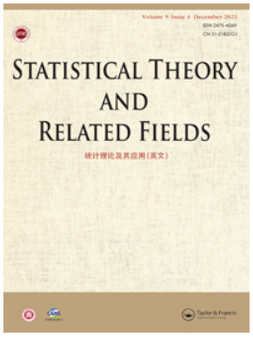

Statistical Theory and Related Fields

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

# Covariance estimation via fiducial inference

W. Jenny Shi, Jan Hannig, Randy C. S. Lai & Thomas C. M. Lee

**To cite this article:** W. Jenny Shi, Jan Hannig, Randy C. S. Lai & Thomas C. M. Lee (2021) Covariance estimation via fiducial inference, Statistical Theory and Related Fields, 5:4, 316-331, DOI: 10.1080/24754269.2021.1877950

**To link to this article:** https://doi.org/10.1080/24754269.2021.1877950

STATISTICAL THEORY AND RELATED FIELDS
2021, VOL. 5, NO. 4, 316–331
https://doi.org/10.1080/24754269.2021.1877950



# Covariance estimation via fiducial inference

W. Jenny Shi [a], Jan Hannig [b], Randy C. S. Lai [c] and Thomas C. M. Lee [c]

[a]Financial Planning & Analysis, MassMutual, Boston, MA, USA; [b]Department of Statistics & Operations Research, University of North Carolina, Chapel Hill, NC, USA; [c]Department of Statistics, University of California, Davis, CA, USA

**ABSTRACT**

As a classical problem, covariance estimation has drawn much attention from the statistical community for decades. Much work has been done under the frequentist and Bayesian frameworks. Aiming to quantify the uncertainty of the estimators without having to choose a prior, we have developed a fiducial approach to the estimation of covariance matrix. Built upon the Fiducial Berstein–von Mises Theorem, we show that the fiducial distribution of the covariate matrix is consistent under our framework. Consequently, the samples generated from this fiducial distribution are good estimators to the true covariance matrix, which enable us to define a meaningful confidence region for the covariance matrix. Lastly, we also show that the fiducial approach can be a powerful tool for identifying clique structures in covariance matrices.

**ARTICLE HISTORY**
Received 23 May 2020
Revised 5 January 2021
Accepted 10 January 2021

**KEYWORDS**
covariance estimation; sparsity; fiducial inference; cliques

**2010 MATHEMATICS SUBJECT CLASSIFICATIONS**
Primary 62J10; 62E20; 62F25; Secondary 62F12

## 1. Introduction

Estimating covariance matrices has historically been a challenging problem. Many regression-based methods have emerged in the last few decades, especially in the concept of 'large $p$ small $n$'. Among the notable methods, there are the graphical LASSO algorithms (Friedman et al., 2008, 2010; Rothman, 2012). Pourahmadi provided a detailed overview on the progress of covariance estimation (Pourahmadi, 2011). The Positive Definite Sparse Covariance Estimators (PDSCE) method (Rothman, 2012) has grained great popularity due to its performance comparing to other current methods, although it only produces a point estimator.

Aiming to have a distribution of good covariance estimators, we propose a generalised fiducial approach. The ideas underpinning fiducial inference were introduced by Fisher (1922,1930,1933,1935), whose intention was to overcome the need for priors and other issues with Bayesian methods perceived at the time. The procedure of fiducial inference allows to obtain a measure on the parameter space without requiring priors and defines approximate pivots for parameters of interest. It is ideal when *a priori* information about the parameters is unavailable. The key recipe of the fiducial argument is the data generating equation. Roughly, the generalised fiducial likelihood is defined as the distribution of the functional inverse of the data generating mechanism.

One great advantage of the fiducial approach to covariance matrix estimation is that, without specifying a prior, it produces a family of matrices that are close to the true covariance with a probabilistic characterisation using the fiducial likelihood function. This attractive property enables a meaningful definition for matrix confidence regions.

We are particularly interested in a high-dimensional multivariate linear model setting with possibly an atypical sparsity constraint. Instead of classical sparsity assumptions on the covariance matrix, we consider a type of experimental design that enforces sparsity on the covariate matrix. This phenomenon often arises in the studies of metabolomics and proteomics. One example of this setup is modelling the relationship between a set of gene expression levels and a list of metabolomic data. The expression levels of the genes serve as the predictor variables while the response variables are a variety of metabolite levels, such as sugar and triglycerides. It is known that only a small subset of genes contribute to each metabolite level, and each gene can be responsible for just a few metabolite levels.

Under the sparse covariate setting, we derive the generalised fiducial likelihood of the covariate matrix based on given observations and prove its asymptotic consistency as the sample size increases. For the covariance with community structures (cliques), we prove the necessary conditions for achieving accurate clique structure estimation. Samples from the fiducial distribution of a covariate matrix can be generated using Monte Carlo methods. In the general case, a reversible jump Markov chain Monte Carlo (RJMCMC)

**CONTACT** Jan Hannig jan.hannig@unc.edu Department of Statistics & Operations Research, University of North Carolina, 330 Hanes Hall, Chapel Hill, NC 27599, USA .

Supplemental data for this article can be accessed here. https://doi.org/10.1080/24754269.2021.1877950

algorithm may be needed. Similar to the classic likelihood functions, fiducial distributions favour models with more parameters. Therefore, in the case where the exact sparsity structure of the covariate is unclear, a penalty term needs to be added. To obtain a family of covariance estimators in the general case, we adapt a zeroth-order method and develop an efficient RJMCMC algorithm that samples from the penalised fiducial distribution.

The rest of the paper is arranged as follows. In Section 2, we will provide a brief background and development on fiducial inference. Then we will introduce the fiducial model for covariance estimation and derive the Generalised Fiducial Distribution (GFD) for the covariate and covariance matrices and examine the asymptotic property of the GFD of the covariance matrix under minor assumptions in Section 3. Some toy examples on sampling from GFD will also be shown. Section 4 focuses on the clique model, where we show some theoretical results for the clique model and how the fiducial approach can be applied to uncover clique structures. Finally, Section 5 concludes the paper with a summary and a short discussion on the relationship of our approach to Bayesian methods.

## 2. Generalised fiducial inference

### 2.1. Brief background

Fiducial inference was first proposed by Fisher (1930) when he introduced the concept of a fiducial distribution of a parameter. In the case of a single parameter family of distributions, Fisher gave the following definition for a *fiducial density* $f(\theta \mid x)$ of the parameter based on a single observation $x$ for the case where the cumulative distribution function $F(x \mid \theta)$ is a monotonic decreasing function of $\theta$:

$$f(\theta \mid x) \propto -\frac{\partial F(x \mid \theta)}{\partial \theta}. \tag{1}$$

A fiducial distribution can be viewed as a Bayesian posterior distribution without hand picking priors. In many single parameter distribution families, Fisher's fiducial intervals coincide with classical confidence interval. For families of distributions with multiple parameters, the fiducial approach leads to confidence set. The definition of fiducial inference has been generalised in the past decades. Hannig et al. (2016) provide a detailed review on the philosophy and current development on the subject.

The generalised fiducial approach has been applied to a variety of models, both parametric and nonparametric, both continuous and discrete. These applications include bioequivalence (Hannig et al., 2006), variance components (Cisewski & Hannig, 2012; Lidong et al., 2008; Li et al., 2018), problems of metrology (Hannig et al., 2007,2003; Wang et al., 2012; Wang & Iyer, 2005, 2006a, 2006b), inter laboratory experiments and international key comparison experiments (Hannig et al., 2018; Iyer et al., 2004), maximum mean of a multivariate normal distribution (Wandler & Hannig, 2011), multiple comparisons (Wandler & Hannig, 2012), extreme value estimation (Wandler & Hannig, 2012), mixture of normal and Cauchy distributions (Glagovskiy, 2006), wavelet regression (Hannig & Lee, 2009), high-dimensional regression (Lai et al., 2015; Williams & Hannig, 2018), item response models (Liu & Hannig, 2016,2017), non-parametric survival function estimation with censoring (Cui & Hannig, 2019), Other related approaches include Martin and Liu (2015); Schweder and Hjort (2016); Xie and Singh (2013).

### 2.2. Generalised fiducial distribution

The idea underlying generalised fiducial inference is built upon a *data generating algorithm* $G(\cdot, \cdot)$ expressing the relationship between the data $X$ and the parameters $\theta$:

$$X = G(U, \theta), \tag{2}$$

where $U$ is the random component of this data generating algorithm whose distribution is known. The data $X$ are assumed to be created by generating a random variable $U$ and plugging it into the data generating algorithm above.

The GFD inverts Equation (2). Assume that $x \in \mathbb{R}^n$ is continuous, and the parameter $\theta \in \mathbb{R}^p$. Under the conditions provided in Hannig et al. (2016), fiducial distribution is shown to have density

$$r(\theta \mid x) = \frac{f(x, \theta) J(x, \theta)}{\int_\Theta f(x, \theta') J(x, \theta') \, \mathrm{d}\theta'}, \tag{3}$$

where $f(x, \theta)$ is the likelihood, and

$$J(x, \theta) = D\left(\nabla_\theta G(u, \theta)|_{u = G^{-1}(x, \theta)}\right). \tag{4}$$

Here $\nabla_\theta G(u, \theta)$ is the $n \times p$ Jacobian matrix. The exact form of $D(\cdot)$ depends on the choices made in the process of inverting (2). In this manuscript, we concentrate on what Hannig et al. (2016) calls the $\ell_2$-norm choice:

$$D(M) = \sqrt{\det(M^{\mathrm{T}} M / n)}, \tag{5}$$

where $M^{\mathrm{T}}$ denotes the matrix transpose of $M$. Other choices, in particular the $\ell_\infty$-norm that was often used in the past, leads to similar results is studied in detail in Shi (2015).

## 3. A fiducial approach to covariance estimation

In this section, we will derive the GFD for the covariance matrix of a multivariate normal random variable. For this problem, various regularised estimators

were proposed under the assumption that the true covariance matrix is sparse (Avella-Medina et al., 2018; Bickel & Levina, 2008a, 2008b; Cai & Liu, 2011; Furrer & Bengtsson, 2007; Huang & Lee, 2016; Huang et al., 2006; Lam & Fan, 2009; Levina et al., 2008; Rothman et al., 2009, 2010; Wu & Pourahmadi, 2003). While many of these estimators have been shown to enjoy excellent rates of convergence, so far little work has been done to quantify the uncertainties of their corresponding estimates.

Let $Q^{\mathrm{T}}$ denote the transpose of a matrix/vector $Q$. Denote a collection of $n$ observed $p$ dimensional objects $\mathbf{Y} = \{Y_i : i = 1, \ldots, n\}$. For the rest of the paper, we assume $p$ is fixed, unless stated otherwise. Consider the following data generating equation:

$$Y_i = AZ_i, \quad i = 1, \ldots, n; \tag{6}$$

where $A$ is a $p \times p$ matrix of full rank; $\boldsymbol{Z} = \{Z_i = (z_{i1}, \ldots, z_{ip})^{\mathrm{T}},\ i = 1, \ldots, n\}$ are independent and identically distributed (i.i.d) $p \times 1$ random vectors following multivariate normal distribution $N(0, I)$. Hence, $Y_i$'s are i.i.d random vectors centred at 0 with covariance matrix $AA^{\mathrm{T}}$,

$$\text{i.e.} Y_i \overset{\text{i.i.d}}{\sim} N(0, \Sigma), \text{where } \Sigma = AA^{\mathrm{T}}. \tag{7}$$

Consequently, we have the likelihood for observations $\boldsymbol{y}$:

$$f(\boldsymbol{y}, A) = (2\pi)^{-\frac{np}{2}} |\det(A)|^{-n} \times \exp\left[-\frac{1}{2}\mathrm{tr}\{nS_n(AA^{\mathrm{T}})^{-1}\}\right], \tag{8}$$

where $S_n = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^{\mathrm{T}}$ is the corresponding sample covariance matrix and $\mathrm{tr}\{\cdot\}$ is the trace operator.

