ABSTRACT

SHI, WENLI. Bayesian Inference in Gaussian Graphical Models under Measurement Error. (Under the direction of Subhashis Ghoshal and Ryan Martin.)

Estimation of sparse, high-dimensional precision matrices is an important and challenging problem. Most of the existing methods assume that the observations can be made precisely but, in practice, this often is not the case; for example, the instruments used to measure the response may have limited precision. This thesis focuses on developing new Bayesian methodology for inference on structured, high-dimensional precision matrices under measurement error with known scale.

In Chapter 2, we propose a Bayesian method for inference on a sparse precision matrix in a Gaussian graphical model with data corrupted by Gaussian measurement error whose variance is assumed known. In particular, we establish a general result with sufficient conditions under which the posterior contraction rates that hold in the no-measurement-error case carry over the measurement-error case. Interestingly, this result does not require the measurement error variance to be small. Some discussions of the case with repeated measure and unknown variance are presented following the method and the theory. We apply the general result to several cases with well-known prior distributions for sparse precision matrices and also to a case with a newly-constructed prior for precision matrices with a sparse factor-loading form. Also, a simulation study highlights the empirical benefits of accounting for the measurement error as opposed to ignoring it, even when that measurement error is relatively small.

In Chapter 3, we consider the Gaussian graphical model under general additive measurement error, which is not necessarily Gaussian. We derive two results of the posterior contraction rates in terms of Frobenius norm that are consistent with the ones without measurement error under two different conditions, respectively. The first theorem restricts the tail decay of the measurement error distribution, while the second requires that the variability of the measurement error decays sufficiently fast and allows any distribution on the measurement error. Two simulation studies illustrate the positive influence of adjusting for the measurement error, which either has a uniform distribution or a Student t-distribution. We also explore the finite-sample performance on the selection, but the result is not ideal for the surveyed structures on the true precision matrices.

We generalize the measurement error as the kernel of a mixture with the Gaussian graphical model in Chapter 4. We introduce a Bayesian method to estimate the precision matrix with the MCMC algorithm for producing samples from the posterior. The posterior contraction rate of the likelihood functions with respect to the Rényi divergence is obtained regardless of the distribution of the measurement error. A simulation study discloses the remarkable gain on the accuracy of estimation after correcting for the measurement error, which follows the Poisson distribution.

In Chapter 5, we conclude the results and discuss some open questions related to the Gaussian graphical model under measurement error. Besides changing the distribution of the measurement error, we consider the nonparanormal graphical model as another extension and introduce a general

Bayesian approach to adjust for the measurement error with Gibbs sampling algorithm when the measurement error is Gaussian.

Bayesian Inference in Gaussian Graphical Models under Measurement Error

by
Wenli Shi

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2021

APPROVED BY:

Leonard Stefanski	Minh Tang
Subhashis Ghoshal Co-chair of Advisory Committee	Ryan Martin Co-chair of Advisory Committee

DEDICATION

To my family.

BIOGRAPHY

Wenli was born in 1992 in Jinan, Shandong, China. In 2015, He obtained the Bachelor degree in Mathematics from Shandong University. During the undergraduate studies, he discovered statistics matching his dominant intention as developing theory and methodology which could be directly implemented on practical problems. Then, he received a Masters degree in Statistics from Duke University, where he first learned about and developed a passion for Bayesian inference. In 2017, he joined North Carolina State University to pursue a Ph.D. in Statistics. After his doctoral defense, he is scheduled to join Facebook as a research scientist.

ACKNOWLEDGEMENTS

First of all, I wish to express the deepest appreciation to my advisors, Dr. Subhashis Ghoshal and Dr. Ryan Martin, for their continuous support and intellectual guidance during my Ph.D. study. Their motivation and enthusiasm for research always encourage me to explore the unknowns and think outside the box. Their constructive advice and kind encouragement are so valuable that I can never finish this dissertation without them.

I would like to convey my sincere gratitude to all the other committee members, Dr. Leonard Stefanski and Dr. Minh Tang, for their insightful comments and suggestions that improve the results in this thesis and Dr. Mihai Diaconeasa for taking precious time to attend the committee as the graduate school representative. A special gratitude goes to Dr. Surya Tokdar from the Department of Statistical Science at Duke University for showing me the beauty of Bayesian statistics and leading me to the research of it.

My real appreciation belongs to the faculty and staff of the Department of Statistics at North Carolina State University for providing the helpful resources and maintaining the academic environment. I would like to thank Dr. Marcia Gumpertz as my academic advisor at the beginning of my Ph.D. journey. Special thanks to Dr. Dennis Boos, Dr. Eric Chi, Dr. Sujit Ghosh, Dr. Jacqueline Hughes-Oliver, Dr. Xinge Jeng, Dr. Soumendra Lahiri, Dr. Wenbin Lu, Dr. Brian Reich for teaching excellent courses, which broaden my knowledge of statistics.

I also want to thank all my colleagues, fellows and friends that have helped, supported and encouraged me through this tough journey. Parting is such sweet sorrow, the Bard wrote. And we will set sail with this wealthy asset of memory.

I am grateful to Bayer and Facebook, Inc. for offering me the fruitful internships. Special thanks to Radha Mohanty and Subash Kashyap from Bayer and Tony Lee, Yu Luo and Shi Qiu from Facebook, Inc. for mentoring me and sharing their experience.

Last but not least, I would like to dedicate this work to my beloved family, who has always been there for me. I am deeply grateful to my parents, Jianguo Shi and Yong Liu, for their unconditional love, endless support and unwavering belief in me and to my fianée, Beilin Jia, for staying by my side and brightening my life.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER

1

INTRODUCTION

1.1 Overview and Outline

Learning a dependence structure has recently received increasing interest, and the so-called *graphical model* provides an undirected graph-based probabilistic tool for describing the intrinsic relationship in multivariate distributions (Lauritzen, 1996). The treatments of learning the structure of the graph and estimating the level of dependence are needed in many scientific problems, ranging from genetic studies to political science. The problems in biology and medical science are the core engines driving the development of such research in recent decades. Biologists aim to infer the genomic network of regulatory relationships through the gene expression data (Friedman, 2004), while the neuroscientists strive to discover functional brain networks and detect the coherence of the activities among different brain regions through fMRI images (Fan, Han & Liu, 2014). Banerjee & Ghosal (2015) explored the grouping relationship of the S&P 500 stocks and compared the results with the 10 Global Industry Classification Standard. Banerjee, Ghaoui & d'Aspremont (2008) introduced another interesting application of graphical models using senate voting records data, which reveals the connections among Democrats and Republicans.

Given an undirected graph, each single node represents an individual random variable and the edge between each pair of nodes encodes their conditionally stochastic dependent relationship. These random variables along with their dependence structure form an undirected graphical model, which is also called a Markov random field or a Markov network, if the Markov property holds. The Markov property is defined as that any two subsets of variables are conditionally independent when some interim nodes (variables) of every path connecting these two subsets are given. For example, the node 1 and node 4 are conditionally independent if node 3 and node 5 are given as shown in

the left diagram of Figure 1.1. However, node 1 and node 2 cannot be conditionally independent as there is no interim nodes on this single path, similar for the pair of node 1 and node 3 (or node 6).

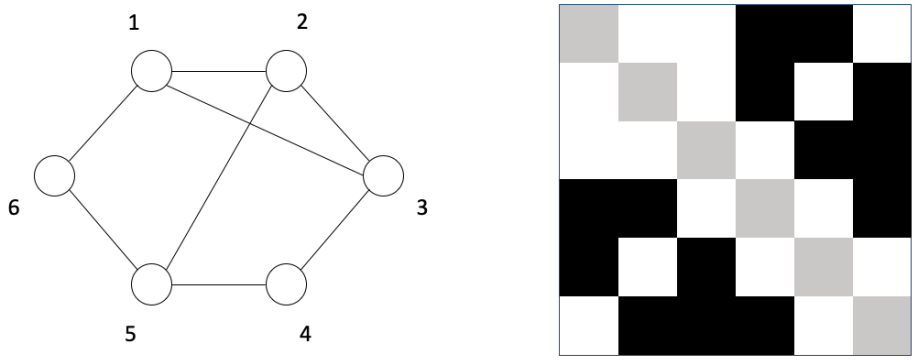

Figure 1.1 Gaussian graphical model and its pattern of the zeros and non-zeros in the inverse covariance matrix. The blacks denote the zero entries and the whites are the non-zero ones.

Although numerous inferential methods were proposed in nonparametric and semi-parametric statistical literature summarized in Drton & Maathuis (2017), Gaussian graphical model, which assumes a multivariate joint Gaussian distribution on the nodes, is still an intuitive and common parametric choice for continuous observations among all the undirected graphical models. The most attractive property of a Gaussian graphical model is that its precision matrix describes the conditional dependence structure. In other words, the conditional independence of two random variables given the remaining ones implies the corresponding off-diagonal entry in the precision matrix to be zero and the converse holds as well. For instance, by comparing the two diagrams in Figure 1.1, all the directly-connected nodes from the left diagram have the white corresponding blocks (non-zero entries) in the right and the not-directly-connected nodes are reflected by the black blocks (zero entries). In other words, these two diagrams in Figure 1.1 are equivalent to represent the graph pattern under Gaussian assumption. Therefore estimating the precision matrix, namely, the inverse of the covariance matrix of a Gaussian random vector, is the key to learning the structure underlying the Gaussian graphical model.

While the maximum likelihood estimator (MLE) is commonly utilized in low-dimensional problems, it is not useful when the dimension is greater than the sample size since the sample covariance matrix is not invertible. Even when the dimension is modest and the sample size is slightly larger, the MLE is not admissible since the information contained in the data may not be sufficient for an accurate estimation. Bayesian methods face similar challenges. Although the Bayesian estimator with the Wishart prior for the precision matrix, and deploying the Gaussian-Wishart conjugacy is always applicable for whatever the sample size and the dimension are, the estimation can hardly

detect the zero entries and is not recommended when the dimension is high because of the insufficient information. Therefore, it is crucial to reduce the number of effective parameters in this high-dimensional context by imposing sparsity, which assumes only a few intrinsic dependent relationships in the graph. Guided by this idea, approaches for estimating sparse precision matrices with a reduced number of free parameters, have attracted considerable attention in the recent literature.

Beyond the challenges of high dimensionality and complex dependence structures, it may happen that the data are also corrupted in some way. A classical example is that measurements taken on sample units can only be done with a low-precision device. In such a case, the natural sample variation is compounded by independent measurement errors. A more recent example, commonly found in medical applications, is where the data are corrupted intentionally to maintain privacy. In any case, the addition of a measurement error on top of the natural sampling variability creates new challenges. While there is an extensive body of literature on the subject of *measurement error* in statistics (Cook & Stefanski, 1994; Carroll et al., 2006; Freedman et al., 2008; Fuller, 2009), there is only little work that has investigated in the context of structured precision matrix estimation given the presence of the Gaussian measurement error. Moreover, to the best of our knowledge, this problem has not been explored under a more general type of measurement error. The present thesis aims to fill that gap.

The thesis is organized as follows. The reminder of this chapter will review some state-of-the-art Bayesian and non-Bayesian techniques for structure learning in Gaussian graphical models, explore the case with additional measurement error and the consequence of ignoring it, and introduce the theory of posterior convergence rate to lay a solid foundation for the upcoming results. In Chapter 2, we propose a Bayesian method to adjust for Gaussian measurement error with known variance, and to estimate the precision matrix of Gaussian graphical models. We obtain the posterior convergence rate in terms of the Frobenius norm when the true precision matrix has a certain structure and has eigenvalues bounded away from 0. In Chapter 3, we assume the measurement error to be additive but not necessarily Gaussian, and maintain the same contraction rate under a mild condition on the magnitude of measurement error or the tail probability of the measurement error distribution, respectively. In Chapter 4, we consider a broader type of measurement error distribution, for example the exponential family, and propose a method to estimate the true precision matrix along with a posterior contraction rate result under the Rényi divergence. Chapter 5 contains conclusions of the results and discussions on the open problems of inference on the precision matrix under measurement error. Particularly, the nonparanormal graphical model is considered as another extension of current model and a Bayesian method is introduced to adjust for the Gaussian measurement error.

1.2 Background

1.2.1 Structure Learning

The precision matrix of a Gaussian random vector is a key object in multivariate analysis because of its role in describing the conditional distributions, unlike the covariance matrix, which expresses the marginal dependence structure. Under the Gaussian assumption, we consider the autoregressive model with lag 1 (AR(1)) for instance to illustrate the difference between marginal and conditional independence, that is, the off-diagonal entries in the covariance and precision matrix intrinsically. Suppose a multivariate Gaussian distribution whose covariance matrix and precision matrix are

$$\Sigma = \begin{pmatrix} 1.0 & 0.7 & 0.7^2 & 0.7^3 & 0.7^4 \\ 0.7 & 1.0 & 0.7 & 0.7^2 & 0.7^3 \\ 0.7^2 & 0.7 & 1.0 & 0.7 & 0.7^2 \\ 0.7^3 & 0.7^2 & 0.7 & 1.0 & 0.7 \\ 0.7^4 & 0.7^3 & 0.7^2 & 0.7 & 1.0 \end{pmatrix}, \quad \Omega = \begin{pmatrix} 1.96 & -1.37 & 0 & 0 & 0 \\ -1.37 & 2.92 & 1.37 & 0 & 0 \\ 0 & -1.37 & 2.92 & -1.37 & 0 \\ 0 & 0 & -1.37 & 2.92 & -1.37 \\ 0 & 0 & 0 & -1.37 & 1.96 \end{pmatrix}. \quad (1.2.1)$$

It is obvious that the precision matrix contains many zeros, which indicates that the corresponding graph is sparse, while the covariance matrix is dense. Every pair is marginally dependent, while some (the entries that the difference of their row number and the column number is greater than 1) are conditionally independent given the others. The covariance matrix and the precision matrix encode different types of dependent information and the techniques for estimating them are distinct. Besides that the precision matrix conveys the particular messages of relation, there is another important reason forcing us to focus on estimating the precision matrix, which will be introduced detailedly in Section 2.1.

Then, the goal is to make inference on the unknown precision matrix Ω, especially in the high-dimensional situation when the dimension p is large. Even for relatively modest p, the information available in the data may be insufficient because the number of unknown parameters to be estimated is in the order of p^2, which can exceed the sample size n. The problem can be often addressed if the precision matrix has a certain structure that allows a significant reduction in the number of free parameters in the model. For example, in Gaussian graphical models (see Lauritzen (1996)), it is common to assume that the underlying graph describing the dependence structure is sparse, thus leading to a precision matrix with many zeros on the off-diagonal. Therefore, sparsity simultaneously simplifies the dependence structure and effectively reduces the dimension of Ω, potentially paving the way for accurate estimation.

1.2.1.1 Literature Review

Regularization methods are often used to incorporate the intended sparse structure into the estimator. Yuan & Lin (2007) and Banerjee, Ghaoui & d'Aspremont (2008) proposed to add an ℓ_1-type penalty to the negative log-likelihood, leading to the so-called graphical LASSO estimator. A fast

computational method using the coordinate descent algorithm was introduced by Friedman, Hastie & Tibshirani (2008). Inspired by the desirable properties of the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) which uses folded concave penalties to avoid the known problem of bias due to excessive shrinkage of large non-zero entries, Fan, Feng & Wu (2009) proposed the graphical SCAD. In a different vein, motivated by the relation between the linear regression and the conditional distribution of multivariate Gaussian random variables, a selection procedure for neighborhood emerged by Meinshausen & Bühlmann (2006) using column-by-column regression with an ℓ_1-penalty for pursuing the sparsity. Although this procedure seems to be logical, it may not guarantee the symmetry of the estimated precision matrix, which causes a contradictory relationship and loss of efficiency. Peng et al. (2009) ameliorated the method to implement a joint sparse regression, which could preserve the symmetry of the estimation and perform the neighborhood selection simultaneously. Yuan (2010) borrowed this thought of regression but replaced the ℓ_1-penalty by the Dantzig selector (Candes & Tao, 2007). Cai, Liu & Luo (2011) designed a procedure based on the Dantzig selector which minimizes the ℓ_1-norm of the precision matrix, while it constrains on the sup-norm between the identity matrix and the product of sample covariance matrix with the precision matrix. Cai, Liu & Zhou (2016) improved this method to be adaptive and the estimator achieves the optimal rate.

In the Bayesian literature, several priors were considered for a sparse precision matrix and resulting computational procedures were developed. Wang (2012) proposed the Bayesian graphical LASSO, which specifies a Laplace prior on the off-diagonal entries of the precision matrix and an exponential prior on the diagonal entries independently. He also developed a clever computational trick, known as *scaling-it-up*, to cancel out the normalizing constant in each posterior sampling stage. Since a Laplace prior, although peaked at zero, does not yield the exact zero value with positive probability, a post-estimation thresholding mechanism is needed to learn the sparsity structure using Wang's method. Banerjee & Ghosal (2015) proposed an adjustment with a mixture of a point mass and a Laplace prior to induce exact sparsity, and also derived the optimal posterior contraction rate with respect to the Frobenius norm. To compute the posterior, they devised a Laplace approximation method, which is a scale of magnitude faster than Markov Chain Monte Carlo (MCMC) methods, but relies on a large sample approximation. Another type of shrinkage prior is introduced by Li, Craig & Bhadra (2019) using the horseshoe prior (Carvalho, Polson & Scott, 2010) on the off-diagonal entries of precision matrix, which cannot set off-diagonal elements to be exactly zero and hence needs additional ad hoc thresholding. More focus can be given on selection than shrinkage using a graphical Wishart (G-Wishart in short) prior, which sets some off-diagonal entries to exact zeros guided by the chosen graph and retains conjugacy with the Gaussian likelihood. Lenkoski & Dobra (2011) and Mohammadi & Wit (2015) proposed useful computational methods that allow MCMC moves across possible graphs. Banerjee & Ghosal (2014) assumed a banded structure on the precision matrix and derived the posterior contraction rate with a G-Wishart prior. Liu & Martin (2019) proposed an empirical G-Wishart prior and demonstrated its optimal posterior contraction rate and strong performance in terms of computational speed and accuracy.

Du & Ghosal (2018) considered a high-dimensional discriminant analysis, where they implemented both the mixture prior and a horseshoe shrinkage prior on the off-diagonal entries in a sparse modified Cholesky decomposition. Xiang, Khare & Ghosh (2015) derived posterior convergence rates for a decomposable graphical model using the G-Wishart prior and Cholesky decomposition in high-dimensional setup.

1.2.1.2 Background on Some Existing Methods

We review several selective methods along with their computational techniques and theoretical results from the literature of estimating the sparse precision matrix. Without loss of generality, suppose that there are observations $X_1, \ldots, X_n$, independent and identically distributed (i.i.d.) from a p-dimensional, mean-zero Gaussian distribution, $\mathsf{N}_p(0, \Sigma^\star)$, where $\Sigma^\star$ denotes the true $p \times p$ positive definite covariance matrix, with corresponding true precision matrix $\Omega^\star = \Sigma^{\star -1}$. Let $\det(\cdot)$ and $\mathrm{tr}(\cdot)$ denote the determinant and trace of a matrix, respectively. Then, the corresponding negative log-likelihood is equivalent to $L(\Omega) = -\log\det(\Omega) + \mathrm{tr}(\Omega S)$, where $S = \sum_{i=1}^n X_i X_i^{\mathrm{T}}/n$ denotes the sample covariance matrix. By taking the derivative of the negative log-likelihood with respect to Ω, the MLE of Ω is obtained as S^{-1}. This estimator is not always available, particularly in the high-dimensional problem, since S is not a full rank matrix and is not invertible when $p > n$. Moreover, the information contained in data may not be sufficient for a satisfactory estimation result even for modest p compared with the $(p+1)p/2$ unknown parameters. The main strategy is to reduce the number of the unknowns by assuming the truth to be sparse. Let s denote the sparsity, that is, the number of non-zero off-diagonal entries in $\Omega^\star$, and then several influential approaches were proposed in the literature following this idea with large p but moderate s.

1.2.1.2.1 Graphical LASSO

Tibshirani (1996) first introduced the ℓ_1-penalty as the method LASSO, which can serve for the shrinkage and selection purpose simultaneously. This thought was deployed to estimate the precision matrix called the graphical LASSO, which adds the ℓ_1-penalty to the negative log-likelihood as the loss function and minimizes it for the estimator, that is,

$$\widehat{\Omega}^L = \underset{\Omega \in \mathscr{M}^+}{\arg\min} \left\{ -\log\det(\Omega) + \mathrm{tr}(\Omega S) + \sum_{i=1}^{p}\sum_{j=1}^{p} \lambda_{i,j}|\Omega_{i,j}| \right\}, \tag{1.2.2}$$

where $\mathscr{M}^+$ denotes the space of all the $p \times p$ positive definite matrices, $|\Omega_{i,j}|$ is the absolute value of the (i,j)th entry of Ω and $\lambda_{i,j} > 0$ is the tuning parameter of each entry for $1 \le i, j \le p$. Based on different choices of λ, several variants of graphical LASSO were proposed. Yuan & Lin (2007) omitted the penalty on the diagonal entries by requiring $\lambda_{i,i} = 0$ for all i and Banerjee, Ghaoui & d'Aspremont (2008) included them, while both methods simply assume an identical tuning parameter λ instead of making it element-wise. The adaptive graphical LASSO treats the regularization parameter $\lambda_{i,j} = 1/|\tilde{\Omega}_{i,j}|^\gamma$ for some hyperparameter $\gamma > 0$ where $\tilde{\Omega}_{i,j}$ are the entries of $\tilde{\Omega}$, which is assumed as a

consistent estimator of Ω (Fan, Feng & Wu, 2009). When $p < n$, a common choice of $\tilde{\Omega}$ is the MLE S^{-1}. An alternate to $\tilde{\Omega}$ is the graphical LASSO estimator when the MLE is unavailable or inconsistent. Since the objective function in (1.2.2) is globally convex, the solution is unique for all positive definite S with $\lambda_{i,j} \geq 0$ or the semi-positive definite S with $\lambda_{i,j} > 0$.

For computing the estimation in (1.2.2), a coordinate-descent type of algorithm was introduced by Friedman, Hastie & Tibshirani (2008). This procedure first estimates the covariance matrix and then utilizes it to compute the estimation of the precision matrix with relatively modest cost. This algorithm also bridges the conceptual gap between the exact problem of graphical LASSO and the approximate proposal in Meinshausen & Bühlmann (2006). Consider the situation where all the $\lambda_{i,j}$ are identical for all $1 \leq i, j \leq p$. To simplify the notation, we use $-$ sign in the subscript to denote the rest of the matrix after removing the corresponding row or column. For instance, $W_{-i,-j}$ denotes the $(p-1)\times(p-1)$ matrix, in which the ith row and jth column of the $p \times p$ matrix W are removed and $W_{-i,j}$ denotes the $(p-1)\times 1$ vector which is the jth column of W after removing its ith element. Then the coordinate-descent algorithm of the graphical LASSO is executed as follows.

- Step 1: Initialize $W^{(0)} = S + \lambda I$ and a $(p-1)\times p$ matrix B.
- Step 2: For $(t+1)$th iteration, update $W_{-i,i}^{(t+1)}$ and $W_{i,-i}^{(t+1)}$ for $i = 1, \ldots, p$ recursively:
  - Solve this LASSO problem, $\hat{\beta} = \arg\min_{\beta}(\beta^{\mathrm{T}} W_{-i,-i}^{(t)} \beta/2 - \beta^{\mathrm{T}} S_{-i,i} + \lambda\|\beta\|_1)$, through the coordinate descent algorithm of LASSO regression and the soft-threshold operator.
  - Update $W_{-i,i}^{(t+1)} = W_{-i,-i}^{(t)}\hat{\beta}$, $W_{i,-i}^{(t+1)} = (W_{-i,-i}^{(t)}\hat{\beta})^{\mathrm{T}}$ and the ith column of B by $B_i = \hat{\beta}$.
- Step 3: Repeat Step 2 until W converges.
- Step 4: Compute $\widehat{\Omega}_{i,i} = 1/(W_{i,i} - W_{-i,i}^{\mathrm{T}} B_i)$ and $\widehat{\Omega}_{i,-i} = \widehat{\Omega}_{-i,i} = -\widehat{\Omega}_{i,i} B_i$ iteratively for $i = 1, \ldots, p$, which are then combined together as the graphical LASSO estimator of Ω.

Rothman et al. (2008) showed the convergence rates of the estimator in terms of Frobenius norm and spectral norm as $\sqrt{n^{-1}(p+s)\log p}$, while another slightly different estimator called SPICE in the same paper improves the convergence rate to $\sqrt{n^{-1} s \log p}$ under spectral norm. Ravikumar et al. (2011) established the model selection consistency and computed the convergence rates in element-wise supremum norm, Frobenius norm and spectral norm under more general conditions, where the random vectors are allowed with heavier tail behavior than Gaussian distribution. The rates in the latter two norms are the minimum of $\sqrt{n^{-1}(p+s)\log p}$ and $\sqrt{n^{-1} d^2 \log p}$, where d denotes the maximum number of nonzero entries in each column of $\Omega^{\star}$.

1.2.1.2.2 Graphical SCAD

Instead of considering the ℓ_1-penalty as (1.2.2), Fan, Feng & Wu (2009) replaced it with the non-concave SCAD penalty (Fan & Li, 2001) and the estimator is computed as

$$\widehat{\Omega}^S = \underset{\Omega \in \mathscr{M}^+}{\arg\min} \left\{ -\log\det(\Omega) + \mathrm{tr}(\Omega S) + \sum_{i=1}^{p}\sum_{j=1}^{p} P_{\lambda,a}(|\Omega_{i,j}|) \right\}, \tag{1.2.3}$$

where the SCAD penalty $P_{\lambda,a}(|\Omega_{i,j}|)$, for the (i,j)th entry of Ω, is formulated as

$$\lambda|\Omega_{i,j}|\mathbb{1}(|\Omega_{i,j}|\leq\lambda)+\frac{2a\lambda|\Omega_{i,j}|-|\Omega_{i,j}|^2-\lambda^2}{2(a-1)}\mathbb{1}(\lambda<|\Omega_{i,j}|\leq a\lambda)+\lambda^2\frac{(a+1)}{2}\mathbb{1}(a\lambda\leq|\Omega_{i,j}|),$$

in which $\mathbb{1}(\cdot)$ denotes the indicator function and $\lambda>0$ and $a>2$ are two tuning parameters. Fan, Feng & Wu (2009) recommended to choose $a=3.7$, while λ could be computed through a grid search based on cross validation. The SCAD penalty has the first derivative with respect to $|\Omega_{i,j}|$ as

$$P'_{\lambda,a}(|\Omega_{i,j}|)=\lambda\left\{\mathbb{1}(|\Omega_{i,j}|\leq\lambda)+\frac{\max(0,a\lambda-|\Omega_{i,j}|)}{(a-1)\lambda}\mathbb{1}(|\Omega_{i,j}|>\lambda)\right\}.$$

Replacing the SCAD penalty by its first-order Taylor expansion at $|\Omega^\star_{i,j}|$, $P_{\lambda,a}(|\Omega_{i,j}|)\approx P_{\lambda,a}(|\Omega^\star_{i,j}|)+P'_{\lambda,a}(|\Omega^\star_{i,j}|)(|\Omega_{i,j}|-|\Omega^\star_{i,j}|)$, we then approximate the solution in (1.2.3) by

$$\underset{\Omega\in\mathscr{M}^+}{\arg\min}\left\{-\log\det(\Omega)+\operatorname{tr}(\Omega S)+\sum_{i=1}^{p}\sum_{j=1}^{p}P'_{\lambda,a}(|\Omega^\star_{i,j}|)|\Omega_{i,j}|\right\},$$

which has the form of graphical LASSO as (1.2.2). Therefore, the algorithm for solving the graphical LASSO is borrowed here by changing $\lambda_{i,j}$ to $P'_{\lambda,a}(|\Omega^{(t-1)}_{i,j}|)$ in the tth coordinate-descent iteration.

Lam & Fan (2009) studied model selection consistency as well as the convergence rates in terms of the Frobenius norm and the spectral norm of the graphical SCAD as $\sqrt{n^{-1}(p+s)\log p}$, while an improvement in the rate in spectral norm to $\sqrt{n^{-1}s\log p}$ was presented for the estimator similar to SPICE (Rothman et al., 2008).

1.2.1.2.3 CLIME

The constrained ℓ_1-minimization for inverse matrix estimation (CLIME) was introduced by Cai, Liu & Luo (2011), which deploys the Dantzig selector (Candes & Tao, 2007) for specifying the loss function and the estimator as

$$\widehat{\Omega}^C=\underset{\Omega\in\mathscr{M}^+}{\arg\min}\|\Omega\|_1,\ \text{subject to}\ \|S\Omega/n-I_p\|_\infty\leq\lambda, \tag{1.2.4}$$

where $\lambda>0$ is a tuning parameter and I_p is the identity matrix of order p. This loss function could be further decomposed into an optimization problem of p dimensional vector and solved by iterating each column of Ω as

$$\min\|\Omega_i\|_1,\ \text{subject to}\ \|S\Omega_i/n-e_i\|_\infty\leq\lambda,\ \text{for}\ i=1,\ldots,p,$$

where Ω_i is the ith column of Ω and e_i denotes the unit column vector with ith element as 1 and the others as 0. Since (1.2.4) and its modification cannot produce a guaranteed symmetric matrix, each pair of the entries symmetric with respect to the diagonal in $\widehat{\Omega}^C$ is replaced by the smaller absolute value after the CLIME is computed. Note that the vector-form CLIME is equivalent to a

linear programming problem, which can be computed through the primal dual interior method (Cai, Liu & Luo, 2011) or an alternating direction method of multiplier (Wang et al., 2013).

Let M denote the upper bound of the ℓ_1-norm of the precision matrix in the parameter space and d denote the maximum number of nonzero entries in each column of $\Omega^\star$. Cai, Liu & Luo (2011) studied the graphical model selection consistency and derived the convergence rate $M^2 d\sqrt{n^{-1}\log p}$ in terms of the spectral norm, the rate $M^2\sqrt{n^{-1}\log p}$ in terms of the supremum norm and the rate $M^2\sqrt{p\,d\,n^{-1}\log p}$ in terms of Frobenius norm. These results are comparable or sometimes even better than the results of the graphical LASSO or graphical SCAD as derived in Ravikumar et al. (2011) or Lam & Fan (2009). Furthermore, Cai, Liu & Zhou (2016) proposed the adaptive CLIME and showed that the aforementioned rates are minimax optimal.

1.2.1.2.4 Bayesian Graphical LASSO

Park & Casella (2008) proposed the Bayesian LASSO for linear model, which suggests that constraint provided by the Laplace prior has the same analytic form as the penalty of LASSO. By assigning such prior on the regression parameters independently, the posterior mode serves as the Bayesian LASSO estimator. Wang (2012) implemented this thought and introduced the Bayesian graphical LASSO, which specified the exponential distribution and Laplace distribution on the diagonal and off-diagonal entries of Ω, respectively, given the Gaussian likelihood and a fixed tuning parameter λ, that is,

$$\pi(\Omega|\lambda) \propto \prod_{i=1}^{p-1}\prod_{j=i+1}^{p}\left\{\frac{\lambda}{2}\exp(-\lambda|\Omega_{i,j}|)\right\}\prod_{i=1}^{p}\left\{\frac{\lambda}{2}\exp\left(-\frac{\lambda}{2}\Omega_{i,i}\right)\right\}\mathbb{1}\left(\Omega\in\mathscr{M}^+\right). \tag{1.2.5}$$

Let Υ denote the $p\times p$ matrix of $\Upsilon_{i,j}$, which is an interim variable. Through the data augmentation of $\Upsilon_{i,j}$ on the priori of $\Omega_{i,j}$, the Laplace prior could be equivalently replaced by a scale mixture of Gaussians, that is, for any $1\le i<j\le p$,

$$\pi(\Omega_{i,j}|\Upsilon_{i,j},\lambda)\pi(\Upsilon_{i,j}|\lambda) \propto \frac{1}{\sqrt{2\pi\Upsilon_{i,j}}}\exp\left(-\frac{\Omega_{i,j}^2}{2\Upsilon_{i,j}}\right)\cdot\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\Upsilon_{i,j}\right)$$

where integrating out $\Upsilon_{i,j}$ over the positive real line gives the Laplace distribution in (1.2.5).

Based on this data augmentation, Wang (2012) proposed a block Gibbs sampling scheme for the Bayesian graphical LASSO posterior. Let $\mathrm{diag}(v)$ denote the diagonal matrix formed by the vector v. With the initial values for Ω and Υ, the following MCMC sampling steps are executed iteratively, that is, for the $(t+1)$th iteration,

- Step 1: Sample $\Omega_{-i,i}^{(t+1)}$, $\Omega_{i,-i}^{(t+1)}$ and $\Omega_{i,i}^{(t+1)}$ for $i=1,\ldots,p$:
  - Given $\Omega^{(t)}$ from the last iteration, sample $\gamma\sim\mathsf{Gamma}(n/2+1,(S_{i,i}+\lambda)/2)$ and $\beta\sim\mathsf{N}_{p-1}(-C\,S_{-i,i},C)$ where $C=\{(S_{i,i}+\lambda)\Omega_{-i,-i}^{(t)-1}+\mathrm{diag}(\Upsilon_{-i,i})^{-1}\}^{-1}$.
  - Update $\Omega_{-i,i}^{(t+1)}=\beta$, $\Omega_{i,-i}^{(t+1)}=\beta^{\mathrm{T}}$ and $\Omega_{i,i}^{(t+1)}=\gamma+\beta^{\mathrm{T}}\Omega_{-i,-i}^{(t)-1}\beta$.
- Step 2: For $1\le i<j\le p$, update $\Upsilon_{i,j}=1/u_{i,j}$, where $u_{i,j}\sim\mathsf{Inv\text{-}Gaussian}(|\lambda/\Omega_{i,j}^{(t+1)}|,\lambda^2)$.

The fully Bayesian method specifies a prior on the tuning parameter λ and makes it a parameter to be estimated instead of fixing it. Wang (2012) suggested the prior $\lambda \sim \mathsf{Gamma}(a, b)$, which leads to the full conditional posterior for λ as $\mathsf{Gamma}(a + p(p+1)/2, b + ||\Omega||_1/2)$.

Different from the exact sparsity obtained from the graphical LASSO, the Bayesian graphical LASSO never gives an estimation with exactly zero entries. Banerjee & Ghosal (2015) extended this method by adding a point mass at 0 in the prior and changed the prior of $\Omega_{i,j}$ to a mixture, which induces the posterior giving exact sparsity and restricts the number of the effective edges in the estimated graph. More importantly, they derived the optimal posterior convergence rate in Frobenius norm as $\sqrt{n^{-1}(p+s)\log p}$, which coincides with the graphical LASSO.

1.2.1.2.5 Bayesian Method with G-Wishart Prior

For Bayesian analysis, a priori probability is usually specified, which will sometimes put a constraint on the support or the number of free parameters. Extended from the traditional Wishart distribution, the G-Wishart prior is proposed, which induces exact sparsity given in the priori of the graph and maintains conjugacy with Gaussian likelihood (Mohammadi & Wit, 2015). More precisely, the prior of Ω has the G-Wishart distribution $W_G(b, D)$ as

$$\pi(\Omega|G) \propto |\Omega_{i,j}|^{(b-2)/2} \exp\{-\mathrm{tr}(D\Omega)/2\}\, \mathbb{1}(\Omega \in \mathscr{M}_G^+), \tag{1.2.6}$$

where G denotes the generic graph, which has some discrete prior on the $2^{p(p-1)/2}$ potential graphs and $\mathscr{M}_G^+$ denotes all the positive definite matrices corresponding to the graph G. The common choices of the priori on G include uniform distribution and truncated Poisson distribution. Another prior suggested by Mohammadi & Wit (2015) is to specify the prior probabilities for the existence of each edge in G, which are suggested as 0.5 uniformly.

By conjugacy, the full conditional posterior of Ω given the graph is still the G-Wishart distribution $W_G(b^\star, D^\star)$ where $b^\star = b+n$ and $D^\star = D+S$. The sampling procedure from the posterior of precision matrix is first simulating a graph and then sampling a precision matrix according to the graph. For the first step, the birth-death Markov chain Monte Carlo process computes the probabilities for the birth and death event in a continuous time birth-death Markov process and traverses the addition (birth) or deletion (death) of an edge based on the birth and death probabilities (Mohammadi & Wit, 2015). Given the current state of graph, a direct sampler from the G-Wishart distribution is applied in the second step for simulating the precision matrix (Lenkoski, 2013). This algorithm outperforms the stochastic search and is more computationally feasible for high-dimensional problems.

Although this method seems to be straightforward, the theoretical results about the convergence of the posterior are not completely established, especially when the dimension is high. Banerjee & Ghosal (2014) studied the posterior convergence rate of a banded precision matrix with band k in spectral norm as $\max(k^{5/2}\sqrt{n^{-1}\log p}, k^{3/2}\gamma(k))$, where $\gamma(k)$ is the banding approximation rate. Now, let k denote the maximum number of nonzero entries in each column of $\Omega^\star$, Xiang, Khare & Ghosh (2015) derived the exactly same rate for the decomposable graphical model using the G-Wishart

prior and Cholesky decomposition in high-dimensional framework. The identical convergence rate is derived by Banerjee (2017) for a wider range of arbitrary decomposable graphical models under similar assumptions and G-Wishart prior, where the complication of using a Cholesky factor is avoided. Liu & Martin (2019) proposed an empirical G-Wishart prior and demonstrated the superior performance in computing speed and accuracy through numerical results, while the theoretical result shows the optimal convergence rate is achieved under Frobenius norm.

1.2.1.2.6 Bayesian Method on Cholesky Structure

Since there exists a unique Cholesky decomposition for every positive definite matrix, this structure is applicably implemented to specify priors for a generic precision matrix, especially when the lower triangular matrix of the Cholesky decomposition is supposed to be sparse. Du & Ghosal (2018) proposed a spike-and-slab prior and a horseshoe shrinkage prior on the Cholesky decomposition $\Omega = LDL^{\mathrm{T}}$, respectively, where L denotes the lower triangular matrix and D denotes the diagonal scale matrix. Then, the prior probability $\pi(\Omega_{i,j} \neq 0)$ depends on its position, that is, the entry with larger row number or column number tends to have a higher probability of being non-zero, if the prior probability $\pi(L_{i,j} \neq 0)$ is uniform over different $1 < j < i < p$. Therefore, to guarantee an approximate uniformity of the prior probability of each $\Omega_{i,j}$ not equal to 0, Du & Ghosal (2018) proposed that $\pi(L_{i,j} \neq 0)$ should decay with the row number at rate $1/\sqrt{i}$. Then, consider the following prior setup for $i = 1, 2, \ldots, p$ and $j = 1, 2, \ldots, i$,

$$D_{i,i} \sim \mathsf{Gamma}(\alpha_1, \beta_1), \quad L_{i,j} | \Upsilon_{i,j} \sim (1 - \Upsilon_{i,j}) \mathsf{N}(0, \sigma_0^2) + \Upsilon_{i,j} \mathsf{N}(0, \sigma_1^2), \quad \Upsilon_{i,j} \sim \mathrm{B}(C_p / \sqrt{i}), \tag{1.2.7}$$

where $\mathrm{B}(\cdot)$ denote the Bernoulli distribution, α_1, β_1, σ_0^2, σ_1^2 and C_p are pre-specified parameters and $\Upsilon_{i,j} = \mathbb{1}(L_{i,j} \neq 0)$ in matrix Υ. The parameter σ_0^2 is commonly specified as a small positive number, ideally 0, while σ_1^2 being large, to constitute the spike and the slab. C_p is a constant regularizing the prior sparsity only depending on p, while the simulation study shown in Du & Ghosal (2018) indicates that its value may not be influential to the estimation results. By substituting the prior on $L_{i,j}$ and $\Upsilon_{i,j}$ with $L_{i,j} | \Upsilon_{i,j} \sim \mathsf{N}(0, \Upsilon_{i,j}^2 \sigma_1^2)$ and $\Upsilon_{i,j} \sim \mathrm{C}^+(0,1)$, where $\mathrm{C}^+(0,1)$ is the standard positive half-Cauchy distribution, the shrinkage prior setting is proposed.

