Electronic Journal of Statistics
Vol. 18 (2024) 1–46
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2196

# Precision matrix estimation under the horseshoe-like prior–penalty dual

**Ksheera Sagar**

*Department of Statistics, Purdue University,*
*150 N. University St.,*
*West Lafayette, Indiana 47907, U.S.A.*
*e-mail:* kkeralap@purdue.edu

**Sayantan Banerjee***

*Operations Management & Quantitative Techniques Area,*
*Indian Institute of Management Indore,*
*Rau-Pithampur Road, Indore, Madhya Pradesh 453556, India*
*e-mail:* sayantanb@iimidr.ac.in

**Jyotishka Datta**

*Department of Statistics, Virginia Tech,*
*250 Drillfield Drive,*
*Blacksburg, Virginia 24061, U.S.A.*
*e-mail:* jyotishka@vt.edu

**Anindya Bhadra**

*Department of Statistics, Purdue University,*
*150 N. University St.,*
*West Lafayette, Indiana 47907, U.S.A.*
*e-mail:* bhadra@purdue.edu

**Abstract:** Precision matrix estimation in a multivariate Gaussian model is fundamental to network estimation. Although there exist both Bayesian and frequentist approaches to this, it is difficult to obtain good Bayesian and frequentist properties under the same prior–penalty dual. To bridge this gap, our contribution is a novel prior–penalty dual that closely approximates the graphical horseshoe prior and penalty, and performs well in both Bayesian and frequentist senses. A chief difficulty with the graphical horseshoe prior is a lack of closed form expression of the density function, which we overcome in this article. In terms of theory, we establish posterior convergence rate of the precision matrix that matches the convergence rate of the frequentist graphical lasso estimator, in addition to the frequentist consistency of the MAP estimator at the same rate. In addition, our results also provide theoretical justifications for previously developed approaches that have been unexplored so far, e.g. for the graphical horseshoe prior. Computationally efficient EM and MCMC algorithms are developed respectively for the penalized likelihood and fully Bayesian estimation problems. In numerical experiments, the horseshoe-based approaches echo their superior theoretical properties by comprehensively outperforming the competing methods. A protein–protein interaction network estimation in B-cell lymphoma is considered to validate the proposed methodology.

*Corresponding author.

**MSC2020 subject classifications:** Primary 62H12; secondary 62F12, 62F15.
**Keywords and phrases:** Graphical models, non-convex optimization, posterior concentration, posterior consistency, sparsity.

Received March 2023.

## Contents



## 1. Introduction

High-dimensional precision matrix estimation under a multivariate normal model is a fundamental building block for network estimation, and a common thread connecting disparate applications such as inference on gene regulatory networks [28], econometrics [20, 12], and neuroscience [49]. The frequentist solution to this problem is now relatively well understood and several useful algorithms exist; see Pourahmadi [44] for a detailed review. However, interested readers will quickly

discern that the Bayesian literature on this problem is still sparse, barring some notable exceptions described in Section 1.1. The reason for this is simple: the focus of a fully Bayesian analysis is on the entire posterior and quantification of uncertainty using the said posterior; a problem fundamentally more demanding computationally. Consequently, the virtues of probabilistic uncertainty quantification notwithstanding, the Bayesian treatment to precision matrix estimation has received relatively scant attention from practitioners. Furthermore, a penalized likelihood estimate with good frequentist properties need not correspond to good Bayesian posterior concentration properties under the corresponding prior. A notable example of this in linear regression models is the lasso penalty [52], and its Bayesian counterpart using the double exponential prior [42], for which Castillo et al. [15] assert: *"the LASSO is essentially non-Bayesian, in the sense that the corresponding full posterior distribution is a useless object."* We address this gap in the literature in the context of graphical models. Our contribution is a novel prior–penalty dual that makes both fully Bayesian and fast penalized likelihood estimation feasible. The key distinguishing feature of our work is that we provide theoretical and empirical support for *both* Bayesian and frequentist solutions to the problem under the same prior–penalty dual. It is shown that the Bayesian posterior as a whole concentrates around the truth and the penalized likelihood point estimate is consistent. To our knowledge, ours is the first work to establish these results using continuous shrinkage priors under an *arbitrary sparsity pattern* in the true precision matrix. This is at a contrast to the current state of the art in theory that imposes additional constraints on the graph, e.g., banded-ness or more general decomposable structures [2, 61, 37, 35]. Typically these assumptions are made for computational and theoretical tractability rather than any intrinsic subject matter knowledge. The reason our work is able to avoid these restrictive assumptions is because we work with continuous "global-local" shrinkage priors and impose sparsity in a weak sense [see, e.g., 7].

The motivating data set arises from a biological application. Protein–protein interaction networks have been found to play a crucial role in cancer [25]. One such significant effort in this direction is "The Cancer Genome Atlas" program [60] that has collected data from over 7,700 patients across 32 different tumor types. From this repository, we retrieve proteomic data of 33 patients with "Lymphoid Neoplasm Diffuse Large B-cell Lymphoma," which is a cancer that starts in white blood cells and spreads to lymph nodes. Our findings are contrasted with that of Ha et al. [25].

### *1.1. The current state of the art and our contributions in context*

A Gaussian graphical model (GGM) remains popular as a fundamental building block for network estimation because of the ease of interpretation of the resulting precision matrix estimate: an inferred off-diagonal zero corresponds to conditional independence of the two corresponding nodes given the rest [see, e.g., 32].

Among the most popular frequentist approaches for estimating GGMs are the graphical lasso [21] and the graphical SCAD [17], which are respectively the

graphical extensions of the lasso [52] and SCAD [18] penalties in linear models. Similarly, the CLIME estimator of Cai et al. [11] is a graphical application of the Dantzig penalty [13]. Fan et al. [20] propose factor-based models for estimation of precision matrices, which are particularly attractive in financial applications where the precision matrix of outcome variables conditioned on some common factors being sparse is a sensible assumption. Alternatively, Callot et al. [12] opt for a node-wise regression approach using $\ell_1$ penalty for minimizing the risk of a Markowitz portfolio. The positive definiteness of their estimate is guaranteed asymptotically, which nevertheless remains hard to establish in finite samples; a common issue with node-wise regression approaches. Zhang and Zou [64] propose a new empirical loss termed the *D-trace loss* to avoid computing the log determinant term in the $\ell_1$ penalized loss. Under certain conditions they also prove that the resulting estimate is identical to the CLIME estimate [11]. A ridge type estimate for precision matrix termed ROPE is proposed by Kuismin et al. [30], who use the squared Frobenius norm of the precision matrix as a penalty function; see Van Wieringen and Peeters [55] for other kinds of ridge estimators. Another distribution free version of the ridge estimate is proposed by Wang et al. [56]. An elastic net penalty [65] is used to determine the functional connectivity among brain regions by Ryali et al. [49]. A comprehensive theoretical treatment for the rate of convergence of precision matrix estimates is given by Lam and Fan [31].

The frequentist approaches listed above generally enjoy faster and more scalable computation, owing to being point estimates. Nevertheless, from a Bayesian perspective, a common theme with these penalized approaches is that the posterior concentration properties of the corresponding priors remain completely unexplored. Bayesian methods for GGMs initially explored *structured precision matrices*, assuming an approximate banded-structure of the underlying graph, or more generally, a decomposable graph; see, for example, [2, 61, 35]. For arbitrary graphs including both decomposable and non-decomposable graph structures, the $G$-Wishart prior has been developed as a conjugate prior for the precision matrix under a Gaussian framework; see, for example, [47, 48, 37, 53], and references therein. For a comprehensive review of Bayesian methods for GGMs under structured sparsity along with computational approaches, we refer the readers to Banerjee et al. [1]. Moving now to Bayesian methodologies for *unstructured precision matrices*, the literature is relatively scant. Wang [57] proposes a Bayesian version of the graphical lasso and uses a clever decomposition of the precision matrix to facilitate block Gibbs sampling and to guarantee the positive definiteness of the resulting estimate. Banerjee and Ghosal [3] consider a similar prior structure as the Bayesian graphical lasso, with the exception that they put a large point-mass at zero for the off-diagonal elements of the precision matrix. Under assumptions of sparsity, they derive posterior convergence rates in the Frobenius norm, and also provide a Laplace approximation method for computing marginal posterior probabilities of models. Spike-and-slab variants with double exponential priors are proposed by Gan et al. [23], Wang [59]. A common issue with the spike-and-slab approach is the presence of binary indicator variables, which typically hinder posterior exploration and the Bayesian lasso

estimate is known to be biased for large signals [15]. Both of these issues are addressed by the graphical horseshoe estimate proposed by Li et al. [36], which is an application of the popular global-local horseshoe prior [14] in GGMs. Li et al. [36] provide considerable empirical evidence of superior performance over several competing Bayesian and frequentist approaches. Nevertheless, their theoretical results are limited to upper bounds on some Kullback–Leibler risk properties and the bias of the resulting estimate. Consequently, whether the graphical horseshoe posterior has correct concentration properties has remained an open question. Similarly, its frequentist dual: the penalized likelihood estimate, also remains unavailable, mainly because there is no closed form of the horseshoe prior or penalty; only a normal scale mixture representation. Both of these issues are resolved in the present paper. We propose a novel prior–penalty dual that closely approximates the graphical horseshoe prior with the density being available explicitly as well as a normal scale mixture, which has important implications in theory and in practice, and in both Bayesian and frequentist settings. Moreover, as a corollary to one of our main results, the posterior concentration properties of the graphical horseshoe are also established, for the first time.

### *1.2. Notations and preliminaries*

For positive real-valued sequences $\{a_n\}$ and $\{b_n\}$, $a_n = O(b_n)$ means that $a_n/b_n$ is bounded, and $a_n = o(b_n)$ means that $a_n/b_n \to 0$ as $n \to \infty$; $a_n \lesssim b_n$ implies that $a_n = O(b_n)$, and $a_n \asymp b_n$ means that both $a_n = O(b_n)$ and $b_n = O(a_n)$ hold. For a sequence of random variables $\{X_n\}$, $X_n = O_P(\epsilon_n)$ means that $\mathrm{P}(|X_n| \leq M\epsilon_n) \to 1$ for some constant $M > 0$.

Vectors are represented in bold lowercase English or Greek letters, with corresponding components denoted by non-bold letters, for example, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. For a vector $\boldsymbol{x} \in \mathbb{R}^p$, the $L_r$-norm, for $0 < r < \infty$, is defined as $\|\boldsymbol{x}\|_r = (\sum_{i=1}^p |x_j|^r)^{1/r}$, and the $L_\infty$-norm is defined as $\|\boldsymbol{x}\|_\infty = \max_{1\leq j\leq p} |x_j|$. The zero-vector is denoted by $\boldsymbol{0}$. Matrices are represented in bold uppercase English or Greek letters, for example, $\boldsymbol{A} = ((a_{ij}))$, where $a_{ij}$ denotes the $(i, j)$th entry of $\boldsymbol{A}$. We denote the identity matrix by $\boldsymbol{I}_p$. For a symmetric matrix $\boldsymbol{A}$, $\mathrm{eig}_1(\boldsymbol{A}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{A})$ denote the ordered eigenvalues of $\boldsymbol{A}$, and its trace and determinant are denoted by $\mathrm{tr}(\boldsymbol{A})$ and $\det \boldsymbol{A}$ respectively. The $L_r$ and $L_\infty$-norms on $p \times p$ matrices are respectively defined as $\|\boldsymbol{A}\|_r = (\sum_{i=1}^p \sum_{j=1}^p |a_{ij}|^r)^{1/r}$, $0 < r < \infty$, and $\|\boldsymbol{A}\|_\infty = \max_{1\leq i,j\leq p} |a_{ij}|$. In particular, the $L_2$-norm, or the Frobenius norm can be expressed as $\|\boldsymbol{A}\|_2 = \{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})\}^{1/2}$. The $L_r$-operator norm of $\boldsymbol{A}$ is defined as $\|\boldsymbol{A}\|_{(r,r)} = \sup\{\|\boldsymbol{A}\boldsymbol{x}\|_r : \|\boldsymbol{x}\|_r = 1\}$. This gives the $L_1$-operator norm as $\|\boldsymbol{A}\|_{(1,1)} = \max_{1\leq j\leq p} \sum_{i=1}^p |a_{ij}|$, and the $L_2$-operator norm as $\|\boldsymbol{A}\|_{(2,2)} = [\max_{1\leq i\leq p}\{\mathrm{eig}_i(\boldsymbol{A}^T\boldsymbol{A})\}]^{1/2}$, so that, for symmetric matrices, $\|\boldsymbol{A}\|_{(2,2)} = \max_{1\leq i\leq p} |\mathrm{eig}_i(\boldsymbol{A})|$. For a symmetric $p$-dimensional matrix $\boldsymbol{A}$, we have, $\|\boldsymbol{A}\|_\infty \leq \|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_2 \leq p\|\boldsymbol{A}\|_\infty$, and $\|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_{(1,1)}$. For a positive definite matrix $\boldsymbol{A}$, $\boldsymbol{A}^{1/2}$ denotes its unique positive square root. The diagonal matrix with the same diagonal as a matrix $\boldsymbol{A}$ is denoted by $\boldsymbol{A}^+$, and $\boldsymbol{A}^-$ denotes the matrix $\boldsymbol{A} - \boldsymbol{A}^+$. The linear space of $p \times p$ symmetric matrices

is denoted by $\mathcal{M}_p$, and $\mathcal{M}_p^+ \subset \mathcal{M}_p$ is the cone of symmetric positive definite matrices of dimension $p \times p$.

The indicator function is denoted by $\mathbb{1}$. We denote the cardinality of a finite set $\mathcal{S}$ by $\#\mathcal{S}$. The Hellinger distance between two probability densities $f$ and $g$ is defined as $h(f,g) = \|f^{1/2} - g^{1/2}\|_2$.

## 2. Formulation of the prior–penalty dual

Let $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ be a random sample from a $p$-dimensional normal distribution with mean $\mathbf{0}$ and a positive definite covariance matrix $\boldsymbol{\Sigma}$. The corresponding precision matrix, or the inverse covariance matrix $\boldsymbol{\Omega} = (\!(\omega_{ij})\!)$ is defined as $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$. The natural estimator of $\boldsymbol{\Sigma}$ is $\boldsymbol{S} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$. We assume that $\boldsymbol{\Omega}$ is sparse, in the sense that the number of non-zero off-diagonal elements is small. We utilize the duality between a Bayesian prior and penalty, where the penalized likelihood estimate is understood to correspond to the maximum a posteriori (MAP) estimate under a given prior. Hence, for fully Bayesian inference on $\boldsymbol{\Omega}$, we need a suitable prior that also results in a penalty function with good frequentist properties; a non-trivial problem even in linear models [15]. We put independent horseshoe-like priors [6] on the off-diagonal and non-informative priors on the diagonal elements of $\boldsymbol{\Omega}$, while restricting the prior mass to positive definite matrices. A key benefit of the horseshoe-like prior, which closely mimics the sparsity-inducing global-local horseshoe prior [14] is that the prior density, and hence the penalty, is available in closed form under the former, unlike under the latter. This allows one to study the penalty (equivalently, the negative logarithm of the prior density) directly, and to establish important properties concerning convexity (see, e.g., Lemma 4.8), which remain much more difficult under the horseshoe prior. For the fully Bayesian model, the element-wise prior specification induced by the horseshoe-like prior is,

$$\begin{aligned} \pi(\omega_{ij} \mid a) &= \log\bigl(1 + a/\omega_{ij}^2\bigr)/\bigl(2\pi a^{1/2}\bigr),\ 1 \le i < j \le p,\ a > 0, \\ \omega_{ii} &\propto 1,\ 1 \le i \le p, \end{aligned} \tag{2.1}$$

where $\pi(\omega_{ij} \mid a)$ gives the horseshoe-like density function for $\omega_{ij}$. The motivation for using this density is two-fold: it has a sharp spike near zero, encoding the Bayesian prior belief that most signals are ignorable; and it also possesses very heavy, polynomially decaying tails, allowing for identification of signals. These two properties closely mimic the popular horseshoe prior for sparse signals [14], and, in fact, one achieves the same origin and tail rates for the density function in terms of $\omega_{ij}$ as in the original horseshoe. The crucial advantage with the horseshoe-like, then, is that there is a closed form to the density function, unlike the horseshoe prior. Nevertheless, similar to the original horseshoe prior, the horseshoe-like prior also admits a convenient latent variable representation as a Gaussian scale-mixture [6]. To be precise, one can write,

$$\omega_{ij} \mid \nu_{ij}, a \sim \mathcal{N}\left(0, \frac{a}{2\nu_{ij}}\right),\ \pi(\nu_{ij}) = \frac{1 - \exp(-\nu_{ij})}{2\pi^{1/2}\nu_{ij}^{3/2}}, \tag{2.2}$$

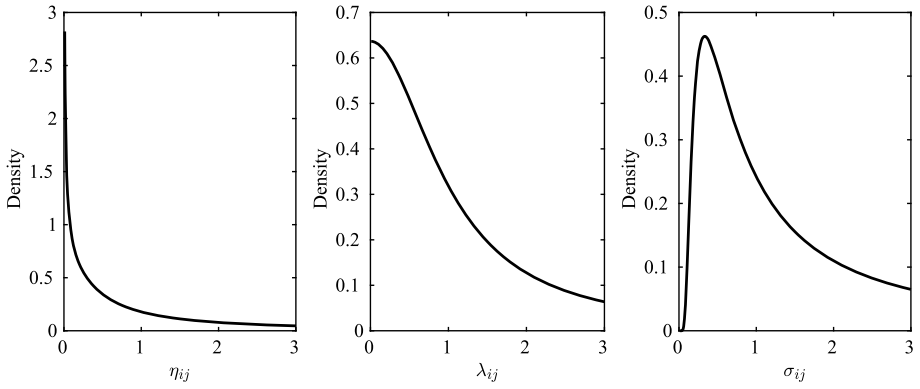


FIG 1. *Densities of local scale parameters. Left panel:* $\pi(\eta_{ij}) = 1/\sqrt{\pi\eta_{ij}}(1 - \exp(-1/\eta_{ij}))$, *where* $\eta_{ij} = 1/\nu_{ij}$ *and* $\nu_{ij}$ *is as defined in Equation* (2.2). *Center panel: standard half-Cauchy density of* $\lambda_{ij}$, *see horseshoe prior hierarchy in Equation* (2.5). *Right panel: If the local scale parameter* ($\sigma_{ij}$) *in a normal scale mixture had an Inverse Gamma prior,* $\sigma_{ij} \sim$ InvGamma$(1/2, 1/2)$.