We propose to estimate the covariance matrix $\Sigma$ through the GFD of covariate matrix $A$:

$$r(A \mid \boldsymbol{y}) \propto J(\boldsymbol{y}, A) f(\boldsymbol{y}, A). \tag{9}$$

Define the stacked observation vector $\mathbf{w} = (y_1^{\mathrm{T}}, \ldots, y_n^{\mathrm{T}})^{\mathrm{T}} = (w_1, \ldots, w_{np})^{\mathrm{T}}$. Denote $\mathbf{u} = (u_1, \ldots, u_n)$, such that $y_i = G(u_i, A),\ \forall i$. Let $a_{ij}$ be the $(i, j)$-th entry of matrix $A$, i.e., $A = [a_{ij}]_{1 \le i,j \le p}$. The corresponding Jacobian $J(\boldsymbol{y}, A)$ derived from (4) is then

$$J(\boldsymbol{y}, A) = D\left(\nabla_A \mathbf{w}|_{\mathbf{u}=G^{-1}(\boldsymbol{y},A)}\right), \tag{10}$$

where $\nabla_A \mathbf{w}$ is an $np \times p^2$ matrix

$$\nabla_A \mathbf{w} = \begin{pmatrix} \frac{\partial w_1}{\partial a_{11}} & \frac{\partial w_1}{\partial a_{12}} & \cdots & \frac{\partial w_1}{\partial a_{pp}} \\ \frac{\partial w_2}{\partial a_{11}} & \frac{\partial w_2}{\partial a_{12}} & \cdots & \frac{\partial w_2}{\partial a_{pp}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial w_{np}}{\partial a_{11}} & \frac{\partial w_{np}}{\partial a_{12}} & \cdots & \frac{\partial w_{np}}{\partial a_{pp}} \end{pmatrix}$$

and $D(\cdot)$ is given by (5).

Often some $a_{kl}$ are known to be zero; a common example is the lower triangular matrix $A$ for which $a_{kl} = 0$ for $l > k$. Additionally, sparsity on the covariate model can be introduced by having most of the $a_{kl}$ known to be zero as a part of the model. Note that if $a_{kl}$ is known to be zero, as implied by model, then the corresponding $(k, l)$th column is dropped. Therefore, depending on the sparsity model, the dimension of $\nabla_A \mathbf{w}$ varies.

Recall, that there is a one-to-one mapping between positive definite matrices $\Sigma$ and lower triangular matrices $A$ with positive entries on the main diagonal. While we are not assuming $A$ is lower triangular, in order to alleviate some identifiability issues we will assume that all diagonal entries of $A$ are positive, i.e., $a_{kk} > 0, k = 1, \ldots, p$.

### 3.1. Jacobian for full models

Suppose that none of the entries of $A$ is fixed at zero, namely, the parameter space $\Theta$ for $A$ is $\mathbb{R}^{p \times p}$. We will refer this to a full model. Under a full model, $\nabla_A \mathbf{w}$ consists of $p$ blocks, each of dimension $np \times p$. Every row of $\nabla_A \mathbf{w}$ has non-zero entries in only one block.

By swapping rows in the matrix $\nabla_A \mathbf{w}$ and plugging $\mathbf{u} = G^{-1}(\boldsymbol{y}, A)$, we obtain the $np \times p^2$ matrix $P$:

$$P = \begin{pmatrix} U & & \\ & \ddots & \\ & & U \end{pmatrix}, \tag{11}$$

where $U = (A^{-1}y_1, \cdots, A^{-1}y_n)^{\mathrm{T}} = V(A^{-1})^{\mathrm{T}}, V = (y_1; \cdots; y_n)^{\mathrm{T}}$. Notice that $P$ breaks into $p$ blocks, $B_1, \ldots, B_p$, where $B_i = \begin{pmatrix} O_{(ni-n)\times p} \\ U \\ O_{(np-ni)\times p} \end{pmatrix}$, $O_{a\times b}$ denotes a zero matrix with dimension $a \times b$.

Since as a consequence of Cauchy–Binnet formula (see also Hannig et al. (2016)), swapping rows do not change the value of the Jacobian function (10). Therefore $J(\boldsymbol{y}, A)$ can be expressed using matrix $P$:

$$J(\boldsymbol{y}, A) = D(P) = |\det(S_n)|^{\frac{p}{2}} |\det(A)|^{-p}, \tag{12}$$

where $S_n = n^{-1}\sum_{i=1}^{n} y_i y_i^{\mathrm{T}}$ is the MLE estimator of the covariance matrix.

By (9), the GFD is proportional to

$$r(A \mid \boldsymbol{y}) \propto |\det(S_n)|^{\frac{p}{2}} (2\pi)^{-\frac{np}{2}} |\det(A)|^{-(n+p)} \times \exp\left[-\frac{1}{2}\mathrm{tr}\{nS_n(AA^{\mathrm{T}})^{-1}\}\right]. \tag{13}$$

By transforming the GFD of $A$, we conclude that the GFD of $\Sigma = AA^{\mathrm{T}}$ has the inverse Wishart distribution with $n$ degrees of freedom and parameter $nS_n$.

### 3.2. Jacobian for the general case

While having a closed form for the GFD of $\Sigma$ for the full model, the covariance estimation requires sufficient

number of observations (roughly at least $n > 15(p+1)$) to maintain reasonable power. In the cases where $n$ is small, we reduce the parameter space by introducing a sparse structure $\mathcal{M}$, which determines which entries of $A$ are known to be zero. Recall, that we only consider $A$ with positive diagonal entries.

Now assume the general case with a sparse model $\mathcal{M}$, where some entries of $A$ are known to be zero. Denote the $(i,j)$th entry of $A$ as $A_{ij}$. Define the zero index set for the $i$th row as

$$S_i = \{j : A_{ij} \equiv 0, j = 1, \ldots, p\}, \quad i = 1, \ldots, p. \tag{14}$$

The set $S_i$ indicates which entries of $A$ in the $i$th row are fixed at zero.

Then Equation (10) becomes

$$J(\boldsymbol{y}, A) = D\left(\tilde{P}\right), \tag{15}$$

where $\tilde{P} = (\tilde{B}_1, \ldots, \tilde{B}_p)$ is the matrix $P$ with correct corresponding columns dropped, i.e., block $\tilde{B}_i$ is obtained from block $B_i$ with $S_i$ columns removed.

Let $p_i$ be the number of nonzero entries in the $i$th row of $A$, and $U_i$ be the sub-matrix of $U$ excluding columns in $S_i$, i.e., $U_i = U_{[:,-S_i]}$. Consequently, Equation (15) becomes

$$J(\boldsymbol{y}, A) = \sqrt{\prod_{i}^{p} \det(U_i^{\mathrm{T}} U_i / n)}. \tag{16}$$

### 3.3. Consistency of fiducial distribution

In general, there is no one-to-one correspondence between the covariance matrix $\Sigma$ and the covariate matrix $A$. However, if $A$ is sparse enough, e.g., a lower triangular matrix with positive diagonal entries, the identifiability problem vanishes. In this section, we will show that, if there is one-to-one correspondence between $\Sigma$ and $A$, then the GFD of the covariate matrix achieves a fiducial Bernstein–von Mises Theorem (Theorem 3.1), which provides theoretical guarantees of asymptotic normality and asymptotic efficiency for the GFD (Hannig et al., 2016).

The results here are derived based on FM-distance (Förstner & Moonen, 1999). For two symmetric positive definite matrices $M$ and $N$, with the eigenvalues $\lambda_i(M,N)$ from $\det(\lambda M - N) = 0$, the FM-distance between the two matrices $M$ and $N$ is

$$\mathbf{d}(M,N) = \sqrt{\sum_{i=1}^{n} \log^2 \lambda_i(M,N)}. \tag{17}$$

This distance measure is a metric and invariant with respect to both affine transformations of the coordinate system and an inversion of the matrices (Förstner & Moonen, 1999).

The Bernstein–von Mises Theorem provides conditions under which the Bayesian posterior distribution is asymptotically normal (van der Vaart, 1998; Ghosh & Ramamoorthi, 2003). The fiducial Bernstein–von Mises Theorem is an extension that includes a list of conditions under which the GFD is asymptotically normal (Sonderegger & Hannig, 2012). Those conditions can be divided into three parts to ensure each of the following:

(a) the Maximum Likelihood Estimator (MLE) is asymptotically normal;
(b) the Bayesian posterior distribution becomes close to that of the MLE;
(c) the fiducial distribution is close to the Bayesian posterior.

It is clear that the MLE of $\Sigma$ is asymptotically normal. Under our model, the conditions for (b) hold due to Proposition A.1 and the construction of the Jacobian formula; the conditions for (c) are satisfied by Propositions A.2, 3.1. Statements and proofs of the propositions are included in Appendix A.1. Here we state only Proposition 3.1 that contains notation needed in the statement of the main Theorem.

**Proposition 3.1:** *The Jacobian function* $J(\boldsymbol{y}, A) \xrightarrow{\text{a.s.}} \pi_{\Sigma_0}(A)$ *uniformly on compacts in A, where* $\pi_{\Sigma_0}(A)$ *is a function of A, independent of the sample size and observations, but depending on the true* $\Sigma_0$. *Moreover* $\pi_{\Sigma_0}(A)$ *is continuous.*

Closely following Sonderegger Hannig (2012), we arrive at Theorem 3.1.

**Theorem 3.1 (Asymptotic Normality):** *Let* $\mathcal{R}_A$ *be an vectorized observation from the fiducial distribution* $r(A \mid \boldsymbol{y})$ *and denote the density of* $B = \sqrt{n}(\mathcal{R}_A - \hat{A}_n)$ *by* $\pi^*(B, \boldsymbol{y})$, *where* $\hat{A}_n$ *is the vectorized version of a maximum likelihood estimator. Let* $I(A)$ *be the Fisher information matrix of the vectorized version of matrix (A). If the sparsity structure is such, that there is one-to-one correspondence between the true covariance matrix* $\Sigma_0$ *and the covariate matrix* $A_0$, $I(A_0)$ *is positive definite,* $\pi_{\Sigma_0}(A_0) > 0$, *then*

$$\int_{\mathbb{R}^{p^2}} \left| \pi^*(B, \boldsymbol{y}) - \frac{\sqrt{\det |I(A_0)|}}{(2\pi)^p} \times \exp\{-B^{\mathrm{T}} I(A_0) B / 2\} \right| \mathrm{d}B \xrightarrow{P_{A_0}} 0. \tag{18}$$

See Appendix A.2 for the proof.

**Remark 3.1:** Since we assume that the diagonal entries of $A$ are positive, the assumption of one-to-one correspondence between $\Sigma_0$ and $A_0$ is satisfied if rows and columns of $A$ can be permuted so that the resulting matrix is lower triangular matrix with positive entries on diagonal.

There are other highly sparse matrices for which there might be a finite number of different $A_{0,r}$ so that

$\Sigma_0 = A_{0,r}A_{0,r}^{\mathrm{T}}$. Of course in this case we cannot distinguish between these $A_{0,r}$ based on data. However, Theorem 3.1 will still be true if we restrict the domain of $A$ to a small enough Euclidean neighbourhood of any of the $A_{0,r}$. Each of these neighbourhoods being selected with a chance proportional to $\pi_{\Sigma_0}(A_{0,r})$.

### 3.4. Sampling in the general case

Given the true model $\mathcal{M}_0$, standard Markov chain Monte Carlo (MCMC) methods can be utilised for the estimation of the covariance matrix. Under the full model and clique model, the GFD of $\Sigma$ follows either an inverse Wishart distribution or a composite of inverse Wishart distributions (see Section 3). Sampling from the GFD becomes straight forward and it can be done through one of the inverse Wishart random generation functions, e.g., InvWishart (MCMCpack, R) or iwishrnd (Matlab).

When $p$ is small and $n$ is large, the estimation of $\Sigma$ can always be done through this setting, regardless if there are zero entries in $A$. The concept of having entries of $A$ fixed at zero is to impose sparsity structure and allow estimation under a high dimensional setting without requiring large number of observations. As in practice the true sparse structure is often unobserved, we will focus on the cases where $\mathcal{M}_0$ is not given.

For the general case, if the sparse model is unknown, we propose to utilise a reversible jump MCMC (RJMCMC) method to efficiently sample from Equation (20) and simultaneously update $\mathcal{M}$.

