frontiers

Supplementary Material

S.1 DATA PREPROCESSING

For the environmental covariates, we considered precipitation of coldest quarter, annual precipitation, annual mean temperature, mean temperature of the wettest quarter, and topographic predictors (i.e. slope) at 100 m resolution. For the purpose of the analysis in this paper, we used the two main PCA axes for summarizing the environmental covariates (see Fig. S1), which are related to 80% of the variation. These two first axes are included as covariates with a quadratic term (using orthogonal polynomials). After retaining the plots with at least 60% of non-missing entries, we obtained 1 139 plots. We used for analysis 111 most abundant species out of 136 dominant species, that were present at least at 2% of plots. The remaining species were pooled into the 'other' category, using the build-in gjamTrim function in the GJAM package. However, we discarded these rare species for further analysis.

S.2 SELECTION OF THE NUMBER OF LATENT FACTORS

We use the low-rank matrix $\mathbf{\Lambda}$ with rank r for the approximation of the residual covariance matrix. The rank of matrix $\mathbf{\Lambda}$ determines the resolution with which JSDMs models the residual covariance. The higher the rank, the closer we are to the full covariance matrix, while at the same time it means that we need to estimate more parameters. In the GJAM model, we need to specify this parameter before fitting the model. We have chosen $r = 5$ using the Deviance information criterion (DIC) implemented in the model (Figure S2). This value is coherent with the values used in Chen et al. (2018); Taylor-Rodriguez et al. (2017); Warton et al. (2015), where the low-rank matrices were able to well approximate the full-rank residual covariance matrices. For the identifiability issue, the matrix $\mathbf{\Lambda}$ have to be full column rank (Taylor-Rodriguez et al., 2017; Geweke and Singleton, 1980). So in our case, $\text{rank}(\mathbf{\Lambda}) = \min\{r, K\}$, where r is the number of factors and K is the number of repeated rows (number of clusters) in matrix $\mathbf{\Lambda}$, so r would be upper bounded by K. In our case, we have a prior guess on K and r satisfies the inequality for the prior guess. However, as we do not know if the prior guess is correct, we need to inspect the posterior distribution of the number of clusters. From the analysis of the posterior distribution on the number of clusters, we can confirm that the number of factors is smaller than the posterior estimate $\hat{K}$ (the smallest estimate is 18).

S.3 DISTRIBUTION OF SPECIES TRAIT RATIO FOR DIFFERENT TRAITS

We computed distribution of species trait ratio for different 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), leaf nitrogen concentration (LNC)) for all models (**DP**, $\mathbf{DP}_c$, $\mathbf{PY}_c$) (Figure S3).

S.4 TECHNICAL BACKGROUND ON DIRICHLET AND PITMAN–YOR PROCESSES

S.4.1 Dirichlet process

The main purpose of this section is to give a brief description of the *Dirichlet process* (DP) and the *Pitman–Yor process* (PY) and clarify the differences between them. For a complete introduction, we refer

to classical literature in Bayesian nonparametrics such as Hjort et al. (2010); Ghosal and Van der Vaart (2017). The distributions of both processes are often used as Bayesian nonparametric priors. The Dirichlet process could be defined through a constructive representation proposed by Sethuraman (1994), so-called stick-breaking construction. Start by fixing parameter $\alpha > 0$ and a probability measure H. Then consider two independent families of random variables:

$$V_k \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha) \qquad \text{and} \qquad \eta_k^\star \overset{\text{iid}}{\sim} H, \quad k = 1, 2, \ldots \tag{S1}$$

Define the random weights as:

$$\begin{aligned} p_1 &= V_1, \\ p_k &= V_k \prod_{j=1}^{k-1} (1 - V_j), \qquad k = 2, 3, \ldots \end{aligned} \tag{S2}$$

If

$$G := \sum_{k=1}^{\infty} p_k \delta_{\eta_k^\star}, \tag{S3}$$

then $G \sim \text{DP}(\alpha, H)$, i.e G is a Dirichlet process with parameters α, H. The construction of $p = \{p_k\}$ can be understood metaphorically as follows. Starting with a stick of length 1, we break it at V_1, assigning p_1 to be the length of stick we just broke off. Now recursively break the other portion to obtain p_2, p_3 and so forth. Notice that $\sum_{k=1}^{\infty} p_k = 1 \; a.s.$

In the stick-breaking representation (S3) of the Dirichlet process, the concentration parameter tunes the distribution on the random weights. When parameter α is small, the Beta-distributed variables V_i tend to be closer to 1 than to 0, implying that the random weights rapidly decrease (in expectation). And the opposite, when α is large, the random weights decrease slowly in expectation. Therefore, realisations of G will be more concentrated in a few clusters when α is small, and the number of clusters increases with α.

The Dirichlet process is an infinite dimensional process and there are two main approaches for sampling from this distribution. One could sample from the Dirichlet process using marginal sample or sample from the approximation of the Dirichlet process with finite-dimensional representation. In the first case, so-called marginal methods exhibit slow mixing and are not considered for Gibbs sampling in the GJAM model. Approximation of the Dirichlet process based on truncating the stick-breaking representation explored by Ishwaran and James (2001) is a common choice, where the truncation number N has to be defined based on the desired approximation error. Another possibility is the finite-dimensional representation by Dirichlet multinomial process, which approximates the Dirichlet process in the limit (Muliere and Secchi, 2003), and could be used as well as finite dimensional prior for some finite N (Müller et al., 2015). In the **DP** model, the latter representation is used, which allows tractable computation with Gibbs sampling. The sampling scheme for the **DP** is described in Section S.6.

S.4.2 Pitman–Yor process

The Pitman–Yor process is a generalization of the Dirichlet process and is also a special case of a larger class of priors known as Gibbs-type priors, which were introduced in the seminal works of Pitman (2003) and Gnedin and Pitman (2006) (see De Blasi et al., 2015, for a review).

The Pitman–Yor process was introduced by Pitman and Yor (1997). Similarly to the Dirichlet process, the Pitman–Yor could be defined using stick-breaking construction. Fix scalar parameters $\alpha > 0$ and $\sigma \in (0, 1)$ and a probability measure H. Then consider two independent families of random variables:

$$V_k \sim \text{Beta}(1 - \sigma, \alpha + k\sigma) \qquad \text{and} \qquad \eta_k^\star \overset{\text{iid}}{\sim} H, \quad k = 1, 2, \ldots \tag{S4}$$

Define random weights as:

$$p_1 := V_1, \qquad p_k := V_k \prod_{j=1}^{k-1} (1 - V_j). \tag{S5}$$

If

$$G = \sum_{k=1}^{\infty} p_k \delta_{\eta_k^\star}, \tag{S6}$$

then $G \sim \text{PY}(\alpha, \sigma, H)$, i.e G is the Pitman–Yor process with concentration parameter α, discount (or diversity) parameter σ and base measure H. Notice that when $\sigma = 0$ it holds: $PY(\alpha, 0, H) = DP(\alpha, H)$.

The difference in terms of the stick-breaking representation is in the sampling of the Beta distributed variables V_k defining the stick-breaking weights, they now depend on two parameters. These variables are no more identically distributed, and when k increases, the second parameter increases, leading to a decrease in the mean value of V_k. Parameter V_k decreases (in mean) when k grows, meaning that the weights p_k will decrease (in mean), slower than in the Dirichlet process. That implies that the Pitman–Yor process has a greater number of distinct clusters. Indeed, when the data of sample size S is modelled with the Dirichlet process, then the number of distinct clusters K_S has logarithmic growth with S, while for the Pitman–Yor process K_S grows as a power law with S (specifically, S^σ).

Similarly to the Dirichlet process, there exist two main approaches for sampling the Pitman–Yor process, marginal and conditional. However, the flexibility of this process comes with some cost. Sampling of the Pitman–Yor process based on truncation is computationally expensive, especially for large values of σ. Lijoi et al. (2020) define the finite-dimensional Pitman–Yor multinomial process which approximates the Pitman–Yor process and allows tractable computation. Pitman–Yor multinomial process is a finite parameter prior and we use this prior for our model $\mathbf{PY_c}$.

