HHS Public Access

Author manuscript

Biometrika. Author manuscript; available in PMC 2022 September 13.

Published in final edited form as:

Biometrika. 2022 September ; 109(3): 817–835. doi:10.1093/biomet/asab056.

Generalized infinite factorization models

L. Schiavon,

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy

A. Canale,

Department of Statistical Sciences, University of Padova, Via Cesare Battisti 241, 35121 Padova, Italy

D. B. Dunson

Department of Statistical Science, Duke University, Durham, North Carolina 27708, U.S.A.

Summary

Factorization models express a statistical object of interest in terms of a collection of simpler objects. For example, a matrix or tensor can be expressed as a sum of rank-one components. However, in practice, it can be challenging to infer the relative impact of the different components as well as the number of components. A popular idea is to include infinitely many components having impact decreasing with the component index. This article is motivated by two limitations of existing methods: (1) lack of careful consideration of the within component sparsity structure; and (2) no accommodation for grouped variables and other non-exchangeable structures. We propose a general class of infinite factorization models that address these limitations. Theoretical support is provided, practical gains are shown in simulation studies, and an ecology application focusing on modelling bird species occurrence is discussed.

Keywords

Adaptive Gibbs sampling; Bird species; Ecology; Factor analysis; High-dimensional data; Increasing shrinkage; Structured shrinkage

1. Introduction

Factorization models are used routinely to express matrices, tensors or other statistical objects based on simple components. The likelihood for data y under a general class of factorization models can be expressed as $L(y; \Lambda, \Psi, \Sigma)$, with $\Lambda = (\Lambda_h, h = 1, \ldots, k\}$ a $p \times k$ matrix, $\Lambda_h = (\lambda_{1h}, \ldots, \lambda_{ph})^{\mathrm{T}}$ the hth column vector of Λ, Ψ and Σ additional parameters, and k a positive integer. This class includes Gaussian linear factor models (Roweis & Ghahramani, 1999), exponential family factor models (Jun & Tao, 2013), Gaussian copula

lorenzo.schiavon@phd.unipd.it .

SUPPLEMENTARY MATERIAL

Supplementary material available at Biometrika online includes the statement and proof of Proposition S1 and the proofs of Proposition 1, Lemmas 1–2, and Corollaries 1–3. The Gibbs sampling algorithm, settings, and additional results of the simulations and ecology data analysis are reported, including trace plots and a sensitivity analysis to varying hyperparameters.

factor models (Murray et al., 2013), latent factor linear mixed models (An et al., 2013), probabilistic matrix factorization (Mnih & Salakhutdinov, 2008), underlying Gaussian factor models for mixed scale data (Reich & Bandyopadhyay, 2010), and functional data factor models (Montagna et al., 2012). A fundamental problem is how to choose weights for the components and the number of components k. This article proposes a general class of Bayesian methods to address this problem.

Although there is a rich literature, selection of k is far from a solved problem. In unsupervised settings, it is common to fit the model for different choices of k and then choose the value with the best goodness-of-fit criteria. For likelihood models, the Bayesian information criteria is particularly popular. It is also common to use an informal elbow rule, selecting the smallest k such that the criteria improves only a small amount for $k+1$. In specific contexts, formal model selection methods have been developed. For example, taking a Bayesian approach, one can choose a prior for k and attempt to approximate the posterior distribution of k using Markov chain Monte Carlo; see Lopes & West (2004) for linear factor models, Miller & Harrison (2018) for mixture models and Yang et al. (2018) for matrix factorization. Although such methods are conceptually appealing, computation can be prohibitive outside of specialized settings.

Due to these challenges it has become popular to rely on over-fitted factorization models, which include more than enough components, but with shrinkage priors adaptively removing unnecessary ones by shrinking their coefficients close to zero. Such approaches were proposed by Rousseau & Mengersen (2011) for mixture models and Bhattacharya & Dunson (2011) for Gaussian linear factor models. The latter approach specifically assumes an increasing shrinkage prior on the columns of the factor loadings matrix Λ. Legramanti et al. (2020) recently modified this approach using a spike and slab structure (Mitchell & Beauchamp, 1988) that increases the mass on the spike for later columns.

Although over-fitted factorizations are widely used, there are two key gaps in the literature. The first is a careful development of the shrinkage properties of increasing shrinkage priors (Durante, 2017). Outside of the factorization context and mostly motivated by high-dimensional regression, there is a rich literature recommending specific desirable properties for shrinkage priors. These include high concentration at zero to favor shrinkage of small coefficients and heavy tails to avoid over shrinking large coefficients. Motivated by this thinking, popular shrinkage priors have been developed including the Dirichlet-Laplace (Bhattacharya et al., 2015) and horseshoe (Carvalho et al., 2010). Current increasing shrinkage priors, such as those of Bhattacharya & Dunson (2011), were not designed to have the desirable shrinkage properties of these priors. For this reason, ad hoc truncation and use of the horseshoe/Dirichlet-Laplace can outperform increasing shrinkage priors in some contexts; for example, this was the case in Ferrari & Dunson (2020).

A second gap in the literature on over-fitted factorization priors is the lack of structured shrinkage. The focus has been on priors for Λ that are exchangeable within columns, with the level of shrinkage increasing with the column index. However, it is common in practice to have *meta covariates* encoding features of the rows of Λ. For example, the rows may correspond to different genes in genomic applications or species in ecology. There is a

rich literature on incorporating gene ontology in statistical analyses of genomic data; refer, for example to Thomas et al. (2009). In ecology it is common to include species traits in species distribution models (Ovaskainen & Abrego, 2020). Beyond the Bayesian literature, it is common to include structured penalties, with the grouped Lasso (Yuan & Lin, 2006) a notable example.

Motivated by these deficiencies of current factorizations priors, this article proposes a broad class of generalized infinite factorization priors, along with corresponding theory and algorithms for routine Bayesian implementation.

2. Generalized infinite factor models

2.1. Model specification

Suppose that an $n \times p$ data matrix y is available. In our motivating application, y_{ij} is a binary indicator of occurrence of bird species j ($j = 1, \ldots, p$) in sample i ($i = 1, \ldots, n$). Consider the following general class of models,

$$y_{ij} = t_j(z_{ij}), \quad z_i = \Lambda \eta_i + \epsilon_i, \quad \epsilon_i \sim f_\epsilon, \tag{1}$$

with Λ a $p \times k$ loadings matrix, η_i a k dimensional factor with diagonal covariance matrix $\Psi = \text{diag}(\psi_{11}, \ldots, \psi_{kk})$, ϵ_i a p-dimensional error term independent of η_i, and the function $t_j : \mathcal{R} \to \mathcal{R}$, for $j = 1 \ldots, p$. We refer to this class as generalized factorization models. Class (1) includes most of the cases mentioned in Section 1. When ϵ_i and η_i are Gaussian random vectors and t_j is the identity function, model (1) is a Gaussian linear factor model. With similar assumptions for ϵ_i and η_i, and assuming $t_j = F_j^{-1}\{\Phi(z_{ij})\}$, with $\Phi(z_{ij})$ the Gaussian cumulative distribution function, model (1) is a Gaussian copula factor model (Murray et al., 2013). Exponential family factor models (Jun & Tao, 2013), probabilistic matrix factorization (Mnih & Salakhutdinov, 2008) and underlying Gaussian models for mixed scale data (Reich & Bandyopadhyay, 2010) can be obtained by appropriately defining the elements in (1), whereas multivariate response regression models belong to this framework when η_i is known.

The matrix $\Omega = \text{var}(z_i)$ can be expressed as $\Omega = \Lambda \Psi \Lambda^T + \Sigma$, where $\Sigma = \text{var}(\epsilon_i)$. Following common practice in Bayesian factor analysis (Bhattacharya & Dunson, 2011), we avoid imposing identifiability constraints on Λ and assume Ψ is pre-specified. Our focus is on a new class of generalized infinite factor models induced through a novel class of priors for Λ that allows infinitely many factors, $k = \infty$. In particular, we let

$$\lambda_{jh} | \theta_{jh} \sim N(0, \theta_{jh}), \quad \theta_{jh} = \tau_0 \gamma_h \phi_{jh}, \quad \tau_0 \sim f_{\tau_0}, \quad \gamma_h \sim f_{\gamma_h}, \quad \phi_{jh} \sim f_{\phi_j}, \tag{2}$$

where f_{τ_0}, f_{γ_h}, and f_{ϕ_j} are supported on $[0, \infty)$ with positive probability mass on $(0, \infty)$. The local ϕ_{jh}, column-specific γ_h, and global τ_0 scales are all independent *a priori*. We let $N(0, 0)$ denote a degenerate distribution with all its mass at zero. Expression (2) induces a class of scale-mixture of Gaussian shrinkage priors (Polson & Scott, 2010) for the loadings. Although we allow infinitely many columns in Λ, (2) induces a prior for Ω supported on the

set of $p \times p$ positive semi-definite matrices under mild conditions reported in Proposition S1 in the Supplementary Material.

Differently from most of the existing literature on shrinkage priors, we want to define a non-exchangeable structure that includes *meta covariates* x informing the sparsity structure of Λ. In our context, meta covariates provide information to distinguish the p different variables as opposed to traditional covariates that serve to distinguish the n subjects. Letting x denote a $p \times q$ matrix of such meta covariates, we choose f_{ϕ_j} not depending on the index h and such that

$$E(\phi_{jh}|\beta_h) = g(x_j^T \beta_h), \quad \beta_h = (\beta_{1h}, \ldots, \beta_{qh})^{\mathrm{T}}, \quad \beta_{mh} \sim f_\beta \quad (m = 1, \ldots, q) \tag{3}$$

where $g: \mathscr{R} \to \mathscr{A} \subset \mathscr{R}_+$ is a known smooth one-to-one differentiable link function, $x_j = (x_{j1}, \ldots, x_{jq})^T$ denotes the jth row vector of x, and β_h are coefficients controlling the impact of the meta covariates on shrinkage of the factor loadings in the hth column of Λ.