RJMCMC is an extension of standard Markov chain Monte Carlo methods that allows simulation of the target distribution on spaces of varying dimensions (Green, 1995). The 'jumps' refers to moves between models with possibly different parameter spaces. More details on RJMCMC can be found in Shi (2015). Since $\mathcal{M}$ is unknown, namely the number and the locations of fixed zeros in the matrix $A$ are unknown, the property of jumping between parameter spaces with different dimension is desired for estimating $\Sigma = AA^{\mathrm{T}}$. Because the search space for RJMCMC is both within parameter space and between spaces, it is known for slower convergence. To improve efficiency of the algorithm, we adapt the zeroth-order method (Brooks et al., 2003) and impose additional sparse constrains.

Assuming that there are fixed zeros in $A$, then for a $p \times p$ matrix $A$, the number needed to be estimated is less than $p^2$. If there are many fixed zeros, then this number is much smaller, hence the estimation is feasible even if the number of observations $n$ is less than $p$. In other words, the sparsity assumption on $A$ allows estimations under a large $p$ small $n$ setting. Suppose the zero entry locations of $A$ are known. The rest of $A$ can be obtain via standard MCMC techniques, such as Metropolis–Hastings.

Figure 1 considers a case with $p = 15$, $n = 30$. It shows the confidence curve plot per Markov chain for each statistic of interest. In addition to D2Sig, LogD

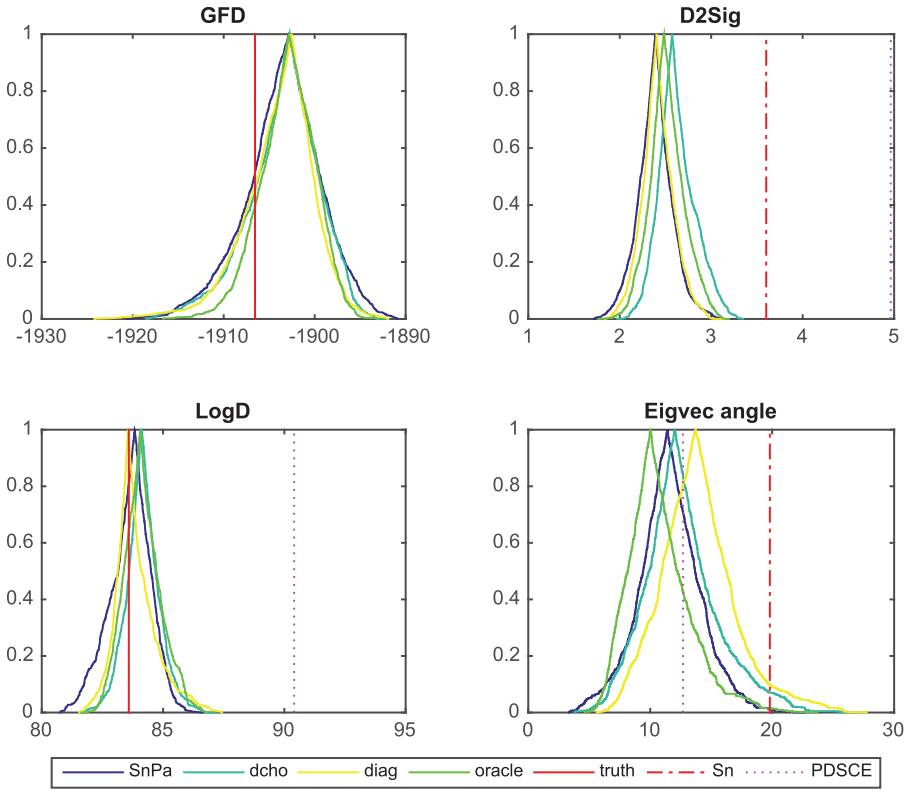


**Figure 1.** All the chains show good estimation of covariance matrix. The estimators are better than both the sample covariance matrix and the PDSCE estimator.

and Eigvec angle as before, we have GFD ($\log(r_p(A \mid \boldsymbol{y}))$ without the normalising constant). The initial states for the four Markov chains are SnPa ($S_n$ restricted to *maxC* (see Section 3.6), in blue), dcho (diagonal matrix of Cholesky decomposition, in cyan), diag (diagonal matrix of $S_n$, in yellow) and oracle (true $A$, in green). In addition, we include the statistics for $\Sigma$, $S_n$, and the PDSCE estimator in comparison with the confidence curves. They are shown as vertical lines as in the previous example.

The fiducial estimators have confidence curves peak around the truth in Panels GFD and LogD. In the right two panels, the (majority of) fiducial estimators lie on the left of the dotted-dashed lines, indicating that the estimators are closer to the truth than the sample covariance. The PDSCE estimator falls on the right edge of the Panel D2Sig shows that it is not as close to the truth. As before, the PDSCE estimator overestimates the covariance determinant. Here, burn in $=$ 5000, window $=$ 10,000.

### 3.5. Model selection for the general case

Often time in practice, to obtain enough statistical power or simply for feasibility, sparse covariates/covariances assumptions are imposed. The exact sparse structure is usually unknown, model selection is required to determine the appropriate parameter space.

Since GFD behaves like the likelihood function, in order to avoid over-fitting, a penalty term on the parameter space needs to be included in the model selection process (Hannig et al., 2016).

For the general case, we propose the following penalty function that is based on the Minimum Description Length (MDL) Rissanen (1978) for a model $\mathcal{M}$:

$$q_{\mathcal{M}}(n) = \exp\left\{-\sum_{i=1}^{p}\left[\frac{1}{2}p_i \log(np) + \log\binom{p}{p_i}\right]\right\}, \tag{19}$$

where $\mathcal{M}$ corresponds to a $p \times p$ matrix with $p_i$ many non-fixed-zero elements in its $i$th row, and $n$ is the number of observations.

The penalised GFD of $A$ is therefore

$$r_p(A \mid \mathcal{M}, \boldsymbol{y}) \propto r(A \mid \mathcal{M}, \boldsymbol{y}) \times \exp\left\{-\sum_{i=1}^{p}\left[\frac{1}{2}p_i \log(np) + \log\binom{p}{p_i}\right]\right\}. \tag{20}$$

### 3.6. Sampling in the general case with sparse locations unknown

In the general case with sparse locations unknown, we further assume that there is a maximum number of nonzeros per column allowed, denoted as *maxC*. This additional constraint can be viewed as each predictor only contribute to few tuples of the multivariate response. This assumption has been implemented to reduce the search space for RJMCMC. The starting states include MaxC ($S_n^{0.5}$, restricted to *maxC*, in blue) along with chol (in cyan), dcho (in artichoke), diag (in yellow) and true (in green) as before. We will revisit the example discussed in Section 3.4.

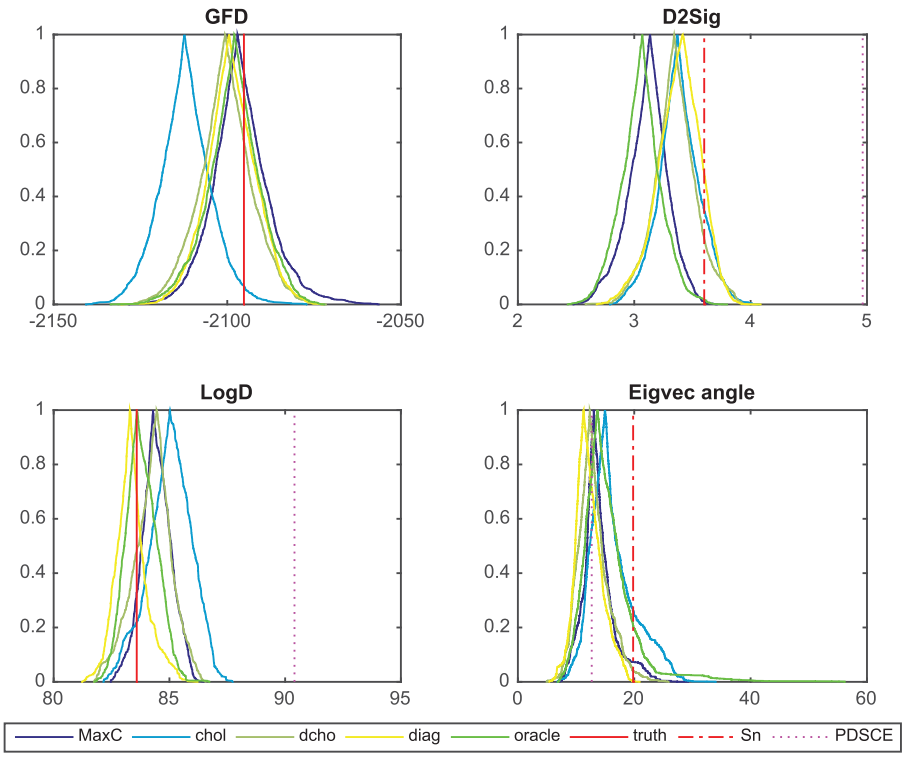


**Figure 2.** Similar to Figure 1, the fiducial estimators are better than both the sample covariance matrix and the PDSCE estimator in this case.

(See Figure 2). In the left two panels, the fiducial estimators peak at the true fiducial likelihood and covariance determinant. The distance comparison plot (top right) show that the estimators are closer to the truth than both the sample covariance matrix and the PDSCE estimator. Bottom right panel shows that the leading eigenvector of the estimators are as close to the truth as for sample covariance and the PDSCE estimator as in Figure 1. Here, burn in = 50,000, window = 10,000.

Additional simulations are included in the supplementary document.

## 4. Clique model

### 4.1. Jacobian for the clique model

Assume that the coordinates of $\boldsymbol{y}$ are broken into cliques, i.e., coordinates $i$ and $j$ are correlated if $i, j$ belong to the same clique and independent otherwise. By simply swapping rows and columns of the covariate matrix, we can arrive at a block diagonal form. Without loss of generality, suppose that $A$ is a block diagonal matrix with block sizes $g_1, \ldots, g_k$. Then its model $\mathcal{M}$ defines the parameter space $\otimes_{i=1}^{k} \mathbb{R}^{g_i \times g_i}$. Given $\mathcal{M}$, as an extension of the full model, the GFD function in this case becomes a composite of inverse Wishart distributions:

$$r(\Sigma \mid \boldsymbol{y}, \mathcal{M}) = \prod_{i=1}^{k} \frac{|nS_n^i|^{\frac{n}{2}}}{2^{\frac{ng_i}{2}} \Gamma_{g_i}\left(\frac{n}{2}\right)} |\Sigma^i|^{-\frac{n+g_i+1}{2}} \times \exp\left\{-\frac{1}{2}\mathrm{tr}\left(nS_n^i(\Sigma^i)^{-1}\right)\right\}, \quad (21)$$

where $S_n^i$ and $\Sigma^i$ are the sample covariance and covariance component of the $i$th clique, and $\Gamma_{g_i}(\cdot)$ is the $g_i$ dimensional multivariate gamma function.

### 4.2. Theoretic results for the clique models

Recall that under the full model,

$$r(A \mid \boldsymbol{y}) \propto |\det(S_n)|^{\frac{p}{2}} (2\pi)^{-\frac{np}{2}} |\det(A)|^{-(n+p)} \times \exp\left[-\frac{1}{2}\mathrm{tr}\{nS_n(AA^{\mathrm{T}})^{-1}\}\right].$$

For clique model selection, we need to evaluate the normalising constant.

$$\int J(\boldsymbol{y}, A) f(\boldsymbol{y} \mid A)\, \mathrm{d}A = \frac{\pi^{(p^2-np)/2} |\det(S_n)|^{\frac{p}{2}} \Gamma_p\left(\frac{n}{2}\right)}{|\det(nS_n)|^{n/2} \Gamma_p\left(\frac{p}{2}\right)}. \quad (22)$$

The detailed derivation is provided in Appendix A.3.

Let us denote by $\mathcal{M}$ a clique model; a collection of $k$ cliques – sets of indexes that are related to each other. The coordinates are assumed independent if they are not in the same cliques. For any positive-definite symmetric matrix $S$, whose dimension is compatible with $\mathcal{M}$, we denote $S^{\mathcal{M}}$ as the matrix obtained from $S$ by setting the off-diagonal entries that corresponds to pairs of indexes not in the same clique within $\mathcal{M}$ to zero. Note that $S^{\mathcal{M}}$ is a block diagonal (after possible permutations of rows and columns) positive-definite symmetric matrix.

