Electronic Journal of Statistics
Vol. 8 (2014) 2111–2137
ISSN: 1935-7524
DOI: 10.1214/14-EJS945

Posterior convergence rates for estimating large precision matrices using graphical models*

Sayantan Banerjee

Department of Biostatistics
The University of Texas MD Anderson Cancer Center
1400 Pressler Street
Houston, TX 77030
USA
e-mail: SBanerjee@mdanderson.org

and

Subhashis Ghosal

Department of Statistics
North Carolina State University
4276 SAS Hall, 2311 Stinson Drive
Raleigh, NC 27695-8203
USA
e-mail: sghosal@stat.ncsu.edu

Abstract: We consider Bayesian estimation of a $p \times p$ precision matrix, when p can be much larger than the available sample size n. It is well known that consistent estimation in such ultra-high dimensional situations requires regularization such as banding, tapering or thresholding. We consider a banding structure in the model and induce a prior distribution on a banded precision matrix through a Gaussian graphical model, where an edge is present only when two vertices are within a given distance. For a proper choice of the order of graph, we obtain the convergence rate of the posterior distribution and Bayes estimators based on the graphical model in the L_∞-operator norm uniformly over a class of precision matrices, even if the true precision matrix may not have a banded structure. Along the way to the proof, we also compute the convergence rate of the maximum likelihood estimator (MLE) under the same set of condition, which is of independent interest. The graphical model based MLE and Bayes estimators are automatically positive definite, which is a desirable property not possessed by some other estimators in the literature. We also conduct a simulation study to compare finite sample performance of the Bayes estimators and the MLE based on the graphical model with that obtained by using a Cholesky decomposition of the precision matrix. Finally, we discuss a practical method of choosing the order of the graphical model using the marginal likelihood function.

AMS 2000 subject classifications: Primary 62H12; secondary 62F12, 62F15.
Keywords and phrases: Precision matrix, G-Wishart, convergence rate.

Received November 2013.

*Research is partially supported by NSF grant number DMS-1106570.

1. Introduction

Estimating a covariance matrix or a precision matrix (inverse covariance matrix) is one of the most important problems in multivariate analysis. Of special interest are situations where the number of underlying variables p is much larger than the sample size n. These situations are common in gene expression data, fMRI data and in several other modern applications. Special care needs to be taken for tackling such high-dimensional scenarios. Conventional estimators like the sample covariance matrix or maximum likelihood estimator behave poorly when the dimension is much higher than the sample size.

Different regularization based methods have been proposed and developed in the recent years for dealing with high-dimensional data. These include banding, thresholding, tapering and penalization based methods to name a few; see, for example, [20, 16, 31, 2, 3, 17, 13, 28, 18, 27, 7, 5]. Most of these regularization based methods for high dimensional models impose a sparse structure in the covariance or the precision matrix, as in [2], where a rate of convergence has been derived for the estimator obtained by "banding" the sample covariance matrix, or by banding the Cholesky factor of the inverse sample covariance matrix, as long as $n^{-1}\log p \to 0$. Cai et al. [7] obtained the minimax rate under the operator norm and constructed a tapering estimator which attains the minimax rate over a smoothness class of covariance matrices. Cai and Liu [4] proposed an adaptive thresholding procedure. More recently, Cai and Yuan [6] introduced a data-driven block-thresholding estimator which is shown to be optimally rate adaptive over some smoothness class of covariance matrices.

There are only a few convergence results available in the Bayesian setting for estimating large covariance or precision matrices. Ghosal [14] studied asymptotic normality of posterior distributions for exponential families (which include the multivariate normal scale family) when the dimension $p \to \infty$, but restricting to the situation $p \ll n$. Recently, Pati et al. [25] 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.

Graphical models [19] provide an excellent tool for sparse covariance or inverse covariance estimation; see [12, 23, 31, 13], as they capture the conditional dependence between the variables by means of a graph. Bayesian methods for inference using graphical models have also been developed, as in [29, 1, 22]. For a complete graph corresponding to the saturated model, clearly the Wishart distribution is the conjugate prior for the precision matrix $\mathbf{\Omega}$. For an incomplete decomposable graph, a conjugate family of priors is given by the G-Wishart prior [29]. The equivalent prior on the covariance matrix is termed as the hyper inverse Wishart distribution in [10]. Letac and Massam [22] 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 distribution were further explored in [26]. Rajarat-

nam et al. [26] also obtained expressions for Bayes estimators under different loss functions.

In this paper, we consider Bayesian estimation of the precision matrix working with a G-Wishart prior induced by a Gaussian graphical model, which has a Markov property with respect to a decomposable graph G. More specifically, we work with a Gaussian graphical model structure which induces banding in the corresponding precision matrix. Approximate banding structure for precision matrix can arise in certain possibly non-stationary time series framework. Suppose that $\{X_t : t = 1, \ldots, p\}$ is a possibly non-stationary time series with approximately Markov dependence on neighborhoods in the sense that off-diagonal elements of its precision matrix decay sufficiently fast with the lag. The covariances cannot be estimated based on a single time series due to lack of stationarity. However, if we have replications $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, even when n is much smaller than p, it is still possible to estimate the entire precision matrix assuming the approximate Markov structure. The graphical model based on the banding structure ensures the decomposability of the graph, along with the presence of a perfect set of cliques, as explained in Section 2. For a G-Wishart prior, we can compute the explicit expression of the normalizing constant of the corresponding marginal distribution of the graph (see Section 5). For arbitrary decomposable graphs, the computation of the normalizing constant requires Markov chain Monte-Carlo (MCMC) based methods; see [1, 8, 9, 21, 11]. We obtain posterior convergence rate and convergence rate of the Bayes estimators and the MLE for the graphical model based on banding on the precision matrix. However, we allow the true precision matrix to be outside this class, provided it is well-approximated by banded matrices in an appropriate sense.

The paper is organized as follows. In the next section, we discuss some preliminaries on graphical models. In Section 3, we formulate the estimation problem and the describe the corresponding model assumptions. Section 4 deals with the main results related to posterior convergence rates. A method for selecting the banding parameter using the explicit form of the marginal likelihood of a graph is discussed in Section 5. In Section 6, we compare the performance of the Bayesian estimators with that of the graphical maximum likelihood estimator (MLE) and the banding estimator proposed in [3]. Proofs of the results are presented in Section 7. Some auxiliary lemmas and their proofs are given in the Appendix.

2. Notations and preliminaries on graphical models

We first describe the notations to be used in this paper. 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 $\mathrm{P}(|X_n| \leq M\delta_n) \to 1$ for some constant M (respectively, $\mathrm{P}(|X_n| < \epsilon\delta_n) \to 1$ for all $\epsilon > 0$). For numerical sequences r_n and s_n, by $r_n \ll s_n$ (or, $r_n \gg s_n$) we mean that $r_n = o(s_n)$, while by $s_n \gtrsim r_n$ we mean that $r_n = O(s_n)$. By $r_n \asymp s_n$, we mean that $r_n = O(s_n)$ and $s_n = O(r_n)$, while $r_n \sim s_n$ stands for $r_n/s_n \to 1$. The indicator function is denoted by $\mathbb{1}$.

We denote vectors by bold lowercase English or Greek letters. The components of a vector are represented by the corresponding non-bold letters, that is, for $\boldsymbol{x} \in \mathbb{R}^p$, $\boldsymbol{x} = (x_1, \ldots, x_p)^T$. We define the following norms for a vector $\boldsymbol{x} \in \mathbb{R}^p$: $\|\boldsymbol{x}\|_r = (\sum_{j=1}^p |x_j|^r)^{1/r}$, $\|\boldsymbol{x}\|_\infty = \max_j |x_j|$. Matrices are denoted by bold uppercase English or Greek letters, like $\boldsymbol{A} = (\!(a_{ij})\!)$, where a_{ij} stands for the (i,j)th entry of $\boldsymbol{A}$. If $\boldsymbol{A}$ is a symmetric $p \times p$ matrix, let $\mathrm{eig}_1(\boldsymbol{A}) \leq \cdots \leq \mathrm{eig}_p(\boldsymbol{A})$ stand for its ordered eigenvalues. We consider the following norms on $p \times p$ matrices

$$\|\boldsymbol{A}\|_r = \left(\sum_{i=1}^p |a_{ij}|^r\right)^{1/r},\ 1 \leq r < \infty, \quad \|\boldsymbol{A}\|_\infty = \max_{i,j} |a_{ij}|,$$
$$\|\boldsymbol{A}\|_{(r,s)} = \sup\{\|\boldsymbol{A}\boldsymbol{x}\|_s \colon \|\boldsymbol{x}\|_r = 1\},$$

by respectively viewing $\boldsymbol{A}$ as a vector in $\mathbb{R}^{p^2}$ and an operator from $(\mathbb{R}^p, \|\cdot\|_r)$ to $(\mathbb{R}^p, \|\cdot\|_s)$, where $1 \leq r, s \leq \infty$. This gives

$$\|\boldsymbol{A}\|_{(1,1)} = \max_j \textstyle\sum_i |a_{ij}|, \quad \|\boldsymbol{A}\|_{(\infty,\infty)} = \max_i \textstyle\sum_j |a_{ij}|$$
$$\|\boldsymbol{A}\|_{(2,2)} = \{\max(\mathrm{eig}_i(\boldsymbol{A}^T\boldsymbol{A}) \colon 1 \leq i \leq p)\}^{1/2},$$

and that for symmetric matrices, $\|\boldsymbol{A}\|_{(2,2)} = \max\{|\mathrm{eig}_i(\boldsymbol{A})| \colon 1 \leq i \leq p\}$, and $\|\boldsymbol{A}\|_{(1,1)} = \|\boldsymbol{A}\|_{(\infty,\infty)}$. The norm $\|\cdot\|_{(r,r)}$ will be referred to as the L_r-operator norm. For two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we say that $\boldsymbol{A} \geq \boldsymbol{B}$ (respectively, $\boldsymbol{A} > \boldsymbol{B}$) if $\boldsymbol{A} - \boldsymbol{B}$ is nonnegative definite (respectively, positive definite). Thus $\boldsymbol{A} > \boldsymbol{0}$ for a positive definite matrix $\boldsymbol{A}$, where $\boldsymbol{0}$ stands for the zero matrix. The identity matrix of order p will be denoted by $\boldsymbol{I}_p$. A vector of 1's is denoted by $\boldsymbol{1}$.

Sets are denoted by non-bold uppercase English letters. For a set T, we denote the cardinality, that is, the number of elements in T, by $\#T$. We denote the submatrix of the matrix $\boldsymbol{A}$ induced by the set $T \subset \{1, 2, \ldots, p\}$ by $\boldsymbol{A}_T$, i.e., $\boldsymbol{A}_T = (\!(a_{ij} \colon i, j \in T)\!)$. By $\boldsymbol{A}_T^{-1}$, we mean the inverse $(\boldsymbol{A}_T)^{-1}$ of the submatrix $\boldsymbol{A}_T$. For a $p \times p$ matrix $\boldsymbol{A} = (\!(a_{ij})\!)$, let $(\boldsymbol{A}_T)^0 = (\!(a_{ij}^*)\!)$ denote a p-dimensional matrix such that $a_{ij}^* = a_{ij}$ for $(i,j) \in T \times T$, and 0 otherwise. Also we denote the "banded" version of $\boldsymbol{A}$ by $B_k(\boldsymbol{A}) = (\!(a_{ij}\mathbb{1}\{|i-j| \leq k\})\!)$ corresponding to banding parameter k, $k < p$.

2.1. Preliminaries on graph theory

Now we discuss some preliminaries on graph theory and undirected graphical models needed to describe our results. Further details are available in [19, 22].

