Electronic Journal of Statistics
Vol. 0 (0000) 1
ISSN: 1935-7524
DOI: 10.1214/154957804100000000

Bayesian estimation of sparse precision matrices in the presence of Gaussian measurement error

Wenli Shi, Subhashis Ghosal* **and Ryan Martin**†

Department of Statistics
North Carolina State University
Raleigh, NC 27695, U.S.A.
e-mail: wshi4@ncsu.edu*;* sghosal@ncsu.edu*;* rgmarti3@ncsu.edu

Abstract: Estimation of sparse, high-dimensional precision matrices is an important and challenging problem. Existing methods all assume that 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. The present paper incorporates measurement error in the context of estimating a sparse, high-dimensional precision matrix. In particular, for a Gaussian graphical model with data corrupted by Gaussian measurement error with unknown variance, we establish a general result which gives sufficient conditions under which the posterior contraction rates that hold in the no-measurement-error case carry over to the measurement-error case. Interestingly, this result does not require that the measurement error variance be small. We apply our 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. Two different simulation studies highlight the empirical benefits of accounting for the measurement error as opposed to ignoring it, even when that measurement error is relatively small.

Keywords and phrases: high-dimensional inference, Gaussian graphical model, measurement error, posterior contraction rate, sparsity.

1. Introduction

The precision matrix, namely, the inverse of the covariance matrix of a Gaussian random vector, is a key object in multivariate analysis because of its role in describing conditional distributions. Let there be observations $X_1, \ldots, X_n$, independent and identically distributed (i.i.d.) from a p-dimensional, mean-zero Gaussian distribution, $\mathsf{N}_p(0, \Sigma)$, where Σ denotes a $p \times p$ positive definite covariance matrix, with corresponding precision matrix $\Omega = \Sigma^{-1}$. The goal is to make inference on the unknown Ω, especially in the high-dimensional situation when 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

*Research is partially supported by Faculty Research and Professional Development Grant from the College of Sciences, North Carolina State University and Army Research Office grant 76643–MA.

†Research is partially supported by U.S. National Science Foundation grant DMS–1811802.

is of the order of p^2, which can exceed 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 [24], it is common to assume that there are only a few intrinsic dependent relationships in the graph and the underlying graph describing the dependence structure is sparse, thus leading to a precision matrix Ω with many zeros on the off-diagonal. Therefore, the sparsity simultaneously simplifies the dependence structure and effectively reduces the dimension of Ω, potentially paving the way for accurate estimation. Regularization methods are often used to incorporate the intended sparse structure into the estimator. Yuan and Lin [33] and Banerjee, Ghaoui and d'Aspremont [1] 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 and Tibshirani [18]. Inspired by the desirable properties of the smoothly clipped absolute deviation (SCAD) penalty [15] which uses folded concave penalties to avoid the known problem of bias due to excessive shrinkage of large non-zero entries, Fan, Feng and Wu [14] proposed the graphical SCAD. Cai, Liu and Luo [6] designed a procedure based on the Dantzig selector [7]. The procedure 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. In the Bayesian literature, several priors were considered for a sparse precision matrix and resulting computational procedures were developed. Wang [32] developed 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 value zero with positive probability, a post-estimation thresholding mechanism is needed to learn the sparsity structure using Wang's method. Banerjee and Ghosal [3] 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 large sample approximations. Selection of the edges in the graph corresponding to selecting nonzero off-diagonal entries may be more important than the shrinkage and estimation. 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, the focus on strcture selection my be exploited. Lenkoski and Dobra [25] and Mohammadi and Wit [28] proposed useful computational methods that allows MCMC moves across possible graphs. Banerjee and Ghosal [2] assumed a banding structure on the precision matrix and derived the posterior contraction rate with a G-Wishart prior. Liu and Martin [27] proposed an empirical G-Wishart prior and demonstrated its optimal posterior contraction rate and strong performance in terms of computational speed and accuracy. Du and Ghosal [12] 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.

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 where 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 [11, 8, 17, 19], very little work has been done in the context of structured precision matrix estimation in the presence of Gaussian measurement errors. Byrd, Nghiem and McGee [5] assumed the variance of measurement error to be known, 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 [26] and Bayesian regularization for graphical models with unequal shrinkage [20] to formulate a new procedure and prove its consistency. Their results also revealed the necessity of adjusting for measurement error when present. Here we propose a fully Bayesian framework for handling measurement error and give general sufficient conditions for establishing the posterior contraction rate.

Our main goal in this paper is to understand the effect of measurement error on Bayesian methods for estimating a high-dimensional structured precision matrix of a multi-dimensional Gaussian random vector. We focus here on a Gaussian measurement error model, for $i = 1, \dots, n$ and $j = 1, \dots, m$,

$$Y_{ij} = X_i + Z_{ij}, \quad X_i \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \Omega^{-1}), \quad Z_{ij} \overset{\text{iid}}{\sim} \mathsf{N}_p(0, \nu I_p) \tag{1.1}$$

where the X and Z samples are mutually independent, I_p is the identity matrix of order p and m is the number of replicates for each X. 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. If ν is unknown, then $m \geq 2$ replicates for each outcome X is required to guarantee identifiability of ν and Ω. On the other hand, for the special case that ν is known, the convergence results hold also for $m = 1$. Let $Y_i = (Y_{i1}^{\mathrm{T}}, \dots, Y_{im}^{\mathrm{T}})^{\mathrm{T}}$. Then, the marginal distribution of the Y's is available in closed-form,

$$Y_i \overset{\text{iid}}{\sim} \mathsf{N}_{mp}(0, \Sigma_\nu), \quad i = 1, \dots, n, \tag{1.2}$$

where Σ_ν is an $mp \times mp$ block matrix, with $\Omega^{-1} + \nu I_p$ as the diagonal blocks and Ω^{-1} as the off-diagonal blocks. Write $\Omega_\nu = \Sigma_\nu^{-1}$, which is an $mp \times mp$ block matrix with $\nu^{-1}\{I_p - (\nu\Omega + mI_p)^{-1}\}$ as the diagonal blocks and $-\nu^{-1}(\nu\Omega + mI_p)^{-1}$ as the off-diagonal blocks.

In contrast to the covariance matrix, on which the effect of the measurement error is simply additive, the inverse $(\nu\Omega + mI_p)^{-1}$ from the above expression reveals how even the simple linear measurement error model leads to a very nonlinear 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 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 concentration 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 troublesome inverse in Ω_ν, it is crucial to have extra control on the prior distribution of the smallest eigenvalue of Ω. To address this, we express Ω as $\kappa I_p + \Theta$, and put independent priors on the scalar $\kappa > 0$ and the $p \times p$ matrix Θ. Then the prior on κ gives us a control over the lower eigenvalue of Ω, while the prior for Θ can be any of those from the literature.

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 and Lv [13], Fan, Liao and Mincheva [16] and Pati et al. [30] have used such a model, with Θ having a sparse factor structure [4], for a high-dimensional *covariance matrix*. To our knowledge, this prior formulation has not been developed for inference on a precision matrix. Our posterior concentration rate result simultaneously covers the measurement error and no-measurement-error cases, and the rate attained parallels that obtained by Pati et al. [30] for the covariance matrix with respect to the Frobenius norm, with some improvements.

The remainder of this paper is organized as follows. In Section 2, 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). Our general framework for incorporating Gaussian measurement error into existing Bayesian procedures for inference on structured, high-dimensional precision matrices is presented in Section 3 along with a general result on posterior contraction rates. 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. Examples of the obtained rates with measurement error based on priors commonly used in the no-measurement-error literature are discussed in Section 4. A new prior for the estimation of a precision matrix with a sparse factor structure is proposed and the corresponding posterior concentration rate is illustrated in Section 5. 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 6, showing that the adjustment towards to the measurement error leads to a lower estimation error in terms of the Frobenius norm. All proofs are presented in the appendix.

2. Effect of ignoring 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?* We show below that failing to account for the measurement error creates a large bias and, therefore, certain adjustments are necessary to account for the presence of measurement error and to ensure accurate estimation of Ω. For the sake of simplicity, we assume that ν is known and $m = 1$ throughout this section.

To develop some intuition, 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 sparsity assumptions. Let $\widehat{\Omega}_n = \widehat{\Omega}(Y_1, \ldots, Y_n)$ denote an asymptotically unbiased estimator of Ω based on the corrupted data $Y_1, \ldots, Y_n$, e.g., $\widehat{\Omega}_n = S_n^{-1}$, 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

$$\|\mathsf{E}_{\Omega^\star,\nu} \widehat{\Omega}_n - (\Omega^{\star -1} + \nu I_p)^{-1}\|_F = o(1), \quad n \to \infty, \tag{2.1}$$

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 2.1. *For a case of fixed dimension p, let $\widehat{\Omega}_n$ be an estimator that ignores the measurement error and satisfies (2.1). If ν is fixed and known, then for all large n,*

$$\mathsf{E}_{\Omega^\star,\nu} \|\widehat{\Omega}_n - \Omega^\star\|_F^2 \geq \tfrac{1}{2} \|\Omega^\star (\nu^{-1} I_p + \Omega^\star)^{-1} \Omega^\star\|_F^2. \tag{2.2}$$

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

$$\mathsf{E}_{\Omega^\star,\nu} \|\widehat{\Omega}_n - \Omega^\star\|_F^2 \geq \tfrac{1}{8} \nu^2 p \lambda_{\min}^4(\Omega^\star), \tag{2.3}$$

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

The proof is given in Appendix A. From the theorem, we can see 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, unknown ν and increasing p.

