(2022), *pp.* 1–21

January 11, 2022

Supplementary material for “Generalized infinite factorization models”

BY L. SCHIAVON AND A. CANALE

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

lorenzo.schiavon@phd.unipd.it, canale@stat.unipd.it

D. B. DUNSON

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

dunson@duke.edu

SUMMARY

This 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 with respect to varying hyperparameters.

S1. PROPOSITIONS AND PROOFS

PROPOSITION S1. *Let* $\Pi_\Lambda \otimes \Pi_\Sigma$ *denote the prior on* (Λ, Σ)*. Let* Θ_Λ *and* Θ_Σ *denote the sample spaces of the matrices* Λ *and* Σ*, respectively. If* $E(\phi_{jh}) = E(\phi_{lh})$ *for every* $h, l \in \{1, \ldots, \infty\}$ *and* $\sum_{h=1}^{\infty} E(\gamma_h) < \infty$*, then,* $\Pi_\Lambda \otimes \Pi_\Sigma(\Theta_\Lambda \times \Theta_\Sigma) = 1$.

Proof. Assume $\Sigma \in \Theta_\Sigma$ and $(\Psi, \Lambda) \in \Theta_\Psi \times \Theta_\Lambda$, with Θ_Σ the set of $p \times p$ positive semi-definite matrices with finite elements, and

$$\Theta_\Psi \times \Theta_\Lambda = \left\{ \Lambda = (\lambda_{jh}),\ \Psi = (\psi_{hh}) : \sum_{h=1}^{\infty} \lambda_{jh}\psi_{hh}\lambda_{sh} < \infty \ \forall\, j, s \in (1, \ldots, p) \right\}.$$

Due to independence, we can study the prior on Σ and Λ separately. The prior on Σ is defined on the set of positive semi-definite matrices. Therefore, it is sufficient to prove that the elements of $\Lambda\Psi\Lambda^{\mathrm{T}}$ are finite almost surely. Using Cauchy-Schwartz, it is straightforward to show that all the entries of $\Lambda\Psi\Lambda^{\mathrm{T}}$ are finite if and only if $\sum_{h=1}^{\infty} \psi_{hh}\lambda_{jh}^2 < \infty$ $(j = 1, \ldots, p)$. Let c satisfy $c > \max_{h=1,\ldots,\infty} \psi_{hh}$. Since

$$E(\lambda_{jh}^2) = E\{E(\lambda_{jh}^2 \mid \phi_{jh}, \gamma_h, \tau_0)\} = E(\phi_{jh})E(\gamma_h)E(\tau_0),$$

and $E(\phi_{jh}) = E(\phi_{j1})$ $(j = 1, \ldots, p;\ h = 1, \ldots, \infty)$, it is sufficient that $\sum_{h=1}^{\infty} E(\gamma_h) < \infty$ to prove that $\sum_{h=1}^{\infty} E(\lambda_{jh}^2) = E(\phi_{j1})E(\tau_0)\sum_{h=1}^{\infty} E(\gamma_h) < \infty$ and then $\sum_{h=1}^{\infty} \psi_{hh}\lambda_{jh}^2 < c\sum_{h=1}^{\infty} \lambda_{jh}^2 < \infty$. □

Proof of Proposition 1. The trace of Ω is $\mathrm{tr}(\Sigma) + \mathrm{tr}(\Lambda_H\Psi_H\Lambda_H^{\mathrm{T}}) + \mathrm{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})$, where $\Lambda_{\Delta_H} = \Lambda - \Lambda_H$ and $\Psi_{\Delta_H} = \Psi - \Psi_H$. Hence, it is equivalent to rewrite the probability of interest as

$$\operatorname{pr}\left\{\frac{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})}{\operatorname{tr}(\Omega)} \geq 1-T\right\}.$$

By Markov's Inequality

$$\operatorname{pr}\left\{\frac{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})}{\operatorname{tr}(\Omega)} \geq 1-T\right\} \leq E\left\{\frac{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})}{\operatorname{tr}(\Omega)}\right\}/(1-T).$$

The expected ratio of two random variables u and v is $E(u/v) = \operatorname{cov}(u, 1/v) + E(u)E(1/v)$, which allows us to write $E(u/v) \leq E(u)E(1/v)$ if $\operatorname{cov}(u, 1/v) \leq 0$. Then, since the covariance between $\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})$ and $\operatorname{tr}(\Omega)$ is non-negative, the following inequality holds

$$E\left\{\frac{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})}{\operatorname{tr}(\Omega)}\right\} \leq E\{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})\}E\left(\frac{1}{\operatorname{tr}(\Omega)}\right).$$

The trace $\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})$ is equal to $\sum_{j=1}^{p}\sum_{h=H+1}^{\infty}\psi_{hh}\lambda_{jh}^2$. The variance of λ_{jh} is $E(\lambda_{jh}^2) = E(\phi_{j1})E(\gamma_h)E(\tau_0)$. Let c satisfy $c \geq \max_{h=1,\dots,\infty}\psi_{hh}$. Since $E(\phi_{j1})$ is finite and $E(\gamma_h) = ab^{h-1}$ with a, b positive constants and $b < 1$, then

$$E\{\operatorname{tr}(\Lambda_{\Delta_H}\Psi_{\Delta_H}\Lambda_{\Delta_H}^{\mathrm{T}})\} \leq cE(\tau_0)a\frac{b^H}{1-b}\sum_{j=1}^{p}E(\phi_{j1}).$$

Since $\operatorname{tr}(\Omega) = \operatorname{tr}(\Lambda\Psi\Lambda^{\mathrm{T}}) + \operatorname{tr}(\Sigma)$, we know that $\operatorname{tr}(\Omega) \geq \sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2 + \sigma_j^2$ for any j in $1, \dots, p$, where σ_j^2 is the j-th diagonal element of Σ. Then, for any j in $1, \dots, p$, we obtain

$$\frac{1}{\operatorname{tr}(\Omega)} \leq \frac{1}{\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2 + \sigma_j^2},$$

and, consequently,

$$E\left\{\frac{1}{\operatorname{tr}(\Omega)}\right\} \leq E\left(\sigma_j^{-2}\right), \qquad E\left\{\frac{1}{\operatorname{tr}(\Omega)}\right\} \leq E\left\{\left(\sum_{h=1}^{\infty}\psi_{hh}\lambda_{jh}^2\right)^{-1}\right\}.$$

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

$$\operatorname{pr}\left\{\frac{\operatorname{tr}(\Lambda_H\Psi_H\Lambda_H^{\mathrm{T}}) + \operatorname{tr}(\Sigma)}{\operatorname{tr}(\Omega)} \leq T\right\} \leq \left(\frac{1}{1-T}\right)a\,c\,\frac{b^H}{1-b}\,m_\Omega\,E(\tau_0)\sum_{j=1}^{p}E(\phi_{j1}),$$

as stated by the Theorem. □

Proof of Lemma 1. Consistently with Proposition 2, $(\lambda_{jh} \mid \Lambda_{-jh})$ has power law tail if $(\theta_{jh} \mid \Lambda_{-jh})$ has power law tail. Furthermore, $\operatorname{pr}(|\lambda_{jh}| > \lambda \mid \Lambda_{-jh})$ has power law tail for large λ if and only if $\operatorname{pr}(|\lambda_{jh}| > \lambda \mid \Lambda_{-jh}, \theta_{jh} > 0)$ has power law tail and $\operatorname{pr}(\theta_{jh} > 0 \mid \Lambda_{-jh}) > 0$. The latter inequality is always true when the marginal $\operatorname{pr}(\theta_{jh} > 0)$ is positive. To prove $(\theta_{jh} \mid \Lambda_{-jh})$ has power law tail, we apply Lemma 3. We first focus on proving the lemma when ϕ_{jh} satisfies the power law tail condition with $\tau_0\gamma_h = w_h$. As the local scale ϕ_{jh} is independent from (Λ_{-jh}, w_h) given β_h, its conditional density is

$$f_{\phi_{jh}|w_h,\Lambda_{-(jh)}}(\phi) = \int_{\Re} f_{\phi_{jh}|\beta_h}(\phi)\, f_{\beta_h|w_h,\Lambda_{-jh}}(\beta)\,\mathrm{d}\beta.$$

As the tail conditions hold for any possible prior on β, we have

$$f_{\phi_{jh}}(\phi) = \int_{\Re} f_{\phi_{jh}|\beta_h}(x)\, f(\beta)\, \mathrm{d}\beta, \qquad f_{\phi_{jh}}(\tilde{\phi}) \propto \tilde{\phi}^{-\alpha}, \qquad \tilde{\phi} = \{\phi : \phi > l\}, \qquad L \gg 0,$$

for any prior density f defined on $\Re$. Hence, $(\phi_{jh} \mid w_{jh}, \Lambda_{-jh})$ is power law tail distributed. We now focus on proving the lemma when τ_0 or γ_h are power law tail distributed. Let $r_h^* = (r_h \mid r_h > 0)$ and $w_{jh}^* = (w_{jh} \mid w_{jh} > 0)$, where r_h is the scale parameter with power law tail and w_h is the product of the remaining two scale parameters, respectively. By Bayes' Theorem

$$f_{r_h^*|w_{jh}^*,\Lambda_{-jh}}(r) = \frac{f_{\Lambda_{-jh}|w_{jh}^*,r_h^*}(\Lambda_{-jh}; r) f_{r_h^*|w_{jh}^*}(r)}{f_{\Lambda_{-jh}|w_{jh}^*}(\Lambda_{-jh})}.$$

