TIGER: A tuning-insensitive approach for optimally estimating Gaussian graphical models

Han Liu[*]

*Department of Operations Research and Financial Engineering, Princeton University,
Princeton, NJ 08544, USA
e-mail:* hanliu@princeton.edu

and

Lie Wang[†]

*Department of Mathematics, Massachusetts Institute of Technology,
Cambridge, MA 02139, USA
e-mail:* liewang@math.mit.edu

Abstract: We propose a new procedure for optimally estimating high dimensional Gaussian graphical models. Our approach is asymptotically tuning-free and non-asymptotically tuning-insensitive: It requires very little effort to choose the tuning parameter in finite sample settings. Computationally, our procedure is significantly faster than existing methods due to its tuning-insensitive property. Theoretically, the obtained estimator simultaneously achieves minimax lower bounds for precision matrix estimation under different norms. Empirically, we illustrate the advantages of the proposed method using simulated and real examples. The R package `camel` implementing the proposed methods is also available on the Comprehensive R Archive Network.

1. Introduction

We consider the problem of learning high dimensional Gaussian graphical models: let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n data points from a d-dimensional random vector $\boldsymbol{X} = (X_1, ..., X_d)^T$ with $\boldsymbol{X} \sim N_d(\boldsymbol{0}, \boldsymbol{\Sigma})$. We want to estimate an undirected graph denoted by $G = (V, E)$, where V contains nodes corresponding to the d variables in $\boldsymbol{X}$, and the edge set E describes the conditional independence relationships between $X_1, ..., X_d$. Let $\boldsymbol{X}_{\setminus \{j,k\}} := \{X_\ell : \ell \neq j, k\}$. We say the joint distribution of $\boldsymbol{X}$ is Markov to G if X_j is independent of X_k given $\boldsymbol{X}_{\setminus \{j,k\}}$ for all $(j, k) \notin E$. For Gaussian distributions, the graph G is encoded by the precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$. More specifically, there is no edge connects X_j and

X_k if and only if $\Theta_{jk} = 0$. The graph estimation problem is then reduced to the estimation of the precision matrix Θ. This problem is also called *covariance selection* (Dempster, 1972).

In low dimensions where $d < n$, Drton and Perlman (2007, 2008) develop a multiple testing procedure for identifying the sparsity pattern of the precision matrix. In high dimensions where $d \gg n$, Meinshausen and Bühlmann (2006) propose a neighborhood pursuit approach for estimating Gaussian graphical models by solving a collection of sparse regression problems using the Lasso (Tibshirani, 1996; Chen et al., 1998). This approach can be viewed as a pseudo-likelihood approximation of the full likelihood. A related approach is to directly estimate Θ by penalizing the likelihood using the L_1-penalty (Banerjee et al., 2008; Yuan and Lin, 2007; Friedman et al., 2008). To further reduce the estimation bias, Lam and Fan (2009); Jalali et al. (2012); Shen et al. (2012) propose either greedy algorithms or non-convex penalties for sparse precision matrix estimation. Under certain conditions, Ravikumar et al. (2011); Rothman et al. (2008) study the theoretical properties of the penalized likelihood methods. Yuan (2010) and Cai et al. (2011) also propose the graphical Dantzig selector and CLIME respectively, which can be solved by linear programming and are more amenable to theoretical analysis than the penalized likelihood approach. More recently, Liu and Luo (2015) and Sun and Zhang (2013) propose the SCIO and scaled-Lasso methods, which estimate the sparse precision matrix in a column-by-column fashion and have good theoretical properties.

Besides Gaussian graphical models, Liu et al. (2012) propose the nonparanormal model family. Instead of assuming X to be Gaussian, they assume there exists a set of monotone functions $f_1, \ldots, f_d$, such that the transformed data $f(X) := (f_1(X_1), \ldots, f_d(X_d))^T$ is Gaussian. More details can be found in Liu et al. (2012) and Lafferty et al. (2012). Zhao et al. (2012) developed a scalable software package to implement the nonparanormal algorithms. Other nonparametric graph estimation methods include forest graphical models (Liu et al., 2011) or conditional graphical models (Liu et al., 2010a).

Most of these methods require choosing some tuning parameters that control the bias-variance tradeoff. Theoretical justifications of these methods are usually built upon some oracle choices of tuning parameters that cannot be implemented in practice. It remains an challenging problem on choosing the regularization parameter in a data-dependent way. Popular techniques include the C_p-statistic (Mallows, 1973), AIC (Akaike information criterion, Akaike (1973)), BIC (Bayesian information criterion, Schwarz (1978)), extended BIC (Chen and Chen, 2008, 2012; Foygel and Drton, 2010), RIC (Risk inflation criterion, Foster and George (1994)), cross validation (Efron, 1982), and covariance penalization (Efron, 2004). Most of these methods require data splitting and have been only justified for low dimensional settings. Significant progress has been made recently on developing likelihood-free regularization selection techniques, including permutation methods (Wu et al., 2007; Boos et al., 2009; Lysen, 2009) and subsampling methods (Lange et al., 2004; Ben-david et al., 2006; Meinshausen and Bühlmann, 2010; Bach, 2008). Meinshausen and Bühlmann (2010) and Bach (2008) and Liu et al. (2010b) also propose to select the tuning parameters using

subsampling. However, these subsampling based methods are computationally expensive and are still lack of theoretical guarantees.

In this paper we propose a new procedure for estimating high dimensional Gaussian graphical models. Our method, named TIGER (Tuning-Insensitive Graph Estimation and Regression), owes a "tuning-insensitive property": it automatically adapts to the unknown sparsity pattern and is asymptotically tuning-free. In finite sample settings, we only need to pay very little effort to tune the regularization parameter. The main idea is to estimate the precision matrix Θ in a column-by-column fashion. For each column, the computation is reduced to a sparse regression problem. This idea has been adopted by many methods, including the neighborhood pursuit (Meinshausen and Bühlmann, 2006), graphical Dantzig selector (Yuan, 2010), CLIME (Cai et al., 2011), SCIO (Liu and Luo, 2015), and the scaled-Lasso method (Sun and Zhang, 2013). These methods differ from each other mainly by how they solve the sparse regression subproblem: the graphical Dantzig selector and CLIME use the Dantzig selector, the SCIO and neighborhood pursuit use the Lasso, while Sun and Zhang (2013) use the scaled-Lasso (Sun and Zhang, 2013). Unlike these methods, the TIGER solves this sparse regression problem using the SQRT-Lasso (Belloni et al., 2012). The main advantage of the TIGER over existing methods is its asymptotic tuning-free property, which allows us to use the entire dataset to efficiently learn and select the model. In contrast, it is well known that the cross-validation and subsampling methods are computationally expensive. Moreover, they may potentially waste valuable data which could otherwise be exploited to learn a better model (Bishop et al., 2003).

Another advantage of the TIGER is its computational simplicity and scalability for large datasets. For problems with large dimensionality d, the TIGER divides the whole problem into many subproblems, in each subproblem it estimates one column of the precision matrix by solving a simple SQRT-Lasso problem. The final matrix estimator is formed by combining the vector solutions into a matrix. This procedure can be solved in a parallel fashion and achieves a linear scale up with the number of CPU cores. An additional performance improvement comes from the tuning-insensitive property of the TIGER. Most existing methods exploit cross-validation to choose tuning parameters, which requires computing the solutions over a full regularization path. In contrast, the TIGER solves the SQRT-Lasso subproblem only for several fixed tuning parameters. For sparse problems, this is significantly faster than existing methods.

In the current paper, we prove theoretical guarantees of the TIGER esitmator $\hat{\Theta}$ of the true precision matrix Θ. In particular, let $\|\Theta\|_{\max} := \max_{jk} |\Theta_{jk}|$ and $\|\Theta\|_1 := \max_j \sum_k |\Theta_{jk}|$. Under the assumption that the condition number of Θ is bounded by a constant, we establish the elementwise sup-norm rate of convergence:

$$\big\|\hat{\Theta} - \Theta\big\|_{\max} = O_P\bigg(\|\Theta\|_1 \sqrt{\frac{\log d}{n}}\bigg). \tag{1.1}$$

If we further assume $\big\|\Theta\big\|_1 \le M_d$ where M_d may scale with d, the obtained

rate in (1.1) is minimax optimal over the model class consisting of precision matrices with bounded condition numbers. This result allows us to effectively conduct graph estimation without the need of very restrictive irrepresentable-type conditions.

Let $\|\boldsymbol{\Theta}\|_2$ be the largest eigenvalue of $\boldsymbol{\Theta}$ (i.e., $\|\boldsymbol{\Theta}\|_2$ is the spectral norm of $\boldsymbol{\Theta}$) and $k := \max_{i=1,\ldots,d} \sum_j I(\boldsymbol{\Theta}_{ij} \neq 0)$. We also prove that

$$\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_2 \le \left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_1 = O_P\left(k\|\boldsymbol{\Theta}\|_2\sqrt{\frac{\log d}{n}}\right). \tag{1.2}$$

Under the same condition that $\|\boldsymbol{\Theta}\|_1 \le M_d$ where M_d may scale with d, this spectral norm rate in (1.2) is also minimax optimal over the same model class as before.

Besides these theoretical results, we also establish a relationship between the SQRT-Lasso and the scaled-Lasso proposed by Sun and Zhang (2013). More specifically, the objective function of the scaled-Lasso can be viewed as a variational upper bound of the SQRT-Lasso. This relationship allows us to develop an efficient algorithm for the TIGER.

Computationally, the TIGER is significantly faster than existing methods since very few tunings are needed. In particular, we propose an iterative algorithm with initial values searched by the Alternating Direction Method of Multipliers (ADMM). For each reduced sparse regression subproblem, the computational complexity is of the same order as solving one single Lasso with a sparse solution. Empirically, we present thorough numerical results to compare the graph recovery and parameter estimation performance of our method with other approaches. Some real data experiments on gene expression datasets are also provided to back up our theory. The R package `camel` implementing the proposed methods is also available on the Comprehensive R Archive Network.

Notations. Let $\mathbf{v} := (v_1, \ldots, v_d)^T \in \mathbb{R}^d$ and $I(\cdot)$ be the indicator function, for $0 < q < \infty$, we define $\|\mathbf{v}\|_q := \left(\sum_{j=1}^d |v_j|^q\right)^{1/q}$, $\|\mathbf{v}\|_0 := \sum_{j=1}^d I(v_j \neq 0)$, and $\|\mathbf{v}\|_\infty := \max_j |v_j|$.

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be a symmetric matrix and $I, J \subset \{1, \ldots, d\}$ be two sets, we denote $\mathbf{A}_{I,J}$ to be the submatrix of $\mathbf{A}$ with rows and columns indexed by I and J. Let $\mathbf{A}_{*j}$ be the jth column of $\mathbf{A}$ and $\mathbf{A}_{*\backslash j}$ be the submatrix of $\mathbf{A}$ with the jth column $\mathbf{A}_{*j}$ removed.

We define the following matrix norms:

$$\|\mathbf{A}\|_q := \max_{\|v\|_q=1} \|\mathbf{A}\mathbf{v}\|_q \ \ , \|\mathbf{A}\|_{\max} := \max_{jk} |\mathbf{A}_{jk}|, \ \text{ and } \ \|\mathbf{A}\|_\mathsf{F} = \left(\sum_{j,k} |\mathbf{A}_{jk}|^2\right)^{1/2}.$$

It is easy to see that when $q = \infty$, $\|\mathbf{A}\|_\infty = \|\mathbf{A}\|_1$. We also denote $\Lambda_{\max}(\mathbf{A})$ and $\Lambda_{\min}(\mathbf{A})$ to be the largest and smallest eigenvalues of $\mathbf{A}$.

The rest of the paper is organized as follows. In Section 2, we introduce some background on Gaussian graphical models. In Section 3, we describe the TIGER estimator and its computational algorithm. In Section 4, we present the theoretical properties including the rates of convergence for parameter estimation and

graph recovery. We also provide further discussions on the connections and differences of our results with other related methods. In Section 5, we demonstrate the numerical performance of our methods through synthetic and real datasets. All proofs are delegated to the appendix.

2. Background on Gaussian graphical models

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be n data points from a d-dimensional Gaussian random vector $\mathbf{X} := (X_1, \ldots, X_d)^T \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$. We denote $\boldsymbol{X}_i := (X_{i1}, \ldots, X_{id})^T$. As has been discussed in the previous section, we define precision matrix to be $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$. In this section, we introduce some background on solving Gaussian graphical models in a column-by-column fashion.