An undirected graph $G = (V, E)$ consists of a non-empty vertex set $V = \{1, 2, \ldots, p\}$ along with an edge-set $E \subseteq \{(i,j) \in V \times V \colon i < j\}$. Two vertices $v, v' \in V$ are said to be adjacent if there is an edge between v and v'. A graph is called complete if all the vertices are adjacent to each other. A graph $G' = (V', E')$ is a subgraph of $G = (V, E)$, denoted by $G' \subseteq G$ if $V' \subseteq V$ and $E' \subseteq E$. For a subgraph $G' \subseteq G$, if $E' = (V' \times V') \cap E$, then G' is called an induced subgraph of G. For a subset $V' \subseteq V$, we denote the subgraph $G_{V'} = (V', (V' \times V') \cap E)$ to be the graph induced by V'. We shall only consider

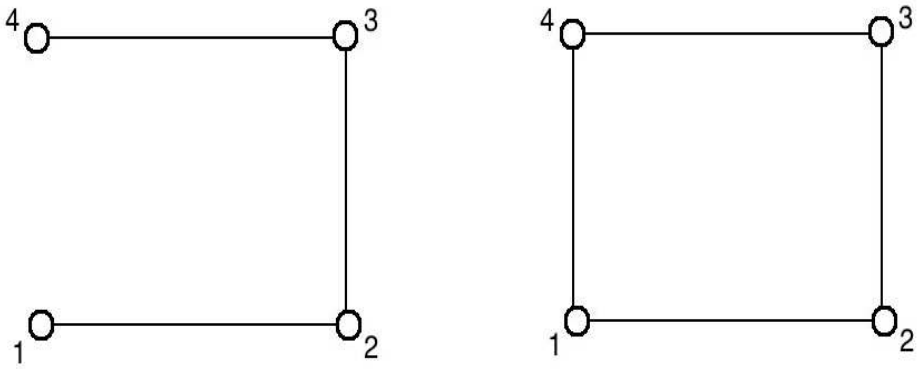


FIG 1. *[Left] An example of a decomposable graph with vertex set* $V = \{1, 2, 3, 4\}$. $\{1, 2\}$, $\{2, 3\}$ *and* $\{3, 4\}$ *are the cliques, whereas* $\{2\}$ *and* $\{3\}$ *are the separators. [Right] A non-decomposable graph with the same vertex set* $V = \{1, 2, 3, 4\}$. *There are four cliques* $\{1, 2\}$, $\{1, 4\}, \{2, 3\}, \{3, 4\}$, *but they cannot be arranged in a perfect order, violating the decomposability condition.*

induced subgraphs henceforth when we refer to subgraphs of a graph. A subset $V' \subseteq V$ is said to be a clique of G if the subgraph $G_{V'}$ is a maximal complete subgraph of G, that is, $G_{V'}$ is not contained in any other complete subgraph of G.

A path in a graph is a finite collection of adjacent edges. If G_1, G_2 and G_3 are subgraphs of G, then G_3 is said to separate G_1 and G_2 if every path from $j \in G_1$ to $k \in G_2$ contains a vertex in G_3. A graph G decomposes in to disjoint subgraphs G_1, G_2 and G_3 if (i) $G_1 \cup G_2 \cup G_3 = G$, (ii) G_3 is complete, and (iii) G_3 separates G_1 and G_2. The decomposition of a graph is proper if neither G_1 nor G_2 is empty. A graph is decomposable if it is complete, or if there exists a proper decomposition (G_1, G_2, G_3) in decomposable subgraphs induced by the vertices in $G_1 \cup G_3$ and $G_2 \cup G_3$. This is a recursive definition which ultimately gives a sequence of cliques and separators of the graph. One of the most important results in this context of decomposability of a graph is that of perfect ordering of cliques. A set of cliques $\mathcal{C}_G = \{C_1, C_2, \ldots, C_r\}$ is said to be in perfect order, if the following holds: For

$$\begin{aligned} H_1 = R_1 = C_1, \quad & H_j = C_1 \cup \cdots \cup C_j, \\ R_j = C_j \setminus H_{j-1}, \quad & S_j = H_{j-1} \cap C_j,\ j = 2, \ldots, r, \end{aligned} \tag{2.1}$$

$\mathcal{S} = \{S_j,\ j = 2, \ldots, p\}$ is the set of minimal separators in G. The sets H_j, R_j and S_j are termed as histories, residuals and separators of the sequence respectively. For a decomposable graph, a perfect order of the cliques always exists. Figure 1 illustrates a decomposable and a non-decomposable graph.

2.2. Undirected Gaussian graphical models

An undirected graph G equipped with a probability distribution P such that the vertex set $V = \{1, \ldots, p\}$ of G corresponds to a p-dimensional random

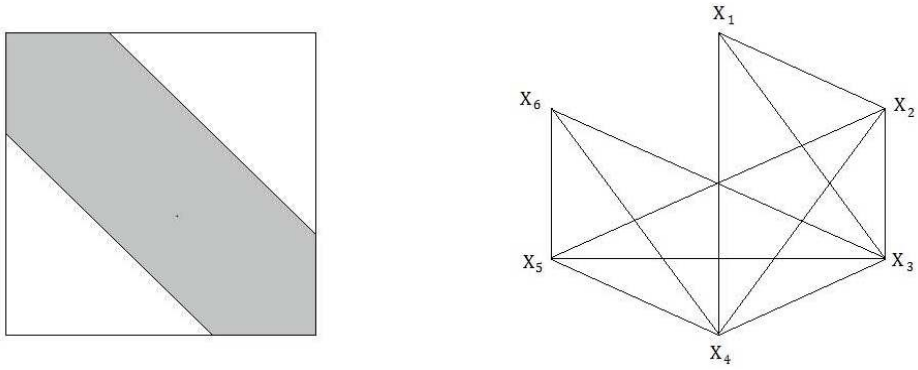


FIG 2. *[Left] Structure of a banded precision matrix with shaded non-zero entries. [Right] The graphical model corresponding to a banded precision matrix of dimension 6 and banding parameter 3.*

variable $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)^T \sim P$, and for any pair $(i, j) \notin E$, $i \neq j$, the random variables X_i and X_j are conditionally independent given all X_k, $k \neq i, j$, is referred to as an undirected graphical model (G, P). The conditional independence property is also called the Markov property of $\boldsymbol{X}$ with respect to the graph G. If $\boldsymbol{X}$ has a multivariate normal distribution, the graphical model is called a Gaussian graphical model (GGM). If $\boldsymbol{X}$ has mean $\mathbf{0}$ (without loss of generality) and positive definite covariance matrix $\boldsymbol{\Sigma}$, then $\boldsymbol{X}$ has Markov property with respect to G if and only if $\omega_{ij} = 0$ for any pair $(i, j) \notin E$, $i \neq j$, where ω_{ij} is the (i, j)th entry of $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$; see [19] for a proof. Thus, for a GGM, absence of an edge between any two vertices is equivalent to a zero entry in $\boldsymbol{\Omega}$. Figure 2 illustrates the connection between a banded precision matrix and the corresponding graphical model. In general in a graphical model for a non-Gaussian vector with finite second moment, for any pair $(i, j) \notin E$, $i \neq j$, we have $\omega_{ij} = 0$, but $\omega_{ij} = 0$ does not imply that $(i, j) \notin E$.

Let us denote the linear space of p-dimensional symmetric matrices by $\mathcal{M}_p$ and $\mathcal{M}_p^+ \subset \mathcal{M}_p$ to be the cone of positive definite matrices of order p. Following the notation in [22], we can restrict the canonical parameter $\boldsymbol{\Omega}$ in $\mathcal{P}_G$, where $\mathcal{P}_G$ is the cone of positive definite symmetric matrices of order p having zero entry corresponding to each pair $(i, j) \notin E$, $i \neq j$, that is,

$$\mathcal{P}_G = \{\boldsymbol{\Omega} = (\!(\omega_{ij})\!) \in \mathcal{M}_p^+ \colon \omega_{ij} = 0 \text{ whenever } (i, j) \notin E, i \neq j\}. \tag{2.2}$$

The linear space of symmetric incomplete matrices $\boldsymbol{A} = (\!(a_{ij})\!)$ with missing entries a_{ij}, $(i, j) \notin E$, will be denoted by $\mathcal{I}_G$. Any such matrix $\boldsymbol{A} \in \mathcal{I}_G$ is said to be partially positive definite over G if for every clique $C \in \mathcal{C}_G$, the corresponding submatrix $\boldsymbol{A}_C$ is positive definite. We denote the cone of partially positive definite matrices by

$$\mathcal{Q}_G = \{\boldsymbol{B} \in \mathcal{I}_G \colon \boldsymbol{B}_{C_i} > \mathbf{0}, C_i \in \mathcal{C}_G\}. \tag{2.3}$$

Gröne et al. [15] proved that there is a bijection between the spaces $\mathcal{P}_G$ and $\mathcal{Q}_G$ for decomposable graphs G. Note that, for $(i,j) \notin E$, the corresponding entries in $\mathbf{\Sigma}$ are not free parameters of the Gaussian model (see [26]). For decomposable graphs, Gröne et al. [15] also showed that for every $\mathbf{\Sigma} \in \mathcal{Q}_G$ there exists a unique positive definite matrix $\mathbf{\Sigma}^*$ such that $\mathbf{\Sigma}^*_{ij} = \mathbf{\Sigma}_{ij}$ for $(i,j) \in E$ such that $\mathbf{\Sigma}^*$ corresponds to the covariance matrix of a Gaussian distribution which is Markov with respect to the underlying graphical structure. Thus when G is decomposable, the parameter space for $\mathbf{\Sigma}$ can be defined by the set of incomplete matrices which are partially positive definite, that is, we can restrict the covariance matrix $\mathbf{\Sigma}$ in $\mathcal{Q}_G \subset \mathcal{I}_G$. More precisely, we can define the parameter space for the GGM as the space of incomplete matrices $\mathbf{\Sigma} \in \mathcal{Q}_G$ such that $\mathbf{\Sigma} = \kappa(\mathbf{\Omega}^{-1})$, $\mathbf{\Omega} \in \mathcal{P}_G$, where $\kappa\colon \mathcal{M}_p \to \mathcal{I}_G$ is the projection of $\mathcal{M}_p$ into $\mathcal{I}_G$.

To give a simple example to illustrate the parameter spaces, consider the decomposable graph as in the left side of Figure 1, such that the underlying random variables $\boldsymbol{X} = (X_1, \ldots, X_4)$ corresponding to the vertices of respective indices follow an autoregressive process of order 1, with covariance between X_i and X_j given by $0.7^{|i-j|}$, for $i = 1, \ldots, 4$, $j = 1, \ldots, 4$. The corresponding precision matrix $\mathbf{\Omega} = (\!(\omega_{ij})\!)$ is given by

$$\mathbf{\Omega} = \begin{pmatrix} 1.961 & -1.373 & 0 & 0 \\ -1.373 & 2.922 & -1.373 & 0 \\ 0 & -1.373 & 2.922 & -1.373 \\ 0 & 0 & -1.373 & 1.961 \end{pmatrix}.$$

Note that in the above example, $\omega_{ij} = 0$ whenever $(i,j) \notin E$, $i \neq j$, E being the edge-set in the underlying graph. For the space of covariance matrices, it is enough to consider the incomplete matrix $\mathbf{\Sigma} = (\!(\sigma_{ij})\!)$, where $\sigma_{ij} = 0.7^{|i-j|}$ for $(i,j) \in E$ and σ_{ij} is a missing entry otherwise. Clearly the submatrices of $\mathbf{\Sigma}$ corresponding to the cliques of the underlying graph are positive definite. This incomplete matrix can be uniquely extended to the positive definite matrix $\mathbf{\Sigma}^* = (\!(\sigma^*_{ij})\!)$, where $\sigma^*_{ij} = 0.7^{|i-j|}$ for all $i, j \in \{1, \ldots, 4\}$.

We shall work only with decomposable Gaussian graphical models in this paper.

3. Model assumption and prior specification

Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ be independent and identically distributed (i.i.d.) random p-vectors with mean $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$. Write $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$, and assume that the $\boldsymbol{X}_i$, $i = 1, \ldots, n$, are multivariate Gaussian. Consistent estimators for the covariance matrix were obtained in [3] by banding the sample covariance matrix, assuming a certain sparsity structure on the true covariance. Our aim is to obtain convergence rates of the graphical MLE and Bayes estimators of the precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ under the condition $n^{-1} \log p \to 0$ where $\mathbf{\Omega}$ ranges over some fairly natural families. For a given positive sequence $\gamma(k) \downarrow 0$,

we consider the class of positive definite symmetric matrices $\mathbf{\Omega} = ((\omega_{ij}))$ as

$$\mathcal{U}(\varepsilon_0, \gamma) = \left\{ \mathbf{\Omega} : \max_j \sum_i \{|\omega_{ij}| : |i-j| > k\} \le \gamma(k) \text{ for all } k > 0, \right. \\ \left. 0 < \varepsilon_0 \le \operatorname{eig}_1(\mathbf{\Omega}) \le \operatorname{eig}_p(\mathbf{\Omega}) \le \varepsilon_0^{-1} < \infty \right\}. \tag{3.1}$$

The sequence $\gamma(k)$ which bounds $\|\mathbf{\Omega} - B_k(\mathbf{\Omega})\|_{(\infty,\infty)} = \max_j \sum_i \{|\omega_{ij}| : |i-j| > k\}$ has been kept flexible so as to include a number of matrix classes.