Since r_h^* is independent from w_{jh}^* for any parameter scale, it is sufficient to prove that the function $f_{\Lambda_{-jh}|w_{jh}^*,r_h^*}(\Lambda_{-jh}; r)$ decreases slower than $c\, r^{-\alpha}$, for $c, \alpha >$ positive constants, when $r \to \infty$. Denoting $F_{\tau_0,\phi_{11}\dots\phi_{pk},\gamma_1,\dots,\gamma_{h-1},\gamma_{h+1},\dots,\gamma_k|w_{jh}^*,r_h^*}$ the probability measure for conditional density $f_{\tau_0,\phi_{11}\dots\phi_{pk},\gamma_1,\dots,\gamma_{h-1},\gamma_{h+1},\dots,\gamma_k|w_{jh}^*,r_h^*}$, we can write

$$\begin{aligned} f_{\Lambda_{-jh}|w_{jh}^*,r_h^*}(\Lambda_{-jh}; r) &= \int f_{\Lambda_{-jh}|\tau_0,\phi_{11}\dots\phi_{pk},\gamma_1,\dots,\gamma_k}(\Lambda_{-jh}; r)\, \mathrm{d}F_{\tau_0,\phi_{11}\dots\phi_{pk},\gamma_1,\dots,\gamma_{h-1},\gamma_{h+1},\dots,\gamma_k|w_{jh}^*,r_h^*} \\ &= \int \prod_{(s,m)\neq(j,h)} f_{\lambda_{sm}|\theta_{sm}}(\lambda_{sm}; r)\, \mathrm{d}F_{\tau_0,\phi_{11},\dots,\gamma_k|w_{jh}^*,r_h^*} \\ &= E\Bigg\{ \prod_{(s,m)\neq(j,h)} f_{\lambda_{sm}|\theta_{sm}}(\lambda_{sm}; r) \;\Bigg|\; w_{jh}^*, r_h^*, \Lambda_{-jh} \Bigg\} \end{aligned}$$

The product inside the expectation is zero when there is a pair of indices (s, m) such that $\lambda_{sm} \neq 0$ and $\theta_{sm} = 0$. However, since the probability $\mathrm{pr}(\theta_{sm} = 0 \mid \lambda_{sm} \neq 0) = 0$, we know that the expected value of the product between the functions $f_{\lambda_{sm}|\theta_{sm}}(\lambda_{sm}; r)$, given $w_{jh}^*, r_h^*, \Lambda_{-jh}$, is strictly positive. We first focus on the case $\gamma_h = r_h$ and prove that $f_{\Lambda_{-jh}|w_{jh}^*,\gamma_h^*}(\Lambda_{-jh}; \gamma)$ decreases slower than $c\gamma^{-\alpha}$ for $c, \alpha > 0$. In this case, we can write the above expectation as

$$E\Bigg\{ \prod_{s=1,m\neq h}^{p} f_k(\lambda_{sm}) \prod_{s\neq j} f_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh}; \gamma_h^*) \;\Bigg|\; w_{jh}^*, \gamma_h^*, \Lambda_{-jh} \Bigg\},$$

where $\prod_{s=1,m\neq h}^{p} f_k(\lambda_{sm})$ is a product between $(k-1) \times p$ strictly positive random variables that does not depend on w_{jh}^* and γ_h^*, while $\prod_{s\neq j} f_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh}; \gamma_h^*)$ is a product between p strictly positive random variables. In particular, if $w_{sh} = 0$, then $f_{\lambda_{sm}|\theta_{sh}}(\lambda_{sh}; \gamma_h^*) = \mathbb{1}(\lambda_{sh} = 0)$. If $w_{sh} > 0$, then

$$f_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh}; \gamma_h^*) = (2\pi w_{sh}^* \gamma_h^*)^{-0.5} \exp\left(-\frac{\lambda_{sh}^2}{2 w_{sh}^* \gamma_h^*} \right) > (2\pi w_{sh}^* \gamma_h^*)^{-0.5} \exp\left(-\frac{\lambda_{sh}^2}{2 w_{sh}^*} \right).$$

Therefore, the upper bound

$$f_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh}; \gamma_h^*) \geq \begin{cases} \min\{1,\ (2\pi w_{sh}^* \gamma_h^*)^{-0.5}\}, & \text{if } \lambda_{sh=0} \\ (2\pi w_{sh}^* \gamma_h^*)^{-0.5} \exp\{-\lambda_{sh}^2/(2w_{sh}^*)\} & \text{if } \lambda_{sh\neq 0}, \end{cases}$$

holds with probability equal to 1. For $\gamma > 1$, we note that $f_{\lambda_{sh}|\theta_{sh}}(\lambda_{sh};\gamma) \geq \gamma^{-0.5} u_{\lambda_{sh}}$ with

$$u_{\lambda_{sh}} = \begin{cases} \min\{1,\ (2\pi w^*_{sh})^{-0.5}\}, & \text{if } \lambda_{sh} = 0, \\ (2\pi w^*_{sh})^{-0.5} \exp\{-\lambda^2_{sh}/(2w^*_{sh})\} & \text{if } \lambda_{sh} \neq 0. \end{cases}$$

Then,

$$E\left\{ \prod_{(s,m)\neq(j,h)} f_{\lambda_{sm}|\theta_{sm}}(\lambda_{sm};\gamma^*_h) \;\middle|\; w^*_{jh}, \gamma^*_h, \Lambda_{-jh} \right\} \geq$$

$$E\left\{ \prod_{s=1,m\neq h}^{p} f_k(\lambda_{sm}) \prod_{s\neq j} \gamma_h^{*-0.5} u_{\lambda_{sh}} \;\middle|\; w^*_{jh}, \gamma^*_h, \Lambda_{-jh} \right\} =$$

$$\gamma_h^{*-0.5\,(p-1)} E\left\{ \prod_{s=1,m\neq h}^{p} f_k(\lambda_{sm}) \prod_{s\neq j} u_{\lambda_{sh}} \;\middle|\; w^*_{jh}, \Lambda_{-jh} \right\},$$

where the expectation is strictly positive and not depending on γ_h. Therefore, for γ sufficiently large, $f_{\Lambda_{-jh}|w^*_{jh},\gamma^*_h}(\Lambda_{-jh};\gamma) \geq c\gamma^{-\alpha}$ holds, with $c, \alpha > 0$, so that $(\gamma_h \mid w_{jh}, \Lambda_{-jh})$ is power law tail distributed. Similarly, if $\tau_0 = r_h$ $(h = 1, \ldots, H)$,

$$E\left\{ \prod_{(s,m)\neq(j,h)} f_{\lambda_{sm}|\theta_{sm}}(\lambda_{sm};\tau^*_0) \mid w^*_{jh}, \tau^*_0, \Lambda_{-jh} \right\} \geq$$

$$E\left(\prod_{(s,m)\neq(j,h)} \tau_0^{*-0.5} u_{\lambda_{sm}} \mid w^*_{jh}, \tau^*_0, \Lambda_{-jh} \right) =$$

$$\tau_0^{*-0.5\,(pH-1)} E\left(\prod_{(s,m)\neq(j,h)} u_{\lambda_{sm}} \mid w^*_{jh}, \Lambda_{-jh} \right),$$

where

$$u_{\lambda_{sm}} = \begin{cases} \min\{1,\ (2\pi w^*_{sm})^{-0.5}\}, & \text{if } \lambda_{sm} = 0, \\ (2\pi w^*_{sm})^{-0.5} \exp\{-\lambda^2_{sm}/(2w^*_{sm})\}, & \text{if } \lambda_{sm} \neq 0. \end{cases}$$

is strictly positive and does not depend on τ_0. Then, if the number H of columns of Λ_H is finite, $f_{\Lambda_{-jh}|w^*_{jh},\tau^*_0}(\Lambda_{-jh};\tau) \geq c\tau^{-\alpha}$ with $c, \alpha > 0$ and τ sufficiently large, implying $(\tau_0 \mid w_{jh}, \Lambda_{-jh})$ is power law tail distributed. Hence, if any of the scale parameters is power law tail distributed for any prior on β, then its distribution, conditionally on Λ_{-jh} and on the product of the other two parameters, is power law tail distributed and, as a consequence, $(\lambda_{jh} \mid \Lambda_{-jh})$ is power law tail distributed. Since $f_{\lambda_{jh}|\Lambda_{-jh}}(\lambda) \geq c|\lambda|^{-\alpha}$ for certain c, α positive constants and $|\lambda| > L$ sufficiently large, in the same settings, we can write

$$f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda) = c|\lambda|^{-\alpha}\{1 + t(|\lambda|)\},$$

where $t(|\lambda|)$ is a positive function. Then,

$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} = -\frac{\alpha}{\lambda} + \frac{\partial t(\lambda)}{\partial\lambda} \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$

$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} = \frac{\alpha}{\lambda} + \frac{\partial\{-t(\lambda)\}}{\partial\lambda} \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

We now consider the sign of the derivative of $t(|\lambda|)$. If $t(|\lambda|)$ is not decreasing,

$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq -\frac{\alpha}{\lambda}, \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$
$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq \frac{\alpha}{\lambda}, \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

whereas if $t(|\lambda|)$ is decreasing, its derivative goes to zero when $|\lambda|$ goes to infinity. Therefore,

$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq f'_{lb}(\lambda) \qquad \text{for } \lambda > L \quad \text{and } L \gg 0,$$
$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq -f'_{lb}(|\lambda|) \qquad \text{for } \lambda < -L \quad \text{and } L \gg 0,$$

where $f'_{lb}(\lambda) < 0 \; \forall \, \lambda > 0$ and $\lim_{\lambda\to\infty} f'_{lb}(|\lambda|) = 0$. The proof is concluded by using this result along with the fact that $f_{\lambda_{jh}|\Lambda_{-(jh)}}(|\lambda|)$ is decreasing when $\lambda \to \infty$,

$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \leq 0 \qquad \text{for } \lambda > L \quad \text{and } L \gg 0$$
$$\frac{\partial[\log\{f_{\lambda_{jh}|\Lambda_{-(jh)}}(\lambda)\}]}{\partial\lambda} \geq 0 \qquad \text{for } \lambda > -L \quad \text{and } L \gg 0,$$

showing that the limit of the derivative for $|\lambda| \to \infty$ is equal to zero. □