Let $\boldsymbol{X} \sim N_d(\mathbf{0}, \boldsymbol{\Sigma})$, the conditional distribution of X_j given $\boldsymbol{X}_{\setminus j}$ satisfies

$$X_j \mid \boldsymbol{X}_{\setminus j} \sim N_{d-1}\left(\boldsymbol{\Sigma}_{\setminus j, j}(\boldsymbol{\Sigma}_{\setminus j, \setminus j})^{-1}\boldsymbol{X}_{\setminus j}\,,\; \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{\setminus j, j}(\boldsymbol{\Sigma}_{\setminus j, \setminus j})^{-1}\boldsymbol{\Sigma}_{\setminus j, j}\right).$$

Let $\boldsymbol{\alpha}_j := (\boldsymbol{\Sigma}_{\setminus j, \setminus j})^{-1}\boldsymbol{\Sigma}_{\setminus j, j} \in \mathbb{R}^{d-1}$ and $\sigma_j^2 := \boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{\setminus j, j}(\boldsymbol{\Sigma}_{\setminus j, \setminus j})^{-1}\boldsymbol{\Sigma}_{\setminus j, j}$. We have

$$X_j = \boldsymbol{\alpha}_j^T \boldsymbol{X}_{\setminus j} + \epsilon_j, \tag{2.1}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $\boldsymbol{X}_{\setminus j}$. By the block matrix inversion formula, we have

$$\boldsymbol{\Theta}_{jj} = (\mathrm{Var}(\epsilon_j))^{-1} = \sigma_j^{-2}, \tag{2.2}$$

$$\boldsymbol{\Theta}_{\setminus j, j} = -(\mathrm{Var}(\epsilon_j))^{-1}\boldsymbol{\alpha}_j = -\sigma_j^{-2}\boldsymbol{\alpha}_j. \tag{2.3}$$

Therefore, we can recover $\boldsymbol{\Theta}$ in a column by column manner by regressing X_j on $\boldsymbol{X}_{\setminus j}$ for $j = 1, 2, \cdots, d$. For example, let

$$\mathbf{X} := \begin{pmatrix} X_{11} & \cdots & X_{1d} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{nd} \end{pmatrix} \in \mathbb{R}^{n \times d}$$

be the data matrix. We denote by $\boldsymbol{\alpha}_j := (\alpha_{j1}, \ldots, \alpha_{j(d-1)})^T \in \mathbb{R}^{d-1}$. Meinshausen and Bühlmann (2006) propose to iteratively estimate each $\boldsymbol{\alpha}_j$ by solving the Lasso regression:

$$\hat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \frac{1}{2n}\big\|\mathbf{X}_{*j} - \mathbf{X}_{*\setminus j}\boldsymbol{\alpha}_j\big\|_2^2 + \lambda_j\big\|\boldsymbol{\alpha}_j\big\|_1,$$

where λ_j is a tuning parameter. Once $\hat{\boldsymbol{\alpha}}_j$ is given, we get the neighborhood edges by reading out the nonzero coefficients of $\boldsymbol{\alpha}_j$. The final graph estimate $\hat{G}$ is obtained by either the "AND" or "OR" rule on combining the neighborhoods for all the d nodes. However, the neighborhood pursuit method of Meinshausen and Bühlmann (2006) only estimates the graph G and does not estimate the inverse covariance matrix $\boldsymbol{\Theta}$.

Instead of using the Lasso estimator as in Meinshausen and Bühlmann (2006), Yuan (2010) proposes to estimate $\boldsymbol{\alpha}_j$ by solving the Dantzig selector:

$$\hat{\boldsymbol{\alpha}}_j = \operatorname*{argmin}_{\boldsymbol{\alpha}_j \in \mathbb{R}^{d-1}} \left\| \boldsymbol{\alpha}_j \right\|_1 \ \text{ subject to } \ \left\| \hat{\boldsymbol{\Sigma}}_{\backslash j, j} - \hat{\boldsymbol{\Sigma}}_{\backslash j, \backslash j} \boldsymbol{\alpha}_j \right\|_\infty \leq \gamma_j,$$

where $\hat{\boldsymbol{\Sigma}} := \mathbf{X}\mathbf{X}^T/n$ is the sample covariance matrix and γ_j is a tuning parameter. Once $\hat{\boldsymbol{\alpha}}_j$ is given, we can estimate σ_j^2 by

$$\hat{\sigma}_j^2 = \left[1 - 2\hat{\boldsymbol{\alpha}}_j^T \hat{\boldsymbol{\Sigma}}_{\backslash j, j} + \hat{\boldsymbol{\alpha}}_j^T \hat{\boldsymbol{\Sigma}}_{\backslash j, \backslash j} \hat{\boldsymbol{\alpha}}_j \right]^{-1}.$$

We then get the estimate $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ by plugging $\hat{\boldsymbol{\alpha}}_j$ and $\hat{\sigma}_j^2$ into (2.2) and (2.3). Yuan (2010) analyzes the L_1-norm error $\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\|_1$ and shows its minimax optimality over certain model space. However, no graph estimation result is provided for this approach.

In another work, Sun and Zhang (2013) propose to estimate $\boldsymbol{\alpha}_j$ and σ_j by solving a scaled-Lasso problem:

$$\hat{\boldsymbol{b}}_j, \hat{\sigma}_j = \operatorname*{argmin}_{\boldsymbol{b}=(b_1,\ldots,b_d)^T, \sigma} \left\{ \frac{\boldsymbol{b}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{b}}{2\sigma} + \frac{\sigma}{2} + \lambda \sum_{k=1}^d \hat{\boldsymbol{\Sigma}}_{kk} |b_k| \ \text{ subject to } b_j = -1 \right\}.$$

Once $\hat{\boldsymbol{b}}_j$ is obtained, $\boldsymbol{\alpha}_j = \hat{\boldsymbol{b}}_{\backslash j}$. Sun and Zhang (2013) analyze the spectral-norm rate of convergence of the obtained precision matrix estimator. They did not investigate the elementwise sup-norm and graph recovery performance. In the next section, we will show that the scaled-Lasso estimator is highly related to our proposed procedure and will discuss the relationship in more details.

To estimate both precision matrix $\boldsymbol{\Theta}$ and graph G, Cai et al. (2011) proposes the CLIME estimator, which directly estimates the jth column of $\boldsymbol{\Theta}$ by solving

$$\hat{\boldsymbol{\Theta}}_{*j} = \operatorname*{argmin}_{\boldsymbol{\Theta}_{*j}} \left\| \boldsymbol{\Theta}_{*j} \right\|_1 \ \text{ subject to } \ \left\| \hat{\boldsymbol{\Sigma}} \boldsymbol{\Theta}_{*j} - \mathbf{e}_j \right\|_\infty \leq \delta_j, \ \text{ for } j = 1, \ldots, d,$$

where $\mathbf{e}_j$ is the jth canonical vector and δ_j is a tuning parameter. Cai et al. (2011) show that this convex optimization can be formulated into a linear program. Once $\hat{\boldsymbol{\Theta}}$ is obtained, we use another tuning parameter τ to threshold $\boldsymbol{\Theta}$ to estimate the graph G. In a follow-up work, Liu and Luo (2015) propose the SCIO estimator, which estimates the jth column of $\boldsymbol{\Theta}$ by solving

$$\hat{\boldsymbol{\Theta}}_{*j} = \operatorname*{argmin}_{\boldsymbol{\Theta}_{*j}} \left\{ \frac{1}{2} \boldsymbol{\Theta}_{*j}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\Theta}_{*j} - \mathbf{e}_j^T \boldsymbol{\Theta}_{*j} + \lambda_j \left\| \boldsymbol{\Theta}_{*j} \right\|_1 \right\}.$$

The theory of most of these graph estimation methods builds upon some theoretical choices of tuning parameters that cannot be implemented in practice. For example, in the neighborhood pursuit method and the graphical Dantzig selector, the tuning parameter λ_j and γ_j depend on σ_j^2, which is unknown. Practically, we usually set $\lambda = \lambda_1 = \cdots = \lambda_d$ and $\gamma = \gamma_1 = \cdots = \gamma_d$ to reduce the number of tuning parameters. However, as we will illustrate in later sections,

such a choice makes the estimating procedure non-adaptive to inhomogeneous graphs. The tuning parameters of the CLIME and SCIO estimators depend on $\|\boldsymbol{\Theta}\|_1$, which is unknown. In general, these methods employ cross validation to conduct data-dependent tuning parameter selection and Liu and Luo (2015) provide some theoretical analysis of the cross-validation method. However, as we discussed before, cross-validation is computationally expensive and a waste of valuable training data.

3. Method

In this section we introduce the use of the SQRT-Lasso from Belloni et al. (2012) for simultaneously estimating the graph G and precision matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$.

The SQRT-Lasso is a penalized optimization algorithm for solving high dimensional linear regression problems. More specifically, for a linear model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^d$ is the vector of unknown coefficients, and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is the noise vector, the SQRT-Lasso estimates $\boldsymbol{\beta}$ by solving

$$\hat{\boldsymbol{\beta}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^d} \Big\{ \frac{1}{\sqrt{n}} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\},$$

where λ is the tuning parameter. It is shown in Belloni et al. (2012) that the choice of λ for the SQRT-Lasso method is asymptotically universal and does not depend on any unknown parameter. In contrast, most of other methods, including the Lasso and Dantzig selector, rely heavily on the unknown standard deviation of the noise. Moreover, the SQRT-Lasso method achieves near oracle performance for the estimation of $\boldsymbol{\beta}$.

3.1. TIGER for graph and precision matrix estimation

In the discussion of this section, we always condition on the observed data $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. Let $\hat{\boldsymbol{\Gamma}} := \operatorname{diag}(\hat{\boldsymbol{\Sigma}})$ be a d-dimensional diagonal matrix with the diagonal elements be the same as those in $\hat{\boldsymbol{\Sigma}}$. We define

$$\boldsymbol{Z} := (Z_1, \ldots, Z_d)^T = \boldsymbol{X} \hat{\boldsymbol{\Gamma}}^{-1/2}.$$

By (2.1), we have

$$Z_j \hat{\boldsymbol{\Gamma}}_{jj}^{1/2} = \boldsymbol{\alpha}_j^T \hat{\boldsymbol{\Gamma}}_{\backslash j, \backslash j}^{1/2} \boldsymbol{Z}_{\backslash j} + \epsilon_j, \tag{3.1}$$

Define $\boldsymbol{\beta}_j := \hat{\boldsymbol{\Gamma}}_{\backslash j, \backslash j}^{1/2} \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \boldsymbol{\alpha}_j$ and $\tau_j^2 = \sigma_j^2 \hat{\boldsymbol{\Gamma}}_{jj}^{-1}$, we then have

$$Z_j = \boldsymbol{\beta}_j^T \boldsymbol{Z}_{\backslash j} + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \epsilon_j. \tag{3.2}$$

We define $\hat{\mathbf{R}}$ to be the sample correlation matrix: $\hat{\mathbf{R}} := \big(\operatorname{diag}(\hat{\boldsymbol{\Sigma}})\big)^{-1/2} \times \hat{\boldsymbol{\Sigma}} \big(\operatorname{diag}(\hat{\boldsymbol{\Sigma}})\big)^{-1/2}$.

Motivated by the model in (3.2), we propose the following precision matrix estimator.

For $j = 1, \ldots, d$, we estimate the jth column of $\boldsymbol{\Theta}$ by solving :

$$\hat{\boldsymbol{\beta}}_j := \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{ \sqrt{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j} + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \qquad (3.3)$$

$$\hat{\tau}_j := \sqrt{1 - 2\hat{\boldsymbol{\beta}}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \hat{\boldsymbol{\beta}}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \hat{\boldsymbol{\beta}}_j},$$

$$\hat{\boldsymbol{\Theta}}_{jj} = \hat{\tau}_j^{-2} \hat{\boldsymbol{\Gamma}}_{jj}^{-1} \quad \text{and} \quad \hat{\boldsymbol{\Theta}}_{\backslash j, j} = -\hat{\tau}_j^{-2} \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \hat{\boldsymbol{\Gamma}}_{\backslash j, \backslash j}^{-1/2} \hat{\boldsymbol{\beta}}_j.$$

For the estimator in (3.3), λ is a tuning parameter. In the next section, we show that by choosing $\lambda = \pi \sqrt{\dfrac{\log d}{2n}}$, the obtained estimator achieves the optimal rates of convergence in the asymptotic setting. Therefore, our procedure is asymptotically tuning-parameter free. For finite samples, we set

$$\lambda := \zeta \pi \sqrt{\frac{\log d}{2n}}, \qquad (3.4)$$

and ζ can be chosen from a range $[\sqrt{2}/\pi, 1]$. Since the choice of ζ does not depend on any unknown parameters or quantities, we call the procedure *tuning-insensitive*. Empirically, we found that in most cases we encountered, simply setting $\zeta = \sqrt{2}/\pi$ works well in finite sample settings. More details can be found in the simulation section.

