Bayesian Analysis (2023) **18**, Number 2, pp. 465–490

A Bayesian Approach for Partial Gaussian Graphical Models With Sparsity*

Eunice Okome Obiang†, Pascal Jézéquel‡ and Frédéric Proïa§

Abstract. We explore various Bayesian approaches to estimate partial Gaussian graphical models. Our hierarchical structures enable to deal with single-output as well as multiple-output linear regressions, in small or high dimension, enforcing either no sparsity, sparsity, group sparsity or even sparse-group sparsity for a bi-level selection through partial correlations (direct links) between predictors and responses, thanks to spike-and-slab priors corresponding to each setting. Adaptative and global shrinkages are also incorporated in the Bayesian modeling of the direct links. An existing result for model selection consistency is reformulated to stick to our sparse and group-sparse settings, providing a theoretical guarantee under some technical assumptions. Gibbs samplers are developed and a simulation study shows the efficiency of our models which give very competitive results, especially in terms of support recovery. To conclude, a real dataset is investigated.

MSC2020 subject classifications: Primary 62A09, 62F15; secondary 62J05.

Keywords: high-dimensional linear regression, partial graphical model, partial correlation, Bayesian approach, sparsity, spike-and-slab, Gibbs sampler.

1 Introduction and motivations

This paper is devoted to the Bayesian estimation of the partial Gaussian graphical models. Graphical models are now widespread in many contexts, like image analysis, economics or biological regulation networks, neural models, etc. A graphical model for the d-dimensional Gaussian vector $Z \sim \mathcal{N}_d(\mu, \Sigma)$ is a model where the conditional dependencies between the coordinates of Z are represented by means of a graph. We refer the reader to the handbook recently edited by Maathuis et al. (2018) for a very complete survey of graphical models theory, or to Chap. 7 of Giraud (2014) for a wide introduction to the subject. It is well-known that the partial correlation between Z_i and Z_j satisfies

$$\mathbb{C}\mathrm{orr}(Z_i,\, Z_j \,|\, Z_{\neq\, i,\, j}) = -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\,\Omega_{jj}}}$$

*The authors thank ALM (Angers Loire Métropole) and the ICO (Institut de Cancérologie de l'Ouest) for the financial support. This work is partially financed through the ALM grant and the "Programme opérationnel régional FEDER-FSE Pays de la Loire 2014-2020" noPL0015129 (EPICURE).

†Univ Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France, eunice.okomeobiang@univ-angers.fr

‡1 Unité de Bioinfomique, Institut de Cancérologie de l'Ouest, Bd Jacques Monod, 44805 Saint Herblain Cedex, France. 2 SIRIC ILIAD, Nantes, Angers, France. 3 CRCINA, INSERM, CNRS, Université de Nantes, Université d'Angers, Institut de Recherche en Santé-Université de Nantes, 8 Quai Moncousu - BP 70721, 44007, Nantes Cedex 1, France, pascal.jezequel@ico.unicancer.fr

§Univ Angers, CNRS, LAREMA, SFR MATHSTIC, F-49000 Angers, France, frederic.proia@univ-angers.fr

 https://doi.org/10.1214/22-BA1315

where $\Omega = \Sigma^{-1} \in \mathbb{S}_{++}^d$ is the precision matrix of Z (the notation $\mathbb{S}_{++}^d$ for the cone of symmetric positive definite matrices of dimension d is used in all the paper). A fundamental consequence of this is that there is a partial correlation between Z_i and Z_j if and only if the (i, j)-th element of Ω is non-zero. The sparse estimation of Ω is therefore a major issue for variable selection in high-dimensional studies, which has given rise to a substantial literature, see *e.g.* the seminal work of Meinshausen and Bühlmann (2006). This logically led numerous authors to investigate interesting properties under various kind of hypotheses, estimation procedures and penalties. Let us mention for example the optimality results obtained by Cai and Zhou (2012) and the penalized estimations of Yuan and Lin (2007), Rothman et al. (2008), Banerjee et al. (2008), Cai et al. (2011) or Ravikumar et al. (2011), all coming with theoretical guarantees, algorithmic considerations and real world examples. Besides, the famous graphical Lasso of Friedman et al. (2008) has become an essential tool for dealing with precision matrix estimation. Perhaps more attractive to us since focusing on each entry of the precision matrix (no longer taken as a whole), the approach of Ren et al. (2015) is remarkable and will serve as a basis for comparison in our simulation study. The Bayesian inference counterpart has been developed as well, it is *e.g.* the subject of Chap. 10 of Maathuis et al. (2018) where various Wishart-type priors are considered for Ω, see also Li et al. (2019) or Gan et al. (2019) for spike-and-slab approaches and all references within.

Suppose now that we deal with a multivariate linear regression of the form

$$\mathbb{Y} = \mathbb{X}B + E$$

where $\mathbb{Y} \in \mathbb{R}^{n\times q}$ is a matrix of q-dimensional responses of which k-th row is ${Y_k}^t$, $\mathbb{X} \in \mathbb{R}^{n\times p}$ is a matrix of p-dimensional predictors of which k-th row is X_k^t, $B \in \mathbb{R}^{p\times q}$ contains the regression coefficients and $E \in \mathbb{R}^{n\times q}$ is a matrix-variate Gaussian noise. The Partial Gaussian Graphical Model (PGGM), developed *e.g.* by Sohn and Kim (2012) or Yuan and Zhang (2014), appears as a powerful tool to exhibit relations between predictors and responses that exist through partial correlations (called from now on 'direct links', as opposed to 'indirect links' resulting from correlations). Indeed, assume that the couple $(Y_k, X_k) \in \mathbb{R}^{q+p}$ is jointly normally distributed with zero mean, covariance Σ and precision Ω. Then, the block decomposition given by

$$\Omega = \begin{pmatrix} \Omega_y & \Delta \\ \Delta^t & \Omega_x \end{pmatrix}$$

with $\Omega_y \in \mathbb{S}_{++}^q$, $\Delta \in \mathbb{R}^{q\times p}$ and $\Omega_x \in \mathbb{S}_{++}^p$ leads to $Y_k \,|\, X_k \sim \mathcal{N}_q(-\Omega_y^{-1}\Delta\, X_k, \Omega_y^{-1})$. This is a crucial remark because one can see that $Y_k = B^t\, X_k + E_k$ with Gaussian noise $E_k \sim \mathcal{N}_q(0, R)$ is a multiple-output regression that may be reparametrized with

$$B = -\Delta^t\, \Omega_y^{-1} \quad \text{and} \quad R = \Omega_y^{-1}. \tag{1.1}$$

A large volume of literature exists for the sparse estimation of B with arbitrary group structures (see *e.g.* Li et al. (2015) or Chap. 6 of Giraud (2014)), but we will not tackle this issue in our study. At least not frontally but indirectly, since the latter relations show that an estimation of B is possible through the one of the pair (Ω_y, Δ). Whereas B contains direct and indirect links between the predictors and the responses (due

e.g. to strong correlations among the variables), Δ is clearly more interesting from an inferential point of view for it only contains direct links. However, while the estimation of (Ω_y, Δ) appears to be essential, it usually depends on the accuracy of the estimation of the whole precision matrix, which, in turn, may be strongly affected by the one of Ω_x. For example, the graphical Lasso of Friedman et al. (2008) involves maximizing the log-likelihood penalized by the elementwise ℓ_1 norm of Ω. For multiple-output high-dimensional regressions where generally $p \gg q$, we understand that a significant bias is likely to result from the large-scale shrinkage. Another substantial advantage of the partial model is that we can override this issue by computing a new objective function in which Ω_x has disappeared, *i.e.* the penalized log-likelihood

$$\begin{aligned} L_n(\Omega_y, \Delta) \;=\; & -\ln\det(\Omega_y) + \operatorname{tr}(S_y\,\Omega_y) + 2\operatorname{tr}(S_{yx}^t\,\Delta) \\ & + \operatorname{tr}(S_x\,\Delta^t\,\Omega_y^{-1}\,\Delta) + \lambda\,\mathrm{pen}(\Omega_y) + \mu\,\mathrm{pen}(\Delta) \end{aligned} \tag{1.2}$$

where $S_x \in \mathbb{S}_{++}^p$ and $S_y \in \mathbb{S}_{++}^q$ are the empirical variances of the responses and the predictors, respectively, and where $S_{yx} \in \mathbb{R}^{q\times p}$ is the empirical covariance, computed on the basis of a set of n observations. This can be obtained either by considering the multiple-output Gaussian regression scheme, or, as it is done by Yuan and Zhang (2014), by eliminating Ω_x thanks to a first optimization step in the objective function of the graphical model. The usual convex penalties are elementwise ℓ_1 norms, possibly deprived of the diagonal terms for Ω_y. This paved the way to the recent study of Chiquet et al. (2017) where the authors replace the penalty on Ω_y by a structuring one enforcing various kind of sparsity patterns in Δ, and to the one of Okome Obiang et al. (2021) in which some theoretical guarantees are provided for a slightly more general estimation procedure.

However, to the best of our knowledge, the Bayesian approach for the PGGM is a new research topic. Given the outputs gathered in $\mathbb{Y}$ and the predictors gathered in $\mathbb{X}$, the objective of this paper is the Bayesian estimation of the direct links and the precision matrix of the responses. This is inspired by the ideas of Xu and Ghosh (2015) for the single-output setting ($q = 1$), and by the ones of Liquet et al. (2017) for the multiple-output setting ($q > 1$). Taking advantage of the relations (1.1), we consider that a Gaussian prior for B must remain Gaussian for Δ (with a correctly updated variance), and that an inverse Wishart prior for R merely becomes a Wishart one for Ω_y. Yet, despite these seemingly small changes in the design of the priors, we will see that the resulting distributions are completely different. The hierarchical models that we are going to study all come from this working base, but let us point out that a wide variety of refinements exists in the recent literature for Bayesian sparsity, like the grouped 'horseshoe' of Xu et al. (2016), the 'aggressive' multivariate Dirichlet-Laplace prior of Wei et al. (2020), the theoretical results for group selection consistency of Yang and Narisetty (2020) or even the extension of the Bayesian spike-and-slab group selection to generalized additive models of Bai et al. (2020), all related to the regression setting but that might also be investigated for PGGMs. To enforce various types of sparsity in Δ for high-dimensional problems, we decided to make use of spike-and-slab priors, with a spike probability guided by a conjugate Beta distribution.

The paper is organized as follows. Sections 2, 3 and 4 are dedicated to the study of our hierarchical models enforcing either no sparsity, sparsity, group sparsity or sparse-group sparsity in the direct links, respectively, according to the terminology of Sec. 2.1 of Giraud (2014). In particular, we will see that our bi-level selection clearly diverges from the strategy of Liquet et al. (2017). We also adapt the reasoning of Yang and Narisetty (2020) to establish group selection consistency under some technical assumptions and an appropriate amount of sparsity. The Gibbs samplers are tested in Section 5. This empirical section is focused on a simulation study first, to evaluate and compare the efficiency of the models, then a real dataset is treated, and a short conclusion ends the paper. All the technical developments are postponed to the supplementary material (Okome Obiang et al., 2022). But, firstly, let us give some examples of what exactly we mean by 'sparse', 'group-sparse' and 'sparse-group-sparse' settings, and let us summarize the definitions that we have chosen to retain for the well-known distributions as well as for the less usual ones, in order to avoid any misinterpretation of our results and proofs.