where marginalizing over the latent $\nu_{ij}$ leads to the desired $\pi(\omega_{ij} \mid a)$ identified above. The density of $\eta_{ij}(= 1/\nu_{ij})$, in comparison with commonly used densities on the latent parameters, in normal scale mixtures, is presented in Figure 1. We can observe the aggressive shrinkage towards zero under the horseshoe family. For modeling valid precision matrices, we must restrict the prior mass on the space of symmetric positive definite matrices $\mathcal{M}_p^+$. Combining the unrestricted prior as in (2.1) and (2.2), along with the above restriction, leads to the joint prior specification on $\boldsymbol{\Omega}$ as,

$$\pi(\boldsymbol{\Omega} \mid \nu; a)\pi(\nu) \propto \prod_{i,j:i<j} \big\{1 - \exp(-\nu_{ij})\big\}\nu_{ij}^{-1} \exp\left(\frac{-\nu_{ij}\omega_{ij}^2}{a}\right)\mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega}), \qquad (2.3)$$

where $\nu = \{\nu_{ij}\}_{i<j}$. In this formulation, the latent parameters $\nu_{ij}$ are component-specific, or local, and the shared parameter $a$ is global, situating the horseshoe-like in the broader category of global-local priors [7]. Further details on the induced marginal prior on $\boldsymbol{\Omega}$ are presented in Appendix A. Although it is possible to put a further hyperprior on $a$, it is considered fixed for point estimation approaches, and is estimated by the effective model size approach of [43] to avoid a collapse to zero. We defer the details to Section 3.3. With the prior specification as in (2.3), the log-posterior $\mathcal{L}$ thus becomes,

$$\mathcal{L} \propto \frac{n}{2}\log\det\boldsymbol{\Omega} - \frac{n}{2}\mathrm{tr}(\boldsymbol{S\Omega}) + \sum_{i,j:i<j}\left\{\log\big(1-\exp(-\nu_{ij})\big) - \log\nu_{ij} - \frac{\nu_{ij}\omega_{ij}^2}{a}\right\}. \quad (2.4)$$

At this point, the corresponding hierarchy of the horseshoe prior, which the horseshoe-like closely approximates, is well worth mentioning. The horseshoe prior [14], recognized as a state-of-the-art for sparse signal recovery [6], was

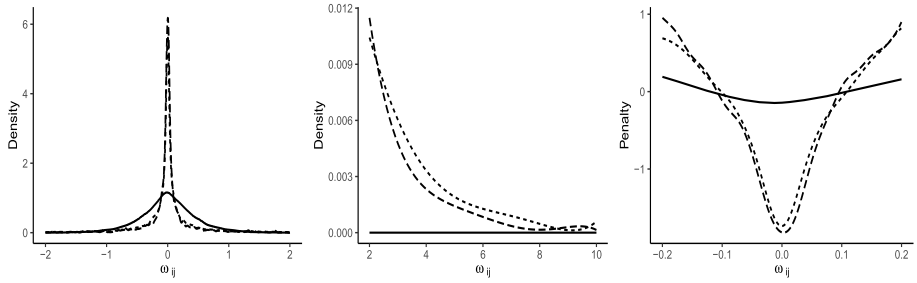


FIG 2. *Smoothed density estimates of a randomly chosen off-diagonal element based on* $10^4$ *Markov chain Monte Carlo samples for the graphical horseshoe-like (dashes), the graphical horseshoe (small dashes) and the Bayesian graphical lasso (solid) priors; providing a visual comparison of (left) spikes near the origin, (middle) heaviness of the tails, and (right) the induced penalty functions for* $p = 10$.

deployed in estimating GGMs by Li et al. [36] with the following hierarchy:

$$\omega_{ij} \mid \lambda_{ij}, \tau \sim \mathcal{N}\big(0, \lambda_{ij}^2 \tau^2\big),\ \pi\big(\lambda_{ij}^2\big) \sim \mathcal{C}^+(0,1),\ \pi\big(\tau^2\big) \sim \mathcal{C}^+(0,1), \tag{2.5}$$

where $\mathcal{C}^+(0,1)$ denotes the standard half-Cauchy distribution. Figure 2 plots the smoothed histogram of prior densities of a randomly chosen off-diagonal element near the origin and at the tails for the graphical horseshoe-like along with two of its relatives: the graphical horseshoe [36] and the Bayesian graphical lasso [57]. The corresponding penalties, given by the negative of the logarithm of the densities, are also shown. The key observations are: (a) the graphical horseshoe and graphical horseshoe-like densities are very similar and (b) both have far sharper spikes near the origin and heavier tails compared to the Laplace priors used in the Bayesian graphical lasso, providing an intuitive basis for the superiority of the horseshoe-family in sparse signal recovery. Extensive formal support for these observations are available in linear models [7], but barring some empirical evidence, the corresponding theoretical support is lacking in graphical models.

Some comments on the desirability of a Gaussian scale mixture representation are also in order. First, the latent mixing variables make it easier to derive fully Bayesian computational strategies via data augmentation. A similar observation is true for point estimates via the expectation–maximization algorithm. Second, using a result of Barndorff-Nielsen et al. [4], it is possible to derive a precise connection between the densities of the mixing variables and that of the resultant mixture. In particular, if the mixing densities are regularly varying in the tails, then so is the resultant Gaussian mixture. Since regular variation is closed under many nonlinear transformations, the heavier tails of global-local priors impart crucial robustness properties for estimating nonlinear, many-to-one functions of the parameters of interest in multi-parameter problems [5], and help avoid marginalization paradoxes of Dawid et al. [16].

## 3. Estimation procedure

### *3.1. ECM algorithm for MAP estimation*

We utilize the Gaussian mixture representation of the horseshoe-like prior with latent scale parameters to devise an Expectation Conditional Maximization (ECM) [39] approach to MAP estimation, building on the calculations for linear models by Bhadra et al. [6]. For updating the elements of the precision matrix, we use the coordinate descent technique proposed by Wang [58], which guarantees the positive definiteness of the precision matrix at each update.

**E Step:** Following Bhadra et al. [6], we calculate the conditional expectation of the latent variable $\nu_{ij}, 1 \leq i < j \leq p$, at current iteration $(t)$ as follows:

$$\nu_{ij}^{(t)} = \mathbb{E}\big(\nu_{ij} \mid \omega_{ij}^{(t)}, a\big) = \left(\log\left(1 + \frac{a}{(\omega_{ij}^{(t)})^2}\right)\right)^{-1} \frac{a^2}{((\omega_{ij}^{(t)})^2 + a)((\omega_{ij}^{(t)})^2)}. \quad (3.1)$$

**CM Steps:** Having updated the latent parameters in the E-Step, the coordinate descent approach of [58] is used to update one column of the precision matrix at a time. Without loss of generality, we present the steps for updating the $p$th column. First we divide the precision matrix $\boldsymbol{\Omega}$ and the sample covariance matrix $\boldsymbol{S}$ into blocks as follows:

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{12}^T & \Omega_{22} \end{bmatrix}, \quad n\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}^T & S_{22} \end{bmatrix},$$

where, $\boldsymbol{\Omega}_{11}$ is a matrix of dimension $(p-1) \times (p-1)$ of the top left block of $\boldsymbol{\Omega}$; $\Omega_{22}$ is the $p$th diagonal element and $\boldsymbol{\Omega}_{12}$ is a $(p-1) \times 1$ dimensional vector of the remaining elements in the $p$th column. The decomposition of $n\boldsymbol{S}$ is analogous. We define $\gamma = \Omega_{22} - \boldsymbol{\Omega}_{12}^T \boldsymbol{\Omega}_{11}^{-1} \boldsymbol{\Omega}_{12}$ and $\boldsymbol{\beta} = \boldsymbol{\Omega}_{12}$. With these transformations, we simplify (2.4) to update the $p$th column. We have,

$$\begin{aligned} \log\det \boldsymbol{\Omega} &= \log(\gamma) + c_1, \\ \operatorname{tr}(n\boldsymbol{S}\boldsymbol{\Omega}) &= 2\boldsymbol{S}_{12}^T\boldsymbol{\beta} + S_{22}\gamma + S_{22}\boldsymbol{\beta}^T\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta} + c_2, \\ \sum_{i,j:i<j} \frac{\nu_{ij}^{(t)}}{a} \cdot \omega_{ij}^2 &= \boldsymbol{\beta}^T \boldsymbol{\Lambda}^{(t)} \boldsymbol{\beta} + c_3, \\ \boldsymbol{\Lambda}^{(t)} &= \frac{1}{a} \operatorname{diag}\big(\nu_{1p}^{(t)}, \ldots, \nu_{p-1,p}^{(t)}\big), \end{aligned} \quad (3.2)$$

where $c_1,\ c_2,\ c_3$ are constants independent of $\boldsymbol{\beta},\ \gamma$. Now the log-posterior with the transformed variables is given by,

$$\mathcal{L} \propto \frac{n}{2}\log(\gamma) - \frac{1}{2}\big(2\boldsymbol{S}_{12}^T\boldsymbol{\beta} + S_{22}\gamma + S_{22}\boldsymbol{\beta}^T\boldsymbol{\Omega}_{11}^{-1}\boldsymbol{\beta}\big) - \boldsymbol{\beta}^T\boldsymbol{\Lambda}^{(t)}\boldsymbol{\beta}.$$

Maximizing the above over $\boldsymbol{\beta}, \gamma$ gives the required update as:

**Algorithm 1** ECM algorithm for MAP estimation (GHS-LIKE-ECM)

**function** ECM FOR GRAPHICAL HORSESHOE-LIKE PENALTY($\boldsymbol{\Omega}_s, \boldsymbol{S}, n, p$)
  $\boldsymbol{\Omega}_s = ((\omega_{s,ij}))$: starting point; $\boldsymbol{\Omega}_u = ((\omega_{u,ij}))$: updated precision matrix; initially set to $\boldsymbol{I}_p$.
  $n\boldsymbol{S} = \boldsymbol{X}^{(n)T}\boldsymbol{X}^{(n)}$; $\boldsymbol{N} = ((N_{ij}))$: A matrix of dimension $p \times p$ which stores $\mathbb{E}(\nu_{ij} \mid \omega_{s,ij}, a)$; $(n, p)$: Sample size and number of variables respectively
  **while** $\boldsymbol{\Delta} = \|\boldsymbol{\Omega}_u - \boldsymbol{\Omega}_s\|_2 <$ tolerance $(= 10^{-3})$ **do**
    **for** $j = 2 \to p$ **do**
      **for** $i = 1 \to (j-1)$ **do**
        $N_{ij} := \mathbb{E}(\nu_{ij} \mid \omega_{s,ij}, a) = (\log(1 + \frac{a}{(\omega_{s,ij})^2}))^{-1} \frac{(a)^2}{((\omega_{s,ij})^2+a)((\omega_{s,ij})^2)}$.
      **end for**
    **end for**
    $\boldsymbol{N} \leftarrow \boldsymbol{N} + \boldsymbol{N}^T$. This is required to compute $\boldsymbol{\Lambda}^{(t)}$ in display (3.2); Set $\boldsymbol{\Omega}_u = \boldsymbol{\Omega}_s$
    **for** $i = 1 \leftarrow p$ **do**
      Update $i^{th}$ column of $\boldsymbol{\Omega}_u$ using coordinate descent algorithm of Wang [58] described above.
    **end for**
  **end while**
  **return** $\hat{\boldsymbol{\Omega}}^{\text{MAP}} = \boldsymbol{\Omega}_u$
**end function**

$$\hat{\gamma} = \frac{n}{S_{22}}, \quad \hat{\boldsymbol{\beta}} = -\big(S_{22}\boldsymbol{\Omega}_{11}^{-1} + 2 \cdot \boldsymbol{\Lambda}^{(t)}\big)^{-1}\boldsymbol{S}_{12}^T. \tag{3.3}$$

Having updated $\boldsymbol{\beta}$, $\gamma$ from (3.3), the $p$th column update of the precision matrix for the next iteration $(t+1)$ becomes

$$\hat{\boldsymbol{\Omega}}_{12}^{(t+1)} = \hat{\boldsymbol{\beta}}, \quad \hat{\boldsymbol{\Omega}}_{12}^{T(t+1)} = \hat{\boldsymbol{\beta}}^T, \quad \hat{\Omega}_{22}^{(t+1)} = \hat{\gamma} + \hat{\boldsymbol{\beta}}^T\boldsymbol{\Omega}_{11}^{-1}\hat{\boldsymbol{\beta}}. \tag{3.4}$$

We repeat the above steps for the remaining $(p-1)$ columns to complete the CM Step updates for $\boldsymbol{\Omega}$, until convergence to the MAP estimator $\hat{\boldsymbol{\Omega}}^{\text{MAP}}$. The procedure is summarized in Algorithm 1. The most computationally expensive step is the required inverse of a $(p-1) \times (p-1)$ matrix to compute $\hat{\boldsymbol{\beta}}$ in (3.3), which needs to be repeated for each of the $p$ columns, giving a per iteration complexity of $O(p^4)$ for the algorithm.

### *3.2. Posterior sampling for the fully Bayesian estimate*

For fully Bayesian estimation, we also outline the MCMC sampling procedure. With substitutions $2\nu_{ij} \mapsto t_{ij}^2$ and $a \mapsto \tau^2$, the prior in (2.2) can be written with a different hierarchy as follows:

$$\omega_{ij} \mid \nu_{ij}, \tau \sim \mathcal{N}\big(0, \tau^2/t_{ij}^2\big),\ \pi(t_{ij}) = \frac{1 - \exp(-t_{ij}^2/2)}{(2\pi)^{1/2}t_{ij}^2},\ t_{ij} \in \mathbb{R},\ \tau^2 > 0,$$

where $\pi(t_{ij})$ above is known as the the slash normal density, expressed as $(\phi(0) - \phi(t_{ij}))/t_{ij}^2$, where $\phi(\cdot)$ is the standard normal density [6]. Introducing a further local latent variable $r_{ij}$, the density for $t_{ij}$ can also be written as a normal scale mixture, where the scale follows a Pareto distribution, that is,

$$t_{ij} \mid r_{ij} \sim \mathcal{N}(0, r_{ij}),\ r_{ij} \sim \text{Pareto}(1/2).$$

For efficient sampling, the above Pareto scale mixture can be represented as a product of an exponential density and an indicator function as follows:

$$\begin{aligned}\pi(t_{ij}) &= \frac{1}{2}\int_1^\infty \frac{1}{(2\pi r_{ij})^{1/2}}\exp\left(-\frac{t_{ij}^2}{2r_{ij}}\right) r_{ij}^{-3/2} dr_{ij} \\ &= \frac{1}{2(2\pi)^{1/2}}\int_0^1 \exp\left(-\frac{t_{ij}^2 m_{ij}}{2}\right) dm_{ij},\end{aligned}$$

that is,

$$\pi(t_{ij}, m_{ij}) = \frac{1}{2(2\pi)^{1/2}}\exp\left(\frac{-t_{ij}^2 m_{ij}}{2}\right)\mathbb{1}(0 < m_{ij} < 1).$$

Different choices of prior for the global scale parameter are possible, but we consider $\tau \sim \mathcal{C}^+(0,1)$. Makalic and Schmidt [38] observed that: if $\tau^2 \mid \xi \sim \text{InvGamma}(1/2, 1/\xi)$ and $\xi \sim \text{InvGamma}(1/2,1)$ then marginally $\tau \sim \mathcal{C}^+(0,1)$. Using this, we can write the posterior updates of $\tau$, $\xi$ as follows:

$$\begin{aligned}\tau^2 \mid \xi, \boldsymbol{X}^{(n)}, \boldsymbol{\Omega}, \{t_{ij}\}_{i<j}, \{m_{ij}\}_{i<j} \sim \text{InvGamma}\Bigg(&\big(p(p-1)/2+1\big)/2, \\ &1/\xi + \sum_{i,j:i<j} t_{ij}^2\omega_{ij}^2/2\Bigg), \\ \xi \mid \tau \sim \text{InvGamma}&\big(1 + 1/\tau^2\big).\end{aligned} \tag{3.5}$$

Following the remaining updates from the graphical horseshoe sampler of Li et al. [36], the complete MCMC scheme for the graphical horseshoe-like is as outlined in Algorithm 2. The per iteration complexity of the algorithm is $O(p^4)$. Diagnostic plots for both the ECM and MCMC algorithms are given in Appendix C.

### *3.3. Estimating the global scale parameter*

We adapt the technique of Piironen and Vehtari [43] for choosing a suitable value of the global shrinkage parameter $a$ by estimating the *effective model size*. The method of Piironen and Vehtari [43] was developed in the context of linear regression, and proceeds using a prior guess on the number of non-zero regression coefficients. In our context, this suggests a prior guess on the number of non-zero edges in the upper (or lower) triangle of $\boldsymbol{\Omega}$. To adapt this method, we use the partial regressions induced by a multivariate Gaussian model, as described next.

Consider the partial regressions induced by the multivariate Gaussian model of Section 2, where each variable is regressed on the remaining ones. Proposition C.5 of Lauritzen [32] yields:

$$\boldsymbol{X}_i = \boldsymbol{X}_{-i}\boldsymbol{\theta}_{-i} + \boldsymbol{\varepsilon}_i;\ \boldsymbol{\varepsilon}_i \sim \mathcal{N}(0, \mathbf{I}_n/\omega_{ii}),$$

for $i = 1, \ldots, p$, where $\boldsymbol{X}_i \in \mathbb{R}^n$, the matrix $\boldsymbol{X}_{-i} \in \mathbb{R}^{n\times(p-1)}$ is formed using the matrix $\boldsymbol{X}^{(n)}$ with its $i$th column removed, and $\boldsymbol{\theta}_{-i} = \{-\omega_{ij}/\omega_{ii}\}_{j\neq i} \in \mathbb{R}^{p-1}$.