S.5 SPECIFICATION OF THE HYPERPARAMETERS FOR THE PRIOR DISTRIBUTION

For all the models (**DP**, $\mathbf{DP_c}$, $\mathbf{PY_c}$) we used N terms in the finite-dimensional representations of the prior processes equal to the number of species S. This is a natural choice because N defines an upper bound of the possible number of clusters, which is the number of species in our definition. However, in order to further reduce the dimension in the case of large number of species S, N is bounded by 150 for **DP** and $\mathbf{DP_c}$, so the finite dimension representations well approximate the infinite process. For the $\mathbf{PY_c}$ model, the value of N could be chosen by computational tractability. Then, on each step of Gibbs sampling, we sample the $n = N$ (i.e, in our case $n = S$) times from the given process to obtain the rows in matrix $\boldsymbol{A}$.

S.5.1 Dirichlet process model ($\mathbf{DP_c}$)

For each value of parameter α and number of terms in the Dirichlet multinomial process N, we can use formula (4) with $n = S$ to compute expected number of clusters $\mathbb{E}[K_S](\alpha, N)$. As N,S are fixed for all our models and α is a random variable with distribution $\text{Ga}(\nu_1, \nu_2)$, $\mathbb{E}[K_S](\alpha, N)$ would be a function of (ν_1, ν_2): $\mathbb{E}[K_S] = f(\nu_1, \nu_2)$. We search such pair (ν_1, ν_2) that $f(\nu_1, \nu_2) = K^*$. For identifiability, we

fix the variance of Gamma distribution at 20, so we can solve the equation for the mean value of this distribution (ν_1/ν_2) and then recover ν_1, ν_2 from the variance relation $\nu_1 = 20\nu_2^2$. As there is no closed-form solution, we used a simulation-based approach for approximating function f. This is done by Monte Carlo as follows:

$$\alpha_i \overset{\text{iid}}{\sim} \text{Ga}(\nu_1, \nu_2), \qquad i = 1, \ldots, m \tag{S7}$$

$$\mathbb{E}[K_S] = \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}[K_S](\alpha_i). \tag{S8}$$

For each pair of (ν_1, ν_2) we compute the prior expected number of clusters $\mathbb{E}[K_S]$ as: (1) sampling m times the value of α_i from $\text{Ga}(\nu_1, \nu_2)$, (2) computing $\mathbb{E}[K_S](\alpha_i)$, (3) $\mathbb{E}[K_S]$ is the average over m values of $\mathbb{E}[K_S](\alpha_i)$. We solve numerically the approximation equation $\mathbb{E}[K_S] = f^*(\nu_1, \nu_2)$ along with constraint $\nu_1 = 20\nu_2^2$ and obtain the pair (ν_1, ν_2).

S.5.2 Pitman–Yor process model (PY_c)

For the Pitman–Yor multinomial prior parameters (α, σ) are fixed. For this process, the prior number of clusters is given in Theorem 3 of Lijoi et al. (2020) and it is equal to:

$$\mathbb{P}(K_{n,N} = k) = \frac{N!}{(N-k)!\sigma(\alpha+1)_{n-1}} \sum_{l=k}^{n} \frac{1}{N^l} \frac{\Gamma(\alpha/\sigma + l)}{\Gamma(\alpha/\sigma + 1)} \mathscr{S}_{l,k} \mathcal{C}_{n,l} \tag{S9}$$

for any $k \leq \min\{N, n\}$, N is the number of terms in finite-representation, n is the number of samples, $\mathscr{S}_{l,k}$ is the Stirling number of the second kind and $\mathcal{C}_{n,k}$ is the following generalized factorial coefficient (see Charalambides, 2005):

$$\mathcal{C}_{n,k} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^i \binom{k}{i} (-i\sigma)_n \tag{S10}$$

Using Equation (S9) for $n = S$ for the distribution of K_S, we can compute the expectation $\mathbb{E}[K_S]$ and variance $V[K_S]$. Specifying $\mathbb{E}[K_S] = K^*$ and defining some value for the variance $V[K_S]$, we can find some pair (α, σ) by numerically solving the system of equations. However, in our implementation, we firstly define desired value σ, and find suitable α solving numerically the equation for $\mathbb{E}[K_S](\alpha) = K^*$.

S.5.3 Hyperparameters for the priors

The hyperparameters used in the model fitting are reported in Table S1.

S.5.4 Prior distribution of the number of clusters

The prior distribution of the number of clusters in the **DP** model could be calibrated by varying N while keeping parameter α fixed (and equal to the number of species as done by Taylor-Rodriguez et al. (2017)). Another way to calibrate the **DP** model could be done by selecting a suitable parameter α. While this approach is not accessible in the **DP** model, we provide here this prior distribution on the number of clusters for comparison. The difference between these two approaches is that they imply two different prior distributions for the number of clusters. Indeed, the prior distribution induced by the Dirichlet process in the **DP** model for parameters in our case study, chosen by suitably fixing N (DPM_1 in Figure S4) is extremely peaked compared to the one induced by the Dirichlet process in the **DP** model when we instead suitably fix α (DPM_2 in Figure S4). An additional hierarchical level for parameter α leads to a higher variance of the prior distribution on the number of clusters in $\textbf{DP}_\text{c}$ model (see Figure S4).

S.6 GIBBS SAMPLING

In this section, we first describe the Gibbs sampling scheme for the original GJAM model, and then provide the description for our modified version of the Gibbs sampler. For identifiability reasons, we need to reparametrize $\boldsymbol{\Sigma}^*$: we work instead with the correlation matrix $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{\Sigma}^*\boldsymbol{D}^{-1/2}$, where $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^T + \sigma_\epsilon^2 \boldsymbol{I}_S$ and $\boldsymbol{\Lambda}$ in eq (1), and $\boldsymbol{D}$ is the diagonal matrix containing $\text{diag}(\boldsymbol{\Sigma}^*)$. Reparametrization of $\boldsymbol{\Sigma}^*$ leads to the reparametrization of latent variable (see details in Taylor-Rodriguez et al., 2017). Here $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iS})$ vector of species observations at plot i and $\Gamma(y_{ij})$ is $(-\infty, 0]$ if $y_{ij} = 0$ and $(0, \infty)$ if $y_{ij} = 1$.

The matrix $\boldsymbol{A}$ is represented as $\boldsymbol{A} = \boldsymbol{Q}(\boldsymbol{k})\boldsymbol{V}$, where $\boldsymbol{V} = (\boldsymbol{v}'_j)_{j=1}^N, \boldsymbol{v}'_j \sim H$, $\boldsymbol{V}$ is $N \times r$ matrix whose rows are all potential atoms (H is the base distribution for the Bayesian nonparametric prior). $\boldsymbol{k}$ is the vector of cluster labels $k = (k_1, \ldots, k_S), (1 \leq k \leq N)$ and $a_l = v_{k_l}$. $\boldsymbol{Q}(k)$ is the $S \times N$ matrix, such that $\boldsymbol{Q}(\boldsymbol{k}) = (e_{k_1}, \ldots, e_{k_S})$ and $\boldsymbol{e}_{k_l}$ is the N-dimensional vector with 1 in the position k_l and 0's elsewhere. An outline of the pseudo-code for Gibbs sampling algorithm in the original GJAM model is provided in Algorithm 1.

In the above scheme, we modified the sampling from the posterior distribution $[\boldsymbol{p} \mid \boldsymbol{k}]$, where $\boldsymbol{k}$ is the vector of cluster labels. In the original model (**DP**), the Dirichlet multinomial process is used as a finite approximation of the Dirichlet process with concentration parameter α and base measure $N(0, \boldsymbol{D}_z)$. Then the posterior distribution of the weights is defined by stick-breaking representation.

S.6.1 Dirichlet process with gamma prior on α ($\mathbf{DP_c}$)

In our $\mathbf{DP_c}$ model, we have employed the same representation of the Dirichlet multinomial process as in the original **DP** model. We added the hyperprior distribution for α parameter. The hyperprior distribution is defined as $\text{Ga}(\nu_1, \nu_2)$ and hyperparameters ν_1, ν_2 are defined in section S.5.1. In the case of the Dirichlet multinomial process, there is no conjugacy for the Gamma distribution, so we used the Metropolis-Hastings step with adaptation for the Gibbs sampling. The conditional density π is

$$\alpha | p \propto \frac{\Gamma(\alpha)}{\Gamma(\alpha/N)^N} p_1^{\alpha/N-1} \ldots p_N^{\alpha/N-1} \alpha^{\nu_1 - 1} e^{-\nu_2 \alpha}. \tag{S11}$$

For the sample from this distribution, we need a Metropolis–Hasting step in our Gibbs sampler. We implemented a Random Walk Metropolis–Hasting, with a truncated normal as proposal since our target distribution, has support on $\mathbb{R}^+$.