3. Accounting for measurement error

3.1. Prior and posterior distributions

For technical reasons that will be made clear below, when 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 to automatically satisfy the requirement, namely, adding a scalar multiple of the identity matrix to the precision matrix. That is, we express the precision matrix as

$$\Omega = \Theta + \kappa I_p, \tag{3.1}$$

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 independent prior distributions to Θ and κ. 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 6.1. Since ν is typically unknown, we assign it a prior distribution. Details about the specific priors for κ, Θ and ν 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

$$\Pi_\kappa(\kappa > t) + \Pi_\kappa(\kappa < t^{-1}) \lesssim e^{-Ct}, \quad \text{for all large } t > 0. \tag{3.2}$$

A common distribution that satisfies this requirement is the inverse Gaussian distribution [10] 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. A generalized inverse Gaussian density, proportional to $t^a e^{-b(t-\xi)^2/t}$ with any $b > 0$ and $a \in \mathbb{R}$, can also be used.

Prior for Θ. Since the structure in Ω is determined by the structure in Θ, we choose Π_Θ to induce the desired structure in Ω. Fortunately, most of the existing priors on a precision matrix could be directly 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 prior [25] or a mixture thereof [3]. Similarly, structures like a sparse Cholesky decomposition or a sparse factor model 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 4 and Section 5 below. Roughly, our technical requirement is that Π_Θ satisfies the sufficient conditions originally laid out in Ghosal, Ghosh and van der Vaart [21] 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.

Prior for ν. We require that the support of the prior distribution for ν is bounded by some large positive constant M, and that it has exponential lower tail, i.e., for some constant $C > 0$,

$$\Pi_\nu(\nu < t^{-1}) \lesssim e^{-Ct}, \quad \text{for all large } t > 0. \tag{3.3}$$

We also require suitable prior concentration around the true-but-unknown measurement error variance $\nu^\star$. A common distribution that satisfies this requirement is the truncated inverse Gaussian distribution or a two-sided truncated distribution. Alternatively, a point mass at the adjusted maximum likelihood estimator of ν, which is

$$\hat{\nu} = \frac{\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^{\mathrm{T}} (Y_{ij} - \bar{Y}_i)}{npm - 1}, \tag{3.4}$$

where $\bar{Y}_i = \sum_{j=1}^m Y_{ij}/m$, can be used, which corresponds to an empirical Bayesian method.

Given a prior for Ω as described above, we update to the posterior distribution via Bayes's theorem. For the measurement error model (1.2), define the likelihood function as

$$L_n(\Omega, \nu) \propto |\det(\Omega^{-1} + \nu I_p)|^{-1/2} \exp[-n \operatorname{tr}\{S_n(\Omega^{-1} + \nu I_p)^{-1}\}/2], \tag{3.5}$$

where S_n is the sample covariance matrix of Y as in Section 2 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

$$\Pi_n(d\Omega, d\nu) = \Pi(d\Omega, d\nu \mid Y_1, \ldots, Y_n) \propto L_n(\Omega, \nu)\, \Pi(d\Omega, d\nu). \tag{3.6}$$

A consequence of this indirect formulation is that the posterior distribution cannot be computed in closed-form. Therefore, MCMC methods are needed to obtain samples from Π_n. Fortunately, these methods can be developed by modifying the existing algorithms available in the no-measurement-error literature; see Section 6.1.

3.2. 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. Even under maximal sparsity, there are p unrestricted diagonal entries, so it is essential that the dimension p is of a smaller order of n, and in particular $\log p$ is the same order of $\log n$, or less. 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_ν for ν and prior Π_κ for κ are reasonable in some sense, then the same posterior contraction rate prevails in the presence of measurement error. The following three conditions make this setup more precise.

Conditions on the prior.

(a) The prior Π_κ has a continuous density on $(0, \infty)$, and satisfies the tail condition (3.2).

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

   (i) there exists a sieve $\mathcal{S}_n$ of precision matrices, having the same posited low-dimensional structure as $\Omega^\star$, with entropy bound
$$\log N(\delta_n, \mathcal{S}_n, \|\cdot\|_2) \lesssim n\epsilon_n^2, \tag{3.7}$$
   where $\delta_n = (2Knp)^{-1}$ for the K presented below;

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

   (iii) for any constant $c > 0$, there exists another constant $C > 0$, such that
$$\Pi_\Theta(\{\Theta : \|\Theta - \Theta^\star\|_F \le c\epsilon_n\}) \gtrsim e^{-Cn\epsilon_n^2} \tag{3.8}$$
   for any $\Theta^\star$ having the same posited structure as $\Omega^\star$.

(c) The prior Π_ν has the support on $(0, M)$ with a large constant M, satisfies a tail condition like in (3.3), and satisfies
$$\Pi_\nu(\{\nu : |\nu - \nu^\star| \le \epsilon_n p^{-1/2}\}) \gtrsim e^{-Cn\epsilon_n^2}, \tag{3.9}$$
for some constant $C > 0$, where $\nu^\star$ is the true measurement error variance.

The conditions related to Π_Θ look complicated but, for the most part, these are the now-classical sufficient conditions from Ghosal, Ghosh and van der Vaart [21] for establishing posterior concentration rate results. Therefore, other authors who have investigated concentration rate properties of posterior distributions under various low-dimensional structures and priors, like in Section 4, would likely have checked these conditions already. One noticeable difference is in the entropy bound in Condition (b)(i), where the radius is proportional to $\delta_n = (2Knp)^{-1}$, 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 conditions for Π_ν are satisfied by a continuous measure or a point mass prior located close enough to $\nu^\star$. For continuous measures, many distributions can be used, in which a common example is the inverse Gaussian distribution. Further, for the point mass prior, a typical choice is $\hat{\nu}$ in (3.4), which satisfies the conditions for Π_ν since now $|\nu - \nu^\star| \lesssim (np)^{-1/2}$ if $\epsilon_n \gtrsim n^{-1/2}$. The prior concentration condition (3.9) can be guaranteed if Π_ν has a continuous support and $\nu^\star > 0$ is fixed, since the prior density has a lower bound around $\nu^\star$ such that
$$\Pi_\nu(\{\nu : |\nu - \nu^\star| \le \epsilon_n p^{-1/2}\}) \gtrsim e^{\log p - \log \epsilon_n}, \tag{3.10}$$

where $-\log \epsilon_n$ is of the order of $\log n$, which is smaller than $n\epsilon_n^2$ up to a constant. However, it will be a little complicated if $\nu^\star = 0$ or it varies in a sequence approaching 0 fast. Then, the lower bound (3.9) may not hold if the prior density rapidly decays at 0, for example, in an inverse Gaussian density. Therefore, the assumption on the prior concentration (3.9) is sometimes useful.

Theorem 3.1. *Assume that $\Omega^\star$ satisfies a specific low-dimensional structure, and 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, where the prior for Θ is based on the same low-dimensional structure as that posited for $\Omega^\star$. Assume that there exists a sequence $\epsilon_n \gtrsim n^{-1/2}$ with $\epsilon_n \to 0$ and $n\epsilon_n^2 \gtrsim \log n$ and a constant M such that Conditions (a)–(c) are satisfied by Π_κ, Π_Θ and Π_ν, respectively. Under the model in (1.2), for any fixed $\nu^\star \geq 0$, the posterior distribution Π_n in (3.6) contracts at the rate ϵ_n, that is, there exists a constant $L > 0$, depending on $\|\Omega^\star\|_2$ and ν, such that*

$$\mathsf{E}_{\Omega^\star,\nu^\star}\Pi_n(\{(\Omega,\nu) : \|\Omega - \Omega^\star\|_F > L\epsilon_n, |\nu - \nu^\star| > L\bar{\epsilon}_n\}) \to 0 \quad \text{as } n \to \infty, \tag{3.11}$$

where $\bar{\epsilon}_n = \epsilon_n p^{-1/2}$. If $\|\Omega^\star\|_2$ is not bounded, then the conclusion of ν remains the same but the conclusion of Ω holds with the rate $\epsilon_n' = \|\Omega^\star\|_2^2\epsilon_n$.

The proof of the theorem is given in Appendix A. A remarkable consequence of Theorem 3.1 is that the posterior contraction rate is not affected by measurement error even when it is not small.

4. Examples

In this section, we investigate some existing Bayesian methods for structured, high-dimensional precision matrix estimation and show how measurement error can be accommodated in these models. Since the prior for ν is regulated to satisfy Condition (c), we consider a prior for Θ from the literature and verify the requirements of Theorem 3.1 for each case. The prior for κ will be assumed to satisfy Condition (a), e.g., by choosing an inverse Gaussian distribution. Therefore, the discussion below will focus on the prior for Θ and on verifying Conditions (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 4.1–4.3 are given in Appendix A.

4.1. General sparsity

Banerjee and Ghosal [3] 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 Conditions (b) (i)–(iii) are satisfied.

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

$$\begin{aligned} \pi(\Theta \mid \Gamma) &\propto \prod_{\Gamma_{ij}=1} \exp(-\lambda|\Theta_{ij}|) \prod_{i=1}^{p} \exp(-\lambda\Theta_{ii}/2), \\ \pi(\Gamma \mid R) &\propto q^{\gamma}(1-q)^{-\gamma+p(p-1)/2}\mathbb{1}\{\gamma \leq R\}, \end{aligned} \tag{4.1}$$

where λ is a hyperparameter 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

$$\Pi(R > M) \leq \exp(-aM \log M) \tag{4.2}$$

for some $a > 0$ and large enough constant M. Such distributions include the Poisson and the binomial distributions.

