Electronic Journal of Statistics
Vol. 11 (2017) 3485–3521
ISSN: 1935-7524
DOI: 10.1214/17-EJS1331

Graphical model selection with latent variables

Changjing Wu

School of Mathematical Sciences, Peking University, Beijing 100871, P.R.C.
e-mail: wcj@pku.edu.cn

Hongyu Zhao

School of Mathematical Sciences, Peking University, Beijing 100871, P.R.C.

Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut 06520, U.S.A.
e-mail: hongyu.zhao@yale.edu

Huaying Fang

Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, U.S.A.
e-mail: hyfang@stanford.edu

and

Minghua Deng

School of Mathematical Sciences, Peking University, Beijing 100871, P.R.C.

Center for Statistical Science, Peking University, Beijing 100871, P.R.C.

Center for Quantitative Biology, Peking University, Beijing 100871, P.R.C.
e-mail: dengmh@pku.edu.cn

Abstract: Gaussian graphical models are commonly used to characterize the conditional dependence among variables. However, ignorance of the effect of latent variables may blur the structure of a graph and corrupt statistical inference. In this paper, we propose a method for learning Latent Variables graphical models via ℓ_1 and trace penalized D-trace loss (LVD), which achieves parameter estimation and model selection consistency under certain identifiability conditions. We also present an efficient ADMM algorithm to obtain the penalized estimation of the sparse precision matrix. Using simulation studies, we validate the theoretical properties of our estimator and show its superior performance over other methods. The usefulness of the proposed method is also demonstrated through its application to a yeast genetical genomic data.

Keywords and phrases: ADMM, Gaussian graphical models, latent variable, low rank, model selection consistency, sparsity.

Received October 2016.

Contents

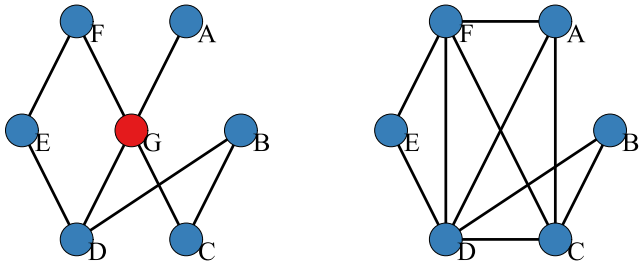
1. Introduction

Model selection for high-dimensional data has attracted much attention in recent years due to the need to analyze and interpret various types of high-dimensional data resulting from technological advances, including the analysis of gene expression data, spectroscopic imaging, fMRI data, and weather forecasting. A model selection problem that is of great importance is the estimation of a high-dimensional covariance matrix and its inverse, also known as the precision matrix. A number of papers have studied this problem in the context of Gaussian graphical models (Cox and Wermuth, 1996), which are represented by an undirected graph $G = (V, E)$, where V contains p nodes corresponding to a collection of joint Gaussian random variables and the edges $E = (e_{ij})_{1 \le i < j \le p}$ describe the conditional independence relationships among variables. Every pair of variables not included in E is conditionally independent given all the other variables and corresponds to a zero entry in the precision matrix (Lauritzen, 1996).

To deal with the singularity problem presented by high-dimensional data to infer precision matrix, regularization methods have been proposed to impose

sparsity constraint on the precision matrix and infer the sparse precision matrix, where an ℓ_1 penalty is often applied to induce sparsity (Yuan and Lin, 2007; Friedman et al., 2008; Rothman et al., 2008; Cai et al., 2011; Zhang and Zou, 2014). Some explicit rates of convergence of such estimators have been obtained (Rothman et al., 2008; Lam and Fan, 2009; Cai et al., 2011; Ravikumar et al., 2011). An alternative approach of estimating a sparse Gaussian model, first proposed by Meinshausen and Bühlmann (2006), is to perform an ℓ_1-regularized regression on every variable. It can asymptotically recover the true graph although does not provide estimates of the precision matrix. Other neighbourhood-based methods have been then introduced to estimate the precision matrix, see Yuan (2010), Sun and Zhang (2013) and Ren et al. (2015). For a comprehensive review on theoretical properties and optimalities of the estimation of structured covariance and precision matrices, see a recent review paper (Cai et al., 2016) and the references therein.

In many applications, however, the graph structure among variables can be decomposed into intrinsic local connections and external global effects. One such example is related to the analysis of genomic data, where Gaussian graphical models have been applied to infer the relationship between genes at the transcriptional level (Segal et al., 2005; Li and Gui, 2006). Although a direct application provides some insights into the gene regulatory network, it ignores the effects of covariates such as sex, age, race, genetic variants on gene expression, and those unmeasured confoundings that may blur the statistical inference (Cheung and Spielman, 2002). Some methods have been proposed to model the conditional Gaussian graphical models to adjust for the effect of covariates (Li et al., 2012; Cai et al., 2013; Chen et al., 2016). Nevertheless, these approaches are not applicable when we either do not observe all relevant variables, or do not incorporate them into the model. To further illustrate the impact of latent variables, Figure 1 shows a simple example of a graph with 7 nodes. Node G is a latent variable and Figure 1(a) is the full graph with all variables. Recall that an edge between two nodes indicates conditional dependence between these two



(a) Full graph (b) Reduced graph

FIG 1. *An illustrative example. The left panel is the full graph with all variables observed and the right panel is the reduced graph without node G.*

variables conditioning on all other variables *in the graph*. Without adjusting for the effect of latent variable G, the reduced graph demonstrated in Figure 1(b) is much denser, where many spurious conditional dependence are observed.

The problem of Gaussian graphical model with latent variables was first studied by Chandrasekaran et al. (2012). Under the joint Gaussian assumption, the authors proposed a regularized maximum likelihood approach and described conditions under which such graphical model was identifiable and the estimators were consistent. A number of alternative estimators were then proposed by other researchers in the accompanied discussion papers. Yuan (2012), Lauritzen and Meinshausen (2012) and Agakov et al. (2012) described an EM-based estimator, which could be treated as an alternative algorithm as well as the ADMM algorithm proposed by Ma et al. (2013). The error bounds for the estimated precision matrix in the Frobenius norm were analyzed by Meng et al. (2014) under more stringent conditions. Giraud and Tsybakov (2012) proposed two approaches based on the Dantzig selector (Candès and Tao, 2007) and the neighborhood selection approach (Meinshausen and Bühlmann, 2006). Ren and Zhou (2012) further analyzed a CLIME-like (Cai et al., 2011) estimator, and the results were improved using a regression method (Ren et al., 2015). The maximum likelihood approach with non-convex penalties are also studied by Xu et al. (2017). Finally, Städler and Bühlmann (2012) considered Gaussian graphical models with missing values, which could also be applied to this problem if we treat hidden variables as missing data.

In this paper, we consider a factor model where the unobserved and observed variables follow a linear relationship. Under such regime, the sample covariance matrix can be decomposed into the sum of a low-rank matrix and the inverse of a sparse matrix, which correspond to the effect of latent variables and the true conditional dependence between manifest variables, respectively. This problem is analogous to the "sparse plus low-rank" matrix decomposition problem (Chandrasekaran et al., 2011; Candès et al., 2011; Hsu et al., 2011; Agarwal et al., 2012) after reformulation. Fan et al. (2013) considered a similar factor model, but they focused on the estimation of a high dimensional covariance matrix and assumed a sparse error covariance matrix. Besides, Kalaitzis and Lawrence (2012) developed the residual component analysis, which also included the low-rank plus inverse sparse matrix decomposition as a special case. Nevertheless, they adopted a Bayesian approach and had to solve an intractable problem in an EM fashion. In this article, we propose an ℓ_1 and trace penalized approach to estimating the precision matrix so as to recover the true graph after adjusting for the effect of hidden variables. Following Chandrasekaran et al. (2012), we present conditions under which such "inverse sparse plus low-rank" decomposition is identifiable. Theoretical analysis reveals that our LVD estimator enjoys estimation and model selection consistency under such identifiability conditions. We also develop an efficient alternating direction method of multipliers (ADMM) for solving our optimization problem. Simulation studies and a real data analysis validate our theoretical results and show its supremacy over other competitors.

The rest of the paper is organized as follows. In Section 2, after basic notation is introduced, we present the formation of our method. Then in Section 3 we in-

troduce the associated computational algorithm. Theoretical analysis including the identifiability results and rates of convergence is established in Section 4. Numerical performance of our method is presented through simulation studies and a real data analysis in Sections 5 and 6, respectively. Section 7 concludes with a summary and discussion. The proofs of main results are given in the Appendix.

2. Methodology

2.1. Notation

For a vector $\boldsymbol{a} = (a_1, \cdots, a_p)^T \in \mathbb{R}^p$, define $\|\boldsymbol{a}\|_1 = \sum_{i=1}^p |a_i|$ and $\|\boldsymbol{a}\|_2 = \sqrt{\sum_{i=1}^p a_i^2}$. For a symmetric matrix $\boldsymbol{A} = (a_{ij}) \in \mathbb{R}^{p\times p}$, we write $\mathrm{tr}(\boldsymbol{A})$ for the trace of $\boldsymbol{A}$, and $\phi_{\min}(\boldsymbol{A})$ and $\phi_{\max}(\boldsymbol{A})$ for the minimum and maximum eigenvalues of $\boldsymbol{A}$. Define the element-wise ℓ_∞ norm $\|\boldsymbol{A}\|_\infty = \max_{1\le i,j\le p} |a_{ij}|$, the element-wise ℓ_1 norm $\|\boldsymbol{A}\|_1 = \sum_{i,j} |a_{ij}|$, the Frobenius norm $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$, the spectral norm $\|\boldsymbol{A}\|_2 = \phi_{\max}^{1/2}(\boldsymbol{A}^T\boldsymbol{A})$, and the nuclear norm $\|\boldsymbol{A}\|_* = \sqrt{\mathrm{tr}(\boldsymbol{A}^T\boldsymbol{A})}$. We use the the notation $\boldsymbol{A} \succ \boldsymbol{0}$ or $\boldsymbol{A} \succeq \boldsymbol{0}$ to denote that $\boldsymbol{A}$ is positive definite/semidefinite.

2.2. The LVD procedure

Suppose $\boldsymbol{Y} = (Y_1, \cdots, Y_p)^T \in \mathbb{R}^p$ is a manifest random vector (gene expression levels) and $\boldsymbol{X} \in \mathbb{R}^r$ is a hidden random vector (confounding factors), we consider the following model

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{B}$ is a $p \times r$ unknown coefficient matrix representing the effect of latent variables on $\boldsymbol{Y}$, and $\boldsymbol{\varepsilon}$ is a $p\times 1$ random Gaussian error vector independent of $\boldsymbol{X}$ with mean zero, covariance matrix $\boldsymbol{\Sigma}^*$ and precision matrix $\boldsymbol{S}^* = (\boldsymbol{\Sigma}^*)^{-1}$. Hence $\boldsymbol{S}^*$ is the parameter of main interest that characterizes the gene-gene interaction network conditioning on latent variables. Without loss of generality, we assume that $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centered at zero and $\mathsf{E}(\boldsymbol{X}\boldsymbol{X}^T) = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Taking variance on both sides of (2.1) and writing $\boldsymbol{\Sigma}_Y = \mathsf{Var}(\boldsymbol{Y})$, we obtain

$$\boldsymbol{\Sigma}_Y = \boldsymbol{B}\boldsymbol{B}^T + (\boldsymbol{S}^*)^{-1}. \tag{2.2}$$

If we only observe $\boldsymbol{Y}$, we only have access to $\boldsymbol{\Sigma}_Y$. The two terms that compose $\boldsymbol{\Sigma}_Y$ can be interpreted as follows: The matrix $\boldsymbol{B}\boldsymbol{B}^T$ serves as a summary of the effect of marginalization over latent variables $\boldsymbol{X}$, which is low-rank if the number of latent variables is small compared to that of observed variables. The precision matrix $\boldsymbol{S}^*$ has an interpretation of conditional independence between observed variables given latent variables. Specifically, $Y_i \perp Y_j | \{\boldsymbol{X}, Y_k : k \ne i, j\}$

if and only if $S^*_{ij} = 0$. If the interaction network can be depicted by a sparse graphical model, then $\boldsymbol{S}^*$ is sparse. Thus, the covariance matrix $\boldsymbol{\Sigma}_Y$ can be decomposed into the sum of a low-rank matrix and the inverse of a sparse matrix. We are interested in detecting the nonzero entries of $\boldsymbol{S}^*$ in order to construct a conditional independence graph for $\boldsymbol{Y}$ after the effect of hidden variables $\boldsymbol{X}$ on $\boldsymbol{Y}$ is removed.

Remark 2.1. We note that $\boldsymbol{X}$ and $\boldsymbol{\varepsilon}$ can be correlated and the "row-rank plus sparse inverse" relationship still holds, since in this case we have

$$\boldsymbol{\Sigma}_Y = \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{B}\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{\varepsilon}) + \mathsf{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{X})\boldsymbol{B}^T + (\boldsymbol{S}^*)^{-1}.$$

Now that $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$, the sum $\boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{B}\mathsf{Cov}(\boldsymbol{X}, \boldsymbol{\varepsilon}) + \mathsf{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{X})\boldsymbol{B}^T$ is still a low-rank matrix as long as r is small.

Suppose that we have n independent and identically distributed observations $\boldsymbol{Y}_{(1)}, \cdots, \boldsymbol{Y}_{(n)}$ from (2.1). Let $\boldsymbol{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \boldsymbol{Y}_{(i)}\boldsymbol{Y}_{(i)}^T$ be the sample covariance matrix. Unlike the joint Gaussian distribution on $(\boldsymbol{X}, \boldsymbol{Y})$ assumed by Chandrasekaran et al. (2012), we do not impose any specific distribution on $\boldsymbol{X}$, thus a maximum likelihood approach is inapplicable. Note that $(\boldsymbol{S}^*)^{-1}$ in (2.2) is not a convex function of $\boldsymbol{S}^*$, we take inverse on both sides of (2.2) and use the Sherman-Morrison-Woodbury formula to obtain

$$(\boldsymbol{\Sigma}_Y)^{-1} = \boldsymbol{S}^* - \boldsymbol{S}^*\boldsymbol{B}(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{S}^*\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{S}^* = \boldsymbol{S}^* - \boldsymbol{L}^*, \tag{2.3}$$

where we set $\boldsymbol{L}^* = \boldsymbol{S}^*\boldsymbol{B}(\boldsymbol{I} + \boldsymbol{B}^T\boldsymbol{S}^*\boldsymbol{B})^{-1}\boldsymbol{B}^T\boldsymbol{S}^*$. Inspired by Zhang and Zou (2014), we consider the following quadratic loss function based on (2.2):

$$\mathrm{LOSS}(\boldsymbol{S}, \boldsymbol{L}; \boldsymbol{\Sigma}_n) = \frac{1}{2}\mathrm{tr}((\boldsymbol{S} - \boldsymbol{L})\boldsymbol{\Sigma}_n(\boldsymbol{S} - \boldsymbol{L})) - \mathrm{tr}(\boldsymbol{S} - \boldsymbol{L}). \tag{2.4}$$

This is called the D-trace loss in Zhang and Zou (2014), since it is expressed as the difference of two trace operators. To see the intuition of this loss, if we multiply $\boldsymbol{\Sigma}_Y^{1/2}$ on both sides of (2.3), where $\boldsymbol{\Sigma}_Y^{1/2}$ is the square root of $\boldsymbol{\Sigma}_Y$ such that $\boldsymbol{\Sigma}_Y^{1/2}\boldsymbol{\Sigma}_Y^{1/2} = \boldsymbol{\Sigma}_Y$, we have

$$(\boldsymbol{\Sigma}_Y)^{-1/2} = \boldsymbol{\Sigma}_Y^{1/2}(\boldsymbol{S}^* - \boldsymbol{L}^*).$$

Now considering the Frobenious norm of the difference of these two terms lead to the D-trace loss in (2.4) if we discard terms independent of $\boldsymbol{S}$ or $\boldsymbol{L}$ and replace $\boldsymbol{\Sigma}_Y$ with its sample version $\boldsymbol{\Sigma}_n$.