The Gibbs sampling regimes for both methods are similar. To start the algorithm, initialize $\Omega^{(0)}$ and the corresponding $\Upsilon^{(0)}$ and for $(t+1)$th MCMC sampling,

- Step 1: Sample $L_{i,j}^{(t+1)} \sim \mathsf{N}(-C D_{j,j}^{(t)} L_{-i,j}^{(t)\mathrm{T}} S_{-i,i}, C)$ for $1 \leq j < i \leq p$, where
  - $C = [\{\sigma_0^2 (1 - \Upsilon_{i,j}^{(t)}) + \sigma_1^2 \Upsilon_{i,j}^{(t)}\}^{-1} + D_{j,j}^{(t)} S_{i,i}]^{-1}$ for the spike-and-slab prior,
  - $C = \{(\tau^2 \Upsilon_{i,j}^{(t)})^{-1} + D_{j,j}^{(t)} S_{i,i}\}^{-1}$ for the horseshoe prior.
- Step 2: Sample $D_{i,i}^{(t+1)} \sim \mathsf{Gamma}(\alpha_1 + n/2, \beta_1 + (L^{(t+1)\mathrm{T}} S L^{(t+1)})_{i,i}/2)$ for $1 \leq j < i \leq p$.
- Step 3: Sample $\Upsilon_{i,j}^{(t+1)}$ for $1 \leq j < i \leq p$, from the full conditional posterior

– $\Upsilon_{i,j}^{(t+1)} \sim \mathrm{B}(P_{i,j})$ where

$$P_{i,j} = \frac{\left(C_p/\sqrt{i}\right) f_N\left(L_{i,j}^{(t+1)}; 0, \sigma_1^2\right)}{\left(C_p/\sqrt{i}\right) f_N\left(L_{i,j}^{(t+1)}; 0, \sigma_1^2\right) + \left(1 - C_p/\sqrt{i}\right) f_N\left(L_{i,j}^{(t+1)}; 0, \sigma_0^2\right)},$$

and f_N denotes the density of Gaussian distribution.

– $\Upsilon_{i,j}^{(t+1)} \sim \mathrm{F}$ where F has the density proportional to

$$\frac{1}{\Upsilon_{i,j}\left(1+\Upsilon_{i,j}^2\right)} \exp\left(-\frac{L_{i,j}^{(t+1)2}}{2\Upsilon_{i,j}^2 \tau^2}\right),$$

for the horseshoe prior, which could be sampled by the Metropolis-Hastings algorithm.

Note that if $\sigma_0^2 = 0$ holds, we actually have $P_{i,j} = \mathbb{1}(L_{i,j} \neq 0)$ for the spike-and-slab prior, where the MCMC update will be trapped in the spike or slab forever. Therefore, in practice, a highly concentrated Gaussian distribution with a small enough σ_0^2 instead of the point mass δ_0 is recommended for overcoming this obstacle and $L_{i,j}$, whose $\Upsilon_{i,j} = 0$, should be changed to 0 after the MCMC procedure. Du & Ghosal (2018) also derived the optimal posterior convergence rate in Frobenius norm as $\sqrt{n^{-1}(p+s)\log p}$ with this sparse Cholesky structure.

1.2.2 Measurement Error

In some applications, the variables of interest are not directly observed but are contaminated by some additional measurement error. Therefore, the observations are actually obtained with random impurity from the data generating procedure and some statistical modeling techniques are proposed for estimating the parameters. An introductory example is ordinary linear regression, where the response is measured with error from the Gaussian distribution. More generally, every mixture could be understood as a true model, randomly contaminated by a measurement error model. For example, the error is binary in logistic regression, while it is discrete in Poisson regression. In linear discriminant analysis, the Gaussian distribution is mixed with the Bernoulli distribution, which could be viewed as the resource of error. Assume that some kernel function is mixed with the multivariate Gaussian distribution and we want to estimate the precision matrix of the Gaussian, which formulates the problem of Gaussian graphical model in the presence of measurement error. It is the key problem of this thesis to develop the Bayesian methods to adjust for the different type of measurement error while estimating the precision matrix.

Errors-in-variables problems are commonly encountered in many scientific studies, especially when the predictors or the independent variables are measured coarsely or cannot be directly measured with desired accuracy. Although some treatments for the measurement error in graphical model has hardly been explored in literature yet, some aspects of the errors-in-variables models were summarized in Carroll et al. (2006), which introduces some resources and examples of measurement

error: inaccurate long-term memory, indirect measurement, scientific estimations and personal privacy. We briefly review some examples here corresponding to each of these resources.

- Inaccurate long-term memory: In some cohort studies of medical science, some nutrition levels are imprecisely measured especially in long-term by inaccurate memory, like daily carbohydrate or saturated fat intake.
- Indirect measurement: In some bioassay experiments in botany, the predictors are the amount of the herbicide actually absorbed by the plant that cannot be directly measured. Instead, the researchers control the amount of the herbicide applied to the plants and use it as the predictors, which are generally larger than the absorbed herbicide depending on absorbing ability of each different plant.
- Scientific estimations: In radiology, it is impossible to determine the true radiation dose received by each human body and thus, the researchers estimate it by some other factors, for instance, in an analysis from Hiroshima and Nagasaki explosions survivor data.
- Personal privacy: Sometimes, a random measurement error is artificially included to prevent the leakage of personal records but also to convey some useful information, if there are some methods to adjust for the corruption and find the true insight. Such a method is commonly adopted in medical applications, and in some other fields as well. For instance, some personal records like the monthly income or spending through survey or some public data may not be provided accurately or even be intentionally contaminated due to privacy protection or some other restrictions in sociological or econometric studies.

1.2.2.1 Effects of Measurement Error

For a situation in which the data analyst is either unaware of the measurement error or simply chooses to ignore it, a natural question is *what can go wrong?* As illustrated in Carroll et al. (2006), the measurement error could induce some effects under certain situations. Most importantly, failing to account for the measurement error creates an unacceptably large bias in the estimation and, therefore, certain adjustments are necessary to account for the presence of measurement error and to ensure an accurate estimation. We start with introducing this phenomenon of bias with the basic ordinary linear regression and then discuss about the other two major influences by the measurement error with examples.

Consider the simple linear regression, where $X_1, \ldots, X_n$ are independently generated from $\mathsf{N}(0, \sigma_1^2)$. Let the response $Y_i = \alpha + \beta X_i + \epsilon_i$ with $\epsilon_1, \ldots, \epsilon_n \overset{\text{iid}}{\sim} \mathsf{N}(0, \sigma_2^2)$ independently of $X_1, \ldots, X_n$. Instead of obtaining X_i directly, we observe $W_i = X_i + U_i$, for $i = 1, \ldots, n$, where $U \sim \mathsf{N}(0, \sigma_1^2)$, which shares the same variance as X for the sake of simplicity, and U is independent with ϵ and X. Then, if the measurement error is ignored, by fitting the ordinary linear regression of Y on W for estimating β, we obtain the estimator as the sample covariance of W and Y divided by the sample variance of

W, whose limit is half of the truth given that $\mathsf{V}(X) = \mathsf{V}(U)$, i.e.,

$$\frac{\mathsf{Cov}(W,Y)}{\mathsf{V}(W)} = \frac{\mathsf{Cov}(X+U, \beta X+\epsilon)}{\mathsf{V}(X+U)} = \frac{\beta \mathsf{V}(X)}{\mathsf{V}(X)+\mathsf{V}(U)} = \frac{1}{2}\beta.$$

Comparing this result with the truth β, the bias is non-negligible especially when the magnitude of β is large.

Moreover, this example also shows that the measurement error in the covariates may cause a loss of power, which is profound sometimes for discovering the relationship between the variables of interest. Comparing the estimation $\beta/2$ with the true parameter β, the result is shrunk towards zero, which will increase the p-value for testing whether $\beta = 0$. As shown in Section 1.8 of Carroll et al. (2006), the power of that test decreases as the variance of the measurement error increasing from a simulation study where $\beta = 0.69$ and $\sigma^2 = 1$. Intuitively, consider the case where the variance of X is small but the variance of the measurement error U is large. Then, the fitted line becomes flat and the real trend is hidden if the measurement error is ignored. Therefore, we conclude that the measurement error may mask the features; see another example in Section 1.1 of Carroll et al. (2006) for details.

On the other hand, the inconsistent estimator computed from the contaminated data may not be able to reveal the structure under the true model especially when the sparsity is pursued. Consider the AR(1) model, where $\theta \sim \mathsf{N}_5(0, \Sigma)$ with covariance and precision matrix shown in (1.2.1), and we want to estimate the precision matrix without knowing the existence of measurement error. Suppose that our observations $X_1, \ldots, X_n$ are generated by $X_i = \theta_i + \epsilon_i$ where the θ's are the true outcomes i.i.d. from $\mathsf{N}_5(0, \Sigma)$ and the ϵ's are the measurement errors i.i.d. from $\mathsf{N}_5(0, 0.1 I_5)$. Based on the additive property of the Gaussian distribution, the marginal distribution of X is $\mathsf{N}_5(0, \Sigma + 0.1 I_5)$. Suppose that a consistent estimator of the precision matrix by X is obtained, then the limit of it is indeed

$$(\Omega^{-1} + 0.1 I_5)^{-1} = \begin{pmatrix} 1.535 & -0.910 & -0.098 & -0.011 & -0.001 \\ -0.910 & 2.074 & -0.852 & -0.092 & -0.011 \\ -0.098 & -0.852 & 2.080 & -0.852 & -0.098 \\ -0.011 & -0.092 & -0.852 & 2.074 & -0.910 \\ -0.001 & -0.011 & -0.098 & -0.910 & 1.535 \end{pmatrix}, \tag{1.2.8}$$

which is clearly not close to Ω in (1.2.1). Furthermore, there is no sparsity (zero values) discovered in the estimation (1.2.8), while there are only 4 non-zero entries in the lower triangular of the truth (1.2.1) except the diagonal. In other words, the estimator indicates that all pairs of the variables are conditionally dependent given others, though the truth is that most of them are not.

1.2.2.2 Bias by Measurement Error in Gaussian Graphical Model

As discussed above, the estimation bias is one of the influences due to ignoring the measurement error, which also happens remarkably in the Gaussian graphical model. To see this, we investigate what will happen when the measurement error is ignored, i.e., when a misspecified no-measurement-

error model is fit to the corrupted data $Y_1, \ldots, Y_n$ in (1.2.9). For the sake of simplicity, we use a rather simple measurement error model to explore this, namely,

$$Y_i = X_i + Z_i, \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad Z_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \nu I_p), \quad i = 1, \ldots, n,$$

where the X and Z samples are mutually independent, ν is assumed to be known throughout for the sake of identifiability and the Y's are the only observations. The marginal distribution of the Y's is available in closed-form given by

$$Y_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1} + \nu I_p), \quad i = 1, \ldots, n. \tag{1.2.9}$$

To develop some intuition for the magnitude of the estimation bias, consider the case where the dimension p is fixed, so that the precision matrix can be estimated directly, at least for large n, without imposing any structural or sparse assumptions. Let $\widehat{\Omega}_n = \widehat{\Omega}(Y_1, \ldots, Y_n)$ denote an asymptotically unbiased estimator of the precision matrix ignoring the measurement error based on the corrupted data $Y_1, \ldots, Y_n$, for instance, $\widehat{\Omega}_n = S_n^{-1}$, the inverse of the sample covariance matrix $S_n = n^{-1} \sum_{i=1}^n Y_i Y_i^{\mathrm{T}}$. By asymptotically unbiased, we mean that

$$\| \mathsf{E}_{\Omega^\star, \nu} \widehat{\Omega}_n - (\Omega^{\star -1} + \nu I_p)^{-1} \|_F = o(1), \quad n \to \infty, \tag{1.2.10}$$

where $\Omega^\star$ denotes the true $p \times p$ precision matrix and $\|A\|_F = \{\mathrm{tr}(A^{\mathrm{T}} A)\}^{1/2}$ denotes the Frobenius norm of a matrix A, with $\mathrm{tr}(\cdot)$ the trace operator. Also, let $\|A\|_2$ denote the spectral norm, i.e., the square root of the maximum eigenvalue of $A^{\mathrm{T}} A$.

Theorem 1.2.1. *For a case of fixed dimension p, let $\widehat{\Omega}_n$ be an estimator that ignores the measurement error and satisfies* (1.2.10)*. If ν is fixed, then for all sufficiently large n,*

$$\mathsf{E}_{\Omega^\star, \nu} \| \widehat{\Omega}_n - \Omega^\star \|_F^2 \geq \tfrac{1}{2} \| \Omega^\star (\nu^{-1} I_p + \Omega^\star)^{-1} \Omega^\star \|_F^2. \tag{1.2.11}$$

Moreover, if $\nu \leq \|\Omega^\star\|_2^{-1}$, then the bound can be simplified to

$$\mathsf{E}_{\Omega^\star, \nu} \| \widehat{\Omega}_n - \Omega^\star \|_F^2 \geq \tfrac{1}{8} \nu^2 p \lambda_{\min}^4 (\Omega^\star), \tag{1.2.12}$$

where $\lambda_{\min}(\cdot)$ stands for the minimum eigenvalue.

Proof. A simple bias–variance decomposition yields

$$\mathsf{E}_{\Omega^\star, \nu} \| \widehat{\Omega}_n - \Omega^\star \|_F^2 = \mathsf{E}_{\Omega^\star, \nu} \| \widehat{\Omega}_n - \mathsf{E}_{\Omega^\star, \nu} \widehat{\Omega}_n \|_F^2 + \| \mathsf{E}_{\Omega^\star, \nu} \widehat{\Omega}_n - \Omega^\star \|_F^2.$$

Since the first term is non-negative, we get

$$\mathsf{E}_{\Omega^\star, \nu} \| \widehat{\Omega}_n - \Omega^\star \|_F^2 \geq \| \mathsf{E}_{\Omega^\star, \nu} \widehat{\Omega}_n - \Omega^\star \|_F^2 = \| (\Omega^{\star -1} + \nu I_p)^{-1} - \Omega^\star \|_F^2 + o(1),$$

where the first term dominates as $n \to \infty$ when ν is fixed. By the Woodbury formula $(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}$, the right hand side equals

$$\|\Omega^\star(\nu^{-1}I_p + \Omega^\star)^{-1}\Omega^\star\|_F^2 + o(1) \geq \tfrac{1}{2}\|\Omega^\star(\nu^{-1}I_p + \Omega^\star)^{-1}\Omega^\star\|_F^2,$$

for large enough n, which proves the first assertion.

Now consider $\nu \leq \|\Omega^\star\|_2^{-1}$, the smallest eigenvalue of $\Omega^{\star-1}$. Using a spectral decomposition UDU^{T} of $\Omega^\star$, where U is an orthogonal matrix and $D = \mathrm{diag}(D_{1,1}, \ldots, D_{p,p})$, we obtain $\nu D_{j,j} \leq 1$ for all $j = 1, \ldots, p$. Hence

$$\|\Omega^\star(\nu^{-1}I_p + \Omega^\star)^{-1}\Omega^\star\|_F^2 = \nu^2\|UD(I_p + \nu D)^{-1}DU^{\mathrm{T}}\|_F^2 \geq \tfrac{\nu^2}{4}\|\Omega^{\star 2}\|_F^2.$$

Since $\|\Omega^{\star 2}\|_F^2 = \|D^2\|_F^2 \geq p \cdot \min(D_{j,j}^4 : j = 1, \ldots, p)$, the second assertion follows. □

From the theorem, we can see that the lower bound on the bias will vanish as $\nu \to 0$ but will increase monotonically to $\|\Omega^\star\|_F^2$ as $\nu \to \infty$. Therefore, even in a relatively low-dimensional setting with fixed p, unless ν is vanishingly small, the mean squared error associated to any asymptotically unbiased estimator of the precision matrix is bounded away from 0 as $n \to \infty$.

Moreover, the same proof would apply to a case of increasing dimension if $p = O(n)$ and $\Omega^\star$ is *known* to be diagonal. Since both the fixed-p and known-to-be diagonal $\Omega^\star$ cases are simpler than the general high-dimensional structured precision matrix estimation problem, and the effect of ignoring measurement error is already profound, we conjecture that the estimation bias result will become even worse in the more general setup involving a complex structure and increasing p, which motivates us to explore the approaches to adjust for different types and magnitudes of measurement error in Gaussian graphical model.

1.2.2.3 Treatments for Measurement Error

Under different scenarios, numerous approaches have been proposed to treat for the measurement error. We will briefly review some techniques for the classical measurement error model at first, where the observations are measured with additive error, usually with zero mean and constant variance. Then, we march towards some very recent methods designed specifically towards the Gaussian measurement error in the Gaussian graphical model, which we are not aware of before composing this thesis.

When validation data is available, the contaminated observations could be treated as missing values and imputed using some methods of missing data, such as multiple imputation (Freedman et al., 2008). The validation dataset is supposed to provide the purified observations and these uncontaminated values are sometimes calculated through an approximation by the internal replicates of the impure observations on the same measurement. Given the validation data, one common method adjusting for the measurement error is regression calibration, where the contaminated covariates are regressed with some functional form on the other covariates from the validation

dataset and then the contaminated (missing) data is replaced by the fitted values. It is necessary to adjust the estimated standard error by bootstrap or sandwich method because the imputation introduces extra variation.

When the variance of the measurement error is available or some contaminated replicates are present to estimate the variance, the simulation extrapolation method could be implemented (Cook & Stefanski, 1994). As the variance of measurement error is known or estimated, some simulated data can be generated by adding extra error with different amount of variance on the contaminated covariates. Then the same response is fitted by the simulated dataset of different level of artificial contamination and a series of estimations are obtained. Based on these estimations, an extrapolation is performed to estimate the truth at the point where the bias induced by the extra error is eliminated. We consider simple linear regression example in Section 1.2.2.1 for better illustration. By sequentially adding the noise with variance ξ_j, the estimations $\hat{\beta}_j = \mathsf{V}(X)/\{\mathsf{V}(X)+(1+\xi_j)\mathsf{V}(U)\}$ are obtained. If an extrapolation of $\hat{\beta}$ at $\xi = -1$ can be computed, the extra noise is removed and the unbiased estimation of the truth is recovered.

Consider a more general condition where an instrumental variable T is available instead of the replicates, which provides another direction for correcting the measurement error. This instrumental variable T is allowed to be generated from another independent and probably biased measurement of the unobserved X, while it is extremely important to identify the relationship between X and T carefully and correctly. Failure to recognize this could lead to the erroneous inference regardless of the sample size and the variance of the measurement error. A unified approach using this instrumental variable has not been presented and we recommend to explore a specific method from Chapter 6 in Carroll et al. (2006) or the literatures of instrumental variable.

If the measurement error is assumed to have a normal distribution with mean 0 and known covariance matrix, the score function method could be applied to estimate parameters; see Chapter 7 in Carroll et al. (2006) for more details and analysis in linear regression and logistic regression. If the distribution of the unobservable variable is assumed, the likelihood or quasilikelihood is available, in which case the MLE or some Bayesian methods could be implemented with reference from Chapter 8 and 9 in Carroll et al. (2006), while it is hard to describe these ad hoc methods detailedly.

Regarding the Gaussian graphical model in the presence of Gaussian measurement error, Byrd, Nghiem & McGee (2021) treated the unobservable outcomes as missing data and recently proposed a method to impute them and estimate the precision matrix iteratively. They combined the imputation–regularized optimization algorithm (Liang et al., 2018) and Bayesian regularization for graphical models with unequal shrinkage (Gan, Narisetty & Liang, 2019) to form the whole procedure and proved the consistency of their approach. Their results also revealed the necessity to adjust for the measurement error when it is present. Different from those results, we propose a fully Bayesian method with the theoretical results of the posterior contraction rate. Furthermore, we consider the more general type of measurement error model, which is not necessarily Gaussian.

There are mainly two situations that we considered, where either the variance of the measure-

ment error is known or we have some replicates on the same outcome from the Gaussian graphical model to estimate that variance accurately. Otherwise, if we only have one observation for each uncontaminated outcome and the variance of the measurement error is unknown, the precision matrix will be non-identifiable. Considering the case that m repeated measures are present for each outcome from the Gaussian graphical model, we could then estimate the variance of the measurement error using analysis of variance (ANOVA). Since this estimator typically has a high accuracy if p or n is large, it can be treated as the truth and deployed for estimating the precision matrix using the methods with known variance. Note that this adjustment for the measurement error with repeated measure requires slight change over the case that the variance of the measurement error is known. Therefore, we will assume that the variance of the measurement error is known throughout this thesis and comment on the case with repeated measure in the discussions following the results.

1.2.2.4 Mixture Model

Beyond the Gaussian measurement error, a more general type of measurement error could be considered, especially when the contamination is intentional for preventing the information leakage. However, the simple form of the marginal distribution as shown in (1.2.9) will not be available due to the lack of conjugacy with the Gaussian graphical model. Assume that the observations are obtained from a mixture distribution of the multivariate Gaussian distribution and the distribution of measurement error, which is not necessarily Gaussian. Some common examples include the Student t-distribution from the location-scale family or the Poisson distribution from the exponential family. Some other truncated distributions such as uniform distribution could be deployed as well.

Without loss of generality, assume that $Y_1, \ldots, Y_n$ have the marginal density of the form

$$\int f(y|x, \nu) dG(x; \theta),$$

where the kernel function $f(\cdot|x, \nu)$ is the density of the measurement error model with location parameter x, which is generated from the mixing distribution function $G(\cdot; \theta)$ with some to-be-estimated parameter θ. Many statistical methods emerged from the mixture model in the literature for various of problems according to the different choices of f and G.

A common choice of the kernel function f is Gaussian and some research was conducted on it with different mixing distributions. Pearson (1894) first considered the Gaussian mixture model and used the method of moments to fit it. After the invention of expectation-maximization (EM) algorithm (Dempster, Laird & Rubin, 1977), handling the likelihood of the Gaussian mixture model became easier, which facilitated the computation of MLE. Diebolt & Robert (1994) developed a data augmentation procedure and proposed the Bayesian estimator using the MCMC algorithm. Another generalization of the Gaussian mixture model is the use of the multinomial distribution as the mixing distribution with a Dirichlet hyperpriors on the parameters (Lavine & West, 1992). However, the curse of dimensionality is imposed similarly to the Gaussian graphical model, especially in the high-dimensional problems, and solutions are still to reduce the dimensionality by regularization (Pan &

Shen, 2007; Witten & Tibshirani, 2010) or a lower dimensional structure (Du & Ghosal, 2018). A more robust model using the kernel of the Student t-distribution instead of the Gaussian distribution is proposed by McLachlan & Peel (1998), which suggests a modification of EM algorithm for computing the estimation. Some other popular choices of the kernel function include Beta distribution (Ghosal, 2001; Rousseau, 2010), location-scale mixtures (Jonge & Zanten, 2010; Kruijer, Rousseau & van der Vaart, 2010) and Gamma distribution (Wiper, Insua & Ruggeri, 2001).

In most previous work of mixture model, the domain of the mixing distribution is commonly considered as finite or at least countable for the purpose of classification with a continuous kernel function. However, different from this popular setup, we would like to specify the mixing distribution as the multivariate Gaussian distribution to align with the Gaussian graphical model and explore more general types of the kernel functions, which serves as the measurement error model. To the best of our knowledge, such a problem to estimate the precision matrix has not been explored with the measurement error beyond Gaussian distribution. Even when the measurement error is Gaussian, a fully Bayesian method is still lacking along with its general theoretical results about the convergence rate. In this thesis, we intent to develop the fully Bayesian methods to account for different types of measurement error and obtain the theoretical results of posterior convergence rate with respect to those methods.

1.2.3 Posterior Contraction Rate

For each $n \geq 1$, consider the observations $Y_1, \ldots, Y_n$ sampled from some distribution parameterized by $\theta^\star$, which lies in a parameter space Θ. Given some prior on θ, let $\Pi(\theta | Y_1, \ldots, Y_n)$ denote the corresponding posterior. Then, the posterior contraction rate, or called posterior convergence rate, with respect to the semi-metric d at the truth $\theta^\star$ of the Bayesian procedure is defined as the sequence ϵ_n such that

$$\mathsf{E}_{\theta^\star}\left[\Pi\left\{\theta \in \Theta : d(\theta, \theta^\star) \geq M\epsilon_n \mid Y_1, \ldots, Y_n\right\}\right] \to 0,$$

for some constant $M > 0$. Consistency is automatically obtained for a fixed parameter space if the rate goes to 0 as the sample size $n \to \infty$. The posterior contraction rate is the Bayesian analog of the rate of convergence by a frequentist, which describes the spread of the distribution of the estimator with increasing number of observations. This concept of posterior contraction rate is now a standard judgement of estimations, which has been studied towards different specific models. For the Bayesian nonparametric models, Ghosal, Ghosh & van der Vaart (2000) and Ghosal & van der Vaart (2007) laid out the general sufficient conditions to establish posterior concentration rate results for the i.i.d. and non-i.i.d. observations, respectively.

Since the observations are assumed to be i.i.d. in this thesis, we focus on the general theory by Ghosal, Ghosh & van der Vaart (2000), which gives the rate by verifying certain conditions with respect to a vanilla semi-metric on the space of densities. This vanilla semi-metric is often chosen to be the Hellinger distance, while some equivalent or weaker semi-metrics are treatable as well, such as the Rényi divergence or the total variation metric. Referring to Figure 1.2, there are four

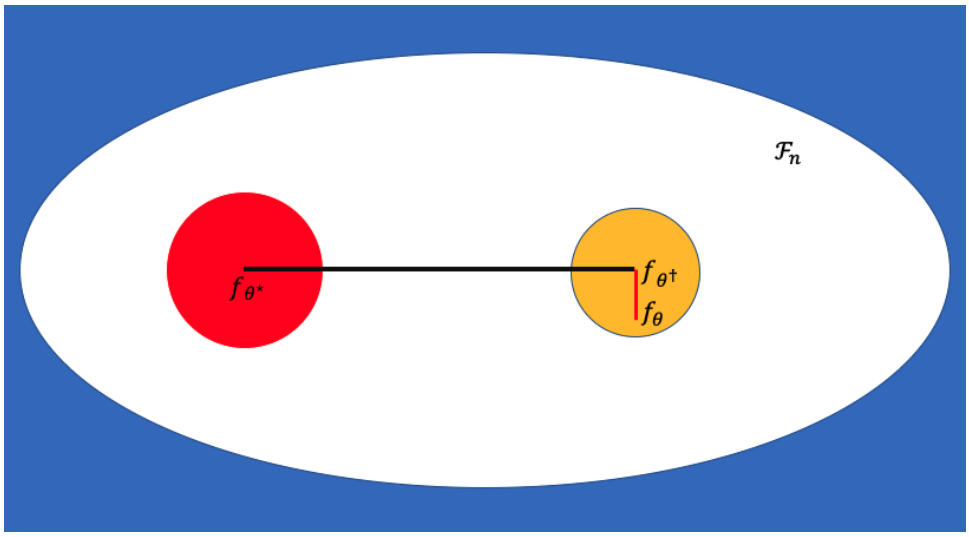

Figure 1.2 An illustration of general theory of posterior contraction rate. The big rectangular denotes the whole parameter space and the white oval denotes the sieve $\mathcal{F}_n$.

conditions to quantify that rate shown as follows:

- A sufficient amount of prior mass around the truth $f_{\theta^\star}$ (red circle).
- There is a test to separate $f_{\theta^\star}$ from every d-ball (yellow circle) centering at $f_{\theta^\dagger}$ with exponentially small error probabilities; The further $f_{\theta^\star}$ and $f_{\theta^\dagger}$ are, the smaller the probabilities are.
- The number of such d-balls in the sieve $\mathcal{F}_n$ (white oval) are at most exponentially large.
- The complement of the sieve $\mathcal{F}_n$ (blue region) has exponentially small prior probability.

The test in the second condition is needed due to the special desire of the semi-metric stronger than the Hellinger distance. For example, in the context of estimating the precision matrix, the vanilla semi-metrics on the space of functions may not be of much interest. Instead, obtaining the rate with respect to a concrete metric such as the L_2-norm for vectors or the Frobenius norm for matrices is more desirable. A contraction rate with respect to such a metric may be obtained after a contraction rate for the density with respect to the Hellinger distance is established by relating the Hellinger distance on the density to the concrete metric. Suppose that we already obtained the posterior contraction rate ϵ_n under a semi-metric d_1 and we wish to obtain that under d_2. By the definition of posterior contraction rate at the true value θ^*, if we can prove the statement that $d_2(\theta, \theta^\star) \leq \epsilon_n$ *is implied by* $d_1(\theta, \theta^\star) \leq C\epsilon_n$ *for some constant* $C > 0$ *and any* $\theta \in \Theta$ under some extra conditions on θ^*, then the same rate is transferred to the new metric d_2. More generally, the bounding relations between the metrics may be non-linear and the two rates may not be the same.

CHAPTER

2

INFERENCE ON A PRECISION MATRIX UNDER GAUSSIAN MEASUREMENT ERROR

2.1 Introduction

As discussed in Section 1.2.2.2, the bias caused by the Gaussian measurement error is not negligible, even when the measurement error variance v is relatively small, especially in the high-dimensional setup. Therefore, our main goal in this chapter is to understand and alleviate the effect of measurement error on Bayesian methods for estimating a high-dimensional structured precision matrix. Since the Bayesian methods for inference on a precision matrix are already rather complicated, for the sake of simplicity and interpretability, we focus here on the additive Gaussian measurement error model, namely,

$$Y_i = X_i + Z_i, \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad Z_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, vI_p), \quad i = 1, \ldots, n, \tag{2.1.1}$$

where the X and Z samples are mutually independent and I_p is the identity matrix of order p. Since the X samples carry information about Ω and the Z samples do not, the observable Y's are "corrupted" by the convolution of informative and non-informative inputs. Therefore, it is equivalent to express (2.1.1) as a hierarchical model, that is,

$$Y_i \mid X_i \sim \mathsf{N}_p(X_i, vI_p), \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad i = 1, \ldots, n. \tag{2.1.2}$$

On the other hand, the convolution can be directly computed through integrating out X and the marginal distribution of the Y's is available in closed-form given by

$$Y_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1} + \nu I_p), \quad i = 1, \ldots, n. \tag{2.1.3}$$

Like in other "deconvolution models", for the sake of identifiability, ν is assumed to be known throughout for most of the results in this chapter.

However, ν is generally unknown in practice. To handle the case where both Ω and ν are unknown, it is necessary to have replication in one form or another. For simplicity, here we assume that we have repeated measurements of each X_i, i.e.,

$$Y_{i,j} = X_i + Z_{i,j}, \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad Z_{i,j} \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \nu I_p), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{2.1.4}$$

where m is the number of replicates for each X. The other two expressions like (2.1.2) and (2.1.3) are available for each replication. We will discuss how to modify the proposed approach to handle the case of unknown ν in Section 2.2.3.

In contrast to the covariance matrix on which the effect of the measurement error is simply additive, the "iterated inverses" $(\Omega^{-1} + \nu I_p)^{-1}$ from the above expression reveals how even the simple linear measurement error model leads to a very non-linear corruption when the goal is estimating the precision matrix. An important quantity in this model is the measurement error variance, ν, which characterizes the magnitude of the measurement errors or the *degree of corruption*. Intuitively, if the measurement error is ignored and ν is not small, then the estimation accuracy of Ω will be negatively affected. Here we develop a general strategy that allows the user to incorporate additive Gaussian measurement error into existing Bayesian procedures for inference on structured, high-dimensional precision matrices in such a way that the posterior contraction rates are preserved and minimal changes to posterior computations are required. To accommodate the measurement error in our theoretical analysis, so that the posterior can effectively undo the "iterated inverses" in (2.1.3), it is crucial to have extra control on the prior distribution of the smallest eigenvalue of Ω. Towards this, we express Ω as $\kappa I_p + \Theta$, and induce a prior on Ω via independent priors on the scalar $\kappa > 0$ and the $p \times p$ matrix Θ. Then the prior on κ gives us eigenvalue control, while the prior for Θ can be any of those from the literature, which is originally specified on the precision matrix since now Θ and Ω share the same structure.

Expressing the matrix of interest as a sum "$\kappa I_p + \Theta$" is a strategy that has appeared already in the literature. Indeed, Fan, Fan & Lv (2008), Fan, Liao & Mincheva (2011), and Pati et al. (2014) used such a decomposition, with Θ having a sparse factor structure (Bartholomew, Knott & Moustaki (2011)), to construct a prior for a high-dimensional *covariance matrix*. To our knowledge, this method has not been used to put a prior on a precision matrix. We obtain the posterior contraction rate of the precision matrix for a prior with such a structure. Our posterior contraction rate result simultaneously covers the measurement error and no-measurement-error cases, and the attained rate parallels that obtained by Pati et al. (2014) for the covariance matrix with respect to the Frobenius

norm, with some improvements.

The remainder of this chapter is organized as follows. Our general framework for incorporating Gaussian measurement error into existing Bayesian procedures for inference on structured, high-dimensional precision matrices is presented in Section 2.2 along with a general result on posterior contraction rates in Gaussian graphical models with measurement error. The main conclusion from the result is that the rate in the absence of measurement error remains in force even when a substantial measurement error is present. The rates obtained from the existing priors for Bayesian structure learning methods without measurement error are discussed in Section 2.3 in the present context. A new prior for the estimation of a precision matrix with a sparse factor structure is proposed and the corresponding posterior contraction rate is illustrated in Section 2.4. The result is new even in the context of a Gaussian graphical model without measurement error. An extensive simulation study for investigating the numerical performance of the proposed model under different magnitudes of measurement error is conducted in Section 2.5, showing that the adjustment towards the measurement error leads to a lower estimation error in terms of Frobenius norm. Finally, proofs of all the theoretical results below are presented in the last two sections.

2.2 Accounting for Measurement Error

2.2.1 Prior and Posterior Distributions

For technical reasons that will be made clear below, when a measurement error is present, it is crucial that we have precise control on the prior distribution of the smallest and the largest eigenvalues of Ω. We introduce a simple device of adding a scalar matrix with an undetermined coefficient κ to the precision matrix to automatically satisfy the requirement. We express the precision matrix as

$$\Omega = \Theta + \kappa I_p, \tag{2.2.1}$$

where Θ is a positive semi-definite matrix and $\kappa > 0$ serves as a lower bound on the smallest eigenvalue of Ω. The strategy is to specify a prior distribution for Ω by assigning prior distributions to Θ and κ independently. That is, the prior Π for Ω is induced from independent priors Π_Θ and Π_κ for Θ and κ, respectively, by the mapping $(\Theta, \kappa) \mapsto \Theta + \kappa I_d$. This term κ is introduced only to automatically assure a lower bound for eigenvalues of the precision matrix Ω in the theoretical results. This structure of the prior, though, is not convenient for computation. Computational issues will be discussed in Section 2.5.1. Details about the specific priors for Θ and κ are presented below.

Prior for κ. As mentioned above, control on the prior distribution of eigenvalues is crucial, so the tails of the prior for κ need to be carefully chosen. In particular, we require exponential tails in both directions, i.e., there exists a constant $C > 0$ such that

$$\Pi_\kappa(\kappa > t) + \Pi_\kappa(\kappa < t^{-1}) \lesssim e^{-Ct}, \quad \text{for all large } t > 0. \tag{2.2.2}$$

A common distribution that satisfies this requirement is the inverse Gaussian (Chhikara, 1988) with density function, in the one-parameter form, given by $\pi_\kappa(t) \propto t^{-3/2} e^{-(t-\xi)^2/(2t)}$, $t > 0$, where $\xi > 0$ plays the role of the mean and variance. More generally, a generalized inverse Gaussian density proportional to $t^a e^{-b(t-\xi)^2/t}$ with any $b > 0$ and $a \in \mathbb{R}$ may be used.

Prior for Θ. Since the structure in Ω is determined by the structure in Θ, we choose Π_Θ in order to induce the desired structure in Ω. Fortunately, most of the existing priors on the precision matrix could be explicitly applied on Θ here. For example, if we believe that Ω has a general sparsity structure, then we could take Π_Θ to be a suitable G-Wishart Lenkoski & Dobra (2011) or a mixture thereof Banerjee & Ghosal (2015). Similarly, structures like a sparse Cholesky decomposition or a sparse factorization can be imposed on Ω with a suitable choice of prior on Θ. Details will be given for a number of special cases in Section 2.3 and Section 2.4 below. Roughly, our technical requirement is that Π_Θ satisfies the sufficient conditions originally laid out in Ghosal, Ghosh & van der Vaart (2000) for posterior contraction at the target rate in the no-measurement-error context. These sufficient conditions have already been verified for various low-dimensional structures and commonly used priors that induce them, so our main focus here can be on the effects of measurement error.

Given a prior for Ω as described above, we update to a posterior distribution via Bayes' theorem. For the measurement error model (2.1.3), define the likelihood function as

$$L_n(\Omega; \nu) \propto |\det(\Omega^{-1} + \nu I_p)|^{-1/2} \exp[-n \operatorname{tr}\{S_n(\Omega^{-1} + \nu I_p)^{-1}\}/2],$$

where $S_n = n^{-1} \sum_{i=1}^n Y_i Y_i^{\mathrm{T}}$ is the sample covariance matrix of Y and det denotes the determinant operator. Then the corresponding posterior distribution, which depends on the data $Y_1, \ldots, Y_n$ and the known measurement error variance, is given by

$$\Pi_n^\nu(d\Omega) = \Pi(d\Omega \mid Y_1, \ldots, Y_n; \nu) \propto L_n(\Omega; \nu)\Pi(d\Omega). \tag{2.2.3}$$

A consequence of this indirect formulation is that the posterior distribution cannot be computed in closed-form. Therefore, MCMC methods are needed to produce samples from Π_n^ν. These can be developed by modifying the corresponding algorithm for the case of measurement error; see Section 2.5.1.

2.2.2 Posterior Contraction Rates

In this subsection, we characterize the posterior contraction rate with respect to the Frobenius distance in terms of the characteristics of the model, the prior, and the true precision matrix. Since even under the maximum sparsity, there are p unrestricted diagonal entries, it is essential that the dimension p is of a smaller order of n, and in particular $\log p$ is bounded by $\log n$. The most interesting case is that $\log p$ and $\log n$ have the same order, for example, $p = n^\alpha$ for some $0 < \alpha < 1$. As discussed above, the intuition here is that if the prior Π_Θ for Θ is such that the posterior would achieve the desired contraction rate without measurement error, and if the prior Π_κ for κ is reasonable in some sense, then the same posterior contraction rate prevails in the presence of measurement error.