To illustrate the usefulness of (3), consider the previously introduced ecological study and suppose $x_j = \{1, \mathbb{1}(\kappa_j = 2), \ldots, \mathbb{1}(\kappa_j = q)\}^T$, where $\kappa_j \in \{1, \ldots, q\}$ denotes the phylogenetic order of species j. Species of the same order may tend to have similarities that can be expressed in terms of a shared pattern of high or low loadings on the same latent factors. To illustrate this situation, we simulate a loadings matrix, displayed in Fig. 1, sampling from the prior introduced in Section 3 where $\mathrm{pr}(\lambda_{jh} = 0) > \mathrm{pr}(\phi_{jh} = 0) > 0$. The loadings within each column are penalized basing on the group structure identified by the $q = 3$ phylogenetic orders (Passeriformes, Charadriiformes, and Piciformes) of the $p = 10$ birds species considered. Our proposed prior allows for the possibility of such structure while not imposing it. In the bird ecology application, x can be defined to include not just phylogenetic placement of each bird species but also species traits, such as size or diet (Tikhonov et al., 2020). Related meta covariates are widely available, both in other ecology applications (Miller et al., 2019) and in other fields such as genomics (Thomas et al., 2009).

2.2. Properties

In this section we present some properties motivating the shrinkage process in (2) and provide insight into prior elicitation. It is important to relate the choice of hyperparameters to the signal-to-noise ratio, expressed as the proportion of variance explained by the factors. Section S2.4 of the Supplementary Material provides a study of the posterior distribution of the proportion of variance explained; the posterior tends to be robust to hyperparameter choice. Below we study key properties of our prior, including an increasing shrinkage property, the ability of the induced marginal prior to accommodate both sparse and large signals, and control of the multiplicity problem in sparse settings. Proofs are included in the Appendix and in Section S1 of Supplementary Material. This theory illuminates the role of hyperparameters; specific recommendations of hyperparameter choice in practice are illustrated under the model settings of Section 3.1.

To formalize the increasing shrinkage property, we introduce the following definition.

DEFINITION 1. *Letting* Π_Λ *denote a shrinkage prior on* Λ, Π_Λ *is a weakly increasing shrinkage prior if* $\text{var}(\lambda_{j(h-1)}) > \text{var}(\lambda_{jh})$ *for* j *in* $1, \ldots, p$ *and* $h = 2, \ldots, \infty$. Π_Λ *is a strongly increasing shrinkage prior if* $\text{var}(\lambda_{s(h-1)}) > \text{var}(\lambda_{jh})$, *for* j, s *in* $\{1, \ldots, p\}$ *and* $h = 2, \ldots, \infty$.

Weakly increasing shrinkage corresponds to the prior variance increasing across columns within each row of Λ, while strongly increasing shrinkage implies that the prior variance of any loading element is larger than all elements with a higher column index. In the following Theorem, we show that the process in (2) induces weakly increasing shrinkage under a simple sufficient condition.

THEOREM 1. *Expression* (2) *is a weakly increasing shrinkage prior under Definition* 1 *if* $E(\gamma_h) > E(\gamma_{h+1})$ *for any* h.

Increasing shrinkage priors favor a decreasing contribution of higher indexed columns of Λ to the covariance Ω. In addition to inducing a flexible shrinkage structure that allows different factors to have a different sparsity structure in their loadings, this allows one to accurately approximate the likelihood $L(y; \Lambda, \Psi, \Sigma)$ by $L(y; \Lambda_H, \Psi_H, \Sigma)$, with Λ_H containing the first H columns of the infinite matrix Λ and Ψ_H the first H rows and columns of Ψ. To measure the induced truncation error of $\Omega_H = \Lambda_H \Psi_H \Lambda_H^{\mathrm{T}} + \Sigma$, we use the trace of Ω. The trace is justified by the fact that the maximum error occurring in an element of Ω due to truncation always lies along the diagonal and by the relation between difference of traces and the nuclear norm, routinely used to approximate low rank minimization problems (Liu & Vandenberghe, 2010). The following Proposition provides conditions on prior (2) so that the under-estimation of Ω that occurs by truncating decreases exponentially fast as H increases.

PROPOSITION 1. *Let* $E(\tau_0)$ and $E(\phi_{jh})$ *be finite for* $j = 1, \ldots, p$ *and* $h = 1, \ldots, \infty$ *and* $E(\gamma_h) = ab^{h-1}$ *with* $a > 0$ *and* $b \in (0, 1)$ *for all* $h = 1, \ldots, \infty$. *Let* $c > 0$ *be a sufficiently large number such that* $c \geq \max_{h=1,\ldots,\infty} \psi_{hh}$. *If*

$$m_\Omega = \min_{j=1,\ldots,p} \left[E\left(\sigma_j^{-2}\right),\ E\left\{ \left(\sum_{h=1}^{\infty} \psi_{hh} \lambda_{jh}^2 \right)^{-1} \right\} \right] < \infty,$$

then for any $T \in (0, 1)$,

$$\text{pr}\left\{ \frac{\text{tr}(\Omega_H)}{\text{tr}(\Omega)} \leq T \right\} \leq \left(\frac{1}{1-T} \right) ac \frac{b^H}{1-b} m_\Omega E(\tau_0) \sum_{j=1}^{p} E\left(\phi_{j1}\right).$$

The above increasing shrinkage properties can be satisfied by naive priors that over-shrink the elements of Λ. It is important to avoid such over-shrinkage and allow not only many elements that are ≈ 0 but also a small proportion of large coefficients. A similar motivation applies in the literature on shrinkage priors in regression (Carvalho et al., 2010). Borrowing from that literature, the marginal prior for λ_{jh} should be concentrated at zero to reduce

mean square error by shrinking small coefficients to zero but with heavy tails to avoid over-shrinking the signal.

To quantify the prior concentration of (2) in an ϵ neighbourhood of zero, we can obtain

$$\mathrm{pr}(|\lambda_{jh}| > \epsilon) \leq \frac{E(\tau_0)E(\gamma_h)E(\phi_{jh})}{\epsilon^2} \tag{4}$$

as a consequence of Markov's inequality. Common practice in local-global shrinkage priors chooses $E(\tau_0)$ small while assigning a heavy-tailed density to the local or column scales. In our case, (3) allows the bound in (4) to be regulated by meta covariates x, while, under the condition in Theorem 1, decreasing $E(\gamma_h)$ with column index causes an increasing concentration near zero, since $E(\phi_{jh}) = E(\phi_{jl})$ for every $h, l \in \{1, \ldots, \infty\}$. The means of the column and the local scales control prior concentration near zero, while over-shrinkage can be ameliorated by choosing f_{ϕ_j} or f_{γ_h} $(h = 1, \ldots, \infty)$ heavy tailed. The following Proposition provides a condition on the prior to guarantee a heavy tailed marginal distribution for λ_{jh}. A random variable has power law tails if its cumulative distribution function F has $1 - F(t) \geq ct^{-a}$ for constants $c > 0$, $a > 0$, and for any $t > L$ for L sufficiently large.

Proposition 2. *If at least one scale parameter among τ_0, γ_h or ϕ_{jh} is characterized by a power law tail prior distribution, then the prior marginal distribution of λ_{jh} has power law tails.*

An important consequence of the heavy tailed property is avoidance of over-shrinkage of large signals. This is often formalized via a tail robustness property (Carvalho et al., 2010). As an initial result, key to showing sufficient conditions for a type of local tail robustness, we provide the following Lemma on the derivative of the log prior in the limit as the value of $\lambda_{jh} \to \infty$.

Lemma 1. *If at least one scale parameter among τ_0, γ_h or ϕ_{jh} has a prior with power law tails for any possible prior distribution of β_h, then for any finite truncation level H,*

$$\lim_{\lambda \to \infty} \frac{\partial \log\left\{f_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)\right\}}{\partial \lambda} = 0$$

where $f_{\lambda_{jh}|\Lambda_{-jh}}(\lambda)$ is the conditional distribution of λ_{jh} given the other elements of Λ_H.

The following definition introduces a type of local tail robustness property.

Definition 2. *Consider model* (1) *with factors η known. Let $f_{\lambda_{jh}|y,\eta,\Lambda_{-jh}(\lambda)}$ denote the posterior density of λ_{jh}, given the data, conditional on any possible value of the other elements of Λ_H for any finite H, and let $\hat{\lambda}_{jh}$ denote the conditional maximum likelihood estimate of λ_{jh} for any possible value of the other elements of Λ_H. We say that the prior on λ_{jh} is tail robust if*

$$\lim_{\hat{\lambda}_{jh} \to \infty} |\hat{\lambda}_{jh} - \arg\max_{\lambda} f_{\lambda_{jh}|y,\eta,\Lambda_{-jh}}(\lambda)| = 0.$$

For a given sample, $\hat{\lambda}_{jh}$ is a fixed quantity; the above limit should be interpreted as what happens as the data support a larger and larger maximum likelihood estimate. In order for tail robustness to hold, we need the data to be sufficiently informative about the parameter λ_{jh} and the likelihood to be sufficiently regular; this is formalized as follows.

Assumption 1. Let $L(y; \Lambda, \eta, \Sigma)$ denote the likelihood for data y conditionally on latent variables η, let $l_s(\lambda)$ denote the derivative function of the log-likelihood with respect to λ_{jh}, and let $\mathcal{J}(\hat{\lambda}_{jh})$ denote the negative of the second derivative of the log-likelihood with respect to λ_{jh}, evaluated at the conditional maximum likelihood estimate $\hat{\lambda}_{jh}$. Then $l_s(\lambda)$ is a continuous function for every $\lambda \in \mathcal{R}$ and $\mathcal{J}(\hat{\lambda}_{jh}) \geq v(\hat{\lambda}_{jh})$, where $v(\hat{\lambda}_{jh})$ is of order $O(1)$ as $\hat{\lambda}_{jh} \to \infty$.

This assumption can be verified for most of the cases mentioned in Section 1; for example, for Gaussian linear factor models $\mathcal{J}(\hat{\lambda}_{jh})$ is of order $O(1)$ with respect to $\hat{\lambda}_{jh}$.

THEOREM 2. *Under Assumption* 1, *if at least one scale parameter among* τ_0, γ_h *or* ϕ_{jh} *is power law tail distributed for any possible prior distribution of* β_h, *then the prior on* λ_{jh} *is tail robust under Definition* 2.