Theorem 4.1. *Assume the same setup as in Theorem 3.1 with the priors (4.1) and (4.2) or fixed $R = R_0$ for general sparsity type of structure. Under the model in (1.2), there exists a constant $L > 0$ such that the posterior distribution Π_n in (3.6) contracts at the rate ϵ_n around Ω^*, 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 $\Omega^\star$.*

4.2. 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 and Ghosal [12]. Let U_{ij} denote the entry at the ith row and jth column of U and D_{ii} denote the ith diagonal entry of D. Let Γ denote the matrix formed by the indicators $\Gamma_{ij} = \mathbb{1}\{U_{ij} \neq 0\}$. Following Proposition 1 in Du and Ghosal [12], for $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, i$, consider the prior

$$\begin{aligned} &(U_{ij} \mid \Gamma_{ij}) \sim (1 - \Gamma_{ij})\mathsf{N}_p(0, \sigma_0^2) + \Gamma_{ij}\mathsf{N}_p(0, \sigma_1^2), \\ &\Gamma_{ij} \sim \mathsf{Bernoulli}(C_p/\sqrt{i}), \\ &D_{ii} \sim \mathsf{Gamma}(\alpha_1, \beta_1), \end{aligned} \tag{4.3}$$

where α_1, β_1, σ_0^2 and σ_1^2 are some hyperparameters and C_p is a constant going to zero as $p \to \infty$ polynomially in p^{-1}.

An alternative to the above prior is to consider more general positive real-valued Γ_{ij}, and induce a prior on U_{ij} through the hierarchical scheme

$$\begin{aligned} &(U_{ij} \mid \Gamma_{ij}) \sim \mathsf{N}_p(0, \Gamma_{ij}^2\sigma_1^2), \\ &\Gamma_{ij} \sim \mathsf{Cauchy}^+(0, 1), \\ &D_{ii} \sim \mathsf{Gamma}(\alpha_1, \beta_1), \end{aligned} \tag{4.4}$$

for $i = 1, 2, \ldots, p$, and $j = 1, 2, \ldots, i$, where $\mathsf{Cauchy}^{+}(0, 1)$ is the positive half-Cauchy distribution and σ_1^2 is a pre-specified global shrinkage parameter.

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

$$\eta = \Pi(|U_{ij}| > \epsilon_n p^{-1}) \leq p^{-b} \quad \text{and} \quad \eta \geq p^{-a},$$

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 (4.3) going to zero as $p \to \infty$ polynomially in p^{-1} or modifying the prior Γ_{ij} in (4.4) to be truncated above by $1/\tau$, where $\tau < p^{-b-2}\epsilon_n^2$. Although the above condition on η is required for the theoretical result, in practice, we choose $C_p = 1$ for the convenience of computation. In our simulation studies, the estimation results are not sensitive to the choice of C_p.

Theorem 4.2. *Consider the setup of Theorem 3.1 with the priors given by (4.3) or (4.4) for the sparse Cholesky decomposition. Then under the model in (1.2), the posterior distribution Π_n in (3.6) contracts at the rate ϵ_n around Ω^*, 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 $U^\star$.*

4.3. Banded structure using G-Wishart prior

Following Banerjee and Ghosal [2], assume that the sparse precision matrix has a banded structure. 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 on the rate of the Frobenius norm in the presence of measurement error.

Suppose that the true $p \times p$ dimensional precision matrix $\Theta^\star$ is k-banded, that is, $\Theta^\star_{ij} = 0$ 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 $S_j = \{j, \ldots, j+k-1\}$, $j = 2, \ldots, p-k$ [2]. 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 $S \subset \{1, \ldots, p\}$, $M_S = (\!(M_{ij} : i, j \in S)\!)$, the principal minor of M formed by the entries of S. Let $\mathcal{P}_G$ 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 4.3. *Consider the setup of Theorem 3.1 with prior $\Theta \sim \mathsf{G\text{-}Wish}(\delta, I_p)$. Assume that the eigenvalues of Ω^* are bounded and bounded away from zero, and for a sufficiently small $\epsilon > 0$, $\{\Omega : \|\Omega - \Omega^*\|_\infty < \epsilon\} \subset \mathcal{P}_G$. Under the model in (1.2), the posterior distribution Π_n in (3.6) contracts at the rate ϵ_n around Ω^*, where $\epsilon_n = n^{-1/2}(p \log n)^{1/2}$.*

5. 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. [30]. 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 (1.1), where the possibility $\nu = 0$ (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 3.1, 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 $\Gamma_{ij} = \mathbb{1}\{\Lambda_{ij} \neq 0\}$, and let $\gamma \leq ks$ denote the total number of non-zero entries of Λ.

We follow Pati et al. [30] 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 $\Omega^\star = \Lambda^\star\Lambda^{\star\top} + \kappa^\star I$ 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 $\mathbb{1}(\Lambda^\star_{ij} \neq 0)$ and $\gamma^\star = \sum_{i,j} \Gamma^\star_{ij} \leq k^\star s^\star$.

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) $(c^{\star 5} s^\star k^\star)^{1/2}(\log n) \lesssim n^{1/2}$;
(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. Assumption (A2) is introduced to control the final convergence rate appropriately; that assumption can be relaxed to $(c^\star s^\star k^\star)^{1/2} \log n \lesssim n^{1/2}$ when $\nu = 0$ is known. Assumption (A3) controls the sparsity, which is crucial in high-dimensional problems.

For the Bayesian model formulation, let Λ_{ij} 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. [30], except that we make certain choices to meet Condition (b) such as the inverse Gaussian distribution. For $i = 1, 2, \ldots, p$, and

$j = 1, 2, \ldots, k$, let the priors

$$\begin{aligned} &\kappa \sim \mathsf{invGaussian}(\mu_1, \lambda_1), \\ &k \sim \mathsf{Pois}(\theta_1), \\ &(\Lambda_{ij} \mid \Gamma_{ij}) \sim (1 - \Gamma_{ij})\delta_0 + \Gamma_{ij} \mathsf{N}_p(0, \sigma_1^2), \\ &\Gamma_{ij} \sim \mathsf{Bernoulli}(\eta), \end{aligned} \tag{5.1}$$

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 $\mathsf{Bin}(pk^\star, \eta)$, where $\eta = \pi(|\Lambda_{ij}| > 0)$ for any $0 < i \leq p$ and $0 < j \leq k^\star$. The choice of η is made to guarantee that $\eta \asymp (pk^\star)^{-1}$. Moreover, with such a specification, we have the prior probability $\eta/2 \leq \pi(|\Lambda_{ij}| > \epsilon_n/4\sqrt{c^{\star 3} pk^\star}) \leq \eta$ with ϵ_n introduced in Theorem 5.1.

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:

$$(\eta \mid k) \sim \mathsf{Beta}(1, akp + 1), \tag{5.2}$$

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 concentration rate in Theorem 5.1. With such a hyper-prior on η, given $k = k^\star$, we can calculate that

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

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

Theorem 5.1. *Suppose that the data are generated from (1.1) and Assumptions (A1), (A2) and (A3) hold for the true precision matrix $\Omega^\star$. Consider the prior given by (5.1). Then the posterior distribution Π_n in (3.6) contracts at the rate ϵ_n around $\Omega^\star$, where*

- $\epsilon_n = n^{-1/2}(c^\star s^\star k^\star)^{1/2}(\log n)^{1/2}$ *if* $\nu^\star = 0$ *is known,*
- *and* $\epsilon_n = n^{-1/2}(c^\star)^{5/2}(s^\star k^\star)^{1/2}(\log n)^{1/2}$ *if* $\nu^\star$ *is fixed and positive.*

If the prior (5.2) is imposed on η, then the rates are

- $\epsilon_n = n^{-1/2}(c^\star s^\star k^\star)^{1/2}(\log n)$ *if* $\nu^\star = 0$ *is known,*
- *and* $\epsilon_n = n^{-1/2}(c^\star)^{5/2}(s^\star k^\star)^{1/2}(\log n)$ *if* $\nu^\star$ *is fixed and positive.*

6. Numerical results

6.1. 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 (X, Y, Ω, ν), it is possible to write down the full conditionals as

$$(X_i \mid Y_{i1}, \ldots, Y_{im}, \Omega, \nu) \overset{\text{ind}}{\sim} \mathsf{N}_p\big(m(mI_p + \nu\Omega)^{-1}\bar{Y}_i,\ \nu(mI_p + \nu\Omega)^{-1}\big), \tag{6.1}$$

$$(\Omega \mid X_1, \ldots, X_n) \sim \Pi(\Omega \mid X_1, \ldots, X_n), \tag{6.2}$$

$$(\nu \mid X_1, \ldots, X_n, Y_{11}, \ldots, Y_{nm}) \sim \Pi(\nu \mid X_1, \ldots, X_n, Y_{11}, \ldots, Y_{nm}), \tag{6.3}$$

where $\Pi(\Omega \mid X_1, \ldots, X_n)$ is the posterior based on the no-measurement-error model, with the augmented dataset $X_1, \ldots, X_n$, and $\bar{Y}_i = m^{-1} \sum_{j=1}^m Y_{i,j}$. Note that $Y_1, \ldots, Y_n$ do not appear in (6.2) 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 literature—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 (6.2) 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 3 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 inverse, and simulations give sensible results. The numerical results presented below employ this simplification.

If ν is known, step (6.3) can simply be ignored by using the true ν in the other steps. However, when ν is unknown, under the model in (1.1), a prior is needed for ν to execute the algorithm. The conditions imposed on the prior for ν in the theoretical results are sufficient but are not necessary. So, for practical implementation, one could feel reasonably safe in taking any suitable prior for ν that simplifies the computation. For example, one might use the non-informative Jeffreys prior, that is, $\pi(\nu) \propto \nu^{-3/2}$, or the inverse-Gamma prior. Consider the non-informative Jeffreys prior and the full conditional posterior of ν is in a closed form as inverse-Gamma distribution as

$$(\nu \mid X_1, \ldots, X_n, Y_{11}, \ldots, Y_{nm}) \sim \mathsf{IG}\left(\frac{mpn}{2}, \frac{\sum_{i=1}^n \sum_{j=1}^m (Y_{ij} - X_i)^{\mathrm{T}}(Y_{ij} - X_i)}{2}\right).$$

Another popular method to deal with the unknown nuisance parameter is using an estimator of ν and pretending it as the truth. This empirical Bayesian method is equivalent to specify a point mass prior located at the estimator. A typical choice of such estimator is that in (3.4). The resulting procedure is sensible as long as the estimator of ν is sufficiently accurate. The following simulation study makes use of this simple plug-in method.