The classical Fischer–Hadamard inequality (Fischer, 1908) implies that for any positive definite symmetric matrix $S$ and any clique model $\det(S) \leq \det(S^{\mathcal{M}})$. Ipsen Lee (2011) provides a useful lower bound. Let $\rho$ be the spectral radius and $\lambda$ be the smallest eigenvalue of $(S^{\mathcal{M}})^{-1}(S - S^{\mathcal{M}})$,

$$\mathrm{e}^{-\frac{p\rho^2}{1+\lambda}} \det(S^{\mathcal{M}}) \leq \det(S) \leq \det(S^{\mathcal{M}}). \quad (23)$$

Assume the clique sizes are $g_1, \ldots, g_k$. Then the GFD of the model is

$$r(\mathcal{M} \mid \boldsymbol{y}) \propto \frac{\pi^{\frac{\sum_{i=1}^{k} g_i^2}{2}}}{|\det S_n^{\mathcal{M}}|^{\frac{n}{2}}} \prod_{i=1}^{k} C_{\mathcal{M},i}(\boldsymbol{y}) \frac{\Gamma_{g_i}\left(\frac{n}{2}\right)}{\Gamma_{g_i}\left(\frac{g_i}{2}\right)}, \quad (24)$$

where $C_{\mathcal{M},i}(\boldsymbol{y})$ denotes the Jacobian constant term $|\det(S_{n,i})|^{\frac{g_i}{2}}$ computed only using the observations in the $i$th clique.

In the remaining part of this section, we consider the dimension of $\boldsymbol{y}$ as a fixed number $p$ and the sample size $n \to \infty$. Similar arguments could be extended to $p \to \infty$ with $p/\sqrt{n} \to 0$.

Given two clique models $\mathcal{M}_1$ and $\mathcal{M}_2$. We write $\mathcal{M}_1 \subset \mathcal{M}_2$ if cliques in $\mathcal{M}_2$ are obtained by merging cliques in $\mathcal{M}_1$. Consequently, $\mathcal{M}_2$ has fewer cliques and these cliques are larger than $\mathcal{M}_1$. Let $\mathcal{M}_0, \Sigma_0$ be the 'true' clique model and covariance matrix used to generate the observed data. We will call all the clique models $\mathcal{M}$ satisfying $\Sigma_0^{\mathcal{M}} = \Sigma_0$ compatible with the true covariance matrix. We assume that $\mathcal{M}_0 \subset \mathcal{M}$ for all clique models compatible with $\Sigma_0$.

The following theorem provides some guidelines for choosing penalty function $q_{\mathcal{M}}(n)$. Its proof is included in the appendix. Define the penalised GFD of the model as $r_p(\mathcal{M} \mid \boldsymbol{y}) = r(\mathcal{M} \mid \boldsymbol{y}) q_{\mathcal{M}}(n)$.

**Theorem 4.1:** *For any clique model $\mathcal{M}$ that is not compatible with $\Sigma_0$, assume $\det(\Sigma_0) < \det(\Sigma_0^{\mathcal{M}})$ and the penalty $\mathrm{e}^{-an} q_{\mathcal{M}}(n)/q_{\mathcal{M}_0}(n) \to 0$ for all $a > 0$ as $n \to \infty$.*

*For any clique model $\mathcal{M}$ compatible with $\Sigma_0$ assume that $q_{\mathcal{M}}(n)/q_{\mathcal{M}_0}(n)$ is bounded.*

*Then as $n \to \infty$ with $p$ held fixed $r_p(\mathcal{M}_0 | \boldsymbol{Y}) \xrightarrow{P} 1$.*

The exact form of the penalty function depends on the norm choice for the Jacobian. Under the $\ell_2$-norm,

the following penalty function works well.

$$q_{\mathcal{M}}(n) = \exp\left\{-\sum_{i=1}^{k}\left[\frac{1}{4}g_i^2\log(n) - \frac{1}{2}g_i^2\log(g_i)\right]\right\}. \tag{25}$$

It is easy to check that Equation (25) satisfies Theorem 4.1.

### 4.3. Sampling from a clique model

The estimation of cliques is closely related to applications in network analysis, such as communities of people in social networks and gene regulatory network. Recall the penalised clique model GFD introduced in Section 4.2,

$$r_p(\mathcal{M}\,|\,\boldsymbol{y}) \propto \frac{\pi^{\frac{\sum_{i=1}^{k} g_i^2}{2}}}{|\det S_n^{\mathcal{M}}|^{\frac{n}{2}}}\prod_{i=1}^{k} C_{\mathcal{M},i}(\boldsymbol{y})\frac{\Gamma_{g_i}\left(\frac{n}{2}\right)}{\Gamma_{g_i}\left(\frac{g_i}{2}\right)}q_{\mathcal{M}}(n).$$

Assuming that both the number of cliques $k$ and the clique sizes $g_k$'s are unknown, the clique structure can be estimated via Gibbs sampler. The first example shows the simulation result for a $200 \times 200$ covariance matrix (Figure 3). We consider the covariance matrix to be with 1's on the diagonal and $(i,j)$th entry being 0.5 if the coordinate $(i,j)$ belongs to a clique. From top down, left to right, Figure 3 shows the trace plot for $\log(r_p(\mathcal{M}\,|\,\boldsymbol{y}))$ without normalising constant, true covariance $\Sigma$, sample covariance $S_n$, and the fiducial probability of the estimated cliques based on the 10 Gibbs sampler Markov chains with random initial states. The trace plot helps to monitor the convergence. The fiducial probability of cliques panel reveals the clique structure precisely. The last panel is the aggregate result of 4000 iterations with burn in $=$ 1000 from the 10 Markov chains.

The covariance estimators can be obtained by sampling from inverse Wishart distributions based on the estimated clique structure. Figure 4 shows the confidence curves of four statistics for estimated covariance matrix $\hat{\Sigma}$: log-transformed generalised fiducial likelihood (SlogGFD), distance to $\Sigma$ (D2Sig), log-determinant (LogD), and angle between the leading eigenvectors of $\hat{\Sigma}$ and $\Sigma$ (Eigvec angle). The truth for SlogGFD and LogD is shown as red solid vertical lines. In D2Sig and Eigvec angle panels, we include comparisons to sample covariance as red dotted-dashed vertical lines. In addition, we compute the point estimation via the Positive Definite Sparse Covariance Estimators (PDSCE) method introduced in Rothman (2012). Its corresponding statistics are shown as magenta dotted vertical lines. In this example, the fiducial estimates peak near the truth in Panels SlogGFD and LogD. The estimated covariance matrices all appear to be more similar to $\Sigma$ than $S_n$ as shown in panels D2Sig and Eigvec angle. The PDSCE estimator is even closer to $\Sigma$ in terms of FM-distance; it however greatly overestimates $\det\Sigma$.

The PDSCE method produces a good point estimator to the covariance matrix. It is worth noting that our

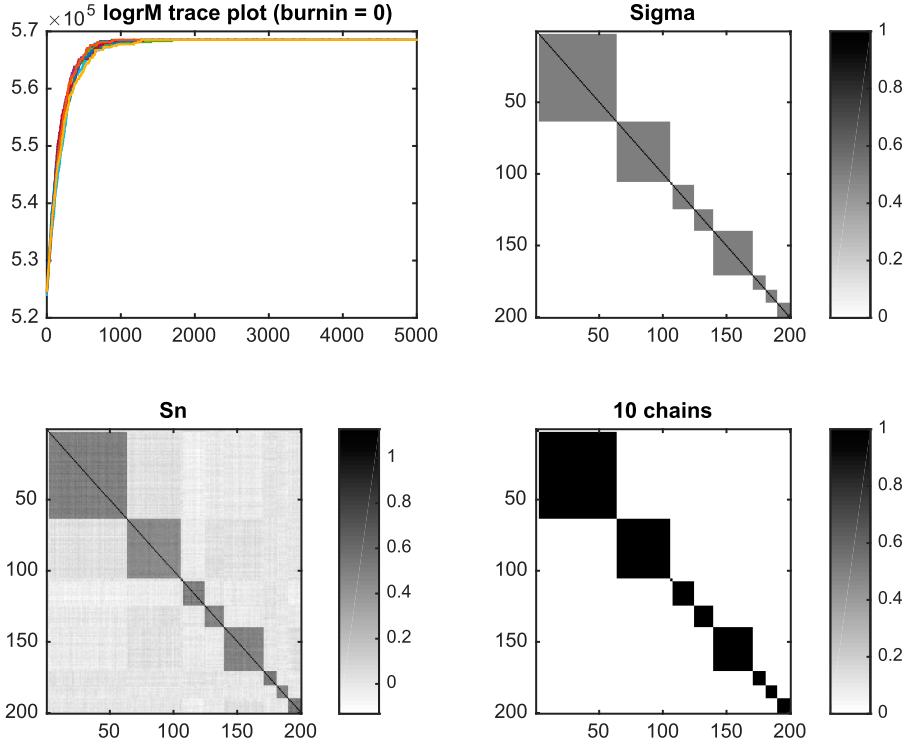


**Figure 3.** Result for $k = 10, p = 200, n = 1000$. The trace plot (top left) shows that the chains converge quickly. Although $\frac{n}{p}$ is small, the sample covariance (bottom left) roughly captures the shape of true covariance (top right). The last panel (bottom right) shows that the fiducial estimate captures the true clique structure perfectly.

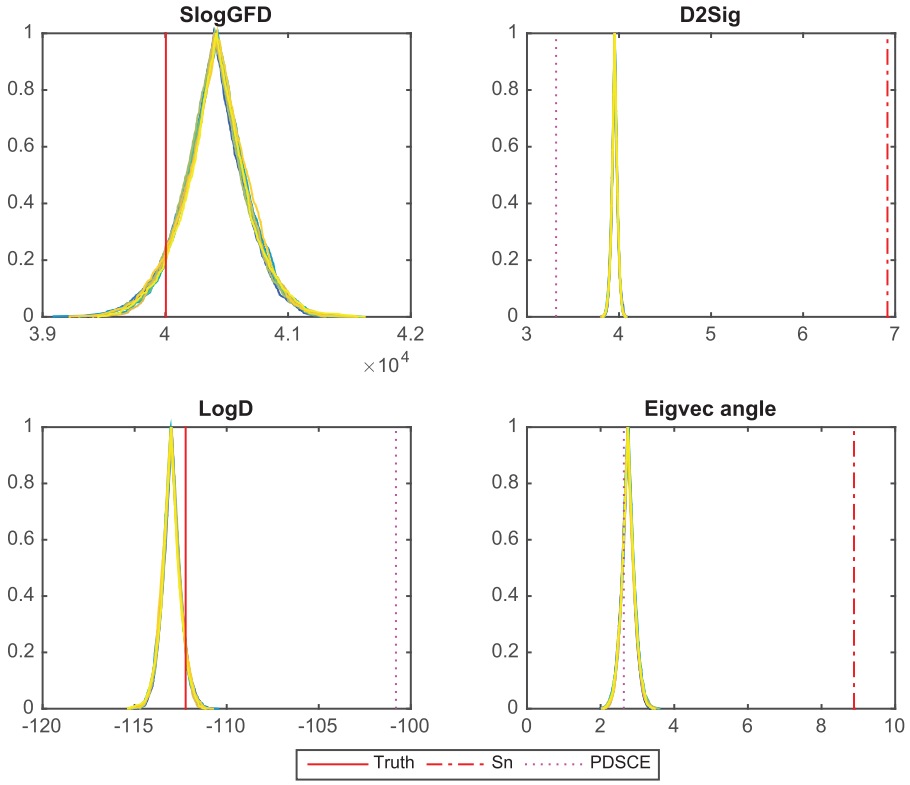


**Figure 4.** Confidence curve plots for estimated covariance matrix. k = 10,p = 200,n = 1000. Comparing to the sample covariance, the estimators are closer to $\Sigma$. The PDSCE estimator shows even smaller FM-distance to $\Sigma$, it, however, greatly overestimates $\det\Sigma$.

method shows similar performance with the benefit of producing a distribution of estimators.