As an additional desirable property, we would like to control for the multiplicity problem within each column λ_h of the loadings matrix, corresponding to increasing numbers of false signals as dimension p increases. This can be accomplished by imposing an asymptotically increasingly sparse property on the prior, which is defined as follows.

DEFINITION 3. *Let* $|\text{supp}_\epsilon(\lambda_h)|$ *denote the cardinality of* $\text{supp}_\epsilon(\lambda_h) = (j : |\lambda_{jh}| > \epsilon)$. *Let* $s_p = o(p)$ *such that* $s_p \geq c_s \log(p)/p$ *for some constant* $c_s > 0$. *We say that the prior on* Λ *defined in* (2) *is an asymptotically increasingly sparse prior if*

$$\lim_{p \to \infty} \text{pr}\{|\text{supp}_\epsilon(\lambda_h)| > a\, s_p | \gamma_h, \tau_0\} = 0, \quad \text{for some constant } a > 0.$$

The quantity $|\text{supp}_\epsilon(\lambda_h)|$ represents an approximate measure of model size for continuous shrinkage priors and, conditionally on β_h, γ_h, and τ_0, it is *a priori* distributed as a sum of independent Bernoulli random variables $\text{Ber}(\zeta_{\epsilon jh})$, where

$$\zeta_{\epsilon jh} = \text{pr}(|\lambda_{jh}| > \epsilon | \beta_h, \gamma_h, \tau_0) \leq \frac{\tau_0 \gamma_h g(x_j^{\mathrm{T}} \beta_h)}{\epsilon^2}.$$

We now provide sufficient conditions for an asymptotically increasingly sparse prior, allowing regulation of the sparsity behaviour of the prior of the columns of Λ for increasing dimension p.

THEOREM 3. *Consider prior* (2) *with* ϕ_{jh} $(j = 1, \ldots, p)$ *a priori independent given* β_h. *If* $\text{pr}\{g(x_j^{\mathrm{T}} \beta_h) \leq v_j(p)\} = 1$, *with* $v_j(p) = O\{\log(p)/p\}$, $(j = 1, \ldots, p)$, *then the prior on* Λ *is asymptotically increasingly sparse under Definition* 3.

The condition of the theorem is easily satisfied, for example, if g is the multiplication of a bounded function and a suitable offset depending on p as assumed in Section 3.1. The multiplicative gamma process (Bhattacharya & Dunson, 2011) and cumulative shrinkage process (Legramanti et al., 2020) do not satisfy the sufficient conditions of Theorem 3, and, furthermore, the following lemma holds.

LEMMA 2. *The multiplicative gamma process prior (*Bhattacharya & Dunson, 2011*) and the cumulative shrinkage process prior (*Legramanti et al., 2020*) are not asymptotically increasing sparse under Definition* 3.

Although this Section has focused on properties of the prior, we find empirically that these properties tend to carry over to the posterior, as will be illustrated in the subsequent sections. For example, the posterior exhibits asymptotic increasing sparsity; refer to Table 2 of Section 4, which shows results for a novel process in our proposed class that is much more effective than current approaches in identifying the true sparsity structure, particularly when p is large.

3. STRUCTURED INCREASING SHRINKAGE PROCESS

3.1. Model specification

In this section we propose a structured increasing shrinkage process prior for generalized infinite factor models satisfying all the sufficient conditions in Propositions 1–2 and Theorems 2–3. Let Ga(a, b) denote the gamma distribution with mean a/b and variance a/b^2. Following the notation of Section 2.1, we specify

$$\tau_0 = 1,\ \gamma_h = \vartheta_h \rho_h,\ \phi_{jh} | \beta_h \sim \text{Ber}\left\{\text{logit}^{-1}\left(x_j^{\mathrm{T}} \beta_h\right) c_p\right\}, \tag{5}$$

$$\vartheta_h^{-1} \sim \text{Ga}(a_\theta, b_\theta),\ a_\theta > 1,\ \rho_h = \text{Ber}(1 - \pi_h),\ \beta_h \sim N_q\left(0, \sigma_\beta^2 I_q\right),$$

where we assume the link $g(x) = \text{logit}^{-1}(x) c_p$, with $\text{logit}^{-1}(x) = e^x/(1 + e^x)$ and $c_p \in (0, 1)$ a possible offset. The parameter $\pi_h = \text{pr}(\gamma_h = 0)$ follows a stick-breaking construction,

$$\pi_h = \sum_{l=1}^{h} w_l, \quad w_l = v_l \prod_{m=1}^{l-1} (1 - v_m), \quad v_m \sim \text{Be}(1, \alpha),$$

with Be(a, b) the beta distribution with mean $a/(a + b)$, such that $\pi_{h+1} > \pi_h$ is guaranteed for any $h = 1, \ldots, \infty$ and $\lim_{h \to \infty} \pi_h = 1$ almost surely. The prior expected number of non degenerate Λ columns is $E\left(\sum_{h=1}^{\infty} \rho_h\right) = \alpha$ (Legramanti et al., 2020), suggesting setting α equal to the expected number of active factors. The prior specification is completed assuming $\Sigma = \text{diag}\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ with $\sigma_j^{-2} \sim \text{Ga}(a_\sigma, b_\sigma)$ for $j = 1, \ldots, p$, consistently with the literature. The hyperparameters can be chosen based on one's prior expectation of the signal-to-noise ratio, as σ_j^2 is the contribution of the noise component to the total variance of the jth variable. A sensitivity study in Section S2.4 of the Supplementary Material, however,

shows that posterior distributions tend to be robust to the specification of a_σ, b_σ. Regarding prior elicitation, we recommend setting $b_\theta \geq a_\theta$ to induce a high enough proportion of variance explained by the factor model. In Section S2.4 in the Supplementary Materials we report empirical evidence of the effect of different prior parameters on this quantity.

The above specification respects (2) and, consequently, the following corollary holds.

COROLLARY 1. *The structured increasing shrinkage process defined in (5)*

- **i.** *is a strongly increasing shrinkage prior under Definition* 1*;*
- **ii.** *for any* $T \in (0, 1)$,

$$\text{pr}\left\{\frac{\text{tr}(\Omega_H)}{\text{tr}(\Omega)} \leq T\right\} \leq \left(\frac{1}{1-T}\right)\frac{b^H}{1-b}\theta_0\frac{a_\sigma}{b_\sigma}\sum_{j=1}^{p} E(\phi_{j1}),$$

with $b = \{\alpha(1+\alpha)\}^{-1}$ *and* $\theta_0 = E(\vartheta_h)$.

We conducted a simulation study on the posterior distribution of $\{\text{tr}(\Omega_H)/\text{tr}(\Omega) \leq T\}$ for varying hyperparameters, and found that the results, reported in Section S2.4 of the Supplementary Material, were quite consistent with our prior truncation error bounds.

The prior concentration of the structured increasing shrinkage process in (5) follows from (4):

$$\text{pr}(|\lambda_{jh}| > \epsilon) \leq \frac{E(\vartheta_h)\{1 - E(\pi_h)E(\phi_{jh})\}}{\epsilon^2} = \frac{\theta_0\{\alpha/(1+\alpha)\}^h}{\epsilon^2}\frac{c_p}{2}.$$

In addition, the inverse gamma prior on ϑ_h implies a power law tail distribution on γ_h inducing robustness properties on λ_{jh} as formalized by the next corollary of Proposition 2 and Theorem 2.

COROLLARY 2. *Under the structured increasing shrinkage process defined in (5)*

- **i.** *the marginal prior distribution on* λ_{jh} *(*$j = 1, \ldots, p$*;* $h = 1, 2, \ldots$*) has power law tails;*
- **ii.** *under Assumption* 1*, the prior on* λ_{jh} *(*$j = 1, \ldots, p$*;* $h = 1, 2, \ldots$*) is tail robust under Definition* 2.

Finally, it is important to assess the joint sparsity properties of the prior on each column of Λ. This is formalized in the following corollary of Theorem 3.

COROLLARY 3. *If* $c_p = O\{\log(p)/p\}$ *the structured increasing shrinkage process defined in (5) is asymptotically increasingly sparse under Definition* 3.

3.2. Posterior computations

Posterior inference is conducted via Markov chain Monte Carlo sampling. Following common practice in infinite factor models (Bhattacharya & Dunson, 2011; Legramanti et

al., 2020; Schiavon & Canale, 2020) we use an adaptive Gibbs algorithm, which attempts to infer the best truncation level H while drawing from the posterior distribution of the parameters. The value of H is adapted only at some Gibbs iterations by discarding redundant factors and, if no redundant factors are identified, by adding a new factor by sampling its parameters from the prior distribution. Convergence of the Markov chain is guaranteed by satisfying the diminishing adaptation condition in Theorem 5 of Roberts & Rosenthal (2007), by specifying the probability of occurrence of an adaptive iteration t as equal to $p(t) = \exp(\alpha_0 + \alpha_1 t)$, where α_0 and α_1 are negative constants, such that frequency of adaptation decreases.

The decomposition of γ_h into two parameters ρ_h and ϑ_h allows one to identify the inactive columns of Λ, corresponding to the redundant factors, as those with $\rho_h = 0$, while H_a indicates the number of active columns of Λ. Consequently, at the adaptive iteration $t + 1$, the truncation level H is set to $H^{(t+1)} = H_a^{(t)} + 1$ if $H_a^{(t)} < H^{(t)} - 1$, and $H^{(t+1)} = H^{(t)} + 1$ otherwise. Given $H^{(t+1)}$, the number of factors of the truncated model at iteration $t + 1$, the sampler draws the model parameters from the corresponding posterior full conditional distributions. The detailed steps of the adaptive Gibbs sampler for the structured increasing shrinkage prior in case of Gaussian or binary data are reported in the Supplementary Material, as well as trace plots of the posterior samples for some parameters of the model in Section 5 (see Section S3.2), showing good mixing.