1. Exact banding: $\gamma(k) = 0$ for all $k \ge k_0$, which means that the true precision matrix is banded, with banding parameter k_0. For instance, any autoregressive process has such a form of precision matrix.
2. Exponential decay: $\gamma(k) = e^{-ck}$. For instance, any moving average process has such a form of precision matrix.
3. Polynomial decay: $\gamma(k) = \gamma k^{-\alpha}$, $\alpha > 0$. This class of matrices was considered in [3, 6].

We shall estimate the precision matrix which belongs $\mathcal{U}(\varepsilon_0, \gamma)$ for some ε_0 and $\gamma(\cdot)$ by restricting the prior on a sieve of banded matrices. A banding structure in the precision matrix can be induced by a Gaussian graphical model. Since $\omega_{ij} = 0$ implies that the components X_i and X_j of $\boldsymbol{X}$ are conditionally independent given the others, we can thus define a Gaussian graphical model $G = (V, E)$, where $V = \{1, \ldots, p\}$ indexing the p components $X_1, X_2, \ldots, X_p$, and E is the corresponding edge set defined by $E = \{(i,j) : 0 < |i-j| \le k\}$, and k is the size of the band. This describes a parameter space for precision matrices consisting of k-banded matrices, and can be used for the maximum likelihood or the Bayesian approach, where for the latter, a prior distribution on these matrices must be specified.

It is not difficult to check that G is an undirected, decomposable graphical model for which a perfect order of cliques exist, given by $\mathcal{C} = \{C_1, C_2, \ldots, C_{p-k}\}$, $C_j = \{j, j+1, \ldots, j+k\}$, $j = 1, 2, \ldots, p-k$. The corresponding separators are given by $\mathcal{S} = \{S_2, S_3, \ldots, S_{p-k}\}$, $S_j = \{j, j+1, \ldots, j+k-1\}$, $j = 2, 3, \ldots, p-k$. The choice of the perfect set of cliques is not unique, but the estimator for the precision matrix $\mathbf{\Omega}$ under all choices of the order remains the same. We shall work with the G-Wishart distribution $W_G(\delta, \boldsymbol{D})$ as a conjugate prior for $\mathbf{\Omega} \in \mathcal{P}_G$. The prior density, derived by [29], is given by

$$p(\mathbf{\Omega}|G) = (I_G(\delta, \boldsymbol{D}))^{-1} (\det(\mathbf{\Omega}))^{(\delta-2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{D}\mathbf{\Omega})\right], \tag{3.2}$$

where $\boldsymbol{D}$ is a symmetric positive definite matrix and

$$I_G(\delta, \boldsymbol{D}) = \int_{\mathbf{\Omega} \in \mathcal{P}_G} (\det(\mathbf{\Omega}))^{(\delta-2)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{D}\mathbf{\Omega})\right] d\mathbf{\Omega} \tag{3.3}$$

is the normalizing constant, which is finite for $\delta > 2$. Letac and Massam [22] introduced a more general family of conjugate priors, called the W_{P_G}-Wishart family, as a prior distribution for $\mathbf{\Omega}$. The W_{P_G}-Wishart distribution $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D})$

has three set of parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{D}$, where $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are suitable functions defined on the cliques and separators of the graph respectively, and $\boldsymbol{D}$ is a scaling matrix. The G-Wishart distribution $W_G(\delta, \boldsymbol{D})$ is a special case of the W_{P_G}-Wishart family where

$$\begin{aligned} \alpha_i &= -\frac{\delta + \#C_i - 1}{2},\ i = 1, 2, \ldots, p-k, \\ \beta_i &= -\frac{\delta + \#S_i - 1}{2},\ i = 2, 3, \ldots, p-k. \end{aligned} \tag{3.4}$$

If the prior distribution on $\frac{1}{2}\boldsymbol{\Omega}$ is $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{D})$, then the posterior distribution of $\frac{1}{2}\boldsymbol{\Omega}$ given the sample covariance $\boldsymbol{S} = n^{-1}\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i^T$ is given by $W_{P_G}(\boldsymbol{\alpha} - \frac{n}{2}\mathbf{1}, \boldsymbol{\beta} - \frac{n}{2}\mathbf{1}, \boldsymbol{D} + \kappa(n\boldsymbol{S}))$. For the G-Wishart distribution in our case, $\#C_i = k+1$ for all $i = 1, 2, \ldots, p-k$, and $\#S_j = k$ for all $j = 2, 3, \ldots, p-k$. Thus

$$\begin{aligned} \alpha_i &= -\frac{\delta + k}{2},\ i = 1, 2, \ldots, p-k, \\ \beta_j &= -\frac{\delta + k - 1}{2},\ j = 2, 3, \ldots, p-k. \end{aligned} \tag{3.5}$$

The posterior mean of $\boldsymbol{\Omega}$ was derived in [26], and is given by

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = -2 \Bigg[&\sum_{j=1}^{p-k} (\alpha_j - \frac{n}{2}) \left((\boldsymbol{D} + \kappa(n\boldsymbol{S}))^{-1}_{C_j} \right)^0 \\ &- \sum_{j=2}^{p-k} (\beta_j - \frac{n}{2}) \left((\boldsymbol{D} + \kappa(n\boldsymbol{S}))^{-1}_{S_j} \right)^0 \Bigg]. \end{aligned} \tag{3.6}$$

Taking $\boldsymbol{D} = \boldsymbol{I}_p$, the p dimensional indicator matrix, and plugging in the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we get the posterior mean with respect to the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$ as,

$$\begin{aligned} \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = \frac{\delta + k + n}{n} \Bigg[&\sum_{j=1}^{p-k} \left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right)^0 \\ &- \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right)^0 \Bigg] + n^{-1} \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right)^0. \end{aligned} \tag{3.7}$$

For a sample of size n from a p-dimensional Gaussian distribution with mean $\mathbf{0}$ and precision matrix $\boldsymbol{\Omega}$, we consider the following two loss functions:

$$\begin{aligned} \text{Stein's loss: } L_1(\widehat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}) &= \frac{1}{2}\mathrm{tr}(\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}) - \log|\widehat{\boldsymbol{\Omega}}\boldsymbol{\Omega}^{-1}| - p, \\ \text{Squared-error loss: } L_2(\widehat{\boldsymbol{\Omega}}, \boldsymbol{\Omega}) &= \mathrm{tr}(\widehat{\boldsymbol{\Omega}} - \boldsymbol{\Omega})^2, \end{aligned} \tag{3.8}$$

for an arbitrary estimator $\widehat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$. The Bayes estimators corresponding to the above two loss functions were derived in [26]. Under the G-Wishart prior

$W_G(\delta, \boldsymbol{I}_p)$, the Bayes estimator $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ corresponding to Stein's loss function is given by

$$\frac{\delta+n-2}{n}\left[\sum_{j=1}^{p-k}\left((n^{-1}\boldsymbol{I}_{k+1}+\boldsymbol{S}_{C_j})^{-1}\right)^0-\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_{k}+\boldsymbol{S}_{S_j})^{-1}\right)^0\right]. \tag{3.9}$$

For the squared-error loss function, the corresponding Bayes estimator is clearly the posterior mean of $\boldsymbol{\Omega}$ given in (3.7). We denote this estimator by $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$. Some other loss functions for estimation of $\boldsymbol{\Omega}$ have also been considered in the literature; see [30].

The graphical MLE for $\boldsymbol{\Omega}$ under the graphical model with banding parameter k is given by (see [19]),

$$\widehat{\boldsymbol{\Omega}}^{\mathrm{M}}=\sum_{j=1}^{p-k}(\boldsymbol{S}^{-1}_{C_j})^0-\sum_{j=2}^{p-k}(\boldsymbol{S}^{-1}_{S_j})^0. \tag{3.10}$$

4. Main results

In this section, we determine the convergence rate of the posterior distribution of the precision matrix. The following theorem describes the behavior of the entire posterior distribution.

Theorem 4.1. *Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be random samples from a p-dimensional Gaussian distribution with mean zero and true precision matrix $\boldsymbol{\Omega}_0\in\mathcal{U}(\varepsilon_0,\gamma)$ for some $\varepsilon_0>0$ and $\gamma(\cdot)$ such that $k^{3/2}\gamma(k)\to 0$ as $k\to\infty$. Suppose that $\boldsymbol{\Omega}$ is given the G-Wishart prior $W_G(\delta,\boldsymbol{I}_p)$, where the graph G has banding of order k. Then posterior distribution of the precision matrix $\boldsymbol{\Omega}$ satisfies*

$$\mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega}-\boldsymbol{\Omega}_0\|_{(\infty,\infty)}>M\epsilon_{n,k}|\boldsymbol{X}_1,\dots,\boldsymbol{X}_n\right\}\right]\to 0 \tag{4.1}$$

for $\epsilon_{n,k}=k^{5/2}(n^{-1}\log p)^{1/2}+k^{3/2}\gamma(k)$ and a sufficiently large constant $M>0$.

In particular, the posterior distribution is consistent in the L_∞-operator norm if $k\to\infty$ such that $k^5 n^{-1}\log p\to 0$.

An important step towards the proof of the above result is to find the convergence rate of the graphical MLE, which is also of independent interest. For high-dimensional situations, even when the sample covariance matrix is singular, the graphical MLE will be positive definite if the number of elements in the cliques of the corresponding graphical model is less than the sample size.

Convergence results for banded empirical covariance (or precision) matrix or estimators based on thresholding approaches are typically given in terms of the L_2-operator norm in the literature. We however use the stronger L_∞-operator norm (or equivalently, L_1-operator norm), so the implication of a convergence rate in our theorems is stronger.

Proposition 4.2. *Let $\boldsymbol{X}_1,\dots,\boldsymbol{X}_n$ be random samples from a p-dimensional Gaussian distribution with mean zero and precision matrix $\boldsymbol{\Omega}_0\in\mathcal{U}(\epsilon_0,\gamma)$ for some $\epsilon_0>0$ and $\gamma(\cdot)$ such that $k^{3/2}\gamma(k)\to 0$ as $k\to\infty$. Then the graphical*

MLE $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ *of* $\mathbf{\Omega}$*, corresponding to the Gaussian graphical model with banding parameter* k*, has convergence rate given by*

$$\|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - \mathbf{\Omega}_0\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2} + k^{3/2}\gamma(k)\right). \tag{4.2}$$

In particular, $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ *is consistent in the* L_∞*-operator norm if* $k \to \infty$ *such that* $k^5 n^{-1}\log p \to 0$*.*

The proof will use the explicit form of the graphical MLE and proceed by bounding the mean squared error of each component and using relations between matrix norms. However, as the graphical MLE involves $(k+1)(p-k/2)$ many terms, a naive approach will lead to a factor p in the estimate, which will not be able to establish a convergence rate in the truly high dimensional situations $p \gg n$. We overcome this obstacle by looking more carefully at the structure of the graphical MLE, and note that for any i, the number of terms in (3.10) which have non-zero ith row is only at most $(2k+1) \ll p$. This along with the description of L_∞-operator norm in terms of row sums give rise to a much smaller factor than p.

Now we treat the Bayes estimators. Consider the G-Wishart prior $W_G(\delta, \boldsymbol{I}_p)$ for $\mathbf{\Omega}$, where the graph G has banding of order k and δ is a positive integer. The following result bounds the difference between $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ and the estimators $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_1}$ and $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_2}$.

Lemma 4.3. *Assume the conditions of Proposition 4.2 and suppose that* $\mathbf{\Omega}$ *is given the* G*-Wishart prior* $W_G(\delta, \boldsymbol{I}_p)$*, where the graph* G *has banding of order* k*. Then* $\|\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_1} - \widehat{\mathbf{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P(k^2/n)$*,* $\|\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_2} - \widehat{\mathbf{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P(k^{5/2}/n)$*.*

The proof of the above lemma is given in the Section 7. Proposition 4.2 and Lemma 4.3 together lead to the following result for the convergence rate of the Bayes estimators under the G in the L_∞-operator norm.