If a symmetric precision matrix estimate is preferred, we could conduct post-processing step as in Cai et al. (2011), which does not affect the theoretical analysis.

Let $\mathbf{Z} \in \mathbb{R}^{n \times d}$ be the normalized data matrix, i.e., $\mathbf{Z}_{*j} = \mathbf{X}_{*j} \boldsymbol{\Sigma}_{jj}^{-1/2}$ for $j = 1, \ldots, d$. An equivalent form of (3.3) is

$$\hat{\boldsymbol{\beta}}_j = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{ \frac{1}{\sqrt{n}} \|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j\|_2 + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \qquad (3.5)$$

$$\hat{\tau}_j = \frac{1}{\sqrt{n}} \|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \hat{\boldsymbol{\beta}}_j\|_2. \qquad (3.6)$$

Once $\hat{\boldsymbol{\Theta}}$ is obtained, the estimated graph $\hat{G} := (V, \hat{E})$ where $(j, k) \in \hat{E}$ if $\hat{\boldsymbol{\Theta}}_{jk} \hat{\boldsymbol{\Theta}}_{kj} \neq 0$.

3.2. Relationship with the scaled-Lasso estimator

In this subsection, we show that the scaled-Lasso from Sun and Zhang (2013) can be viewed as a variational upper bound of the objective function of the SQRT-Lasso and they are solving the same problem[1].

---

[1] This relationship is proposed and kindly provided by Cun-hui Zhang.

More specifically, we consider the following optimization:

$$\hat{\boldsymbol{\beta}}_j := \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}, \tau_j \geq 0}{\operatorname{argmin}} \left\{ \frac{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j}{2\tau_j} + \frac{\tau_j}{2} + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \quad (3.7)$$

Proposition 3.1. *The problems in* (3.3) *and* (3.7) *have the same solution* $\hat{\boldsymbol{\beta}}_j$.

Proof. For any $a, b > 0$, we have $a^2 + b^2 \geq 2ab$ and the equality is attained if and only $a = b$. Therefore, we have

$$\frac{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j}{2\tau_j} + \frac{\tau_j}{2} \geq \sqrt{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j}.$$

This shows that the objective function in (3.7) is an upper bound of the objective function in (3.3). The equality is attained if and only if

$$\tau_j = \sqrt{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j}.$$

This finishes the proof. □

This relationship between the TIGER and scaled-Lasso provides an efficient algorithm as described in the next subsection.

3.3. Computational algorithm

Equation (3.3) is jointly convex with respect to $\boldsymbol{\beta}_j$ and τ_j and can be solved by a coordinate-descent procedure. In the tth iteration, for a given $\tau_j^{(t)}$, we first solve the subproblem

$$\boldsymbol{\beta}_j^{(t+1)} := \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{ \frac{1 - 2\boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, j} + \boldsymbol{\beta}_j^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \boldsymbol{\beta}_j}{2\tau_j^{(t)}} + \lambda \|\boldsymbol{\beta}_j\|_1 \right\}, \quad (3.8)$$

This is a Lasso problem and can be efficiently solved by the coordinate descent algorithm developed by Friedman et al. (2007). Once $\boldsymbol{\beta}_j^{(t+1)}$ is obtained, we can calculate $\tau_j^{(t+1)}$ as

$$\tau_j^{(t+1)} = \sqrt{1 - 2\big(\boldsymbol{\beta}_j^{(t+1)}\big)^T \hat{\mathbf{R}}_{\backslash j, j} + \big(\boldsymbol{\beta}_j^{(t+1)}\big)^T \hat{\mathbf{R}}_{\backslash j, \backslash j} \big(\boldsymbol{\beta}_j^{(t+1)}\big)}. \quad (3.9)$$

We iterate these two steps until the algorithm converges.

On thing to note is that the above algorithm converges fast if a good initial value of τ_j is provided. For example, if τ_j is close to $\hat{\tau}_j$, the algorithm converges in only 3 to 5 iterations. In fact, with a good initial value of τ_j, the computation is roughly the same as running one single tuning parameter of the Lasso with a sparse solution. However, with an inappropriate initial value of τ_j, the computational complexity can be as heavy as calculating the full regularization path of a Lasso problem, which is inefficient.

To obtain a good initial estimate of τ_j, we propose an augmented Lagrange method as follows: We first reparameterize (3.5) and (3.6) as

$$\hat{\boldsymbol{\beta}}_j, \hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}, \boldsymbol{\gamma} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{\sqrt{n}} \|\boldsymbol{\gamma}\|_2 + \lambda \|\boldsymbol{\beta}_j\|_1 \ \ \text{subject to } \boldsymbol{\gamma} = \mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j \right\},$$

(3.10)

$$\hat{\tau}_j = \frac{1}{\sqrt{n}} \|\hat{\boldsymbol{\gamma}}\|_2.$$

(3.11)

We consider the following augmented Lagrangian function

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\beta}_j, \boldsymbol{\gamma}, \boldsymbol{\alpha}) \ := \ & \frac{1}{\sqrt{n}} \|\boldsymbol{\gamma}\|_2 + \lambda \|\boldsymbol{\beta}_j\|_1 + \rho \langle \boldsymbol{\alpha}, \boldsymbol{\gamma} - \mathbf{Z}_{*j} + \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j \rangle \\
& + \frac{\rho}{2} \|\boldsymbol{\gamma} - \mathbf{Z}_{*j} + \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j\|_2^2
\end{aligned}
$$

where $\rho > 0$ is the penalty parameter for the violation of the linear constraints. For simplicity, throughout this paper we assume that $\rho > 0$ is fixed (In practical implementations, we simply set $\rho = 1$). $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the Lagrange multiplier vector but rescaled by ρ for computational and notational convenience. This reparametrization decouples the computational dependency in the optimization problem. Therefore a complicated problem can be split into multiple simpler sub-problems, each of which can be solved easily.

The augmented Lagrangian method works in an iterative fashion. Suppose we have the solution $\boldsymbol{\beta}_j^{(t)}, \boldsymbol{\gamma}^{(t)}, \boldsymbol{\alpha}^{(t)}$ at the t-th iteration, the algorithm proceeds as follows:

Step 1. Update $\boldsymbol{\beta}_j$ by

$$\boldsymbol{\beta}_j^{(t+1)} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{ \lambda \|\boldsymbol{\beta}_j\|_1 + \frac{\rho}{2} \|\boldsymbol{\alpha}^{(t)} + \boldsymbol{\gamma}^{(t)} - \mathbf{Z}_{*j} + \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j\|_2^2 \right\}.$$

(3.12)

Let $\boldsymbol{u}^{(t)} := \mathbf{Z}_{*j} - \boldsymbol{\alpha}^{(t)} - \boldsymbol{\gamma}^{(t)}$. and $\lambda_\rho := \lambda/\rho$. The problem in (3.12) is equivalent to

$$\boldsymbol{\beta}_j^{(t+1)} = \underset{\boldsymbol{\beta}_j \in \mathbb{R}^{d-1}}{\operatorname{argmin}} \left\{ \lambda_\rho \|\boldsymbol{\beta}_j\|_1 + \frac{1}{2} \|\boldsymbol{u}^{(t)} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j\|_2^2 \right\}.$$

This is a Lasso subproblem which can be efficiently solved by the coordinate descent algorithm (Friedman et al., 2007).

Step 2. Given $\boldsymbol{\beta}_j^{(t+1)}$, we then update $\boldsymbol{\gamma}$ by

$$\boldsymbol{\gamma}^{(t+1)} = \underset{\boldsymbol{\gamma} \in \mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{\sqrt{n}} \|\boldsymbol{\gamma}\|_2 + \frac{\rho}{2} \|\boldsymbol{\gamma} - \mathbf{Z}_{*j} + \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j^{(t+1)} + \boldsymbol{\alpha}^{(t)}\|_2^2 \right\}.$$

(3.13)

The problem (3.13) a the closed-form solution via the soft-thresholding operator,

$$\boldsymbol{\gamma}^{(t+1)} = \left(\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\alpha}^{(t)} \right) \cdot \left(1 - \frac{1}{\rho \sqrt{n} \|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\alpha}^{(t)}\|_2} \right)_+,$$

where $(x)_+ := \max\{0, x\}$ is the hinge function.

Step 3. Given $\boldsymbol{\gamma}^{(t+1)}$ and $\boldsymbol{\beta}_j^{(t+1)}$, we update the Lagrange multiplier $\boldsymbol{\alpha}$ by

$$\boldsymbol{\alpha}^{(t+1)} = \boldsymbol{\alpha}^{(t)} + \boldsymbol{\gamma}^{(t+1)} - \mathbf{Z}_{*j} + \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j^{(t+1)}.$$

Since we have rescaled the Lagrange multiplier $\boldsymbol{\alpha}$ by ρ, the updating equation for α is independent of ρ.

The algorithm stops when the following convergence criterion is satisfied

$$\max\left\{\frac{\|\boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\beta}_j^{(t)}\|_2}{\|\boldsymbol{\beta}_j^{(t)}\|_2}, \frac{\|\boldsymbol{\beta}_j^{(t+1)} - \boldsymbol{\gamma}^{(t)}\|_2}{\|\boldsymbol{\beta}_j^{(t)}\|_2}\right\} \leq \epsilon,$$

where $\epsilon > 0$ is a precision tolerance parameter. Theoretically, we can set ϵ to be a very small value, e.g., $\epsilon = 10^{-6}$. This directly solves $\hat{\beta}_j$. However, empirically, we found it is more efficient to set $\epsilon = 10^{-3}$ and only obtain a crude initial estimate $\hat{\tau}_j^{\text{crude}}$. We then use $\hat{\tau}_j^{\text{crude}}$ as the initial value and alternatively solve (3.8) and (3.9). Such a hybrid procedure delivers the best empirical performance.

3.4. Fine-tune the regularization parameter

To achieve the best finite sample performance, we could also find-tune the parameter ζ in (3.4) on a small interval $[\sqrt{2}/\pi, 1]$. Practically, since our procedure is tuning-insensitive, it suffices to only cross-validate 3 values and pick the best one: $\zeta \in \{\sqrt{2}/\pi, 0.6, 1\}$. In general, the algorithm runs very efficiently.

4. Theoretical properties

In this section we investigate the theoretical properties of the proposed method. We begin with some notations and assumptions. We define a matrix class $\mathcal{M}(\xi_{\max}, k)$:

$$\mathcal{M}(\xi_{\max}, k) \quad := \quad \left\{\boldsymbol{\Theta} = \boldsymbol{\Theta}^T \in \mathbb{R}^{d\times d} : \boldsymbol{\Theta} \succ \mathbf{0}, \frac{\Lambda_{\max}(\boldsymbol{\Theta})}{\Lambda_{\min}(\boldsymbol{\Theta})} \leq \xi_{\max}, \right.$$

$$\left. \max_i \sum_j I(\boldsymbol{\Theta}_{ij} \neq 0) \leq k \right\},$$

where $\xi_{\max}$ is a constant and k may scale with the sample size n. We first list down three required assumptions:

(A1) $\boldsymbol{\Theta} \in \mathcal{M}(\xi_{\max}, k)$,
(A2) $k^2 \log d = o(n)$,
(A3) $\limsup_{n\to\infty} \max_{1\leq j\leq d} \boldsymbol{\Sigma}_{jj}^2 \frac{\log d}{n} < \frac{1}{4}$.

All these assumptions are mild. Assumption (A1) only requires the inverse co-variance matrix $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1}$ to have a bounded condition number. Assumption (A2) is equivalent to

$$\lim_{n \to \infty} k\sqrt{\frac{\log d}{n}} = 0. \tag{4.1}$$

In later analysis, we will show that this condition is necessary to secure the consistency of the precision matrix estimation under different matrix norms. Assumption (A3) constrains that the marginal variance of X_j should not diverge too fast.

4.1. Precision matrix estimation consistency

We study the estimation error of precision matrix $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$ under different norms, including spectral norm, matrix L_1-norm, elementwise sup-norm, and Frobenius norm. The rate under the elementwise sup-norm is important for graph recovery. The rate under the spectral norm is important since it leads to the consistency of the estimation of eigenvalues and eigenvectors, which may further be utilized to analyze the theoretical properties of downstream statistical inference. We analyze the rate of spectral norm by bounding the L_1-norm rate. The rate of Frobenius is also a useful measure on the accuracy of the estimation of $\boldsymbol{\Theta}$ and can be viewed as the sum of squared errors for estimating individual rows. Our main results indicate that the TIGER procedure simultaneously achieves the optimal rates of convergence under all these different matrix norms. We present these results in two main theorems and compare our results with related work in the literature. The proofs of these theorems can be found in the appendix.