3.3. Identifiability and posterior summaries

Non-identifiability of the latent structure creates problems in interpretation of the results from Markov chain Monte Carlo samples. Indeed, both Λ and η are only identifiable up to an arbitrary rotation P with $PP^T = I_k$. This is a well known problem in Bayesian factor models and there is a rich literature proposing post-processing algorithms that align posterior samples $\Lambda^{(t)}$, so that one can then obtain interpretable posterior summaries. Refer to McParland et al. (2014), Aßmann et al. (2016), and Roy et al. (2019) for alternative post-processing algorithms in related contexts.

Unfortunately, such post-hoc alignment algorithms destroy the structure we have carefully imposed on the loadings in terms of sparsity and dependence on meta covariates. Therefore, we propose a different solution to obtain a point estimate of Λ based on finding a representative Monte Carlo draw $\Lambda^{(t)}$ consistently with the proposals of Dahl (2006) and Wade et al. (2018) in the context of Bayesian model-based clustering. Specifically, we summarize Λ and $\beta = (\beta_1, \beta_2, \ldots)$ through $\Lambda^{(t^*)}$ and $\beta^{(t^*)}$ sampled at iteration t^*, characterized by the highest marginal posterior density function $f(\Lambda, \beta, \Sigma \mid y)$ obtained by integrating out the scale parameters τ_0, γ_h, ϕ_{jh} ($j = 1, \ldots, p$, $h = 1, \ldots$) and the latent factors η_i ($i = 1, \ldots, n$) from the posterior density function. Formally, we select the iteration $t^* \in \{1, \ldots, T\}$ such that

$$f\left(\Lambda^{(t^*)}, \beta^{(t^*)}, \Sigma^{(t^*)} | y\right) > f\left(\Lambda^{(t)}, \beta^{(t)}, \Sigma^{(t)} | y\right) \quad (t = 1, \ldots, T),$$

where $t = 1, \ldots, T$ indexes the posterior samples. Under the structured increasing shrinkage prior described in Section 3.1, these computations are straightforward. The matrices $\Lambda^{(t^*)}$, $\beta^{(t^*)}$, $\Sigma^{(t^*)}$ are Monte Carlo approximations of the maximum *a posteriori* estimator, which corresponds to the Bayes estimator under L_∞ loss. Although one can argue that L_∞ is not an ideal choice of loss philosophically in continuous parameter problems, it nonetheless is an appealing pragmatic choice in our context and is broadly used in other sparse estimation contexts, as in the algorithm proposed by Ročková & George (2016) that similarly aims to recover a strongly sparse posterior mode of an over-parameterized factor model.

4. Simulation experiments

We assess the performance of our structured increasing shrinkage prior compared with current approaches (Bhattacharya & Dunson, 2011; Ročková & George, 2016; Legramanti et al., 2020) through a simulation study. We have a particular interest in inferring sparse and interpretable loadings matrices Λ, but also assess performance in estimating the induced covariance matrix Ω and number of factors. We generate synthetic data from four scenarios based on different loadings structures. For each scenario we simulate $R = 25$ data sets with $n = 250$ observations from $y_i \sim N_p\left(0, \Lambda_0\Lambda_0^T + I_p\right)$ $(i = 1, \ldots, n)$. In Scenario a, we assume non sparse Λ_0, sampling the loadings λ_{jh} from a Gaussian distribution with mean zero, variance equal to $\sigma_\lambda^2 = 1$ and ordering them to obtain decreasing variance over the columns. To ensure that each element λ_{jh} represents a signal, we shifted them away from zero by $\sigma_\lambda^2/3$. In Scenario b we remove the decreasing behaviour and introduce a random sparsity pattern characterized by an increasing number of zero entries over the column index. The loadings matrix for Scenario c is characterized by both the decreasing behaviour over the columns of Scenario a and the random sparsity structure of Scenario b. Finally, in Scenario d, while the decreasing behaviour is kept, we induce a sparsity pattern dependent on a categorical and two continuous meta covariates x_0. Details are reported in Section S2.2 of the Supplementary Material.

For each scenario we consider four combinations of dimension and sparsity level of Λ_0. We let $(p, k, s) \in \{(16, 4, 0.6), (32, 8, 0.4), (64, 12, 0.3), (128, 16, 0.2)\}$, where s is the proportion of non-zero entries of Λ, with the exception of Scenario a where $s = 1$. In these settings the algorithm takes from 0.07 to 0.73 seconds of computational time per iteration depending on the dimension p and considering an R implementation on an Intel Core i5-6200U CPU laptop with 15.8 GB of RAM. To estimate the structured increasing shrinkage model, we set x equal to the p-variate column vector of 1s, $\sigma_\beta = 1$ and, consistently with Corollary 3, $c_p = 2e\log(p)/p$. In Scenario d we also estimate and compare a correctly specified structured increasing model with $x = x_0$. For the method proposed by Ročková & George (2016), we set the hyperparameters as suggested by the authors, while for the remaining approaches, we follow the hyperparameter specification and factor selection guidelines in Section 4 of Schiavon & Canale (2020).

Scenario a is a worst case for the proposed method since there is no sparsity, no structure, and the elements of the loadings matrix are similar in magnitude. However, even in this case, structured increasing shrinkage performs essentially identically to the best competitor,

as illustrated by the results in Table 1. The results of Ročková & George (2016) are not reported as they are not competitive, as can be seen in table S2 in the Supplementary Material. We report the median and interquartile range over the R replicates of the logarithm of the pseudo-marginal likelihood (Gelfand & Dey, 1994) and of the estimated posterior mean of the number of factors $E(H_a \mid y)$.

Scenario b judges performance in detecting sparsity. The proposed approach shows better performance in the logarithm of the pseudo-marginal likelihood and mean squared error of the covariance matrix, particularly as sparsity increases, as displayed in Fig. 2. Consistently with (Legramanti et al., 2020), the covariance mean squared error is estimated in each simulation by $\sum_{j,l}^{p}\sum_{t=1}^{S}\left(\omega_{jl}^{(t)}-\omega_{jl0}\right)^2/\{p(p+1)/2\}$, where ω_{jl0} and $\omega_{jl}^{(t)}$ are the elements jl of $\Omega_0=\Lambda_0\Lambda_0^{\mathrm{T}}+I_p$ and $\Omega^{(t)}=\Lambda^{(t)}\Lambda^{(t)\mathrm{T}}+I_p$, respectively. The proposed approach allows exact zeros in the loadings, while the competitors require thresholding to infer sparsity. Following the thresholding approach described in Section S2.2 of the Supplementary Material, we evaluate performance in inferring the sparsity pattern via the mean classification error:

$$MCE=\frac{1}{S}\sum_{t=1}^{S}\frac{\sum_{j=1}^{p}\sum_{h=1}^{k^{*(t)}}\left|\mathbb{1}\left(\lambda_{jh0}=0\right)-\mathbb{1}\left(\lambda_{jh}^{(t)}=0\right)\right|}{pk},$$

where $k^{*(t)}$ is the maximum between the true number of factors k and $H_a^{(t)}$, and λ_{jh0} and $\lambda_{jh}^{(t)}$ are the elements jh of Λ_0 and $\Lambda^{(t)}$, respectively. If $H_a^{(t)}$ or k are smaller than k^*, we fix the higher indexed columns at zero, possibly leading to a mean classification error bigger than one. The results reported in Table 2 show that the proposed structured increasing shrinkage prior is much more effective in identifying sparsity in Λ, maintaining good performance even with large p and in strongly sparse contexts. Also, more accurate estimation of the number of factors is obtained, as reported in Table S1 in the Supplementary Material.

Similar comments apply in Scenarios c and d reported in Fig. S2 in the Supplementary Material. The superior performance of the structured increasing shrinkage model is only partially mitigated in Scenario c for large p for the logarithm of the pseudo-marginal likelihood. In Scenario d, the use of meta covariates has a mild benefit in identifying the sparsity pattern. In lower signal-to-noise settings, meta covariates have a bigger impact, and they also aid interpretation, as illustrated in the next section. Additional details, tables, and plots for all scenarios are reported in Section S2.3 of the Supplementary Material.

5. Finnish bird co-occurrence application

We illustrate our approach by modelling co-occurrence of the fifty most common bird species in Finland (Lindström et al., 2015), focusing on data in 2014. Response y is an $n \times p$ binary matrix denoting occurrence of $p = 50$ species in $n = 137$ sampling areas. An $n \times c$ environmental covariate matrix w is available, including a 5-level habitat type, ‘spring temperature’ (mean temperature in April and May), and (spring temperature)^2, leading to c

= 7. We consider a meta covariate $p \times q$ matrix x of species traits: logarithm of typical body mass, migratory strategy (short-distance migrant, resident species, long-distance migrant), and a 7-level superfamily index. We model species presence or absence via a multivariate probit regression model:

$$y_{ij} = \mathbb{1}(z_{ij} > 0), \quad z_{ij} = w_i^T \mu_j + \epsilon_{ij}, \quad \epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^T \sim N_p(0, \Lambda\Lambda^T + I_p), \tag{6}$$

where μ_j characterizes impact of environmental covariates on species occurrence probabilities, and covariance in the latent z_i vector is characterized through a factor model. To borrow information across species while incorporating species traits, we let

$$\mu_j \sim N_c(bx_j, \sigma_\mu^2 I_c), \quad b = (b_1, \ldots, b_q), \quad b_m \sim N_c(0, \sigma_b^2 I_c), \tag{7}$$

where b is a $c \times q$ coefficient matrix with column vectors b_m given Gaussian priors.

Model (6)–(7) is consistent with popular joint species distribution models (Ovaskainen et al., 2016; Tikhonov et al., 2017; Ovaskainen & Abrego, 2020), with current standard practice using a multiplicative gamma process for Λ. We compare this approach to an analysis that instead uses our proposed structured increasing shrinkage prior to allow the species traits x to impact Λ and hence the covariance structure across species. After standardizing w and x, we set $a = 4$, $a_\theta = b_\theta = 2$ and $\sigma_\mu = \sigma_b = 1$. Posterior sampling is straightforward via a Gibbs sampler reported in Section S3.1 of the Supplementary Material.

Figure S8 in the Supplementary Material displays the posterior means of μ and b. A first investigation shows large heterogeneity of the habitat type effects across different species. Matrix b shows that covariate effects tend to not depend on migratory strategy or body mass, with the exception of urban habitats tending to have more migratory birds.