Proposition 4.4. *In the setting of Lemma 4.3, for* $\widehat{\mathbf{\Omega}}^{\mathrm{B}}$ *either* $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_1}$ *or* $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_2}$*, we have*

$$\|\widehat{\mathbf{\Omega}}^{\mathrm{B}} - \mathbf{\Omega}_0\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2} + k^{3/2}\gamma(k)\right). \tag{4.3}$$

In particular, the Bayes estimators $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_1}$ *and* $\widehat{\mathbf{\Omega}}^{\mathrm{B}}_{L_2}$ *are consistent in the* L_∞*-operator norm if* $k \to \infty$ *such that* $k^5 n^{-1}\log p \to 0$*.*

Remarks on the convergence rates. Observe that the convergence rates of the graphical MLE, the Bayes estimators and the posterior distribution obtained above are the same. The obtained rates can be optimized by choosing k appropriately as in a bias-variance trade-off. The fastest possible rates obtained from the theorems may be summarized for the different decay rates of $\gamma(k)$ as follows: If the true precision matrix is banded with banding parameter k_0, then the optimal rate of convergence $n^{-1/2}(\log p)^{1/2}$ is obtained by choosing any fixed $k \geq k_0$. When $\gamma(k)$ decays exponentially, the rate of convergence $n^{-1/2}(\log p)^{1/2}(\log n)^{5/2}$ can be obtained by choosing k approximately proportional to $\log n$ with some sufficiently large constant of proportionality. If $\gamma(k)$

decays polynomially with index $\alpha > 3/2$ as in [3], we get the convergence rate of $(n^{-1}\log p)^{(2\alpha-3)/(4\alpha+4)}$ corresponding $k \asymp (n/\log p)^{1/(2\alpha+2)}$.

It is to be noted that we have not assumed that the true structure of the precision matrix arises from a graphical model. The graphical model is a convenient tool to generate useful estimators through the maximum likelihood and Bayesian approach, but the graphical model itself may be a misspecified model. Further, it can be inspected from the proof of the theorems that the Gaussianity assumption on true distribution of the observations is not essential, although the graphical model assumes Gaussianity to generate estimators. The Gaussianity assumption is used to control certain probabilities by applying the probability inequality Lemma A.3 of [3]. However, it was also observed in [3] that one only requires bounds on the moment generating function of X_i^2, $i = 1, \ldots, p$. In particular, any thinner tailed distribution, such as one with a bounded support, will allow the arguments to go through.

4.1. Estimation using a reference prior

A reference prior for the covariance matrix $\boldsymbol{\Sigma}$, obtained in [26], can also be used to induce a prior on $\boldsymbol{\Omega}$. This corresponds to an improper $W_{P_G}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{0})$ distribution for $\frac{1}{2}\boldsymbol{\Omega}$ with

$$\begin{aligned} &\alpha_i = 0, \qquad i = 1, 2, \ldots, r, \\ &\beta_2 = \frac{1}{2}(c_1 + c_2) - s_2,\ \beta_j = \frac{1}{2}(c_j - s_j), \qquad j = 2, 3, \ldots, r. \end{aligned} \tag{4.4}$$

By Corollary 4.1 in [26], the posterior mean $\widehat{\boldsymbol{\Omega}}^{\mathrm{R}}$ of the precision matrix is given by

$$\begin{aligned} \widehat{\boldsymbol{\Omega}}^{\mathrm{R}} = &\sum_{j=1}^{r} (\boldsymbol{S}_{C_j}^{-1})^0 - \{1 - n^{-1}(c_1 + c_2 - 2s_2)\}(\boldsymbol{S}_{S_2}^{-1})^0 \\ &- \sum_{j=3}^{r} \{1 - n^{-1}(c_j - s_j)\}(\boldsymbol{S}_{S_j}^{-1})^0. \end{aligned} \tag{4.5}$$

Similar to the conclusion of Lemma 4.3, using the reference prior, the L_∞-operator norm of the difference between the Bayes estimator $\widehat{\boldsymbol{\Omega}}^{\mathrm{R}}$ and the graphical MLE $\widehat{\boldsymbol{\Omega}}^{\mathrm{M}}$ satisfies

$$\|\widehat{\boldsymbol{\Omega}}^{\mathrm{R}} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} = O_P\left(k^2/n\right). \tag{4.6}$$

A sketch of the proof is given in Section 7.

5. Estimation of banding parameter

In this section, we propose a method of selecting the banding parameter k of the graphical model using the marginal posterior probabilities of the graph induced by banding k, $k = 1, 2, \ldots$. For the G-Wishart prior $W_G(\delta, \boldsymbol{D})$ for $\boldsymbol{\Omega}$, the posterior is given by $W_G(\delta + n, \boldsymbol{D} + n\boldsymbol{S})$. We can get the marginal likelihood

for G as

$$p(\boldsymbol{X}|G) = (2\pi)^{-np/2} \frac{I_G(\delta+n, \boldsymbol{D}+n\boldsymbol{S})}{I_G(\delta, \boldsymbol{D})}, \tag{5.1}$$

where $I_G(\delta, \boldsymbol{D})$ is the normalizing constant in the density (3.2) of $W_G(\delta, \boldsymbol{D})$, given by (3.3). For a complete graph G, the expression for $I_G(\delta, \boldsymbol{D})$ is given by

$$I_G(\delta, \boldsymbol{D}) = \frac{2^{(\delta+p-1)p/2} \pi^{p(p-1)/4} \prod_{i=0}^{p-1} \Gamma\left(\frac{\delta+p-1-i}{2}\right)}{(\det(\boldsymbol{D}))^{\frac{\delta+p-1}{2}}}; \tag{5.2}$$

see [24]. Roverato [29] showed that for a decomposable graph G,

$$I_G(\delta, \boldsymbol{D}) = \frac{\prod_{j=1}^{r} I_{C_j}(\delta, \boldsymbol{D}_{C_j})}{\prod_{j=2}^{r} I_{S_j}(\delta, \boldsymbol{D}_{S_j})}, \tag{5.3}$$

where $\{C_1, \ldots, C_r\}$ and $\{S_2, \ldots, S_r\}$ denote the set of cliques and separators respectively corresponding to G.

In our case, the model which is fit has a banded structure in the precision matrix. The graphs associated with the banding structure are linearly indexed by the banding parameter k. We denote the graphical model induced by banding parameter k by G^k. Let ρ_k be a prior on k. Then the posterior distribution of k is given by

$$p(k \mid \boldsymbol{X}) = \frac{J_{G^k}(\delta, n, \boldsymbol{D}, n\boldsymbol{S})\rho_k}{\sum_{k'} J_{G^{k'}}(\delta, n, \boldsymbol{D}, n\boldsymbol{S})\rho_{k'}}, \tag{5.4}$$

where for a graph G

$$J_G(\delta, n, \boldsymbol{D}, n\boldsymbol{S}) = \frac{I_G(\delta+n, \boldsymbol{D}+n\boldsymbol{S})}{I_G(\delta, \boldsymbol{D})}. \tag{5.5}$$

Let the cliques and separators be respectively denoted by $C_j^k = \{j, j+1, \ldots, j+k\}$, $j = 1, \ldots, p-k$, and $S_j^k = \{j, j+1, \ldots, j+k-1\}$, $j = 2, \ldots, p-k$. Note that the sub-graphs corresponding to the cliques and separators are complete, with respective dimensions $k+1$ and k, and $r = p-k$. Therefore (5.2) and (5.3) together reduces the expression for $I_{G^k}(\delta, \boldsymbol{D})$ to

$$\frac{\prod_{j=1}^{p-k} 2^{(\delta+k)(k+1)/2} \pi^{k(k+1)/4} \prod_{i=0}^{k} \Gamma\left(\frac{\delta+k-i}{2}\right)}{\prod_{j=2}^{p-k} 2^{(\delta+k-1)k/2} \pi^{k(k-1)/4} \prod_{i=0}^{k-1} \Gamma\left(\frac{\delta+k-1-i}{2}\right)} \frac{(\det(\boldsymbol{D}_{S_j}))^{(\delta+k-1)/2}}{(\det(\boldsymbol{D}_{C_j}))^{(\delta+k)/2}}. \tag{5.6}$$

Now, with the choice $\boldsymbol{D} = \boldsymbol{I}_p$ used in the prior $W_{G^k}(\delta, \boldsymbol{I})$, (5.5) gives

$$\begin{aligned} J_{G^k}(\delta, n, \boldsymbol{I}_p, n\boldsymbol{S}) \quad = \quad & \frac{\prod_{j=1}^{p-k} 2^{n(k+1)/2}}{\prod_{j=2}^{p-k} 2^{nk/2}} \left(\prod_{i=0}^{k} \frac{\Gamma\left(\frac{\delta+n+i}{2}\right)}{\Gamma\left(\frac{\delta+i}{2}\right)} \right) \left(\frac{\Gamma\left(\frac{\delta+n+k}{2}\right)}{\Gamma\left(\frac{\delta+k}{2}\right)} \right)^{p-k-1} \\ & \times \frac{\prod_{j=2}^{p-k} (\det((\boldsymbol{I}_p + n\boldsymbol{S})_{S_j^k}))^{(\delta+k+n-1)/2}}{\prod_{j=1}^{p-k} (\det((\boldsymbol{I}_p + n\boldsymbol{S})_{C_j^k}))^{(\delta+n+k)/2}}. \end{aligned} \tag{5.7}$$

Substituting this expression in (5.4), we get an explicit expression for the posterior distribution of k.

A natural method of selecting k is to consider the posterior mode. In the next section, we investigate the performance of the posterior mode of G^k through a simulation study.

6. Numerical results

We check the performance of the Bayes estimators of the precision matrix and compare with the graphical MLE and the banded estimator as proposed in [3].

Data is simulated from $\mathrm{N}_p(0, \mathbf{\Sigma})$, assuming specific structures of the covariance $\mathbf{\Sigma}$ or the precision $\mathbf{\Omega}$. For all simulations, we compute the L_∞-operator norm, L_2-operator norm, L_2-norm and L_∞-norm of the difference between the estimate and the true parameter for sample sizes $n = 100, 200, 500$ and $p = 50, 100, 200, 500$, representing cases like $p < n$, $p \sim n$, $p > n$ and $p \gg n$. We simulate 100 replications in each cases. Some of the simulation models are the same as those in [3].

Example 6.1 (Autoregressive process: AR(1) covariance structure)**.** Let the true covariance matrix have entries given by

$$\sigma_{ij} = \rho^{|i-j|},\ 1 \le i, j \le p, \tag{6.1}$$

with $\rho = 0.3$ in our simulation experiment. The precision matrix is banded in this case, with banding parameter 1.

Example 6.2 (Autoregressive process: AR(4) covariance structure)**.** The elements of true precision matrix are given by

$$\begin{aligned}\omega_{ij} = {} & \mathbb{1}(|i-j| = 0) + 0.4\,\mathbb{1}(|i-j| = 1) + 0.2\,\mathbb{1}(|i-j| = 2) \\ & + 0.2\,\mathbb{1}(|i-j| = 3) + 0.1\,\mathbb{1}(|i-j| = 4).\end{aligned} \tag{6.2}$$

This precision matrix corresponds to an AR(4) process.

Example 6.3 (Long range dependence)**.** We consider a fractional Gaussian Noise process, that is, the increment process of fractional Brownian motion. The elements of the true covariance matrix are given by

$$\sigma_{ij} = \frac{1}{2}[||i-j|+1|^{2H} - 2|i-j|^{2H} + ||i-j|-1|^{2H}],\ 1 \le i, j \le p, \tag{6.3}$$

where $H \in [0.5, 1]$ is the Hurst parameter. We take $H = 0.7$ in the simulation example. This precision matrix does not fall in the polynomial smoothness class used in the theorems. We include this example in the simulation study to check how the proposed method is performing when the assumptions of the theorems are not met.

