Journal of Multivariate Analysis 136 (2015) 147–162

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Bayesian structure learning in graphical models

Sayantan Banerjee [a,*], Subhashis Ghosal [b]

[a] The University of Texas MD Anderson Cancer Center, United States

[b] North Carolina State University, United States

ARTICLE INFO

Article history:
Received 7 April 2014
Available online 23 January 2015

AMS subject classifications:
primary 62H12
secondary 62F12
62F15

Keywords:
Graphical lasso
Graphical models
Laplace approximation
Posterior convergence
Precision matrix

ABSTRACT

We consider the problem of estimating a sparse precision matrix of a multivariate Gaussian distribution, where the dimension p may be large. Gaussian graphical models provide an important tool in describing conditional independence through presence or absence of edges in the underlying graph. A popular non-Bayesian method of estimating a graphical structure is given by the graphical lasso. In this paper, we consider a Bayesian approach to the problem. We use priors which put a mixture of a point mass at zero and certain absolutely continuous distribution on off-diagonal elements of the precision matrix. Hence the resulting posterior distribution can be used for graphical structure learning. The posterior convergence rate of the precision matrix is obtained and is shown to match the oracle rate. The posterior distribution on the model space is extremely cumbersome to compute using the commonly used reversible jump Markov chain Monte Carlo methods. However, the posterior mode in each graph can be easily identified as the graphical lasso restricted to each model. We propose a fast computational method for approximating the posterior probabilities of various graphs using the Laplace approximation approach by expanding the posterior density around the posterior mode. We also provide estimates of the accuracy in the approximation.

1. Introduction

Statistical inference on a large covariance or precision matrix (inverse of covariance matrix) is a topic of growing interest in recent times. Often the dimension p grows with the sample size n and even p can exceed n. Data of this type are frequently encountered in fMRI, spectroscopy, gene array expressions and so on. Estimation of a covariance or precision matrix is of special interest because of its importance in methods like principal component analysis (PCA), linear discriminant analysis (LDA), etc. In cases where $p > n$, the sample covariance matrix is necessarily singular, and hence an estimator of the precision matrix cannot be obtained by inverting it. Therefore we need to resort to other techniques for handling the high-dimensional problems.

Regularization methods for estimation of the covariance or precision matrix have been proposed and studied in recent literature for high-dimensional problems. These include banding, thresholding, tapering and penalization based methods; for example, see Ledoit and Wolf [22], Huang et al. [18], Yuan and Lin [35], Bickel and Levina [4,5], Karoui [19], Friedman et al. [13], Rothman et al. [29], Lam and Fan [20], Rothman et al. [30], Cai et al. [8,7]; see also Banerjee and Ghosal [3] for a Bayesian method based on banding. The primary goal of these regularization based methods is to impose a sparsity structure in the

* Correspondence to: Department of Biostatistics, The University of Texas MD Anderson Cancer Center, 1400 Pressler Street, Houston, TX 77030, United States.

E-mail address: SBanerjee@mdanderson.org (S. Banerjee).

http://dx.doi.org/10.1016/j.jmva.2015.01.015
0047-259X/

matrix. Most of these methods are applicable to situations where there is a natural ordering in the underlying variables, for example in data from time series, spatial data, etc., so that variables which are far off from each other have smaller correlations or partial correlations. In high-dimensional situations for data arising from genetics or econometrics, a natural ordering of the underlying variables may not always be readily available and hence estimation methods which are invariant to the ordering of the variables are desirable.

For estimation of a sparse inverse covariance matrix, graphical models [21] provide an excellent tool, as the conditional dependence between the component variables is captured by an undirected graph; see Dobra et al. [12], Meinshausen and Bühlmann [25], Yuan and Lin [35], Friedman et al. [13]. There are several methods in the frequentist literature for the estimation of the precision matrix through graphical models. These methods include minimization of the penalized log-likelihood of the data with a lasso type penalty on the elements of the precision matrix. Several algorithms have been developed in the literature to solve the above optimization problem, including coordinate descent based algorithm for the lasso, which is popularly known as the graphical lasso [25,13,2,35,16,33]. Other methods include the Sparse Permutation Invariant Covariance Estimator (SPICE) [29].

Frequentist behavior of Bayesian methods in the context of high dimensional covariance matrix estimation have been studied only by a few authors. Ghosal [14] studied asymptotic normality of posterior distributions for exponential families, which include the normal model with unknown covariance matrix, when the dimension $p \to \infty$, but restricting to $p \ll n$. Recently, Pati et al. [27] considered sparse Bayesian factor models for dimensionality reduction in high dimensional problems and showed consistency in the L_2-operator norm (also known as the spectral norm) by using a point mass mixture prior on the factor loadings, assuming such a factor model representation for the true covariance matrix.

Bayesian methods for inference using graphical models have also been developed, as in Roverato [31], Atay-Kayis and Massam [1], Letac and Massam [23]. A conjugate family of priors, known as the G-Wishart prior [31] have been developed for incomplete decomposable graphs. The equivalent prior on the covariance matrix is termed as the hyper inverse Wishart distribution in Dawid and Lauritzen [11]. Letac and Massam [23] introduced a more general family of conjugate priors for the precision matrix, known as the W_{P_G}-Wishart family of distributions, which also has the conjugacy property. The properties of this family of distributions, including expressions for the Bayes estimators were further explored in Rajaratnam et al. [28]. Recently Banerjee and Ghosal [3] studied posterior convergence rates for a G-Wishart prior inducing a banding structure, where the true precision matrix need not have the banding structure.

Wang [32] developed a Bayesian version of the graphical lasso, by putting Laplace priors on the off-diagonal elements of the precision matrix and exponential priors on the diagonals. Similar in lines with the Bayesian lasso [26], the posterior mode in this case coincides with the graphical lasso estimate. A block Gibbs sampler is also developed for sampling from the resulting posterior. However, the Bayesian graphical lasso does not introduce any sparsity in the graphical structure because of the absence of a point mass at zero in the prior distribution for the off-diagonal elements. On the other hand, if point masses are introduced, the resulting posterior distribution on the structure of the graph becomes extremely difficult to compute based on the traditional reversible jump Markov chain Monte Carlo method.

In this paper, we derive posterior convergence rates for the Bayesian graphical lasso prior in terms of the Frobenius norm under appropriate sparsity conditions when the dimension p grows with the sample size n. For computing the posterior distribution, we propose a Laplace approximation based method to compute the posterior probability of different graphical structures. Such Laplace approximations based methods have been developed for variable selection in regression models; for example, see Yuan and Lin [34], Curtis et al. [10]. The lasso type penalty on the elements lead to non-differentiability of the integrand, when the graphical lasso sets an off-diagonal entry to zero, but the model includes that off-diagonal entry as a free variable. We shall call such models non-regular following the terminology used by Yuan and Lin [34] for variable selection in linear regression models. We show that the posterior probability of non-regular models are substantially smaller than their regular counterparts and hence in comparison may be ignored from consideration. We also estimate the error in the Laplace approximation for regular models.

The paper is organized as follows. In the next section, we introduce notations and discuss preliminaries on graphical models required for the other sections of the paper. In Section 3, we state model assumptions and specify the prior distribution on the underlying parameters, derive the form of the posterior and obtain the posterior convergence rate using the general theory developed in Ghosal et al. [15]. In Section 4, we develop the approximation of the posterior probabilities for different graphical models and discuss the issue of non-regular graphical models. We also show that the error in approximation of the posterior probabilities using the Laplace approximation is asymptotically negligible under appropriate conditions. A simulation study is performed in the Section 5 followed by a real data example in Section 6. Proofs of main results and additional lemmas are included in the Appendix.

2. Notations and preliminaries

An undirected graph G comprises of a non-empty set V of p vertices indexing the components of a p-dimensional random vector along with an edge set $E \subset \{(i, j) \in V \times V : i < j\}$. Let $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ be distributed as $\mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Omega}^{-1})$, where the precision matrix $\boldsymbol{\Omega} = ((\omega_{ij}))$ is such that $(i, j) \notin E$ implies $\omega_{ij} = 0$. We then say that $\boldsymbol{X}$ follows a Gaussian graphical model (GGM) with respect to the graph G. Since the absence of an edge between i and j implies conditional independence of X_i and X_j given $(X_r : r \neq i, j)$, a GGM serves as an excellent tool in representing the sparsity structure in the precision matrix. Following the notation in Letac and Massam [23], the canonical parameter $\boldsymbol{\Omega}$ is restricted to $\mathcal{P}_G$, where $\mathcal{P}_G$ is the cone of

positive definite symmetric matrices of order p having zero entry corresponding to each missing edge in E. We also denote the linear space of symmetric matrices of order p by $\mathcal{M}$, and $\mathcal{M}^+ \subset \mathcal{M}$ to be the cone of positive definite matrices of order p. Corresponding to each GGM $G = (V, E)$, we define the set $\bar{E} = \{(i, j) \in V \times V : i = j, \text{ or } (i, j) \in E\}$.

By $t_n = O(\delta_n)$ (respectively, $o(\delta_n)$), we mean that t_n/δ_n is bounded (respectively, $t_n/\delta_n \to 0$ as $n \to \infty$). For a random sequence X_n, $X_n = O_P(\delta_n)$ (respectively, $X_n = o_P(\delta_n)$) means that $P(|X_n| \le M\delta_n) \to 1$ for some constant M (respectively, $P(|X_n| < \epsilon\delta_n) \to 1$ for all $\epsilon > 0$). For numerical sequences r_n and s_n, by $r_n \ll s_n$ (or, $s_n \gg r_n$) we mean that $r_n = o(s_n)$, while by $r_n \lesssim s_n$ (or $s_n \gtrsim r_n$) we mean that $r_n = O(s_n)$. By $r_n \asymp s_n$ we mean that both $r_n \lesssim s_n$ and $s_n \lesssim r_n$ hold, while $r_n \sim s_n$ stands for $r_n/s_n \to 1$. The indicator function is denoted by $\mathbb{1}$. Non-stochastic vectors are represented in bold lowercase English or Greek letters with the components of a vector by the corresponding non-bold letters, that is, for $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. For a vector $\boldsymbol{x} \in \mathbb{R}^p$, we define the following vector norms: $\|\boldsymbol{x}\|_r = \left(\sum_{j=1}^p |x_j|^r\right)^{1/r}$, $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. Matrices are denoted in bold uppercase English or Greek letters, like $\boldsymbol{A} = ((a_{ij}))$, where a_{ij} stands for the (i, j)th entry of $\boldsymbol{A}$. The identity matrix of order p will be denoted by $\boldsymbol{I}_p$. If $\boldsymbol{A}$ is a symmetric $p \times p$ matrix, let $\text{eig}_1(\boldsymbol{A}) \le \cdots \le \text{eig}_p(\boldsymbol{A})$ stand for its eigenvalues and let the trace of $\boldsymbol{A}$ be denoted by $\text{tr}(\boldsymbol{A})$. Viewing $\boldsymbol{A}$ as a vector in $\mathbb{R}^{p^2}$, we define L_r, $1 \le r < \infty$ and L_∞-norms on $p \times p$ matrices as

$$\|\boldsymbol{A}\|_r = \left(\sum_{i=1}^p \sum_{j=1}^p |a_{ij}|^r\right)^{1/r}, \quad 1 \le r < \infty, \ \|\boldsymbol{A}\|_\infty = \max_{i,j} |a_{ij}|.$$

Note that $\|\boldsymbol{A}\|_2 = \sqrt{\text{tr}(\boldsymbol{A}^T\boldsymbol{A})}$, the Frobenius norm. Viewing $\boldsymbol{A}$ an operator from $(\mathbb{R}^p, \|\cdot\|_r)$ to $(\mathbb{R}^p, \|\cdot\|_s)$, where $1 \le r, s \le \infty$, we can also define, $\|\boldsymbol{A}\|_{(r,s)} = \sup(\|\boldsymbol{A}\boldsymbol{x}\|_s : \|\boldsymbol{x}\|_r = 1)$. We refer to the norm $\|\cdot\|_{(r,r)}$ as the L_r-operator norm. This gives the L_2-operator norm of $\boldsymbol{A}$ as

$$\|\boldsymbol{A}\|_{(2,2)} = [\max\{\text{eig}_i(\boldsymbol{A}^T\boldsymbol{A}) : 1 \le i \le p\}]^{1/2}.$$

For symmetric matrices, $\|\boldsymbol{A}\|_{(2,2)} = \max\{|\text{eig}_i(\boldsymbol{A})| : 1 \le i \le p\}$. For symmetric matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of order p, we have the following:

$$\begin{aligned} &\|\boldsymbol{A}\|_\infty \le \|\boldsymbol{A}\|_{(2,2)} \le \|\boldsymbol{A}\|_2 \le p\|\boldsymbol{A}\|_\infty, \\ &\|\boldsymbol{A}\boldsymbol{B}\|_2 \le \|\boldsymbol{A}\|_{(2,2)}\|\boldsymbol{B}\|_2, \qquad \|\boldsymbol{A}\boldsymbol{B}\|_2 \le \|\boldsymbol{A}\|_2\|\boldsymbol{B}\|_{(2,2)}. \end{aligned} \tag{2.1}$$

Let $\boldsymbol{A}^{1/2}$ stand for the unique positive definite square root of a positive definite matrix $\boldsymbol{A}$. We say that $\boldsymbol{A} > \boldsymbol{0}$ if $\boldsymbol{A}$ is positive definite, where $\boldsymbol{0}$ stands for the zero matrix. We denote sets in non-bold uppercase English letters. The cardinality of a set T, that is, the number of elements in T is denoted by $\#T$. We define the symmetric matrix $\boldsymbol{E}_{(i,j)} = ((\mathbb{1}_{\{(i,j),(j,i)\}}(l, m)))$.