With the same underlying clique model, we generate 200 data sets. Then we apply our method with a random Markov chain starting point and compute the one-sided $p$-values for the estimate covariance log determinant. With the same true covariance matrix, a new set of 1000 observations are generated for each simulation. Figure 5 shows the quantile–quantile plot of the $p$-values against the uniform [0,1] distribution. The dotted-dashed envelope is the 95% coverage band. It shows a well-calibrated 95% confidence interval. The $p$-value curve (in green) is well enclosed by the envelope, indicating good calibration of the coverage.

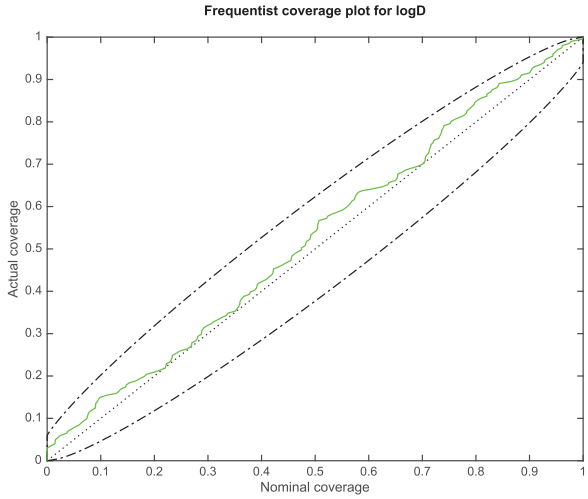


**Figure 5.** 95% coverage plots for 200 repeated simulations. k = 10,p = 200,n = 1000. The $p$-values (in green) roughly follow a uniform [0,1] distribution, and they lie inside of the envelope.

## 5. Discussion

Covariance estimation is an important problem in statistics. In this manuscript, we propose to look into this classical problem via a generalised fiducial approach. We demonstrate that, under mild assumptions, the GFD of the covariate matrix is asymptotically normal. In addition, we discuss the clique model and show that the fiducial approach is a powerful tool for identifying clique structures, even when the dimension of the parameter space is large and the ratio $n/p$ is small. To identify the covariance structure for non-clique models, in contrast to typical sparse covariance/precision matrix assumptions, we look at cases where the ratio $n/p$ is small and the covariate matrix is sparse. This 'unusual' sparsity assumption arises in applications where multiple dependent variables contribute to several response variables collaboratively. The fiducial approach allows us to obtain a distribution of covariance estimators that are better than sample covariance and comparable to the PDSCE estimator. The distances to true covariance matrix show that as dimension increases, the fiducial estimators become closer to the true covariance matrix.

Similar to Bayesian approaches, generalised fiducial inference produces a distribution of estimators, yet the two methods differ fundamentally. Bayesian methods rely on prior distributions on the parameter of interest, while fiducial approaches depend on the data generating equation. In the framework discussed here, the data generating mechanism is natural to establish than choosing appropriate priors while some other times priors are easier to construct.

Estimating sparse covariance matrix without knowing the fixed zeros is a hard problem. While our approach shows promising results for the clique model, for the general case it still suffers from a few drawbacks: (1) due to the nature of RJMCMC, the computational burden can be significant if the matrix is not very sparse; (2) to limit the search space, a row/column-wise sparsity upper bound needs to be chosen based on prior knowledge of the data type; (3) the results presented in this manuscript assume a squared covariate matrix, which can be limited to direct applications to high-throughput data. Furthermore, a more sophisticated way of choosing initial states and mixing method can improve the efficiency of our algorithm. It is possible and well worth it to extend our current work to more general cases.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

Shi's research was supported in part by the National Library of Medicine Institutional Training Grant T15 LM009451. Hannig's research was supported in part by the National Science Foundation (NSF) under Grant Nos. 1512945, 1633074, and 1916115. Lee's research was supported in part by the NSF under Grant No. 1512945 and 1513484.

## Notes on contributors

***W. Jenny Shi*** obtained her PhD in Statistics from the University of North Carolina. From 2015 to 2018, she was a National Institute of Health postdoctoral fellow at the University of Colorado. She is now a Quantitative Strategist at MassMutual, specializing in financial modeling and strategic initiatives.

***Jan Hannig*** received his Mgr (MS equivalent) in mathematics in 1996 from the Charles University, Prague, Czech Republic. He received Ph.D. in statistics and probability in 2000 from Michigan State University under the direction of Professor A.V. Skorokhod. From 2000 to 2008 he was on the faculty of the Department of Statistics at Colorado State University where he was promoted to an Associate Professor. He has joined the Department of Statistics and Operation Research at the University of North Carolina at Chapel Hill in 2008 and was promoted to Professor in 2013. He is an elected member of International Statistical Institute and a fellow of the American Statistical Association and Institute of Mathematical Statistics.

***Randy C. S. Lai*** obtained his Ph.D. in Statistics from the University of California, Davis (UC Davis). In 2015–2019 he was an Assistant Professor at the University of Maine. He is now a Visiting Assistant Professor at UC Davis, and he will join Google as a Data Scientist in Spring 2021. His research interests include fiducial inference and statistical computing.

***Thomas C. M. Lee*** is Professor of Statistics and Associate Dean for the Faculty in Mathematical and Physical Sciences at the University of California, Davis. He is an elected Fellow of the American Association for the Advancement of Science (AAAS), the American Statistical Association (ASA), and the Institute of Mathematical Statistics (IMS). From 2013 to 2015 he served as the editor-in-chief for the Journal of Computational and Graphical Statistics, and from 2015 to 2018 he served as the Chair of the Department of Statistics at UC Davis. His recent research interests include astrostatistics, fiducial inference, machine learning, and statistical image and signal processing.

## ORCID

*W. Jenny Shi* http://orcid.org/0000-0002-5564-0246
*Jan Hannig* http://orcid.org/0000-0002-4164-0173
*Randy C. S. Lai* http://orcid.org/0000-0002-4291-9256
*Thomas C. M. Lee* http://orcid.org/0000-0001-7067-405X

## References

Abramowitz, M., & Stegun, I. A. (1964). *Handbook of mathematical functions: with formula, graphs and mathematical tables*. Courier Corporation.

Avella-Medina, M., Battey, H. S., Fan, J., & Li, Q. (2018). Robust estimation of high-dimensional covariance and precision matrices. *Biometrika*, *105*, 271–284. https://doi.org/10.1093/biomet/asy011

Bickel, P. J., & Levina, E. (2008a). Covariance regularization by thresholding. *The Annals of Statistics*, *36*, 2577–2604. https://doi.org/10.1214/08-AOS600

Bickel, P. J., & Levina, E. (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics*, *36*, 199–227. https://doi.org/10.1214/009053607000000758

Brooks, S. P., Giudici, P., & Roberts, G. O. (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society*, *65*(1), 3–39. https://doi.org/10.1111/rssb.2003.65.issue-1

Cai, T., & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association*, *106*, 672–684. https://doi.org/10.1198/jasa.2011.tm10560

Cisewski, J., & Hannig, J. (2012). Generalized fiducial inference for normal linear mixed models. *The Annals of Statistics*, *40*(4), 2102–2127. https://doi.org/10.1214/12-AOS1030

Cui, Y., & Hannig, J. (2019). Nonparametric generalized fiducial inference for survival functions under censoring, with discussion and rejoinder by the author. *Biometrika*, *106*, 501–518. https://doi.org/10.1093/biomet/asz016

Fischer, E. (1908). Uber den hadamardschen determinantensatz. *Archiv D Math U Phys*, *13*, 32–40.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, *222*, 309–368. https://doi.org/10.1098/rsta.1922.0009

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, *26*, 528–535. https://doi.org/10.1017/S0305004100016297

Fisher, R. A. (1933). The concepts of inverse probability and fiducial probability referring to unknown parameters. *Proceedings of the Royal Society of London Series A*, *139*, 343–348.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *The Annals of Eugenics*, *6*, 91–98.

Förstner, W., & Moonen, B. (1999). A metric for covariance matrices. In *Quo vadis geodesia? Festschrift for Erik W. Grafarend on the occasion of his 60th birthday, Schriftenreihe der Institute des Studiengangs Geodäsie und Geoinformatik* (pp. 113–128). IAGB, Stuttgart.

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics (Oxford, England)*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Applications of lasso and grouped lass to estimation of sparse graphical models. *Technical Report*, Standford University.

Furrer, R., & Bengtsson, T. (2007). Estimation of high-dimensional prior and posterior covariance matrices in Kalman filter variants. *Journal of Multivariate Analysis*, *98*, 227–255. https://doi.org/10.1016/j.jmva.2006.08.003

Ghosh, J. K., & Ramamoorthi, R. V. (2003). *Bayesian nonparametrics*. Springer Series in Statistiscs. Springer-Verlag.

Glagovskiy, Y. S. (2006). *Construction of fiducial confidence intervals for the mixture of cauchy and normal distributions*(Master's thesis). Colorado State University.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. https://doi.org/10.1093/biomet/82.4.711

Hannig, J., Feng, Q., Iyer, H., Wang, C., & Liu, X. (2018). Fusion learning for inter-laboratory comparisons. *Journal of Statistical Planning and Inference*, *195*, 64–79. https://doi.org/10.1016/j.jspi.2017.09.011

Hannig, J., Iyer, H. K., Lai, R. C. S., & Lee, T. C. M. (2016). Generalized fiducial inference: a review and new results. *Journal of the American Statistical Association*, *111*, 1346–1361. https://doi.org/10.1080/01621459.2016.1165102

Hannig, J., Iyer, H. K., & Wang, J. C. M. (2007). Fiducial approach to uncertainty assessment: account for error due to instrument resolution. *Metrologia*, *44*, 476–483. https://doi.org/10.1088/0026-1394/44/6/006

Hannig, J., & Lee, T. C. M. (2009). Generalized fiducial inference for wavelet regression. *Biometrika*, *96*, 847–860. https://doi.org/10.1093/biomet/asp050

Hannig, J., Lidong, E., Abdel-Karim, A., & Iyer, H. K. (2006). Simultaneous fiducial generalized confidence intervals for ratios of means of lognormal distributions. *Austrian Journal of Statistics*, *35*, 261–269. https://doi.org/10.17713/ajs.v35i2&3.372

Hannig, J., Wang, J. C. M., & Iyer, H. K. (2003). Uncertainty calculation for the ratio of dependent measurements. *Metrologia*, *4*, 177–186. https://doi.org/10.1088/0026-1394/40/4/306

Huang, H.-C., & Lee, T. C. M. (2016). High-dimensional covariance estimation under the presence of outliers. *Statistics and Its Interface*, *9*, 461–468. https://doi.org/10.4310/SII.2016.v9.n4.a6

Huang, J. Z., Liu, N., Pourahmadi, M., & Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, *93*, 85–98. https://doi.org/10.1093/biomet/93.1.85

Ipsen, I. C., & Lee, D. J. (2011). Determinant approximations. *arXiv preprint* arXiv:1105.0437.

Iyer, H. K., Wang, C. M. J., & Mathew, T. (2004). Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, *99*, 1060–1071. https://doi.org/10.1198/016214504000001682

Jameson, G. (2013). Inequalities for gamma function ratios. *The American Mathematical Monthly*, *120*(10), 936–940. https://doi.org/10.4169/amer.math.monthly.120.10.936

Lai, R. C. S., Hannig, J., & Lee, T. C. M. (2015). Generalized fiducial inference for ultrahigh dimensional regression. *Journal of American Statistical Association*, *110*, 760–772. https://doi.org/10.1080/01621459.2014.931237

Lam, C., & Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, *37*, 42–54. https://doi.org/10.1214/09-AOS720

Levina, E., Rothman, A. J., & Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics*, *2*, 245–263. https://doi.org/10.1214/07-AOAS139

Li, X., Su, H., & Liang, H. (2018). Fiducial generalized *p*-values for testing zero-variance components in linear mixed-effects models. *Science China Mathematics*, *61*(7), 1303–1318. https://doi.org/10.1007/s11425-016-9068-8

Lidong, E., Hannig, J., & Iyer, H. K. (2008). Fiducial intervals for variance components in an unbalanced two-component normal mixed linear model. *Journal of the American Statistical Association*, *103*, 854–865. https://doi.org/10.1198/016214508000000229

Liu, Y., & Hannig, J. (2016). Generalized fiducial inference for binary logistic item response models. *Psychometrica*, *81*, 290–324. https://doi.org/10.1007/s11336-015-9492-7

Liu, Y., & Hannig, J. (2017). Generalized fiducial inference for logistic graded response models. *Psychometrica*, *82*, 1097–1125. https://doi.org/10.1007/s11336-017-9554-0

Martin, R., & Liu, C. (2015). *Inferential models: reasoning with uncertainty*. CRC Press.