Theorem 4.1 provides the rates of convergence under the matrix L_1 and spectral norms.

Theorem 4.1 (L_1 and spectral norm rates)**.** *We choose the regularization parameter λ as in (3.4) with $\zeta = 1$. Under Assumptions (A1), (A2), and (A3), we have*

$$\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_1 = O_P\Big(k\left\|\boldsymbol{\Theta}\right\|_2\sqrt{\frac{\log d}{n}}\Big), \tag{4.2}$$

$$\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_2 = O_P\Big(k\left\|\boldsymbol{\Theta}\right\|_2\sqrt{\frac{\log d}{n}}\Big). \tag{4.3}$$

Proof. The proof of (4.2) can be found in Appendix D.1. The proof of (4.3) follows from the inequality that $\left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_2 \le \left\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\right\|_1$. □

If we further assume $\left\|\boldsymbol{\Theta}\right\|_1 \le M_d$ where M_d may scale with d, i.e., we define the following new matrix class $\mathcal{M}(M_d, \xi_{\max}, k)$:

$$\mathcal{M}(M_d, \xi_{\max}, k) := \Big\{\boldsymbol{\Theta} \in \mathcal{M}(\xi_{\max}, k) : \left\|\boldsymbol{\Theta}\right\|_1 \le M_d\Big\}. \tag{4.4}$$

The result of Theorem 4.1 implies that

$$\sup_{\boldsymbol{\Theta} \in \mathcal{M}(M_d, \xi_{\max}, k)} \left\| \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\|_2 = O_P\left(k M_d \sqrt{\frac{\log d}{n}}\right). \tag{4.5}$$

Based on the results of Cai et al. (2016), Liu and Luo (2015), and Yuan (2010), this rate of convergence is minimax optimal on model class $\mathcal{M}(M_d, \xi_{\max}, k)$.

The next theorem provides the rates of convergence under the elementwise sup-norm, which is useful for the graph recovery from the precision matrix $\boldsymbol{\Theta}$.

Theorem 4.2 (Elementwise sup-norm rate). *We choose the regularization parameter λ as in (3.4) with $\zeta = 1$. Let s be the total number of nonzero off-diagonal elements of $\boldsymbol{\Theta}$. Under Assumptions (A1), (A2), and (A3), we have*

$$\left\| \hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta} \right\|_{\max} = O_P\left(\left\| \boldsymbol{\Theta} \right\|_1 \sqrt{\frac{\log d}{n}} \right). \tag{4.6}$$

Proof. The proof of (4.6) can be found in Appendix D.2. □

Again, based on the results in Cai et al. (2016) and Liu and Luo (2015), we know that the TIGER achieves the minimax optimal rates of convergence under both elementwise sup-norm and Frobenius norm on model class $\mathcal{M}(M_d, \xi_{\max}, k)$.

In summary, the TIGER simultaneously achieves the optimal rates of convergence under spectral norm, matrix L_1-norm, and elementwise sup-norm.

4.2. Graph recovery consistency

Next, we study the graph recovery property of the TIGER. Let $\hat{\boldsymbol{\Theta}}$ be the estimated precision matrix. Recall that we define the estimated graph $\hat{G} := (V, \hat{E})$ where $(j, k) \in \hat{E}$ if and only if $\hat{\boldsymbol{\Theta}}_{jk} \hat{\boldsymbol{\Theta}}_{kj} \neq 0$. Similarly, the true graph $G := (V, E)$ where $(j, k) \in E$ if and only if $\boldsymbol{\Theta}_{jk} \neq 0$. We have the following theorem on graph recovery consistency.

Theorem 4.3 (Graph recovery consistency). *We choose the regularization parameter λ as in (3.4) with $\zeta = 1$. We assume Assumptions (A1), (A2), and (A3) hold and for a sufficiently large constant K, such that the smallest nonzero element of $\boldsymbol{\Theta}$ satisfies*

$$\boldsymbol{\Theta}_{\text{crit}} := \min_{(j,k) \in E} \left| \boldsymbol{\Theta}_{jk} \right| \geq K \left\| \boldsymbol{\Theta} \right\|_1 \sqrt{\frac{\log d}{n}}.$$

We have $E \subset \hat{E}$ when n is large enough.

Theoretically, it is possible that the TIGER delivers some precision matrix estimates with very small nonzero values. To achieve exact recovery, we can threshold the estimated precision matrix $\hat{\boldsymbol{\Theta}}$ to get a sparse precision matrix

estimate $\hat{\mathbf{\Theta}}^A$. Specifically, let

$$
\hat{\mathbf{\Theta}}_{jk}^A := \left\{
\begin{array}{ll}
\hat{\mathbf{\Theta}}_{jk} & \text{if} \quad |\hat{\mathbf{\Theta}}_{jk}| > A\sqrt{\dfrac{\log d}{n}} \\[2ex]
0 & \text{if} \quad |\hat{\mathbf{\Theta}}_{jk}| \le A\sqrt{\dfrac{\log d}{n}}
\end{array}
\right. .
$$

It can be seen that under the same conditions of Theorem 4.3, there exists a constant A (A may depend on K) such that the above hresholded estimator achieves exact recovery. More discussions can be found in Cai et al. (2011). Unlike the CLIME and graphical Dantzig selector where the linear program solver may deliver very small nonzero values (but not exact zero). Our algorithm defined in Section 3 explicitly exploit soft-thresholding operator and is capable of delivering exact zero. Empirically, we found the TIGER procedure works very effectively in graph estimation even without this hard-thresholding step. So this hard-thresholding step is more of theoretical interest and is not necessary in applications.

It is worth pointing out that if we are only interested in estimating the graph, assumption (A2) can be straightforwardly relaxed to $k \log d = o(n)$, see for example Meinshausen and Bühlmann (2006) and Jalali et al. (2012).

4.3. Discussion

In this subsection, we briefly discuss the theoretical properties of the TIGER estimator and compare them with other existing results.

The SCIO method proposed in Liu and Luo (2015) also provides the rates of convergence for precision matrix estimation under various norms. One can see that the TIGER estimator and SCIO estimator achieve the same rates of convergence in terms of spectral norm, matrix L_1-norm, and elementwise sup-norm. However, in this paper we consider a much larger matrix class since we only bound the condition number of the precision matrix while the SCIO bounds the largest and smallest eigenvalues from above and below, respectively. More-over, the SCIO requires the irrepresentable condition, which is much stronger than our condition. In fact, it is still an open problem on whether the SCIO estimator can achieve the same rates as the TIGER on the same model class $\mathcal{M}(M_d, \xi_{\max}, k)$.

The graphical Dantzig selector (Yuan, 2010) also considers the model class where the largest and smallest eigenvalues are bounded from above and below. Therefore the results of the Graphical Dantzig selector are on a smaller model class than the TIGER estimator. Moreover, the graphical Dantzig does not have any result on graph recovery consistency.

When compared with the CLIME in Cai et al. (2011), we see that the rate of convergence of the TIGER is faster since the spectral norm rate of the CLIME is $O_P\!\left(kM_d^2\sqrt{\frac{\log d}{n}}\right)$, and the spectral norm rate of the TIGER is $O_P\!\left(kM_d\sqrt{\frac{\log d}{n}}\right)$.

When compared with the glasso method, see for example Rothman et al. (2008) and Ravikumar et al. (2011), the TIGER estimator achieves the minimax optimal rates of convergence under spectral norm, matrix L_1-norm and elementwise sup-norm. In contrast, the only theoretical result of the glasso is in terms of Frobenius norm. it is still an open problem on whether the glasso can achieve the same spectral norm rate of convergence as the TIGER method on the model class $\mathcal{M}(M_d, \xi_{\max}, k)$.

The scaled-Lasso estimator (Sun and Zhang, 2013) provides the rates of convergence under spectral norm and matrix L_1 norm of the precision matrix estimation. However, the scale-Lasso estimator considers a different model class where the smallest eigenvalue of the correlation matrix is bounded away from zero by a constant. It is not clear on the optimality of the obtained rates of convergence over that model class. Besides, Sun and Zhang (2013) does not provide the elementwise sup-norm analysis and hence does not have the graph recovery result. Though the TIGER has a close relationship with the scaled-Lasso, the theoretical analysis of the current paper is fundamentally different from that in Sun and Zhang (2013). Another related work is Meinshausen and Bühlmann (2006), where they only focus on the graph recovery and do not have precision matrix estimation result.

5. Experimental results

We compare the performance of the TIGER with other methods (glasso and CLIME) in parameter estimation and graph recovery using simulated and real datasets. The TIGER and CLIME algorithms are implemented in the R package `camel`, and it is publicly available through CRAN. The glasso method is implemented in the R package `huge` (ver. 1.2.3).

5.1. Numerical simulations

In our numerical simulations, we consider 6 settings to compare these methods: (i) $n = 200, d = 100$; (ii) $n = 200, d = 200$; (iii) $n = 200, d = 400$; (iv) $n = 400, d = 100$; (v) $n = 400, d = 200$; (vi) $n = 400, d = 400$. We adopt the following models for generating undirected graphs and precision matrices. A typical run of the generated graphs and the heatmaps of the precision matrices are illustrated in Figure 1.

∗ **Scale-free graph**. The degree distribution of the scale-free graph follows a power law. The graph is generated by the preferential attachment mechanism. The graph begins with an initial small chain graph of 2 nodes. New nodes are added to the graph one at a time. Each new node is connected to one existing node with a probability that is proportional to the number of degrees that the existing node already has. Formally, the probability p_i that the new node is connected to an existing node i is, $p_i = \frac{k_i}{\sum_j k_j}$, where k_i is the degree of node i. The resulting graph has d edges ($d = 200$ or $d = 400$). Once the graph

(a) Scale-free   (b) Erdös-Rényi   (c) Hub

(d) Cluster   (e) Band   (f) Block

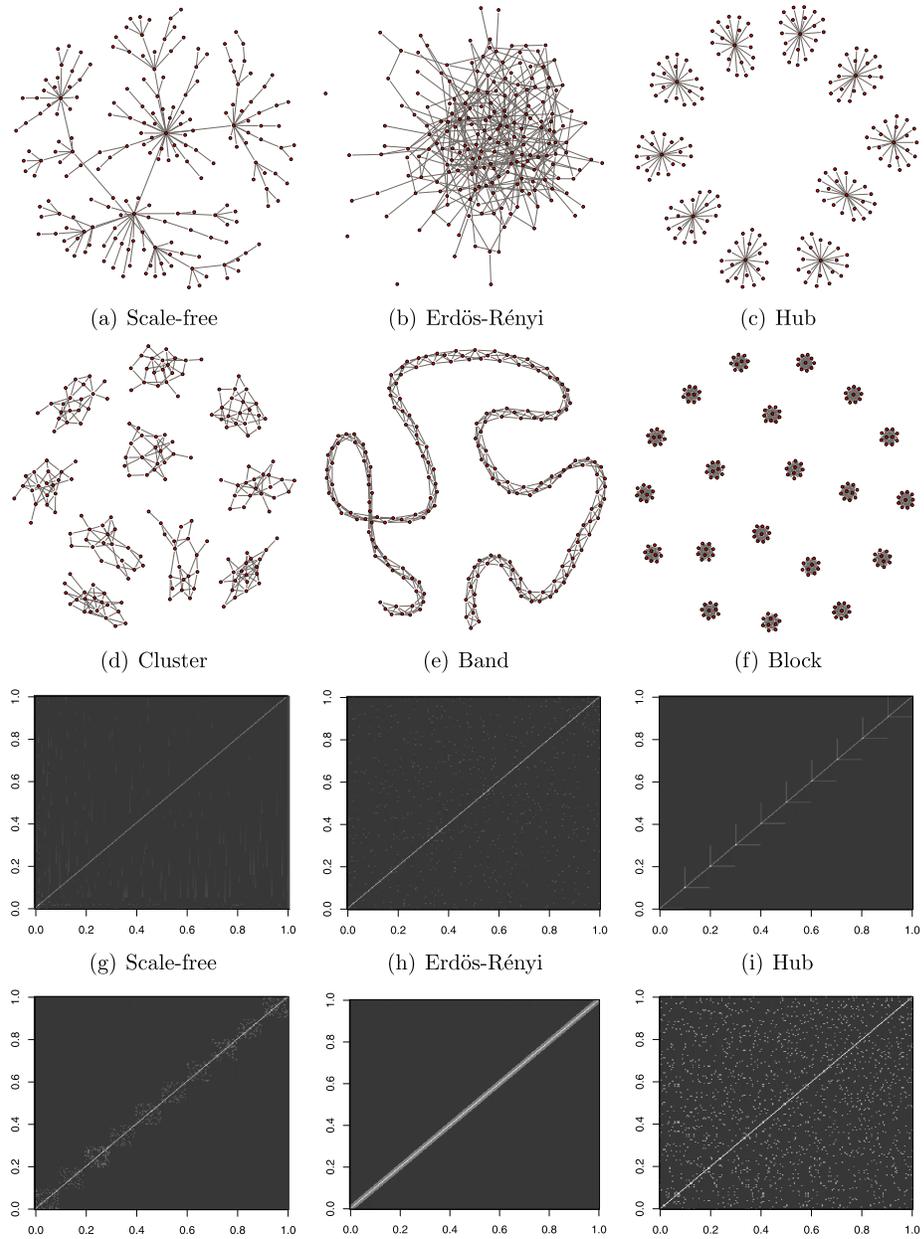(g) Scale-free   (h) Erdös-Rényi   (i) Hub