The Hellinger distance between two probability densities q_1 and q_2 is given by $h(q_1, q_2) = \|\sqrt{q_1} - \sqrt{q_2}\|_2$. For a subset A of a metric space (S, d), $N(\epsilon, A, d)$ denotes the ϵ-covering number of A with respect to d, that is, the minimum number of d-balls of size ϵ in S needed to cover A.

3. Model, prior and posterior concentration

Consider n independent random samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is nonsingular and the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is sparse. The problem is to estimate $\boldsymbol{\Omega}$ and to learn the underlying graphical structure. Let $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i\boldsymbol{X}_i^T$, the natural estimator of $\boldsymbol{\Sigma}$.

The graphical lasso produces sparse solutions for the precision matrix, as the lasso does for linear regression. The graphical lasso estimator minimizes two times the penalized average negative log-likelihood

$$-\log\det(\boldsymbol{\Omega}) + \text{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) + \frac{\lambda}{n}\|\boldsymbol{\Omega}\|_1, \tag{3.1}$$

over the class of positive definite matrices, and $\lambda \ge 0$ acts as the penalty parameter. Rothman et al. [29] derived frequentist convergence rates of the penalized estimator under some sparsity assumptions on the true precision matrix. More specifically, consider the following class of positive definite matrices of order p:

$$\mathcal{U}(\varepsilon_0, s) = \left\{\boldsymbol{\Omega} : \#\{(i, j) : 1 \le i < j \le p, \ \omega_{ij} \ne 0\} \le s, 0 < \varepsilon_0 \le \text{eig}_1(\boldsymbol{\Omega}) \le \text{eig}_p(\boldsymbol{\Omega}) \le \varepsilon_0^{-1} < \infty\right\}. \tag{3.2}$$

Though Rothman et al. [29] considered penalizing only the off-diagonal elements of $\boldsymbol{\Omega}$, a minor modification of their proof leads to the same convergence rate for the graphical lasso estimator, obtained by additionally penalizing the diagonal elements. Let us denote $\boldsymbol{\Omega}^*$ as the graphical lasso obtained by minimizing (3.1) based on a sample of size n from a p-dimensional Gaussian distribution with precision matrix $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$. Then, it follows from Theorem 1 in Rothman et al. [29] that the rate of convergence of $\boldsymbol{\Omega}^*$ is $n^{-1/2}(p+s)^{1/2}(\log p)^{1/2}$ in the Frobenius norm. In particular, this implies that $\|\boldsymbol{\Omega}^* - \boldsymbol{\Omega}_0\|_2$ tends to zero in probability whenever $n^{-1}(p + s)\log p \to 0$. Under this condition, we have that

$$\|\boldsymbol{\Omega}^*\|_{(2,2)} = O_P(1), \qquad \|\boldsymbol{\Omega}^{*-1}\|_{(2,2)} = O_P(1). \tag{3.3}$$

The first relation follows by the triangle inequality $\|\mathbf{\Omega}^*\|_{(2,2)} \leq \|\mathbf{\Omega}_0\|_{(2,2)} + \|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}$ and the norm inequality $\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)} \leq \|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_2$, while for the second relation observe that

$$\begin{aligned}\|\mathbf{\Omega}^{*-1}\|_{(2,2)} &\leq \|\mathbf{\Omega}_0^{-1}\|_{(2,2)} + \|\mathbf{\Omega}^{*-1} - \mathbf{\Omega}_0^{-1}\|_{(2,2)} \\ &\leq \|\mathbf{\Omega}_0^{-1}\|_{(2,2)} + \|\mathbf{\Omega}_0^{-1}\|_{(2,2)}\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}\|\mathbf{\Omega}^{*-1}\|_{(2,2)},\end{aligned}$$

which leads to

$$\|\mathbf{\Omega}^{*-1}\|_{(2,2)} \leq \frac{\|\mathbf{\Omega}_0^{-1}\|_{(2,2)}}{1 - \|\mathbf{\Omega}_0^{-1}\|_{(2,2)}\|\mathbf{\Omega}^* - \mathbf{\Omega}_0\|_{(2,2)}}. \tag{3.4}$$

In the Bayesian context, Wang [32] introduced the graphical lasso prior, which uses exponential distributions on diagonal elements and Laplace density $\lambda e^{-\lambda|x|}/2$ on off-diagonal elements, all independently of each other, and finally imposes a positive definiteness constraint. The graphical lasso prior has a drawback that it puts absolutely continuous priors on the elements of the precision matrix, and hence the posterior probabilities of the event $\{\omega_{ij} = 0\}$ is always exactly zero.

Wang [32] also mentioned an extension of the graphical lasso by putting an additional level of prior on the underlying graphical model structure using point mass priors on the events corresponding to the absence of an edge in the edge-set E, but did not pursue the method. We put point-mass prior on the events $\{\omega_{ij} = 0\}$ to make posterior inference about the sparse structure of the underlying graphical model. Define $\mathbf{\Gamma} = (\gamma_{ij} : 1 \leq i < j \leq p)$ to be a $\binom{p}{2}$-dimensional vector of edge-inclusion indicator, that is,

$$\gamma_{ij} = \mathbb{1}\{(i,j) \in E\}, \quad 1 \leq i < j \leq p. \tag{3.5}$$

Identifying $\mathbf{\Gamma}$ with the set of indices $\{(i,j) : \gamma_{ij} = 1\}$, we denote by $\#\mathbf{\Gamma}$ the number of non-zero elements in $\mathbf{\Gamma}$. Similar to the Bayesian graphical lasso prior, given the underlying graphical structure, we put a Laplace prior on the non-zero off-diagonal elements of the precision matrix and for the diagonal elements we have an exponential prior, overall maintaining the positive definiteness of the parameter. In order to establish convergence rates, we in fact need to keep $\mathbf{\Omega}$ and its inverse away from singular matrices by imposing a restriction on their eigenvalues. Let $\mathcal{M}_0^+$ be a subset of $\mathcal{M}^+$ whose elements have eigenvalues bounded between two fixed positive numbers. Then the joint prior density on $\mathbf{\Omega}$ is given by,

$$p(\mathbf{\Omega}|\mathbf{\Gamma}) \propto \prod_{\gamma_{ij}=1} \exp(-\lambda|\omega_{ij}|) \prod_{i=1}^{p} \exp\left(-\lambda\omega_{ii}/2\right) \mathbb{1}_{\mathcal{M}_0^+}(\mathbf{\Omega}). \tag{3.6}$$

We propose two different priors on the graphical structure indicator $\mathbf{\Gamma}$. The edge indicators γ_{ij}, $1 \leq i < j \leq p$, are considered to be independent and identically distributed (i.i.d) Bernoulli(q) random variables, but then conditioned to the restriction that the model size $\sum_{1\leq i<j\leq p} \gamma_{ij}$ does not exceed $\bar{R}$. For some $a_1, a_2 > 0$, the prior distribution on $\bar{R}$ is assumed to satisfy

$$\mathrm{P}(\bar{R} > a_1 m) \leq e^{-a_2 m \log m}, \quad m = 1, 2, \ldots. \tag{3.7}$$

This prior is similar to that used by Castillo and van der Vaart [9], which chooses the model size first according to a distribution with a similar tail decay and then subsets are selected randomly with equal probability. We can also specify the individual priors on γ_{ij} the same as above, but now truncating the model size to some fixed $\bar{r}$, where $\bar{r}$ may depend on n and is chosen to satisfy a metric entropy condition required for posterior convergence.

Thus, in the first situation, the prior on the graphical structure indicator $\mathbf{\Gamma}$, given $\bar{R}$, is given by,

$$p(\mathbf{\Gamma} \mid \bar{R}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2} - \#\mathbf{\Gamma}} \mathbb{1}(\#\mathbf{\Gamma} \leq \bar{R}), \tag{3.8}$$

leading to

$$p(\mathbf{\Gamma}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2} - \#\mathbf{\Gamma}} \mathrm{P}(\bar{R} \geq \#\mathbf{\Gamma}). \tag{3.9}$$

In the second case, the prior on $\mathbf{\Gamma}$ is simply given by

$$p(\mathbf{\Gamma}) \propto q^{\#\mathbf{\Gamma}}(1-q)^{\binom{p}{2} - \#\mathbf{\Gamma}} \mathbb{1}(\#\mathbf{\Gamma} \leq \bar{r}). \tag{3.10}$$

Smaller values of q prefer graphical models with fewer number of edges, and hence induce more sparsity in the precision matrix.

Due to the positive definiteness constraint on the parameter $\mathbf{\Omega}$, the normalizing constant corresponding to the posterior distribution of the graphical model is intractable. One possible solution is to employ a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm, which jumps between models of varying dimensions to evaluate the posterior probabilities. As there are as many as $2^{\binom{p}{2}}$ possible models, the posterior model probabilities estimated by RJMCMC visits are extremely unreliable. We consider a radically different approach to posterior computation based on Laplace approximations, elaborated in the next section.

Under the above prior specifications, the joint posterior distribution of $\boldsymbol{\Omega}$ and $\boldsymbol{\Gamma}$ given the data $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ is given by

$$\begin{aligned} p(\boldsymbol{\Omega}, \boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}) &\propto p(\boldsymbol{X}^{(n)}|\boldsymbol{\Omega}, \boldsymbol{\Gamma}) p(\boldsymbol{\Omega}|\boldsymbol{\Gamma}) p(\boldsymbol{\Gamma}) \\ &= (2\pi)^{-np/2} \{\det(\boldsymbol{\Omega})\}^{n/2} \exp\{-n \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})/2\} \\ &\quad \times \prod_{\gamma_{ij}=1} \left\{\lambda \exp(-\lambda|\omega_{ij}|)/2\right\} \prod_{i=1}^{p} \{\lambda \exp(-\lambda\omega_{ii}/2)/2\} \times p(\boldsymbol{\Gamma}) \mathbb{1}_{\mathcal{M}_0^+}(\boldsymbol{\Omega}). \end{aligned} \tag{3.11}$$

Thus,

$$p(\boldsymbol{\Omega}, \boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}) \propto C_{\boldsymbol{\Gamma}} Q(\boldsymbol{\Omega}, \boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}), \tag{3.12}$$

where

$$C_{\boldsymbol{\Gamma}} = (2\pi)^{-np/2} q^{\#\boldsymbol{\Gamma}} (1-q)^{\binom{p}{2} - \#\boldsymbol{\Gamma}} (\lambda/2)^{p+\#\boldsymbol{\Gamma}} \beta(\boldsymbol{\Gamma}),$$

$$\beta(\boldsymbol{\Gamma}) = \begin{cases} \mathrm{P}(\bar{R} \geq \#\boldsymbol{\Gamma}), & \text{for prior as in (3.9)}, \\ \mathbb{1}\{\#\boldsymbol{\Gamma} \leq \bar{r}\}, & \text{for prior as in (3.10)}, \end{cases}$$

$$Q(\boldsymbol{\Omega}, \boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}) = \{\det(\boldsymbol{\Omega})\}^{n/2} \exp\{-n \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega})/2\} \prod_{\gamma_{ij}=1} \exp(-\lambda|\omega_{ij}|) \times \prod_{i=1}^{p} \exp(-\lambda\omega_{ii}/2)\, \mathbb{1}_{\mathcal{M}_0^+}(\boldsymbol{\Omega}). \tag{3.13}$$

The following result gives posterior convergence rate as $n \to \infty$. We assume that the true model is sparse, as described by (3.2).