Pourahmadi, M. (2011). Covariance estimation: the GLM and regularization perspectives. *Statistical Science*, *26*(3), 369–387. https://doi.org/10.1214/11-STS358

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–658. https://doi.org/10.1016/0005-1098(78)90005-5

Rothman, A. (2012). Positive definite estimators of large covariance matrices. *Biometrika*, *99*(3), 733–740. https://doi.org/10.1093/biomet/ass025

Rothman, A. J., Levina, E., & Zhu, J. (2009). Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, *104*, 177–186. https://doi.org/10.1198/jasa.2009.0101

Rothman, A. J., Levina, E., & Zhu, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika*, *97*, 539–550. https://doi.org/10.1093/biomet/asq022

Schweder, T., & Hjort, N. L. (2016). *Confidence, likelihood, probability*. Cambridge University Press.

Shi, W. J. (2015). *Bayesian modeling for viral sequencing and covariance estimation via fiducial inference* (PhD thesis). University of North Carolina at Chapel Hill.

Sonderegger, D., & Hannig, J. (2012). Bernstein–von Mises theorem for generalized fiducial distributions with application to free knot splines. *Preprint*.

van der Vaart, A. W. (1998). *Asymptotic statistics, volume 3 of Cambridge series in Statistical and Probabilistic Mathematics*. Cambridge University Press.

Wandler, D. V., & Hannig, J. (2011). Fiducial inference on maximum mean of a multivariate normal distribution. *Journal of Multivariate Analysis*, *102*, 87–104. https://doi.org/10.1016/j.jmva.2010.08.003

Wandler, D. V., & Hannig, J. (2012). A fiducial approach to multiple comparisons. *Journal of Statistical Planning and Inference*, *142*, 878–895. https://doi.org/10.1016/j.jspi.2011.10.011

Wandler, D. V., & Hannig, J. (2012). Generalized fiducial confidence intervals for extremes. *Extremes*, *15*, 67–87. https://doi.org/10.1007/s10687-011-0127-9

Wang, J. C. M., Hannig, J., & Iyer, H. K. (2012). Pivotal methods in the propagation of distributions. *Metrologia*, *49*, 382–389. https://doi.org/10.1088/0026-1394/49/3/382

Wang, J. C. M., & Iyer, H. K. (2005). Propagation of uncertainties in measurements using generalized inference. *Metrologia*, *42*, 145–153. https://doi.org/10.1088/0026-1394/42/2/010

Wang, J. C. M., & Iyer, H. K. (2006a). A generalized confidence interval for a measurand in the presence of type-a and type-b uncertainties. *Measurement*, *39*, 856–863. https://doi.org/10.1016/j.measurement.2006.04.011

Wang, J. C. M., & Iyer, H. K. (2006b). Uncertainty of analysis of vector measurands using fiducial inference. *Metrologia*, *43*, 486–494. https://doi.org/10.1088/0026-1394/43/6/002

Williams, J., & Hannig, J. (2018). Non-penalized variable selection in high-dimensional linear model settings via generalized fiducial inference. *Annals of Statistics*, *47*(3), 1723–1753. https://doi.org/10.1214/18-AOS1733

Wu, W. B., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, *90*, 831–844. https://doi.org/10.1093/biomet/90.4.831

Xie, M., & Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: a review. *International Statistical Review*, *81*, 3–39. https://doi.org/10.1111/insr.2013.81.issue-1

## Appendix

### A.1 Regularity conditions and Jacobian formula

Before proving the theorem on consistency of the GFD, we will first define the $\delta$-neighbourhood of $A_0$ and establish some regularity conditions on the likelihood function and Jacobian formula (Propositions A.1, A.2, 3.1).

**Definition A.1:** For a fixed covariate matrix $A_0$ and $\delta \geq 0$, define the $\delta$-neighbourhood of $A_0$ as the set $B(A_0, \delta) = \{A : \mathbf{d}(AA^{\mathrm{T}}, A_0A_0^{\mathrm{T}}) \leq \delta\}$. Recall that $\mathbf{d}$ is the FM-distance (17).

**Proposition A.1:** *For any $\delta > 0$ there exists $\epsilon > 0$ such that*

$$P_{A_0}\left\{\sup_{A\notin B(A_0,\delta)} \frac{1}{n}(L_n(A) - L_n(A_0)) \leq -\epsilon\right\} \to 1,$$

*where $L_n(A) = \log f(\boldsymbol{y}, A) = \sum_{i=1}^n \log f(y_i, A)$.*

***Proof:*** Let $\Sigma = AA^{\mathrm{T}}$, $\Sigma_0 = A_0A_0^{\mathrm{T}}$. Denote $S_n$ as the sample covariance matrix as before, $n \in \mathbb{N}$. Since $S_n$ is the maximum likelihood estimator, we have

$$S_n \xrightarrow{P_{A_0}} \Sigma_0, \quad \text{i.e.,}$$
$$\forall\, r > 0, \quad P_{A_0}(\{\omega : \mathbf{d}(S_n(\omega), \Sigma_0) \geq r\}) \to 0.$$

Define $L_{\delta,n} = \{\omega : \mathbf{d}(S_n(\omega), \Sigma_0) < \delta/2\}$. For an arbitrary $\omega \in L_{\delta,n}$, assume that $\lambda_i^{\dagger}$'s and $\lambda_i^{*}$'s are the eigenvalues of $S_n(\omega)\Sigma^{-1}$ and $S_n(\omega)\Sigma_0^{-1}$, respectively. Suppose that $A \notin B(A_0, \delta)$, then

$$\begin{aligned}\delta < \mathbf{d}(\Sigma, \Sigma_0) &\leq \mathbf{d}(\Sigma, S_n(\omega)) + \mathbf{d}(S_n(\omega), \Sigma_0)\\ &< \mathbf{d}(\Sigma, S_n(\omega)) + \delta/2\\ &\Rightarrow \mathbf{d}(\Sigma, S_n(\omega)) = \sqrt{\sum_{i=1}^p \log^2 \lambda_i^{\dagger}} > \delta/2.\end{aligned}$$

So there exists $k \in \{1, 2, \ldots, p\}$, such that $\ln^2 \lambda_k^{\dagger} > \frac{\delta^2}{4p}$, then

$$\begin{aligned}\ln \lambda_k - \lambda_k < \max\Bigg\{&\frac{\delta}{2\sqrt{p}} - \exp\left(\frac{\delta}{2\sqrt{p}}\right),\\ &-\frac{\delta}{2\sqrt{p}} - \exp\left(-\frac{\delta}{2\sqrt{p}}\right)\Bigg\} := m_\delta\end{aligned}$$

due to the fact that the function $g(\lambda) = \ln \lambda - \lambda$ is concave with unique maxima $\lambda = 1; g(1) = -1$.

Meanwhile,

$$\begin{aligned}&\frac{1}{n}(L_n(A) - L_n(A_0))(\omega)\\ &\quad= -\ln|\det(A)| - \frac{1}{2}\mathrm{tr}\{S_n(\omega)\Sigma^{-1}\}\\ &\qquad+ \ln|\det(A_0)| + \frac{1}{2}\mathrm{tr}\{S_n(\omega)\Sigma_0^{-1}\}\\ &\quad= \frac{1}{2}\ln(S_n(\omega)\Sigma^{-1}) - \frac{1}{2}\mathrm{tr}\{S_n(\omega)\Sigma^{-1}\}\\ &\qquad- \frac{1}{2}\ln(S_n(\omega)\Sigma_0^{-1}) + \frac{1}{2}\mathrm{tr}\{S_n(\omega)\Sigma_0^{-1}\}\\ &\quad= \frac{1}{2}\left\{\sum_{i=1}^p(\ln\lambda_i^{\dagger} - \lambda_i^{\dagger}) - \sum_{i=1}^p(\ln\lambda_i^{*} - \lambda_i^{*})\right\}\\ &\quad< \frac{1}{2}\left\{-(p-1) + m_\delta + p\right\}\\ &\quad= \frac{1}{2}(m_\delta + 1).\end{aligned}$$

This implies

$$\sup_{A\notin B(A_0,\delta)} \frac{1}{n}(L_n(A) - L_n(A_0))(\omega) \leq \frac{1}{2}(m_\delta + 1) < 0.$$

Let $\epsilon = -\frac{1}{2}(m_\delta + 1)$, $U_{\delta,n} = \{\omega : \sup_{A\notin B(A_0,\delta)} \frac{1}{n}(L_n(A) - L_n(A_0))(\omega) \leq -\epsilon\}$. Then $L_{\delta,n} \subseteq U_{\delta,n}$. Notice that

$$\begin{aligned}1 &= \lim_{n\to\infty} P_{A_0}(L_{\delta,n}) = \liminf_{n\to\infty} P_{A_0}(L_{\delta,n})\\ &\leq \liminf_{n\to\infty} P_{A_0}(U_{\delta,n}) \leq \limsup_{n\to\infty} P_{A_0}(U_{\delta,n}) \leq 1.\end{aligned}$$

Therefore, $\lim_{n\to\infty} P_{A_0}(U_{\delta,n}) = 1$. ∎

**Proposition A.2:** *Let $L_n(\cdot)$ be as above. Then for any $\delta > 0$*

$$\inf_{A\notin B(A_0,\delta)} \frac{\min_{\mathbf{i}=\{i_1,\ldots,i_p\}\atop 1\leq i_1<\cdots<i_p\leq n} \log f(A, \boldsymbol{y}_{\mathbf{i}})}{|L_n(A) - L_n(A_0)|} \xrightarrow{A_0} 0,$$

*where $f(A, \boldsymbol{y}_{\mathbf{i}})$ is the joint likelihood of $p$ observations $y_{i_1}, \ldots, y_{i_p}$.*

***Proof:*** Note that

$$\inf_{A\notin B(A_0,\delta)} \frac{\min_{\mathbf{i}=\{i_1,\ldots,i_p\}\atop 1\leq i_1<\cdots<i_p\leq n} \log f(A, \boldsymbol{y}_{\mathbf{i}})}{|L_n(A) - L_n(A_0)|}$$

$$\leq \frac{\inf_{A\notin B(A_0,\delta)} \min_{\substack{\mathbf{i}=\{i_1,\ldots,i_p\}\\ 1\leq i_1<\cdots<i_p\leq n}} \log f(A,\boldsymbol{y}_{\mathbf{i}})}{\inf_{A\notin B(A_0,\delta)} |L_n(A)-L_n(A_0)|}.$$

For any $A \notin B(A_0,\delta)$, denote $\Sigma = AA^{\mathrm{T}}$, $\Sigma_0 = A_0A_0^{\mathrm{T}}$ and let $t>0$, we have

$$\begin{aligned}
&P_{A_0}\left(\min_{\substack{\mathbf{i}=\{i_1,\ldots,i_p\}\\ 1\leq i_1<\cdots<i_p\leq n}} \log f(A,\boldsymbol{y}_{\mathbf{i}}) \leq -t\log n\right)\\
&\leq P_{A_0}\left(\min_{i=1,\ldots,n} \log f(A,Y_i) \leq -\frac{t\log n}{p}\right)\\
&= 1-\left[1-P_{A_0}\left(-\log f(A,Y_i)\geq -\frac{t\log n}{p}\right)\right]^n\\
&\leq 1-\left[1-\frac{p\mathrm{E}_{A_0}(-\log f(A,Y_i))}{t\log n}\right]^n \text{ (Markov inequality)}\\
&= 1-\left[1-\frac{p(\log(2\pi)+\log\det(\Sigma)+\mathrm{tr}\{\Sigma^{-1}\Sigma_0\})}{2t\log n}\right]^n\\
&\to 0, \quad \text{as } n\to\infty.
\end{aligned}$$

Note that the numerator goes to $-\infty$ at most as fast as $-t\log n$. Meanwhile, for a fixed $n$ and any $\omega\in L_{\delta,n}=\{\omega: \mathbf{d}(S_n(\omega),\Sigma_0)<\delta/2\}$,

$$\begin{aligned}
&\inf_{A\notin B(A_0,\delta)} |L_n(A)-L_n(A_0)|\\
&= -\sup_{A\notin B(A_0,\delta)} L_n(A)-L_n(A_0)\geq \epsilon n.
\end{aligned}$$

By Proposition (A.1),

$$\lim_{n\to\infty} P_{A_0}\left(\inf_{A\notin B(A_0,\delta)} |L_n(A)-L_n(A_0)|\geq \epsilon n\right)=1,$$

i.e., the denominator goes to infinity at least as fast as $\epsilon n$. ■

***Proof of Proposition 3.1:*** Given an ordered index vector $\mathbf{r}=(r_1,\ldots,r_l)$, let $E_{\mathbf{r}}=(e_{r_1};\cdots;e_{r_l})$, where each $e_{r_j}$ is a $1\times p$ vector with 1 in the $r_j$th tuple and 0 everywhere else. Denote $-\mathbf{r}=\{1,\ldots,p\}\setminus\mathbf{r}$.