**Algorithm 2** The Graphical Horseshoe-Like MCMC Sampler (GHS-LIKE-MCMC)

**function** GHS-LIKE($\boldsymbol{S}, n, burnin, nmc$) ▷ where $n\boldsymbol{S} = \boldsymbol{X}^{(n)T}\boldsymbol{X}^{(n)}$, $n$ = sample size
Set $p$ to be number of rows (or columns) in $\boldsymbol{S}$
Set initial values $\boldsymbol{\Omega} = \boldsymbol{I}_p$, $\boldsymbol{\Sigma} = \boldsymbol{I}_p$, $\boldsymbol{T} = \mathbf{1}\mathbf{1}^T$, $\boldsymbol{M} = \mathbf{1}\mathbf{1}^T$, $\tau = 1$, where $\mathbf{1}$ is a $p$-dimensional vector with all elements equal to 1, $\boldsymbol{T} = (\!(t_{ij}^2)\!)$, $\boldsymbol{M} = (\!(m_{ij})\!)$. $s_{ii}$ is the $i^{\text{th}}$ diagonal element of $n\boldsymbol{S}$ and $\boldsymbol{S}_{(-i)i}$ is the $i^{\text{th}}$ column of $n\boldsymbol{S}$ excluding $s_{ii}$.
**for** $iter = 1$ to $(burnin + nmc)$ **do**
**for** $i = 1$ to $p$ **do**
$\gamma \sim \text{Gamma}(\text{shape} = n/2 + 1, \text{rate} = 2/s_{ii})$ ▷ sample $\gamma$
$\boldsymbol{\Omega}^{-1}_{(-i)(-i)} = \boldsymbol{\Sigma}_{(-i)(-i)} - \boldsymbol{\sigma}_{(-i)i}\boldsymbol{\sigma}'_{(-i)i}/\sigma_{ii}$
$\boldsymbol{C} = [s_{ii}\boldsymbol{\Omega}^{-1}_{(-i)(-i)} + (\text{diag}(\tau^2/t_{(-i)i}))^{-1}]^{-1}$
$\boldsymbol{\beta} \sim \text{Normal}(-\boldsymbol{C}\boldsymbol{S}_{(-i)i}, \boldsymbol{C})$ ▷ sample $\boldsymbol{\beta}$
$\boldsymbol{\omega}_{(-i)i} = \boldsymbol{\beta}$, $\omega_{ii} = \gamma + \boldsymbol{\beta}^T\boldsymbol{\Omega}^{-1}_{(-i)(-i)}\boldsymbol{\beta}$ ▷ variable transformation
$\boldsymbol{t}_{(-i)i} \sim \text{Gamma}(\text{shape} = 3/2, \text{rate} = m_{(-i)i}/2 + \boldsymbol{\omega}^2_{(-i)i}/2\tau^2)$ ▷ sample $t$, where $\boldsymbol{t}_{(-i)i}$ is a vector of length $(p-1)$ with entries $t^2_{ji}, j \neq i$. Entries of $\boldsymbol{t}_{(-i)i}$ greater than $10^{15}$ are not updated (for numerical stability while sampling $\boldsymbol{\beta}$).
$\boldsymbol{m}_{(-i)i} \sim \text{Exponential}(\text{rate} = \boldsymbol{t}_{(-i)i}/2)\ \mathbb{1}(0 < \boldsymbol{m}_{(-i)i} < 1)$ ▷ sample $m$
Save updated $\boldsymbol{\Omega}$
$\boldsymbol{\Sigma}_{(-i)(-i)} = \boldsymbol{\Omega}^{-1}_{(-i)(-i)} + (\boldsymbol{\Omega}^{-1}_{(-i)(-i)}\boldsymbol{\beta})(\boldsymbol{\Omega}^{-1}_{(-i)(-i)}\boldsymbol{\beta})'/\gamma$, $\boldsymbol{\sigma}_{(-i)i} = -(\boldsymbol{\Omega}^{-1}_{(-i)(-i)}\boldsymbol{\beta})/\gamma$, $\sigma_{ii} = 1/\gamma$
Save updated $\boldsymbol{\Sigma}, \boldsymbol{T}, \boldsymbol{M}$.
**end for**
Sample $\tau$, $\xi$ as in (3.5). ▷ Sample $\tau$, $\xi$
**end for**
Return MC samples $\boldsymbol{\Omega}$
**end function**

There are $p$ such partial regressions, each corresponding to a column of $\boldsymbol{X}^{(n)}$. In order to use the *effective model size* approach of Piironen and Vehtari [43], one needs to have a prior guess ($p_0$) for the number of non-zero coefficients in each regression. In our case, we consider $p_0 = 2$ for every partial regression. The rationale behind this consideration is a prior belief that there would be approximately $p$ number of non-zero terms in the off-diagonal of the true precision matrix $\boldsymbol{\Omega}_0$, which amounts to roughly 2 non-zero coefficients in each partial regression (because of the symmetry of $\boldsymbol{\Omega}$). For the sake of simplicity, we set $\omega_{ii} = 1/\sigma^2$ for $i = 1, \ldots, p$.

When the horseshoe-like prior (2.2) is imposed on the elements of $\boldsymbol{\theta}_{-i}$, the shrinkage estimates of the elements of $\boldsymbol{\theta}_{-i}$ can be written as, $\bar{\omega}_{ij} = (1 - \kappa_{ij})\hat{\omega}_{ij}$, where, $\kappa_{ij}$ is called shrinkage coefficient and $\hat{\omega}_{ij}$ is the ordinary least squares (OLS) estimate of $\omega_{ij}$. Further, under a simplifying assumption that the design matrices for the partial regressions are approximately orthogonal, one has: $\kappa_{ij} = (1 + n\sigma^{-2}a(2\nu_{ij})^{-1})^{-1}$, where $\sigma^2$ is the variance of the noise terms $\boldsymbol{\varepsilon}_i$. In the context of $i^{th}$ partial regression, the effective model size (as defined by Piironen and Vehtari [43]) can be written as, $m^{(i)}_{\text{eff}} = \sum_{j \neq i}(1 - \kappa_{ij})$. In order to compute the global scale parameter $a$ or to decide a prior for it, Piironen and Vehtari [43] set $\mathbb{E}(m^{(i)}_{\text{eff}}) = p_0$, which is the expected number of non-zero ele-

ments in $\boldsymbol{\theta}_{-i}$. Hence, we need to find an expression for $\mathbb{E}(\kappa_{ij})$ in order to solve for $a$ in $\mathbb{E}(m_{\text{eff}}^{(i)}) = p_0$. Using the standard Jacobian technique we get the density of $\kappa_{ij}$ as:

$$\pi(\kappa_{ij}) = \frac{1}{2\pi^{1/2}}\kappa_{ij}^{-3/2}(1-\kappa_{ij})^{-1/2}\left(\frac{na}{2\sigma^2}\right)^{-1/2}\left\{1-\exp\left(-\frac{\kappa_{ij}}{1-\kappa_{ij}}\frac{na}{2\sigma^2}\right)\right\}.$$

After some trivial variable transforms, $\mathbb{E}(\kappa_{ij} \mid a)$ can be written as:

$$\mathbb{E}(\kappa_{ij} \mid a) = \frac{2\sigma}{(2\pi na)^{1/2}}\int_0^{\pi/2}\left\{1-\exp\left(-\frac{na}{2\sigma^2}\tan^2\eta\right)\right\}d\eta. \tag{3.6}$$

Hence, the consideration $\mathbb{E}(m_{\text{eff}}^{(i)}) = p_0$ leads to the estimating equation:

$$\sum_{\substack{j=1\\ j\neq i}}^{p}\big(1-\mathbb{E}(\kappa_{ij} \mid a)\big) = p_0. \tag{3.7}$$

Note that the expression $\mathbb{E}(\kappa_{ij} \mid a)$ as in (3.6) is independent of the indices $i$, $j$ due to the symmetry of the prior. Further, $\mathbb{E}(\kappa_{ij} \mid a)$ depends only on the ratio $a/\sigma^2$ and hence, to uniquely solve for $a$, we set $\sigma^2 = 1$. As a closed form expression of the integral in (3.6) is not available, we approximate the exponential function in the integral with a high order polynomial and numerically solve for $a$ using (3.7) for given $n$ and $p$ and the prior guess that $p_0 = 2$.

## 4. Theoretical properties

### *4.1. Posterior concentration results*

In this section, we present our main result on the posterior contraction rate of the precision matrix $\boldsymbol{\Omega}$ around the true precision matrix $\boldsymbol{\Omega}_0$ with respect to the Frobenius norm under the graphical horseshoe-like prior. The technique of our proofs uses the general theory on posterior convergence rates as outlined in [24], which establishes the desired convergence with respect to the Hellinger distance. However, from the perspective of a precision matrix, the Frobenius norm is easier to interpret in comparison to the Hellinger distance. Under suitable assumptions on the eigenspace of the precision matrices, [3] showed that these two distances are equivalent, and hence the same posterior contraction rates hold with respect to the Frobenius norm as well. We assume that the true underlying graph is sparse, so that the corresponding true precision matrix $\boldsymbol{\Omega}_0$ has $s$ non-zero off-diagonal elements. The total number of non-zero elements in $\boldsymbol{\Omega}_0$ is $p+s$, which gives the effective dimension of the parameter $\boldsymbol{\Omega}_0$. To establish the desired posterior concentration results, we shall need to control both the actual dimension and the effective dimension of the true precision matrix. Overall, our theoretical analyses depend on certain assumptions on the true precision matrix, the dimension and sparsity, and the prior space. We present the details of these assumptions along with relevant discussions below.

**Assumption 4.1.** *The prior is restricted to a subspace of symmetric positive definite matrices, $\mathcal{M}_p^+(L)$, where*

$$\mathcal{M}_p^+(L) = \big\{\boldsymbol{\Omega} \in \mathcal{M}_p^+ : 0 < L^{-1} \leq \mathrm{eig}_1(\boldsymbol{\Omega}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{\Omega}) \leq L < \infty\big\}. \quad (4.1)$$

**Assumption 4.2.** *The actual dimension $p$ satisfies the condition $p = n^b$, $b \in (0,1)$, and the effective dimension $p+s$ satisfies $(p+s)\log p/n = o(1)$.*

**Assumption 4.3.** *The true precision matrix $\boldsymbol{\Omega}_0$ belongs to the parameter space given by*

$$\mathcal{U}(\varepsilon_0, s) = \Big\{\boldsymbol{\Omega} \in \mathcal{M}_p^+ : \sum_{1\leq i\neq j\leq p} \mathbb{1}(\omega_{ij} \neq 0) \leq s, \\ 0 < \varepsilon_0^{-1} \leq \mathrm{eig}_1(\boldsymbol{\Omega}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{\Omega}) \leq \varepsilon_0 < \infty\Big\}.$$

**Assumption 4.4.** *The bound $[L^{-1}, L]$ on the eigenvalues of $\boldsymbol{\Omega}$ as specified in (4.1) satisfies $L > \varepsilon_0$, or, in other words, $\varepsilon_0 = cL$, for some $c \in (0,1)$.*

**Assumption 4.5.** *The global shrinkage parameter $a$ satisfies the condition, $a^{1/2} < n^{-1/2}p^{-b_1}(s\log p)^{1/2}$, for some sufficiently large constant $b_1 > 0$.*

The condition on the eigenvalues of $\boldsymbol{\Omega}$ as specified in Assumption 4.1 is necessary for arriving at the theoretical results involving the posterior convergence rate of $\boldsymbol{\Omega}$. In this paper, we assume that $L$ is a fixed constant, which can be arbitrarily large, so that for practical implementation, we work with $\boldsymbol{\Omega} \in \mathcal{M}_p^+$ as specified in (2.3). However, this condition does not affect the practical implementation of our proposed method, and is used purely as a technical requirement, so that we only can work with $\boldsymbol{\Omega}$ and $\boldsymbol{\Sigma}$ that are away from singular matrices. Beyond this, no structural assumptions such as decomposability are placed on either $\boldsymbol{\Omega}$ or $\boldsymbol{\Sigma}$. Similar assumptions have been made in related works; see [37] and [34]. Assumption 4.2 implies that the dimension grows to infinity as the sample size $n \to \infty$, but at a slower rate than $n$. Additionally, the condition on the effective dimension ensures that the posterior convergence rate goes to zero as $n \to \infty$. Similar conditions are necessary in proving the contraction results in other related works, for example, see [3, 37, 34]. Assumption 4.3 implies that the true precision matrix $\boldsymbol{\Omega}_0$ is sparse, and has eigenvalues that are bounded away from zero or infinity. Similar conditions are common in the literature on large precision matrix estimation problems; see, for example, [2, 3, 37, 34]; among others. Assumption 4.4 is crucial in learning the precision matrix in a high-dimensional framework. This condition ensures that $\boldsymbol{\Omega}_0 \in \mathcal{M}_p^+(L)$, that is, the prior space contains the true precision matrix, which is necessary in efficient learning of the same. Assumption 4.5 ensures that the prior puts sufficient mass around the true zero elements in the precision matrix. The condition on the global scale parameter $a$ is a sufficient one, and is required to obtain the desired posterior convergence rate. We present the main theoretical result for posterior convergence now. A proof can be found in Section 7.

**Theorem 4.6.** *Let* $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$ *be a random sample from a* $p$*-dimensional normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\boldsymbol{\Sigma}_0 = \boldsymbol{\Omega}_0^{-1}$*, where* $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$*. Consider the prior specification as given by (2.3) with an additional restriction on the prior space as outlined in Assumption 4.1. Under the remainder of the assumptions as given in Assumptions 4.2–4.5, the posterior distribution of* $\boldsymbol{\Omega}$ *satisfies*

$$\mathbb{E}_0\big[\Pr\big\{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 > M\epsilon_n \mid \boldsymbol{X}^{(n)}\big\}\big] \to 0,$$

*for* $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ *and a sufficiently large constant* $M > 0$*.*

**Corollary 4.7.** *Under similar conditions as in Theorem 4.6 above, the posterior distribution of* $\boldsymbol{\Omega}$ *has the posterior convergence rate* $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ *around* $\boldsymbol{\Omega}_0$ *with respect to the Frobenius norm under the graphical horseshoe prior as specified in (2.5) with an additional restriction on the prior space as outlined in Assumption 4.1.*

A proof of Corollary 4.7 is in Section 7 and settles the question of posterior concentration for the graphical horseshoe which Li et al. [36] did not address. The posterior convergence rate above directly compares with the rate of convergence of the frequentist graphical lasso estimator [46], and is identical to the posterior convergence rates obtained by Banerjee and Ghosal [3] and Liu and Martin [37]. However, our work is the first to address *unstructured precision matrices*, apart of a mild assumption of sparsity, using computationally efficient *continuous shrinkage priors*. This is at a contrast with previous theoretical analyses that imposed restrictive assumptions such as decomposability.

### *4.2. Properties of the MAP estimator*

The MAP estimator of $\boldsymbol{\Omega}$ can be found by maximizing the following objective function:

$$\begin{aligned} Q(\boldsymbol{\Omega}) = \log \pi\big(\boldsymbol{\Omega} \mid \boldsymbol{X}^{(n)}\big) &= \ell(\boldsymbol{\Omega}) + \sum_{i,j:i<j} \log \pi(\omega_{ij} \mid a) + C \\ &= \frac{n}{2}\big(\log\det \boldsymbol{\Omega} - \operatorname{tr}(\boldsymbol{S}\boldsymbol{\Omega})\big) - \sum_{i,j:i<j} pen_a(\omega_{ij}) + C, \end{aligned} \tag{4.2}$$

where $pen_a(\omega) = -\log\log(1 + a/\omega^2)$, $a > 0$, is the horseshoe-like penalty. We start by proving $pen_a(\omega)$ is strictly concave in the following lemma, with a proof in Section 7.

**Lemma 4.8.** *The extended real-valued penalty function* $pen_a(x) = -\log\log(1 + a/x^2)$, $a > 0$*, is strictly concave for all* $x \in dom(pen_a)$*, separately for* $x > 0$ *and* $x < 0$*.*

A direct consequence of Lemma 4.8 is as follows. Let $\boldsymbol{\Omega}^{(t)}$ be the $t$th iterate of a local linear approximation (LLA) algorithm [66], that is,

$$\boldsymbol{\Omega}^{(t+1)} = \operatorname{argmax}\left\{ \ell(\boldsymbol{\Omega}) - \sum_{i,j:i<j} pen'_a\big(|\omega_{ij}^{(t)}|\big)\ |\omega_{ij}| \right\}, \quad t = 1, 2, \ldots.$$

Then Theorem 1 of Zou and Li [66], together with the strict concavity of horsehoe-like penalty function from Lemma 4.8, guarantees that the LLA algorithm will satisfy an ascent property, that is, $Q(\boldsymbol{\Omega}^{(t+1)}) > Q(\boldsymbol{\Omega}^{(t)})$, and hence the LLA algorithm will be a special case of minorize–maximize algorithms. Further, the MAP estimate gives exactly zero off-diagonal values. The reason for this is that the horseshoe-like penalty is non-convex (Lemma 4.8), and penalized likelihood point estimates under non-convex penalties enjoy the property of sparsity, as established by Fan and Li [19].

We now present the result on consistency of the MAP estimate using the graphical horseshoe-like prior via an ECM algorithm, with a proof in Section 7.

**Theorem 4.9.** *Under the conditions of Theorem 4.6, the MAP estimator of* $\boldsymbol{\Omega}$*, given by* $\hat{\boldsymbol{\Omega}}^{\mathrm{MAP}}$ *is consistent, in the sense that*

$$\|\hat{\boldsymbol{\Omega}}^{\mathrm{MAP}} - \boldsymbol{\Omega}_0\|_2 = O_P(\epsilon_n),$$

*where* $\epsilon_n$ *is the posterior convergence rate as defined in Theorem 4.6.*

The above result guarantees that the MAP estimator also converges to the true precision matrix $\boldsymbol{\Omega}_0$ at the same rate as the posterior convergence rate in the Frobenius norm. By triangle inequality, Theorem 4.6 and Theorem 4.9 together imply that $\|\boldsymbol{\Omega} - \hat{\boldsymbol{\Omega}}^{\mathrm{MAP}}\|_2 = O_P(\epsilon_n)$, so that the posterior probability of an $\epsilon_n$-neighborhood around the MAP estimator with respect to the Frobenius norm converges to one. This pleasing correspondence between the fully Bayesian and MAP estimates under the same prior–penalty dual is far from guaranteed, in the face of possible contradictions pointed out by Castillo et al. [15] for the lasso in linear models.

## 5. Numerical experiments

We compare the MAP and MCMC estimates under the horseshoe-based methods (GHS, GHS-LIKE-MCMC and GHS-LIKE-ECM) with two frequentist approaches: GLASSO, GSCAD and one Bayesian approach: the Bayesian GLASSO (BGL). We consider two problem dimensions: $(n, p) = (120, 100)$ and $(120, 200)$. For each dimension, we perform simulations under four different structures of the true precision matrix $\boldsymbol{\Omega}_0$ as in Li et al. [36] and Friedman et al. [21]. These are: Random, Hubs, Cliques positive and Cliques negative, as detailed below.

1. *Random.* The off-diagonal entries of $\boldsymbol{\Omega}_0$ are non-zero with probability 0.01 when $p = 100$ and 0.002 when $p = 200$. The non-zero entries are then sampled uniformly from $(-1, -0.2)$. A simple Erdős-Rényi model is an example of generating a Random graph. This structure serves as a useful test case.

2. *Hubs.* The rows/columns are partitioned into $K$ disjoint groups $G_1, \ldots, G_K$. The off-diagonal entries $\omega_{ij}^0$ are set to 0.25 if $i \neq j$ and $i, j \in G_k$ for $k = 1, \ldots, K$. In our simulations we consider $p/10$ groups with equal number of elements in each group. From a practical viewpoint, Hubs are essential in protein–protein interaction networks [27, 29].
3. *Cliques positive and Cliques negative.* Same as Hubs, except for setting all $\omega_{ij}^0$, $i \neq j$ and $i, j \in G_k$, we select 3 members within each group, $g_k \subset G_k$, and set $\omega_{ij}^0 = 0.75$, $i \neq j$ and $i, j \in g_k$ for 'Cliques positive' and set $\omega_{ij}^0 = -0.45$, $i \neq j$ and $i, j \in g_k$ for 'Cliques negative'. In terms of application, groups of functionally associated proteins called modules (cliques), are responsible for cellular functions [26, 45].

For each setting of $(n, p)$ and $\boldsymbol{\Omega}_0$, we generate 50 data sets (repetitions) and estimate the precision matrices by the methods stated above. We generate 6000 MCMC samples for all the fully Bayesian methods, with initial 1000 samples as burn-in. All three horseshoe based methods are implemented in MATLAB, GSCAD is as implemented by Wang [57] and GLASSO is implemented in R package 'glasso' [22]. Starting points for GHS-LIKE-ECM are randomly chosen in order to avoid getting stuck in a local minimum (see details in Appendix C) and its global shrinkage parameter is chosen as in Section 3.3. The values of global scale parameter hence obtained for simulations are as follows:

(i) **Tables 1, 2**: $(n, p) = (120, 100)$, estimate of the global scale parameter $a = 0.0143$.
(ii) **Tables 3, 4**: $(n, p) = (120, 200)$, estimate of the global scale parameter $a = 0.0169$.