Example 1.1. To explain a set of phenotypic traits, suppose that we investigate a large collection of genetic markers spread over twenty chromosomes. For coordinate sparsity ('sparse' setting), only a few markers are active. For group sparsity ('group-sparse' setting), the markers are clustered into groups (formed by chromosomes) and only a few of them are active. For sparse-group sparsity ('sparse-group-sparse' setting), only a few chromosomes are active and they are sparse, the result is a bi-level selection (chromosomes and markers). This will be the context of our example on real data (Section 5.2).

Definition 1.1 (Gaussian)**.** *The density of* $X \in \mathbb{R}^{d_1 \times d_2}$ *following the matrix normal distribution* $\mathcal{MN}_{d_1 \times d_2}(M, \Sigma_1, \Sigma_2)$ *is given by*

$$p(X) = \frac{1}{(2\pi)^{\frac{d_1 d_2}{2}} |\Sigma_1|^{\frac{d_2}{2}} |\Sigma_2|^{\frac{d_1}{2}}} \exp\left(-\frac{1}{2} \operatorname{tr}\left(\Sigma_2^{-1}(X-M)^t \Sigma_1^{-1}(X-M)\right)\right)$$

where $M \in \mathbb{R}^{d_1 \times d_2}$, $\Sigma_1 \in \mathbb{S}_{++}^{d_1}$ *and* $\Sigma_2 \in \mathbb{S}_{++}^{d_2}$. *When* $d_2 = 1$, *this is a multivariate normal distribution* $\mathcal{N}_d(\mu, \Sigma)$ *with* $d = d_1$, $\mu = M$ *and* $\Sigma = \Sigma_2^{-1}\Sigma_1$, *having density*

$$p(X) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (X-\mu)^t \Sigma^{-1}(X-\mu)\right)$$

where $\mu \in \mathbb{R}^d$ *and* $\Sigma \in \mathbb{S}_{++}^d$.

Definition 1.2 (Generalized Inverse Gaussian)**.** *The density of* $X \in \mathbb{S}_{++}^d$ *following the matrix generalized inverse Gaussian distribution* $\mathcal{MGIG}_d(\nu, A, B)$ *is given by*

$$p(X) = \frac{|X|^{\nu - \frac{d+1}{2}}}{\left|\frac{A}{2}\right|^{\nu} B_\nu\left(\frac{A}{2}, \frac{B}{2}\right)} \exp\left(-\frac{1}{2} \operatorname{tr}\left(A X^{-1} + B X\right)\right) \mathbb{1}_{\{X \in \mathbb{S}_{++}^d\}}$$

where $\nu \in \mathbb{R}$, $A \in \mathbb{S}_{++}^d$, $B \in \mathbb{S}_{++}^d$ *and* B_ν *is a Bessel-type function of order* ν. *When* $d = 1$, *this is a generalized inverse Gaussian distribution* $\mathcal{GIG}(\nu, a, b)$ *with* $a = A$ *and* $b = B$, *having density*

$$p(X) = \frac{X^{\nu-1}}{\left(\frac{a}{2}\right)^{\nu} B_\nu\left(\frac{a}{2}, \frac{b}{2}\right)} \mathrm{e}^{-\frac{a}{2X} - \frac{bX}{2}} \mathbb{1}_{\{X > 0\}}$$

where $\nu \in \mathbb{R}$, $a > 0$ and $b > 0$.

Definition 1.3 (Wishart/Gamma/Exponential)**.** *The density of $X \in \mathbb{S}_{++}^d$ following the matrix Wishart distribution $\mathcal{W}_d(u, V)$ is given by*

$$p(X) = \frac{|X|^{\frac{u-d-1}{2}}}{2^{\frac{d\,u}{2}}\,\Gamma_d\big(\frac{u}{2}\big)\,|V|^{\frac{u}{2}}}\,\exp\!\left(-\frac{1}{2}\,\mathrm{tr}\big(V^{-1}X\big)\right)\mathbb{1}_{\{X\,\in\,\mathbb{S}_{++}^d\}}$$

where $u > d-1$, $V \in \mathbb{S}_{++}^d$ and Γ_d is the multivariate Gamma function of order d. When $d = 1$, this is a Gamma distribution $\Gamma(a,\, b)$ with $a = \frac{u}{2}$ and $\frac{1}{b} = 2\,V$, having density

$$p(X) = \frac{b^{\,a}\,X^{a-1}}{\Gamma(a)}\,\mathrm{e}^{-b\,X}\,\mathbb{1}_{\{X\,>\,0\}}$$

where $a > 0$ and $b > 0$. The exponential distribution $\mathcal{E}(\ell)$ is then defined as the $\Gamma(1,\, \ell)$ distribution, for $\ell > 0$.

Definition 1.4 (Beta)**.** *The density of $X \in [0, 1]$ following the Beta distribution $\beta(a,\, b)$ is given by*

$$p(X) = \frac{X^{a-1}\,(1-X)^{\,b-1}}{\beta(a, b)}\,\mathbb{1}_{\{0\,\leqslant\,X\,\leqslant\,1\}}$$

where $a > 0$, $b > 0$ and β is the Beta function.

In all the paper, data and parameters are gathered in $\Theta = \{\mathbb{Y}, \mathbb{X}, \Delta, \Omega_y, \nu, \lambda, \pi\}$ and, to standardize, for any $e \in \Theta$, we note $\Theta_e = \Theta \backslash \{e\}$.

2 The sparse setting

In this section, $\lambda_i \in \mathbb{R}$ is the i-th component of $\lambda \in \mathbb{R}^p$, $\Delta_i \in \mathbb{R}^q$ is the i-th column of Δ and $\mathbb{X}_i \in \mathbb{R}^n$ stands for the i-th column of $\mathbb{X}$ ($1 \leqslant i \leqslant p$). Let us consider the hierarchical Bayesian model, where the columns of Δ are assumed to be independent, given by

$$\left\{\begin{array}{lcl} \mathbb{Y}\,|\,\mathbb{X}, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-\mathbb{X}\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \Delta_i\,|\,\Omega_y, \lambda_i, \pi & \overset{\perp\!\!\!\perp}{\sim} & (1-\pi)\,\mathcal{N}_q(0,\, \lambda_i\,\Omega_y) + \pi\,\delta_0 \\ \lambda_i & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha,\, \ell_i) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi & \sim & \beta(a,\, b) \end{array}\right. \tag{2.1}$$

for $i \in [\![1, p]\!]$, with hyperparameters $\alpha = \frac{1}{2}(q+1)$, $\ell_i > 0$, $u > q-1$, $V \in \mathbb{S}_{++}^q$, $a > 0$ and $b > 0$. A general ungrouped sparsity is promoted in the columns of Δ through the spike-and-slab prior. In this mixture model, π is the prior spike probability and λ is an adaptative shrinkage factor acting at the predictor scale (λ_i is associated with the direct links between predictor i and all the responses). When $\ell_i = \ell$ for all i, we will rather speak of global shrinkage. The degree of sparsity will be characterized by the number N_0 of zero columns of Δ, that is

$$N_0 = \mathrm{Card}(i,\, \Delta_i = 0) = \sum_{i=1}^{p} \mathbb{1}_{\{\Delta_i\,=\,0\}}. \tag{2.2}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (2.1), we may use the conditional distributions given in the proposition below.

Proposition 2.1. *In the hierarchical model* (2.1)*, the conditional posterior distributions are as follows.*

– *The parameter* Δ *satisfies, for* $i \in [\![1, p]\!]$,

$$\Delta_i \,|\, \Theta_{\Delta_i} \;\sim\; (1-p_i)\,\mathcal{N}_q(-s_i\,H_i,\, s_i\,\Omega_y) + p_i\,\delta_0$$

where

$$H_i = \Omega_y\,\mathbb{Y}^t\,\mathbb{X}_i + \sum_{j\neq i}\langle \mathbb{X}_i,\, \mathbb{X}_j\rangle\,\Delta_j, \qquad s_i = \frac{\lambda_i}{1+\lambda_i\,\|\mathbb{X}_i\|^2}$$

and

$$p_i = \frac{\pi}{\pi + (1-\pi)\,(1+\lambda_i\,\|\mathbb{X}_i\|^2)^{-\frac{q}{2}}\,\exp\!\Big(\frac{s_i\,H_i^t\,\Omega_y^{-1}H_i}{2}\Big)}.$$

– *The parameter* Ω_y *satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\!\left(\frac{n-p+N_0+u}{2},\, \Delta\,(\mathbb{X}^t\,\mathbb{X} + D_\lambda^{-1})\,\Delta^t,\, \mathbb{Y}^t\,\mathbb{Y} + V^{-1}\right)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

– *The parameter* λ *satisfies, for* $i \in [\![1, p]\!]$,

$$\lambda_i \,|\, \Theta_{\lambda_i} \;\sim\; \mathbb{1}_{\{\Delta_i \neq 0\}}\,\mathcal{GIG}\!\left(\frac{1}{2},\, \Delta_i^t\,\Omega_y^{-1}\Delta_i,\, 2\,\ell_i\right) + \mathbb{1}_{\{\Delta_i = 0\}}\,\Gamma(\alpha,\, \ell_i).$$

– *The parameter* π *satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(N_0 + a,\, p - N_0 + b\big).$$

Proof. See Section B.1 of the supplementary material. □

Remark 2.1. The Bayesian Lasso, as introduced *e.g.* in Sec. 6.1 of Hastie et al. (2015) or in Park and Casella (2008), assumes a prior Laplace distribution for the regression coefficients conditional on the noise variance. In our case, $\Delta_i \,|\, \Omega_y, \pi$ is still a multivariate spike-and-slab (after integrating over λ_i), with a slab following a so-called multivariate K-distribution (see Eltoft et al. (2006)), which is a generalization of the multivariate Laplace distribution. See *e.g.* Sec 2.1 of Liquet et al. (2017). From this point of view, our study is in line with the usual Bayesian regression schemes. Perhaps more interesting, going on with the idea of the authors, suppose that, for all $1 \leqslant i \leqslant p$, $\Delta_i = b_i\,\Delta_i^*$ where Δ_i^* follows the multivariate K-distribution described above and $b_i \,|\, \pi \sim \mathcal{B}(1-\pi)$ is independent of Δ_i^*. Now, the sparsity in Δ is not induced by a spike-and-slab strategy anymore but, equivalently, by multiplying the slab part by an independent Bernoulli

variable being 0 with probability π. Then, it is possible to show that the negative log-likelihood of this alternative hierarchical model is given, up to an additive constant that does not depend on Δ, by

$$\frac{1}{2}\left\|(\mathbb{Y}+\mathbb{X}\,\Delta^t\,\Omega_y^{-1})\,\Omega_y^{\frac{1}{2}}\right\|_F^2+\sum_{i=1}^{p}c_i\left\|\Omega_y^{-\frac{1}{2}}\Delta_i^*\right\|_F+\ln\left(\frac{1-\pi}{\pi}\right)\sum_{i=1}^{p}b_i$$

where $c_i>0$. We first recognize an ℓ_2-type penalty but also an ℓ_0-type penalty on Δ (provided that $\pi<\frac{1}{2}$) since summing the b_i amounts to counting the number of non-zero columns in Δ. Consequently, there is a close connection between our hierarchical Bayesian model and the regressions penalized by ℓ_2 and ℓ_0 norms, problems that are known to be very hard to solve due to combinatorial optimization.