6.2. Simulations

Since the proposed method covers both cases that the true value $\nu^\star$ of ν is known with $m = 1$ or it is unknown with $m > 1$, we conduct two separate simulation studies to explore the performance of adjusting for the measurement error.

6.2.1. When ν is known and $m = 1$

We conduct a simulation study over different structures of true precision matrix and known magnitudes of measurement error with $m = 1$. We fix the dimension $p = 50$ and sample size $n = 100$. Let $\Omega^\star_{ij}$ 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): $\Omega^\star_{ii} = 10$ and $\Omega^\star_{i,i-1} = \Omega^\star_{i,i+1} = 5$ for $1 \le i \le p$; $\Omega^\star_{ij} = 0$ otherwise.
- AR(2): $\Omega^\star_{ii} = 10$, $\Omega^\star_{i,i-1} = \Omega^\star_{i,i+1} = 5$ and $\Omega^\star_{i,i-2} = \Omega^\star_{i,i+2} = 2.5$ for $1 \le i \le p$; $\Omega^\star_{ij} = 0$ otherwise.
- Block(2): $\Omega^\star_{ii} = 10$, $\Omega^\star_{ij} = 5$ for $(k-1)p/2 + 1 \le i \ne j \le kp/2$ and $1 \le k \le 2$; $\Omega^\star_{ij} = 0$ otherwise.
- Block(5): $\Omega^\star_{ii} = 10$, $\Omega^\star_{ij} = 5$ for $(k-1)p/5 + 1 \le i \ne j \le kp/5$ and $1 \le k \le 5$; $\Omega^\star_{ij} = 0$ otherwise.

For each sparse structure, 100 replicates are run. We consider the priors assigned on the Cholesky decomposition structure in Section 4.2 and the setup introduced as (4.3) 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 6.1 to sample from the posterior and choose Y_i's as the initial values of X_i's for $i = 1, \dots, n$, 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 1, 2, 3, and 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

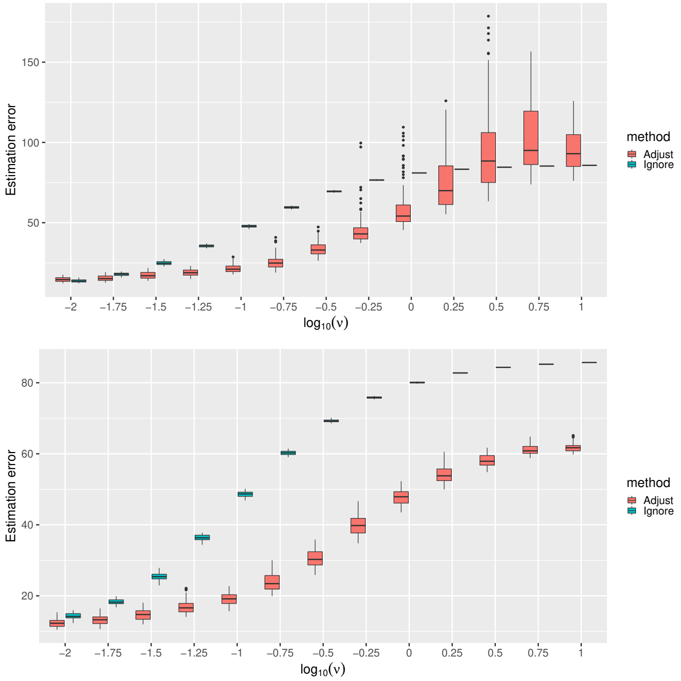

Fig 1: Frobenius norm estimation error in the AR(1) model versus the magnitude of measurement error using diffuse prior (top) and informative prior (bottom).

other words, ν varies from 10^{-2} to 10 over 13 different values, which are equally spaced on the log-scale.

For the results of AR(1) model in Figure 1, the estimator that corrects for the measurement error has better accuracy in terms of the 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 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 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 approximates to a fixed value 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

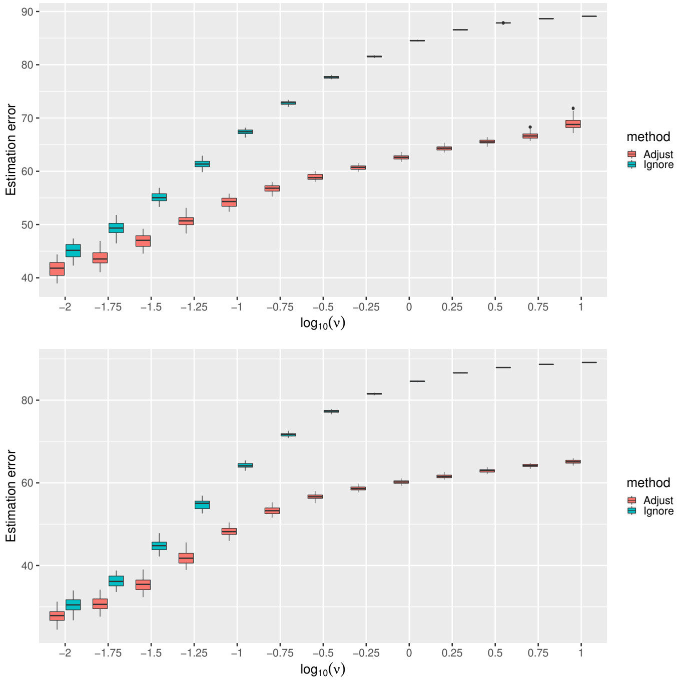

Fig 2: Frobenius norm estimation error in the AR(2) model versus the magnitude of measurement error using diffuse prior (top) and informative prior (bottom).

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. Similar phenomena are observed in the block structures in Figure 3 and 4.

6.2.2. When ν is unknown and $m > 1$

We consider the same simulation settings as in Section 6.2.1, except that now ν is unknown and we have $m = 2$ replications for each outcomes to estimation ν and Ω. Since the effectiveness of adjusting for the measurement error in those four structures are similar, only AR(1) and AR(2) structures are considered. Further, the empirical Bayesian method using $\hat{\nu}$ as in (3.4) is employed to estimate ν and only the informative prior is used for estimating Ω since its performance is much better than the diffuse prior.

For both AR(1) and AR(2) structures, as shown in Figure 5, correcting for the measurement error improves the accuracy of the estimation in terms of the Frobenius norm for each choice of ν. Again, the estimators based on the proposed

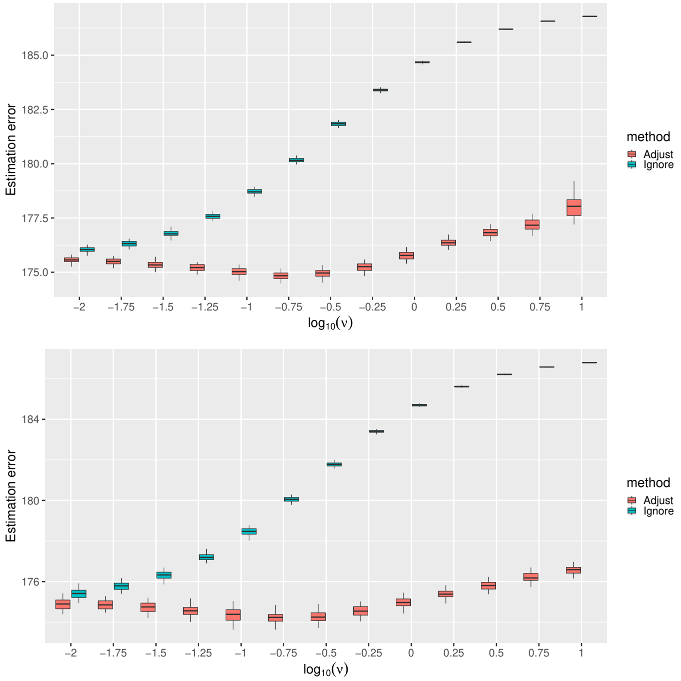

Fig 3: Frobenius norm estimation error in the Block(2) model versus the magnitude of measurement error using diffuse prior (top) and informative prior (bottom).

method have larger variance than the baseline model since the unknown ν and the extra layer in the hierarchical model introduce more variability. Despite having slightly larger variability, the estimation error with the proposed method is still far superior to that using the naive method that ignores measurement error. There is a peculiar initial downward trend in the proposed method's estimation error as $\nu^\star$ increases. We believe that this is because, when $\nu^\star$ is very small, i.e., close to the boundary $\nu^\star = 0$, the plug-in estimator $\hat{\nu}$ loses accuracy. However, when $\nu^\star$ is away from the boundary, the expected trend emerges, namely, the estimation error is increasing but more slowly than for the naive method.