FIG 1. *An illustration of the 6 graph patterns and the heatmaps of their corresponding precision matrices adopted in the simulations. To ease visualization, we only present example graphs with $d = 200$ nodes.*

is obtained, we generate an adjacency matrix $\mathbf{A}$ by setting the nonzero off-diagonal elements to be 0.3 and the diagonal elements to be 0. We calculate its smallest eigenvalue $\Lambda_{\min}(\mathbf{A})$. The precision matrix is constructed as

$$\mathbf{\Theta} = \mathbf{D}\big[\mathbf{A} + \big(\big|\Lambda_{\min}(\mathbf{A})\big| + 0.2\big)\cdot\mathbf{I}_d\big]\mathbf{D}, \tag{5.1}$$

where $\mathbf{D}\in\mathbb{R}^{d\times d}$ is a diagonal matrix with $\mathbf{D}_{jj} = 1$ for $j = 1,\ldots,d/2$ and $\mathbf{D}_{jj} = 3$ for $j = d/2+1,\ldots,d$. The covariance matrix $\mathbf{\Sigma} := \mathbf{\Theta}^{-1}$ is then computed to generate the multivariate normal data: $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n \sim N_d(\mathbf{0},\mathbf{\Sigma})$.

∗ **Erdös-Rényi random graph**. We add an edge between each pair of nodes with probability 0.02 independently. The resulting graph has approximately 400 edges when $d = 200$ and $1,596$ edges when $d = 400$. Once the graph is obtained, we construct the adjacency matrix $\mathbf{A}$ and generate the precision matrix $\mathbf{\Theta}$ using (5.1) but setting $\mathbf{D}_{jj} = 1$ for $j = 1,\ldots,d/2$ and $\mathbf{D}_{jj} = 1.5$ for $j = d/2+1,\ldots,d$. We then invert $\mathbf{\Theta}$ to get the covariance matrices $\mathbf{\Sigma}$ and generate the multivariate normal data: $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n \sim N_d(\mathbf{0},\mathbf{\Sigma})$.

∗ **Hub graph**. The d nodes are evenly partitioned into $d/20$ disjoint groups with each group contains 20 nodes. Within each group, one node is selected as the hub and we add edges between the hub and the other 19 nodes in that group. The resulting graph has 190 edges when $d = 200$ and 380 edges when $d = 400$. Once the graph is obtained, we generate the precision matrix in the same way as the Erdös-Rényi random graph model.

∗ **Cluster graph**. Similar to the hub model, the d nodes are evenly partitioned into $d/20$ disjoint groups with each group contains 20 nodes. The subgraph of each group is an Erdös-Rényi random graph with the probability parameter 0.2. The resulting graph has approximately 380 edges when $d = 200$ and 760 edges when $d = 400$. Once the graph is obtained, we generate the precision and covariance matrices in the same way as the Erdös-Rényi random graph model.

∗ **Band graph**. Each node is assigned a coordinate j with $j = 1,...,d$. Two nodes are connected by an edge whenever the corresponding coordinates are at distance less than or equal to 3. The resulting graph has approximately 594 edges when $d = 200$ and $1,194$ edges when $d = 400$. Once the graph is obtained, we generate the precision and covariance matrices in the same way as the Erdös-Rényi random graph model.

∗ **Block graph**. The precision matrix $\mathbf{\Theta}$ is a block diagonal matrix with block size $d/20$. Each block has off-diagonal entries equal to 0.5 and diagonal entries equal to 1. Such a matrix is guaranteed to be positive definite. The resulting matrix is then randomly permuted by rows and columns. The resulting graph has approximately 900 edges when $d = 200$ and $3,800$ edges when $d = 400$. The covariance matrix $\mathbf{\Sigma} := \mathbf{D}^{-1}\mathbf{\Theta}^{-1}\mathbf{D}^{-1}$ is then computed to generate multivariate normal data, where $\mathbf{D}$ is a diagonal matrix with $\mathbf{D}_{jj} = 1$ for $j = 1,\ldots,d/2$ and $\mathbf{D}_{jj} = 1.5$ for $j = d/2+1,\ldots,d$.

5.1.1. *Graph recovery performance*

We first compare the TIGER with the CLIME and glasso on their graph recovery performance. Let $G = (V, E)$ be a d-dimensional graph. We denote by $|E|$ the number of edges in the graph G. We use the false positive and false negative rates to evaluate the graph recovery performance. Let $\hat{G}^\lambda = (V, \hat{E}^\lambda)$ be an estimated graph using a regularization parameter λ in any of these procedures. The number of false positives when using the regularization parameter λ is defined as $\mathrm{FP}(\lambda) :=$ the number of edges in $\hat{E}^\lambda$ but not in E. The number of false negatives with λ is defined as $\mathrm{FN}(\lambda) :=$ the number of edges in E but not in $\hat{E}^\lambda$. We further define the false negative rate (FNR) and false positive rate (FPR) as

$$\mathrm{FNR}(\lambda) := \frac{\mathrm{FN}(\lambda)}{|E|} \quad \text{and} \quad \mathrm{FPR}(\lambda) := \mathrm{FP}(\lambda) / \left[\binom{d}{2} - |E| \right]. \tag{5.2}$$

To illustrate the overall performance of the TIGER, CLIME and glasso methods over the full paths, the receiver operating characteristic (ROC) curves are drawn using $\big(\mathrm{FNR}(\lambda), 1 - \mathrm{FPR}(\lambda)\big)$. For the TIGER method, we found that using or without using the second hard-thresholding step provides the same graph estimates. So, the presented result does not use the second hard-thresholding step. The ROC curves for these models are presented in Figures 2, 3, 4.

From the ROC curves on the scale-free model in Figure 2, we see that the graph recovery performance of the TIGER is significantly better than those of the CLIME and glasso in higher dimensional settings (when $d = 200$ or $d = 400$ and $d \geq n$). This result suggests that the TIGER is more adaptive to inhomogeneous noise models. From Figure 2, we also see that the TIGER has similar graph recovery performance as the CLIME on the band model. Both the TIGER and CLIME significantly outperform the glasso. In particular, in the high dimensional setting when $d = 400$ and $n = 200$, the TIGER outperforms both CLIME and glasso. This result suggests that the TIGER is more reliable when facing higher dimensional problems on this model.

For the other models, from the ROC curves in Figures 3 and 4, we see that the three methods perform similarly on these settings, and for Erdös-Rényi random graph models, the TIGER outperforms the CLIME in the settings when $n \leq d$. This means that the TIGER is effective for a wide range of models. In summary, the above numerical results suggest that the TIGER is a very competitive graph estimation method in high dimensions.

5.1.2. *Tuning-insensitive regularization path*

We are interested in studying the tuning-insensitive property of the TIGER. For conciseness, we only discuss the TIGER and CLIME in this section and compare their regularization paths. Recall that

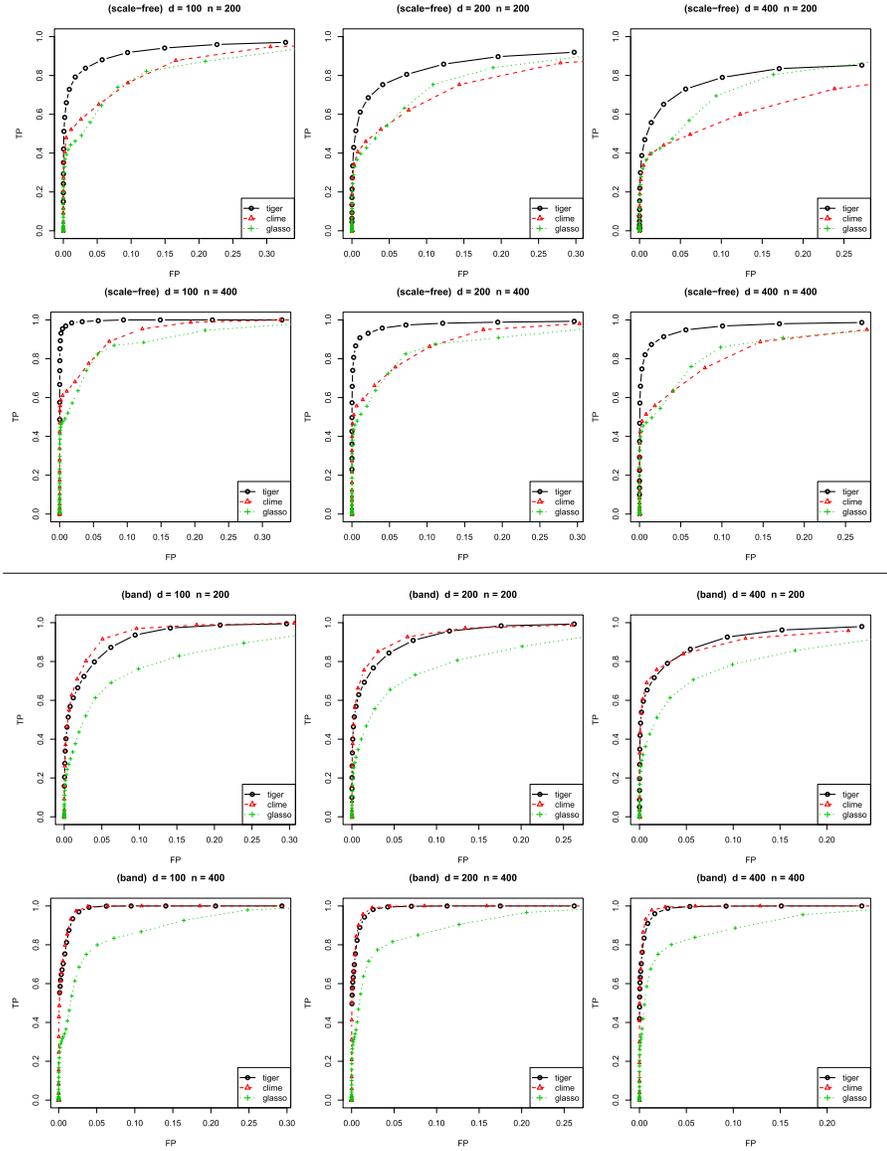$$\lambda = \zeta \cdot \frac{\pi}{\sqrt{2}} \sqrt{\frac{\log d}{n}}, \tag{5.3}$$