Remark 2.2. As noted in Zhang and Zou (2014), the D-trace norm loss bears some resemblance to the one proposed in Liu and Luo (2015). Indeed, they considerer the same loss for $\boldsymbol{S} - \boldsymbol{L}$ as in (2.4), except in a column-by-column fashion. Specifically, let $\boldsymbol{s}_i$ and $\boldsymbol{l}_i$ be the ith column of $\boldsymbol{S}$ and $\boldsymbol{L}$ respectively, then the loss function for $\boldsymbol{s}_i - \boldsymbol{l}_i$ in Liu and Luo (2015) takes the form

$$\frac{1}{2}(\boldsymbol{s}_i - \boldsymbol{l}_i)^T\boldsymbol{\Sigma}_n(\boldsymbol{s}_i - \boldsymbol{l}_i) - \boldsymbol{e}_i^T(\boldsymbol{s}_i - \boldsymbol{l}_i),$$

where $\boldsymbol{e}_i$ is the ith standard orthonormal basis of $\mathbb{R}^p$. Nevertheless as pointed out in Zhang and Zou (2014) the positive-definite property of a matrix that should be taken into account in graphical models, can only be properly dealt with using a loss viewing $\boldsymbol{S}-\boldsymbol{L}$ as a whole rather than columnwisely. This is even more crucial in our setting, as the information in the low-rank matrix $\boldsymbol{L}$ will be lost if we consider $\boldsymbol{l}_i, i=1,\ldots,p$, separately.

For consideration of a low-rank $\boldsymbol{L}$ and a sparse $\boldsymbol{S}$, we propose to estimate $\boldsymbol{S}$ and $\boldsymbol{L}$ by solving the following ℓ_1 and trace regularized D-trace loss

$$\begin{array}{ll}\min_{\boldsymbol{R},\boldsymbol{S},\boldsymbol{L}} & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R})-\mathrm{tr}(\boldsymbol{R})+\lambda_n(\gamma\|\boldsymbol{S}\|_1+\mathrm{tr}(\boldsymbol{L}))\\ \text{subject to} & \boldsymbol{R}=\boldsymbol{S}-\boldsymbol{L}, \boldsymbol{R}\succ\boldsymbol{0}, \boldsymbol{L}\succeq\boldsymbol{0},\end{array} \tag{2.5}$$

where λ_n is a tuning parameter that controls the strength of regularization and γ is a tuning parameter that provides a trade-off between the ℓ_1 and trace penalties. These penalties are used to encourage sparsity and low-rankness as the ℓ_1 and nuclear norm (trace when $\boldsymbol{L}$ is positive semidefinite) are convex relaxations of the sparsity level and rank, respectively; see for example in Donoho (2006); Candès and Tao (2010). Finally, the positive (semi)definite constraints are imposed on $\boldsymbol{S}-\boldsymbol{L}$ and $\boldsymbol{L}$ according to model (2.2) and decomposition (2.3).

2.3. Relations to other models

2.3.1. Penalized likelihood approach

Chandrasekaran et al. (2012) assumed that $(\boldsymbol{Y},\boldsymbol{X})$ are joint Gaussian with a covariance matrix $\boldsymbol{\Sigma}_{(Y,X)}=\left[\begin{smallmatrix}\boldsymbol{\Sigma}_Y & \boldsymbol{\Sigma}_{YX}\\ \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_X\end{smallmatrix}\right]$ and precision matrix $\boldsymbol{\Omega}_{(Y,X)}=\boldsymbol{\Sigma}_{(YX)}^{-1}$. Then by the Schur complement (Horn and Johnson, 2012) with respect to block $\boldsymbol{\Omega}_{\boldsymbol{X}}$, we have

$$\boldsymbol{\Sigma}_{\boldsymbol{Y}}^{-1}=\boldsymbol{\Omega}_Y-\boldsymbol{\Omega}_{YX}\boldsymbol{\Omega}_X^{-1}\boldsymbol{\Omega}_{XY}. \tag{2.6}$$

They proposed a regularized maximum likelihood approach to estimate the sparse structure $\boldsymbol{\Omega}_Y$ and the low-rank term $\boldsymbol{\Omega}_{YX}\boldsymbol{\Omega}_X^{-1}\boldsymbol{\Omega}_{XY}$. We see that criterion (2.6) is the same as (2.3) and is also used by other authors, see Ren and Zhou (2012); Giraud and Tsybakov (2012); Xu et al. (2017) among others. Since the joint Gaussian distribution implies a linear conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{X}$, our assumption is weaker and arrives at the same matrix decomposition criterion. Nevertheless, we emphasize that allowing for $\boldsymbol{X}$ to take distributions other than Gaussian, especially discrete distributions, is of particular interest in practice. The global dependence between variables is often due to factors taking discrete values such as batch effects, hence the joint Gaussian assumption is sometimes unrealistic.

As to different forms of loss function, we make two remarks here. First, one might think that the log likelihood is the optimal loss in terms of estimating the precision matrix under the joint Gaussian assumption due to its likelihood explanation. Under our regime, however, even though $\boldsymbol{\varepsilon}$ in (2.1) follows a Gaussian distribution to ensure the conditional independence interpretation, we do

not make any distribution assumption for $\boldsymbol{X}$. Hence our loss function, which is analogous to the least squares loss, is more convenient under such model. Second, similar to the discussion presented in Zhang and Zou (2014), both loss functions have an optimum occurs at the true value if we replace $\hat{\boldsymbol{\Sigma}}_n$ with its true value $\boldsymbol{\Sigma}_Y$. The preference for these two loss functions may alter from case to case and we will explore more about the performances of these two approaches in numerical studies.

2.3.2. Factor analysis

The factor analysis (FA) model can be expressed in the same form as in (2.1), but with different assumptions on the random error $\boldsymbol{\varepsilon}$. We state here explicitly as

$$\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{B}$ is a $p \times r$ unknown factor loading matrix, $\boldsymbol{X}$ is the $r \times 1$ hidden factors satisfying $\mathsf{E}\boldsymbol{X} = 0$ and $\mathsf{E}\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}$. Last, $\boldsymbol{\varepsilon}$ is a $p \times 1$ Gaussian random error independent of X, with zero mean and covariance matrix $\boldsymbol{\Sigma}$ satisfying $\boldsymbol{\Sigma} = \operatorname{diag}\{\sigma_1, \ldots, \sigma_p\}$.

The main difference between our model and the classical factor analysis model is the goal of inference. Our interest lies in the detection of conditional dependence between observed variables after adjusting for the effect of latent factors, and we do not pay much attention to the global dependence induced by hidden factors. On the other hand, the underlying assumption in FA is that the variability among manifest variables can be captured by a small number of unobserved factors, and parameters of interest include $\boldsymbol{B}, \boldsymbol{X}$ and $\{\sigma_i, i = 1, \ldots, p\}$. In general, the decomposition of correlations among observed variables are the same for these two models, where the hidden variables describe the global dependence while the random error depicts a more subtle local dependence. With different goals in mind, there are substantial differences both on the assumptions made on the linear model and the strategies used to make statistical inference.

Fan et al. (2013) considered the same factor model as in (2.1), but assumed a sparse error covariance $\boldsymbol{\Sigma}^*$ instead of a sparse error precision matrix. Based on a fast diverging eigenvalues assumption, they introduced the principal orthogonal complement thresholding estimator for estimating the covariance and derived its convergence rate. Although the "low-rank plus sparse" decomposition bears some resemblance to our model, interpretations of these two models differ significantly. Further, their main interest lies in estimating the high-dimensional covariance matrix of observed variables, i.e., $\boldsymbol{\Sigma}_Y$, which is also different from our method.

3. Algorithm and tuning parameter selection

3.1. Numerical algorithm

In this section we develop an efficient algorithm for solving the optimization problem (2.5). Since $\boldsymbol{R} \succ \boldsymbol{0}$ is not a closed convex cone, we instead consider the

following problem:

$$\begin{array}{ll}\min_{\boldsymbol{R},\boldsymbol{S},\boldsymbol{L}} & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \lambda_n(\gamma\|\boldsymbol{S}\|_1 + \mathrm{tr}(\boldsymbol{L})),\\ \text{subject to} & \boldsymbol{R} = \boldsymbol{S} - \boldsymbol{L}, \boldsymbol{R} = \boldsymbol{R}^T, \boldsymbol{L} \succeq \boldsymbol{0}.\end{array} \tag{3.1}$$

Note that we drop the positive-definite constraint for $\boldsymbol{S}-\boldsymbol{L}$. Although $\boldsymbol{S}-\boldsymbol{L}$ is an estimate of $\boldsymbol{\Sigma}_Y$, which should be positive-definite according to its interpretation, solving (3.1) is more convenient than the one that imposes positive-definiteness on $\boldsymbol{S}-\boldsymbol{L}$. Furthermore, we find in our numerical experiments that in all cases the solution to (3.1) does satisfy the constraint.

Remark 3.1. For the sake of completeness, we also develop methods for solving the following problem:

$$\begin{array}{ll}\min_{\boldsymbol{R},\boldsymbol{S},\boldsymbol{L}} & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \lambda_n(\gamma\|\boldsymbol{S}\|_1 + \mathrm{tr}(\boldsymbol{L})),\\ \text{subject to} & \boldsymbol{R} = \boldsymbol{S} - \boldsymbol{L}, \boldsymbol{R} \succeq \varepsilon\boldsymbol{I}, \boldsymbol{L} \succeq \boldsymbol{0},\end{array} \tag{3.2}$$

where ε is a pre-defined threshold. The algorithm for solving this problem is deferred in Appendix C.

We next introduce the ADMM for solving (3.1). For notational simplicity, let

$$\begin{aligned} f(\boldsymbol{R}) &= \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}),\\ g(\boldsymbol{S}) &= \lambda_n\gamma\|\boldsymbol{S}\|_1,\\ h(\boldsymbol{L}) &= \lambda_n\mathrm{tr}(\boldsymbol{L}) + \mathcal{I}(\boldsymbol{L} \succeq \boldsymbol{0}),\end{aligned}$$

where the indicator function $\mathcal{I}(\boldsymbol{L} \succeq \boldsymbol{0})$ is defined as

$$\mathcal{I}(\boldsymbol{L} \succeq \boldsymbol{0}) = \begin{cases} 0, & \text{if } \boldsymbol{L} \succeq \boldsymbol{0},\\ +\infty, & \text{otherwise.}\end{cases}$$

We now rewrite the problem (3.1) into one with two blocks: let $\boldsymbol{Z} = (\boldsymbol{R}, \boldsymbol{S}, \boldsymbol{L})$, $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{L}})$, we aim to solve

$$\begin{array}{ll}\min_{\boldsymbol{Z},\tilde{\boldsymbol{Z}}} & f(\boldsymbol{R}) + g(\boldsymbol{S}) + h(\boldsymbol{L}) + \phi(\tilde{\boldsymbol{Z}}),\\ \text{subject to} & \boldsymbol{Z} = \tilde{\boldsymbol{Z}},\end{array} \tag{3.3}$$

where $\phi(\tilde{\boldsymbol{Z}}) = \mathcal{I}(\tilde{\boldsymbol{R}} - \tilde{\boldsymbol{S}} + \tilde{\boldsymbol{L}} = \boldsymbol{0})$. Its Lagrange form is then given by

$$f(\boldsymbol{R}) + g(\boldsymbol{S}) + h(\boldsymbol{L}) + \phi(\tilde{\boldsymbol{Z}}) + \left\langle \boldsymbol{\Lambda}, \boldsymbol{Z} - \tilde{\boldsymbol{Z}} \right\rangle + \frac{\rho}{2}\left\|\boldsymbol{Z} - \tilde{\boldsymbol{Z}}\right\|_F^2,$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ is the multiplier of the linear constraint $\tilde{\boldsymbol{R}} = \boldsymbol{R}$, $\rho > 0$ is the penalty parameter for the violation of the constraint and $\langle\cdot,\cdot\rangle$ denotes the standard trace inner product. Hence the ADMM algorithm can be written as

$$\boldsymbol{Z}^{k+1} = \arg\min_{\boldsymbol{Z}} f(\boldsymbol{R}) + g(\boldsymbol{S}) + h(\boldsymbol{L}) + \left\langle \boldsymbol{\Lambda}^k, \boldsymbol{Z} - \tilde{\boldsymbol{Z}}^k \right\rangle + \frac{\rho}{2}\left\|\boldsymbol{Z} - \tilde{\boldsymbol{Z}}^k\right\|_F^2, \tag{3.4}$$

$$\tilde{\boldsymbol{Z}}^{k+1} = \arg\min_{\tilde{\boldsymbol{Z}}} \left\langle \boldsymbol{\Lambda}^k, \boldsymbol{Z}^{k+1} - \tilde{\boldsymbol{Z}} \right\rangle + \frac{\rho}{2} \left\| \boldsymbol{Z}^{k+1} - \tilde{\boldsymbol{Z}} \right\|_F^2 + \phi(\tilde{\boldsymbol{Z}}), \tag{3.5}$$

$$\boldsymbol{\Lambda}^{k+1} = \boldsymbol{\Lambda}^k + \rho(\boldsymbol{Z}^{k+1} - \tilde{\boldsymbol{Z}}^{k+1}). \tag{3.6}$$

Step (3.6) is trivial, we now elaborate on strategies for solving subproblems (3.4) and (3.5). As for (3.4), let $\boldsymbol{W}^k = \tilde{\boldsymbol{Z}}^k - \frac{1}{\rho}\boldsymbol{\Lambda}^k$ and partition it into three blocks $\boldsymbol{W}^k = (\boldsymbol{W}_R^k, \boldsymbol{W}_L^k, \boldsymbol{W}_S^k)$, then $(\boldsymbol{S}^{k+1}, \boldsymbol{L}^{k+1}, \boldsymbol{R}^{k+1})$ can be solved separately as follows. To update $\boldsymbol{S}^{k+1}$, we have

$$\boldsymbol{S}^{k+1} = \arg\min_{\boldsymbol{S}} \lambda_n \gamma \|\boldsymbol{S}\|_1 + \frac{\rho}{2} \left\| \boldsymbol{S} - \boldsymbol{W}_S^k \right\|_F^2,$$

which implies

$$\boldsymbol{S}^{k+1} = \text{Shrink}(\boldsymbol{W}_S^k, \lambda_n \gamma / \rho), \tag{3.7}$$

where the shrink operator is defined as

$$\text{Shrink}(\boldsymbol{S}, \tau) = \begin{cases} S_{ij} - \tau, & \text{if } S_{ij} > \tau, \\ S_{ij} + \tau, & \text{if } S_{ij} < -\tau, \\ 0, & \text{if } |S_{ij}| \le \tau. \end{cases}$$

To update $\boldsymbol{L}^{k+1}$, we have

$$\boldsymbol{L}^{k+1} = \arg\min_{\boldsymbol{L}} \lambda_n \text{tr}(\boldsymbol{L}) + \mathcal{I}(\boldsymbol{L} \succeq \boldsymbol{0}) + \frac{\rho}{2} \left\| \boldsymbol{L} - \boldsymbol{W}_L^k \right\|_F^2.$$