Under the $\ell_2$-norm,

$$\begin{aligned}
J(\boldsymbol{y},A) &= \sqrt{\prod_{i=1}^{p}\det(U_i^{\mathrm{T}}U_i/n)}\\
&= \sqrt{\prod_{i=1}^{p}\det\left(E_{-S_i}^{\mathrm{T}}A^{-1}S_n(A^{-1})^{\mathrm{T}}E_{-S_i}\right)},
\end{aligned}$$

where $S_i$ is the list of indexes of fixed zeros in the $i$th row of $A$.

By the Strong Law of Large Numbers for $S_n$ and continuity of $J(\boldsymbol{y},A)$,

$$\begin{aligned}
J(\boldsymbol{y},A) &\longrightarrow \sqrt{\prod_{i=1}^{p}\det\left(E_{-S_i}^{\mathrm{T}}A^{-1}\Sigma_0(A^{-1})^{\mathrm{T}}E_{-S_i}\right)}\\
&:= \pi_{\Sigma_0}(A) \quad \text{a.s.}
\end{aligned}$$

Note that both $P_n=J(\boldsymbol{y},A)$ and $P_0=\pi_{\Sigma_0}(A)$ are polynomials of entries of $A^{-1}$. If the domain of $A$ is in compact, the coefficients of $P_n$ converge to the coefficients of $P_0$ uniformly. Furthermore, the derivative is bounded, hence $P_n$ is equicontinuous. We have $J(\boldsymbol{y},A)\xrightarrow{\text{a.s.}}\pi_{\Sigma_0}(A)$ uniformly on compacts in $A$. ■

## A.2 Proof of Theorem 3.1

***Proof:*** Proposition 3.1 implies

$$\sup_{A\in B(A_0,\delta)} |J(\boldsymbol{y},A)-\pi_{\Sigma_0}(A)|\to 0 \quad \text{a.s. } P_{A_0}.$$

$$\begin{aligned}
\pi^*(B,\boldsymbol{y}) &= \frac{J\left(\boldsymbol{y},\hat{A}_n+\frac{B}{\sqrt{n}}\right)f\left(\boldsymbol{y}|\hat{A}_n+\frac{B}{\sqrt{n}}\right)}{\int_{\mathbb{R}^{p^2}} J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)f\left(\boldsymbol{y}\,|\,\hat{A}_n+\frac{C}{\sqrt{n}}\right)\mathrm{d}C}\\
&= \frac{J\left(\boldsymbol{y},\hat{A}_n+\frac{B}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{A}_n+\frac{B}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]}{\int_{\mathbb{R}^{p^2}} J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C}.
\end{aligned}$$

Notice that

$$H=-\frac{1}{n}\frac{\partial^2}{\partial A\partial A}(\hat{A}_n)\to I(A_0) \quad \text{a.s. } P_{A_0}.$$

It suffices to show that

$$\begin{aligned}
&\int_{\mathbb{R}^{p^2}}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\right.\\
&\left.-\pi_{\Sigma_0}(A_0)\exp\left[\frac{-C^TI(A_0)C}{2}\right]\right|\mathrm{d}C\xrightarrow{P_{A_0}}0.
\end{aligned}\tag{A1}$$

Let $C_x$ be the $ij$th entry of C, where $x=i+(p-1)j$. By Taylor Theorem,

$$\begin{aligned}
L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right) &= L_n(\hat{A}_n)+\sum_{x=1}^{p^2}\left(\frac{C_x}{\sqrt{(n)}}\right)\frac{\partial}{\partial A_x}L_n(\hat{A}_n)\\
&\quad+\frac{1}{2}\sum_{x=1}^{p^2}\sum_{y=1}^{p^2}\left(\frac{C_xC_y}{(\sqrt{(n)})^2}\right)\\
&\quad\times\frac{\partial^2}{\partial A_x\partial A_y}L_n(\hat{A}_n)\\
&\quad+\frac{1}{6}\sum_{x=1}^{p^2}\sum_{y=1}^{p^2}\sum_{z=1}^{p^2}\left(\frac{C_xC_yC_z}{(\sqrt{(n)})^3}\right)\\
&\quad\times\frac{\partial^3}{\partial A_x\partial A_y\partial A_z}L_n(A')\\
&= L_n(\hat{A}_n)-\frac{C^{\mathrm{T}}HC}{2}+R_n
\end{aligned}$$

for some $A'\in[\hat{A}_n,\hat{A}_n+\frac{C}{\sqrt{n}}]$. Notice that $R_n=\mathcal{O}p(n^{-3/2}\times||C||)$. Given any $0<\delta<\delta_0$ and $t>0$, the parameter space $\mathbb{R}^{p^2}$ can be partitioned into three regions:

$$\begin{aligned}
S_1 &= \{C: ||C||<t\log\sqrt{n}\};\\
S_2 &= \{C: t\log\sqrt{n}<||C||<\delta\sqrt{n}\};\\
S_3 &= \{C: ||C||>\delta\sqrt{n}\}.
\end{aligned}$$

On $S_1\cup S_2$,

$$\int_{S_1\cup S_2}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\right.$$

$$
\begin{aligned}
&-\pi_{\Sigma_0}(A_0)\exp\left[\frac{-C^{\mathrm{T}}I(A_0)C}{2}\right]\Bigg|\,\mathrm{d}C\\
&\le \int_{S_1\cup S_2}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)-\pi_{\Sigma_0}\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right|\\
&\quad\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C\\
&\quad+\int_{S_1\cup S_2}\left|\pi_{\Sigma_0}\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right.\\
&\quad\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\\
&\quad-\pi_{\Sigma_0}(A_0)\exp\left[\frac{-C^{\mathrm{T}}I(A_0)C}{2}\right]\Bigg|\,\mathrm{d}C.
\end{aligned}
$$

Since $\pi_{\Sigma_0}(\cdot)$ is a proper prior on the region $S_1\cup S_2$, the second term goes to zero by the Bayesian Bernstein–von Mises Theorem (see the proof of Theorem 1.4.2 in Ghosh and Ramamoorthi (2003)).

Next we notice that

$$
\begin{aligned}
&\int_{S_1\cup S_2}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)-\pi_{A_0}\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right|\\
&\qquad\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C\\
&\quad\le\sup_{C\in S_1\cup S_2}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)-\pi_{A_0}\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right|\\
&\qquad\times\int_{S_1\cup S_2}\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C.
\end{aligned}
$$

Since $\sqrt{n}(\hat{A}_n-A_0)\xrightarrow{\mathcal{D}}N(0,I(A_0)^{-1})$, we have

$$
P_{A_0}\left[\left\{\hat{A}_n+\frac{C}{\sqrt{n}};\ C\in S_1\cup S_2\right\}\subset B(A_0,\delta_0)\right]\to 1.
$$

Furthermore,

$$
L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n\left(\hat{A}_n\right)=-\frac{C^{\mathrm{T}}HC}{2}+R_n,
$$

so the integral converges in probability to 1. Since $\max_{C\in S_1\cup S_2}\le\delta$ and $J_n\to\pi_{\Sigma_0}$, the term goes to 0 in probability.

Turning our attention to $S_3$, notice that

$$
\begin{aligned}
&\int_{S_3}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\right.\\
&\qquad\left.-\pi_{\Sigma_0}(A_0)\exp\left[\frac{-C^{\mathrm{T}}I(A_0)C}{2}\right]\right|\mathrm{d}C\\
&\le\int_{S_3}J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\\
&\qquad\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C\\
&\qquad+\int_{S_3}\pi_{\Sigma_0}(A_0)\exp\left[\frac{-C^{\mathrm{T}}I(A_0)C}{2}\right]\mathrm{d}C.
\end{aligned}
$$

The last integral goes to zero in $P_{A_0}$ because $\min_{S_3}||C||\to\infty$.

For each $\boldsymbol{y}$, let $\mathbf{i}$ be

$$
\begin{aligned}
\mathbf{i}&=\operatorname*{argmin}_{\tilde{\mathbf{i}}}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)-J\left(\boldsymbol{y}_{\tilde{\mathbf{i}}},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right|\\
&=\operatorname*{argmin}_{\tilde{\mathbf{i}}}h(\boldsymbol{y},C,\tilde{\mathbf{i}}).
\end{aligned}
$$

$$
\begin{aligned}
&\int_{S_3}J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\mathrm{d}C\\
&\quad\le\int_{S_3}\left\{h(\boldsymbol{y},C,\mathbf{i})f\left(\boldsymbol{y}_{\mathbf{i}}|\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right.\\
&\qquad\left.+J\left(\boldsymbol{y}_{\mathbf{i}},\hat{A}_n\frac{C}{\sqrt{n}}\right)f\left(\boldsymbol{y}_{\mathbf{i}}|\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right\}\\
&\qquad\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right.\\
&\qquad\left.-L_n(\hat{A}_n)-\log f\left(\boldsymbol{y}_{\mathbf{i}}|\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right]\mathrm{d}C.
\end{aligned}
$$

Note that as $n$ goes to infinity, the first two product terms, $h(\cdot)f(\cdot)$ and $J(\cdot)f(\cdot)$, are both bounded; the exponent term goes to $-\infty$ by Proposition A.2, so the integral goes to zero in probability.

Having shown Equation (A1), we now follow Ghosh and Ramamoorthi (2003) and let

$$
\begin{aligned}
D_n=\int_{\mathbb{R}^{p^2}}&\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{C}{\sqrt{n}}\right)\right.\\
&\left.\times\exp\left[L_n\left(\hat{A}_n+\frac{C}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\right|\mathrm{d}C.
\end{aligned}
$$

Then the main result to be proven Equation (18) becomes

$$
\begin{aligned}
&D_n^{-1}\left\{\int_{\mathbb{R}^{p^2}}\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{B}{\sqrt{n}}\right)\right.\right.\\
&\quad\times\exp\left[L_n\left(\hat{A}_n+\frac{B}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\\
&\quad\left.\left.-D_n\frac{\sqrt{\det(I(A_0))}}{(2\pi)^p}\exp\left(-\frac{B^{\mathrm{T}}I(A_0)B}{2}\right)\right|\right\}\mathrm{d}B\xrightarrow{P_{A_0}}0.
\end{aligned}
\tag{A2}
$$

Because

$$
\begin{aligned}
&\int_{\mathbb{R}^{p^2}}J(\boldsymbol{y},\hat{A}_n)\exp\left(-\frac{B^{\mathrm{T}}I(A_0)B}{2}\right)\mathrm{d}B\\
&\quad=J(\boldsymbol{y},\hat{A}_n)\int_{\mathbb{R}^{p^2}}\exp\left(-\frac{B^{\mathrm{T}}I(A_0)B}{2}\right)\mathrm{d}B\\
&\quad=J(\boldsymbol{y},\hat{A}_n)\frac{(2\pi)^p}{\sqrt{\det(H)}}\\
&\quad\xrightarrow{\text{a.s.}}\pi(A_0)\frac{(2\pi)^p}{\sqrt{\det(H)}},
\end{aligned}
$$

and (A1) implies that $D_n\xrightarrow{P}\pi(A_0)\frac{(2\pi)^p}{\sqrt{\det(H)}}$. It is sufficient to show that the integral in Equation (A2) goes to 0 in probability. This integral is less than $I_1+I_2$, where

$$
\begin{aligned}
I_1=\int_{\mathbb{R}^{p^2}}&\left|J\left(\boldsymbol{y},\hat{A}_n+\frac{B}{\sqrt{n}}\right)\right.\\
&\times\exp\left[L_n\left(\hat{A}_n+\frac{B}{\sqrt{n}}\right)-L_n(\hat{A}_n)\right]\\
&\left.-J\left(\boldsymbol{y},\hat{A}_n\right)\exp\left(-\frac{B^{\mathrm{T}}I(A_0)B}{2}\right)\right|\mathrm{d}B
\end{aligned}
$$

and

$$
\begin{aligned}
I_2=\int_{\mathbb{R}^{p^2}}&\left|J\left(\boldsymbol{y},\hat{A}_n\right)\exp\left(-\frac{B^{\mathrm{T}}HB}{2}\right)\right.\\
&\left.-D_n\frac{\sqrt{\det(I(A_0))}}{(2\pi)^p}\exp\left(-\frac{B^{\mathrm{T}}I(A_0)B}{2}\right)\right|\mathrm{d}B.
\end{aligned}
$$

Equation (A1) shows that $I_1\to 0$ in probability.