S.6.2 Pitman–Yor model process ($\mathbf{PY_c}$)

For the $\mathbf{PY_c}$ model, we have used the Pitman–Yor multinomial process introduced by Lijoi et al. (2020) as the finite dimensional approximation for the Pitman–Yor process. In Theorem 4, authors define posterior distribution for the Pitman-Yor multinomial process as a linear combination of the Dirichlet and ratio-stable distributions. We have followed the proposed sampling scheme for posterior distribution, which is described in details in the paper. Following the notations in Lijoi et al. (2020):

We denote samples from the Pitman–Yor multinomial process as $Z_1, ..Z_n \mid p_N \overset{iid}{\sim} p_N$, $p_N \sim PYM(\sigma, \alpha, N, H)$, where n is the number of samples, N is the number of terms in the Pitman–Yor multinomial process, H is the base measure. (in our case $n = N = S$). Then, $Z^{(N)} = (Z_1, .., Z_N)$ have k distinct values $k < N$ $(Z_1^*, ..Z_k^*)$, $(\overline{Z}_{k+1}, \ldots, \overline{Z}_N)$ point masses in p_N. The posterior distribution is

Algorithm 1: Gibbs sampling GJAM (Taylor-Rodriguez et al., 2017)

Input: Training data $\{\boldsymbol{x}_i, \boldsymbol{y}_i\}_{i=1,\ldots,n}$

Hyperparameters truncation number N

Initialization of parameters $\boldsymbol{B}, \boldsymbol{Z}, \boldsymbol{w}, \boldsymbol{p}, \boldsymbol{k}, \boldsymbol{A}$

for $t \leq T$ **do**

sample $\boldsymbol{z}_i^* \sim \text{tr}N_S(\boldsymbol{B^*x}_i + \boldsymbol{Aw}_i, \sigma_\epsilon^2 I_S; \Gamma(\boldsymbol{y}_i))$

sample $\boldsymbol{v}_j$ for $j = 1, \ldots, N$:

if $j \notin \boldsymbol{k}$ **then**

sample $\boldsymbol{v}_j \sim N_r(0, \boldsymbol{D}_v)$

end

else

denote $S_j = \{l = 1, \ldots, S, \text{s.t } k_l = j\}$

sample $\boldsymbol{v_j} \sim N_r(\mu_v, \Sigma_{v_j})$ where $\Sigma_{v_j} = (\frac{|S_j|}{\sigma_\epsilon^2}\boldsymbol{W}^T\boldsymbol{W} + \boldsymbol{D_v}^{-1})^{-1}$ and $\mu_v = \boldsymbol{\Sigma}_{v_j}\boldsymbol{W}^t\frac{1}{\sigma_\epsilon^2}\sum_{l\in S_j}(\boldsymbol{z}^{(l)} - \boldsymbol{X}\boldsymbol{\beta}_l)$

end

sample $\boldsymbol{w}_i \sim N_r(\Sigma_W \boldsymbol{A}\frac{1}{\sigma_\epsilon^2}(\boldsymbol{z}_i - \boldsymbol{B}\boldsymbol{x}_i))$, where $\boldsymbol{\Sigma}_W = \left(\frac{1}{\sigma_\epsilon^2}\boldsymbol{A}^t\boldsymbol{A} + \boldsymbol{I}_r\right)$

sample $\boldsymbol{k} \sim \prod_{l=1}^S \sum_{j=1}^N p_{lj}\delta_j(k_l)$, where $p_{lj} \propto p_j \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\|\boldsymbol{z}^{(l)} - \boldsymbol{X}\boldsymbol{\beta}_l - \boldsymbol{W}\boldsymbol{v}_j\|^2\right\}$

sample $\boldsymbol{p}$:

sample $v_j \sim \text{Beta}(\alpha/N + \sum_{l=1}^S I_{k_l=j}, (N-j)\alpha/N + \sum_{s=j+1}^S \sum_{l=1}^S I_{k_l=s})$, where $j = 1, \ldots, N-1$ and $\sum_{l=1}^S I_{k_k=j} = n_j$

$p_1 = v_1, p_k = v_k\prod_{l=1}^{k-1}(1-v_j), p_N = 1 - \sum_{j=1}^{N-1} p_j$

sample $\sigma_\epsilon^2 \sim \text{IG}\left(\frac{nS+\nu}{2} + 1, \frac{\sum_{i=1}^n ||\boldsymbol{z}_i - \boldsymbol{B}\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{w}_i||^2}{2} + \frac{\nu}{G^2}\right)$

sample $\boldsymbol{D}_v \sim \text{IW}(\boldsymbol{D}_v \mid 2 + r + N - 1, \boldsymbol{V}^t\boldsymbol{V} + 4\text{diag}\{1/\eta_1, \ldots, 1/\eta_r\})$

sample $\boldsymbol{B}^* \sim N_{Sp}((\boldsymbol{Z}^* - \boldsymbol{W}\boldsymbol{A}^T)^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \sigma_\epsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}, \boldsymbol{I}_S)$

Compute the variables on the correlation scale:

$\boldsymbol{z}_i = \boldsymbol{D}^{-1/2}\boldsymbol{z}_i^*$, $\boldsymbol{B} = \boldsymbol{D}^{-1/2}\boldsymbol{B}^*$ and $\boldsymbol{R} = \boldsymbol{D}^{-1/2}(\boldsymbol{A}\boldsymbol{A}^T + \sigma_\epsilon^2\boldsymbol{I})\boldsymbol{D}^{-1/2}$

end

defined by $\boldsymbol{k}$ and latent variables $\ell^k = (\ell_1, \ldots, \ell_k)$:

$$p_N \mid Z^{(N)}, \ell^{(k)} = \sum_{j=1}^k (w_j + w_{k+1}r_j)\delta_{Z^*j} + w_{k+1}\sum_{j=k+1}^N r_j\delta_{\overline{Z}_j},$$

$$(w_1, \ldots, w_k \mid Z^{(N)}, \ell^{(k)}) \sim \text{Dir}(n_1 - \ell_1\sigma, \ldots, n_k - \ell_k\sigma, \alpha + |\ell^{(k)}|\sigma).$$

$$(r_1, \ldots, r_{N-1} \mid Z^{(N)}, \ell^{(k)}) \sim \text{RS}(\sigma, \alpha + |\ell^{(k)}|\sigma, 1/N, \ldots, 1/N),$$

Algorithm 2: Metropolis–Hastings algorithm for α in $\mathbf{DP_c}$

Initialization of parameters $\alpha^{(0)}$
for $t \leq T$ **do**
  propose: $\alpha^c \sim \text{tr}N(\alpha, \sigma_p^2)$, where $\sigma_p^2 = (2.38)^2 \text{var}(\alpha^0, \ldots, \alpha^{(t-1)})$
  $p(\alpha^c \mid \alpha^t) = \min\{1, \frac{\text{tr}N(\alpha;\alpha^c)\pi(\alpha^c)}{\text{tr}N(\alpha^c;\alpha)\pi(\alpha)}\}$
  sample $u \sim \text{Uniform}(u; 0, 1)$
  if $u < p$ **then**
    $\alpha^t \leftarrow \alpha^{(c)}$
  end
  else
    $\alpha^t \leftarrow \alpha^{(t-1)}$
  end
end

where $\ell^{(k)} = (\ell_1, \ldots, \ell_k), \ell_j \in \{1, \ldots, n_j\}$, $|\ell^{(k)}| = \ell_1 + \ldots + \ell_k$ and it follows the distribution:

$$\Pr(\ell_1 = \mathrm{l}_1, \ldots, \mathrm{l}_k \mid \theta^{(n)}) \propto \Gamma(\alpha/\sigma + |\mathrm{l}^{(k)}|) \prod_{j=1}^{k} \frac{\mathcal{C}_{n_j,\mathrm{l}_j}}{N^{\mathrm{l}_j}}$$

Then sampling ℓ_k is simplified by introducing latent variable V, such that:

$$\Pr(\ell_1 = \mathrm{l}_1, \ldots, \mathrm{l}_k \mid Z^{(N)}, V) \propto \prod_{j=1}^{k} \left(\frac{V}{N}\right)^{\mathrm{l}_j} \mathcal{C}_{n_j,\mathrm{l}_j} \tag{S12}$$

$$V \mid Z^{(N)} \propto e^{-v} v^{\alpha/\sigma - 1} \prod_{j=1}^{k} \sum_{l_j=1}^{n_j} \left(\frac{V}{N}\right)^{\mathrm{l}_j} \mathcal{C}_{n_j,\mathrm{l}_j} \tag{S13}$$

The pseudo-code for Gibbs sampling algorithm used in the $\mathbf{PY_c}$ model is provided in Algorithm 3.

