Bayesian Scalable Precision Factor Analysis for Massive Sparse Gaussian Graphical Models

Noirrit Kiran Chandra[a] (noirrit.chandra@utdallas.edu)
Peter Müller[b,c] (pmueller@math.utexas.edu)
Abhra Sarkar[b] (abhra.sarkar@utexas.edu)

[a]Department of Mathematical Sciences,
The University of Texas at Dallas,
800 W. Campbell Rd, Richardson, TX 75080-3021, USA

[b]Department of Statistics and Data Sciences,
The University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA

[c]Department of Mathematics,
The University of Texas at Austin,
2515 Speedway, PMA 8.100, Austin, TX 78712-1823, USA

Abstract

We propose a novel approach to estimating the precision matrix of multivariate Gaussian data that relies on decomposing them into a low-rank and a diagonal component. Such decompositions are very popular for modeling large covariance matrices as they admit a latent factor based representation that allows easy inference. The same is however not true for precision matrices due to the lack of computationally convenient representations which restricts inference to low-to-moderate dimensional problems. We address this remarkable gap in the literature by building on a latent variable representation for such decomposition for precision matrices. The construction leads to an efficient Gibbs sampler that scales very well to high-dimensional problems far beyond the limits of the current state-of-the-art. The ability to efficiently explore the full posterior space also allows the model uncertainty to be easily assessed. The decomposition crucially additionally allows us to adapt sparsity inducing priors to shrink the insignificant entries of the precision matrix toward zero, making the approach adaptable to high-dimensional small-sample-size sparse settings. Exact zeros in the matrix encoding the underlying conditional independence graph are then determined via a novel posterior false discovery rate control procedure. A near minimax optimal posterior concentration rate for estimating precision matrices is attained by our method under mild regularity assumptions. We evaluate the method's empirical performance through synthetic experiments and illustrate its practical utility in data sets from two different application domains.

Key Words: Factor models, False discovery rate control, Gaussian graphical models, Markov chain Monte Carlo, Posterior concentration, Precision matrix estimation, Scalable computation, Shrinkage priors

Short/Running Title: Precision Factor Analysis

Corresponding Author: Abhra Sarkar (abhra.sarkar@utexas.edu)

1 Introduction

For multivariate Gaussian distributed data $\mathbf{y} = (y_1, \ldots, y_d)^{\mathrm{T}} \sim \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$, all conditional dependence information is contained in the inverse covariance matrix $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega} = ((\omega_{j,j'}))$, or the precision. Two 'nodes' y_j and $y_{j'}$ are conditionally independent given the rest if and only if $\omega_{j,j'} = 0$. The underlying conditional (in)dependence graph is then obtained by connecting the pairs of nodes $\{(j, j') : \omega_{j,j'} \neq 0\}$ by undirected 'edges'. Estimating the precision matrix, including especially its sparsity patterns, for such data is therefore an important statistical problem (Lauritzen, 1996; Koller and Friedman, 2009).

In this article, we model $\boldsymbol{\Omega}$ via a *low-rank and diagonal* (LRD) decomposition. Bhattacharya *et al.* (2016) introduced a representation for efficient sampling from Gaussian distributions with known LRD structured precision matrices. In this article, we adapt the representation in a different way to obtain a novel factor analytic framework for unknown precision matrices modeled via LRD decompositions with applications to graphical models. While such decomposition is already a widely used standard tool for modeling high-dimensional $\boldsymbol{\Sigma}$, to our knowledge, the construction has not been utilized before for modeling $\boldsymbol{\Omega}$. Our research addresses this remarkable gap in the literature. Although any positive definite matrix can always be factorized this way, the main challenge is to introduce such constructions for $\boldsymbol{\Omega}$ in a way that allows efficient and scalable posterior inference. Going significant steps further, we also adapt this approach to sparse high-dimensional settings. Noting that any sparse $\boldsymbol{\Omega}$ can always be represented with a sparse factorization, we impose sparsity in $\boldsymbol{\Omega}$ by using sparsity-inducing priors for the factorization.

Existing Methods for Covariance and Precision Matrix Estimation: The existing literature on sparse covariance and precision matrix estimation is vast. In

sparse covariance matrix estimation problems, the entire matrix $\boldsymbol{\Sigma}$ is often directly penalized (Levina *et al.*, 2008; Bien and Tibshirani, 2011). In an alternative approach, $\boldsymbol{\Sigma}$ is assumed to admit a LRD structure $\boldsymbol{\Sigma} = \widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^{\mathrm{T}} + \widetilde{\boldsymbol{\Delta}}$ where $\widetilde{\boldsymbol{\Lambda}}$ is a $d \times q$ order matrix and $\widetilde{\boldsymbol{\Delta}}$ is a diagonal matrix with all positive entries. In theory, all positive definite matrices admit such a representation for some $0 \leq q \leq d$, and in practice, $q \ll d$ often suffice to produce good approximations. Also importantly, this model admits the latent variable representation $\mathbf{y} = \widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{u}} + \widetilde{\mathbf{v}}$ where $\widetilde{\mathbf{u}} \sim \mathrm{N}_q(\mathbf{0}, \mathbf{I}_q)$ and $\widetilde{\mathbf{v}} \sim \mathrm{N}_d(\mathbf{0}, \widetilde{\boldsymbol{\Delta}})$. The formulation allows massive scalability in computation, making LRD based methods popular in the high-dimensional covariance matrix estimation literature (Fan *et al.*, 2011, 2018; Daniele *et al.*, 2019), especially in Bayesian settings (Bhattacharya and Dunson, 2011; Pati *et al.*, 2014; Zhu *et al.*, 2014; Kastner, 2019, and others). A sparse $\boldsymbol{\Sigma}$, however, does not usually produce a sparse $\boldsymbol{\Omega}$ and the strategy of inverting the estimated $\boldsymbol{\Sigma}$ to obtain an estimate of $\boldsymbol{\Omega}$ tends to exhibit poor empirical performance (Pourahmadi, 2013).

As in covariance matrix estimation problems, penalized likelihood based methods that directly penalize the number and/or absolute values of non-zero entries in $\boldsymbol{\Omega}$ have also been developed in the frequentist setting (Yuan and Lin, 2007; Banerjee *et al.*, 2008; Rothman *et al.*, 2008; d'Aspremont *et al.*, 2008; Friedman *et al.*, 2008; Witten *et al.*, 2011; Mazumder and Hastie, 2012; Zhang and Zou, 2014). Alternatively, Meinshausen and Bühlmann (2006); Peng *et al.* (2009) developed neighborhood selection methods that learn the edges by regressing each variable on the rest with penalties on large regression coefficients.

When the main focus is on estimating the underlying dependence graph, Bayesian approaches instead rely on defining a hierarchical prior on $\boldsymbol{\Omega}$ preceded by a prior on the graph. Choices for the latter include uniform priors over graph sizes (Armstrong

et al., 2009), priors centered around some informed location (Mitra *et al.*, 2013), priors with edge inclusions following a binomial distribution (Dobra *et al.*, 2004; Carvalho and Scott, 2009), a hyper-Markov distribution on decomposable graphs (Dawid and Lauritzen, 1993), its generalizations to non-decomposable settings (Roverato, 2002; Khare *et al.*, 2018), etc. However, such hierarchical construction with a separate model layer for the underlying graph structure makes posterior exploration quite challenging. Markov chain Monte Carlo (MCMC) algorithms have been designed specifically for such models (Dellaportas *et al.*, 2003; Atay-Kayis and Massam, 2005; Carvalho *et al.*, 2007; Dobra *et al.*, 2011; Green and Thomas, 2013; Lenkoski, 2013; Mohammadi and Wit, 2015, and others) but these strategies still rely on expensive local exploration moves, often involving trans-dimensional proposals in the graph space and/or approximations of intractable normalizing constants, hence remaining computationally infeasible beyond only a few tens of dimensions (Jones *et al.*, 2005).

Bayesian methods that directly penalize $\mathbf{\Omega}$, thereby avoiding to have to specify a separate prior for the underlying graph, have started to get some attention but the literature remains sparse. Yoshida and West (2010) proposed a factor model with complex constraints enforcing identical sparsity patterns in the covariance and precision matrices which may be restrictive in practice while also having scalability issues. Banerjee and Ghosal (2015) studied spike and slab type priors (Ishwaran and Rao, 2005) to shrink irrelevant off-diagonals to zero. For point mass mixture priors, MCMC based model space exploration can generally be daunting and may lead to slow mixing and convergence even in simple mean and linear regression problems. Continuous shrinkage priors (Polson and Scott, 2010) that allow fast and efficient posterior exploration have also been adapted for precision matrices. Wang (2012); Khondker *et al.* (2013) and Li *et al.* (2019a) designed block Gibbs samplers that allow updating entire

columns of $\boldsymbol{\Omega}$ at once. Mohammadi *et al.* (2021) proposed an approximated sampler that can scale up to a few hundreds but the problem remains infeasible for modern applications with many thousands of nodes. More recent developments along these lines (Gan *et al.*, 2019; Li *et al.*, 2019b; Deshpande *et al.*, 2019) have focused on fast deterministic Expectation-Maximization (EM) algorithms instead, which scale well to problems with a few hundred dimensional nodes but only estimate the posterior mode (MAP) and not the full posterior. Ksheera Sagar *et al.* (2021) studied both MCMC and MAP estimation for element-wise horseshoe like priors on the precision matrix and studied their convergence properties. The approach, however, suffers from similar scalability issues. Also, in many of these approaches (Friedman *et al.*, 2008; Peng *et al.*, 2009), the estimated $\boldsymbol{\Omega}$ is not guaranteed to be positive definite, requiring post-hoc analysis to fix the estimate.

Owing to the lack of computationally tractable hierarchical structures, in high-dimensions, posterior explorations in existing Bayesian precision matrix and graph estimation methods thus still remain prohibitively expensive if not entirely impossible. With the few exceptions, existing Bayesian approaches also do not come with rigorous theoretical guarantees.

Our Proposed Precision Factor Model: In contrast to covariance matrices, there are currently no flexible LRD decomposition methods that admit easily tractable latent variable representations for precision matrices, posing significant methodological and computational challenges, especially for MCMC based Bayesian inference.

In this article, we derive a factor analytic framework building on a representation for Gaussian distributions with an LRD decomposed precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta}$ for some $d \times q$ order matrix $\boldsymbol{\Lambda}$ and some diagonal matrix $\boldsymbol{\Delta}$ (Bhattacharya *et al.*, 2016). The representation is immediately useful in facilitating efficient posterior simulation

in Bayesian inference of $\boldsymbol{\Omega}$ and easily scales to problems with dimensions d far beyond the limits of the current state-of-the-art.

Since all positive definite matrices admit LRD representations ($q = 0$ and $q = d$ being the two extremes), the proposed approach imposes no restrictive assumptions on the precision matrix or the underlying graph. Additionally, we discuss a constructive approach to find an LRD representation of arbitrary sparse $\boldsymbol{\Omega}$. In simulation examples, we show that $q \ll d$ often suffices to produce good approximations of $\boldsymbol{\Omega}$ even when they do not exactly admit an LRD for such q. Conversely, since the representation always produces a positive definite matrix, unlike many existing procedures such as Meinshausen and Bühlmann (2006); Friedman *et al.* (2008); Gan *et al.* (2019), we obtain positive definite estimates simply by design.

As the off-diagonals elements of $\boldsymbol{\Omega}$ are contributed entirely by $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$, a sparse $\boldsymbol{\Lambda}$ is expected to produce a sparse $\boldsymbol{\Omega}$ (see Figure 3 and Section 2.2.2). A suitably chosen penalty on $\boldsymbol{\Lambda}$ therefore allows to flexibly adapt its non-zero elements such that insignificant off-diagonals of $\boldsymbol{\Omega}$ are shrunk towards zero. The strategy has been successfully employed in high-dimensional sparse covariance matrix estimation literature in both Bayesian (Pati *et al.*, 2014) and frequentist paradigms (Daniele *et al.*, 2019).

In this article, we adapt the Dirichlet-Laplace shrinkage priors (Bhattacharya *et al.*, 2015) on $\boldsymbol{\Lambda}$ for their theoretical and computational tractability. As an artifact of Bayesian methods with continuous shrinkage priors, exact zeroes do not appear in the posterior samples. Thus, we address the problem of non-zero off-diagonal/edge selection from the posterior MCMC samples through a novel multiple hypothesis testing approach that allows the posterior false edge discovery rate (FDR) (Chandra and Bhattacharya, 2019) to be controlled at any desired level. This is another salient feature of our proposed method that properly accommodates posterior uncertainty in

reporting a point estimate for the graph.

With our model and prior specifications, we establish near minimax-optimal contraction rates of the posterior around the true $\mathbf{\Omega}$ under mild assumptions when the number of nodes increases exponentially with the sample size. We evaluate the proposed method's finite sample efficacy through simulation experiments where it either outperforms or is competitive with previously existing methods in moderately high-dimensional problems while also scaling to dimensions far beyond the reach of many of those methods. We illustrate our method's practical utility in real data sets from two different application domains.

Our Key Contributions: Overall, our main contributions to the literature include (a) proposing a likelihood based approach built on a low-rank and diagonal decomposition of the precision matrix, (b) build on a latent factor representation for such decomposition that provides new interpretations for such models, (c) designing a Gibbs sampling algorithm that exploits this latent factor representation and scales very well to high-dimensional problems, (d) adapting shrinkage priors for the low-rank component that further makes the method applicable to high-dimensional small-sample-size sparse settings, (e) developing a novel FDR control procedure for graph selection from the posterior samples of the precision matrix, and (f) establishing rigorous asymptotic properties of the posterior of the proposed approach.

Outline of the Article: The rest of this article is organized as follows. Section 2 details our matrix decomposition based model. Section 2.1 discusses our novel factor analytic representation; Section 2.2 discusses how a low-rank approach can be used to learn arbitrary sparse precision matrices; Section 2.3 discusses the priors; Section 2.4 describes the posterior sampling algorithm; Section 2.5 presents our graph selection procedure via FDR control; Section 2.6 discusses the posterior's asymptotic

properties. Section 3 summarizes the results of simulation experiments. Section 4 presents the results for two real data sets. Section 5 contains concluding remarks.

2 Gaussian Precision Factor Models

We consider a random sample of n observations assumed to be independently and identically distributed (iid) d-dimensional random vectors following a multivariate Normal distribution with mean zero and precision matrix $\mathbf{\Omega}$ as

$$\mathbf{y}_i \overset{\text{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \mathbf{\Omega}^{-1}), \quad i = 1, \ldots, n, \tag{1}$$

where $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,d})^{\mathrm{T}}$. The primary goal is to estimate $\mathbf{\Omega}$, especially identifying its sparsity pattern that characterizes conditional independence relationships between different components of $\mathbf{y}$.

To this end, we consider an LRD decomposition of $\mathbf{\Omega}$ as

$$\mathbf{\Omega} = \mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} + \mathbf{\Delta}, \tag{2}$$

where $\mathbf{\Lambda}^{d\times q} = ((\lambda_{r,c}))$ and $\mathbf{\Delta} = \mathrm{diag}(\delta_1^2, \ldots, \delta_d^2)$. Factorization (2) is completely flexible in the sense that a matrix is positive definite if and only if it admits such a representation for some $0 \leq q \leq d$. Any precision matrix $\mathbf{\Omega}$ can therefore be written as (2) for a sufficiently large value of q. For most practical cases, values of $q \ll d$ suffice to approximate $\mathbf{\Omega}$ well, resulting in a huge reduction in dimensions and allowing massive scalability in computation. In model (2), $\mathbf{\Lambda}$ is not strictly identifiable since, for example, $\mathbf{\Lambda}\mathbf{\Lambda}^{\mathrm{T}} = \mathbf{\Lambda}\mathbf{Q}\mathbf{Q}^{\mathrm{T}}\mathbf{\Lambda}^{\mathrm{T}}$ for any orthogonal matrix $\mathbf{Q}$. Inference on $\mathbf{\Omega}$ being the primary interest, individual identifiability or interpretation of the model parameters in (2) is, however, not required.

2.1 Factor Analytic Representation

One main advantage of modeling covariance matrices via LRD decompositions is

the existence of a latent factor representation that greatly facilitates computation. Tractable latent factor representations are, however, not known to exist for precision matrices, presenting major barriers against efficient inference. Recently Bhattacharya *et al.* (2016) introduced a representation that greatly facilitates fast sampling from multivariate Gaussian distributions with known precision matrices with an LRD structure. We show that, by exploiting this representation in a different way, a latent factor model can in fact be obtained for LRD decomposed precision matrices as well that allows efficient posterior inference of its unknown components in much the same way as classical factor models facilitate computation for unknown LRD decomposed covariance matrices. The representation, formalized in Proposition 1 below and referred to in this article as the 'precision factor model', also provides novel insights into the very construction of Gaussian graphical models.

Proposition 1. *The model* $\mathbf{y}_i \overset{iid}{\sim} \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Omega}^{-1}), i = 1, \ldots, n$, *with* $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta}$, *where* $\boldsymbol{\Lambda}$ *is* $d \times q$ *with* $q \leq d$ *and* $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1^2, \ldots, \delta_d^2)$, *admits the equivalent representation*

$$\mathbf{y}_i = -\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda})^{-1}\mathbf{u}_i + \mathbf{v}_i, \tag{3}$$

$$\textit{where} \quad \begin{bmatrix} \mathbf{u}_i \\ \mathbf{v}_i \end{bmatrix} \overset{iid}{\sim} \mathrm{N}_{q+d}\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda} & \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1} \\ \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda} & \boldsymbol{\Delta}^{-1} \end{bmatrix}\right). \tag{4}$$

The proposition follows straightforwardly using Sherman-Woodbury identity for the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$ given by

$$\boldsymbol{\Omega}^{-1} = (\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta})^{-1} = \boldsymbol{\Delta}^{-1} - \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}(\mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}.$$

Note that $\mathrm{cov}(\mathbf{y}_i, \mathbf{u}_i) = \mathbf{0}$ in (3). A reverse-engineered construction that is particularly useful for posterior simulation in Bayesian settings utilizes this fact and generates $\mathbf{u}_i \overset{\mathrm{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{P})$ with $\mathbf{P} = (\mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda})$ independent of $\mathbf{y}_{1:n}$ first and then sets $\mathbf{v}_i = \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i + \mathbf{y}_i$. Clearly then, $\mathbf{v}_i \overset{\mathrm{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Delta}^{-1})$. Also,

$$\begin{bmatrix}\mathbf{u}_i\\ \mathbf{y}_i\end{bmatrix} \overset{\text{iid}}{\sim} \mathrm{N}_{q+d}\left(\mathbf{0}, \begin{bmatrix}\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \boldsymbol{\Omega}^{-1}\end{bmatrix}\right) \text{ then implies } \begin{bmatrix}\mathbf{u}_i\\ \mathbf{v}_i\end{bmatrix} \overset{\text{iid}}{\sim} \mathrm{N}_{q+d}\left(\mathbf{0}, \begin{bmatrix}\mathbf{P} & \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\\ \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda} & \boldsymbol{\Delta}^{-1}\end{bmatrix}\right),$$

as in (4) in Proposition 1. Importantly, we can also write

$$\mathbf{u}_i = \boldsymbol{\Lambda}^{\mathrm{T}}\mathbf{v}_i + \boldsymbol{\varepsilon}_i, \text{ where } \boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{I}_q). \tag{5}$$

Although mathematically simple, Proposition 1 has far-reaching implications. An efficient and highly scalable Gibbs sampler follows immediately from the construction by first generating the latent vectors $\mathbf{u}_i, \mathbf{v}_i$ given $\boldsymbol{\Lambda}, \boldsymbol{\Delta}$ and $\mathbf{y}_i$ as described above, and then updating the rows of $\boldsymbol{\Lambda}$ given $\mathbf{u}_{1:n}$ and $\mathbf{v}_{1:n}$ using (5).

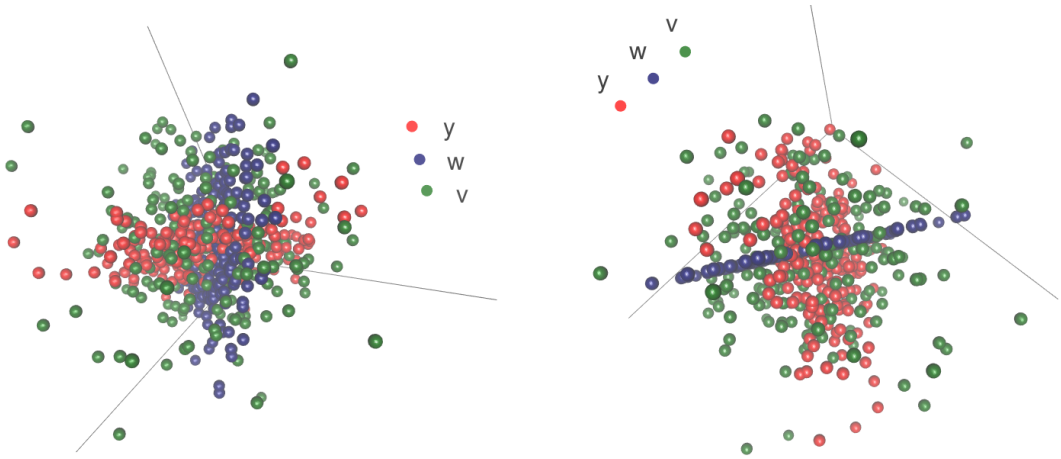


Figure 1: Graphical view of the precision factor model: Plots of observed $\mathbf{y}_i$, latent $\mathbf{w}_i = \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i$ and latent $\mathbf{v}_i = \mathbf{w}_i + \mathbf{y}_i$ for a model with dimension $d = 3$ and rank $q = 2$ from different viewing angles. The d-dimensional vectors $\mathbf{w}$ are supported on a lower q-dimensional plane. Also, $\mathbf{u}$ and $\mathbf{y}$ are independent of each other. Since $\mathbf{w}$ are linear transformations of $\mathbf{u}$, they are also independent of $\mathbf{y}$. In the plots, $\mathbf{w}$ and the average values of $\mathbf{y}$ can be seen to be living on planes that are orthogonal to each other. The vectors $\mathbf{v}$ are more scattered than $\mathbf{y}$ and, in obtaining the variance-covariance of $\mathbf{y} = \mathbf{v} - \mathbf{w}$, the larger variances of $\mathbf{v}$ are perfectly compensated by the negative covariances between $\mathbf{v}$ and $\mathbf{w}$.