The particular case $q=1$ is a very useful corollary of the proposition (see Corollary A.1 in the supplementary material). Note that we can also easily derive the Bayesian counterpart of the standard PGGM adapted to the small-dimensional case, with no sparsity, by taking $\pi=0$.

Corollary 2.1. *In the hierarchical model (2.1) with $\pi=0$, the conditional posterior distributions are as follows.*

– *The parameter Δ satisfies, for $i\in[\![1,p]\!]$,*

$$\Delta_i\,|\,\Theta_{\Delta_i}\;\sim\;\mathcal{N}_q(-s_i\,H_i,\,s_i\,\Omega_y)$$

where

$$H_i=\Omega_y\,\mathbb{Y}^t\,\mathbb{X}_i+\sum_{j\neq i}\langle\mathbb{X}_i,\,\mathbb{X}_j\rangle\,\Delta_j\qquad\text{and}\qquad s_i=\frac{\lambda_i}{1+\lambda_i\,\|\mathbb{X}_i\|^2}.$$

– *The parameter Ω_y satisfies*

$$\Omega_y\,|\,\Theta_{\Omega_y}\;\sim\;\mathcal{MGIG}_q\left(\frac{n-p+u}{2},\,\Delta\,(\mathbb{X}^t\,\mathbb{X}+D_\lambda^{-1})\,\Delta^t,\,\mathbb{Y}^t\,\mathbb{Y}+V^{-1}\right)$$

where $D_\lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_p)$.

– *The parameter λ satisfies, for $i\in[\![1,p]\!]$,*

$$\lambda_i\,|\,\Theta_{\lambda_i}\;\sim\;\mathcal{GIG}\left(\frac{1}{2},\,\Delta_i^t\,\Omega_y^{-1}\Delta_i,\,2\,\ell_i\right).$$

Proof. This is a consequence of Proposition 2.1. □

In the simulation study of Section 5.1, Scen. 0, 1 and 2 are dedicated to the sparse setting. The next section discusses the group sparsity in Δ.

3 The group-sparse setting

The predictors are now ordered in m groups of sizes $\kappa_1 + \ldots + \kappa_m = p$. For the g-th group ($1 \leqslant g \leqslant m$), $\lambda_g \in \mathbb{R}$ is the g-th component of $\lambda \in \mathbb{R}^m$, the covariate submatrix is $\underline{\mathbb{X}}_g \in \mathbb{R}^{n \times \kappa_g}$ and the corresponding slice of Δ is $\underline{\Delta}_g \in \mathbb{R}^{q \times \kappa_g}$. Let us consider the hierarchical Bayesian model, where the columns of Δ are assumed to be independent both within and between the groups, given by

$$\left\{ \begin{array}{lcl} \mathbb{Y} \,|\, \mathbb{X}, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-\mathbb{X}\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \underline{\Delta}_g \,|\, \Omega_y, \lambda_g, \pi & \perp\!\!\!\sim & (1-\pi)\,\mathcal{MN}_{q\times\kappa_g}(0,\, \lambda_g\,\Omega_y, I_{\kappa_g}) + \pi\,\delta_0 \\ \lambda_g & \perp\!\!\!\sim & \Gamma(\alpha_g,\, \ell_g) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi & \sim & \beta(a,\, b) \end{array} \right. \tag{3.1}$$

for $g \in [\![1, m]\!]$, with hyperparameters $\alpha_g = \frac{1}{2}(q\,\kappa_g + 1)$, $\ell_g > 0$, $u > q-1$, $V \in \mathbb{S}_{++}^q$, $a > 0$ and $b > 0$. A general group sparsity is promoted in the columns of Δ through the spike-and-slab prior at the group level. In this mixture model, π is the prior spike probability and λ is an adaptative shrinkage factor acting at the group scale (λ_g is associated with the direct links between the predictors of group g and all the responses). Likewise, when $\ell_g = \ell$ for all g, we will rather speak of global shrinkage. Now, the degree of sparsity will be characterized by N_0 given in (2.2), but also by the number G_0 of zero groups of Δ, that is

$$G_0 = \mathrm{Card}(g,\, \underline{\Delta}_g = 0) = \sum_{g=1}^{m} \mathbb{1}_{\{\underline{\Delta}_g = 0\}}. \tag{3.2}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (3.1), we may use the conditional distributions given in the proposition below.

Proposition 3.1. *In the hierarchical model (3.1), the conditional posterior distributions are as follows.*

– *The parameter Δ satisfies, for $g \in [\![1, m]\!]$,*

$$\underline{\Delta}_g \,|\, \Theta_{\underline{\Delta}_g} \;\sim\; (1-p_g)\,\mathcal{MN}_{q\times\kappa_g}(-H_g\,S_g,\, \Omega_y,\, S_g) + p_g\,\delta_0$$

where

$$H_g = \Omega_y \mathbb{Y}^t\,\underline{\mathbb{X}}_g + \sum_{j\neq g} \underline{\Delta}_j\,\underline{\mathbb{X}}_j^t\,\underline{\mathbb{X}}_g, \qquad S_g = \lambda_g\left(I_{\kappa_g} + \lambda_g\,\underline{\mathbb{X}}_g^t\,\underline{\mathbb{X}}_g\right)^{-1}$$

and

$$p_g = \frac{\pi}{\pi + (1-\pi)\,|I_{\kappa_g} + \lambda_g\,\underline{\mathbb{X}}_g^t\,\underline{\mathbb{X}}_g|^{-\frac{q}{2}}\,\exp\!\left(\frac{\mathrm{tr}(H_g^t\,\Omega_y^{-1}H_g\,S_g)}{2}\right)}.$$

– *The parameter Ω_y satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\!\left(\frac{n-p+N_0+u}{2},\, \Delta\,(\mathbb{X}^t\,\mathbb{X} + D_\lambda^{-1})\,\Delta^t,\, \mathbb{Y}^t\,\mathbb{Y} + V^{-1}\right)$$

where $D_\lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_1, \ldots, \lambda_m, \ldots, \lambda_m)$ with each λ_g duplicated κ_g times.

– *The parameter* λ *satisfies, for* $g \in [\![1, m]\!]$,

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\left(\frac{1}{2},\, \mathrm{tr}(\underline{\Delta}_g^t\, \Omega_y^{-1} \underline{\Delta}_g),\, 2\,\ell_g\right) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \ell_g).$$

– *The parameter* π *satisfies*

$$\pi \,|\, \Theta_\pi \;\sim\; \beta\big(G_0 + a,\, m - G_0 + b\big).$$

Proof. See Section B.2 of the supplementary material. □

Note that Remark 2.1 still applies to this configuration, after some adjustments (the ℓ_0-like penalty is on the number of non-zero groups). Here again, the particular case $q = 1$ is a very useful corollary (see Corollary A.2 in the supplementary material). In the simulation study of Section 5.1, Scen. 3 and 4 are dedicated to the group-sparse setting. To conclude this section, a theoretical guarantee is provided (given Ω_y and with $\lambda = \lambda_n$ and $\pi = \pi_n$ depending on n). It is possible to obtain a model selection consistency property for this approach when both the number of observations n and the number of groups $m = m_n$ tend to infinity, by adapting the reasoning of Yang and Narisetty (2020) related to the linear regression (with $q = 1$). Indeed, when Ω_y is known, Δ reduces to a linear transformation of B. Thus, it is not surprising that a similar result follows under the same kind of hypotheses. In the sequel, we denote by $\mathbb{X}_{(k)} \in \mathbb{R}^{n \times |k|}$ the design matrix of rank r_k corresponding to the submodel indexed by the binary vector $k \in \{0,1\}^m$ having $|k|$ non-zero values ($k_g = 1$ means that the g-th group is included in the model), and by $\Pi_{(k)} \in \mathbb{R}^{n \times n}$ the projection matrix onto the column-space of $\mathbb{X}_{(k)}$. Similarly, Δ restricted to k is $\Delta_{(k)} \in \mathbb{R}^{q \times |k|}$. The true model is called t and $t^{\pm g}$ are submodels of t that contain only the g-th group or that are deprived of it, respectively. Let

$$\delta_1 = \inf_{1 \leqslant g \leqslant |t|} \left\| (I_n - \Pi_{(t^{-g})})\, \mathbb{X}_{(t^{+g})}\, \Delta_{(t^{+g})}^t\, \Omega_y^{-\frac{1}{2}} \right\|_F^2$$

and, for some $K > 0$,

$$\delta_2^K = \inf_{k \,\in\, E_K} \left\| (I_n - \Pi_{(k)})\, \mathbb{X}_{(t)}\, \Delta_{(t)}^t\, \Omega_y^{-\frac{1}{2}} \right\|_F^2$$

with $E_K = \{k \text{ such that } t \not\subset k \text{ and } r_k \leqslant K r_t\}$. Let also,

$$\mu_{n,\,\min}^K = \inf_{k \,\in\, F_K} \mu^+\left(\frac{\mathbb{X}_{(k)}^t\, \mathbb{X}_{(k)}}{n}\right) \quad \text{and} \quad \bar{\mu}_n = \inf_{k \,\in\, F} \mu^*\left(\frac{\mathbb{X}_{(k)}^t\, (I_n - \Pi_{(k \cap t)})\, \mathbb{X}_{(k)}}{n}\right)$$

with $F_K = \{k \text{ such that } t \subset k \text{ and } r_k \leqslant (K+1)\, r_t\}$ and $F = \{k \text{ such that } |k \backslash t| > 0\}$, and where, for a square matrix A, $\mu^+(A)$ is the minimum non-zero eigenvalue of A and $\mu^*(A)$ is the geometric mean of the non-zero eigenvalues of A. The hypotheses are those of Yang and Narisetty (2020) that we have to slightly adapt. By $f_n \asymp g_n$ we mean that there is a constant $c \neq 0$ such that $f_n / g_n \to c$ as n tends to infinity.

(H.1) There exists a rate such that $m_n = \mathrm{e}^{v_n}$ with $v_n \to +\infty$ and $v_n = o(n)$.

(H.2) The prior slab probability satisfies $1 - \pi_n \asymp 1/m_n$.

(H.3) The shrinkage factors satisfy $n\lambda_n^\sharp \asymp m_n^{2+\eta}\,\bar{\mu}_n^{-\eta}$ and $\mu_{n,\min}^K\, n\lambda_n^\sharp \to +\infty$ for some $\eta > 0$, where $\lambda_n^\sharp = \max_i \lambda_{n,i}$.

(H.4) There exists $\epsilon_1 > 0$ such that $\delta_1 > (1+\epsilon_1)\, r_t\, [(4+\eta)\ln m_n - \eta \ln \bar{\mu}_n]$.

(H.5) There exists $\epsilon_2 > 0$ such that $\delta_2^K > (1+\epsilon_2)\, r_t\, [(4+\eta)\ln m_n - \eta \ln \bar{\mu}_n]$ for some $K > \max(8/\eta + 1, \eta/(\eta-1))$.

We refer the reader to p. 917 of Yang and Narisetty (2020) where the authors give very clarifying comments on the interpretation to be given to these technical assumptions. In particular, while (H.1), (H.2) and (H.3) control the behavior of m_n, π_n and λ_n as n tends to infinity, (H.4) and (H.5) are related to sensitivity and specificity and are therefore in connection with the true model t.