Algorithm 3: Algorithm for sampling $\boldsymbol{p}$ in $\mathbf{PY_c}$

Hyperparameters truncation number N
Initialization of parameters $\boldsymbol{\ell}, \boldsymbol{p}, v, \boldsymbol{k}$
for $t \leq T$ **do**
  $V \sim f(v)$
  $\boldsymbol{\ell} \sim \prod_{j=1}^{k} \left(\frac{V}{N}\right)^{\mathrm{l}_j} \mathcal{C}_{n_j,\mathrm{l}_j}$
  $\boldsymbol{W} \sim \text{Dir}(n_1 - \ell_1\sigma, \ldots, n_k - \ell_k\sigma, \alpha + |\ell^{(k)}|\sigma)$
  $\boldsymbol{R} \sim \text{RS}(\sigma, \alpha + |\ell^{(k)}|\sigma, 1/N, \ldots, 1/N)$
  $\boldsymbol{p} = (w_1 + w_{k+1}r_1, \ldots, w_k + w_{k+1}r_k, w_{k+1}r_{k+1}, \ldots, w_{k+1}r_N)$
end

Remark

1. For sampling $V \mid \boldsymbol{k}$ we used the ratio of uniforms algorithm for the density $f(v)$. We have used the **rust** package (Northrop, 2019), which use the generalized ratio-of-uniforms algorithm, based on the work of Wakefield et al. (1991).
2. For sampling ratio stable distribution we have used the approach proposed in Remark 1 in Lijoi et al. (2020). For sampling the tempered-stable distribution we used the function **rlaptrans** from Ridout (2009).

S.7 CLUSTER ESTIMATE

Bayesian nonparametric models described in this paper induce clustering on the rows of matrix $\boldsymbol{\Lambda}$. In Bayesian modeling we obtain the posterior distribution of random partition on rows of $\boldsymbol{\Lambda}$. The challenging question is how to summarize this posterior distribution and obtain cluster estimate. We follow the Bayesian method based on decision and information theoretic approaches proposed in Wade et al. (2018) and Rastelli and Friel (2018). These approaches are based on specifying the loss function $L(c, \hat{c})$, which measures the loss of estimating true clustering c with $\hat{c}$. Optimal cluster estimate is then defined through the minimization of this expected posterior loss, given the data $\boldsymbol{Y}_{1:n}$

$$c^* = \arg\min_{\hat{c}} \mathbb{E}[L(c, \hat{c}) \mid \boldsymbol{Y}_{1:n}] = \arg\min_{\hat{c}} \sum_{c} L(c, \hat{c}) p(c \mid \boldsymbol{Y}_{1:n}), \tag{S14}$$

where $p(c \mid \boldsymbol{Y}_{1:n})$ is posterior distribution of partition c. For defining the loss function, we need to specify the distance on the partition space. The two distances are generally used: the Binder and Variation of Information. In both articles, the authors, propose to use VI distance for defining the posterior clustering. One of the properties of the Binder loss, that was shown in their works, is that the Binder loss overestimates the number of clusters. For this reason, in this paper we discuss the results obtained with VI distance. Two approaches differ in the algorithm used for finding optimal solution for (S14). We have used the approach for estimation of the optimal clustering from Rastelli and Friel (2018), for computational efficiency.

It is important to mention that, for finding solution in (S14), we need to explore all possible partitions c. Partitions c that we obtain through MCMC sampling could not contain the optimal c that minimizes (S14). For exploring the partition spaces beside the obtained MCMC samples the 'greedy search' algorithm is employed. Exploring the whole partition space is computationally too expensive, so the algorithm uses the approximations scheme. In addition, the algorithm should be run several times with different starting points.

In our analysis, for each model, we used three restarts with different starting points: every point in a separate cluster, random partition with sixteen clusters, and PFG partition. From all restarts for each model, we have chosen the optimal cluster with the smallest value of the posterior expected loss.

S.8 SUPPLEMENTARY MATERIALS FOR RESULTS SECTION

S.8.1 Prediction

We evaluated model fit at the species level. Table S2 provides the predictive performances measured by calculating the area under the receiver operating characteristic curve (AUC) on both training (AUC_{in}) and testing datasets ((AUC_{out})).

S.8.2 Comparison clustering results with random partition

To complete the comparison between the clusters, estimated by our models and PFGs, we compared the cluster estimated by our models with the two different random partitions. First random partition is

randomly assign species to 16 groups (RU). Second partition assigns the species randomly to the 16 groups, where the size of the groups is the same as in PFGs (RW). Clusters estimated by **DP**, $\mathbf{DP}_c$, $\mathbf{PY}_c$ are slightly closer in terms of adjusted Rand indices (ARI) to PFGs, than to random partition.

S.8.3 Sensitivity to prior specification

In this section we provide the parameters for $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models, that were used for studying the sensitivity for the prior calibration (see Table S.10).

Moreover, we can consider the posterior distribution of the α parameter in the $\mathbf{DP}_c$ model. The expected number of clusters of the prior distribution is set to 8 and 56 correspondingly (Fig. S6). We can see that when the posterior distribution is close when the prior expected number of clusters is set to 8 and 16. While for the large prior number of clusters (56), the posterior distribution of α is concentrated at higher values.

S.9 RESIDUAL CORRELATION MATRICES

In Figures S7-S9 we provide residual correlation matrices for models **DP**, $\mathbf{DP}_c$, $\mathbf{PY}_c$.

S.10 CONVERGENCE ASSESSMENT

For each model, we ran two MCMC chains of 80 000 iterations, with 30 000 burn-in iterations. Convergence was assessed through the calculating Gelman–Rubin diagnostics or visual inspection of the traceplots. We provide here the effective sample size (Figure S10) and potential scale reduction (Figure S11) for regression coefficients and coefficients of the covariance matrix. In Figure S11 we provide only coefficients with the value of potential scale reduction factor less than 1.1, the convergence of a few coefficients with the values higher than 1.1 was confirmed through visual inspection of their traceplots. Potential scale reduction is less than 1.1 and the effective sample size is relatively high, so we can confirm satisfactory convergence. Convergence for regression coefficients is better than for coefficients of the residual covariance matrix (lower potential scale reduction factor and larger effective sample size).

REFERENCES

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

Taylor-Rodriguez D, Kaufeld K, Schliep EM, Clark JS, Gelfand AE. Joint species distribution modeling: dimension reduction using Dirichlet processes. *Bayesian Analysis* **12** (2017) 939–967.

Warton DI, Blanchet FG, O'Hara RB, Ovaskainen O, Taskinen S, Walker SC, et al. So many variables: joint modeling in community ecology. *Trends in Ecology & Evolution* **30** (2015) 766–779.

Geweke JF, Singleton KJ. Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association* **75** (1980) 133–137.

Hjort NL, Holmes C, Müller P, Walker SG. *Bayesian nonparametrics* (Cambridge University Press) (2010).

Ghosal S, Van der Vaart A. *Fundamentals of nonparametric Bayesian inference* (Cambridge University Press) (2017).

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

Ishwaran H, James L. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96** (2001) 161–173.

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

Müller P, Quintana FA, Jara A, Hanson T. *Bayesian nonparametric data analysis* (Springer) (2015).

Pitman J. Poisson-Kingman partitions. Statistics and science: a Festschrift for Terry Speed, 134. *IMS Lecture Notes Monogr. Ser* **40** (2003).

Gnedin A, Pitman J. Exchangeable Gibbs partitions and Stirling triangles. *Journal of Mathematical Sciences* **138** (2006) 5674–5685.

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

Pitman J, Yor M. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25** (1997) 855–900.

Lijoi A, Prünster I, Rigon T. The Pitman–Yor multinomial process for mixture modeling. *Biometrika, forthcoming* (2020).

Charalambides CA. *Combinatorial methods in discrete distributions*, vol. 600 (John Wiley & Sons) (2005).

Northrop PJ. *rust: Ratio-of-Uniforms Simulation with Transformation* (2019). R package version 1.3.8.