Proof of Lemma 2. In both the multiplicative gamma process and cumulative shrinkage process, priors on Λ are exchangeable within columns, that is $\mathrm{pr}(|\lambda_{jh}| > \epsilon \mid \gamma_h, \tau_0) = \zeta_{\epsilon h}$ does not depend on j. Then, the prior density of $|\mathrm{supp}_\epsilon(\lambda_h)|$, conditionally on γ_h and τ_0 is *a priori* distributed as a sum of independent and identically distributed Bernoulli random variables $\mathrm{Ber}(\zeta_{\epsilon h})$. Furthermore, $\zeta_{\epsilon h}$ does not depend on p. By applying the Chernoff's method, we obtain

$$\mathrm{pr}\{|\mathrm{supp}_\epsilon(\lambda_h)| < as_p \mid \gamma_h, \tau_0\} \leq \exp\left\{ats_p + p\zeta_{\epsilon h}(e^{-t} - 1)\right\},$$

for any $t > 0$ and with $1 - e^{-t} > 0$. Hence,

$$\mathrm{pr}\{|\mathrm{supp}_\epsilon(\lambda_h)| > as_p \mid \gamma_h, \tau_0\} \geq 1 - \exp[-p\{(1 - e^{-t})\zeta_{\epsilon h} - ats_p/p\}],$$

where the limit of the lower bound is $\lim_{p\to\infty} 1 - \exp[-p\{(1 - e^{-t})\zeta_{\epsilon h} - ats_p/p\}] = 1$, which concludes the proof. □

Proof of Corollary 1. i. It is sufficient to prove the conditions required by Theorem 1. We have $E(\gamma_h) = E(\vartheta_h)E(\rho_h) = E(\rho_h)\, b_\theta/(a_\theta - 1)$, where

$$E(\rho_h) = 1 - \sum_{l=1}^{h} E(w_l) = 1 - \sum_{l=1}^{h-1} E(w_l) - E(w_h) = E(\rho_{h-1}) - E(w_h).$$

Since the random variable w_l is obtained as a product of positive random variables, $E(w_l) > 0$ for every $l = 1, \dots, h$. Therefore $E(\gamma_h) < E(\gamma_{h-1})$ for each $h = 2, \dots, H$.

ii. It is sufficient to prove the conditions required by Proposition 1. It is straightforward to verify $E(\tau_0) = 1$ and $E(\phi_{jh}) \leq 1$ for $j = 1, \dots, p$ and $h = 1, \dots, \infty$. The column scale expectation is

$$E(\gamma_h) = E(\vartheta_h)\left(\frac{\alpha}{1+\alpha}\right)\left(\frac{\alpha}{1+\alpha}\right)^{h-1},$$

which can be written in a form ab^{h-1}. The elements σ_j^{-2} are gamma distributed guaranteeing finite expectation for all $j = 1, \ldots, p$. □

Proof of Corollary 2. It is sufficient to prove the conditions required by Theorem 2. The probability density function of the column scale γ_h $(h = 1, 2, \ldots)$ of the structured increasing shrinkage prior evaluated at a certain $\gamma > 0$ is

$$f_{\gamma_h}(\gamma) = \text{pr}(\rho_h = 1) f_{\vartheta_h}(\gamma) \propto \gamma^{-a_\theta - 1} \exp(-b_\theta/\gamma),$$

where $f_{\vartheta_h}(\gamma)$ is the inverse gamma probability density function evaluated at γ. The function $\gamma^{-a_\theta-1}\exp(-b_\theta/\gamma)$ is of order $O(\gamma^{-a_\theta-1})$ as γ goes to infinity. Since $a_\theta > 0$, we conclude that the column scale γ_h is power law tail distributed. The independence between γ_h and β_h $(h = 1, 2, \ldots)$ guarantees that the latter result hold for any possible prior distribution f_β on β. □

Proof of Corollary 3. It is sufficient to prove the conditions required by Theorem 3. The structured increasing shrinkage prior is such that, for every $j = 1, \ldots, p$ and $h \geq 1$, we have $g(x_j^T \beta_h) \leq c_p < 1$. The proof is obtained under the assumption $c_p = O\{\log(p)/p\}$. □

S2. SIMULATION EXPERIMENTS

S2.1. *Gibbs sampler for structured increasing shrinkage model for Gaussian data*

We can rewrite the model for y_{ij} for the specific case of the structured increasing shrinkage process and Gaussian data as

$$y_{ij} = \sum_{h=1}^{\infty} \sqrt{\rho_h}\sqrt{\phi_{jh}}\,\lambda^*_{jh}\eta_{ih} + \epsilon_{ij} \qquad \lambda^*_{jh} \sim N(0, \vartheta_h),$$

where λ^*_{jh} is a continuous random variable and we let $\beta_{mh} \sim N(0, \sigma^2_\beta)$. The notation $(x \mid -)$ denotes the full conditional distribution of x conditionally on everything else. Given H the number of factors of the truncated model, the sampler cycles through the following steps.

Step S1. Update, for $i = 1, \ldots, n$, the factor η_i according to the posterior full conditional

$$(\eta_i \mid -) \sim N_H\big\{(I_H + \Lambda_H^{\mathrm{T}}\Sigma^{-1}\Lambda_H)^{-1}\Lambda_H^{\mathrm{T}}\Sigma^{-1}y_i,\, (I_H + \Lambda_H^{\mathrm{T}}\Sigma^{-1}\Lambda_H)^{-1}\big\}.$$

Step S2. Update, for j in $1, \ldots, p$, the elements of Σ, by sampling

$$(\sigma_j^{-2} \mid -) \sim \text{Ga}\left\{a_\sigma + \frac{n}{2},\, b_\sigma + \frac{1}{2}\sum_{i=1}^{n}(y_{ij} - \lambda_j^{\mathrm{T}}\eta_i)^2\right\}.$$

Step S3. Update β_h $(h = 1, \ldots, H)$ exploiting the Pólya-Gamma data-augmentation strategy (Polson et al., 2013) and the decompostition $\phi_{jh} = \phi_{jh}^{(L)}\phi_{jh}^{(C)}$, with $\phi_{jh}^{(L)}\phi_{jh}^{(C)}$ independent a priori and distributed as Ber$\{\text{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)\}$ and Ber(c_p), respectively.

Substep S3.1. Update $\phi_{jh}^{(L)}$, for $j = 1, \ldots, p$ and $h = 1, \ldots, H$, setting $\phi_{jh}^{(L)} = 1$ if $\phi_{jh} = 1$ and sampling from the full conditional distribution

$$\text{pr}(\phi_{jh}^{(L)} = l) \propto \begin{cases} 1 - \text{logit}^{-1}(x_j^{\mathrm{T}}\beta_h) & \text{for } l = 0, \\ \text{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)(1 - c_p) & \text{for } l = 1, \end{cases}$$

if $\phi_{jh} = 0$.

Substep S3.2. Let $f(y) \propto \sum_{n=0}^{\infty}(-1)^n A_n (2\pi y^3)^{-0.5} \exp\{-(2n+b)^2(8y)^{-1} - 0.5c^2 y\}$ indicate the probability density function of a Pólya-Gamma distributed random variable $y \sim \text{PG}(b,c)$. For each $h = 1,\dots,H$, generate p independent random variables $d_{j(h)}$ sampling from the full conditional distribution $(d_{j(h)} \mid -) \sim \text{PG}(1, x_j^{\mathrm{T}}\beta_h)$. Let $D_{(h)}$ denote the $p \times p$ diagonal matrix with entries $d_{j(h)}$ $(j = 1,\dots,p)$.

Substep S3.3. Define the $q \times q$ diagonal matrix $B = \sigma_\beta^2 I_q$. For each $h = 1,\dots,H$, update β_h sampling from

$$(\beta_h \mid -) \sim N_q\{(x^{\mathrm{T}} D_{(h)} x + B^{-1})^{-1}(x^{\mathrm{T}}\kappa_h),\, (x^{\mathrm{T}} D_{(h)} x + B^{-1})^{-1}\},$$

where κ_h is a p-dimensional vector with the j-th entry equal to $\phi_{jh}^{(L)} - 0.5$.

Step S4. Update the elements λ_{jh}^* by sampling from the independent full conditional posterior distributions of the row vectors $\lambda_j^* = (\lambda_{j1}^*,\dots,\lambda_{jH}^*)$, for $j = 1,\dots,p$,

$$(\lambda_j^* \mid -) \sim N_H\big\{(D^{-1} + \sigma_j^{-2}\eta_{(j)}^{\mathrm{T}}\eta_{(j)})^{-1}\sigma_j^{-2}\eta_{(j)}^{\mathrm{T}} y^{(j)},\, (D^{-1} + \sigma_j^{-2}\eta_{(j)}^{\mathrm{T}}\eta_{(j)})^{-1}\big\},$$

where $\eta_{(j)}$ is the $n \times H$ matrix such that the generic element is $\eta_{(j)ih} = \eta_{ih}\sqrt{\rho_h}\sqrt{\phi_{jh}}$, $D^{-1} = \text{diag}(\vartheta_1^{-1},\dots,\vartheta_H^{-1})$ and $y^{(j)} = (y_{1j},\dots,y_{nj})^{\mathrm{T}}$. Set $\lambda_{jh} = \lambda_{jh}^*\sqrt{\rho_h}\sqrt{\phi_{jh}}$.

Step S5. Update the column scales γ_h (for $h = 1,\dots,H$), following the substeps below and setting $\gamma_h = \vartheta_h\rho_h$. Consistently with Legramanti et al. (2020), define the independent indicators z_h $(h = 1,\dots,p)$ with prior $\text{pr}(z_h = l) = w_l$.

Substep S5.1. Update the augmented data z_h by sampling from the full conditional distribution

$$\text{pr}(z_h = l) \propto \begin{cases} w_l \prod_{i=1}^n \prod_{j=1}^p N(y_{ij}; \mu_{ijh}^{(0)}, \sigma_j^2) & \text{for} \quad l = 1,\dots,h \\ w_l \prod_{i=1}^n \prod_{j=1}^p N(y_{ij}; \mu_{ijh}^{(1)}, \sigma_j^2) & \text{for} \quad l = h+1,\dots,H, \end{cases} \tag{S1}$$

where $N(x;\mu,\sigma^2)$ indicates the Gaussian probability density function with mean μ and variance σ^2. The mean values $\mu_{ijh}^{(0)}$ and $\mu_{ijh}^{(1)}$ are defined according to $\mu_{ijh}^{(z)} = \sum_{l\neq h}^H \sqrt{\rho_l}\,\sqrt{\phi_{jl}}\lambda_{jl}^*\eta_{il} + \sqrt{z}\,\sqrt{\phi_{jh}}\lambda_{jh}^*\eta_{ih}$. Set $\rho_h = 1$ if $z_h > h$, else $\rho_h = 0$.