The estimated Λ and meta covariate coefficients β, following the guidelines of Section 3.3, are displayed in Fig. 3. The loadings matrix is quite sparse, indicating that each latent factor impacts a small group of species. Positive sign of the loadings means that high levels of the corresponding factors increase the probability of observing birds from those species. Lower elements of $\beta^{(t^*)}$, represented with light cells on the right panel, induce higher shrinkage on the corresponding group of birds. To facilitate interpretation, we re-arrange the rows of $\Lambda^{(t^*)}$ according to the most relevant species traits in terms of shrinkage, which are migration strategy and body mass. The species influenced by the first factor are fairly homogeneous, characterized by short distance or resident migratory strategies and a larger body mass. The strongly negative value of $\beta_{(t^*)42}$ suggests heavier species of birds tend to have loadings close to zero for the second factor. This is also true for the third factor, which also does not impact short-distance migrants.

Figure S9 in the Supplementary Material shows a spatial map of the sampling units coloured accordingly to the values of the first and the third latent factors. We can interpret these latent factors as unobserved environmental covariates. We find that the species traits included in our analysis only partially explain the loadings structure; this is as expected and provides motivation for the proposed approach. Sparsity in the loadings matrix helps in interpretation.

Species may load on the same factor not just because they have similar traits but also because they tend to favor similar habitats for reasons not captured by the measured traits.

The induced covariance matrix $\Omega = \Lambda\Lambda^T + I_p$ across species is of particular interest. We compare estimates of Ω under the multiplicative gamma process, estimated using the R package hmsc (Tikhonov et al., 2020), and our proposed structured increasing shrinkage model. Figure 4 reports the posterior mean of the correlation matrices under the two competing models. The network graph based on the posterior mean of the partial correlation matrices, reported in Fig. 5, reveals several communities of species under the proposed structured increasing shrinkage prior that are not evident under the multiplicative gamma.

We also find that the multiplicative gamma process provides a slightly worse fit to the data. The logarithm of the pseudo marginal likelihood computed on the posterior samples of the structured increasing shrinkage model is equal to −21.06, higher than that achieved by the competing model, which is −21.36. Using 4-fold cross-validation, we compared the log-likelihood evaluated in the held-out data, with μ and Ω estimated by the posterior mean in the training set. The mean of the log-likelihood was −22.62 under the structured increasing shrinkage and −23.22 under the multiplicative gamma process prior.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement

The authors thank Daniele Durante, Sirio Legramanti, Otso Ovaskainen, and Gleb Tikhonov for useful comments on an early version of this manuscript. This project has received funding from the University of Padova under the STARS Grants programme, the United States National Institutes of Health under grant R01ES027498 and the European Research Council under the European Unions Horizon 2020 research and innovation programme (grant agreement No 856506).

Appendix: Lemmas and proofs

Proof of Theorem 1. The variance of λ_{jh} is

$$\mathrm{var}(\lambda_{jh}) = E\left\{E\left(\lambda_{jh}^2|\phi_{jh}, \gamma_h, \tau_0\right)\right\} = E\left\{E\left(\theta_{jh}|\phi_{jh}, \gamma_h, \tau_0\right)\right\}.$$

By construction, $E\left(\theta_{jh}|\phi_{jh}, \gamma_h, \tau_0\right) = \phi_{jh}\gamma_h\tau_0$. Then,

$$\mathrm{var}(\lambda_{jh}) = E(\phi_{jh}\gamma_h\tau_0) = E(\phi_{j1})E(\gamma_h)E(\tau_0) > E(\phi_{j1})E(\gamma_{h+1})E(\tau_0) = \mathrm{var}(\lambda_{jh+1}),$$

since the scale parameters are independent and the local scale ϕ_{jh} is equally distributed over the column index h. □

To prove Proposition 2 we need to introduce the following Lemma.

Lemma 3. *Let u, v denote two real positive random variables. If at least one among $(u \mid v)$ and $(v \mid u)$ is power law tail distributed, then the product uv is power law tail distributed.*

Proof. For a positive value w, we can write

$$\mathrm{pr}(uv > w) = \int_0^\infty \mathrm{pr}(u > w/v|v) f(v)\mathrm{d}v = E\left\{F^C_{u|v}(w/v)\right\},$$

where $F^C_{u|v}(w) = \mathrm{pr}(u > w|v)$ and $f(v)$ is the probability density function of v. If $F^C_{u|v}(w) \geq cw^{-\alpha}$ with c, α positive constants and w greater than a sufficiently large number L, then

$$\mathrm{pr}(uv > w) \geq E\left\{c(w/v)^{-\alpha}\right\} = cw^{-\alpha}E(v^\alpha) \quad w > L \gg 0\,.$$

If $E(v^\alpha) = \infty$, then $\mathrm{pr}(uv > w) > cw^{-\alpha} = O(w^{-\alpha})$, otherwise $\mathrm{pr}(uv > w) \geq v(w)$ for $w > L$, with $v(w)$ a function of order $O(w^{-\alpha})$ as w goes to infinity. This shows that the right tail of the distribution of the random variable uv follows a power law behaviour. □

Proof of Proposition 2. Consider the strictly positive random variables $\theta^*_{jh} = (\theta_{jh}|\theta_{jh} > 0)$, $\tau^*_0 = (\tau_0|\tau_0 > 0)$, $\gamma^*_h = (\gamma_h|\gamma_h > 0)$, and $\phi^*_{jh} = (\phi_{jh}|\phi_{jh} > 0)$. Since θ^*_{jh} is equal to the product $\tau^*\gamma^*_h\phi^*_{jh}$ of independent positive random variables, Lemma 3 ensures that if at least one of those scale parameters follows a power law tail distribution, then θ^*_{jh} is power law tail distributed, so that $\mathrm{pr}(\theta^*_{jh} > \theta) \geq c\theta^{-\alpha}$ for c, α positive constants and $\theta > L$. Without loss of generality, we focus on the right tail of λ_{jh}. Let

$$\mathrm{pr}(\lambda_{jh} > \lambda) = \mathrm{pr}(\lambda_{jh} > \lambda|\theta_{jh} > 0)\mathrm{pr}(\theta_{jh} > 0) + \mathrm{pr}(\lambda_{jh} > \lambda|\theta_{jh} = 0)\mathrm{pr}(\theta_{jh} = 0)\,. \quad (1)$$

It is straightforward to observe that λ_{jh} marginally has a power law tail if and only if $(\lambda_{jh}\,|\,\theta_{jh} > 0)$ is power law tail distributed and $\mathrm{pr}(\theta_{jh} > 0)$ is strictly positive. Since $\mathrm{pr}(\tau_0 > 0) > 0$, $\mathrm{pr}(\gamma_h > 0) > 0$, and $\mathrm{pr}(\phi_{jh} > 0) > 0$, then $\mathrm{pr}(\theta_{jh} > 0) > 0$, given independence between the scale parameters. Focusing on $\theta_{jh} > 0$ in the first term of the right hand side of (1), we have

$$\mathrm{pr}(\lambda_{jh} > \lambda|\theta^*_{jh}) = 1 - \Phi(\lambda\theta^{*\,-0.5}_{jh}),$$

and we want to prove that the marginal $F^c_{\lambda_{jh}}(\lambda) = \mathrm{pr}(\lambda_{jh} > \lambda)$ is sub-exponential as $\lambda \to \infty$. Using the lower bound for the right tail of the standard Gaussian of Abramowitz & Stegun (1948),

$$1 - \Phi(\lambda\theta^{*\,-0.5}_{jh}) \geq \left(\frac{2}{\pi}\right)^{0.5} \frac{\theta^{*0.5}_{jh}}{\lambda + \left(\lambda^2 + 4\theta^*_{jh}\right)^{0.5}} e^{-\lambda^2/\left(2\theta^*_{jh}\right)}.$$

Marginalizing over θ^*_{jh}, we obtain

$$\mathrm{pr}\left(\lambda_{jh} > \lambda | \theta^*_{jh}\right) \geq E\left\{\left(\frac{2}{\pi}\right)^{0.5} \frac{\theta^{*0.5}_{jh}}{\lambda + \left(\lambda^2 + 4\theta^*_{jh}\right)^{0.5}} e^{-\lambda^2/\left(2\theta^*_{jh}\right)}\right\} = E\left\{t_\lambda\left(\theta^*_{jh}\right)\right\},$$

where $t_\lambda\left(\theta^*_{jh}\right)$ is a monotonically increasing nonnegative function defined on the positive real line. Applying Markov's inequality, we have $E\left\{t_\lambda\left(\theta^*_{jh}\right)\right\} > \mathrm{pr}\left(\theta^*_{jh} > \epsilon\right) t_\lambda(\epsilon)$, and letting $\epsilon = \lambda^2$

$$E\left\{t_\lambda\left(\theta^*_{jh}\right)\right\} > \mathrm{pr}\left(\theta^*_{jh} > \lambda^2\right) \frac{e^{-0.5}}{1 + 5^{0.5}} \left(\frac{2}{\pi}\right)^{0.5}.$$

If $\mathrm{pr}\left(\theta^*_{jh} > \lambda\right) \geq c\lambda^{-\alpha}$ for certain α, c positive constants and λ sufficiently large, then

$$\mathrm{pr}\left(\lambda_{jh} > \lambda | \theta^*_{jh}\right) \geq \frac{e^{-0.5}}{1 + 5^{0.5}} \left(\frac{2}{\pi}\right)^{0.5} c\lambda^{-2\alpha} = \tilde{c}\lambda^{-\tilde{\alpha}},$$

where $\tilde{c} = e^{-0.5}\left(1 + 5^{0.5}\right)^{-1} (2/\pi)^{0.5} c > 0$ and $\tilde{\alpha} = \alpha/2 > 0$. By symmetry, $\mathrm{pr}\left(\lambda_{jh} < -\lambda | \theta_{jh} > 0\right) \geq \tilde{c}\lambda^{-\tilde{\alpha}}$ for $\lambda > L$ sufficiently large. It is sufficient that the marginal distribution of θ^*_{jh} has power law right tail to guarantee that $\left(\lambda_{jh} | \theta_{jh} > 0\right)$ has power law tail and then that marginally λ_{jh} has power law tail. □