Fɪɢ 2. *ROC curves for the Scale-free and Band models (Best visualized in color).*

we see that ζ and λ have a one-to-one mapping. In Figure 5 (a) and (b), we plot the curves of Frobenius-norm errors vs. the tuning parameter ζ for the TIGER and CLIME. We define $\mathrm{FNR}(\zeta)$ and $\mathrm{FPR}(\zeta)$ in the same way as in (5.2). We also define the graph recovery accuracy as

$$\mathrm{Accuracy}(\zeta) := 1 - \mathrm{FPR}(\zeta) - \mathrm{FNR}(\zeta). \tag{5.4}$$
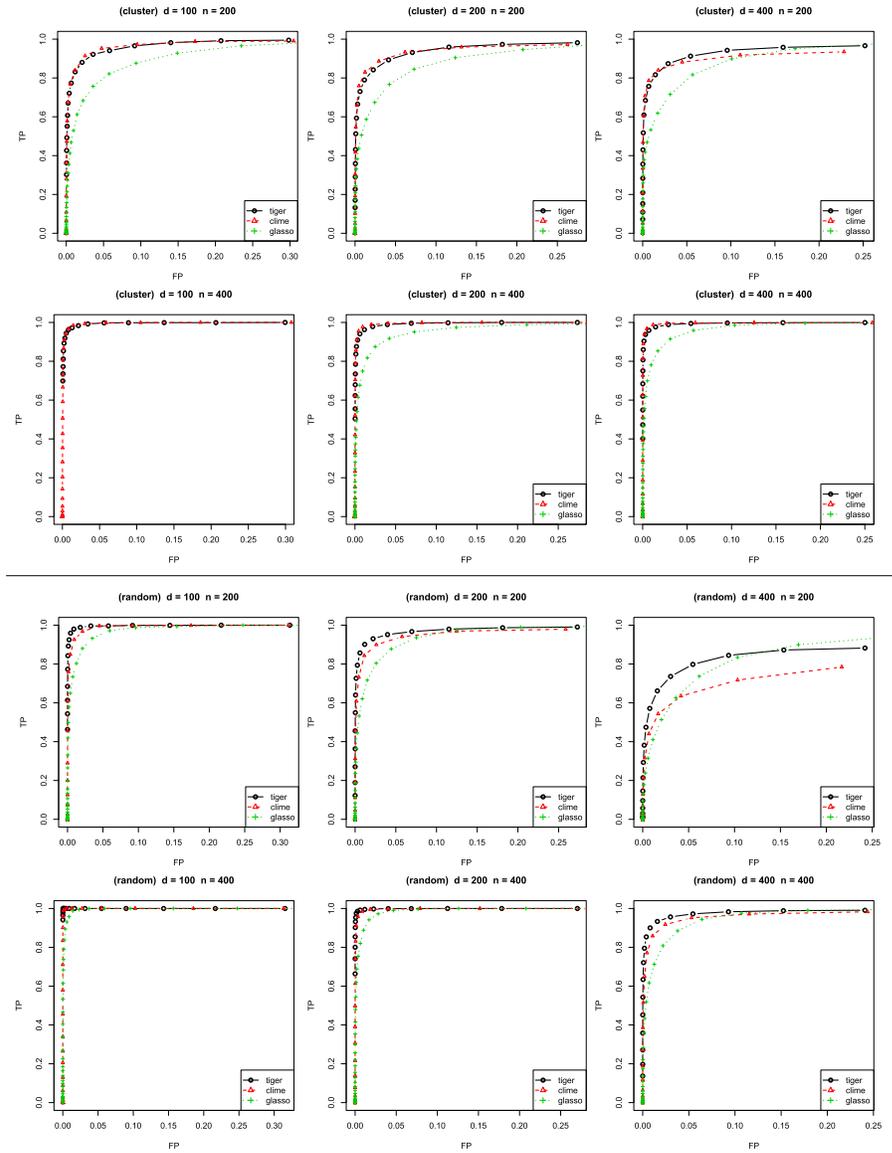
FIG 3. *ROC curves for the cluster and Erdös-Rényi random graph models (Best visualized in color).*

In Figure 5 (c) and (d), we plot the curves of the graph recovery accuracy vs. the tuning parameter ζ for the TIGER and CLIME. The vertical axis of these plots are calibrated so that the results are directly comparable. These plots illustrate the tuning-insensitive property of the TIGER regularization path. For the TIGER, we found it is empirically safe to only consider the regularization path
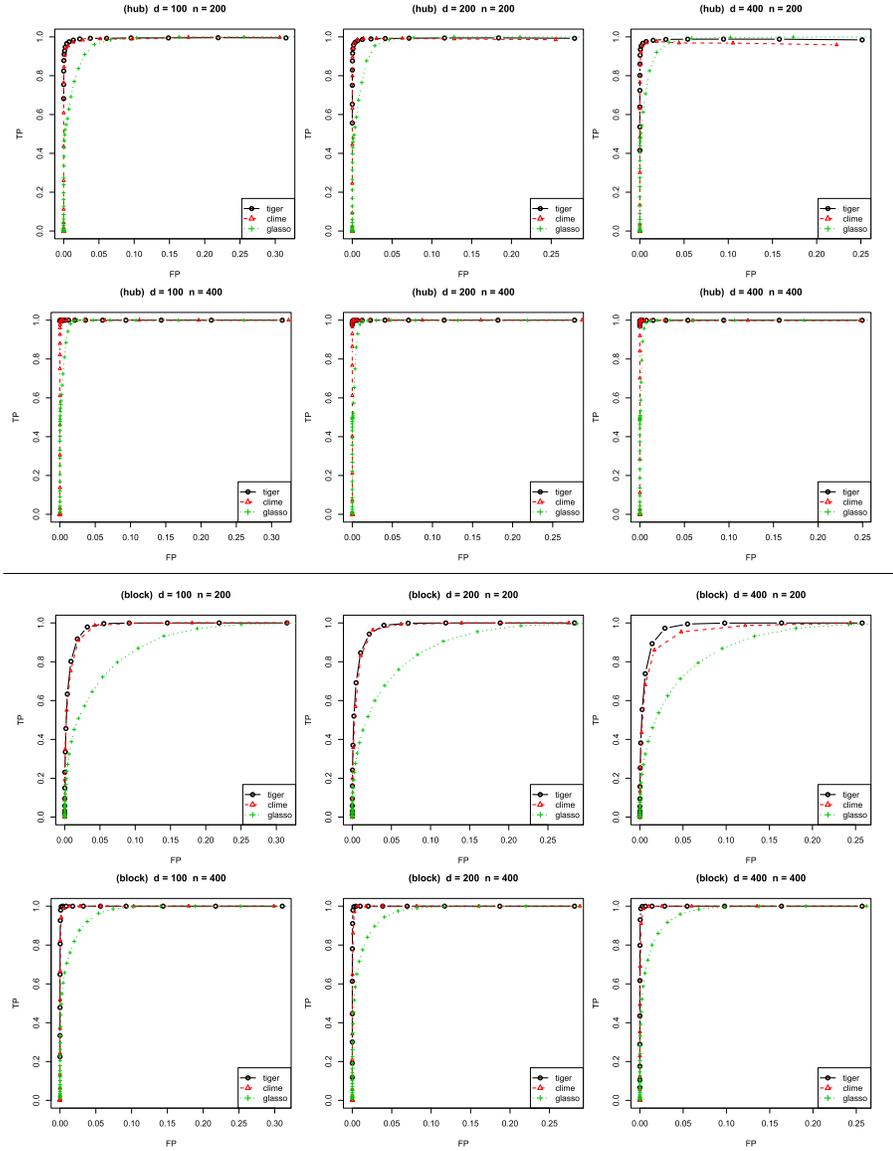
Fig 4. *ROC curves for the hub and block models (Best visualized in color).*

over the range $\zeta \in [\sqrt{2}/\pi, 1]$. From both subplots (a) and (c), the regularization paths are flat and do not change dramatically with the change of ζ (In another word, the procedure is insensitive to the tuning parameter). In contrast, for the CLIME, we need to search over a much larger range of ζ to find the optimal value and the paths are more irregular. In the subplots (b) and (d), we visual-

Fig 5. *Comparison of the regularization paths of the TIGER and CLIME on the hub graph model. The vertical axis of these plots (Forebius divide by the dimensionality d) are calibrated so that the results are directly comparable. These plots illustrate the tuning-insensitive property of the TIGER regularization path. For the TIGER, we only need to consider the range $\zeta \in [\sqrt{2}/\pi, 1]$. From bo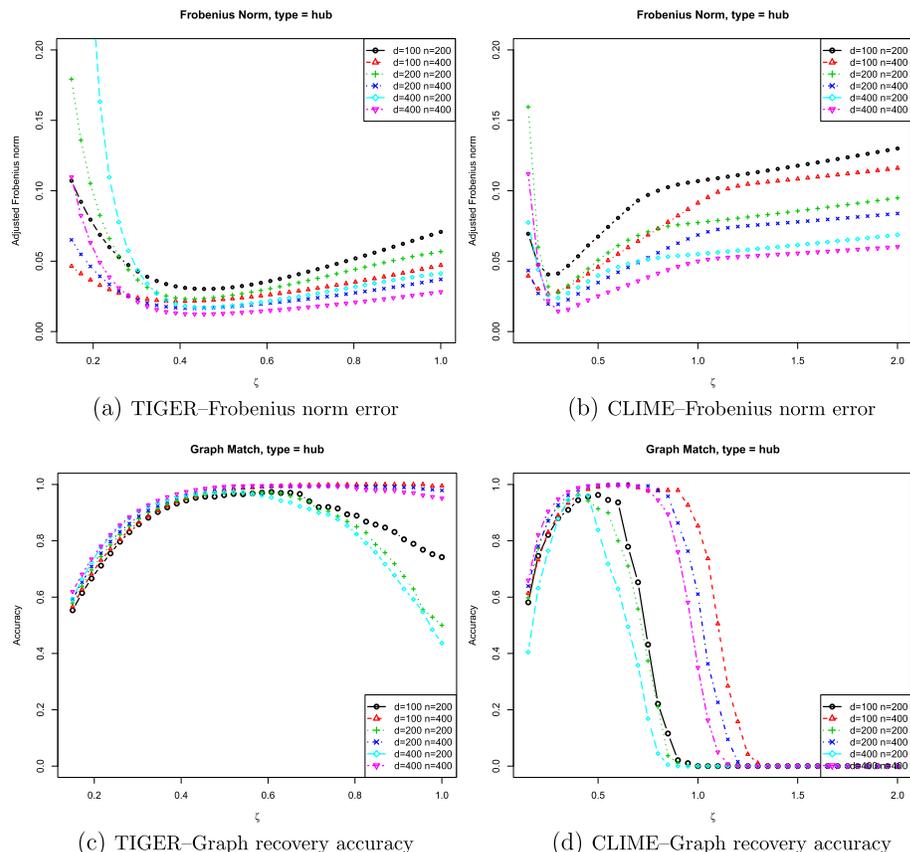th subplots (a) and (c), the regularization paths are flat and do not change dramatically with ζ. In contrast, for the CLIME, we need to search over a much larger range of ζ to find the optimal value and the paths are more irregular (the subplots (b) and (d) only show a part of this range).*

ize the regularization paths of the CLIME over $\zeta \in [0.125, 2]$, these are only a sub-fraction of the whole regularization paths of the CLIME and the paths are irregular (or more sensitive to the choice of ζ). Therefore, it is much easier to choose a reasonable tuning parameter for the TIGER than for the CLIME. In most cases, the choice of $\zeta = 1$ for TIGER provides a sparser graph estimate than the true graph. This is due to the fact that the asymptotic analysis in the previous section involves many union bounds, which may be too conservative in finite sample settings. As will be illustrated in the next subsection, we found that in most settings, simply setting $\zeta = \sqrt{2}/\pi$ yields a reasonably good graph

and precision matrix estimates. Such a choice is used as a default in the `camel` package on CRAN.

5.1.3. Parameter estimation performance

We then compare the TIGER with CLIME and glasso on their parameter estimation performance. For each model described before, we generate a training sample of n. The tuning parameters of the glasso and CLIME are automatically chosen in a data-dependent way. More specifically, for a given sample size n, we generate the same number of independent data points from the same distribution as a validation set. For each tuning parameter, the glasso or CLIME estimated precision matrix $\hat{\boldsymbol{\Theta}}$ is calculated on training data. The optimal tuning parameter is chosen by minimizing the held-out negative log-likelihood loss

$$\mathcal{L}_n(\hat{\boldsymbol{\Theta}}) := \mathrm{tr}\big(\hat{\boldsymbol{\Sigma}}\hat{\boldsymbol{\Theta}}\big) - \log\big|\hat{\boldsymbol{\Theta}}\big|$$

on the validation set. For the TIGER, the tuning parameter ζ in (3.4) is simply chosen to be $\zeta = \sqrt{2}/\pi$ so that the procedure is completely tuning-free. For dimensionality $d = 100, 200, 400$, we consider the spectral norm error $\big\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\big\|_2$ and Frobenius norm error $\big\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}\big\|_{\mathsf{F}}$ of all the 6 models described in the previous subsections.

The results are reported in Tables 1 and 2. In these tables we present the mean and standard deviation (in the parenthesis) of the spectral and Frobenius norm errors based on 50 random trials. We see that in almost all cases, the TIGER and CLIME outperform the glasso. In most cases, the TIGER outperforms the CLIME. Our results, though obtained in different experimental settings, are consistent with the results in Sun and Zhang (2013) for the scaled-Lasso based matrix inversion method.