The following two conditions make this setup more precise.

Condition 2.2.1. *We specify the following conditions on the prior.*

(a) The prior Π_κ has a continuous density on $(0,\infty)$, and satisfies the tail condition (2.2.2).

(b) Given ϵ_n and posited structure in $\Omega^\star$,

(i) there exists a sieve $\mathscr{S}_n$ of precision matrices, having the same posited low-dimensional structure as $\Omega^\star$, with entropy bound

$$\log N(\delta_n, \mathscr{S}_n, \|\cdot\|_2) \lesssim n\epsilon_n^2, \text{ where } \delta_n = \eta(np)^{-1}, \text{ for any } \eta > 0;$$

(ii) the prior Π_Θ for Θ satisfies $\Pi_\Theta(\mathscr{S}_n^c) \lesssim e^{-Kn\epsilon_n^2}$ for some sufficiently large $K > 0$;

(iii) for any constant $c > 0$, there exists another constant $C > 0$, such that

$$\Pi_\Theta(\{\Theta : \|\Theta - \Theta^\star\|_F \leq c\epsilon_n\}) \gtrsim e^{-Cn\epsilon_n^2}$$

for any $\Theta^\star$ having the same posited structure as $\Omega^\star$.

The latter conditions related to Π_Θ look complicated but, for the most part, these are the now-classical sufficient conditions from Ghosal, Ghosh & van der Vaart (2000) for establishing posterior contraction rate results. Therefore, other authors who have investigated contraction rate properties of posterior distributions under various low-dimensional structures and priors, like in Section 2.3, would likely have checked these conditions already. One noticeable difference is in the entropy bound in Condition 2.2.1 (b)(i), where the radius is proportional to $\delta_n = \eta(np)^{-1}$, which is rather small. However, the dimension is what drives the entropy's magnitude, while the radius only impacts the logarithmic term, so the small δ_n has no significant effect. The following theorem characterizes the contraction rate in an abstract setting.

Theorem 2.2.1. *Assume that $\Omega^\star$ has eigenvalues bounded away from 0. Consider a prior distribution for $\Omega = \Theta + \kappa I_p$ induced from independent prior distributions Π_κ for κ and Π_Θ for Θ. Assume that there exists a sequence ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \gtrsim \log n$ such that Π_Θ satisfies Condition 2.2.1 (b)(i)–(iii) above for $\Omega^\star$, and Π_κ satisfies Condition 2.2.1 (a). Under the model in* (2.1.3)*, for any known and fixed $\nu \geq 0$, the posterior distribution Π_n^ν in* (2.2.3) *contracts at the rate ϵ_n, that is, there exists a constant $L > 0$, depending on $\|\Omega^\star\|_2$ and ν, such that*

$$\mathsf{E}_{\Omega^\star,\nu}\Pi_n^\nu(\{\Omega : \|\Omega - \Omega^\star\|_F > L\epsilon_n\}) \to 0 \quad \text{as } n \to \infty.$$

If $\|\Omega^\star\|_2$ is not bounded, then the conclusion holds with the rate $\epsilon_n' = \|\Omega^\star\|_2^2\epsilon_n$ for any fixed $\nu > 0$ and $\epsilon_n' = \|\Omega^\star\|_2\epsilon_n$ if $\nu\|\Omega^\star\|_2 = O(1)$.

It is remarkable that the posterior contraction rate is not affected by measurement error even when it is not small. The proof of the theorem is given in Section 2.6.1.

2.2.3 Handling Unknown ν

When ν is unknown, that is, under the model in (2.1.4), estimation of ν is necessary. A common choice of such an unbiased estimator is the bias-corrected MLE by ANOVA, which is expressed as

$$\hat{\nu} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{i,j}-\bar{Y}_i)^{\mathrm{T}}(Y_{i,j}-\bar{Y}_i)}{npm-1}, \tag{2.2.4}$$

where $\bar{Y}_i = \sum_{j=1}^{m} Y_{i,j}/m$. From a practical point of view, we could simply replace the unknown ν in the previous developments with the estimate $\hat{\nu}$ defined above, akin to an empirical Bayes approach. Alternatively, we could introduce a prior on ν and implementing the hierarchical Bayesian method, such as the Jeffreys prior, whose maximum a posteriori (MAP) estimation coincides with the estimation by ANOVA. Since the unknown ν changes the model significantly, the theoretical results of posterior contraction rate is currently under exploration as some steps in the proof to be revised. Intuitively, the rate would not change much under the model in (2.1.4), since the estimations from both methods are extremely close to the truth with the error of order $1/\sqrt{np}$ given the constant m. A rigorous proof of this rate requires a justification that the posterior contraction rate of Ω is obtained with a given ν in a shrinking neighborhood of the truth ν^*. Now that ν and $\nu^\star$ need not be the same, ν is assumed to be close to ν^* appropriately and the contraction result is uniform over ν lying in that neighborhood. Such a strategy has been used by Yoo & Ghosal (2016) to obtain the contraction results of both the empirical Bayesian or the hierarchical Bayesian approaches for the nonparametric multivariate regression with unknown error variance.

2.3 Examples

Here we investigate some existing Bayesian methods for structured, high-dimensional precision matrix estimation and show how measurement error can be accommodated in these models. For each case, we consider a prior for Θ from the literature and verify the requirements of Theorem 2.2.1. The prior on κ will be assumed to satisfy Condition 2.2.1 (a), e.g., by choosing an inverse Gaussian distribution. Therefore, the discussion below will focus on the prior for Θ and on verifying Condition 2.2.1 (b) (i)–(iii) for Π_Θ. We assume that both the smallest and largest eigenvalues of the true precision matrix $\Omega^\star$ are bounded away from 0 and ∞ for all the examples in this section. Proofs of the rates derived in Theorems 2.3.1–2.3.3 are given in Section 2.6.2–2.6.4, respectively.

2.3.1 General Sparsity

Banerjee & Ghosal (2015) proposed a Bayesian method to estimate a precision matrix with a general sparse structure in a Gaussian graphical model. In the first example, we adopt their setting as a prior for Θ and verify that all required Condition 2.2.1 (b) (i)–(iii) are satisfied.

Let $\Theta_{i,j}$ denote the entry at the ith row and jth column of Θ and Γ denote the matrix with the (i,j)th entry $\Gamma_{i,j} = \mathbb{1}(\Theta_{i,j} \neq 0)$. The cardinality of $\{(i,j) : 1 \leq i < j \leq p, \Gamma_{i,j} = 1\}$ will be denoted by γ.

Consider the following prior

$$\begin{aligned}\pi(\Theta \mid \Gamma) &\propto \prod_{\Gamma_{i,j}=1} \exp(-\lambda|\Theta_{i,j}|) \prod_{i=1}^{p} \exp(-\lambda\Theta_{i,i}/2), \\ \pi(\Gamma \mid R) &\propto q^{\gamma}(1-q)^{-\gamma+p(p-1)/2}\mathbb{1}(\gamma \leq R),\end{aligned} \tag{2.3.1}$$

where λ is a hyper-parameter and q is a pre-specified probability controlling the sparsity. The smaller q is, the more sparse Θ is. Another factor controlling the sparsity, R, is either given a prior or is taken to be a large enough constant. Since the latter is a trivial case of the former, we demonstrate the main result of posterior contraction rate in the former setting. The prior of R should satisfy

$$\Pi(R > M) \leq \exp(-aM\log M) \tag{2.3.2}$$

for some $a > 0$ and large enough constant M. Such distributions include the Poisson and the binomial distributions. Since the prior (2.3.1) deploys the mixture of the Laplace distribution and a point mass, the posterior computation is so troublesome that the Laplace approximation is implemented.

Theorem 2.3.1. *Assume the same setup as in Theorem 2.2.1 with the priors* (2.3.1) *and* (2.3.2) *or fixed* $R = R_0$ *for general sparsity type of structure. Under the model in* (2.1.3)*, for any known and fixed* $v \geq 0$*, there exists a constant* $L > 0$ *such that the posterior distribution* Π_n^v *in* (2.2.3) *contracts at the rate* ϵ_n *around* $\Omega^\star$*, where* $\epsilon_n = n^{-1/2}(p+s^\star)^{1/2}(\log n)^{1/2}$*, with* $s^\star$ *denoting the number of nonzero off-diagonal entries in* $\Omega^\star$*.*

2.3.2 Sparse Cholesky Decomposition

Assume that the true precision matrix has a sparse Cholesky decomposition $\Theta = UDU^{\mathrm{T}}$, where U is a lower-triangular matrix and D is a diagonal matrix, and we specify a prior on Θ through U and D as in Du & Ghosal (2018). Let $U_{i,j}$ denote the entry at the ith row and jth column of U and $D_{i,i}$ denote the ith diagonal entry of D. Let Γ denote the matrix formed by the indicators $\Gamma_{i,j} = \mathbb{1}(U_{i,j} \neq 0)$. Following Proposition 1 in Du & Ghosal (2018), for $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, i$, consider the prior

$$\begin{aligned}(U_{i,j} \mid \Gamma_{i,j}) &\sim (1-\Gamma_{i,j})\mathsf{N}_p(0, \sigma_0^2) + \Gamma_{i,j}\mathsf{N}_p(0, \sigma_1^2), \\ \Gamma_{i,j} &\sim \mathsf{Bernoulli}(C_p/\sqrt{i}), \\ D_{i,i} &\sim \mathsf{Gamma}(\alpha_1, \beta_1),\end{aligned} \tag{2.3.3}$$

where α_1, β_1, σ_0^2 and σ_1^2 are some hyper-parameters and C_p is a constant only depending on p. Du & Ghosal (2018) states that the result is not sensitive to the choice of C_p from the simulation study, which also matches with our observation here.

An alternative to the above prior is to consider more general $\Gamma_{i,j}$ positive real-valued, and induce

a prior on $U_{i,j}$ through the hierarchical scheme

$$\begin{aligned}
&(U_{i,j} \mid \Gamma_{i,j}) \sim \mathsf{N}_p(0, \Gamma_{i,j}^2 \tau^2), \\
&\Gamma_{i,j} \sim \mathsf{Cauchy}^+(0,1), \\
&D_{i,i} \sim \mathsf{Gamma}(\alpha_1, \beta_1),
\end{aligned} \tag{2.3.4}$$

for $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, i$, where $\mathsf{Cauchy}^+(0,1)$ is the positive half-Cauchy distribution and τ is a pre-specified global shrinkage parameter.

For both setups, let γ denote the number of $U_{i,j}$'s greater than ϵ_n/p, where ϵ_n is introduced as below, and γ will have a binomial distribution as $\mathsf{Bin}(p(p-1)/2, \eta)$, where

$$\eta = \Pi(|U_{i,j}| > \epsilon_n/p) \leq p^{-b} \quad \text{and} \quad \eta \geq p^{-a}, \tag{2.3.5}$$

with some constants $a, b > 2$ as we assumed for any $0 < j < i \leq p$. This restriction on η can be achieved by either choosing C_p in (2.3.3) going to zero as $p \to \infty$ polynomially in $1/p$ or modifying the prior $\Gamma_{i,j}$ in (2.3.4) to be truncated above by $1/\tau$, where $\tau < p^{-b-2}\epsilon_n^2$. Although the condition (2.3.5) is required for the theoretical result, it is not suggested to specify the hyper-parameters in practice for the convenience of computation.

Theorem 2.3.2. *Consider the setup of Theorem 2.2.1 with the priors given by* (2.3.3) *or* (2.3.4) *satisfying* (2.3.5) *for the sparse Cholesky decomposition. Then under the model in* (2.1.3)*, for any known and fixed $v \geq 0$, the posterior distribution Π_n^v in* (2.2.3) *contracts at the rate ϵ_n around $\Omega^\star$, where $\epsilon_n = n^{-1/2}(p+s^\star)^{1/2}(\log n)^{1/2}$, with $s^\star$ denoting the number of nonzero off-diagonal entries in $U^\star$.*

2.3.3 Banded Structure Using G-Wishart Prior

Following Banerjee & Ghosal (2014), they assumed a k-banded structure on the precision matrix and used a G-Wishart prior. They derived the posterior convergence rate under such structure and prior with respect to the spectral norm. In this example, we consider the same structure and prior but move our attention to the rate of Frobenius norm in the presence of measurement error.

Suppose that the true $p \times p$ dimensional, k-banded precision matrix $\Theta^\star$, that is, $\Theta^\star_{i,j} = 0$ for all $i, j = 1, \ldots, p$, such that $|i-j| > k$, with a fixed known value of k. A graphical Wishart distribution prior $\Theta \sim \mathsf{G\text{-}Wish}(\delta, I_p)$, is assigned on Θ, where the graph G is induced by the k-banding. It is easy to conclude that the graph is decomposable with cliques $C_j = \{j, \ldots, j+k\}$, $j = 1, \ldots, p-k$, and separators $S_j = \{j, \ldots, j+k-1\}$, $j = 2, \ldots, p-k$; see Banerjee & Ghosal (2014). An important property we use is that given $\Theta_{S_2}, \ldots, \Theta_{S_{p-k}}$, the matrices $\Theta_{C_1}, \ldots, \Theta_{C_{p-k}}$ are conditionally independent and are Wishart distributed with δ degrees of freedom; here and elsewhere for a matrix M and $S \subset \{1, \ldots, p\}$, $M_S = (\!(M_{i,j} : i, j \in S)\!)$, the principal minor of M formed by the entries of S. Let $\mathscr{P}_G$ stand for the cone of positive definite matrices compliant with the graphical structure, that is, the (i,j)th entry is 0 if (i,j) is not an edge of the graph.

Theorem 2.3.3. *Consider the setup of Theorem 2.2.1 with prior* $\Theta \sim \mathsf{G\text{-}Wish}(\delta, I_p)$. *Assume that the eigenvalues of* $\Omega^\star$ *are bounded and bounded away from zero, and for a sufficiently small* $\epsilon > 0$, $\{\Omega : \|\Omega - \Omega^\star\|_\infty < \epsilon\} \subset \mathscr{P}_G$. *Under the model in* (2.1.3), *for any known and fixed* $v \geq 0$, *the posterior distribution* Π_n^v *in* (2.2.3) *contracts at the rate* ϵ_n *around* $\Omega^\star$, *where* $\epsilon_n = n^{-1/2}(p \log n)^{1/2}$.

2.4 Sparse Factor-model Structure

The sparse factor-model structure has been used in the literature to develop prior distributions for structured, high-dimensional covariance matrices, e.g., in Pati et al. (2014). However, to our knowledge, such a prior has not been proposed for a structured precision matrix, even when no measurement error is present. So we separate it from the examples in the previous section because of the novel use of the prior for estimating the precision matrix and new results on posterior contraction rate even in a model without measurement error.

Consider the model (2.1.1), where the possibility $v = 0$ (i.e. the no-measurement error model $X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1})$) is not ruled out. Following our discussion in Section 2.2.1, we assume the precision matrix Ω to be of the form

$$\Omega = \Theta + \kappa I, \quad \Theta = \Lambda\Lambda^{\mathrm{T}},$$

where Λ is a $p \times k$ matrix with $k \leq p$ and at most s non-zero entries on each of the k columns. For a given k, let Γ denote the matrix with the (i, j)th entry $\Gamma_{i,j} = \mathbb{1}(\Lambda_{i,j} \neq 0)$, and let $\gamma \leq ks$ denote the total number of non-zero entries of Λ.

We follow Pati et al. (2014) and impose the following assumptions on the true precision matrix $\Omega^\star$, the corresponding factor-loading matrix $\Lambda^\star$, its dimension $k^\star$, and the inherent error $\kappa^\star$. We assume that the true precision matrix $\Omega^\star$ has also the factor model structure of the form $\Omega^\star = \Lambda^\star\Lambda^{\star\mathrm{T}} + \kappa^\star I$ where $\Lambda^\star \in \mathbb{R}^{p \times k^\star}$ and $k^\star \ll p$. In high-dimensional setup, we typically assume there are two sequences bounding the column sparsity of the true loading matrix $\Lambda^\star$ as $s^\star$ and the largest eigenvalue of true precision matrix $\Omega^\star$ as $c^\star$. Furthermore, we let $\Gamma^\star$ denote the matrix of $\mathbb{1}(\Lambda^\star_{i,j} \neq 0)$ and $\gamma^\star = \sum_{i,j} \Gamma^\star_{i,j} \leq k^\star s^\star$.

Assumptions. Suppose that there exist $c^\star$, $k^\star$ and $s^\star$ such that the following hold:

(A1) $1/c^\star \leq \kappa^\star \leq c^\star/2$ and $\|\Lambda^\star\|_2^2 \leq c^\star/2$ such that $1/c^\star \leq 1/\|\Omega^{\star-1}\|_2^{-1} \leq \|\Omega^\star\|_2 \leq c^\star$;

(A2) $(c^{\star 5} s^\star k^\star)^{1/2}(\log n) \lesssim n^{1/2}$ (improved to $(c^{\star 3} s^\star k^\star)^{1/2}(\log n) \lesssim n^{1/2}$ when $v = 0$);

(A3) each column of $\Lambda^\star$ has at most $s^\star$ non-zero entries.

Assumption (A1) is to give control on the upper and lower bound of the true precision matrix. The Assumption (A2) is introduced to control the final convergence rate appropriately. Assumption (A3) controls the sparsity, which is crucial in high-dimensional problems.

For the Bayesian model formulation, let $\Lambda_{i,j}$ denote the entry at the ith row and jth column of the factor-loading matrix Λ. Then, we consider the spike-and-slab prior similar to Pati et al.

(2014), except that we make certain choices to meet Condition 2.2.1 (a) such as the inverse Gaussian distribution. For $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, k$, let the priors

$$\begin{aligned} &\kappa \sim \mathsf{invGaussian}(\mu_1, \lambda_1), \\ &k \sim \mathsf{Pois}(\theta_1), \\ &(\Lambda_{i,j} \mid \Gamma_{i,j}) \sim (1-\Gamma_{i,j})\delta_0 + \Gamma_{i,j}\mathsf{N}_p(0, \sigma_1^2), \\ &\Gamma_{i,j} \sim \mathsf{Bernoulli}(\eta), \end{aligned} \tag{2.4.1}$$

where δ_0 represents the Dirac distribution at zero and μ_1, λ_1, θ_1, $\sigma_1^2 \geq 1$ and η are all pre-specified hyper-parameters. The condition on variance σ_1^2 is natural because with the point mass at zero, large variation is preferable. In such a prior setup, given $k = k^\star$, γ will have a binomial distribution as $\mathsf{Bin}(p\,k^\star, \eta)$, where $\eta = \pi(|\Lambda_{i,j}| > 0)$ for any $0 < i \leq p$ and $0 < j \leq k^\star$. The choice of η is made to guarantee that $\eta \asymp (p\,k^\star)^{-1}$. Moreover, with such a specification, we have the prior probability $\eta/2 \leq \pi(|\Lambda_{i,j}| > \epsilon_n/4\sqrt{c^{\star 3} p\,k^\star}) \leq \eta$ with ϵ_n introduced in Theorem 2.4.1.

In general, it is not easy to specify a bound $k^\star$ that controls the sparsity of $\Lambda^\star$ and, hence, it is difficult to specify an appropriate η. The problem can be addressed by putting a further prior on η:

$$(\eta \mid k) \sim \mathsf{Beta}(1, a\,k\,p + 1), \tag{2.4.2}$$

where a is the only new extra pre-specified hyper-parameter. However, when this deeper hierarchical model is utilized, there is a slight loss in terms of the posterior contraction rate in Theorem 2.4.1. With such a hyper-prior on η, given $k = k^\star$, we can calculate that

$$\begin{aligned} \pi(|\Lambda_{i,j}| > 0) &= (a\,k^\star p + 2)^{-1} \\ \pi(|\Lambda_{i,j}| > \epsilon_n/4\sqrt{c^{\star 3} p\,k^\star}) &\asymp (a\,k^\star p + 2)^{-1}, \end{aligned}$$

for any $0 < i \leq p$ and $0 < j \leq k^\star$.

Theorem 2.4.1. *Suppose that the data are generated from* (2.1.1) *and Assumptions (A1), (A2) and (A3) hold for the true precision matrix* $\Omega^\star$. *Consider the prior given by* (2.4.1). *Then the posterior distribution* Π_n^ν *in* (2.2.3) *contracts at the rate* ϵ_n *around* $\Omega^\star$, *where*

- $\epsilon_n = n^{-1/2}(c^\star)^{3/2}(s^\star k^\star)^{1/2}(\log n)^{1/2}$ *if* $\nu c^\star = O(1)$,
- *and* $\epsilon_n = n^{-1/2}(c^\star)^{5/2}(s^\star k^\star)^{1/2}(\log n)^{1/2}$ *if* ν *is fixed and positive.*

If the prior (2.4.2) *is imposed on* η, *then the rates are*

- $\epsilon_n = n^{-1/2}(c^\star)^{3/2}(s^\star k^\star)^{1/2}(\log n)$ *if* $\nu c^\star = O(1)$,
- $\epsilon_n = n^{-1/2}(c^\star)^{5/2}(s^\star k^\star)^{1/2}(\log n)$ *if* ν *is fixed and positive.*

2.5 Numerical Results

2.5.1 Computation

The existing literature often provides algorithms for sampling from the posterior distribution of Ω in the no-measurement-error context, and there is a simple and intuitive way to leverage these tools for sampling in cases with measurement error. Because the Bayes model in the measurement error case provides a joint distribution for the triple (X, Y, Ω), it is possible to write down the full conditionals as

$$(X_i \mid Y_i, \Omega) \overset{\text{ind}}{\sim} \mathsf{N}_p((I_p + \nu\Omega)^{-1} Y_i, \nu(I_p + \nu\Omega)^{-1}), \quad i = 1, \ldots, n, \tag{2.5.1}$$

$$(\Omega \mid X_1, \ldots, X_n) \sim \Pi(\Omega \mid X_1, \ldots, X_n), \tag{2.5.2}$$

where $\Pi(\Omega \mid X_1, \ldots, X_n)$ is the posterior based on the no-measurement-error model, with the augmented dataset $X_1, \ldots, X_n$. Note that the $Y_1, \ldots, Y_n$ disappear in (2.5.2) because Ω is conditionally independent of Y, given X. Therefore, if we know how to sample from the no-measurement-error posterior distribution—e.g., using the algorithms available in the published literature on this topic—then we can easily embed this into a Gibbs sampling framework wherein we iteratively sample from this set of full conditionals and obtain a posterior sample of precision matrices that accommodates the known measurement error.

To execute step (2.5.2) efficiently, the prior on Ω needs to be convenient to work with. The assumed structure of $\Omega = \Theta + \kappa I_p$ in the theoretical results in Section 2.2 is undoubtedly not convenient for computation. The role of κ is solely as a technical device to ensure a lower bound for the eigenvalues of Ω. If the prior on Θ already ensures a bound on the eigenvalues, then the additional term is not needed even for the theory. For practical applications, the additional term κI may not make a noticeable numerical difference and may sometimes be dropped, provided that this does not cause any instability in inverses, and simulations give sensible results. The numerical results presented below employ this simplification.

When ν is unknown under the model in (2.1.4), we can implement the hierarchical Bayesian or the empirical Bayesian method as discussed in Section 2.2.1. For the hierarchical Bayesian method, the prior of ν is chosen as the non-informative Jeffreys prior, that is, $\pi(\nu) \propto 1/\nu^{3/2}$. According to this prior, the full conditional posterior of ν is in a closed form as inverse-Gamma distribution as

$$(\nu \mid X_1, \ldots, X_n, Y_{1,1}, \ldots, Y_{n,m}) \sim \mathsf{IG}\left(\frac{mpn}{2}, \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{i,j} - X_i)^{\mathrm{T}}(Y_{i,j} - X_i)}{2}\right), \tag{2.5.3}$$

from which we could obtain the posterior samples of ν iteratively by incorporating it into the MCMC algorithm with steps (2.5.1) and (2.5.2). Since now we have replicates of each outcome X_i, its full

conditional posterior is changed and step (2.5.1) should now become

$$(X_i \mid Y_{i,1},\ldots,Y_{i,m},\Omega,\nu) \overset{\text{ind}}{\sim} \mathsf{N}_p\left((mI_p+\nu\Omega)^{-1}\sum_{j=1}^{m} Y_{i,j},\, \nu(mI_p+\nu\Omega)^{-1}\right), \quad i=1,\ldots,n. \tag{2.5.4}$$

Given the initial values, we iteratively sample ν, X_i's and Ω from the posteriors by Gibbs sampler through steps (2.5.3), (2.5.4) and (2.5.2). If the empirical Bayesian method is preferred, the unknown ν in step (2.5.4) is simply replaced by $\hat{\nu}$ in (2.2.4) and then the algorithm is executed through steps (2.5.4) and (2.5.2).

2.5.2 Simulations

We conduct a simulation study over different structures of true precision matrix and magnitudes of measurement error. We consider the situation with known ν and fix the dimension $p=50$ and sample size $n=100$. Let $\Omega^\star_{i,j}$ denote the entry in the ith row and jth column of the true precision matrix $\Omega^\star$, and consider the following four sparse structures for $\Omega^\star$:

- AR(1): $\Omega^\star_{i,i}=10$ and $\Omega^\star_{i,i-1}=\Omega^\star_{i,i+1}=5$ for $1\le i\le p$; $\Omega^\star_{i,j}=0$ otherwise.
- AR(2): $\Omega^\star_{i,i}=10$, $\Omega^\star_{i,i-1}=\Omega^\star_{i,i+1}=5$ and $\Omega^\star_{i,i-2}=\Omega^\star_{i,i+2}=2.5$ for $1\le i\le p$; $\Omega^\star_{i,j}=0$ otherwise.
- Block(2): $\Omega^\star_{i,i}=10$, $\Omega^\star_{i,j}=5$ for $(k-1)p/2+1\le i\ne j\le kp/2$ and $1\le k\le 2$; $\Omega^\star_{i,j}=0$ otherwise.
- Block(5): $\Omega^\star_{i,i}=10$, $\Omega^\star_{i,j}=5$ for $(k-1)p/5+1\le i\ne j\le kp/5$ and $1\le k\le 5$; $\Omega^\star_{i,j}=0$ otherwise.

For each sparse structure, 100 replicates are run. We consider the priors assigned on the Cholesky decomposition structure in Section 2.3.2 and the setup introduced as (2.3.3) to estimate the precision matrix, since all the the true precision matrices $\Omega^\star$ described above have a sparse Cholesky decomposition and this prior will induce a posterior with fast and simple MCMC algorithm. Since our main interest is about the influence of the measurement error, we will not survey any other types of priors in this simulation study. The hyper-parameters are specified as $\alpha_1=\beta_1=1/2$ and $C_p=1$, since it is unrealistic to specify a too large C_p. To show the different magnitude of influence by the priors, we consider two combinations of the spike-and-slab prior:

- Diffuse prior: $\sigma_0^2=0.01$ and $\sigma_1^2=10$.
- Informative prior: $\sigma_0^2=0.0001$ and $\sigma_1^2=1$.

We use the Gibbs sampling technique introduced in Section 2.5.1 to sample from the posterior and choose Y_i's as the initial values of X_i's for $i=1,\ldots,n$, respectively. To show the effectiveness of the proposed method that adjusts for the measurement error, we compare the estimator after adjustment with that ignoring the measurement error. The estimation error is noted as "adjust" and "ignore" in the graphs. The estimator is the posterior mean and the Frobenius norm estimation errors are given in Figures 2.1, 2.2, 2.3, and 2.4 for the four structures, respectively. In these figures, only

the central 90% of the estimation errors are displayed to remove some outliers, which are caused by the extra variation from measurement error in such a finite sample situation. Note that the x-axis in these figures denotes $\log_{10} \nu$, where ν is the measurement error variance. In other words, ν varies from 10^{-2} to 10 over 13 different values, which are equally spaced on the log-scale.

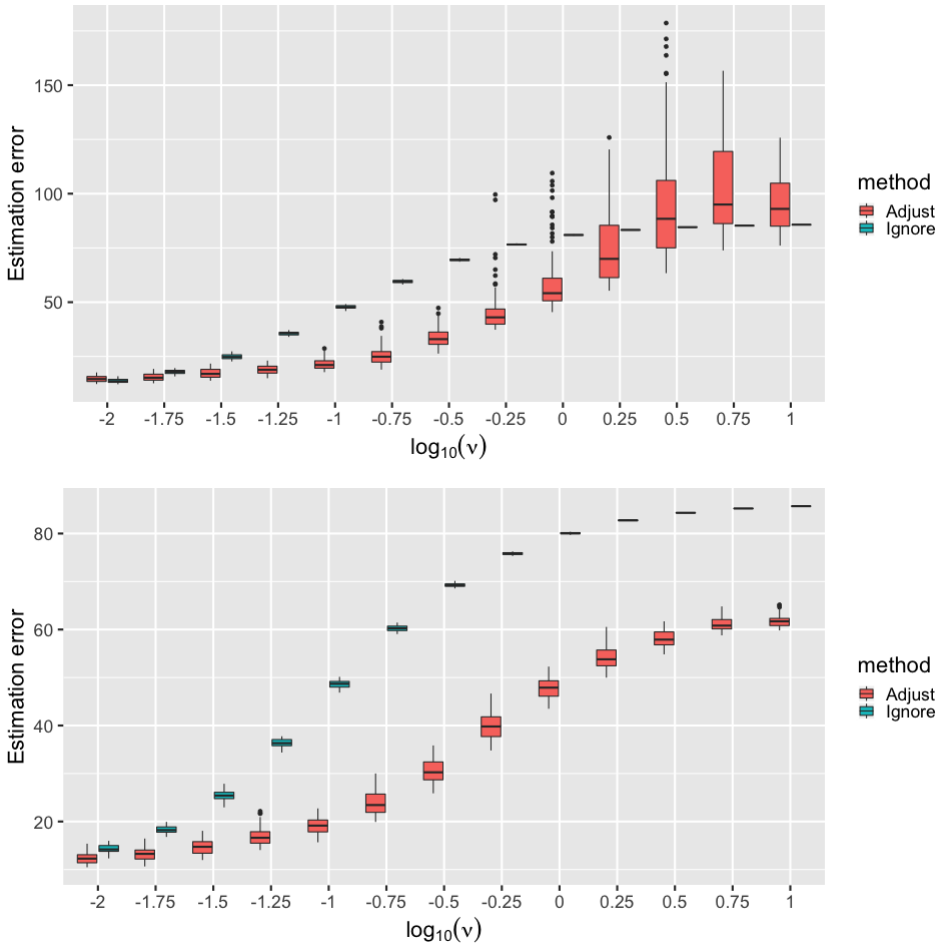

Figure 2.1 Boxplots of the estimation error in terms of the Frobenius norm using diffuse prior (top) and informative prior (bottom) in the AR(1) model over different magnitude of measurement error following the uniform distribution.

For the results of AR(1) model in Figure 2.1, the estimator that corrects for the measurement error has better accuracy in terms of Frobenius norm except when ν is relatively large using the diffuse prior. At the same time, the variance of the estimation error becomes larger when $\nu > 1$, compared with the variance of the baseline model, which even ignores the measurement error. This inflation of the variance in the adjusting model is due to the relatively small sample size compared to the relatively large dimension, as well as the extra variation introduced by the measurement error. This large variance means a lack of sufficient information about the signal in the data, so the estimation

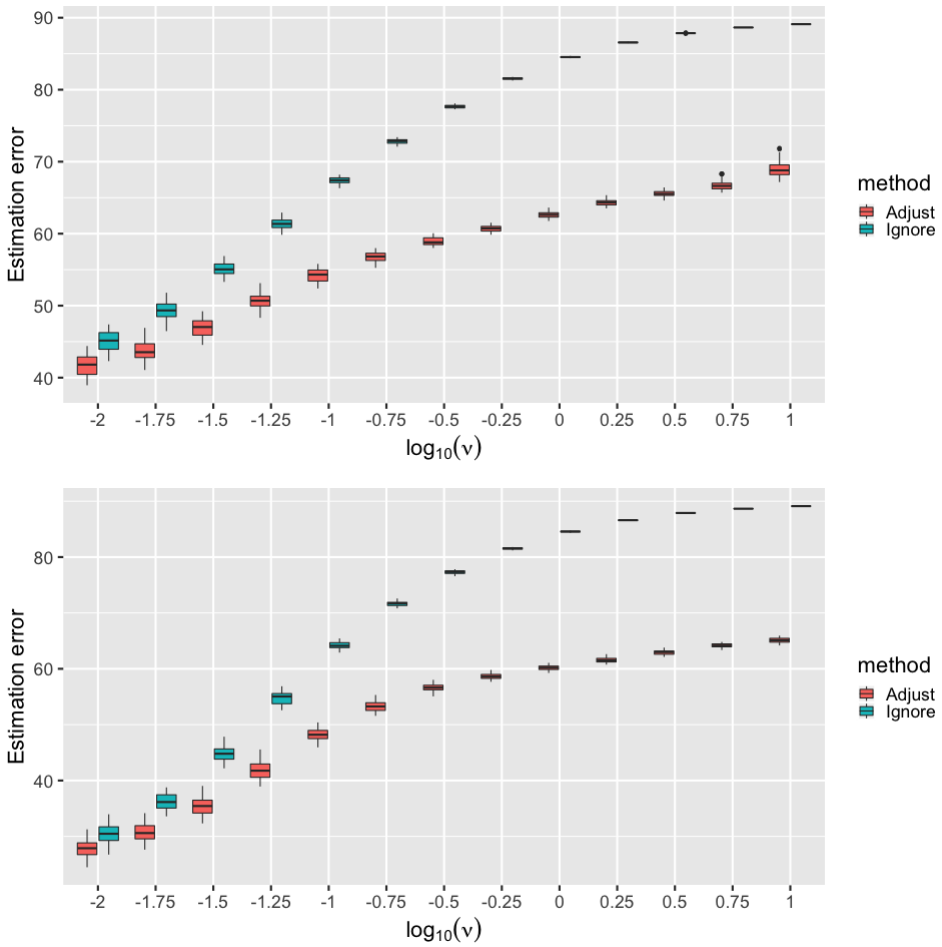

Figure 2.2 Boxplots of the estimation error in terms of the Frobenius norm using diffuse prior (top) and informative prior (bottom) in the AR(2) model over different magnitude of measurement error following the uniform distribution.

with the diffuse prior becomes problematic when ν is large. When a more informative prior is used, the variance is more stable and our procedure beats the naive method uniformly as shown in the right plot of Figure 2.1. On the other hand, since the estimator is close to the zero matrix when the magnitude of the measurement error is large, its error becomes more stable and the variance is tiny. This assertion can be verified by comparing the error with large ν and the Frobenius norm of the true precision matrix listed above. This is the reason why we choose the entries of the true precision matrix relatively large such that the bias could be more dominant when the measurement error is ignored even with a small variance.

When the structure of the precision matrix is more complex, such as in AR(2), the proposed method performs uniformly better than the naive one, and the error stabilizes over different measurement error scenarios in Figure 2.2. Similar phenomena are observed in the block structures in Figures 2.3 and 2.4.

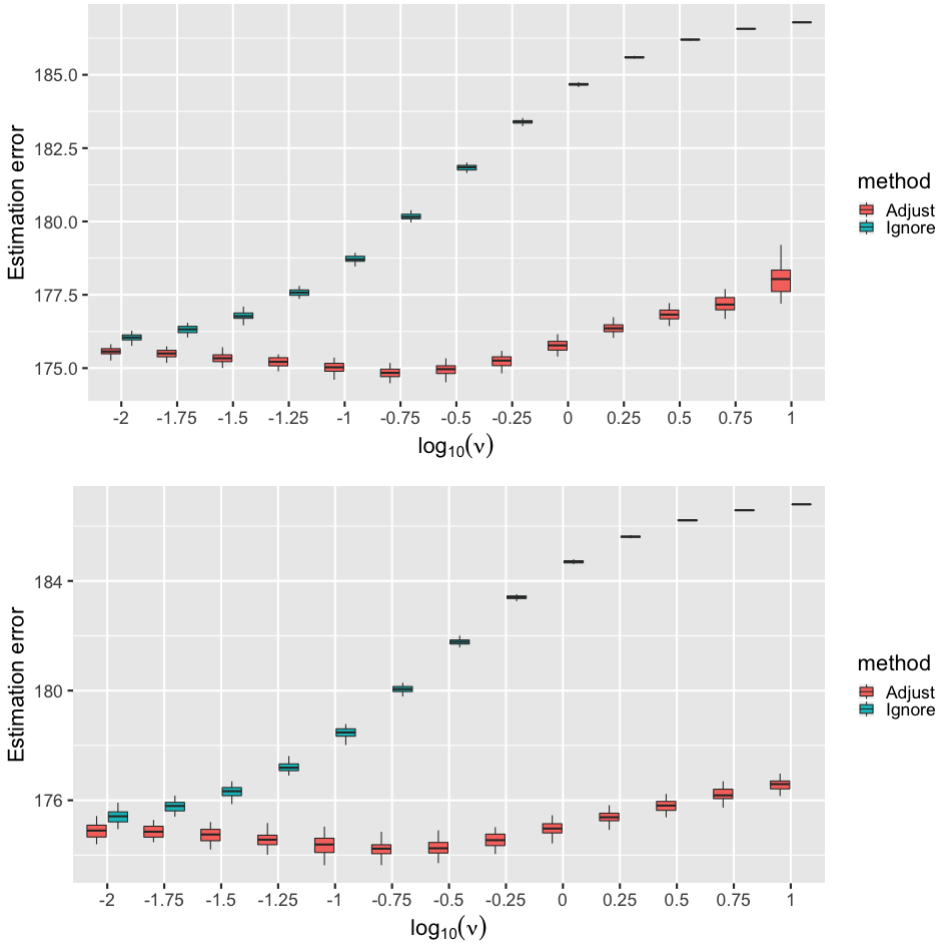

Figure 2.3 Boxplots of the estimation error in terms of the Frobenius norm using diffuse prior (top) and informative prior (bottom) in the Block(2) model over different magnitude of measurement error following the uniform distribution.

2.6 Proofs of the Theorems

2.6.1 Proof of Theorem 2.2.1

The proof will proceed in a sequence of steps driven by those sufficient conditions in Ghosal, Ghosh & van der Vaart (2000) for bounding the posterior contraction rate.