Theorem 3.1. *Let $\boldsymbol{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ be a random sample from a p-dimensional Gaussian distribution with mean $\boldsymbol{0}$ and precision matrix $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$ for some $0 < \varepsilon_0 < \infty$ and $0 \leq s \leq p(p-1)/2$. Also assume that the prior distributions $p(\boldsymbol{\Omega} \mid \boldsymbol{\Gamma})$ and $p(\boldsymbol{\Gamma})$ as in (3.6) and (3.9) or (3.10) with $q < 1/2$ and the range of eigenvalues of matrices in $\mathcal{M}_0^+$ is sufficiently broad to contain $[\varepsilon_0, \varepsilon_0^{-1}]$. Then the posterior distribution of $\boldsymbol{\Omega}$ satisfies*

$$\mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 > M\epsilon_n \mid \boldsymbol{X}^{(n)}\right\}\right] \to 0, \tag{3.14}$$

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

The proof uses the general theory of posterior convergence of Ghosal et al. [15] and will be given in the Appendix. The above posterior convergence rate matches exactly with the convergence rate of the graphical lasso obtained in Rothman et al. [29].

Note that, Theorem 3.1 gives $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2 = O(\epsilon_n)$ with posterior probability tending to one in probability and from Rothman et al. [29] it follows that, $\|\boldsymbol{\Omega}^* - \boldsymbol{\Omega}_0\|_2 = O_P(\epsilon_n)$, where $\boldsymbol{\Omega}^*$ is the graphical lasso. Hence, by the triangle inequality, $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}^*\|_2 = O(\epsilon_n)$ with posterior probability tending to one in probability. This gives,

$$\frac{\int_{\|\boldsymbol{\Omega}-\boldsymbol{\Omega}^*\|_2 \leq \epsilon_n} f(\boldsymbol{\Omega}) \prod_{(i,j)\in\bar{E}} d\omega_{ij}}{\int_{\boldsymbol{\Omega}\in\mathcal{M}_0^+} f(\boldsymbol{\Omega}) \prod_{(i,j)\in\bar{E}} d\omega_{ij}} \to 1, \tag{3.15}$$

where $f(\boldsymbol{\Omega})$ is a bounded and positive measurable function of $\boldsymbol{\Omega}$. If the model is restricted to $\boldsymbol{\Gamma}$, then the posterior and the graphical lasso will concentrate around the projection of the true precision matrix on the model at the rate ϵ_n, so that the posterior probability of an ϵ_n-Frobenius neighborhood around the graphical lasso in model $\boldsymbol{\Gamma}$ will go to one.

4. Posterior computation

Let $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}} = ((\omega_{\boldsymbol{\Gamma},ij}))$ denote the precision matrix in model $\boldsymbol{\Gamma}$. The marginal posterior density of the graphical structure indicator $\boldsymbol{\Gamma}$ can be obtained by integrating out elements of the precision matrix in the joint posterior density in (3.11), to get

$$p(\boldsymbol{\Gamma}|\boldsymbol{X}^{(n)}) \propto C_{\boldsymbol{\Gamma}} \int_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}\in\mathcal{M}_0^+} \exp\{-n\, h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}})/2\} \prod_{(i,j)\in\bar{E}_{\boldsymbol{\Gamma}}} d\omega_{\boldsymbol{\Gamma},ij}, \tag{4.1}$$

where

$$h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}) = -\log\det(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}) + \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} |\omega_{\boldsymbol{\Gamma},ij}| + \frac{\lambda}{n} \sum_{i=1}^{p} \omega_{\boldsymbol{\Gamma},ii}, \tag{4.2}$$

and $\bar{E}_{\boldsymbol{\Gamma}} = \{(i,j) \in V \times V : i = j, \text{ or } \gamma_{ij} = 1 \text{ for } i \neq j\}$. In other words, $\bar{E}_{\boldsymbol{\Gamma}}$ refers to the indices of the diagonal elements and the non-zero off-diagonal elements corresponding to the edges in the graphical model defined by the edge-inclusion indicator vector $\boldsymbol{\Gamma}$, to be referred to as model $\boldsymbol{\Gamma}$ below.

Note that $h_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})$ is minimized at $\mathbf{\Omega}_{\mathbf{\Gamma}} = \mathbf{\Omega}^*_{\mathbf{\Gamma}}$, the graphical lasso corresponding to the penalty parameter λ/n and model $\mathbf{\Gamma}$.

The marginal posterior of $\mathbf{\Gamma}$ is, however, intractable. We give an approximate method for the posterior probability computations of various models using Laplace approximation. The Laplace approximation requires expanding the integrand in (4.1) around the maximum, which in this case, coincides with the graphical lasso in model $\mathbf{\Gamma}$. Laplace approximation technique is well accepted in Bayesian computation, most notably in deriving the Bayesian Information Criterion (BIC) as an approximation to logarithm of posterior model probabilities.

4.1. Approximating model posterior probabilities

Define $\mathbf{\Delta}_{\mathbf{\Gamma}} = \mathbf{\Omega}_{\mathbf{\Gamma}} - \mathbf{\Omega}^*_{\mathbf{\Gamma}} = ((u_{\mathbf{\Gamma},ij}))$, where $\mathbf{\Omega}^*_{\mathbf{\Gamma}} = ((\omega^*_{\mathbf{\Gamma},ij}))$ is the graphical lasso solution corresponding to the underlying graphical model structure $\mathbf{\Gamma}$ and penalty parameter λ/n. Then,

$$p(\mathbf{\Gamma}|\boldsymbol{X}) \propto C_{\mathbf{\Gamma}} \exp\{-n\, h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})/2\} \left\{\det(\mathbf{\Omega}^*_{\mathbf{\Gamma}})\right\}^{-n/2} \times \int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}} \in \mathcal{M}_0^+} \exp\{-n\, g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})/2\} \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij}, \tag{4.3}$$

where $g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ is given by

$$-\log\det(\mathbf{\Delta}_{\mathbf{\Gamma}} + \mathbf{\Omega}^*_{\mathbf{\Gamma}}) + \mathrm{tr}(\widehat{\mathbf{\Sigma}}\mathbf{\Delta}_{\mathbf{\Gamma}}) + \frac{2\lambda}{n} \sum_{\gamma_{ij}=1} (|u_{\mathbf{\Gamma},ij} + \omega^*_{\mathbf{\Gamma},ij}| - |\omega^*_{\mathbf{\Gamma},ij}|) + \frac{\lambda}{n} \sum_{i=1}^{p} u_{\mathbf{\Gamma},ii}. \tag{4.4}$$

Clearly $g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ is minimized at $\mathbf{\Delta}_{\mathbf{\Gamma}} = \mathbf{0}$ by the definition of $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$, so the first derivative of $g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ vanishes at $\mathbf{0}$, provided that it is differentiable at $\mathbf{0}$. Define the matrix $\boldsymbol{H}_{\boldsymbol{B}} = ((h_{\boldsymbol{B}}\{(i,j),(l,m)\}))$, where

$$h_{\boldsymbol{B}}\{(i,j),(l,m)\} = \mathrm{tr}\left\{\boldsymbol{B}^{-1}\boldsymbol{E}_{(i,j)}\boldsymbol{B}^{-1}\boldsymbol{E}_{(l,m)}\right\}. \tag{4.5}$$

Using standard matrix calculus (for example, see Section 15.9 of Harville [17]), we can find that the Hessian of $g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ is the $\#\bar{E}_{\mathbf{\Gamma}} \times \#\bar{E}_{\mathbf{\Gamma}}$ matrix $\boldsymbol{H}_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$, whose $\{(i,j),(l,m)\}$th entry for $(i,j),(l,m) \in \bar{E}_{\mathbf{\Gamma}}$ is given by

$$\frac{\partial^2 g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})}{\partial u_{\mathbf{\Gamma},ij}\partial u_{\mathbf{\Gamma},lm}} = \mathrm{tr}\left\{(\mathbf{\Delta}_{\mathbf{\Gamma}} + \mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\boldsymbol{E}_{(i,j)}(\mathbf{\Delta}_{\mathbf{\Gamma}} + \mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\boldsymbol{E}_{(l,m)}\right\}. \tag{4.6}$$

Thus the Laplace approximation $p^*(\mathbf{\Gamma} \mid \boldsymbol{X}^{(n)})$ to the posterior probability $p(\mathbf{\Gamma} \mid \boldsymbol{X}^{(n)})$ is given by

$$\begin{aligned} p^*(\mathbf{\Gamma}|\boldsymbol{X}^{(n)}) &\propto C_{\mathbf{\Gamma}} \exp\{-n\, h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})/2\} \left\{\det(\mathbf{\Omega}^*_{\mathbf{\Gamma}})\right\}^{-n/2} \exp\{-n\, g(\mathbf{0})/2\}(2\pi)^{\#\bar{E}_{\mathbf{\Gamma}}/2}(n/2)^{-\#\bar{E}_{\mathbf{\Gamma}}/2} \left[\det\left\{\frac{\partial^2 g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})}{\partial \mathbf{\Delta}_{\mathbf{\Gamma}}\partial \mathbf{\Delta}_{\mathbf{\Gamma}}^T}\Big|_{\mathbf{0}}\right\}\right]^{-1/2} \\ &= C_{\mathbf{\Gamma}} \exp\{-n\, h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})/2\}(\pi/n)^{\#\bar{E}_{\mathbf{\Gamma}}/2}\{\det(\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}})\}^{-1/2}. \end{aligned} \tag{4.7}$$

The approximation in (4.7) is meaningful only if all entries of the graphical lasso in model $\mathbf{\Gamma}$ are non-zero; otherwise the derivative of $g_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ does not exist. A similar situation arises in the context of regression models; see Yuan and Lin [34] and Curtis et al. [10]. In the next section, we show that such "non-regular models" can essentially be ignored for the purpose of posterior probability evaluation. Also to satisfy the regularity conditions for Laplace approximation, the determinant of the Hessian above needs to be bounded away from zero. We have shown this in Lemma A.3 by arguing that the minimum eigenvalue of the Hessian is bounded away from zero. The partial derivatives as defined in Eq. (4.6) need to be bounded in a neighborhood of $\mathbf{\Delta}_{\mathbf{\Gamma}} = \mathbf{0}$, which follows from continuity, since $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$ is close to the projection of the true $\mathbf{\Omega}_0$ on $\mathbf{\Gamma}$ with high probability where entries of the inverse are bounded.