Proposition 3.2. *Suppose that (H.1)–(H.5) are satisfied. Then, as n tends to infinity,*

$$\mathbb{P}(\mathcal{T} \,|\, \mathbb{Y}, \mathbb{X}, \Omega_y) \xrightarrow{\mathbb{P}} 1$$

where $\mathcal{T} = \{t \text{ is selected}\}$ and t is the true model.

Proof. The result is obtained by following the same lines as the proof of Thm 2.1 of Yang and Narisetty (2020). One just has to clarify a few points to solve the issues arising from $q \geqslant 1$ and from the adaptative shrinkage, which is done in Section B.4 of the supplementary material. □

Remark 3.1. Obviously, Proposition 3.2 also holds for the sparse setting (with $m = p$) and in that case, it is instructive to draw the parallel with Thm. 1 of Ren et al. (2015) even if the estimation procedure is very different. The authors show that, to obtain a $\sqrt{n}$-consistent estimation of the precision matrix Ω in a GGM, Ω must contain at most $\asymp \sqrt{n}/\ln p$ non-zero columns. In the Gibbs sampler (see Proposition 2.1), the slab probability $1 - \pi$ is generated according to a distribution that satisfies

$$\mathbb{E}[1 - \pi \,|\, \Theta_\pi] = \frac{p - N_0 + b}{p + a + b} \quad \text{and} \quad \mathbb{V}(1 - \pi \,|\, \Theta_\pi) = \frac{(N_0 + a)(p - N_0 + b)}{(p + a + b)^2\,(p + a + b + 1)}.$$

Thus, if the model selects $\asymp \sqrt{n}/\ln p$ predictors, it follows that the posterior expectation of $1 - \pi$ is $\asymp \sqrt{n}/(p \ln p) = 1/p$ when $p = \mathrm{e}^{\sqrt{n}}$. In that case, the posterior variance of $1 - \pi$ is $\asymp 1/p^2$. To sum up, in a model with $\asymp \sqrt{n}/\ln p$ predictors selected, the posterior distribution of $1 - \pi$ is very concentrated around $1/p$ which conforms to (H.1) and (H.2). This is not directly comparable due to the different procedures, but it seems interesting to observe that the same orders of magnitude are involved to reach theoretical guarantees for the estimation of Δ.

In the next section, an approach is suggested to deal with sparse-group sparsity in Δ, for a bi-level selection.

4 The sparse-group-sparse setting

To produce a sparse model both at the variable level (for variable selection) and at the group level (for group selection), it seems natural to carry on with our strategy by introducing another spike-and-slab effect into the first one. The predictors are still ordered in m groups of sizes $\kappa_1 + \ldots + \kappa_m = p$. For the g-th group ($1 \leqslant g \leqslant m$), $\lambda_g \in \mathbb{R}$ is the g-th component of $\lambda \in \mathbb{R}^m$ and, for the i-th predictor of this group ($1 \leqslant i \leqslant \kappa_g$), $\nu_{gi} \in \mathbb{R}$ is the i-th component of $\nu_g \in \mathbb{R}^{\kappa_g}$. The i-th column of the covariate submatrix $\underline{\mathbb{X}}_g$ is $\mathbb{X}_{gi} \in \mathbb{R}^n$ and the corresponding slice of $\underline{\Delta}_g$ is $\Delta_{gi} \in \mathbb{R}^q$ while $\underline{\Delta}_{g\setminus i} \in \mathbb{R}^{q\times(\kappa_g-1)}$ is $\underline{\Delta}_g$ deprived of Δ_{gi}. Here our approach diverges from Xu and Ghosh (2015) and Liquet et al. (2017). The bi-level selection of the authors is made through spike-and-slab effects both at the group scale and on the individual variances, considered as truncated Gaussians, generating zero groups and (almost surely) zero coefficients within the groups. Let us suggest instead the Bayesian hierarchical model given by

$$\left\{\begin{array}{lcl} \mathbb{Y} \,|\, \mathbb{X}, \Delta, \Omega_y & \sim & \mathcal{MN}_{n\times q}(-\mathbb{X}\,\Delta^t\,\Omega_y^{-1},\, I_n,\, \Omega_y^{-1}) \\ \underline{\Delta}_g \,|\, \nu_g, \lambda_g, \pi & \overset{\perp\!\!\!\perp}{\sim} & (1-\pi_1)\left[(1-\pi_2)\mathcal{N}_q(0,\, \lambda_g\,\nu_{gi}\,\Omega_y) + \pi_2\,\delta_0\right]^{\otimes\,\kappa_g} + \pi_1\,\delta_0 \\ \nu_{gi} & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha,\, \ell_{gi}) \\ \lambda_g & \overset{\perp\!\!\!\perp}{\sim} & \Gamma(\alpha_g,\, \gamma_g) \\ \Omega_y & \sim & \mathcal{W}_q(u,\, V) \\ \pi_j & \overset{\perp\!\!\!\perp}{\sim} & \beta(a_j,\, b_j) \end{array}\right. \tag{4.1}$$

for $g \in [\![1, m]\!]$, $i \in [\![1, \kappa_g]\!]$ and $j \in [\![1, 2]\!]$, with hyperparameters $\alpha = \frac{1}{2}(q+1)$, $\alpha_g = \frac{1}{2}(q\,\kappa_g + 1)$, $\ell_{gi} > 0$, $\gamma_g > 0$, $u > q-1$, $V \in \mathbb{S}_{++}^q$, $a_j > 0$, and $b_j > 0$. In this mixture model, π_1 is the prior spike probability on the groups whereas π_2 is the prior spike probability within the non-zero groups, for a bi-level selection. In terms of cumulative shrinkage effects, λ is an adaptative shrinkage factor acting at the group scale and ν is an adaptative shrinkage factor acting at the predictor scale (λ_g is associated with the direct links between the predictors of group g and all the responses whereas ν_{gi} is associated with the direct links between predictor i of group g and all the responses). In this way, (4.1) opens up many perspectives for dealing with bi-level shrinkage. We can set $\gamma_g = \gamma$ for all g, for a global shrinkage at the group scale. At the predictor scale, when $\ell_{gi} = \ell_g$ for all i, this is a global shrinkage in the g-th group but we might even consider a full global shrinkage $\ell_{gi} = \ell$. However, an identifiability issue may result from the product $\lambda_g\,\nu_{gi}$ between group and within-group effects. Even if the posterior distributions depend on different levels of data that shall resolve it, one can for example fix $\lambda_g = 1$ (for adaptative) or $\nu_{gi} = 1$ (for global) and let the shrinkage entirely rely on the other parameter. Although it achieves the same objectives as those of Xu and Ghosh (2015) and Liquet et al. (2017), this hierarchy seems more consistent with our previous sections (take $\pi_2 = 0$ and $\nu_{gi} = 1$ to remove the within-group effect and recover the group-sparse setting of Section 3, take $\pi_1 = 0$ and $\lambda_g = 1$ to remove the group effect and recover the sparse setting of Section 2). In this context, the degree of sparsity is still characterized by N_0 given in (2.2) for the predictor scale, by G_0 given in (3.2) for the group scale, but also, for the within-group scale, by the number N_{0g} of zero columns in

each particular group g, that is, for all $1 \leqslant g \leqslant m$,

$$N_{0g} = \text{Card}(i,\, \Delta_{gi} = 0) = \sum_{i=1}^{\kappa_g} \mathbb{1}_{\{\Delta_{gi}=0\}}. \tag{4.2}$$

We also need to define the number J_0 of zero columns in the non-zero groups, that is

$$J_0 = \text{Card}(i,\, \Delta_{gi} = 0 \text{ and } \underline{\Delta}_g \neq 0) = \sum_{g=1}^{m} N_{0g}\, \mathbb{1}_{\{\Delta_g \neq 0\}}. \tag{4.3}$$

To implement a Gibbs sampler from the full posterior distribution stemming from (4.1), we may use the conditional distributions given in the proposition below.

Proposition 4.1. *In the hierarchical model (4.1), the conditional posterior distributions are as follows.*

– *The parameter Δ_{gi} satisfies, for $g \in [\![1, m]\!]$ and $i \in [\![1, \kappa_g]\!]$,*

$$\Delta_{gi} \,|\, \Theta_{\Delta_{gi}} \;\sim\; (1 - p_{gi})\mathcal{N}_q(-s_{gi}\, H_{gi},\, s_{gi}\, \Omega_y) + p_{gi}\, \delta_0$$

where

$$H_{gi} = \Omega_y\, \mathbb{Y}^t\, \mathbb{X}_{gi} + \sum_{h,j \neq g,i} \langle \mathbb{X}_{gi},\, \mathbb{X}_{hj} \rangle\, \Delta_{hj}, \qquad s_{gi} = \frac{\nu_{gi}\, \lambda_g}{1 + \nu_{gi}\, \lambda_g\, \|\mathbb{X}_{gi}\|^2}$$

and

$$p_{gi} = \frac{\rho_{gi}}{\rho_{gi} + (1 - \pi_1)\,(1 - \pi_2)\,(1 + \nu_{gi}\, \lambda_g\, \|\mathbb{X}_{gi}\|^2)^{-\frac{q}{2}}\, \exp\!\Big(\frac{s_{gi}\, H_{gi}^t\, \Omega_y^{-1}\, H_{gi}}{2}\Big)}$$

in which $\rho_{gi} = (1 - \pi_1)\, \pi_2\, \mathbb{1}_{\{\Delta_{g\setminus i} \neq 0\}} + \pi_1\, \mathbb{1}_{\{\Delta_{g\setminus i} = 0\}}$.

– *The parameter Ω_y satisfies*

$$\Omega_y \,|\, \Theta_{\Omega_y} \;\sim\; \mathcal{MGIG}_q\Big(\frac{n - p + N_0 + u}{2},\, \Delta\,(\mathbb{X}^t\, \mathbb{X} + D_{\lambda\nu}^{-1})\, \Delta^t,\, \mathbb{Y}^t\, \mathbb{Y} + V^{-1}\Big)$$

where $D_{\lambda\nu} = \text{diag}(\nu_{11}\, \lambda_1, \ldots, \nu_{1\kappa_1}\, \lambda_1, \ldots, \nu_{m1}\, \lambda_m, \ldots, \nu_{m\kappa_m}\, \lambda_m)$.

– *The parameter ν satisfies, for $g \in [\![1, m]\!]$ and $i \in [\![1, \kappa_g]\!]$,*

$$\nu_{gi} \,|\, \Theta_{\nu_{gi}} \;\sim\; \mathbb{1}_{\{\Delta_{gi} \neq 0\}}\, \mathcal{GIG}\Big(\frac{1}{2},\, \frac{\Delta_{gi}^t\, \Omega_y^{-1} \Delta_{gi}}{\lambda_g},\, 2\,\ell_{gi}\Big) + \mathbb{1}_{\{\Delta_{gi} = 0\}}\, \Gamma(\alpha,\, \ell_{gi}).$$

– *The parameter λ satisfies, for $g \in [\![1, m]\!]$,*

$$\lambda_g \,|\, \Theta_{\lambda_g} \;\sim\; \mathbb{1}_{\{\Delta_g \neq 0\}}\, \mathcal{GIG}\Big(\frac{qN_{0g} + 1}{2},\, \text{tr}(D_{\nu_g}^{-1} \underline{\Delta}_g^t\, \Omega_y^{-1} \underline{\Delta}_g),\, 2\,\gamma_g\Big) + \mathbb{1}_{\{\Delta_g = 0\}}\, \Gamma(\alpha_g,\, \gamma_g)$$

where $D_{\nu_g} = \text{diag}(\nu_{g1}, \ldots, \nu_{g\kappa_g})$.