Tables 1–3 show the simulation results for the different scenarios and compare the performance of the Bayes estimators with the graphical MLE and the

TABLE 1. *Simulation results for AR(1) model based on 100 replications; figures in parentheses indicate standard errors*

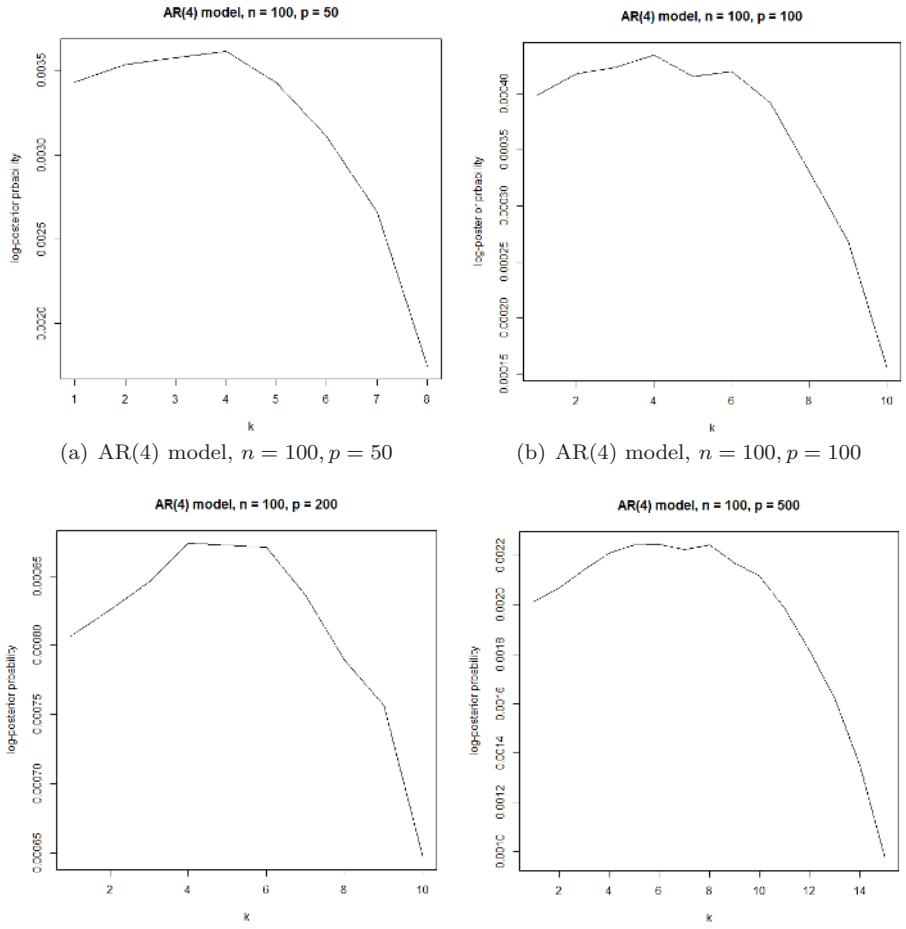
		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.252	1.295	1.175	1.249	0.799	0.820	0.773	0.797	0.477	0.485	0.470	0.477
		(0.029)	(0.029)	(0.027)	(0.029)	(0.018)	(0.018)	(0.017)	(0.018)	(0.009)	(0.009)	(0.008)	(0.008)
	$L_{2,2}$	1.003	1.044	0.940	0.999	0.644	0.663	0.623	0.642	0.374	0.381	0.368	0.374
		(0.029)	(0.023)	(0.021)	(0.023)	(0.016)	(0.016)	(0.015)	(0.016)	(0.007)	(0.007)	(0.007)	(0.007)
50	L_2	2.374	2.454	2.275	2.366	1.609	1.643	1.575	1.607	0.976	0.986	0.968	0.975
		(0.026)	(0.027)	(0.023)	(0.026)	(0.017)	(0.017)	(0.016)	(0.016)	(0.008)	(0.008)	(0.008)	(0.008)
	L_∞	0.729	0.767	0.688	0.726	0.450	0.468	0.437	0.450	0.272	0.278	0.268	0.272
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.010)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
	$L_{\infty,\infty}$	1.378	1.420	1.295	1.374	0.889	0.912	0.861	0.889	0.525	0.534	0.516	0.525
		(0.029)	(0.029)	(0.027)	(0.029)	(0.018)	(0.018)	(0.017)	(0.018)	(0.009)	(0.009)	(0.009)	(0.009)
	$L_{2,2}$	1.112	1.152	1.042	1.107	0.712	0.734	0.687	0.711	0.408	0.416	0.401	0.408
		(0.022)	(0.022)	(0.021)	(0.023)	(0.015)	(0.015)	(0.015)	(0.015)	(0.007)	(0.007)	(0.007)	(0.007)
100	L_2	3.365	3.482	3.223	3.354	2.264	2.310	2.217	2.263	1.383	1.397	1.371	1.382
		(0.027)	(0.030)	(0.024)	(0.027)	(0.016)	(0.017)	(0.015)	(0.016)	(0.010)	(0.010)	(0.009)	(0.010)
	L_∞	0.814	0.852	0.768	0.808	0.500	0.518	0.484	0.499	0.298	0.305	0.293	0.298
		(0.018)	(0.018)	(0.017)	(0.018)	(0.010)	(0.011)	(0.010)	(0.010)	(0.005)	(0.005)	(0.005)	(0.005)
	$L_{\infty,\infty}$	1.558	1.602	1.463	1.557	1.002	1.027	0.967	1.001	0.582	0.593	0.572	0.582
		(0.028)	(0.028)	(0.027)	(0.029)	(0.019)	(0.019)	(0.019)	(0.019)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.237	1.276	1.160	1.233	0.791	0.814	0.763	0.790	0.453	0.463	0.445	0.453
		(0.022)	(0.021)	(0.021)	(0.022)	(0.015)	(0.015)	(0.015)	(0.015)	(0.007)	(0.007)	(0.007)	(0.007)
200	L_2	4.750	4.915	4.548	4.738	3.211	3.277	3.143	3.209	1.971	1.987	1.955	1.970
		(0.024)	(0.026)	(0.022)	(0.025)	(0.017)	(0.018)	(0.017)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)
	L_∞	0.923	0.961	0.872	0.918	0.564	0.584	0.545	0.564	0.327	0.334	0.321	0.326
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)	(0.006)	(0.006)	(0.006)	(0.006)
	$L_{\infty,\infty}$	1.765	1.805	1.657	1.763	1.109	1.134	1.069	1.108	0.642	0.653	0.631	0.643
		(0.028)	(0.028)	(0.027)	(0.028)	(0.017)	(0.017)	(0.017)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.407	1.443	1.321	1.406	0.887	0.909	0.856	0.886	0.504	0.514	0.495	0.504
		(0.022)	(0.022)	(0.021)	(0.022)	(0.014)	(0.013)	(0.013)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)
500	L_2	7.527	7.783	7.209	7.505	5.079	5.177	4.975	5.074	3.133	3.160	3.107	3.131
		(0.029)	(0.030)	(0.026)	(0.029)	(0.015)	(0.016)	(0.015)	(0.015)	(0.010)	(0.010)	(0.009)	(0.010)
	L_∞	1.030	1.066	0.974	1.028	0.626	0.646	0.606	0.625	0.359	0.367	0.354	0.359
		(0.017)	(0.017)	(0.016)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)	(0.006)	(0.006)	(0.006)	(0.006)

TABLE 2. *Simulation results for AR(4) model based on 100 replications; figures in parentheses indicate standard errors*

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.836	2.066	1.758	1.821	1.078	1.177	1.053	1.076	0.642	0.673	0.636	0.641
		(0.040)	(0.041)	(0.038)	(0.038)	(0.020)	(0.021)	(0.019)	(0.020)	(0.011)	(0.011)	(0.011)	(0.011)
	$L_{2,2}$	1.158	1.340	1.101	1.149	0.672	0.754	0.654	0.672	0.399	0.426	0.394	0.399
		(0.027)	(0.028)	(0.025)	(0.025)	(0.014)	(0.015)	(0.014)	(0.014)	(0.008)	(0.009)	(0.008)	(0.008)
50	L_2	2.539	2.951	2.463	2.526	1.635	1.789	1.612	1.631	0.988	1.030	0.983	0.987
		(0.027)	(0.030)	(0.025)	(0.026)	(0.015)	(0.018)	(0.014)	(0.015)	(0.008)	(0.009)	(0.008)	(0.008)
	L_∞	0.574	0.692	0.554	0.572	0.326	0.378	0.320	0.325	0.180	0.196	0.179	0.180
		(0.014)	(0.015)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)	(0.003)	(0.004)	(0.003)	(0.004)
	$L_{\infty,\infty}$	1.993	2.231	1.907	1.973	1.210	1.315	1.182	1.209	0.702	0.738	0.694	0.702
		(0.037)	(0.038)	(0.035)	(0.036)	(0.018)	(0.019)	(0.017)	(0.018)	(0.010)	(0.010)	(0.010)	(0.010)
	$L_{2,2}$	1.263	1.451	1.200	1.255	0.761	0.849	0.738	0.760	0.440	0.471	0.434	0.440
		(0.025)	(0.025)	(0.023)	(0.024)	(0.012)	(0.013)	(0.012)	(0.012)	(0.008)	(0.008)	(0.007)	(0.008)
100	L_2	3.626	4.220	3.518	3.607	2.337	2.561	2.303	2.331	1.408	1.466	1.400	1.407
		(0.028)	(0.032)	(0.026)	(0.028)	(0.014)	(0.016)	(0.013)	(0.014)	(0.008)	(0.009)	(0.008)	(0.008)
	L_∞	0.626	0.749	0.605	0.625	0.357	0.411	0.351	0.356	0.196	0.215	0.194	0.196
		(0.015)	(0.015)	(0.014)	(0.014)	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.003)	(0.003)	(0.003)
	$L_{\infty,\infty}$	2.165	2.413	2.069	2.145	1.324	1.435	1.292	1.319	0.763	0.802	0.754	0.762
		(0.034)	(0.035)	(0.032)	(0.033)	(0.018)	(0.019)	(0.018)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)
	$L_{2,2}$	1.376	1.569	1.307	1.363	0.841	0.932	0.816	0.838	0.479	0.512	0.472	0.478
		(0.022)	(0.022)	(0.021)	(0.021)	(0.013)	(0.013)	(0.012)	(0.013)	(0.008)	(0.008)	(0.008)	(0.008)
200	L_2	5.145	5.988	4.992	5.116	3.332	3.652	3.283	3.324	1.995	2.079	1.984	1.994
		(0.028)	(0.032)	(0.026)	(0.028)	(0.015)	(0.017)	(0.014)	(0.015)	(0.007)	(0.008)	(0.007)	(0.007)
	L_∞	0.695	0.821	0.671	0.689	0.393	0.449	0.386	0.393	0.215	0.235	0.213	0.215
		(0.013)	(0.014)	(0.013)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)	(0.003)	(0.004)	(0.003)	(0.003)
	$L_{\infty,\infty}$	2.476	2.732	2.362	2.447	1.482	1.599	1.444	1.480	0.833	0.875	0.823	0.830
		(0.035)	(0.036)	(0.034)	(0.034)	(0.018)	(0.018)	(0.018)	(0.019)	(0.010)	(0.011)	(0.010)	(0.010)
	$L_{2,2}$	1.579	1.778	1.498	1.562	0.946	1.039	0.918	0.945	0.526	0.561	0.518	0.524
		(0.023)	(0.024)	(0.022)	(0.023)	(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.006)	(0.006)
500	L_2	8.205	9.553	7.957	8.159	5.287	5.793	5.210	5.273	3.161	3.296	3.143	3.158
		(0.026)	(0.030)	(0.024)	(0.026)	(0.016)	(0.018)	(0.015)	(0.016)	(0.007)	(0.008)	(0.006)	(0.007)
	L_∞	0.787	0.916	0.759	0.779	0.434	0.491	0.426	0.434	0.244	0.265	0.242	0.244
		(0.013)	(0.013)	(0.012)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)	(0.004)	(0.004)	(0.004)	(0.004)

TABLE 3. *Simulation results for Fractional Gaussian noise model based on 100 replications; figures in parentheses indicate standard errors*