Wakefield J, Gelfand A, Smith A. Efficient generation of random variates via the ratio-of-uniforms method. *Statistics and Computing* **1** (1991) 129–133.

Ridout MS. Generating random numbers from a distribution specified by its Laplace transform. *Statistics and Computing* **19** (2009) 439.

Wade S, Ghahramani Z, et al. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis* **13** (2018) 559–626.

Rastelli R, Friel N. Optimal Bayesian estimators for latent variable cluster models. *Statistics and Computing* **28** (2018) 1169–1186.

LIST OF FIGURES

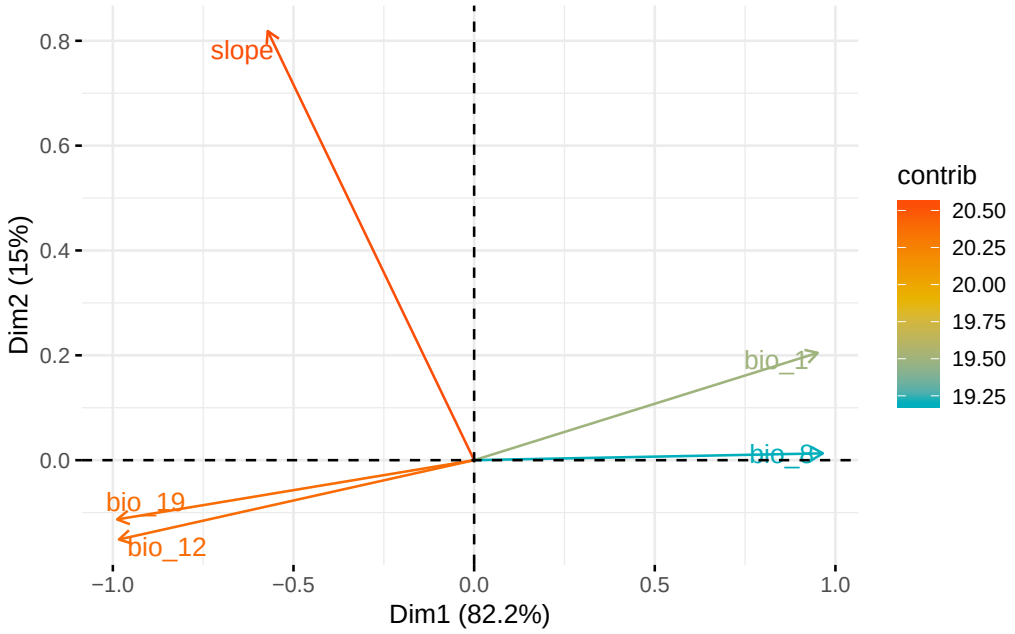

Figure S1. Biplot representation of Principle Component Analysis (PCA) performed on the environmental variables: precipitation of coldest quarter (*bio_8*), annual precipitation (*bio_12*), annual mean temperature (*bio_1*), mean temperature of the wettest quarter (*bio_19*) and topographic predictors (i.e. slope) (*slope*) at 100 m resolution.

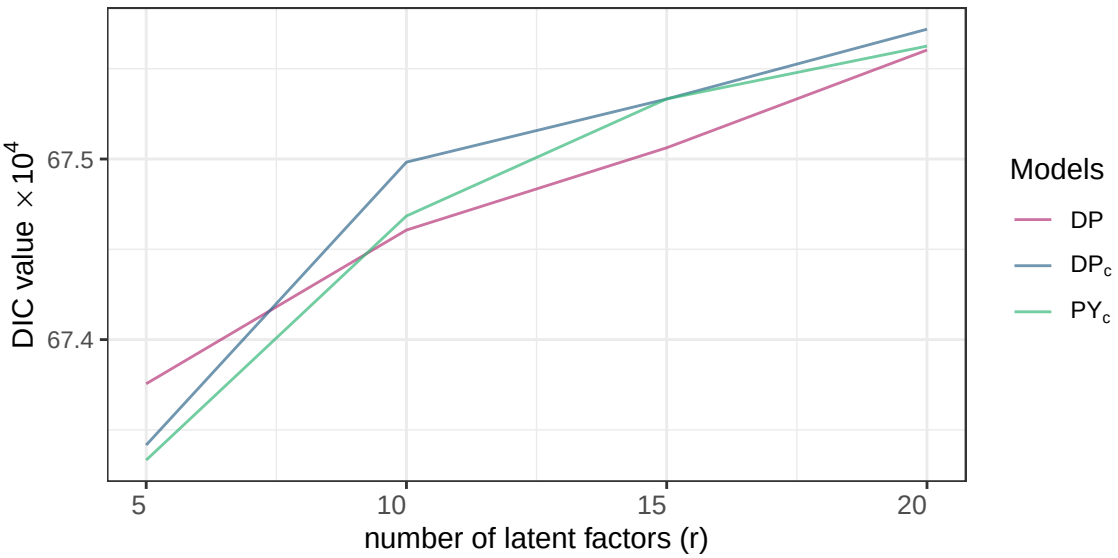

Figure S2. Deviance information criterion (DIC) values for different values r, $r = \{5, 10, 15, 20\}$ for all models **DP**, **DP**$_c$, **PY**$_c$

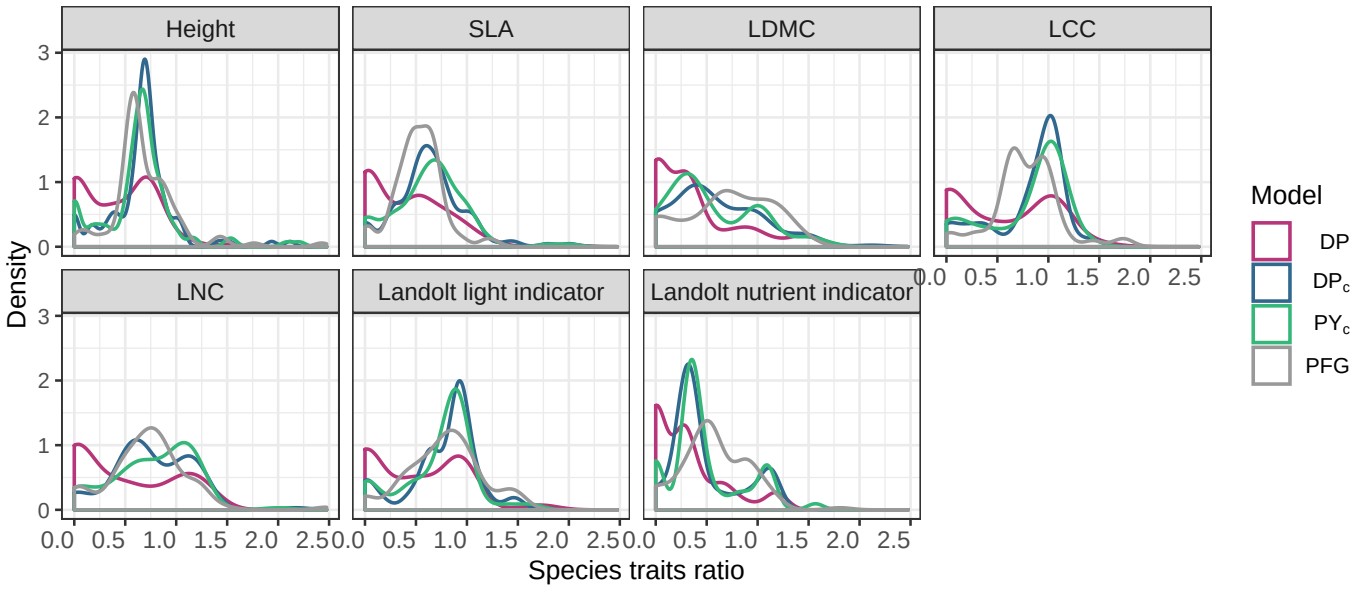

Figure S3. Distribution of species trait ratio for different traits and for all clustering methods. The reference curve is the distribution of species trait ratio of PFGs. (**DP**, **DP**$_c$, **PY**$_c$)