Recall that the prior for Ω is based on independent priors for the ingredients Θ and κ in the representation $\Omega = \Theta + \kappa I_p$ in (2.2.1). For a given true $\Omega^\star$, in the proof we consider the corresponding representation

$$\Omega^\star = \Theta^\star + \kappa^\star I_p.$$

But this decomposition is not unique—there are many $\Theta^\star$ and $\kappa^\star$ that would satisfy this equation, e.g., fix a weight $w \in (0,1)$, set $\kappa^\star = w\lambda_{\min}(\Omega^\star)$, and then $\Theta^\star = \Omega^\star - \kappa^\star I_p$. Fortunately, this non-uniqueness does not affect us here, since the same conclusion is reached for every choice of $(\Theta^\star, \kappa^\star)$ that satisfy

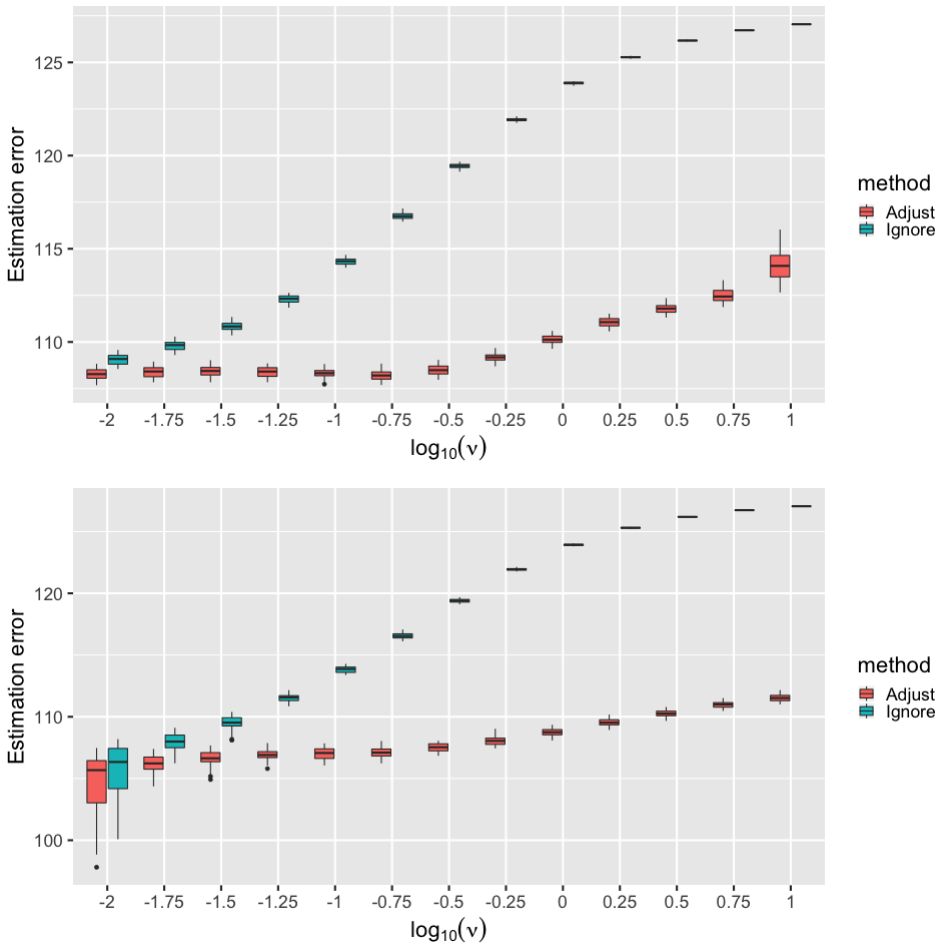

Figure 2.4 Boxplots of the estimation error in terms of the Frobenius norm using diffuse prior (top) and informative prior (bottom) in the Block(5) model over different magnitude of measurement error following the uniform distribution.

the above display, provided the corresponding assumptions hold there. Indeed, recall that, e.g., Condition 2.2.1 (b) requires that concentration of the prior Π_Θ hold around *some* $\Theta^\star$ sharing the same assumed structure in $\Omega^\star$.

Step 1: *Prior concentration.* For a generic Ω, let g_Ω and f_Ω denote the $\mathsf{N}_p(0,\Omega^{-1})$ and $\mathsf{N}_p(0,\Omega^{-1}+\nu I_p)$ densities, respectively, so that the data in (2.1.3) are i.i.d. with density $f_{\Omega^\star}$. For the target rate ϵ_n, we aim to show that for some $C>0$,

$$\Pi(\{\Omega : K(f_{\Omega^\star}, f_\Omega) \le \epsilon_n^2, V(f_{\Omega^\star}, f_\Omega) \le \epsilon_n^2\}) \gtrsim e^{-Cn\epsilon_n^2}, \tag{2.6.1}$$

where, as defined above, $K(f,g)=\int f\log(f/g)$ and $V(f,g)=\int f\log^2(f/g)$ denote the Kullback–Leibler (KL) divergence and corresponding KL variation, respectively, for any two probability densities f and g. From the information loss inequalities in Ghosal & van der Vaart (2017, Lemma B.11),

and Lemma 2.7.1 below, the above prior probability can be lower-bounded by

$$\Pi(\{\Omega : K(g_{\Omega^\star}, g_\Omega) \le \epsilon_n^2/5, V(g_{\Omega^\star}, g_\Omega) \le \epsilon_n^2/5\}). \tag{2.6.2}$$

By our assumption that the eigenvalues of $\Omega^\star$ are bounded away from 0 and infinity, and Lemma 2.7.2, there exists $c > 0$ such that

$$\Pi(\{\Omega : K(g_{\Omega^\star}, g_\Omega) \le \epsilon_n^2/5, V(g_{\Omega^\star}, g_\Omega) \le \epsilon_n^2/5\}) \gtrsim \Pi(\{\Omega : \|\Omega - \Omega^\star\|_F \le c\,\epsilon_n\}). \tag{2.6.3}$$

Replacing Ω by (Θ, κ) and $\Omega^\star$ by $(\Theta^\star, \kappa^\star)$, the triangle inequality gives

$$\|\Omega - \Omega^\star\|_F \le \|\Theta - \Theta^\star\|_F + \|\kappa I_p - \kappa^\star I_p\|_F \le \|\Theta - \Theta^\star\|_F + p^{1/2}|\kappa - \kappa^\star|. \tag{2.6.4}$$

By Condition 2.2.1 (a) on Π_κ and $n\epsilon_n^2 \gtrsim \log n$, we get that for some constant $G > 0$,

$$\Pi_\kappa(\{\kappa : |\kappa - \kappa^\star| \le c\,p^{-1/2}\epsilon_n/2\}) \gtrsim p^{-1/2}\epsilon_n = \exp(\log \epsilon_n - \tfrac{1}{2}\log p) \gtrsim e^{-Gn\epsilon_n^2}. \tag{2.6.5}$$

Combining (2.6.3)–(2.6.5) and using Condition 2.2.1 (b) on Π_Θ, (2.6.2) follows, establishing (2.6.1).

Step 2: *Sieve, test construction, and error rates.* In order to apply the theory of posterior contraction in Ghosal & van der Vaart (2017) to show that the contraction rate at the truth with respect to a distance d is ϵ_n, we need to establish a test for the true value against most of the complement of the d-neighborhood of size ϵ_n around the truth with error probabilities decaying exponentially in $n\epsilon_n^2$. This test is obtained by combining tests for the truth against small balls with centers separated from the truth by at least ϵ_n. Whether such a test for the truth against a small ball exists depends on the metric. If d is the Hellinger metric on the corresponding densities, then such a test exists by celebrated existence theorems. For other metrics, either such a test has to be constructed directly in the given situation, or the distance has to be dominated by a multiple of the Hellinger distance near the true value. In the present context, the distance of interest is the Frobenius distance on the precision matrix, which is not directly comparable with the Hellinger distance for arbitrary pairs of matrices. Here, we construct the required test directly from the likelihood ratio tests of the truth against separated simple alternatives, which can be elegantly quantified by the Rényi divergence, similar to the strategy pursued by Ning, Jeong & Ghosal (2020) and Jeong & Ghosal (2021). The structure of multivariate normality allows a useful control over the size of the likelihood ratios essential for this approach to work.

After the basic tests are constructed, these need to be combined, which is possible if their number can be controlled appropriately, and in particular, there are only finitely many covering sets with centers separated from the truth. This necessitates defining a sieve, a sequence of increasing subsets of the parameter space, on which the number of covering sets can be controlled, and the complement of the sieve has an exponentially small prior probability. We shall work with the sieve

$$\mathscr{T}_n = \{\Theta + \kappa I_p : \Theta \in \mathscr{S}_n, M_n^{-1} \le \kappa \le M_n\}, \tag{2.6.6}$$

with $\mathscr{S}_n$ as in the statement of the theorem and we choose $M_n = Kn\epsilon_n^2 \to \infty$, for some sufficiently large multiple K. What makes this sieve appropriate for our purposes here is that every $\Omega \in \mathscr{T}_n$ has eigenvalues lower-bounded by M_n^{-1}, which is not too small. This eigenvalue control is critical to our demonstration below that the combined likelihood ratio test has suitable bounds on its Type I/II errors in the presence of measurement error.

Recall that the Rényi divergence (of order 1/2), or the negative log-affinity, between two densities f and g is given by $R(f,g) = -\log \int (fg)^{1/2}$. In the present context, the densities are those of $\mathsf{N}_p(0, (\Omega^{-1} + \nu I_p)^{-1})$, to be denoted by f_Ω, indexed by Ω, and $R(f_{\Omega^\star}, f_\Omega)$ can be abbreviated by $R_\nu(\Omega^\star, \Omega)$. By simple calculations,

$$R_\nu(\Omega^\star, \Omega) = -\log\left(\frac{|\Omega^{\star -1} + \nu I_p|^{-1/4} |\Omega^{-1} + \nu I_p|^{-1/4}}{|\frac{1}{2}(\Omega^{\star -1} + \nu I_p)^{-1} + \frac{1}{2}(\Omega^{-1} + \nu I_p)^{-1}|^{1/2}} \right).$$

For $\Omega^\star$ the true precision matrix, fix another $\Omega^\dagger \in \mathscr{T}_n$ so that $R_\nu(\Omega^\star, \Omega^\dagger) \geq \epsilon_n^2$. A most powerful Neyman–Pearson test is then given by $\phi_n = \mathbb{1}(f^n_{\Omega^\dagger} \geq f^n_{\Omega^\star})$, where f^n_Ω denotes the joint density function for n i.i.d. samples from f_Ω. By Markov's inequality, the Type I error probability is bounded by

$$\begin{aligned}
\mathsf{E}_{\Omega^\star, \nu} \phi_n &= \int \mathbb{1}\left[\left\{ \frac{f^n_{\Omega^\dagger}(y^n)}{f^n_{\Omega^\star}(y^n)} \right\}^{1/2} \geq 1 \right] f^n_{\Omega^\star}(y^n)\, dy^n \\
&\leq \left[\int \{ f_{\Omega^\dagger}(y) f_{\Omega^\star}(y) \}^{1/2}\, dy \right]^n \\
&= e^{-nR_\nu(\Omega^\star, \Omega^\dagger)} \\
&\leq e^{-n\epsilon_n^2}.
\end{aligned}$$

By reversing the roles of $\Omega^\star$ and Ω, it follows that the Type I error probability $\mathsf{E}_{\Omega^\dagger, \nu}(1-\phi_n) \leq e^{-n\epsilon_n^2}$ as well. Next, take a generic Ω such that $\|\Omega - \Omega^\dagger\|_2 \leq \delta_n$, where $\delta_n = (2np)^{-1}$. Then

$$\begin{aligned}
\mathsf{E}_{\Omega, \nu}(1-\phi_n) &= \mathsf{E}_{\Omega^\dagger, \nu}\left\{ (1-\phi_n) f^n_\Omega / f^n_{\Omega^\dagger} \right\} \\
&\leq \{ \mathsf{E}_{\Omega^\dagger, \nu}(1-\phi_n) \}^{1/2} \left[\int \{ f_\Omega(y) / f_{\Omega^\dagger}(y) \}^2 f_{\Omega^\dagger}(y)\, dy \right]^{n/2}. \qquad (2.6.7)
\end{aligned}$$

By Lemma 2.7.3, the second factor can be bounded by

$$\int \left\{ \frac{g_\Omega(x)}{g_{\Omega^\dagger}(x)} \right\}^2 g_{\Omega^\dagger}(x)\, dx = \frac{|\Omega|}{|\Omega^\dagger|^{1/2} |2\Omega - \Omega^\dagger|^{1/2}} = \frac{|B|^{1/2}}{|2I_p - B^{-1}|^{1/2}},$$

where g_Ω is the density of $\mathsf{N}_p(0, \Omega^{-1})$ and $B = \Omega^{1/2} \Omega^{\dagger -1} \Omega^{1/2}$. By the choice of the sieve in (2.6.6), we get $\|(\Omega^\dagger)^{-1}\|_2 \leq M_n$ and $\|\Omega - \Omega^\dagger\|_2 \leq \delta_n$, which implies that, on the sieve,

$$\|B - I\|_2 \leq \|(\Omega^\dagger)^{-1}\|_2 \|\Omega - \Omega^\dagger\|_2 \leq M_n \delta_n.$$

By Weyl's inequality, the eigenvalues of B are between $1 - M_n\delta_n$ and $1 + M_n\delta_n$. Applying the inequality

$1-x^{-1} < \log x < x-1$ for any $x > 0$, we find that

$$\frac{|B|^{1/2}}{|2I_p - B^{-1}|^{1/2}} \le \exp[p\{\log(1+M_n\delta_n) - \log(2-1/(1-M_n\delta_n))\}/2] \le e^{pM_n\delta_n}$$

and consequently,

$$\mathsf{E}_{\Omega,\nu}(1-\phi_n) \le \exp(-\tfrac{n}{2}\epsilon_n^2 + \tfrac{1}{2}pM_n\delta_n) = \exp(-n\epsilon_n^2/4),$$

as $\delta_n = \eta(np)^{-1}$ and $M_n = Kn\epsilon_n^2$. This gives the Type II error bound (2.6.7) uniformly over the set $\{\Omega : \|\Omega - \Omega^\dagger\|_2 \le \delta_n\}$.

Step 3: *Entropy bound.* In the above analysis, the $\Omega^\dagger$ separated from $\Omega^\star$ was fixed but arbitrary. So we can repeat that argument for finitely many different $\Omega^\dagger$'s and construct a test for the complement of the Rényi neighborhood of $\Omega^\star$ by taking the maximum of those $\Omega^\dagger$-specific tests. The logarithm of the number of such tests is therefore bounded by the δ_n-entropy of $\mathscr{T}_n$ with respect to spectral norm $\log N(\delta_n, \mathscr{T}_n, \|\cdot\|_2)$. It suffices to show that this is bounded by a constant multiple of $n\epsilon_n^2$. Towards this, if we take $\Omega_1 = \Theta_1 + \kappa_1 I_p$ and $\Omega_2 = \Theta_2 + \kappa_2 I_p$ in $\mathscr{T}_n$, then by the triangle inequality, for the given δ_n,

$$\log N(\delta_n, \mathscr{T}_n, \|\cdot\|_2) \le \log N(\delta_n/2, \mathscr{S}_n, \|\cdot\|_2) + \log(M_n\delta_n^{-1}) \lesssim n\epsilon_n^2 + \log n \lesssim n\epsilon_n^2.$$

Step 4: *Prior probability of the complement of the sieve.* In view of Condition 2.2.1 (a) and (b)(ii) on the prior and the choice $M_n = Kn\epsilon_n^2$, we estimate $\Pi(\mathscr{T}_n^c) \le \Pi_\Theta(\mathscr{S}_n^c) + \Pi_\kappa([M_n^{-1}, M_n]^c) \lesssim e^{-Gn\epsilon_n^2}$, where the constant $G > 0$ can be made as large as we wish by choosing K large enough.

Step 5: *Convert from Rényi to Frobenius.* From the previous steps, and the general result of Theorem 2.1 in Ghosal, Ghosh & van der Vaart (2000), we obtain a contraction rate in terms of Rényi divergence $\mathsf{E}_{\Omega^\star,\nu}\Pi_n^\nu(\{\Omega : R_\nu(\Omega^\star,\Omega) > L'\epsilon_n^2\}) \to 0$ for some $L' > 0$ sufficiently large. Under the assumption that $\|\Omega^\star\|_2$ is bounded, we shall conclude that $\mathsf{E}_{\Omega^\star,\nu}\Pi_n^\nu(\{\Omega : \|\Omega^\star - \Omega\|_F > L\epsilon_n\}) \to 0$ for some $L > 0$. Towards this, define $A = \Omega_\nu^{\star -1/2}\Omega_\nu\Omega_\nu^{\star -1/2}$, where $\Omega_\nu = (\Omega^{-1} + \nu I_p)^{-1}$ and $\Omega_\nu^\star = (\Omega^{\star -1} + \nu I_p)^{-1}$. Let $\alpha_1 \le \cdots \le \alpha_p$ denote the eigenvalues of A in the increasing order. It follows from Lemma A.2(ii) of Banerjee & Ghosal (2015) that, if the Hellinger distance or the Rényi divergence of $f_{\Omega^\star}$ from f_Ω is sufficiently small, then $\max\{|\alpha_j - 1| : j = 1, \ldots, p\} \le 1$ and, therefore, $\alpha_p \le 2$. Since $4\alpha(1+\alpha)^{-2} < 1$ for all $\alpha \in (0, 2]$, and $-\log x \ge 1 - x$ for all $x \in (0, 1)$, we get that the Rényi divergence $R_\nu(\Omega^\star, \Omega)$ can be written as

$$\begin{aligned} -\frac{1}{4}\log\frac{|A|}{|\frac{1}{2}I_p + \frac{1}{2}A|^2} &= -\frac{1}{4}\sum_{j=1}^p \log\frac{4\alpha_j}{(1+\alpha_j)^2} \\ &\ge \frac{1}{4}\sum_{j=1}^p \left\{1 - \frac{4\alpha_j}{(1+\alpha_j)^2}\right\} = \frac{1}{4}\sum_{j=1}^p \left(\frac{1-\alpha_j}{1+\alpha_j}\right)^2. \end{aligned}$$

Since $1 + \alpha_j \le 1 + \alpha_p \le 3$ for all j, and $\|A - I_p\|_F^2 = \sum_{j=1}^p (1-\alpha_j)^2$, we have that $R_\nu(\Omega^\star, \Omega) \gtrsim \|A - I_p\|_F^2$.

Next, observe that

$$\|\Omega_\nu - \Omega_\nu^\star\|_F^2 = \|\Omega_\nu^{\star 1/2}\Omega_\nu^{\star -1/2}(\Omega_\nu - \Omega_\nu^\star)\Omega_\nu^{\star -1/2}\Omega_\nu^{\star 1/2}\|_F^2 \le \|\Omega_\nu^\star\|_2^2 \|A - I\|_F^2.$$

Combining these, we conclude that

$$R_\nu(\Omega^\star, \Omega) \gtrsim \|\Omega_\nu - \Omega_\nu^\star\|_F^2 / \|\Omega_\nu^\star\|_2^2 = (1 + \nu\|\Omega^\star\|_2)^2 \|\Omega_\nu - \Omega_\nu^\star\|_F^2 / \|\Omega^\star\|_2^2,$$

by the fact that $\|\Omega_\nu^\star\|_2 = (\|\Omega^\star\|_2^{-1} + \nu)^{-1} = \|\Omega^\star\|_2(1 + \nu\|\Omega^\star\|_2)^{-1}$. For $R_\nu(\Omega^\star, \Omega)$ small, $\|\Omega_\nu - \Omega_\nu^\star\|_F \lesssim \|\Omega^\star\|_2$, so it follows from Lemma 2.7.4 that

$$R_\nu(\Omega^\star, \Omega) \gtrsim \frac{(1 + \nu\|\Omega^\star\|_2)^2 \|\Omega - \Omega^\star\|_F^2}{\|\Omega^\star\|_2^2 (1 + \nu\|\Omega^\star\|_2)^4} \ge \frac{\|\Omega - \Omega^\star\|_F^2}{\|\Omega^\star\|_2^2 (1 + \nu\|\Omega^\star\|_2)^2}.$$

We assume that $\|\Omega^\star\|_2 \lesssim 1$, so $R_\nu(\Omega^\star, \Omega) \gtrsim \|\Omega - \Omega^\star\|_F^2$ follows immediately.

Finally, note that if $\|\Omega^\star\|_2$ is not bounded, then, from the penultimate display,

$$\|\Omega - \Omega^\star\|_F^2 \lesssim \|\Omega^\star\|_2^4 R_\nu(\Omega^\star, \Omega).$$

Therefore, the result in Theorem 2.2.1 holds with the modified rate $\epsilon_n' = \|\Omega^\star\|_2^2 \epsilon_n$.

2.6.2 Proof of Theorem 2.3.1

As Condition 2.2.1 (a) is directly assumed, we only need to verify the conditions of Theorem 2.2.1 for the sieve

$$\mathscr{S}_n = \{\Theta : \gamma \le R_n, \|\Theta\|_\infty \le M_n\}, \quad R_n = K n \epsilon_n^2 / \log n, \quad M_n = K n \epsilon_n^2,$$

for some sufficiently large constant $K > 0$. Since $\|\Theta\|_2^2 \le \|\Theta\|_F^2 \le R_n \|\Theta\|_\infty^2$, we have

$$\log N(\delta_n, \mathscr{S}_n, \|\cdot\|_2) \le \log N(\delta_n, \mathscr{S}_n, \|\cdot\|_F) \le \log N(\delta_n R_n^{-1/2}, \mathscr{S}_n, \|\cdot\|_\infty),$$

where the last expression is no more than

$$\log \left\{ \sum_{j=1}^{R_n} \binom{p(p-1)/2}{j} \left(\frac{R_n^{1/2} M_n}{\delta_n} \right)^j \right\} \lesssim R_n \log\left(R_n^{3/2} p^2 M_n / \delta_n\right) \lesssim R_n \log n.$$

Since $R_n \asymp n\epsilon_n^2 / \log n$, this verifies Condition 2.2.1 (b)(i).

Next, for $\Theta \in \mathscr{S}_n^c$, either $|\Theta_{i,j}| > M_n$ for some (i, j), or $\gamma > R_n$. The probability of this set is less than $p^2 \Pi(\|\Theta\|_\infty > M_n) + \Pi(R > R_n)$. Since the entries of Θ have exponential or Laplace distribution, both of which have an exponentially small tail probability, the first term is bounded by a multiple of $\exp(-\lambda M_n) \lesssim \exp(-\lambda K n \epsilon_n^2)$. The second term is bounded by $\Pi(R > R_n) \lesssim \exp(-a R_n \log R_n) \lesssim \exp(-a K n \epsilon_n^2)$ by the assumption (2.3.2) and the choice $R_n \asymp n\epsilon_n^2 / \log n$. By taking K sufficiently large, we verify Condition 2.2.1 (b)(ii).

To verify Condition 2.2.1 (b)(iii), observe that

$$\begin{aligned}\Pi(\|\Theta-\Theta^\star\|_F \le c\epsilon_n) &\gtrsim \Pi(\|\Theta-\Theta^\star\|_\infty \le c\epsilon_n/p)\\ &\gtrsim (c\epsilon_n/p)^{p+s^\star}\\ &\gtrsim \exp\{-c'(p+s^\star)\log n\},\end{aligned}$$

for some $c'>0$. By equating $n\epsilon_n^2$ with $(p+s^\star)\log n$, we obtain the advertised rate of $\epsilon_n = n^{-1/2}(p+s^\star)^{1/2}(\log n)^{1/2}$.

2.6.3 Proof of Theorem 2.3.2

We consider the sieve

$$\mathscr{S}_n = \{\Theta = UDU^{\mathrm{T}} : \gamma \le R_n, \|D\|_\infty \le M_n, \|U\|_\infty \le \sqrt{M_n}\},$$

where $R_n = Kn\epsilon_n^2/\log n$ and $M_n = Kn\epsilon_n^2$ for some sufficiently large constant $K>0$, and verify Condition 2.2.1 (b)(i)–(iii).

For any two precision matrices $\Theta_1, \Theta_2 \in \mathscr{S}_n$ with Cholesky decompositions $\Theta_1 = U_1D_1U_1^{\mathrm{T}}$ and $\Theta_2 = U_2D_2U_2^{\mathrm{T}}$, we obtain

$$\begin{aligned}\|\Theta_1-\Theta_2\|_2 &\le \|U_1D_1U_1^{\mathrm{T}} - U_1D_1U_2^{\mathrm{T}}\|_2 + \|U_1D_1U_2^{\mathrm{T}} - U_1D_2U_2^{\mathrm{T}}\|_2 + \|U_1D_2U_2^{\mathrm{T}} - U_2D_2U_2^{\mathrm{T}}\|_F\\ &\le \|U_1\|_2\|D_1\|_2\|U_1-U_2\|_F + \|U_1\|_2\|U_2\|_2\|D_1-D_2\|_F + \|U_2\|_2\|D_2\|_2\|U_1-U_2\|_F\\ &= (\|D_1\|_2\|U_1-U_2\|_F + \|D_1-D_2\|_F + \|D_2\|_2\|U_1-U_2\|_F)\\ &\le M_n p(\|U_1-U_2\|_\infty + \|D_1-D_2\|_\infty),\end{aligned}$$

since $\|U_1\|_2 = \|U_2\|_2 = 1$ and $\|D_1\|_2 = \|D_1\|_\infty \le M_n$. Hence

$$\log N(\delta_n, \mathscr{S}_n, \|\cdot\|_2) \le \log\left\{\sum_{j=1}^{R_n}\binom{p(p-1)/2}{j}\left(\frac{M_n}{\delta_n p M_n}\right)^{j+p}\right\} \lesssim (R_n+p)\log n.$$

Since $R_n \asymp n\epsilon_n^2/\log n$ and $p \lesssim n\epsilon_n^2/\log n$, we have verified Condition 2.2.1 (b)(i).

To verify Condition 2.2.1 (b)(ii), we observe that

$$\Pi(\mathscr{S}_n^c) \lesssim \Pi(\gamma > R_n) + p\Pi(\|D\|_\infty > M_n) + p^2\Pi(\|U\|_\infty > \sqrt{M_n}).$$

Under both (2.3.3) and (2.3.4), γ has a binomial distribution, and therefore the first term on the right hand side is bounded by $\exp(-aR_n\log R_n) \lesssim \exp(-aKn\epsilon_n^2)$. By the tail probabilities of gamma and normal distributions, we have an upper bound for the remaining two terms as $\exp(-\lambda M_n) \lesssim \exp(-\lambda Kn\epsilon_n^2)$. Choosing K sufficiently large ensures the required bound.

To verify Condition 2.2.1 (b)(iii) about prior concentration, we have for some constant $c > 0$,

$$\begin{aligned}\Pi(\|D - D^\star\|_F \le \epsilon_n) &\ge \Pi(\|D - D^\star\|_\infty \le \epsilon_n/\sqrt{p}) \\ &\gtrsim (\epsilon_n/p)^p \\ &\gtrsim \exp\{-c\,p\log(\epsilon_n/p)\} \end{aligned} \tag{2.6.8}$$

since all the diagonal values of $D^\star$ are bounded away from 0 and the prior density around $D^\star$ is lower bounded, and

$$\begin{aligned}\Pi(\|U - U^\star\|_F \le \epsilon_n) &\ge \Pi(\|U - U^\star\|_\infty \le \epsilon_n/p) \\ &= \{\Pi(|U_{i,j}| < \epsilon_n/p)\}^{p(p-1)/2 - s^\star} \\ &\quad \times \{\Pi(|U_{i,j} - U^\star_{i,j}| \le \epsilon_n/p \,||U_{i,j}| > \epsilon_n/p)\Pi(|U_{i,j}| > \epsilon_n/p)\}^{s^\star} \\ &\gtrsim (1-p^{-b})^{p(p-1)/2-s^\star}(\epsilon_n/p)^{s^\star} p^{-a s^\star}\end{aligned}$$

which shares the same lower bound (2.6.8). This follows because $\Pi(\|U - U^\star\|_\infty \le \epsilon_n/p)$ is equal to the probability that $\Pi(|U_{i,j}| < \epsilon_n/p) \gtrsim (1-p^{-b})$ when $|U^\star_{i,j}| = 0$, and that is the case for $p(p-1)/2 - s^\star$ many pairs. Now by the triangle inequality, the facts that $\|U^\star\|_2$ and $\|D^\star\|_2$ are assumed to have a constant upper bound, and the prior independence of U and D, it follows that $-\log\Pi(\|\Theta - \Theta^\star\|_F \le \epsilon) \lesssim (p + s^\star)\log(p/\epsilon_n) \lesssim (p+s^\star)\log n$, so the rate $\epsilon_n = [\{(p+s^\star)\log n\}n^{-1}]^{1/2}$ satisfies the required condition.

2.6.4 Proof of Theorem 2.3.3

By the arguments given at the beginning of the proof of Theorem 2.2.1, we may assume that our choice of $\Theta^\star$ meets the two conditions assumed about $\Omega^\star$, namely, has eigenvalues bounded and bounded away from 0, and a fixed, sufficiently small-size $\|\cdot\|_\infty$-neighborhood of $\Theta^\star$ is contained in $\mathscr{P}_G$.

We consider the sieve

$$\mathscr{S}_n = \{\Theta : \|\Theta\|_\infty \le M_n\},$$

where $M_n = K n\epsilon_n^2$ with K to be chosen a suitably large constant. On the sieve, $\|\Theta\|_2^2 \le \|\Theta\|_F^2 \le 2pk\|\Theta\|_\infty^2$, which leads to the entropy estimate

$$\begin{aligned}\log N(\delta_n, \mathscr{S}_n, \|\cdot\|_2) &\le \log N(\delta_n/\sqrt{2pk}, \mathscr{S}_n, \|\cdot\|_\infty) \\ &\le \log(k\cdot(\sqrt{2pk}M_n/\delta_n)^{pk}) \\ &\lesssim pk\log n. \end{aligned} \tag{2.6.9}$$

Note that $\Theta \in S_n^c$ if $|\Theta_{i,j}| > M_n$ for some pair (i,j). The positive definiteness of Θ implies that the largest entry of Θ in absolute value occurs at a diagonal position. By the property of the G-Wishart distribution, each diagonal entry is distributed as a chi-square distribution with δ degrees

of freedom. From the tail estimate of a chi-square random variable Y, it then follows that

$$\Pi(S_n^c) \leq p\,\mathsf{P}(Y > M_n) \leq \exp(-cM_n + \log p) \leq \exp(-c'n\epsilon_n^2) \tag{2.6.10}$$

for some $c' > 0$ which can be made as large we please by choosing K large enough.

It remains to verify the prior concentration condition with $\epsilon_n = \{n^{-1}(p\log n)\}^{1/2}$. There are at most pk free arguments in Θ due to the k-banding structure. By Roverato (2000), the G-Wishart density at the true value $\Theta^\star$ is bounded below by a constant multiple of the product of a power of $\det(\Theta^\star)$, $e^{-\mathrm{tr}(\Theta^\star)/2}$ and e^{-cp} for some $c > 0$; see Equations (3.2), (3.3), (5.2), and (5.3) of Banerjee & Ghosal (2014). Clearly, $\mathrm{tr}(\Theta^\star) = O(p)$ and $|\log\det(\Theta^\star)| = O(p)$ by the boundedness of the eigenvalues of $\Theta^\star$ and its inverse. Since $\Theta^\star$ stays away from the boundary of $\mathscr{P}_G$ by the assumption, it follows that

$$\Pi(\|\Omega-\Omega^\star\|_F \leq \epsilon_n) \gtrsim \Pi(\|\Omega-\Omega^\star\|_\infty \leq \epsilon_n/\sqrt{kp}) \gtrsim e^{-c''p}(\epsilon_n/\sqrt{kp})^{kp} \tag{2.6.11}$$

for some $c'' > 0$. From (2.6.9)–(2.6.11), the rate $\epsilon_n = \{n^{-1}(p\log n)\}^{1/2}$ follows by an application of Theorem 2.2.1.

2.6.5 Proof of Theorem 2.4.1

We apply Theorem 2.2.1 by verifying Conditions (b)(i)–(iii) on Π_Θ with $\Theta = \Lambda\Lambda^{\mathrm{T}}$. We prove the first assertion only; the second assertion can be established by a minor adjustment of the argument described in the end. Observe that

$$\|\Theta-\Theta^\star\|_F \leq \|\Lambda\Lambda^{\mathrm{T}} - \Lambda\Lambda^{\star\mathrm{T}}\|_F + \|\Lambda\Lambda^{\star\mathrm{T}} - \Lambda^\star\Lambda^{\star\mathrm{T}}\|_F \leq (\|\Lambda\|_2 + \|\Lambda^\star\|_2)\|\Lambda-\Lambda^\star\|_F.$$

We can lower bound $\Pi(\|\Lambda-\Lambda^\star\|_F \leq \epsilon/4\sqrt{c^{\star 3}})$ by

$$\Pi(\|\Lambda-\Lambda^\star\|_\infty \leq \epsilon/4\sqrt{c^{\star 3}pk^\star}\,|k = k^\star)\Pi(k = k^\star),$$

where the second factor is bounded below by $\exp(-K_1 k^\star\log k^\star)$ up to a constant multiple, with some constant K_1, by well-known properties of the Poisson distribution. For the first term, we consider the supremum over all the entries of $0 < i \leq p$ and $0 < j \leq k^\star$ and let $\delta = \epsilon/(4\sqrt{c^{\star 3}pk^\star})$. Then it can be bounded below by

$$\{1-\Pi(|\Lambda_{i,j}| > a^\star)\}^{(p-s^\star)k^\star}\{\Pi(|\Lambda_{i,j}-\Lambda^\star_{i,j}| \leq a^\star \mid |\Lambda_{i,j}| > a^\star)\Pi(|\Lambda_{i,j}| > a^\star)\}^{s^\star k^\star}.$$

The first factor is proportional to $\{1-(pk^\star)^{-1}\}^{(p-s^\star)k^\star} = O(1)$ since $s^\star \lesssim p$. For the second factor, we know that $\Pi(|\Lambda_{i,j}-\Lambda^\star_{i,j}| \leq \delta \mid |\Lambda_{i,j}| > \delta) \gtrsim \delta\exp(-c^\star/2)$ since the probability density is lower bounded on a closed region and $\Pi(|\Lambda_{i,j}| > \delta) \asymp (pk^\star)^{-1}$. Therefore, we conclude that

$$\begin{aligned}\Pi(\|\Lambda-\Lambda^\star\|_\infty \leq a^\star \mid k = k^\star) &\gtrsim (\epsilon/4\sqrt{c^{\star 3}p^3k^{\star 3}})^{s^\star k^\star}\exp(-c^\star s^\star k^\star/2)\\ &\gtrsim \exp(-c^\star s^\star k^\star\log n).\end{aligned}$$

Then Condition (b)(iii) is verified for $\epsilon_n = \{n^{-1}(c^\star s^\star k^\star \log n)\}^{1/2}$.

Next, let

$$\mathscr{S}_n = \{\Lambda\Lambda^{\mathrm{T}} : \gamma \le \gamma_n, k \le k_n, \|\Lambda\|_\infty \le \sqrt{M_n/(2\gamma_n)}\},$$

where γ_n is a large constant multiple of $n\epsilon_n^2/\log n$, $M_n = n\epsilon_n^2\gamma_n$ and $k_n = \gamma_n$. To bound $\Pi(\mathscr{S}_n^c)$ so that the theorem applies to give the rate ϵ_n, we need to establish that for some sufficiently large constant $C > 0$,

$$\begin{aligned} \Pi(k > k_n) &\lesssim e^{-Cn\epsilon_n^2}, \\ \Pi(\gamma > \gamma_n) &\lesssim e^{-Cn\epsilon_n^2}, \\ \max \Pi(|\Lambda_{i,j}| > \sqrt{M_n/(2\gamma_n)}) &\lesssim e^{-Cn\epsilon_n^2}. \end{aligned} \tag{2.6.12}$$

This is so because then conditioning on $k \le k_n$ and $\gamma \le \gamma_n$, to obtain the desired bound $e^{-Cn\epsilon_n^2}$ for $\Pi(\mathscr{S}_n^c)$, the maximum number of (i,j) pairs to be considered is bounded by $\gamma_n^2 k_n^2$, which can be absorbed in the exponent appearing in the bound for $\max\Pi(|\Lambda_{i,j}| > \sqrt{M_n/(2\gamma_n)})$. The first relation follows by the tail estimate

$$\sum_{k=k_n}^{\infty} \frac{\theta_1^k}{k!} \exp(-\theta_1) \le \frac{\theta_1^{k_n}}{k_n!} \sum_{k=0}^{\infty} \frac{\theta_1^k}{k!} \exp(-\theta_1) = \frac{\theta_1^{k_n}}{k_n!} \le \exp(-\tfrac{1}{2} k_n \log k_n),$$

for sufficiently large n. To derive the second relation, it is now sufficed to condition on k with $k \le k_n$. By the tail probability of binomial distribution, $\Pi(\gamma > \gamma_n)$ is bounded by $e^{-C'\gamma_n \log n}$ for some constant $C' > 0$, which gives the desired bound.

For the third inequality in (2.6.12), the tail estimate of a normal distribution gives $e^{-C'M_n/\gamma_n}$, giving the desired bound in view of the choices of M_n and γ_n.

By the relations

$$\|\Theta_1 - \Theta_2\|_2^2 \le \|\Theta_1 - \Theta_2\|_F^2 \le (\|\Lambda_1\|_2^2 + \|\Lambda_2\|_2^2)\|\Lambda_1 - \Lambda_2\|_F^2,$$

the δ_n-metric entropy of $\mathscr{S}_n$ with respect to the spectral norm is bounded by

$$\begin{aligned} \log \left\{ \sum_{k=1}^{k_n} \sum_{\gamma=1}^{\gamma_n} \binom{pk}{\gamma} \left(\frac{\sqrt{2M_n\gamma_n}\sqrt{M_n/(2\gamma_n)}}{\delta_n} \right)^{\gamma} \right\} &\lesssim \log \left\{ k_n \gamma_n (pk_n)^{\gamma_n} (M_n/\delta_n)^{\gamma_n} \right\} \\ &\lesssim \gamma_n \log n, \end{aligned}$$

which is of the order of $n\epsilon_n^2$. This gives the rate $\epsilon_n = \{n^{-1}(\log n)c^\star s^\star k^\star\}^{1/2}$ in terms of the Rényi divergence. By the last assertion of Theorem 2.2.1, since $\|\Omega^\star\|_2 \le c^\star$, the contraction rate in terms of the Frobenius distance is

$$c^\star n^{-1/2} (\log n)^{1/2} (c^\star s^\star k^\star)^{1/2} = n^{-1/2} (\log n)^{1/2} (c^\star)^{3/2} (s^\star k^\star)^{1/2}$$

in the case $\nu = 0$ of no-measurement error (or more generally if $\nu \|\Omega^\star\|_2$ is bounded) and to

$$(c^\star)^2 n^{-1/2} (\log n)^{1/2} (c^\star s^\star k^\star)^{1/2} = n^{-1/2} (\log n)^{1/2} (c^\star)^{5/2} (s^\star k^\star)^{1/2}$$

for fixed measurement error.

When a prior (2.4.2) is put on η, the only change in the calculation comes from the fact that then γ has a beta–binomial distribution instead of binomial. By the tail probability of beta–binomial distribution in Castillo & van der Vaart (2012), $\Pi(\gamma > \gamma_n)$ is bounded by

$$k_n(pk_n - \gamma_n) \frac{\binom{(1+a_4)pk_n - \gamma_n}{a_4 pk_n}}{\binom{(1+a_4)pk_n + 1}{a_4 pk_n}} \lesssim pk_n^2 \left(1 - \frac{\gamma_n + 1}{(1+a_4)pk_n + 1}\right)^{a_4 pk_n} \lesssim pk_n^2 e^{-C'\gamma_n}$$

for some constant $C' > 0$, which can be bounded as desired; here we have used the fact that $\gamma_n / pk_n \to 0$. Since the tail estimate is weaker than $e^{-C'\gamma_n \log n}$ obtained in the case of a fixed η, this implies that the contraction rate in terms of the Rényi divergence weakens to $n^{-1/2}(\log n)\sqrt{c^\star s^\star k^\star}$ and that in terms of the Frobenius distance weakens to $\{n^{-1}(\log n) c^{\star 3} s^\star k^\star\}^{1/2}$ and $\{n^{-1}(\log n) c^{\star 5} s^\star k^\star\}^{1/2}$ respectively, depending on $\nu = 0$ and ν is positive and fixed.