Substep S5.2. For $h = 1,\dots,H$, update ϑ_h^{-1} sampling from $\text{Ga}(a_\theta + 0.5p, b_\theta + 0.5\sum_{j=1}^p \lambda_{jh}^{*2})$.

Substep S5.3. For $l = 1,\dots,H-1$, sample v_l from

$$(v_l \mid -) \sim \text{Be}\big\{1 + \sum_{h=1}^H \mathbb{1}(z_h = l), \alpha + \mathbb{1}(z_h > l)\big\},$$

set $v_H = 1$ and update $w_l = v_l \prod_{m=1}^{l-1}(1 - v_m)$, for $l = 1,\dots,H$.

Table S1: *Median and interquartile range of the LPML, Cov. MSE and of $E(H_a \mid y)$ computed in 25 replications assuming Scenario b and several combinations of (p, k, s)*

	(p, k, s)	MGP		CUSP		SIS	
		$Q_{0.5}$	IQR	$Q_{0.5}$	IQR	$Q_{0.5}$	IQR
LPML	(16,4,0.6)	-28.20	0.33	-28.20	0.33	-28.17	0.32
	(32,8,0.4)	-56.95	0.53	-57.00	0.51	-56.80	0.49
	(64,12,0.3)	-111.35	0.70	-111.71	0.74	-110.76	0.89
	(128,16,0.2)	-211.65	0.74	-215.94	1.57	-210.19	0.86
Cov. MSE	(16,4,0.6)	0.25	0.12	0.25	0.12	0.23	0.10
	(32,8,0.4)	0.32	0.08	0.33	0.10	0.30	0.12
	(64,12,0.3)	0.37	0.10	0.43	0.11	0.22	0.09
	(128,16,0.2)	0.23	0.03	0.32	0.04	0.09	0.01
$E(H_a \mid y)$	(16,4,0.6)	8.91	1.52	4.00	0.00	4.00	0.00
	(32,8,0.4)	11.27	1.48	7.00	1.00	8.00	0.00
	(64,12,0.3)	14.72	1.49	11 .00	0.00	12.00	0.00
	(128,16,0.2)	17.16	0.81	12.00	1.75	16.00	0.00

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

Step S6. Update independently the local scales, for $j = 1, \ldots, p$ and $h = 1, \ldots, H$, by sampling from the full conditional distributions

$$\mathrm{pr}(\phi_{jh} = u) \propto \begin{cases} \{1 - \mathrm{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)\, c_p\} \prod_{i=1}^{n} N(y_{ij}; \mu_{ijh}^{(u)}, \sigma_j^2) & \text{for } u = 0 \\ \mathrm{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)\, c_p \prod_{i=1}^{n} N(y_{ij}; \mu_{ijh}^{(u)}, \sigma_j^2) & \text{for } u = 1. \end{cases}$$

with $\mu_{ijh}^{(u)} = \sum_{l \neq h}^{H} \sqrt{\rho_l}\, \sqrt{\phi_{jl}} \lambda_{jl}^{*} \eta_{il} + \sqrt{\rho_h}\, \sqrt{u} \lambda_{jh}^{*} \eta_{ih}$.

S2.2. *Simulation settings*

The results reported in Section 4 are obtained running the algorithms for 25000 iterations discarding the first 10000 iterations. Then, we thin the Markov chain, saving every 5-th sample. We adapt the number of active factors at iteration t with probability $p(t) = \exp(-1 - 5\, 10^{-4} t)$. We set $a_\sigma = 1$ and $b_\sigma = 0.3$. In the structured increasing shrinkage algorithm, we choose the offset constant $c_p = 2e \log(p)/p$ which belongs to $(0, 1)$ for every $p \geq 15$.

In scenario d, the meta covariates in matrix x_0 are a categorical variable with four balanced categories, a continuous variable sampled from a multivariate Gaussian distribution, and a continuous variable where the p elements are sampled from p gamma distributions.

To infer the structural zeros within each column of Λ in the cumulative shrinkage process and in the multiplicative gamma process, we set λ_{jh} to zero when $|\lambda_{jh}|$ $(j = 1, \ldots, p)$ is under a certain threshold. We choose the threshold equal to 0.05, which is consistent with the value of the hyperparameter θ_∞ used in the cumulative shrinkage process.

To address column order ambiguity and label switching, we compute the mean classification error only after having ordered the columns of $\Lambda^{(t)}$ (for $t = 1, \ldots, S$), for each model, increasingly with respect to the number of zero entries identified.

S2.3. *Simulation results*

We report additional results for the simulation study of Section 4 of the main paper.

In scenario b, we also apply the method proposed by Ročková & George (2016), which is referred to as parameter expanded likelihood expected maximization. Hyperparameters are set

Table S2: *Median and interquartile range of the Cov. MSE and CE computed applying the parameter expanded likelihood expected maximization method of Ročková & George (2016) in 25 replications under Scenario b and several combinations of* (p, k, s)

(p, k, s)	Cov. MSE		CE	
	$\mathrm{Q}_{0.5}$	IQR	$\mathrm{Q}_{0.5}$	IQR
(16,4,0.6)	0.55	0.16	0.77	0.14
(32,8,0.4)	0.66	0.11	0.76	0.10
(64,12,0.3)	0.64	0.11	0.90	0.10
(128,16,0.2)	0.42	0.03	1.06	0.20

Cov. MSE, covariance mean squared error; CE, classification error; PXLEM, parameter expanded likelihood expected maximization; $\mathrm{Q}_{0.5}$, median; IQR, interquartile range.

Table S3: *Median and interquartile range of the LPML, Cov. MSE,* $E(H_a \mid y)$ *and MCE computed in 25 replications assuming Scenario c and several combinations of* (p, k, s)

	(p, k, s)	MGP		CUSP		SIS	
		$\mathrm{Q}_{0.5}$	IQR	$\mathrm{Q}_{0.5}$	IQR	$\mathrm{Q}_{0.5}$	IQR
LPML	(16,4,0.6)	-27.62	0.24	-27.62	0.25	-27.59	0.24
	(32,8,0.4)	-56.16	0.64	-56.22	0.51	-55.89	0.59
	(64,12,0.3)	-109.64	0.69	-110.67	0.71	-109.06	0.65
	(128,16,0.2)	-209.57	0.88	-214.19	1.76	-208.34	1.04
Cov. MSE	(16,4,0.6)	0.30	0.10	0.29	0.09	0.26	0.11
	(32,8,0.4)	0.77	0.26	0.72	0.18	0.80	0.43
	(64,12,0.3)	1.01	0.35	0.94	0.21	1.20	1.22
	(128,16,0.2)	0.78	0.18	0.87	0.21	0.35	0.48
$E(H_a \mid y)$	(16,4,0.6)	8.38	1.80	3.44	1.00	4.00	0.00
	(32,8,0.4)	10.38	1.12	5.05	0.91	8.00	1.00
	(64,12,0.3)	13.67	1.20	8.00	0.92	12.00	0.00
	(128,16,0.2)	16.56	0.83	9.00	0.00	16.00	0.00
MCE	(16,4,0.6)	0.98	0.17	0.53	0.20	0.24	0.06
	(32,8,0.4)	0.65	0.07	0.44	0.08	0.19	0.07
	(64,12,0.3)	0.59	0.04	0.48	0.04	0.18	0.06
	(128,16,0.2)	0.48	0.02	0.44	0.01	0.06	0.10

LPML, logarithm of the pseudo-marginal likelihood; Cov. MSE, covariance mean squared error; MCE, mean classification error; CUSP, cumulative increasing shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; $\mathrm{Q}_{0.5}$, median; IQR, interquartile range.

as suggested by the authors. This approach focuses on finding a sparse mode based on an overparameterized factor model. The performance in terms of mean squared error in covariance estimation and classification error in detecting sparsity in Λ is reported in Table S2. The results are not competitive with the other approaches we have considered.

Table S4: *Median and interquartile range of the LPML, Cov. MSE, $E(H_a \mid y)$ and MCE computed in 25 replications assuming Scenario d and several combinations of* (p, k, s)

	(p, k, s)	MGP		CUSP		SIS		SIS$_{mc}$	
		Q$_{0.5}$	IQR	Q$_{0.5}$	IQR	Q$_{0\cdot5}$	IQR	Q$_{0.5}$	IQR
LPML	(16,4,0.6)	-27.74	0.43	-27.75	0.43	-27.71	0.40	-27.73	0.40
	(32,8,0.4)	-56.25	0.69	-56.35	0.72	-56.16	0.68	-56.12	0.65
	(64,12,0.3)	-109.72	0.61	-110.54	0.88	-109.27	0.46	-109.16	0.57
	(128,16,0.2)	-209.60	0.48	-213.50	1.21	-208.11	0.42	-208.03	0.47
Cov. MSE	(16,4,0.6)	0.31	0.11	0.30	0.14	0.28	0.14	0.27	0.16
	(32,8,0.4)	0.70	0.25	0.71	0.18	0.75	0.22	0.79	0.78
	(64,12,0.3)	1.03	0.43	0.91	0.29	1.51	0.59	1.16	0.84
	(128,16,0.2)	0.93	0.49	0.90	0.33	1.49	1.21	1.28	1.81
$E(H_a \mid y)$	(16,4,0.6)	8.60	0.64	3.96	0.80	4.00	0.00	4.00	0.00
	(32,8,0.4)	10.71	1.24	5.75	1.00	7.00	1.00	8.00	0.00
	(64,12,0.3)	13.93	1.37	8.00	0.92	12.00	0.00	12.00	0.00
	(128,16,0.2)	16.56	0.88	9.00	0.00	16.00	1.00	16.00	1.00
MCE	(16,4,0.6)	0.94	0.13	0.64	0.19	0.26	0.08	0.23	0.13
	(32,8,0.4)	0.67	0.10	0.49	0.09	0.20	0.08	0.20	0.10
	(64,12,0.3)	0.58	0.05	0.47	0.04	0.21	0.06	0.21	0.08
	(128,16,0.2)	0.49	0.02	0.43	0.02	0.18	0.11	0.17	0.11

LPML, logarithm of the pseudo-marginal likelihood; Cov. MSE, covariance mean squared error; MCE, mean classification error; CUSP, cumulative increasing shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; SIS$_{mc}$, structured increasing shrinkage process with meta covariates; Q$_{0.5}$, median; IQR, interquartile range.