Hence

$$\boldsymbol{L}^{k+1} = \text{Proj}(\boldsymbol{W}_L^k - \lambda_n \boldsymbol{I}/\rho), \tag{3.8}$$

where the projection operator is defined in the following way: for a real symmetric matrix $\boldsymbol{L}$, let $\boldsymbol{L} = \boldsymbol{U}_L \text{diag}(\boldsymbol{\sigma}_L) \boldsymbol{U}_L^T$ be its eigenvalue decomposition, define $\text{Proj}(\boldsymbol{L})$ as

$$\text{Proj}(\boldsymbol{L}) = \boldsymbol{U}_L \text{diag}(\boldsymbol{\pi}_L) \boldsymbol{U}_L^T,$$

where $\boldsymbol{\pi}$ is given by

$$\pi_{L,i} = \max\{\sigma_{L,i}, 0\}, \quad i = 1, \ldots, p.$$

Last, to update $\boldsymbol{R}^{k+1}$, we write

$$\boldsymbol{R}^{k+1} = \arg\min_{\boldsymbol{R} = \boldsymbol{R}^T} \frac{1}{2} \text{tr}(\boldsymbol{R} \boldsymbol{\Sigma}_n \boldsymbol{R}) - \text{tr}(\boldsymbol{R}) + \frac{\rho}{2} \left\| \boldsymbol{R} - \boldsymbol{W}_R^k \right\|_F^2. \tag{3.9}$$

The first-order optimality condition is then given by

$$\frac{1}{2}(\boldsymbol{R}^{k+1} \boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n \boldsymbol{R}^{k+1}) - \boldsymbol{I} + \rho \left(\boldsymbol{R}^{k+1} - \boldsymbol{W}_R^k \right) = \boldsymbol{0}.$$

Following Zhang and Zou (2014), the above problem has an explicit form solution. Let $\boldsymbol{\Sigma}_n = \boldsymbol{U}_\Sigma \text{diag}(\boldsymbol{\sigma}_\Sigma) \boldsymbol{U}_\Sigma^T$ be the eigenvalue decomposition of $\boldsymbol{\Sigma}_n$, where $\boldsymbol{U}_\Sigma \in \mathbb{R}^{p \times p}$, then we have

$$\boldsymbol{R}^{k+1} = \boldsymbol{U}_\Sigma \left\{ \boldsymbol{U}_\Sigma^T (\rho \boldsymbol{W}_R^k + \boldsymbol{I}) \boldsymbol{U}_\Sigma \circ \boldsymbol{\Pi}_\Sigma \right\} \boldsymbol{U}_\Sigma^T, \tag{3.10}$$

where $\circ$ denotes the Hadamard product of matrices and $\mathbf{\Pi}_\Sigma$ is defined as

$$\Pi_{\Sigma,ij} = \frac{2}{\sigma_{\Sigma,i} + \sigma_{\Sigma,j} + 2\rho}, \quad i,j = 1,\ldots,p.$$

The proofs of (3.7) and (3.8) are obvious, and we provide a proof of (3.10) in Appendix B for completeness.

We next derive solution to the sub-problem (3.5). Note that solving (3.5) is equivalent to

$$\min_{\tilde{\boldsymbol{Z}}} \frac{1}{2} \left\| \tilde{\boldsymbol{Z}} - \boldsymbol{Z}^{k+1} - \frac{1}{\rho}\mathbf{\Lambda}^k \right\|_F^2, \quad \text{s.t.,} \quad \tilde{\boldsymbol{R}} - \tilde{\boldsymbol{S}} + \tilde{\boldsymbol{L}} = \mathbf{0}.$$

Let $\boldsymbol{T}^k = \boldsymbol{Z}^{k+1} + \frac{1}{\rho}\mathbf{\Lambda}^k$ and partition $\boldsymbol{T}^k$ similarly into three blocks $\boldsymbol{T}^k = (\boldsymbol{T}_R^k, \boldsymbol{T}_S^k, \boldsymbol{T}_L^k)$ in the same form as $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{L}})$. Then the first-order optimality condition is given by

$$\left(\tilde{\boldsymbol{R}}, \tilde{\boldsymbol{S}}, \tilde{\boldsymbol{L}}\right) - \left(\boldsymbol{T}_R^k, \boldsymbol{T}_S^k, \boldsymbol{T}_L^k\right) + (\mathbf{\Gamma}, -\mathbf{\Gamma}, \mathbf{\Gamma}) = \mathbf{0},$$

where $\mathbf{\Gamma}$ is the Lagrange multiplier associated with $\tilde{\boldsymbol{Z}}$. Therefore, we have

$$\tilde{\boldsymbol{R}} = \boldsymbol{T}_R^k - \mathbf{\Gamma}, \quad \tilde{\boldsymbol{S}} = \boldsymbol{T}_S^k + \mathbf{\Gamma}, \quad \tilde{\boldsymbol{L}} = \boldsymbol{T}_L^k - \mathbf{\Gamma}. \tag{3.11}$$

Combining the above display with the equality constraint $\tilde{\boldsymbol{R}} - \tilde{\boldsymbol{S}} + \tilde{\boldsymbol{L}} = \mathbf{0}$ yields

$$\mathbf{\Gamma} = (\boldsymbol{T}_R^k + \boldsymbol{T}_L^k - \boldsymbol{T}_S^k)/3,$$

which completes the update of $\tilde{\boldsymbol{Z}}$. We summarize the proposed algorithm as follows.

Algorithm 1 ADMM for solving (3.1)

Require: $\mathbf{\Sigma}_n$ and parameters λ_n, γ.
Ensure: $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{L}}$.
1. Initialize $\boldsymbol{Z}^0, \tilde{\boldsymbol{Z}}^0$ and $\mathbf{\Lambda}^0$ with some possible values.
While not converge do
2. Update $\boldsymbol{Z}^{k+1}$ by solving equations (3.7), (3.8) and (3.10).
3. Update $\tilde{\boldsymbol{Z}}^{k+1}$ as in (3.11).
4. Update $\mathbf{\Lambda}^{k+1}$ as in (3.6).
End while.

The stopping criterion of our algorithm is when both the primal error $(\boldsymbol{Z}^{k+1} - \tilde{\boldsymbol{Z}}^{k+1})$ and the dual error $(\tilde{\boldsymbol{Z}}^{k+1} - \tilde{\boldsymbol{Z}}^k)$ is small enough with a tolerance level of 10^{-4}. In view of (3.3), the proposed ADMM is a special case of the consensus problem (Boyd et al., 2011, Chap 7) with two blocks. The composition of quadratic form and affine mapping $(\boldsymbol{L} - \boldsymbol{S})$ imply that the D-trace loss is convex as well as the two penalty terms and the positive semidefinite constraint, which guarantees the global convergence of our algorithm. From the view of computational complexity, as it suffices to do the SVD of $\mathbf{\Sigma}_n$ in (3.9) for once, the most

time-consuming step is the SVD for solving (3.8) in each iteration, whose complexity is $O(p^3)$. Such decomposition can be further accelerated by only solving leading singular values and vectors when λ_n is not too small. Besides, ADMM would usually converge to a moderate accuracy in only tens of iterations (Boyd et al., 2011, Chap 3.2.2), which is also observed in our numerical experiments and proves to be sufficient. Using a laptop with Intel Core I5-5200 2.20GHz and 8 GB of RAM, our codes written in MATLAB 9.1.0 is able to converge in seconds for a few hundreds of nodes and in tens of seconds when $p = 1000$.

3.2. Choices of tuning parameters

We have two tuning parameters to be tuned: λ_n that controls the strength of regularization and γ that provides a trade-off between regularizations on $\boldsymbol{S}$ and $\boldsymbol{L}$. We will show later in Theorem 4.1 that the consistency results hold for a range of values of γ, which means our LVD procedure is robust with respect to γ. This phenomenon is also observed by Chandrasekaran et al. (2012) and in our numerical studies.

As for λ_n, we tune it via K-fold cross validation for a fine gird of λ. Specifically, we denote the cross validation error for λ by

$$\begin{aligned}\mathrm{CV}_\lambda(\boldsymbol{Y}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\mathrm{tr}\left(\left(\hat{\boldsymbol{S}}_\lambda^{(-k)} - \hat{\boldsymbol{L}}_\lambda^{(-k)}\right)\boldsymbol{\Sigma}_n^{(k)}\left(\hat{\boldsymbol{S}}_\lambda^{(-k)} - \hat{\boldsymbol{L}}_\lambda^{(-k)}\right)\right)\\ - \mathrm{tr}\left(\hat{\boldsymbol{S}}_\lambda^{(-k)} - \hat{\boldsymbol{L}}_\lambda^{(-k)}\right)\end{aligned} \tag{3.12}$$

where $\boldsymbol{\Sigma}_n^{(k)}$ is the sample covariance matrix for the kth part sample, and $\hat{\boldsymbol{S}}_\lambda^{(-k)}$ and $\hat{\boldsymbol{L}}_\lambda^{(-k)}$ are the estimates under λ with the kth part sample removed.

Another useful practice is to apply a warm-start method on decreasing sequences of values for λ. Exactly, if we want to compute a sequence of solutions of (3.1) for $\lambda_1 > \lambda_2 > \cdots > \lambda_L$, the initial values of the solution at λ_i can be chosen as the solution at λ_{i-1}. Note that when λ_1 is large, $\hat{\boldsymbol{S}}$ would be very sparse, thus the algorithm is not very sensitive to the initial values for λ_1. Subsequently, for other values of λ_i, the initial values would be close to the resulting estimates, hence the number of iterations is greatly reduced. The proposed algorithm would usually stop after tens of iterations and the performance of this warm start strategy proves to be satisfactory in practice.

4. Theoretical analysis

4.1. Identifiability

Before we investigate theoretical properties of the proposed LVD estimator, it is important to first cope with the identifiability issue. Recall that the whole procedure stems from the decomposition (2.3), that is, the inverse of the population covariance can be decomposed into the sum of a sparse matrix and a

low-rank matrix. Obviously this decomposition problem is ill-posed if we only have access to the sample covariance matrix without any further assumptions. To deal with this problem, we begin by introducing some notion that plays an important role in the identifiability issue.

Indeed there are cases where such a decomposition is unidentifiable. For example, if the sparse matrix $\boldsymbol{S}^*$ itself has a low rank (after subtracting its diagonal entries), then we can subtract these entries from $\boldsymbol{S}^*$ and add them to $\boldsymbol{L}^*$. Figure 2 illustrates an example of the "sparse plus low-rank" decomposition that seems to be identifiable, as entries of the low-rank matrix $\boldsymbol{L}^*$ ($\operatorname{rank}(\boldsymbol{L}) = 1$) spread out over all entries of the matrix.

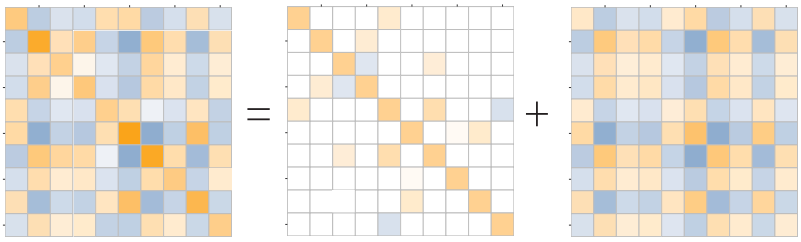


FIG 2. *An illustrative example of the decomposition* $(\boldsymbol{\Sigma}_Y)^{-1} = \boldsymbol{S}^* + (-\boldsymbol{L}^*)$.

Such identifiability problem has been studied in Candès et al. (2011); Chandrasekaran et al. (2011, 2012); Hsu et al. (2011). We follow the framework developed by Chandrasekaran et al. (2012) to deal with this problem. Since the set of sparse matrices and the set of low-rank matrices can be viewed as algebraic varieties (sets of solutions to systems of polynomial equations), we denote the tangent space at the sparse matrix $\boldsymbol{S}^*$ by

$$\Omega(\boldsymbol{S}^*) = \left\{\boldsymbol{N} \in \mathbb{R}^{p\times p} : \operatorname{supp}(\boldsymbol{N}) \subset \operatorname{supp}(\boldsymbol{S}^*)\right\}.$$

Similarly, the tangent space at the low-rank matrix $\boldsymbol{L}^*$ is defined as follows: suppose $\boldsymbol{L}^*$ is a rank-r square matrix with its singular value decomposition given by $\boldsymbol{L}^* = \boldsymbol{U}_L \operatorname{diag}(\boldsymbol{\sigma}_L)\boldsymbol{V}_L^T$, where $\boldsymbol{U}_L, \boldsymbol{V}_L \in \mathbb{R}^{p\times r}$ and $\boldsymbol{\sigma}_L \in \mathbb{R}^r$, then the tangent space at $\boldsymbol{L}^*$ is given by

$$T(\boldsymbol{L}^*) = \left\{\boldsymbol{U}_L\boldsymbol{Y}_1^T + \boldsymbol{Y}_2\boldsymbol{V}_L^T : \boldsymbol{Y}_1, \boldsymbol{Y}_2 \in \mathbb{R}^{p\times r}\right\}.$$

Denote by $\Omega^* = \Omega(\boldsymbol{S}^*)$ and $T^* = T(\boldsymbol{L}^*)$. Since both Ω^* and T^* are subspaces in $\mathbb{R}^{p\times p}$, a sufficient and necessary condition for $\boldsymbol{S}^*$ and $\boldsymbol{L}^*$ to be identifiable with prior knowledge of $\boldsymbol{\Sigma}_Y, \Omega^*$ and T^* is that these subspaces intersect transversally:

$$\Omega^* \cap T^* = \{\boldsymbol{0}\}.$$

To quantify transversality between these two tangent spaces, Chandrasekaran et al. (2012) introduced the following quantity with respect to the tangent space Ω^* and T^*:

$$\xi(T^*) = \max_{\boldsymbol{N}\in T^*, \|\boldsymbol{N}\|_2=1} \|\boldsymbol{N}\|_\infty,$$

$$\mu(\Omega^*) = \max_{\boldsymbol{N}\in\Omega^*, \|\boldsymbol{N}\|_\infty=1} \|\boldsymbol{N}\|_2.$$

According to definitions, a small $\xi(T^*)$ means that any element in the tangent space T^* cannot have its support concentrated in a few locations; a small $\mu(\Omega^*)$ implies that the spectrum of any element in Ω^* is not too concentrated. Hence these two quantities serve as tools for measuring "incoherence" of the row/column spaces of the low-rank matrix and the spectrum of the sparse matrix respectively. As discussed in Chandrasekaran et al. (2011), we always have $\xi(T^*) \leq 1$. Further if $\text{rank}(\boldsymbol{L}^*) = r$ and it has almost maximally incoherent row/column spaces, $\xi(T^*)$ can be as small as $\sim \sqrt{\frac{r}{p}}$. Similarly, $\mu(\Omega^*)$ is upper-bounded by the maximal degree of $\boldsymbol{S}^*$, which implies that if $\boldsymbol{S}^*$ has at most $\deg(\boldsymbol{S}^*)$ nonzero entries per row/column, then we must have $\mu(\Omega^*) \leq \deg(\boldsymbol{S}^*)$.