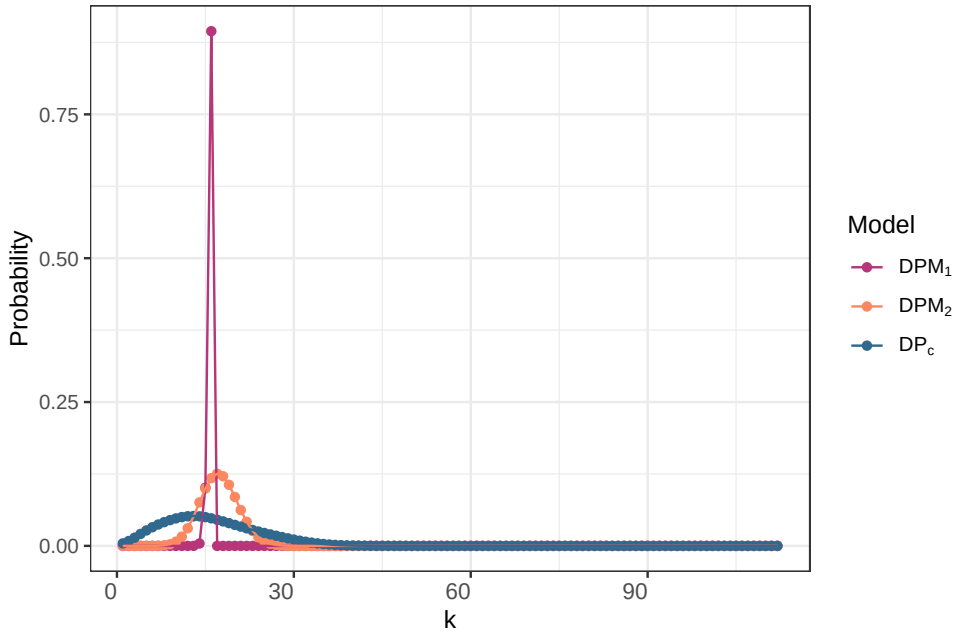

Figure S4. Prior distribution of the number of clusters K_S for **DP** model with parameters $N = 16$ and $\alpha = S$ (DPM_1), and **DP** with parameter $N = S$ and $\alpha = 6.23$ (DPM_2) and for $\textbf{DP}_\text{c}$ model specified such that $E[K_S] = 16$, $n = S$ for all the distributions.

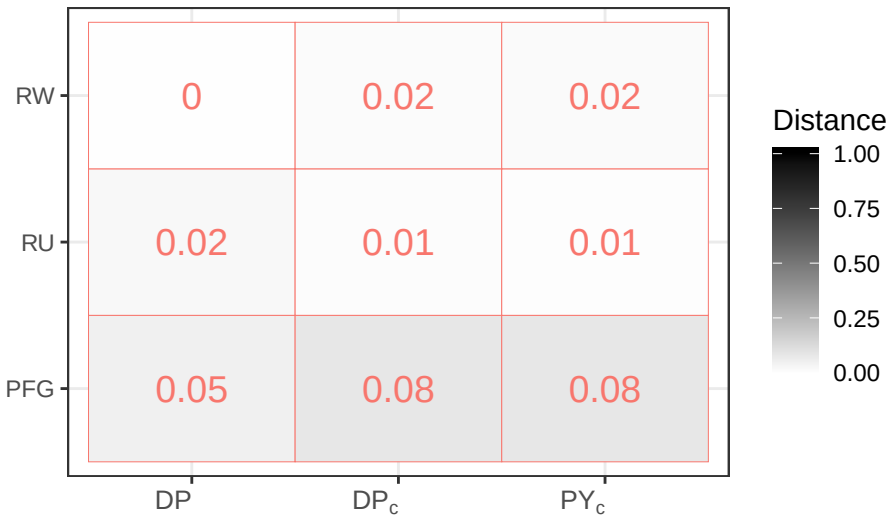

Figure S5. Pairwise adjusted Rand indices (ARI) between the the clusters estimated by the models (**DP**, **DP**$_c$, **PY**$_c$), PFGs and two random partitions of species RU and RW.

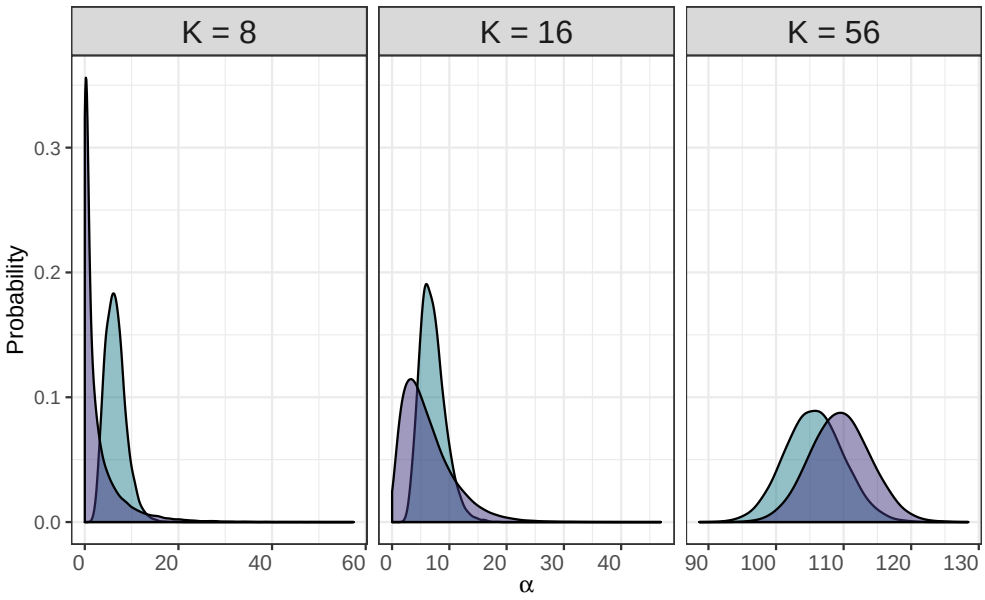

Figure S6. Prior (violet) and posterior (green) distribution of the parameter α for the model $\mathbf{DP}_{\mathrm{c}}$ and different prior specification of the number of clusters. Prior expected number of clusters $\mathbb{E}[K_S] = K$, where K take values in $\{8, 16, 56\}$.