4.2. Ignorability of non-regular models

As discussed in the previous section, the objective function of the graphical lasso problem in model $\mathbf{\Gamma}$ is not differentiable if the graphical lasso has a zero off-diagonal entry, that is, $\omega^*_{ij} = 0$ for at least one $\gamma_{ij} = 1$. Let us assume, for notational simplicity, that the vector $\mathbf{\Gamma}$ has been arranged in such a way that the first t elements of $\mathbf{\Gamma}$ are 1 and the rest are 0. Also, among those t 1s, the last r of them have corresponding graphical lasso solution equal to zero. For such a non-regular model, we argue that the submodel $\mathbf{\Gamma}'$, with first $(t-r)$ 1s and rest 0s, provides the same graphical lasso solution for the non-zero elements as the bigger model $\mathbf{\Gamma}$. This means that for (i,j) such that $\gamma_{ij} = \gamma'_{ij} = 1$, the graphical lasso solution corresponding to $\mathbf{\Gamma}$, given by $\omega^*_{\mathbf{\Gamma},ij}$ is identical with that corresponding to $\mathbf{\Gamma}'$, given by $\omega^*_{\mathbf{\Gamma}',ij}$. We refer to such a submodel $\mathbf{\Gamma}'$ as the corresponding regular submodel of the non-regular model $\mathbf{\Gamma}$.

Lemma 4.1. *For the corresponding regular submodel $\mathbf{\Gamma}'$ of $\mathbf{\Gamma}$, the graphical lasso solution corresponding to models $\mathbf{\Gamma}$ and $\mathbf{\Gamma}'$ are identical.*

We give a proof of the above lemma in the Appendix. Recall that $\mathbf{\Omega}_{\mathbf{\Gamma}}$ denotes the precision matrix in model $\mathbf{\Gamma}$ and $\mathbf{\Delta}_{\mathbf{\Gamma}} = \mathbf{\Omega}_{\mathbf{\Gamma}} - \mathbf{\Omega}^*_{\mathbf{\Gamma}} = ((u_{\mathbf{\Gamma},ij}))$. Also, the graphical lasso in model $\mathbf{\Gamma}$ be denoted by $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$. Note that $\mathbf{\Omega}^*_{\mathbf{\Gamma}'} = \mathbf{\Omega}^*_{\mathbf{\Gamma}}$ by the definition of the regular submodel $\mathbf{\Gamma}'$. The ratio of the posterior model probabilities of the two model is given by,

$$\frac{p(\mathbf{\Gamma}|\mathbf{X}^{(n)})}{p(\mathbf{\Gamma}'|\mathbf{X}^{(n)})} = \frac{C_{\mathbf{\Gamma}} \int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}} \in \mathcal{M}_0^+} \exp\{-n\, h_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})/2\} \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij}}{C_{\mathbf{\Gamma}'} \int_{\mathbf{\Delta}_{\mathbf{\Gamma}'}+\mathbf{\Omega}^*_{\mathbf{\Gamma}'} \in \mathcal{M}_0^+} \exp\{-n\, h_{\mathbf{\Gamma}'}(\mathbf{\Delta}_{\mathbf{\Gamma}'})/2\} \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma}',ij}}. \tag{4.8}$$

The following result shows the ignorability of the non-regular models.

Theorem 4.2. *Consider the prior on $\mathbf{\Gamma}$ as given in (3.9) or (3.10) with $q < 1/2$. The posterior probability of a non-regular model $\mathbf{\Gamma}$ is always less than that of the corresponding regular submodel $\mathbf{\Gamma}'$.*

Proof. Using (3.15), we have,

$$\frac{p(\mathbf{\Gamma}|\mathbf{X}^{(n)})}{p(\mathbf{\Gamma}'|\mathbf{X}^{(n)})} = \frac{C_{\mathbf{\Gamma}} \int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 \le \epsilon_n} \exp\{-n\, h_{\mathbf{\Gamma}}(\mathbf{\Delta}_{\mathbf{\Gamma}})/2\} \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij} + o(1)}{C_{\mathbf{\Gamma}'} \int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}'}\|_2 \le \epsilon_n} \exp\{-n\, h_{\mathbf{\Gamma}'}(\mathbf{\Delta}_{\mathbf{\Gamma}'})/2\} \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma}',ij} + o(1)}.$$

Now, note that for (i, j) such that $\gamma_{ij} = \gamma'_{ij} = 1$, we have,

$$\{u_{\mathbf{\Gamma},ij} : \|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 \le \epsilon_n\} \subset \{u_{\mathbf{\Gamma}',ij} : \|\mathbf{\Delta}_{\mathbf{\Gamma}'}\|_2 \le \epsilon_n\}.$$

Hence, using Lemma A.2, we get

$$\begin{aligned}
\frac{p(\mathbf{\Gamma}|\mathbf{X}^{(n)})}{p(\mathbf{\Gamma}'|\mathbf{X}^{(n)})} &\le \frac{C_{\mathbf{\Gamma}}}{C_{\mathbf{\Gamma}'}} \int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 \le \epsilon_n} \exp\left(-\frac{n}{2}\frac{2\lambda}{n} \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\mathbf{\Gamma},ij}|\right) \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}} \cap \bar{E}^c_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma},ij} \\
&\le \frac{C_{\mathbf{\Gamma}}}{C_{\mathbf{\Gamma}'}} \int \exp\left(-\lambda \sum_{\gamma_{ij}=1,\gamma'_{ij}=0} |u_{\mathbf{\Gamma},ij}|\right) \prod_{(i,j)\in \bar{E}_{\mathbf{\Gamma}} \cap \bar{E}^c_{\mathbf{\Gamma}'}} du_{\mathbf{\Gamma},ij} \\
&= \frac{C_{\mathbf{\Gamma}}}{C_{\mathbf{\Gamma}'}} \left(\frac{2}{\lambda}\right)^{\#\mathbf{\Gamma} - \#\mathbf{\Gamma}'} \\
&= \left(\frac{q}{1-q}\right)^r \frac{\beta(\mathbf{\Gamma})}{\beta(\mathbf{\Gamma}')} \\
&\le \left(\frac{q}{1-q}\right)^r.
\end{aligned} \tag{4.9}$$

The last inequality follows from the fact that if the prior as in (3.9) is used, then $\mathrm{P}(\bar{R} \ge \#\mathbf{\Gamma}) \le \mathrm{P}(\bar{R} \ge \#\mathbf{\Gamma}')$ since $\#\mathbf{\Gamma} > \#\mathbf{\Gamma}'$. For the other prior as in (3.10), the inequality follows trivially as it involves the ratio of two indicator variables only.

For $q < 1/2$, the above ratio is less than 1. This completes the proof. □

The above result is particularly important in the sense that we can focus on regular models only, ignoring non-regular ones especially if q is chosen to be small. While approximating the posterior probabilities of the regular models, we re-normalize the values considering the regular models only.

4.3. Error in Laplace approximation

The approximation of the posterior probability of a model $\mathbf{\Gamma}$ is based on a Taylor series expansion of the function $h_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})$ defined in (4.2) around the graphical lasso in model $\mathbf{\Gamma}$. Let $\mathbf{\Delta}_{\mathbf{\Gamma}} = \mathbf{\Omega}_{\mathbf{\Gamma}} - \mathbf{\Omega}^*_{\mathbf{\Gamma}}$, and $\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ denote the vectorized version of $\mathbf{\Delta}_{\mathbf{\Gamma}}$, but excluding the entries set to zero by the model $\mathbf{\Gamma}$. Thus $\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})$ is a vector of dimension at most $p + \#\mathbf{\Gamma}$. The following result gives the bound on the remainder term of the Taylor series expansion under the above assumptions.

Lemma 4.3. *For any regular model $\mathbf{\Gamma}$, with probability tending to one, the remainder term in the expansion of the function $h_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})$, as defined in (4.2), around $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$, is bounded by $(p + \#\mathbf{\Gamma})\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2^2 \left(C_1\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2 + C_2\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2^2\right)/2$ for some positive constants C_1, C_2.*

This result can be used to find a bound for the error in Laplace approximation of the posterior probabilities of the graphical model structures. The following result gives the condition for which the error in approximation is asymptotically negligible.

Theorem 4.4. *With probability tending to one, the error in Laplace approximation of the posterior probability of any regular model $\boldsymbol{\Gamma}$ is asymptotically negligible if $(p + \#\boldsymbol{\Gamma})^2\epsilon_n \to 0$ in probability, where ϵ_n is the posterior convergence rate, that is, the error in the Laplace approximation tends to zero in probability if $n^{-1/2}(p + \#\boldsymbol{\Gamma})^{5/2}(\log p)^{1/2} \to 0$ in probability.*

The proof of the above result depends on several additional results, including Lemma 4.3 involving the bound on the remainder term in the Taylor series expansion of $h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}})$. We give a proof of the above result along with these additional results in the Appendix.

5. Simulation results

We perform a simulation study to assess the performance of the Bayesian method for graphical structure learning. We use 4 different models for our simulations, and we specify these models in terms of the elements of the covariance matrix $\boldsymbol{\Sigma} = ((\sigma_{ij}))$ or the precision matrix $\boldsymbol{\Omega} = ((\omega_{ij}))$, as follows:

1. Model 1: AR(1) model, $\sigma_{ij} = 0.7^{|i-j|}$.
2. Model 2: AR(2) model, $\omega_{ii} = 1$, $\omega_{i,i-1} = \omega_{i-1,i} = 0.5$, $\omega_{i,i-2} = \omega_{i-2,i} = 0.25$.
3. Model 3: Star model, where every node is connected to the first node, and $\omega_{ii} = 1$, $\omega_{1,i} = \omega_{i,1} = 0.1$, and $\omega_{ij} = 0$ otherwise.
4. Model 4: Circle model, $\omega_{ii} = 2$, $\omega_{i-1,i} = \omega_{i,i-1} = 1$, $\omega_{1,p} = \omega_{p,1} = 0.9$.

Corresponding to each model, we generate samples of size $n = 100, 200$ and dimension $p = 30, 50, 100$. The penalty parameter λ for the graphical lasso algorithm is chosen such that $\lambda/n = 0.5$ and the value of q appearing in the prior of the graphical structure indicator is taken to be 0.4. In computation, there is implicit restriction on the size by the size of the graphical lasso. For the theory, there is a restriction on the model size through the variable $\bar{R}$ which has sharp tails (that is, $\bar{R}$ is unlikely to be big). Thus in either way, a restriction on size is imposed meaning larger values of q are acceptable without problem (subject to $q < 1/2$ for ignorability of non-regular models). Thus if we need to fix a value of q, it makes sense to choose relatively big to avoid low sensitivity. Lower values will be justified if we have strong prior information. Choosing q from data by putting a prior and calculating its posterior is of course sensible but cannot be done in our setting, as we are not computing the full posterior.

We run 100 replications for each of the models and find the median probability model for each replication. To assess the performance of the median probability model (denoted by 'MPP'), we compute the specificity (SP), sensitivity (SE) and Matthews Correlation Coefficient (MCC) averaged across the replications as defined below and also compute the same for the graphical lasso (denoted by 'GL'). The results are presented in Table 1.

$$\text{SP} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \qquad \text{SE} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$
$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}, \tag{5.1}$$

where TP, TN, FP and FN respectively denote the true positives (edges included which are present in the true model), true negatives (edges excluded which are absent in the true model), false positives (edges included which are absent in the true model) and false negatives (edges excluded which are present in the true model) in the selected model. In this context we also define the False Positive Rate (FPR) as FPR = FP/(TN+FP), that is, FPR = 1−SP. We show the ROC curves corresponding to the various models with values of the penalty parameter ranging between 0.1 and 1 plotting Sensitivity (SE) against False Positive Rate (FPR). The curves are shown in Figs. 1 and 2.

In all the cases, the Bayesian method performs slightly better than the graphical lasso in terms of specificity, but suffers a bit in sensitivity. This is expected from a theoretical viewpoint as the approximate posterior probabilities are computed for the graphs which are sub-graphs of the graphical structure identified by the graphical lasso. The remaining graphs are not explored owing to non-regularity. Choosing $q < 1/2$ ensures that those structures can be ignored safely.

The sensitivity results for both the methods are not good for AR(2) and Star models. The ROC curves corresponding to these two models reveal that higher values of the penalty parameter result in better sensitivity, but at the cost of higher false positive rate. Overall, in terms of MCC, the median probability model performs better than the model selected by the graphical lasso.