We next present the irrepresentability condition for establishing the model selection consistency of our LVD estimator. As our loss function is itself a quadratic loss, a key quantity is its derivative with respect to $\boldsymbol{S}$ and $\boldsymbol{L}$ (c.f., the Gram matrix and its role in the analysis of Lasso). Specifically, we define the following linear operator

$$h_\Sigma(\boldsymbol{R}) = \frac{1}{2}(\boldsymbol{\Sigma R} + \boldsymbol{R\Sigma}), \tag{4.1}$$

where both $\boldsymbol{\Sigma}$ and $\boldsymbol{R}$ are $p \times p$ matrices. For any linear subspace T of matrices, denote by $\mathcal{P}_T$ the projection onto T. To measure transversality between the tangent spaces Ω^* and T^* in terms of the operator $h_{\Sigma_Y}(\cdot)$, we need to analyze the following quantities. First, the minimum gain of $h_{\Sigma_Y}(\cdot)$ restricted to Ω^* and the maximum effect of elements in Ω^* on the orthogonal complement $(\Omega^*)^\perp$ are given by

$$\begin{aligned}
\alpha_\Omega &= \min_{\boldsymbol{M}\in\Omega^*, \|\boldsymbol{M}\|_\infty=1} \|\mathcal{P}_{\Omega^*} h_{\Sigma_Y}(\boldsymbol{M})\|_\infty, \\
\delta_\Omega &= \max_{\boldsymbol{M}\in\Omega^*, \|\boldsymbol{M}\|_\infty=1} \|\mathcal{P}_{(\Omega^*)^\perp} h_{\Sigma_Y}(\boldsymbol{M})\|_\infty.
\end{aligned}$$

Similarly, the minimum gain of $h_{\Sigma_Y}(\cdot)$ restricted to T^* and the maximum effect of elements in T^* on the orthogonal direction $(T^*)^\perp$ are defined as

$$\begin{aligned}
\alpha_T &= \min_{\boldsymbol{M}\in T^*, \|\boldsymbol{M}\|_2=1} \|\mathcal{P}_{T^*} h_{\Sigma_Y}(\boldsymbol{M})\|_2, \\
\delta_T &= \max_{\boldsymbol{M}\in T^*, \|\boldsymbol{M}\|_2=1} \|\mathcal{P}_{(T^*)^\perp} h_{\Sigma_Y}(\boldsymbol{M})\|_2.
\end{aligned}$$

Finally, we also need to control the behavior of $h_{\Sigma_Y}(\cdot)$ restricted to Ω^* in the spectral norm and its behavior restricted to T^* in the ℓ_∞ norm, that is

$$\begin{aligned}
\beta_T &= \max_{\boldsymbol{M}\in T^*, \|\boldsymbol{M}\|_\infty=1} \|h_{\Sigma_Y}(\boldsymbol{M})\|_\infty, \\
\beta_\Omega &= \max_{\boldsymbol{M}\in\Omega^*, \|\boldsymbol{M}\|_2=1} \|h_{\Sigma_Y}(\boldsymbol{M})\|_2.
\end{aligned}$$

Note that the two sets of quantities $(\alpha_\Omega, \delta_\Omega)$ and (α_T, δ_T) measures the effect of $h_{\Sigma_Y}(\cdot)$ restricted to spaces Ω^* and T^* in the natural norm, respectively. For notational simplicity, set

$$\alpha = \min\{\alpha_\Omega, \alpha_T\}, \quad \beta = \max\{\beta_T, \beta_\Omega\}, \quad \delta = \max\{\delta_\Omega, \delta_T\}.$$

Then the following irrepresentability condition is assumed in our theoretical analysis: there exists a $\nu \in (0, \frac{1}{2})$ such that

$$\frac{\delta}{\alpha} < 1 - 2\nu. \tag{4.2}$$

We note that this assumption can be viewed as a generalization of the irrepresentability condition assumed in high-dimension regression literature (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) and graphical model literature (Ravikumar et al., 2011; Zhang and Zou, 2014). It also bears some resemblance to the one imposed in Chandrasekaran et al. (2012), although the operators lie in the core of the analysis are different owing to different loss functions.

Finally, the superposition of penalizations can be seen as a norm,

$$f_\gamma(\boldsymbol{S}, \boldsymbol{L}) = \gamma\|\boldsymbol{S}\|_1 + \|\boldsymbol{L}\|_*,$$

which implies that its dual norm is given by

$$g_\gamma(\boldsymbol{S}, \boldsymbol{L}) = \max\left\{\frac{\|\boldsymbol{S}\|_\infty}{\gamma}, \|\boldsymbol{L}\|_2\right\}.$$

This dual norm $g_\gamma(\cdot, \cdot)$ will serve as a tool for measuring consistency in our following analysis.

4.2. Estimation and model selection consistency

In this section, we analyze the performance of the proposed LVD estimator in a nonasymptotic framework. Let $(\hat{\boldsymbol{S}}, \hat{\boldsymbol{L}})$ be the solution to optimization problem (2.5). We establish rates of convergence under the assumption that $\boldsymbol{Y}_{(1)}, \ldots, \boldsymbol{Y}_{(n)}$ are independent and identically distributed samples from a sub-Gaussian distribution with covariance $\boldsymbol{\Sigma}_Y$. That is, we assume that there exists $K > 0$ such that

$$\mathsf{P}(|\boldsymbol{Y}_{(i)}| > t) \leq \exp(1 - t^2/K^2) \quad \text{for all } i \text{ and } t > 0. \tag{4.3}$$

This assumption implies that the observed variable $\boldsymbol{Y}$ have a sub-Gaussian tail, which is used to control the deviation between the sample covariance matrix and the population counterpart measured in the spectral norm, i.e., $\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_Y\|_2$.

Remark 4.1. If this condition is relaxed to that $\boldsymbol{Y}$ has finite fourth moment, Vershynin (2012) proved a suboptimal rate that only differed from the optimal one by a logarithmic factor $\log p$ and conjectured the same convergence rate. For the sake of simplicity, we stick to the sub-Gaussian tails. Still, we emphasize

that it is weaker than the joint Gaussian distribution in Chandrasekaran et al. (2012). One possible choice of the scaling is when the number of latent variables and the magnitude of coefficient matrix $\boldsymbol{B}$ are fixed, $\boldsymbol{X}$ could follow a discrete distribution with finite values or a sub-Gaussian distribution.

We are now ready to present results that establish estimation and model selection consistency of our LVD estimator. Denote by $\psi_1 = \phi_{\min}(\boldsymbol{\Sigma}_Y) \leq \phi_{\max}(\boldsymbol{\Sigma}_Y) = \psi_2$ and d the maximum node degree in $\boldsymbol{S}^*$. We further let θ be the minimum magnitude of nonzero entry of the sparse matrix $\boldsymbol{S}^*$ and σ be the minimum nonzero eigenvalue of the low-rank matrix $\boldsymbol{L}^*$.

Theorem 4.1 (Estimation and model selection consistency)**.** *Suppose that the irrepresentability condition (4.2) holds, and $\mu(\Omega^*)$ and $\xi(T^*)$ satisfy the following condition for identifiability*

$$\mu(\Omega^*)\xi(T^*) \leq \frac{1}{2}\left(\frac{\nu\alpha}{(2-\nu)\beta}\right)^2. \tag{4.4}$$

If we choose γ and λ_n such that

$$\gamma \in \left[\frac{\xi(T^*)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega^*)\beta(2-\nu)}\right], \quad \lambda_n = \max\left\{1, \frac{1}{\gamma}\right\}\frac{(3-2\nu)C_K}{\psi_1}\sqrt{\frac{p}{n}}, \tag{4.5}$$

where C_K is an absolute constant that only depends on K, and assume that

$$\sigma > \frac{3}{\alpha}\lambda_n \quad \textit{and} \quad \frac{\psi_1}{\psi_2} > \max\left\{1, \frac{1}{\gamma}\right\}\frac{3(3-2\nu)}{\alpha}C_K\sqrt{\frac{p}{n}},$$

then with probability at least $1 - 2\exp\{-p\}$, we have

$$g_\gamma\left(\hat{\boldsymbol{S}} - \boldsymbol{S}^*, \hat{\boldsymbol{L}} - \boldsymbol{L}^*\right) \leq \frac{3}{\alpha}\lambda_n. \tag{4.6}$$

Hence if we further assume $\theta > \frac{3\gamma}{\alpha}\lambda_n$, then under the same probability we have

$$sign(\hat{\boldsymbol{S}}) = sign(\boldsymbol{S}^*). \tag{4.7}$$

The proof of Theorem 4.1 is given in Appendix D. Theorem 4.1 says that under the identifiability condition (4.4), the irrepresentability condition (4.2) and some conditions on the smallest eigenvalue on $\boldsymbol{L}^*$ and $\boldsymbol{\Sigma}_Y$, then with proper chosen tuning parameters γ and λ_n, our proposed LVD procedure is consistent with certain rates of convergence. Model selection consistency is also achieved by further assuming a minimal signal strength condition on the precision matrix $\boldsymbol{S}^*$.

Our results look similar to those of Chandrasekaran et al. (2012) (hereafter referred to as CPW by taking initials of authors). As both methods require irrepresentable and identifiability conditions, it is interesting to compare our conditions (4.2) and (4.4) with theirs. Although taking the same form, specific

TABLE 1
Comparison of key quantities used in irrepresentabiliy and identifiability conditions.

	LVD	CPW
α_T	$1+\frac{\pi}{2(1-\pi)}$	$1+\frac{\pi}{1-\pi}$
α_Ω	$1+\frac{\pi}{p(1-\pi)}$	$1+\frac{2\pi}{p(1-\pi)}+\left\{1-(p-1)\left\{\frac{1+\frac{2\pi}{p(1-\pi)}+2\left(\frac{\pi}{p(1-\pi)}\right)^2}{1+\frac{2\pi}{p(1-\pi)}+2(p-1)\left(\frac{\pi}{p(1-\pi)}\right)^2}\right\}\right\}\left(\frac{\pi}{p(1-\pi)}\right)^2$
δ_T	0	0
δ_Ω	$\frac{\pi}{p(1-\pi)}$	$\frac{\pi(2-\pi)}{p(1-\pi)^2}$
β_T	<1	<1
β_Ω	$1+\frac{\pi}{1-\pi}$	$1+\frac{2\pi}{1-\pi}+\left(\frac{\pi}{1-\pi}\right)^2$

definitions of some quantities, such as α, β, and δ, differ from each other. The key difference is the operator (4.1) and the counterpart of theirs, which is

$$h_{\Sigma_Y}^{\mathrm{CPW}}(\boldsymbol{R}) = \boldsymbol{\Sigma}_Y \boldsymbol{R} \boldsymbol{\Sigma}_Y.$$

All six quantities defined with a $h_{\Sigma_Y}(\cdot)$ in our approach should be replaced with $h_{\Sigma_Y}^{\mathrm{CPW}}(\cdot)$ in their analysis. It is in general difficult to give a thorough comparison between these two sets of conditions. Instead, we consider a simple and special case in which these quantities could be computed explicitly.

Let $\boldsymbol{S}^* = \boldsymbol{I}_p$ be the $p \times p$ identity matrix and $\boldsymbol{L} = \pi(\mathbf{1}_p/\sqrt{p})(\mathbf{1}_p/\sqrt{p})^T = \pi \boldsymbol{J}_p/p$ be the rank-1 matrix with the maximal incoherence with standard orthonormal basis, where $\boldsymbol{J}_p$ is the $p \times p$ matrix with all ones and π is a parameter. Note that to guarantee $\boldsymbol{L}^*$ and $\boldsymbol{S}^* - \boldsymbol{L}^*$ are both positive definite, we must have $0 < \pi < 1$. In this specific circumstances, all six quantities for both methods are given in Table 1. Hence for the irrepresentability condition, as long as $p \geq 2$, our method requires $\frac{\pi}{p(1-\pi)} \Big/ \left(1+\frac{\pi}{p(1-\pi)}\right) < 1$, which always holds regardless of π. In contrast, the CPW approach requires $\frac{\max\{\delta_T^{\mathrm{CPW}}, \delta_\Omega^{\mathrm{CPW}}\}}{\min\{\alpha_T^{\mathrm{CPW}}, \alpha_\Omega^{\mathrm{CPW}}\}} = \frac{\delta_\Omega^{\mathrm{CPW}}}{\alpha_\Omega^{\mathrm{CPW}}} < 1$. For a fixed $0 < \pi < 1$, the solution of the above inequality (as a function of p) is the largest root of a cubic function and its approximate solution is given by $p \approx \frac{\sqrt{2}\pi}{(1-\pi)^2}$. This means if π is close to 1, the dimensionality should be large enough to guarantee the incoherence of the low-rank matrix. For example when $\pi = 0.9$, the identifiability condition fails for the CPW method when $p < 118$ (118 is the exact number and 127 is the approximation).

Similarly, we can compare the identifiability condition between these two methods, and it suffices to focus on the right hand side of equation (4.4). Denote by $\zeta = \frac{\nu\alpha}{(2-\nu)\beta}$ and set $\nu = \frac{1}{2}(1 - \frac{\delta}{\alpha})$, we have $\zeta = \frac{\nu\alpha}{(2-\nu)\beta} = \frac{\alpha(\alpha-\delta)}{(3\alpha+\delta)\beta}$. It can be shown after tedious calculation that $\zeta^{\mathrm{LVD}} > \zeta^{\mathrm{CPW}}$ always holds when $p \geq 2$ and $0 < \pi < 1$. For any fixed $0 < \pi < 1$, the ratio $\zeta^{\mathrm{LVD}}/\zeta^{\mathrm{CPW}}$ has a limit greater than 1 when $p \to \infty$. Figure 3 shows the ratio when $30 \leq p \leq 1000$ and $0 < \pi \leq 0.75$. Note that we do not include large π and small p as the irrepresentability condition for CPW barely or does not hold in this scenario.

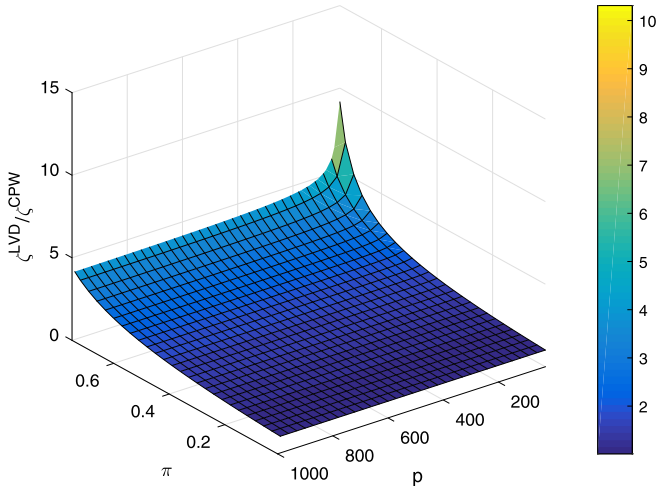


FIG 3. *Comparison of identifiability condition between LVD and CPW.*

Besides, the limit of $\zeta^{\mathrm{LVD}}/\zeta^{\mathrm{CPW}}$ goes to 1 when π is small and can be large as π increases.

We make a few more remarks on Theorem 4.1. First, the maximum degree of $\boldsymbol{S}^*$ and the rank of $\boldsymbol{L}$ do not explicitly appear in the theorem, but are hidden in conditions and other factors. To better illustrate their impacts on our results, we focus on $\mu(\Omega^*)$ and $\xi(T^*)$ and assume temporarily that $\alpha, \beta, \delta, \nu, \psi_1, \psi_2$ are of the order $O(1)$. Recall that $\mu(\Omega^*) \le d := \deg(\boldsymbol{S}^*)$ and $\xi(T^*) \sim \sqrt{r/p}$ if $\boldsymbol{L}^*$ is nearly maximally incoherent. In this circumstance, condition (4.4) essentially requires $d\sqrt{\frac{r}{p}} = O(1)$ and λ_n is of the order somewhere between $O(d\sqrt{\frac{p}{n}})$ and $O(\frac{p}{\sqrt{rn}})$ (depending on the choice of γ). These results match the rate of CPW, but hopefully improve other factors appearing in the rate and are derived after removing the joint Gaussian assumption.

Second, the consistency result (as well as CPW) holds for a range of values of γ, which is preferable in practice since we do not need to tune two parameters. If we choose γ at the upper end, we have the following corollary.