Continuing the thread at the beginning of this subsection, parallels can be drawn with latent factor models for covariance matrices where a similar decomposition $\boldsymbol{\Sigma} = \widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{\Lambda}}^{\mathrm{T}} + \widetilde{\boldsymbol{\Delta}}$ arises from the model $\mathbf{y}_i = \widetilde{\boldsymbol{\Lambda}}\widetilde{\mathbf{u}}_i + \widetilde{\mathbf{v}}_i$ with independent latent components $\widetilde{\mathbf{u}}_i$ and $\widetilde{\mathbf{v}}_i$, where the latent factors $\widetilde{\mathbf{u}}_i \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{I}_q)$, with q typically $\ll d$, and the

errors $\widetilde{\mathbf{v}}_i \overset{\text{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \widetilde{\boldsymbol{\Delta}})$. The covariance between the components of $\mathbf{y}_i$ and a part of the variance of $\mathbf{y}_i$ are thus explained by (a) the variance-covariance of the latent factors $\widetilde{\mathbf{u}}_i$ (b) while the remaining unexplained variance is attributed to the errors $\widetilde{\mathbf{v}}_i$.

In contrast, for the precision factor model depicted in Figure 1, we have from model (3) that $\mathbf{y}_i = -\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i + \mathbf{v}_i$ with dependent latent components $\mathbf{u}_i$ and $\mathbf{v}_i$, where the latent factors $\mathbf{u}_i \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{P})$, with q expected again to be $\ll d$, and the 'errors' $\mathbf{v}_i \overset{\text{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Delta}^{-1})$. The variance-covariance of $\mathbf{y}_i$ is thus explained by (a) the variance-covariance of the latent factors $\mathbf{u}_i$, (b) the variance of the 'errors' $\mathbf{v}_i$, and (c) the covariance between $\mathbf{u}_i$ and $\mathbf{v}_i$.

If independence between the variance contributing components is desired, an alternative view $\mathbf{v}_i = \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i + \mathbf{y}_i$ of model (3) is to see the latent $\mathbf{v}_i$'s be composed of independent components $\mathbf{u}_i$ and $\mathbf{y}_i$, where the latent factors $\mathbf{u}_i \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{P})$, with $q \ll d$ as before, and the 'errors' $\mathbf{y}_i \overset{\text{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Sigma})$. The $\mathbf{v}_i$'s can thus be represented in an orthogonal decomposition with components $\mathbf{w}_i = \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i$ and $\mathbf{y}_i$. The larger variances $\boldsymbol{\Delta}^{-1}$ of $\mathbf{v}_i$ are now being explained by (a) the variance-covariance of the latent factors $\mathbf{u}_i$ and (b) the variance-covariance of the 'errors' $\mathbf{y}_i$, the off-diagonal covariance terms of these components perfectly cancelling each other to produce the diagonal matrix $\boldsymbol{\Delta}^{-1}$.

In yet another view based on equation (5), the vectors $\mathbf{v}_i$ can instead be interpreted as independent heterogeneous latent factors and $\mathbf{u}_i$ the associated response vectors subject to white noises $\boldsymbol{\varepsilon}_i$. Contrary to the classical factor model, in this view, we have the factor dimension $d \gg$ the response dimension q. Clearly, $\boldsymbol{\Omega}^{-1}$ is now the conditional variance-covariance of $\mathbf{v}_i$ given $\mathbf{u}_i$, that is, given the response variables $\mathbf{u}_i$, we can study the behavior of the latent factors $\mathbf{v}_i$ by examining $\boldsymbol{\Omega}$.

Although the precision factor model described here has immediate practical im-

plications for Bayesian inference of Gaussian precision matrices considered in this article, the representation itself is not specific to the chosen inferential paradigm but is a mathematical identity that may be of broad general interest in understanding the basic construction of such models.

2.2 Low-rank Modeling of Sparse Precision Matrices

In this section we discuss how a sparse precision matrix can indeed admit an LRD decomposition. Then we discuss our strategy of inducing sparsity in high-dimensional precision matrix estimation problems.

2.2.1 LRD Decomposition of Sparse Precision Matrices

We first discuss how a sparse $\mathbf{\Omega}$ can admit an LRD decomposition. To get some insights, consider the stylized example from Figure 2 where $d = 5$ and $\mathbf{\Omega}$ has $\widetilde{s} = 3$ non-zero off-diagonals. We recall the factor analytic representation of our model from Section 2.1 where we write $\mathbf{u} = \mathbf{\Lambda}^{\mathrm{T}}\mathbf{v} + \boldsymbol{\varepsilon}$ in equation (5) and interpret $\mathbf{\Omega} = ((\omega_{j,h}))$ as the precision matrix of $\mathbf{v}$ conditionally on $\mathbf{u}$. Under this representation, $\omega_{j,h} \neq 0$ if and only if $\boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_h \neq 0$, that is, if v_j and v_h are connected to the same u_ℓ, $1 \leq \ell \leq q$.

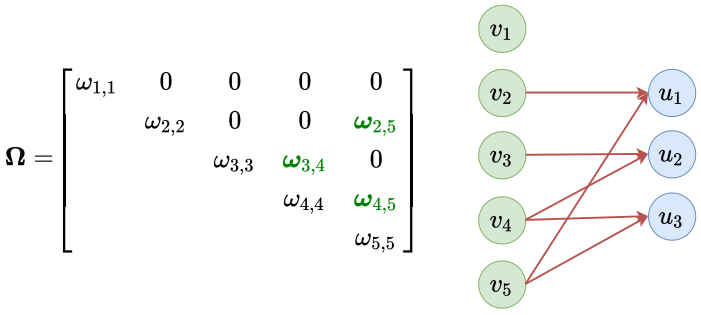


Figure 2: Constructing an LRD decomposition of a sparse $\mathbf{\Omega}^{5\times5}$: For each non-zero $\omega_{j,h}$, we connect v_j and v_h to u_ℓ. Notably, y_1 is marginally independent to every other variable and therefore we do not connect v_1 to any u_ℓ.

Likewise, we can construct a sparse $\mathbf{\Lambda}$ as follows

$$\underbrace{\begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}}_{\mathbf{u}} = \underbrace{\begin{bmatrix} 0 & \lambda_{2,1} & 0 & 0 & \lambda_{5,1} \\ 0 & 0 & \lambda_{3,2} & \lambda_{4,2} & 0 \\ 0 & 0 & 0 & \lambda_{4,3} & \lambda_{5,3} \end{bmatrix}}_{\mathbf{\Lambda}^{\mathrm{T}}} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \end{bmatrix}}_{\mathbf{v}} + \underbrace{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix}}_{\boldsymbol{\varepsilon}},$$

subject to $\lambda_{2,1}\lambda_{5,1} = \omega_{2,5}$, $\lambda_{3,2}\lambda_{4,2} = \omega_{3,4}$, $\lambda_{4,3}\lambda_{5,3} = \omega_{4,5}$ and $\delta_j^2 + \boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_j = \omega_{j,j}$ for all $j = 1, \ldots, d$. Although this construction is not unique, we see that a sparse LRD decomposition indeed exists for our concerned $\mathbf{\Omega}$.

We can generalize this argument for arbitrary sparse $\mathbf{\Omega}$ with $\widetilde{s}$ number of non-zero off-diagonals. We start with a $d \times q$ order $\mathbf{\Lambda}$ with all entries equal to zero. Let $\mathcal{E}$ be the set of edges in the conditional dependence graph. Note that each edge corresponds to a partial correlation between Y_j and Y_h, $1 \leq j, h \leq d$. For an edge $\omega_{j,h} \in \mathcal{E}$, we set λ_{j,ℓ_1} and λ_{h,ℓ_1} to be non-zero for some ℓ_1 so that $\boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_h \neq 0$. For the next edge $\omega_{j',h'} \in \mathcal{E}$, suppose $\omega_{j,j'} = \omega_{j,h'} = \omega_{h,h'} = 0$. Then we set λ_{j,ℓ_2} and λ_{h,ℓ_2} to be non-zero for some $\ell_2 \neq \ell_1$ so that $\boldsymbol{\lambda}_{j'}^{\mathrm{T}}\boldsymbol{\lambda}_{h'} \neq 0$ but $\boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_{j'} = \boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_{h'} = \boldsymbol{\lambda}_h^{\mathrm{T}}\boldsymbol{\lambda}_{h'} = 0$, and so on. Note that in this construction (which need not be unique), for each edge $\omega_{j,j'} \neq 0$, we need to connect v_j and $v_{j'}$ to a u_ℓ. Therefore, we do not need q to be larger than $\widetilde{s}$. The existence of such a $\mathbf{\Lambda}$ and $\mathbf{\Delta}$ is subject to the solution of the following system of quadratic equations on $\{\mathrm{vec}(\mathbf{\Lambda}), \boldsymbol{\delta}_{1:d}\}$ with $dq + d$ unknowns and $d + \widetilde{s}$ equations

$$\boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_h = \omega_{j,h} \text{ for all } \omega_{j,h} \in \mathcal{E}, \quad \delta_j^2 + \boldsymbol{\lambda}_j^{\mathrm{T}}\boldsymbol{\lambda}_j = \omega_{j,j} \text{ for all } j = 1, \ldots, d.$$

In practical high-dimensional applications, even though d is large, the conditional dependence graph is often sparse, that is, $\widetilde{s} \ll d$. Barvinok and Rudelson (2022) noted that when the number of unknown variables are considerably larger than the number of equations, a (not necessarily unique) solution to quadratic system of equations exists under very generic conditions.

2.2.2 Inducing Sparsity in $\boldsymbol{\Omega}$ via Sparsity in $\boldsymbol{\Lambda}$

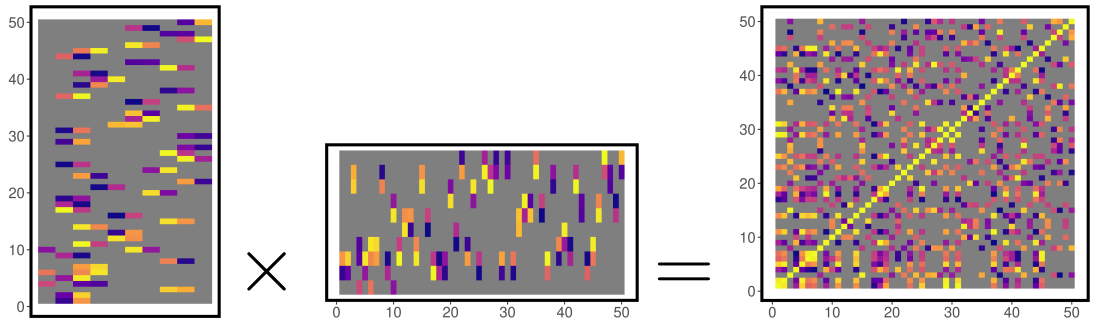


Figure 3: Sparse $\boldsymbol{\Lambda}$ producing a sparse $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$. Here, the gray cells represent exact zeros, and purple to yellow represent smaller (negative) to larger (positive) values.

The off-diagonals elements of $\boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta}$ are contributed entirely by $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}}$. This allows to achieve sparsity in $\boldsymbol{\Omega}$ by inducing sparsity in $\boldsymbol{\Lambda}$. Following the discussion in Section 2.2.1, a carefully designed data-adaptive penalty on $\boldsymbol{\Lambda}$ (e.g., a shrinkage prior on $\boldsymbol{\Lambda}$ that sufficiently increases the probability of obtaining a sparse $\boldsymbol{\Omega}$) would induce sparsity in $\boldsymbol{\Lambda}$ in such a (not necessarily unique) way that the sparsity patterns in $\boldsymbol{\Omega}$ are also accurately recovered. In this section, we show that by inducing sparsity in $\boldsymbol{\Lambda}^{d\times q} = ((\lambda_{j,h}))$, sparsity is also induced in $\boldsymbol{\Omega}^{d\times d} = ((\omega_{j,j'}))$. We begin with the spike-and-slab prior (Ishwaran and Rao, 2005) that allows exact zeros via a spike at zero and large nonzero elements via a continuous slab $g(\cdot)$ as

$$\lambda_{j,h} \mid \pi, \theta_{j,h} \sim \pi \mathbb{1}_{\{0\}}(\cdot) + (1-\pi) g(\cdot), \quad \pi \sim \mathrm{Beta}(a_\pi, b_\pi),$$

where π is the prior probability of observing a zero for $\lambda_{j,h}$. For such priors $\Pi(\omega_{j,j'} = 0 \mid \pi)$ can be analytically reduced to a simple form that helps gain insights into how inducing sparsity in $\boldsymbol{\Lambda}$ can induce sparsity in $\boldsymbol{\Omega}$ with high probability. Specifically, since $\omega_{j,j'} = \boldsymbol{\lambda}_j \boldsymbol{\lambda}_{j'}^{\mathrm{T}}$, where $\boldsymbol{\lambda}_j$ is the j^{th} row of $\boldsymbol{\Lambda}$, $\omega_{j,j'} = 0$ can only happen when the r^{th} entries of $\boldsymbol{\lambda}_j$ and $\boldsymbol{\lambda}_{j'}$ are both not from the slab distribution $g()$ for all $r = 1, \ldots, q$. Therefore, we have

$$\Pi(\omega_{j,j'} = 0 | \pi) = \textstyle\sum_{r=0}^{q} \left\{ \binom{q}{r} \pi^r (1-\pi)^{q-r} \right\} \times \pi^{q-r} = \pi^q (2-\pi)^q.$$

The behavior of $\Pi(\omega_{j,j'} = 0 \mid \pi)$ for varying values of π and q are shown in Figure 4. It can be seen that by controlling π, it is possible to induce any desired level of sparsity in $\boldsymbol{\Omega}$. The hierarchical Beta prior on π allows for data-adaptive learning and shrink the elements of $\boldsymbol{\Omega}$ accordingly (Scott and Berger, 2010).

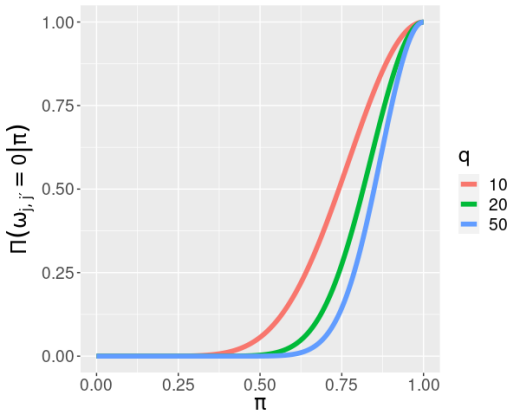


Figure 4: π versus $\Pi(\omega_{j,j'} = 0 \mid \pi)$ for different values of q. Clearly, by controlling π, it is possible to induce desired level of sparsity in $\boldsymbol{\Omega}$. A hierarchical prior on π adaptively learns from the data and shrinks the elements of $\boldsymbol{\Omega}$ accordingly.

Implementation of spike-and-slab type mixture priors is computationally challenging. Thus, continuous global-local shrinkage priors have gained popularity in the Bayesian sparse estimation literature as they often greatly simplify posterior computation (Polson and Scott, 2010) while retaining almost similar statistical properties of the classical spike-and-slab prior. In particular, we use the Dirichlet-Laplace (DL, Bhattacharya *et al.*, 2015) prior that has equivalent asymptotic properties with respect to the spike-and-slab in a similar context (Pati *et al.*, 2014, Theorem 5.1) while being amenable to scalable posterior computation. The results for the spike-and-slab discussed here provides the intuitions on how continuous shrinkage priors like the DL induce sparsity in $\boldsymbol{\Omega}$ by inducing shrinkage on $\boldsymbol{\Lambda}$.

2.2.3 Advantages of the LRD Decomposition

For high-dimensional sparse covariance matrix estimation, this LRD decomposition strategy is extensively used in both Bayesian (Pati *et al.*, 2014) and frequentist paradigms (Daniele *et al.*, 2019) where the associated latent factor formulation allows scalable computation in big data problems (Bhattacharya and Dunson, 2011; Fan *et al.*, 2011; Sabnis *et al.*, 2016; Fan *et al.*, 2018; Kastner, 2019, and others). Penalizing $\mathbf{\Lambda}$ is thus a sensible approach to induce sparsity in $\mathbf{\Omega}$. It also comes with many practical advantages described below.

Cholesky factorization based covariance and precision matrix estimation methods also use a similar strategy and penalize the lower triangular matrix $\mathbf{L}$ to induce sparsity in $\mathbf{\Omega} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$ (Dallakyan and Pourahmadi, 2020). However, for undirected graphs, the estimates can vastly differ depending on the ordering of the variables (Kang and Deng, 2020). In contrast, our proposed LRD decomposition based approach with a penalty on $\mathbf{\Lambda}$ is invariant to the ordering of the variables.

In most existing approaches that directly penalize $\mathbf{\Omega}$, frequentist or Bayesian, ensuring the positive definiteness of $\mathbf{\Omega}$ is also non-trivial, particularly in high dimensional settings. In frequentist regimes, for example, the graphical lasso (Friedman *et al.*, 2008) does not guarantee positive definiteness of $\mathbf{\Omega}$ (Mazumder and Hastie, 2012); similar behavior has also been observed for neighborhood selection methods (Meinshausen and Bühlmann, 2006; Peng *et al.*, 2009) and post-hoc treatments are required for positive definiteness. In Bayesian MAP approaches such as Gan *et al.* (2019), special care is needed in designing the optimization algorithm to ensure positive definiteness. For many existing Bayesian methods relying on MCMC (Wang, 2012; Khondker *et al.*, 2013; Li *et al.*, 2019a), constrained update of $\mathbf{\Omega}$ in each step is difficult and expensive. Applications of these methods thus remain fairly limited to

small to moderate dimensional problems. In contrast, the LRD formulation is always guaranteed to produce a positive definite estimate by construction.

As also discussed in the Introduction, posterior exploration is extremely challenging in Bayesian approaches that penalize the underlying the dense graphs first and then assign a prior on $\boldsymbol{\Omega}$ conditional on the graph, often requiring restrictive assumptions on the graph (Dawid and Lauritzen, 1993) while still remaining computationally infeasible beyond only a few tens of dimensions (Jones *et al.*, 2005). Our proposed approach on the other hand, works directly in the $\boldsymbol{\Omega}$ space, albeit via the LRD decomposition, avoiding having to define separate complex priors on the graph space thereby also avoiding restrictive assumptions on the graph while also achieving scalability far beyond such approaches via its latent factor representation.

2.3 Prior Specification

To accommodate high-dimensional data, with $d \gg n$, it is crucial to reduce the effective number of parameters in the $d \times q$ loadings matrix $\boldsymbol{\Lambda}$. A wide variety of sparsity inducing shrinkage priors for $\boldsymbol{\Lambda}$ can be considered for this purpose. In this article, we employ a two-parameter generalization of the original Dirichlet-Laplace (DL) prior from Bhattacharya *et al.* (2015) that allows more flexible tail behavior. On a d-dimensional vector $\boldsymbol{\theta}$, our DL prior with parameters a and b, denoted by $\text{DL}(a, b)$, can be specified in the following hierarchical manner

$$\theta_j | \boldsymbol{\psi}, \boldsymbol{\phi}, \tau \overset{\text{ind}}{\sim} \text{N}(0, \psi_j \phi_j^2 \tau^2), \quad \psi_j \overset{\text{iid}}{\sim} \text{Exp}(1/2), \quad \boldsymbol{\phi} \sim \text{Dir}(a, \ldots, a), \quad \tau \sim \text{Ga}(da, b),$$

where θ_j is the j^{th} element of $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ is a vector of same length as $\boldsymbol{\theta}$, $\text{Exp}(a)$ is an exponential distribution with mean $1/a$, $\text{Dir}(a_1, \ldots, a_d)$ is the d-dimensional Dirichlet distribution and $\text{Ga}(a, b)$ is the gamma distribution with mean a/b and variance a/b^2. The original DL prior is a special case with $b = 1/2$. We let $\text{vec}(\boldsymbol{\Lambda}) \sim \text{DL}(a, b)$.

The column dimension q of $\mathbf{\Lambda}$ will almost always be unknown. Assigning a prior on q and implementing a reversible jump MCMC (Green, 1995) type algorithm can be inefficient and expensive. In this paper, we adopt an empirical Bayes type approach to set q to a large value determined from the data and let the prior shrink the extra columns to zeros, substantially simplifying the computation. The strategy is discussed in greater details later in this section. We show in Section 2.6 that this approach is sufficient for the recovery of true $\mathbf{\Omega}$ under very mild conditions.

When $d \gg n$, d distinct δ_j^2's also result in over-parametrization. To reduce the number of parameters, we assume the δ_j^2's to comprise a small number of unique values. To achieve this in a data adaptive way, we use a Dirichlet process (DP) prior (Ferguson, 1973) on the δ_j^2's as

$$\delta_j^2 | G \overset{\text{iid}}{\sim} G, \quad G|\alpha \sim \text{DP}(\alpha, G_0) \text{ with } G_0 = \text{Ga}(a_\delta, b_\delta), \quad \alpha \sim \text{Ga}(a_\alpha, b_\alpha), \tag{6}$$

where α is the concentration parameter and G_0 is the base measure. Samples drawn from a DP are almost surely discrete, inducing a clustering of the δ_j^2's and thus reducing the dimension. Integrating out G, model (6) leads to a recursive Polya urn scheme illustrative of the clustering mechanism while also being convenient for posterior computation (Escobar and West, 1995; Neal, 2000). Specifically, we have

$$\delta_{j+1}^2 | \alpha, \boldsymbol{\delta}_{1:j}^2 \propto \textstyle\sum_{\ell=1}^{k_j} d_{j,\ell} \mathbb{1}(\delta_{j+1}^2 = \delta_\ell^{*2}) + \alpha G_0(\delta_{j+1}^2),$$

where $\{\delta_1^{*2}, \ldots, \delta_{k_j}^{*2}\}$ denote the unique values among $\boldsymbol{\delta}_{1:j}^2$ and $d_{j,r} = \sum_{\ell=1}^{j} \mathbb{1}(c_\ell = r)$ denote their multiplicities, $c_1, \ldots, c_d$ being the latent variables such that $c_j = r$ if δ_j^2 belongs to the r^{th} cluster.

