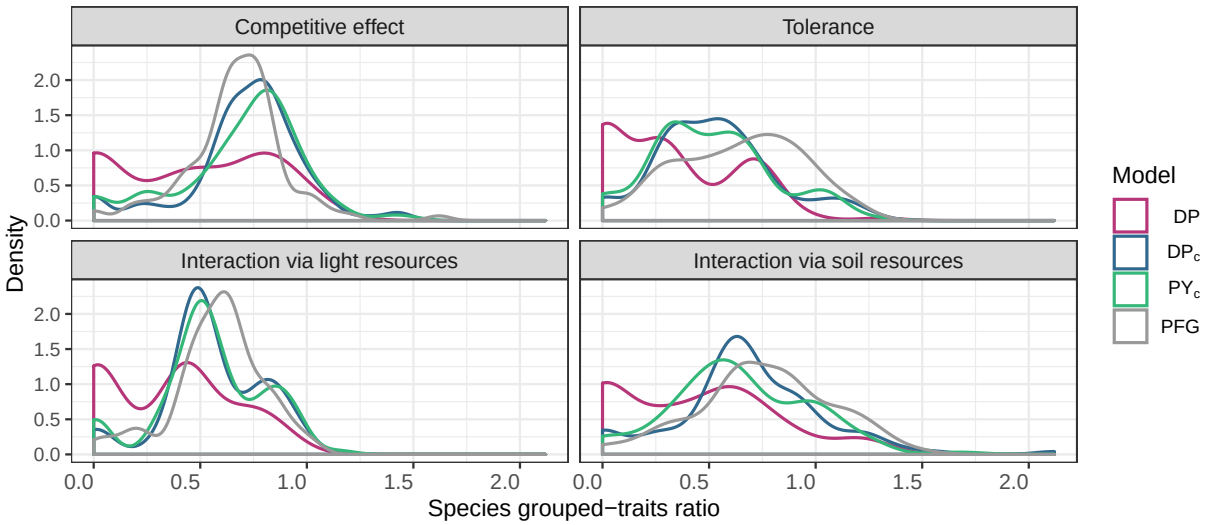

Figure 4. Distribution of species grouped-trait ratio for different trait categories and for all clustering methods. The reference curve is the distribution of species grouped-trait ratio of PFGs. (**DP**,**DP**$_c$, **PY**$_c$)

Table 1. Specification of concentration parameter α for the **DP**, **DP**$_c$, **PY**$_c$ models. K^* is the prior ecological belief on the number of groups of species with the same residual correlation structure.

Model	**Concentration parameter** α	**Reference**
DP	Fixed (number of species)	[48]
DP$_c$	Ga(ν_1, ν_2) s.t $\mathbb{E}[K_S] = K^*$	Ours
PY$_c$	Fixed, s.t $\mathbb{E}[K_S] = K^*$	Ours + [27]