Corollary 4.1. *Under the assumptions of Theorem 4.1 and assume α, β, ν are constants. If we choose $\gamma \asymp \frac{1}{d}$ and $\lambda_n \asymp d\sqrt{\frac{p}{n}}$, then we have*

$$g_\gamma\left(\hat{\boldsymbol{S}} - \boldsymbol{S}^*, \hat{\boldsymbol{L}} - \boldsymbol{L}^*\right) = O_p\left(d\sqrt{\frac{p}{n}}\right)$$

as long as $\sigma, \frac{\psi_1}{\psi_2} \gtrsim d\sqrt{\frac{p}{n}}$. Further, model selection consistency for $\boldsymbol{S}^*$ holds if $\theta \gtrsim \sqrt{\frac{p}{n}}$.

Finally, our technical analysis is different from CPW. While they resorted to Brouwer's fixed point theorem in proofs, we use a more direct approach to

analyze the LVD procedure owing to its simple quadratic form. Since there is no log-determinant term in our loss function, we additionally assume the spectrum of observed covariance matrix $\mathbf{\Sigma}_Y$ is well-conditioned to guarantee positive-definiteness of our estimates.

5. Simulation studies

In this section simulation studies are carried out to compare the numerical performance of our LVD estimator $\hat{\boldsymbol{S}}_{\mathrm{LVD}}$, the regularized maximum likelihood estimator $\hat{\boldsymbol{S}}_{\mathrm{CPW}}$ from Chandrasekaran et al. (2012), and the graphical lasso $\hat{\boldsymbol{S}}_{\mathrm{glasso}}$ from Friedman et al. (2008). We are particularly interested in investigating how these methods perform differently in relation to different settings of latent variables and how the dimensionality affects their performance.

We consider nine models as shown in Table 2. For each model, we generate a $r \times p$ coefficient matrix $\boldsymbol{B}$ with its entries independently following $\mathrm{Unif}([-1.5,-0.5] \cup [0.5,1.5])$. Each element of the $n \times r$ design matrix $\boldsymbol{X}$ is independently drawn from a Bernoulli(0.5) or a standard Normal distribution according to different model set-ups. We generate the $p \times p$ precision matrix with $\mathsf{P}(S^*_{ij} \neq 0 | i \neq j)$ shown in Table 2 and diagonal entries equaling to 1. If $S^*_{ij} \neq 0, i \neq j$, we set $|S_{ij}| = 0.25$ and its sign with equal probability. Finally, we generate a $n \times p$ random error matrix $\boldsymbol{E}$ so that each row $\boldsymbol{E}_{i\cdot} \sim \mathcal{N}_p(0, (\boldsymbol{S}^*)^{-1})$. The $n \times p$ outcome matrix $\boldsymbol{Y}$ is set to be $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}$.

The nine models are divided into three groups in terms of dimensionality. While the sample size are fixed at $n = 200$, the number of observed variables ranges from $p = 20$ in the first three models to $p = 50$ in Models 4–6, and we consider a moderately high dimension with $p = 100$ in Models 7–9. Within each group, the latent variables follow from Bernoulli distributions in the first model and Normal distributions in the second model. We also set $r = 0$ in the third model, which corresponds to no latent variable at all. When $\boldsymbol{X}$ is normal, $\boldsymbol{Y}$ also has a normal distribution, which may give an edge to $\hat{\boldsymbol{S}}_{\mathrm{CPW}}$ since it adopts the maximum likelihood approach. If $\boldsymbol{X}$ has binary values, on the other hand,

TABLE 2
Parameters for nine models.

	Parameters		
	(n, p, r)	type of X	$\mathsf{P}(S^*_{ij} \neq 0 \mid i \neq j)$
Model 1	$(200, 20, 2)$	Bernoulli	0.2
Model 2	$(200, 20, 2)$	Normal	0.2
Model 3	$(200, 20, 0)$	–	0.2
Model 4	$(200, 50, 2)$	Bernoulli	0.1
Model 5	$(200, 50, 2)$	Normal	0.1
Model 6	$(200, 50, 0)$	–	0.1
Model 7	$(200, 100, 5)$	Bernoulli	0.05
Model 8	$(200, 100, 5)$	Normal	0.05
Model 9	$(200, 100, 0)$	–	0.05

$\boldsymbol{Y}$ will follow a mixture Gaussian distribution. This setting is very common in real data analysis when $\boldsymbol{X}$ is used to model batch effects or other categorical covariates such as sex and race. Finally, we expect $\hat{\boldsymbol{S}}_{\text{glasso}}$ to perform well when there is no latent variable. The number of latent variables is set to be $r = 2$ in Models 1–2 and 4–5, and increases to $r = 5$ in Models 7–8.

For each model, we choose the tuning parameter λ_n using 10-fold cross validation as described in Section 3.2. As for γ, we have shown that our estimate is insensitive to a wide choice of low-rank tuning parameters, which is also observed by Chandrasekaran et al. (2012); Yuan (2012) and in our numerical experiments, hence we report here the results with a pre-chosen fixed constant that may vary across models. The tuning parameters for CPW and glasso are also chosen using 10-fold cross validation, except that the cross validation error is defined as the negative log likelihood according to their loss functions. Further, the tuning parameter γ for CPW is also fixed at a constant such that the low-rank matrices estimated by the LVD and CPW have approximately the same rank for the sake of fairness.

Since our main goal is to estimate $\boldsymbol{S}^*$, especially its non-zero pattern that carries information on direct interactions among manifest variables after adjusting for the effect of latent variables, we use six measures on estimation and model selection quantities to assess the performance of each method. The estimation error $\hat{\boldsymbol{S}} - \boldsymbol{S}^*$ are evaluated by the spectral norm, the matrix ℓ_1 norm and the Frobenius norm. Further, the model selection performance is characterized by the true positive rate (TPR), the true negative rate (TNR), and the Matthews correlation coefficient (MCC). Specifically, these three measures are defined as

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

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives in identifying the non-zero elements in the off-diagonal precision matrix.

The simulation results for Models 1–9 are summarized in Table 3. We see that when p is small, our proposed LVD procedure outperforms the other two methods in terms of all measures, especially when $\boldsymbol{X}$ has a discrete distribution. When the dimensionality is moderately high, the performance of our estimator and that of $\hat{\boldsymbol{S}}_{\text{CPW}}$ on TPR, TNR and MCC are comparable to each other. This is because the marginal distribution of $\boldsymbol{Y}$ does not deviate significantly from a multivariate normal distribution owing to dimensionality, and the fact that the penalized likelihood can be viewed as a penalized log-determinant Bregman divergence (Ravikumar et al., 2011). Nevertheless, we note that $\hat{\boldsymbol{S}}_{\text{LVD}}$ has the smallest estimation error even if when p is moderately high. Finally, it can be seen that the graphical lasso estimator only performs well in the absence of latent variables as expected. Notice that glasso may not achieve the best estimation performance even in the case without latent variables, owing to different choices

TABLE 3
Simulation results for Models 1–9. Each performance measure is averaged over 100 replications with standard deviations shown in parentheses.

Method	$\|\cdot\|_2$	$\|\cdot\|_{\ell_1}$	$\|\cdot\|_F$	TPR	TNR	MCC	rank($\hat{\boldsymbol{L}}$)
Model 1, $(n,p,r)=(200,20,2)$, Bernoulli							
LVD	0.96 (0.01)	1.71 (0.03)	1.93 (0.02)	0.75 (0.01)	0.89 (0.00)	0.61 (0.01)	5.14 (0.06)
CPW	1.12 (0.01)	1.97 (0.02)	2.36 (0.02)	0.53 (0.02)	0.96 (0.00)	0.57 (0.01)	5.05(0.07)
glasso	1.34 (0.01)	2.19 (0.02)	2.59 (0.02)	0.79 (0.01)	0.41 (0.00)	0.16 (0.01)	–
Model 2, $(n,p,r)=(200,20,2)$, Normal							
LVD	1.00 (0.01)	1.76 (0.03)	1.96 (0.03)	0.72 (0.02)	0.90 (0.01)	0.59 (0.01)	4.94 (0.08)
CPW	1.13 (0.01)	1.97 (0.03)	2.36 (0.03)	0.56 (0.02)	0.95 (0.00)	0.56 (0.01)	4.70 (0.07)
glasso	1.35 (0.01)	2.23 (0.02)	2.61 (0.02)	0.78 (0.01)	0.41 (0.00)	0.15 (0.01)	–
Model 3, $(n,p,r)=(200,20,0)$							
LVD	0.76 (0.01)	1.24 (0.02)	1.51 (0.02)	0.97 (0.01)	0.88 (0.01)	0.76 (0.01)	0 (0)
CPW	0.93 (0.01)	1.49 (0.02)	1.92 (0.02)	0.95 (0.01)	0.87 (0.01)	0.71 (0.01)	0.03 (0.02)
glasso	0.93 (0.01)	1.49 (0.02)	1.91 (0.02)	0.95 (0.01)	0.86 (0.01)	0.71 (0.01)	–
Model 4, $(n,p,r)=(200,50,2)$, Bernoulli							
LVD	0.86 (0.01)	1.68 (0.02)	2.55 (0.03)	0.75 (0.01)	0.89 (0.00)	0.51 (0.01)	4.67 (0.10)
CPW	0.97 (0.01)	1.84 (0.02)	3.02 (0.02)	0.64 (0.02)	0.94 (0.00)	0.53 (0.01)	4.70 (0.13)
glasso	1.25 (0.01)	2.32 (0.02)	3.39 (0.03)	0.64 (0.01)	0.59 (0.00)	0.14 (0.01)	–
Model 5, $(n,p,r)=(200,50,2)$, Normal							
LVD	0.82 (0.01)	1.63 (0.02)	2.50 (0.02)	0.74 (0.01)	0.89 (0.00)	0.49 (0.00)	4.55 (0.08)
CPW	0.97 (0.01)	1.81 (0.02)	3.06 (0.02)	0.56 (0.02)	0.95 (0.00)	0.51 (0.01)	4.22 (0.13)
glasso	1.23 (0.01)	2.27 (0.02)	3.40 (0.03)	0.60 (0.01)	0.59 (0.00)	0.12 (0.01)	–
Model 6, $(n,p,r)=(200,50,0)$							
LVD	0.81 (0.01)	1.59 (0.02)	2.45 (0.02)	0.74 (0.01)	0.94 (0.00)	0.63 (0.00)	2.88 (0.10)
CPW	0.92 (0.01)	1.76 (0.02)	2.89 (0.03)	0.61 (0.01)	0.97 (0.00)	0.65 (0.01)	2.76 (0.13)
glasso	0.90 (0.01)	1.70 (0.02)	2.84 (0.03)	0.69 (0.01)	0.95 (0.00)	0.63 (0.00)	–
Model 7, $(n,p,r)=(200,100,5)$, Bernoulli							
LVD	0.95 (0.01)	1.95 (0.02)	4.04 (0.01)	0.58 (0.00)	0.96 (0.00)	0.46 (0.00)	10.22 (0.09)
CPW	0.99 (0.01)	2.03 (0.02)	4.36 (0.04)	0.57 (0.01)	0.95 (0.00)	0.44 (0.00)	9.52 (0.23)
glasso	1.28 (0.00)	2.79 (0.01)	4.89 (0.03)	0.68 (0.01)	0.64 (0.00)	0.14 (0.00)	–
Model 8, $(n,p,r)=(200,100,5)$, Normal							
LVD	0.96 (0.00)	1.96 (0.02)	4.07 (0.01)	0.57 (0.00)	0.96 (0.00)	0.46 (0.00)	10.15 (0.08)
CPW	1.01 (0.01)	2.04 (0.02)	4.42 (0.04)	0.51 (0.01)	0.96 (0.00)	0.45 (0.00)	10.35 (0.24)
glasso	1.28 (0.00)	2.81 (0.02)	4.92 (0.03)	0.67 (0.01)	0.64 (0.00)	0.14 (0.00)	–
Model 9, $(n,p,r)=(200,100,0)$							
LVD	0.86 (0.01)	1.75 (0.02)	3.60 (0.03)	0.72 (0.01)	0.96 (0.00)	0.58 (0.01)	4.06 (0.13)
CPW	0.92 (0.01)	1.83 (0.02)	3.98 (0.04)	0.69 (0.01)	0.96 (0.00)	0.58 (0.01)	3.48 (0.19)
glasso	0.89 (0.01)	1.80 (0.01)	3.84 (0.04)	0.76 (0.01)	0.94 (0.00)	0.55 (0.01)	–

of tuning parameters caused by different loss functions used in cross validation. While we use the Frobenious norm as in (3.12), the other two methods adopt the log determinant loss. Our method tend to choose a smaller tuning parameter compared with those from CPW and glasso, which may lead to better estimation performance. Similar results have also be observed in Zhang and Zou (2014), and such difference is negligible when tuning parameters are changing; see later in Figure 4.

To further investigate the performance on graph structure recovery, we obtain the receiver operating characteristic (ROC) curve for each simulated data set by

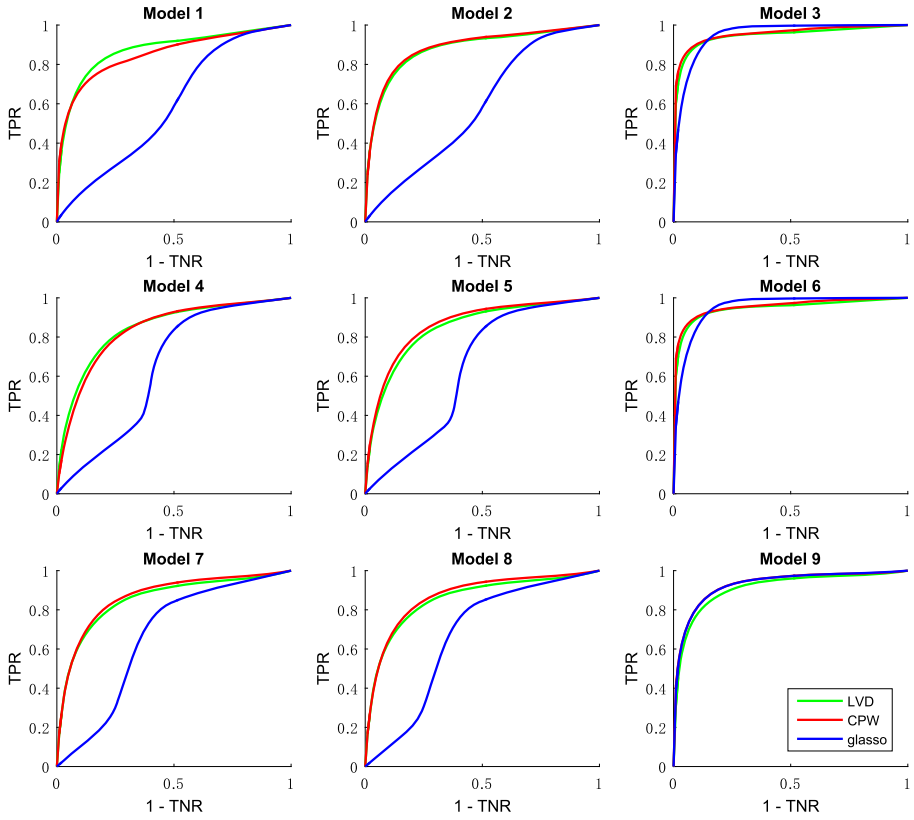


FIG 4. *The average ROC curves obtained by varying the tuning parameter over* 100 *replications.*

varying the turning parameter for the sparse matrix. Figure 4 shows the ROC curves averaged over 100 replications. Figures on the same row correspond to same dimensionality and different set-ups of latent variable $\boldsymbol{X}$. The performance of our LVD estimator is comparable to that of the penalized likelihood method in most cases and better than the other two methods when $p = 20$ and $\boldsymbol{X}$ follows a Bernoulli distribution. The graphical lasso estimator, on the other hand, performs poorly when there exists latent variables, that is, in Models 1–2, 4–5, and 7–8. We also note that in these cases, there is a "kink" in the ROC curve for glasso. Below this critical point, the graphical lasso estimator performs no better than random guess; as p and r increase (thus the effects of latent variables decrease), this critical point moves towards the original point. This phenomenon implies that a direct application of graphical lasso in real data analysis may lead to unreliable results if there is concern over latent variables, especially when the number of node is not large. Finally, we emphasize that the two methods accounting for the effect of latent variables, $\hat{\boldsymbol{S}}_{\text{LVD}}$ and $\hat{\boldsymbol{S}}_{\text{CPW}}$, perform no worse than $\hat{\boldsymbol{S}}_{\text{glasso}}$ in all models.