Tuning parameters for GLASSO and GSCAD are chosen by 5-fold cross validation. Owing to the signal-adaptive credible interval length and the conservative nature of variable selection, observed in the horseshoe prior [54], the middle 50% posterior credible intervals are used for variable selection for the Bayesian approaches. This choice helps to reduce the number of false negatives due to wider credible intervals [36], which is corroborated by the higher MCC values, observed for the horseshoe-based methods in simulations (Tables 1–4). We provide results on: Stein's loss $(= \operatorname{tr}(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}_0) - \log\det(\hat{\boldsymbol{\Omega}}\boldsymbol{\Sigma}_0) - p)$, Frobenius norm (F norm $= \|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_2$), true and false positive rates for detecting non-zero off-diagonal entries (resp., TPR and FPR), the Matthews Correlation Coefficient (MCC), and average CPU time. Note that for the fully Bayesian estimate, our theory concerns posterior concentration properties, and connections with convergence in Frobenius norm is established in Banerjee and Ghosal [3]. However, for the sake of completeness and comparisons with point estimation approaches, we provide variable selection results for all approaches as well, in addition to Stein's loss (an empirical measure of Kullback–Leibler divergence) and F norm that focus more directly on the entire distribution. Comparison of the estimates by graphical horseshoe-like MCMC, under 50 vs. 100 repetitions, with $\boldsymbol{\Omega}_0$ having hub structure and $(n, p) = (120, 100)$, is presented in Appendix D, Table 9. As similar results were observed in other representative settings, we present our

TABLE 1

*Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates, Matthews Correlation Coefficient of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix $\mathbf{\Omega}_0$ (Random and Cliques negative structures), where $n = 120$ and $p = 100$. The precision matrix is estimated by frequentist graphical lasso with penalized diagonal elements (GL1) and with unpenalized diagonal elements (GL2), graphical SCAD (GSCAD), Bayesian graphical lasso (BGL), the graphical horseshoe (GHS), graphical horseshoe-like ECM (ECM) and graphical horseshoe-like MCMC (MCMC). The best performer in each row is shown in bold. Average CPU time is in seconds.*

| | | | Random<br>35 nonzero pairs out of 4950<br>nonzero elements $\sim -\text{Unif}(0.2, 1)$ | | | | |
|---|---|---|---|---|---|---|---|
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 5.245 | 6.785 | 5.21 | 42.997 | **2.176** | 3.758 | 2.63 |
| | (0.254) | (0.464) | (0.242) | (0.898) | (0.278) | (0.282) | (0.306) |
| F norm | 3.348 | 4.084 | 3.333 | 3.952 | **1.194** | 2.224 | 2.033 |
| | (0.115) | (0.143) | (0.117) | (0.139) | (0.144) | (0.108) | (0.138) |
| TPR | 0.951 | 0.882 | **0.998** | 0.979 | 0.819 | 0.948 | 0.827 |
| | (0.03) | (0.038) | (0.009) | (0.023) | (0.041) | (0.032) | (0.037) |
| FPR | 0.101 | 0.045 | 0.994 | 0.166 | **0.0005** | 0.071 | 0.001 |
| | (0.013) | (0.007) | (0.005) | (0.0007) | (0.0003) | (0.005) | (0.001) |
| MCC | 0.232 | 0.321 | 0.005 | 0.181 | **0.869** | 0.275 | 0.832 |
| | (0.018) | (0.024) | (0.001) | (0.007) | (0.031) | (0.016) | (0.037) |
| Avg CPU time | 4.988 | 4.719 | 53.977 | 550.422 | 252.84 | 5.94 | 326.33 |
| | | | Cliques negative<br>30 nonzero pairs out of 4950<br>nonzero elements $= -0.45$ | | | | |
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 4.607 | 7.134 | 4.567 | 42.618 | **1.862** | 3.417 | 2.284 |
| | (0.223) | (0.529) | (0.231) | (0.896) | (0.263) | (0.251) | (0.278) |
| F norm | 2.823 | 3.851 | 2.813 | 3.814 | **1.969** | 2.107 | 2.003 |
| | (0.117) | (0.138) | (0.112) | (0.165) | (0.212) | (0.124) | (0.181) |
| TPR | **1** | **1** | **1** | **1** | 0.983 | **1** | 0.992 |
| | (0) | (0) | (0) | (0) | (.024) | (0) | (0.019) |
| FPR | 0.1 | 0.028 | 0.983 | 0.158 | **0.0004** | 0.073 | 0.001 |
| | (0.01) | (0.006) | (0.013) | (0.007) | (0.0003) | (0.005) | (0) |
| MCC | 0.232 | 0.42 | 0.01 | 0.177 | **0.936** | 0.268 | 0.932 |
| | (0.014) | (0.036) | (0.003) | (0.005) | (0.024) | (0.009) | (0.03) |
| Avg CPU time | 2.962 | 3.2648 | 24.792 | 550.768 | 253.04 | 5.282 | 325.99 |

simulation results (Tables 1–4) with 50 repetitions. Codes for implementing our procedures are available at https://github.com/sagarknk/Graphical_HSL.

It can be clearly seen from Tables 1–4 that the horseshoe based methods generally perform the best. GHS has the smallest Stein's loss in all settings expect in Hubs when $(n, p) = (120, 100)$. This corroborates the finding of Li et al. [36] that GHS results in improved Kullback–Leibler risk properties (of which Stein's loss is an empirical measure) when compared to prior densities that are bounded above at the origin, e.g., BGL, and it is apparent from both tables that BGL has the worst Stein's loss. For GHS-LIKE-ECM and MCMC, the measures of Stein's loss are generally close to that of GHS, and much better compared to the other competing methods. A similar pattern emerges in the results for F norm, with the horseshoe-based methods once again outperforming

TABLE 2

*Performance measures of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix* $\mathbf{\Omega}_0$ *(Hubs and Cliques positive structures), where* $n = 120$ *and* $p = 100$*. All other abbreviations and definitions follow from Table 1.*

| | Hubs<br>90 nonzero pairs out of 4950<br>nonzero elements = 0.25 | | | | | | |
|---|---|---|---|---|---|---|---|
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 5.255 | 6.328 | 5.213 | 43.042 | 5.101 | **4.22** | 5.121 |
| | (0.263) | (0.414) | (0.261) | (0.802) | (0.455) | (0.369) | (0.467) |
| F norm | 3.018 | 3.432 | 3.003 | 4.295 | 2.544 | **2.415** | 2.574 |
| | (0.091) | (0.112) | (0.093) | (0.156) | (0.126) | (0.103) | (0.131) |
| TPR | 0.995 | 0.986 | **0.998** | 0.995 | 0.872 | 0.985 | 0.846 |
| | (0.007) | (0.017) | (0.002) | (0.008) | (0.04) | (0.014) | (0.039) |
| FPR | 0.101 | 0.045 | 0.983 | 0.186 | **0.003** | 0.062 | **0.003** |
| | (0.016) | (0.008) | (0.012) | (0.007) | (0.001) | (0.005) | (0.001) |
| MCC | 0.373 | 0.523 | 0.016 | 0.27 | **0.85** | 0.458 | 0.832 |
| | (0.027) | (0.039) | (0.006) | (0.006) | (0.027) | (0.015) | (0.03) |
| Avg CPU time | 1.739 | 1.76 | 48.54 | 549.196 | 252.94 | 5.811 | 328.659 |
| | Cliques positive<br>30 nonzero pairs out of 4950<br>nonzero elements = 0.75 | | | | | | |
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 6.010 | 7.48 | 5.98 | 44.163 | **1.781** | 3.753 | 2.386 |
| | (0.212) | (0.45) | (0.21) | (0.790) | (0.232) | (0.275) | (0.281) |
| F norm | 4.96 | 5.7 | 4.95 | 4.916 | **1.888** | 2.411 | 2.131 |
| | (0.1) | (0.13) | (0.107) | (0.103) | (0.184) | (0.142) | (0.177) |
| TPR | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| | (0) | (0) | (0) | (0) | (0) | (0) | (0) |
| FPR | 0.11 | 0.042 | 0.972 | 0.177 | **0.0008** | 0.068 | 0.002 |
| | (0.013) | (0.0011) | (0.014) | (0.006) | (0.005) | (0.006) | (0.001) |
| MCC | 0.22 | 0.353 | 0.013 | 0.166 | **0.94** | 0.277 | 0.879 |
| | (0.013) | (0.041) | (0.003) | (0.004) | (0.031) | (0.012) | (0.037) |
| Avg CPU time | 1.997 | 2.157 | 83.852 | 553.743 | 252.46 | 5.903 | 326.915 |

the competitors and performing similarly among themselves. It is worth noting, however, that the fully Bayesian approaches (GHS-LIKE–MCMC and GHS) generally result in the best statistical performance, at the expense of a considerably longer computing time, making the trade-off between fully Bayesian and penalized likelihood approaches apparent.

Coming next to variable selection results, one may expect the penalized likelihood approaches to really shine; since these methods produce exact zeros, unlike the Bayesian approaches that necessitate some form of post-processing. Nevertheless, the Bayesian approaches offer the advantage of controlling the trade-off between TPR and FPR, by varying the width of the credible interval, for example. With our chosen mechanism (i.e., a variable is considered not to be selected if the middle 50% credible interval includes zero), the GHS-LIKE-MCMC and GHS have the smallest TPR. Nevertheless, the penalized methods also have higher FPR in general (except for GHS-LKE-ECM), which results in lower MCC overall. In particular, the GSCAD estimate, which is not guaranteed to be positive definite in finite samples [20], seems not to work well in general.

TABLE 3

*Performance measures of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix $\mathbf{\Omega}_0$ (Random and Cliques negative structures), where $n = 120$ and $p = 200$. All other abbreviations and definitions follow from Table 1.*

| | | | | Random<br>29 nonzero pairs out of 19900<br>nonzero elements $\sim -\text{Unif}(0.2, 1)$ | | | |
|---|---|---|---|---|---|---|---|
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 10.06 | 15.578 | 9.975 | 117.092 | **3.073** | 11.109 | 3.354 |
| | (0.4) | (1.12) | (0.4) | (1.563) | (0.305) | (0.562) | (0.323) |
| F norm | 4.469 | 5.929 | 4.44 | 6.803 | **2.468** | 3.917 | 2.471 |
| | (0.151) | (0.176) | (0.156) | (0.162) | (0.137) | (0.108) | (0.129) |
| TPR | 0.944 | 0.845 | **0.999** | 0.982 | 0.848 | 0.97 | 0.842 |
| | (0.036) | (0.036) | (0.005) | (0.024) | (0.038) | (0.028) | (0.038) |
| FPR | 0.052 | 0.163 | 0.984 | 0.103 | 0.0001 | 0.066 | **0.000** |
| | (0.007) | (0.002) | (0.011) | (0.003) | (0.00007) | (0.003) | (0) |
| MCC | 0.152 | 0.242 | 0.004 | 0.11 | **0.882** | 0.138 | 0.866 |
| | (0.011) | (0.015) | (0.002) | (0.004) | (0.029) | (0.005) | (0.037) |
| Avg CPU time | 38.759 | 43.486 | 510.703 | 4484.22 | 1866.47 | 80.939 | 1879.155 |
| | | | | Cliques negative<br>60 nonzero pairs out of 19900<br>nonzero elements $= -0.45$ | | | |
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 11.604 | 18.088 | 11.541 | 125.138 | **3.985** | 12.467 | 4.827 |
| | (0.401) | (0.993) | (0.396) | (1.714) | (0.403) | (0.626) | (0.457) |
| F norm | 4.443 | 6.024 | 4.439 | 6.299 | **2.861** | 3.8 | 2.885 |
| | (0.088) | (0.143) | (0.076) | (0.168) | (0.209) | (0.126) | (0.182) |
| TPR | **1** | **1** | **1** | **1** | .975 | **1** | 0.983 |
| | (0) | (0) | (0) | (0) | (.173) | (0) | (0.014) |
| FPR | 0.066 | 0.016 | 0.998 | 0.099 | 0.0002 | 0.084 | **0.000** |
| | (0.004) | (0.002) | (0.005) | (0.002) | (0.0001) | (0.003) | (0) |
| MCC | 0.202 | 0.395 | 0.006 | 0.164 | **0.944** | 0.178 | 0.949 |
| | (0.006) | (0.027) | (0.001) | (0.002) | (0.16) | (0.003) | (0.02) |
| Avg CPU time | 32.936 | 36.49 | 548.26 | 4492.67 | 1876.96 | 70.683 | 1927.943 |

Additional numerical results investigating the choice of starting values for the GHS-LIKE-ECM algorithm is given in Appendix C. Performance measures corresponding to precision matrix estimates, estimated using the Bayesian structure learning framework of Mohammadi and Wit [40], are presented in Appendix D, Table 8. Except for the true positive rate in some cases, poor performance is observed in all other metrics.

## 6. Protein–protein interaction network in B-cell lymphoma

We analyze Reverse Phase Protein Array (RPPA) data of 33 patients with lymphoid neoplasm "Diffuse Large B-cell Lymphoma" to infer the protein interaction network. The data set consists of protein expressions for 67 genes across 12 pathways for all patients. As in simulations, we use 50% posterior credible intervals for variable selection in GHS, BGL and GHS-LIKE-MCMC. The estimated sparsity (% of zero elements) and number of non zeros in the lower triangle of the estimates are given in Table 5. The estimate of the global shrink-

TABLE 4

*Performance measures of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix* $\mathbf{\Omega}_0$ *(Hubs and Cliques positive structures), where* $n = 120$ *and* $p = 200$*. All other abbreviations and definitions follow from Table 1.*

| | | | | Hubs | | | |
|---|---|---|---|---|---|---|---|
| | | | | 180 nonzero pairs out of 19900 | | | |
| | | | | nonzero elements = 0.25 | | | |
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 12.407 | 15.243 | 12.331 | 123 | **11.692** | 12.825 | 12.355 |
| | (0.491) | (0.819) | (0.465) | (1.31) | (0.781) | (0.623) | (0.772) |
| F norm | 4.594 | 5.3 | 4.583 | 7.129 | **3.763** | 4.209 | 3.868 |
| | (0.01) | (0.152) | (0.084) | (0.16) | (0.132) | (0.107) | (0.127) |
| TPR | 0.99 | 0.976 | **1** | 0.991 | 0.779 | 0.986 | 0.792 |
| | (0.007) | (0.137) | (0) | (0.006) | (0.034) | (0.009) | (0.034) |
| FPR | 0.065 | 0.024 | 0.999 | 0.119 | **0.001** | 0.066 | 0.002 |
| | (0.005) | (0.006) | (0.001) | (0.003) | (0.0003) | (0.003) | (0) |
| MCC | 0.336 | 0.515 | 0.01 | 0.248 | **0.82** | 0.332 | 0.804 |
| | (0.015) | (0.043) | (0.003) | (0.003) | (0.024) | (0.006) | (0.025) |
| Avg CPU time | 17.847 | 19.917 | 517.33 | 4499.30 | 1870.57 | 74.808 | 1902.72 |
| | | | | Cliques positive | | | |
| | | | | 60 nonzero pairs out of 19900 | | | |
| | | | | nonzero elements = 0.75 | | | |
| | GL1 | GL2 | GSCAD | BGL | GHS | ECM | MCMC |
| Stein's loss | 14.523 | 17.262 | 14.477 | 126.487 | **3.797** | 13.512 | 4.951 |
| | (0.339) | (0.692) | (0.333) | (1.41) | (0.35) | (0.522) | (0.410) |
| F norm | 7.59 | 8.553 | 7.596 | 7.936 | **2.733** | 4.248 | 3.009 |
| | (0.1) | (0.115) | (0.091) | (0.109) | (0.181) | (0.142) | (0.177) |
| TPR | **1** | **1** | **1** | **1** | **1** | **1** | **1** |
| | (0) | (0) | (0) | (0) | (0) | (0) | (0) |
| FPR | 0.065 | 0.024 | 0.991 | 0.115 | **0.0004** | 0.08 | 0.001 |
| | (0.005) | (0.004) | (0.007) | (0.002) | (0.0001) | (0.002) | (0) |
| MCC | 0.205 | 0.335 | 0.01 | 0.15 | **0.959** | 0.184 | 0.899 |
| | (0.007) | (0.028) | (0.002) | (0.002) | (0.19) | (0.003) | (0.022) |
| Avg CPU time | 23.768 | 25.3 | 880.46 | 4497.97 | 1872.55 | 80.652 | 1929.587 |

TABLE 5

*Percentage of zeros (% Sparsity) and number of non-zero entries (NNZ) in the lower triangle of the precision matrix estimate of RPPA data for the competing approaches.*

| | MCMC | ECM | GHS | BGL | GL1 | GL2 | GSCAD |
|---|---|---|---|---|---|---|---|
| % Sparsity | 95.66 | 88.6 | 91.59 | 73.72 | 69.88 | 73.67 | $9.05 \times 10^{-4}$ |
| NNZ | 96 | 252 | 186 | 581 | 666 | 582 | 2209 |

age parameter for GHS-LIKE-ECM corresponding to $(n, p) = (33, 67)$, as found via solving equation (3.7) is $a = 0.0519$. We note that the GHS-LIKE-MCMC gives the sparsest estimate, almost 4% sparser than the GHS. This is consistent with prior studies that found robust gene networks are typically sparse [33]. As in the simulations, GSCAD performs the worst. To compare with a prior analysis of the same data set, we use the PRECISE framework of Ha et al. [25]. This method can infer directed edges, but we ignore the directionality since we are interested in interactions and not causation. The proportions of edges in the estimates that 'agree' (inferred by the estimate and the PRECISE framework) and 'do not agree' (inferred by the estimate, but not by the PRECISE frame-

TABLE 6
*Proportion of edges that 'agree' (AE) and 'do not agree' (NE) with the edges inferred using the PRECISE framework.*

| | MCMC | ECM | GHS | BGL | GL1 | GL2 | GSCAD |
|---|---|---|---|---|---|---|---|
| AE | 0.275 | 0.412 | 0.325 | 0.575 | 0.638 | 0.6 | 1 |
| NE | 0.034 | 0.101 | 0.074 | 0.247 | 0.284 | 0.247 | 0.984 |

work), are presented in Table 6. Protein networks realized from the estimates are presented in Figure 3. It can be seen that the GHS-LIKE-MCMC has the sparsest estimate among the methods that allow for interaction across all proteins, unlike the PRECISE framework that ignores interactions among proteins in different pathways, which may not be biologically justifiable.

## 7. Proofs of main results

### 7.1. *Proof of Theorem 4.6*

We use the general theory of posterior convergence rate as outlined in Theorem 2.1 of [24]. We also refer to several auxiliary lemmas from Appendix B throughout the proof. We need to show the following:

(i) the prior concentration rate of Kullback–Leibler $\epsilon_n^2$-neighborhoods is at least $\exp(-cn\epsilon_n^2)$ for some constant $c > 0$,
(ii) for a suitably chosen sieve of densities $\mathcal{P}_n$, the $\epsilon_n$-metric entropy of $\mathcal{P}_n$ is bounded by a constant multiple of $n\epsilon_n^2$,
(iii) the probability of the complement of the above sieve is exponentially small, that is, $\Pi(\mathcal{P}_n^c) \leq \exp(-c'n\epsilon_n^2)$, for some constant $c' > 0$.

The above three parts together give the posterior convergence rate $\epsilon_n$ with respect to the Hellinger distance on the space of densities of the precision matrix. Owing to the intrinsic relationship between the Hellinger distance and the Frobenius distance for precision matrices as given by Lemma A.1 of [3], we get the desired posterior convergence rate.