Proof of Theorem 2. The mode of the conditional posterior density of λ_{jh} is $\tilde{\lambda}_{jh}$ such that

$$l_s\left(\tilde{\lambda}_{jh}; y, \eta\right) + \frac{\partial}{\partial \lambda} \log\left\{f_{\lambda_{jh} | \Lambda_{-jh}}(\lambda)\right\} | \lambda = \tilde{\lambda}_{jh} = 0, \tag{2}$$

where $l_s\left(\tilde{\lambda}_{jh}; y, \eta\right)$ is the jhth element of the score function of the likelihood for the data y conditionally on the latent variables η, and $f_{\lambda_{jh} | \Lambda_{-jh}}$ is the conditional prior density function of $\left(\lambda_{jh} | \Lambda_{-jh}\right)$. Given prior symmetry, without loss of generality, we focus on $\hat{\lambda}_{jh} > 0$. In a neighbourhood $\left(\hat{\lambda}_{jh} - \varepsilon,\ \hat{\lambda}_{jh} + \varepsilon\right)$ of the conditional maximum likelihood estimate $\hat{\lambda}_{jh}$ of λ_{jh}, we can approximate the score function using a Taylor expansion: $l_s(\lambda; y) = -\mathcal{J}\left(\hat{\lambda}_{jh}\right)\left(\lambda - \hat{\lambda}_{jh}\right) + \epsilon_\varepsilon$, where $\mathcal{J}\left(\hat{\lambda}_{jh}\right) > 0$ is the negative of the derivative of $l_s(\lambda; y)$ evaluated at $\lambda = \hat{\lambda}_{jh}$, and ϵ_ε is an approximation error term such that $\lim_{\varepsilon \to 0} \epsilon_\varepsilon / \varepsilon = 0$. For $\hat{\lambda}_{jh}$ large enough, such that $\hat{\lambda}_{jh} - \varepsilon > L$ with $L \gg 0$, Lemma 1 holds for every λ in $\left(\hat{\lambda}_{jh} - \varepsilon, \hat{\lambda}_{jh} + \varepsilon\right)$, leading to the lower bound

$$-\mathcal{J}\left(\hat{\lambda}_{jh}\right)\left(\lambda - \hat{\lambda}_{jh}\right) + f'_{lb}(\lambda) + \epsilon_\varepsilon \leq l_s(\lambda; y) + \frac{\partial}{\partial \lambda} \log\left\{f_{\lambda_{jh} | \Lambda_{-jh}}(\lambda)\right\},$$

where $f'_{lb}(\lambda)$ is a non positive continuous function for every $\lambda > 0$, $\lim_{\lambda \to +\infty} f'_{lb}(\lambda) = 0$. Let ε be a function of $\hat{\lambda}_{jh}$ such that $\lim_{\hat{\lambda}_{jh} \to \infty} \varepsilon = 0$ and $\lim_{\hat{\lambda}_{jh} \to \infty} f'_{lb}\left(\hat{\lambda}_{jh}\right)/\varepsilon = 0$. The limit for $\hat{\lambda}_{jh} \to \infty$ of the lower bound evaluated in $\hat{\lambda}_{jh} - \varepsilon$ is

$$\lim_{\hat{\lambda}_{jh} \to \infty} \mathcal{J}\left(\hat{\lambda}_{jh}\right)\varepsilon + f'_{lb}\left(\hat{\lambda}_{jh} - \varepsilon\right) + \epsilon_\varepsilon = \lim_{\hat{\lambda}_{jh} \to \infty} |\varepsilon|\left\{\mathcal{J}\left(\hat{\lambda}_{jh}\right) + f'_{lb}\left(\hat{\lambda}_{jh} - \varepsilon\right)/|\varepsilon| + \epsilon_\varepsilon/|\varepsilon|\right\}.$$

Under Assumption 1, $\lim_{\hat{\lambda}_{jh} \to \infty} \mathcal{J}\left(\hat{\lambda}_{jh}\right) + f'_{lb}\left(\hat{\lambda}_{jh} - \varepsilon\right)/|\varepsilon| + \epsilon_\varepsilon/|\varepsilon| \geq 0$, which guarantees $\hat{\lambda}_{jh} - \varepsilon \leq \tilde{\lambda}_{jh} \leq \hat{\lambda}_{jh}$, and, hence $\lim_{\hat{\lambda}_{jh} \to \infty} |\tilde{\lambda}_{jh} - \hat{\lambda}_{jh}| = 0$, which proves the theorem. □

Proof of Theorem 3. Since the local scales are independent, conditionally on β, we can apply the Chernoff's method and obtain the following upper bound

$$\mathrm{pr}\left\{|\mathrm{supp}_\epsilon(\lambda_h)| > a s_p | \beta_h, \gamma_h, \tau_0\right\} \leq \exp\left(-s_p a t\right) \exp\left\{\left(e^t - 1\right) \sum_{j=1}^{p} \zeta_{\epsilon jh}\right\},$$

for every $t > 0$ and $\zeta_{\epsilon jh} = \left\{\tau_0 \gamma_h g\left(x_j^{\mathrm{T}} \beta_h\right)\right\}/\epsilon^2$ a function of β_h. Since $g\left(x_j^{\mathrm{T}} \beta_h\right)$ is of order $\leq O(\log(p)/p)$ by assumption and is limited above with respect to β_h, we can deduce $g\left(x_j^{\mathrm{T}} \beta_h\right) \leq c_j \log(p)/p$ for p sufficiently large and for some constant $c_j > 0$ that does not depend on β_h and is asymptotically of order $O(1)$ with respect to p. Then, for $p \gg 0$,

$$\sum_{j=1}^{p} g\left(x_j^{\mathrm{T}} \beta_h\right) \leq \sum_{j=1}^{p} c_j \log(p)/p \leq p \log(p)/p \max_{1 \leq j \leq p} c_j = m \log(p),$$

where $m = \max_{1 \leq j \leq p} c_j$ does not depend on β_h. Then, the upper bound is

$$\mathrm{pr}\left\{|\mathrm{supp}_\epsilon(\lambda_h)| > a s_p | \beta_h, \gamma_h, \tau_0\right\} \leq \exp\left\{-s_p a t + \left(e^t - 1\right) \frac{\tau_0 \gamma_h}{\epsilon^2} m \log(p)\right\}.$$

Let us choose $t = \log\left\{\epsilon^2/(\tau_0 \gamma_h m) + 1\right\}$. Since $s_p \geq \log(p) c_s$ for a certain $c_s > 0$, then, for any $a > (c_s t)^{-1}$, we can write

$$\mathrm{pr}\left\{|\mathrm{supp}_\epsilon(\lambda_h)| > a s_p | \beta_h, \gamma_h, \tau_0\right\} \leq \exp\{-\log(p)\tilde{a}\},$$

where $\tilde{a}$ is a positive constant such that $a = (1 + \tilde{a})(c_s t)^{-1}$. The upper bound does not depend on β_h, so

$$\mathrm{pr}\left\{|\mathrm{supp}_\epsilon(\lambda_h)| > a s_p | \gamma_h, \tau_0\right\} \leq \upsilon(p)$$

with $\upsilon(p)$ of order $O(p^{-1})$ that goes to zero. □

References

Abramowitz M & Stegun IA (1948). Handbook of mathematical functions with formulas, graphs, and mathematical tables, vol. 55. US Government printing office.

An X, Yang Q & Bentler PM (2013). A latent factor linear mixed model for high-dimensional longitudinal data analysis. Statistics in medicine 32, 4229–4239. [PubMed: 23640746]

Assmann C, Boysen-Hogrefe J & Pape M (2016). Bayesian analysis of static and dynamic factor models: An ex-post approach towards the rotation problem. Journal of Econometrics 192, 190–206.

Bhattacharya A & Dunson DB (2011). Sparse Bayesian infinite factor models. Biometrika 98, 291–306. [PubMed: 23049129]

Bhattacharya A, Pati D, Pillai NS & Dunson DB (2015). Dirichlet-Laplace priors for optimal shrinkage. Journal of the American Statistical Association 110, 1479–1490. [PubMed: 27019543]

Carvalho CM, Polson NG & Scott JG (2010). The horseshoe estimator for sparse signals. Biometrika 97, 465–480.

Dahl DB (2006). Model-based clustering for expression data via a Dirichlet process mixture model. Bayesian inference for gene expression and proteomics 4, 201–218.

Durante D (2017). A note on the multiplicative gamma process. Statistics and Probability Letters 122, 198–204.

Ferrari F & Dunson DB (2020). Bayesian factor analysis for inference on interactions. Journal of the American Statistical Association, 1–12.

Gelfand AE & Dey DK (1994). Bayesian model choice: asymptotics and exact calculations. Journal of the Royal Statistical Society: Series B (Methodological) 56, 501–514.

Jun L & Tao D (2013). Exponential Family Factors for Bayesian Factor Analysis. IEEE Transactions on neural networks and learning systems 24, 964–976. [PubMed: 24808477]

Legramanti S, Durante D & Dunson DB (2020). Bayesian cumulative shrinkage for infinite factorizations. Biometrika 107, 745–752. [PubMed: 32831355]

Lindström Å, Green M, Husby M, KÅlÅs JA & Lehikoinen A (2015). Large-scale monitoring of waders on their boreal and arctic breeding grounds in northern Europe. Ardea 103, 3–15.

Liu Z & Vandenberghe L (2010). Interior-point method for nuclear norm approximation with application to system identification. SIAM Journal on Matrix Analysis and Applications 31, 1235–1256.

Lopes HF & West M (2004). Bayesian model assessment in factor analysis. Statistica Sinica, 41–67.

McParland D, Gormley IC, McCormick TH, Clark SJ, Kabudula CW & Collinson MA (2014). Clustering south African households based on their asset status using latent variable models. The annals of applied statistics 8, 747. [PubMed: 25485026]

Miller JE, Li D, LaForgia M & Harrison S (2019). Functional diversity is a passenger but not driver of drought-related plant diversity losses in annual grasslands. Journal of Ecology 107, 2033–2039.