6. Analysis of a yeast data set

To demonstrate our method in real data analysis, we present results from the analysis of a yeast genetical genomics data set generated by Brem and Kruglyak (2005). They used BY4716 and the wild isolate RM11-1a as parent strains to grow 112 yeast segregants. Then they isolated RNA and hybridized cDNA to microarrays that had 6216 yeast genes assayed on each array. It is nearly impossible to build a gene-gene interaction network for all the genes owing to the small sample size and restricted perturbation in biological systems. We instead apply our method to a set of 56 genes that belong to the yeast mitogen-activated protein kinase (MAPK) signaling pathway provided by the KEGG database (Kanehisa et al., 2010).

The S. cerevisiae genome encodes multiple MAP kinase orthologs. Fus3 mediates cellular response to peptide pheromones, Kss1 permits adjustment to nutrient limiting conditions and Hog1 is necessary for survival under hyperosmotic conditions. Besides, Slt2/Mpk1 is required for repair of injuries to the cell wall. Figure 5 displays the illustrative pathway structure. Since several genes such as Ste20, Ste12 and Ste11 appear in multiple locations, this graph cannot be treated as the "true graph" for evaluating or comparing different methods. Our goal is to construct a conditional independent network among these genes at the expression levels and compare the results to this reference graph in the hope of gaining some biological interpretations.

We apply the above methods to this set of 56 genes and use 10-fold cross-validation to choose tuning parameters for all approaches. The model selected by cross-validation include 44 (LVD), 508 (CPW) and 536 (glasso) links among the 56 genes, respectively. While the graphical lasso estimator tends to choose a large number of interactions as expected, we see that $\hat{\boldsymbol{S}}_{\mathrm{CPW}}$ also selects too

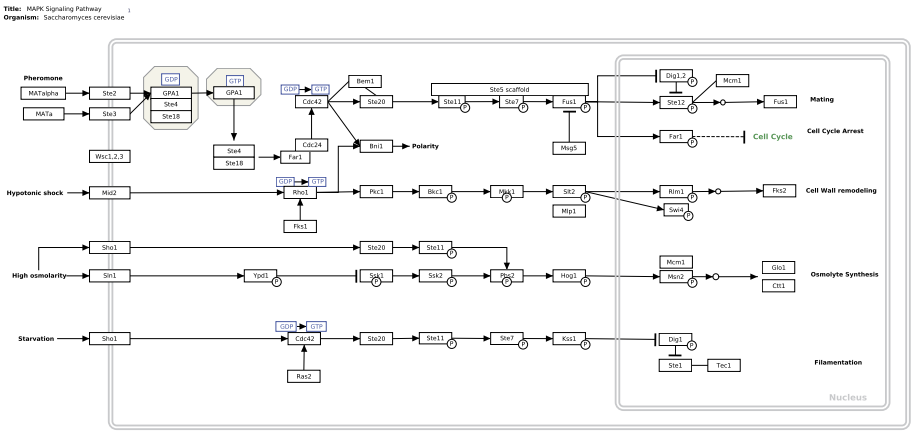


FIG 5. *The yeast MAPK pathway downloaded from* http://www.wikipathways.org/index.php/Pathway:WP510 *(Kelder et al., 2012).*

many links to interpret. On the other hand, our LVD estimator results in a much sparser graph, which implies that the yeast data set we are analyzing may be subject to the effect of latent variables. We emphasize that we choose γ such that the low-rank matrices estimated by LVD and CPW have the same rank. Further, although tuning γ will change the rank of the estimated low-rank matrix, the resulting graph of these two methods are stable across different values of γ.

Figure 6 shows the undirected graph for 37 linked genes on the MAPK pathway constructed base on our estimator. Although we do not expect the resulting graph to fully recover the original MAPK pathway, it indeed has some biological meanings. For example, DIG1, FUS1, FUS3, GPA1, FAR1, STE2, STE3, STE12, STE18 and STE20 are linked together, which suggests a strong interactions between these genes because they are all involved in the yeast pheromone and mating process. Similarly, PKC1 and MKK1, MLP1 and SWI4 are linked together, while Slt2, RLM1, FKS1 and MID2 are linked through CTT1 and MSN4 owing to their interaction in the cell wall remodelling process. Finally, CTT1, SHO1, MSN4, YPD1, MCM1 and SLN1 are connected via SLT2 and RLM1 since they participate in osmolyte synthesis. Our method also estimates a low-rank matrix that summarizes the effects of latent variables; some methods have been developed to interpret latent variables (Taeb and Chandrasekaran, 2016) and factorize this low-rank matrix, for example, into sparse factors (Witten et al., 2009).

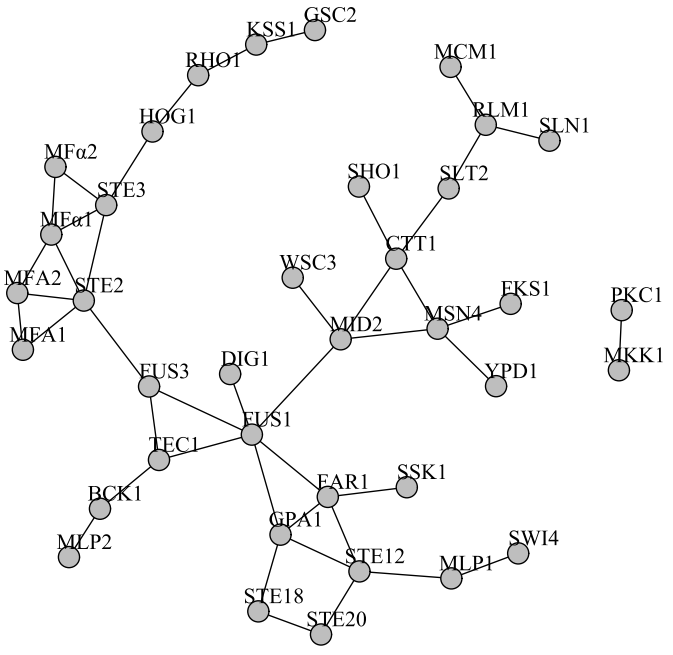


FIG 6. *Latent variables' effects adjusted conditional independence graph constructed based on the estimated precision matrix for the 37 genes on the yeast MAPK pathway.*

To compare across different methods, we follow the stability selection approach (Meinshausen and Bühlmann, 2010) and fit the graph for all three methods based on 100 bootstrap subsamples of size $\lfloor n/2 \rfloor$ with tuning parameters fixed at the values chosen before. An extra cut-off threshold π_{thr} that determines whether an interaction is stable varies across methods and is derived according to Theorem 1 in Meinshausen and Bühlmann (2010) such that the expected number of wrong edges is less than 30 among the 1540 possible edges. Since the cut-off thresholds π_{thr} for stability selection have respective values for different methods, the resulting graphs have comparable number of edges with 36 for LVD, 39 for CPW and 50 for glasso. Among these edges, 26 of them are selected by all three methods, with another 2 links selected simultaneously by our method and glasso/CPW. Although our estimator selects the least number of edges in the graph, it involves the most number of genes whose interactions can be interpreted by the real MAPK pathway illustrated in Figure 5.

7. Conclusion

In this paper, we study the problem of modeling the statistical dependence of random variables as a sparse graphical model conditioning on a few additional latent components. Since it is very common that the observed variables are correlated with some hidden variables in real world data, considering such model is of great importance. We develop the LVD procedure, an ℓ_1 and trace regularized minimization method for latent-variable graphical model selection, and prove its estimation and model selection consistency under certain identifiability conditions. We also propose a computationally efficient algorithm that can be applied to high-dimensional settings. Our simulation studies verify our theoretical results and show the superior performance of our method over other approaches. Finally we demonstrate the effectiveness of our method in a yeast gene expression data set.

Several improvements and extensions of our method deserve further analysis. Both joint Gaussian distribution and our linear factor model (2.1) assume a Gaussian distribution for the random error to facilitate the estimation of conditional independence among variables. Graphical models for non-Gaussian data have attracted a lot of attention recently, including discrete variables (Ravikumar et al., 2010; Loh and Wainwright, 2013), exponential families (Yang et al., 2015), and mixed data (Chen et al., 2015; Cheng et al., 2016; Fan et al., 2017). It is of great interest to develop statistically consistent methods for latent-variable modeling with such non-Gaussian variables. As a matter of fact, Tan et al. (2016) have already studied semiparametric exponential family graphical models with latent variables when replicates within each subject are available. Another important issue that is worthwhile to pursue concerns about controlling the false discovery rate (FDR) and constructing confidence intervals for latent variable graphical models. Similar problems for Gaussian graphical models have been considered by Liu (2013); Wasserman et al. (2014); Janková and van de Geer

(2015); Ren et al. (2015). Particularly, results on entrywise confidence interval for latent variables graphical models under the joint Gaussian assumption have been established in Ren et al. (2015). It is promising to extend these methodologies to the more general distribution assumptions, with an emphasize on controlling the FDR of detected edges.

Appendix A: Additional simulation studies

A.1. Stability of γ

We have already showed in Theorem 4.1 that estimation and model selection consistency is possible for a range of γ. In this section we investigate the performance of our LVD estimator when γ is varied. For simplicity, we only consider Model 4 in Section 5, that is, $n = 200, p = 50, r = 2$ and the latent variables take binary values. Recall that in Table 3, the average rank of the estimated low-rank matrix is 4.67 for our LVD estimator. Table 4 summarizes results when we vary γ. For each γ fixed, we choose λ_n using 10-fold cross validation. Note that $\gamma = 0.15$ corresponds to results in Table 3.

We see that as γ varies from 0.12 to 0.35, the average estimated rank of the low-rank matrix ranges from 2.85 to 13.11. Nevertheless, the estimation errors measured in all three quantities remain stable regardless of different choices of γ, especially when $0.15 \leq \gamma \leq 0.3$. Besides, the Matthews Correlation Coefficient

TABLE 4
Simulation results for a range of γ. Each performance measure is averaged over 100 replications with standard deviations shown in parentheses.

γ	$\|\cdot\|_2$	$\|\cdot\|_{\ell_1}$	$\|\cdot\|_F$	TPR	TNR	MCC	rank($\hat{\boldsymbol{L}}$)
0.12	0.86 (0.01)	1.70 (0.02)	2.54 (0.03)	0.78 (0.01)	0.85 (0.00)	0.46 (0.00)	2.85 (0.08)
0.13	0.86 (0.01)	1.68 (0.02)	2.53 (0.03)	0.77 (0.01)	0.86 (0.00)	0.47 (0.01)	3.36 (0.08)
0.14	0.86 (0.01)	1.67 (0.02)	2.53 (0.03)	0.77 (0.01)	0.87 (0.00)	0.49 (0.01)	4.07 (0.08)
0.15	0.86 (0.01)	1.68 (0.02)	2.55 (0.03)	0.75 (0.01)	0.89 (0.00)	0.51 (0.01)	4.67 (0.10)
0.16	0.85 (0.01)	1.67 (0.02)	2.54 (0.02)	0.74 (0.01)	0.90 (0.00)	0.52 (0.01)	5.24 (0.11)
0.17	0.85 (0.01)	1.67 (0.02)	2.53 (0.02)	0.73 (0.01)	0.91 (0.00)	0.52 (0.01)	6.03 (0.11)
0.18	0.85 (0.01)	1.67 (0.02)	2.53 (0.02)	0.71 (0.01)	0.91 (0.00)	0.53 (0.01)	6.69 (0.13)
0.19	0.85 (0.01)	1.67 (0.02)	2.53 (0.02)	0.70 (0.01)	0.92 (0.00)	0.54 (0.01)	7.26 (0.13)
0.20	0.85 (0.01)	1.68 (0.02)	2.54 (0.02)	0.68 (0.01)	0.93 (0.00)	0.55 (0.01)	7.83 (0.13)
0.22	0.84 (0.01)	1.69 (0.02)	2.54 (0.02)	0.65 (0.01)	0.94 (0.00)	0.55 (0.01)	8.99 (0.13)
0.24	0.84 (0.01)	1.70 (0.02)	2.54 (0.02)	0.62 (0.02)	0.95 (0.00)	0.55 (0.01)	10.11 (0.15)
0.26	0.83 (0.01)	1.71 (0.02)	2.54 (0.02)	0.58 (0.02)	0.96 (0.00)	0.55 (0.01)	11.22 (0.18)
0.28	0.84 (0.01)	1.73 (0.02)	2.59 (0.03)	0.54 (0.02)	0.97 (0.00)	0.54 (0.01)	11.91 (0.22)
0.30	0.85 (0.01)	1.75 (0.03)	2.64 (0.03)	0.49 (0.02)	0.97 (0.00)	0.53 (0.01)	12.41 (0.25)
0.32	0.85 (0.01)	1.77 (0.03)	2.69 (0.04)	0.45 (0.02)	0.98 (0.00)	0.51 (0.01)	12.70 (0.28)
0.35	0.87 (0.01)	1.80 (0.03)	2.76 (0.04)	0.38 (0.02)	0.98 (0.00)	0.48 (0.01)	13.11 (0.31)

is even larger when we increase γ, i.e., when our LVD method estimates the $\hat{\boldsymbol{L}}$ with a moderate high rank. Consequently, the estimated sparse matrix becomes more sparse (TPR decreases and TNR increases) since γ provides a trade-off between the sparse and the low-rank matrices.

A.2. Effect of r

To understand how the global effects caused by latent variables affect the performance of different methods, we extend the simulation study in Section 5 by varying r. In order to better illustrate the effect of r, we consider the low-dimensional scenario, that is, $n = 200$ and $p = 20$. As is explained in Section 5, we fix γ at a pre-chosen constant so that the low-rank matrices estimated by LVD and CPW have approximately the same rank, and we only set $\boldsymbol{X}$ taking binary values for simplicity.

Table 5 summarizes results when r is varied from 0 to 4. We have already seen in Section 5 that glasso only performs well when there is no latent variable at all. Although the performance of all three methods deteriorates as r increases, our LVD procedure outperforms other approaches in all cases.

TABLE 5
Simulation results for a range of r. Each performance measure is averaged over 100 replications with standard deviations shown in parentheses.