2.7 Auxiliary Lemmas

Recall that $K(f_1, f_2) = \int f_1 \log(f_1/f_2)$ and $V(f_1, f_2) = \int f_1 \log^2(f_1/f_2)$ denote the Kullback–Leibler (KL) divergence and corresponding KL variation for any two densities f_1, f_2. We have the following Lemma as an extension of Lemma B.11 in Ghosal & van der Vaart (2017) to the KL variation.

Lemma 2.7.1. *Let f and g be two probability densities on a common sample space and let $\tilde{f}$ and $\tilde{g}$ be the corresponding densities of the measurable function $T(X, U)$, where $X \sim f$ and $X \sim g$, respectively, where $U \sim \mathsf{Unif}(0,1)$ independent of X. Then the KL variation is $V(\tilde{f}, \tilde{g}) \le V(f, g) + 4K(f, g)$.*

Proof. Define the positive and negative parts of the KL variation as $V^{\pm}(f, g) = \int f(\log_{\pm}(f/g))^2$, respectively and therefore $V(f, g) = V^+(f, g) + V^-(f, g)$. By Lemma B.13 and Lemma B.11 in Ghosal & van der Vaart (2017), $V^-(\tilde{f}, \tilde{g}) \le 4K(\tilde{f}, \tilde{g}) \le 4K(f, g)$.

For the positive part, note that $h(u, v) = u\{\log(u/v)\}^2$ is convex in (u, v) when $\log(u/v) \ge 0$, $u > 0$ and $v > 0$. This can be verified by checking that the Hessian matrix is positive semi-definite, where

$$\begin{aligned}
(\partial^2 h / \partial u^2) &= 2u^{-1}\{1 + \log(u/v)\} \\
(\partial^2 h / \partial u \, \partial v) &= -2v^{-1}\{1 + \log(u/v)\} \\
(\partial^2 h / \partial v^2) &= 2uv^{-2}\{1 + \log(u/v)\}.
\end{aligned}$$

Then, following the proof of Lemma B.11 in Ghosal & van der Vaart (2017), we have $V^+(\tilde{f}, \tilde{g}) \le V^+(f, g) \le V(f, g)$ and we obtain the desired result. □

Let g_Ω denote the densities of $\mathsf{N}_p(0,\Omega^{-1})$ as usual and we have the following Lemma, which is inspired by the proof of Theorem 3.1 in Banerjee & Ghosal (2015) and the proof of Theorem 2 in Du & Ghosal (2018).

Lemma 2.7.2. *Suppose the eigenvalues of $\Omega^\star$ lie in $[1/M, M]$ for some $M > 0$. If $\|\Omega - \Omega^\star\|_F \leq \epsilon/M$ for some sufficiently small $\epsilon > 0$, then $K(g_{\Omega^\star}, g_\Omega) \lesssim \epsilon^2$ and $V(g_{\Omega^\star}, g_\Omega) \lesssim \epsilon^2$.*

Proof. By the definition of the KL divergence,

$$K(g_{\Omega^\star}, g_\Omega) = \tfrac{1}{2}\log|\Omega^\star\Omega^{-1}| + \tfrac{1}{2}\mathsf{E}_{\Omega^\star}(X^{\mathrm{T}}(\Omega-\Omega^\star)X) = \tfrac{1}{2}\log|\Omega^\star\Omega^{-1}| + \tfrac{1}{2}\mathrm{tr}(\Omega\Omega^{\star-1} - I_p),$$

since $\mathsf{E}_{\Omega^\star}(X^{\mathrm{T}}AX) = \mathrm{tr}(A\Omega^{\star-1})$ if $X \sim \mathsf{N}_p(0,\Omega^{\star-1})$ for any $p \times p$ dimensional symmetric matrix A. Furthermore, twice the right-hand side equivalent to

$$-\log|\Omega^{\star-1/2}\Omega\Omega^{\star-1/2}| + \mathrm{tr}(\Omega^{\star-1/2}\Omega\Omega^{\star-1/2} - I_p) = \sum_{i=1}^{p}(1-\lambda_i - \log\lambda_i), \tag{2.7.1}$$

where $\lambda_1, \ldots, \lambda_p$ denotes the eigenvalues of $\Omega^{\star-1/2}\Omega\Omega^{\star-1/2}$. Note that $|1-\lambda_i-\log\lambda_i| \lesssim (1-\lambda_i^2)$ as $|1-\lambda_i| \leq \epsilon$ for all $i = 1, \ldots, p$. Therefore, the expression in (2.7.1) is bounded by a constant multiple of

$$\sum_{i=1}^{p}(1-\lambda_i)^2 = \|I - \Omega^{\star-1/2}\Omega\Omega^{\star-1/2}\|_F^2 \leq \epsilon^2,$$

since $\|I - \Omega^{\star-1/2}\Omega\Omega^{\star-1/2}\|_F \leq \|\Omega^{\star-1}\|_2\|\Omega-\Omega^\star\|_F \leq \epsilon$. This establishes the first assertion.

Similarly, for the corresponding KL variation,

$$\begin{aligned} V(g_{\Omega^\star}, g_\Omega) &= \tfrac{1}{4}\log^2|\Omega^\star\Omega^{-1}| + \tfrac{1}{2}\log|\Omega^\star\Omega^{-1}|\mathsf{E}_{\Omega^\star}(X^{\mathrm{T}}(\Omega-\Omega^\star)X) + \tfrac{1}{4}\mathsf{E}_{\Omega^\star}(X^{\mathrm{T}}(\Omega-\Omega^\star)X)^2 \\ &= \tfrac{1}{4}\mathsf{V}_{\Omega^\star}(X^{\mathrm{T}}(\Omega-\Omega^\star)X) + K^2(g_{\Omega^\star}, g_\Omega), \end{aligned}$$

by adding and subtracting $\mathsf{E}^2_{\Omega^\star}(X^{\mathrm{T}}(\Omega-\Omega^\star)X)$. The latter term has been already bounded by a constant multiple of ϵ^4. The first term is equal to a constant times

$$\mathrm{tr}((\Omega-\Omega^\star)\Omega^{\star-1}(\Omega-\Omega^\star)\Omega^{\star-1}) = \mathrm{tr}(I_p - \Omega^{\star-1/2}\Omega\Omega^{\star-1/2})^2 = \sum_{i=1}^{p}(1-\lambda_i)^2 \lesssim \epsilon^2,$$

which proves the second assertion. □

Lemma 2.7.3. *Let $y \mapsto f_\theta(y)$ be a probability density function with respect to Lebesgue measure on a space $\mathbb{Y}$ for each $\theta \in \Theta$. Similarly, let g_1 and g_2 be probability densities with respect to Lebesgue measure on a space Θ. Then,*

$$\int_{\mathbb{Y}} \frac{\{\int_\Theta f_\theta(y)g_2(\theta)\,d\theta\}^2}{\int_\Theta f_\theta(y)g_1(\theta)\,d\theta}\,dy \leq \int_\Theta \frac{g_2^2(\theta)}{g_1(\theta)}\,d\theta.$$

Proof. First, note that

$$\int_\Theta f_\theta(y)g_2(\theta)\,d\theta = \int_\Theta f_\theta^{1/2}(y)g_1^{1/2}(\theta)\cdot\frac{f_\theta^{1/2}(y)g_2(\theta)}{g_1^{1/2}(\theta)}\,d\theta.$$

By the Cauchy–Schwarz inequality, the square of the above integral is upper bounded by

$$\int_\Theta f_\theta(y)g_1(\theta)\,d\theta\cdot\int_\Theta\frac{f_\theta(y)g_2^2(\theta)}{g_1(\theta)}\,d\theta.$$

Divide by $\int_\Theta f_\theta(y)g_1(\theta)\,d\theta$ and integrate with respect to y to get

$$\int_{\mathbb{Y}}\int_\Theta\frac{f_\theta(y)g_2^2(\theta)}{g_1(\theta)}\,d\theta\,dy.$$

Now change the order of integration and use the fact that f_θ is a probability density for each θ to complete the proof. □

Lemma 2.7.4. *Given Ω and $\Omega^\star$, set $\Omega_\nu=(\Omega^{-1}+\nu I_p)^{-1}$ and $\Omega_\nu^\star=(\Omega^{\star-1}+\nu I_p)^{-1}$. Then*

$$\|\Omega_\nu-\Omega_\nu^\star\|_F\le\|\Omega-\Omega^\star\|_F$$

and, if there exist a constant K such that $\|\Omega_\nu-\Omega_\nu^\star\|_F\le K\|\Omega^\star\|_2$, then

$$\|\Omega_\nu-\Omega_\nu^\star\|_F\gtrsim(1+\nu\|\Omega^\star\|_2)^{-2}\|\Omega-\Omega^\star\|_F. \tag{2.7.2}$$

Proof. Since $\Omega(I_p+\nu\Omega)^{-1}=(I_p-(I_p+\nu\Omega)^{-1})/\nu$ for any Ω, for the first inequality, write

$$\begin{aligned}\|\Omega_\nu-\Omega_\nu^\star\|_F&=\|\Omega(I_p+\nu\Omega)^{-1}-\Omega^\star(I_p+\nu\Omega^\star)^{-1}\|_F\\&=\|\nu^{-1}(I_p+\nu\Omega)^{-1}-\nu^{-1}(I_p+\nu\Omega^\star)^{-1}\|_F\\&=\|(I_p+\nu\Omega)^{-1}(\Omega-\Omega^\star)(I_p+\nu\Omega^\star)^{-1}\|_F\\&\le\|\Omega-\Omega^\star\|_F,\end{aligned}$$

where the latter inequality holds because the eigenvalues of $(I_p+\nu\Omega)^{-1}$ and $(I_p+\nu\Omega^\star)^{-1}$ are all upper-bounded by 1. By the same reason, for the inequality in (2.7.2), note that

$$\begin{aligned}\|\Omega-\Omega^\star\|_F&=\|(I_p+\nu\Omega)(I_p+\nu\Omega)^{-1}(\Omega-\Omega^\star)(I_p+\nu\Omega^\star)^{-1}(I_p+\nu\Omega^\star)\|_F\\&\le\|I_p+\nu\Omega\|_2\|I_p+\nu\Omega^\star\|_2\|\Omega_\nu-\Omega_\nu^\star\|_F,\end{aligned}$$

which implies

$$\|\Omega_\nu-\Omega_\nu^\star\|_F\ge\frac{\|\Omega-\Omega^\star\|_F}{\|I_p+\nu\Omega\|_2\|I_p+\nu\Omega^\star\|_2}. \tag{2.7.3}$$

Next we seek an upper bound on $\|I_p+\nu\Omega\|_2$. By the triangle inequality and the relation $\|\cdot\|_2\le\|\cdot\|_F$,

if we set $\epsilon = \|\Omega_\nu - \Omega^\star_\nu\|_F$, then we get $\|\Omega_\nu\|_2 \leq \|\Omega_\nu - \Omega^\star_\nu\|_2 + \|\Omega^\star_\nu\|_2 \leq \epsilon + \|\Omega^\star_\nu\|_2$.

With some relatively straightforward algebra, it follows that if $\|\Omega_\nu\|_2 \leq \epsilon + \|\Omega^\star_\nu\|_2$, then $\|\Omega\|_2 \lesssim \epsilon + \|\Omega^\star\|_2$. By assumption, we have $\epsilon + \|\Omega^\star\|_2 \lesssim \|\Omega^\star\|_2$, which immediately leads to

$$\|I_p + \nu\Omega\|_2 = 1 + \nu\|\Omega\|_2 \lesssim 1 + \nu\|\Omega^\star\|_2.$$

Finally, the bound in (2.7.2) follows from the above display and (2.7.3). □

CHAPTER

3

INFERENCE ON A PRECISION MATRIX UNDER ADDITIVE MEASUREMENT ERROR

3.1 Introduction

Although the Gaussian distribution is commonly used to model the errors-in-variables problems caused by inaccurate device or indirect measurement, there are still some potential applications that the measurement error is better modeled by some other distributions. Some examples include the Student t-distribution for its robustness to the outliers and the uniform distribution for truncated measurement error. One modern example of such application is to prevent the leakage of personal information. Tons of data are created, stored and analyzed nowadays, containing personal information exposed to the risk of hacking or leaking. Therefore, it is helpful to propose some methods to encrypt the data randomly but maintain the worthwhile information for further analysis, which can also maintain the privacy of information. In such case, this artificial measurement error may not necessarily follow a Gaussian distribution and therefore, we are open to explore any type of distributions as the measurement error and evaluate the estimation under this artificial corruption. In this chapter, we particularly consider the general type of additive measurement error and suppose that the "corrupted" observations Y have the distribution

$$Y_i = X_i + \sqrt{\nu} Z_i, \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p\left(0, \Omega^{-1}\right), \quad Z_i \overset{\text{iid}}{\sim} \mathsf{H}_p, \quad i = 1, \ldots, n, \tag{3.1.1}$$

where H_p denotes the distribution of the p-dimensional measurement error. Assume that each entry of Z_i is independent with the others and therefore, the density of H_p can be written as the product of p identical density functions of each entry $Z_{i,j}$. For simplicity, let H_p denote the distribution for the p-dimensional measurement error Z and assume that it has mean 0 and constant variance. Since this model is more complicated, we assume that ν is known throughout in this and the following chapters. We also assume that the X and Z samples are mutually independent. Under all the assumptions, the marginal distribution of the Y is computed by convolution, i.e.,

$$f_{\Omega,\nu}(y) = \int \nu^{-p/2} h_p\big((y-x)/\sqrt{\nu}\big) g_\Omega(x)\, dx, \tag{3.1.2}$$

where h_p is the density of H_p. Particularly, if H_p is Gaussian, it is available in the closed-form and we are back to the discussion in Chapter 2. However, this convolution cannot be explicitly computed in general, which would cause some difficulty as we will demonstrate later. Moreover, an alternative sampling technique is required as the Gibbs sampling introduced in Section 2.5.1 does not hold without the conjugacy of the measurement error distribution and the Gaussian distribution.

This chapter is organized as follows. In Section 3.2, we describe the proposed method to adjust for the additive measurement error based on the existing Bayesian approaches for inference on the structured, high-dimensional precision matrices, which includes the prior specification, two theoretical results on posterior contraction rates and the computation technique. Due to the lack of an explicit form of the marginal distribution in (3.1.2), we need a stronger condition to transfer the rate from Rényi divergence to the Frobenius norm and maintain the same rate as in Theorem 2.2.1. The condition is to control the spread of the measurement error, by either imposing a restriction on the tail probability or requiring the magnitude ν to decay to 0 fast enough. Two simulation studies are conducted in Section 3.3 to investigate the finite-sample performance of the proposed model under various magnitudes of measurement error, which follows the uniform distribution or the Student t-distribution with degree of freedom 5. Finally, the proofs of all the theorems are provided in Section 3.4.

3.2 Main Results

3.2.1 Prior and Posterior Distributions

While there are some new challenges we will face in this more general context, some of the ideas developed in Chapter 2 carry over directly. For example, to avoid ill-conditioned precision matrices with small eigenvalues, like in Section 2.2.1, we express the precision matrix as

$$\Omega = \Theta + \kappa I_p, \tag{3.2.1}$$

where Θ is a positive semi-definite matrix and $\kappa > 0$ serves as a lower bound on the smallest eigenvalue of Ω. By assigning priors to Θ and κ independently, we automatically induce a prior on

Ω via the mapping (3.2.1). With such a prior specification, a control of the smallest eigenvalue of the precision matrix Ω by exponentially small prior probability can be obtained in the theoretical results of posterior contraction rate. To satisfy the conditions in the posterior contraction rate theory (Ghosal, Ghosh & van der Vaart, 2000) as shown in Section 1.2.3, the prior specification is required to satisfy Condition 2.2.1 on both κ and Θ. Furthermore, to accommodate the indirect formulation in (3.1.2), a control on the largest entry of Ω in the prior is required besides these conditions on the priors inherited from the Gaussian measurement error. With the additive formulation of the precision matrix as shown in (3.2.1), this bound can be directly imposed on Θ if Condition 2.2.1 (a) is satisfied.

Condition 3.2.1. *Given the same ϵ_n as shown in Condition 2.2.1 (b), there exists a sequence $M_n \geq n\epsilon_n^2$ such that the prior probability $\Pi_\Theta(||\Theta||_\infty \geq M_n/2) \leq \exp(-D\, n\epsilon_n^2)$ for some constant $D > 0$.*

Note that the following inequality holds for the M_n in Condition 3.2.1 through (3.2.1), i.e.,

$$\Pi_\Omega(||\Omega||_\infty \leq M_n) \geq \Pi_\Theta(||\Theta||_\infty \leq M_n/2)\Pi_\kappa(\kappa \leq M_n/2),$$

since the priors of Ω and κ are independent. Therefore, by considering the complement, we have the control on the prior probability of $||\Omega||_\infty$ as

$$\begin{aligned}\Pi_\Omega(||\Omega||_\infty \geq M_n) &\leq \Pi_\Theta(||\Theta||_\infty \geq M_n/2)\Pi_\kappa(\kappa \leq M_n/2) + \Pi_\kappa(\kappa \geq M_n/2) \\ &\leq \Pi_\Theta(||\Theta||_\infty \geq M_n/2) + \Pi_\kappa(\kappa \geq M_n/2),\end{aligned}$$

which has an upper bound in the order of $\exp\left(-\min(D, 1/2)\, n\epsilon_n^2\right)$, provided that Condition 2.2.1 (a) and Condition 3.2.1 are satisfied. For κ, a common example is the inverse Gaussian distribution and for Θ, the priors introduced in Section 2.3 and 2.4 could be considered. Specifically, the priors directly imposed on Ω such as the ones discussed in Section 2.3.1 and Section 2.3.3 maintain the bound in Condition 3.2.1 with $M_n = K n\epsilon_n^2$ for some constant K. The proof of it is identical to the step for verifying Condition 2.2.1 (b)(ii) in Section 2.6.2 and Section 2.6.4 because the sieve $\mathscr{S}_n$ for such prior mostly has a control of $||\Theta||_\infty$. Considering priors on the structured decompositions, for instance, the ones introduced in Section 2.3.2 and Section 2.4, we can verify the Condition 3.2.1 with a larger M_n due to the more components involved to compute $||\Theta||_\infty$. According to the sieve S_n in Section 2.6.3, a crude choice of M_n for the prior as shown in Section 2.3.2 to satisfy the Condition 3.2.1 is in the order of $p\, n^2\epsilon_n^4$. For the prior specified on the factor-model structure introduced in Section 2.4, the Condition 3.2.1 can be verified if M_n is equal to $p\, n\epsilon_n^2$ up to some constant by checking the proof for Condition 2.2.1 (b)(ii) in Section 2.6.5.

After the prior for Ω is determined, the posterior distribution is obtained using Bayes' theorem. The likelihood function is defined by the measurement error model (3.1.1) as

$$L_n(\Omega; \nu) \propto \prod_{i=1}^{n} \int \nu^{-p/2} h_p\left((Y_i - X_i)/\sqrt{\nu}\right) \sqrt{\det(\Omega)} \exp\left(-X_i^{\mathrm{T}} \Omega X_i/2\right) dX_i.$$

Note that X is integrated out in the likelihood, which is supposed to be unobservable and irrelevant for estimating Ω in the model. Depending on the corrupted data $Y_1, \ldots, Y_n$ and the known measurement error variance ν, it gives the corresponding posterior distribution as

$$\Pi_n^{\nu}(d\Omega) = \Pi(d\Omega \mid Y_1, \ldots, Y_n; \nu) \propto L_n(\Omega; \nu)\, \Pi(d\Omega). \tag{3.2.2}$$

3.2.2 Theoretical Results

Given the prior setup and the associated posterior, two theorems of the contraction rate are stated as follows with some extra assumptions on the measurement error model. To maintain the same rate in the no-measurement-error context, some assumptions are imposed to control the spread of the measurement error, due to the lack of Gaussian conjugacy as in Theorem 2.2.1. This can be achieved by either restriction the tail probability of measurement error distribution or requiring the magnitude ν decaying to 0 in a desired rate. We first consider the sub-Gaussian distributions on the measurement error.

Theorem 3.2.1. *Assume that $\Omega^\star$ has eigenvalues bounded away from 0. Consider a prior distribution for $\Omega = \Theta + \kappa I_p$ induced from independent prior distributions for κ and Θ. Assume that there exist sequences ϵ_n and M_n with $\epsilon_n \to 0$, $n\epsilon_n^2 \gtrsim \log n$ and $M_n \to \infty$ such that Condition 2.2.1 and Condition 3.2.1 are satisfied. Under the model in* (3.1.1)*, if $\nu p^2 M_n \lesssim \epsilon_n^2$ and the measurement error model H_p is sub-Gaussian with mean equal to 0 and variance equal to 1 that*

$$\Pi(|Z_{1,1}| > z_1) \le c_1 \exp\left(-c_2 z_1^2\right),$$

the posterior distribution Π_n^{ν} in (3.2.2) *contracts at the rate ϵ_n, that is, there exists a constant $L > 0$, depending on $\|\Omega^\star\|_2$, such that*

$$\mathsf{E}_{\Omega^\star,\nu} \Pi_n^{\nu}(\{\Omega : \|\Omega - \Omega^\star\|_F > L\epsilon_n\}) \to 0 \quad \text{as } n \to \infty.$$

The proof of the theorem is given in Section 3.4.1. Even though h is assumed to have variance equal to 1 for the sake of simplicity, the proof easily generalizes for any constant variance, by a slight modification.

Theorem 3.2.1 indicates that if the Gaussianity assumption of the measurement error model as in Theorem 2.2.1 cannot hold, it costs more to guarantee the posterior contraction rate. The whole parameter space in Theorem 3.2.1 is restricted to maintain a control over the eigenvalues, which is sometimes a necessary condition and is also assumed in Banerjee & Ghosal (2015) and Du & Ghosal (2018). Although the sub-Gaussianity assumption imposes a considerable restriction, it does include a broader class of distributions such as uniform distribution and all the truncated distributions. The scale ν of the measurement error is assumed to converge to 0, which indicates that the more precise observations are desired to maintain the contraction rate. On the other hand, considering each observation Y as an estimator of the unobservable outcome X and ν as the reciprocal of the sample size, this condition tells us that we need more conditions to get a precise estimate.

However, without the sub-Gaussianity assumption, we need a stronger condition on ν compared with the result in Theorem 3.2.1, as shown in the following theorem.

Theorem 3.2.2. *Assume that $\Omega^\star$ has eigenvalues bounded away from 0. Consider a prior distribution for $\Omega = \Theta + \kappa I_p$ induced from independent prior distributions for κ and Θ. Assume that there exist sequences ϵ_n and M_n with $\epsilon_n \to 0$, $n\epsilon_n^2 \gtrsim \log n$ and $M_n \to \infty$ such that Condition 2.2.1 and Condition 3.2.1 are satisfied. Under the model in (3.1.1), if $\nu p M_n \lesssim \epsilon_n^4$ and the measurement error model H_p has mean 0 and variance 1, the posterior distribution Π_n^ν in (3.2.2) contracts at the rate ϵ_n, that is, there exists a constant $L > 0$, depending on $\|\Omega^\star\|_2$, such that*

$$\mathsf{E}_{\Omega^\star,\nu}\Pi_n^\nu(\{\Omega : \|\Omega - \Omega^\star\|_F > L\epsilon_n\}) \to 0 \quad \text{as } n \to \infty.$$

It is remarkable to notice that the conditions on the measurement error model H_p are relatively weak and any distribution with finite variance may be assumed on it, while we need a stronger accuracy requirement on the "corrupted" data. Although the condition on ν seems to be restrictive, it may not be very unnatural. Suppose that $p \le \sqrt{n}$ and the number of non-zero entries in $\Omega^\star$ has the order of p, that is, $s \lesssim p$, for example the banded structure or some other sparse structures. Then, from the examples in Section 2.3 and Section 2.4, the contraction rate ϵ_n is supposed to be of the order of $\sqrt{p/n}$ if the logarithmic term is ignored. Then, the estimation error per entry of the precision matrix is of the order ϵ_n^2/p^2 and that of the estimation error of the true observation from the "corrupted" observation is of the order ν, which is almost ϵ_n^4/p. Because

$$\epsilon_n^2/p^2 = 1/pn \ge p/n^2 = \epsilon_n^4/p,$$

the condition really means that the measurement error needs to be smaller than the estimation error for the contraction rate from the no-measurement error case to be maintained. Although the Gaussian case did not need any restriction on the scale of the measurement error, the condition seems to be natural in general. The proof of the theorem is given in Section 3.4.2.

3.2.3 Computation

Like most of the modern Bayesian methods, the formulation of the posterior (3.2.2) is indirect and we cannot obtain the posterior distribution in closed-form. Therefore, MCMC is necessary to produce samples from Π_n^ν. Similar to what we did for the Gaussian measurement error as in Section 3.2.3, we can modify the corresponding algorithms to adjust for this more general type of measurement error.

To utilize the existing algorithms and tools for sampling from the posterior distribution of Ω in the no-measurement-error context, the idea of data augmentation is again applied. Since the measurement error model is not necessarily Gaussian and the benefit of the conjugacy is lost, the extra step to adjust for the measurement error is now replaced by the Metropolis-Hastings algorithm. The full MCMC sampling is listed as: for the $(k+1)$-th iteration,

Step 1: for each i, sample $X_i^{(k+1)}$ from the full conditional posterior

$$(X_i^{(k+1)} \mid Y_i, \Omega^{(k)}) \overset{\text{ind}}{\sim} \nu^{-p/2} h_p\left((Y_i - X_i)/\sqrt{\nu}\right) g_{\Omega^{(k)}}(X_i),$$

using the Metropolis-Hastings algorithm with some proposal distribution described below.

Step 2: Sample $\Omega^{(k+1)}$ from its posterior

$$(\Omega \mid X_1^{(k+1)}, \ldots, X_n^{(k+1)}) \sim \Pi(\Omega \mid X_1^{(k+1)}, \ldots, X_n^{(k+1)}),$$

where $\Pi(\Omega \mid X_1^{(k+1)}, \ldots, X_n^{(k+1)})$ is the posterior based on the no-measurement-error model, with the augmented dataset $X_1^{(k+1)}, \ldots, X_n^{(k+1)}$ for the k-th iteration.

In Step 1, the random walk Metropolis algorithm is recommended with the Gaussian proposal distribution, while the choice of its variance has to be careful. If the variance of the proposal distribution is much larger than ν, the acceptance ratio of the Metropolis algorithm will become extremely small, which means that the X's are mostly stuck at the initial values (which are usually Y's) and would not be updated. Therefore, a small variance compared to ν is recommended for the Gaussian proposal distribution, for instance, $\mathsf{N}_p(\cdot, \nu I_p/10)$ or $\mathsf{N}_p(\cdot, \nu I_p/100)$.

It is worthwhile to notice that the second step is irrelevant to Y given all the X, which we could treat as the posterior samples of the uncontaminated outcomes. From there, we can implement the method to sample from the no-measurement-error posterior in the literature by inserting that algorithm into the second step and sampling iteratively to obtain the samples from the posterior of precision matrices, in which the known measurement error is adjusted.

However, the current prior of Ω induced by the independent prior distribution on Θ and κI_p is not convenient for the MCMC sampling, which is also an obstacle for applying the no-measurement-error method directly. Note that Θ is supposed to share the same structure as Ω and κ is just a technical device to ensure a lower bound for the eigenvalues of Ω. Therefore, dropping the κ part in computation is allowable if such a bound is guaranteed or loss of such a bound does not cause any instability when calculating the inverses. The numerical results presented below employ this simplification.

3.3 Simulation Studies

We consider the uniform distribution and the Student t-distribution for the measurement error model to cover the scenarios of truncated distribution and the fat-tail distribution, respectively. At the end of this section, we use the estimation from the second case to explore the numerical accuracy of variable selection for the precision matrix.

3.3.1 Measurement Error Following the Uniform Distribution

To explore the finite-sample performance of our adjustment for the sub-Gaussian measurement error, let the measurement error H_p be truncated and have a uniform distribution in $[-1,1]$. As in the simulation study for the Gaussian measurement error in Section 2.5.2, four structures of true precision matrix (AR(1), AR(2), Block(2) and Block(5)) are considered sequentially with the dimension $p=50$ and sample size $n=100$. Since the performance of the proposed method is not significantly different from that of the baseline method when ν is extremely small as shown in Section 2.5.2, we focus on the larger ν in this simulation study, where the ν varies from $10^{-0.5}$ to 10 over 7 different values, which are equally spaced on the log-scale. For each sparse structure and each choice of ν, 100 replicates are run.

For each replicate, the data are generated by following these steps:

- Obtain the true precision matrix with one of the aforementioned structures and compute the covariance matrix $\Sigma^\star = \Omega^{\star-1}$;
- Generate $X_1,\ldots,X_n$ from $\mathsf{N}_p(0,\Sigma^\star)$ and $Z_1,\ldots,Z_n$ from $\mathsf{Unif}[-1,1]$ independently;
- For each X_i and Z_i, generate the observation by $Y_i = X_i + \sqrt{\nu}Z_i$.

We implement the MCMC sampling scheme introduced in Section 3.2.3 with the two proposal distributions $\mathsf{N}_p(\cdot, \nu I_p/10)$ and $\mathsf{N}_p(\cdot, \nu I_p/100)$ in Step 1, respectively and consider the prior setup (2.3.3) on the Cholesky decomposition structure in Section 2.3.2 in the Step 2 of the the MCMC sampling to get the posterior samples. We consider the same hyper-parameter specification for α_1, β_1, C_p, σ_0^2 and σ_1^2 as that in Section 2.5.2 only with the informative choice of priors, since it produces a better estimation compared to the diffuse priors suggested by the Gaussian case. We assign the initial values of X_i's as Y_i's for $i=1,\ldots,n$, respectively, such that the initial values would not make a difference to the performance of the methods.

We also compare the proposed method with the model when the measurement error is totally ignored in the estimation and mark the methods in the graphs as "adjust" and "ignore". The estimator is the posterior mean and the Frobenius norm estimation errors are given in Figures 3.1, 3.2, 3.3, and 3.4 for the four structures, respectively, where we only show the central 90% of the estimation errors to remove some outliers. Note that the measurement error ν is displayed in its logarithm with basis 10 to flatten the curve.

From the top panel of the Figure 3.1, the estimation gains some improvement over all the magnitudes of measurement error we considered. That difference becomes more conspicuous when ν is smaller but fade gradually as ν increases. Eventually, when $\nu=10$, the improvement by adjusting for the measurement error is almost negligible compared with the result of ignoring the measurement error. We believe that the reason for this fading difference is the smaller acceptance ratio for a larger ν, which will cause fewer updates, when the proposal distribution is $\mathsf{N}_p(\cdot, \nu I_p/10)$. This also demonstrates the reason why the variance of the estimation error from the proposed method decreases as ν increases. When the proposal distribution is switched to $\mathsf{N}_p(\cdot, \nu I_p/100)$ as

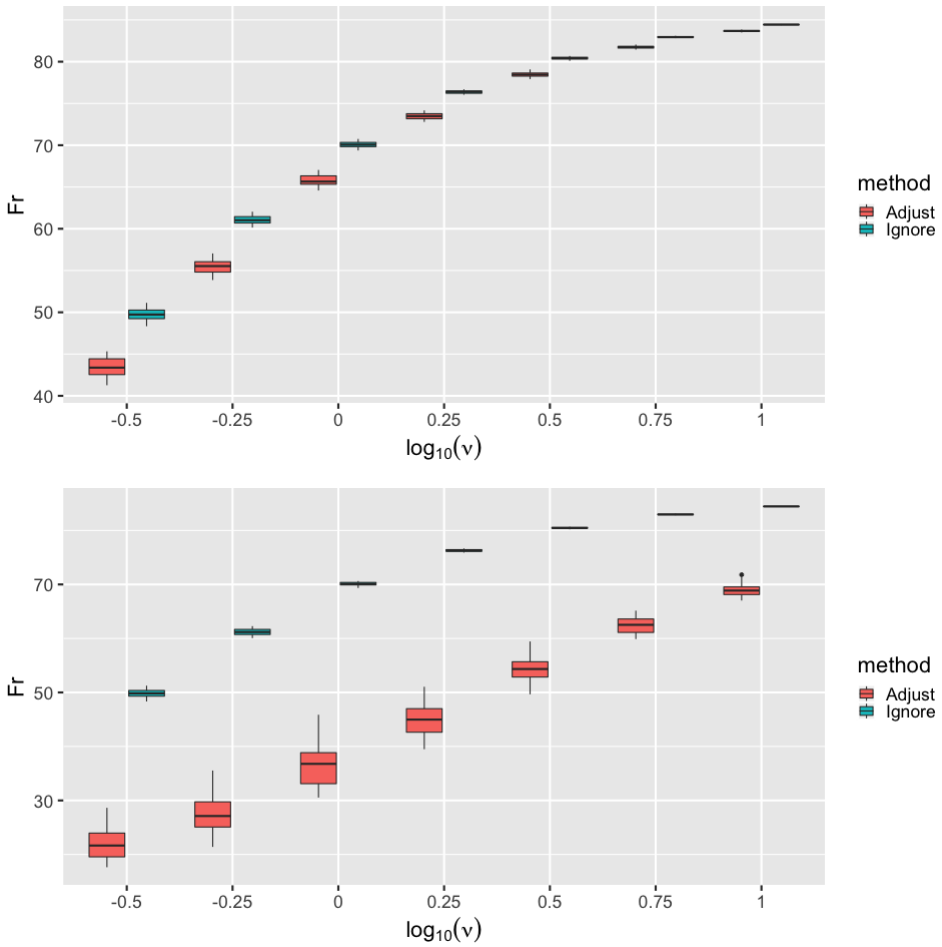

Figure 3.1 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(1) model over different magnitude of measurement error following the uniform distribution.

shown in the bottom panel of the Figure 3.1, the acceptance ratio increases for every ν and the estimation becomes better in terms of Frobenius norm uniformly. On the other hand, the estimation error from the method adjusting for the measurement error in the top panel obtains a smaller variance compared to that in the bottom panel uniformly because of the fewer updates by the Metropolis algorithm with a lower acceptance ratio. The same observation can be found in the rest three more complex structures in Figure 3.2, 3.3 and 3.4, which indicates that the estimation error stabilizes over different structures of true precision matrix when the measurement error follows the uniform distribution.

3.3.2 Measurement Error Following the Student t-distribution

Considering the uniform distribution may not be a suitable candidate for encrypting data or modeling the data with much outliers, the fat-tail distribution may be another direction to explore its

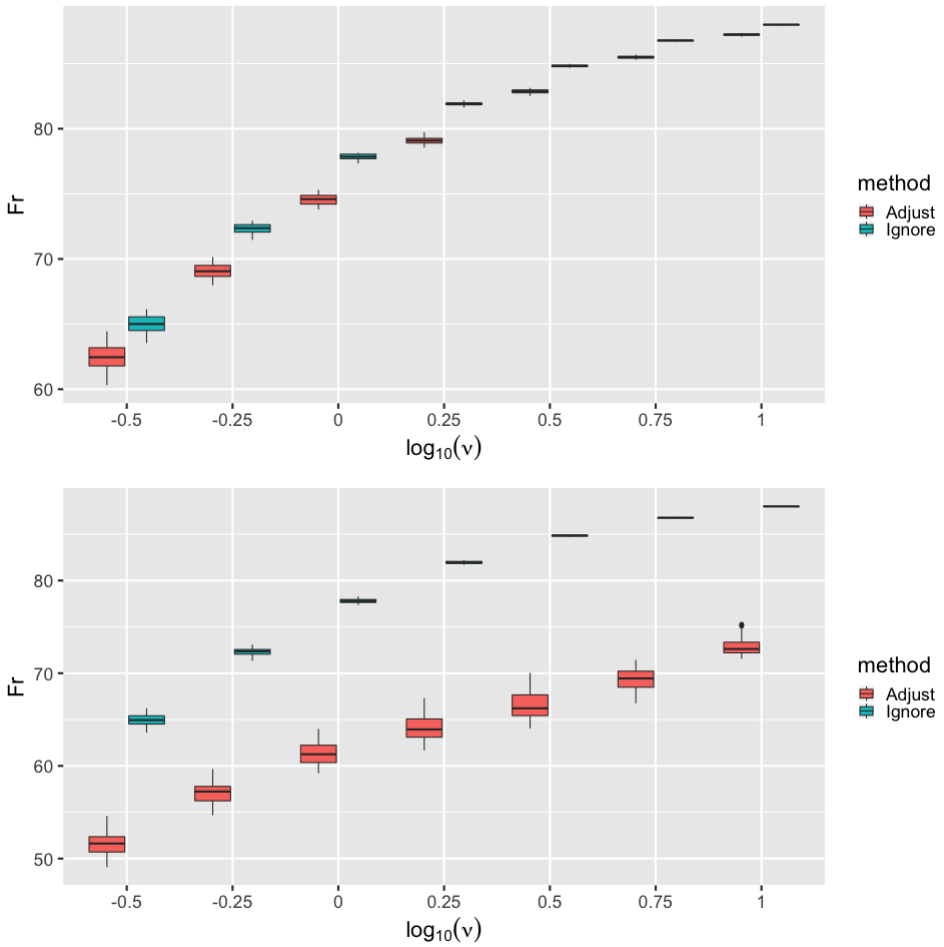

Figure 3.2 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(2) model over different magnitude of measurement error following the uniform distribution.

finite-sample performance. In this simulation study, we assume the measurement error H_p to have the Student t-distribution with degree of freedom 5. We still consider the four structures for the true precision matrix discussed in Section 2.5.2 and fixed the dimension $p = 50$ and sample size $n = 100$. Only the larger ν's are considered, which varies from $10^{-0.5}$ to 10 over 7 different values equally spaced on the log-scale, since the benefit of adjusting for the measurement error fades with small ν. We run 100 replicates for each combination of the sparse structure of the true precision matrix and the choices of ν. The procedure for generating the data is identical to that in Section 3.3.1 except that the uniform distribution is now replaced by the Student t-distribution with degree of freedom 5.