– *The parameter* π *satisfies, for* $j \in [\![1, 2]\!]$,

$$\pi_j \,|\, \Theta_{\pi_j} \;\sim\; \beta\big(A_j + a_j,\, B_j + b_j\big).$$

where $A_1 = G_0$, $B_1 = m - G_0$, $A_2 = J_0$ *and* $B_2 = p - N_0$.

Proof. See Section B.3 of the supplementary material. □

It only remains to evoke the explicit results for the particular case $q = 1$ (see Corollary A.3 in the supplementary material). In the simulation study of Section 5.1, Scen. 5 and 6 are dedicated to the sparse-group-sparse setting.

5 Empirical results

In this empirical section,[1] let us call (s), (gs) and (sgs) the related settings, and let us denote by (ad) the adaptative shrinkage and by (gl) the global shrinkage. First of all, these models contain many hyperparameters that have to be carefully tuned. Our experiments showed that, unsurprisingly, the results are strongly impacted by the prior amount of shrinkage on Δ, driven by ℓ and even by γ for (sgs). Apart from the usual cross-validation procedures, we could stay in line with our Bayesian approach and suggest conjugate Gamma hyperpriors. This is very easy to implement, but the hyperparameters are now replaced by other hyperparameters and the same questions arise. Instead, like in Xu and Ghosh (2015) and Liquet et al. (2017), we follow the idea of Park and Casella (2008) and we use a Monte-Carlo EM algorithm. By way of example, from the full posterior probability (B.1 in suppl. mat.) and since $\lambda_i \sim \Gamma(\alpha, \ell_i)$ for all i, it is not hard to see that, with (s),

$$\ln p(\Delta, \Omega_y, \lambda, \pi \,|\, \mathbb{Y}, \mathbb{X}) = \sum_{i=1}^{p} (\alpha \ln \ell_i - \ell_i\, \lambda_i) + T_{\neq \ell}$$

where the term $T_{\neq \ell}$ does not depend on ℓ. Thus, the k-th iteration of the EM algorithm should lead to

$$\ell_i^{(k)} = \frac{\frac{1}{2}(q+1)}{\mathbb{E}^{(k-1)}[\lambda_i \,|\, \mathbb{Y}, \mathbb{X}]} \quad \text{and} \quad \ell^{\,(k)} = \frac{\frac{p}{2}(q+1)}{\sum_{i=1}^{p} \mathbb{E}^{(k-1)}[\lambda_i \,|\, \mathbb{Y}, \mathbb{X}]}$$

for the adaptative shrinkage and the global shrinkage ($\lambda_i = \lambda$), respectively. The intractable conditional expectations are then estimated with the help of the Gibbs samples. For (gs), the results are mainly the same as above (replace $q + 1$ by $q\kappa_g + 1$ in the first case, $p(q + 1)$ by $qp + m$ in the second case and consider $1 \leqslant g \leqslant m$ instead of $1 \leqslant i \leqslant p$), and similar results also follow with (sgs). Recall that our definitions of the adaptative and global shrinkages are given in the corresponding sections, in the description of the hierarchical models. The tuning of u and V (or v) is actually trickier. Because $\mathbb{E}[\Omega_y] = uV$, we set $V = \frac{1}{u} I_q$ and u is conveniently chosen to be the smallest

[1] The codes and the dataset are available at `https://github.com/FredericProia/BayesPGGM`.

integer such that Ω_y is (almost surely) invertible, that is $u = q$ (see *e.g.* Brown et al. (1998)). This is particularly adapted when the dataset is standardized. Finally, a and b reflect the degree of sparsity to introduce in the direct links. We can set $a \gg b$ to promote sparse settings, which is potentially interesting when $p \gg n$, but $a = b = 1$ is a standard non-informative choice and $a < b$ may also be useful for variable selection (see *e.g.* the real dataset of Section 5.2). They can be chosen from a cross-validation step (for prediction purposes) or to enforce some degree of sparsity (for selection purposes), just like a practitioner manages the tuning parameter of the Lasso. The posterior median is used to estimate Δ and get sparsity whereas the posterior mean is used to estimate Ω_y. Indeed, we don't want to impose any sparsity on Ω_y (q is small), so we decided to retain this standard choice. But the concern is much greater for Δ because some coordinates must be exactly zero. This is the reason why the posterior median seemed a more appropriate choice (in particular, it suffices for the sampler to generate zeros more than half the time for the empirical posterior median to be zero). Due to the huge amount of calculations in the simulations, the estimations are made on the basis of 3000 iterations of the sampler in which the first 2000 are burn-ins. This is revised upwards for the real data (10000 iterations with 5000 burn-ins).

Remark 5.1. To the best of our knowledge, there is no simple way to sample from the $\mathcal{MGIG}_d$ distribution as soon as $d > 1$. The recent method described in Sec. 3.3.2 of Fang et al. (2020), relying on the Matsumoto-Yor property (see Thm. 3.1 of Massam and Wesołowski (2006)) to get a $\mathcal{MGIG}_d$ sample from the very standard $\mathcal{GIG}$ and $\mathcal{W}_d$ distributions, is unfortunately inapplicable in our context. Indeed, for example in the sparse setting, that would require finding $z \in \mathbb{R}^q$ such that $\mathbb{Y}^t\,\mathbb{Y} + V^{-1} = b\,zz^t$ for some $b > 0$, which is clearly impossible since $\mathbb{Y}^t\,\mathbb{Y} + V^{-1}$ has full rank. In Fazayeli and Banerjee (2016), the authors show that $\mathcal{MGIG}_d(\nu, A, B)$ is a unimodal distribution of which mode $M \in \mathbb{S}_{++}^d$ is the unique solution of the algebraic Riccati equation $(d + 1 - 2\,\nu)\,M + MBM = A$, and a standard importance sampling approach follows for the mean of the distribution. Our fallback solution is to solve this Riccati equation at each step and to replace all $\mathcal{MGIG}_d$ random variables by the (unique) mode of the consecutive distributions. To assess the credibility of this *ad hoc* sampling, the 'oracle' models in which Ω_y and the shrinkage parameters are known are added to the simulations. We will see that, despite an unavoidable loss, the results remain pretty consistent. In particular, the support recovery does not appear to be impacted.

5.1 A simulation study

In this section, the matrix of order $d \geqslant 1$ given by

$$C_d = \left(\rho^{|i-j|}\right)_{1 \leqslant i,j \leqslant d}$$

will be used as a typical covariance structure, for some $0 \leqslant \rho < 1$. Thus, the precision matrices will be chosen as a multiple of C_d^{-1} to keep the same guideline in our simulations. The responses

$$Y_k = B^t\,X_k + E_k$$

are generated through relations (1.1) where, for all $1 \leqslant k \leqslant n$, $E_k \sim \mathcal{N}_q(0, R)$. Because our models assume prior independence (or group-independence) in the columns of Δ,

it seems necessary to look at the influence of correlation among the predictors. So the standard choice $X_k \sim \mathcal{N}_p(0, I_p)$ is first considered, but in some cases we will also test $X_k \sim \mathcal{N}_p(0, C_p)$ for $\rho = 0.5$ and $\rho = 0.9$ to introduce a significant correlation between close predictors (see Figure 1). For each experiment, the support recovery of Δ is evaluated thanks to the so-called F-score given by

$$F = \frac{2\, p_r\, r_e}{p_r + r_e} \quad \text{where} \quad p_r = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad \text{and} \quad r_e = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

are the precision and the recall, respectively, and where T/F and P/N stand for true/false and positive/negative. To assess prediction skills, n_e randomly chosen observations are used for estimation (for different n_e) and the remaining $n_v = n - n_e = 100$ independent observations serve to compute the mean squared prediction error (MSPE). The results are compared to the ones obtained *via* the penalized maximum of likelihood (PML) approach of Yuan and Zhang (2014) thanks to the correctly adapted implementations of Chiquet et al. (2017) and Okome Obiang et al. (2021), with a cross-validated tuning parameter. In addition, we compute the sparse precision matrix estimations given by the graphical Lasso (GLasso) of Friedman et al. (2008), and by the CLIME algorithm of Cai et al. (2011), using the `R` packages `glasso` and `fastclime`, respectively. Note that we always keep a small value for q, so Δ is penalized but not Ω_y when possible (PML and GLasso). Finally, the recent approach of Ren et al. (2015), called ANT and based on the individual estimations of the partial correlations, is also implemented. Unlike PML, GLasso and CLIME, sparsity is not the result of penalizations for ANT but, instead, a threshold is deduced from the asymptotic normality of the estimates to decide which are significant and which can be set to zero. Let us add some preliminary comments about the methods compared in these simulations, all related to high-dimensional precision matrix estimation.

- There is a important advantage in favor of our Bayesian approaches, PML and ANT because they do not need the estimation of $\Omega_x \in \mathbb{S}^p_{++}$. Indeed, extracting the estimation of $\Delta \in \mathbb{R}^{q\times p}$ and $\Omega_y \in \mathbb{S}^q_{++}$ from that of the full precision matrix $\Omega \in \mathbb{S}^{q+p}_{++}$ may generate a drastic bias when $p \gg q$, and that explains in particular why GLasso and CLIME give pretty bad results in what follows.
- In its standard version, ANT is not designed to produce column-sparsity or group-sparsity in Δ. So, by considering multiple testing at the column or even group level, we allow groups of coefficients to be zeroed simultaneously. We have observed that this modified ANT method (called ANT* in the simulations) loses a bit in prediction quality but is greatly improved for support recovery.
- Unfortunately, this is not appropriate for PML, GLasso and CLIME. It is therefore not surprising that they are largely outperformed by our Bayesian models and ANT* for (gs) and (sgs). Using group-penalties, which to the best of our knowledge still does not exist, should improve the results of these methods to some extent.

The seven scenarios below, from Scen. 0 to Scen. 6, as heterogeneous as possible, represent the diversity of the situations (high-dimensionality, kind of sparsity, dimension

of the responses, coefficients hard to detect, etc.). We repeat each one $N = 100$ or $N = 50$ times, depending on the computation times involved, and the numerical results for $n_e = 400$ and uncorrelated predictors are summarized in Table 1. In addition, the evolution of MSPE is represented on Figure 1 for Scen. 1, 3 and 5, when n_e grows from 100 to 500, both for uncorrelated and correlated predictors. The three configurations (s), (gs) and (sgs) are tested on the grouped scenarios (from Scen. 3 to Scen. 6) with the adaptative shrinkage.