Appendix A: Proofs of the theorems

A.1. Proof of Theorem 2.1

A simple bias–variance decomposition yields

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

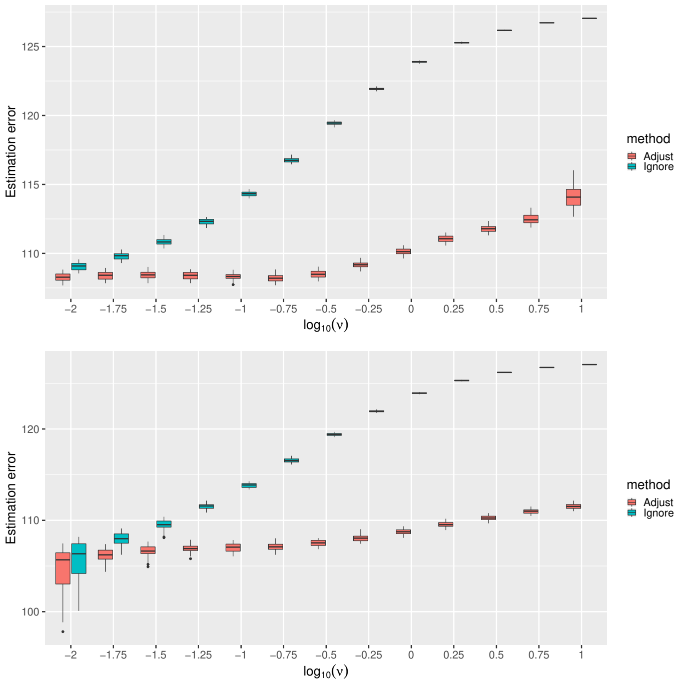

Fig 4: Frobenius norm estimation error in the Block(5) model versus the magnitude of measurement error using diffuse prior (top) and informative prior (bottom).

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 $D = \operatorname{diag}(D_{11}, \ldots, D_{pp})$, we obtain $\nu D_{jj} \leq 1$ 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 $\|\Omega^{\star 2}\|_F^2 = \|D^2\|_F^2 \geq p \cdot \min(D_{jj}^4 : j = 1, \ldots, p)$, the second assertion follows.

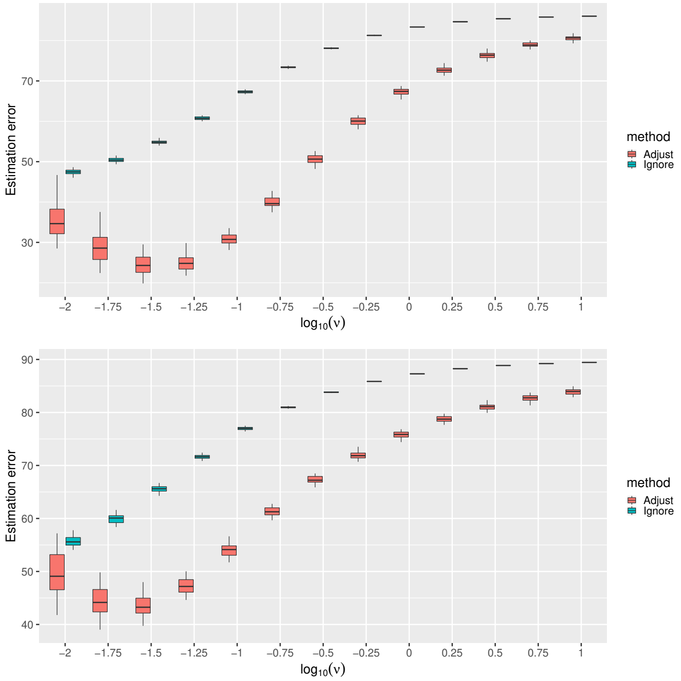

Fig 5: Frobenius norm estimation error in the AR(1) model (top) and AR(2) model (bottom) versus the magnitude of measurement error.

A.2. Proof of Theorem 3.1

The proof will proceed in a sequence of steps driven by those sufficient conditions in Ghosal, Ghosh and van der Vaart [21] for bounding the posterior concentration rate since now the Y_i's are independent and identically distributed.

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 (3.1). 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 the above display, provided the corresponding assumptions hold there. Indeed, recall that, e.g., Condition (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 generic Ω and ν, let g_Ω and $g_{\Omega,\nu}$ 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 (1.2) are i.i.d. with density $g_{\Omega,\nu}$. For the target rate ϵ_n, we aim to show that for some $C > 0$,

$$\Pi(\{(\Omega,\nu) : K(g_{\Omega^\star,\nu^\star}, g_{\Omega,\nu}) \le \epsilon_n^2, V(g_{\Omega^\star,\nu^\star}, g_{\Omega,\nu}) \le \epsilon_n^2\}) \gtrsim e^{-Cn\epsilon_n^2}, \tag{A.1}$$

where $K(f_1, f_2) = \int f_1 \log(f_1/f_2)$ 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 of two densities f_1 and f_2, respectively, as defined above. By our assumption that the eigenvalues of $\Omega^\star$ are bounded away from 0 and infinity, and Lemma B.1 in Appendix B, there exists $c > 0$ such that

$$\begin{aligned}\Pi(\{(\Omega,\nu) : K(g_{\Omega^\star,\nu^\star}, g_{\Omega,\nu}) \le \epsilon_n^2, V(g_{\Omega^\star,\nu^\star}, g_{\Omega,\nu}) \le \epsilon_n^2\})& \\ \gtrsim \Pi(\{\Omega,\nu : \|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F \le c\epsilon_n\}).&\end{aligned} \tag{A.2}$$

Then, from the first result in Lemma B.2 in Appendix B, there exists $c_1, c_2 > 0$ such that

$$\begin{aligned}\Pi(\{(\Omega,\nu) : \|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F \le c\epsilon_n\})& \\ \gtrsim \Pi(\{\Omega : \|\Omega - \Omega^\star\|_F \le c_1\epsilon_n\}) \cdot \Pi(\{\nu : |\nu - \nu^\star| \le c_2\epsilon_n p^{-1/2}\}).&\end{aligned} \tag{A.3}$$

For the first term on the right hand side, replacing Ω by (Θ,κ) and $\Omega^\star$ by $(\Theta^\star,\kappa^\star)$, the triangle inequality gives

$$\|\Omega - \Omega^\star\|_F \le \|\Theta - \Theta^\star\|_F + \|\kappa I_p - \kappa^\star I_p\|_F \le \|\Theta - \Theta^\star\|_F + p^{1/2}|\kappa - \kappa^\star|. \tag{A.4}$$

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

$$\Pi_\kappa(\{\kappa : |\kappa - \kappa^\star| \le cp^{-1/2}\epsilon_n/2\}) \gtrsim p^{-1/2}\epsilon_n = \exp\{\log\epsilon_n - \tfrac{1}{2}\log p\} \gtrsim e^{-Gn\epsilon_n^2}. \tag{A.5}$$

For the second term in (A.3), it is automatically guaranteed by Condition (c) on Π_ν. Combining (A.3)–(A.5) and using Conditions (b) on Π_Θ, (A.2) follows, establishing (A.1).

Step 2: *Sieve, test construction, and error rates.* In order to apply the theory of posterior contraction in Ghosal and van der Vaart [22] to show that the contraction rate at the truth with respect 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 and Ghosal [29] and Jeong and Ghosal [23]. 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

$$\mathcal{T}_n = \{\Omega = \Theta + \kappa I_p, \nu : \Theta \in \mathcal{S}_n,\ M_n^{-1} \le \kappa \le M_n, M_n^{-1} \le \nu \le M_n\}, \tag{A.6}$$

with $\mathcal{S}_n$ as in the statement of the theorem and we choose $M_n = Kn\epsilon_n^2 \to \infty$, for some sufficiently large multiple K and therefore $M_n \le n$. What makes this sieve appropriate for our purposes here is that every $\Omega \in \mathcal{T}_n$ 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 log-affinity, between two densities f_1 and f_2 is given by $R(f_1, f_2) = -\log \int (f_1 f_2)^{1/2}$. In the present context, the densities are those of $\mathsf{N}_{mp}(0, \Omega_\nu^{-1})$, to be denoted by $g_{\Omega,\nu}$, indexed by Ω, and $R(g_{\Omega^*,\nu^\star}, g_{\Omega,\nu})$ can be abbreviated by $R(\Omega^\star_{\nu^\star}, \Omega_\nu)$. By simple calculations,

$$R(\Omega^\star_{\nu^\star}, \Omega_\nu) = -\log\Big(\frac{|\Omega^\star_{\nu^\star}|^{1/4}|\Omega_\nu|^{-1/4}}{|\frac{1}{2}\Omega^\star_{\nu^\star} + \frac{1}{2}\Omega_\nu|^{1/2}}\Big).$$

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

$$\begin{aligned}
\mathsf{E}_{\Omega^\star,\nu^\star}\phi_n &= \int \mathbb{1}\Big[\Big\{\frac{g^n_{\Omega^\dagger,\nu^\dagger}(y^n)}{g^n_{\Omega^\star,\nu^\star}(y^n)}\Big\}^{1/2} \ge 1\Big] g^n_{\Omega^\star,\nu^\star}(y^n)\, dy^n \\
&\le \Big[\int \{g_{\Omega^\dagger,\nu^\dagger}(y) g_{\Omega^\star,\nu^\star}(y)\}^{1/2}\, dy\Big]^n \\
&= e^{-nR(\Omega^\star_{\nu^\star}, \Omega^\dagger_{\nu^\dagger})} \\
&\le e^{-n\epsilon_n^2}.
\end{aligned}$$

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

$$\begin{aligned}\mathsf{E}_{\Omega,\nu}(1-\phi_n) &= \mathsf{E}_{\Omega^\dagger,\nu}\big\{(1-\phi_n)g^n_{\Omega,\nu}/g^n_{\Omega^\dagger,\nu^\dagger}\big\}\\ &\le \{\mathsf{E}_{\Omega^\dagger,\nu}(1-\phi_n)\}^{1/2}\Big[\int \{g_{\Omega,\nu}(y)/g_{\Omega^\dagger,\nu^\dagger}(y)\}^2\, g_{\Omega^\dagger,\nu^\dagger}(y)\,dy\Big]^{n/2}.\end{aligned} \tag{A.7}$$