**(i) Prior concentration** We first define $\mathcal{B}(p_{\boldsymbol{\Omega}_0}, \epsilon_n)$, the $\epsilon_n^2$-neighborhoods of the true density in the Kullback–Leibler sense. For

$$K(p_1, p_2) = \int p_1 \log(p_1/p_2),\ V(p_1, p_2) = \int p_1 \log^2(p_1/p_2),$$

this is defined as $\mathcal{B}(p_{\boldsymbol{\Omega}_0}, \epsilon_n) = \{p_{\boldsymbol{\Omega}}: K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2, V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \leq \epsilon_n^2\}$. For $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$, $\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)$, let $d_1, \ldots, d_p$ denote the eigenvalues of $\boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}$. Then, using Lemma B.1, we have,

$$K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) = -\frac{1}{2}\sum_{i=1}^{p} \log d_i - \frac{1}{2}\sum_{i=1}^{n}(1 - d_i),$$

$$V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) = \frac{1}{2}\sum_{i=1}^{n}(1 - d_i)^2 + K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}})^2. \tag{7.1}$$

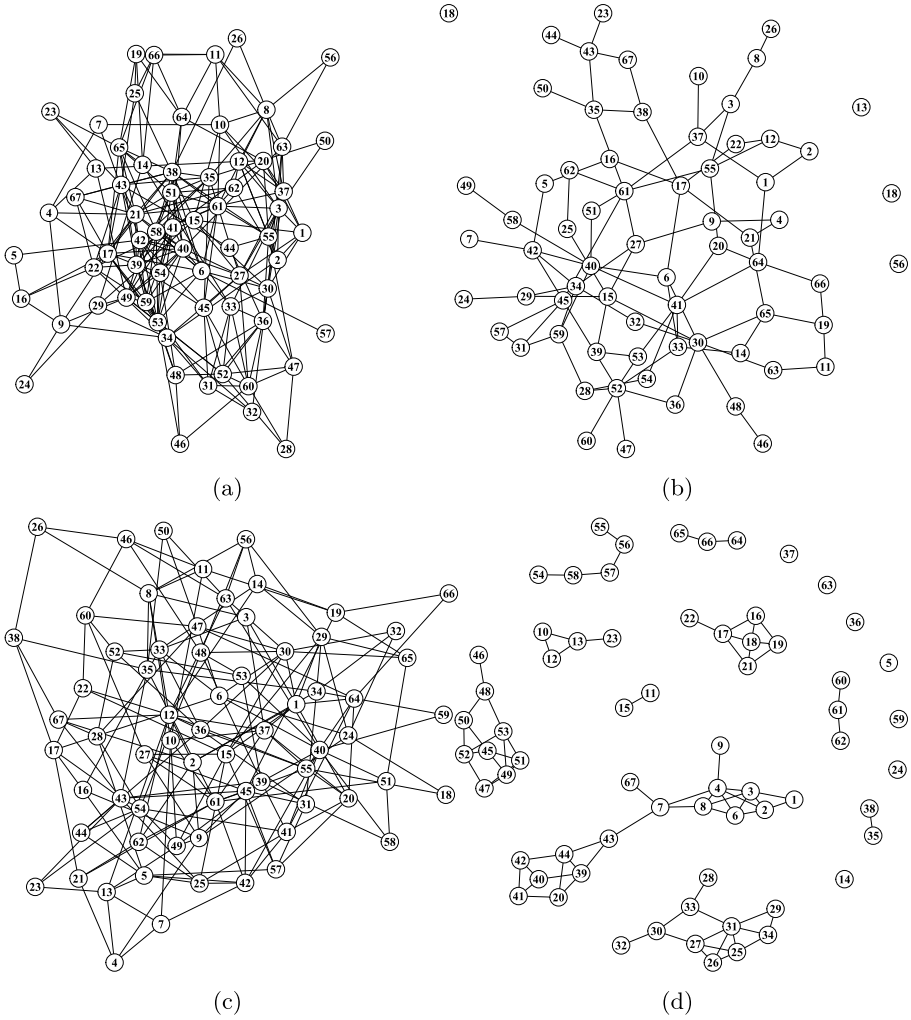

FIG 3. *(a), (b), (c) and (d) correspond to RPPA networks for GHS-LIKE-ECM, GHS-LIKE-MCMC, GHS and PRECISE. The nodes are numbered from 1 to 67, which are proteins. The map between node numbers and protein names is given in the Appendix E, Table 10.*

Note that $\sum_{i=1}^{n}(1-d_i)^2 = \|\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}\|_2^2$, so that, when the expression $\|\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}\|_2^2$ is small, we have, $\max_{1\le i\le p}|1-d_i| < 1$; see [3]. This gives, using (7.1),

$$K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) = -\frac{1}{2}\sum_{i=1}^{p}\log d_i - \frac{1}{2}\sum_{i=1}^{n}(1-d_i) \lesssim \sum_{i=1}^{n}(1-d_i)^2,$$

$$V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \lesssim \sum_{i=1}^{n}(1-d_i)^2.$$

Observe that,

$$\sum_{i=1}^{n}(1-d_i)^2 = \|\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}\|_2^2 = \|\boldsymbol{\Omega}_0^{-1/2}(\boldsymbol{\Omega}-\boldsymbol{\Omega}_0)\boldsymbol{\Omega}_0^{-1/2}\|_2^2$$
$$\leq \|\boldsymbol{\Omega}_0^{-1}\|_2^2\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_2^2 \leq \varepsilon_0^{-2}\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_2^2.$$

Hence, for a sufficiently small constant $c_1 > 0$, we have,

$$\Pi\big(\mathcal{B}(p_0,\epsilon_n)\big) \geq \Pi\{\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_2 \leq c_1\epsilon_n\} \geq \Pi\{\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_\infty \leq c_1\epsilon_n/p\}.$$

The proposed prior on $\boldsymbol{\Omega}$ has a bounded spectral norm. However, such a constraint can only increase the prior concentration, since $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$, $\varepsilon_0 < L$. Hence, we may pretend component-wise independence of the elements of $\boldsymbol{\Omega}$, so that the above expression can be simplified as products of marginal prior probabilities. We have,

$$\Pi(\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_\infty \leq c_1\epsilon_n/p) \gtrsim (c_1\epsilon_n/p)^{(p+s)} \prod_{\{(i,j):\omega_{ij,0}=0\}} \pi(|\omega_{ij}| \leq c_1\epsilon_n/p)$$
$$\geq (c_1\epsilon_n/p)^{(p+s)} \min_{\{(i,j):\omega_{ij,0}=0\}} \big\{\pi(|\omega_{ij}| \leq c_1\epsilon_n/p)\big\}^{\binom{p}{2}-s}.$$

Note that, from equation (B.1) in Lemma B.2, we have, for all $1 \leq i, j \leq p$,

$$\big\{\pi(|\omega_{ij}| \leq c_1\epsilon_n/p)\big\}^{\binom{p}{2}-s} \geq \big\{1-p^{-b_1'}\big\}^{\binom{p}{2}-s} \to 1.$$

Thus, $\Pi(\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_\infty \leq c_1\epsilon_n/p) \gtrsim (c_1\epsilon_n/p)^{(p+s)}$. The prior concentration rate condition thus gives, $(p+s)(\log p + \log(1/\epsilon_n)) \asymp n\epsilon_n^2$, so as to yield $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$.

**(ii) Choosing the sieve and controlling metric entropy** We now carefully choose a sieve in the space of prior densities to control its Hellinger metric entropy. Consider the sieve $\mathcal{P}_n$ such that the maximum number of elements of $\boldsymbol{\Omega}$ exceeding $\delta_n = \epsilon_n/p^\nu, \nu > 0$ is at most $\bar{r}_n$, and the absolute values of the entries of $\boldsymbol{\Omega}$ are at most $B$. Formally, the sieve is thus given by,

$$\mathcal{P}_n = \bigg\{\boldsymbol{\Omega} \in \mathcal{M}_p^+(L): \sum_{j,k} \mathbb{1}(|\omega_{jk}| > \delta_n) \leq \bar{r}_n, \|\boldsymbol{\Omega}\|_\infty \leq B\bigg\},$$

where $\delta_n = \epsilon_n/p^\nu$ and some sufficiently large $B > 0$. We shall choose $B$ in such a way that the metric entropy condition is satisfied. Note that, for $\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2 \in \mathcal{M}_p^+(L)$, $\|\boldsymbol{\Omega}_1-\boldsymbol{\Omega}_2\|_2^2 \leq p^2\|\boldsymbol{\Omega}_1-\boldsymbol{\Omega}_2\|_\infty^2$, so that, if $\|\boldsymbol{\Omega}_1-\boldsymbol{\Omega}_2\|_\infty^2 \leq \epsilon_n^2/p^{2\nu}$, where $\nu$ is such that $B \leq p^{\nu-1}$, we have, $\|\boldsymbol{\Omega}_1-\boldsymbol{\Omega}_2\|_2^2 \leq \epsilon_n^2/p^{2(\nu-1)}$. The $\epsilon_n/p^\nu$-metric entropy w.r.t. the $L_\infty$-norm is given by

$$\log\left\{\left(\frac{Bp^\nu}{\epsilon_n}\right)^p \sum_{j=1}^{\bar{r}_n}\left(\frac{Bp^\nu}{\epsilon_n}\right)^j \binom{\binom{p}{2}}{j}\right\}$$

$$
\begin{aligned}
&= \log\left\{\left(\frac{Bp^{\nu}}{\epsilon_n}\right)^{p}\right\} + \log\left\{\sum_{j=1}^{\bar{r}_n}\left(\frac{Bp^{\nu}}{\epsilon_n}\right)^{j}\binom{\binom{p}{2}}{j}\right\} \\
&\leq \log\left\{\left(\frac{Bp^{\nu}}{\epsilon_n}\right)^{p}\right\} + \log\left\{\bar{r}_n\left(\frac{Bp^{\nu}}{\epsilon_n}\right)^{\bar{r}_n}\binom{p+\binom{p}{2}}{\bar{r}_n}\right\} \\
&\lesssim (\bar{r}_n + p)\big(\log p + \log B + \log(1/\epsilon_n)\big).
\end{aligned}
$$

Choosing $\bar{r}_n \sim k_1 n\epsilon_n^2/\log n$, $k_1 > 0$, and $B \sim k_2 n\epsilon_n^2$, $k_2 > 0$, the above metric entropy is bounded by a constant multiple of $n\epsilon_n^2$. Since $\|\boldsymbol{\Omega}_1\|_{(2,2)} \leq p\|\boldsymbol{\Omega}_1\|_{\infty} \leq pB \leq p^{\nu}$, and $h^2(p_1, p_2) \leq p^2\|\boldsymbol{\Omega}_1\|_{(2,2)}^2\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_{\infty}^2$, the $\epsilon_n$-metric entropy with respect to the Hellinger distance is also bounded by a constant multiple of $n\epsilon_n^2$. Thus, the rate $\epsilon_n$ obtained via the prior concentration rate calculation satisfies the metric entropy condition as well.

**(iii) Controlling probability of the complement of the sieve** The task of controlling the probability of the complement of the sieve can further be divided into two sub-parts. Note that,

$$
\Pi\big(\mathcal{P}_n^c\big) \leq \Pi(N \geq \bar{r}_n + 1) + \Pi(\|\boldsymbol{\Omega}\|_{\infty} > B).
$$

We will calculate the probabilities in the right-hand side of the above display under an unconstrained case, and then take care of the truncation used in the prior for $\boldsymbol{\Omega}$, given by $\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)$, by finding a suitable lower bound for the event $\{0 < L^{-1} < \|\boldsymbol{\Omega}\|_{(2,2)} < L < \infty\}$. Let us denote the prior under the unconstrained case by $\Pi^*$.

Define $N = \#\{(i,j)\colon |\omega_{ij}| > \delta_n\}$. Note that, $N \sim Bin(p_n^*, \nu_n)$, $p_n^* = \binom{p}{2}$, $\nu_n = \Pr(|\omega_{ij}| > \delta_n)$. Using results on bounding the Binomial CDF as in [50], we have,

$$
\begin{aligned}
\Pi^*(N \geq \bar{r}_n + 1) &\leq 1 - \Phi\big\{\big(2p_n^* H\big[\nu_n, \bar{r}_n/p_n^*\big]\big)^{1/2}\big\} \\
&\leq (2\pi)^{-1/2}\big(2p_n^* H\big[\nu_n, \bar{r}_n/p_n^*\big]\big)^{-1/2}\exp\big\{-p_n^* H\big[\nu_n, \bar{r}_n/p_n^*\big]\big\},
\end{aligned}
$$

where

$$
p_n^* H\big[\nu_n, \bar{r}_n/p_n^*\big] = \bar{r}_n \log\left(\frac{\bar{r}_n}{p_n^*\nu_n}\right) + \big(p_n^* - \bar{r}_n\big)\log\left(\frac{p_n^* - \bar{r}_n}{p_n^* - p_n^*\nu_n}\right).
$$

It suffices to prove that $p_n^* H[\nu_n, \bar{r}_n/p_n^*] \geq O(n\epsilon_n^2)$. We have,

$$
p_n^* H\big[\nu_n, \bar{r}_n/p_n^*\big] \approx \bar{r}_n \log\left(\frac{\bar{r}_n}{p_n\nu_n}\right) + \big(p_n^2 - \bar{r}_n\big)\log\left(\frac{p_n^2 - \bar{r}_n}{p_n^2 - p_n^2\nu_n}\right).
$$

For the first term on the RHS above, we have, $\bar{r}_n \log\{\bar{r}_n/(p_n\nu_n)\} \geq \bar{r}_n \log \bar{r}_n + b_1'\bar{r}_n \log p_n$, since $\nu_n \leq p_n^{-b_1'}$ vide (B.1) in Lemma B.2. Note that $\bar{r}_n \log p \sim \bar{r}_n \log n \asymp n\epsilon_n^2$, so as to get $\bar{r}_n \log\{\bar{r}_n/(p_n\nu_n)\} \geq n\epsilon_n^2$. For the second term, we have, $(p_n^2 - \bar{r}_n)\log\{(p_n^2 - \bar{r}_n)/(p_n^2 - p_n^2\nu_n)\} \asymp \bar{r}_n(1 - \bar{r}_n/p_n^2) = o(n\epsilon_n^2)$. Hence, we get, $p_n^* H[\nu_n, \bar{r}_n/p_n^*] \geq O(n\epsilon_n^2)$, which implies,

$$
\Pi^*(N \geq \bar{r}_n + 1) \leq \exp\big\{-C' n\epsilon_n^2\big\}, \tag{7.2}
$$

for some $C' > 0$. From (B.2) in Lemma B.2, for the choice of $B \sim k_2 n\epsilon_n^2$ as outlined in the metric entropy condition above, we have,

$$\Pi^*(\|\mathbf{\Omega}\|_\infty > B) \leq 2p^2 \exp\bigl(-k_2 n\epsilon_n^2\bigr). \tag{7.3}$$

Combining equations (7.2) and (7.3), we get, for a suitable constant $c_3 > 0$,

$$\Pi^*(N \geq \bar{r}_n + 1) + \Pi^*(\|\mathbf{\Omega}\|_\infty > B) \lesssim \exp\bigl(-c_3 n\epsilon_n^2\bigr). \tag{7.4}$$

Combining (7.4) and (B.5), we get, for a suitable constant $c_4 > 0$,

$$\begin{aligned}\Pi\bigl(\mathcal{P}_n^c\bigr) = \frac{\Pi^*(\mathcal{P}_n^c)}{\Pi(\mathbf{\Omega} \in \mathcal{M}_p^+(L))} &\lesssim \exp\bigl(-c_3 n\epsilon_n^2\bigr) L^{-p} \exp\bigl(C_1' p\bigr) \\ &= \exp\bigl(-c_3 n\epsilon_n^2 - p\log L + C_1' p\bigr) \\ &\lesssim \exp\bigl(-c_4 n\epsilon_n^2\bigr).\end{aligned}$$

Hence, the complement of the chosen sieve has exponentially small prior probability. Thus, $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$ is the posterior convergence rate and the result is established.

### 7.2. Proof of Corollary 4.7

The proof of this result is exactly similar to that of Theorem 4.6. The proof of the latter relies on Lemma B.2 and Lemma B.5 that are specific to the graphical horseshoe-like prior, and the corollaries given by Corollary B.3 and Corollary B.6 are respectively their counterparts corresponding to the graphical horseshoe prior. The utilization of the general lemma on Kullback–Leibler distance computations as outlined in Lemma B.1 remains identical in the present case.

### 7.3. Proof of Lemma 4.8

We will prove concavity by proving the second derivative is negative. By direct calculations:

$$\frac{d^2}{dx^2}\bigl(pen_a(x)\bigr) = \frac{d^2}{dx^2}\left(-\log\log\left(1+\frac{a}{x^2}\right)\right) = -\frac{2a((a+3x^2)\log(1+a/x^2) - 2a)}{x^2(a+x^2)^2(\log^2(1+a/x^2))}. \tag{7.5}$$

Since the denominator of the RHS in (7.5) is always positive, we can investigate the sign of the double derivative of the above penalty function by considering only the numerator, and furthermore as $a > 0$, we need the following to hold to prove concavity:

$$\bigl(a+3x^2\bigr)\log\bigl(1+a/x^2\bigr) - 2a \geq 0. \tag{7.6}$$

Substituting $\log(1+a/x^2)$ by $z$, so that $x^2 = a/(\exp(z)-1)$, we have $z \geq 0$, and the RHS of (7.6) is given by,

$$\bigl(a+3x^2\bigr)\log\bigl(1+a/x^2\bigr) - 2a = \left(a + \frac{3a}{\exp(z)-1}\right)z - 2a$$

$$= a\left(\frac{3z + (z-2)(\exp(z)-1)}{\exp(z)-1}\right)$$
$$> a\left(\frac{z(1+z)}{\exp(z)-1}\right) > 0, \quad \text{since}\, \exp(z) > 1 + z.$$

This proves the (strict) concavity of the graphical horseshoe-like penalty function.