One possible reason for the superior empirical performance of the TIGER over CLIME and glasso is that the data-dependent tuning selection procedure described in the previous section does not work well for the CLIME and glasso. To gain more insight, we also report the oracle estimation results in Tables 3 and 4. In these tables we present the mean and standard deviation (in the parenthesis) of the spectral and Frobenius norm errors based on 50 random trials. For all three methods, we draw the full regularization paths and select the best tuning parameter by minimizing the corresponding spectral or Frobenius norm errors to the true precision matrix. From these tables, we see that again the TIGER and CLIME outperform glasso and the TIGER is slightly better than CLIME.

5.1.4. Time benchmark and dense graph estimation

In the above numerical study we did not report the exact wall-clock time complexity of the proposed algorithm. A thorough evaluation of the computational performance of the TIGER estimator is reported in a recent work (Li

TABLE 1

Quantitative comparisons of the TIGER, Glasso, and CLIME on the scale-free, hub, and band models using data-dependent model selection method.

Model	n	d	Spectrum Norm			Frobenius Norm		
			TIGER	CLIME	GLasso	TIGER	CLIME	GLasso
scale-free	200	100	3.71370 (0.4651)	4.41206 (0.4149)	6.00048 (0.1838)	11.5245 (0.7596)	14.3260 (1.0260)	21.1852 (0.4794)
		200	4.11834 (0.6495)	4.19539 (0.3070)	6.54951 (0.3823)	16.3318 (0.8017)	17.0776 (0.5752)	35.1912 (2.4153)
		400	4.43263 (0.6412)	4.58973 (0.5466)	7.21601 (0.2184)	23.4459 (0.7896)	24.1157 (1.0119)	50.9695 (0.3715)
hub	400	100	2.77888 (0.3167)	3.53363 (0.5060)	5.30674 (0.1676)	8.3591 (0.5206)	11.4838 (0.7872)	17.4945 (0.3696)
		200	2.68762 (0.2668)	3.29772 (0.2480)	5.23455 (0.2602)	11.7521 (0.4706)	14.0718 (0.5345)	23.3348 (0.3524)
		400	3.31452 (0.5758)	3.57110 (0.3068)	6.32043 (0.2675)	16.9996 (0.6411)	19.1326 (0.7619)	41.8133 (0.3717)
hub	200	100	2.67040 (0.4446)	4.31716 (0.4979)	6.51814 (0.3070)	5.4347 (0.3356)	7.3545 (0.5204)	10.6739 (0.4329)
		200	3.02307 (0.4151)	5.35713 (0.3096)	7.41337 (0.1374)	8.2277 (0.3049)	12.7079 (0.4182)	18.1376 (0.1653)
		400	3.34315 (0.3017)	6.11506 (0.2093)	7.69167 (0.1061)	12.0676 (0.2801)	19.8038 (0.3058)	26.4038 (0.2002)
	400	100	1.82245 (0.2619)	2.46543 (0.3995)	5.28610 (0.1834)	3.7161 (0.1811)	4.5920 (0.2968)	8.5380 (0.2247)
		200	2.07601 (0.2105)	3.11260 (0.2491)	6.23933 (0.1208)	5.6517 (0.2366)	7.5627 (0.3261)	15.0739 (0.1747)
		400	2.23719 (0.1675)	4.02006 (0.4557)	6.43531 (0.0913)	8.1420 (0.1207)	13.0960 (1.1531)	22.0039 (0.1230)
band	200	100	5.72715 (0.1546)	4.37815 (0.3626)	6.36789 (0.0850)	16.7205 (0.2793)	12.8498 (0.8868)	17.4244 (0.1506)
		200	6.04373 (0.1381)	5.81030 (0.1378)	7.32236 (0.0517)	24.6770 (0.2373)	23.3002 (0.3368)	28.8514 (0.1293)
		400	6.28046 (0.0760)	6.76049 (0.0738)	7.87172 (0.1168)	36.0083 (0.1940)	38.6665 (0.1779)	44.7691 (0.7423)
	400	100	4.34163 (0.1847)	2.92088 (0.2552)	5.53820 (0.0871)	12.5012 (0.2989)	8.2567 (0.5046)	14.7940 (0.1483)
		200	4.69878 (0.1302)	3.62779 (0.1494)	5.71676 (0.0740)	19.0534 (0.2633)	14.3844 (0.2863)	21.9130 (0.1285)
		400	5.01406 (0.0709)	4.24658 (0.4150)	6.82527 (0.0392)	28.6523 (0.2313)	24.1215 (2.0819)	37.6074 (0.0953)

et al., 2016). In addition to the ADMM algorithm described in this paper, Li et al. (2016) also introduce an efficient pathwise iterative smoothing shrinkage thresholding algorithm, along a novel model-based perspective for analyzing the smoothing optimization framework. Thorough wall-clock timing benchmarks of these algorithms on different datasets are reported in Li et al. (2016).

TABLE 2
Quantitative comparisons of the TIGER, Glasso, and CLIME on the block, Erdös-Rényi random, and cluster models using data-dependent model selection method.

Model	n	d	Spectrum Norm			Frobenius Norm		
			TIGER	CLIME	GLasso	TIGER	CLIME	GLasso
block	200	100	3.88080 (0.2123)	2.92791 (0.1907)	4.62777 (0.1221)	12.7803 (0.2231)	9.8588 (0.2728)	14.1474 (0.3767)
		200	4.21258 (0.2073)	3.50847 (0.2014)	5.05219 (0.0408)	19.1984 (0.1841)	16.4653 (0.2043)	21.8517 (0.0455)
		400	4.54196 (0.1456)	4.72388 (0.1344)	5.44260 (0.0387)	28.8940 (0.1471)	29.7674 (0.1169)	34.0288 (0.2657)
random	400	100	2.61796 (0.1746)	1.91879 (0.2812)	3.63244 (0.0969)	8.4651 (0.2232)	6.1005 (0.6898)	10.7808 (0.0964)
		200	2.86024 (0.1613)	2.04201 (0.1219)	4.26378 (0.0476)	12.8697 (0.2201)	9.7477 (0.2173)	17.7396 (0.0674)
		400	3.12185 (0.1423)	3.03083 (0.1365)	4.70495 (0.0333)	19.4939 (0.2277)	19.6861 (0.1791)	28.0988 (0.0522)
random	200	100	1.40361 (0.2093)	1.63446 (0.2246)	2.66316 (0.2626)	4.9173 (0.2158)	5.5390 (0.2072)	7.4979 (0.5315)
		200	1.92515 (0.1352)	2.13847 (0.1984)	2.97502 (0.2074)	9.3623 (0.1845)	10.4678 (0.5063)	13.2284 (0.8328)
		400	3.03486 (0.0710)	3.57549 (0.0417)	4.16129 (0.0313)	17.6548 (0.1165)	20.1956 (0.1937)	23.3014 (0.0777)
	400	100	0.96871 (0.1265)	1.08246 (0.1420)	1.94886 (0.1879)	3.3962 (0.1332)	3.7713 (0.1521)	5.4841 (0.2622)
		200	1.38675 (0.1133)	1.45379 (0.1517)	2.20816 (0.0634)	6.6106 (0.1273)	7.2891 (0.2580)	9.6509 (0.1012)
		400	2.21101 (0.0792)	2.49335 (0.2862)	3.02710 (0.0468)	13.3298 (0.2506)	14.9634 (0.7566)	17.0508 (0.1373)
cluster	200	100	3.84966 (0.3371)	3.45717 (0.3465)	5.35790 (0.1325)	8.9219 (0.2710)	8.3992 (0.3329)	11.7727 (0.1427)
		200	3.66157 (0.2469)	3.90672 (0.3981)	5.11136 (0.2043)	11.6676 (0.2153)	12.1507 (0.7474)	16.0086 (0.5784)
		400	2.99469 (0.1527)	3.34376 (0.1257)	4.11403 (0.0848)	15.1022 (0.1233)	16.5420 (0.1500)	20.2334 (0.1057)
	400	100	2.74935 (0.2604)	2.32058 (0.2737)	4.21725 (0.4733)	6.4102 (0.2067)	5.7983 (0.2324)	8.9115 (0.8694)
		200	2.97759 (0.1958)	2.80625 (0.2788)	4.18294 (0.0969)	8.7524 (0.1863)	8.4317 (0.3450)	11.9685 (0.1067)
		400	2.20812 (0.0627)	2.29522 (0.0675)	3.58522 (0.0630)	11.0885 (0.1962)	11.4516 (0.1888)	16.8512 (0.1091)

As for another concern, many of the graphs (except the block one) under consideration in the above numerical study have a high sparsity level (i.e., edges present represent about 1-2% of the potential ones). It would be interesting to investigate the performance of the proposed method on graphical models with dense graphs (e.g, graphs at the sparsity level 5% or 10%). Such analysis is also conducted and reported in Li et al. (2016).

TABLE 3

Quantitative comparisons of the TIGER, Glasso, and CLIME on the scale-free, hub, and band models using oracle model selection method.

Model	n	d	Spectrum Norm			Frobenius Norm		
			TIGER	CLIME	GLasso	TIGER	CLIME	GLasso
scale-free	200	100	3.49331 (0.2888)	4.20880 (0.3896)	4.44102 (0.2517)	11.4419 (0.7051)	14.0362 (0.8067)	14.6163 (0.4395)
		200	3.55648 (0.4300)	3.97078 (0.3312)	4.06802 (0.3903)	15.5481 (0.5613)	16.8466 (0.4808)	20.6089 (0.2790)
		400	3.93523 (0.5407)	4.50636 (0.4290)	4.36077 (0.2883)	21.2401 (0.5478)	23.6391 (0.2663)	32.9509 (0.1983)
	400	100	2.65555 (0.2640)	2.92806 (0.2593)	3.72758 (0.2702)	8.3378 (0.5277)	10.1875 (0.4220)	11.5142 (0.4212)
		200	2.62826 (0.2368)	3.20446 (0.1861)	3.46773 (0.3240)	11.7350 (0.4554)	13.9074 (0.3791)	16.2752 (0.2982)
		400	3.10365 (0.4002)	3.42918 (0.2338)	3.74972 (0.3319)	16.7337 (0.5986)	18.9504 (0.4950)	25.1616 (0.2071)
hub	200	100	2.42283 (0.3287)	2.66693 (0.3362)	4.13856 (0.2862)	5.4124 (0.3213)	6.6481 (0.3566)	9.6412 (0.2294)
		200	2.80843 (0.2737)	3.37883 (0.2853)	4.18505 (0.2116)	8.2170 (0.3008)	10.9757 (0.3744)	15.5200 (0.1934)
		400	3.18571 (0.2904)	4.51681 (0.3600)	4.06350 (0.0675)	12.0676 (0.2801)	15.5895 (0.2402)	24.0977 (0.1782)
	400	100	1.63670 (0.1863)	1.70506 (0.2001)	3.31738 (0.2704)	3.6970 (0.1763)	4.2848 (0.2483)	7.6101 (0.2076)
		200	1.93008 (0.2119)	2.23970 (0.2799)	3.52276 (0.1672)	5.6341 (0.2342)	6.4934 (0.2451)	12.3937 (0.1873)
		400	2.10478 (0.1478)	2.49148 (0.2029)	3.45920 (0.1311)	8.1420 (0.1207)	10.4606 (0.1850)	19.0323 (0.1332)
band	200	100	3.29816 (0.2127)	3.27817 (0.2519)	4.69050 (0.1342)	12.2957 (0.4519)	11.7620 (0.4908)	14.0160 (0.2475)
		200	3.89225 (0.2489)	4.53169 (0.2191)	4.78844 (0.1114)	19.7902 (0.3364)	20.0180 (0.3778)	22.8946 (0.2232)
		400	4.72424 (0.1313)	5.57854 (0.1061)	4.77270 (0.0857)	31.1798 (0.2101)	34.4919 (0.3571)	37.4700 (0.1044)
	400	100	2.09320 (0.1604)	2.18392 (0.2292)	3.87918 (0.1348)	7.8187 (0.2798)	7.4255 (0.2845)	10.8361 (0.2176)
		200	2.44876 (0.1452)	2.65924 (0.1878)	4.00698 (0.1037)	12.8840 (0.1946)	12.3307 (0.2244)	17.2196 (0.1598)
		400	2.82127 (0.1166)	2.94520 (0.0979)	4.08170 (0.0655)	20.4888 (0.2645)	21.4755 (0.2665)	28.8954 (0.1768)

5.2. *Arabidopsis thaliana gene network analysis*

This dataset includes 118 gene expression arrays from Arabidopsis thaliana originally appeared in Wille et al. (2004). Our analysis focuses on gene expression from 39 genes involved in two isoprenoid metabolic pathways: 16 from the mevalonate (MVA) pathway are located in the cytoplasm, 18 from the plastidial (MEP) pathway are located in the chloroplast, and 5 are located in the mito-

TABLE 4

Quantitative comparisons of the TIGER, Glasso, and CLIME on the block, Erdös-Rényi random, and cluster models using oracle model selection method.