Choice of Hyperparameters: To specify the value of the latent dimension q, we adopt a principal component analysis (PCA) based empirical Bayes type approach (Bai and Ng, 2008). First, we perform a sparse PCA (Baglama and Reichel, 2005) on the data matrix $\mathbf{y}_{1:n}$ and compute the reciprocals of the singular values. We set q

to be the number of inverse singular values in decreasing order that adds up to 95% of the total sum. We set $a = 0.5$ and $b = 2.0$ in all our simulation studies and real data applications. For the prior on the residual variances and the DP concentration parameter, we set $a_\delta = b_\delta = a_\alpha = b_\alpha = 0.1$.

2.4 Posterior Computation

The latent factor construction discussed in Section 2.1 leads to a novel, elegant Gibbs sampler that is operationally simple and free of any tuning parameter. There is also no need for additional constraints to ensure positive definiteness which used to be a major setback for MCMC based methods for precision matrix estimation. These features allow the sampler to be applied to dimensions far beyond the reach of the current state-of-the-art. Scalable posterior computation in LRD decomposed covariance matrix models is pretty standard in the literature but remained a major barrier for precision matrix models. The Gibbs sampler described here addresses this significant gap in the literature.

Starting with some initial values of $\boldsymbol{\Lambda}, \boldsymbol{\Delta}$ and other parameters, our sampler iterates between the following steps. Other parameters and hyperparameters being implicitly understood in the conditioning, Step 1 defines a transition for $\mathbf{u}, \mathbf{v}|\mathbf{y}, \boldsymbol{\Lambda}, \boldsymbol{\Delta}$; Step 2 for $\boldsymbol{\Lambda}|\mathbf{u}, \mathbf{v}$ (which is identical to $\boldsymbol{\Lambda}|\mathbf{y}, \mathbf{u}, \mathbf{v}, \boldsymbol{\Delta}$); Step 3 for $\boldsymbol{\Delta}|\mathbf{v}$ (which is identical to $\boldsymbol{\Delta}|\mathbf{y}, \mathbf{u}, \mathbf{v}, \boldsymbol{\Lambda}$); and Steps 4 and 5 are standard updates for the parameters and hyper-parameters of the DL and DP priors, respectively.

Step 1 Generate $\mathbf{u}_1, \ldots, \mathbf{u}_n \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{P})$ with $\mathbf{P} = (\mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda})$ independently from $\mathbf{y}_{1:n}$ and let $\mathbf{v}_i = \mathbf{y}_i + \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_i$.

Step 2 We have $\mathbf{u}_i = \sum_{r=1}^{d} \boldsymbol{\lambda}_r v_{r,i} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\lambda}_r = (\lambda_{r,1}, \ldots, \lambda_{r,q})$ is the r^{th} row of $\boldsymbol{\Lambda}$ and $\mathbf{v}_i = (v_{1,i}, \ldots, v_{d,i})^{\mathrm{T}}$. Define $\mathbf{u}_i^{(j)} = \mathbf{u}_i - \sum_{r \neq j} \boldsymbol{\lambda}_r v_{r,i}$. Then $\mathbf{u}_i^{(j)} = \boldsymbol{\lambda}_j v_{j,i} + \boldsymbol{\varepsilon}_i$.

Conditioned on $\mathbf{u}_i^{(j)}$, $\mathbf{v}_i$ and the associated hyper-parameters, $\boldsymbol{\lambda}_j$'s can be updated sequentially for $j = 1, \ldots, d$ from the distribution

$$\boldsymbol{\lambda}_j \sim \mathrm{N}_q\{(\mathbf{D}_j^{-1} + \left\|\mathbf{v}^{(j)}\right\|^2 \mathbf{I}_q)^{-1}\mathbf{w}_j, (\mathbf{D}_j^{-1} + \left\|\mathbf{v}^{(j)}\right\|^2 \mathbf{I}_q)^{-1}\},$$

where $\mathbf{D}_j = \tau^2 \mathrm{diag}\left(\psi_{j,1}\phi_{j,1}^2, \ldots, \psi_{j,q}\phi_{j,q}^2\right)$, $\mathbf{v}^{(j)} = (v_{j,1}, \ldots, v_{j,n})^{\mathrm{T}}$ and $\mathbf{w}_j = \sum_{i=1}^n v_{j,i}\mathbf{u}_i^{(j)}$.

Step 3 Sample the δ_j^2's through the following steps.

(i) Let $d_{r,-j} = \sum_{\ell \neq j} \mathbb{1}(c_\ell = r)$ and $\mathbf{v}^{(-j)}$ to be the collection of all $\mathbf{v}^{(\ell)}$'s, $\ell = 1, \ldots, d$, excluding $\mathbf{v}^{(j)}$. For $j = 1, \ldots, d$, sample the cluster indicators sequentially from the distribution

$$p(c_j = r) \propto \begin{cases} d_{r,-j} \int \mathrm{N}(\mathbf{v}^{(j)}; 0, \delta_r^{*-2})\mathrm{d}G_0\left(\delta_r^{*2}|\mathbf{v}^{(-j)}\right) \text{ for } r \in \{c_\ell\}_{\ell \neq j}; \\ \alpha \int \mathrm{N}(\mathbf{v}^{(j)}; 0, \delta_r^{*-2})\mathrm{d}G_0(\delta_r^{*2}) \text{ for } r \neq c_\ell \text{ for all } \ell \neq j. \end{cases}$$

The above integrals are analytically available and involves the density of a multivariate central Student's t-distribution for $G_0 = \mathrm{Ga}(a_\delta, b_\delta)$.

(ii) Let the unique values in $\mathbf{c}_{1:d}$ be $\{1, \ldots, k\}$. For $r = 1, \ldots, k$, set $d_r = \sum_j \mathbb{1}(c_j = r)$ and $\mathbf{V}_r = \sum_{j:c_j=r} \left\|\mathbf{v}^{(j)}\right\|^2$, and independently sample $\delta_r^{*2} \sim \mathrm{Ga}\left(a_\delta + nd_r/2, b_\delta + \mathbf{V}_r/2\right)$.

(iii) Set $\delta_j^2 = \delta_{c_j}^{*2}$.

Step 4 Sample the hyper-parameters in the priors on $\boldsymbol{\Lambda}$ through the following steps.

(i) For $j = 1, \ldots, d$ and $h = 1, \ldots q$ sample $\widetilde{\psi}_{j,h}$ independently from an inverse-Gaussian distribution $\mathrm{iG}\left(\tau\phi_{j,h}/|\lambda_{j,h}|, 1\right)$ and set $\psi_{j,h} = 1/\widetilde{\psi}_{j,h}$.

(ii) Sample the full conditional posterior distribution of τ from a generalized inverse Gaussian $\mathrm{giG}\left\{dq(a-1), 2b, 2\sum_{j,h} |\lambda_{j,h}|/\phi_{j,h}\right\}$ distribution.

(iii) Draw $T_{j,h}$ independently with $T_{j,h} \sim \mathrm{giG}(a-1, 1, 2|\lambda_{j,h}|)$ and set $\phi_{j,h} = T_{j,h}/T$ with $T = \sum_{j,h} T_{j,h}$.

Step 5 Following West (1992), first generate $\varphi \sim \text{Beta}(\alpha+1, d)$, evaluate $\pi/(1-\pi) = (a_\alpha + k - 1)/\{d(b_\alpha - \log\varphi)\}$ and then generate

$$\alpha|\varphi, k \sim \begin{cases} \text{Ga}(\alpha + k, b_\alpha - \log\varphi) \text{ with probability } \pi, \\ \text{Ga}(\alpha + k - 1, b_\alpha - \log\varphi) \text{ with probability } 1 - \pi. \end{cases}$$

Remark 1. *Note that the main strategies underlying the algorithm above are not specific to the DL prior considered here. We are free to choose any other shrinkage prior that admits a conditionally Gaussian hierarchical representation for the entries of* $\boldsymbol{\Lambda}$ *and modify Steps 2 and 4 accordingly. This is still a very large class of priors (Polson and Scott, 2010), including, e.g., horseshoe (Carvalho et al., 2009), multiplicative gamma (Bhattacharya and Dunson, 2011), etc. For non-conjugate priors, Steps 3 and 4 can also be modified with appropriate Metropolis-Hastings schemes. The other steps remain the same, making it a very broadly adaptable algorithm. An MCMC scheme for generic priors is outlined in Section S.1 of the supplementary materials.*

Remark 2. *In the above sampler, the conditional posterior covariance matrix of* $\boldsymbol{\lambda}_j$ *is diagonal which allows us to update it with linear complexity. Thus, in each MCMC iteration, we only need a single small dimensional* $q \times q$ *order matrix factorization operation in Step 1 to simulate* $\mathbf{u}_{1:n}$*. While high-dimensional matrix factorization operations are usally numerically very expensive, we are able to completely avoid that, facilitating substantial scalability.*

Remark 3. *All but Steps 1 and 3 can be divided into parallel operations in a straightforward manner. Additionally, in recent versions of many popular statistical software, including* `R`*, matrix operations are inherently parallelized and hence Step 1 is also highly scalable in any decent computing system.*

We implemented the Gibbs sampler in `C++` and ported to `R` using the `Rcpp` package (Eddelbuettel and Francois, 2011). In each case considered in this article, synthetic

or real, we ran $5,500$ iterations which takes approximately 37 minutes on a system with an i9-10900K CPU and 64GB memory for a $d = 1,000$ dimensional problem with sample size $n = 1,000$. The initial $1,250$ samples were discarded as burn-in and the remaining samples were thinned by an interval of 5. In all our experiments, convergence was swift and mixing was excellent. A comparison of the runtimes of our method and a few other existing methods is presented in Figure 5(c) below.

2.5 Graph Selection

As discussed in Section 2.2.2, the off-diagonals of $\mathbf{\Omega}$ are penalized by inducing shrinkage on $\mathbf{\Lambda}$. The theoretical results in Section 2.6 and numerical experiments in Section 3 show that, with our carefully constructed data adaptive shrinkage priors on $\mathbf{\Lambda}$, the inferred sparsity patterns in $\mathbf{\Lambda}$ are such that a sparse $\mathbf{\Omega}$ is also accurately recovered (see Figures S.1 and S.3(a) for estimates of sparse precision matrices obtained by our method). Exact zero estimates are, however, not obtained even for the insignificant off-diagonal elements of $\mathbf{\Omega}$ for finite samples. This is an artifact of continuous shrinkage priors, since the probabilities of exact zeroes are *almost surely* null for finite samples although the posterior probabilities of arbitrary sets around zeroes are very high.

We address the issue of non-zero edge selection through a novel multiple hypothesis testing based approach. For $i = 1, \ldots, d,\ j = i+1, \ldots, d$ and some $\epsilon > 0$, we consider testing

$$H_{0,i,j} : |\rho_{i,j}| \leq \epsilon \quad \text{versus} \quad H_{1,i,j} : |\rho_{i,j}| > \epsilon, \tag{7}$$

where $\rho_{i,j}$ is the $(i,j)^{th}$ element of $\text{diag}(\mathbf{\Omega})^{-1/2}\ \mathbf{\Omega}\ \text{diag}(\mathbf{\Omega})^{-1/2}$, the partial correlation matrix derived from $\mathbf{\Omega}$. Here we follow Berger (1985, Chapter 4, pp. 148) in replacing the point nulls $H_{0,i,j} : \rho_{i,j} = 0$ by reasonable interval nulls $H_{0,i,j} : |\rho_{i,j}| \leq \epsilon$. If $H_{0,i,j}$ is rejected in favor of $H_{1,i,j}$, we conclude that there is an edge between nodes i and j.

We utilize posterior uncertainty to resolve these testing problems. Specifically, we define $d_{i,j} = \mathbb{1}\{\Pi(H_{1,i,j}|\mathbf{y}_{1:n}) > \beta\}$ as the decision rule which controls the posterior FDR defined as

$$\text{FDR}_{\mathbf{y}} = \frac{\sum_{i,j} d_{i,j}\Pi(H_{0,i,j}|\mathbf{y}_{1:n})}{\max(\sum_{i,j} d_{i,j}, 1)},$$

at the level $1-\beta$. Importantly, the decision rule also incurs the lowest false non-discovery rate (Müller *et al.*, 2004). For a fixed β, the $\text{FDR}_{\mathbf{y}}$ depends on the choice of ϵ. To obtain the optimal ϵ, we compute the $\text{FDR}_{\mathbf{y}}$'s on a grid of ϵ values in $(0,1)$ and then set $\epsilon = \inf_{\epsilon'} \text{FDR}_{\mathbf{y}}(\epsilon') \leq 1-\beta$. This way, we have the FDR, that is, a quantified statistical uncertainty associated with the estimated graph. In all simulation experiments and real data applications in this paper, we control $\text{FDR}_{\mathbf{y}}$ at the 0.10 level of significance.

Remark 4. *Although this FDR control procedure is widely applicable, efficient posterior exploration is crucial for this step and hence may not be achievable or scalable to high dimensional problems when adapted to previously existing Bayesian approaches.*

2.6 Posterior Concentration

Preliminaries and Notation: We let $\|\mathbf{A}\|_F$ and $\|\mathbf{A}\|_2$ denote the Frobenius and spectral norms of a matrix $\mathbf{A}$, respectively; $s^2_{\min}(\mathbf{A})$ be the smallest singular value of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$; and $\lambda_1(\mathbf{A}), \ldots, \lambda_d(\mathbf{A})$ be the eigenvalues of $\mathbf{A}$ in decreasing order when $\mathbf{A}$ is a d-dimensional diagonalizable matrix. Also, $a_n = o(b_n)$ and $a_n = O(b_n)$ imply that $\lim |a_n/b_n| = 0$ and $\limsup |a_n/b_n| < \infty$, respectively. Throughout, C, C' and $\widetilde{C}$ are used to denote positive constants whose values might change from one line to the next but are independent from everything else. A d-dimensional vector $\boldsymbol{\theta}$ is said to be s-sparse if only s among the d elements of $\boldsymbol{\theta}$ are non-zero. We denote the set of all s-sparse vectors in $\mathbb{R}^d$ by $\ell_0[s, d]$.

We allow the model parameters to increase in dimensions with sample size n, indicated by associating them with the suffix n. We let $\Pi_n(\cdot)$ denote the prior and $\Pi_n(\cdot|\mathbf{y}_{1:n})$ the corresponding posterior given data $\mathbf{y}_{1:n}$, respectively.

Assumptions on the Data Generating Process: We assume that $\mathbf{y}_i \overset{\text{iid}}{\sim} \mathrm{N}_{d_n}(\mathbf{0}, \boldsymbol{\Omega}_{0n}^{-1}), i = 1, \ldots, n$. In Section 2.2.1 we discussed how an LRD decomposition of a sparse $\boldsymbol{\Omega}_{0n}$ can be constructed. Hence, we assume that $\boldsymbol{\Omega}_{0n}$ admits the factor representation $\boldsymbol{\Omega}_{0n} = \boldsymbol{\Lambda}_{0n}\boldsymbol{\Lambda}_{0n}^{\mathrm{T}} + \delta_{0n}^2\mathbf{I}_{d_n}$, where $\boldsymbol{\Lambda}_{0n}$ is a $d_n \times q_{0n}$ order sparse matrix and $\delta_{0n}^2 > 0$ is a scalar. To simplify the theoretical analysis, we deviate here slightly from the proposed model and assume that the δ^2's all come from a single cluster with the common value δ_{0n}^2. We show that the precision factor model with an appropriate DL prior can recover $\boldsymbol{\Omega}_{0n}$ in ultra high-dimensional settings. As discussed in Section 2.3, we fix q_n to a liberal large value and let the prior shrink the extra columns to zeros. We show that with appropriate sparsity conditions, the posterior of the precision factor model with the DL prior then concentrates around $\boldsymbol{\Omega}_{0n}$, even when the data dimension d_n increases in exponential order with n. The requisite conditions are stated below.

(C1) Let $\{q_{0n}\}_{n=1}^{\infty}$ and $\{s_n\}_{n=1}^{\infty}$ be increasing sequences of positive integers such that $s_n^2 = O(\log d_n)$, $q_{0n} = O\{s_n \log(d_n q_{0n})\}$ and $s_n q_{0n} \log(d_n q_{0n}) = o(n)$.

(C2) $\boldsymbol{\Lambda}_{0n}$ is a $d_n \times q_{0n}$ order full rank matrix such that each column of $\boldsymbol{\Lambda}_{0n}$ belongs to $\ell_0[s_n, d_n]$, $\liminf s_{\min}(\boldsymbol{\Lambda}_{0n}) > 0$ and $\|\boldsymbol{\Lambda}_{0n}\|_2 = O(1)$.

(C3) The scalar δ_{0n}^2 lies in some compact set $[\delta_{\min}^2, \delta_{\max}^2]$.

Specifics of the Postulated Precision Factor Model: Let $\{q_n\}_{n=1}^{\infty}$ be an increasing sequence of positive integers such that $q_{0n} \leq q_n = o(n)$ and $s_n q_{0n} \log(d_n q_n) = o(n)$. We assume that $\mathbf{y}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_{d_n}(\mathbf{0}, \boldsymbol{\Omega}_n^{-1})$, where $\boldsymbol{\Omega}_n = \boldsymbol{\Lambda}_n\boldsymbol{\Lambda}_n^{\mathrm{T}} + \delta_n^2\mathbf{I}_{d_n}$ and $\boldsymbol{\Lambda}_n$ is a

$d_n \times q_n$ matrix. We consider a DL(a_n, b_n) prior on vec$(\mathbf{\Lambda}_n)$ with $a_n = 1/d_n q_n$ and $b_n = \log^{3/2}(d_n q_n)$. We assume a gamma prior on $\delta_n^2 \sim \text{Ga}(a_\delta, b_\delta)$ independent of $\mathbf{\Lambda}_n$ and truncated to the compact set $[\delta^2_{\min}, \delta^2_{\max}]$.

Condition (C1) specifies the requisite sparsity conditions. It also imposes a condition on d_n. Specifically, it can be seen that d_n can be of exponential order of n, allowing the recovery of massive precision matrices based on relatively small sample sizes. Note that, subject to appropriate orthogonal transformation, marginally $\text{var}(y_j) = 1/\left\{\lambda_j\left(\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Lambda}_{0n}\right) + \delta_{0n}^2\right\}$. Conditions (C2) and (C3) ensure that these variances lie in a compact set. The prior specification provides a set of sufficient conditions on the class of proposed models. The true number of latent factors q_{0n} is also assumed smaller than the number of latent factors q_n in the postulated models.

We let $\mathcal{P}_{0n}$ denote the class of precision matrices satisfying (C1)-(C3) and $\mathcal{P}_n$ denote the class of positive definite matrices parametrized by $(q_n, \delta_n^2, \mathbf{\Lambda}_n)$ in our precision factor model. Notably, $\mathcal{P}_{0n}$ can also be parametrized by $(q_{0n}, \delta_{0n}^2, \mathbf{\Lambda}_{0n})$. However, since $q_{0n} \leq q_n$, for a fixed q_n, $\mathcal{P}_n$ may be overparametrized for $\mathcal{P}_{0n}$. In Theorem 1, we show that with increasing sample size, the posterior distribution supported on $\mathcal{P}_n$ still concentrates around the true $\mathbf{\Omega}_{0n}$ from $\mathcal{P}_{0n}$. Using general results from Ghosal and van der Vaart (2017), we establish a convergence rate in terms of the operator norm. A rigorous proof is detailed in Section S.2 in the supplementary materials.

Theorem 1. *For any* $\mathbf{\Omega}_{0n} \in \mathcal{P}_{0n}$, $\epsilon_n = (s_n q_{0n})^4 \log(d_n q_n)\sqrt{q_{0n} s_n \log(d_n q_n)/n}$ *and any sequence* $M_n \to \infty$,

$$\lim_{n\to\infty} \mathbb{E}_{\mathbf{\Omega}_{0n}} \Pi_n \left(\|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 > M_n \epsilon_n | \mathbf{y}_{1:n}\right) = 0.$$

In Section 2.2.1 we discussed a constructive way to find a sparse LRD decomposition of arbitrary sparse $\mathbf{\Omega}_{0n}$ where the column dimension q_{0n} of $\mathbf{\Lambda}_{0n}$ need not exceed the number of non-zero off-diagonals of $\mathbf{\Omega}_{0n}$ or edges in the conditional dependence

graph. Theorem 1 implies that the precision matrix is recoverable if (C1) holds among others. Note that (C1) implies $q_{0n} = o(n)$. Hence, $\mathbf{\Omega}_{0n}$ can be learned from the data using the precision factor model as long as the number of edges is bounded by the sample size. Such assumptions are required to recover massive dimensional precision matrices from relatively smaller amount of data (Meinshausen and Bühlmann, 2006; Ksheera Sagar *et al.*, 2021).