Let $h_\nu(\cdot|X)$ denote the density of $\mathsf{N}_p(X, \nu I_p)$ and from the Cauchy–Schwarz inequality, the second factor in the square brackets equals

$$\begin{aligned}&\int\cdots\int \frac{\left(\int\prod_{i=1}^m h_\nu(y_i|x)g_\Omega(x)dx\right)^2}{\int\prod_{i=1}^m h_{\nu^\dagger}(y_i|x)g_{\Omega^\dagger}(x)dx}dy_1\cdots dy_m\\ &\quad\le \int\int\cdots\int\prod_{i=1}^m \frac{\left(h_\nu(y_i|x)\right)^2}{h_{\nu^\dagger}(y_i|x)}dy_1\cdots dy_m\frac{\left(g_\Omega(x)\right)^2}{g_{\Omega^\dagger}(x)}dx\\ &\quad= \left(\frac{\nu^\dagger}{\sqrt{\nu}\sqrt{2\nu^\dagger-\nu}}\right)^{mp}\int\frac{\left(g_\Omega(x)\right)^2}{g_{\Omega^\dagger}(x)}dx\\ &\quad= \left(1+\frac{(\nu^\dagger-\nu)^2}{\nu(2\nu^\dagger-\nu)}\right)^{mp/2}\int\frac{\left(g_\Omega(x)\right)^2}{g_{\Omega^\dagger}(x)}dx\end{aligned}$$

By the choice of $M_n \le n$ and δ_n, the first term can be bounded by

$$\left(1+\frac{2M_n^2}{4K^2n^2p^2}\right)^{mp/2} \le \left(1+\frac{1}{2K^2p^2}\right)^{mp/2} \le \exp(m/4K^2p).$$

Then, the second factor equals

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

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 (A.6), we get $\|(\Omega^\dagger)^{-1}\|_2 \le M_n$ and $\|\Omega-\Omega^\dagger\|_2 \le \delta_n$, which implies that, on the sieve,

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

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,

$$\mathsf{E}_{\Omega,\nu}(1-\phi_n) \le \exp\{-n\epsilon_n^2/2 + npM_n\delta_n/2 + m/4K^2p\} \lesssim \exp\{-n\epsilon_n^2/4\},$$

as $\delta_n = (2Knp)^{-1}$ and $M_n = Kn\epsilon_n^2$. This gives the Type II error bound (A.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 $\mathcal{T}_n$ with respect to spectral norm $\log N(\delta_n, \mathcal{T}_n, (\|\cdot\|_2, |\cdot|))$. It suffices to show that this is bounded by a constant multiple of $n\epsilon_n^2$. To this end, if we take $\Omega_1 = \Theta_1 + \kappa_1 I_p$, ν_1 and $\Omega_2 = \Theta_2 + \kappa_2 I_p$, ν_2 in $\mathcal{T}_n$, then by the triangle inequality, for the given δ_n,

$$\begin{aligned}\log N(\delta_n, \mathcal{T}_n, (\|\cdot\|_2, |\cdot|)) &\leq \log N(\delta_n/2, \mathcal{S}_n, \|\cdot\|_2) + 2\log(M_n\delta_n^{-1}) \\ &\lesssim n\epsilon_n^2 + 2\log n \lesssim n\epsilon_n^2.\end{aligned}$$

Step 4: *Prior probability of the complement of the sieve.* In view of Conditions (a), (b)(ii) and (c) on the prior distributions and the choice $M_n = Kn\epsilon_n^2$, we estimate $\Pi(\mathcal{T}_n^c) \leq \Pi_\Theta(\mathcal{S}_n^c) + \Pi_\kappa([M_n^{-1}, M_n]^c) + \Pi_\nu([M_n^{-1}, M_n]^c) \lesssim e^{-Gn\epsilon_n^2}$, 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 and van der Vaart [21], we obtain a concentration rate in terms of Rényi divergence $\mathsf{E}_{\Omega^\star,\nu^\star}\Pi_n(\{(\Omega,\nu) : R(\Omega^\star_{\nu^\star}, \Omega_\nu) > L'\epsilon_n^2\}) \to 0$ for some $L' > 0$ sufficiently large. Under the assumption that $\|\Omega^\star\|_2$ is bounded, we shall conclude that $\mathsf{E}_{\Omega^\star,\nu^\star}\Pi_n(\{\Omega, \nu : \|\Omega^\star - \Omega\|_F > L\epsilon_n^2, |\nu^\star - \nu| > L\epsilon_n^2\}) \to 0$ for some $L > 0$. Towards this, define $A = \Omega^{\star-1/2}_{\nu^\star}\Omega_\nu\Omega^{\star-1/2}_{\nu^\star}$. Let $\alpha_1 \leq \cdots \leq \alpha_{mp}$ denote the eigenvalues of A in the increasing order. It follows from Lemma A.2(ii) of Banerjee and Ghosal [3] that, if the Hellinger distance or the Rényi divergence of $g_{\Omega^\star,\nu^\star}$ from $g_{\Omega,\nu}$ is sufficiently small, then $\max\{|\alpha_j - 1| : j = 1, \ldots, mp\} \leq 1$ and, therefore, every $\alpha_j \leq 2$. 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 $R(\Omega^\star_{\nu^\star}, \Omega_\nu)$ can be written as

$$\begin{aligned}-\frac{1}{4}\log\frac{|A|}{|\frac{1}{2}I_{mp} + \frac{1}{2}A|^2} &= -\frac{1}{4}\sum_{j=1}^{mp}\log\frac{4\alpha_j}{(1+\alpha_j)^2} \\ &\geq \frac{1}{4}\sum_{j=1}^{mp}\Big\{1 - \frac{4\alpha_j}{(1+\alpha_j)^2}\Big\} = \frac{1}{4}\sum_{j=1}^{mp}\Big(\frac{1-\alpha_j}{1+\alpha_j}\Big)^2.\end{aligned}$$

Since $1 + \alpha_j \leq 1 + \alpha_p \leq 3$ for all j, and $\|A - I_{mp}\|_F^2 = \sum_{j=1}^p (1-\alpha_j)^2$, we have that $R(\Omega^\star_{\nu^\star}, \Omega_\nu) \gtrsim \|A - I_{mp}\|_F^2$. Next, observe that

$$\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F^2 = \|\Omega^{\star 1/2}_{\nu^\star}\Omega^{\star-1/2}_{\nu^\star}(\Omega_\nu - \Omega^\star_\nu)\Omega^{\star-1/2}_{\nu^\star}\Omega^{\star 1/2}_{\nu^\star}\|_F^2 \leq \|\Omega^\star_{\nu^\star}\|_2^2\,\|A - I\|_F^2.$$

Combining these, we conclude that

$$R(\Omega^\star_{\nu^\star}, \Omega_\nu) \gtrsim \|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F^2,$$

by the fact that $\|\Omega^\star_{\nu^\star}\|_2 \leq 1/\nu^\star$. For $R(\Omega^\star_{\nu^\star}, \Omega_\nu) \lesssim \epsilon_n^2$, it follows from the second part of Lemma B.2 in Appendix B that

$$|\nu - \nu^\star| \lesssim \epsilon_n/\sqrt{p}, \quad \|\Omega - \Omega^\star\|_F \lesssim \|\Omega^\star\|_2^2\epsilon_n.$$

We assume that $\|\Omega^\star\|_2 \lesssim 1$, so $\|\Omega - \Omega^\star\|_F \lesssim \epsilon_n$ follows immediately.

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

$$\|\Omega - \Omega^\star\|_F^2 \lesssim \|\Omega^\star\|_2^4 \epsilon_n^2.$$

Therefore, the result of Ω in Theorem 3.1 holds with the modified rate $\epsilon_n' = \|\Omega^\star\|_2^2 \epsilon_n$.

A.3. Proof of Theorem 4.1

As Condition (a) is directly assumed, we only need to verify the conditions of Theorem 3.1 for the sieve

$$\mathcal{S}_n = \{\Theta : \gamma \leq R_n, \|\Theta\|_\infty \leq M_n\}, \quad R_n = Kn\epsilon_n^2/\log n, \quad M_n = Kn\epsilon_n^2,$$

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

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

where the last expression is no more than

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

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

Next, for $\Theta \in \mathcal{S}_n^c$, either $|\Theta_{ij}| > M_n$ 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 $\Pi(R > R_n) \lesssim \exp(-aR_n \log R_n) \lesssim \exp(-aKn\epsilon_n^2)$ by the assumption (4.2) and the choice $R_n \asymp n\epsilon_n^2/\log n$. By taking K sufficiently large, we verify Condition (b)(ii).

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

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

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

A.4. Proof of Theorem 4.2

We consider the sieve

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

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 Conditions (b)(i)–(iii).

For any two precision matrices $\Theta_1, \Theta_2 \in \mathcal{S}_n$ with Cholesky decompositions $\Theta_1 = U_1 D_1 U_1^{\mathrm{T}}$ and $\Theta_2 = U_2 D_2 U_2^{\mathrm{T}}$, we obtain $\|\Theta_1 - \Theta_2\|_2$ less than or equal to

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

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

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

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

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

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

Under both (4.3) and (4.4), 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 $\exp(-\lambda M_n) \lesssim \exp(-\lambda K n\epsilon_n^2)$. Choosing K sufficiently large ensures the required bound.