		$n = 100$				$n = 200$				$n = 500$			
p	Norm	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky	MLE	$\mathbf{\Omega}^{\mathrm{B}}_{L_2}$	$\mathbf{\Omega}^{\mathrm{B}}_{L_1}$	Cholesky
	$L_{\infty,\infty}$	1.530	1.588	1.493	1.527	1.184	1.213	1.170	1.182	0.969	0.981	0.965	0.969
		(0.025)	(0.025)	(0.024)	(0.024)	(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)
	$L_{2,2}$	0.849	0.902	0.820	0.846	0.587	0.614	0.577	0.586	0.412	0.422	0.409	0.412
		(0.019)	(0.019)	(0.018)	(0.019)	(0.012)	(0.012)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
50	L_2	2.169	2.271	2.116	2.164	1.666	1.706	1.646	1.665	1.329	1.340	1.323	1.329
		(0.020)	(0.022)	(0.020)	(0.020)	(0.012)	(0.013)	(0.012)	(0.012)	(0.006)	(0.006)	(0.006)	(0.006)
	L_∞	0.570	0.615	0.552	0.569	0.360	0.376	0.355	0.359	0.221	0.226	0.219	0.221
		(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)	(0.003)	(0.004)	(0.003)	(0.003)
	$L_{\infty,\infty}$	1.687	1.745	1.647	1.682	1.316	1.345	1.301	1.315	1.058	1.069	1.053	1.058
		(0.024)	(0.024)	(0.023)	(0.024)	(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)
	$L_{2,2}$	0.939	0.992	0.907	0.937	0.645	0.672	0.633	0.645	0.441	0.451	0.438	0.441
		(0.018)	(0.018)	(0.017)	(0.018)	(0.011)	(0.011)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
100	L_2	3.076	3.222	3.000	3.068	2.351	2.406	2.322	2.351	1.886	1.902	1.876	1.885
		(0.021)	(0.023)	(0.020)	(0.022)	(0.013)	(0.014)	(0.012)	(0.013)	(0.006)	(0.007)	(0.006)	(0.006)
	L_∞	0.634	0.680	0.614	0.632	0.395	0.413	0.388	0.394	0.238	0.243	0.236	0.238
		(0.015)	(0.016)	(0.015)	(0.015)	(0.008)	(0.008)	(0.007)	(0.008)	(0.003)	(0.003)	(0.003)	(0.003)
	$L_{\infty,\infty}$	1.855	1.914	1.811	1.857	1.446	1.475	1.430	1.446	1.134	1.145	1.129	1.134
		(0.022)	(0.022)	(0.021)	(0.022)	(0.014)	(0.014)	(0.014)	(0.014)	(0.008)	(0.008)	(0.008)	(0.008)
	$L_{2,2}$	1.024	1.078	0.989	1.025	0.705	0.733	0.693	0.705	0.471	0.481	0.468	0.471
		(0.016)	(0.016)	(0.016)	(0.016)	(0.011)	(0.011)	(0.011)	(0.011)	(0.005)	(0.005)	(0.005)	(0.005)
200	L_2	4.348	4.555	4.239	4.340	3.335	3.414	3.294	3.334	2.659	2.680	2.645	2.659
		(0.020)	(0.021)	(0.019)	(0.020)	(0.013)	(0.014)	(0.013)	(0.013)	(0.006)	(0.006)	(0.006)	(0.006)
	L_∞	0.712	0.760	0.689	0.711	0.439	0.460	0.431	0.439	0.257	0.262	0.255	0.256
		(0.015)	(0.015)	(0.014)	(0.015)	(0.008)	(0.008)	(0.008)	(0.008)	(0.003)	(0.004)	(0.003)	(0.004)
	$L_{\infty,\infty}$	2.059	2.116	2.008	2.056	1.565	1.595	1.548	1.566	1.219	1.231	1.214	1.219
		(0.022)	(0.022)	(0.022)	(0.022)	(0.012)	(0.012)	(0.012)	(0.013)	(0.007)	(0.007)	(0.007)	(0.007)
	$L_{2,2}$	1.156	1.208	1.116	1.151	0.773	0.800	0.759	0.773	0.507	0.517	0.504	0.507
		(0.017)	(0.016)	(0.016)	(0.017)	(0.010)	(0.010)	(0.010)	(0.010)	(0.004)	(0.005)	(0.004)	(0.005)
500	L_2	6.865	7.190	6.694	6.850	5.266	5.386	5.201	5.263	4.215	4.249	4.194	4.215
		(0.022)	(0.023)	(0.021)	(0.022)	(0.012)	(0.013)	(0.012)	(0.012)	(0.007)	(0.007)	(0.007)	(0.007)
	L_∞	0.802	0.851	0.776	0.801	0.482	0.505	0.474	0.482	0.282	0.288	0.280	0.282
		(0.014)	(0.014)	(0.014)	(0.014)	(0.007)	(0.007)	(0.007)	(0.007)	(0.004)	(0.004)	(0.004)	(0.004)



(a) AR(4) model, $n = 100, p = 50$ (b) AR(4) model, $n = 100, p = 100$

(c) AR(4) model, $n = 100, p = 200$ (d) AR(4) model, $n = 100, p = 500$

FIG 3. *Figures showing log-posterior probabilities of graphs corresponding to different banding parameters k. The graphs are trimmed for larger values of k as the log-posterior probabilities decay further.*

banded estimator obtained in [3] based on a modified Cholesky decomposition. The banding parameter k is chosen using the bandwidth estimation method discussed in Section 5. Figure 3 shows the log-posterior probabilities of the graphs corresponding to banding parameter k for prior distribution of k given by $\rho_k \propto \exp(-k^4)$.

7. Proofs

In this section we provide the proofs of the theorems and lemmas stated in Section 4. Proofs of these results will require some additional lemmas and propositions, which we include in the Appendix.

Proof of Proposition 4.2. Let the true covariance matrix be denoted by $\mathbf{\Sigma}_0$, that is, $\mathbf{\Omega}_0 = \mathbf{\Sigma}_0^{-1}$ and $\mathbf{\Gamma}_0$ is the inverse of the banded version of $\mathbf{\Omega}_0$, that is $\mathbf{\Gamma}_0 = (B_k(\mathbf{\Omega}_0))^{-1}$. This is well defined for sufficiently large k since $B_k(\mathbf{\Omega}_0)$ is close to the non-singular matrix $\mathbf{\Omega}_0$ (even in L_∞ operator norm). The L_∞-operator norm of the difference between the graphical MLE $\widehat{\mathbf{\Omega}}^{\mathrm{M}}$ and the true precision matrix $\mathbf{\Omega}_0$ can be written as

$$\|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - \mathbf{\Omega}_0\|_{(\infty,\infty)} \leq \|\widehat{\mathbf{\Omega}}^{\mathrm{M}} - B_k(\mathbf{\Omega}_0)\|_{(\infty,\infty)} + \|\mathbf{\Omega}_0 - B_k(\mathbf{\Omega}_0)\|_{(\infty,\infty)}. \tag{7.1}$$

As shown in [19], in a decomposable Gaussian graphical model with precision matrix $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$, we have,

$$\sum_{j=1}^{p-k} \left(\mathbf{\Omega}_{C_j}\right)^0 - \sum_{j=2}^{p-k} \left(\mathbf{\Omega}_{S_j}\right)^0 = \sum_{j=1}^{p-k} \left(\mathbf{\Sigma}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k} \left(\mathbf{\Sigma}_{S_j}^{-1}\right)^0.$$

In our case we shall be working with $B_k(\mathbf{\Omega}_0)$ and the corresponding inverse matrix $\mathbf{\Gamma}_0$. Using the above representation and also the form of the graphical MLE, the first term on the right side of the bound in equation (7.1) can be written as,

$$\begin{aligned}
&\left\|\sum_{j=1}^{p-k}\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k}\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \sum_{j=1}^{p-k}\left(\mathbf{\Omega}_{0,C_j}\right)^0 + \sum_{j=2}^{p-k}\left(\mathbf{\Omega}_{0,S_j}\right)^0\right\|_{(\infty,\infty)} \\
&= \left\|\sum_{j=1}^{p-k}\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k}\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \sum_{j=1}^{p-k}\left(\mathbf{\Gamma}_{0,C_j}^{-1}\right)^0 + \sum_{j=2}^{p-k}\left(\mathbf{\Gamma}_{0,S_j}^{-1}\right)^0\right\|_{(\infty,\infty)} \\
&\leq \left\|\sum_{j=1}^{p-k}\left(\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\mathbf{\Gamma}_{0,C_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} + \left\|\sum_{j=2}^{p-k}\left(\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \left(\mathbf{\Gamma}_{0,S_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)}.
\end{aligned}$$

Using the fact that there are only $(2k+1)$ terms in above expressions inside the norms which have a given row non-zero, it follows that

$$\begin{aligned}
&\left\|\sum_{j=1}^{p-k}\left\{\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\mathbf{\Gamma}_{0,C_j}^{-1}\right)^0\right\}\right\|_{(\infty,\infty)} \\
&\quad= \max_l \sum_{l'}\left|\left[\sum_{j=1}^{p-k}\left\{\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\mathbf{\Gamma}_{0,C_j}^{-1}\right)^0\right\}\right]_{(l,l')}\right| \\
&\quad\leq \max_l \sum_{j=1}^{p-k}\sum_{l'}\left|\left[\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\mathbf{\Gamma}_{0,C_j}^{-1}\right)^0\right]_{(l,l')}\right| \\
&\quad\leq (2k+1)\max_j \max_l \sum_{l'}\left|\left[\boldsymbol{S}_{C_j}^{-1} - \mathbf{\Gamma}_{0,C_j}^{-1}\right]_{(l,l')}\right| \\
&\quad= (2k+1)\max_j \left\|\boldsymbol{S}_{C_j}^{-1} - \mathbf{\Gamma}_{0,C_j}^{-1}\right\|_{(\infty,\infty)}
\end{aligned}$$

$$\lesssim k^{3/2} \max_j \left\| \boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Gamma}_{0,C_j}^{-1} \right\|_{(2,2)}, \tag{7.2}$$

where the subscript (l, l') on the matrices above stand for their respective (l, l')th entries. Using the multiplicative inequality $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|\|\boldsymbol{B}\|$ of operator norms, we have

$$\begin{aligned} &\max_j \left\| \boldsymbol{S}_{C_j}^{-1} - \boldsymbol{\Gamma}_{0,C_j}^{-1} \right\|_{(2,2)} \\ &\quad = \max_j \left\| \boldsymbol{\Gamma}_{0,C_j}^{-1} (\boldsymbol{\Gamma}_{0,C_j} - \boldsymbol{S}_{C_j}) \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \\ &\quad \leq \max_j \left\{ \left\| \boldsymbol{\Gamma}_{0,C_j}^{-1} \right\|_{(2,2)} \left\| \boldsymbol{\Gamma}_{0,C_j} - \boldsymbol{S}_{C_j} \right\|_{(2,2)} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \right\}. \end{aligned} \tag{7.3}$$

The middle term $\|\boldsymbol{\Gamma}_{0,C_j} - \boldsymbol{S}_{C_j}\|_{(2,2)}$ in the above expression is bounded by

$$\left\| \boldsymbol{\Sigma}_{0,C_j} - \boldsymbol{S}_{C_j} \right\|_{(2,2)} + \left\| \boldsymbol{\Gamma}_{0,C_j} - \boldsymbol{\Sigma}_{0,C_j} \right\|_{(2,2)}. \tag{7.4}$$

We now estimate the norm difference between the matrices $\boldsymbol{\Gamma}_0$ and $\boldsymbol{\Sigma}_0$. We have,

$$\begin{aligned} \|\boldsymbol{\Gamma}_0 - \boldsymbol{\Sigma}_0\|_{(2,2)} &= \left\| \boldsymbol{\Gamma}_0 (\boldsymbol{\Sigma}_0^{-1} - \boldsymbol{\Gamma}_0^{-1}) \boldsymbol{\Sigma}_0 \right\|_{(2,2)} \\ &= \left\| \boldsymbol{\Gamma}_0 \left(\boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0) \right) \boldsymbol{\Sigma}_0 \right\|_{(2,2)} \\ &\leq \|\boldsymbol{\Gamma}_0\|_{(2,2)} \left\| \boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0) \right\|_{(2,2)} \|\boldsymbol{\Sigma}_0\|_{(2,2)}. \end{aligned} \tag{7.5}$$

Now, $\|\boldsymbol{\Sigma}_0\|_{(2,2)} = (\mathrm{eig}_1(\boldsymbol{\Omega}_0))^{-1} \leq \varepsilon_0^{-1}$ by assumption and

$$\left\| \boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0) \right\|_{(2,2)} \leq \left\| \boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0) \right\|_{(\infty,\infty)} \leq \gamma(k)$$

since $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, \gamma)$. Also, $\|\boldsymbol{\Gamma}_0\|_{(2,2)} = (\mathrm{eig}_1(B_k(\boldsymbol{\Omega}_0)))^{-1}$. Note that $B_k(\boldsymbol{\Omega}_0) \geq \boldsymbol{\Omega}_0 - \gamma(k)\boldsymbol{I}$, which has minimum eigenvalue bounded away from zero by the assumption on $\boldsymbol{\Omega}_0$ and since $\gamma(k) \to 0$. Thus, equation (7.5) gives,

$$\|\boldsymbol{\Gamma}_0 - \boldsymbol{\Sigma}_0\|_{(2,2)} = O(\gamma(k)). \tag{7.6}$$