Method	$\|\cdot\|_2$	$\|\cdot\|_{\ell_1}$	$\|\cdot\|_F$	TPR	TNR	MCC	rank($\hat{\boldsymbol{L}}$)
				$r = 0$			
LVD	0.76 (0.01)	1.24 (0.02)	1.51 (0.02)	0.97 (0.01)	0.88 (0.01)	0.76 (0.01)	0 (0)
CPW	0.93 (0.01)	1.49 (0.02)	1.92 (0.02)	0.95 (0.01)	0.87 (0.01)	0.71 (0.01)	0.03 (0.02)
glasso	0.93 (0.01)	1.49 (0.02)	1.91 (0.02)	0.95 (0.01)	0.86 (0.01)	0.71 (0.01)	–
				$r = 1$			
LVD	0.96 (0.01)	1.69 (0.03)	1.94 (0.03)	0.74 (0.02)	0.95 (0.00)	0.72 (0.01)	4.13 (0.07)
CPW	1.11 (0.01)	1.94 (0.03)	2.34 (0.02)	0.50 (0.02)	0.99 (0.00)	0.62 (0.01)	4.04(0.08)
glasso	1.26 (0.01)	2.03 (0.02)	2.36 (0.02)	0.76 (0.01)	0.45 (0.01)	0.17 (0.01)	–
				$r = 2$			
LVD	0.96 (0.01)	1.71 (0.03)	1.93 (0.02)	0.75 (0.01)	0.89 (0.00)	0.61 (0.01)	5.14 (0.06)
CPW	1.12 (0.01)	1.97 (0.02)	2.36 (0.02)	0.53 (0.02)	0.96 (0.00)	0.57 (0.01)	5.05(0.07)
glasso	1.34 (0.01)	2.19 (0.02)	2.59 (0.02)	0.79 (0.01)	0.41 (0.00)	0.16 (0.01)	–
				$r = 3$			
LVD	1.01 (0.01)	1.86 (0.02)	2.09 (0.02)	0.55 (0.01)	0.93 (0.00)	0.51 (0.01)	6.68 (0.06)
CPW	1.16 (0.01)	2.06 (0.02)	2.47 (0.02)	0.26 (0.01)	0.98 (0.00)	0.38 (0.01)	6.46 (0.08)
glasso	1.40 (0.01)	2.36 (0.02)	2.80 (0.02)	0.79 (0.01)	0.36 (0.00)	0.13 (0.01)	–
				$r = 4$			
LVD	1.06 (0.01)	1.92 (0.03)	2.20 (0.02)	0.47 (0.01)	0.91 (0.00)	0.41 (0.01)	7.38 (0.07)
CPW	1.18 (0.01)	2.09 (0.03)	2.55 (0.02)	0.21 (0.02)	0.98 (0.00)	0.31 (0.01)	7.10 (0.09)
glasso	1.44 (0.01)	2.44 (0.02)	2.95 (0.02)	0.80 (0.01)	0.31 (0.00)	0.10 (0.01)	–

Appendix B: Derivation of (3.10)

Proof. We rewrite the first-order optimality condition for (3.9) as

$$\frac{1}{2}(\boldsymbol{R}^{k+1}\boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n\boldsymbol{R}^{k+1}) - \boldsymbol{I} + \rho\left(\boldsymbol{R}^{k+1} - \boldsymbol{W}_R^k\right) = \boldsymbol{0}. \tag{B.1}$$

Since $\boldsymbol{\Sigma}_n$ is positive-semidefinite, let $\boldsymbol{\Sigma}_n = \boldsymbol{U}_\Sigma \mathrm{diag}(\boldsymbol{\sigma}_\Sigma)\boldsymbol{U}_\Sigma^T$ be the eigenvalue decomposition of $\boldsymbol{\Sigma}_n$, where $\boldsymbol{U}_\Sigma \in \mathbb{R}^{p\times p}$. Write $\dot{\boldsymbol{R}}^{k+1} = \boldsymbol{U}_\Sigma^T\boldsymbol{R}^{k+1}\boldsymbol{U}_\Sigma$ and multiply $\boldsymbol{U}_\Sigma^T$ from the left and $\boldsymbol{U}_\Sigma$ from the right in (B.1), we obtain

$$\frac{1}{2}\left(\dot{\boldsymbol{R}}^{k+1}\mathrm{diag}(\boldsymbol{\sigma}_\Sigma) + \mathrm{diag}(\boldsymbol{\sigma}_\Sigma)\dot{\boldsymbol{R}}^{k+1}\right) + \rho\dot{\boldsymbol{R}}^{k+1} = \rho\boldsymbol{U}_\Sigma^T(\boldsymbol{W}_R^k + \boldsymbol{I})\boldsymbol{U}_\Sigma.$$

To solve $\dot{\boldsymbol{R}}^{k+1}$ from the above display, all we need to do is to make sure that for each $i, j = 1, \ldots, p$, we have

$$\frac{1}{2}\left(\dot{R}_{ij}^{k+1}\sigma_{\Sigma,j} + \sigma_{\Sigma,i}\dot{R}_{ij}^{k+1}\right) + \rho\dot{R}_{ij}^{k+1} = \left(\rho\boldsymbol{U}_\Sigma^T(\boldsymbol{W}_R^k + \boldsymbol{I})\boldsymbol{U}_\Sigma\right)_{ij},$$

which establishes (3.10). □

Appendix C: Numerical algorithm for solving (3.2)

The ADMM described in Algorithm 1 for solving (3.1) need to be slightly modified. Specifically, we need to update $\boldsymbol{R}^{k+1}$ by solving the following optimization problem instead of (3.9):

$$\boldsymbol{R}^{k+1} = \underset{\boldsymbol{R}\succeq\varepsilon\boldsymbol{I}}{\arg\min}\ \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \frac{\rho}{2}\left\|\boldsymbol{R} - \boldsymbol{W}_R^k\right\|_F^2. \tag{C.1}$$

As there is no explicit form solution for this problem, we could solve it via an inner loop of ADMM. After introducing an auxiliary variable $\check{\boldsymbol{R}}$ such that $\check{\boldsymbol{R}} = \boldsymbol{R}$, we turn to solve

$$\begin{array}{ll} \min & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \frac{\rho}{2}\left\|\boldsymbol{R} - \boldsymbol{W}_R^k\right\|_F^2 + \mathcal{I}(\check{\boldsymbol{R}}\succeq\varepsilon\boldsymbol{I}) \\ \text{subject to} & \check{\boldsymbol{R}} = \boldsymbol{R}. \end{array}$$

It can be easily seen that the updates of $\boldsymbol{R}$ and $\check{\boldsymbol{R}}$ are similar to (3.10) and (3.8). We finally note that it is sufficient to solve (3.1) in practice, which is much simpler and faster.

Appendix D: Proof of Theorem 4.1

We start from a Lemma that characterizes the deviation of the sample covariance matrix from its true value, whose proof can be found in Vershynin (2010).

Lemma D.1. *Under assumption (4.3), the empirical covariance matrix satisfies*

$$\mathsf{P}\left\{\|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_Y\|_2 \leq C_K\sqrt{\frac{p}{n}}\right\} \geq 1 - 2\exp(-p)$$

where C_K is a constant that only depends on K.

Proof of Theorem 4.1. In what follows we will condition on the event that the result in Lemma D.1 holds, which occurs with probability at least $1 - 2\exp(p)$. The key idea used in the proof is a technique known as the primal-dual witness method used previously in analysis of the Lasso (Wainwright, 2009) and graphical model (Ravikumar et al., 2011). The proof is summarized in the following three steps.

Step 1. Let $(\check{\boldsymbol{S}}, \check{\boldsymbol{L}}, \check{\boldsymbol{R}})$ be the solution to the following problem:

$$\begin{array}{ll} \min_{\boldsymbol{S},\boldsymbol{L},\boldsymbol{R}} & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \lambda_n(\gamma\|\boldsymbol{S}\|_1 + \|\boldsymbol{L}\|_*), \\ \text{subject to} & \boldsymbol{R} = \boldsymbol{S} - \boldsymbol{L}, \boldsymbol{R} = \boldsymbol{R}^T, \boldsymbol{L} = \boldsymbol{L}^T. \end{array} \tag{D.1}$$

Note that there is no positive-(semi)definite constraint in (D.1). It will be shown later in *Step 2* that under conditions assumed in Theorem 4.1, $\check{\boldsymbol{R}}$ and $\check{\boldsymbol{L}}$ are positive-(semi)definite with high probability, hence it suffices to study problem (D.1). We will show in this step that if we solve (D.1) subject to additional constraints that $\boldsymbol{S}$ and $\boldsymbol{L}$ belong to tangent spaces Ω^* and T^*, that is,

$$\begin{array}{ll} \min_{\boldsymbol{S},\boldsymbol{L},\boldsymbol{R}} & \frac{1}{2}\mathrm{tr}(\boldsymbol{R}\boldsymbol{\Sigma}_n\boldsymbol{R}) - \mathrm{tr}(\boldsymbol{R}) + \lambda_n(\gamma\|\boldsymbol{S}\|_1 + \|\boldsymbol{L}\|_*), \\ \text{subject to} & \boldsymbol{R} = \boldsymbol{S} - \boldsymbol{L}, \boldsymbol{R} = \boldsymbol{R}^T, \boldsymbol{L} = \boldsymbol{L}^T, \boldsymbol{S} \in \Omega^*, \boldsymbol{L} \in T^*, \end{array} \tag{D.2}$$

then these two sets of solutions are the same.

Step 2. We analyze the solution to optimization problem (D.2) (denoted by $(\bar{\boldsymbol{S}}, \bar{\boldsymbol{L}}, \bar{\boldsymbol{R}})$) and show that $(\bar{\boldsymbol{S}}, \bar{\boldsymbol{L}}, \bar{\boldsymbol{R}})$ enjoys estimation consistency measured in terms of $g_\gamma(\bar{\boldsymbol{S}} - \boldsymbol{S}^*, \bar{\boldsymbol{L}} - \boldsymbol{L}^*)$ as claimed in Theorem 4.1. We further show that under suitable conditions on minimum magnitude of nonzero entry of $\boldsymbol{S}^*$ and minimum nonzero singular value of $\boldsymbol{L}^*$, $\check{\boldsymbol{R}}$ and $\check{\boldsymbol{L}}$ are positive-(semi)definite with high probability, thus model selection consistency for $\boldsymbol{S}^*$ is achieved. Therefore, we have $(\hat{\boldsymbol{S}}, \hat{\boldsymbol{L}}, \hat{\boldsymbol{R}}) = (\check{\boldsymbol{S}}, \check{\boldsymbol{L}}, \check{\boldsymbol{R}}) = (\bar{\boldsymbol{S}}, \bar{\boldsymbol{L}}, \bar{\boldsymbol{R}})$, which completes the proof of Theorem 4.1.

Proof of Step 1.

The first-order optimality condition for (D.1) is given by

$$\begin{aligned} \frac{1}{2}\left(\check{\boldsymbol{R}}\boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n\check{\boldsymbol{R}}\right) - \boldsymbol{I} &\in -\lambda_n\gamma\partial\|\check{\boldsymbol{S}}\|_1, \\ \frac{1}{2}\left(\check{\boldsymbol{R}}\boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n\check{\boldsymbol{R}}\right) - \boldsymbol{I} &\in \lambda_n\partial\|\check{\boldsymbol{L}}\|_*. \end{aligned}$$

Similarly, the first-order optimality condition for the constrained problem (D.2) is given by

$$\begin{array}{l} \frac{1}{2}\left(\bar{\boldsymbol{R}}\boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n\bar{\boldsymbol{R}}\right) - \boldsymbol{I} + \boldsymbol{Q}_{(\Omega^*)^\perp} \in -\lambda_n\gamma\partial\|\bar{\boldsymbol{S}}\|_1, \\ \frac{1}{2}\left(\bar{\boldsymbol{R}}\boldsymbol{\Sigma}_n + \boldsymbol{\Sigma}_n\bar{\boldsymbol{R}}\right) - \boldsymbol{I} + \boldsymbol{Q}_{(T^*)^\perp} \in \lambda_n\partial\|\bar{\boldsymbol{L}}\|_*, \end{array} \tag{D.3}$$

where $\boldsymbol{Q}_{(\Omega^*)^\perp}$ and $\boldsymbol{Q}_{(T^*)^\perp}$ are Lagrange multipliers such that $\boldsymbol{Q}_{(\Omega^*)^\perp} \in (\Omega^*)^\perp$ and $\boldsymbol{Q}_{(T^*)^\perp} \in (T^*)^\perp$. Recall that

$$h_\Sigma(\boldsymbol{R}) = \frac{1}{2}(\boldsymbol{\Sigma R} + \boldsymbol{R\Sigma}).$$

Restricting (D.3) to the space $\mathcal{Y}^*$, we have

$$\begin{aligned}\mathcal{P}_{\Omega^*}\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right) &= -\lambda_n\gamma\,\mathrm{sign}(\bar{\boldsymbol{S}}),\\ \mathcal{P}_{T^*}\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right) &= \lambda_n\bar{\boldsymbol{U}}\bar{\boldsymbol{V}}^T,\end{aligned}$$

where $\bar{\boldsymbol{L}} = \bar{\boldsymbol{U}}\mathrm{diag}(\bar{\boldsymbol{\sigma}})\bar{\boldsymbol{V}}^T$ is the SVD of $\bar{\boldsymbol{L}}$. Let $\mathcal{A}$ and $\mathcal{A}^\dagger$ be the addition and adjoint of the addition operator, that is, for matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have

$$\mathcal{A}(\boldsymbol{A}, \boldsymbol{B}) = \boldsymbol{A} + \boldsymbol{B} \quad \text{and} \quad \mathcal{A}^\dagger(\boldsymbol{A}) = (\boldsymbol{A}, \boldsymbol{A}).$$

Further denote by $\mathcal{P}_{\mathcal{Y}}$ the projection onto $\mathcal{Y}$. Setting

$$\boldsymbol{Z} = \mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right),$$

we then have $g_\gamma(\boldsymbol{Z}) = \lambda_n$. Since $(\bar{\boldsymbol{S}}, \bar{\boldsymbol{L}}, \bar{\boldsymbol{R}})$ satisfies the optimality condition for (D.1) on $\mathcal{Y}^*$, we need to show that

$$g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right)\right) < \lambda_n.$$

Let $\boldsymbol{E}_n = \boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}^*, \boldsymbol{\Delta}_S = \bar{\boldsymbol{S}} - \boldsymbol{S}^*$ and $\boldsymbol{\Delta}_L = \boldsymbol{L}^* - \bar{\boldsymbol{L}}$. Rewriting $h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}$ in terms of $\boldsymbol{E}_n, \boldsymbol{\Delta}_S$ and $\boldsymbol{\Delta}_L$, we obtain

$$\begin{aligned}h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I} &= \frac{1}{2}\left((\boldsymbol{\Sigma}^* + \boldsymbol{E}_n)(\boldsymbol{R}^* + \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) + (\boldsymbol{R}^*\right.\\ &\quad \left. + \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)(\boldsymbol{\Sigma}^* + \boldsymbol{E}_n)\right) - \boldsymbol{\Sigma}^*\boldsymbol{R}^*\\ &= h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) + h_{E_n}(\boldsymbol{R}^*).\end{aligned}$$

Therefore, we need to prove

$$\begin{aligned}g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right)\right) &\le g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\right)\\ &\quad + g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{E_n}(\boldsymbol{R}^*)\right)\\ &= T_1 + T_2 < \lambda_n.\end{aligned}$$

Let $m = \max\{1/\gamma, 1\}$, we first note that term T_2 can be bounded as

$$g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{E_n}(\boldsymbol{R}^*)\right) \le g_\gamma\left(\mathcal{A}^\dagger h_{E_n}(\boldsymbol{R}^*)\right) \le m\|\boldsymbol{E}_n\boldsymbol{R}^*\|_2 \le \frac{mC_K}{\psi_1}\sqrt{\frac{p}{n}},$$

where the last inequality follows from $\psi_1 \le \lambda_{\min}(\boldsymbol{\Sigma}_Y)$ and Lemma D.1. To bound term T_1, we first show that under the irrepresentability condition (4.2), we have

$$g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\right) \le (1-\delta)g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\right). \tag{D.4}$$

For $\boldsymbol{S} \in \Omega^*, \boldsymbol{L} \in T^*$ with $\|\boldsymbol{S}\|_\infty = \gamma$ and $\|\boldsymbol{L}\|_2 = 1$, it follows from definitions of α, β, γ and $\xi(T^*)$ that

$$\begin{aligned}\|\mathcal{P}_{\Omega^*} h_{\Sigma^*}(\boldsymbol{S}+\boldsymbol{L})\|_\infty &\geq \|\mathcal{P}_{\Omega^*} h_{\Sigma^*}(\boldsymbol{S})\|_\infty - \|\mathcal{P}_{\Omega^*} h_{\Sigma^*}(\boldsymbol{L})\|_\infty \\ &\geq \alpha\gamma - \|h_{\Sigma^*}(\boldsymbol{L})\|_\infty \\ &\geq \alpha\gamma - \beta\xi(T^*).\end{aligned}$$