### *7.4. Proof of Theorem 4.9*

Consider the MAP estimator of the precision matrix corresponding to the graphical horseshoe-like prior $\hat{\boldsymbol{\Omega}}^{\text{MAP}}$ as outlined in Section 4.2. Define $\boldsymbol{\Delta} = (\!(\delta_{ij})\!) = \boldsymbol{\Omega} - \boldsymbol{\Omega}_0$ such that $\|\boldsymbol{\Delta}\|_2 = M\epsilon_n$, $M > 0$ is a large constant. Here, $\boldsymbol{\Omega}_0 = (\!(\omega_{ij,0})\!)$ is the true precision matrix. The true covariance matrix is $\boldsymbol{\Sigma}_0 = (\!(\sigma_{ij,0})\!)$, and the natural estimator of the covariance is $\boldsymbol{S} = (\!(s_{ij})\!)$. Consider $Q(\boldsymbol{\Omega})$ as defined in (4.2). If we can show that for some small $\varepsilon > 0$,

$$\mathbb{P}\Big(\sup_{\|\boldsymbol{\Delta}\|_2 = M\epsilon_n} Q(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) < Q(\boldsymbol{\Omega}_0)\Big) \geq 1 - \varepsilon,$$

then there exists a local maximizer $\hat{\boldsymbol{\Omega}}$ such that $\|\hat{\boldsymbol{\Omega}} - \boldsymbol{\Omega}_0\|_2 = O_P(\epsilon_n)$. We have,

$$Q(\boldsymbol{\Omega}) = l(\boldsymbol{\Omega}) - \sum_{i<j} pen(\omega_{ij}) = \frac{n}{2}\log\det(\boldsymbol{\Omega}) - \frac{n}{2}\text{tr}(\boldsymbol{S}\boldsymbol{\Omega}) - \sum_{i<j} pen(\omega_{ij})$$
$$= \frac{n}{2}\left\{\log\det(\boldsymbol{\Omega}) - \text{tr}(\boldsymbol{S}\boldsymbol{\Omega}) - \frac{2}{n}\sum_{i<j} pen(\omega_{ij})\right\} = \frac{n}{2}h(\boldsymbol{\Omega}),\ \text{say.}$$

Let us denote as $(2/n)\, pen(\omega_{ij})$ as $p_n(\omega_{ij})$. This gives,

$$h(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) - h(\boldsymbol{\Omega}) = \log\det(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) - \text{tr}\big(\boldsymbol{S}(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta})\big) - \log\det(\boldsymbol{\Omega}_0)$$
$$+ \text{tr}(\boldsymbol{S}\boldsymbol{\Omega}_0) - \sum_{i<j}\big\{p_n(\omega_{ij,0} + \delta_{ij}) - p_n(\omega_{ij,0})\big\}. \tag{7.7}$$

By Taylor's series expansion of logarithm of the determinant of a matrix, we have,

$$\log\det(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) - \log\det(\boldsymbol{\Omega}_0)$$
$$= \text{tr}(\boldsymbol{\Sigma}_0\boldsymbol{\Delta}) - \text{vec}(\boldsymbol{\Delta})^T\left[\int_0^1 (1-\nu)(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1} \otimes (\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\, d\nu\right]\text{vec}(\boldsymbol{\Delta}).$$

Plugging the above in (7.7), we have the expression for $h(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta}) - h(\boldsymbol{\Omega})$ as

$$\text{tr}\big[(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\boldsymbol{\Delta}\big] - \text{vec}(\boldsymbol{\Delta})^T\left[\int_0^1 (1-\nu)(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1} \otimes (\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\, d\nu\right]\text{vec}(\boldsymbol{\Delta})$$
$$- \sum_{i<j}\big\{p_n(\omega_{ij,0} + \delta_{ij}) - p_n(\omega_{ij,0})\big\}$$

$$= \mathrm{I} + \mathrm{II} + \mathrm{III}, \text{ say.} \tag{7.8}$$

We shall now separately bound the three terms I, II, and III. For bounding I, we have,

$$\left|\mathrm{tr}\left[(\boldsymbol{\Sigma}_0 - \boldsymbol{S})\boldsymbol{\Delta}\right]\right| \leq \left|\sum_{i\neq j}(\sigma_{ij,0} - s_{ij})\delta_{ij}\right| + \left|\sum_{i}(\sigma_{ii,0} - s_{ii})\delta_{ii}\right|. \tag{7.9}$$

Using Boole's inequality and Lemma B.7, we have, with probability tending to one,

$$\max_{i\neq j}|s_{ij} - \sigma_{ij,0}| \leq C_1\left(\frac{\log p}{n}\right)^{1/2}.$$

Hence, the first term in the RHS of (7.9) is bounded by $C_1(\log p/n)^{1/2}\|\boldsymbol{\Delta}^-\|_1$. By Cauchy-Schwarz inequality and Lemma B.7, we have, with probability tending to one,

$$\begin{aligned}\left|\sum_{i}(\sigma_{ii,0} - s_{ii})\delta_{ii}\right| &\leq \left\{\sum_{i}(\sigma_{ii,0} - s_{ii})^2\right\}^{1/2}\|\boldsymbol{\Delta}^+\|_2 \\ &\leq p^{1/2}\max_{1\leq i\leq p}|s_{ii} - \sigma_{ii,0}|\|\boldsymbol{\Delta}^+\|_2 \\ &\leq C_2\left(\frac{p\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^+\|_2 \leq C_2\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^+\|_2.\end{aligned}$$

Thus, combining the bounds above, with probability approaching one, a bound for expression I is,

$$\mathrm{I} \leq C_1\left(\frac{\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^-\|_1 + C_2\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^+\|_2. \tag{7.10}$$

Now we proceed to find suitable bounds for expression II. Note that II is upper bounded by the negative of the minimum of

$$\mathrm{vec}(\boldsymbol{\Delta})^T\left[\int_0^1(1-\nu)(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\otimes(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\,d\nu\right]\mathrm{vec}(\boldsymbol{\Delta}).$$

Using the result that $\min_{\|\boldsymbol{x}\|_2=1}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \mathrm{eig}_1(\boldsymbol{A})$, we have,

$$\begin{aligned}&\min\left\{\mathrm{vec}(\boldsymbol{\Delta})^T\left[\int_0^1(1-\nu)(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\otimes(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\,d\nu\right]\mathrm{vec}(\boldsymbol{\Delta})\right\} \\ &= \|\boldsymbol{\Delta}\|_2^2\,\mathrm{eig}_1\left[\int_0^1(1-\nu)(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\otimes(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}\,d\nu\right] \\ &\geq \|\boldsymbol{\Delta}\|_2^2\int_0^1(1-\nu)\mathrm{eig}_1^2(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1}d\nu \geq \frac{1}{2}\|\boldsymbol{\Delta}\|_2^2\min_{0\leq\nu\leq1}\mathrm{eig}_1^2(\boldsymbol{\Omega}_0 + \nu\boldsymbol{\Delta})^{-1} \\ &\geq \frac{1}{2}\min\left\{\mathrm{eig}_1^2(\boldsymbol{\Omega}_0 + \boldsymbol{\Delta})^{-1}\colon \|\boldsymbol{\Delta}\|_2 \leq M\epsilon_n\right\}.\end{aligned}$$

Note that, $\text{eig}_1^2(\boldsymbol{\Omega}_0+\boldsymbol{\Delta})^{-1} = \text{eig}_p^{-2}(\boldsymbol{\Omega}_0+\boldsymbol{\Delta}) \geq (\|\boldsymbol{\Omega}_0\|_{(2,2)}+\|\boldsymbol{\Delta}\|_{(2,2)})^{-2} \geq \varepsilon_0^2/2$, with probability tending to one. The last inequality follows from the fact that $\|\boldsymbol{\Delta}\|_{(2,2)} \leq \|\boldsymbol{\Delta}\|_2 = o(1)$. Hence, with probability tending to one, we have the bound for expression II as

$$\text{II} \leq -\frac{1}{4}\varepsilon_0^2\|\boldsymbol{\Delta}\|_2^2. \tag{7.11}$$

Finally, we proceed to find suitable bounds for expression III. Let us denote the set $\mathcal{S} = \{(i,j)\colon \omega_{ij,0} = 0, i<j\}$. This set comprises of the indices in the uppper triangle of the true precision matrix that are exactly equal to zero. The complement of $\mathcal{S}$ consists of the indices with non-zero entries in the uppper triangle of the same. We can partition expression III (without the negative sign) as

$$\begin{aligned}
&\sum_{i<j}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} \\
&= \sum_{(i,j)\in\mathcal{S}}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} + \sum_{(i,j)\in\mathcal{S}^c}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} \\
&= \sum_{(i,j)\in\mathcal{S}}\big\{p_n(\delta_{ij}) - p_n(0)\big\} + \sum_{(i,j)\in\mathcal{S}^c}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} \\
&> \frac{M'}{n} + \sum_{(i,j)\in\mathcal{S}^c}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\}.
\end{aligned}$$

The last inequality follows from the fact that $pen(\theta) \to -\infty$ as $|\theta| \to 0$, and hence the first term in the above expression is larger than $M'/n$ for some large $M' > 0$. This implies:

$$\sum_{i<j}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} > \sum_{(i,j)\in\mathcal{S}^c}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\}.$$

By Taylor's series expansion of $p_n(\omega_{ij}+\delta_{ij})$ around $\omega_{ij,0}(\neq 0)$, we have,

$$p_n(\omega_{ij}+\delta_{ij}) - p_n(\omega_{ij,0}) = \delta_{ij}p_n'(\omega_{ij,0}) + \frac{1}{2}\delta_{ij}^2 p_n''(\omega_{ij,0})\big(1+o(1)\big).$$

Since $-x \leq |x|$, we can write,

$$\begin{aligned}
&-\sum_{(i,j)\in\mathcal{S}^c}\big\{p_n(\omega_{ij,0}+\delta_{ij}) - p_n(\omega_{ij,0})\big\} \\
&\leq \max\big\{|p_n'(\omega_{ij,0})|\big\}\sum_{(i,j)\in\mathcal{S}^c}|\delta_{ij}| + \frac{1}{2}\max\big\{|p_n''(\omega_{ij,0})|\big\}\sum_{(i,j)\in\mathcal{S}^c}\delta_{ij}^2\big(1+o(1)\big) \\
&\leq \max\big\{|p_n'(\omega_{ij,0})|\big\}\|\boldsymbol{\Delta}^{-1}\|_1 + \frac{1}{2}\max\big\{|p_n''(\omega_{ij,0})|\big\}\|\boldsymbol{\Delta}^{-}\|_2^2\big(1+o(1)\big) \\
&\leq s^{1/2}\max\big\{|p_n'(\omega_{ij,0})|\big\}\|\boldsymbol{\Delta}\|_2 + \frac{1}{2}\max\big\{|p_n''(\omega_{ij,0})|\big\}\|\boldsymbol{\Delta}\|_2^2\big(1+o(1)\big).
\end{aligned} \tag{7.12}$$

Now, note that,

$$|p_n'(\omega_{ij,0})| = \frac{2}{n}|pen'(\omega_{ij,0})| = \frac{2}{n}\frac{2a/|\omega_{ij,0}|^3}{(1+\frac{a}{\omega_{ij,0}^2})\log(1+\frac{a}{\omega_{ij,0}^2})}.$$

Since $(1+x)\log(1+x) > x$ for $x > -1, x \neq 0$, we have, $|p_n'(\omega_{ij,0})| < 4/(n|\omega_{ij,0}|)$. We now arrive at a suitable bound for the double derivative of the penalty. Note that, for $\theta \neq 0$,

$$\begin{aligned} pen''(\theta) &= -\frac{2a\{(a+3\theta^2)\log(1+\frac{a}{\theta^2}) - 2a\}}{\theta^6(1+\frac{a}{\theta^2})^2\log^2(1+\frac{a}{\theta^2})} \\ &\leq -\frac{2a\{(a+3\theta_0^2)\log(1+\frac{a}{\theta_0^2}) - 2a\}}{\theta^6(1+\frac{a}{\theta^2})^2\log^2(1+\frac{a}{\theta^2})}, \end{aligned}$$

where $\theta_0 = \arg\max_\theta\{-(a+3\theta^2)\log(1+a/\theta^2)\} = (ak)^{1/2}$, $k = \{\exp(z_0) - 1\}^{-1}$, $z_0 \approx 1.0356$. Hence,

$$\begin{aligned} |pen''(\theta)| &\leq \frac{2a|(a+3\theta_0^2)\log(1+\frac{a}{\theta_0^2}) - 2a|}{\theta^6(1+\frac{a}{\theta^2})^2\log^2(1+\frac{a}{\theta^2})} \\ &\leq \frac{2a|(a+3\theta_0^2)\log(1+\frac{a}{\theta_0^2}) - 2a|}{\theta^6\frac{a^2}{\theta^4}} \\ &= \frac{2|(a+3\theta_0^2)\log(1+\frac{a}{\theta_0^2}) - 2a|}{a\theta^2} \\ &= \frac{2|(a+3ak)\log(1+\frac{a}{ak}) - 2a|}{a\theta^2} \approx \frac{C_3}{\theta^2}, \end{aligned}$$

where $C_3 > 0$ is a constant not depending on $n$ or $a$. This gives,

$$|p_n''(\omega_{ij,0})| = \frac{2}{n}|pen''(\omega_{ij,0})| < \frac{2C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}.$$

Thus, expression III can be bounded as follows:

$$\text{III} \leq s^{1/2}\|\boldsymbol{\Delta}\|_2\frac{4}{n\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|} + \frac{C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}\|\boldsymbol{\Delta}\|_2^2\big(1+o(1)\big). \quad (7.13)$$

Combining Equations (7.10), (7.11), and (7.13), we have, with probability tending to one,

$$Q(\boldsymbol{\Omega}_0+\boldsymbol{\Delta}) - Q(\boldsymbol{\Omega}_0)$$

$$
\begin{aligned}
&\leq C_1\left(\frac{\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^-\|_1 + C_2\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}^+\|_2 - \frac{1}{4}\varepsilon_0^2\|\boldsymbol{\Delta}\|_2^2 \\
&+ s^{1/2}\|\boldsymbol{\Delta}\|_2\frac{4}{n\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|} + \frac{C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}\|\boldsymbol{\Delta}\|_2^2\big(1+o(1)\big) \\
&\leq C_1\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}\|_2 + C_2\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}\|_2 - \frac{1}{4}\varepsilon_0^2\|\boldsymbol{\Delta}\|_2^2 \\
&+ s^{1/2}\|\boldsymbol{\Delta}\|_2\frac{4}{n\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|} + \frac{C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}\|\boldsymbol{\Delta}\|_2^2\big(1+o(1)\big) \\
&\leq \|\boldsymbol{\Delta}\|_2^2\left\{C_1\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}\|_2^{-1} + C_2\left(\frac{(p+s)\log p}{n}\right)^{1/2}\|\boldsymbol{\Delta}\|_2^{-1} - \frac{1}{4}\varepsilon_0^2\right. \\
&\left.+ (p+s)^{1/2}\|\boldsymbol{\Delta}\|_2^{-1}\frac{4}{n\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|} + \frac{C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}\big(1+o(1)\big)\right\} \\
&= \|\boldsymbol{\Delta}\|_2^2\left\{\frac{C_1}{M} + \frac{C_2}{M} - \frac{1}{4}\varepsilon_0^2 + \frac{4}{(n\log p)^{1/2}\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|}\right. \\
&\qquad\left.+\frac{C_3}{n\min_{(i,j)\in\mathcal{S}^c}\omega_{ij,0}^2}\big(1+o(1)\big)\right\} < 0,
\end{aligned}
$$

for $M$ sufficiently large, and owing to the fact that the last two terms inside the bracket in the above display are $o(1)$ as $\min_{(i,j)\in\mathcal{S}^c}|\omega_{ij,0}|$ are bounded away from zero. This completes the proof.

## 8. Concluding remarks

Our main contribution in this paper is twofold: first, we propose a fully analytical prior–penalty dual termed the *graphical horseshoe-like* for inference in graphical models, and second, we provide the first ever optimality results for both the frequentist point estimate as well as the fully Bayesian posterior. Consequently, we also establish the first Bayesian optimality results for the graphical horseshoe prior of Li et al. [36]. Our simulation studies clearly establish that the family of horseshoe based priors perform the best among state-of-the-art competitors across a wide range of data generating mechanisms, and suggest a potential trade-off between computational burden and statistical performance vis-à-vis penalized likelihood and fully Bayesian procedures. Our analysis of the RPPA data establishes the proposed approach as an effective regularizer of a gene interaction network; useful for identifying the key interactions in the disease etiology of cancer.

Although we focus on the estimation of $\boldsymbol{\Omega}$, two other important aspects of network inference are edge selection and the associated uncertainty quantification. Here we use posterior credible intervals for edge selection, but it might be interesting to incorporate other methods that have been proposed for variable selection with shrinkage priors, such as 2-means [8] or shrinkage factor

thresholding [51], with appropriate modifications. On a related note, it will be interesting to establish the Bayes risk and the oracle under $0-1$ loss and we conjecture that global-local shrinkage priors will attain such oracular risk with suitable assumptions on the prior tails and the global shrinkage parameter. Finally, it will be worth investigating whether one can extend the methods for generalized linear models, e.g. graphical models with exponential families as node-conditional distributions [63]. It has been shown that while restricting the response distribution to natural exponential families with quadratic variance functions, shrinkage estimators enjoy certain optimality properties [62], and it remains to be settled whether similar properties hold true for graphical models as well.

## Appendix A: The marginal graphical horseshoe-like prior and implications for estimation algorithms

The graphical horseshoe-like prior on the individual off-diagonal elements $\omega_{ij}$ has a nice Gaussian scale mixture representation as outlined in Section 2. However, the marginal prior on these elements are not horseshoe-like, owing to the positive definite constraint on the precision matrix $\mathbf{\Omega}$. In this section, we argue that the hierarchical representation based on the scale-mixtures induces the proposed marginal prior on $\mathbf{\Omega}$ and all the related marginal and conditional distributions are proper. Alongside this, we also argue that the intractable normalizing constant in the marginal prior of $\mathbf{\Omega}$ does not affect the conditional expectation calculations for executing the expectation conditional maximization steps in our computations.

The marginal prior on $\mathbf{\Omega}$ given the global scale parameter $a$ can be written as,

$$\pi(\mathbf{\Omega} \mid a) = C(a)^{-1} \prod_{i<j} \pi(\omega_{ij} \mid a) \mathbb{1}_{\mathcal{M}_p^+}(\mathbf{\Omega}), \tag{A.1}$$

where $C(a)$ is the normalizing constant depending on $a$. Using the Gaussian scale-mixture representation, we have a hierarchical representation of the above prior as,

$$\pi(\mathbf{\Omega} \mid \nu, a) = C(\nu, a)^{-1} \prod_{i<j} \pi(\omega_{ij} \mid \nu_{i,j}, a) \mathbb{1}_{\mathcal{M}_p^+}(\mathbf{\Omega}), \tag{A.2}$$

where $C(\nu, a)$ is an intractable constant depending on $\nu$ and $a$. The prior on $\nu$ is,

$$\pi(\nu) \propto C(\nu, a) \prod_{i<j} \pi(\nu_{ij}) = C_2(a)^{-1} C(\nu, a) \prod_{i<j} \pi(\nu_{ij}), \tag{A.3}$$

where $C_2(a)$ is a constant such that

$$C_2(a) = \int C(\nu, a) \prod_{i<j} \pi(\nu_{ij})\, d\nu.$$

The constant $C(a)$ in (A.1) is finite because,

$$C(a) = \int \prod_{i<j} \pi(\omega_{ij} \mid a) \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega})\, d(\omega_{ij})_{i\leq j} < 2K \int \prod_{i<j} \pi(\omega_{ij} \mid a)\, d(\omega_{ij})_{i<j} < \infty,$$

where $K < \infty$ is such that $|\omega_{ii}| < K, (i = 1, \ldots, p)$, since $\boldsymbol{\Omega}$ is restricted to be positive definite and hence the diagonal elements are finite. Also,

$$\begin{aligned} C(\nu, a) &= \int \prod_{i<j} \pi(\omega_{ij} \mid \nu_{i,j}, a) \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega})\, d(\omega_{ij})_{i\leq j} \\ &< 2K \int \prod_{i<j} \pi(\omega_{ij} \mid \nu_{i,j}, a)\, d(\omega_{ij})_{i<j} < \infty. \end{aligned}$$

The induced marginal prior on $\boldsymbol{\Omega}$ based on the hierarchical representation as in (A.2) and (A.3) is,

$$\pi^*(\boldsymbol{\Omega} \mid a) = C_2(a)^{-1} \prod_{i<j} \pi(\omega_{ij} \mid a) \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega}).$$

Since $\int \pi(\boldsymbol{\Omega} \mid a)\, d\boldsymbol{\Omega} = \int \pi^*(\boldsymbol{\Omega} \mid a)\, d\boldsymbol{\Omega} = 1$, it immediately implies that $C(a) = C_2(a)$. Thus, the intractable constant $C(\nu, a)$ in (A.2) and (A.3) cancels out in the hierarchical representation, so as to arrive at the induced marginal prior (A.1). The above results also establish that the priors $\pi(\boldsymbol{\Omega} \mid a), \pi(\boldsymbol{\Omega} \mid \nu, a)$, and $\pi(\nu)$ are proper.