Miller JW & Harrison MT (2018). Mixture models with a prior on the number of components. Journal of the American Statistical Association 113, 340–356. [PubMed: 29983475]

Mitchell TJ & Beauchamp JJ (1988). Bayesian variable selection in linear regression. Journal of the American Statistical Association 83, 1023–1036.

Mnih A & Salakhutdinov RR (2008). Probabilistic matrix factorization. In Advances in neural information processing systems.

Montagna S, Tokdar ST, Neelon B & Dunson DB (2012). Bayesian latent factor regression for functional and longitudinal data. Biometrics 68, 1064–1073. [PubMed: 23005895]

Murray JS, Dunson DB, Carin L & Lucas JE (2013). Bayesian Gaussian copula factor models for mixed data. Journal of the American Statistical Association 108, 656–665. [PubMed: 23990691]

Ovaskainen O & Abrego N (2020). Joint Species Distribution Modelling: With Applications in R. Cambridge University Press.

Ovaskainen O, Abrego N, Halme P & 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.

Polson NG & Scott JG (2010). Shrink globally, act locally: Bayesian sparsity and regularization. Bayesian Statistics 9, 1–16.

Reich BJ & Bandyopadhyay D (2010). A latent factor model for spatial data with informative missingness. The annals of applied statistics 4, 439. [PubMed: 20628551]

Roberts GO & Rosenthal JS (2007). Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. Journal of applied probability 44, 458–475.

Ročková V & George EI (2016). Fast bayesian factor analysis via automatic rotations to sparsity. Journal of the American Statistical Association 111, 1608–1622.

Rousseau J & Mengersen K (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73, 689–710.

Roweis S & Ghahramani Z (1999). A unifying review of linear Gaussian models. Neural computation 11, 305–345. [PubMed: 9950734]

Roy A, Schaich-Borg J & Dunson DB (2019). Bayesian time-aligned factor analysis of paired multivariate time series. arXiv preprint arXiv:1904.12103.

Schiavon L & Canale A (2020). On the truncation criteria in infinite factor models. Stat 9, e298.

Thomas DC, Conti DV, Baurley J, Nijhout F, Reed M & Ulrich CM (2009). Use of pathway information in molecular epidemiology. Human genomics 4, 21. [PubMed: 21072972]

Tikhonov G, Abrego N, Dunson D & Ovaskainen O (2017). Using joint species distribution models for evaluating how species-to-species associations depend on the environmental context. Methods in Ecology and Evolution 8, 443–452.

Tikhonov G, Opedal ÁH, Abrego N, Lehikoinen A, de Jonge MM, Oksanen J & Ovaskainen O (2020). Joint species distribution modelling with the R-package hmsc. Methods in ecology and evolution 11, 442–447. [PubMed: 32194928]

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

Yang L, Fang J, Duan H, Li H & Zeng B (2018). Fast low-rank Bayesian matrix completion with hierarchical Gaussian prior models. IEEE Transactions on Signal Processing 66, 2804–2817.

Yuan M & Lin Y (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68, 49–67.

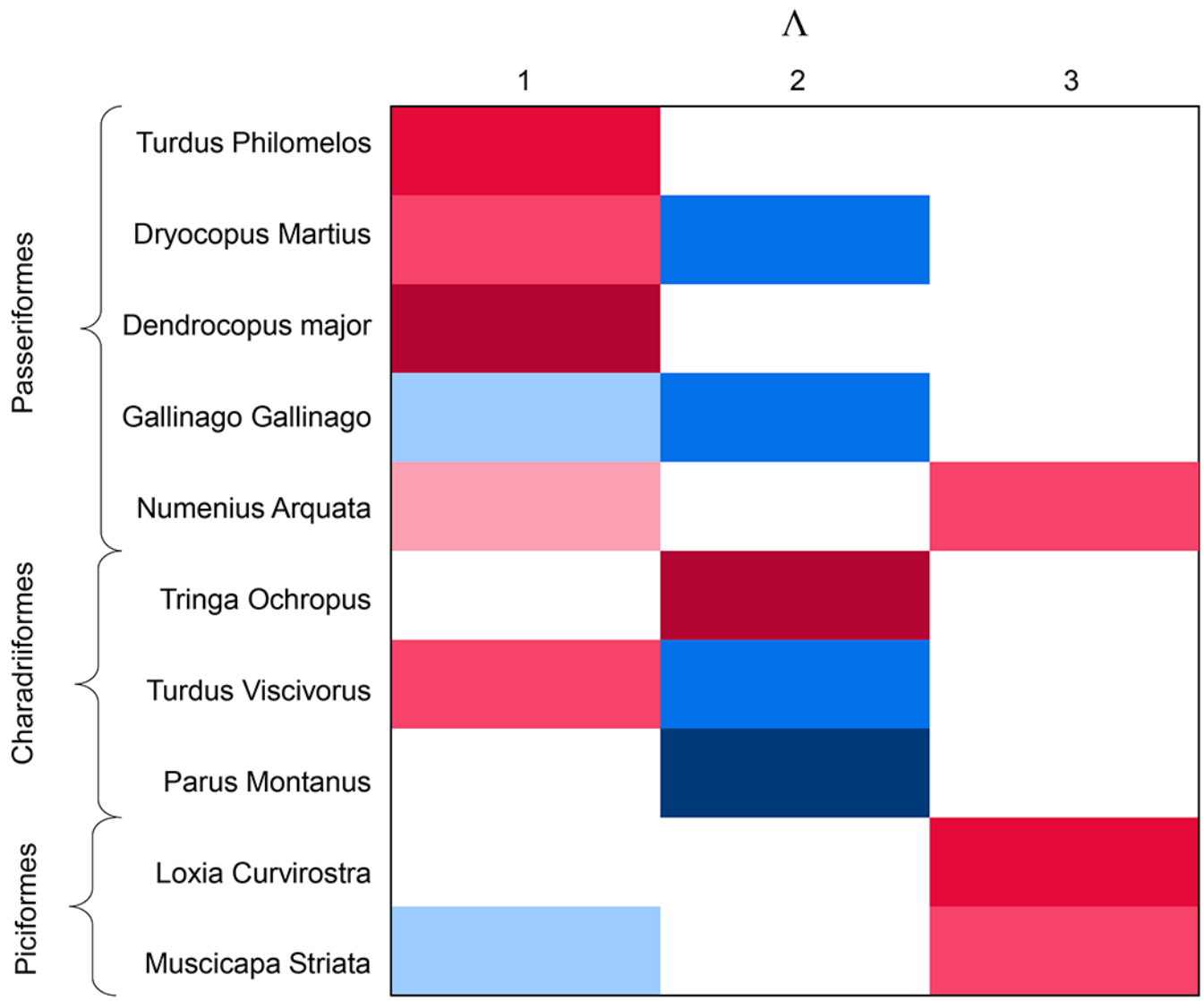

Fig. 1:

Illustrative loadings matrix of an ecology application, where the rows refer to ten bird species belonging to three phylogenetic orders. White cells represent the elements of Λ equal to zero, while blue and red cells represent negative and positive values, respectively.

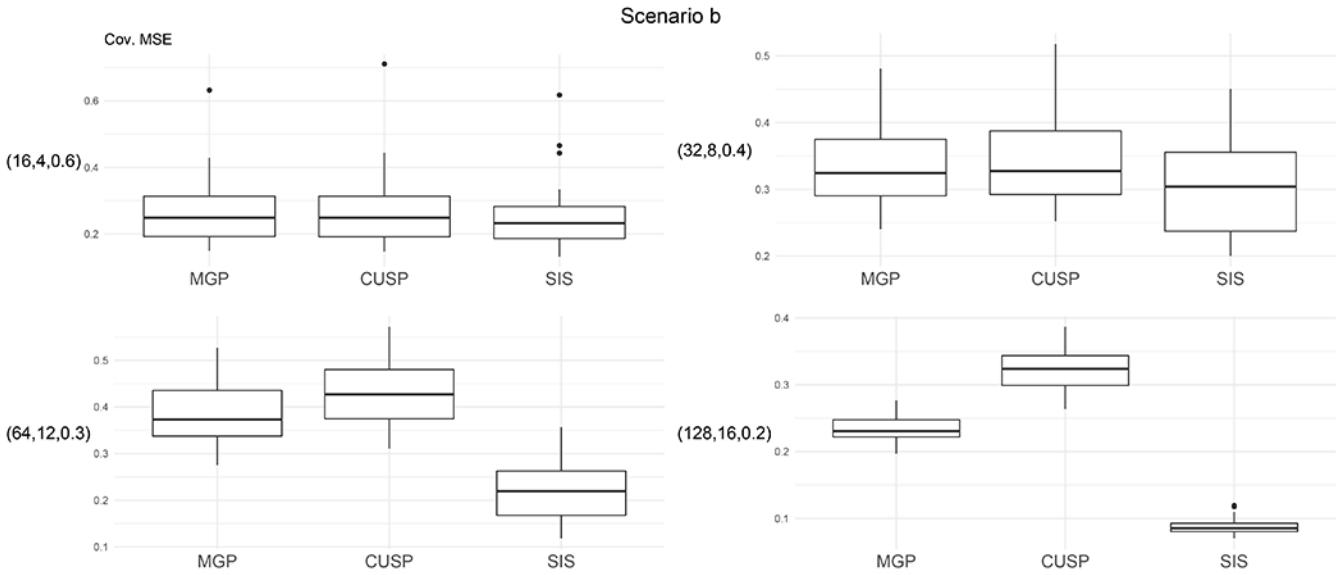

Fig. 2:

Boxplots of mean squared error of the covariance matrix of each model for different combinations of (*p, k, s*) in Scenario b. Cov. MSE, covariance mean squared error; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process.