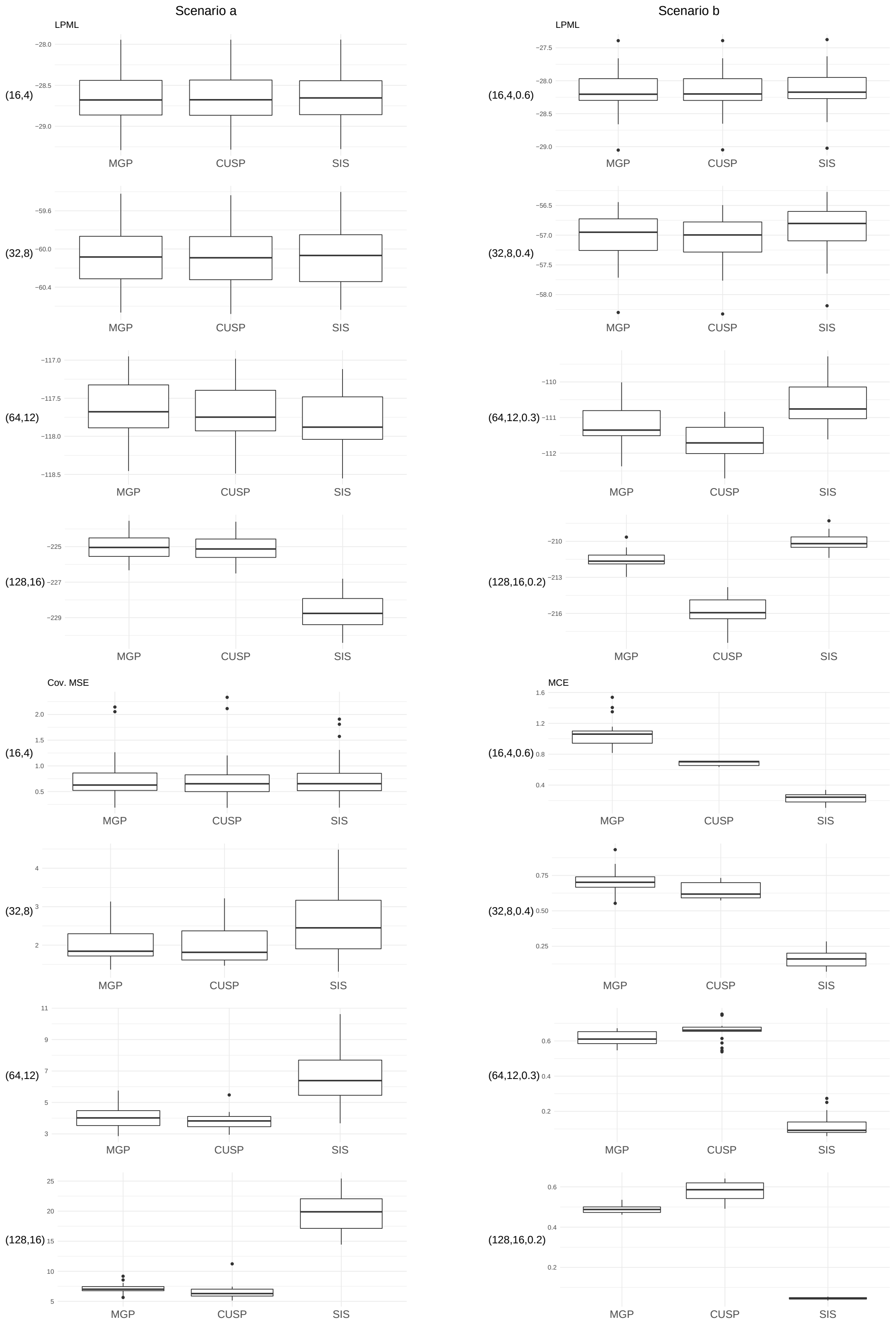

Fig. S1: Boxplots of the logarithm of the pseudo-marginal likelihood for all combinations (p, k) in scenario a (top left panel) and scenario b (top right panel), of the covariance mean square error in scenario a (bottom left panel), and of the mean classification error in scenario b (bottom right panel). LPML, logarithm of the pseudo-marginal likelihood; Cov. MSE, covariance mean squared error; MCE, mean classification error; CUSP, cumulative increasing shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process.

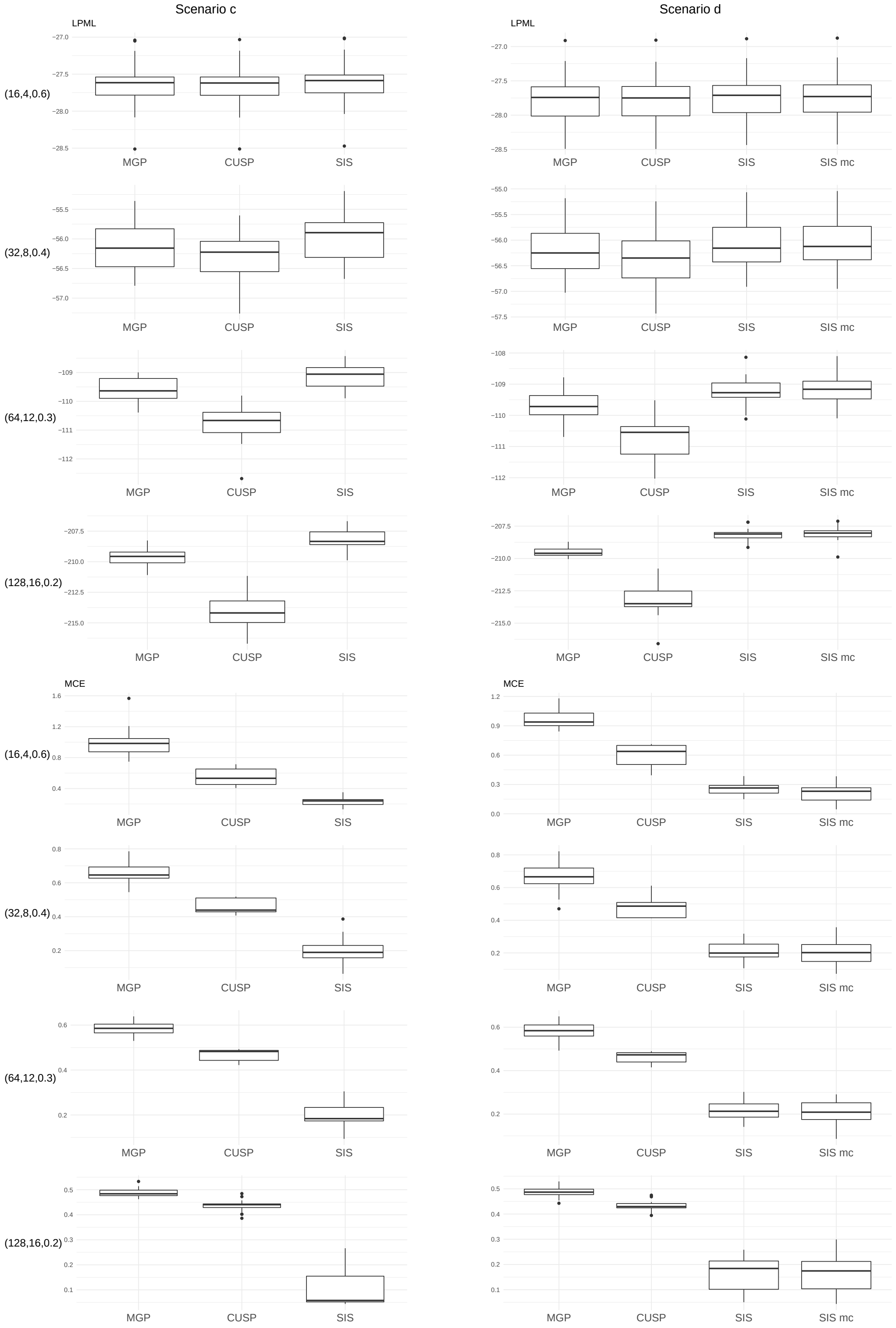

Fig. S2: Boxplots of the logarithm of the pseudo-marginal likelihood and of the mean classification error of each model for all combinations of (p, k, s) in Scenario c (left panel) and Scenario d (right panel). LPML, logarithm of the pseudo-marginal likelihood; MCE, mean classification error; CUSP, cumulative shrinkage process; MGP, multiplicative gamma process; SIS, structured increasing shrinkage process; SIS mc, structured increasing shrinkage process with meta covariates.

S2.4. *Simulation study of sensitivity to hyperparameters and truncation level*

We conduct further simulation experiments to assess the impact of some hyperparameters on relevant prior and posterior summaries. Figure S3 displays the prior distribution, obtained simulating 10,000 samples from the prior, of the proportion of variance explained by the structured increasing shrinkage factor model for varying α, $\{E(\sigma^{-2}), \mathrm{var}(\sigma^{-2})\}$, and $\{E(\vartheta_h^{-1}), \mathrm{var}(\vartheta_h^{-1})\}$. The hyperparameter α, representing the expected number of factors, positively affects the proportion of explained variance. The influence of the hyperparameters regulating the distribution of ϑ_h is even clearer, with concentrated prior on a large value of $E(\vartheta^{-1})$, inducing a smaller proportion of variance explained by the factor model. The role of $\{E(\sigma^{-2}), \mathrm{var}(\sigma^{-2})\}$ is less clear, but suggests that sufficiently large mean and variance can guarantee higher flexibility.