The MCMC sampling scheme described in Section 3.2.3 is utilized to sample from the posterior, where the proposal distributions are chosen as $\mathsf{N}_p(\cdot, \nu I_p/10)$ and $\mathsf{N}_p(\cdot, \nu I_p/100)$ in Step 1, respectively and the priors of Ω are specified on the Cholesky decomposition as that in Section 2.3.2. The hyperparameters are assigned as follows: $\alpha_1 = \beta_1 = 1/2$, $C_p = 1$, $\sigma_0^2 = 0.0001$ and $\sigma_1^2 = 1$, identical to that in Section 3.3.1. The initial values for the X's in Monte Carlo chains are chosen as same as Y for

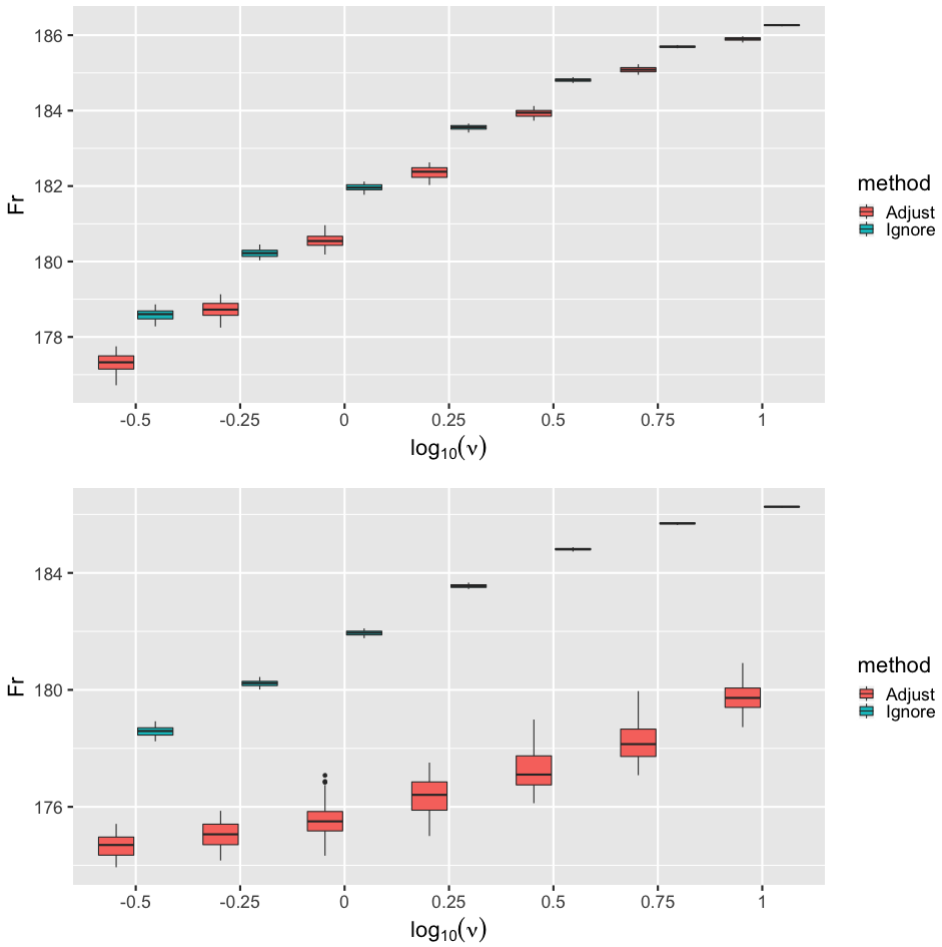

Figure 3.3 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(2) model over different magnitude of measurement error following the uniform distribution.

each replicate such that they would not cause any difference between methods.

The posterior mean is used as the estimator for comparing the estimation error of the method to adjust for the measurement error and the one ignoring it, which are noted as "adjust" and "ignore" in the graphs. The estimation error is computed in terms of the Frobenius norm, as shown in Figures 3.5, 3.6, 3.7, and 3.8 for the four structures, respectively, where only the central 90% of the errors are displayed to remove some outliers and the measurement error ν is presented in its logarithm with basis 10.

For these four structures, compared with the baseline model, correcting for the measurement error is beneficial for every magnitude of measurement error, while this positive effect becomes weaker as ν increases. This is natural because it is more difficult to adjust for the measurement error with larger scale. However, the trend of the proposed method in the top panels in Figures 3.5, 3.6, 3.7, and 3.8 indicates that the current proposal distribution does not produce a satisfactory acceptance ratio, which causes the fewer updates in Step 1 of the MCMC algorithm and the smaller variance with

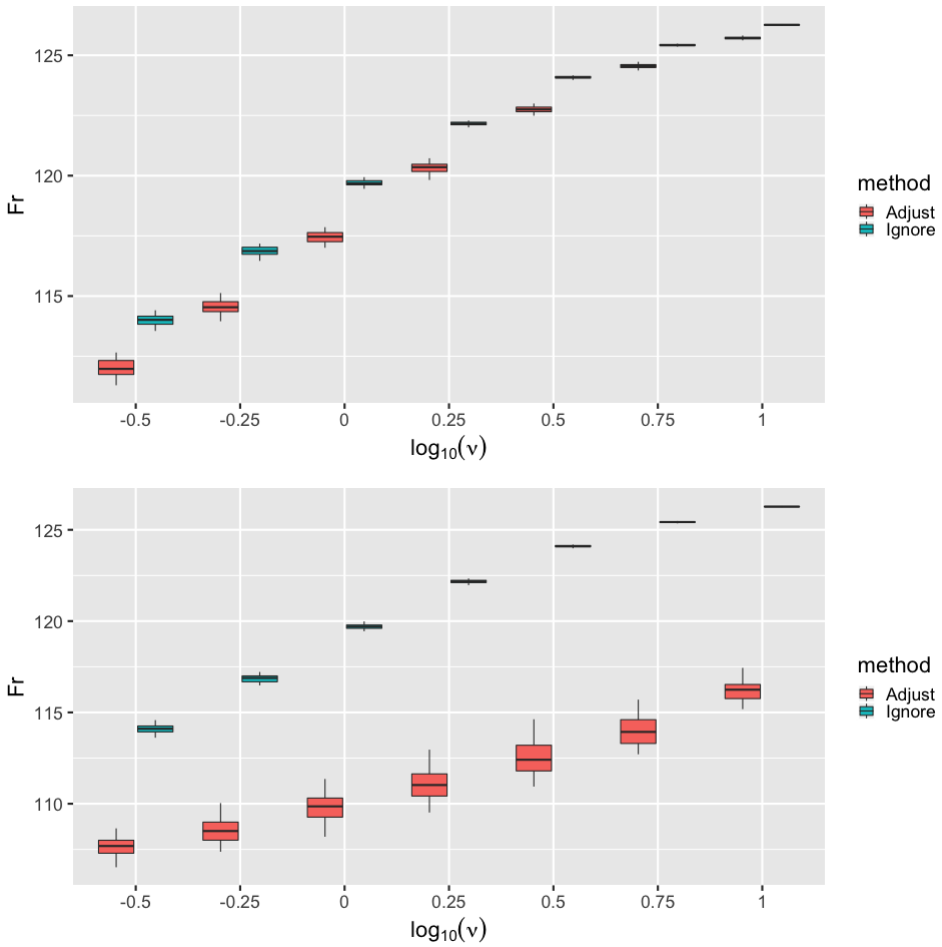

Figure 3.4 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(5) model over different magnitude of measurement error following the uniform distribution.

a larger ν. Therefore, when the proposal distribution is changed to $\mathsf{N}_p(\cdot, \nu I_p/100)$, the estimation errors become smaller uniformly, although the trend maintains. Furthermore, the variances of the estimation error over difference ν are comparable, which reflects that the acceptance ratio by this narrow proposal distribution is more beneficial and adaptive to different magnitude of ν. On the other hand, the estimation error from the naive method has a smaller variance compared with the proposed method, which includes more variation since it considered the extra layer of measurement error in the model. By comparing among these four structures, we cannot discover any significant difference in the trend, which suggests that the result does not change much among different structures of true precision matrix when the measurement error follows the t_5 distribution.

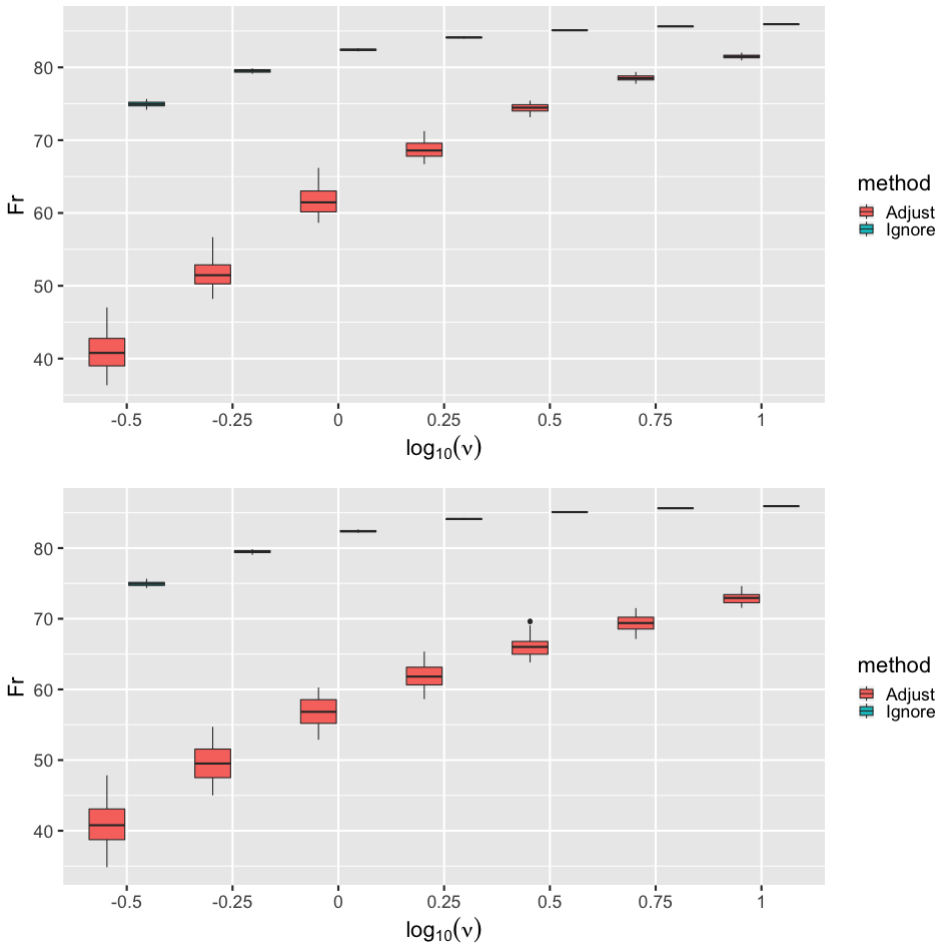

Figure 3.5 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(1) model over different magnitude of measurement error following the t_5 distribution.

3.3.3 Accuracy of Selection

All the former discussion is about the accuracy of the estimation, which is measured between the estimation and the true precision matrix under Frobenius norm. Moreover, there is another important question remain unsolved in the literature of the structure learning even under the no-measurement-error case, which is the selection consistency, that is, to find the non-zero entries more accurately as the sample size goes to infinity. Even though we will not investigate that theoretically, we would like to explore the finite-sample performance of the selection problem using the estimation whose measurement error model is the t_5 distribution.

Since the spike-and-slab prior is specified on the lower-triangular matrix of the Cholesky decomposition, we compute the selection result from the estimation of this lower-triangular matrix. Let Γ denote the selection estimator of this lower-triangular matrix, where $\Gamma_{i,j} = 1$ if the corresponding entry is estimated as non-zero and $\Gamma_{i,j} = 0$ if the corresponding entry is estimated as zero, for

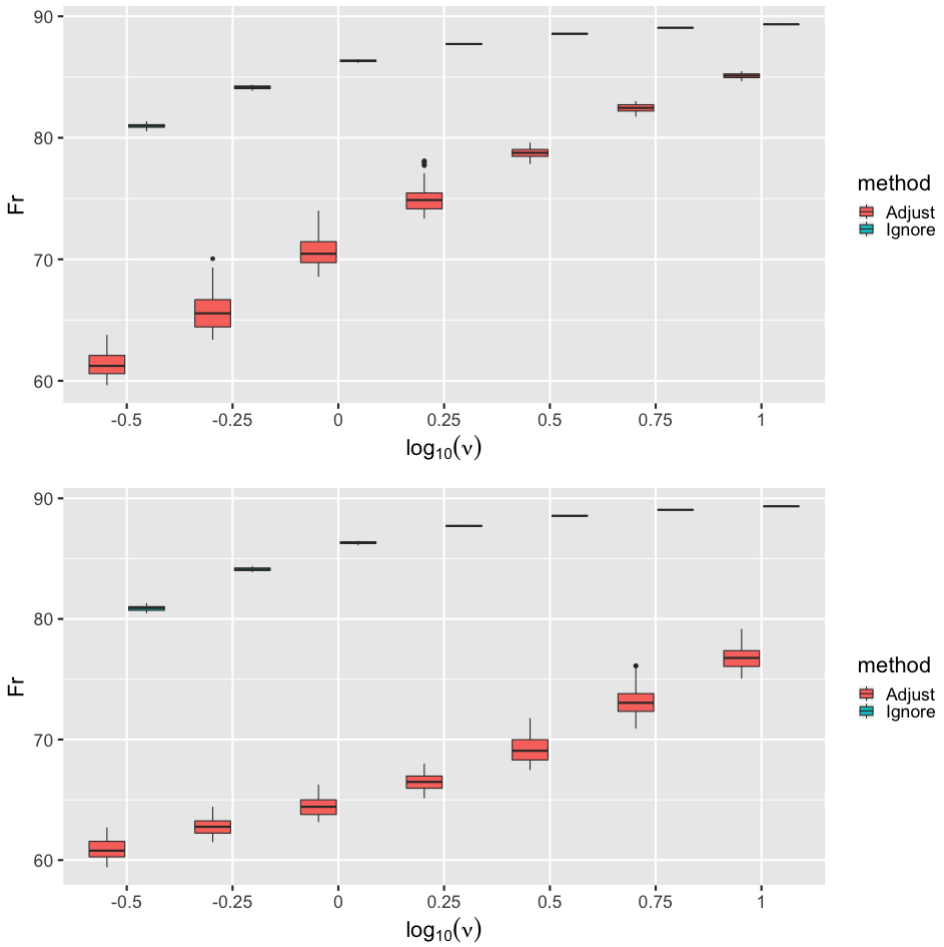

Figure 3.6 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(2) model over different magnitude of measurement error following the t_5 distribution.

$1 \le j < i \le p$. Thanks to the selection property of the spike-and-slab prior, we estimate $\Gamma_{i,j} = 1$ if its posterior probability of staying at the slab side is greater than 0.5, that is, $\Pi_n^{\nu}\left(L_{i,j} = 1\right) > 0.5$, where the posterior probability is calculated by the MCMC samples. Because the sensitivity and specificity may not be comprehensive when the precision matrices are sparse, we consider the Matthews correlation coefficient (MCC) as the justification, which is formulated as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP is the number of true positives, TN the number of true negatives, FP the number of false positives and FN the number of false negatives. We draw the boxplots of these MCCs from the 100 replicates over different magnitude of measurement error and different structures of the true precision matrix as shown in Figures 3.9, 3.10, 3.11, and 3.12 to compare the selection results from the proposed method to adjust for the measurement error and the naive method to ignore the

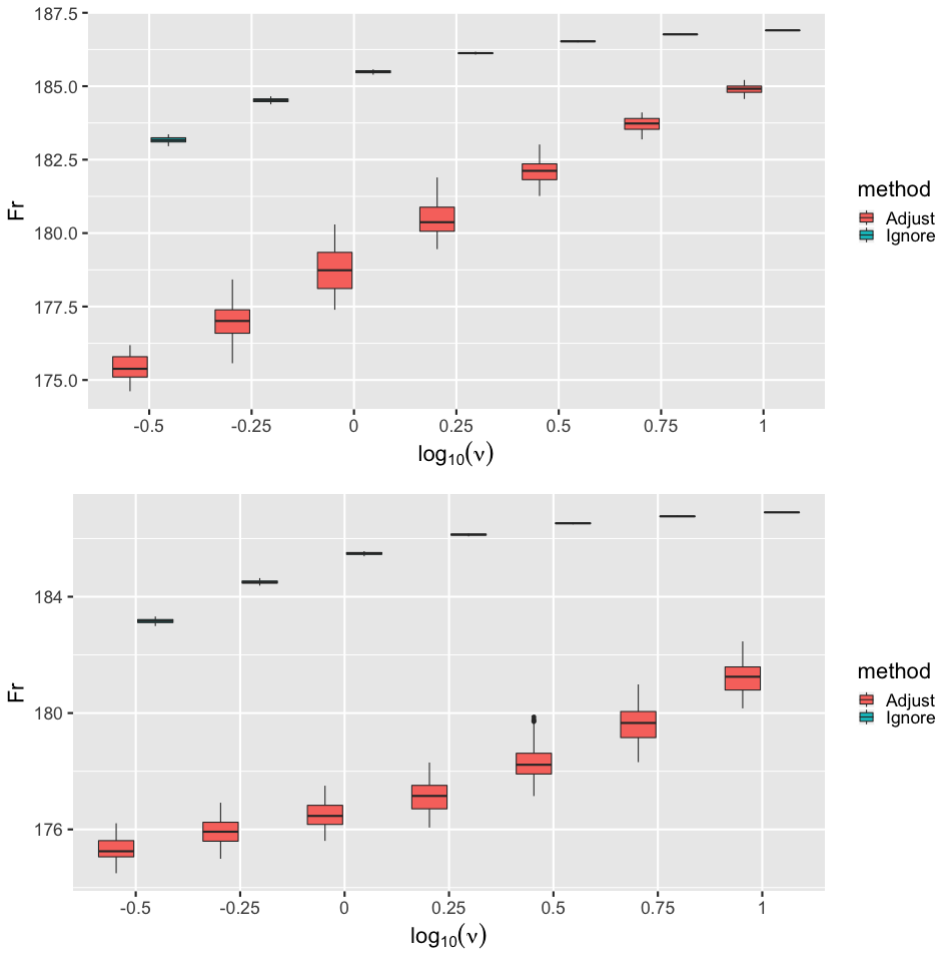

Figure 3.7 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(2) model over different magnitude of measurement error following the t_5 distribution.

measurement error.

For the AR(1) model as shown in Figure 3.9, the selection of both methods is more accurate when the magnitude of measurement error is relatively small and it becomes worse until nearly a random guess when $\nu = 10$. Even though the proposed method performs better when $\nu = 10^{-0.5}$, the selection accuracy in terms of MCC from the proposed method is almost comparable to that from the baseline model for the other choices of ν. However, neither of them offers a satisfactory selection result when the large measurement error is involved. While the selection result of the proposed method is improved when the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/100)$ is deployed in the bottom panel, the trend does not change. When the structure becomes more complex as shown in Figures 3.10, 3.11 and 3.12, neither of the proposed method or the baseline method gives significantly better selection result compared with the random prediction. However, notice that the proposed method performs slightly better than the baseline for every magnitude ν, which indicates that the measurement error still needs to be adjusted.

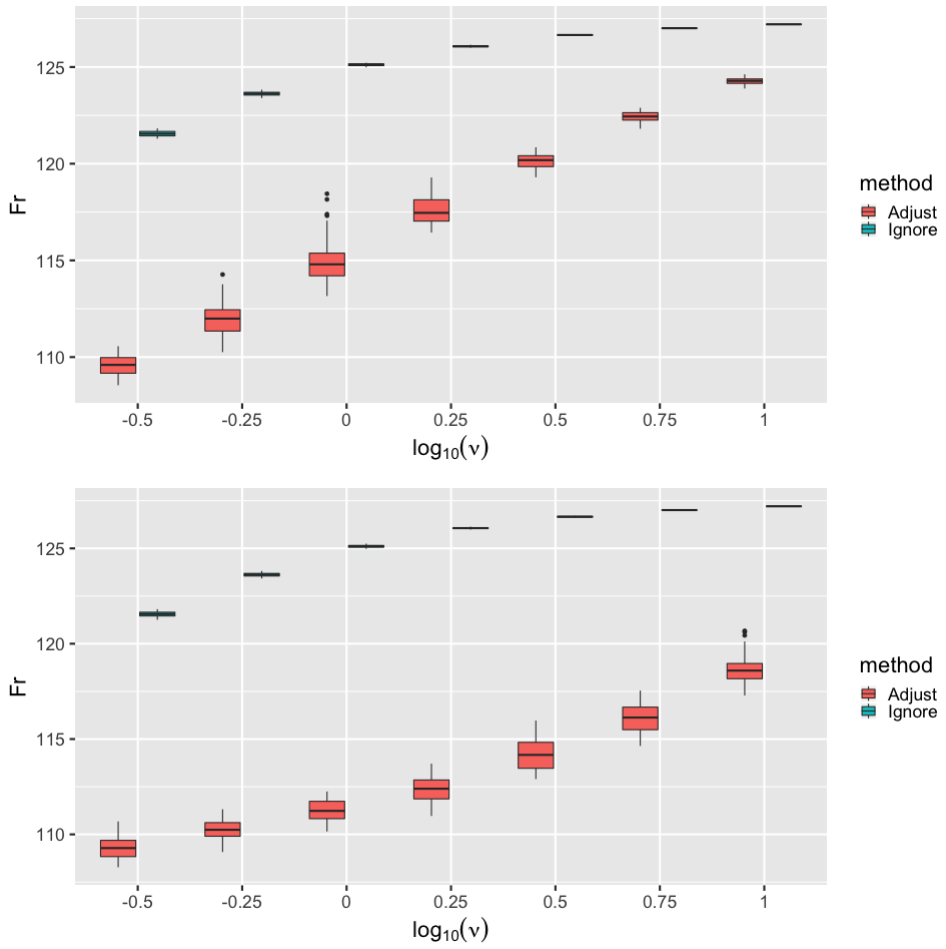

Figure 3.8 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(5) model over different magnitude of measurement error following the t_5 distribution.

3.4 Proofs of the Theorems

3.4.1 Proof of Theorem 3.2.1

We will prove the theorem by the same strategy as the proof of Theorem 2.6.1, which is to first obtain the posterior contraction rate in terms of the Rényi divergence and then to transfer the result to that under the Frobenius norm. Luckily, since the first part of the proof only replies on the Condition 2.2.1, which is already assumed in Theorem 3.2.1, we can simply repeat the Steps 1–4 of the proof in Section 2.6.1 to get

$$\mathsf{E}_{\Omega^*,\nu}\Pi_n^{\nu}\big(\{\Omega : R(f_{\Omega^*,\nu}, f_{\Omega,\nu}) > L\epsilon_n\}\big) \to 0 \quad \text{as } n \to \infty,$$

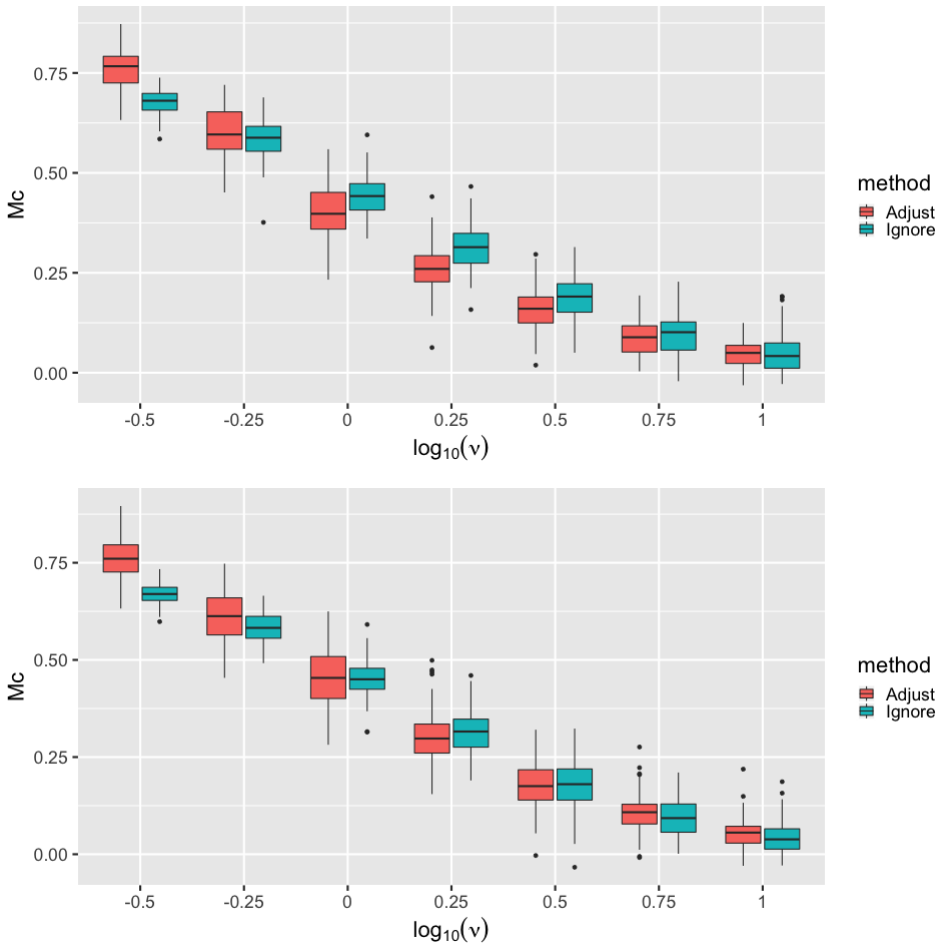

Figure 3.9 Boxplots of the MCC using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(1) model over different magnitude of measurement error following the t_5 distribution.

where the Rényi divergence is defined as

$$R(f_{\Omega^\star,\nu}, f_{\Omega,\nu}) = -\log \int \left(f_{\Omega^\star,\nu} f_{\Omega,\nu}\right)^{1/2}.$$

For the event that $A = \{\Omega : R(f_{\Omega^\star,\nu}, f_{\Omega,\nu}) > L\epsilon_n\}$, the basic probability addition rule and the conditional probability tell us that

$$\begin{aligned}\Pi_n^\nu(A) &= \Pi_n^\nu(A \mid \|\Omega\|_\infty \le M_n)\Pi_\Omega(\|\Omega\|_\infty \le M_n) + \Pi_n^\nu(A \mid \|\Omega\|_\infty \ge M_n)\Pi_\Omega(\|\Omega\|_\infty \ge M_n) \\ &\le \Pi_n^\nu(A \mid \|\Omega\|_\infty \le M_n) + \Pi_\Omega(\|\Omega\|_\infty \ge M_n),\end{aligned}$$

where the later decays to 0 in the rate of $\exp\left(-n\epsilon_n^2\right)$ by the Condition 3.2.1. Therefore, to complete the transfer of the norms, it is suffice to prove that $\|\Omega - \Omega^\star\|_F \lesssim \epsilon_n$ given $R(f_{\Omega^\star,\nu}, f_{\Omega,\nu}) \lesssim \epsilon_n^2$ for any Ω with $\|\Omega\|_\infty \le M_n$ under the assumptions in Theorem 3.2.1.

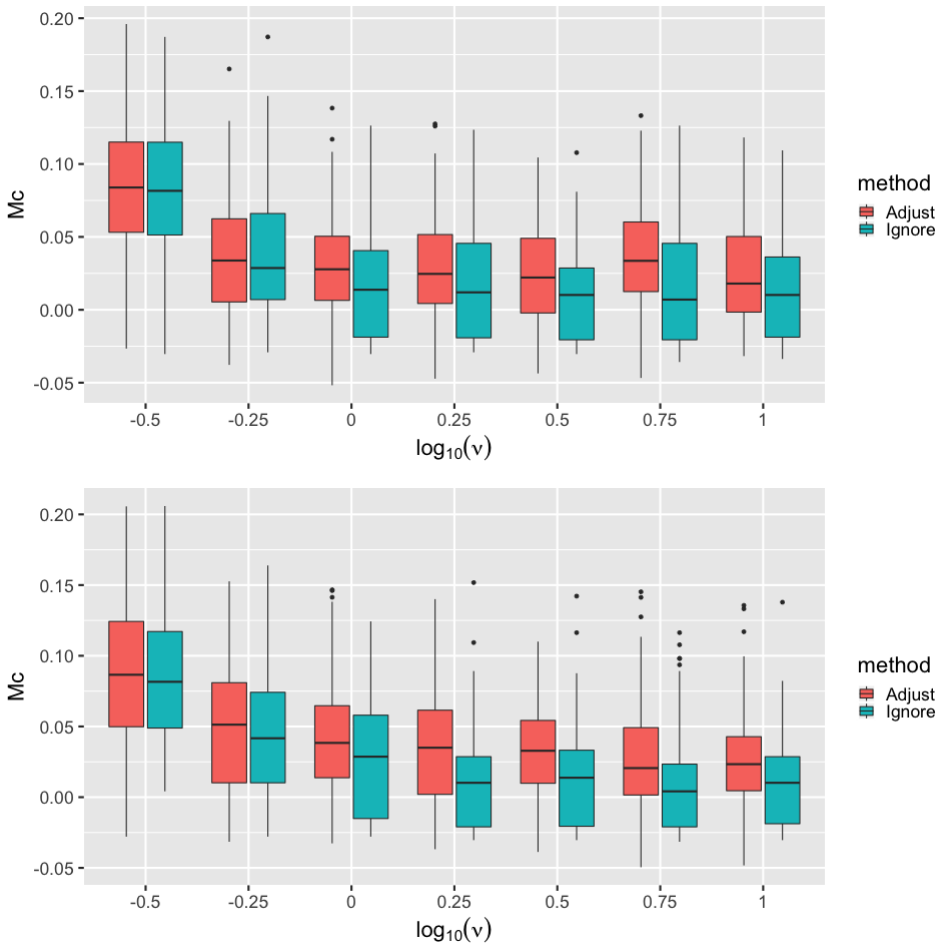

Figure 3.10 Boxplots of the MCC using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(2) model over different magnitude of measurement error following the t_5 distribution.

To achieve this, we first do a change of variable as

$$f_{\Omega,\nu}(y) = \int \nu^{-p/2} h_p\left(\nu^{-1/2}(y-x)\right) g_\Omega(x)\, dx = \int h_p(z)\, g_\Omega\left(y - \nu^{1/2} z\right) dz.$$

Then, expand the Gaussian density and we have

$$\begin{aligned} g_\Omega\left(y - \nu^{1/2} z\right) &= \sqrt{\frac{\det(\Omega)}{(2\pi)^p}} \exp\left(-\frac{y^{\mathrm{T}}\Omega y - 2\sqrt{\nu} y^{\mathrm{T}}\Omega z + \nu z^{\mathrm{T}}\Omega z}{2}\right) \\ &= g_\Omega(y) \exp\left(\frac{2\sqrt{\nu} y^{\mathrm{T}}\Omega z - \nu z^{\mathrm{T}}\Omega z}{2}\right). \end{aligned}$$

Let $u_\Omega(z, y) = \sqrt{\nu} y^{\mathrm{T}}\Omega z - \nu z^{\mathrm{T}}\Omega z/2$ and we use the Taylor expansion on the exponential term at

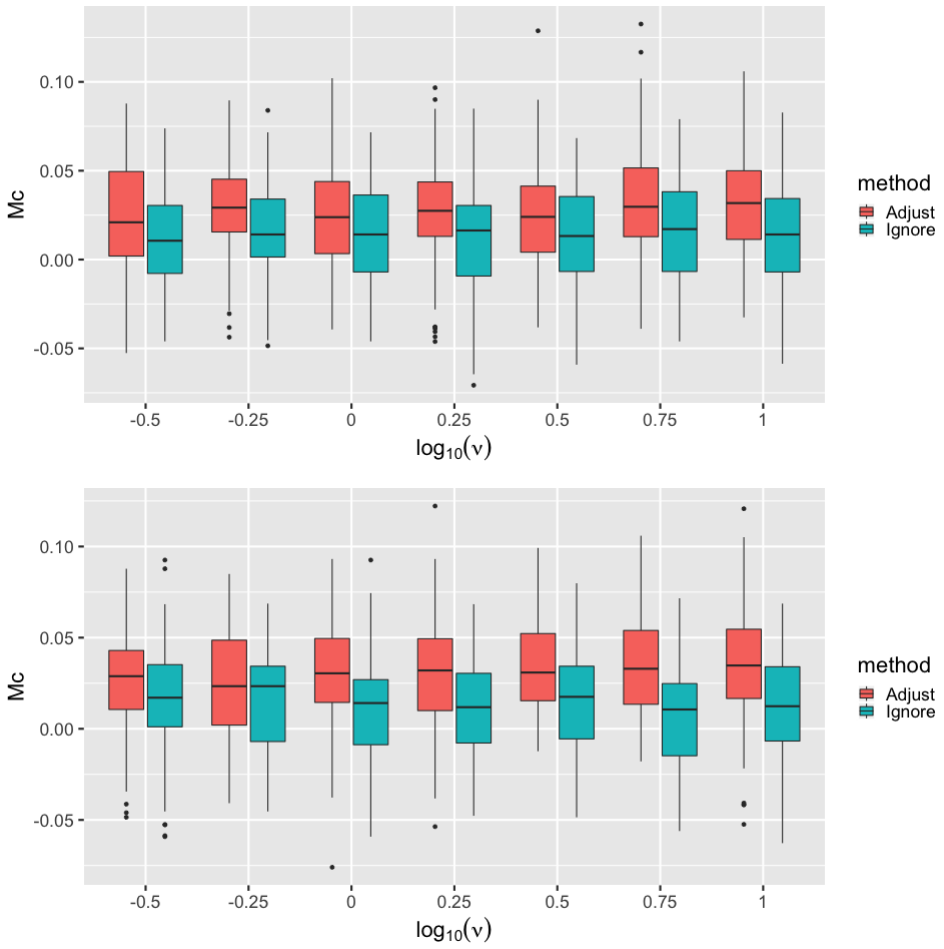

Figure 3.11 Boxplots of the MCC using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(2) model over different magnitude of measurement error following the t_5 distribution.

$u_\Omega(z,y)=0$ serving as an upper bound, that is,

$$\exp\big(u_\Omega(z,y)\big)\le 1+u_\Omega(z,y)+\frac{u_\Omega^2(z,y)}{2}+\frac{u_\Omega^2(z,y)}{2}\exp\big\{u_\Omega(z,y)\big\}.$$

For the ease of writing, we express the upper bound of $f_{\Omega,\nu}$ as

$$f_{\Omega,\nu}(y)\le g_\Omega(y)\big\{1+r_\Omega(y)+s_\Omega(y)+t_\Omega(y)\big\}, \tag{3.4.1}$$

where

$$r_\Omega(y)=\int u_\Omega(z,y)\,h_p(z)\,dz=\int\big(\sqrt{\nu}y^{\mathrm{T}}\Omega z-\nu z^{\mathrm{T}}\Omega z/2\big)h_p(z)\,dz=-\frac{\nu}{2}\mathrm{tr}(\Omega),$$

$$s_\Omega(y)=\int\frac{u_\Omega^2(z,y)}{2}h_p(z)\,dz,\quad t_\Omega(y)=\int\frac{u_\Omega^2(z,y)}{2}\exp\big\{u_\Omega(z,y)\big\}\,h_p(z)\,dz,$$

since the mean of h is 0 and the variance of it is 1. Since $r_\Omega(y)$ is a constant in y and strictly less

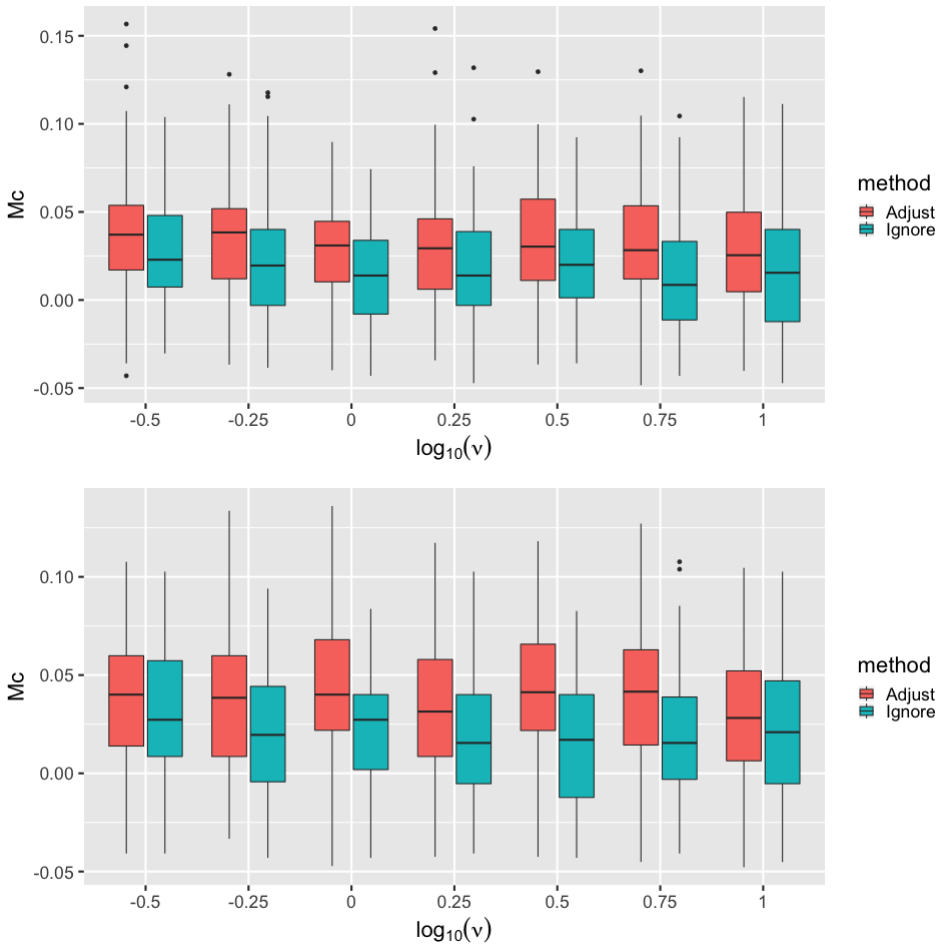

Figure 3.12 Boxplots of the MCC using the proposal distribution $\mathsf{N}_p(\cdot, vI_p/10)$ (top) and $\mathsf{N}_p(\cdot, vI_p/100)$ (bottom) in the Block(5) model over different magnitude of measurement error following the t_5 distribution.

than 0, ignoring it from the right hand side of (3.4.1) will increase the value and therefore maintain the inequality. We then have the inequality as

$$f_{\Omega,v}(y) \leq g_{\Omega}(y)\{1+s_{\Omega}(y)+t_{\Omega}(y)\},$$

where we shall keep all the other terms as what they are for now and compute it later.