We now show that it suffices to consider the component-wise scale-mixture representation of the horseshoe-like prior to find the conditional expectation of the latent parameters $\nu_{ij}$ in the expectation step (see equation (3.1)) of the expectation conditional maximization algorithm. The conditional distribution of $\nu$ given $\boldsymbol{\Omega}$ and $a$ can be written as,

$$\begin{aligned} \pi(\nu \mid \boldsymbol{\Omega}, a) &= \frac{\pi(\boldsymbol{\Omega}, \nu \mid a)}{\pi(\boldsymbol{\Omega} \mid a)} = \frac{\pi(\boldsymbol{\Omega} \mid \nu, a)\pi(\nu)}{\pi(\boldsymbol{\Omega} \mid a)} \\ &= \frac{\prod_{i<j} \pi(\omega_{ij} \mid \nu_{ij}, a) \prod_{i<j} \pi(\nu_{ij}) \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega})}{\prod_{i<j} \pi(\omega_{ij} \mid a) \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega})}. \end{aligned}$$

This gives,

$$\pi(\nu_{ij} \mid \boldsymbol{\Omega}, a) = \frac{\pi(\omega_{ij} \mid \nu_{ij}, a)\pi(\nu_{ij})}{\pi(\omega_{ij} \mid a)} \mathbb{1}_{\mathcal{M}_p^+}(\boldsymbol{\Omega}).$$

Thus, the expectation step (3.1) holds given that the conditional maximization step produces positive definite estimates of $\boldsymbol{\Omega}$ in each iteration.

## Appendix B: Auxiliary lemmas

**Lemma B.1.** *Let $p_k$ denote the density of a $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Sigma}_k)$ random variable, $k = 1, 2$. Denote the corresponding precision matrices by $\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k^{-1}, k = 1, 2$. Then,*

$$\mathbb{E}_{p_1}\left\{\log \frac{p_1}{p_2}(\boldsymbol{X})\right\} = \frac{1}{2}\left\{\log\det \boldsymbol{\Omega}_1 - \log\det \boldsymbol{\Omega}_2 + \operatorname{tr}\left(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2\right) - d\right\},$$

$$\mathrm{Var}_{p_1}\left\{\log\frac{p_1}{p_2}(\boldsymbol{X})\right\} = \frac{1}{2}\,\mathrm{tr}\big\{\big(\boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1/2} - \boldsymbol{I}_d\big)^2\big\}.$$

*Proof.* Let us define $\boldsymbol{A} = \boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}^{-1/2}$. Note that, for a random variable $\boldsymbol{Z} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma})$, we have,

$$\mathbb{E}\big(\boldsymbol{Z}^T\boldsymbol{\Lambda}\boldsymbol{Z}\big) = \mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}),\ \mathrm{Var}\big(\boldsymbol{Z}^T\boldsymbol{\Lambda}\boldsymbol{Z}\big) = 2\,\mathrm{tr}(\boldsymbol{\Lambda}\boldsymbol{\Sigma}\boldsymbol{\Lambda}\boldsymbol{\Sigma}).$$

Then, for $\boldsymbol{X} \sim \mathcal{N}_d(\boldsymbol{0}, \boldsymbol{\Sigma}_1)$,

$$\begin{aligned}
\mathbb{E}_{p_1}\left\{\log\frac{p_1}{p_2}(\boldsymbol{X})\right\} &= \frac{1}{2}\big\{\log\det\boldsymbol{\Omega}_1 - \log\det\boldsymbol{\Omega}_2 + \mathbb{E}_{p_1}\big(\boldsymbol{X}^T(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{X}\big)\big\} \\
&= \frac{1}{2}\big\{\log\det\boldsymbol{\Omega}_1 - \log\det\boldsymbol{\Omega}_2 + \mathrm{tr}\big[(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{\Sigma}_1\big]\big\} \\
&= \frac{1}{2}\big\{\log\det\boldsymbol{\Omega}_1 - \log\det\boldsymbol{\Omega}_2 + \mathrm{tr}\big(\boldsymbol{\Omega}_1^{-1}\boldsymbol{\Omega}_2\big) - d\big\}.
\end{aligned}$$

Also,

$$\begin{aligned}
\mathrm{Var}_{p_1}\left\{\log\frac{p_1}{p_2}(\boldsymbol{X})\right\} &= \mathbb{E}_{p_1}\left[\log\frac{p_1}{p_2}(\boldsymbol{X}) - \mathbb{E}_{p_1}\left\{\log\frac{p_1}{p_2}(\boldsymbol{X})\right\}\right]^2 \\
&= \frac{1}{4}\mathbb{E}_{p_1}\big\{\boldsymbol{X}^T(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{X} - \mathbb{E}_{p_1}\big(\boldsymbol{X}^T(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{X}\big)\big\}^2 \\
&= \frac{1}{4}\,\mathrm{Var}_{p_1}\big\{\boldsymbol{X}^T(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{X}\big\} \\
&= \frac{1}{4}2\,\mathrm{tr}\big\{(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{\Omega}_1^{-1}(\boldsymbol{\Omega}_2 - \boldsymbol{\Omega}_1)\boldsymbol{\Omega}_1^{-1}\big\} \\
&= \frac{1}{2}\,\mathrm{tr}\big\{\big(\boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1/2} - \boldsymbol{I}_d\big)^2\big\}.
\end{aligned}$$

□

**Lemma B.2.** *Consider the horseshoe-like prior $\pi(\theta \mid a)$. Then, for the global shrinkage parameter $a$ satisfying the condition $a^{1/2} < n^{-1/2}p^{-b_1}(s\log p)^{1/2}$ for some sufficiently large constant $b_1 > 0$, and a constant $\nu > 0$, we have,*

$$1 - \int_{-\epsilon_n/p^\nu}^{\epsilon_n/p^\nu} \pi(\theta \mid a)\,d\theta \le p^{-b_1'}, \tag{B.1}$$

*for some constants $\nu, b_1' > 0$. Additionally, for some sufficiently large constant $B \sim b_2 n\epsilon_n^2$, if the global scale parameter satisfies the condition $a/B^2 < p^{-2b_3}$ for some constant $b_3 > 0$, we have,*

$$-\log\left(\int_{|\theta|>B} \pi(\theta \mid a)\,d\theta\right) \gtrsim B. \tag{B.2}$$

*Proof.* We have,

$$1 - \int_{-\epsilon_n/p^\nu}^{\epsilon_n/p^\nu} \pi(\theta \mid a)\,d\theta = \int_{|\theta|>\epsilon_n/p^\nu} \pi(\theta \mid a)\,d\theta$$

$$
\begin{aligned}
&= \int_{|\theta|>\epsilon_n/p^\nu} \frac{1}{2\pi a^{1/2}} \log\left(1+\frac{a}{\theta^2}\right) d\theta \\
&= 2\int_{\epsilon_n/p^\nu}^{\infty} \frac{1}{2\pi a^{1/2}} \log\left(1+\frac{a}{\theta^2}\right) d\theta \\
&\leq \int_{\epsilon_n/p^\nu}^{\infty} \frac{1}{\pi a^{1/2}} \frac{a}{\theta^2}\, d\theta = \frac{2}{\pi}\frac{a^{1/2}p^\nu}{\epsilon_n}.
\end{aligned}
$$

Note that, for $a^{1/2} < n^{-1/2}p^{-b_1}(s\log p)^{1/2}$, the right hand side of the display above is bounded by $p^{-b_1'}$, for $0 < b_1' \leq b_1 - \nu$. This proves the first part of the lemma. For the second part, note that,

$$
\int_{|\theta|>B} \pi(\theta \mid a)\, d\theta \leq \frac{2}{\pi}\frac{a^{1/2}}{B}.
$$

Hence, for the condition $a^{1/2}/B < p^{-b_3}$, we have,

$$
\int_{|\theta|>B} \pi(\theta \mid a)\, d\theta \lesssim p^{-b_3} = \exp(-b_3 \log p) \lesssim \exp\bigl(-b_2 n\epsilon_n^2\bigr),
$$

which implies that, for $B \sim b_2 n\epsilon_n^2$,

$$
-\log\left(\int_{|\theta|>B} \pi(\theta \mid a)\, d\theta\right) \gtrsim B.
$$

This completes the proof. □

**Corollary B.3.** *The above lemma holds true under the same conditions on the global shrinkage parameter for the horseshoe prior as well.*

*Proof.* Note that the prior density of the horseshoe prior satisfies

$$
p_{HS}(\theta \mid a) < \frac{2}{a^{1/2}(2\pi)^{3/2}} \log\left(1+\frac{2a}{\theta^2}\right), \tag{B.3}
$$

which implies that, retracing the steps in the proof of Lemma B.2 above,

$$
\int_{|\theta|>t} p_{HS}(\theta \mid a)\, d\theta \lesssim \frac{a^{1/2}}{t}. \tag{B.4}
$$

The result thus follows immediately. □

We now present the Gershgorin Circle Theorem [10], that will be required in the proof of our main result on posterior convergence rate. The actual theorem holds for complex matrices, but we only need the result for real matrices.

**Theorem B.4** (Gershgorin Circle Theorem for real matrices)**.** *Let $\boldsymbol{A} = ((a_{ij}))$ be a $p$-dimensional real-valued matrix with real eigenvalues. Define $R_i = \sum_{j\neq i} |a_{ij}|$,*

*$i = 1, \ldots, p$, the row sums of the absolute entries of $\boldsymbol{A}$ excluding the diagonal element. Then, each eigenvalue of $\boldsymbol{A}$ is in at least one of the disks*

$$\mathcal{D}_i(\boldsymbol{A}) = \{z : |z - a_{ii}| \leq R_i\},\ 1 \leq i \leq p.$$

*Equivalently, the $p$ eigenvalues of $\boldsymbol{A}$ are contained in the region in the real plane determined by*

$$\mathcal{D}(\boldsymbol{A}) = \bigcup_{i=1}^{p} \mathcal{D}_i(\boldsymbol{A}).$$

*Proof.* The eigenvalue equation for $\boldsymbol{A}$ is given by $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$, where $\lambda$ is an eigenvalue of $\boldsymbol{A}$ and $\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$ is the corresponding non-zero eigenvector. Let us consider $1 \leq m \leq p$ such that $|x_m| = \|\boldsymbol{x}\|_\infty$. Then, the above eigenvalue equation implies that, $\sum_{j=1}^{p} a_{mj}x_j = \lambda x_m$. Rearranging the terms, we get, $\sum_{j \neq m} a_{mj}x_j = (\lambda - a_{mm})x_m$, which implies that,

$$|\lambda - a_{mm}||x_m| = \left|\sum_{j \neq m} a_{mj}x_j\right| \leq \sum_{j \neq m} |a_{mj}||x_j| \leq |x_m| \sum_{j \neq m} |a_{mj}|.$$

Hence, for any eigenvalue $\lambda$ of $\boldsymbol{A}$, we have, $|\lambda - a_{mm}| \leq \sum_{j \neq m} |a_{mj}|$. Thus, each of the $p$ eigenvalues of $\boldsymbol{A}$ must lie in at least one of the disks $\mathcal{D}_i(\boldsymbol{A})$ as defined in the theorem above. This completes the proof. □

**Lemma B.5.** *For the graphical horseshoe-like prior (2.2), under the assumption that the global scale parameter satisfies the condition $a^{1/2} < L^{-1}n^{-1/2}p^{-(2+u)} \times (s \log p)^{1/2}$, $u > 0$, the prior probability owing to the constraint $\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)$ has the lower bound*

$$\Pi\big(\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)\big) \gtrsim L^p \exp\big(-C_1' p\big), \tag{B.5}$$

*for some suitable constant $C_1' > 0$.*

*Proof.* We shall use the Gershgorin Circle theorem presented in Theorem B.4. Each of the eigenvalues of $\boldsymbol{\Omega}$, given by $\mathrm{eig}_1(\boldsymbol{\Omega}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{\Omega})$, lies in the interval $\bigcup_{j=1}^{p}[\omega_{jj} \mp \sum_{k=1,k\neq j}^{p} |\omega_{kj}|]$. This implies,

$$\Pi\big(\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)\big) \geq \Pi\left(\min_j \left(\omega_{jj} - \sum_{k=1,k\neq j}^{p} |\omega_{kj}|\right) > 0, \boldsymbol{\Omega} \in \mathcal{M}_p^+(L)\right).$$

For the constraint that $\min_j(\omega_{jj} - \sum_{k=1,k\neq j}^{p} |\omega_{kj}|) > 0$,

$$\mathrm{eig}_p(\boldsymbol{\Omega}) = \|\boldsymbol{\Omega}\|_{(2,2)} \leq \|\boldsymbol{\Omega}\|_{(1,1)} = \max_j \left(\omega_{jj} + \sum_{k=1,k\neq j}^{p} |\omega_{kj}|\right) \leq 2 \max_j \omega_{jj},$$

and,

$$\mathrm{eig}_1(\boldsymbol{\Omega}) \geq \min_j \left(\omega_{jj} - \sum_{k=1,k\neq j}^{p} |\omega_{kj}|\right).$$

Thus,

$$\begin{aligned}
&\Pi\big(\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)\big) \\
&\geq \Pi\Bigg(L^{-1} \leq \min_j \Bigg(\omega_{jj} - \sum_{k=1,k\neq j}^{p} |\omega_{kj}|\Bigg) \leq 2\max_j \omega_{jj} \leq L\Bigg) \\
&\geq \Pi\Big(L^{-1} \leq \min_j \big(\omega_{jj} - L^{-1}\big) \leq 2\max_j \omega_{jj} \leq L \mid \max_{k\neq j} |\omega_{kj}| \\
&< (Lp)^{-1}\Big)\Pi\Big(\max_{k\neq j} |\omega_{kj}| < (Lp)^{-1}\Big) \\
&= \Pi\Big(L^{-1} \leq \min_j \big(\omega_{jj} - L^{-1}\big) \leq 2\max_j \omega_{jj} \leq L\Big)\Pi\Big(\max_{k\neq j} |\omega_{kj}| < (Lp)^{-1}\Big).
\end{aligned} \tag{B.6}$$

Note that,

$$\begin{aligned}
&\Pi\Big\{L^{-1} \leq \min_j \big(\omega_{jj} - L^{-1}\big) \leq 2\max_j \omega_{jj} \leq L\Big\} \\
&\geq \Pi\big(2L^{-1} \leq \omega_{jj} \leq L/2,\ 1 \leq j \leq p\big) \\
&= \prod_{j=1}^{p} \Pi\big(2L^{-1} \leq \omega_{jj} \leq L/2\big) \sim L^p.
\end{aligned} \tag{B.7}$$

Also, from (B.1) in Lemma B.2, we get,

$$\begin{aligned}
\Pi\Big(\max_{k\neq j} |\omega_{kj}| < (Lp)^{-1}\Big) &= \prod_{k\neq j} \big\{1 - \Pi\big(|\omega_{kj}| > (Lp)^{-1}\big)\big\} \\
&\geq \big(1 - C_0 a^{1/2} Lp\big)^{p^2} \geq \exp\big(-C_1 a^{1/2} Lp^3\big) \\
&\geq \exp\big(-C_1' p\big).
\end{aligned} \tag{B.8}$$

The last inequality follows from the fact that $a^{1/2} < L^{-1} n^{-1/2} p^{-(2+u)} (s \log p)^{1/2}$, $u > 0$, and that $s \log p = O(n)$. Therefore, combining (B.6), (B.7) and (B.8), we get, $\Pi(\boldsymbol{\Omega} \in \mathcal{M}_p^+(L)) \gtrsim L^p \exp(-C_1' p)$, thus completing the proof. □

**Corollary B.6.** *The above lemma holds true for the graphical horseshoe prior as well under the same conditions on the global shrinkage parameter.*

*Proof.* The proof of this result is exactly similar to that of Lemma B.5. The lower bound on the off-diagonal entries follows immediately from Corollary B.3. The rest of the arguments remain intact. □

**Lemma B.7** (Lemma A.3 in Bickel and Levina [9])**.** *Let* $\boldsymbol{Z}_i \overset{iid}{\sim} \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathrm{eig}_p(\boldsymbol{\Sigma}) \leq \varepsilon_0 < \infty$. *Then, if* $\boldsymbol{\Sigma} = (\!(\sigma_{ij})\!)$,

$$\Pr\left[\left|\sum_{i=1}^{n} Z_{ij} Z_{ik} - \sigma_{jk}\right| \geq nt\right] \leq c_1 \exp\big(-c_2 nt^2\big),\ |t| \leq \delta,$$

*where* $c_1, c_2$ *and* $\delta$ *depend on* $\varepsilon_0$ *only.*

## Appendix C: Diagnostics: choice of starting values for the ECM algorithm and trace plots for the ECM and MCMC algorithms

Since the likelihood surface under the GHS-Like prior is likely highly multimodal, and the ECM algorithm is only guaranteed to find a local mode, we provide additional numerical results investigating the effect of starting values on the estimates. Given a true precision matrix $\mathbf{\Omega}_0$ and $(n, p) = (50, 100)$ we generate 50 data sets, and perform estimation with 1, 10, 20 and 50 randomly chosen starting points. The accuracy measures of these estimates are represented as 1*, 10*, 20* and 50* in the Table 7. In general, we observe that the estimates from 50 different starting points perform the best in terms of Stein's loss, Frobenius norm, and TPR; while being slightly worse in terms of FPR and MCC.