Remark 5. *We derive a minimax lower bound of* $\sqrt{s_n \log d_n / n}$ *for the precision matrix estimation problem in Theorem 4 in the supplementary materials when* $q_{0n} \leq q_n = O(1)$ *which is a special case of (C1). The rate in Theorem 1 is* $s_n^4 \log d_n$ *times the lower bound. If we further let* $d_n = O(n^r)$ *for some fixed positive integer* r*, the rate attains the minimax lower bound up to a* $(\log n)^3$ *term.*

Remark 6. *An interesting implication of our theoretical results is the robustness with respect to the choice of the latent dimension* q *under very mild conditions. For the class of our postulated precision factor models, we err on the side of overestimating* q *and let the DL prior shrink the extra factors to zero. The theoretical results imply that this strategy is sufficient to recover* $\mathbf{\Omega}_{0n}$ *efficiently. The precise recovery rate, however, naturally depends on the choice of* q_n*, a key parameter that distinguishes the class of postulated models from the class of true models (Shalizi, 2009). The closer* q_n *is to* q_{0n}*, the smaller the space to search for the truth, and the better the rate.*

3 Simulation Studies

In this section, we discuss the results of some synthetic numerical experiments. We evaluate the performance of estimating the precision matrix $\mathbf{\Omega}$ itself as well as the underlying graph. We simulate data from three different cases - (i) a Gaussian au-

toregressive (AR) process of order 2, (ii) a multivariate Gaussian distribution with banded precision matrix, and (iii) a multivariate Gaussian distribution with randomly generated arbitrarily structured sparse (RSM) precision matrix. In all these cases, we take the mean vector to be zero. For plotting purposes in the RSM case, we assume two nodes to have an edge between them if the absolute value of the associated partial correlation exceeds 0.1 and refer to this as the 'true' graph. We perform experiments for dimensions $d = 50$, 100, 200 and 1,000. For $d \leq 200$ we take $n = 100$ and for $d = 1,000$ we take $n = 1,000$. We plot the true precision matrices and the associated true graphs for $d = 50$ in Figure S.1 in the supplementary materials and in Figure 6(a) here in the main paper. For $d \leq 200$ and $d = 1,000$ we consider 50 and 20 independent replications for each scenario, respectively.

We apply our precision factor (PF) model to recover the precision matrices, using the posterior mean as our Bayesian point estimate. We compare it with the methods Bayesian graphical model under shrinkage (Bagus, Gan *et al.*, 2019), the graphical Lasso (Glasso, Friedman *et al.*, 2008) and the neighborhood selection by Meinshausen and Bühlmann (M&B, 2006). We do not consider any previously existing Bayesian posterior sampling based methods here as they do not scale well beyond only a few tens of dimensions. We assess the comparative efficacy of these methods using qualitative graphical summaries as well as quantitative performance measures. For the proposed PF method, the graphs are estimated following the FDR control procedure outlined in Section 2.5. For Bagus, we follow the prescription in Section 4.3 of Gan *et al.* (2019). For $d = 1,000$ in the banded case, the codes for Bagus did not work and we could not consider it as a competitor in that particular setup. For the other methods, the non-zero entries in the estimated precision matrix are considered as edges.

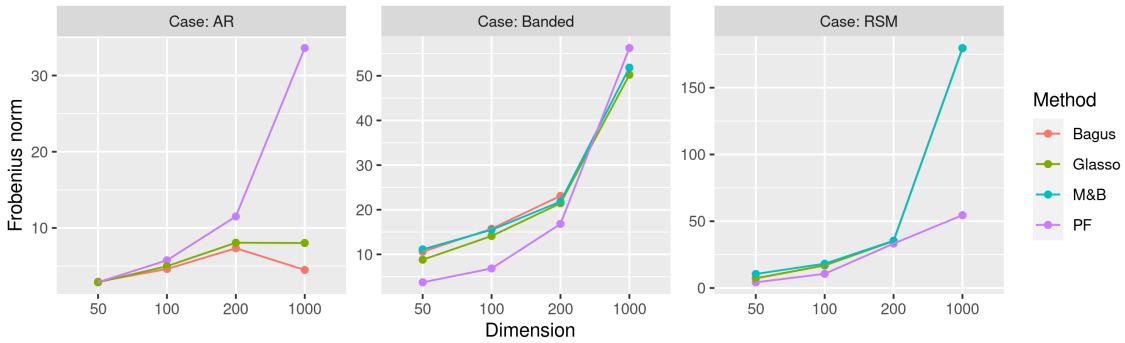
We compute the Frobenius norm between the true and estimated precision corre-

lation matrices. To assess the accuracy of the derived graphs, we compute specificity $= \frac{TN}{TN+FP}$ and sensitivity $= \frac{TP}{TP+FN}$, where TP (true positives), FP (false positives), TN (true negatives) and FN (false negatives) are based on the detection of the edges in the estimated graphs and comparing them with the corresponding true graphs. In Figure 5, the average values of these criteria across the replications are plotted for the competing methods. We also compare the execution times of the different methods.

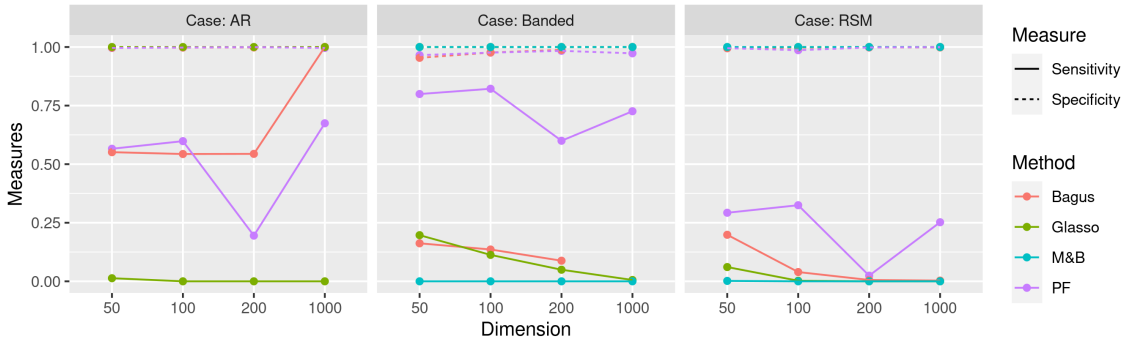
From Figure 5, we see that, for banded precision matrices, the Frobenius norms obtained by our proposed PF method are very similar to those produced by the competitors. In the case of the arbitrarily structured precision matrix, however, the PF method performs much better, especially in higher dimensions. Also, the PF method yields much higher sensitivity in almost all the scenarios at the expense of slightly smaller specificity in some cases. This is expected since we are allowing a small margin of error by controlling the FDR at 0.10. Notably, the PF method is much more powerful in detecting the true edges. The other methods seem to be quite conservative in that regard, as seen in Figure 5(b). This explains the high Frobenius norm produced by the PF method in the AR(2) case with $d = 1,000$. The true precision matrix being extremely sparse with only two off-diagonals being non-zero, the conservative competitors are performing better in recovering the precision matrix in this particular setup. We see in Figure 5(c) that the PF method is slower than M&B and Glasso but is much faster than Bagus. Note however that, unlike the competitors, we are exploring the full posterior, not just providing a point estimate. On a related important note, since we are able to sample from the full posterior, unlike the other methods, we are able to provide a natural way of quantifying posterior uncertainty. Specifically, posterior credible intervals for each element of the precision matrix can be obtained from the MCMC samples. For $d = 50$ and sample size $n = 100$, we plot

the lower 2.5% and upper 97.5% quantiles of the entry-wise partial correlations along with the simulation truths in Figure S.1 in the supplementary materials. The plots indicate that the simulation truths are well-within the confidence bounds.

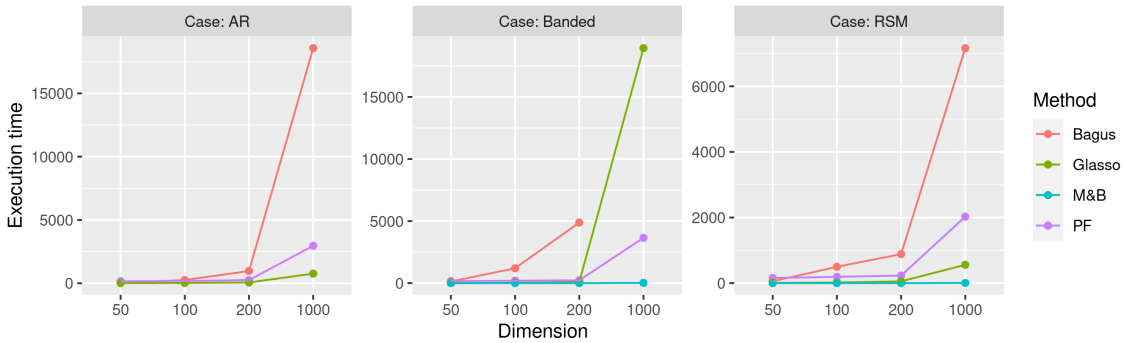
Circos plots (Gu *et al.*, 2014) of the true and estimated graphs are shown in Figure 6. For the AR(2) case, none of the methods are doing well in recovering the graph. For the other cases, the proposed PF method outperforms the competitors. Figures S.2 and S.3 in the supplementary materials provide alternate graphical representations useful in visually discerning the superior performance of the PF method.



(a) Frobenius norm

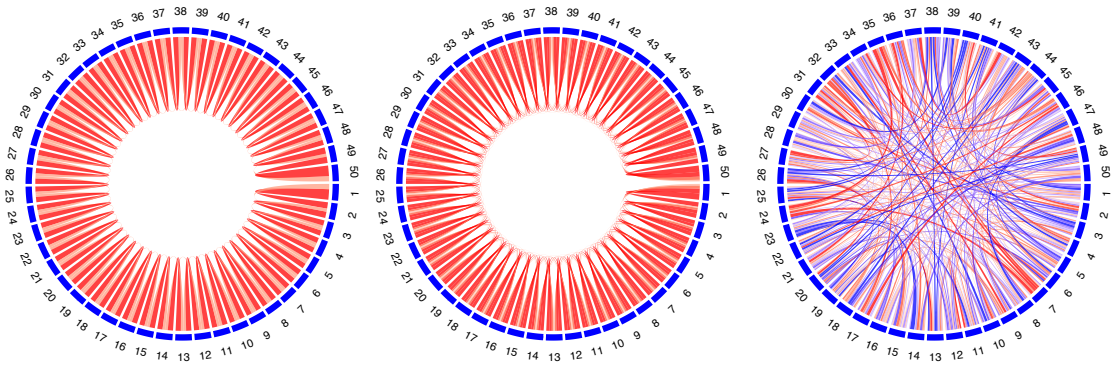


(b) Sensitivity and specificity

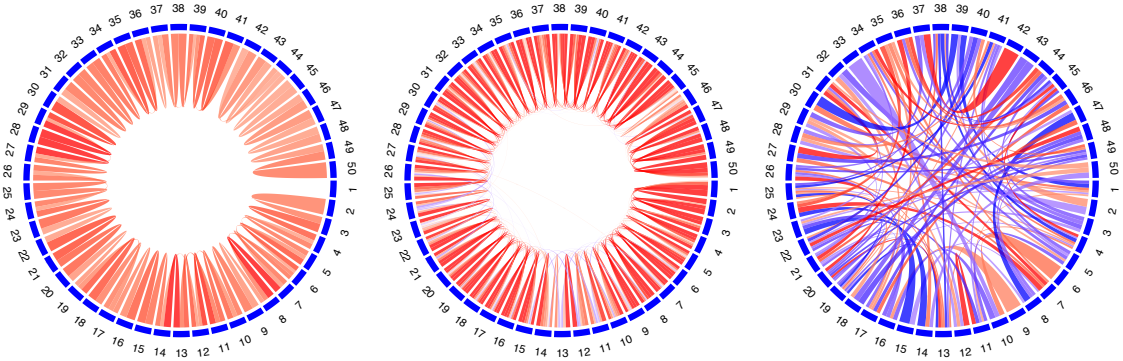


(c) Execution time

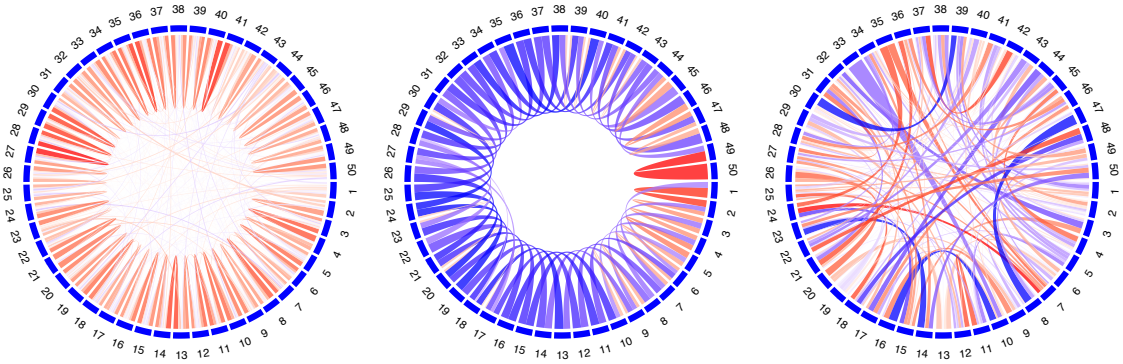
Figure 5: Results of simulation experiments: Panel (a) shows the Frobenius norms between the true and estimated partial correlation matrices; panel (b) shows the sensitivity and specificity; and panel (c) shows the execution times in seconds.



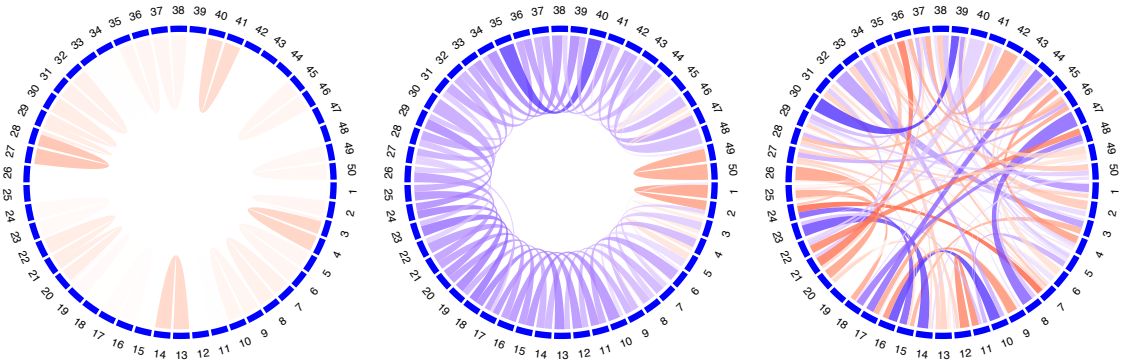
(a) True graphs.



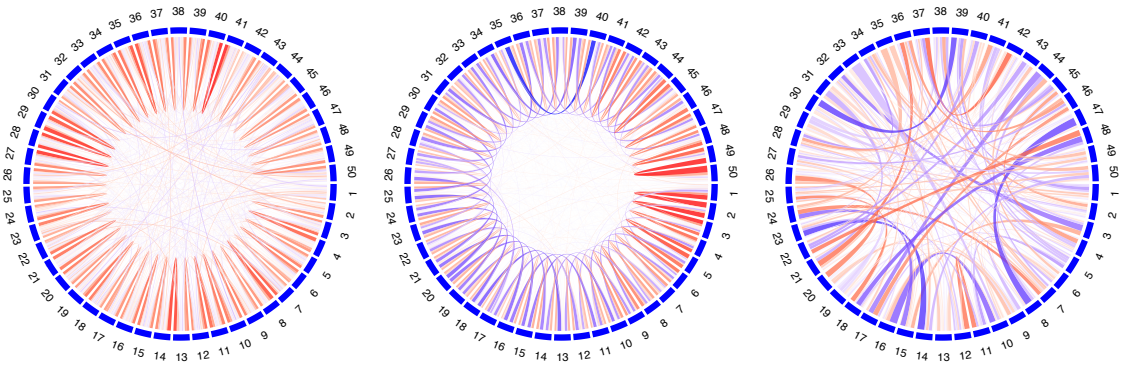
(b) Estimated graphs by the PF method.



(c) Estimated graphs by Bagus.



(d) Estimated graphs by M&B.



(e) Estimated graphs by Glasso.

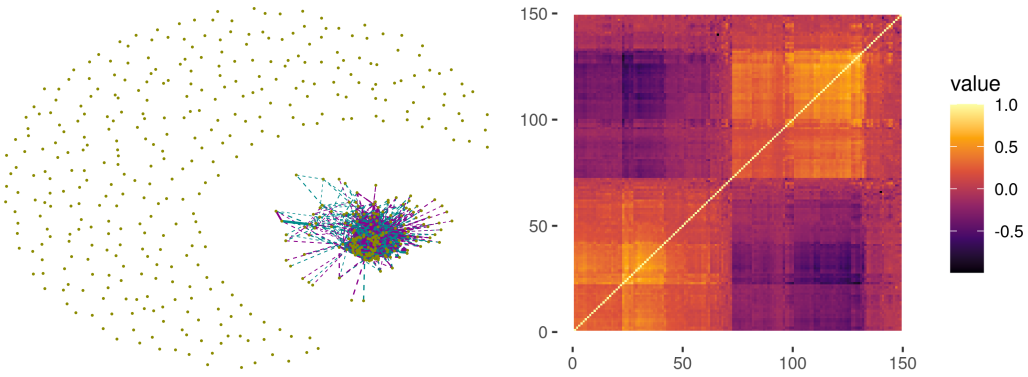
Figure 6: Results of simulation experiments: Graph recovery: Panel (a) shows the true graphs for AR(2), banded and RSM structures from left to right; panels (b), (c), (d) and (e) show the corresponding estimated graphs for our proposed PF and Bagus, M&B and Glasso methods, respectively. Positive (negative) associations are represented by blue (red) links, their opacities being proportional to the corresponding association strengths. The link widths are inversely proportional to the number of edges associated with the corresponding nodes. Figure S.2 in the supplementary materials provides a useful alternative graphical representation.

4 Application

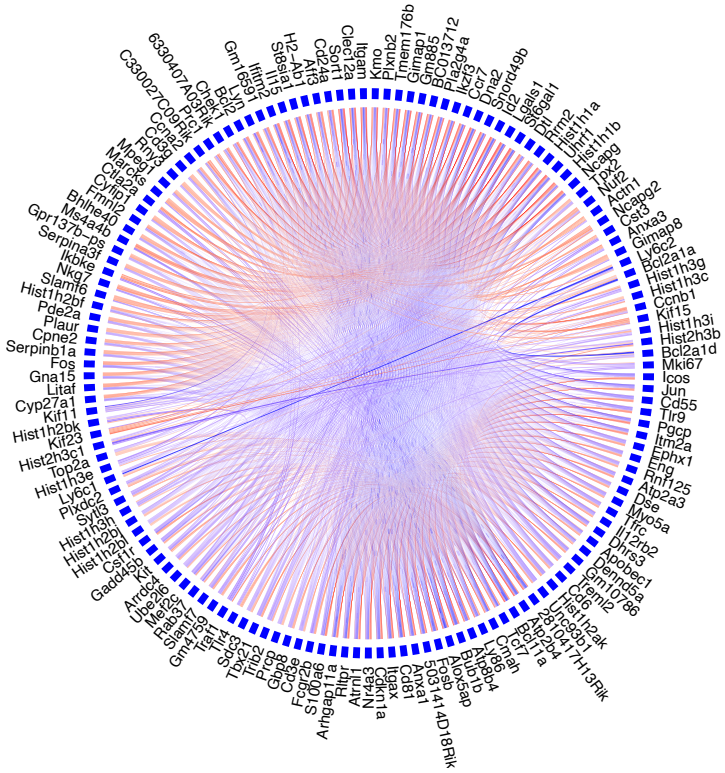
We applied our proposed method to two real data sets from two different application domains, namely genomics and finance. To meet space constraints, the finance application is presented separately in Section S.4 in the supplementary materials.

Immune cells serve specialized roles in innate and adaptive bodily responses to eliminate antigens. To understand the cell biology of carcinogenic processes, study of immune cells and the genes therein are thus of immense importance.

We obtain the data from the Immunological Genome Project (ImmGen) data browser (Heng *et al.*, 2008). ImmGen is a collaborative scientific research project that is currently building a gene-expression database for all characterized immune cells in mice. In particular, we use the GSE15907 microarray dataset (Painter *et al.*, 2011; Desch *et al.*, 2011) comprising of multiple immune cell lineages which were isolated ex-



(a) Graph of all genes. (b) Heatmap of the connected genes.



(c) Graph of the connected genes.

Figure 7: Results for ImmGen microarray data: Panel (a) shows the graph for all genes; panel (b) shows the heatmap of the partial correlation matrix for the genes having at least 5 edges; panel (c) zooms into the graph of these genes. Positive (negative) associations are represented by blue (red) links, their opacities being proportional to the corresponding association strengths. The link widths are inversely proportional to the number of edges associated with the corresponding nodes.

vivo, primarily from young adult B6 male mice and double-sorted to $> 99\%$ purity. The cell population includes all adaptive and innate lymphocytes (B, abT, gdT, Innate-Like Lymphocytes), myeloid cells (dendritic cells, macrophages, monocytes),

mast cells and neutrophils. The already normalized dataset has more than $21,000$ gene expressions from $n = 653$ immune cells. We made a $\log_2$ transformation of the data and filtered the top 2.5% genes with highest variances using the `genefilter R` package (Gentleman *et al.*, 2020), resulting in a $d = 544$ dimensional problem. Since different cell-types exhibit very different gene expression profiles, we centered the gene-expressions separately within each cell type. The estimated graph corresponding to all genes is shown in Figure 7(a). We can see clear evidence towards conditional independence between most of the genes, indicating that only a small subset of genes are functionally responsible for the variability.