6. Illustration with real data

In this section we illustrate the Bayesian graphical structure learning method with the stock price data from Yahoo! Finance. Description of the data set can be found in Liu et al. [24] and available in the `huge` package on `CRAN` [36] as `stockdata`. The data set consists of closing prices of stocks that were consistently included in the S&P 500 index in the time period January 1, 2003 to January 1, 2008 for a total of 1258 days. The stocks are also categorized into 10 Global Industry Classification Standard (GICS) sectors, namely, "Health Care", "Materials", "Industrials", "Consumer Staples", "Consumer Discretionary", "Utilities", "Information Technology", "Financials", "Energy", "Telecommunication Services".

Table 1
Simulation results for different structures of precision matrices. Figures in parentheses indicate standard errors.

Model	p	$n = 100$						$n = 200$					
		MPP			GL			MPP			GL		
		SP	SE	MCC	SP	SE	MCC	SP	SE	MCC	SP	SE	MCC
AR(1)	30	0.977 (0.003)	0.941 (0.019)	0.831 (0.015)	0.961 (0.003)	0.983 (0.010)	0.784 (0.013)	0.986 (0.002)	0.996 (0.003)	0.907 (0.014)	0.969 (0.002)	1.000 (0.000)	0.823 (0.013)
	50	0.987 (0.002)	0.953 (0.013)	0.841 (0.010)	0.977 (0.001)	0.986 (0.004)	0.785 (0.010)	0.991 (0.001)	0.992 (0.004)	0.903 (0.008)	0.980 (0.001)	1.000 (0.000)	0.823 (0.006)
	100	0.992 (0.001)	0.967 (0.008)	0.837 (0.007)	0.989 (0.001)	0.991 (0.003)	0.804 (0.006)	0.994 (0.001)	0.995 (0.002)	0.890 (0.008)	0.991 (0.001)	0.999 (0.001)	0.827 (0.006)
AR(2)	30	0.975 (0.003)	0.470 (0.014)	0.546 (0.013)	0.964 (0.002)	0.535 (0.013)	0.558 (0.012)	0.987 (0.002)	0.495 (0.008)	0.617 (0.008)	0.982 (0.002)	0.517 (0.009)	0.610 (0.007)
	50	0.983 (0.001)	0.462 (0.013)	0.541 (0.011)	0.971 (0.002)	0.508 (0.010)	0.522 (0.009)	0.993 (0.001)	0.489 (0.005)	0.629 (0.007)	0.987 (0.001)	0.534 (0.001)	0.622 (0.006)
	100	0.989 (0.001)	0.470 (0.006)	0.537 (0.006)	0.980 (0.001)	0.531 (0.007)	0.514 (0.007)	0.995 (0.001)	0.484 (0.006)	0.624 (0.004)	0.993 (0.001)	0.529 (0.009)	0.624 (0.005)
Star	30	0.947 (0.004)	0.289 (0.038)	0.228 (0.036)	0.937 (0.003)	0.310 (0.043)	0.224 (0.036)	0.995 (0.001)	0.210 (0.032)	0.378 (0.041)	0.993 (0.001)	0.252 (0.036)	0.402 (0.038)
	50	0.945 (0.003)	0.492 (0.034)	0.332 (0.025)	0.934 (0.003)	0.514 (0.035)	0.317 (0.023)	0.993 (0.000)	0.475 (0.034)	0.585 (0.024)	0.990 (0.001)	0.514 (0.032)	0.577 (0.022)
	100	0.939 (0.002)	1.000 (0.000)	0.485 (0.007)	0.927 (0.002)	1.000 (0.000)	0.452 (0.005)	0.988 (0.000)	1.000 (0.000)	0.792 (0.008)	0.984 (0.001)	1.000 (0.000)	0.748 (0.007)
Circle	30	0.733 (0.004)	1.000 (0.000)	0.399 (0.003)	0.694 (0.006)	1.000 (0.000)	0.369 (0.004)	0.719 (0.005)	1.000 (0.000)	0.388 (0.004)	0.674 (0.004)	1.000 (0.000)	0.354 (0.003)
	50	0.831 (0.003)	1.000 (0.000)	0.409 (0.003)	0.822 (0.002)	1.000 (0.000)	0.398 (0.003)	0.833 (0.002)	1.000 (0.000)	0.411 (0.003)	0.814 (0.002)	1.000 (0.000)	0.390 (0.002)
	100	0.891 (0.001)	1.000 (0.000)	0.378 (0.002)	0.894 (0.001)	1.000 (0.000)	0.383 (0.002)	0.903 (0.008)	1.000 (0.000)	0.399 (0.002)	0.902 (0.001)	1.000 (0.000)	0.397 (0.002)

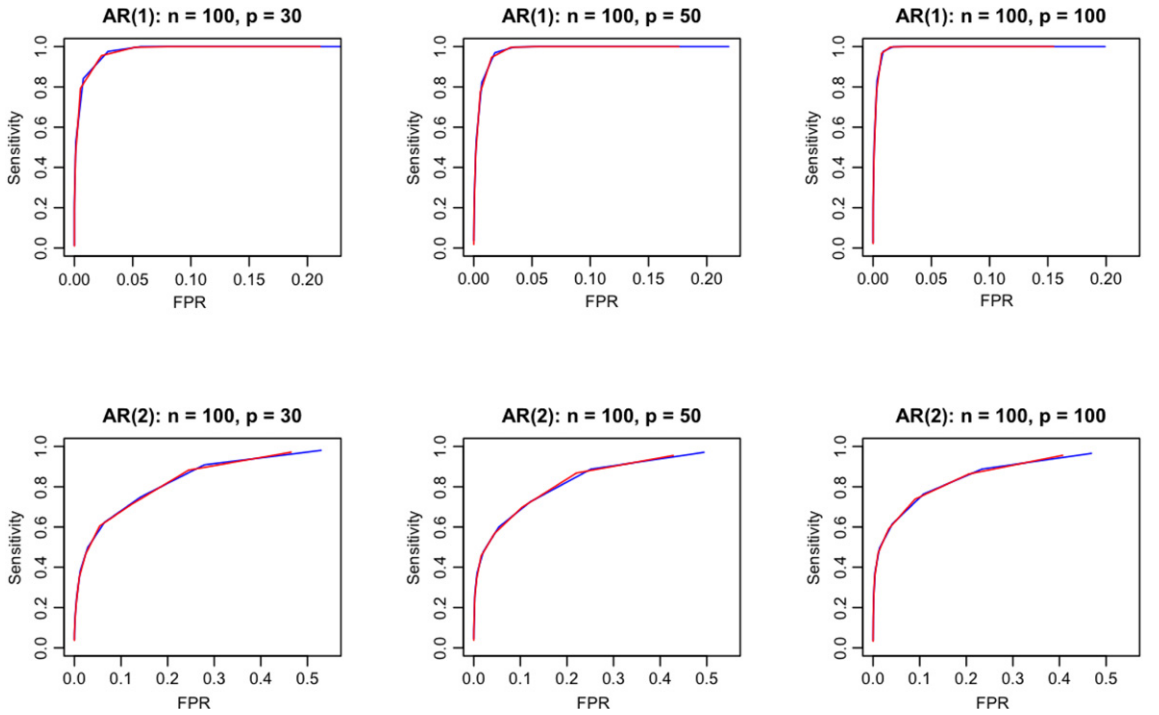

Fig. 1. ROC curves for AR(1) and AR(2) structures of the precision matrix corresponding to sample size $n = 100$ and matrix dimensions $p = 30, 50, 100$. The penalty parameter in the graphical lasso algorithm varies between 0.1 and 1. Red curve corresponds to the median probability model (MPP) and the blue curve corresponds to graphical lasso (GL). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Denoting Y_{tj} as the closing stock price for the jth stock on day t, we construct the 1257×452 data matrix $\boldsymbol{S}$ with entries $s_{tj} = \log(Y_{(t+1)j}/Y_{tj})$, $t = 1, \ldots, 1257$, $j = 1, \ldots, 452$. For analysis, we construct the data matrix $\boldsymbol{X}$ by standardizing $\boldsymbol{S}$, so that each stock has mean zero and standard deviation one. We find the median probability model as selected by the Bayesian

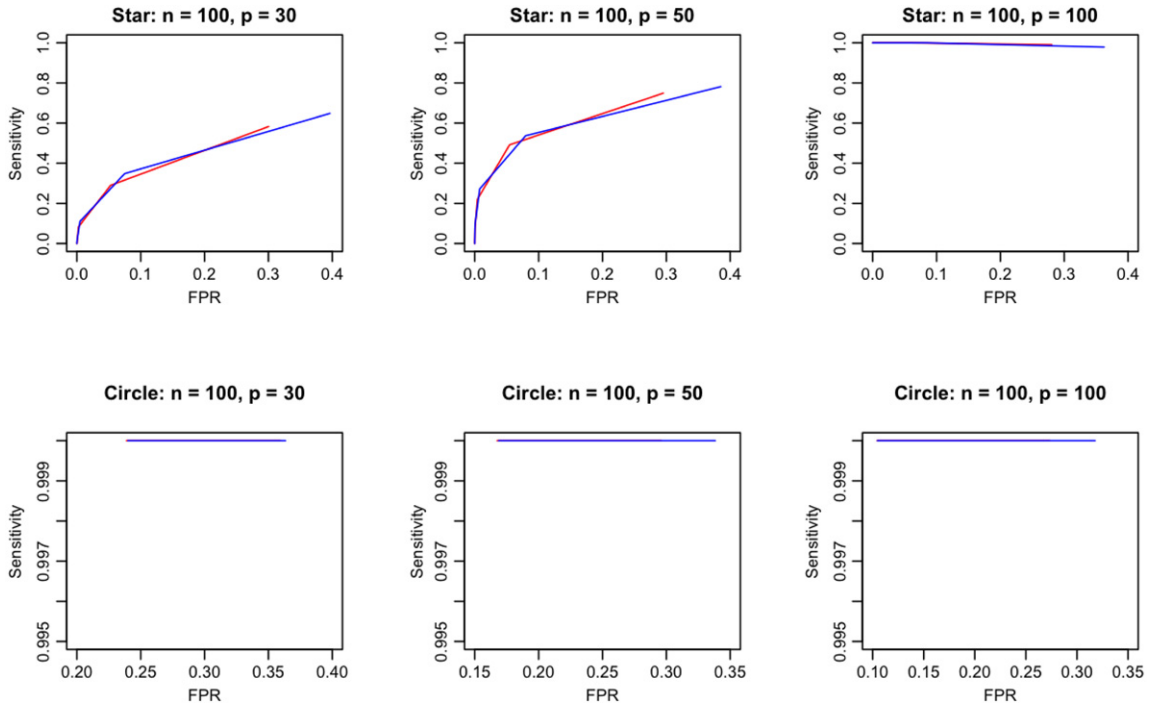

Fig. 2. ROC curves for Star and Circle structures of the precision matrix corresponding to sample size $n = 100$ and matrix dimensions $p = 30, 50, 100$. The penalty parameter in the graphical lasso algorithm varies between 0.1 and 1. Red curve corresponds to the median probability model (MPP) and the blue curve corresponds to graphical lasso (GL). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

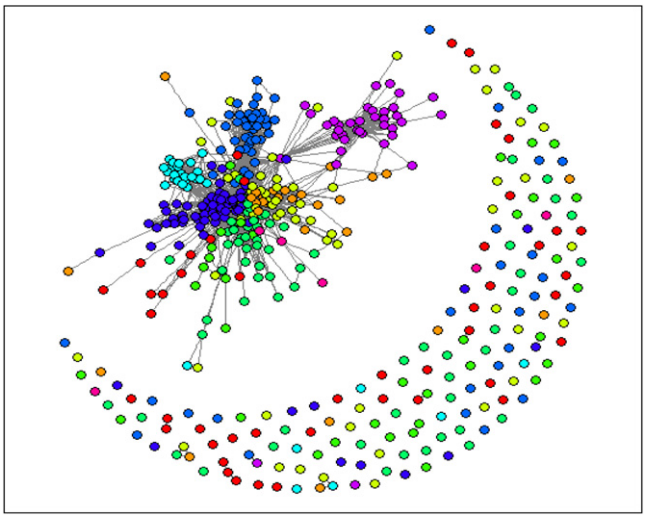

Fig. 3. Graphical structure of the median probability model selected by the Bayesian graphical structure learning method for the stock price data. Vertices of the graph are colored corresponding to different GICS sectors. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

graphical structure learning method. The corresponding graphical structure is displayed in Fig. 3. The vertices of the graph are colored corresponding to the different GICS sectors. We find that stocks from the same sectors tend to be related with other members from that category, and generally not related across different sectors, though there are some connections. The grouping of the stocks corresponding to their sectors is expected, implying that the stock prices for a particular sector are conditionally independent of those of other sectors.