Then, let $\bar{\Omega} = (\Omega+\Omega^{\star})/2$ and we consider the affinity $\int \sqrt{f_{\Omega,v}(y) f_{\Omega^{\star},v}(y)}\, dy$, which is equal to

$$\begin{aligned}
&\int \sqrt{g_{\Omega}(y) g_{\Omega^{\star}}(y)} \cdot \sqrt{1+s_{\Omega}(y)+t_{\Omega}(y)} \cdot \sqrt{1+s_{\Omega^{\star}}(y)+t_{\Omega^{\star}}(y)} dy \\
&= \frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^{\star})}{\det^{1/2}(\bar{\Omega})} \int g_{\bar{\Omega}}(y) \cdot \sqrt{1+s_{\Omega}(y)+t_{\Omega}(y)} \cdot \sqrt{1+s_{\Omega^{\star}}(y)+t_{\Omega^{\star}}(y)} dy \\
&\leq \frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^{\star})}{\det^{1/2}(\bar{\Omega})} \int g_{\bar{\Omega}}(y) \cdot \left\{1+\frac{s_{\Omega}(y)+t_{\Omega}(y)}{2}+\frac{s_{\Omega^{\star}}(y)+t_{\Omega^{\star}}(y)}{2}\right\} dy. \qquad (3.4.2)
\end{aligned}$$

There are two integrals that need to be computed on the right hand side and we shall first calculate the one involving $s_\Omega(y)$. By definition, we have

$$\int g_{\bar{\Omega}}(y)\, s_\Omega(y)\, dy = \frac{1}{8}\int\int \left\{ \nu^2 \left(z^{\mathrm{T}}\Omega z\right)^2 - 4\nu\sqrt{\nu} z^{\mathrm{T}}\Omega z z^{\mathrm{T}}\Omega y + 4\nu z^{\mathrm{T}}\Omega y\, y^{\mathrm{T}}\Omega z \right\} g_{\bar{\Omega}}(y)\, dy\, h_p(z)\, dz.$$

Since the mean of $g_{\bar{\Omega}}$ is 0 and its covariance matrix is $\bar{\Omega}^{-1}$, use the assumption that $\|\Omega\|_\infty \le M_n$ and we further have the right hand side equal to

$$\frac{1}{8}\nu^2 \left\{ c_4 \sum_{i=1}^{p} \Omega_{i,i}^2 + 2\sum_{i=1}^{p}\sum_{j\neq i} \left(\Omega_{i,j}^2 + \Omega_{i,i}\Omega_{j,j}\right) \right\} + \frac{1}{2}\nu \mathrm{tr}\left(\Omega^2\bar{\Omega}^{-1}\right) \le C_1\nu^2 p^2 M_n^2 + \nu p M_n \lesssim \epsilon_n^2,$$

where c_i denotes the i-th moment of h by some constants C_1. Then, we consider the integration involving $t_\Omega(y)$ as

$$\int g_{\bar{\Omega}}(y)\, t_\Omega(y)\, dy = \frac{1}{2}\int\int u_\Omega^2(z,y) \exp\left(\sqrt{\nu} y^{\mathrm{T}}\Omega z - \nu z^{\mathrm{T}}\Omega z/2\right) g_{\bar{\Omega}}(y)\, dy\, h_p(z)\, dz,$$

and by some operations on the Gaussian kernel, the last expression is equivalent to

$$\frac{1}{2}\int\int u_\Omega^2(z,y)\, g_{\bar{\Omega}}\left(y - \sqrt{\nu}\bar{\Omega}^{-1}\Omega z\right) dy\, \exp\left\{ \frac{\nu}{2} z^{\mathrm{T}}\Omega\left(\bar{\Omega}^{-1}\Omega - I\right) z \right\} h_p(z)\, dz.$$

Then, replacing $u_\Omega(z,y)$ by its definition and calculating the integration in terms of y, we obtain that the right hand side is equal to

$$\frac{1}{8}\int \left\{ \nu^2 \left(z^{\mathrm{T}}\Omega z\right)^2 - 4\nu^2 z^{\mathrm{T}}\Omega z z^{\mathrm{T}}\Omega\bar{\Omega}^{-1}\Omega z + 4\nu z^{\mathrm{T}}\Omega\bar{\Omega}^{-1}\Omega z \right\} \exp\left\{ \frac{\nu}{2} z^{\mathrm{T}}\Omega\left(\bar{\Omega}^{-1}\Omega - I\right) z \right\} h_p(z)\, dz,$$

which is less than or equal to

$$\frac{1}{8}\nu^2 \int \left(z^{\mathrm{T}}\Omega z\right)^2 \exp\left(\frac{\nu}{2} z^{\mathrm{T}}\Omega z\right) h_p(z)\, dz + \frac{1}{2}\nu \int z^{\mathrm{T}}\Omega\bar{\Omega}^{-1}\Omega z \exp\left(\frac{\nu}{2} z^{\mathrm{T}}\Omega z\right) h_p(z)\, dz.$$

By the Cauchy-Schwarz inequality, there exists an upper bound for the last expression as

$$\nu^2 \left\{ \int \left(z^{\mathrm{T}}\Omega z\right)^4 h_p(z)\, dz \right\}^{1/2} + \nu \left\{ \int \exp\left(\nu z^{\mathrm{T}}\Omega z\right) h_p(z)\, dz \right\}^{1/2} + \nu \left\{ \int \left(z^{\mathrm{T}}\Omega\bar{\Omega}^{-1}\Omega z\right)^2 h_p(z)\, dz \right\}^{1/2}, \tag{3.4.3}$$

where the first term and the last term are bounded by $\nu^2 p^2 M_n^2$ and $\nu p M_n$ up to some constant, respectively. Since h is sub-Gaussian with mean 0 and variance 1, we obtain an upper bound for the

middle term, that is, by the independence of the entries of Z,

$$\begin{aligned}
\nu\left\{\int \exp\left(\nu z^{\mathrm{T}}\Omega z\right) h_p(z)\, dz\right\}^{1/2} &\leq \nu\left(\mathsf{E}\left(\exp\left(\nu p M_n Z_{1,1}^2\right)\right)\right)^{p/2} \\
&\leq \nu\left\{1+\int_0^\infty 2\nu p M_n z_1 \exp\left(\nu p M_n z_1^2\right)\Pi\left(|Z_{1,1}|>z_1\right) dz_1\right\}^{p/2} \\
&\leq \nu\left\{1+2c_1\nu p M_n \int_0^\infty z_1 \exp\left(\left(\nu p M_n - c_2\right) z_1^2\right) dz_1\right\}^{p/2} \\
&= \nu\left(1+\frac{c_1\nu p M_n}{c_2-\nu p M_n}\right)^{p/2} \leq \nu\left(1+\frac{c_3\epsilon_n^2}{p}\right)^{p/2} \lesssim \nu\exp(c_3\epsilon_n^2/2),
\end{aligned}$$

which has a lower order than $\nu p M_n$. Therefore, we find a constant upper bound of (3.4.3). Therefore, repeating all the procedure on $s_{\Omega^\star}(y)$ and $t_{\Omega^\star}(y)$ and applying the upper bounds for those four quantities with respect to y in (3.4.2), it follows that

$$\int \sqrt{f_{\Omega,\nu}(y) f_{\Omega^\star,\nu}(y)}\, dy \leq \frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)}\left(1+C_2\epsilon_n^2\right),$$

for some constant $C_2>0$. It follows from the definition of Rényi divergence and some basic inequality of logarithm that

$$\begin{aligned}
-\log\int \sqrt{f_{\Omega,\nu}(y) f_{\Omega^\star,\nu}(y)}\, dy &\geq -\log\frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} - \log\left(1+C_2\epsilon_n^2\right) \\
&\geq -\log\frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} - C_2\epsilon_n^2,
\end{aligned}$$

where the right hand side is lower bounded by ϵ_n^2 up to some negative constant. Therefore, given that $R(f_{\Omega^\star,\nu}, f_{\Omega,\nu}) \lesssim \epsilon_n^2$, it tells us that

$$-\log\frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} \lesssim \epsilon_n^2,$$

where the left hand side is exactly the Rényi divergence of g_Ω and $g_{\Omega^\star}$, which is written as $R(\Omega^\star,\Omega)$ for simplicity. We will use this result to show that $\|\Omega-\Omega^\star\|_F \lesssim \epsilon_n$.

To this end, let $A=\Omega^{\star-1/2}\Omega\Omega^{\star-1/2}$ and $\alpha_1\leq\cdots\leq\alpha_p$ denote the eigenvalues of A in the increasing order. According to the result from Lemma A.2(ii) of Banerjee & Ghosal (2015), if the Rényi divergence (equivalently, the Hellinger distance) of $g_{\Omega^\star}$ from g_Ω is sufficiently small, then $\max\{|\alpha_j-1| : j=1,\ldots,p\}\leq 1$ and, therefore, $\alpha_p\leq 2$. Since $4\alpha(1+\alpha)^{-2}<1$ for all $\alpha\in(0,2]$, and $-\log x\geq 1-x$ for all

$x \in (0,1)$, we get that the Rényi divergence of g_Ω and $g_{\Omega^\star}$, that is $R(\Omega^\star, \Omega)$, can be written as

$$-\frac{1}{4}\log\frac{|A|}{|\frac{1}{2}I_p+\frac{1}{2}A|^2} = -\frac{1}{4}\sum_{j=1}^{p}\log\frac{4\alpha_j}{(1+\alpha_j)^2} \geq \frac{1}{4}\sum_{j=1}^{p}\Big\{1-\frac{4\alpha_j}{(1+\alpha_j)^2}\Big\} = \frac{1}{4}\sum_{j=1}^{p}\left(\frac{1-\alpha_j}{1+\alpha_j}\right)^2.$$

Since $1+\alpha_j \leq 1+\alpha_p \leq 3$ for all j, and $\|A-I_p\|_F^2 = \sum_{j=1}^{p}(1-\alpha_j)^2$, we have that $R(\Omega^\star,\Omega) \gtrsim \|A-I_p\|_F^2$. Furthermore, since $\Omega^\star$ is assumed to have bounded eigenvalues, it follows from that last inequality that

$$\|\Omega-\Omega^\star\|_F^2 \leq \|\Omega^\star\|_2^2\|A-I_p\|_F^2 \lesssim \epsilon_n^2.$$

We then prove the assertion at the beginning and finish this proof.

3.4.2 Proof of Theorem 3.2.2

This proof proceeds as in Section 3.4.1 with some change in a main step to alleviate the requirement on the tail probability of h. We first repeat all the steps and calculations in the proof of Section 3.4.1 until we obtain the inequality

$$f_{\Omega,\nu}(y) \leq g_\Omega(y)\big\{1+s_\Omega(y)+t_\Omega(y)\big\},$$

where we again let $u_\Omega(z,y) = \sqrt{\nu}y^{\mathrm{T}}\Omega z - \nu z^{\mathrm{T}}\Omega z/2$ and define the following quantities

$$s_\Omega(y) = \int \frac{u_\Omega^2(z,y)}{2}h_p(z)\,dz, \quad t_\Omega(y) = \int \frac{u_\Omega^2(z,y)}{2}\exp\big\{u_\Omega(z,y)\big\}\,h_p(z)\,dz.$$

Then, we consider the affinity $\int\sqrt{f_{\Omega,\nu}(y)\,f_{\Omega^\star,\nu}(y)}\,dy$, which is equal to

$$\begin{aligned}
&\int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\cdot\sqrt{1+s_\Omega(y)+t_\Omega(y)}\cdot\sqrt{1+s_{\Omega^\star}(y)+t_{\Omega^\star}(y)}\,dy \\
&\quad\leq \int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\cdot\Big(1+\sqrt{s_\Omega(y)+t_\Omega(y)}\Big)\cdot\Big(1+\sqrt{s_{\Omega^\star}(y)+t_{\Omega^\star}(y)}\Big)\,dy, \\
&\quad\leq \int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\sqrt{s_\Omega(y)+t_\Omega(y)}\,dy + \int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\sqrt{s_{\Omega^\star}(y)+t_{\Omega^\star}(y)}\,dy \\
&\qquad+\int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\,dy + \int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\cdot\sqrt{s_\Omega(y)+t_\Omega(y)}\cdot\sqrt{s_{\Omega^\star}(y)+t_{\Omega^\star}(y)}\,dy. \qquad (3.4.4)
\end{aligned}$$

since both $s_\Omega(y)$ and $t_\Omega(y)$ are positive for any positive semi-definite Ω. Note that there are three terms that need an upper bound except the third term of (3.4.4), since it is the essential term leading us to the Frobenius norm as shown in the last step of the proof in Section 3.4.1. Since the first two quantities are symmetric by switching Ω with $\Omega^\star$, the upper bound for the first term will be shown and the second follows. By the Cauchy-Schwarz inequality, an upper bound of the first term in

(3.4.4) is displayed as

$$\int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\sqrt{s_\Omega(y)+t_\Omega(y)}dy \leq \left[\int \{s_\Omega(y)+t_\Omega(y)\}g_\Omega(y)dy\right]^{1/2},$$

where there are two quantities that need to be computed on the right hand side and we shall first work on the first one involving $s_\Omega(y)$. By definition, we have

$$\int g_\Omega(y)s_\Omega(y)dy = \frac{1}{8}\int\int \left\{\nu^2(z^T\Omega z)^2 - 4\nu\sqrt{\nu}z^T\Omega z z^T\Omega y + 4\nu z^T\Omega y y^T\Omega z\right\}g_\Omega(y)dy\, h_p(z)dz.$$

Since the mean of g_Ω is 0 and its covariance matrix is Ω^{-1}, use the assumption that $\|\Omega\|_\infty \leq M_n$ and we further have the right hand side equal to

$$\frac{1}{8}\nu^2\left\{c_4\sum_{i=1}^p \Omega_{i,i}^2 + 2\sum_{i=1}^p\sum_{j\neq i}\left(\Omega_{i,j}^2+\Omega_{i,i}\Omega_{j,j}\right)\right\} + \frac{1}{2}\nu\mathrm{tr}\left(\Omega^2\Omega^{-1}\right) \leq C_1\nu^2p^2M_n^2 + \nu pM_n \lesssim \epsilon_n^4,$$

where c_i denotes the i-th moment of h by some constants C_1. Then, we consider the integration involving $t_\Omega(y)$ as

$$\int g_\Omega(y)t_\Omega(y)dy = \frac{1}{2}\int\int u_\Omega^2(z,y)\exp\left(\sqrt{\nu}y^T\Omega z - \nu z^T\Omega z/2\right)g_\Omega(y)dy\, h_p(z)dz,$$

and by some operations on the Gaussian kernel and replacing $u_\Omega(z,y)$ by its definition, the last expression is equivalent to

$$\frac{1}{8}\int\int \left\{\nu^2(z^T\Omega z)^2 - 4\nu\sqrt{\nu}z^T\Omega z z^T\Omega y + 4\nu z^T\Omega y y^T\Omega z\right\}g_\Omega(y-\sqrt{\nu}z)dy\, h_p(z)dz.$$

Then, calculating the integration in terms of y and ignoring the strictly negative term, we obtain that the right hand side is less than or equal to

$$\begin{aligned}&\frac{1}{8}\nu^2\int (z^T\Omega z)^2 h_p(z)dz + \frac{1}{2}\nu\int z^T\Omega z h_p(z)dz \\ &\quad= \frac{1}{8}\nu^2\left\{c_4\sum_{i=1}^p\Omega_{i,i}^2 + 2\sum_{i=1}^p\sum_{j\neq i}\left(\Omega_{i,j}^2+\Omega_{i,i}\Omega_{j,j}\right)\right\} + \frac{1}{2}\nu\mathrm{tr}(\Omega) \leq C_1\nu^2p^2M_n^2 + \nu pM_n \lesssim \epsilon_n^4.\end{aligned}$$

Summarizing all these results and we obtain that in (3.4.4),

$$\int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\sqrt{s_\Omega(y)+t_\Omega(y)}dy \lesssim \epsilon_n^2, \quad \int \sqrt{g_\Omega(y)g_{\Omega^\star}(y)}\sqrt{s_{\Omega^\star}(y)+t_{\Omega^\star}(y)}dy \lesssim \epsilon_n^2.$$

By the Cauchy-Schwarz inequality, the fourth term in (3.4.4) is less than or equal to

$$\left[\int \left\{s_{\Omega}(y)+t_{\Omega}(y)\right\} g_{\Omega}(y)\, dy\right]^{1/2} \cdot \left[\int \left\{s_{\Omega^\star}(y)+t_{\Omega^\star}(y)\right\} g_{\Omega^\star}(y)\, dy\right]^{1/2},$$

which has a lower order than ϵ_n^2 as $\epsilon_n \to 0$. Therefore, the fourth term in (3.4.4) is negligible compared with the remaining terms, and we obtain an upper bound for the affinity as

$$\int \sqrt{f_{\Omega,\nu}(y) f_{\Omega^\star,\nu}(y)}\, dy \le \int \sqrt{g_{\Omega}(y) g_{\Omega^\star}(y)}\, dy + C_3\epsilon_n^2 \le \frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} + C_3\epsilon_n^2,$$

where $\bar{\Omega} = (\Omega+\Omega^\star)/2$ by some constant C_3.

Given that $R(f_{\Omega^\star,\nu}, f_{\Omega,\nu}) \lesssim \epsilon_n^2$, it follows from the definition of Rényi divergence that

$$-\log\left\{\frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} + C_3\epsilon_n^2\right\} \le -\log\int \sqrt{f_{\Omega,\nu}(y) f_{\Omega^\star,\nu}(y)}\, dy \le C_4\epsilon_n^2,$$

for some constant C_4. Through some basic calculations of the logarithm and the fact that there exists some constants C_5 and C_6 such that $\exp\left(-C_4\epsilon_n^2\right) \ge 1 - C_5\epsilon_n^2$ and $-\log\left(1-(C_3+C_5)\epsilon_n^2\right) \le C_6\epsilon_n^2$ when ϵ_n is small enough, we then obtain the following inequality that

$$-\log\frac{\det^{1/4}(\Omega)\det^{1/4}(\Omega^\star)}{\det^{1/2}\left(\bar{\Omega}\right)} \le -\log\left\{\exp\left(-C_4\epsilon_n^2\right) - C_3\epsilon_n^2\right\} \le C_6\epsilon_n^2,$$

where the left hand side is exactly the Rényi divergence of g_{Ω} and $g_{\Omega^\star}$, which is written as $R(\Omega^\star,\Omega)$ for simplicity. We will use this result to show that $\|\Omega-\Omega^\star\|_F \lesssim \epsilon_n$.

To prove this, let $A = \Omega^{\star-1/2}\Omega\Omega^{\star-1/2}$ and $\alpha_1 \le \cdots \le \alpha_p$ denote the eigenvalues of A in the increasing order. According to the result from Lemma A.2(ii) of Banerjee & Ghosal (2015), if the Rényi divergence (equivalently, the Hellinger distance) of $g_{\Omega^\star}$ from g_{Ω} is sufficiently small, then $\max\{|\alpha_j - 1| : j = 1,\ldots,p\} \le 1$ and, therefore, $\alpha_p \le 2$. Since $4\alpha(1+\alpha)^{-2} < 1$ for all $\alpha \in (0,2]$, and $-\log x \ge 1-x$ for all $x \in (0,1)$, we get that the Rényi divergence of g_{Ω} and $g_{\Omega^\star}$, that is $R(\Omega^\star,\Omega)$, can be written as

$$-\frac{1}{4}\log\frac{|A|}{|\frac{1}{2}I_p + \frac{1}{2}A|^2} = -\frac{1}{4}\sum_{j=1}^{p}\log\frac{4\alpha_j}{(1+\alpha_j)^2} \ge \frac{1}{4}\sum_{j=1}^{p}\left\{1 - \frac{4\alpha_j}{(1+\alpha_j)^2}\right\} = \frac{1}{4}\sum_{j=1}^{p}\left(\frac{1-\alpha_j}{1+\alpha_j}\right)^2.$$

Since $1+\alpha_j \le 1+\alpha_p \le 3$ for all j, and $\|A-I_p\|_F^2 = \sum_{j=1}^{p}(1-\alpha_j)^2$, we have that $R(\Omega^\star,\Omega) \gtrsim \|A-I_p\|_F^2$. Furthermore, since $\Omega^\star$ is assumed to have bounded eigenvalues, it follows from that last inequality that

$$\|\Omega-\Omega^\star\|_F^2 \le \|\Omega^\star\|_2^2\|A-I_p\|_F^2 \lesssim \epsilon_n^2.$$

We then prove the necessary condition to convey the rate result from the Rényi divergence to the Frobenius norm by this inequality.

CHAPTER

4

INFERENCE ON A PRECISION MATRIX UNDER GENERALIZED MEASUREMENT ERROR

4.1 Introduction

As discussed in Section 1.2.2.4, any density function from a parametric distribution family serving as the kernel of the mixture model could be recognized as a generalized measurement error model. Consider the Gaussian graphical model as the mixing distribution and the additive measurement error discussed in Chapter 3 specifies this mixture as the convolution using the additive structure. If we abandon the additive relation of the true outcomes and the measurement error but maintain the hierarchical mixture structure, the Gaussian graphical model under generalized measurement error is proposed.

Assume that the "corrupted" observations Y_i's have the hierarchical distribution

$$Y_i \mid X_i \overset{\text{ind}}{\sim} \mathsf{H}_p^{\nu}(X_i), \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0,\Omega^{-1}), \quad i=1,\ldots,n, \tag{4.1.1}$$

where H_p^{ν} denotes the joint distribution of the p independent measurement errors. In other words, each coordinate $Y_{i,j} \mid X_{i,j}$ is independent with $Y_{i,k} \mid X_{i,k}$ for $j \neq k$. Moreover, a scale parameter ν is imposed on the distribution H_p^{ν}, which has mean $f_1(X_{i,j})$ and variance $\nu f_2(X_{i,j})$ for the j-th covariate. For the sake of simplicity, write the mean of $Y_i \mid X_i$ as $f_1(X_i)$ and its variance as $\nu f_2(X_i)$. Apparently, the distribution H_p^{ν} will shrink down to a point mass as $\nu \to 0$, which corresponds to the

no-measurement-error case. Let h_p^ν denote the density of H_p^ν and the marginal distribution of the Y can be indirectly expressed by an integration, i.e.,

$$f_{\Omega,\nu}(y) = \int h_p^\nu(y \mid x) g_\Omega(x)\, dx. \tag{4.1.2}$$

This formulation indicates that the Gaussian measurement error and the additive measurement error are just special cases to this generalized measurement error model by specifying H_p^ν as the multivariate Gaussian distribution or letting H_p^ν belong to the location-scale family. Some other examples of generalized measurement error include the exponential dispersion models (Jørgensen, 1987), which extend the exponential family by an extra scale parameter to account for the dispersion. The dispersed Poisson distribution is an important candidate in such models, where the "corrupted" observations Y_i's have the distribution

$$X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad Z_{i,j} \mid X_{i,j} \overset{\text{iid}}{\sim} \mathsf{Pois}(\exp(X_{i,j})/\nu), \quad Y_i = \nu Z_i, \tag{4.1.3}$$

for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Now the measurement error density h_p^ν has the specific expression

$$h_p^\nu(Y_i \mid X_i) = \prod_{j=1}^{p} \frac{\exp(-\exp(X_{i,j})/\nu)}{(Y_i/\nu)!} \left(\frac{\exp(X_{i,j})}{\nu} \right)^{Y_{i,j}/\nu},$$

where ! denotes the factorial operation. Since $Z_{i,j}$'s are sampled from Poisson distribution, $Y_{i,j}/\nu$'s are natural numbers and the factorials are meaningful. Then, the observations conditioning on the true outcomes have the mean $\mathsf{E}(Y_{i,j} \mid X_{i,j}) = \exp(X_{i,j})$ and the variance $\mathsf{V}(Y_{i,j} \mid X_{i,j}) = \nu \exp(X_{i,j})$. Since the integration (4.1.2) cannot be computed explicitly or even approximately in the general form, the theoretical results will focus on the closeness of the likelihood functions instead of the variable-of-interest directly.

The rest of this chapter is organized as follows. In Section 4.2, the Bayesian method of adjusting for the generalized measurement error in Gaussian graphical model is introduced along with the posterior contraction rate on the closeness of likelihood functions and the MCMC algorithm to sample from the posterior. We conduct a simulation study to illustrate the improvement of adjusting for the measurement error, when it has the Poisson distribution, in Section 4.3.

4.2 Main Results

4.2.1 Prior and Posterior Distributions

Due to the change of the interested metric, the basic conditions for the Gaussian measurement error could meet the demands of this general model. By the same reasons as introduced in Section 2.2.1, the precision matrix is expressed as

$$\Omega = \Theta + \kappa I_p, \tag{4.2.1}$$

where Θ is a positive semi-definite matrix and $\kappa > 0$ serves as a lower bound on the smallest eigenvalue of Ω. A major benefit of such specification is to provide a control on the smallest eigenvalue of the generic Ω and also to guarantee the positive definiteness. Impose independent priors on κ and Θ respectively to satisfy the Condition 2.2.1 and we obtain the prior of Ω. The part (b) of these conditions are now-classical to derive the posterior contraction rate based on the general theory as introduced in Ghosal, Ghosh & van der Vaart (2000).

According to the data generating process in (4.1.1), the likelihood function of the observations is defined as

$$L_n(\Omega; \nu) \propto \prod_{i=1}^{n} \int h_p^{\nu}(Y_i \mid X_i) \sqrt{\det(\Omega)} \exp\left(-X_i^{\mathrm{T}} \Omega X_i / 2\right) dX_i.$$

With the prior of Ω as introduced, the posterior distribution is updated via Bayes's theorem as

$$\Pi_n^{\nu}(d\Omega) = \Pi(d\Omega \mid Y_1, \ldots, Y_n; \nu) \propto L_n(\Omega; \nu) \Pi(d\Omega), \tag{4.2.2}$$

which depends on the data $Y_1, \ldots, Y_n$ and the known scale parameter.

4.2.2 Posterior Contraction Rate

Note that all the previous posterior contraction rate results are obtained with respect to the Frobenius norm, which is an explicit metric on the precision matrices, while such a result is difficult to derive with the generalized measurement error (4.1.1). Particularly, the abstract formulation (4.1.2) impedes the conversion from the closeness of densities to a direct distance of matrices. Therefore, we take a step back and explore the closeness of the likelihood (4.1.2) under the posterior distribution of Ω. Define the Rényi divergence as

$$R(f_{\Omega^\star, \nu}, f_{\Omega, \nu}) = -\log \int \left(f_{\Omega^\star, \nu} f_{\Omega, \nu}\right)^{1/2},$$

and we characterize the posterior contraction rate in terms of the Rényi divergence.

Theorem 4.2.1. *Assume that $\Omega^\star$ has eigenvalues bounded away from 0. Consider a prior distribution for $\Omega = \Theta + \kappa I_p$ induced from independent prior distributions Π_κ for κ and Π_Θ for Θ. Assume that there exists a sequence ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \gtrsim \log n$ such that Π_Θ satisfies Condition 2.2.1 (b)(i)–(iii) for $\Omega^\star$, and Π_κ satisfies Condition 2.2.1 (a). Under the model in* (4.1.1)*, for any measurement error model H_p^{ν}, the posterior distribution Π_n^{ν} in* (4.2.2) *contracts at the rate ϵ_n, that is, there exists a constant $L > 0$, depending on $\|\Omega^\star\|_2$, such that*

$$\mathsf{E}_{\Omega^\star, \nu} \Pi_n^{\nu}(\{\Omega : R(f_{\Omega^\star, \nu}, f_{\Omega, \nu}) > L\epsilon_n\}) \to 0 \quad \text{as } n \to \infty.$$

The proof of this theorem is identical to the one in Section 2.6.1 by abandoning the last step. Even though this convergence is not in terms of a familiar matrix norm, the result is still insightful when considering the generality of this model. In particular, since the observations are generated

with respect to the true density $f_{\Omega^*,\nu}$, this contraction rate can be understood to describe the errors of predictions from the true model and the estimated model, which uses the estimation of Ω. Moreover, Theorem 4.2.1 is valid, even when the measurement error model H_p^ν does not meet the restrictions on the mean or variance.

4.2.3 Computation

Since the formulation of the posterior (4.2.2) is indirect, some MCMC methods are needed to produce samples from Π_n^ν. Luckily, the MCMC algorithm introduced in Section 3.2.3 can be revised to suit for this generalized measurement error and also to leverage the existing methods in literature. For the $(k+1)$-th iteration of the MCMC algorithm,

Step 1: for each i, sample $X_i^{(k+1)}$ from the full conditional posterior

$$(X_i^{(k+1)} \mid Y_i, \Omega^{(k)}) \overset{\text{ind}}{\sim} h_p^\nu(Y_i \mid X_i) g_{\Omega^{(k)}}(X_i),$$

using the Metropolis-Hastings algorithm with some proposal distribution.

Step 2: Sample $\Omega^{(k+1)}$ from its posterior

$$(\Omega \mid X_1^{(k+1)}, \ldots, X_n^{(k+1)}) \sim \Pi(\Omega \mid X_1^{(k+1)}, \ldots, X_n^{(k+1)}).$$

where the second step can be substituted by the MCMC algorithms for the no-measurement-error situation, for instance, the models discussed in Section 2.3, depending on the augmented data $X_1^{(k+1)}, \ldots, X_n^{(k+1)}$, if the technical device κ can be dropped from the expression of Ω. In fact, the technical device κ is an obstacle to efficiently apply these algorithms and dropping it is applicable if the samples of Ω are guaranteed to be positive definite, which is mostly true in finite-sample situations. Regarding the proposal distribution in Step 1, we consider the Gaussian distribution for the simple computation. To avoid an inefficient updates and obtain a satisfactory acceptance ratio, we recommend a small variance that is adaptive to ν, for instance, $\mathsf{N}_p(\cdot, \nu I_p/10)$ or $\mathsf{N}_p(\cdot, \nu I_p/100)$.

4.3 Simulation Study

There are some important generalized measurement error models that we want to analyze numerically, for example, Poisson distribution, and to compare the estimation accuracy with the method ignoring the measurement error over different magnitude of ν. The Poisson distribution is of particular interest not only because it is a very important candidate from the exponential family but also due to the discrete outcomes it generates. All of the measurement errors we discussed in the aforementioned chapters are continuous, where the observations are close to the true outcomes, especially when ν is small. Thus, it would be more efficient to encrypt the personal information as discussed in Section 1.2.2 if the discrete Poisson measurement error is employed. Furthermore, the Poisson measurement error can randomly stratify the personal records and disclose less information

naturally. Therefore, we will conduct a simulation study to explore the effect of adjusting for the Poisson measurement error over ignoring it, where the "corrupted" observations Y_i's are generated from model (4.1.3).

Same as the simulation study in Section 2.5.2, we consider the following four structures on the true precision matrix: AR(1), AR(2), Block(2) and Block(5), with the dimension $p = 50$ and sample size $n = 100$. Since the simulation study for the Gaussian measurement error indicates that adjusting for the measurement error with small ν is not influential, we only consider the ν varying from $10^{-0.5}$ to 10^{1} over 7 different values, which are equally spaced on the log-scale. 100 replicates of experiments are executed for each combination of precision matrix structure and ν.

The observations of each replicate are generated by following these steps:

- Obtain the true precision matrix with one of the aforementioned structures and compute the covariance matrix $\Sigma^\star = \Omega^{\star -1}$;
- Generate $X_1, \ldots, X_n$ from $\mathsf{N}_p(0, \Sigma^\star)$ and $Z_1, \ldots, Z_n$ from the Poisson distribution separately by coordinate, where $Z_{i,j}$ is sampled from $\mathsf{Pois}(\exp(X_{i,j})/\nu)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$;
- For each Z_i, generate the observation by $Y_i = \nu Z_i$.

The MCMC algorithm introduced in Section 4.2.3 is deployed to produce samples from the posterior. The proposal distribution is chosen as $\mathsf{N}_p(\cdot, \nu I_p/10)$ or $\mathsf{N}_p(\cdot, \nu I_p/100)$ respectively and the prior setup (2.3.3) on the Cholesky decomposition structure in Section 2.3.2 is implemented as the Step 2 to get the posterior samples. For the hyper-parameters, we make the same choices as that in Section 2.5.2 that $\alpha_1 = \beta_1 = 1/2$, $C_p = 1$, $\sigma_0^2 = 0.0001$ and $\sigma_1^2 = 1$, since the informative prior performs better than the diffuse prior in the Gaussian measurement error case. Since we know the existence of the Poisson measurement error, the initial values of X_i's are specified as $\log(Y_i)$ respectively for $i = 1, \ldots, n$ in the proposed method to better adjust for this nonlinear corruption, unlike that we simply specify the Y's as the initial values of X's in the additive measurement error cases.

The adjustment for the measurement error is compared with the baseline model where the measurement error is ignored by the estimation error in terms of Frobenius norm between the posterior mean and the truth. These methods are marked as "adjust" and "ignore" in Figures 4.1, 4.2, 4.3, and 4.4 for the four structures, respectively, where ν is expressed on a base-10 logarithmic scale to flatten the curve. In these graphs, the central 90% of the estimation errors are displayed in order to remove the outliers.

For the AR(1) structure as shown in the top panel of the Figure 4.1, adjusting for the measurement error provides a more accurate estimation than ignoring it, while this benefit is reduced as ν increases. Similar trend also happens to the simulation studies of the additive measurement errors. Note that when $\nu = 10$, the variance of the estimation error of the proposed method is large, while the mean of it is still small than that of the baseline model. This can be caused by the insufficient posterior updating using the broad proposal distribution. Thus, when the proposal distribution is changed to $\mathsf{N}_p(\cdot, \nu I_p/100)$, we obtain the same trend of the estimation error along ν and the uncommon

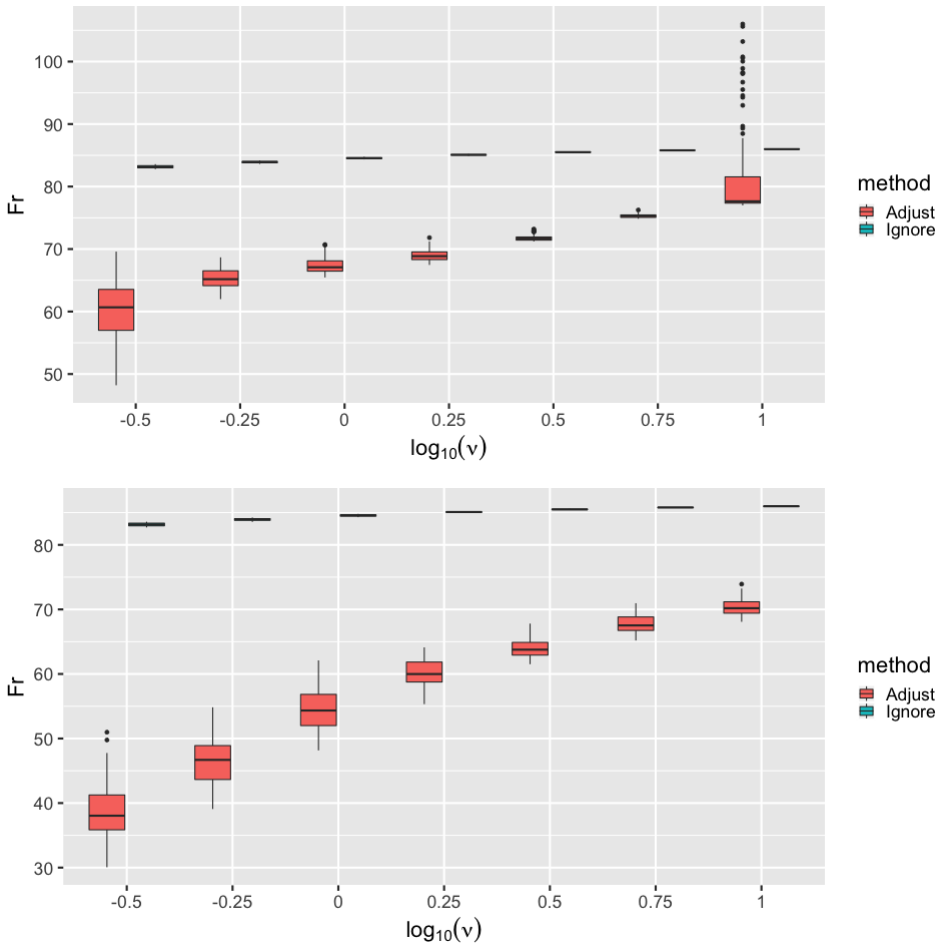

Figure 4.1 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(1) model over different magnitude of measurement error following the Poisson distribution.

large variance disappears. Note that the effect of adjusting for the Poisson measurement error is larger than the former ones, while part of the contribution is thanks to the different initial values. Because of the nonlinear corruption by the Poisson measurement error, ignoring it will further harm the estimation compared with the additive measurement errors. Comparing the size of the boxes, we discover that the estimation error variances of the adjusting method is larger than the ones of the baseline model, which is natural since we consider the extra layer in the hierarchical model to adjust for the measurement error. Similar results can be obtained in the other three structures in Figures 4.2, 4.3 and 4.4, which indicates that adjusting for the measurement error is necessary and beneficial when it is present.

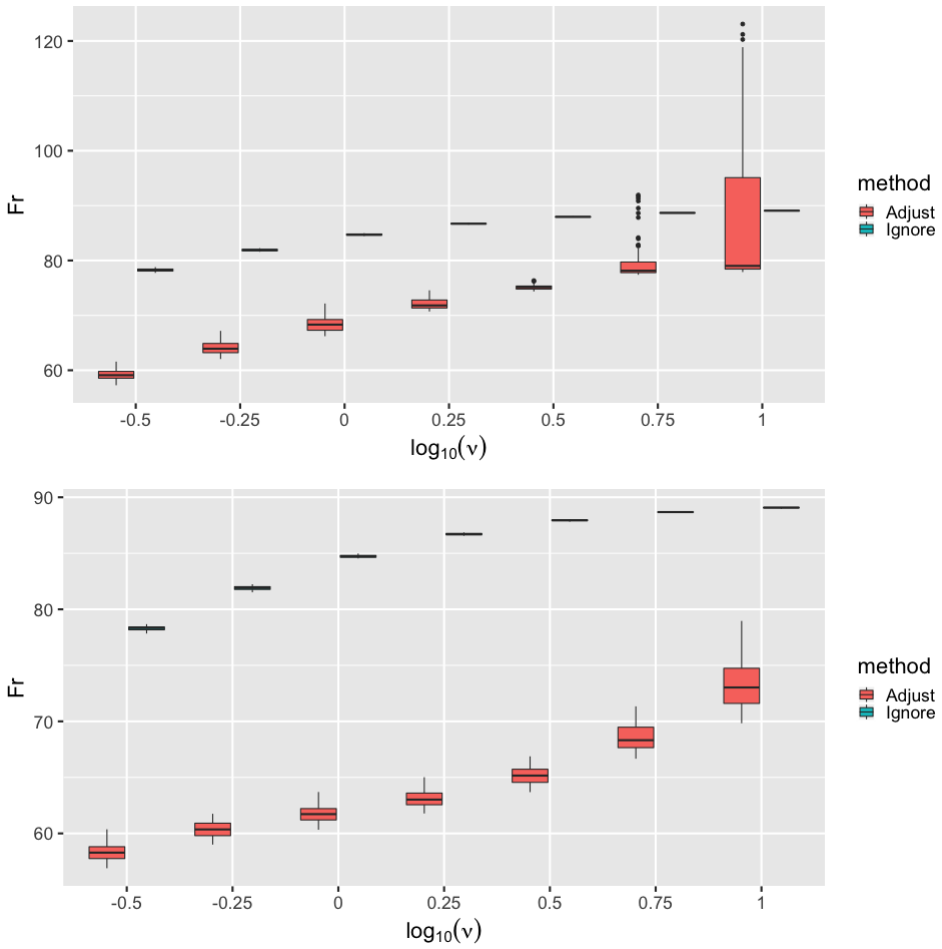

Figure 4.2 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the AR(2) model over different magnitude of measurement error following the Poisson distribution.