The latter comment is confirmed by the study of the impact of α and $\{E(\sigma^{-2}), \mathrm{var}(\sigma^{-2})\}$ on the posterior bound of the truncation error and on the posterior distribution of the proportion of variance explained by the factor model. Specifically, we generate synthetic data sets with $n = 100$ observations with dimension $p = 50$ from the Gaussian linear factor model $y_i \sim N_p(0, \Lambda_0\Lambda_0^T + I_p)$, with Λ_0 a sparse $p \times k$ matrix with $k = 50$. We randomly set two thirds of the elements of Λ_0 equal to zero, drawing the non zero elements from a Gaussian distribution with mean zero and variances θ_h sampled from an inverse gamma distribution $\theta_h^{-1} \sim \mathrm{Ga}(2, 2)$. We keep the number of active factors H fixed at 50, and set $a_\theta = b_\theta = 2$ and $c_p = 2e\log(p)/p$. We run the Gibbs algorithm for the structured increasing shrinkage model for 15000 iterations, discarding the first 5000 iterations. Then, we thin the Markov chain, saving every 5-th sample.

In Figure S4 the sampled posterior distribution of the proportion of variance explained by the factor model $\mathrm{tr}(\Lambda\Lambda^{\mathrm{T}})/\mathrm{tr}(\Omega)$ is reported for varying α and $\{E(\sigma^{-2}), \mathrm{var}(\sigma^{-2})\}$. The same proportion computed on the matrices generating the data is $\mathrm{tr}(\Lambda_0\Lambda_0^{\mathrm{T}})/\mathrm{tr}(\Omega_0) = 0.966$, with $\Omega_0 = \Lambda_0\Lambda_0^{\mathrm{T}} + I_{50}$. A sufficiently concentrated prior on a large value of $E(\sigma^{-2})$ seems more suitable to model such data, even if we have incorrect expectations on the number of factors, i.e. α set small.

Figure S5 displays the Monte Carlo approximation of the posterior probability of truncation error $\mathrm{pr}\{\mathrm{tr}(\Omega_H)/\mathrm{tr}(\Omega) < T\}$ for different values of H and T and varying α and $\{E(\sigma^{-2}), \mathrm{var}(\sigma^{-2})\}$. If Λ_0 is sparse, a small value of α induces good approximations even with H smaller than the true number of factors. The inferred sparsity pattern in Λ is robust to the prior distribution for σ^{-2}.

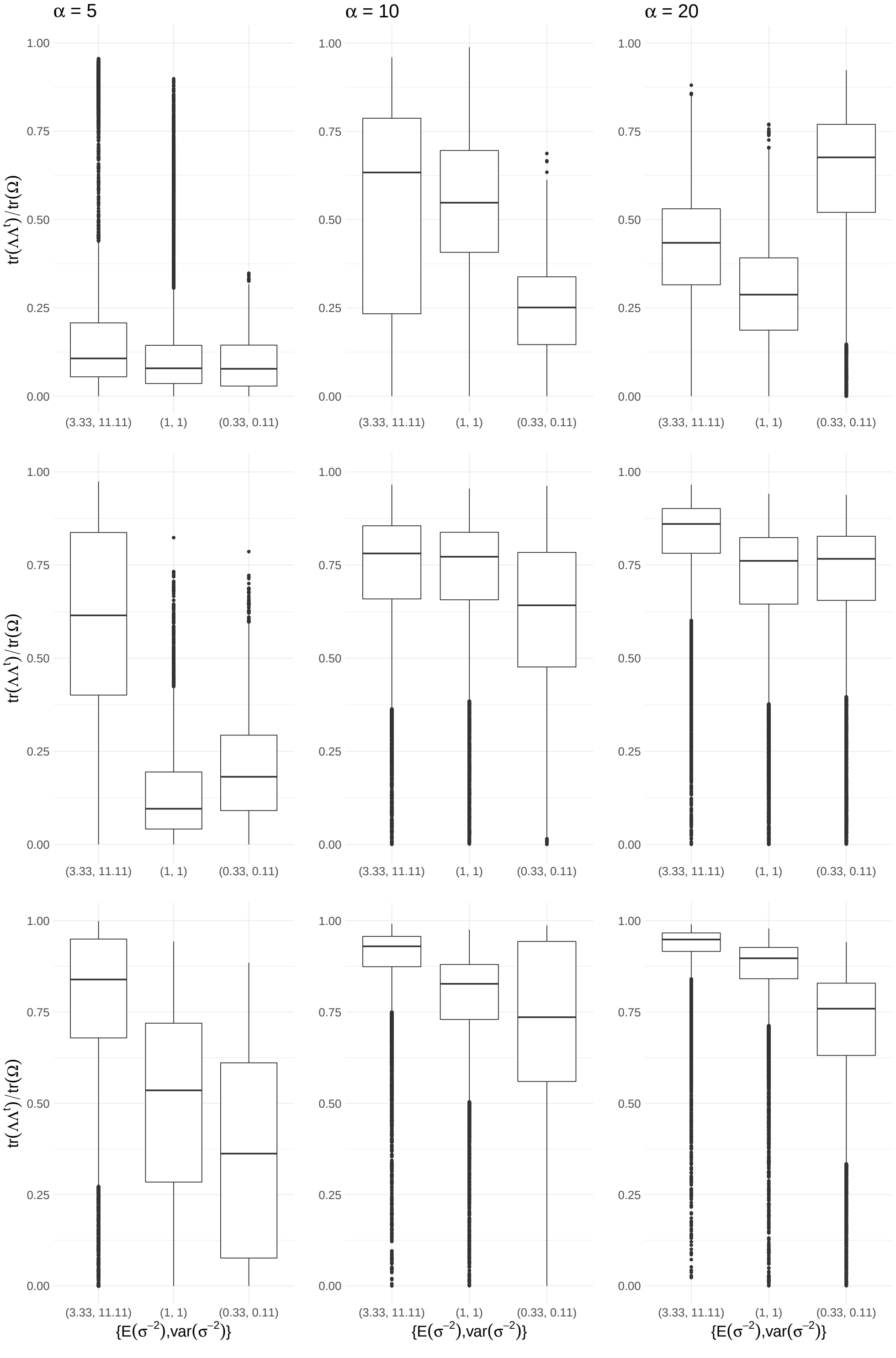

Fig. S3: Boxplots of the prior proportion of variance explained by the factor model $\text{tr}(\Lambda\Lambda^{\text{T}})/\text{tr}(\Omega)$. The quantity is obtained simulating 10,000 samples from the prior distribution with varying values of the parameters. The horizontal axis characterize the effect of $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\}$; differences for $\alpha \in (5, 10, 20)$ are reported in each column; differences for $\{E(\vartheta^{-1}), \text{var}(\vartheta^{-1})\} \in \{(2, 2), (1, 0.5), (0.5, 0.125)\}$ are reported in each row.

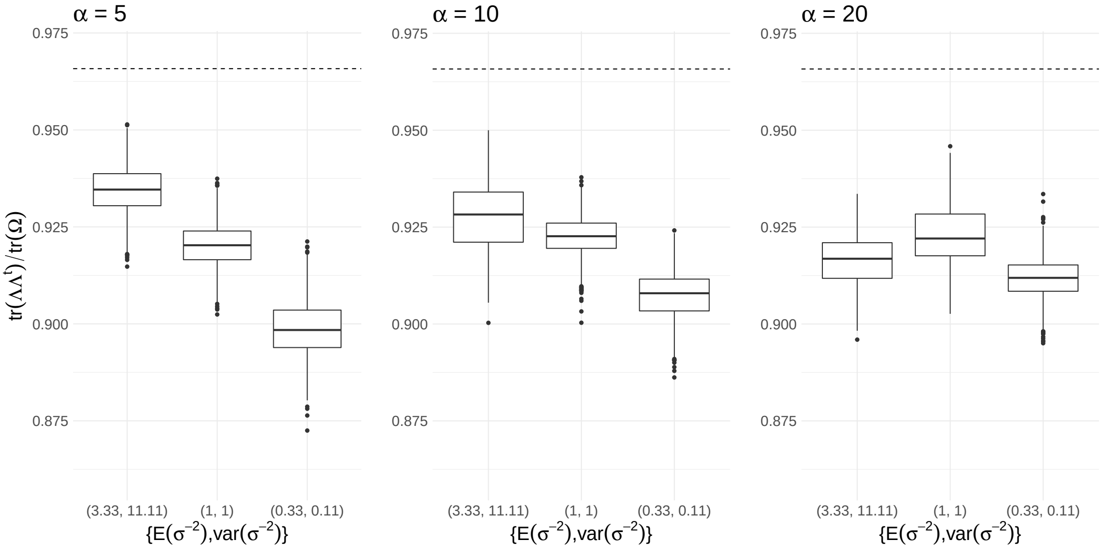

Fig. S4: Boxplots representing the simulated posterior distribution of the proportion of variance explained by the factor model tr$(\Lambda\Lambda^{\mathrm{T}})/tr(\Omega)$ for varying α and $\{E(\sigma^{-2}),\, \text{var}(\sigma^{-2})\}$. The dashed lines represent the proportion computed on the true value of Λ and Ω.