Since

$$J(\boldsymbol{y}, \hat{A}_n) \xrightarrow{P} \pi(A_0) \quad \text{and} \quad D_n \xrightarrow{P} \pi(A_0)\frac{(2\pi)^p}{\sqrt{\det(I(A_0))}},$$

we have

$$I_2 = \left| J\left(\boldsymbol{y}, \hat{A}_n\right) - D_n \frac{\sqrt{\det(I(A_0))}}{(2\pi)^p} \right| \times \int_{\mathbb{R}^{p^2}} \exp\left(-\frac{B^{\mathrm{T}} H B}{2}\right) \mathrm{d}B \xrightarrow{P} 0.$$

■

### A.3 Derivation of the normalising constant (22)

Using a substitution $A^{-1}(nS_n)^{1/2} = \boldsymbol{Z}$ with the Jacobian $\mathrm{d}A = |\det \boldsymbol{Z}|^{-2p} |\det(nS_n)|^{p/2}\, \mathrm{d}\boldsymbol{Z}$ we have

$$\begin{aligned}
&\int J(\boldsymbol{y}, A) f(\boldsymbol{y}|A)\, \mathrm{d}A \\
&= |\det(S_n)|^{\frac{p}{2}} \int \frac{\mathrm{e}^{-\frac{1}{2}\operatorname{tr}(A^{-1}(nS_n)^{1/2})(A^{-1}(nS_n)^{1/2})^{\top}}}{(2\pi)^{np/2}|\det A|^{n+p}}\, \mathrm{d}A \\
&= |\det(S_n)|^{\frac{p}{2}} \int |\det \boldsymbol{Z}|^{n-p} |\det(nS_n)|^{-n/2} \mathrm{e}^{-\frac{1}{2}\operatorname{tr}\boldsymbol{Z}\boldsymbol{Z}^{\top}}\, \mathrm{d}\boldsymbol{Z} \\
&= (2\pi)^{-(n-p)p/2} |\det(S_n)|^{\frac{p}{2}} |\det(nS_n)|^{-n/2} E|\det \boldsymbol{Z}|^{n-p} \\
&= \frac{\pi^{(p^2-np)/2} |\det(S_n)|^{\frac{p}{2}} \Gamma_p\left(\frac{n}{2}\right)}{|\det(nS_n)|^{n/2} \Gamma_p\left(\frac{p}{2}\right)}.
\end{aligned}$$

The last equality follows from the fact that for a $p \times p$ matrix of independent standard normal normal variables $Z$ we have

$$E|\det \boldsymbol{Z}|^n = \frac{2^{np/2}\Gamma_p\left(\frac{n+p}{2}\right)}{\Gamma_p\left(\frac{p}{2}\right)}.$$

### A.4 Lemmas for the clique model

**Lemma A.1:** *Under the $\ell_2$-norm, for any clique model $\mathcal{M}$ with $k$ cliques of sizes $g_i$, $i = 1, \ldots, k$, we have*

$$C_{\mathcal{M},i}(\boldsymbol{y}) = |\det(S_n^{\mathcal{M},i})|^{g_i/2} \to |\det(\Sigma_0^{\mathcal{M},i})|^{\frac{g_i}{2}} \quad \text{a.s.},$$

*where $S_n^{\mathcal{M},i}$ is the sample covariance computed using only observations within clique $i$ under the model $\mathcal{M}$, and $\Sigma_0^{\mathcal{M},i}$ denotes the ith block component of $\Sigma_0^{\mathcal{M}}$.*

***Proof:*** The Strong Law of Large Numbers implies $S_{n,i}^{\mathcal{M}} \to \Sigma_0^{\mathcal{M},i}$ a.s. for each $i = 1, \ldots, k$ and the results follow by continuity. ■

Lemma A.1 provides the limits of the constant $C_{\mathcal{M},i}(\boldsymbol{y})$ as sample size increases. The next lemma shows how the ratio $\frac{\prod_{i=1}^k \Gamma_{g_i}(\frac{n}{2})}{\prod_{j=1}^l \Gamma_{h_j}(\frac{n}{2})}$ behaves when sample size increases.

**Lemma A.2:** *Let $g_i, i = 1, \ldots, k$ and $h_j, j = 1, \ldots, l$ be integers such that $\sum_{i=1}^k g_i = \sum_{j=1}^l h_i$. Then as $n \to \infty$*

$$\frac{\prod_{i=1}^k \Gamma_{g_i}\left(\frac{n}{2}\right)}{\prod_{j=1}^l \Gamma_{h_j}\left(\frac{n}{2}\right)} \sim \left(\frac{\pi}{n}\right)^{\frac{\sum_{i=1}^k g_i^2 - \sum_{j=1}^l h_j^2}{4}}.$$

***Proof:*** It is well known (Abramowitz & Stegun, 1964) that

$$\frac{\Gamma(x+y)}{\Gamma(x)} \sim x^y, \quad \text{as } x \to \infty \quad \text{and} \quad y \text{ is fixed.} \tag{A3}$$

Recall

$$\frac{\prod_{i=1}^k \Gamma_{g_i}\left(\frac{n}{2}\right)}{\prod_{j=1}^l \Gamma_{h_j}\left(\frac{n}{2}\right)} = \frac{\pi^{\sum_{i=1}^k (g_i^2 - g_i)/4} \prod_{i=1}^k \prod_{s=1}^{g_i} \Gamma\left(\frac{n+1-s}{2}\right)}{\pi^{\sum_{i=1}^l (h_i^2 - h_i)/4} \prod_{j=1}^l \prod_{t=1}^{h_j} \Gamma\left(\frac{n+1-t}{2}\right)}.$$

Since both numerator and denominator include a product of $p$ gamma functions, the result of the lemma then follows directly from Equation (A3). Note that Equation (A3) will be sufficient when $p$ is fixed. More precise bounds available in Jameson (2013) could be used when $p$ is growing with $n$. ■

**Lemma A.3:** *Let $\mathcal{M}$ be a clique model.*

(i) *If $\det(\Sigma_0) < \det(\Sigma_0^{\mathcal{M}})$, then there is $a > 0$, such that*

$$\left| \frac{\det(S_n^{\mathcal{M}_0})}{\det(S_n^{\mathcal{M}})} \right|^{n/2} \leq \mathrm{e}^{-an} \quad \textit{eventually a.s.}$$

(ii) *If $\mathcal{M} \neq \mathcal{M}_0$ is compatible with $\Sigma_0$, then as $n \to \infty$*

$$\left| \frac{\det(S_n^{\mathcal{M}_0})}{\det(S_n^{\mathcal{M}})} \right|^{n/2} = \mathcal{O}_P(1).$$

***Proof:*** If $\det(\Sigma_0) < \det(\Sigma_0^{\mathcal{M}})$, set $a = \frac{\log\det\Sigma_0^{\mathcal{M}} - \log\det\Sigma_0}{4}$. By the Strong Law of Large Numbers,

$$S_n^{\mathcal{M}_0} \to \Sigma_0, \quad S_n^{\mathcal{M}} \to \Sigma_0^{\mathcal{M}}, \text{ a.s.}$$

Thus eventually a.s. $\det S_n^{\mathcal{M}_0} / \det S_n^{\mathcal{M}} < \mathrm{e}^{-a}$ and the statement of the lemma follows.

If $\mathcal{M} \neq \mathcal{M}_0$ is compatible with $\Sigma_0$, by the Central Limit Theorem

$$\sqrt{n}(S_n^{\mathcal{M}} - S_n^{\mathcal{M}_0}) \xrightarrow{\mathcal{D}} R.$$

By Slutsky's theorem the spectral radius and minimum eigenvalue of $(S_n^{\mathcal{M}_0})^{-1}(S_n^{\mathcal{M}} - S_n^{\mathcal{M}_0})$ satisfy $\rho = \mathcal{O}_P(n^{-1/2})$ and $\lambda = o_P(1)$ respectively. Consequently by (23)

$$\left| \frac{\det S_n^{\mathcal{M}_0}}{\det S_n^{\mathcal{M}}} \right|^{n/2} \leq \mathrm{e}^{\frac{np\rho^2}{2(1+\lambda)}} = \mathcal{O}_P(1).$$

■

### A.5 Proof of Theorem 4.1

**Theorem A.1:** *For any clique model $\mathcal{M}$ that is not compatible with $\Sigma_0$ assume $\det(\Sigma_0) < \det(\Sigma_0^{\mathcal{M}})$ and the penalty $\mathrm{e}^{-an} q_{\mathcal{M}}(n)/q_{\mathcal{M}_0}(n) \to 0$ for all $a > 0$ as $n \to 0$.*

*For any clique model $\mathcal{M}$ compatible with $\Sigma_0$ assume that $q_{\mathcal{M}}(n)/q_{\mathcal{M}_0}(n)$ is bounded.*

*Then as $n \to \infty$ with $p$ held fixed $r_p(\mathcal{M}_0 \mid \boldsymbol{Y}) \xrightarrow{P} 1$.*

***Proof:*** Because for any fixed $p$ there are finitely many clique models, we only need to prove that for any $\mathcal{M} \neq \mathcal{M}_0$, $\frac{r_p(\mathcal{M} \mid \boldsymbol{Y})}{r_p(\mathcal{M}_0 \mid \boldsymbol{Y})} \xrightarrow{P} 0$.

Denote by $g_i, i = 1, \ldots, k$, the size of cliques in $\mathcal{M}$ and $h_j, j = 1, \ldots, l$, the size of cliques in $\mathcal{M}_0$.

By Lemma (A.1), (A.2) we have as $n \to \infty$

$$\frac{r_p(\mathcal{M} \mid \boldsymbol{Y})}{r_p(\mathcal{M}_0 \mid \boldsymbol{Y})} \sim Kn^{-\frac{\sum_{i=1}^{k} g_i^2 - \sum_{j=1}^{l} h_j^2}{4}} \frac{q_{\mathcal{M}}(n)}{q_{\mathcal{M}_0}(n)} \left| \frac{\det S_n^{\mathcal{M}_0}}{\det S_n^{\mathcal{M}}} \right|^{n/2},$$

where $K$ is a constant independent of $n$.

If $\mathcal{M}$ is not compatible with $\Sigma_0$ by assumption and Lemma A.3(i), we have $\frac{r_p(\mathcal{M}|\boldsymbol{Y})}{r_p(\mathcal{M}_0 \mid \boldsymbol{Y})} \to 0$ a.s.

If $\mathcal{M} \neq \mathcal{M}_0$ is compatible with $\Sigma_0$ notice that $\mathcal{M}$ is obtained by pooling together some cliques of $\mathcal{M}_0$. Therefore $\sum_{i=1}^{k} g_i^2 - \sum_{j=1}^{l} h_j^2 > 4$. Consequently $\frac{r_p(\mathcal{M} \mid \boldsymbol{Y})}{r_p(\mathcal{M}_0 \mid \boldsymbol{Y})} \xrightarrow{P} 0$ by assumption and Lemma A.3(ii). ■