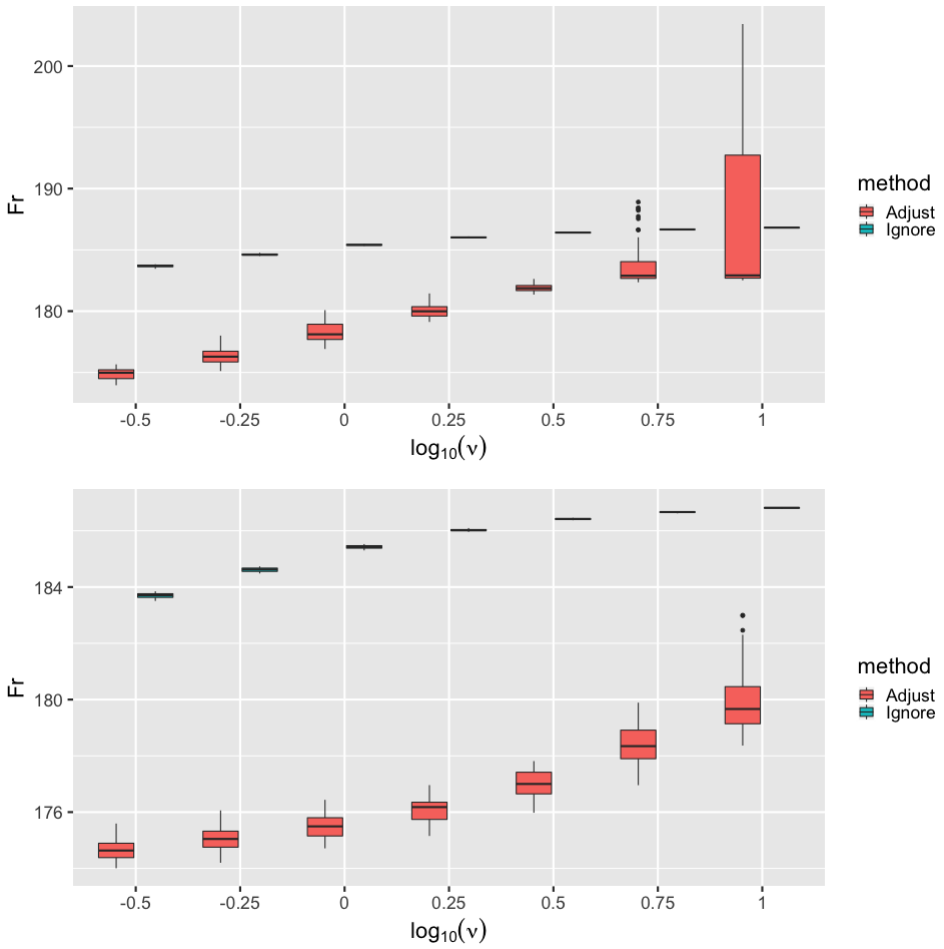

Figure 4.3 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathsf{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathsf{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(2) model over different magnitude of measurement error following the Poisson distribution.

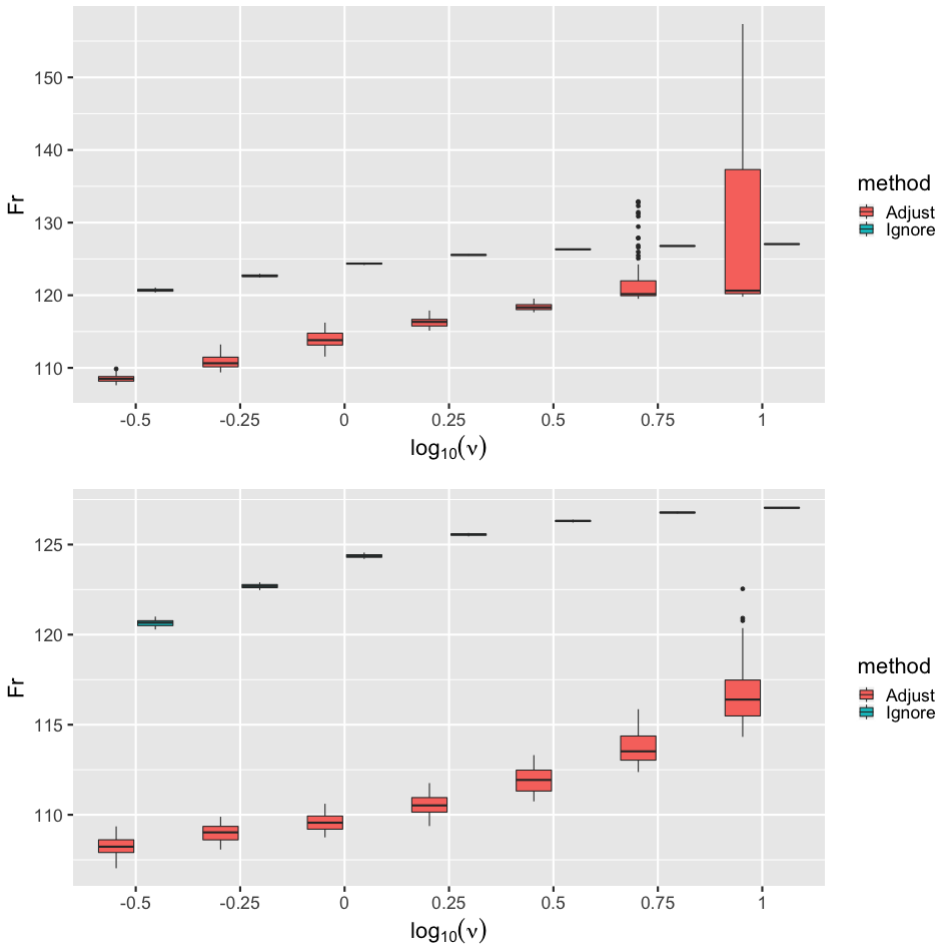

Figure 4.4 Boxplots of the estimation error in terms of the Frobenius norm using the proposal distribution $\mathrm{N}_p(\cdot, \nu I_p/10)$ (top) and $\mathrm{N}_p(\cdot, \nu I_p/100)$ (bottom) in the Block(5) model over different magnitude of measurement error following the Poisson distribution.

CHAPTER

5

CONCLUSIONS AND OPEN PROBLEMS

5.1 Overview

There are still some open questions of Gaussian graphical model that are of much interest, especially in theory, while some of these questions are not necessarily involved with measurement error. Therefore, we will highlight some concluding remarks of the results in the former chapters in Section 5.2 and then outline three interesting open problems. In Section 5.3, we discuss the selection consistency of the Gaussian graphical model, which is lack of a formal theory even in the no-measurement-error case. Section 5.4 contains some extensions of the work about the generalized measurement error as discussed in Chapter 4, particularly in theory. In Section 5.5, we consider the nonparanormal graphical model under measurement error as an extension of the Gaussian graphical model and introduce a Bayesian method to estimate the precision matrix, when the measurement error is Gaussian.

5.2 Conclusions

The Gaussian graphical models under different types of measurement error with known scale parameter have been investigated with the theoretical and numerical results. When the Gaussian measurement error is involved, we propose a Bayesian framework to adjust for the measurement error and estimate the precision matrix by applying the existing methods. The framework is rather simple and easy to apply. It is remarkable to notice that the posterior contraction rate remains unchanged as the rate in the no-measurement-error case if the same prior for the precision matrix is used. Considering the additive measurement error, we revise the former theoretical result and com-

puting algorithm to adjust for it. However, some extra conditions are required to maintain the same posterior contraction rate. A particular one needs the scale parameter to converge to 0 faster than a certain rate, which is reasonable to accommodate the more general distribution. A new sampling technique is proposed using the Metropolis-Hastings algorithm to adjust for the measurement error. The same computation method can be implemented to adjust for the generalized measurement error and estimate the precision matrix, which serves as the kernel of a mixture with the Gaussian graphical model. However, with such an arbitrary choice of the measurement error model, the posterior contraction rate originally under the Frobenius norm in the no-measurement-error case is preserved now with respect to the Rényi divergence. Such a semi-metric measures the prediction error between the estimated model and the true model instead of the estimation error. Several simulation studies highlight the empirical benefits of adjusting for the three types of measurement errors as opposed to ignoring it, respectively.

5.3 Selection Consistency

As discussed in Section 3.3.3, the accuracy of selection is poor when the measurement error is present, especially when the structure is dense. In Gaussian graphical model, an accurate selection result is sometimes more desirable since it reveals the whole structure of the graph. For example, the neurologists are interested in whether the two neurons are conditionally correlated and the strength of that correlation is secondary. Intuitively, we expect that the selection result is consistent to the true graph as the sample size increases. However, to the best of our knowledge, this theoretical problem of selection consistency by a Bayesian method has not been satisfactorily resolved even in the no-measurement-error case, while some literature already started to explore its finite-sample performance by simulation (Banerjee & Ghosal, 2015).

5.4 Generalized Measurement Error

For the Gaussian graphical model under the generalized measurement error, even though the contraction rate is presented in Theorem 4.2.1, it is a bit disappointing that the metric is in terms of densities rather than precision matrices. Intuitively, if ν is small enough, the estimation error should decay to 0 in the desired rate, no matter what the distribution of the measurement error, just like the results in Theorem 3.2.1 and 3.2.2. Therefore, it would be more satisfactory to establish the posterior contraction rate result with respect to the Frobenius norm of the generalized measurement error model with some additional conditions.

A reasonable control of the spread of the measurement error distribution could be helpful, for example, by letting ν decay to 0 in Theorem 3.2.2. With a small enough ν, we know that the measurement error distribution approximates to a Gaussian distribution by the central limit theorem. Furthermore, the corrupted observations Y can be transformed to center at X instead of $f_1(X)$ by the inverse transformation or an approximation of it using the delta method. However, there is

another obstacle that the variance of the generalized measurement error is involved by the true outcome X's, which will be kept after applying the Gaussian approximation and delta method. A common example of such distribution is the Poisson distribution, whose variance is equal to its mean, while most distributions from the exponential family have this property. To handle this difficulty, the variance-stabilizing transformation might be useful in some specific cases, while a general method is still lacking. If the variance after some transformation is irrelative to X's, this generalized measurement error is equivalent to the additive measurement error and the techniques in Chapter 3 can be applied.

5.5 Nonparanormal Graphical Model

5.5.1 Introduction

For the Gaussian graphical model under some measurement error as introduced in the previous chapters, we assume that the kernel function is mixed with the multivariate Gaussian distribution and aim to estimate the precision matrix. Sometimes, the Gaussian distributional assumption may not hold and a more general extension for describing the graphical model is preferred. Such an extension is achieved by assuming that the true outcomes are generated from the nonparanormal distribution. Then, the former problem could be extended to the nonparanormal graphical model under measurement error. However, to the best of our knowledge, no method has been established for inference on the nonparanormal graphical model while adjusting for the measurement error, even in the simplest case with Gaussian measurement error under the high-dimensional setup.

Nonparanormal distribution is a nonparametric extension of the Gaussian distribution, where the outcome X is assumed to have a Gaussian distribution only after a smooth monotone transformation h, i.e., $h(X) = (h_1(X_1), \ldots, h_p(X_p)) \sim \mathsf{N}_p(\mu, \Omega^{-1})$. Sometimes, h is assumed to be known. But more generally, it is unspecified and commonly need to be estimated through the nonparametric approaches. By adding this transformation to the Gaussian graphical model, Liu, Lafferty & Wasserman (2009) designed the nonparanormal graphical model and proposed a two-step method for estimating the precision matrix Ω, which first estimates h by a truncated empirical distribution function and then estimates Ω using the graphical LASSO. The copula Gaussian graphical model is another approach, which uses rank-based methods to transform the observed variables instead of estimating h and use the transformed data to make the inference (Dobra & Lenkoski, 2011; Liu et al., 2012; Pitt, Chan & Kohn, 2006). Mulgrave & Ghosal (2020) proposed a Bayesian method, which uses a finite B-spline basis on h and equips it with a truncated multivariate Gaussian distribution to guarantee the monotonicity. After the prior of h is chosen, the t-spike-and-slab prior is applied for the inference on the precision matrix.

However, all these approaches assume that the observations are accurately measured with desired precision. Intuitively, the measurement error may negatively affect the accuracy of estimation when its scale is large. One of such examples is discussed in Section 1.2.2.2 for the Gaussian graphical model. Since the nonparanormal graphical model is more complicated, the influence of

this corruption is supposed to be more profound and non-ignorable. Then, our question is natural: *how can we adjust for the measurement error in the nonparanormal graphical model?* Since the nonparanormal graphical model is rather complicated, we only focus on the Gaussian measurement error here. But the strategy to adjust for other types of measurement error is similar by referring to the aforementioned chapters and changing the method accordingly.

Suppose that for the "corrupted" observations $X_1, \ldots, X_n$, there are smooth monotonically increasing functions $h = (h_1, \ldots, h_p)$ such that

$$Y_i = h(X_i) \mid Z_i \overset{\text{ind}}{\sim} \mathsf{N}_p(Z_i, \nu I_p), \quad Z_i \overset{\text{iid}}{\sim} \mathsf{N}_p(\mu, \Omega^{-1}), \quad i = 1, \ldots, n, \tag{5.5.1}$$

where $h(X_i) = (h_1(X_{i,1}), \ldots, h_p(X_{i,p}))$, with corresponding precision matrix $\Omega = \Sigma^{-1}$. For the sake of simplicity, assume that the scale of measurement error ν is known throughout. Let $\Omega_\nu = \left(\Omega^{-1} + \nu I_p\right)^{-1}$ and Y_i's are available in closed-form based on the normal-normal conjugacy given by

$$Y_i \overset{\text{iid}}{\sim} \mathsf{N}_p\left(\mu, \Omega_\nu^{-1}\right), \quad i = 1, \ldots, n. \tag{5.5.2}$$

By taking the inverse of h on the Y's, the observations X's are generated, which are used to estimate μ and Ω. Given the conditions of the location and scale as introduced in Section 5.5.2, the identifiability of μ and Ω is guaranteed by the one-to-one transformation h and the known ν, which transfer the information of the mean and the dependence structure to the observations. If h is known, the problem becomes the "corrupted" Gaussian graphical model by taking the transformation h of the observations, which has already been addressed in Chapter 2. On the other hand, if the distribution of X given Z is considered as a whole as the measurement error model given h, this "corrupted" nonparanormal graphical model becomes another special case of the Gaussian graphical model under generalized measurement error. Note that for any continuous random variable, there is always a strictly increasing transformation to transform a Gaussian variable into that. Therefore, this "corrupted" nonparanormal graphical model even with known h considerably expands the border of the distributions for the Gaussian graphical model under measurement error.

The remaining content of this section is organized as follows. In Section 5.5.2, we propose the priors for estimating the precision matrix of the nonparanormal graphical model under Gaussian measurement error and the posterior computation is introduced in Section 5.5.3. Some of the results, especially the part to estimate the unknown transformation h, are inspired by Mulgrave & Ghosal (2020). In Section 5.5.4, we discuss some potential directions for further research according to this nonparametric modification.

5.5.2 Priors

Since the data from model (5.5.1) is produced by three steps (the transformation h, the corruption by ν and the randomness from the original Gaussian distribution), the approach for estimating ν and Ω is divided by these three components. For the transformation h, it is not enough to identify μ and Ω by only assuming that h is monotonically increasing, since a linear transformation of h can

produce another valid estimations of μ and Ω. Therefore, we follow the assumptions of Mulgrave & Ghosal (2020) and require the following condition on h.

Condition 5.5.1. *To guarantee the identifiability of μ and Ω, assume that*

$$h_d(1/2)=0, \quad h_d(3/4)-h_d(1/4)=1, \quad d=1,\ldots,p.$$

This condition restricts the location and scale of the transformation h, which guarantees the uniqueness of the estimation for the tuple (h,μ,Ω) given the observations $X_1,\ldots,X_n$. According to this assumption, we specify the priors for h, μ, and Ω separately.

Following the idea of Mulgrave & Ghosal (2020), the B-spline basis functions are considered to estimate the function h nonparametrically, and the model (5.5.2) is expressed as

$$h(X)=\sum_{j=1}^{J}\theta_j B_j(X) \overset{\text{iid}}{\sim} \mathsf{N}_p\left(\mu,\Omega_\nu^{-1}\right), \tag{5.5.3}$$

where $B_j(\cdot)$ are the B-spline basis functions, θ_j are the p corresponding coefficients in the expansion, and J is the number of B-spline basis functions. The number of basis functions J is selected by the Akaike Information Criterion (AIC).

Prior on θ. To obtain the estimation of the coefficients, let θ_d denote the J-dimensional vector of the coefficients of the B-spline basis for each of the p dimensions and consider the multivariate Gaussian prior, that is,

$$\theta_d \overset{\text{iid}}{\sim} \mathsf{N}_J\left(\zeta,\sigma^2 I_J\right), \quad d=1,\ldots,p, \tag{5.5.4}$$

where ζ and σ^2 are some pre-specified hyper-parameters. Mulgrave & Ghosal (2020) suggested a choice for ζ involving two hyper-parameters as

$$\zeta_j=\zeta_0+\sigma_0\Phi_0^{-1}\left(\frac{j-0.375}{J-0.75+1}\right), \quad j=1,\ldots,J, \tag{5.5.5}$$

where ζ_0 and $\sigma_0>0$ are some constants and Φ_0^{-1} denotes the quantile function of the standard Gaussian distribution.

However, the J-dimensional vector θ_d produced by the prior (5.5.4) is free of the Condition 5.5.1, which only has $J-2$ free parameters. To impose the Condition 5.5.1 on the prior of θ_d, consider the linear restriction

$$\sum_{j=1}^{J}\theta_{d,j}B_j(1/2)=0, \quad \sum_{j=1}^{J}\theta_{d,j}\left\{B_j(3/4)-B_j(1/4)\right\}=1, \quad d=1,\ldots,p,$$

which is denoted by $A\theta_d=c$ for the sake of simplicity. It is demonstrated in Mulgrave & Ghosal (2020) that the prior (5.5.4) conditioned on this restriction is equivalent to the following prior,

$$\overline{\theta}_d \mid \{A\theta_d=c\} \sim \mathsf{N}_{J-2}\left(\overline{\xi},\overline{\Phi}\right), \quad d=1,\ldots,p,$$

where $\xi = \zeta + A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}(c - A\zeta)$, $\Phi = \sigma^2\left(I - A^{\mathrm{T}}(AA^{\mathrm{T}})^{-1}A\right)$ and a bar denotes the $J-2$ dimension reduction, e.g. $\overline{\theta}_d = (\theta_{d,1}, \ldots, \theta_{d,J-2})$. For the rest two elements removed by the dimension reduction, we can compute them through the linear restriction $A\theta_d = c$ by the expression $\left(\theta_{d,J-1}, \theta_{d,J}\right)^{\mathrm{T}} = W_d\overline{\theta}_d + q_d$ with some $2\times(J-2)$-dimensional matrix W_d and 2-dimensional vector q_d.

On the other hand, h is assumed to be monotonically increasing by the definition of nonparanormal distribution, which the prior (5.5.4) should be conditioned on. For the monotonicity restriction, it is equivalent to the condition that the series of inequalities $\theta_{d,2} - \theta_{d,1} > 0, \ldots, \theta_{d,J} - \theta_{d,J-1} > 0$, or in the matrix form $F\theta_d > 0$ with the fixed $(J-1)\times J$-dimensional matrix F. Due to the Condition 5.5.1, it is equivalent to $\overline{F}\overline{\theta}_d + \overline{g} > 0$ for some $(J-1)\times(J-2)$-dimensional matrix $\overline{F}$ and $(J-2)$-length vector $\overline{g}$ calculated by A and c (Mulgrave & Ghosal, 2020).

After imposing the identifiability and monotonicity conditions, the final prior on θ is given by a truncated Gaussian distribution, that is,

$$\overline{\theta}_d \mid \{A\theta_d = c\} \overset{\text{iid}}{\sim} \mathsf{TN}_{J-2}(\overline{\xi}, \overline{\Phi}, \mathscr{T}),$$

where $\mathscr{T} = \{\overline{\theta}_d : \overline{F}\overline{\theta}_d + \overline{g} > 0\}$ and $\mathsf{TN}_J(\mu, \Sigma, \mathscr{T})$ denotes the $\mathsf{N}_J(\mu, \Sigma)$ restricted on a set $\mathscr{T}$.

Prior on μ. The Jefferys improper prior is considered for the prior on μ, that is,

$$\pi(\mu) \propto 1.$$

Prior on Ω. All the priors for Ω in the context of structure learning of the Gaussian graphical model could be deployed to estimation the precision matrix, such as the examples introduced in Section 2.3 and 2.4. Since κ is relatively small in most practical problems, it can be ignored for numerical purposes for simplicity. The only exception is the factor-model structure, where the κ originally stays. Besides these priors, Mulgrave & Ghosal (2020) proposed the t-spike-and-slab prior for all $j < i$, $i = 1, \ldots, n$, that is,

$$\begin{aligned} &(\Omega_{i,j} \mid \Gamma_{i,j}, \tau_{i,j}^2) \sim (1-\Gamma_{i,j})\mathsf{N}(0, c_0\tau_{i,j}^2) + \Gamma_{i,j}\mathsf{N}(0, \tau_{i,j}^2), \quad \Omega_{i,i} \sim \mathsf{Exp}(1/2), \\ &\Gamma_{i,j} \sim \mathsf{Bernoulli}(\pi), \quad \pi \sim \mathsf{Beta}(1, 10), \quad \tau_{i,j}^2 \sim \mathsf{IG}(b_0, b_1), \end{aligned} \tag{5.5.6}$$

where c_0, b_0 and b_1 are hyper-parameters, which are suggested to be selected by BIC and $\Gamma_{i,j} = \mathbb{1}(\Omega_{i,j} \neq 0)$. The first expression in (5.5.6) is equivalent to

$$(\Omega_{i,j} \mid \Gamma_{i,j}, \tau_{i,j}^2) \sim \mathsf{N}(0, V_{i,j}), \quad V_{i,j} = c_0(1-\Gamma_{i,j})\tau_{i,j}^2 + \Gamma_{i,j}\tau_{i,j}^2.$$

By integrating out $\tau_{i,j}$, we get that $(\Omega_{i,j} \mid \Gamma_{i,j})$ has a t-distribution, which makes the prior more robust.

5.5.3 Posterior Computation

Given the priors for θ, μ and Ω as described above, the full posterior distribution is

$$\begin{aligned}&\Pi_n^{\nu}(\theta,\Omega,\mu \mid X)\\ &\quad\propto (\det\Omega)^{n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\big(B(X_i)\theta-\mu\big)^{\mathrm{T}}\big(\Omega^{-1}+\nu I\big)^{-1}\big(B(X_i)\theta-\mu\big)\right\}\times\pi\big(\overline{\theta}\big)\times\pi\big(\mu\big)\times\pi(\Omega).\end{aligned}$$

The calculation of this joint posterior distribution is standard by Bayes' theorem. The following Gibbs sampling algorithm is deployed to evaluate the posterior and produce the estimations.

The whole algorithm can be divided into three steps in each iteration respectively for producing the samples of θ, μ and Ω. These three steps are sequentially iterated until convergence and the posterior samples after the burn-in are used to compute the estimations.

First step. Sample the B-spline coefficients using the exact Hamiltonian Monte Carlo algorithm (Pakman & Paninski, 2014) for every $d=1,\ldots,p$:

$$\begin{aligned}&\pi\big(\overline{\theta}_d \mid Y,\overline{\theta}_{\{1,\ldots,p\}\setminus d},\mu,\Omega\big)\\ &\quad\propto\exp\Bigg[-\frac{1}{2}\overline{\theta}_d^{\mathrm{T}}\left\{\frac{1}{\lambda_d^2}\sum_{i=1}^{n}\big(\overline{B}+\widetilde{B}W_d\big)^{\mathrm{T}}\big(\overline{B}+\widetilde{B}W_d\big)+\overline{\Gamma}^{-1}\right\}\overline{\theta}_d\\ &\qquad+\left\{\overline{\xi}\overline{\Gamma}^{-1}-\frac{1}{\lambda_d^2}\sum_{i=1}^{n}\big(\widetilde{B}q_d-\delta_{d,i}\big)^{\mathrm{T}}\big(\overline{B}+\widetilde{B}W_d\big)\right\}\overline{\theta}_d\Bigg]\mathbb{1}\big(\overline{F}_d\overline{\theta}_d+\overline{g}_d>0\big)\end{aligned}$$

where $\widetilde{B}=\big(B_{J-1}(X_i),B_J(X_i)\big)$ is the reminder after the dimension reduction, $\lambda_d^2=1/\Omega_{\nu,d,d}$ and

$$\delta_{d,i}=\mu_d+\sum_{k\in(1:p\setminus d)}\left(-\frac{\Omega_{\nu,d,k}}{\Omega_{\nu,d,d}}\right)\big(Y_{i,k}-\mu_k\big).$$

Note that the precision matrix for sampling θ is the corrupted one, which is actually involved in (5.5.3).

Second step. The second step in each iteration is to produce the samples of μ and the centered unobservable Z's from the posterior by these three mini-steps:

(a) Compute $Y_{i,d}=\sum_{j=1}^{J}\theta_{d,j}B_j\big(X_{i,d}\big)$;

(b) Sample $\mu\mid(Y,\Omega)\sim\mathsf{N}_p\big(\frac{1}{n}\sum_{i=1}^{n}Y_i,\frac{1}{n}\Omega_{\nu}^{-1}\big)$;

(c) Sample $Z_i\mid(Y_i,\Omega)\sim\mathsf{N}_p\big((I_p+\nu\Omega)^{-1}(Y_i-\mu),\nu(I_p+\nu\Omega)^{-1}\big)$.

Note that aside of the centering, the step (c) is identical to (2.5.1), where the data augmentation is also deployed to produce the uncorrupted samples from the posterior.

Third step. Since now the Z's are centered, which only carry the information of the true precision matrix, the sampling algorithm for the corresponding prior can be employed as the third step to produce the posterior samples for Ω just like (2.5.2). One particular example is the priors using the

Cholesky decomposition that is deployed in the former chapters, whose Gibbs sampling algorithm is reviewed in Section 1.2.1.2.6. If the t-spike-and-slab prior is considered, we follow these four mini-steps to produce the posterior samples of Ω (Mulgrave & Ghosal, 2020):

(a) For every $d = 1, \ldots, p$, sample each column vector of Ω by

$$(\Omega_{-d,d} \mid -) \sim \mathsf{N}_{p-1}\left(-CS_{-d,d}, C\right), (u \mid -) \sim \mathsf{Gamma}\left(\frac{n}{2}+1, \frac{S_{d,d}+1}{2}\right),$$

where $S = Z^{\mathrm{T}}Z$, $C = \left\{\left(S_{d,d}+1\right)\Omega^{-1}_{-d,-d} + \mathrm{diag}\left(V_{-d,d}\right)^{-1}\right\}^{-1}$ and $u = \Omega_{d,d} - \Omega^{\mathrm{T}}_{-d,d}\Omega^{-1}_{-d,-d}\Omega_{-d,d}$. The negative subscripts mean that the corresponding row/column is not selected;

(b) For every $d = 1, \ldots, p$, sample each column vector of Γ by

$$P\left(\Gamma_{d,k} = 1 \mid -\right) = \frac{\phi\left(\Omega_{d,k} \mid 0, \tau^2_{d,k}\right)\pi}{\phi\left(\Omega_{d,k} \mid 0, \tau^2_{d,k}\right)\pi + \phi\left(\Omega_{d,k} \mid 0, c_0\tau^2_{d,k}\right)(1-\pi)},$$

where ϕ stands for the normal density function;

(c) For every $k = 1, \ldots, p$ and $d < k$, update $\tau^2_{d,k}$ by

$$\left(\tau^2_{d,k} \mid -\right) \sim \mathsf{IG}\left(b_0 + \frac{1}{2}, b_1 + \frac{\Omega^2_{d,k}}{2}\left(\Gamma_{d,k} + \frac{1-\Gamma_{d,k}}{c_0}\right)\right);$$

(d) Update π based on the off-diagonal entry $\Gamma_{d,k}$

$$(\pi \mid -) \sim \mathsf{Beta}\left(1 + \sum_{d<k} \mathbb{1}\left(\Gamma_{d,k} = 1\right), 10 + \sum_{d<k} \mathbb{1}\left(\Gamma_{d,k} = 0\right)\right).$$

5.5.4 Discussion

Even though the proposed method should intuitively produce a more accurate estimation of the true precision matrix compared with the method ignoring the measurement error, it is lack of a theoretical foundation to demonstrate the consistency or the convergence rate. Mulgrave & Ghosal (2020) proved the posterior consistency of their method without the measurement error, which provides a start point for the theoretical work. Regarding to the posterior contraction rate, there is no existence of such a theory even in the no-measurement error case. Thus, it is desirable to establish some theoretically large-sample results about the nonparanormal graphical model in the presence of measurement error, like the estimation consistency, the posterior contraction rate or even the selection consistency.

BIBLIOGRAPHY

Banerjee, O., Ghaoui, L. E. & d'Aspremont, A. (2008). "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data". *Journal of Machine Learning Research* **9**.Mar, pp. 485–516.

Banerjee, S. (2017). "Posterior convergence rates for high-dimensional precision matrix estimation using G-Wishart priors". *Stat* **6**.1, pp. 207–217.

Banerjee, S. & Ghosal, S. (2014). "Posterior convergence rates for estimating large precision matrices using graphical models". *Electronic Journal of Statistics* **8**.2, pp. 2111–2137.

— (2015). "Bayesian structure learning in graphical models". *Journal of Multivariate Analysis* **136**, pp. 147–162.

Bartholomew, D. J., Knott, M. & Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Vol. 904. John Wiley & Sons.

Byrd, M., Nghiem, L. H. & McGee, M. (2021). "Bayesian regularization of Gaussian graphical models with measurement error". *Computational Statistics & Data Analysis* **156**, p. 107085.

Cai, T. T., Liu, W. & Zhou, H. H. (2016). "Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation". *The Annals of Statistics* **44**.2, pp. 455–488.

Cai, T., Liu, W. & Luo, X. (2011). "A constrained ℓ_1 minimization approach to sparse precision matrix estimation". *Journal of the American Statistical Association* **106**.494, pp. 594–607.

Candes, E. & Tao, T. (2007). "The Dantzig selector: Statistical estimation when p is much larger than n". *The Annals of Statistics* **35**.6, pp. 2313–2351.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. CRC press.

Carvalho, C. M., Polson, N. G. & Scott, J. G. (2010). "The horseshoe estimator for sparse signals". *Biometrika* **97**.2, pp. 465–480.

Castillo, I. & van der Vaart, A. (2012). "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences". *The Annals of Statistics* **40**.4, pp. 2069–2101.

Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*. Vol. 95. CRC Press.

Cook, J. R. & Stefanski, L. A. (1994). "Simulation-extrapolation estimation in parametric measurement error models". *Journal of the American Statistical Association* **89**.428, pp. 1314–1328.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **39**.1, pp. 1–22.

Diebolt, J. & Robert, C. P. (1994). “Estimation of finite mixture distributions through Bayesian sampling”. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **56**.2, pp. 363–375.

Dobra, A. & Lenkoski, A. (2011). “Copula Gaussian graphical models and their application to modeling functional disability data”. *The Annals of Applied Statistics* **5**.2A, pp. 969–993.

Drton, M. & Maathuis, M. H. (2017). “Structure learning in graphical modeling”. *Annual Review of Statistics and Its Application* **4**, pp. 365–393.

Du, X. & Ghosal, S. (2018). “Bayesian discriminant analysis using a high dimensional predictor”. *Sankhya A. The Indian Journal of Statistics* **80**.1, suppl. S112–S145.

Fan, J., Fan, Y. & Lv, J. (2008). “High dimensional covariance matrix estimation using a factor model”. *Journal of Econometrics* **147**.1, pp. 186–197.

Fan, J., Feng, Y. & Wu, Y. (2009). “Network exploration via the adaptive LASSO and SCAD penalties”. *The Annals of Applied Statistics* **3**.2, p. 521.

Fan, J., Han, F. & Liu, H. (2014). “Challenges of big data analysis”. *National science review* **1**.2, pp. 293–314.

Fan, J. & Li, R. (2001). “Variable selection via nonconcave penalized likelihood and its oracle properties”. *Journal of the American Statistical Association* **96**.456, pp. 1348–1360.

Fan, J., Liao, Y. & Mincheva, M. (2011). “High dimensional covariance matrix estimation in approximate factor models”. *The Annals of Statistics* **39**.6, p. 3320.

Freedman, L. S., Midthune, D., Carroll, R. J. & Kipnis, V. (2008). “A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression”. *Statistics in Medicine* **27**.25, pp. 5195–5216.

Friedman, J., Hastie, T. & Tibshirani, R. (2008). “Sparse inverse covariance estimation with graphical lasso”. *Biostatistics* **9**.3, pp. 432–441.

Friedman, N. (2004). “Inferring cellular networks using probabilistic graphical models”. *Science* **303**.5659, pp. 799–805.

Fuller, W. A. (2009). *Measurement Error Models*. Vol. 305. John Wiley & Sons.

Gan, L., Narisetty, N. N. & Liang, F. (2019). “Bayesian regularization for graphical models with unequal shrinkage”. *Journal of the American Statistical Association* **114**.527, pp. 1218–1231.

Ghosal, S. (2001). “Convergence rates for density estimation with Bernstein polynomials”. *The Annals of Statistics* **29**.5, pp. 1264–1280.

Ghosal, S., Ghosh, J. K. & van der Vaart, A. (2000). “Convergence rates of posterior distributions”. *The Annals of Statistics* **28**.2, pp. 500–531.

Ghosal, S. & van der Vaart, A. (2007). “Convergence rates of posterior distributions for non-i.i.d. observations”. *The Annals of Statistics* **35**.1, pp. 192–223.

Ghosal, S. & van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Vol. 44. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xxiv+646.

Jeong, S. & Ghosal, S. (2021). "Unified Bayesian theory of sparse linear regression with nuisance parameters". *Electronic Journal of Statistics* **15**.1, pp. 3040–3111.

Jonge, R. de & Zanten, J. H. van (2010). "Adaptive nonparametric Bayesian inference using location-scale mixture priors". *The Annals of Statistics* **38**.6, pp. 3300–3320.

Jørgensen, B. (1987). "Exponential dispersion models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **49**.2, pp. 127–145.

Kruijer, W., Rousseau, J. & van der Vaart, A. (2010). "Adaptive Bayesian density estimation with location-scale mixtures". *Electronic Journal of Statistics* **4**, pp. 1225–1257.

Lam, C. & Fan, J. (2009). "Sparsistency and rates of convergence in large covariance matrix estimation". *The Annals of Statistics* **37**.6B, pp. 4254–4278.

Lauritzen, S. L. (1996). *Graphical Models*. Vol. 17. Oxford Statistical Science Series. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York.

Lavine, M. & West, M. (1992). "A Bayesian method for classification and discrimination". *Canadian Journal of Statistics* **20**.4, pp. 451–461.

Lenkoski, A. (2013). "A direct sampler for G-Wishart variates". *Stat* **2**.1, pp. 119–128.

Lenkoski, A. & Dobra, A. (2011). "Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior". *Journal of Computational and Graphical Statistics* **20**.1, pp. 140–157.

Li, Y., Craig, B. A. & Bhadra, A. (2019). "The graphical horseshoe estimator for inverse covariance matrices". *Journal of Computational and Graphical Statistics* **28**.3, pp. 747–757.

Liang, F., Jia, B., Xue, J., Li, Q. & Luo, Y. (2018). "An imputation–regularized optimization algorithm for high dimensional missing data problems and beyond". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **80**.5, p. 899.

Liu, C. & Martin, R. (2019). "An empirical G-Wishart prior for sparse high-dimensional Gaussian graphical models". Unpublished manuscript, `arXiv:1912.03807`.

Liu, H., Han, F., Yuan, M., Lafferty, J. & Wasserman, L. (2012). "High-dimensional semiparametric Gaussian copula graphical models". *The Annals of Statistics* **40**.4, pp. 2293–2326.

Liu, H., Lafferty, J. & Wasserman, L. (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research* **10**.10.

McLachlan, G. J. & Peel, D. (1998). "Robust cluster analysis via mixtures of multivariate t-distributions". *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, pp. 658–666.

Meinshausen, N. & Bühlmann, P. (2006). "High-dimensional graphs and variable selection with the lasso". *The Annals of Statistics* **34**.3, pp. 1436–1462.

Mohammadi, A. & Wit, E. C. (2015). "Bayesian structure learning in sparse Gaussian graphical models". *Bayesian Analysis* **10**.1, pp. 109–138.

Mulgrave, J. J. & Ghosal, S. (2020). "Bayesian inference in nonparanormal graphical models". *Bayesian Analysis* **15**.2, pp. 449–475.

Ning, B., Jeong, S. & Ghosal, S. (2020). "Bayesian linear regression for multivariate responses under group sparsity". *Bernoulli* **26**.3, pp. 2353–2382.

Pakman, A. & Paninski, L. (2014). "Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians". *Journal of Computational and Graphical Statistics* **23**.2, pp. 518–542.

Pan, W. & Shen, X. (2007). "Penalized model-based clustering with application to variable selection". *Journal of Machine Learning Research* **8**.May, pp. 1145–1164.

Park, T. & Casella, G. (2008). "The Bayesian lasso". *Journal of the American Statistical Association* **103**.482, pp. 681–686.

Pati, D., Bhattacharya, A., Pillai, N. S. & Dunson, D. (2014). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices". *The Annals of Statistics* **42**.3, pp. 1102–1130.

Pearson, K. (1894). "Contributions to the mathematical theory of evolution". *Philosophical Transactions of the Royal Society of London. A* **185**, pp. 71–110.

Peng, J., Wang, P., Zhou, N. & Zhu, J. (2009). "Partial correlation estimation by joint sparse regression models". *Journal of the American Statistical Association* **104**.486, pp. 735–746.

Pitt, M., Chan, D. & Kohn, R. (2006). "Efficient Bayesian inference for Gaussian copula regression models". *Biometrika* **93**.3, pp. 537–554.

Ravikumar, P., Wainwright, M. J., Raskutti, G. & Yu, B. (2011). "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence". *Electronic Journal of Statistics* **5**, pp. 935–980.

Rothman, A. J., Bickel, P. J., Levina, E. & Zhu, J. (2008). "Sparse permutation invariant covariance estimation". *Electronic Journal of Statistics* **2**, pp. 494–515.

Rousseau, J. (2010). "Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density". *The Annals of Statistics* **38**.1, pp. 146–180.

Roverato, A. (2000). "Cholesky decomposition of a hyper inverse Wishart matrix". *Biometrika* **87**.1, pp. 99–112.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**.1, pp. 267–288.

Wang, H. (2012). "Bayesian graphical lasso models and efficient posterior computation". *Bayesian Analysis* **7**.4, pp. 867–886.

Wang, H., Banerjee, A., Hsieh, C.-J., Ravikumar, P. K. & Dhillon, I. S. (2013). “Large scale distributed sparse precision estimation”. *Advances in Neural Information Processing Systems*, pp. 584–592.

Wiper, M., Insua, D. R. & Ruggeri, F. (2001). “Mixtures of gamma distributions with applications”. *Journal of Computational and Graphical Statistics* **10**.3, pp. 440–454.

Witten, D. M. & Tibshirani, R. (2010). “A framework for feature selection in clustering”. *Journal of the American Statistical Association* **105**.490, pp. 713–726.

Xiang, R., Khare, K. & Ghosh, M. (2015). “High dimensional posterior convergence rates for decomposable graphical models”. *Electronic Journal of Statistics* **9**.2, pp. 2828–2854.

Yoo, W. W. & Ghosal, S. (2016). “Supremum norm posterior contraction and credible sets for nonparametric multivariate regression”. *The Annals of Statistics* **44**.3, pp. 1069–1102.

Yuan, M. (2010). “High dimensional inverse covariance matrix estimation via linear programming”. *Journal of Machine Learning Research* **11**.Aug, pp. 2261–2286.

Yuan, M. & Lin, Y. (2007). “Model selection and estimation in the Gaussian graphical model”. *Biometrika* **94**.1, pp. 19–35.