Similarly, we have

$$\begin{aligned}\|\mathcal{P}_{T^*} h_{\Sigma^*}(\boldsymbol{S}+\boldsymbol{L})\|_2 &\geq \|\mathcal{P}_{T^*} h_{\Sigma^*}(\boldsymbol{L})\|_2 - \|\mathcal{P}_{T^*} h_{\Sigma^*}(\boldsymbol{S})\|_2 \\ &\geq \alpha - \|h_{\Sigma^*}(\boldsymbol{S})\|_2 \\ &\geq \alpha - 2\beta\gamma\mu(\Omega^*).\end{aligned}$$

Combining these results yields

$$g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S},\boldsymbol{L})\right) \geq \alpha - \beta\max\left\{\frac{\xi(T^*)}{\gamma}, 2\gamma\mu(\Omega^*)\right\}. \tag{D.5}$$

To upper-bound $g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S},\boldsymbol{L})\right)$, we have $g_\gamma\left(\mathcal{P}_{\mathcal{Y}^{*\perp}}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S},\boldsymbol{L})\right)$ for $\boldsymbol{S} \in \Omega^*, \boldsymbol{L} \in T^*$ with $\|\boldsymbol{S}\|_\infty = \gamma$ and $\|\boldsymbol{L}\|_2 = 1$, we have

$$\|\mathcal{P}_{(\Omega^*)^\perp} h_{\Sigma^*}(\boldsymbol{S}+\boldsymbol{L})\|_\infty \leq \|\mathcal{P}_{(\Omega^*)^\perp} h_{\Sigma^*}(\boldsymbol{S})\|_\infty + \|\mathcal{P}_{(\Omega^*)^\perp} h_{\Sigma^*}(\boldsymbol{L})\|_\infty \leq \delta\gamma + \beta\xi(T^*),$$

and

$$\|\mathcal{P}_{(T^*)^\perp} h_{\Sigma^*}(\boldsymbol{S}+\boldsymbol{L})\|_2 \leq \|\mathcal{P}_{(T^*)^\perp} h_{\Sigma^*}(\boldsymbol{L})\|_2 + \|\mathcal{P}_{(T^*)^\perp} h_{\Sigma^*}(\boldsymbol{S})\|_2 \leq \delta + \beta\mu(\Omega^*)\gamma.$$

Therefore, we obtain

$$g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S},\boldsymbol{L})\right) \leq \delta + \beta\max\left\{\frac{\xi(T^*)}{\gamma}, \gamma\mu(\Omega^*)\right\}. \tag{D.6}$$

Note that if we choose γ such that

$$\gamma \in \left[\frac{\xi(T^*)\beta(2-\nu)}{\nu\alpha}, \frac{\nu\alpha}{2\mu(\Omega^*)\beta(2-\nu)}\right],$$

we have

$$\max\left\{\frac{\xi(T^*)}{\gamma}, 2\gamma\mu(\Omega^*)\right\} \leq \frac{\nu\alpha}{\beta(2-\nu)}. \tag{D.7}$$

This observation together with (D.5) and (D.6) implies

$$\begin{aligned}\frac{g_\gamma\left(\mathcal{P}_{\mathcal{Y}^{*\perp}}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S,\boldsymbol{\Delta}_L)\right)}{g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S,\boldsymbol{\Delta}_L)\right)} &\leq \frac{\delta + \beta\max\left\{\frac{\xi(T^*)}{\gamma}, 2\gamma\mu(\Omega^*)\right\}}{\alpha + \beta\max\left\{\frac{\xi(T^*)}{\gamma}, 2\gamma\mu(\Omega^*)\right\}} \\ &\leq \frac{\frac{\delta}{\alpha} + \frac{\nu}{2-\nu}}{1 + \frac{\nu}{2-\nu}}\end{aligned}$$

$$\leq \frac{(1-2\nu)(2-\nu)+\nu}{2}$$
$$< 1-\nu,$$

which completes the proof of (D.4). Finally, recall that $\boldsymbol{Z} = \mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger\left(h_{\Sigma_n}(\bar{\boldsymbol{R}}) - \boldsymbol{I}\right)$ and $g_\gamma(\boldsymbol{Z}) = \lambda_n$, we obtain

$$\begin{aligned} g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\right) &= g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger(\boldsymbol{Z} - h_{E_n}(\boldsymbol{R}^*)\right) \\ &\leq g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger(\boldsymbol{Z})\right) + g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger(h_{E_n}(\boldsymbol{R}^*)\right) \\ &\leq \lambda_n + \frac{2mC_K}{\psi_1}\sqrt{\frac{p}{n}} \end{aligned}$$

Combining these bounds together yields

$$\begin{aligned} g_\gamma\left(\mathcal{P}_{(\mathcal{Y}^*)^\perp}\mathcal{A}^\dagger\left(h_{\Sigma_n}(\hat{\boldsymbol{R}}_{\mathcal{Y}^*}) - \boldsymbol{I}\right)\right) &\leq (1-\nu)\left(\lambda_n + \frac{2mC_K}{\psi_1}\sqrt{\frac{p}{n}}\right) + \frac{mC_K}{\psi_1}\sqrt{\frac{p}{n}} \\ &\leq (1-\nu/2)\lambda_n \end{aligned}$$

as long as we choose λ_n as

$$\lambda_n = \frac{(3-2\nu)mC_K}{\psi_1}\sqrt{\frac{p}{n}},$$

which completes the proof of *Step 1*.

Proof of Step 2.

We investigate estimation performance of the solution to the constrained problem (D.2). Since

$$g_\gamma(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) = g_\gamma(\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)),$$

we aim to control the last term using $g_\gamma(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L))$. For $\boldsymbol{S} \in \Omega^*, \boldsymbol{L} \in T^*$ with $\|\boldsymbol{S}\|_\infty = \gamma$ and $\|\boldsymbol{L}\|_2 = 1$, we have already showed that

$$\begin{aligned} g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S}, \boldsymbol{L})\right) &\geq \alpha - \beta\max\left\{\frac{\xi(T^*)}{\gamma}, 2\gamma\mu(\Omega^*)\right\} \\ &\geq \frac{2\alpha}{3} = \frac{2\alpha}{3}g_\gamma(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L), \end{aligned}$$

where we have used (D.7) and $\nu < \frac{1}{2}$ in the last inequality. Therefore, we obtain

$$\begin{aligned} \max\left\{\frac{\|\boldsymbol{\Delta}_S\|_\infty}{\gamma}, \|\boldsymbol{\Delta}_L\|_2\right\} &= g_\gamma(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) \\ &\leq \frac{3}{2\alpha}g_\gamma\left(\mathcal{P}_{\mathcal{Y}^*}\mathcal{A}^\dagger h_{\Sigma^*}\mathcal{A}\mathcal{P}_{\mathcal{Y}^*}(\boldsymbol{S}, \boldsymbol{L})\right) \\ &\leq \frac{3}{2\alpha}(\lambda_n + 2m\|\boldsymbol{E}_n\|_2/\psi_1) \\ &\leq \frac{3}{\alpha}\lambda_n, \end{aligned}$$

which establishes the estimation error bound. Further, due to additional constraints, we must have

$$\bar{\boldsymbol{S}} \in \Omega^* \quad \text{and} \quad \bar{\boldsymbol{L}} \in T^*.$$

Hence if the minimum nonzero singular value of $\boldsymbol{L}^*$, i.e., σ, satisfies

$$\sigma > \frac{3}{\alpha}\lambda_n,$$

we have $\bar{\boldsymbol{L}} \succeq \boldsymbol{0}$. Besides, since for a symmetric matrix $\boldsymbol{A}$, we have $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_{\ell_\infty} \leq \deg(\boldsymbol{A})\|\boldsymbol{A}\|_\infty$, which implies

$$\begin{aligned}\|\bar{\boldsymbol{R}} - \boldsymbol{R}^*\|_2 &\leq \|\boldsymbol{\Delta}_S\|_2 + \|\boldsymbol{\Delta}_L\|_2 \\ &\leq \frac{3}{\alpha}(d\gamma + 1)\lambda_n.\end{aligned}$$

Hence if the minimum eigenvalue of $\boldsymbol{R}^* = (\boldsymbol{\Sigma}_Y)^{-1}$ satisfies

$$\frac{1}{\psi_2} > \frac{3}{\alpha}(d\gamma + 1)\lambda_n,$$

we have $\bar{\boldsymbol{R}} \succ \boldsymbol{0}$. These results combining with those proved in *Step 1* show that solutions to optimizations problems (2.5), (D.1) and (D.2) are the same. Finally, using the assumption that

$$\theta > \frac{3\gamma}{\alpha}\lambda_n,$$

we obtain model selection consistency

$$\operatorname{sign}(\hat{\boldsymbol{S}}) = \operatorname{sign}(\boldsymbol{S}^*)$$

as claimed. □

Acknowledgments

The authors are grateful to the editor, the associate editor, and two anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (Nos. 31171262, 31428012, 31471246), and the National Key Basic Research Project of China (No. 2015CB910303).

References

Agakov, F. V., Orchard, P., and Storkey, A. J. (2012). Discriminative mixtures of sparse latent fields for risk management. In *International Conference on Artificial Intelligence and Statistics*, pages 10–18.

Agarwal, A., Negahban, S., and Wainwright, M. J. (2012). Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197. MR2985947

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122.

Brem, R. B. and Kruglyak, L. (2005). The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences*, 102(5):1572–1577.

Cai, T. T., Li, H., Liu, W., and Xie, J. (2013). Covariate-adjusted precision matrix estimation with an application in genetical genomics. *Biometrika*, 100(1):139–156. MR3034329

Cai, T. T., Liu, W., and Luo, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*, 106(494):594–607. MR2847973

Cai, T. T., Ren, Z., and Zhou, H. H. (2016). Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electronic Journal of Statistics*, 10(1):1–59. MR3466172

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37. MR2811000

Candès, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351. MR2382644

Candès, E. J. and Tao, T. (2010). The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080. MR2723472

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967. MR3059067

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. (2011). Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596. MR2817479

Chen, M., Ren, Z., Zhao, H., and Zhou, H. (2016). Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. *Journal of the American Statistical Association*, 111(513):394–406. MR3494667

Chen, S., Witten, D. M., and Shojaie, A. (2015). Selection and estimation for mixed graphical models. *Biometrika*, 102(1):47–64. MR3335095

Cheng, J., Li, T., Levina, E., and Zhu, J. (2016). High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378. MR3640193

Cheung, V. G. and Spielman, R. S. (2002). The genetics of variation in gene expression. *Nature Genetics*, 32:522–525.

Cox, D. R. and Wermuth, N. (1996). *Multivariate Dependencies: Models, Analysis and Interpretation*, volume 67. CRC Press. MR1456990

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306. MR2241189

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4):603–680. MR3091653

Fan, J., Liu, H., Ning, Y., and Zou, H. (2017). High dimensional semiparametric latent graphical model for mixed data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(2):405–421. MR3611752

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, 9(3):432–441.

Giraud, C. and Tsybakov, A. (2012). Discussion: latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1984–1988. MR3059071

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. Cambridge University Press. MR2978290

Hsu, D., Kakade, S. M., and Zhang, T. (2011). Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234. MR2883652

Janková, J. and van de Geer, S. (2015). Confidence intervals for high-dimensional inverse covariance estimation. *Electronic Journal of Statistics*, 9:1205–1229. MR3354336

Kalaitzis, A. and Lawrence, N. (2012). Residual component analysis: generalising PCA for more flexible inference in linear-Gaussian models. *arXiv preprint arXiv:1206.4560*.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(suppl 1):D355–D360.

Kelder, T., van Iersel, M. P., Hanspers, K., Kutmon, M., Conklin, B. R., Evelo, C. T., and Pico, A. R. (2012). Wikipathways: building research communities on biological pathways. *Nucleic Acids Research*, 40(D1):D1301–D1307.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278. MR2572459

Lauritzen, S. and Meinshausen, N. (2012). Discussion: latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1973–1977. MR3059069

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press. MR1419991

Li, B., Chun, H., and Zhao, H. (2012). Sparse estimation of conditional graphical models with application to gene networks. *Journal of the American Statistical Association*, 107(497):152–167. MR2949348

Li, H. and Gui, J. (2006). Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks. *Biostatistics*, 7(2):302–317.

Liu, W. (2013). Gaussian graphical model estimation with false discovery rate control. *The Annals of Statistics*, 41(6):2948–2978. MR3161453

Liu, W. and Luo, X. (2015). Fast and adaptive sparse precision matrix estimation in high dimensions. *Journal of Multivariate Analysis*, 135:153–162. MR3306432

Loh, P.-L. and Wainwright, M. J. (2013). Structure estimation for discrete graphical models: generalized covariance matrices and their inverses. *The Annals of Statistics*, 41(6):3022–3049. MR3161456

Ma, S., Xue, L., and Zou, H. (2013). Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Computation*, 25(8):2172–2198. MR3100000

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, 34(3):1436–1462. MR2278363

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473. MR2758523

Meng, Z., Eriksson, B., and Hero III, A. O. (2014). Learning latent variable Gaussian graphical models. *arXiv preprint arXiv:1406.2721*.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. (2010). High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319. MR2662343

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. (2011). High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980. MR2836766

Ren, Z., Sun, T., Zhang, C.-H., and Zhou, H. H. (2015). Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *The Annals of Statistics*, 43(3):991–1026. MR3346695

Ren, Z. and Zhou, H. H. (2012). Discussion: latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1989–1996. MR3059072

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515. MR2417391

Segal, E., Friedman, N., Kaminski, N., Regev, A., and Koller, D. (2005). From signatures to models: understanding cancer using microarrays. *Nature Genetics*, 37:S38–S45.

Städler, N. and Bühlmann, P. (2012). Missing values: sparse inverse covariance estimation and an extension to sparse regression. *Statistics and Computing*, 22(1):219–235. MR2865066

Sun, T. and Zhang, C.-H. (2013). Sparse matrix inversion with scaled Lasso. *The Journal of Machine Learning Research*, 14(1):3385–3418. MR3144466

Taeb, A. and Chandrasekaran, V. (2016). Interpreting latent variables in factor models via convex optimization. *arXiv preprint arXiv:1601.00389*.

Tan, K. M., Ning, Y., Witten, D. M., and Liu, H. (2016). Replicates in high dimensions, with applications to latent variable graphical models. *Biometrika*, 103(4):761–777. MR3620438

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*. MR2963170

Vershynin, R. (2012). How close is the sample covariance matrix to the actual covariance matrix? *Journal of Theoretical Probability*, 25(3):655–686. MR2956207

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55(5):2183–2202. MR2729873

Wasserman, L., Kolar, M., and Rinaldo, A. (2014). Berry-Esseen bounds for estimating undirected graphs. *Electronic Journal of Statistics*, 8(1):1188–1224. MR3263117

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534.

Xu, P., Ma, J., and Gu, Q. (2017). Speeding up latent variable Gaussian graphical model estimation via nonconvex optimizations. *arXiv preprint arXiv:1702.08651*.

Yang, E., Ravikumar, P., Allen, G. I., and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *Journal of Machine Learning Research*, 16(1):3813–3847. MR3450553

Yuan, M. (2010). High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 11(Aug):2261–2286. MR2719856

Yuan, M. (2012). Discussion: latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1968–1972. MR3059068

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35. MR2367824

Zhang, T. and Zou, H. (2014). Sparse precision matrix estimation via Lasso penalized D-trace loss. *Biometrika*, 101(1):103–120. MR3180660

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7(Nov):2541–2563. MR2274449