For more insights, we zoom into the connected genes and plot the corresponding partial correlation matrix and graph, limited to the genes possessing at least 5 edges, in Figures 7(b) and 7(c) respectively. We notice overall positive partial correlations between the histone class of genes such as Hist2h3b, Hist1h3h, Hist2h3c1, Hist1h3c, etc (Wolffe, 2001). The positive correlations between these protein coding genes indicate that these genes are generally expressed together in the mechanism. Bcl2a1a and Bcl2a1d, two functional isoforms of the B cell leukemia 2 family member A1 exerting important pro-survival functions, show strong positive association. We also observe strong positive association between Ly6c1 and Ly6c2 genes. Lee *et al.* (2013) noted these genes to be located adjacent to each other in the mid-section of the Ly6 complex. They share $> 95\%$ similarity in their genomic and protein sequences and these two genes has been considered synonymous to each other and hence exhibit strong positive association. Although positive correlations are often observed between the membrane-spanning 4A class of genes such as Ms4a4c, Ms4a4b, Ms4a6b, etc. (Liang *et al.*, 2001), we find the genes to be conditionally almost independent in our analysis. It indicates that their expressions might be regulated by other genes.

From Figure 7(b), it thus seems that there exist two blocks where within each block the genes are positively associated whereas between the blocks the genes are negatively associated. This might again be an indication that either of these blocks of genes are generally expressed together in immune cells.

5 Discussion

In this article, we proposed a novel flexible statistical model for Gaussian precision matrices that relies on decomposing them into a low-rank and a diagonal component. The decomposition is theoretically and practically highly flexible and is thus often used in covariance matrix models, arising naturally in their computationally tractable latent factor based representations. The approach has, however, not been popular in precision matrix and related graph estimation problems as it poses daunting computational challenges when applied to such settings. We addressed this issue in this article by exploring a previously under-utilized latent variable construction, leading to a highly scalable Gibbs sampler that allows efficient posterior inference. The decomposition based strategy also allowed us to use sparsity inducing priors to shrink insignificant off-diagonal entries toward zero while also making the approach adaptable to high-dimensional sparse settings. We specifically adapted the Dirichlet-Laplace prior for sparse precision matrix estimation. We developed a novel posterior FDR control based method to perform graph selection that properly accommodates posterior uncertainty. We also established theoretical convergence guarantees for the proposed model in high-dimensional sparse settings. In synthetic experiments, the proposed method vastly outperformed its competitors in receiving the true underlying graphs. We illustrated the method's practical utility through real data examples from genomics and finance.

Aside from providing fundamentally new probabilistic perspectives on Gaussian precision matrix and related graph estimation problems, our work also bridges the gap between frequentist penalized likelihood based strategies and Bayesian shrinkage prior based ideas for such models. The simple and highly scalable computational algorithms resulting from our latent factor representation should also free up Bayesians from having to put restrictive assumptions on the graph structures to achieve computational tractability, opening up new opportunities to adapt the basic models to more complex and more realistic data structures and study designs. A few such methodological extensions we are pursuing as topics of separate ongoing research include dynamic Gaussian graphical models (Huang and Chen, 2017), covariate dependent Gaussian graphical models, nonparanormal (Liu *et al.*, 2009) and Gaussian copula graphical models (Pitt *et al.*, 2006), etc.

Supplementary Materials

Supplementary materials discuss a general strategy for posterior computation under broad classes of generic priors, proofs of the theoretical results, some additional figures, and an application to a NASDAQ-100 stock price dataset.

References

Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). Bayesian covariance matrix estimation using a mixture of decomposable graphical models. *Statistics and Computing*, **19**, 303–316.

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

Baglama, J. and Reichel, L. (2005). Augmented implicitly restarted Lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, **27**, 19–42.

Bai, J. and Ng, S. (2008). Large dimensional factor analysis. *Foundations and Trends in Econometrics*, **3**, 89–163.

Banerjee, O., El Ghaoui, L., 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.

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

Barvinok, A. and Rudelson, M. (2022). When a system of real quadratic equations has a solution. *Advances in Mathematics*, **403**, 108391.

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. Springer series in statistics. Springer-Verlag, New York, 2nd edition.

Bhattacharya, A. and Dunson, D. B. (2011). Sparse Bayesian infinite factor models. *Biometrika*, **98**, 291–306.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika*, **103**, 985–991.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, **98**, 807–820.

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

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

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, pages 73–80. PMLR.

Chandra, N. K. and Bhattacharya, S. (2019). Non-marginal decisions: A novel Bayesian multiple testing procedure. *Electronic Journal of Statistics*, **13**, 489–535.

Dallakyan, A. and Pourahmadi, M. (2020). Fused-lasso regularized Cholesky factors of large nonstationary covariance matrices of longitudinal data. *arXiv 2007.11168*.

Daniele, M., Pohlmeier, W., and Zagidullina, A. (2019). Sparse approximate factor estimation for high-dimensional covariance matrices. *arXiv:1906.05545*.

d'Aspremont, A., Banerjee, O., and El Ghaoui, L. (2008). First-order methods for sparse covariance selection. *SIAM Journal on Matrix Analysis and Applications*, **30**, 56–66.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, **21**, 1272–1317.

Dellaportas, P., Giudici, P., and Roberts, G. (2003). Bayesian inference for nondecomposable graphical Gaussian models. *Sankhyā: The Indian Journal of Statistics*, pages 43–55.

Desch, A. N., Randolph, G. J., *et al.* (2011). CD103+ pulmonary dendritic cells preferentially acquire and present apoptotic cell–associated antigen. *Journal of Experimental Medicine*, **208**, 1789–1797.

Deshpande, S. K., Ročková, V., and George, E. I. (2019). Simultaneous variable and covariance selection with the multivariate spike-and-slab lasso. *Journal of Computational and Graphical Statistics*, **28**, 921–931.

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

Dobra, A., Lenkoski, A., and Rodriguez, A. (2011). Bayesian inference for general Gaussian graphical models with application to multivariate lattice data. *Journal of the American Statistical Association*, **106**, 1418–1433.

Eddelbuettel, D. and Francois, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, **40**, 1–18.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**, 577–588.

Fan, J., Liao, Y., and Mincheva, M. (2011). High-dimensional covariance matrix estimation in approximate factor models. *The Annals of Statistics*, **39**, 3320–3356.

Fan, J., Liu, H., and Wang, W. (2018). Large covariance estimation through elliptical factor models. *The Annals of Statistics*, **46**, 1383–1414.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, **1**, 209–230.

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

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.

Gentleman, R., Carey, V., Huber, W., and Hahne, F. (2020). *genefilter: methods for filtering genes from high-throughput experiments*. R package version 1.70.0.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Green, P. J. and Thomas, A. (2013). Sampling decomposable graphs using a Markov chain on junction trees. *Biometrika*, **100**, 91–110.

Gu, Z., Gu, L., Eils, R., Schlesner, M., and Brors, B. (2014). *circlize* implements and enhances circular visualization in R. *Bioinformatics*, **30**, 2811–2812.

Heng, T. S., Painter, M. W., Elpek, K., Lukacs-Kornek, V., Mauermann, N., Turley, S. J., Koller, D., Kim, F. S., Wagers, A. J., Asinovski, N., *et al.* (2008). The immunological genome project: networks of gene expression in immune cells. *Nature Immunology*, **9**, 1091–1094.

Huang, F. and Chen, S. (2017). Learning dynamic conditional Gaussian graphical models. *IEEE Transactions on Knowledge and Data Engineering*, **30**, 703–716.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics*, **33**, 730–773.

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statistical Science*, **20**, 388–400.

Kang, X. and Deng, X. (2020). An improved modified Cholesky decomposition approach for precision matrix estimation. *Journal of Statistical Computation and Simulation*, **90**, 443–464.

Kastner, G. (2019). Sparse Bayesian time-varying covariance estimation in many dimensions. *Journal of Econometrics*, **210**, 98–115.

Khare, K., Rajaratnam, B., and Saha, A. (2018). Bayesian inference for Gaussian graphical models beyond decomposable graphs. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, **80**, 727–747.

Khondker, Z. S., Zhu, H., Chu, H., Lin, W., and Ibrahim, J. G. (2013). The Bayesian covariance lasso. *Statistics and its Interface*, **6**, 243.

Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques*. MIT press.

Ksheera Sagar, K. N., Banerjee, S., Datta, J., and Bhadra, A. (2021). Precision matrix estimation under the horseshoe-like prior-penalty dual. arXiv:2104.10750.

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

Lee, P. Y., Wang, J.-X., *et al.* (2013). Ly6 family proteins in neutrophil biology. *Journal of Leukocyte Biology*, **94**, 585–594.

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

Levina, E., Rothman, A., and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *The Annals of Applied Statistics*, **2**, 245–263.

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

Li, Z., Mccormick, T., and Clark, S. (2019b). Bayesian joint spike-and-slab graphical lasso. In *International Conference on Machine Learning*, pages 3877–3885. PMLR.

Liang, Y., Buckley, T. R., *et al.* (2001). Structural organization of the human MS4A gene cluster on chromosome 11q12. *Immunogenetics*, **53**, 357–368.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, **10**, 2295–2328.

Mazumder, R. and Hastie, T. (2012). The graphical lasso: New insights and alternatives. *Electronic Journal of Statistics*, **6**, 2125–2149.

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

Mitra, R., Müller, P., Liang, S., Yue, L., and Ji, Y. (2013). A Bayesian graphical model for chip-seq data on histone modifications. *Journal of the American Statistical Association*, **108**, 69–80.

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

Mohammadi, R., Massam, H., and Letac, G. (2021). Accelerating Bayesian structure learning in sparse Gaussian graphical models. *Journal of the American Statistical Association*. To appear.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: The case of gene expression microarrays. *Journal of the American Statistical Association*, **99**, 990–1001.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, **9**, 249–265.

Painter, M. W., Davis, S., Hardy, R. R., Mathis, D., Benoist, C., Consortium, I. G. P., *et al.* (2011). Transcriptomes of the B and T lineages compared by multiplatform microarray profiling. *The Journal of Immunology*, **186**, 3047–3057.

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.

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

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

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, **9**, 1–24.

Pourahmadi, M. (2013). *High-dimensional covariance estimation*. John Wiley & Sons.

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

Roverato, A. (2002). Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, **29**, 391–411.

Sabnis, G., Pati, D., Engelhardt, B., and Pillai, N. (2016). A divide and conquer strategy for high dimensional Bayesian factor models. *arXiv preprint arXiv:1612.02875*.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, **38**, 2587–2619.

Shalizi, C. R. (2009). Dynamics of Bayesian updating with dependent data and misspecified models. *Electronic Journal of Statistics*, **3**, 1039–1074.

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

West, M. (1992). *Hyperparameter estimation in Dirichlet process mixture models*. Duke University ISDS Discussion Paper# 92-A03.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, **20**, 892–900.

Wolffe, A. (2001). Histone genes. In *Encyclopedia of Genetics*, pages 948–952. Academic Press, New York.

Yoshida, R. and West, M. (2010). Bayesian learning in sparse graphical factor models via variational mean-field annealing. *Journal of Machine Learning Research*, **11**, 1771–1798.

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

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via lasso penalized D-trace loss. *Biometrika*, **101**, 103–120.

Zhu, H., Khondker, Z., Lu, Z., and Ibrahim, J. G. (2014). Bayesian generalized low rank regression models for neuroimaging phenotypes and genetic markers. *Journal of the American Statistical Association*, **109**, 977–990.

Supplementary Materials for

Bayesian Scalable Precision Factor Analysis for Massive Sparse Gaussian Graphical Models

Noirrit Kiran Chandra[a] (noirrit.chandra@utdallas.edu)
Peter Müller[b,c] (pmueller@math.utexas.edu)
Abhra Sarkar[b] (abhra.sarkar@utexas.edu)

[a]Department of Mathematical Sciences,
The University of Texas at Dallas,
800 W. Campbell Rd, Richardson, TX 75080-3021, USA

[b]Department of Statistics and Data Sciences,
The University of Texas at Austin,
2317 Speedway D9800, Austin, TX 78712-1823, USA

[c]Department of Mathematics,
The University of Texas at Austin,
2515 Speedway, PMA 8.100, Austin, TX 78712-1823, USA

Supplementary materials discuss a general strategy for posterior computation under a broad class of generic priors, proofs of the theoretical results, some additional figures, and an application to a NASDAQ-100 stock price dataset.

S.1 General Strategy for Posterior Computation

In this section, we sketch out a general strategy for posterior inference for the precision factor model not limited to the prior specifications in Section 2.3.

Sampling model: The data generating mechanism, we recall, is

$$\mathbf{y}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_d(\mathbf{0}, \boldsymbol{\Omega}^{-1}) \quad \text{with} \quad \boldsymbol{\Omega} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\mathrm{T}} + \boldsymbol{\Delta},$$

where $\boldsymbol{\Lambda}$ is a $d \times q$ dimensional matrix with $q \leq d$ and $\boldsymbol{\Delta} = \mathrm{diag}(\delta_1^2, \ldots, \delta_d^2)$.

Generic priors: We consider the following broad classes of hierarchical priors on $\boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ with $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ being the respective associated hyperparameters

$$\mathrm{vec}(\boldsymbol{\Lambda})|\boldsymbol{\xi} \sim \mathrm{N}_{dq}\{\mathbf{0}, g(\boldsymbol{\xi})\}, \quad \boldsymbol{\xi} \sim \Pi_{\boldsymbol{\xi}}, \qquad \boldsymbol{\Delta}|\boldsymbol{\zeta} \sim \Pi_{\boldsymbol{\Delta}|\boldsymbol{\zeta}}(\boldsymbol{\Delta}|\boldsymbol{\zeta}), \quad \boldsymbol{\zeta} \sim \Pi_{\boldsymbol{\zeta}}.$$

Varying choices of $g(\boldsymbol{\xi}), \Pi_{\boldsymbol{\xi}}, \Pi_{\boldsymbol{\Delta}|\boldsymbol{\zeta}}, \Pi_{\boldsymbol{\zeta}}$ then produce a broad class of priors for the model parameters.

The Gibbs sampler then iterates through the following steps.

Step 1. Sample the latent variables: Generate $\mathbf{u}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_q(\mathbf{0}, \mathbf{P})$ with $\mathbf{P} = (\mathbf{I}_q + \boldsymbol{\Lambda}^{\mathrm{T}}\boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda})$ independently from $\mathbf{y}_{1:n}$ and let $\mathbf{v}_{1:n} = \mathbf{y}_{1:n} + \boldsymbol{\Delta}^{-1}\boldsymbol{\Lambda}\mathbf{P}^{-1}\mathbf{u}_{1:n}$.

Step 2. Sample the rows of $\boldsymbol{\Lambda}$: We have $\mathbf{u}_i = \sum_{r=1}^{d} \boldsymbol{\lambda}_r v_{r,i} + \boldsymbol{\varepsilon}_i$, where $\boldsymbol{\lambda}_r = (\lambda_{r,1}, \ldots, \lambda_{r,q})$ is the r^{th} row of $\boldsymbol{\Lambda}$ and $\mathbf{v}_i = (v_{1,i}, \ldots, v_{d,i})^{\mathrm{T}}$. Define $\mathbf{u}_i^{(j)} = \mathbf{u}_i - \sum_{r \neq j} \boldsymbol{\lambda}_r v_{r,i}$. Then $\mathbf{u}_i^{(j)} = \boldsymbol{\lambda}_j v_{j,i} + \boldsymbol{\varepsilon}_i$. Conditioned on $\mathbf{u}_i^{(j)}$, $\mathbf{v}_i$ and the associated hyper-parameters, $\boldsymbol{\lambda}_j$'s can be updated sequentially $j = 1, \ldots, d$ from the following posterior distribution

$$\boldsymbol{\lambda}_j \sim \mathrm{N}_q\{(\mathbf{D}_j^{-1} + \left\|\mathbf{v}^{(j)}\right\|^2 \mathbf{I}_q)^{-1}\mathbf{w}_j, (\mathbf{D}_j^{-1} + \left\|\mathbf{v}^{(j)}\right\|^2 \mathbf{I}_q)^{-1}\},$$

where $\mathbf{D}_j$ is the prior covariance matrix of $\boldsymbol{\lambda}_j$, $\mathbf{v}^{(j)} = (v_{j,1}, \ldots, v_{j,n})^{\mathrm{T}}$ and $\mathbf{w}_j = \sum_{i=1}^{n} v_{j,i}\mathbf{u}_i^{(j)}$.

Step 3. Sample $\boldsymbol{\Delta}$: Sample $\boldsymbol{\Delta}$ from $\Pi_{\boldsymbol{\Delta}|\mathbf{v}^{(1:d)},\boldsymbol{\zeta}}$.

Step 4. Sample the hyperparameters of $\boldsymbol{\Lambda}$: Sample $\boldsymbol{\xi}$ from $\boldsymbol{\xi} \sim \Pi_{\boldsymbol{\xi}|\boldsymbol{\Lambda}}$.

Step 5. Sample the hyperparameters of $\boldsymbol{\Delta}$: Sample $\boldsymbol{\zeta}$ from $\boldsymbol{\zeta} \sim \Pi_{\boldsymbol{\zeta}|\boldsymbol{\Delta}}$.

For non-conjugate priors, appropriate Metropolis-within-Gibbs type MCMC algorithms can be employed in Steps 2, 3, 4 or 5.

S.2 Proofs of Theoretical Results

Notations: In what follows, for an operation '$*$', $\overline{a * b}$ is sometimes used for $(a * b)$. Also, for two sequences $a_n, b_n \geq 0$, $a_n \precsim b_n$ implies that $a_n \leq C b_n$ for some constant $C > 0$; $a_n \asymp b_n$ implies that $0 < \liminf |a_n/b_n| \leq \limsup |a_n/b_n| < \infty$. $|\mathbf{A}|$ denotes the determinant of the square matrix $\mathbf{A}$. For a set S, $|S|$ denotes its cardinality. Let $\|\mathbf{x}\|$ is the Euclidean norm of a vector $\mathbf{x}$. We denote the Kullback-Leibler (KL) divergence between two mean zero Gaussian distributions with precision matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}'$ by $\mathbb{KL}\left(\mathbf{\Omega} \parallel \mathbf{\Omega}'\right)$. We borrow the following result from Pati *et al.* (2014) (Lemma 1.1. from the supplement). For brevity of notations, we reuse C, $\widetilde{C}$, C'', etc. in the proofs to denote constants and their values may not be the same throughout the same proof. Nevertheless, we were careful to make sure that these quantities are indeed constants.

Lemma 1. *For any two matrices* $\mathbf{A}$ *and* $\mathbf{B}$*,*

(i) $s_{\min}\left(\mathbf{A}\right)\|\mathbf{B}\|_F \leq \|\mathbf{AB}\|_F \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_F$.

(ii) $s_{\min}\left(\mathbf{A}\right)\|\mathbf{B}\|_2 \leq \|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{B}\|_2$.

(iii) $s_{\min}\left(\mathbf{A}\right)s_{\min}\left(\mathbf{B}\right) \leq s_{\min}\left(\mathbf{AB}\right) \leq \|\mathbf{A}\|_2 s_{\min}\left(\mathbf{B}\right)$.

Theorem 1. Let $\mathcal{P}_n = \mathcal{P}_{1,n} \cup \mathcal{P}_{2,n}$ be an arbitrary partition. Note that,

$$\Pi_n\left(\|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 > M_n\epsilon_n|\mathbf{y}_{1:n}\right) \leq \Pi_n\left(\mathbf{\Omega}_n \in \mathcal{P}_{1,n} : \|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 > M_n\epsilon_n|\mathbf{y}_{1:n}\right) + \Pi_n\left(\mathcal{P}_{2,n}|\mathbf{y}_{1:n}\right).$$

To prove the theorem, we show that, as $n \to \infty$, the posterior distribution on $\mathcal{P}_{1,n}$ concentrates around $\mathbf{\Omega}_{0n}$ while the remaining mass assigned to $\mathcal{P}_{2,n}$ diminishes to zero. We adopt a variation of Theorem 8.22 by Ghosal and van der Vaart (2017) on posterior contraction rates. Define $B_{n,0}(\mathbf{\Omega}_{0n}, \epsilon) = \{\mathbf{\Omega}_n \in \mathcal{P}_n : \mathbb{KL}\left(\mathbf{\Omega}_{0n} \parallel \mathbf{\Omega}_n\right) \leq n\epsilon^2\}$. We now formally state the result tailored towards our application.

Theorem 2. *Let* $\mathcal{P}_n$ *be a class of distributions prametrized by* $\mathbf{\Omega}_n$*,* $\mathbb{P}_{\mathbf{\Omega}_n}$ *the corresponding distribution and* $\mathbf{\Omega}_{0n}$ *be the true data-generating value of the parameter. Let* e_n *be a metric on* $\mathcal{P}_n$ *and* $\mathcal{P}_{1,n} \subset \mathcal{P}_n$*. For constants* τ_n *with* $n\tau_n^2 \geq 1$ *and every sufficiently large* $j \in \mathbb{N}$*, assume the following conditions hold.*

(i) For some $C > 0$*,* $\Pi_n\left\{B_{n,0}(\mathbf{\Omega}_{0n}, \tau_n)\right\} \geq e^{-Cn\tau_n^2}$*;*

(ii) Define the set $G_{j,n} = \{\mathbf{\Omega}_n \in \mathcal{P}_{1,n} : j\tau_n < e_n\left(\mathbf{\Omega}_n, \mathbf{\Omega}_{0n}\right) \leq 2j\tau_n\}$*. There exists tests* ϕ_n *such that, for some* $K > 0$*,*

$$\lim_{n\to\infty} \mathbb{E}_{\mathbf{\Omega}_{0n}}\phi_n = 0; \qquad \sup_{\mathbf{\Omega}_n \in G_{j,n}} \mathbb{E}_{\mathbf{\Omega}_n}(1 - \phi_n) \leq \exp\left(-Knj^2\tau_n^2\right).$$

Then, $\Pi_n\left\{\boldsymbol{\Omega}_n \in \mathcal{P}_{1,n} : e_n\left(\boldsymbol{\Omega}_n, \boldsymbol{\Omega}_{0n}\right) > M_n\tau_n | \mathbf{y}_{1:n}\right\} \to 0$ *in* $\mathbb{P}_{\boldsymbol{\Omega}_{0n}}$*-probability for any* $M_n \to \infty$.