In the presence of a highly multimodal likelihood surface, it is safe to believe that true signal which might be missed for any given starting point. Hence

TABLE 7

*Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates, Matthews Correlation Coefficient of precision matrix estimates for GHS-LIKE-ECM over 50 data sets with $p = 100$ and $n = 50$. The best performer in each row is shown in bold. Average CPU time is in seconds.*

| | Random | | | | Hubs | | | |
|---|---|---|---|---|---|---|---|---|
| nonzero pairs | 35/4950 | | | | 90/4950 | | | |
| nonzero elements | $\sim -\text{Unif}(0.2, 1)$ | | | | 0.25 | | | |
| $p = 100, n = 50$ | 1* | 10* | 20* | 50* | 1* | 10* | 20* | 50* |
| Stein's loss | 9.624 | 8.503 | 8.196 | **8.302** | 12.563 | 10.494 | 10.408 | **10.268** |
| | (0.915) | (0.801) | (0.749) | (0.749) | (0.83) | (0.885) | (0.831) | (0.81) |
| F norm | 3.674 | 3.344 | 3.286 | **3.279** | 4.166 | 3.672 | 3.645 | **3.616** |
| | (0.237) | (0.191) | (0.191) | (0.178) | (0.197) | (0.188) | (0.171) | (0.174) |
| TPR | 0.703 | 0.816 | 0.814 | **0.825** | 0.551 | 0.756 | 0.766 | **0.772** |
| | (0.044) | (0.042) | (0.046) | (0.039) | (0.049) | (0.053) | (0.054) | (0.053) |
| FPR | **0.021** | 0.054 | 0.058 | 0.063 | **0.015** | 0.038 | 0.041 | 0.044 |
| | (0.002) | (0.004) | (0.004) | (0.005) | (0.002) | (0.004) | (0.004) | (0.004) |
| MCC | **0.329** | 0.271 | 0.26 | 0.253 | **0.464** | 0.435 | 0.426 | 0.419 |
| | (0.024) | (0.014) | (0.013) | (0.011) | (0.034) | (0.027) | (0.025) | (0.023) |
| Avg CPU time | 6.735 | ... | ... | ... | 5.378 | ... | ... | ... |
| | Cliques negative | | | | Cliques positive | | | |
| nonzero pairs | 30/4950 | | | | 30/4950 | | | |
| nonzero elements | -0.45 | | | | 0.75 | | | |
| $p = 100, n = 50$ | 1* | 10* | 20* | 50* | 1* | 10* | 20* | 50* |
| Stein's loss | 9.149 | 7.417 | 7.276 | **7.289** | 13.896 | 9.156 | 8.719 | **8.65** |
| | (0.774) | (0.673) | (0.673) | (0.67) | (1.032) | (0.83) | (0.822) | (0.848) |
| F norm | 3.746 | 3.231 | 3.18 | **3.174** | 5.453 | 4.178 | 3.995 | **3.974** |
| | (0.265) | (0.263) | (0.215) | (0.255) | (0.245) | (0.241) | (0.268) | (0.261) |
| TPR | 0.911 | 0.995 | 0.999 | **1** | 0.741 | **0.997** | **0.997** | **0.997** |
| | (0.029) | (0.012) | (0.005) | (0) | (0.048) | (0.001) | (0.01) | (0.01) |
| FPR | **0.021** | 0.053 | 0.058 | 0.064 | **0.023** | 0.051 | 0.055 | 0.059 |
| | (0.002) | (0.004) | (0.005) | (0.005) | (0.002) | (0.003) | (0.004) | (0.005) |
| MCC | **0.433** | 0.312 | 0.3 | 0.287 | **0.344** | 0.318 | 0.308 | 0.298 |
| | (0.016) | (0.013) | (0.012) | (0.011) | (0.024) | (0.009) | (0.01) | (0.01) |
| Avg CPU time | 5.08 | ... | ... | ... | 5.088 | ... | ... | ... |

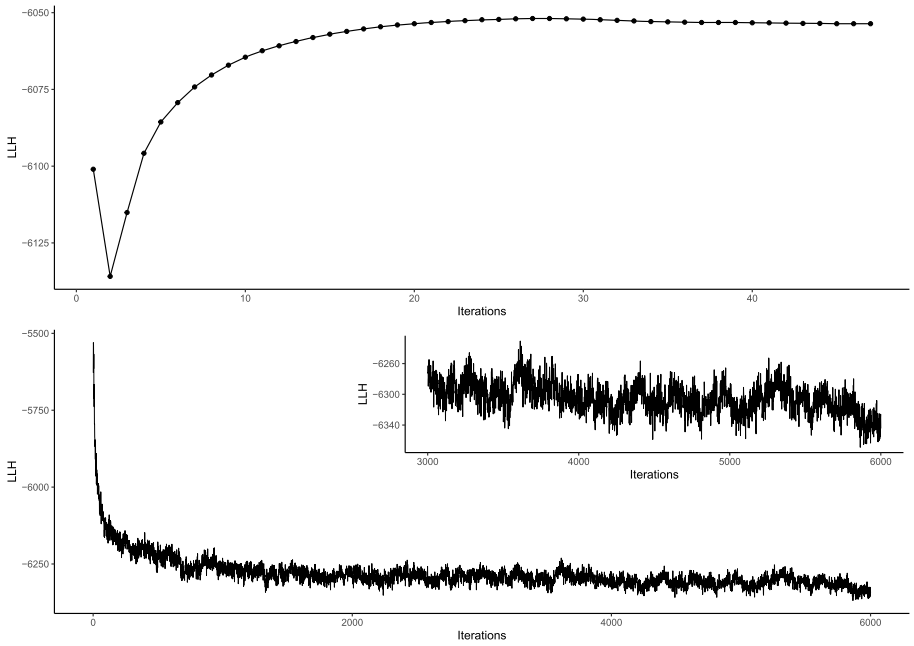


FIG 4. *Top and bottom panels show the plot of log-likelihood (LLH) vs. Iterations when the precision matrix was estimated for a representative data set using GHS-LIKE-ECM and GHS-LIKE-MCMC procedures respectively, for 'Hubs' structure when* $n = 120$, $p = 100$. *The inset plot in the bottom panel shows the zoomed-in version of the plot for the last 3000 samples.*

averaging across different starting values leads to an improvement in terms of most metrics and this what we choose to follow in our examples. Nevertheless, it is reassuring to see the final results are not too sensitive to the choice of starting values.

Further, Figure 4 shows a sample trace plot of log-likelihood when the precision matrix was estimated for a representative data set using GHS-LIKE-ECM and GHS-LIKE-MCMC. It is apparent that convergence to a local maximum (for ECM) and to the stationary distribution (for MCMC) occur relatively quickly. Similar behavior was observed in all other settings.

## Appendix D: Additional simulation results

Performance measures of precision matrix estimates, estimated using Bayesian structure learning framework of Mohammadi and Wit [40], is presented in Table 8. Comparison between performance measures of precision matrix estimates, estimated using graphical horseshoe-like MCMC, for a representative simulation setting, over 50 and 100 replications, is presented in Table 9.

TABLE 8. *Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates, Matthews Correlation Coefficient of precision matrix estimates over 50 data sets generated by multivariate normal distributions with precision matrix* $\mathbf{\Omega}_0$. *The precision matrix is estimated Bayesian structure learning framework of Mohammadi and Wit [40], using the R-package* `BDgraph` *[41]. Average CPU time is in seconds.*

| | $n = 120$, $p = 100$ | | | | $n = 120$, $p = 200$ | | | |
|---|---|---|---|---|---|---|---|---|
| | Structure of $\mathbf{\Omega}_0$ | | | | Structure of $\mathbf{\Omega}_0$ | | | |
| | Hubs | Random | Cliques pos. | Cliques neg. | Hubs | Random | Cliques pos. | Cliques neg. |
| Stein's loss | 12.227 | 10.119 | 9.58 | 9.255 | 142.629 | 139.66 | 142.537 | 148.101 |
| | (0.965) | (1.053) | (0.893) | (0.788) | (1.744) | (2.746) | (2.423) | (2.407) |
| F norm | 5.138 | 4.816 | 4.708 | 4.778 | 9.599 | 10.278 | 10.71 | 9.643 |
| | (0.264) | (0.313) | (0.3) | (0.295) | (0.253) | (0.339) | (0.412) | (0.307) |
| TPR | 0.954 | 0.943 | 0.998 | 0.973 | 0.754 | 0.819 | 0.864 | 0.784 |
| | (0.018) | (0.038) | (0.008) | (0.029) | (0.03) | (0.071) | (0.05) | (0.043) |
| FPR | 0.218 | 0.193 | 0.195 | 0.183 | 0.124 | 0.123 | 0.117 | 0.113 |
| | (0.007) | (0.007) | (0.005) | (0.006) | (0.002) | (0.001) | (0.002) | (0.002) |
| MCC | 0.233 | 0.158 | 0.156 | 0.157 | 0.178 | 0.08 | 0.126 | 0.116 |
| | (0.008) | (0.008) | (0.003) | (0.007) | (0.009) | (0.008) | (0.009) | (0.008) |
| Avg CPU time | 44.942 | 50.749 | 37.634 | 33.605 | 222.083 | 251.553 | 241.138 | 226.941 |

TABLE 9

*Mean (sd) Stein's loss, Frobenius norm, true positive rates and false positive rates, Matthews Correlation Coefficient of precision matrix estimates over 50 data sets (50 replications) and 100 data sets (100 replications) generated by multivariate normal distributions with precision matrix $\mathbf{\Omega}_0$ (Hub structure), where $n = 120$ and $p = 100$. The precision matrix is estimated by graphical horseshoe-like MCMC. Average CPU time is in seconds.*

| | 50 replications | 100 replications |
|---|---|---|
| Stein's loss | 5.121 | 5.12 |
| | (0.467) | (0.493) |
| F norm | 2.574 | 2.576 |
| | (0.131) | (0.134) |
| TPR | 0.846 | 0.84 |
| | (0.039) | (0.04) |
| FPR | 0.003 | 0.003 |
| | (0.001) | (0.001) |
| MCC | 0.832 | 0.826 |
| | (0.03) | (0.031) |
| Avg CPU time | 328.659 | 327.53 |

## Appendix E: Additional details on the proteomics data

Table 10 provides the map between the node numbers and protein names in Figure 3.

## Acknowledgments

The authors would like to thank the Associate Editor and three anonymous referees for their valuable comments and feedback that resulted in substantial improvement in the revised version of the manuscript. S.B. was supported by DST INSPIRE Faculty Award, Grant No. 04/2015/002165, and IIM Indore Young Faculty Research Chair award, J.D. was supported by U.S. National Science Foundation Grant DMS-2015460, K.S. and A.B. were partially supported by U.S. National Science Foundation Grant DMS-2014371.

TABLE 10. *Map between node numbers and protein names in Figure 3.*

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | BAK1 | 11 | MYH11 | 21 | PCNA | 31 | TP53 | 41 | ATK1S1 | 51 | MAPK14 | 61 | MTOR |
| 2 | BAX | 12 | RAB11A, RAB11B | 22 | FOXM1 | 32 | RAD50 | 42 | TSC2 | 52 | RPS6KA1 | 62 | RPS6 |
| 3 | BID | 13 | CTNNB1 | 23 | CDH1 | 33 | RAD51 | 43 | INPP4B | 53 | YBX1 | 63 | RB1 |
| 4 | BCL2L11 | 14 | GADPH | 24 | CLDN7 | 34 | XRCC1 | 44 | PTEN | 54 | EGFR | 64 | ESR1 |
| 5 | CASP7 | 15 | RBM15 | 25 | TP53BP1 | 35 | FN1 | 45 | ARAF | 55 | ERBB2 | 65 | PGR |
| 6 | BAD | 16 | CDK1 | 26 | ATM | 36 | CDH2 | 46 | JUN | 56 | ERBB3 | 66 | AR |
| 7 | BCL2 | 17 | CCNB1 | 27 | CHEK1 | 37 | COL6A1 | 47 | RAF1 | 57 | SHC1 | 67 | GATA3 |
| 8 | BCL2L1 | 18 | CCNE1 | 28 | CHEK2 | 38 | SERPINE1 | 48 | MAPK8 | 58 | SRC | | |
| 9 | BIRC2 | 19 | CCNE2 | 29 | XRCC5 | 39 | ATK1, ATK2, ATK3 | 49 | MAPK1, MAPK3 | 59 | EIF4EBP1 | | |
| 10 | CAV1 | 20 | CDKN1B | 30 | MRE11A | 40 | GKS3A, GKS3B | 50 | MAP2K1 | 60 | RPS6KB1 | | |

## References


[1] Banerjee, S., Castillo, I., and Ghosal, S. (2021). Bayesian inference in high-dimensional models. *arXiv preprint arXiv:2101.04491*. MR3251280

[2] Banerjee, S. and Ghosal, S. (2014). Posterior convergence rates for estimating large precision matrices using graphical models. *Electronic Journal of Statistics*, 8(2):2111–2137. MR3273620

[3] Banerjee, S. and Ghosal, S. (2015). Bayesian structure learning in graphical models. *Journal of Multivariate Analysis*, 136:147–162. MR3321485

[4] Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). Normal variance-mean mixtures and z distributions. *International Statistical Review*, pages 145–159. MR0678296

[5] Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). Default Bayesian analysis with global-local shrinkage priors. *Biometrika*, 103(4):955–969. MR3620450

[6] Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019a). The horseshoe-like regularization for feature subset selection. *Sankhya B*, pages 1–30. MR4256316

[7] Bhadra, A., Datta, J., Polson, N. G., and Willard, B. T. (2019b). Lasso meets horseshoe: A survey. *Statistical Science*, 34(3):405–427. MR4017521

[8] Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490. MR3449048

[9] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 36(1):199–227. MR2387969

[10] Brualdi, R. A. and Mellendorf, S. (1994). Regions in the complex plane containing the eigenvalues of a matrix. *The American Mathematical Monthly*, 101(10):975–985. MR1304322

[11] Cai, T., Liu, W., and Luo, X. (2011). A constrained $\ell_1$ minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607. MR2847973

[12] Callot, L., Caner, M., Önder, A. Ö., and Ulaşan, E. (2019). A nodewise regression approach to estimating large portfolios. *Journal of Business & Economic Statistics*, pages 1–12. MR4235193

[13] Candès, E. J. and Tao, T. (2010). The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080. MR2723472

[14] Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480. MR2650751

[15] Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018. MR3375874

[16] Dawid, A. P., Stone, M., and Zidek, J. V. (1973). Marginalization paradoxes in Bayesian and structural inference. *Journal of the Royal Statistical Society: Series B*, 35(2):189–213. MR0365805

[17] Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive

lasso and SCAD penalties. *The Annals of Applied Statistics*, 3(2):521–541. MR2750671

[18] Fan, J. and Li, R. (2001a). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360. MR1946581

[19] Fan, J. and Li, R. (2001b). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360. MR1946581

[20] Fan, J., Liao, Y., and Liu, H. (2016). An overview of the estimation of large covariance and precision matrices. *The Econometrics Journal*, 19(1):C1–C32. MR3501529

[21] Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

[22] Friedman, J., Hastie, T., and Tibshirani, R. (2018). *glasso: Graphical Lasso: Estimation of Gaussian Graphical Models*. R package version 1.10. MR4107668

[23] Gan, L., Narisetty, N. N., and Liang, F. (2019). Bayesian regularization for graphical models with unequal shrinkage. *Journal of the American Statistical Association*, 114(527):1218–1231. MR4011774

[24] Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2):500–531. MR1790007

[25] Ha, M. J., Banerjee, S., Akbani, R., Liang, H., Mills, G. B., Do, K.-A., and Baladandayuthapani, V. (2018). Personalized integrated network modeling of the cancer proteome atlas. *Scientific Reports*, 8(1):1–14.

[26] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761):C47–C52.

[27] He, X. and Zhang, J. (2006). Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2(6):e88.

[28] Huynh-Thu, V. A. and Sanguinetti, G. (2019). Gene regulatory network inference: an introductory survey. In *Gene Regulatory Networks*, pages 1–23. Springer.

[29] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

[30] Kuismin, M. O., Kemppainen, J. T., and Sillanpää, M. J. (2017). Precision matrix estimation with ROPE. *Journal of Computational and Graphical Statistics*, 26(3):682–694. MR3698677

[31] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254. MR2572459

[32] Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991

[33] Leclerc, R. D. (2008). Survival of the sparsest: robust gene networks are parsimonious. *Molecular Systems Biology*, 4(1):213.

[34] Lee, K., Jo, S., and Lee, J. (2022). The beta-mixture shrinkage prior for sparse covariances with near-minimax posterior convergence rate. *Journal*

*of Multivariate Analysis*, 192:105067. MR4450889

[35] Lee, K. and Lee, J. (2021). Estimating large precision matrices via modified Cholesky decomposition. *Statistica Sinica*, 31(2021):173–196. MR4270383

[36] Li, Y., Craig, B. A., and Bhadra, A. (2019). The graphical horseshoe estimator for inverse covariance matrices. *Journal of Computational and Graphical Statistics*, 28(3):747–757. MR4007755

[37] Liu, C. and Martin, R. (2019). An empirical $g$-Wishart prior for sparse high-dimensional Gaussian graphical models. *arXiv preprint arXiv:1912.03807*. MR4138128

[38] Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

[39] Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278. MR1243503

[40] Mohammadi, A. and Wit, E. C. (2015a). Bayesian structure learning in sparse gaussian graphical models. *Bayesian Analysis*, 10(1):109–138. MR3420899

[41] Mohammadi, R. and Wit, E. C. (2015b). Bdgraph: An R package for bayesian structure learning in graphical models. *arXiv preprint arXiv:1501.05108*.

[42] Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686. MR2524001

[43] Piironen, J. and Vehtari, A. (2017). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, 11(2):5018–5051. MR3738204

[44] Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, 26(3):369–387. MR2917961

[45] Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555.

[46] Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515. MR2417391

[47] Roverato, A. (2000). Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1):99–112. MR1766831

[48] Roverato, A. (2002). Hyper inverse wishart distribution for non-decomposable graphs and its application to bayesian inference for gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411. MR1925566

[49] Ryali, S., Chen, T., Supekar, K., and Menon, V. (2012). Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. *NeuroImage*, 59(4):3852–3861.

[50] Song, Q. and Liang, F. (2017). Nearly optimal Bayesian shrinkage for high dimensional regression. *arXiv preprint arXiv:1712.08964*. MR4535982

[51] Tang, X., Xu, X., Ghosh, M., and Ghosh, P. (2018). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya*

*A*, 80(2):215–246. MR3850065

[52] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58:267–288. MR1379242

[53] van den Boom, W., Beskos, A., and De Iorio, M. (2022). The g-wishart weighted proposal algorithm: Efficient posterior computation for gaussian graphical models. *Journal of Computational and Graphical Statistics*, 31(4):1215–1224. MR4513382

[54] van der Pas, S., Szabó, B., and van der Vaart, A. (2017). Uncertainty quantification for the horseshoe (with discussion). *Bayesian Analysis*, 12(4):1221–1274. MR3724985

[55] Van Wieringen, W. N. and Peeters, C. F. (2016). Ridge estimation of inverse covariance matrices from high-dimensional data. *Computational Statistics & Data Analysis*, 103:284–303. MR3522633

[56] Wang, C., Pan, G., Tong, T., and Zhu, L. (2015). Shrinkage estimation of large dimensional precision matrix using random matrix theory. *Statistica Sinica*, 25:993–1008. MR3409734

[57] Wang, H. (2012). Bayesian graphical lasso models and efficient posterior computation. *Bayesian Analysis*, 7(4):867–886. MR3000017

[58] Wang, H. (2014). Coordinate descent algorithm for covariance graphical lasso. *Statistics and Computing*, 24(4):521–529. MR3223538

[59] Wang, H. (2015). Scaling it up: Stochastic search structure learning in graphical models. *Bayesian Analysis*, 10(2):351–377. MR3420886

[60] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J. M. (2013). The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120.

[61] Xiang, R., Khare, K., and Ghosh, M. (2015). High dimensional posterior convergence rates for decomposable graphical models. *Electronic Journal of Statistics*, 9(2):2828–2854. MR3439186

[62] Xie, X., Kou, S. C., and Brown, L. (2016). Optimal shrinkage estimation of mean parameters in family of distributions with quadratic variance. *The Annals of Statistics*, 44(2):564–597. MR3476610

[63] Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2012). Graphical models via generalized linear models. In *NIPS*, volume 25, pages 1367–1375.

[64] Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via lasso penalized d-trace loss. *Biometrika*, 101(1):103–120. MR3180660

[65] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320. MR2137327

[66] Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36(4):1509–1533. MR2435443