The above norm bound also applies to the difference of corresponding submatrices defined by the cliques and separators of the graph. Also note that

$$\left\| \boldsymbol{\Gamma}_{0,C_j}^{-1} \right\|_{(2,2)} \leq \left\| \boldsymbol{\Omega}_{0,C_j} \right\|_{(2,2)} \leq \|\boldsymbol{\Omega}_0\|_{(2,2)} \leq \varepsilon_0^{-1}$$

as $\boldsymbol{\Gamma}_{0,C_j}^{-1} \leq (B_k(\boldsymbol{\Omega}_0))_{C_j} = \boldsymbol{\Omega}_{0,C_j}$ and $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, \gamma)$. Further, $\|\boldsymbol{\Sigma}_{0,C_j}\|_{(2,2)} \leq \|\boldsymbol{\Sigma}_0\|_{(2,2)} = (\mathrm{eig}_1(\boldsymbol{\Omega}_0))^{-1} \leq \varepsilon_0^{-1}$. Thus applying Lemma A.4 with $r = 2$ and $\boldsymbol{D} = \boldsymbol{\Sigma}_{0,C_j}$, we obtain

$$\begin{aligned} \mathrm{P}\left(\max_j \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \geq M_1 \right) &\leq p \max_j \mathrm{P}\left(\left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} \geq M_1 \right) \\ &\leq M_1' p k^2 \exp[-m_1 n k^{-2}] \end{aligned}$$

for some constant $M_1, M_1', m_1 > 0$, while from Lemma A.3,

$$\mathrm{P}\left(\max_j \left\|\boldsymbol{\Sigma}_{0,C_j} - \boldsymbol{S}_{C_j}\right\|_{(2,2)} \geq t\right) \leq M_2 p k^2 \exp[-m_2 n k^{-2} t^2]$$

for $|t| < m_2'$ for some constants $M_2, m_2, m_2' > 0$. Now choose $t = Ak(n^{-1}\log p)^{1/2}$ for some sufficiently large A to get the bound, using equations (7.2), (7.3), (7.4) and (7.6),

$$\left\|\sum_{j=1}^{p-k}\left(\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \left(\boldsymbol{\Sigma}_{0,C_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2} + k^{3/2}\gamma(k)\right). \tag{7.7}$$

By a similar argument, we can establish that

$$\left\|\sum_{j=2}^{p-k}\left(\left(\boldsymbol{S}_{S_j}^{-1}\right)^0 - \left(\boldsymbol{\Sigma}_{0,S_j}^{-1}\right)^0\right)\right\|_{(\infty,\infty)} = O_P\left(k^{5/2}(n^{-1}\log p)^{1/2} + k^{3/2}\gamma(k)\right). \tag{7.8}$$

Therefore, as $\|\boldsymbol{\Omega}_0 - B_k(\boldsymbol{\Omega}_0)\|_{(\infty,\infty)} \leq \gamma(k)$ for any $\boldsymbol{\Omega}_0 \in \mathcal{U}(\varepsilon_0, \gamma)$, the assertion follows. □

Proof of Lemma 4.3. We shall first prove the result for $\widehat{\boldsymbol{\Omega}}_{L_2}^{\mathrm{B}}$. The L_∞-operator norm of $\widehat{\boldsymbol{\Omega}}_{L_2}^{\mathrm{B}} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}$ can be bounded by

$$\frac{1}{n}\left\|\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right\|_{(\infty,\infty)} \tag{7.9}$$

$$+ \frac{\delta+k+n}{n}\left\|\sum_{j=1}^{p-k}\left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1}\right)^0 - \sum_{j=1}^{p-k}(\boldsymbol{S}_{C_j}^{-1})^0\right\|_{(\infty,\infty)} \tag{7.10}$$

$$+ \frac{\delta+k+n}{n}\left\|\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0 - \sum_{j=2}^{p-k}(\boldsymbol{S}_{S_j}^{-1})^0\right\|_{(\infty,\infty)} \tag{7.11}$$

$$+ \left|\frac{\delta+k+n}{n} - 1\right|\left\|\sum_{j=1}^{p-k}\left(\boldsymbol{S}_{C_j}^{-1}\right)^0 - \sum_{j=2}^{p-k}\left(\boldsymbol{S}_{S_j}^{-1}\right)^0\right\|_{(\infty,\infty)}. \tag{7.12}$$

Now, the expression in (7.9) above is

$$\frac{1}{n}\max_l \sum_{l'}\left|\left[\sum_{j=2}^{p-k}\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right]_{(l,l')}\right|$$

$$\leq \frac{1}{n}\max_l \sum_{j=2}^{p-k}\sum_{l'}\left|\left[\left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1}\right)^0\right]_{(l,l')}\right|$$

$$\leq \frac{2k+1}{n} \max_j \max_l \sum_{l'} \left| \left[(n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right]_{(l,l')} \right|$$
$$= \frac{2k+1}{n} \max_j \left\| (n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right\|_{(\infty,\infty)},$$

which is bounded by a multiple of

$$\frac{k^{3/2}}{n} \max_j \left\| (n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right\|_{(2,2)} \leq \frac{k^{3/2}}{n} \max_j \left\| \boldsymbol{S}_{S_j}^{-1} \right\|_{(2,2)}. \tag{7.13}$$

In view of Lemma A.4, we have that for some $M_3, M_3', m_3 > 0$,

$$\mathrm{P}\left(\max_j \left\| \boldsymbol{S}_{S_j}^{-1} \right\|_{(2,2)} \geq M_3 \right) \leq M_3' p k^2 \exp[-m_3 n k^{-2}],$$

which converges to zero if $k^2(\log p)/n \to 0$. This leads to the estimate

$$n^{-1} \left\| \sum_{j=2}^{p-k} \left((n^{-1}\boldsymbol{I}_k + \boldsymbol{S}_{S_j})^{-1} \right)^0 \right\|_{(\infty,\infty)} = O_P\left(k^{3/2}/n \right). \tag{7.14}$$

For (7.10), we observe that

$$\left\| \sum_{j=1}^{p-k} \left((n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right)^0 - \sum_{j=1}^{p-k} (\boldsymbol{S}_{C_j}^{-1})^0 \right\|_{(\infty,\infty)}$$
$$\leq (2k+1) \max_j \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(\infty,\infty)}$$
$$\lesssim k^{3/2} \max_j \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}$$

and that

$$\left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} - \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}$$
$$\leq \left\| (n^{-1}\boldsymbol{I}_{k+1} + \boldsymbol{S}_{C_j})^{-1} \right\|_{(2,2)} \left\| n^{-1}\boldsymbol{I}_{k+1} \right\|_{(2,2)} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}$$
$$\leq n^{-1} \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)}^2.$$

Now under $k^2(\log p)/n \to 0$, an application of Lemma A.4 leads to the bound $O_P(k^{3/2}/n)$ for (7.10).

A similar argument gives rise to the same $O_P(k^{3/2}/n)$ bound for (7.11).

Finally to consider (7.12). As argued in bounding (7.9), we have that

$$\left\| \sum_{j=1}^{p-k} \left(\boldsymbol{S}_{C_j}^{-1} \right)^0 - \sum_{j=2}^{p-k} \left(\boldsymbol{S}_{S_j}^{-1} \right)^0 \right\|_{(\infty,\infty)}$$
$$\leq k^{1/2}(2k+1) \left[\max_j \left\| \boldsymbol{S}_{C_j}^{-1} \right\|_{(2,2)} + \max_j \left\| \boldsymbol{S}_{S_j}^{-1} \right\|_{(2,2)} \right] = O_P(k^{3/2}),$$

under the assumption $k^2(\log p)/n \to 0$ by another application of Lemma A.4. Since $n^{-1}(\delta + k + n) - 1 = O(k/n)$, it follows that the expression in (7.12) is $O_P(k^{5/2}/n)$, which is the weakest estimate among all terms in the bound for $\|\widehat{\boldsymbol{\Omega}}^B_{L_2} - \widehat{\boldsymbol{\Omega}}^M\|_{(\infty,\infty)}$. The result thus follows for $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}$.

The assertion for the estimator $\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_1}$ follows similarly. □

Proof of Proposition 4.4. The proof follows from Theorem 4.2 and Lemma 4.3 using the triangle inequality. □

Proof of Theorem 4.1. The posterior distribution of the precision matrix $\boldsymbol{\Omega}$ given the data $\boldsymbol{X}$ is a G-Wishart distribution $W_G(\delta + n, \boldsymbol{I}_p + n\boldsymbol{S})$. We can write $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \sum_{j=1}^{p-k} (\boldsymbol{\Omega}_{C_j})^0 - \sum_{j=2}^{p-k} (\boldsymbol{\Omega}_{S_j})^0 = \sum_{j=1}^{p-k} \left(\boldsymbol{\Sigma}^{-1}_{C_j}\right)^0 - \sum_{j=2}^{p-k} \left(\boldsymbol{\Sigma}^{-1}_{S_j}\right)^0. \tag{7.15}$$

The submatrix $\boldsymbol{\Sigma}_{C_j}$ for any clique C_j has a inverse Wishart distribution with parameters $\delta + n$ and scale matrix $(\boldsymbol{I}_p + n\boldsymbol{S})_{C_j}$, $j = 1, \ldots, p-k$. Thus, $W_{C_j} = \boldsymbol{\Sigma}^{-1}_{C_j}$ has a Wishart distribution induced by the corresponding inverse Wishart distribution. In particular, if $i \in C_j$, then $\tau_{in}^{-1} w_{ii}$ has chi-square distribution with $(\delta + n)$ degrees of freedom, where τ_{in} is the (i,i)th entry of $((\boldsymbol{I} + \boldsymbol{S}_{C_j})^{-1})^0$. Fix a clique $C = C_j$ and define $\boldsymbol{T}_n = \mathrm{diag}(w_{ii}\colon i \in C)$. For $i, j \in C$, let $w^*_{ij} = w_{ij}/\sqrt{\tau_{in}\tau_{jn}}$ and $\boldsymbol{W}^*_C = (\!(w^*_{ij}\colon i, j \in C)\!)$. Then $\boldsymbol{W}^*_C$ given $\boldsymbol{X}$ has a Wishart distribution with parameters $\delta + n$ and scale matrix $\boldsymbol{T}_n^{-1/2}(\boldsymbol{I}_{k+1} + n\boldsymbol{S}_C)\boldsymbol{T}_n^{-1/2}$.

We first note that $\max_i \tau_{in} = O_P(n^{-1})$. To see this, observe that $(\boldsymbol{I}_k + n\boldsymbol{S}_C)^{-1} \le n^{-1}\boldsymbol{S}_C^{-1}$, so that

$$\max_i |\tau_{in}| \le \frac{1}{n}\|\boldsymbol{S}_C^{-1}\|_{(2,2)} = O_P(n^{-1})$$

in view of Lemma A.4. On the other hand, from Lemma A.3, it follows that $\max_C \|\boldsymbol{S}_C\|_{(2,2)} = O_P(1)$, so with probability tending to one, $\boldsymbol{S}_C \le L\boldsymbol{I}_C$, and hence $(\boldsymbol{I} + n\boldsymbol{S})^{-1}_C \ge (1 + nL)^{-1}\boldsymbol{I}_C$ simultaneously for all cliques, for some constant $L > 0$. Hence $\max_i \tau_{in}^{-1} = O_P(n)$. Consequently, with probability tending to one, the maximum eigenvalue of $\boldsymbol{T}_n^{-1/2}(\boldsymbol{I}_{k+1} + n\boldsymbol{S}_C)\boldsymbol{T}_n^{-1/2}$ is bounded by a constant depending only on ε_0, simultaneously for all cliques. Hence applying Lemma A.3 of [3], it follows that for all i, j,

$$\mathrm{P}\left[|w_{ij} - \mathrm{E}(w_{ij}|\boldsymbol{X})| \ge t\right] \le M_4 \exp[-m_4(\delta + n)t^2], \quad |t| < m'_4, \tag{7.16}$$

for some constants $M_4, m_4, m'_4 > 0$ depending on ϵ_0 only.