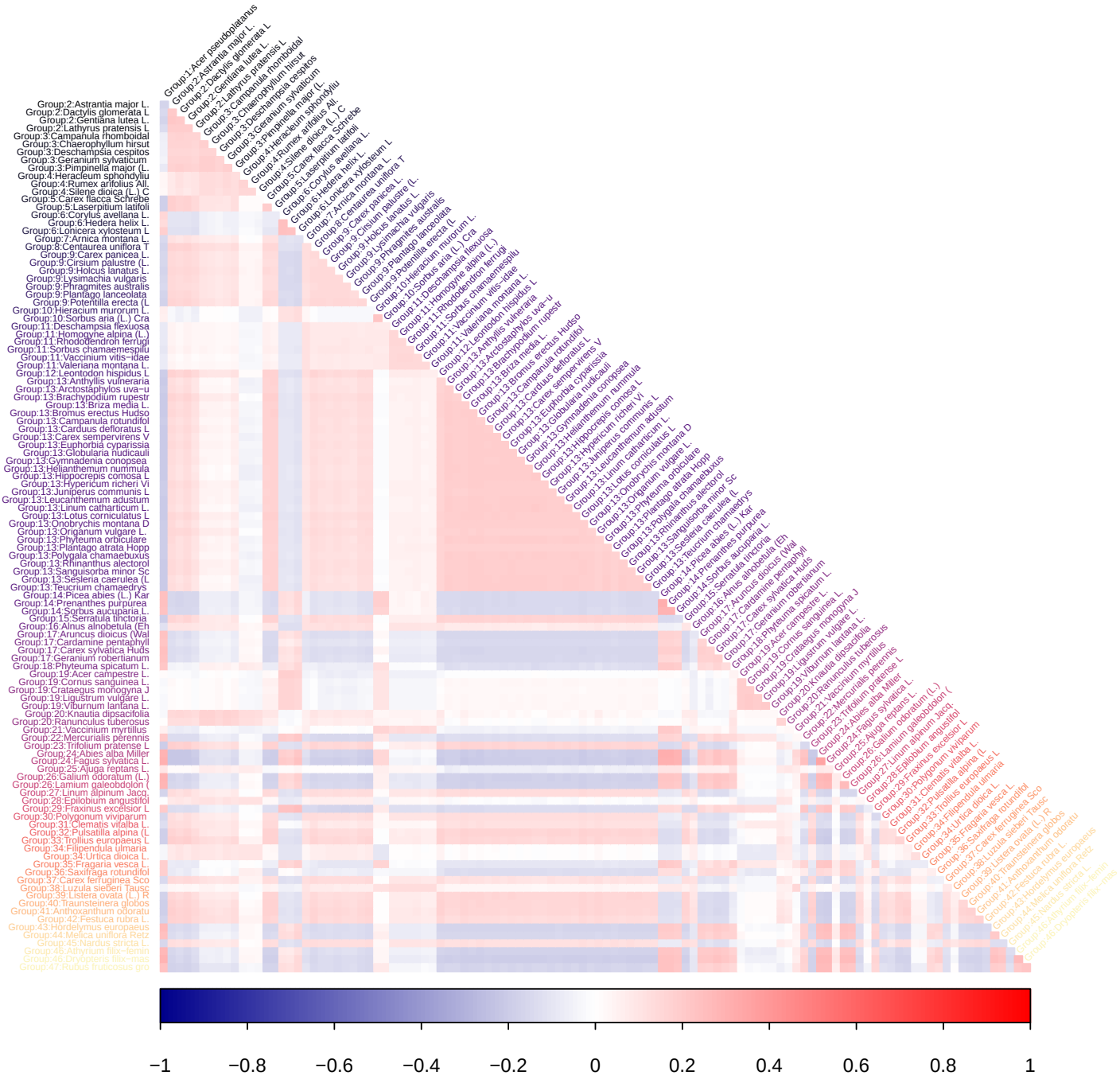

Figure S7. Residual correlation matrix for **DP** model.

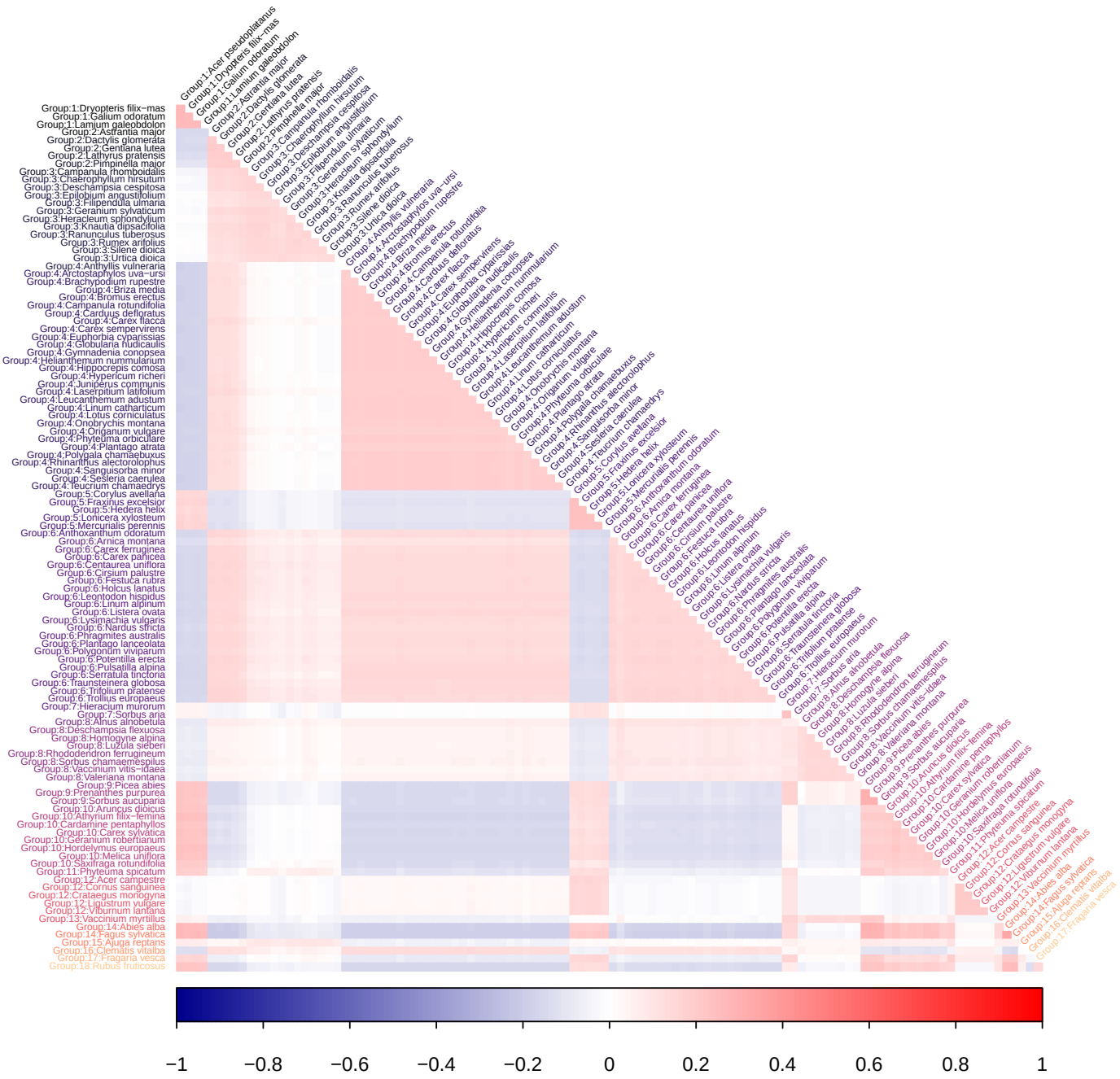

Figure S8. Residual correlation matrix for $\mathbf{DP}_c$ model.

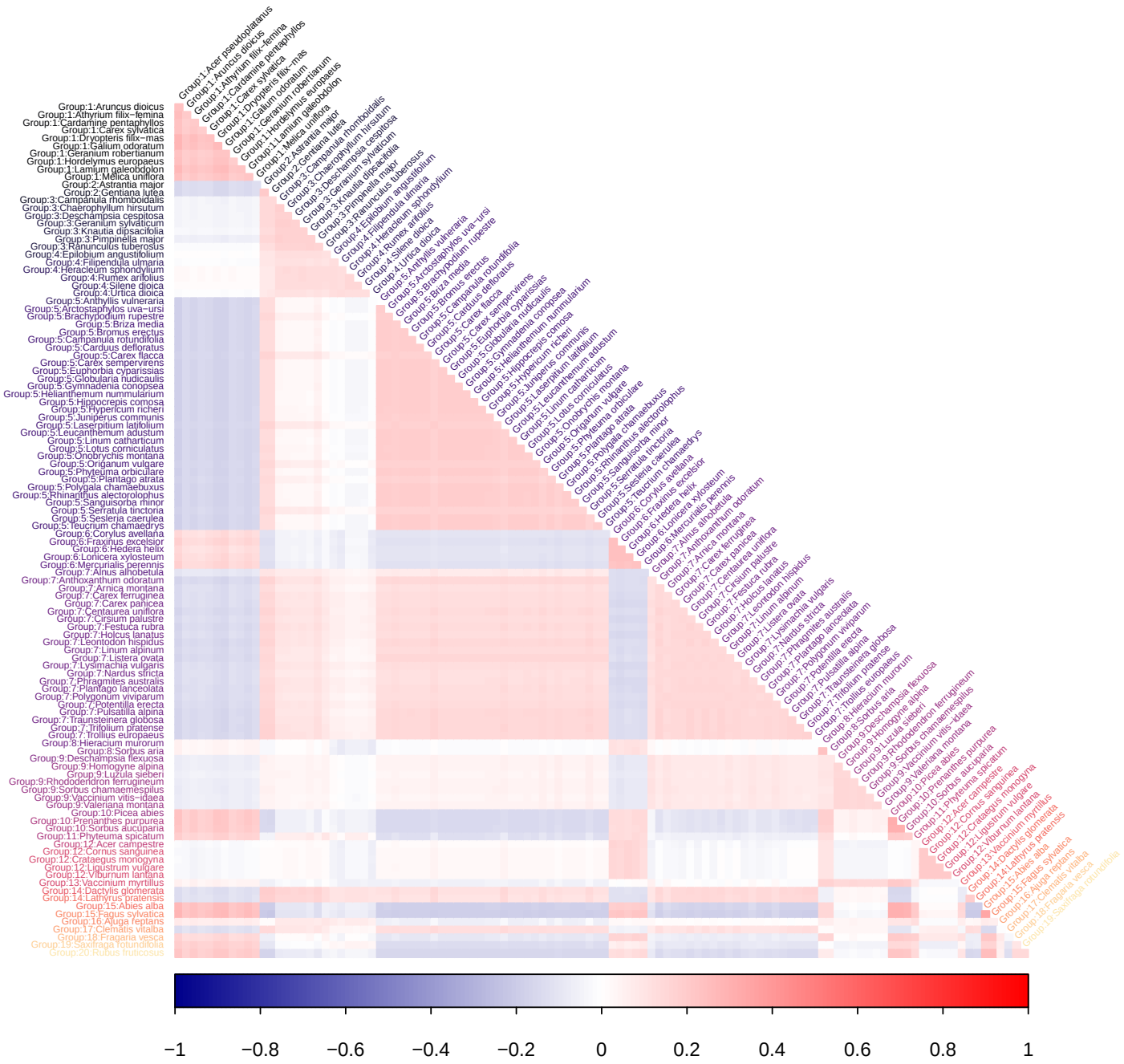

Figure S9. Residual correlation matrix for $\mathbf{PY}_c$ model.