We also individually study data pertaining to some of the specific sectors to have a closer look at the strength of the groupings where perturbations due to latent factors is least expected. For this, we consider the sectors "Utilities" and "Information Technology". The graphical structure is displayed in Fig. 4. The two sectors clearly separate as expected.

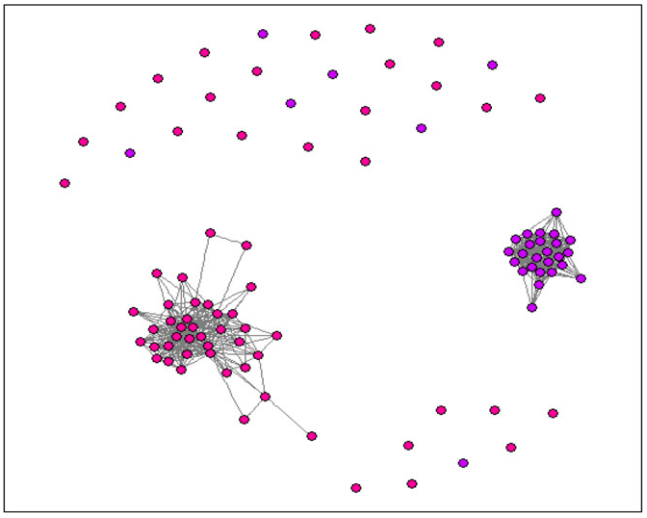

Fig. 4. Graphical structure corresponding to the subgraph corresponding to the sectors "Utilities" [red] and "Information Technology" [violet]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Appendix. Proofs

Proof of Theorem 3.1. We apply the general theory of posterior convergence rate by verifying the conditions in Theorem 2.1 of Ghosal et al. [15]. This requires evaluating the prior concentration rate of Kullback–Leibler neighborhoods, finding a suitable sieve in the space of densities and controlling its Hellinger metric entropy and showing that the complement of the sieve has exponentially small prior probability. Then the posterior convergence rate at the true density is obtained in terms of the Hellinger distance. Finally, by Lemma A.1, within the model the Hellinger distance is equivalent with the Frobenius distance on precision matrices. Hence the entropy calculation can be done in terms of the Frobenius distance and the convergence rate at the true precision matrix will also follow in terms of the Frobenius distance.

To estimate prior concentration, let

$$B(p_{\boldsymbol{\Omega}_0}, \epsilon_n) = \{p : K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \le \epsilon_n^2, V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \le \epsilon_n^2\},$$

where $K(f, g) = \int f \log(f/g)$ and $V(f, g) = \int f \log^2(f/g)$. Recall that, for $\boldsymbol{Z} \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and a $p \times p$ symmetric matrix $\boldsymbol{A}$, we have,

$$\mathrm{E}(\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}), \qquad \mathrm{Var}(\boldsymbol{Z}^T\boldsymbol{A}\boldsymbol{Z}) = 2\,\mathrm{tr}(\boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}\boldsymbol{\Sigma}). \tag{A.1}$$

Then

$$\begin{aligned} K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) &= \frac{1}{2}(\log\det\boldsymbol{\Omega}_0 - \log\det\boldsymbol{\Omega}) - \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - \boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}) \\ &= -\frac{1}{2}\sum_{i=1}^{p}\log d_i - \frac{1}{2}\sum_{i=1}^{p}(1 - d_i), \end{aligned}$$

and

$$\begin{aligned} V(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - 2\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1} + \boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1}) \\ &= \frac{1}{2}\mathrm{tr}(\boldsymbol{I}_p - \boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2})^2 \\ &= \frac{1}{2}\sum_{i=1}^{p}(1 - d_i)^2, \end{aligned}$$

where d_i, $i = 1, \ldots, p$, are the eigenvalues of $\boldsymbol{\Omega}_0^{-1/2}\boldsymbol{\Omega}\boldsymbol{\Omega}_0^{-1/2}$. Now, $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \ge h^2(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}})$, and hence by Lemma A.1 $\max_i|d_i - 1| < 1$. Hence we can expand $\log d_i$ in the powers of $(1 - d_i)$ to get $K(p_{\boldsymbol{\Omega}_0}, p_{\boldsymbol{\Omega}}) \sim \frac{1}{4}\sum_{i=1}^{p}(1 - d_i)^2$. By Lemma A.1, $\sum_{i=1}^{p}(1 - d_i)^2$ is bounded by a constant multiple of $\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_2^2$, and hence a constant multiple of $p^2\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_\infty^2$ in view of (2.1). Therefore $B(p_{\boldsymbol{\Omega}_0}, \epsilon_n) \supset \{p_{\boldsymbol{\Omega}} : \|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_\infty \le c\epsilon_n/p\}$, and hence it suffices to get a lower estimate of the prior probability of the latter set. The components of $\boldsymbol{\Omega}$ are not independently distributed, since the prior for $\boldsymbol{\Omega}$ is truncated to $\mathcal{M}_0^+$. However, as a small neighborhood of $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, s)$ lies within $\mathcal{M}_0^+$, the truncation can only increase prior concentration of $B(p_{\boldsymbol{\Omega}_0}, \epsilon_n)$.

Therefore we can pretend componentwise independence for the purpose of lower bounding the above prior probability which gives the estimate

$$\Pi\left(\|\boldsymbol{\Omega}_0 - \boldsymbol{\Omega}\|_\infty \leq c\epsilon_n/p\right) \gtrsim (c\epsilon_n/p)^{p+s}, \tag{A.2}$$

where Π denotes the prior distribution on $\boldsymbol{\Omega}$. The prior concentration rate condition thus gives,

$$(p+s)(\log p + \log {\epsilon_n}^{-1}) \asymp n\epsilon_n^2, \tag{A.3}$$

so as to get $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$.

Consider the sieve $\mathcal{P}_n$ to be the space of all densities $p_{\boldsymbol{\Omega}}$ such that the graph induced by $\boldsymbol{\Omega}$ has maximum number of edges $\bar{r} < \binom{p}{2}/2$ and each entry of $\boldsymbol{\Omega}$ is at most L in absolute value, where $\bar{r}$ and L depend on n and are to be chosen later. Then the metric entropy of the set of precision matrices with respect to the Frobenius distance is given by

$$\begin{aligned}\log\left\{\sum_{j=1}^{\bar{r}}\left(\frac{L}{\epsilon_n}\right)^j\binom{\binom{p}{2}}{j}\right\} &\leq \log\left\{\bar{r}\left(\frac{L}{\epsilon_n}\right)^{\bar{r}}\binom{p+\binom{p}{2}}{\bar{r}}\right\}\\ &\lesssim \log\bar{r} + \bar{r}\log L + \bar{r}\log{\epsilon_n}^{-1} + \bar{r}\log p,\end{aligned}$$

so the choices $\bar{r} \sim b_1 n\epsilon_n^2/\log n$, and $L \sim b_2 n\epsilon_n^2$ for any choice of constants $b_1, b_2 > 0$ will satisfy the rate equation

$$\log\bar{r} + \bar{r}\log p + \bar{r}\log{\epsilon_n}^{-1} + \bar{r}\log(n\epsilon_n^2) \asymp n\epsilon_n^2. \tag{A.4}$$

Note that under the condition $n\epsilon_n^2/\log n \ll \binom{p}{2}$, the requirement $\bar{r} < \binom{p}{2}/2$ is satisfied as $n \to \infty$.

To bound the prior probability of $\mathcal{P}_n^c$, observe that $p_{\boldsymbol{\Omega}}$ can fall outside $\mathcal{P}_n$ only if either an entry exceeds L or the number of off-diagonal entries exceed $\bar{r}$. The probability of the former event is bounded by $\binom{p}{2}e^{-L}$, which is bounded by $\exp(-b_3 n\epsilon_n^2)$ for some constant $b_3 > 0$ by the choice $L \sim b_2 n\epsilon_n^2$ and b_3 can be chosen to be as large as we like by choosing b_2 sufficiently large. The probability of the latter event is

$$\mathrm{P}(\bar{R} > \bar{r}) \leq \exp(-a_2' b_1 n\epsilon_n^2), \tag{A.5}$$

where $a_2' b_1$ can be made as large as possible by making b_1 large. Hence $\epsilon_n = n^{-1/2}(p+s)^{1/2}(\log n)^{1/2}$ is the posterior convergence rate. □

The following lemma establishes a norm equivalence necessary for finding posterior convergence rate and metric entropy calculations.

Lemma A.1. *If $p_{\boldsymbol{\Omega}_k}$ is the density of $\mathrm{N}_p(\mathbf{0}, \boldsymbol{\Omega}_k^{-1})$, $k = 1, 2$, then for all $\boldsymbol{\Omega}_k \in \mathcal{M}_0^+$, $k = 1, 2$, and d_i, $i = 1, \ldots, p$, eigenvalues of $\boldsymbol{A} = \boldsymbol{\Omega}_1^{-1/2}\boldsymbol{\Omega}_2\boldsymbol{\Omega}_1^{-1/2}$, we have that for some $\delta > 0$ and constant $c_0 > 0$,*

(i) $c_0^{-1}\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 \leq \sum_{i=1}^p |d_i - 1|^2 \leq c_0\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2$,
(ii) $h(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) < \delta$ *implies* $\max_i |d_i - 1| < 1$ *and* $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2 \leq c_0 h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2})$,
(iii) $h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) \leq c_0\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2$,

Proof. Recall that as $\boldsymbol{\Omega}_1 \in \mathcal{M}_0^+$, both $\|\boldsymbol{\Omega}_1\|_2$ and $\|\boldsymbol{\Omega}_1^{-1}\|_2$ are bounded by a constant. As $\boldsymbol{I}_p - \boldsymbol{A}$ have eigenvalues $(1-d_1), \ldots, (1-d_p)$, we have

$$\begin{aligned}\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2 &= \|\boldsymbol{\Omega}_1^{1/2}(\boldsymbol{I}_p - \boldsymbol{A})\boldsymbol{\Omega}_1^{1/2}\|_2^2\\ &\leq \|\boldsymbol{\Omega}_1\|_{(2,2)}^2\|\boldsymbol{I}_p - \boldsymbol{A}\|_2^2\\ &= \|\boldsymbol{\Omega}_1\|_{(2,2)}^2\mathrm{tr}\left(\boldsymbol{I}_p - \boldsymbol{A}\right)^2\\ &= \|\boldsymbol{\Omega}_1\|_{(2,2)}^2\sum_{i=1}^p(d_i - 1)^2\end{aligned}$$

while conversely

$$\begin{aligned}\sum_{i=1}^p(d_i - 1)^2 &= \|\boldsymbol{I}_p - \boldsymbol{A}\|_2^2\\ &= \|\boldsymbol{\Omega}_1^{-1/2}(\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2)\boldsymbol{\Omega}_1^{-1/2}\|_2^2\\ &\leq \|\boldsymbol{\Omega}_1^{-1}\|_{(2,2)}^2\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2^2.\end{aligned}$$

This establishes (i). In particular, if $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2$ is sufficiently small, then $\max_i |d_i - 1| < 1$.