We define $\mathcal{P}_{1,n} = \{\boldsymbol{\Omega}_n : \|\boldsymbol{\Omega}_n\|_2 < C''(q_{0n}s_n)^4 \log(d_nq_n)\}$, $\mathcal{P}_{2,n} = \mathcal{P}_n \backslash \mathcal{P}_{1,n}$, $e_n\left(\boldsymbol{\Omega}_n, \boldsymbol{\Omega}_{0n}\right) = \frac{\|\boldsymbol{\Omega}_n - \boldsymbol{\Omega}_{0n}\|_2}{C''(q_{0n}s_n)^4 \log(d_nq_n)}$ and $n\tau_n^2 = q_{0n}s_n \log(d_nq_n)$ for some large enough constant $C'' > 0$. We verify (i) in Lemma 3 and show the existence of a sequence of test functions satisfying (ii) in Lemma 4. Hence,

$\Pi_n\left\{\boldsymbol{\Omega}_n \in \mathcal{P}_{1,n} : e_n\left(\boldsymbol{\Omega}_n, \boldsymbol{\Omega}_{0n}\right) > M_n\tau_n | \mathbf{y}_{1:n}\right\} \to 0$ in $\mathbb{P}_{\boldsymbol{\Omega}_{0n}}$-probability for every $M_n \to \infty$.

Subsequently applying the dominated convergence theorem (DCT), we get $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\Omega}_{0n}} \Pi_n(\boldsymbol{\Omega}_n \in \mathcal{P}_{1,n} : \|\boldsymbol{\Omega}_n - \boldsymbol{\Omega}_{0n}\|_2 > M_n\tau_n | \mathbf{y}_{1:n}) = 0$. To conclude the proof, we show that the remaining mass assigned to $\mathcal{P}_{2,n}$ goes to 0 in the following theorem.

Theorem 3. $\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\Omega}_{0n}} \Pi_n\left(\mathcal{P}_{2,n} | \mathbf{y}_{1:n}\right) = 0$. □

Theorem 2. The theorem closely resembles Theorem 8.22 from Ghosal and van der Vaart (2017). In the discussion following equation (8.22) in the book, the authors noted that for the iid case, simpler theorems are obtained by using an absolute lower bound on the prior mass and by replacing the local entropy by the global entropy. In particular, (i) implies Theorem 8.19(i) in the book. Similarly, (ii) is the same condition in Theorem 8.20 and hence the proof. □

Lemma 2. *(Kullback-Leibler upper bound) Let* $\boldsymbol{\Omega}_n$ *and* $\boldsymbol{\Omega}_{0n}$ *be* $d_n \times d_n$ *order positive definite matrices. Then,*

$$2\mathbb{KL}\left(\boldsymbol{\Omega}_{0n} \parallel \boldsymbol{\Omega}_n\right) = -\log\left|\boldsymbol{\Omega}_{0n}^{-1}\boldsymbol{\Omega}_n\right| + \text{trace}(\boldsymbol{\Omega}_{0n}^{-1}\boldsymbol{\Omega}_n - \mathbf{I}_{d_n}) \leq \|\boldsymbol{\Omega}_{0n} - \boldsymbol{\Omega}_n\|_F^2 \|\boldsymbol{\Omega}_{0n}\|_2 / 2\delta_{\min}^6.$$

Proof. Let $\mathbf{H} = \boldsymbol{\Omega}_{0n}^{-1/2}\boldsymbol{\Omega}_n\boldsymbol{\Omega}_{0n}^{-1/2}$. Letting $\psi_1, \ldots, \psi_{d_n}$ be the eigenvalues of $\mathbf{H}$, we note that

$$2\mathbb{KL}\left(\boldsymbol{\Omega}_{0n} \parallel \boldsymbol{\Omega}_n\right) = \sum_{j=1}^{d_n} \left\{(\psi_j - 1) - \log\psi_j\right\}. \tag{S.1}$$

Consider the function $h_\beta(x) = \log x - (x-1) + \beta(x-1)^2/2$ on $(0, \infty)$ for $\beta > 1$. Note that $h_\beta(x) \geq 0$ for all $x \geq 1/\beta$. From (C3) and using Lemma 1, $\psi_j \geq s_{\min}\left(\mathbf{H}\right) \geq s_{\min}\left(\boldsymbol{\Omega}_n\right) s_{\min}\left(\boldsymbol{\Omega}_{0n}^{-1}\right) = \delta_{\min}^2 / \|\boldsymbol{\Omega}_{0n}\|_2$. Noting that $\delta_{\min}^2 / \|\boldsymbol{\Omega}_{0n}\|_2 < 1$, we set $\beta = \|\boldsymbol{\Omega}_{0n}\|_2 / \delta_{\min}^2$. Therefore,

$$\sum_{j=1}^{d_n} \left\{\log\psi_j - (\psi_j - 1)\right\} \geq -\frac{\|\boldsymbol{\Omega}_{0n}\|_2}{2\delta_{\min}^2} \sum_{j=1}^{d_n} (\psi_j - 1)^2 = -\frac{\|\boldsymbol{\Omega}_{0n}\|_2}{2\delta_{\min}^2} \|\mathbf{H} - \mathbf{I}_{d_n}\|_F^2. \tag{S.2}$$

Again, using Lemma 1 and (C3), we have

$$\|\mathbf{H}-\mathbf{I}_{d_n}\|_F \leq \|\boldsymbol{\Omega}_{0n}-\boldsymbol{\Omega}_n\|_F \left\|\boldsymbol{\Omega}_{0n}^{-1}\right\|_2 \leq \|\boldsymbol{\Omega}_{0n}-\boldsymbol{\Omega}_n\|_F/\delta_{\min}^2. \tag{S.3}$$

Combing (S.1)-(S.3), we conclude the lemma. □

Lemma 3. *Define the set* $B_{n,0}(\boldsymbol{\Omega}_{0n},\tau) = \{\boldsymbol{\Omega}_n \in \mathcal{P}_n : \mathbb{KL}(\boldsymbol{\Omega}_{0n} \parallel \boldsymbol{\Omega}_n) \leq n\tau^2\}$. *Then,* $\Pi_n\{B_{n,0}(\boldsymbol{\Omega}_{0n},\tau_n)\} \geq e^{-Cn\tau_n^2}$ *for* $n\tau_n^2 = s_n q_{0n}\log(d_n q_n)$ *and some absolute constant* $C>0$.

Proof. From Lemma 2, we note that

$$\begin{aligned}\Pi_n\{B_{n,0}(\boldsymbol{\Omega}_{0n},\tau_n)\} &\geq \Pi_n\left(\|\boldsymbol{\Omega}_{0n}-\boldsymbol{\Omega}_n\|_F^2\|\boldsymbol{\Omega}_{0n}\|_2/4\delta_{\min}^6 \leq n\tau_n^2\right)\\ \geq\ \Pi_n\left(\left\|\boldsymbol{\Lambda}_n\boldsymbol{\Lambda}_n^{\mathrm{T}}-\boldsymbol{\Lambda}_{0n}\boldsymbol{\Lambda}_{0n}^{\mathrm{T}}\right\|_F \leq \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\|\boldsymbol{\Omega}_{0n}\|_2^{1/2}}\right)&\Pi_n\left(\|\boldsymbol{\Delta}_n-\boldsymbol{\Delta}_{0n}\|_F \leq \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\|\boldsymbol{\Omega}_{0n}\|_2^{1/2}}\right).\end{aligned} \tag{S.4}$$

We handle the $\boldsymbol{\Lambda}$ and $\boldsymbol{\Delta}$ parts in (S.4) separately and conclude the proof by showing that each of the terms individually exceeds $e^{-Cn\tau_n^2}$ for some constant $C>0$.

The $\boldsymbol{\Lambda}$ part in (S.4): We define $\widetilde{\boldsymbol{\Lambda}}_{0n} = [\boldsymbol{\Lambda}_{0n} \quad \mathbf{0}^{d_n\times\overline{q_n-q_{0n}}}]$, after augmenting q_n-q_{0n} null columns to $\boldsymbol{\Lambda}_{0n}$. Also $\widetilde{\boldsymbol{\Lambda}}_{0n}\widetilde{\boldsymbol{\Lambda}}_{0n}^{\mathrm{T}} = \boldsymbol{\Lambda}_{0n}\boldsymbol{\Lambda}_{0n}^{\mathrm{T}}$. Invoking (C2) and Lemma 1, we have that $\left\|\boldsymbol{\Lambda}_n-\widetilde{\boldsymbol{\Lambda}}_{0n}\right\|_F < \epsilon$ implies that $\left\|\boldsymbol{\Lambda}_n\boldsymbol{\Lambda}_n^{\mathrm{T}}-\widetilde{\boldsymbol{\Lambda}}_{0n}\widetilde{\boldsymbol{\Lambda}}_{0n}^{\mathrm{T}}\right\|_F \leq C\epsilon^2$ for some $C>0$. Note that, from (C2) and (C3), $\|\boldsymbol{\Omega}_{0n}\|_2 = O(1)$. Let $\mathbf{A}^S$ denote the vector of elements of $\mathbf{A}$ corresponding to an index set S, and S_0 denote the set of nonzero elements in $\widetilde{\boldsymbol{\Lambda}}_{0n}$. Then, for some absolute constant $\widetilde{C}$, we have

$$\begin{aligned}\Pi_n\left(\left\|\boldsymbol{\Lambda}_n\boldsymbol{\Lambda}_n^{\mathrm{T}}-\boldsymbol{\Lambda}_{0n}\boldsymbol{\Lambda}_{0n}^{\mathrm{T}}\right\|_F \leq \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\|\boldsymbol{\Omega}_{0n}\|_2^{1/2}}\right) &\geq \Pi_n\left(\left\|\boldsymbol{\Lambda}_n-\widetilde{\boldsymbol{\Lambda}}_{0n}\right\|_F \leq \widetilde{C}n^{1/4}\tau_n^{1/2}\right)\\ \geq \Pi_n\left(\left\|\boldsymbol{\Lambda}_n^{S_0}-\boldsymbol{\Lambda}_{0n}^{S_0}\right\| \leq \widetilde{C}n^{1/4}\tau_n^{1/2}/2\right)&\Pi_n\left(\left\|\boldsymbol{\Lambda}_n^{S_0^C}\right\| \leq \widetilde{C}n^{1/4}\tau_n^{1/2}/2\right)\end{aligned} \tag{S.5}$$

The first term in (S.5) ensures that the Bayesian model assigns requisite prior mass around the *signals* whereas the second term ensures that appropriate *shrinkage* is envisaged through the prior. We show that each of these products exceeds $e^{-Cn\tau_n^2}$.

The shrinkage part in (S.5): From Bhattacharya *et al.* (2015, equation (10)), we get that $\boldsymbol{\Lambda}_n = \frac{2}{b_n}((\lambda^*_{n,j,h}))$ where $\lambda^*_{n,j,h}$'s are marginally iid random variables with a priori with pdf $\Pi(\lambda_{n,j,h}) = |\lambda_{n,j,h}|^{(a_n-1)/2}K_{1-a_n}(\sqrt{2|\lambda_{n,j,h}|})/\{2^{(1+a_n)/2}\Gamma(a_n)\}$ where $K_\nu(x) = \frac{\Gamma(\nu+1/2)(2x)^\nu}{\sqrt{\pi}}\int_0^\infty \frac{\cos t}{(t^2+x^2)^{\nu+1/2}}\mathrm{d}t$ is the modified Bessel function of the second

kind. Using Lemma 3.2 of the same paper and letting $H_n = b_n n^{1/4}\tau_n^{1/2}$, we have

$$\Pi_n\left(\left\|\mathbf{\Lambda}_n^{S_0^C}\right\| \le \widetilde{C}n^{1/4}\tau_n^{1/2}/2\right) \ge \left\{\Pr\left(|\lambda^*_{n,j,h}| \le \frac{CH_n}{|S_0^C|}\right)\right\}^{|S_0^C|} \ge \left\{1 - \frac{C}{\Gamma(a_n)}\log\frac{|S_0^C|}{H_n}\right\}^{|S_0^C|}.$$

Now, $|S_0^C| = d_nq_n - s_nq_{0n}$ and $\Gamma(a_n) = \Gamma(1+a_n)/a_n$. Since $a_n = 1/(d_nq_n)$ and H_n is an increasing sequence, $\frac{1}{\Gamma(a_n)}\log\frac{|S_0^C|}{H_n} = \frac{1}{d_nq_n\Gamma(1+a_n)}\log\frac{d_nq_n - s_nq_{0n}}{H_n} \le 1$ and hence

$$\begin{aligned}\Pi_n\left(\left\|\mathbf{\Lambda}_n^{S_0^C}\right\| \le \widetilde{C}n^{1/4}\tau_n^{1/2}/2\right) &\ge \left\{1 - \frac{C}{\Gamma(a_n)}\log\frac{|S_0^C|}{H_n}\right\}^{|S_0^C|} \ge \left\{1 - \frac{C\log(d_nq_n)}{d_nq_n\Gamma(1+a_n)}\right\}^{d_nq_n} \\ &\ge e^{-C\log(d_nq_n)}\left\{1 - \frac{C^2\log^2(d_nq_n)}{d_nq_n}\right\} \ge e^{-C'q_{0n}s_n\log(d_nq_n)}.\end{aligned}$$

The second last inequality in the previous equation follows since, for any $n \ge 1$ and $|x| \le n$, we have $\left(1 + \frac{x}{n}\right)^n \ge e^x\left(1 - \frac{x^2}{n}\right)$.

The signal part in (S.5): Let us define $v_q(r)$ to be the q-dimensional Euclidean ball of radius r centered at zero and $|v_q(r)|$ denotes its volume. For the sake of brevity, denote $v_q = |v_q(1)|$, so that $|v_q(r)| = r^q v_q$. Letting $\mathbf{\Lambda}_n^* = ((\lambda^*_{n,j,h}))^{d_n\times q_n}$ ($\lambda^*_{n,j,h}$'s defined earlier) and $t_n = b_n\left(\widetilde{C}n^{1/4}\tau_n^{1/2}/2 + \left\|\mathbf{\Lambda}_{0n}^{S_0}\right\|\right)$, we have

$$\begin{aligned}\Pi_n\left(\left\|\mathbf{\Lambda}_n^{S_0} - \mathbf{\Lambda}_{0n}^{S_0}\right\| \le \widetilde{C}n^{1/4}\tau_n^{1/2}/2\right) &= \Pi_n\left(\left\|\mathbf{\Lambda}_n^{*S_0} - b_n\mathbf{\Lambda}_{0n}^{S_0}\right\| \le \widetilde{C}b_nn^{1/4}\tau_n^{1/2}/2\right) \\ &\ge \left|v_{|S_0|}(t_n)\right| \inf_{v_{|S_0|}(t_n)}\Pi_n(\mathbf{\Lambda}_n^{*S_0}).\end{aligned} \tag{S.6}$$

Note that, $|S_0| = s_nq_{0n}$ and $\left\|\mathbf{\Lambda}_{0n}^{S_0}\right\| = \|\mathbf{\Lambda}_{0n}\|_F$ and using Castillo and van der Vaart (2012, Lemma 5.3), $v_q \asymp (2\pi e)^{q/2}q^{-q+1/2}$. Note also that for $x > 0$, $\frac{\log(1+x)}{x} \le \frac{1}{\sqrt{1+x}}$ which implies that $x^x \le e^{x^{3/2}}$. Hence,

$$\left|v_{|S_0|}(t_n)\right| = t_n^{s_nq_{0n}}v_{s_nq_{0n}} \ge \exp\{s_nq_{0n}\log t_n - C'(q_{0n}s_n)^{3/2}/2\} \ge \exp\{-Cs_nq_{0n}\log(d_nq_n)\}.$$

The last inequality follows from (C1) and the prior specifications. From Bhattacharya *et al.* (2015, Lemma 3.1) we have

$$\begin{aligned}&\inf_{v_{|S_0|}(t_n)}\Pi_n(\mathbf{\Lambda}_n^{*S_0}) = \exp\left[-C\left\{s_nq_{0n}\log(1/a_n) + (s_nq_{0n})^{3/4}\sqrt{t_n}\right\}\right] \\ &= \exp\left[-C\left\{s_nq_{0n}\log(d_nq_n) + (s_nq_{0n})^{3/4}b_n^{1/2}\left(\widetilde{C}n^{1/4}\tau_n^{1/2}/2 + \|\mathbf{\Lambda}_{0n}\|_F\right)^{1/2}\right\}\right] \\ &\ge \exp\left(-Cs_nq_{0n}\log\overline{d_nq_n}\right),\end{aligned}$$

since, $b_n = \log^{3/2}(d_n q_n)$, $n\tau_n^2 = s_n q_{0n} \log \overline{d_n q_n}$ and $\|\mathbf{\Lambda}_{0n}\|_F^2 = O(q_{0n})$ from (C2). Hence from (S.6), $\Pi_n\left(\left\|\mathbf{\Lambda}_n^{S_0} - \mathbf{\Lambda}_{0n}^{S_0}\right\| \le \widetilde{C} n^{1/4} \tau_n^{1/2}/2\right) \ge \exp\left(-C s_n q_{0n} \log \overline{d_n q_n}\right)$.

The $\mathbf{\Delta}$ part in (S.4): Note that, $\Pi_n\left(\|\mathbf{\Delta}_n - \mathbf{\Delta}_{0n}\|_F \le \frac{\sqrt{n}\tau_n \delta_{\min}^3}{\|\mathbf{\Omega}_{0n}\|_2^{1/2}}\right) \ge \Pi_n\left(|\delta_{0n}^2 - \delta_n^2| \le \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\sqrt{d_n\|\mathbf{\Omega}_{0n}\|_2}}\right)$. Now, $\Pi_n(|\delta_{0n}^2 - \delta_n^2| \le x) \ge e^{-\delta_{0n}^2}(1 - e^{-2x})$. Using $1 - e^{-2x} \ge x$ for $x \in (0, 1/2)$ and since $\frac{\sqrt{n}\tau_n\delta_{\min}^3}{\sqrt{d_n\|\mathbf{\Omega}_{0n}\|_2}} \to 0$ as $n \to \infty$, we have for some $C' > 0$

$$\Pi_n\left(|\delta_{0n}^2 - \delta_n^2| \le \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\sqrt{d_n\|\mathbf{\Omega}_{0n}\|_2}}\right) \ge \exp\left(-\delta_0^2 + \log \frac{\sqrt{n}\tau_n\delta_{\min}^3}{\sqrt{d_n\|\mathbf{\Omega}_{0n}\|_2}}\right) \ge \exp\left(-C' s_n q_{0n} \log \overline{d_n q_n}\right).$$

□

Corollary 1. *For arbitrary $\widetilde{\tau}_n > n^{-1/2}$, we have $\Pi_n\{B_{n,0}(\mathbf{\Omega}_{0n}, \widetilde{\tau}_n)\} \ge e^{-C q_{0n} s_n \log(d_n q_n)}$ for some constant $C > 0$.*

Following the proof of Lemma 3, it can be seen that the above corollary holds.

Lemma 4. *Let $\mathbf{\Omega}_{0n} \in \mathcal{P}_{0n}$ and the set $G_{j,n} = \left\{\mathbf{\Omega}_n \in \mathcal{P}_{1,n} : j\tau_n < \frac{1}{\zeta_n}\|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 \le 2j\tau_n\right\}$ denote an annulus of inner radius $j\tau_n\zeta_n$ and outer radius $(j+1)\tau_n\zeta_n$, where $\zeta_n = C''(q_{0n}s_n)^4 \log(d_n q_n)$, in operator norm around $\mathbf{\Omega}_{0n}$ for some integer $j > 1$. Based on iid samples $\mathbf{y}_{1:n}$ from $\mathrm{N}_{d_n}(\mathbf{0}, \mathbf{\Omega}_{0n}^{-1})$, consider the following hypothesis testing problem*

$$H_0 : \mathbf{\Omega}_n = \mathbf{\Omega}_{0n} \text{ versus } H_1 : \mathbf{\Omega}_n \in G_{j,n}.$$

We simulate $\mathbf{u}_{1:n} \overset{iid}{\sim} \mathrm{N}_{q_{0n}}(\mathbf{0}, \mathbf{P}_{0n})$, with $\mathbf{P}_{0n} = (\mathbf{I}_{q_{0n}} + \mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Delta}_{0n}^{-1}\mathbf{\Lambda}_{0n})$ independently from $\mathbf{y}_{1:n}$ and define $\mathbf{v}_i = \mathbf{y}_i + \mathbf{\Delta}_{0n}^{-1}\mathbf{\Lambda}_{0n}\mathbf{P}_{0n}^{-1}\mathbf{u}_i$. Letting $\mathbf{V}^{\mathrm{T}} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)$, we define the following test function $\phi_n = \mathbb{1}\left\{\left\|\mathbf{\Lambda}_{0n}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{\Delta}_{0n}^{-1}\right)\mathbf{\Lambda}_{0n}\right\|_2 > \tau_n\right\}$. Then

$$\lim_{n\to\infty} \mathbb{E}_{\mathbf{\Omega}_{0n}}\phi_n = 0; \qquad \sup_{\mathbf{\Omega}_n \in G_{j,n}} \mathbb{E}_{\mathbf{\Omega}_n}(1 - \phi_n) \le \exp\left(-K n j^2 \tau_n^2\right),$$

for some absolute constant $K > 0$.