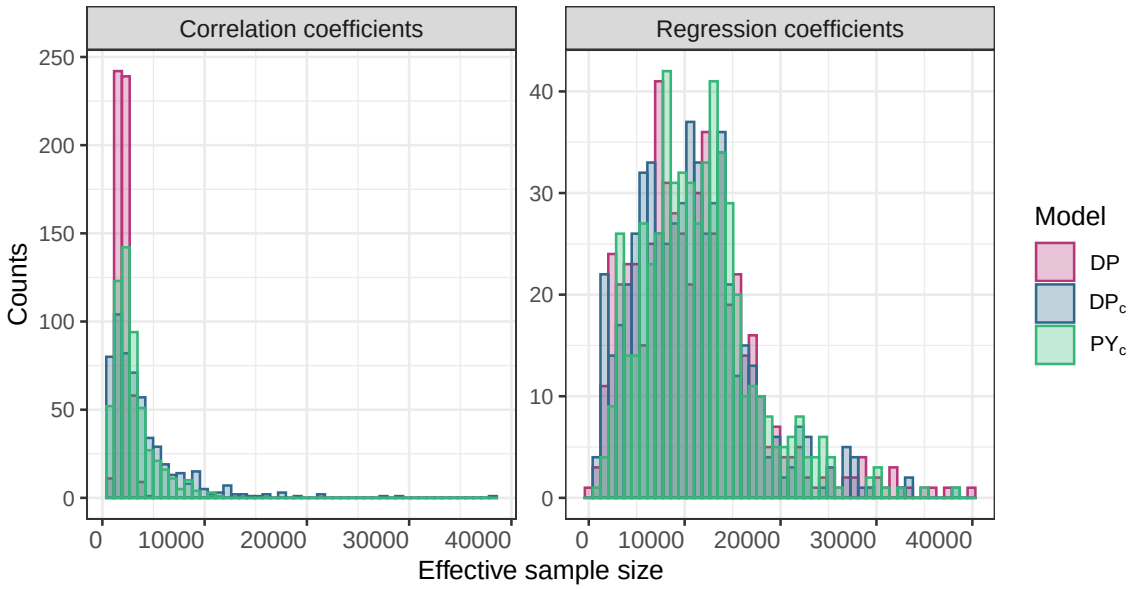

Figure S10. Effective sample size for the coefficients of covariance matrix $\boldsymbol{\Sigma}$ (left) and regression coefficients (elements of $\boldsymbol{B}$ matrix) (right) for all models (**DP**, **DP**$_c$, **PY**$_c$).

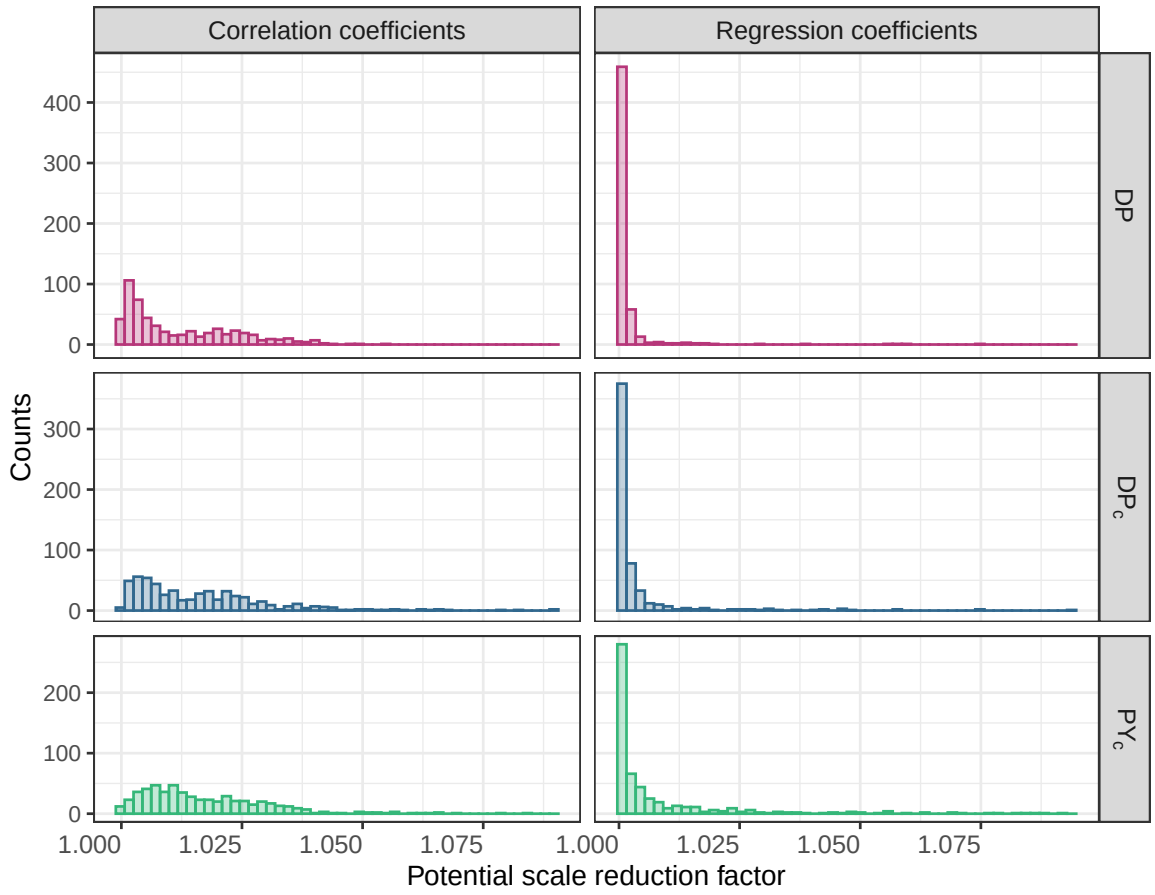

Figure S11. Potential scale reduction factor for coefficient of covariance matrix $\boldsymbol{\Sigma}$ (left) and regression coefficients (right) for all models (**DP**, **DP**$_c$, **PY**$_c$)).

LIST OF TABLES

Table S1. Hyperparameters used for model fitting in models **DP**, **DP**$_c$, **PY**$_c$ (Section S.5).

Parameters		**DP**	**DP**$_c$	**PY**$_c$
α	mean	112	6.23	0.47
	ν_1	-	1.93	-
	ν_2	-	0.31	-
σ		0	0	0.5
$\mathbb{E}[K_S]$		56.2	16	16

Table S2. Prediction performance and model fit for all the models. AUC_{out} corresponds to the model out-of-sample prediction. AUC_{in} is prediction on the train data.

Parameter	**DP**	**DP$_c$**	**PY$_c$**
AUC_{out}	0.745	0.747	0.746
AUC_{in}	0.756	0.758	0.755

Table S3. Parameters for $\mathbf{DP}_c$ and $\mathbf{PY}_c$ models, where expected number of clusters is defined as $\mathbb{E}[K_S]$ take values in $\{8, 56\}$ (Section S.8.3).

Models		$\mathbf{DP}_c$		$\mathbf{PY}_c$	
α	mean	2.7	109.5	0.64	7.7
	ν_1	0.34	600	-	-
	ν_2	0.13	5.4	-	-
σ		0	0	0.25	0.8
$E[K_S]$		8	56	8	56