- *Scenario 0 (small dimension, no sparsity).* Let $q = 1$, $p = 5$ and set $\omega_y = 1$. We fill Δ with $\mathcal{N}(0, 2\,\omega_y)$ coefficients.
- *Scenario 1 (sparse direct links, univariate responses).* Let $q = 1$, $p = 50$ and set $\omega_y = 1$. We randomly choose 10 locations of Δ filled with $\mathcal{N}(0, \omega_y)$ coefficients while the others are zero.
- *Scenario 2 (sparse direct links, multivariate responses).* Let $q = 2$, $p = 80$ and set $\Omega_y = 2\,C_2^{-1}$ with $\rho = 0.5$. We randomly choose 10 columns of Δ filled with $\mathcal{N}_2(0, \Omega_y)$ coefficients while the others are zero.
- *Scenario 3 (group-sparse direct links, univariate responses).* Let $q = 1$, $p = 320$ and set $\omega_y = 1$. We consider $m = 5$ groups of size 100, 10, 100, 10 and 100. The two groups of size 10 are filled with $\mathcal{N}(0, 0.5\,\omega_y)$ and $\mathcal{N}(0, \omega_y)$ coefficients, respectively, while the other groups are zero.
- *Scenario 4 (group-sparse direct links, multivariate responses).* Let $q = 3$, $p = 500$ and set $\Omega_y = 3\,C_3^{-1}$ with $\rho = 0.5$. We divide the columns of Δ into $m = 25$ groups of size 20. We randomly choose 3 groups filled with $\mathcal{N}_3(0, 0.5\,\Omega_y)$, $\mathcal{N}_3(0, \Omega_y)$ and $\mathcal{N}_3(0, 1.5\,\Omega_y)$ coefficients, respectively, while the other groups are zero.
- *Scenario 5 (sparse-group-sparse direct links, univariate responses).* Let $q = 1$, $p = 150$ and set $\omega_y = 1$. We consider $m = 3$ groups of size 50. Only the second group is non-zero, into which we randomly fill 10 locations with $\mathcal{N}(0, \omega_y)$ coefficients.
- *Scenario 6 (sparse-group-sparse direct links, multivariate responses).* Let $q = 5$, $p = 1000$ and set $\Omega_y = 5\,C_5^{-1}$ with $\rho = 0.5$. We divide the columns of Δ into $m = 20$ groups of size 50, and a randomly chosen one is half filled with $\mathcal{N}_5(0, \Omega_y)$ coefficients. The others columns of Δ are zero.

Now, let us try to summarize our observations. In terms of support recovery, the Bayesian spike-and-slab framework and the modified ANT* method give results incomparably better than the sparsity-inducing penalized approaches (PML, GLasso and CLIME). As suggested in Rem. 3.3 of Okome Obiang et al. (2021), this may be a consequence of the fact that the cross-validation steps calibrate the models to reach the best prediction error, sometimes at the cost of support recovery by picking a small penalty level. The superiority of ANT over GLasso and CLIME is recognized and discussed in Ren et al. (2015), but this also highlights the ability of our Bayesian models to reach good results both in prediction and in support recovery. It can also be seen that (s)

Scenario 0

Mod.	Shr.	MSPE	F	p_r	r_e
(s-or)	-	1.01 (0.11)	1.00	1.00	1.00
(s)	(ad)	1.03 (0.13)	1.00	1.00	1.00
(s)	(gl)	1.03 (0.13)	1.00	1.00	1.00
PML	-	1.01 (0.16)	1.00	1.00	1.00
GLasso	-	1.00 (0.15)	1.00	1.00	1.00
CLIME	-	1.00 (0.15)	1.00	1.00	1.00
ANT*	-	1.04 (0.13)	1.00	1.00	1.00

Hyperparam. $\pi = 0$

Scenario 1

Mod.	Shr.	MSPE	F	p_r	r_e
(s-or)	-	1.02 (0.13)	0.95	1.00	0.90
(s)	(ad)	1.04 (0.13)	0.95	1.00	0.90
(s)	(gl)	1.03 (0.13)	0.95	1.00	0.90
PML	-	1.08 (0.15)	0.82	0.69	1.00
GLasso	-	2.37 (0.96)	0.78	0.77	0.80
CLIME	-	2.52 (0.98)	0.79	0.78	0.80
ANT*	-	1.25 (0.22)	0.87	0.85	0.90

Hyperparam. $(25, 1)$

Scenario 2

Mod.	Shr.	MSPE	F	p_r	r_e
(s-or)	-	0.52 (0.09)	0.95	1.00	0.90
(s)	(ad)	0.54 (0.09)	0.95	1.00	0.90
(s)	(gl)	0.55 (0.08)	0.95	1.00	0.90
PML	-	0.77 (0.15)	0.86	1.00	0.75
GLasso	-	1.74 (0.49)	0.72	0.91	0.60
CLIME	-	1.11 (0.35)	0.73	0.76	0.70
ANT*	-	1.04 (0.44)	0.90	0.89	0.91

Hyperparam. $(80, 1)$

Scenario 3

Mod.	Shr.	MSPE	F	p_r	r_e
(gs-or)	-	1.03 (0.27)	1.00	1.00	1.00
(gs)	(ad)	1.04 (0.27)	1.00	1.00	1.00
(gs)	(gl)	1.04 (0.34)	1.00	1.00	1.00
(s)	(ad)	1.16 (0.27)	0.92	1.00	0.85
(sgs)	(ad)	1.07 (0.25)	0.92	1.00	0.86
PML	-	1.80 (0.36)	0.89	1.00	0.80
GLasso	-	4.23 (1.61)	0.58	0.50	0.70
CLIME	-	2.98 (1.22)	0.68	0.90	0.55
ANT*	-	1.52 (0.95)	1.00	1.00	1.00

Hyperparam. $(100, 1) - (5, 1) - (5, 1, 25, 1)$

Scenario 4

Mod.	Shr.	MSPE	F	p_r	r_e
(gs-or)	-	0.40 (0.14)	1.00	1.00	1.00
(gs)	(ad)	0.45 (0.16)	1.00	1.00	1.00
(gs)	(gl)	0.46 (0.17)	1.00	1.00	1.00
(s)	(ad)	0.52 (0.18)	0.98	1.00	0.96
(sgs)	(ad)	0.48 (0.17)	0.99	1.00	0.98
PML	-	3.18 (0.53)	0.75	0.94	0.62
GLasso	-	9.46 (1.38)	0.46	0.66	0.35
CLIME	-	8.32 (1.51)	0.48	0.45	0.52
ANT*	-	6.53 (1.22)	1.00	1.00	1.00

Hyperparam. $(100, 1) - (25, 1) - (50, 1, 50, 1)$

Scenario 5

Mod.	Shr.	MSPE	F	p_r	r_e
(sgs-or)	-	1.00 (0.15)	0.96	1.00	0.92
(sgs)	(ad)	1.04 (0.16)	0.95	1.00	0.91
(sgs)	(gl)	1.03 (0.16)	0.91	1.00	0.84
(s)	(ad)	1.08 (0.14)	0.93	1.00	0.87
(gs)	(ad)	1.24 (0.19)	0.33	0.20	1.00
PML	-	1.92 (0.60)	0.89	1.00	0.80
GLasso	-	3.48 (1.30)	0.78	0.86	0.71
CLIME	-	1.88 (0.92)	0.79	1.00	0.65
ANT*	-	1.26 (0.98)	0.88	0.86	0.90

Hyperparam. $(50, 1) - (3, 1) - (3, 1, 50, 1)$

Scenario 6

Mod.	Shr.	MSPE	F	p_r	r_e
(sgs-or)	-	0.21 (0.13)	1.00	1.00	1.00
(sgs)	(ad)	0.24 (0.32)	1.00	1.00	1.00
(sgs)	(gl)	0.24 (0.33)	1.00	1.00	1.00
(s)	(ad)	0.29 (0.26)	0.98	1.00	0.96
(gs)	(ad)	0.31 (0.30)	0.67	0.50	1.00
PML	-	0.50 (0.17)	0.83	0.95	0.74
GLasso	-	3.83 (0.77)	0.50	0.97	0.34
CLIME	-	2.98 (0.51)	0.51	1.00	0.34
ANT*	-	2.10 (0.72)	1.00	1.00	1.00

Hyperparam. $(100, 1) - (20, 1) - (20, 1, 50, 1)$

Table 1: Medians of the mean squared prediction errors (with standard deviations), F-scores, precisions and recalls after $N = 100$ repetitions of Scen. 0 to Scen. 6 ($N = 50$ for Scen. 4 and Scen. 6), with $n_e = 400$ and uncorrelated predictors. The suffix -or is used to denote 'oracle' settings. The hyperparameters chosen for the prior spike probability are indicated in the last row of each table, from left to right: (a, b) for (s) and (gs), (a_1, b_1, a_2, b_2) for (sgs).

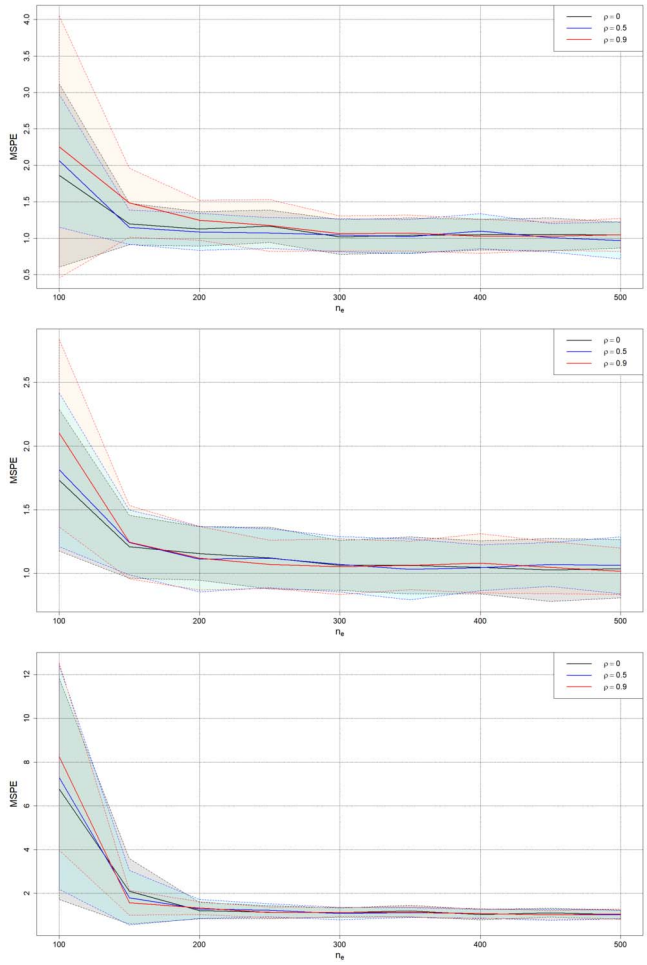

Figure 1: Medians of the mean squared prediction errors obtained after $N = 100$ repetitions of Scen. 1 (top), Scen. 3 (middle) and Scen. 5 (bottom) with ± 1 standard deviation and n_e growing from 100 to 500. The black curves correspond to uncorrelated predictors ($\rho = 0$) while the blue and red curves correspond to correlated predictors ($\rho = 0.5$ and $\rho = 0.9$, respectively).

gives weaker results than (sgs) in the grouped scenarios, probably due to the fact that it does not take into account the group structure, but still better than the penalized methods. However, the computational times involved (see remarks below) make (s) less relevant than (sgs) in these situations, even if the results are not drastically different. Unsurprisingly, (gs) is not suitable in the sparse-group-sparse settings in terms of support recovery. Our experiments show that it is able to identify influential groups without being mistaken but, even though the resulting estimates are small where they should

Chr.	1	2	3	4	5	6	7	8	9	10
Nb.	74	67	63	60	39	45	52	43	31	51
Chr.	11	12	13	14	15	16	17	18	19	20
Nb.	21	26	33	22	15	27	18	30	34	19

Table 2: Number of markers on each chromosome, which correspond to the sizes κ_g of each group for $1 \leqslant g \leqslant 20$ when running (gs) and (sgs).

be zero, it is not designed to be used for bi-level selection. Figure 1 shows that the results are pretty stable from $n_e = 200$ observations in the learning set: for $n_e < 200$ the MSPEs are rather chaotic before stabilizing. The same figure also highlights that the presence of correlation in the predictors does not seem to have a significant effect on the estimation procedure, except for small size samples and high correlation where the degradation is noticeable. Overall, the real strength of the Bayesian spike-and-slab approach is clearly the support recovery of the direct links between predictors and responses but it seems that one can hardly expect to deal with very high-dimensional studies as long as we do not impose a group structure or a huge degree of sparsity. The highly competitive MSPEs obtained confirm the relevance of Bayesian PGGMs not only for variable selection but also for prediction purposes in the context of high-dimensional regressions.