Proof. **Type-I error:** Under H_0, we have $\mathbf{v}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_{d_n}(\mathbf{0}, \mathbf{\Delta}_{0n}^{-1})$. Letting $\mathbf{\Xi}_{0n} = \mathrm{cov}\left(\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{v}_i\right) = \mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Delta}_{0n}^{-1}\mathbf{\Lambda}_{0n}$, we have, $\phi_n \le \mathbb{1}\left(\|\mathbf{\Xi}_{0n}\|_2 \left\|\frac{1}{n}\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i^{\mathrm{T}} - \mathbf{I}_{q_{0n}}\right\|_2 > \tau_n\right)$, where $\|\mathbf{\Xi}_{0n}\|_2 = O(1)$ from (C2). From Vershynin (2012, Corollary 5.50), $\mathbb{E}_{H_0}\phi_n \le 2\exp\left(-\widetilde{C} q_{0n} t_n^2\right)$ for a universal constant $\widetilde{C} > 0$ and any sequence t_n satisfying $q_{0n}t_n^2 \le n\tau_n^2$. Since from (C1), $q_{0n} \to \infty$ and $n\tau_n^2/q_{0n} \to \infty$ as $n \to \infty$, t_n can be constructed such that $t_n \gtrsim 1$ and hence $\lim_{n\to\infty}\mathbb{E}_{H_0}\phi_n = 0$.

Type-II error: If $\mathbf{y}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_{d_n}(\mathbf{0}, \mathbf{\Omega}_n)$ with $\mathbf{\Omega}_n = \mathbf{\Lambda}_n\mathbf{\Lambda}_n^{\mathrm{T}} + \mathbf{\Delta}_n$, then $\mathbf{v}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_{d_n}(\mathbf{0}, \mathbf{\Delta}_n^{-1} + \mathbf{\nabla}_n)$ where $\mathbf{\nabla}_n = \mathbf{\Delta}_{0n}^{-1}\mathbf{\Lambda}_{0n}(\mathbf{I}_{q_{0n}} + \mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Delta}_{0n}^{-1}\mathbf{\Lambda}_{0n})^{-1}\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Delta}_{0n}^{-1} - \mathbf{\Delta}_n^{-1}\mathbf{\Lambda}_n(\mathbf{I}_{q_n} + \mathbf{\Lambda}_n^{\mathrm{T}}\mathbf{\Delta}_n^{-1}\mathbf{\Lambda}_n)^{-1}\mathbf{\Lambda}_n^{\mathrm{T}}\mathbf{\Delta}_n^{-1}$. Notably, $\mathbf{\Delta}_n^{-1} - \mathbf{\Delta}_{0n}^{-1} + \mathbf{\nabla}_n = \mathbf{\Omega}_n - \mathbf{\Omega}_{0n}$. Now,

$$
\begin{aligned}
1 - \phi_n &= \mathbb{1}\left\{\left\|\mathbf{\Lambda}_{0n}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{\Delta}_{0n}^{-1}\right)\mathbf{\Lambda}_{0n}\right\|_2 \le \tau_n\right\} \\
&= \mathbb{1}\left\{\left\|\mathbf{\Lambda}_{0n}^{\mathrm{T}}\left\{\left(\frac{1}{n}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{\Delta}_n^{-1} - \mathbf{\nabla}_n\right) + (\mathbf{\Delta}_n^{-1} - \mathbf{\Delta}_{0n}^{-1} + \mathbf{\nabla}_n)\right\}\mathbf{\Lambda}_{0n}\right\|_2 \le \tau_n\right\} \\
&\le \mathbb{1}\left\{\left\|\mathbf{\Lambda}_{0n}^{\mathrm{T}}\left(\frac{1}{n}\mathbf{V}^{\mathrm{T}}\mathbf{V} - \mathbf{\Delta}_n^{-1} - \mathbf{\nabla}_n\right)\mathbf{\Lambda}_{0n}\right\|_2 > \left\|\mathbf{\Lambda}_{0n}^{\mathrm{T}}(\mathbf{\Omega}_n - \mathbf{\Omega}_{0n})\mathbf{\Lambda}_{0n}\right\|_2 - \tau_n\right\} \\
&\le \mathbb{1}\left\{\left\|\left\{\frac{1}{n}\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{\Lambda}_{0n} - \mathbf{\Lambda}_{0n}^{\mathrm{T}}(\mathbf{\Delta}_n^{-1} + \mathbf{\nabla}_n)\mathbf{\Lambda}_{0n}\right\}\right\|_2 > s_{\min}\left(\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Lambda}_{0n}\right)\|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 - \tau_n\right\} \\
&\le \mathbb{1}\left\{\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i^{\mathrm{T}} - \mathbf{I}_{q_n}\right\|_2 > \frac{s_{\min}\left(\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Lambda}_{0n}\right)}{\left\|\mathbf{\Delta}_n^{-1} + \mathbf{\nabla}_n\right\|_2\|\mathbf{\Lambda}_{0n}\|_2^2}\left(\|\mathbf{\Omega}_n - \mathbf{\Omega}_{0n}\|_2 - \frac{\tau_n}{s_{\min}\left(\mathbf{\Lambda}_{0n}^{\mathrm{T}}\mathbf{\Lambda}_{0n}\right)}\right)\right\} \qquad \text{(S.7)}
\end{aligned}
$$

where $\mathbf{z}_{1:n} \overset{\text{iid}}{\sim} \mathrm{N}_{q_n}(\mathbf{0}, \mathbf{I}_{q_n})$. By construction of $\mathcal{P}_{1,n}$, $\left\|\mathbf{\Delta}_n^{-1} + \mathbf{\nabla}_n\right\|_2 \le \zeta_n$ for $\mathbf{\Omega}_n \in \mathcal{P}_{1,n}$. Hence, for $\mathbf{\Omega}_n \in G_{j,n}$, the RHS of (S.7) is bounded below by $j\tau_n$ for sufficiently large j. Using Vershynin (2012, Eqn 5.26) again, we have for all $\mathbf{\Omega}_n \in G_{j,n}$, $\mathbb{E}_{\mathbf{\Omega}_n}(1 - \phi_n) \le e^{-Knj^2\tau_n^2}$.

Hence the proof. □

Theorem 3. For densities p and q, define $\mathbb{V}(p \parallel q) = \int\left\{\log\frac{p}{q} - \mathbb{KL}(p \parallel q)\right\}^2 \mathrm{d}p$. When p and q are mean zero multivariate Gaussian densities with precision matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}'$, we simply denote $\mathbb{V}(\mathbf{\Omega} \parallel \mathbf{\Omega}')$. We define the set $B_{n,2}(\mathbf{\Omega}_{0n}, \epsilon) = \{\mathbf{\Omega}_n \in \mathcal{P}_n : \mathbb{KL}(\mathbf{\Omega}_{0n} \parallel \mathbf{\Omega}_n) \le n\epsilon^2,\ \mathbb{V}(\mathbf{\Omega}_{0n} \parallel \mathbf{\Omega}_n) \le n\epsilon^2\}$.

From Banerjee and Ghosal (2015, Theorem 3.1) and the proof of Lemma 2, we have $\mathbb{V}(\mathbf{\Omega}_{0n} \parallel \mathbf{\Omega}_n) = \frac{1}{2\delta_{\min}^4}\|\mathbf{\Omega}_{0n} - \mathbf{\Omega}_n\|_F^2$ and therefore $B_{n,2}(\mathbf{\Omega}_{0n}, \epsilon) = \{\mathbf{\Omega}_n \in \mathcal{P}_n : \|\mathbf{\Omega}_{0n} - \mathbf{\Omega}_n\|_F^2 \le C\epsilon\}$ for some constant $C > 0$.

Let $\{t_n\}_{n=1}^{\infty}$ be an increasing sequence of positive numbers. Then, $\Pi_n(\|\mathbf{\Lambda}_n\|_2 \ge t_n) \le \Pi_n(\|\mathbf{\Lambda}_n\|_F \ge t_n) \le \Pi_n(\|\mathrm{vec}(\mathbf{\Lambda}_n^*)\|_{\ell_1} \ge b_n t_n) \le 2e^{-C\sqrt{b_n t_n}}$ for some constant $C > 0$ where $\|\cdot\|_{\ell_1}$ is the ℓ_1 norm. The last inequality follows from Pati *et al.* (2014, Lemma 7.4). Hence, for $\zeta_n = C''(q_{0n}s_n)^4\log(d_nq_n)$,

$$
\Pi_n(\mathcal{P}_{2,n}) \le \Pi_n\left\{\delta_{\max}^2 + \|\mathbf{\Lambda}_n\|_2^2 \ge \zeta_n\right\} \ge \Pi_n\left\{\|\mathbf{\Lambda}_n^*\|_2^2 \ge K'C''(q_{0n}s_n\log\overline{d_nq_n})^4\right\} \le e^{-CC''q_{0n}s_n\log(d_nq_n)}.
$$

From Corollary 1, for arbitrary $\widetilde{\tau}_n > n^{-1/2}$, $\Pi_n\{B_{n,2}(\mathbf{\Omega}_{0n}, \widetilde{\tau}_n)\} \ge e^{-Cq_{0n}s_n\log(d_nq_n)}$. Hence, $\frac{\Pi_n(\mathcal{P}_{2,n})}{\Pi_n\left\{B_{n,2}(\mathbf{\Omega}_{0n}, \widetilde{\tau}_n)\right\}} \le e^{-2Cq_{0n}s_n\log(d_nq_n)} = o(e^{-2n\widetilde{\tau}_n^2})$ for $\widetilde{\tau}_n < \tau_n$. The last display holds for suitable large enough choice of C'', which, however, can be chosen independent of n. Using Ghosal and van der Vaart (2017, Theorem 8.20) and subsequent application of DCT we

conclude the proof. □

Theorem 4 (minimax rate)**.** *If $\widehat{\mathbf{\Omega}}_n$ is a sequence of estimators of $\mathbf{\Omega}_{0n} \in \mathcal{P}_{0n}$ with $q_{0n} = O(1)$, then for some absolute constant $C > 0$*

$$\inf_{\widehat{\mathbf{\Omega}}_n} \sup_{\mathbf{\Omega}_{0n} \in \mathcal{P}_{0n}} \left\| \widehat{\mathbf{\Omega}}_n - \mathbf{\Omega}_{0n} \right\|_2 \geq C \sqrt{\frac{s_n \log d_n}{n}}.$$

Proof. We use Fano's lemma to derive a lower bound for the minimax risk. Let $\mathcal{F} = \{\mathbf{\Omega}_{(1)}, \ldots, \mathbf{\Omega}_{(m_n)}\}$, $m_n \geq 2$ be a finite subset of $\mathcal{P}_{0n}$ and let $\widehat{\mathbf{\Omega}}_n$ be an estimate of $\mathbf{\Omega}_{0n}$ based on iid samples $\mathbf{y}_{1:n}$. Suppose for all $j \neq j'$, we have $e(\mathbf{\Omega}_{(j)}, \mathbf{\Omega}_{(j')}) \geq e_{m_n}$ for some metric e and $\mathbb{KL}\left(\mathbf{\Omega}_{(j)} \parallel \mathbf{\Omega}_{(j')}\right) \leq K_{m_n}$. Letting $\mathbb{E}_j$ denote the expectation under $\mathrm{N}_{d_n}(\mathbf{0}, \mathbf{\Omega}_{(j)}^{-1})$, Fano's lemma (Yu, 1997) implies

$$\max_{1 \leq j \leq m_n} \mathbb{E}_j e(\mathbf{\Omega}_{(j)}, \mathbf{\Omega}_{(j')}) \leq \frac{e_{m_n}}{2} \left(1 - \frac{K_{m_n} + \log 2}{\log m_n}\right). \tag{S.8}$$

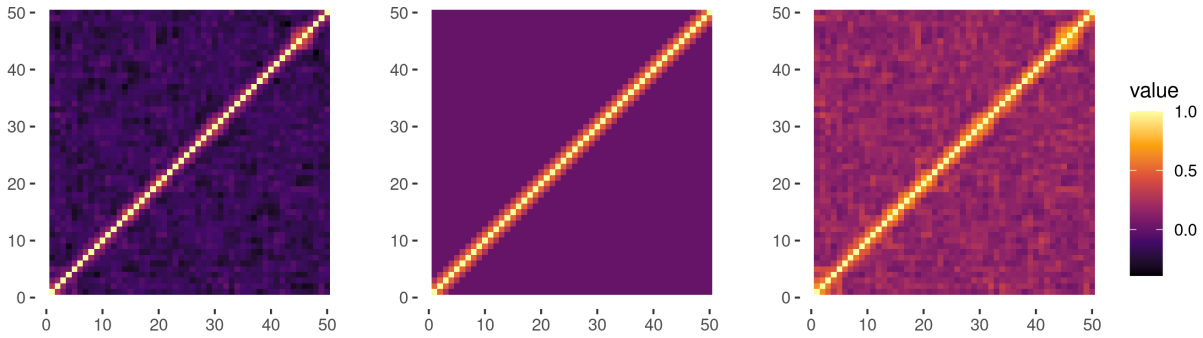
We now construct the finite class $\mathcal{F}$. Let $r_n = d_n - 1$. Define $\mathcal{M} = \{\mathbf{x} \in \mathbb{R}^{r_n} : x_j \in \{0, 1\}$ for all j, $\sum_j x_j = s_n\}$ to be the collection of all binary vectors of length r_n with exactly s_n ones and let $\langle \mathbf{x}, \mathbf{y} \rangle_H$ be the Hamming distance between two binary vectors $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{g}_j = (\mathbf{x}_j, 0)$ denote the d_n-dimensional vector obtained by appending zero at the end of $\mathbf{x}_j$ with $\mathbf{x}_j \in \mathcal{M}$. With this definitions, set $\mathbf{\Omega}_{(j)} = \beta_n \mathbf{I}_{d_n} + \gamma_n \mathbf{g}_j \mathbf{g}_j^{\mathrm{T}} + \kappa_n \boldsymbol{\varepsilon}_{d_n} \boldsymbol{\varepsilon}_{d_n}^{\mathrm{T}}$ where $\boldsymbol{\varepsilon}_{d_n}$ is the vector with 1 in the d_n^{th} coordinate and zero elsewhere, and $\gamma_n \leq \beta_n \leq \kappa_n$ are positive sequences to be chosen below.

Observe that if $\langle \mathbf{g}_j, \mathbf{g}_{j'} \rangle_H = s_n - p_n$ for $j \neq j$, then $\mathbf{g}_j^{\mathrm{T}} \mathbf{g}_{j'} = p_n$. Note also that $\mathbf{\Omega}_{(j)} - \mathbf{\Omega}_{(j')} = \gamma_n (\mathbf{g}_j \mathbf{g}_j^{\mathrm{T}} - \mathbf{g}_{j'} \mathbf{g}_{j'}^{\mathrm{T}})$. The nonzero eigenvalues of the matrix $\mathbf{B} = \mathbf{g}_j \mathbf{g}_j^{\mathrm{T}} - \mathbf{g}_{j'} \mathbf{g}_{j'}^{\mathrm{T}}$ are $(\sqrt{s_n^2 - p_n^2}, -\sqrt{s_n^2 - p_n^2})$, since $\mathrm{rank}(\mathbf{B}) = 2$, $\mathrm{trace}(\mathbf{B}) = 0$ and $\mathrm{trace}(\mathbf{B}^2) = 2(s_n^2 - p_n^2)$. This implies that $\left\| \mathbf{\Omega}_{(j)} - \mathbf{\Omega}_{(j')} \right\|_2 = \gamma_n (s_n^2 - p_n^2)$. Since $\mathbf{g}_j \in \mathcal{M}$ for all j, by symmetry $\left| \mathbf{\Omega}_{(j)} \right| = \left| \mathbf{\Omega}_{(j')} \right|$ for all $j \neq j'$. Hence $\mathbb{KL}\left(\mathbf{\Omega}_{(j)} \parallel \mathbf{\Omega}_{(j')}\right) = \frac{1}{2} \mathrm{trace}\left\{ \mathbf{\Omega}_{(j)} \mathbf{\Omega}_{(j')}^{-1} - d_n \right\}$. Let $\mathbf{A} = \beta_n (\mathbf{A} + t \mathbf{g}_j \mathbf{g}_j^{\mathrm{T}})$, where $\mathbf{A}$ is a diagonal matrix with the first $(d_n - 1)$ diagonals equaling one and the d_n^{th} entry being $(1 + \kappa_n / \beta_n)$. Subsequently, applying the Woodbury matrix inversion identity, we get $(\mathbf{A} + t_n \mathbf{g}_{j'} \mathbf{g}_{j'}^{\mathrm{T}})^{-1} = \mathbf{A}^{-1} - \frac{t_n}{1 + t_n s_n} \mathbf{g}_{j'} \mathbf{g}_{j'}^{\mathrm{T}}$ so that $\mathbf{\Omega}_{(j)} \mathbf{\Omega}_{(j')}^{-1} = \mathbf{I}_{d_n} - \frac{t_n}{1 + t_n s_n} \mathbf{g}_j \mathbf{g}_j^{\mathrm{T}} + t_n \mathbf{g}_{j'} \mathbf{g}_{j'}^{\mathrm{T}} - \frac{t_n^2 p_n}{1 + t_n s_n} \mathbf{g}_j \mathbf{g}_{j'}^{\mathrm{T}}$ with $t_n = \gamma_n / \beta_n$. Observing that $\mathrm{trace}(\mathbf{g}_j \mathbf{g}_j^{\mathrm{T}}) = s_n$ and $\mathrm{trace}(\mathbf{g}_j \mathbf{g}_{j'}^{\mathrm{T}}) = p_n$, we get $\mathbb{KL}\left(\mathbf{\Omega}_{(j)} \parallel \mathbf{\Omega}_{(j')}\right) = \frac{1}{2} \frac{t_n^2}{t_n s_n + 1} (s_n^2 - p_n^2)$.

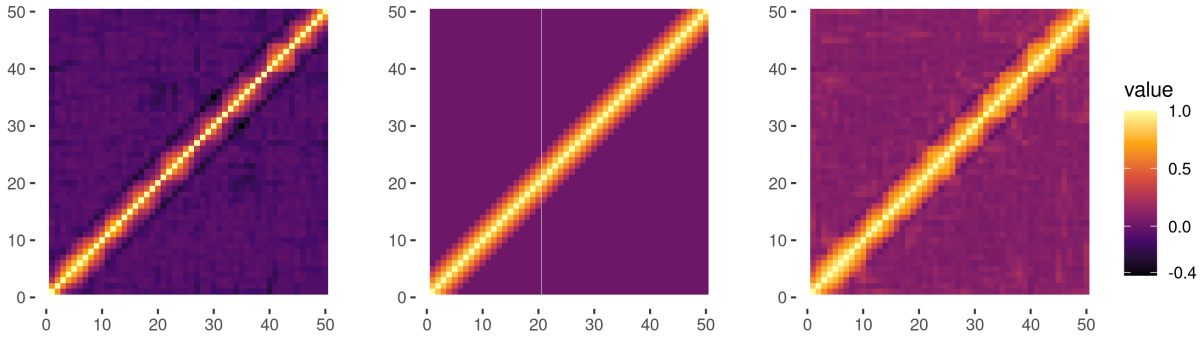
Now, from Pati *et al.* (2014, Lemma 5.6), given $s_n \geq 6$, there exists a subset $\mathcal{M}_0 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{m_n}\}$ of $\mathcal{M}$ with $m_n \asymp \exp(C s_n \log d_n)$ and $\langle \mathbf{x}_j, \mathbf{x}_{j'} \rangle_H \geq s_n / 3$ for all $1 \leq j \neq j' \leq m_n$, where C is a positive constant independent of d_n. We set $\mathcal{F} = \{\mathbf{\Omega}_{(j)} : \mathbf{x}_j \in \mathcal{M}_0\}$. Using the aforementioned lemma and preceding discussions, we have that p_n is bounded above by $2s_n/3$ fo all pairs $j \neq j' \in \mathcal{M}_0$. Hence, we can choose $e_{m_n} \geq c_1 \gamma_n s_n$ and $K_{m_n} = (t s_n)^2 = (\gamma_n s_n / \beta_n)^2$ in (S.8). To obtain e_{m_n} as a lower bound to the minimax risk up to a constant, we set $K_{m_n} / \log m_n = C'$ for some $C' \in (0, 1)$. Since $\log m_n \asymp C s_n \log d_n$, we obtain by choosing $\beta_n, \kappa_n \asymp 1$, that $e_{m_n}^2 = C (\gamma_n s_n)^2 = C \frac{s_n \log d_n}{n}$. □

S.3 Additional Figures

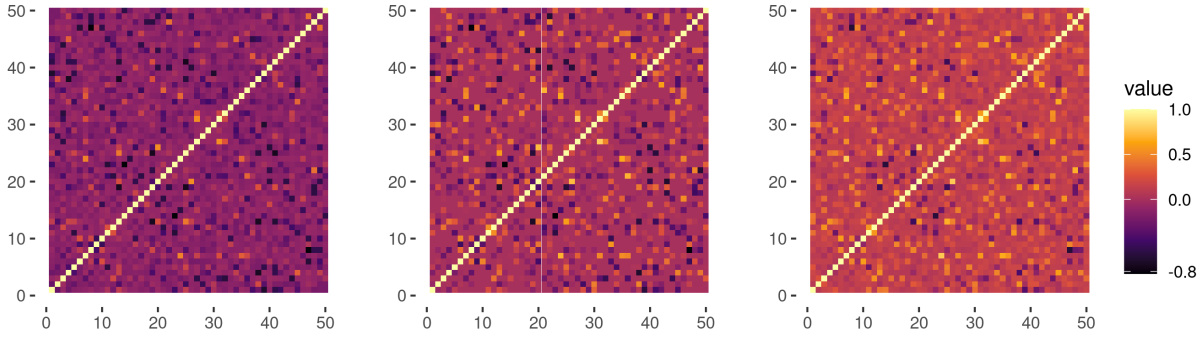
S.3.1 Uncertainty Quantification



(a) Quantiles for the AR(2) case



(b) Quantiles for the banded case

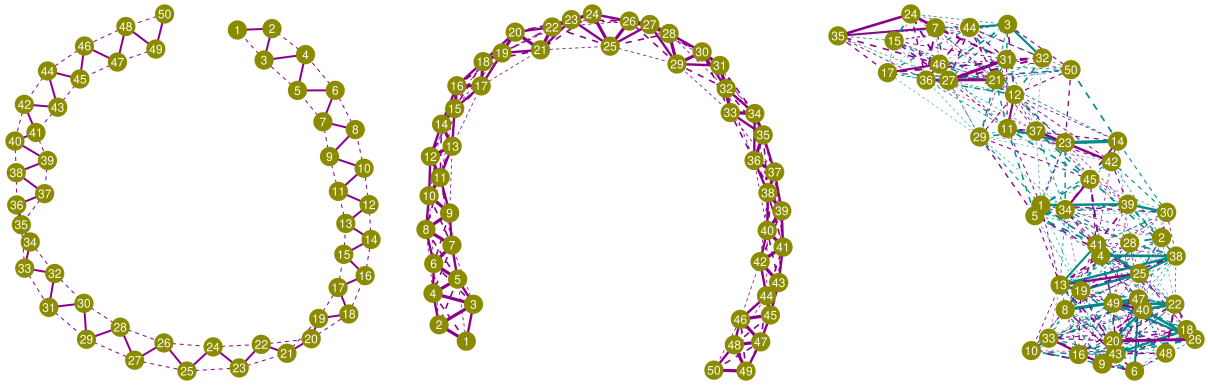


(c) Quantiles for the RSM case

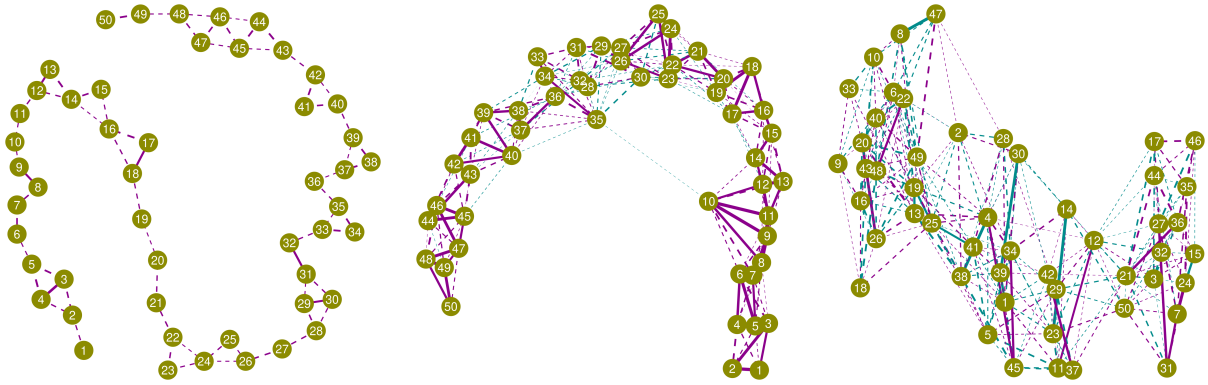
Figure S.1: Results of simulation experiments: Uncertainty quantification: In each panel, the middle heatmap is the simulation truth of $\mathbf{R} = \text{diag}(\boldsymbol{\Omega})^{-1/2}\ \boldsymbol{\Omega}\ \text{diag}(\boldsymbol{\Omega})^{-1/2}$, the left and the right heatmaps show the lower 2.5% and the upper 97.5% quantiles of $\mathbf{R}$, respectively, estimated from the MCMC samples.