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

$$\begin{aligned}
\Pi(\|D - D^\star\|_F \leq \epsilon_n) &\geq \Pi(\|D - D^\star\|_\infty \leq \epsilon_n/\sqrt{p})\\
&\gtrsim (\epsilon_n/p)^p\\
&\gtrsim \exp\{-cp\log(\epsilon_n/p)\} \qquad \text{(A.8)}
\end{aligned}$$

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

$$\begin{aligned}
\Pi(\|U - U^\star\|_F \leq \epsilon_n) &\geq \Pi(\|U - U^\star\|_\infty \leq \epsilon_n/p)\\
&= \{\Pi(|U_{ij}| < \epsilon_n/p)\}^{p(p-1)/2-s^\star}\\
&\quad\times\{\Pi(|U_{ij} - U_{ij}^\star| \leq \epsilon_n/p \mid |U_{ij}| > \epsilon_n/p)\Pi(|U_{ij}| > \epsilon_n/p)\}^{s^\star}\\
&\gtrsim (1 - p^{-b})^{p(p-1)/2-s^\star}(\epsilon_n/p)^{s^\star}p^{-as^\star}
\end{aligned}$$

which shares the same lower bound (A.8). This follows because $\Pi(\|U-U^\star\|_\infty \leq \epsilon_n/p)$ is equal to the probability that $\Pi(|U_{ij}| < \epsilon_n/p) \gtrsim (1-p^{-b})$ when $|U^\star_{ij}| = 0$, and that is the case for $p(p-1)/2 - s^*$ 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 $-\log \Pi(\|\Theta - \Theta^\star\|_F \leq \epsilon) \lesssim (p+s^*)\log(p/\epsilon_n) \lesssim (p+s^*)\log n$, so the rate $\epsilon_n = [\{(p+s^*)\log n\}n^{-1}]^{1/2}$ satisfies the required condition.

A.5. Proof of Theorem 4.3

By the arguments given at the beginning of the proof of Theorem 3.1, 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 $\mathcal{P}_G$.

We consider the sieve

$$\mathcal{S}_n = \{\Theta : \|\Theta\|_\infty \leq M_n\},$$

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

$$\begin{aligned}\log N(\delta_n, \mathcal{S}_n, \|\cdot\|_2) &\leq \log N(\delta_n/\sqrt{2pk}, \mathcal{S}_n, \|\cdot\|_\infty) \\ &\leq \log(k \cdot (\sqrt{2pk} M_n/\delta_n)^{pk}) \\ &\lesssim pk \log n. \end{aligned} \tag{A.9}$$

Note that $\Theta \in S_n^c$ if $|\Theta_{ij}| > M_n$ 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

$$\Pi(S_n^c) \leq p\,\mathsf{P}(Y > M_n) \leq \exp(-cM_n + \log p) \leq \exp(-c'n\epsilon_n^2) \tag{A.10}$$

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

It remains to verify that the prior concentration rate is $\epsilon_n = \{n^{-1}(p\log n)\}^{1/2}$. There are at most pk free arguments in Θ due to the k-banding structure. By Roverato [31], 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 and Ghosal [2]. 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 $\mathcal{P}_G$ by the assumption, it follows that

$$\Pi(\|\Omega - \Omega^*\|_F \leq \epsilon_n) \gtrsim \Pi(\|\Omega - \Omega^*\|_\infty \leq \epsilon_n/\sqrt{kp}) \gtrsim e^{-c''p}(\epsilon_n/\sqrt{kp})^{kp} \tag{A.11}$$

for some $c'' > 0$. From (A.9)–(A.11), the rate $\epsilon_n = \{n^{-1}(p\log n)\}^{1/2}$ follows by an application of Theorem 3.1.

A.6. Proof of Theorem 5.1

We apply Theorem 3.1 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

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

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

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

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 $0 < i \le p$ and $0 < j \le k^\star$ and let $\delta = \epsilon/(4\sqrt{c^{\star 3}pk^\star})$. Then it can be bounded below by

$$\{1-\Pi(|\Lambda_{ij}|>a^\star)\}^{(p-s^\star)k^\star}\{\Pi(|\Lambda_{ij}-\Lambda^\star_{ij}|\le a^\star \mid |\Lambda_{ij}|>a^\star)\Pi(|\Lambda_{ij}|>a^\star)\}^{s^\star k^\star}.$$

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 $\Pi(|\Lambda_{ij}-\Lambda^\star_{ij}| \le \delta \mid |\Lambda_{ij}| > \delta) \gtrsim \delta \exp(-c^\star/2)$ since the probability density is lower bounded on a closed region and $\Pi(|\Lambda_{ij}| > \delta) \asymp (pk^\star)^{-1}$. Therefore, we conclude that

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

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

Next, let

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

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 $\Pi(\mathcal{S}_n^c)$ so that the theorem applies to give the rate ϵ_n, we need to establish that for some sufficiently large constant $C > 0$,

$$\begin{aligned}\Pi(k>k_n) &\lesssim e^{-Cn\epsilon_n^2},\\ \Pi(\gamma>\gamma_n) &\lesssim e^{-Cn\epsilon_n^2},\\ \max\Pi(|\Lambda_{ij}| > \sqrt{M_n/(2\gamma_n)}) &\lesssim e^{-Cn\epsilon_n^2}.\end{aligned} \tag{A.12}$$

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 $\Pi(\mathcal{S}_n^c)$, 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 $\max \Pi(|\Lambda_{ij}| > \sqrt{M_n/(2\gamma_n)})$. 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 now suffices 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 (A.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 $\mathcal{S}_n$ with respect to the spectral norm is bounded by

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

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 3.1, since $\|\Omega^\star\|_2 \le c^\star$ by assumption, the contraction rate in terms of the Frobenius distance is $n^{-1/2}(\log n)^{1/2}(c^\star s^\star k^\star)^{1/2}$ in the case of no-measurement error changes 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 a fixed scale of measurement error.

When a prior (5.2) 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 and van der Vaart [9], $\Pi(\gamma > \gamma_n)$ 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 $\{n^{-1}(\log n) c^\star s^\star k^\star\}^{1/2}$ and $\{n^{-1}(\log n) c^{\star 5} s^\star k^\star\}^{1/2}$ respectively, depending on $\nu^\star = 0$ is known and $\nu^\star$ is positive and fixed.

Appendix B: 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. 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 and Ghosal [3] and the proof of Theorem 2 in Du and Ghosal [12].

Lemma B.1. *Suppose that the eigenvalues of $\Omega^\star$ lie in $[M^{-1}, M]$ for some large enough constant $M > 0$. If $\|\Omega - \Omega^\star\|_F \le \epsilon M^{-1}$ for a sufficiently small $\epsilon > 0$, then* $\max\{K(g_{\Omega^\star}, g_\Omega), V(g_{\Omega^\star}, g_\Omega)\} \lesssim \epsilon^2$.

Proof. By the definition of the KL divergence,

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

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

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

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 $|1 - \lambda_i| \le \epsilon$ for all $i = 1, \ldots, p$. Therefore, the expression in (B.1) is bounded by a constant multiple of

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

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

Similarly, for the corresponding KL variation,

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

by adding and subtracting $[\mathsf{E}_{\Omega^\star}(X^{\mathrm{T}}(\Omega - \Omega^\star)X)]^2$. 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 B.2. *Given Ω, $\Omega^\star$, ν and $\nu^\star$, let Ω_ν denote the precision matrix for model (1.2) involving Ω and ν and $\Omega^\star_{\nu^\star}$ denote the truth of it with $\Omega^\star$ and $\nu^\star$. Then, if $|\nu-\nu^\star| \leq \epsilon/\sqrt{p}$ and $\|\Omega-\Omega^\star\|_F \leq \epsilon$ for small enough ϵ, we have*

$$\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F \leq K_1\epsilon,$$

where K_1 is a constant only depending on m, $\nu^\star$ and $\|\Omega^\star\|_2$. On the other hand, if $\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F \leq \epsilon$ and $\nu \leq M$ for some constant M, we have

$$|\nu-\nu^\star| \leq K_2\epsilon/\sqrt{p}, \quad \|\Omega-\Omega^\star\|_F \leq K_3\|\Omega^\star\|_2^2\epsilon, \tag{B.2}$$

where K_2 and K_3 are constants only depending on m and $\nu^\star$.

Proof. For simplicity, let $A = -(\nu\Omega + mI_p)^{-1}/\nu$ with the truth $A^\star$ involving $\Omega^\star$ and $\nu^\star$. Then, for the first inequality, write

$$\begin{aligned}\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F^2 &= m\|I_p/\nu + A - I_p/\nu^\star - A^\star\|_F^2 + m(m-1)\|A-A^\star\|_F^2 \\ &\leq 2m\|I_p/\nu - I_p/\nu^\star\|_F^2 + m(m+1)\|A-A^\star\|_F^2.\end{aligned}$$

The first term on the right hand side equals $2m\epsilon^2/\nu^2\nu^{\star 2}$, which is less than or equal to $4m\epsilon^2/\nu^{\star 4}$ because $\nu \geq \nu^\star/2$ if ϵ is small enough. Consider the second term except the front constant, which equals

$$\begin{aligned}&\|(\nu^\star\Omega^\star + mI_p)^{-1}/\nu^\star - (\nu\Omega + mI_p)^{-1}/\nu\|_F^2 \\ &\quad\leq 2\|(\nu^\star\Omega^\star + mI_p)^{-1}(1/\nu^\star - 1/\nu)\|_F^2 \\ &\qquad + 2\|(\nu^\star\Omega^\star + mI_p)^{-1}/\nu - (\nu\Omega + mI_p)^{-1}/\nu\|_F^2 \\ &\quad\leq 2\|(\nu^\star\Omega^\star + mI_p)^{-1}\|_2^2\|(I_p/\nu^\star - I_p/\nu)\|_F^2 \\ &\qquad + 4\|(\nu^\star\Omega^\star + mI_p)^{-1} - (\nu\Omega + mI_p)^{-1}\|_F^2/\nu^\star,\end{aligned}$$

where the first term on the right hand side is less than or equal to $4\epsilon^2/\nu^{\star 4}$ by the former calculation. It holds because the eigenvalues of $(\nu^\star\Omega^\star + mI_p)^{-1}$ and $(\nu\Omega + mI_p)^{-1}$ are all upper-bounded by 1. For the key in the second term except the constants, it equals

$$\|(\nu^\star\Omega^\star + mI_p)^{-1}(\nu\Omega - \nu^\star\Omega^\star)(\nu\Omega + mI_p)^{-1}\|_F^2 \leq \|\nu\Omega - \nu^\star\Omega^\star\|_F^2,$$

by the same reason as above. The right hand side is less than or equal to

$$\|\nu\Omega - \nu\Omega^\star\|_F^2 + \|\nu\Omega^\star - \nu^\star\Omega^\star\|_F^2 \leq 2\nu^\star\|\Omega-\Omega^\star\|_F^2 + \|\Omega^\star\|_F^2(\nu-\nu^\star)^2 \leq (2\nu^\star + \|\Omega^\star\|_2^2)\epsilon^2.$$

The first inequality holds because $\nu \leq 2\nu^\star$ if ϵ is small enough. Therefore, we proved the first assertion.