5.2 Identification of a sparse set of predictors in a real dataset

Let us now study the `Hopx` dataset, fully described in Petretto et al. (2010). It contains $p = 770$ genetic markers spread over $m = 20$ chromosomes from $n = 29$ inbred rats. It also contains the corresponding measured gene expression levels of $q = 4$ tissues (adrenal gland, fat, heart and kidney). The goal is to identify a sparse set of predictors that jointly explain the outcomes, with the natural group structure formed by chromosomes (see Table 2). This dataset has already been analyzed in Liquet et al. (2016), using a Bayesian regression without group structure, and later in Liquet et al. (2017) including group and sparse-group structures. So the PGGM is supposed to bring new perspectives about relationships in terms of partial correlations. A particularity of this dataset is that the responses are very correlated, so we should expect an estimation of Ω_y^{-1} with significant non-diagonal elements and a clear advantage in using PGGMs. Indeed, a predictor considered to be influencing all the outcomes could be the result of a direct relation to one tissue propagated to the others by an artificial correlation effect. As can be seen on Figure 2, the predictors are also highly correlated with their neighbors (for the sake of readability, we only represent the correlogram of predictors located on chromosomes 8, 9 and 10).

The small sample size relative to the number of covariates (29/770) weakens the study. To strengthen our conclusions, we decided to run $N = 100$ experiments based on 25 randomly chosen observations and to aggregate the results. We first investigate the selection of predictors at the chromosomes scale, *i.e.* we run (gs) according to the previous protocol with an adaptative shrinkage and we choose $(a, b) = (1, 20)$ in the prior probability π. The empirical distribution of the posterior probability of inclusion for each

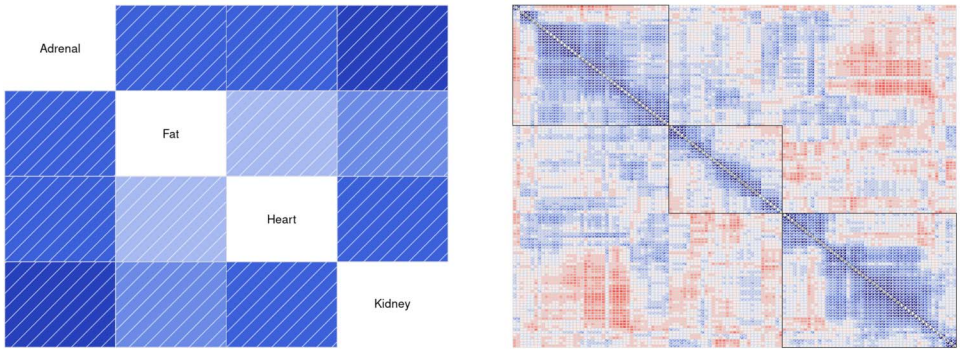

Figure 2: Correlogram of responses (left) and correlogram of predictors located on chromosomes 8, 9 and 10 (right). The colormap associates red with negative correlations and blue with positive correlations.

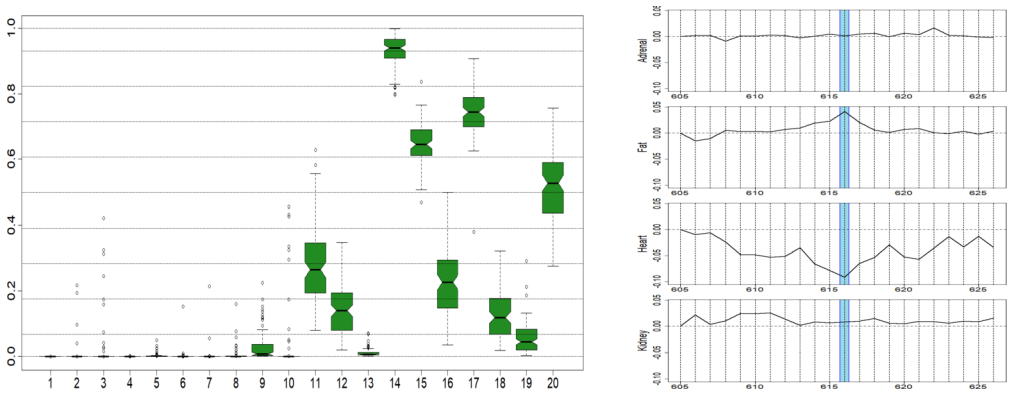

Figure 3: Empirical distribution of the posterior probability of inclusion estimated by (gs) for each chromosome (left). Aggregated (gs) estimation of Δ on chromosome 14 with `D14Mit3` highlighted (right).

chromosome is represented on the left of Figure 3. The selection procedure focuses on chromosomes 14, 15 and 17 (and not just on chromosomes 2 and 3 as in Liquet et al. (2017)) but the estimation process gives an overwhelming advantage to chromosome 14, far ahead of its neighbors. This is undoubtedly the influence of `D14Mit3`, a marker located on chromosome 14 and known to have a very significant effect on this dataset. The main conclusion to be drawn at this stage is that chromosome 14 has a positive effect on `Fat` and a *negative* effect on `Heart`, as can also be seen on the right of Figure 3. Therefore, it is likely that the overall positive influence of `D14Mit3` identified by previous authors is due to the combination of a direct positive link with `Fat`, a direct negative link with `Heart` and a correlation effect from the outcomes. This hypothesis is given additional credibility by the numerical results: from (gs), the corresponding column

of Δ is approximately $(0.00, 0.04, -0.09, 0.00)$ which, through relations (1.1), leads to $(0.15, 0.25, 0.34, 0.21)$ as estimated regression coefficients. This roughly corresponds to the values indicated in Tab. 2 of Liquet et al. (2017), at least for the main effect on `Heart`. Thus for chromosome 14, the numerical results coincide but the interpretations are clearly different. Of course, similar reasonings can be carried out for the less influent chromosomes.

It is perhaps more interesting to look for a bi-level selection in order to identify a sparse set of markers and not only chromosomes. In this regard, (sgs) is launched using the same statistical protocol, adaptative shrinkage and hardly informative hyperparameters $a_1 = 3$, $b_1 = 1$, $a_2 = 1$ and $b_2 = 1$ which happen to be sufficient to generate a huge degree of sparsity. While many chromosomes are excluded from the model given by (gs), with (sgs) we see some contributions localized in certain chromosomes having little influence when taken as a whole. At the markers scale, the randomness of the sampler and the high level of correlation between close predictors probably explain the presence of artifacts which sometimes make it difficult to distinguish the real contributions from the background noise. We therefore use the $N = 100$ experiments to build 95% confidence intervals and keep only significant estimates. By way of example, Figure 4 displays the results obtained on chromosomes 7, 8 and 14. The main markers standing out are summarized in Table 3 together with the kind of direct influences detected. Markers already highlighted in Liquet et al. (2016) or Liquet et al. (2017) are also indicated. One can see that most of our conclusions coincide, but new markers are suggested (especially on chromosome 8) and others have disappeared. Overall, the more stringent statistical protocol that we used led to the retention of fewer predictors with more guarantee. An important consequence of this study is the new interpretations in terms of direct influences allowed by PGGMs. Especially as the residual correlations, hidden in the estimation of $R = \Omega_y^{-1}$ and closely related to the correlations between the responses, are very high (greater than 0.7), as we suspected from Figure 2.

5.3 Discussion and conclusion

To conclude, we would like to draw the attention of the reader to some weaknesses of the study, still under investigation. On the one hand, as soon as p is large (say, $p \geqslant 500$), the Bayesian studies should be conducted with a group structure or by promoting very sparse settings because due to the outline of the sampler, looping over each column of Δ may quickly become intractable. A group structure limits the number of loops (only $m \ll p$ per sampler iteration), although each loop may require the generation of large Gaussian vectors (up to $(q \times \kappa_g)$-dimensional), so compromises are needed. Subdividing the dataset is natural when it is intrinsically equipped with a group structure (*e.g.* that of the previous section), we could suggest otherwise a clustering of the set of predictors to gather similar entries and control the size of the groups. At this stage, our procedures cannot compete with the Lasso-type algorithms (GLasso, CLIME or even ANT) in terms of computational times. This is an issue on which future studies should focus (ongoing works are devoted to translating the samplers into more efficient environments), enhanced MCMC methods may also be useful or novel computational strategies like the 'shotgun' stochastic algorithm of Yang and Narisetty (2020). On

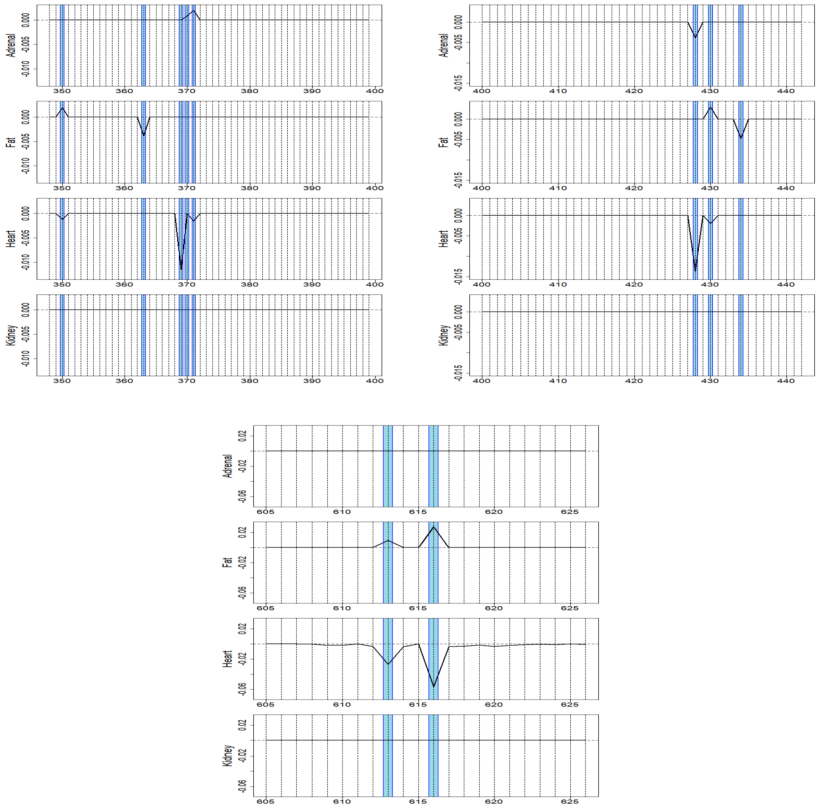

Figure 4: Aggregated (sgs) estimation of Δ on chromosomes 7, 8 and 14, from left to right and bottom. The highlighted markers are `D7Cebr205s3`, `D7Mit6`, `D7Rat19`, `Myc` and `D7Rat17` for chromosome 7, `D8Mgh4`, `D8Rat135` and `Rbp2` for chromosome 8 and `D14Rat8` and `D14Mit3` for chromosome 14.