Now, as a G-Wishart prior gives rise to a k-banded structure, as arguing in the bounding of (7.9) and using (7.15), we have that, for some $M_5, m_5, m'_5 > 0$, and all $|t| < m'_5$,

$$\mathrm{P}\left\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} \ge k^2 t|\boldsymbol{X}\right\} \le M_5 p k^2 \exp[-m_5 n t^2]. \tag{7.17}$$

The reduction in the number of terms in the rows from p to $(2k+1)$ is possible due to the fact that the G-Wishart posterior preserves the banded structure of the precision matrix. Choosing $t = A(n^{-1}\log p)^{1/2}$, with A sufficiently large, we get

$$\mathrm{E}_0[\mathrm{P}\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} \geq Ak^2(n^{-1}\log p)^{1/2}|\boldsymbol{X}\}] \to 0. \tag{7.18}$$

Therefore, using Proposition 4.4,

$$\begin{aligned}&\mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega} - \boldsymbol{\Omega}_0\|_{(\infty,\infty)} > 2\epsilon_n|\boldsymbol{X}\right\}\right]\\ &\quad\leq \mathrm{P}_0\left\{\|\widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2} - \boldsymbol{\Omega}_0\|_{(\infty,\infty)} > \epsilon_n|\boldsymbol{X}\right\} + \mathrm{E}_0\left[\mathrm{P}\left\{\|\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}}^{\mathrm{B}}_{L_2}\|_{(\infty,\infty)} > \epsilon_n|\boldsymbol{X}\right\}\right],\end{aligned}$$

which converges to zero if $\epsilon_n = A(k^{5/2}(n^{-1}\log p)^{1/2} + k^{3/2}\gamma(k))$. □

Proof of (4.6). For the graph induced by banding, the posterior mean under the reference prior is given by the expression

$$\widehat{\boldsymbol{\Omega}}^R = \mathrm{E}(\boldsymbol{\Omega}|\boldsymbol{S}) = \sum_{j=1}^{p-k}(\boldsymbol{S}^{-1}_{C_j})^0 - (1-n^{-1})(\boldsymbol{S}^{-1}_{S_2})^0 - (1-n^{-1})\sum_{j=3}^{p-k}(\boldsymbol{S}^{-1}_{S_j})^0.$$

Therefore

$$\begin{aligned}\|\widehat{\boldsymbol{\Omega}}^{\mathrm{R}} - \widehat{\boldsymbol{\Omega}}^{\mathrm{M}}\|_{(\infty,\infty)} &= \left\|n^{-1}(\boldsymbol{S}^{-1}_{S_2})^0 + n^{-1}\sum_{j=3}^{p-k}(\boldsymbol{S}^{-1}_{S_j})^0\right\|_{(\infty,\infty)}\\ &\leq n^{-1}\left\|\sum_{j=2}^{p-k}(\boldsymbol{S}^{-1}_{S_j})^0\right\|_{(\infty,\infty)} + n^{-1}\left\|(\boldsymbol{S}^{-1}_{S_2})^0\right\|_{(\infty,\infty)}.\end{aligned}$$

The rest of the proof proceeds as in Lemma 4.3. □

Appendix: Proofs of auxiliary results

In this section we give proofs of some lemmas we have used in the paper, which are of some general interest.

The first lemma deals with the various equivalence conditions for matrix norms and is easily found in standard textbooks.

Lemma A.1. *For a symmetric matrix $\boldsymbol{A}$ of order k, we have the following:*

1. $\|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_{(\infty,\infty)} \leq \sqrt{k}\|\boldsymbol{A}\|_{(2,2)}$;
2. $\|\boldsymbol{A}\|_{\infty} \leq \|\boldsymbol{A}\|_{(2,2)} \leq \|\boldsymbol{A}\|_{(\infty,\infty)} \leq k\|\boldsymbol{A}\|_{\infty}$;

Now we show that matrices $\boldsymbol{\Omega}$ belonging to the class $\mathcal{U}(\varepsilon_0,\gamma)$ automatically have bounded L_∞-operator norm.

Lemma A.2. *For every ε_0, there exist K depending only on ε_0 and $\gamma(\cdot)$ such that for all $\Omega \in \mathcal{U}(\varepsilon_0,\gamma)$, we have $\|\boldsymbol{\Omega}\|_{(\infty,\infty)} \leq K$.*

Proof. For any fixed k_0,

$$\begin{aligned}\|\boldsymbol{\Omega}\|_{(\infty,\infty)} &\leq \|\boldsymbol{\Omega} - B_{k_0}(\boldsymbol{\Omega})\|_{(\infty,\infty)} + \|B_{k_0}(\boldsymbol{\Omega})\|_{(\infty,\infty)} \\ &\leq \gamma(k_0) + (2k_0+1)\|\boldsymbol{\Omega}\|_{\infty} \\ &\leq \gamma(k_0) + (2k_0+1)\|\boldsymbol{\Omega}\|_{(2,2)} \\ &\leq \gamma(k_0) + (2k_0+1)\varepsilon_0^{-1}, \end{aligned} \tag{A.19}$$

so the conclusion holds for $K = \gamma(k_0) + (2k_0+1)\varepsilon_0^{-1}$. □

Lemma A.3. *Let $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, be i.i.d. k-dimensional random vectors distributed as $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{D})$ and $\|\boldsymbol{D}\|_{(2,2)} \leq K$. Then for the sample variance $\boldsymbol{S} = n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T$, we have for $r \in \{2, \infty\}$*

$$\mathrm{P}\left[\|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)} \geq t\right] \leq Mk^2 \exp(-mnk^{-2}t^2), \quad |t| \leq m', \tag{A.20}$$

where $M, m, m' > 0$ depend on K only.

In particular, if $k^2(\log k)/n \to 0$, then $\|\boldsymbol{S}\|_{(\infty,\infty)} = O_P(1)$.

Proof. By the estimate of the large deviation probability given in Lemma A.3 of [3], it is immediate that $\mathrm{P}[\|\boldsymbol{S} - \boldsymbol{D}\|_{\infty} \geq t] \leq Mk^2\exp(-mnt^2)$. Now the assertion follows by noting from Lemma A.1 that $\|\boldsymbol{S}-\boldsymbol{D}\|_{(r,r)} \leq k\|\boldsymbol{S}-\boldsymbol{D}\|_{\infty}$. □

Lemma A.4. *Let $\boldsymbol{Z}_i$, $i = 1, \ldots, n$, be i.i.d. k-dimensional random vectors distributed as $\mathrm{N}_k(\boldsymbol{0}, \boldsymbol{D})$ and $\max\{\|\boldsymbol{D}^{-1}\|_{(r,r)}, \|\boldsymbol{D}\|_{(2,2)}\} \leq K$ for $r \in \{2, \infty\}$. Then for the sample variance $\boldsymbol{S} = n^{-1}\sum_{i=1}^n \boldsymbol{Z}_i\boldsymbol{Z}_i^T$, we have*

$$\mathrm{P}\left[\|\boldsymbol{S}^{-1}\|_{(r,r)} \geq M\right] \leq M'k^2\exp(-mnk^{-2}C'^2), \tag{A.21}$$

where $M > K$ and $M', m > 0$ depend on M and K only.

Proof. Note that,

$$\begin{aligned}\|\boldsymbol{S}^{-1}\|_{(r,r)} &\leq \|\boldsymbol{D}^{-1}\|_{(r,r)} + \|\boldsymbol{S}^{-1} - \boldsymbol{D}^{-1}\|_{(r,r)} \\ &= \|\boldsymbol{D}^{-1}\|_{(r,r)} + \|\boldsymbol{D}^{-1}\|_{(r,r)}\|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)}\|\boldsymbol{S}^{-1}\|_{(r,r)} \\ &\leq K(1 + \|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)}\|\boldsymbol{S}^{-1}\|_{(r,r)}). \end{aligned} \tag{A.22}$$

This implies that

$$\|\boldsymbol{S}^{-1}\|_{(r,r)} \leq \frac{K}{1 - \|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)}K}.$$

Thus, using Lemma A.3, we obtain

$$\begin{aligned}\mathrm{P}\left[\|\boldsymbol{S}^{-1}\|_{(r,r)} \geq M\right] &\leq \mathrm{P}\left[\frac{K}{1 - \|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)}K} \geq M\right] \\ &\leq \mathrm{P}\left[\|\boldsymbol{S} - \boldsymbol{D}\|_{(r,r)} \geq K^{-1} - M^{-1}\right] \\ &\leq M'k^2\exp(-mnk^{-2}). \end{aligned} \tag{A.23}$$

□

References

[1] Atay-Kayis, A. and Massam, H. (2005). A Monte-Carlo method for computing the marginal likelihood in nondecomposable Gaussian graphical models. *Biometrika* **92** 317–335. MR2201362

[2] Bickel, P. J. and Levina, E. (2008a). Covariance regularization by thresholding. *Ann. Statist.* **36** 2577–2604. MR2485008

[3] Bickel, P. J. and Levina, E. (2008b). Regularized estimation of large covariance matrices. *Ann. Statist.* **36** 199–227. MR2387969

[4] Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.* **106** 672–684. MR2847949

[5] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1-minimization approach to sparse precision matrix estimation. *J. Amer. Statist. Assoc.* **106** 594–607. MR2847973

[6] Cai, T. T. and Yuan, M. (2012). Adaptive covariance matrix estimation through block thresholding. *Ann. Statist.* **40** 2014–2042. MR3059075

[7] Cai, T. T., Zhang, C. H. and Zhou, H. H. (2010). Optimal rates of convergence for covariance matrix estimation. *Ann. Statist.* **38** 2118–2144. MR2676885

[8] Carvalho, C. M., Massam, H. and West, M. (2007). Simulation of hyper-inverse Wishart distributions in graphical models. *Biometrika* **94** 647–659. MR2410014

[9] Carvalho, C. M. and Scott, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96** 497–512. MR2538753

[10] Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

[11] Dobra, A., Lenkoski, A. and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *J. Amer. Statist. Assoc.* **106** 1418–1433. MR2896846

[12] Dobra, A., Hans, C., Jones, B., Nevins, J. R., Yao, G. and West, M. (2004). Sparse graphical models for exploring gene expression data. *J. Multivariate Anal.* **90** 196–212. MR2064941

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

[14] Ghosal, S. (2000). Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity. *J. Multivariate Anal.* **74** 49–68. MR1790613

[15] Gröne, R., Johnson, C. R., Sá, E. M. and Wolkowicz, H. (1984). Positive definite completions of partial Hermitian matrices. *Linear Algebra Appl.* **58** 109–124. MR0739282

[16] Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika* **93** 85–98. MR2277742

[17] Karoui, N. E. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Ann. Statist.* **36** 2717–2756. MR2485011

[18] Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Statist.* **37** 4254. MR2572459

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

[20] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate Anal.* **88** 365–411. MR2026339

[21] Lenkoski, A. and Dobra, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graphical Statist.* **20** 140–157. MR2816542

[22] Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs. *Ann. Statist.* **35** 1278–1323. MR2341706

[23] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34** 1436–1462. MR2278363

[24] Muirhead, R. (2005). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

[25] Pati, D., Bhattacharya, A., Pillai, N. S. and Dunson, D. (2014). Posterior contraction in sparse Bayesian factor models for massive covariance matrices. *Ann. Statist.* **42** 1102–1130. MR3210997

[26] Rajaratnam, B., Massam, H. and Carvalho, C. M. (2008). Flexible covariance estimation in graphical Gaussian models. *Ann. Statist.* **36** 2818–2849. MR2485014

[27] Rothman, A. J., Levina, E. and Zhu, J. (2009). Generalized thresholding of large covariance matrices. *J. Amer. Statist. Assoc.* **104** 177–186. MR2504372

[28] Rothman, A. J., Bickel, P. J., Levina, E. and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electron. J. Statist.* **2** 494–515. MR2417391

[29] Roverato, A. (2000). Cholesky decomposition of a hyper inverse Wishart matrix. *Biometrika* **87** 99–112. MR1766831

[30] Yang, R. and Berger, J. O. (1994). Estimation of a covariance matrix using the reference prior. *Ann. Statist.* **22** 1195–1211. MR1311972

[31] Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika* **94** 19–35. MR2367824