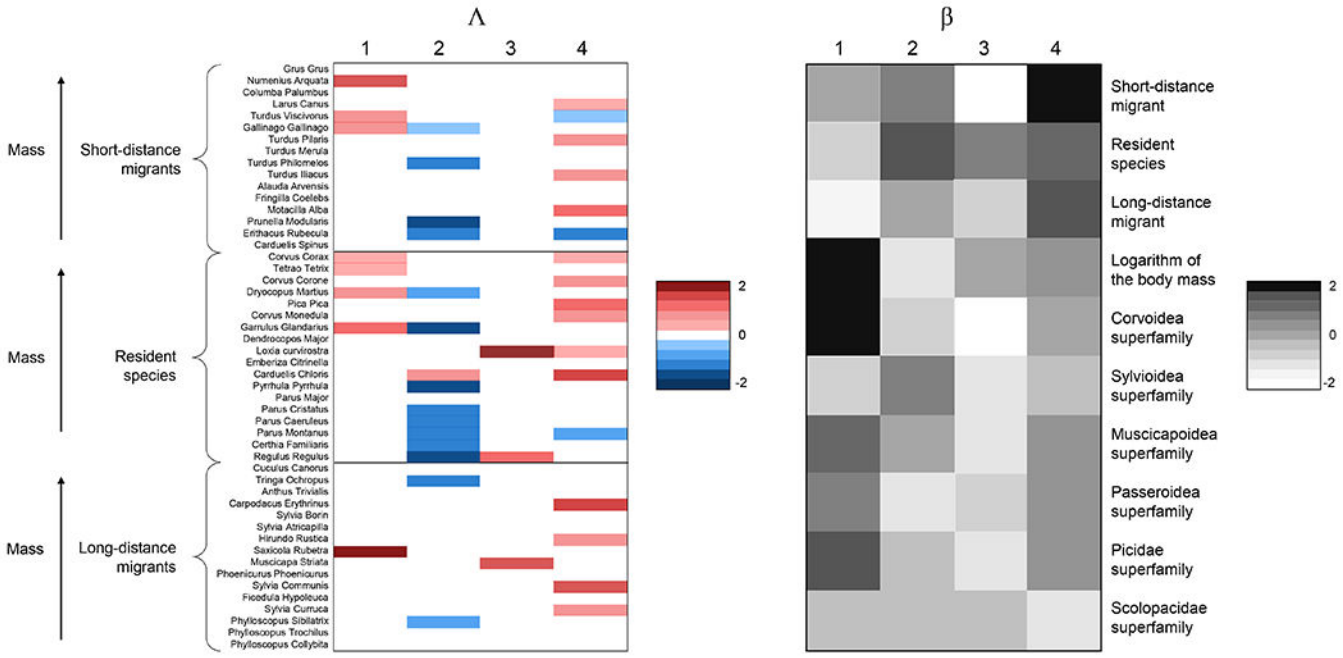

Fig. 3:

Posterior summaries $\Lambda^{(t^*)}$ and $\beta^{(t^*)}$ of the structured increasing shrinkage model; rows of left matrix refer to 50 birds species, and rows of right matrix to ten species traits. Light coloured cells of $\beta^{(t^*)}$ induce shrinkage on corresponding cells of $\Lambda^{(t^*)}$.

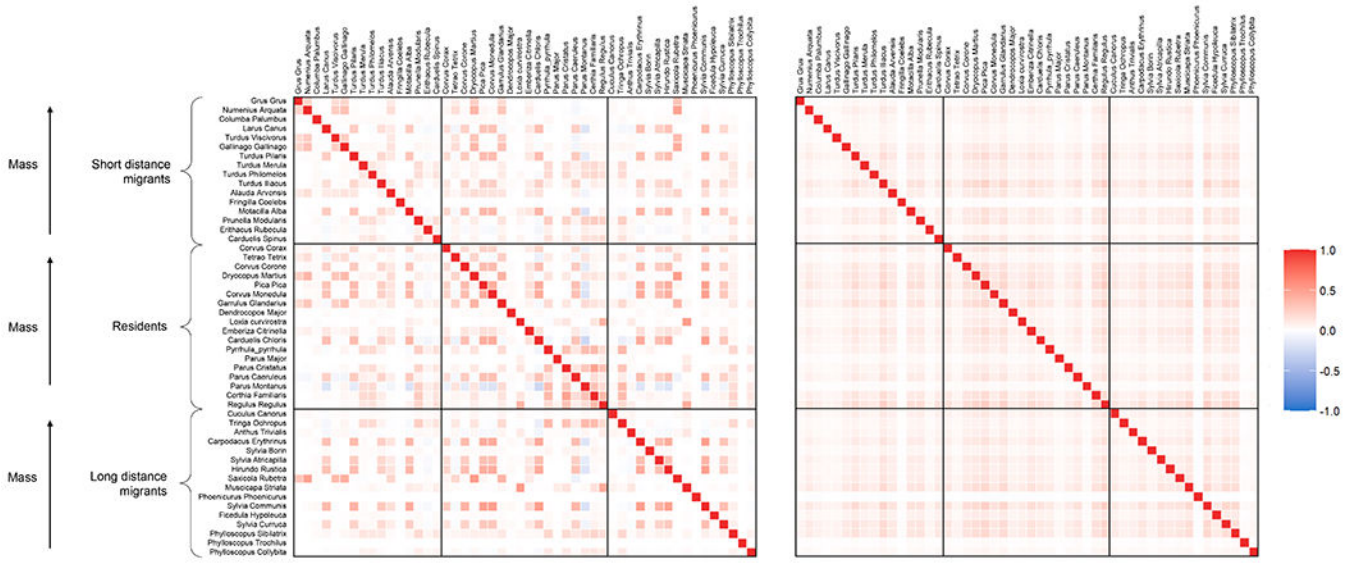

Fig. 4:

Posterior mean of the correlation matrices estimated by the structured increasing shrinkage model (on the left) and the multiplicative gamma process model (on the right).

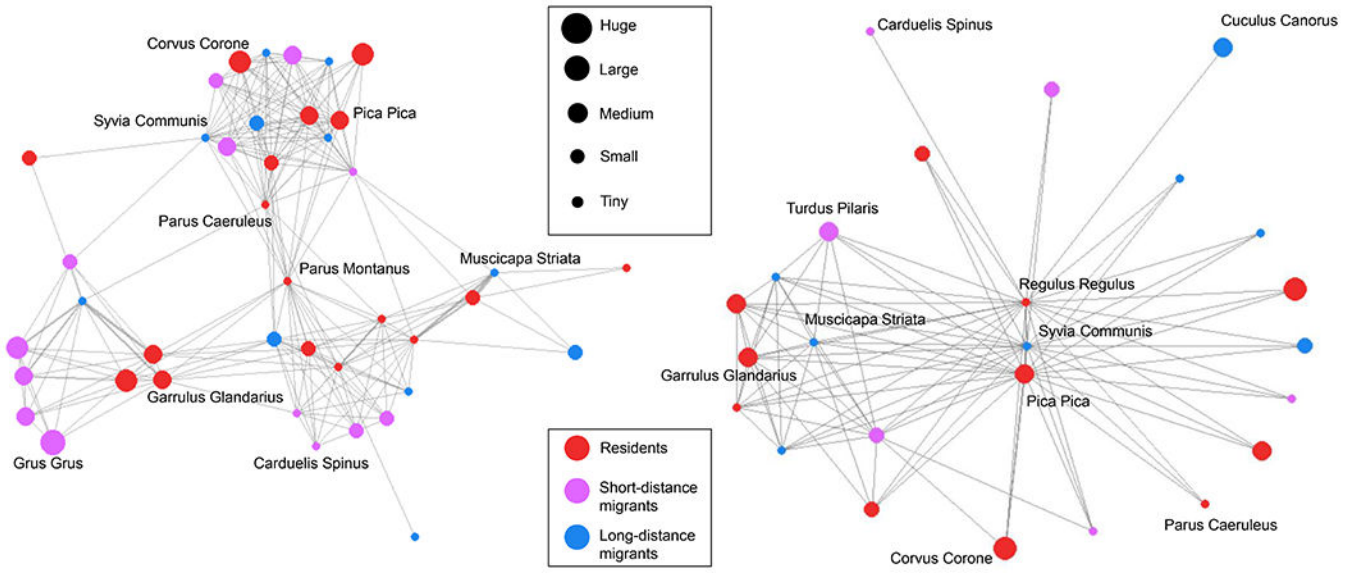

Fig. 5:

Graphical representation based on the inverse of the posterior mean of the correlation matrices estimated by the structured increasing shrinkage model (on the left) and the multiplicative gamma process model (on the right). Edge thicknesses are proportional to the latent partial correlations between species. Values below 0.025 are not reported. Nodes are positioned using a FruchtermanReingold force-direct algorithm.

Table 1:

Median and interquartile range of LPML and E(H_a | y) in 25 replications of Scenario a for different combinations of (p, k); Scenario a is a worst case for the proposed SIS method.

	(*p*, *k*)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
LPML	(16,4)	−28.68	0.42	−28.68	0.43	−28.65	0.41
	(32,8)	−60.08	0.45	−60.09	0.45	−60.07	0.49
	(64,12)	−117.68	0.56	−117.75	0.53	−117.88	0.56
	(128,16)	−225.04	1.04	−225.13	1.04	−228.76	1.47
$E(H_a \mid y)$	(16,4)	8.17	1.44	4.00	0.00	4.00	0.00
	(32,8)	10.68	0.33	8.00	0.00	8.00	0.00
	(64,12)	14.16	1.09	12.00	0.00	12.00	0.00
	(128,16)	17.03	0.47	16.00	0.00	18.00	0.02

LPML, logarithm of the pseudo-marginal likelihood; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; $Q_{0.5}$, median; IQR, interquartile range.

Table 2:

Median and interquartile range of the mean classification error computed in 25 replications assuming Scenario b and several combinations of (p, k, s)

MCE	(*p*, *k*, *s*)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
	(16,4,0.6)	1.06	0.16	0.38	0.01	0.24	0.09
	(32,8,0.4)	0.70	0.07	0.48	0.08	0.16	0.09
	(64,12,0.3)	0.61	0.07	0.58	0.01	0.09	0.06
	(128,16,0.2)	0.49	0.03	0.52	0.08	0.04	0.01

MCE, mean classification error; MGP, multiplicative gamma process; CUSP, cumulative shrinkage process; SIS, structured increasing shrinkage process; $Q_{0.5}$, median; IQR, interquartile range.