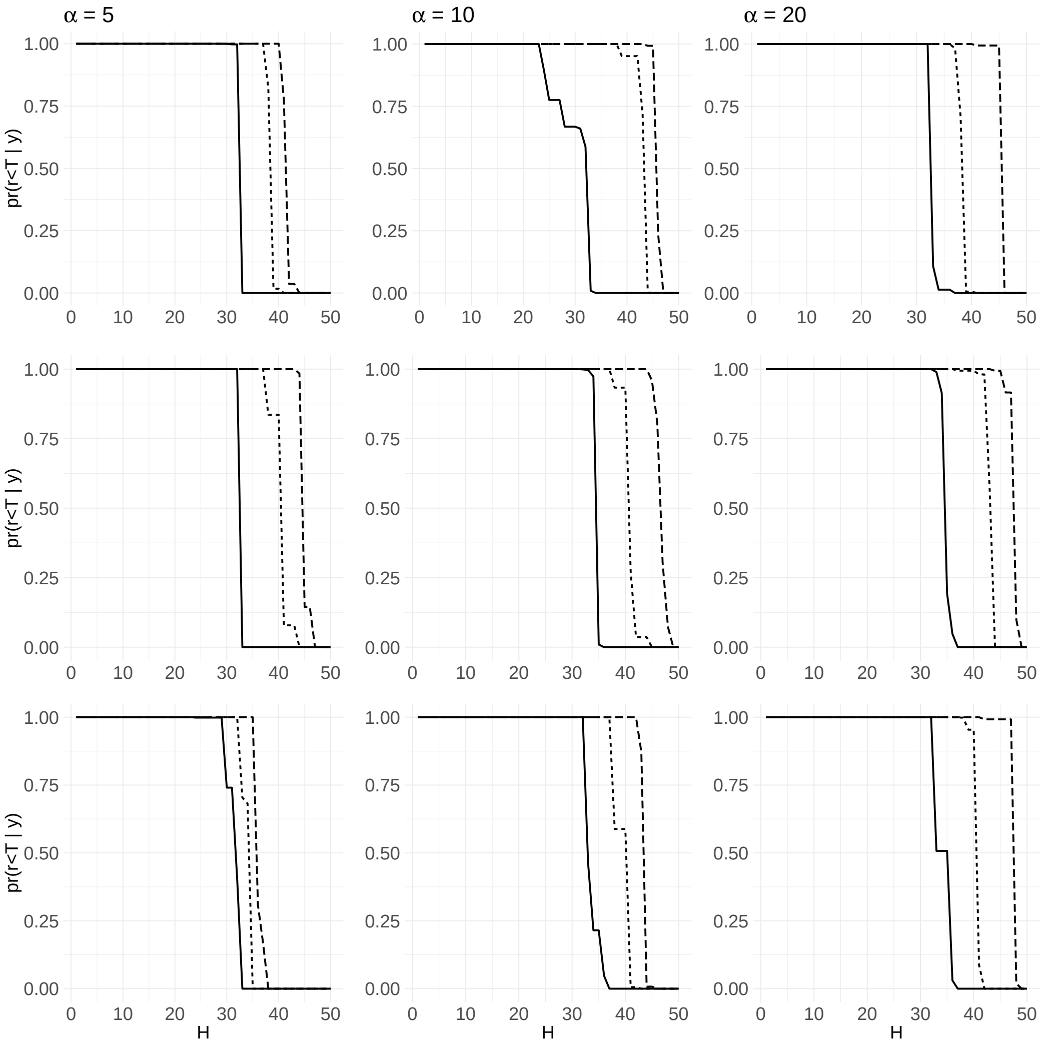

Fig. S5: Monte Carlo approximation of the posterior probability of truncation error pr$(r < T \mid y)$, with $r =$ tr$(\Omega_H)/$tr(Ω), at varying of H. The quantity is computed for T equal to 0.75 (—), 0.9 (- - -), and 0.95 (– – –) and varying $\alpha \in (5, 10, 20)$ over the columns and $\{E(\sigma^{-2}), \text{var}(\sigma^{-2})\} \in \{(3.33, 11.11), (1, 1), (0.33, 0.11)\}$ over the rows of the figure.

S3. FINNISH BIRD CO-OCCURRENCE APPLICATION

230 S3.1. *Gibbs algorithms of probit structured increasing shrinkage model*

In case of probit data (see Section 5 of the main paper) and the structured increasing shrinkage process, we can rewrite the latent model for z_{ij} as

$$z_{ij} = w_i^T \mu_j + \epsilon_{ij},$$

$$\epsilon_{ij} = \sum_{h=1}^{\infty} \sqrt{\rho_h} \sqrt{\phi_{jh}}\, \lambda^*_{jh} \eta_{ih} + \varepsilon_{ij}, \qquad \lambda^*_{jh} \sim N(0, \vartheta_h), \qquad \varepsilon_{ij} \sim N(0, 1),$$

where λ^*_{jh} is an absolutely continuous random variable. Let the notation $(x \mid -)$ denote the full conditional distribution of x conditionally on everything else. Given H the number of factors of the truncated model, the sampler cycles through the following steps.

Step S1. Update μ_j, for every $j = 1, \ldots, p$, by sampling from the independent full conditional posterior distributions

$$(\mu_j \mid -) \sim N_c\big[(\sigma_\mu^{-2} I_c + w^{\mathrm{T}} w)^{-1}\{w^{\mathrm{T}}(z^{(j)} - \eta\lambda_j) + b x_j\}, (\sigma_\mu^{-2} I_c + w^{\mathrm{T}} w)^{-1}\big],$$

where $z^{(j)} = (z_{1j}, \ldots, z_{nj})^{\mathrm{T}}$ and $\eta = (\eta_1, \ldots, \eta_n)^{\mathrm{T}}$.

Step S2. Update b_l $(l = 1, \ldots, c)$ sampling from conditionally independent posteriors

$$(b_l \mid -) \sim N_q\big\{(\sigma_b^{-2} I_q + \sigma_\mu^{-2} x^{\mathrm{T}} x)^{-1} \sigma_\mu^{-2}(x^{\mathrm{T}} \mu^{(l)}),\, (\sigma_b^{-2} I_q + \sigma_\mu^{-2} x^{\mathrm{T}} x)^{-1}\big\},$$

where $\mu^{(l)} = (\mu_{1l}, \ldots, \mu_{pl})^{\mathrm{T}}$

Step S3. Update the elements z_{ij} $(i = 1, \ldots, n;\ j = 1 \ldots, p)$ sampling independently from the truncated normal

$$(z_{ij} \mid -) \sim TN(\lambda_j^{\mathrm{T}} \eta_i + w_i^T \mu_j, 1, l_{ij}, u_{ij}),$$

where the lower bound l_{ij} is equal to 0 if $y_{ij} = 1$ and $-\infty$ otherwise. The upper bound $u_{ij} = 0$ if $y_{ij} = 0$ and ∞ otherwise. Then, set $\epsilon = z - w\mu$.

Step S4. Update, for $i = 1, \ldots, n$, the factor η_i according to the posterior full conditional

$$(\eta_i \mid -) \sim N_H\big\{(I_H + \Lambda_H^{\mathrm{T}} \Lambda_H)^{-1} \Lambda_H^{\mathrm{T}} \epsilon_i,\, (I_H + \Lambda_H^{\mathrm{T}} \Lambda_H)^{-1}\big\}.$$

Step S5. Update β_h $(h = 1, \ldots, H)$ exploiting the Pólya-Gamma data-augmentation strategy (Polson et al., 2013) and the decompostition $\phi_{jh} = \phi_{jh}^{(L)} \phi_{jh}^{(C)}$, with $\phi_{jh}^{(L)} \phi_{jh}^{(C)}$ independent a priori and distributed as Ber$\{\text{logit}^{-1}(x_j^{\mathrm{T}} \beta_h)\}$ and Ber(c_p), respectively.

Substep S5.1. Update $\phi_{jh}^{(L)}$, for $j = 1, \ldots, p$ and $h = 1, \ldots, H$, setting $\phi_{jh}^{(L)} = 1$ if $\phi_{jh} = 1$ and sampling from the full conditional distribution

$$\text{pr}(\phi_{jh}^{(L)} = l) \propto \begin{cases} 1 - \text{logit}^{-1}(x_j^{\mathrm{T}} \beta_h) & \text{for } l = 0, \\ \text{logit}^{-1}(x_j^{\mathrm{T}} \beta_h)(1 - c_p) & \text{for } l = 1, \end{cases}$$

if $\phi_{jh} = 0$.

Substep S5.2. Let $f(y) \propto \sum_{n=0}^{\infty} (-1)^n A_n (2\pi y^3)^{-0.5} \exp\{-(2n + b)^2 (8y)^{-1} - 0.5 c^2 y\}$ indicate the probability density function of a Pólya-Gamma distributed random variable $y \sim$ PG(b, c). For each $h = 1, \ldots, H$, generate p independent random variables $d_{j(h)}$ sampling from the full conditional distribution $(d_{j(h)} \mid -) \sim$ PG$(1, x_j^{\mathrm{T}} \beta_h)$. Let $D_{(h)}$ denote the $p \times p$ diagonal matrix with entries $d_{j(h)}$ $(j = 1, \ldots, p)$.

Substep S5.3. Define the $q \times q$ diagonal matrix $B = \sigma_\beta^2 I_q$. For each $h = 1, \ldots, H$, update β_h sampling from

$$(\beta_h \mid -) \sim N_q\{(x^{\mathrm{T}} D_{(h)} x + B^{-1})^{-1}(x^{\mathrm{T}} \kappa_h),\, (x^{\mathrm{T}} D_{(h)} x + B^{-1})^{-1}\},$$

where κ_h is the p-dimensional vector with the j-th entry equal to $\phi_{jh}^{(L)} - 0.5$.

Step S6. Update the elements λ^*_{jh} by sampling from the independent full conditional posterior distributions of the rows vector $\lambda^*_j = (\lambda^*_{j1}, \dots, \lambda^*_{jH})$, for $j = 1, \dots, p$,

$$(\lambda^*_j \mid -) \sim N_H\big\{(D^{-1} + \eta^{\mathrm{T}}_{(j)}\eta_{(j)})^{-1}\eta^{\mathrm{T}}_{(j)}\epsilon^{(j)},\ (D^{-1} + \eta^{\mathrm{T}}_{(j)}\eta_{(j)})^{-1}\big\},$$

where $\eta_{(j)}$ is the $n \times H$ matrix such that the generic element is $\eta_{(j)ih} = \eta_{ih}\sqrt{\rho_h}\sqrt{\phi_{jh}}$, $D^{-1} = \mathrm{diag}(\vartheta_1^{-1}, \dots, \vartheta_H^{-1})$ and $\epsilon^{(j)} = (\epsilon_{1j}, \dots, \epsilon_{nj})^{\mathrm{T}}$. Set $\lambda_{jh} = \lambda^*_{jh}\sqrt{\rho_h}\sqrt{\phi_{jh}}$.

Step S7. Update the column scales γ_h (for $h = 1, \dots, H$), following the substeps below and setting $\gamma_h = \vartheta_h\rho_h$. Consistently with Legramanti et al. (2020), define the independent indicators u_h ($h = 1, \dots, p$) with prior $\mathrm{pr}(u_h = l) = w_l$.