For the second result, write $\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F^2$ by entries and it equals

$$\begin{aligned}&\sum_{i,j} m(I_p/\nu - I_p/\nu^\star)_{ij}^2 + 2m(I_p/\nu - I_p/\nu^\star)_{ij}(A-A^\star)_{ij} + m^2(A-A^\star)_{ij}^2 \\ &\quad = (m-1)\sum_{i,j}(I_p/\nu - I_p/\nu^\star)_{ij}^2 + \sum_{i,j}[(I_p/\nu - I_p/\nu^\star)_{ij} + m(A-A^\star)_{ij}]^2\end{aligned}$$

$$= (m-1)\|I_p/\nu - I_p/\nu^\star\|_F^2 + \|I_p/\nu - I_p/\nu^\star + m(A - A^\star)\|_F^2. \tag{B.3}$$

If $m > 1$, since both terms on the right hand side are strictly positive, we have that $\|I_p/\nu - I_p/\nu^\star\|_F^2 \leq \epsilon^2/(m-1)$, which implies that

$$|\nu^\star - \nu| \leq \nu\nu^\star\epsilon/\sqrt{p(m-1)} \leq K_2\epsilon/\sqrt{p},$$

where K_2 only depends on m, $\nu^\star$ and M. On the other hand, notice that

$$\|\Omega_\nu - \Omega^\star_{\nu^\star}\|_F^2 = m\|I_p/\nu + A - I_p/\nu^\star - A^\star\|_F^2 + m(m-1)\|A - A^\star\|_F^2,$$

where both terms on the right hand side are positive. Therefore, we also have that $\|A - A^\star\|_F \leq \epsilon/\sqrt{m(m-1)}$, where the left hand side equals

$$\|(\nu^\star\Omega^\star + mI_p)^{-1}/\nu^\star - (\nu\Omega + mI_p)^{-1}/\nu^\star + (\nu\Omega + mI_p)^{-1}/\nu^\star - (\nu\Omega + mI_p)^{-1}/\nu\|_F.$$

By the triangle inequality, it implies that

$$\|(\nu^\star\Omega^\star + mI_p)^{-1}/\nu^\star - (\nu\Omega + mI_p)^{-1}/\nu^\star\|_F \leq \epsilon/\sqrt{m(m-1)} + K_2\epsilon/\sqrt{p}, \tag{B.4}$$

since the eigenvalues of $(\nu\Omega + mI_p)^{-1}$ are strictly upper-bounded by 1. Note that

$$\begin{aligned}
&\|\nu\Omega - \nu^\star\Omega^\star\|_F \\
&\quad = \|(\nu^\star\Omega^\star + mI_p)[(\nu^\star\Omega^\star + mI_p)^{-1} - (\nu\Omega + mI_p)^{-1}](\nu\Omega + mI_p)\|_F \\
&\quad \leq \|\nu^\star\Omega^\star + mI_p\|_2\|\nu\Omega + mI_p\|_2\|(\nu^\star\Omega^\star + mI_p)^{-1} - (\nu\Omega + mI_p)^{-1}\|_F,
\end{aligned}$$

which implies

$$\|(\nu^\star\Omega^\star + mI_p)^{-1} - (\nu\Omega + mI_p)^{-1}\|_F \geq \frac{\|\nu\Omega - \nu^\star\Omega^\star\|_F}{\|\nu\Omega + mI_p\|_2\|\nu^\star\Omega^\star + mI_p\|_2}.$$

By the triangle inequality, $\|\nu^\star\Omega^\star - \nu\Omega\|_F \geq \|\nu^\star\Omega^\star - \nu^\star\Omega\|_F - \|\nu^\star\Omega - \nu\Omega\|_F$ and thus,

$$\begin{aligned}
&\|\Omega - \Omega^\star\|_F \\
\leq& \sqrt{p}\|\Omega\|_2|\nu - \nu^\star|/\nu^\star + \|\nu\Omega + mI_p\|_2\|\nu^\star\Omega^\star + mI_p\|_2(\epsilon/\sqrt{m(m-1)} + K_2\epsilon/\sqrt{p}) \\
\leq& \|\Omega\|_2K_2\epsilon/\nu^\star + (m + 2\nu^\star\|\Omega\|_2)(m + \nu^\star\|\Omega^\star\|_2)[1/\sqrt{m(m-1)} + K_2/\sqrt{p}]\epsilon,
\end{aligned}$$

since $\nu \leq 2\nu^\star$ when ϵ is small enough. It therefore remains only to obtain an upper bound on $\|\Omega\|_2$. By the Woodbury formula, let $\delta = \epsilon/\sqrt{m(m-1)} + K_2\epsilon/\sqrt{p}$ and the left hand side of (B.4) equals

$$\|(\Omega^{-1} + \nu I_p/m)^{-1} - (\Omega^{\star-1} + \nu^\star I_p/m)^{-1}\|_F/m^2 \leq \delta.$$

By the triangle inequality and the relation $\|\cdot\|_2 \leq \|\cdot\|_F$, we get that

$$\|(\Omega^{-1} + \nu I_p/m)^{-1}\|_2 \leq \|(\Omega^{-1} + \nu I_p/m)^{-1} - (\Omega^{\star-1} + \nu^\star I_p/m)^{-1}\|_2$$

$$
\begin{aligned}
&+ \|(\Omega^{\star-1} + \nu^\star I_p/m)^{-1}\|_2 \\
\leq& m^2\delta + \|(\Omega^{\star-1} + \nu^\star I_p/m)^{-1}\|_2.
\end{aligned}
$$

With some relatively straightforward algebra, it follows that

$$
\begin{aligned}
\|\Omega\|_2 &\leq \frac{m^2\delta + \|\Omega^\star\|_2}{1 - m\delta(\nu^\star + K_2\epsilon/\sqrt{p}) - (\nu^\star + K_2\epsilon/\sqrt{p})\|\Omega^\star\|_2/(m + \nu^\star\|\Omega^\star\|_2)} \\
&\lesssim m^2\delta + \|\Omega^\star\|_2,
\end{aligned}
$$

since ϵ can be as small as we need. Finally, the second bound in (B.2) follows from the above display and (B.4) for $m > 1$.

If $m = 1$ and $\nu^\star$ is known, the right hand side of (B.3) gives us that $\|A - A^\star\|_F \leq \epsilon$. By similar calculation thereafter as above, we obtain the same bound in (B.2). □

References

[1] Banerjee, O., Ghaoui, L. E. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* **9** 485–516. MR2417243

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

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

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

[5] Byrd, M., Nghiem, L. H. and McGee, M. (2021). Bayesian regularization of Gaussian graphical models with measurement error. *Computational Statistics & Data Analysis* **156** 107085. MR4182924

[6] Cai, T., Liu, W. and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106** 594–607. MR2847973

[7] Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics* **35** 2313–2351. MR2382644

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

[9] Castillo, I. and van der Vaart, A. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* **40** 2069–2101. MR3059077

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

[11] Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association* **89** 1314–1328. MR3273616

[12] Du, X. and Ghosal, S. (2018). Bayesian discriminant analysis using a high dimensional predictor. *Sankhya A. The Indian Journal of Statistics* **80** S112–S145. MR3968360

[13] Fan, J., Fan, Y. and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics* **147** 186–197. MR2472991

[14] Fan, J., Feng, Y. and Wu, Y. (2009). Network exploration via the adaptive LASSO and SCAD penalties. *The Annals of Applied Statistics* **3** 521. MR2750671

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

[16] Fan, J., Liao, Y. and Mincheva, M. (2011). High dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics* **39** 3320. MR3012410

[17] Freedman, L. S., Midthune, D., Carroll, R. J. and Kipnis, V. (2008). A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine* **27** 5195–5216. MR2516750

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

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

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

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

[22] Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Nonparametric Bayesian Inference. Cambridge Series in Statistical and Probabilistic Mathematics* **44**. Cambridge University Press, Cambridge. MR3587782

[23] Jeong, S. and Ghosal, S. (2021). Unified Bayesian theory of sparse linear regression with nuisance parameters. *Electronic Journal of Statistics* **15** 3040–3111. MR4280166

[24] Lauritzen, S. L. (1996). *Graphical Models. Oxford Statistical Science Series* **17**. The Clarendon Press, Oxford University Press, New York Oxford Science Publications. MR1419991

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

[26] Liang, F., Jia, B., Xue, J., Li, Q. and 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** 899. MR3874303

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

[28] Mohammadi, A. and Wit, E. C. (2015). Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis* **10** 109–138. MR3420899

[29] Ning, B., Jeong, S. and Ghosal, S. (2020). Bayesian linear regression for multivariate responses under group sparsity. *Bernoulli* **26** 2353–2382. MR4091112

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

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

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

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