the other hand, the procedures are obviously very sensitive to the initialization of the sampler, especially when $p \gg n$. For example, the term $|I_{\kappa_g} + \lambda_g \underline{\underline{\mathbb{X}}}_g^t \underline{\underline{\mathbb{X}}}_g|$ is likely to explode when κ_g is large and $\lambda_g > 1$, that is why λ_g has to be carefully controlled *via* an accurate initial choice of ℓ_g. Our heuristic approach is to initialize ℓ_g such that $\mathbb{E}[\lambda_g] < 1$ to control the behavior of $|I_{\kappa_g} + \lambda_g \underline{\underline{\mathbb{X}}}_g^t \underline{\underline{\mathbb{X}}}_g|$ during the first iterations. This works pretty well in practice, but needs to be done on a case-by-case basis, which could be improved. From a theoretical point of view, we should obviously enhance the estimation procedure by sampling from the $\mathcal{MGIG}_q$ distribution for $q > 1$, and not using the mode. Our fallback solution gives satisfactory but not completely rigorous results. In addition, it could be interesting to generalize the support recovery guarantee of Proposition 3.2 to (sgs), which is certainly possible at the cost of a few additional developments. Overall, our study shows that for the moderate values of p (up to 10^3 or 10^4), the Bayesian approach of the partial Gaussian graphical models is a very serious alternative to the

Chromosomes	Markers	Main direct influences
3	D3Mit16*	Adrenal+ Heart−
7	D7Cebr205s3*	Fat+ Heart−
	D7Mit6*	Fat−
	D7Rat19*	Heart−
	Myc*	Adrenal+
	D7Rat17	Adrenal+ Heart−
8	D8Mgh4	Adrenal− Heart−
	D8Rat135	Fat+ Heart−
	Rbp2	Fat−
10	D10Rat33*	Adrenal+
	D10Mit3*	Adrenal+
	D10Rat31*	Fat−
11	D11Rat47	Fat−
14	D14Rat8*	Fat+ Heart−
	D14Mit3*	Fat+ Heart−
15	D15Cebr7s13	Kidney−
	D15Rat21*	Adrenal+ Kidney−
17	Prl	Adrenal− Kidney−
20	D20Rat55	Kidney−

Table 3: Main relations detected by (sgs). X* means that marker X has already been suggested by previous authors in this dataset. Y− (Y+) means that response Y is negatively (positively) influenced by X.

frequentist penalized estimations, for prediction but also and especially for support recovery.

Supplementary Material

A Bayesian approach for partial Gaussian graphical models with sparsity (supplementary material) (DOI: 10.1214/22-BA1315SUPP; .pdf). Single-output corollaries and computational steps related to posterior distributions.

References

Bai, R., Moran, G. E., Antonelli, J. L., Chen, Y., and Boland, M. R. (2020). "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models." *J. Am. Stat. Assoc.*, 1–14. MR4399078. doi: https://doi.org/10.1080/01621459.2020.1765784. 467

Banerjee, O., El Ghaoui, L., and D'Aspremont, A. (2008). "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data." *J. Mach. Learn. Res.*, 9: 485–516. MR2417243. 466

Brown, P. J., Vannucci, M., and Fearn, T. (1998). "Multivariate Bayesian variable selection and prediction." *J. R. Statist. Soc. B.*, 60(3): 627–641. MR1626005. doi: https://doi.org/10.1111/1467-9868.00144. 478

Cai, T., Liu, W., and Luo, X. (2011). "A Constrained ℓ_1 Minimization Approach to Sparse Precision Matrix Estimation." *J. Am. Stat. Assoc.*, 106(494): 594–607. MR2847973. doi: https://doi.org/10.1198/jasa.2011.tm10155. 466, 479

Cai, T. and Zhou, H. (2012). "Optimal rates of convergence for sparse covariance matrix estimation." *Ann. Stat.*, 40(5): 2389–2420. MR3097607. doi: https://doi.org/10.1214/12-AOS998. 466

Chiquet, J., Mary-Huard, T., and Robin, S. (2017). "Structured regularization for conditional Gaussian graphical models." *Stat. Comput.*, 27(3): 789–804. MR3613598. doi: https://doi.org/10.1007/s11222-016-9654-1. 467, 479

Eltoft, T., Kim, T., and Lee, T. (2006). "Multivariate scale mixture of Gaussians modeling." In *Independent Component Analysis and Blind Signal Separation*, 799–806. Springer Berlin Heidelberg. 470

Fang, Y., Karlis, D., and Subedi, S. (2020). "A Bayesian approach for clustering skewed data using mixtures of multivariate normal-inverse Gaussian distributions." *arXiv:2005.02585*. 478

Fazayeli, F. and Banerjee, A. (2016). *The Matrix Generalized Inverse Gaussian distribution: properties and applications*, volume 9851 of *Frasconi P., Landwehr N., Manco G., Vreeken J. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2016. Lecture Notes in Computer Science*. Springer, Cham. 478

Friedman, J., Hastie, T., and Tibshirani, R. (2008). "Sparse inverse covariance estimation with the graphical Lasso." *Biostatistics.*, 9(3): 432–441. 466, 467, 479

Gan, L., Yang, X., Narisetty, N., and Liang, F. (2019). "Bayesian Joint Estimation of Multiple Graphical Models." In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc. 466

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. MR3307991. 465, 466, 468

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman & Hall/CRC Monographs on Statistics and Applied Probability. CRC Press. MR3616141. 470

Li, Y., Nan, B., and Zhu, J. (2015). "Multivariate sparse group lasso for the multivariate multiple linear regression with an arbitrary group structure." *Biometrics.*, 71: 354–363. MR3366240. doi: https://doi.org/10.1111/biom.12292. 466

Li, Z., Mccormick, T., and Clark, S. (2019). "Bayesian Joint Spike-and-Slab Graphical Lasso." In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3877–3885. PMLR. 466

Liquet, B., Bottolo, L., Campanella, G., Richardson, S., and Chadeau-Hyam, M. (2016). "R2GUESS: A Graphics Processing Unit-Based R Package for Bayesian Variable Selection Regression of Multivariate Responses." *J. Stat. Softw.*, 69(2): 1–32. MR3724978. doi: https://doi.org/10.1214/17-BA1081. 483, 485

Liquet, B., Mengersen, K., Pettitt, A. N., and Sutton, M. (2017). "Bayesian Variable Selection Regression of Multivariate Responses for Group Data." *Bayesian Anal.*, 12(4): 1039–1067. MR3724978. doi: https://doi.org/10.1214/17-BA1081. 467, 468, 470, 475, 477, 483, 484, 485

Maathuis, M., Drton, M., Lauritzen, S. L., and Wainwright, M. (2018). *Handbook of Graphical Models*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. MR3889064. 465, 466

Massam, H. and Wesołowski, J. (2006). "The Matsumoto-Yor property and the structure of the Wishart distribution." *J. Multivariate. Anal.*, 97: 103–123. MR2208845. doi: https://doi.org/10.1016/j.jmva.2004.11.008. 478

Meinshausen, N. and Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the Lasso." *Ann. Stat.*, 34(3): 1436–1462. MR2278363. doi: https://doi.org/10.1214/009053606000000281. 466

Okome Obiang, E., Jézéquel, P., and Proïa, F. (2021). "A partial graphical model with a structural prior on the direct links between predictors and responses." *ESAIM Probab. Stat.*, 25: 298–324. MR4283606. doi: https://doi.org/10.1051/ps/2021010. 467, 479, 480

Okome Obiang, E., Jézéquel, P., and Proïa, F. (2022). "Supplementary Material for "A Bayesian approach for partial Gaussian graphical models with sparsity"." *Bayesian Analysis*. doi: https://doi.org/10.1214/22-BA1315SUPP. 468

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *J. Am. Stat. Assoc.*, 103(482): 681–686. MR2524001. doi: https://doi.org/10.1198/016214508000000337. 470, 477

Petretto, E., Bottolo, L., Langley, S. R., Heinig, M., McDermott-Roe, C., Sarwar, R., Pravenec, M., Hübner, N., Aitman, T. J., Cook, S. A., and Richardson, S. (2010). "New Insights into the Genetic Control of Gene Expression using a Bayesian Multitissue Approach." *PLOS Comput. Biol.*, 6(4): 1–13. 483

Ravikumar, P., Wainwright, M., Raskutti, G., and Yu, B. (2011). "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence." *Electron. J. Stat.*, 5: 935–980. MR2836766. doi: https://doi.org/10.1214/11-EJS631. 466

Ren, Z., Sun, T., Zhang, C. H., and Zhou, H. H. (2015). "Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Models." *Ann. Stat.*, 43(3): 991–1026. MR3346695. doi: https://doi.org/10.1214/14-AOS1286. 466, 474, 479, 480

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). "Sparse permutation invariant covariance estimation." *Electron. J. Stat.*, 2: 494–515. MR2417391. doi: https://doi.org/10.1214/08-EJS176. 466

Sohn, K. A. and Kim, S. (2012). "Joint estimation of structured sparsity and output structure in multiple-output regression via inverse-covariance regularization." In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics.*, volume 22 of *Proceedings of Machine Learning Research*, 1081–1089. PMLR. 466

Wei, R., Reich, B. J., Hoppin, J. A., and Ghosal, S. (2020). "Sparse Bayesian additive nonparametric regression with application to health effects of pesticides mixtures." *Statist. Sinica*, 30: 55–79. MR4285485. doi: https://doi.org/10.5705/ss.202017.0315. 467

Xu, X. and Ghosh, M. (2015). "Bayesian variable selection and estimation for Group Lasso." *Bayesian Anal.*, 10(4): 909–936. MR3432244. doi: https://doi.org/10.1214/14-BA929. 467, 475, 477

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). "Bayesian grouped horseshoe regression with application to additive models." In *AI 2016: Advances in Artificial Intelligence*, 229–240. Springer International Publishing. MR3595648. doi: https://doi.org/10.1007/978-3-319-50127-7_19. 467

Yang, X. and Narisetty, N. (2020). "Consistent Group Selection with Bayesian High Dimensional Modeling." *Bayesian Anal.*, 15(3): 909–935. MR4132654. doi: https://doi.org/10.1214/19-BA1178. 467, 468, 473, 474, 485

Yuan, M. and Lin, Y. (2007). "Model selection and estimation in the Gaussian graphical model." *Biometrika.*, 94(1): 19–35. MR2367824. doi: https://doi.org/10.1093/biomet/asm018. 466

Yuan, X. T. and Zhang, T. (2014). "Partial Gaussian graphical model estimation." *IEEE. T. Inform. Theory.*, 60(3): 1673–1687. MR3168429. doi: https://doi.org/10.1109/TIT.2013.2296784. 466, 467, 479

Acknowledgments

The authors warmly thank the two anonymous reviewers for their valuable suggestions and comments which clearly contributed to the improvement of the article. The authors thank Mario Campone (project leader and director of the ICO), Mathilde Colombié (scientific coordinator of EPICURE clinical trial) and Fadwa Ben Azzouz, biomathematician in Bioinfomics, for the initiation, the coordination and the smooth running of the project.