Substep S7.1. Update the augmented data u_h by sequentially sampling from the full conditional distribution

$$\mathrm{pr}(u_h = l) \propto \begin{cases} w_l \prod_{i=1}^n \prod_{j=1}^p N(\epsilon_{ij}; \mu^{(0)}_{ijh}, \sigma^2_j) & \text{for} \quad l = 1, \dots, h \\ w_l \prod_{i=1}^n \prod_{j=1}^p N(\epsilon_{ij}; \mu^{(1)}_{ijh}, \sigma^2_j) & \text{for} \quad l = h+1, \dots, H. \end{cases} \tag{S2}$$

The mean values $\mu^{(0)}_{ijh}$ and $\mu^{(1)}_{ijh}$ are defined according to $\mu^{(u)}_{ijh} = \sum_{l \neq h}^H \sqrt{\rho_l}\,\sqrt{\phi_{jl}}\lambda^*_{jl}\eta_{il} + \sqrt{u}\,\sqrt{\phi_{jh}}\lambda^*_{jh}\eta_{ih}$. Set $\rho_h = 1$ if $u_h > h$, else $\rho_h = 0$.

Substep S7.2. For $h = 1, \dots, H$, update ϑ_h^{-1} sampling from $\mathrm{Ga}(a_\theta + 0.5p, b_\theta + 0.5\sum_{j=1}^p \lambda^{*2}_{jh})$.

Substep S7.3. For $l = 1, \dots, H-1$, sample v_l from

$$(v_l \mid -) \sim \mathrm{Be}\Big\{1 + \sum_{h=1}^H \mathbb{1}(u_h = l), \alpha + \mathbb{1}(u_h > l)\Big\},$$

set $v_H = 1$ and update $w_l = v_l \prod_{m=1}^{l-1}(1 - v_m)$, for $l = 1, \dots, H$.

Step S8. Update the local scales, independently for $j = 1, \dots, p$ and sequentially over $h = 1, \dots, H$, by sampling from the full conditional distributions

$$\mathrm{pr}(\phi_{jh} = u) \propto \begin{cases} \{1 - \mathrm{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)\,c_p\} \prod_{i=1}^n N(\epsilon_{ij}; \mu^{(u)}_{ijh}, 1) & \text{for } u = 0 \\ \mathrm{logit}^{-1}(x_j^{\mathrm{T}}\beta_h)\,c_p \prod_{i=1}^n N(\epsilon_{ij}; \mu^{(u)}_{ijh}, 1) & \text{for } u = 1. \end{cases}$$

with $\mu^{(u)}_{ijh} = \sum_{l \neq h}^H \sqrt{\rho_l}\,\sqrt{\phi_{jl}}\lambda^*_{jl}\eta_{il} + \sqrt{\rho_h}\,\sqrt{u}\lambda^*_{jh}\eta_{ih}$.

The results reported in Section 5 are obtained running the algorithm for 40000 iterations discarding the first 20000 iterations. Then, we thin the Markov Chain, saving every 5-th sample. We adapt the number of active factors at iteration t with probability $p(t) = \exp(-1 - 2.5\,10^{(-4)}t)$ and, given the high value of p considered, we choose the offset constant $c_p = 2e\log(p)/p$ which belongs to $(0, 1)$ for every $p \geq 15$.

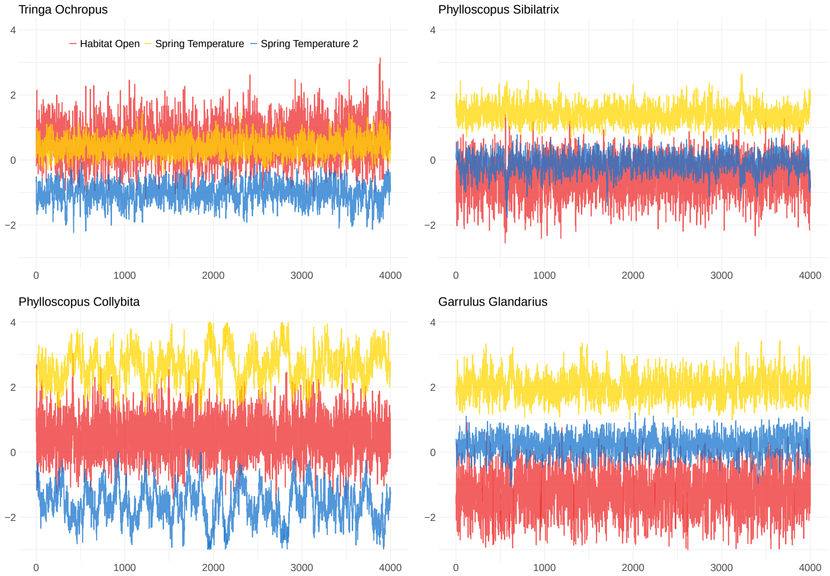

Fig. S6: Chain plots of the marginal posterior samples of 12 mean coefficients of the matrix μ obtained by the Gibbs sampler, discarding the first 20000 iterations and saving every 5-th sample.

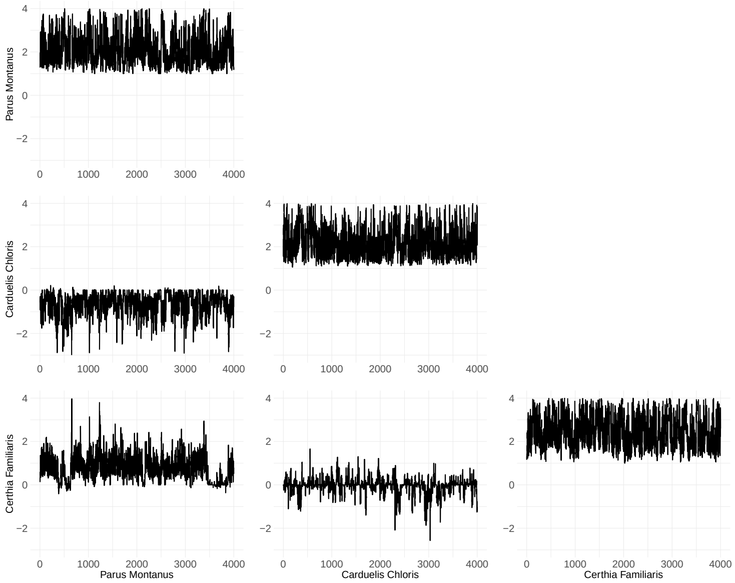

Fig. S7: Chain plots of the marginal posterior samples of six elements of the covariance matrix obtained by the Gibbs sampler, discarding the first 20000 iterations and saving every 5-th sample.

S3.2. *Gibbs chains mixing*

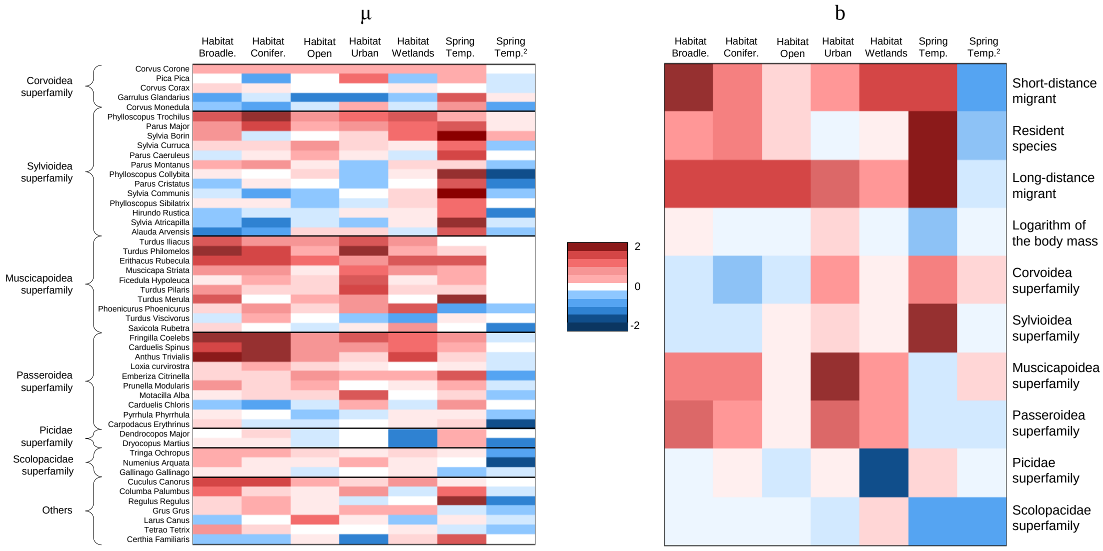

Fig. S8: Posterior mean of μ and b for the structured increasing shrinkage model; rows of left matrix refer to the 50 birds species, and rows of right matrix to the ten species traits. Broadle: broadleaved forests; Conifer: coniferous forests; Temp: temperature.

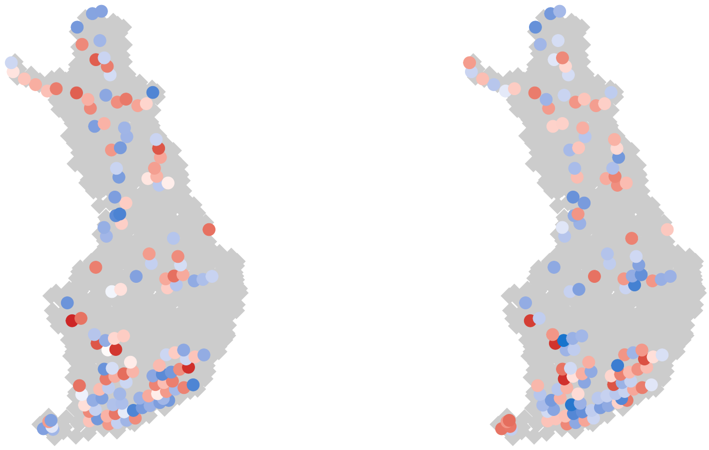

Fig. S9: Maps of the sampling units in Finland coloured accordingly to the values of the first and the third latent factors sampled at iteration t^*. Red and blue spots represent the environments with positive and negative values of the factors, respectively.

S3.3. *Figures*

References

Legramanti, S., Durante, D. & Dunson, D. B. (2020). Bayesian cumulative shrinkage for infinite factorizations. *Biometrika* **107**, 745–752.

Polson, N. G., Scott, J. G. & Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association* **108**, 1339–1349.

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