S.3.2 Graph Estimation

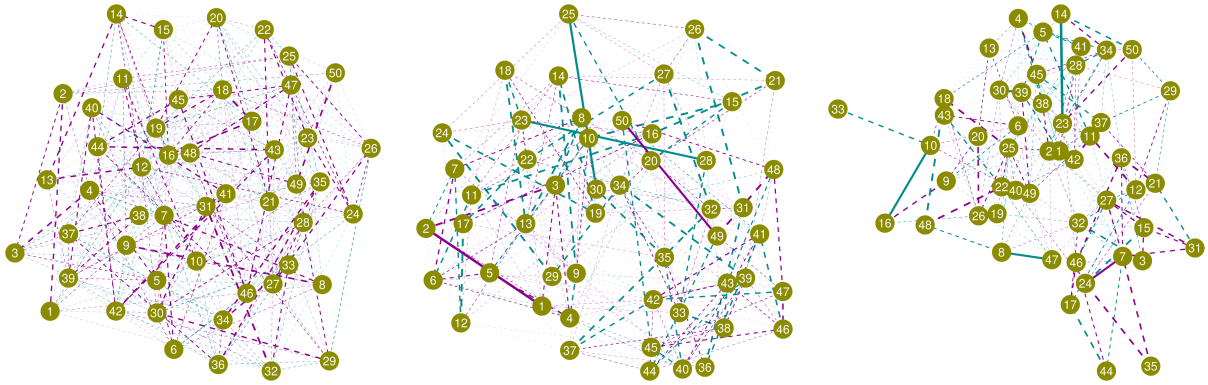
True and estimated graphs derived from the respective precision matrices as described in Section 3 of the main paper using the `GGally` package (Schloerke *et al.*, 2020) in `R`.



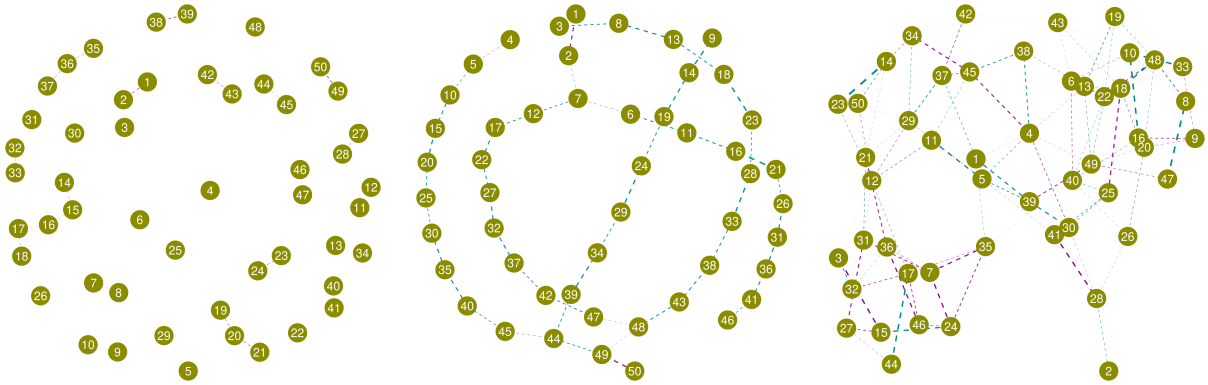
(a) True graphs.



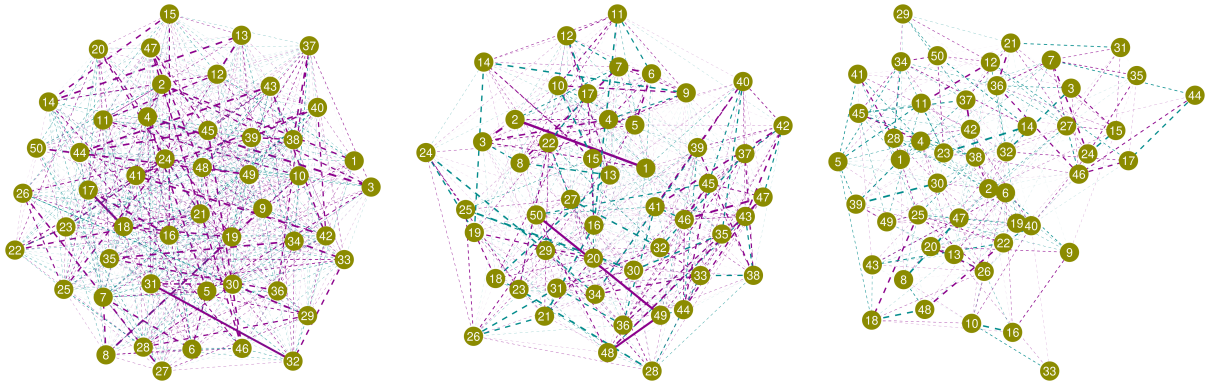
(b) Estimated graphs by the PF method.



(c) Estimated graphs by Bagus.



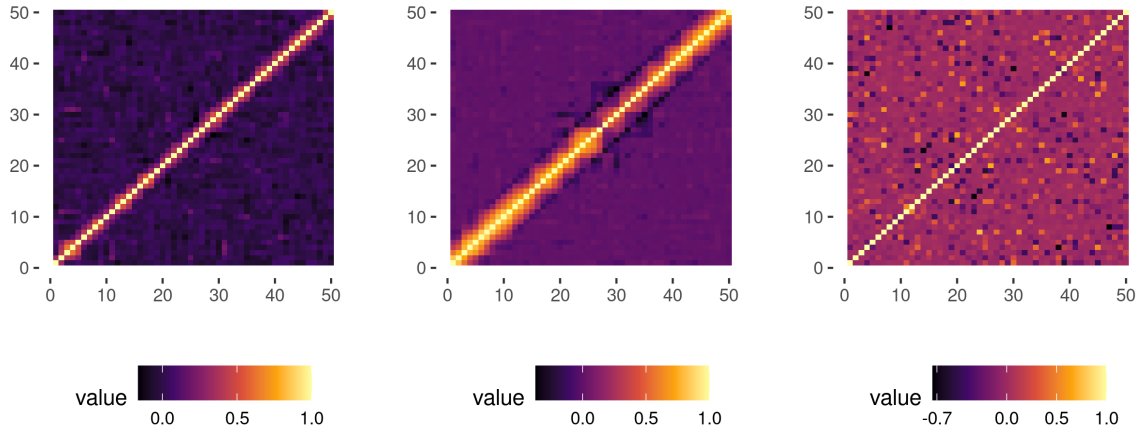
(d) Estimated graphs by M&B.



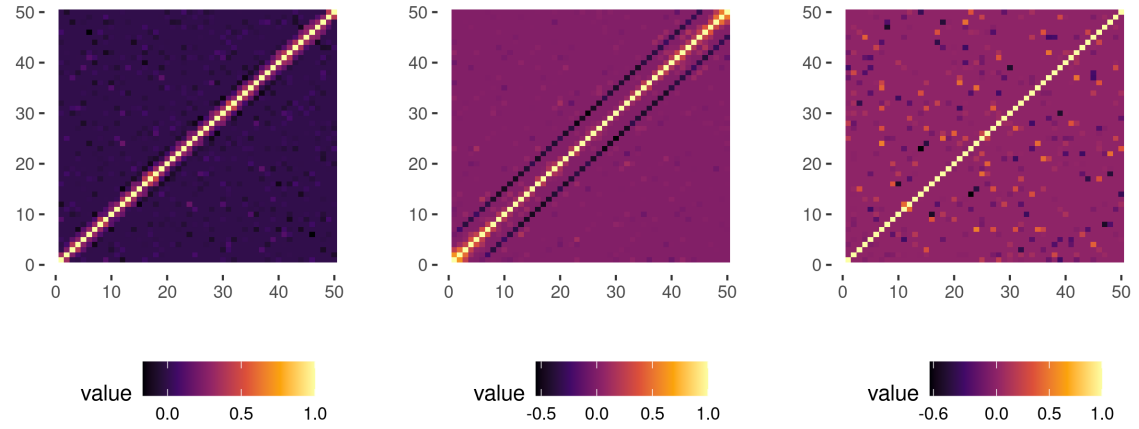
(e) Estimated graphs by Glasso.

Figure S.2: Results of simulation experiments: Graph recovery: Panel (a) shows the true graphs for AR(2), banded and RSM structures from left to right; panels (b), (c), (d) and (e) show the corresponding estimated graphs for the our proposed PF and Bagus, M&B and Glasso methods, respectively. Positive (negative) associations are represented by blue (magenta) edges and edge-widths are proportional to the association strength. If the absolute value of a partial correlation coefficient is greater (less) than 0.5, the corresponding edge is represented by a solid (dotted) line.

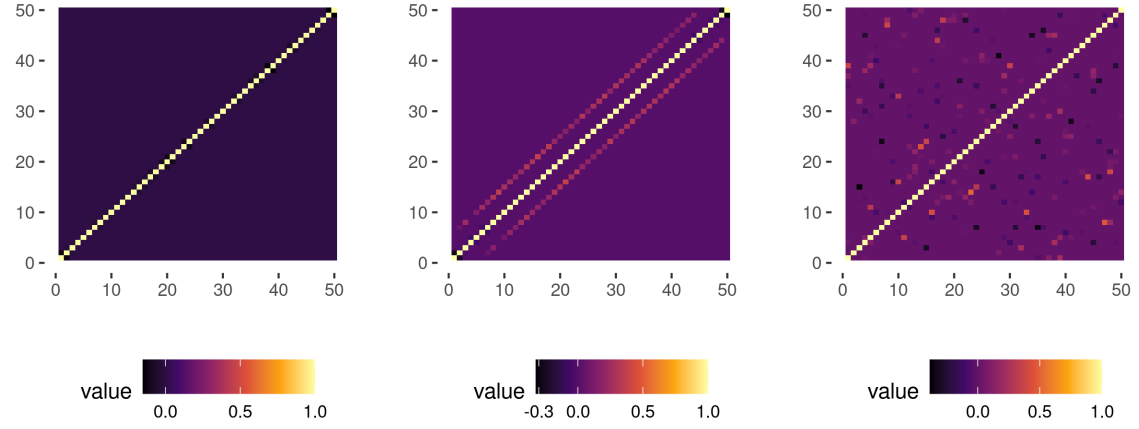
S.3.3 Precision Matrix Estimation



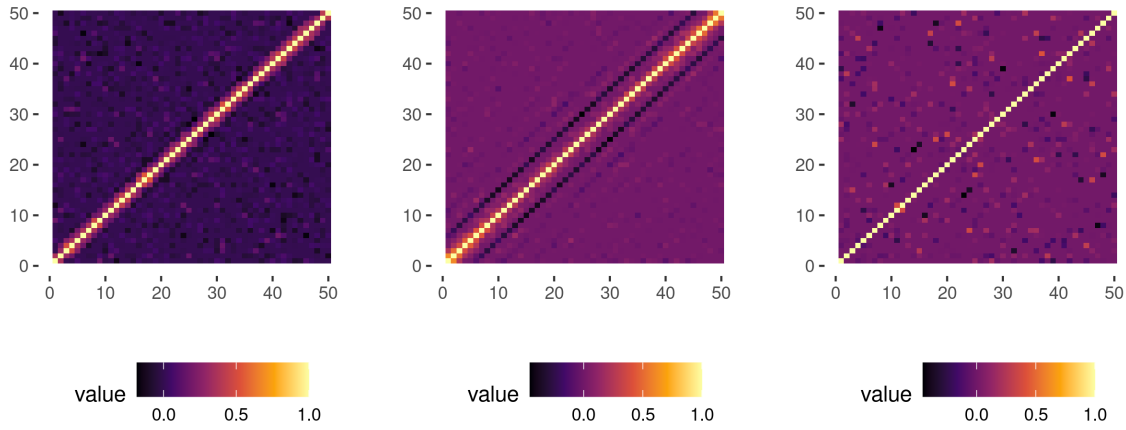
(a) Results obtained by the proposed PF method.



(b) Results obtained by Bagus.



(c) Results obtained by M&B.



(d) Results obtained by Glasso.

Figure S.3: Results of simulation experiments: Recovery of the precision matrices: Panels (a), (b), (c) and (d) show the heatmaps of the estimated scaled precision matrices $\mathbf{R} = \text{diag}(\mathbf{\Omega})^{-1/2}\ \mathbf{\Omega}\ \text{diag}(\mathbf{\Omega})^{-1/2}$ for PF, Bagus, M&B and Glasso, respectively, in left to right order for the AR(2), banded and RSM model.

S.4 NASDAQ-100 Stock Price Data

Here we apply the precision factor model to a stock price dataset comprising the top 100 companies listed in NASDAQ for the period Jan 2015 - Dec 2019 recorded every week. We obtained the data from Yahoo finance. After removing missing records, we ended up with data on 91 companies. First, we removed the trend by linear trend fitting. The estimated graph shown in Figure S.4 indicates a sparse structure.

We highlight some interesting features observed from the analysis. We observe a strong positive association between the rival GPU developers AMD and Nvidia Corporation (NVDA) consistent with the increasing demand of GPU computing. Similarly, a positive association between Qualcomm (QCOM) and Marvell Technology Group (MRVL), both of which develop and produce semiconductors and related technology, can be seen. The electronic design automation company Synopsys (SNPS) and the software company Cadence Design Systems, Inc. (CDNS) exhibit strong positive associations. Electronic Arts (EA) and Activision (ATVI) are video game developers and a Positive association is observed between them. ATVI seem to have strong associations with Amazon.com, Inc. (AMZN) and Apple Inc (AAPL). The semiconductor manufacturer companies Texas Instruments (TXN) and Xilinx (XLNX) seem to have a negative association. We observe positive association between the computer memory and data storage producing company Micron Technology, Inc. (MU) and Match Group (MTCH). Interestingly, the American semiconductor company Skyworks Solutions, Inc. (SWKS) exhibit strong negative association with companies like Microsoft (MSFT), Comcast Corporation (CMCSA), MU, etc. Many of the aforementioned

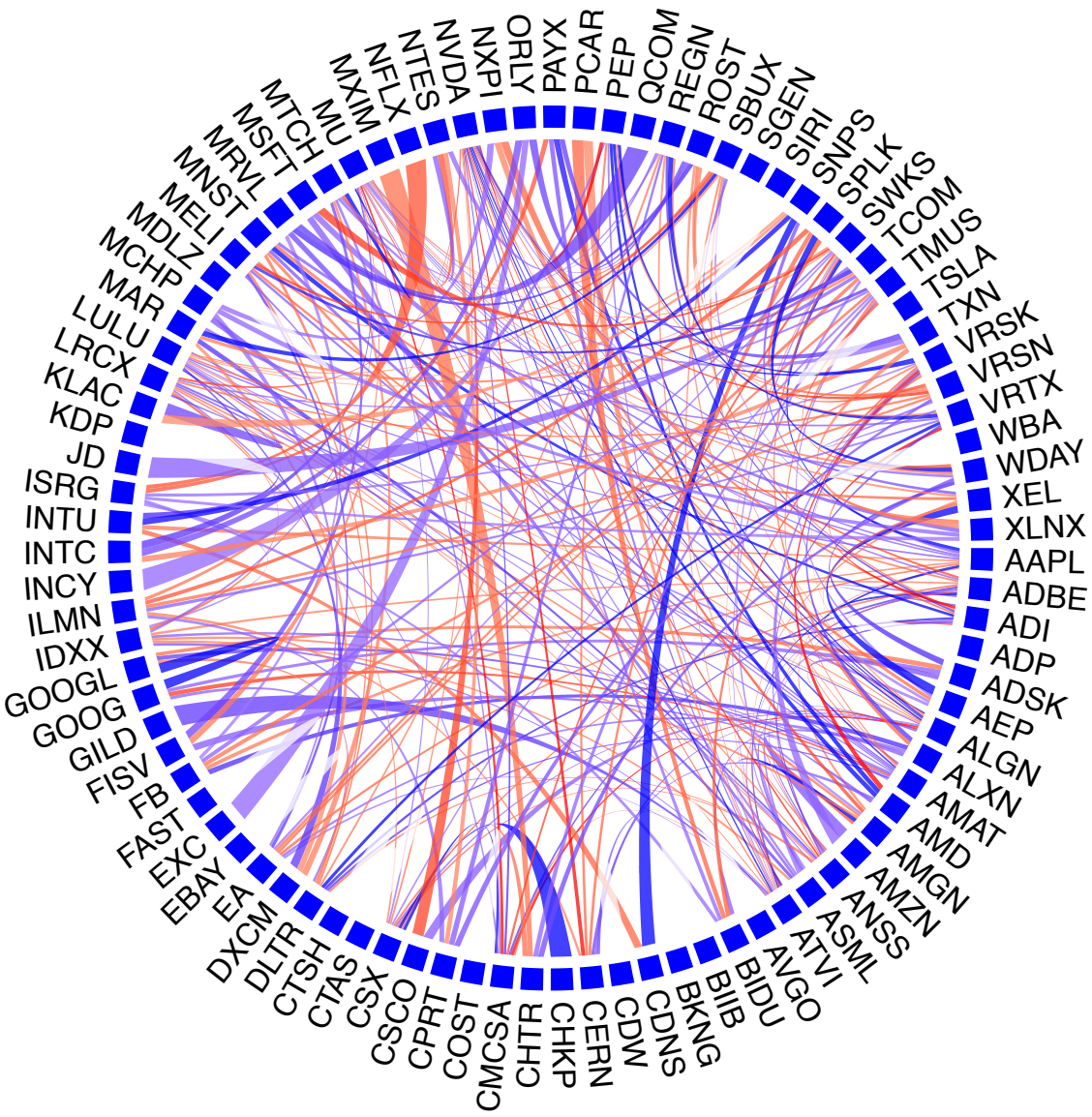


Figure S.4: Results for NASDAQ-100 stock price data: Positive (negative) associations are represented by blue (red) links, their opacities being proportional to the corresponding association strengths. The link widths are inversely proportional to the number of edges associated with the corresponding nodes.

stocks exhibit strong positive association with the Meta Platforms, Inc. (FB) previously known as Facebook Inc. As expected, we observe a strong positive association between GOOGL and GOOG, which are Google shares with and without voting rights, respectively.

Notably, most of the connected nodes in the NASDAQ-100 listing correspond to technology companies and electronics manufacturers. This is in accordance with the ushering in of the 'digital era' over the last decade, with technology giants taking over major shares of the market. Also, the strong negative associations between several companies in similar domains reflect the competitive nature of the market.

References

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

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, **110**, 1479–1490.

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.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

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.

Schloerke, B., Cook, D., Larmarange, J., Briatte, F., Marbach, M., Thoen, E., Elberg, A., and Crowley, J. (2020). *GGally: Extension to 'ggplot2'*. R package version 2.0.0.

Vershynin, R. (2012). *Introduction to the non-asymptotic analysis of random matrices*, page 210–268. Cambridge University Press.

Yu, B. (1997). Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam*, pages 423–435. Springer.