Now by direct calculations, in a normal scale model,

$$\frac{1}{2}h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) = 1 - \{\det(\boldsymbol{A}^{1/2} + \boldsymbol{A}^{-1/2})\}^{-1/2} = 1 - \left\{\prod_{i=1}^{p} \frac{1}{2}(d_i^{1/2} + d_i^{-1/2})\right\}^{-1/2},$$

If $h(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2}) < \delta$, this implies $1-\{\prod_{i=1}^{p} \frac{1}{2}(d_i^{1/2}+d_i^{-1/2})\}^{-1/2} \leq \delta^2/2$. Rearranging the terms, we get, $\prod_{i=1}^{p} \frac{1}{2}(d_i^{1/2}+d_i^{-1/2}) \leq (1-\delta^2/2)^{-2} = 1+\eta$, say. Since every term in the product exceeds 1, we have,

$$\max_i \frac{1}{2}(d_i^{1/2} + d_i^{-1/2}) \leq 1 + \eta. \tag{A.6}$$

The above equation, upon squaring and rearrangement of terms, gives, for all i, $(d_i - 1)^2 \leq 2d_i^{1/2}\eta$. Note that Eq. (A.6) gives that $d_i^{1/2} \leq 2(1+\eta)$. Hence, the above equation implies that $(d_i - 1)^2 \leq 4\eta(1+\eta)$. If $\delta > 0$ is chosen sufficiently small to make $\eta < (\sqrt{2}-1)/2$, then $|d_i - 1| < 1$ for all $i = 1, \ldots, p$.

Abbreviating $h^2(p_{\boldsymbol{\Omega}_1}, p_{\boldsymbol{\Omega}_2})$ by h^2, the expression for the Hellinger distance gives $\prod_{i=1}^{p}(d_i^{1/2} + d_i^{-1/2}) = 2^p(1-h^2)^{-2}$. On the other hand, for some constants $c_1, c_2 > 0$,

$$\left[1 + c_1 \sum_{i=1}^{p}(d_i - 1)^2\right] \leq 2^{-p} \prod_{i=1}^{p}(d_i^{1/2} + d_i^{-1/2}) \leq \left[1 + c_2 \sum_{i=1}^{p}(d_i - 1)^2\right] \tag{A.7}$$

by a Taylor series expansion, which is possible since $\max_i |d_i - 1| < 1$. The lower estimate gives $c_1 \sum_{i=1}^{p}(d_i - 1)^2 \leq (1-h^2)^{-2} \sim 2h^2$, which proves (ii).

Finally, since the Hellinger distance is bounded, to prove (iii), it suffices to consider the case $\|\boldsymbol{\Omega}_1 - \boldsymbol{\Omega}_2\|_2 < \delta$ where $\delta > 0$ is sufficiently small. Then part (i) implies that $\max_i |d_i - 1| < 1$, and hence the upper estimate in (A.7) gives $2h^2 \sim (1-h^2)^{-2} \leq c_2 \sum_{i=1}^{p}(d_i - 1)^2$, which proves (iii) in view of part (i). □

Proof of Lemma 4.1. The graphical lasso for the model $\boldsymbol{\Gamma}$, given by $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^* = ((\omega_{\boldsymbol{\Gamma},ij}^*))$ satisfies

$$\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1} - \widehat{\boldsymbol{\Sigma}} - \lambda \boldsymbol{G} = \boldsymbol{0}, \tag{A.8}$$

where $\boldsymbol{G} = ((g_{ij}))$ is a matrix with elements $g_{ij} = \omega_{\boldsymbol{\Gamma},ij}^*/|\omega_{\boldsymbol{\Gamma},ij}^*|$ if $\omega_{\boldsymbol{\Gamma},ij}^* \neq 0$ and $g_{ij} \in [-1, 1]$ if $\omega_{\boldsymbol{\Gamma},ij}^* = 0$, by the Karush–Kuhn–Tucker (KKT) condition; see, for example, Boyd and Vandenberghe [6], Witten et al. [33]. When the model $\boldsymbol{\Gamma}$ is non-regular and $\boldsymbol{\Gamma}'$ is its corresponding regular submodel, let (i, j) be a pair such that $\gamma_{ij} = 1$ but $\omega_{\boldsymbol{\Gamma},ij}^* = 0$. Then $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*$ automatically satisfies the KKT condition also for the model $\boldsymbol{\Gamma}'$ because $\omega_{\boldsymbol{\Gamma},ij}^* = 0$ for any $\gamma_{ij}' = 0$. □

The following lemma is essential in proving the ignorability of the non-regular models for posterior probability evaluation.

Lemma A.2. *Consider a non-regular model $\boldsymbol{\Gamma}$ and let $\boldsymbol{\Gamma}'$ be its corresponding regular submodel with their common graphical lasso $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*$. Let $\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} = \boldsymbol{\Omega}_{\boldsymbol{\Gamma}} - \boldsymbol{\Omega}^* = ((u_{\boldsymbol{\Gamma},ij}))$ and $\boldsymbol{\Delta}_{\boldsymbol{\Gamma}'} = ((u_{\boldsymbol{\Gamma}',ij}))$ such that $u_{\boldsymbol{\Gamma}',ij} = u_{\boldsymbol{\Gamma},ij}$ if $i = j$ or $\gamma_{ij} = \gamma_{ij}' = 1$ and $u_{\boldsymbol{\Gamma}',ij} = 0$ for pairs (i, j) with $\gamma_{ij}' = 0$. Then for fixed values of $u_{\boldsymbol{\Gamma}',ij}$ for $\gamma_{ij}' = 1$, we have,*

$$\log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}) \leq \log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}'} + \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}'}). \tag{A.9}$$

Proof. Consider maximization of the function

$$f(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}) = \log\det(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*) - \operatorname{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}) \tag{A.10}$$

with respect to the elements $u_{\boldsymbol{\Gamma},ij}$ where $(i, j) \in \{(i, j) : \gamma_{ij} = 1, \gamma_{ij}' = 0\}$. Differentiating the above function for a particular value of u_{ij} gives,

$$\frac{\partial f(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})}{\partial u_{\boldsymbol{\Gamma},ij}} = \operatorname{tr}\left[\left\{(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]. \tag{A.11}$$

The maximizer $\widehat{u}_{\boldsymbol{\Gamma},ij}$ satisfies $\operatorname{tr}\left[\left\{(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}} + \boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^*)^{-1}\boldsymbol{E}_{(i,j)} - \widehat{\boldsymbol{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right] = 0$. Now consider the function $g_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})$ defined in (4.4). The derivative of $g_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})$ with respect to $u_{\boldsymbol{\Gamma},ij}$ satisfies

$$\left.\frac{\partial g_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})}{\partial u_{\boldsymbol{\Gamma},ij}}\right|_{u_{\boldsymbol{\Gamma},ij}=0^+, u_{\boldsymbol{\Gamma},lm}=0, \quad \forall(l,m)\neq(i,j)} \geq 0, \tag{A.12}$$

and

$$\left.\frac{\partial g_{\boldsymbol{\Gamma}}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})}{\partial u_{\boldsymbol{\Gamma},ij}}\right|_{u_{\boldsymbol{\Gamma},ij}=0^-, u_{\boldsymbol{\Gamma},lm}=0, \quad \forall(l,m)\neq(i,j)} \leq 0. \tag{A.13}$$

The above two conditions give,

$$\operatorname{tr}\left[\left\{(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}+\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*})^{-1}\boldsymbol{E}_{(i,j)}-\widehat{\boldsymbol{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]\Big|_{u_{\boldsymbol{\Gamma},ij}=0^{+},u_{\boldsymbol{\Gamma},lm}=0,\forall(l,m)\neq(i,j)}=0\leq\frac{2\lambda}{n},$$

$$\operatorname{tr}\left[\left\{(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}+\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*})^{-1}\boldsymbol{E}_{(i,j)}-\widehat{\boldsymbol{\Sigma}}\boldsymbol{E}_{(i,j)}\right\}\right]\Big|_{u_{\boldsymbol{\Gamma},ij}=0^{-},u_{\boldsymbol{\Gamma},lm}=0,\forall(l,m)\neq(i,j)}=0\geq-\frac{2\lambda}{n}.$$

Since the first derivative of $f(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})$ is continuous at 0, we have $\widehat{u}_{\boldsymbol{\Gamma},ij}=0$. This immediately implies the result stated in the lemma. □

Proof of Lemma 4.3. The Taylor series expansion of $h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}})$ under model $\boldsymbol{\Gamma}$, defined by (4.2) gives,

$$h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}})=h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*})+\frac{1}{2}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{T}\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})+R_{n},\tag{A.14}$$

where R_n is the remainder term in the expansion. Using the integral form of the remainder, we have,

$$h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}})=h_{\boldsymbol{\Gamma}}(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*})+\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{T}\left\{\int_{0}^{1}(1-\nu)\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}d\nu\right\}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}).\tag{A.15}$$

Subtracting (A.15) from (A.14) gives,

$$\begin{aligned}
R_n &= \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{T}\left\{\int_{0}^{1}(1-\nu)\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}d\nu\right\}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})-\frac{1}{2}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{T}\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})\\
&= \operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{T}\left\{\int_{0}^{1}(1-\nu)\left(\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\right)d\nu\right\}\operatorname{vec}(\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})\\
&\leq \|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}\left\|\int_{0}^{1}(1-\nu)\left(\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\right)d\nu\right\|_{(2,2)}\\
&\leq \|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}\int_{0}^{1}(1-\nu)\|\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\|_{(2,2)}d\nu\\
&\leq \frac{1}{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}\max_{0\leq\nu\leq1}\|\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\|_{(2,2)}\\
&\leq \frac{1}{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}(p+\#\boldsymbol{\Gamma})\max_{0\leq\nu\leq1}\|\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\|_{\infty}
\end{aligned}\tag{A.16}$$

since the Hessian is a matrix of order $(p+\#\boldsymbol{\Gamma})\times(p+\#\boldsymbol{\Gamma})$ for a regular model $\boldsymbol{\Gamma}$. The above bound involves the maximum of the absolute differences between the elements of the Hessian matrices $\boldsymbol{H}$ computed at two different values $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}$. Observe that by the matrix norm relations in (2.1) and (3.1) with probability tending to one,

$$\begin{aligned}
\|(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{-1}-\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\|_{\infty} &\leq \|(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{-1}-\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\|_{(2,2)}\\
&= \|\nu\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}(\boldsymbol{I}+\nu\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{-1}\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\|_{(2,2)}\\
&\leq \nu\|\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*-1}\|_{(2,2)}^{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{(2,2)}\\
&\leq K\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}.
\end{aligned}\tag{A.17}$$

For any symmetric matrix $\boldsymbol{A}=((a_{ij}))$ we note that $\operatorname{tr}\left\{\boldsymbol{A}\boldsymbol{E}_{(i,j)}\boldsymbol{A}\boldsymbol{E}_{(l,m)}\right\}$ has the form $2a_{il}a_{jm}+2a_{im}a_{jl}$ for $i\neq j, l\neq m$; $2a_{il}a_{im}$ for $i=j, l\neq m$; a_{il}^{2} for $i=j, l=m$. Hence the elements of $\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}$ have the respective forms $(2a_{il}a_{jm}+2a_{im}a_{jl})-(2b_{il}b_{jm}+2b_{im}b_{jl})$, $2a_{il}a_{im}-2b_{il}b_{im}$ and $a_{il}^{2}-b_{il}^{2}$, where $((a_{ij}))=(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*})^{-1}$ and $((b_{ij}))=(\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}})^{-1}$. Then, using Eq. (A.17), we get, with probability tending to one, for all $((i,j),(l,m))$

$$\sum a_{il}a_{jm}-\sum b_{il}b_{jm}\leq C_{1}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}+C_{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}.\tag{A.18}$$

Since this holds true for any arbitrary element of $\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}$, using (3.3) and (A.18), we get that with probability tending to one,

$$\|\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}+\nu\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}}-\boldsymbol{H}_{\boldsymbol{\Omega}_{\boldsymbol{\Gamma}}^{*}}\|_{\infty}\leq C_{1}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}+C_{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2},\tag{A.19}$$

where C_1 and C_2 are suitable constants. Now using (A.16) and (A.19), with probability tending to one, we have,

$$R_{n}\leq\frac{1}{2}(p+\#\boldsymbol{\Gamma})\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}\left(C_{1}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}+C_{2}\|\boldsymbol{\Delta}_{\boldsymbol{\Gamma}}\|_{2}^{2}\right).\quad\square$$

Proof of Theorem 4.4. Using the Taylor series expansion of $h_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})$ as in (A.14), we can write the posterior probability of model $\mathbf{\Gamma}$ given the data $\boldsymbol{X}^{(n)}$ as in Eq. (4.1) to be proportional to

$$\int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}\in\mathcal{M}_0^+} \exp\left\{-\frac{n}{2}\left(h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})+\frac{1}{2}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})+R_n\right)\right\}\prod_{(i,j)\in\bar{E}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma},ij}.$$

We denote $\prod_{(i,j)\in\bar{E}_{\mathbf{\Gamma}}} du_{\mathbf{\Gamma}ij}$ by $d\mathbf{\Delta}_{\mathbf{\Gamma}}$ for notational simplicity. Using (3.15), we get

$$\frac{\int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2\le\epsilon_n}\exp\left[-n\left\{h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})+\frac{1}{2}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})+R_n\right\}/2\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}}{\int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}\in\mathcal{M}_0^+}\exp\left[-n\left\{h_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})+\frac{1}{2}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})+R_n\right\}/2\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}}\to 1.$$

Also, for $\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2\le\epsilon_n$, $R_n\le(p+\#\mathbf{\Gamma})\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2^2\epsilon_n/2$. Thus, the upper and lower bounds of the integral $\int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2\le\epsilon_n}\exp\{-nh_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})/2\}d\mathbf{\Delta}_{\mathbf{\Gamma}}$ are given by

$$e^{-nh_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})/2}\int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2\le\epsilon_n}\exp\left[-n\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\left\{\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mp(p+\#\mathbf{\Gamma})\epsilon_n\boldsymbol{I}\right\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})/4\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}.$$

Note that,

$$\int_{\|\mathbf{\Delta}_{\mathbf{\Gamma}}\|_2>\epsilon_n}\exp\left[-n\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\left\{\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mp(p+\#\mathbf{\Gamma})\epsilon_n\boldsymbol{I}\right\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})/4\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}\to 0,$$

if $(p+\#\mathbf{\Gamma})\epsilon_n\to 0$ and the minimum eigenvalue of $\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$ is bounded away from zero, which we prove in Lemma A.3 below. Hence, the bounds can be simplified to

$$e^{-nh_{\mathbf{\Gamma}}(\mathbf{\Omega}^*_{\mathbf{\Gamma}})/2}\int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}\in\mathcal{M}_0^+}\exp\left[-\frac{n}{4}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\left\{\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mp(p+\#\mathbf{\Gamma})\epsilon_n\boldsymbol{I}\right\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}.$$

Using the above bounds, the ratio of the actual integral to the approximate integral has upper and lower bounds given by

$$\frac{\int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}\in\mathcal{M}_0^+}\exp\left[-n\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\left\{\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mp(p+\#\mathbf{\Gamma})\epsilon_n\boldsymbol{I}\right\}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})/4\right]d\mathbf{\Delta}_{\mathbf{\Gamma}}}{\int_{\mathbf{\Delta}_{\mathbf{\Gamma}}+\mathbf{\Omega}^*_{\mathbf{\Gamma}}\in\mathcal{M}_0^+}\exp\left\{-n\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})^T\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mathrm{vec}(\mathbf{\Delta}_{\mathbf{\Gamma}})/4\right\}d\mathbf{\Delta}_{\mathbf{\Gamma}}}=\left[\frac{\det\left\{\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\mp(p+\#\mathbf{\Gamma})\epsilon_n\boldsymbol{I}\right\}}{\det(\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}})}\right]^{-1/2}. \quad \text{(A.20)}$$

The above expression lies between $\left[1\mp\{\mathrm{eig}_1(\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}})\}^{-1}(p+\#\mathbf{\Gamma})\epsilon_n\right]^{-(p+\#\mathbf{\Gamma})/2}$. Using Lemma A.3 below, $\mathrm{eig}_1(\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}})\gg 0$, and hence the above bound on the ratio goes to 1 if $(p+\#\mathbf{\Gamma})^2\epsilon_n\to 0$, so that the error in Laplace approximation is asymptotically small. □

We now prove the result that the eigenvalues of the Hessian $\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$ are bounded away from zero.

Lemma A.3. *Given a model $\mathbf{\Gamma}$, the minimum eigenvalue of the Hessian $\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$ corresponding to the function $h_{\mathbf{\Gamma}}(\mathbf{\Omega}_{\mathbf{\Gamma}})$, evaluated at $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$, is bounded away from zero.*

Proof. Note that the Hessian $\boldsymbol{H}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$ evaluated at the graphical lasso $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$ corresponding to the model $\mathbf{\Gamma}$ has the form $\boldsymbol{T}'\boldsymbol{A}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\boldsymbol{T}$, where the $(p+2\#\mathbf{\Gamma})$-dimensional matrix $\boldsymbol{A}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}$ is a principal minor of the $p^2\times p^2$ matrix $(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\otimes(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}$, and $\boldsymbol{T}$ is a $(p+2\#\mathbf{\Gamma})\times(p+\#\mathbf{\Gamma})$ matrix of 0s and 1s having full column rank. Thus $\boldsymbol{T}'\boldsymbol{A}_{\mathbf{\Omega}^*_{\mathbf{\Gamma}}}\boldsymbol{T}$ has full rank if the minimum eigenvalue of $(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\otimes(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}$ is bounded away from zero. Note that, $\mathrm{eig}_1\{(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\otimes(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\}=[\mathrm{eig}_1\{(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\}]^2$. The parameter space $\mathcal{M}_0^+$ of precision matrices insist on fixed bounds for minimum and maximum eigenvalues and thus $\mathbf{\Omega}^*_{\mathbf{\Gamma}}$ in any model $\mathbf{\Gamma}$ has to maintain the eigenvalue restriction. Hence, $[\mathrm{eig}_1\{(\mathbf{\Omega}^*_{\mathbf{\Gamma}})^{-1}\}]^2=1/\|\mathbf{\Omega}^*_{\mathbf{\Gamma}}\|_2^2>0$. □

References

[1] A. Atay-Kayis, H. Massam, A Monte–Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models, Biometrika 92 (2) (2005) 317–335.
[2] O. Banerjee, L. El Ghaoui, A. d'Aspremont, Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data, J. Mach. Learn. Res. 9 (2008) 485–516.
[3] S. Banerjee, S. Ghosal, Posterior convergence rates for estimating large precision matrices using Graphical models, Electron. J. Stat. 8 (2) (2014) 2111–2137.
[4] P. Bickel, E. Levina, Covariance regularization by thresholding, Ann. Statist. 36 (6) (2008) 2577–2604.
[5] P. Bickel, E. Levina, Regularized estimation of large covariance matrices, Ann. Statist. 36 (1) (2008) 199–227.
[6] S.P. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
[7] T. Cai, W. Liu, X. Luo, A constrained ℓ_1-minimization approach to sparse precision matrix estimation, J. Amer. Statist. Assoc. 106 (494) (2011) 594–607.
[8] T. Cai, C. Zhang, H. Zhou, Optimal rates of convergence for covariance matrix estimation, Ann. Statist. 38 (4) (2010) 2118–2144.

[9] I. Castillo, A. van der Vaart, Needles and straw in a haystack: posterior concentration for possibly sparse sequences, Ann. Statist. 40 (4) (2012) 2069–2101.
[10] S.M. Curtis, S. Banerjee, S. Ghosal, Fast Bayesian model assessment for nonparametric additive regression, Comput. Statist. Data Anal. 71 (2014) 347–358.
[11] A. Dawid, S. Lauritzen, Hyper Markov laws in the statistical analysis of decomposable graphical models, Ann. Statist. 21 (3) (1993) 1272–1317.
[12] A. Dobra, C. Hans, B. Jones, J. Nevins, G. Yao, M. West, Sparse graphical models for exploring gene expression data, J. Multivariate Anal. 90 (1) (2004) 196–212.
[13] J. Friedman, T. Hastie, R. Tibshirani, Sparse inverse covariance estimation with the graphical lasso, Biostatistics 9 (3) (2008) 432–441.
[14] S. Ghosal, Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity, J. Multivariate Anal. 74 (1) (2000) 49–68.
[15] S. Ghosal, J.K. Ghosh, A.W. Van Der Vaart, Convergence rates of posterior distributions, Ann. Statist. 28 (2) (2000) 500–531.
[16] J. Guo, E. Levina, G. Michailidis, J. Zhu, Joint estimation of multiple graphical models, Biometrika 98 (1) (2011) 1–15.
[17] D.A. Harville, Matrix Algebra from a Statistician's Perspective, Springer, 2008.
[18] J. Huang, N. Liu, M. Pourahmadi, L. Liu, Covariance matrix selection and estimation via penalised normal likelihood, Biometrika 93 (1) (2006) 85–98.
[19] N. Karoui, Operator norm consistent estimation of large-dimensional sparse covariance matrices, Ann. Statist. 36 (6) (2008) 2717–2756.
[20] C. Lam, J. Fan, Sparsistency and rates of convergence in large covariance matrix estimation, Ann. Statist. 37 (6B) (2009) 4254.
[21] S. Lauritzen, Graphical Models, Clarendon Press, Oxford, 1996.
[22] O. Ledoit, M. Wolf, A well-conditioned estimator for large-dimensional covariance matrices, J. Multivariate Anal. 88 (2) (2004) 365–411.
[23] G. Letac, H. Massam, Wishart distributions for decomposable graphs, Ann. Statist. 35 (3) (2007) 1278–1323.
[24] H. Liu, J. Lafferty, L. Wasserman, The nonparanormal: semiparametric estimation of high dimensional undirected graphs, J. Mach. Learn. Res. 10 (2009) 2295–2328.
[25] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, Ann. Statist. 34 (3) (2006) 1436–1462.
[26] T. Park, G. Casella, The Bayesian lasso, J. Amer. Statist. Assoc. 103 (482) (2008) 681–686.
[27] D. Pati, A. Bhattacharya, N. Pillai, D. Dunson, Posterior contraction in sparse Bayesian factor models for massive covariance matrices, Ann. Statist. 42 (3) (2014) 1102–1130.
[28] B. Rajaratnam, H. Massam, C. Carvalho, Flexible covariance estimation in graphical Gaussian models, Ann. Statist. 36 (6) (2008) 2818–2849.
[29] A. Rothman, P. Bickel, E. Levina, J. Zhu, Sparse permutation invariant covariance estimation, Electron. J. Statist. 2 (2008) 494–515.
[30] A. Rothman, E. Levina, J. Zhu, Generalized thresholding of large covariance matrices, J. Amer. Statist. Assoc. 104 (485) (2009) 177–186.
[31] A. Roverato, Cholesky decomposition of a hyper inverse Wishart matrix, Biometrika 87 (1) (2000) 99–112.
[32] H. Wang, Bayesian graphical lasso models and efficient posterior computation, Bayesian Anal. 7 (4) (2012) 867–886.
[33] D.M. Witten, J.H. Friedman, N. Simon, New insights and faster computations for the graphical lasso, J. Comput. Graph. Statist. 20 (4) (2011) 892–900.
[34] M. Yuan, Y. Lin, Efficient empirical bayes variable selection and estimation in linear models, J. Amer. Statist. Assoc. 100 (472) (2005).
[35] M. Yuan, Y. Lin, Model selection and estimation in the Gaussian graphical model, Biometrika 94 (1) (2007) 19–35.
[36] T. Zhao, H. Liu, K. Roeder, J. Lafferty, L. Wasserman, The huge package for high-dimensional undirected graph estimation in R, J. Mach. Learn. Res. 98888 (2012) 1059–1062.