Model	n	d	Spectrum Norm			Frobenius Norm		
			TIGER	CLIME	GLasso	TIGER	CLIME	GLasso
block	200	100	1.78662 (0.1583)	1.96932 (0.1773)	2.30284 (0.1754)	7.0491 (0.2670)	7.2065 (0.3231)	8.7645 (0.1882)
		200	2.26303 (0.1872)	2.58484 (0.1816)	2.72901 (0.0660)	11.8684 (0.2641)	13.6219 (0.2550)	16.0190 (0.1207)
		400	2.79826 (0.2608)	3.67637 (0.2448)	3.60729 (0.0702)	20.5628 (0.2978)	26.3292 (0.2781)	25.4660 (0.0911)
	400	100	1.18086 (0.1352)	1.22283 (0.1702)	1.89098 (0.1439)	4.5090 (0.1631)	4.2706 (0.1898)	6.2619 (0.1538)
		200	1.33627 (0.1202)	1.48764 (0.1009)	1.94656 (0.1143)	7.4712 (0.1473)	7.9438 (0.1856)	11.0249 (0.1133)
		400	1.58491 (0.1664)	2.04870 (0.1761)	2.58467 (0.0316)	11.8926 (0.1673)	15.1659 (0.1953)	19.7672 (0.1090)
random	200	100	1.30656 (0.1552)	1.54958 (0.1801)	1.93012 (0.1224)	4.8291 (0.2090)	5.4697 (0.1878)	6.4292 (0.1705)
		200	1.74892 (0.1026)	2.05243 (0.0972)	2.23740 (0.0769)	9.1014 (0.1862)	10.2764 (0.2062)	11.3447 (0.1505)
		400	2.55517 (0.0663)	3.49539 (0.1640)	3.01732 (0.0501)	17.0088 (0.1298)	20.1956 (0.1937)	20.0869 (0.0937)
	400	100	0.87331 (0.0838)	1.03685 (0.0957)	1.35098 (0.0712)	3.3014 (0.1317)	3.7713 (0.1521)	4.8714 (0.1361)
		200	1.19717 (0.0783)	1.37938 (0.0963)	1.79435 (0.0578)	6.2812 (0.1174)	7.1464 (0.1425)	8.7373 (0.0880)
		400	1.67683 (0.1227)	2.11043 (0.0694)	2.45761 (0.0596)	12.3481 (0.2473)	14.0788 (0.3149)	15.6806 (0.1436)
cluster	200	100	2.50462 (0.2153)	2.61965 (0.2687)	3.12584 (0.2686)	7.9860 (0.2865)	8.1190 (0.3492)	9.9439 (0.1874)
		200	2.72033 (0.2079)	2.89910 (0.2528)	2.91801 (0.1019)	11.1472 (0.2200)	11.5100 (0.2111)	13.2443 (0.1515)
		400	2.49864 (0.1125)	3.30405 (0.1874)	2.94846 (0.1043)	14.8704 (0.2025)	16.5420 (0.1500)	18.3397 (0.1082)
	400	100	1.63625 (0.1633)	1.74909 (0.1612)	2.47721 (0.2247)	5.3929 (0.2439)	5.4465 (0.2127)	7.1791 (0.1576)
		200	1.84653 (0.1555)	2.04163 (0.1747)	2.29099 (0.1570)	7.7920 (0.1718)	7.7961 (0.1957)	10.4819 (0.1135)
		400	1.66908 (0.1359)	1.75361 (0.0699)	2.46020 (0.0658)	10.3907 (0.1888)	10.8631 (0.1774)	14.6437 (0.1132)

chondria. While the two pathways generally operate independently, crosstalk is known to happen (Wille et al., 2004). Our goal is to recover the gene regulatory network, with special interest in crosstalk. We first examine whether the data actually satisfies the Gaussian distribution assumption. In Figure 6 we plot the histogram and normal QQ plot of the expression levels of a gene named MECPS. From the histogram, we see the distribution is left-skewed compared to the Gaussian distribution. From the normal QQ plot, we see the empirical
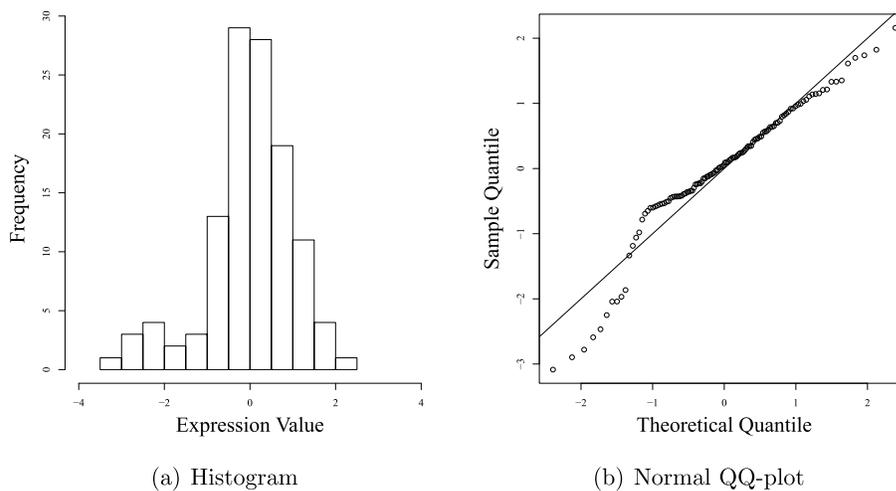
(a) Histogram   (b) Normal QQ-plot

FIG 6. *The histogram and normal QQ plots of the marginal expression levels of the gene MECPS. We see the data are not exactly Gaussian distributed.*

distribution has a heavier tail compared to Gaussian. To suitably apply the TIGER method on this dataset, we need to first transform the data so that its distribution is closer to Gaussian. Therefore, we Gaussianize the marginal expression values of each gene by converting them to the corresponding normal-scores. This is automatically done by the `huge.npn` function in the R package `huge` (Zhao et al., 2012).

We apply the TIGER on the transformed data using the default tuning parameter $\zeta = \sqrt{2}/\pi$. The estimated network is shown in Figure 7. We see the estimated network is very sparse with only 44 edges. We draw the within-pathway connections using solid lines and the between-pathway connections using dashed lines. Our result is consistent with previous investigations, which suggest that the connections from genes AACT1 and HMGR2 to gene MECPS indicate a primary sources of the crosstalk between the MEP and MVA pathways and these edges are presented in the estimated network. MECPS is clearly a hub gene for this pathway.

For the MEP pathway, the genes DXPS2, DXR, MCT, CMK, HDR, and MECPS are connected as in the true metabolic pathway. Similarly, for the MVA pathway, the genes AACT2, HMGR2, MK, MPDC1, MPDC2, FPPS1 and FPP2 are closely connected. Our analysis suggests 11 cross-pathway links, which is consistent to previous investigation in Wille et al. (2004). This result suggests that there might exist rich inter-pathway crosstalks.

5.3. Human gene network analysis

Next, we apply the TIGER method to analyze a publicly available human gene expression data. This data has been described and studied by Bhadra and
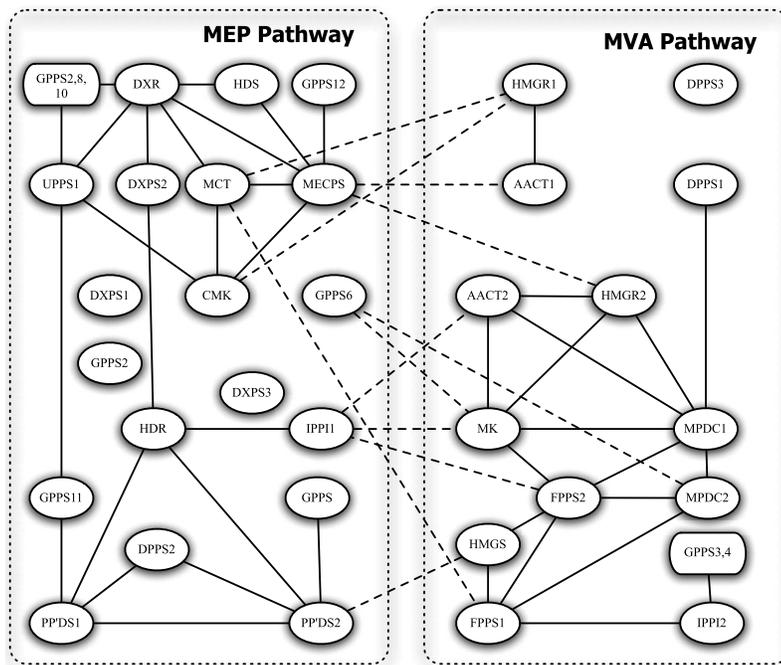
FIG 7. *The estimated gene networks of the Arabadopsis dataset. The within-pathway links are denoted by solid lines and between-pathway links are denoted by dashed lines.*

Mallick (2013); Mohammadi and Wit (2015); Chen et al. (2008); Stranger et al. (2007). This dataset contains 60 unrelated individuals of Northern and Western European ancestry from Utah (CEU), whose genotypes are available from the Sanger Institute website[2]. Illumina's Sentrix Human-6 Expression BeadChips are used to measure gene expression in B-lymphocyte cells for all the individuals (Stranger et al., 2007). The genotype for rare homozygous, heterozygous and homozygous common alleles are coded by 0, 1, and 2, respectively. The raw data were background corrected and then quantile normalized across four replicates of a single indi- vidual and then median normalized across all individuals. 100 most variable probes among the 47, 293 total available probes corresponding to different Illumina TargetID were chased for the analysis (Bhadra and Mallick, 2013; Mohammadi and Wit, 2015). Each selected probe corresponds to a different transcript. The data can be found in the R package BDgraph. The dataset is thus of size $n = 60$ and $d = 100$. Bhadra and Mallick (2013) have analyzed the data, and they found 55 significant interactions among the 100 chosen traits. Mohammadi and Wit (2015) exploited a Bayesian method to infer the gene network and the inferred network corresponding to the highest posterior density contains 281 edges. he identified 281 interactions include almost all the

---

[2]ftp://ftp.sanger.ac.uk/pub/genevar.

significant interactions discovered by Bhadra and Mallick (2013). Among these 281 edges, 86 edges were deemed as significant interactions (with a posterior probability > 0.6).

We apply the TIGER method on the same data using the default parameter $\zeta = \sqrt{2}/\pi$ estimated network for all the 100 chosen traits is shown in Figure 8. The estimated network contains 306 edges. It includes almost all the interactions discovered by Bhadra and Mallick (2013) (40 out of 55), and it also includes the majority of the interactions found by Mohammadi and Wit (2015) (70 out of 86). This shows the efficacy of our proposed tuning-insensitive approach. It is also computationally faster.

6. Conclusions and discussion

We introduce a tuning-insensitive approach named TIGER for estimating high dimensional Gaussian graphical models. Our method is asymptotically tuning-free and simultaneously achieves the minimax optimal rates of convergence in precision matrix estimation under different matrix norms (matrix L_1, spetral, Frobenius, and elementwise sup-norm). Computationally, our procedure is significantly faster than existing methods due to its tuning-insensitive property. The advantages of our estimators are also illustrated using both simulated and real data examples. The TIGER approach is a very competitive alternative for estimating high dimensional Gaussian graphical models.

There are several possible directions to expand the current methods. First, it is interesting to extend the TIGER to the nonparanomral setting for estimating high dimensional Gaussian copula graphical models (Liu et al., 2012). Second, it is also interesting to extend the TIGER to more complex settings where latent variables or missing data might exist.

Appendix A: Proofs

We begin with everal important lemmas, followed by the proof of the main results. For notational simplicity, we use a generic constant C whose value may vary from line to line.

A.1. Preliminaries

For any set $S \subset \{1, 2, \cdots, d\}$ and $|S| \leq k$, let $\boldsymbol{\Delta}_{\bar{c}}^d(S)$ denote a subset of $\mathbb{R}^d$ defined as

$$\boldsymbol{\Delta}_{\bar{c}}^d(S) := \left\{ \boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{\beta}_{S^c}\|_1 \leq \bar{c}\|\boldsymbol{\beta}_S\|_1, \boldsymbol{\beta} \neq \mathbf{0} \right\}.$$

Also, we define

$$\boldsymbol{\Delta}_{\bar{c}}^d(k) := \bigcup_{S \subset \{1, 2, \cdots, d\}, |S| \leq k} \boldsymbol{\Delta}_{\bar{c}}^d(S).$$
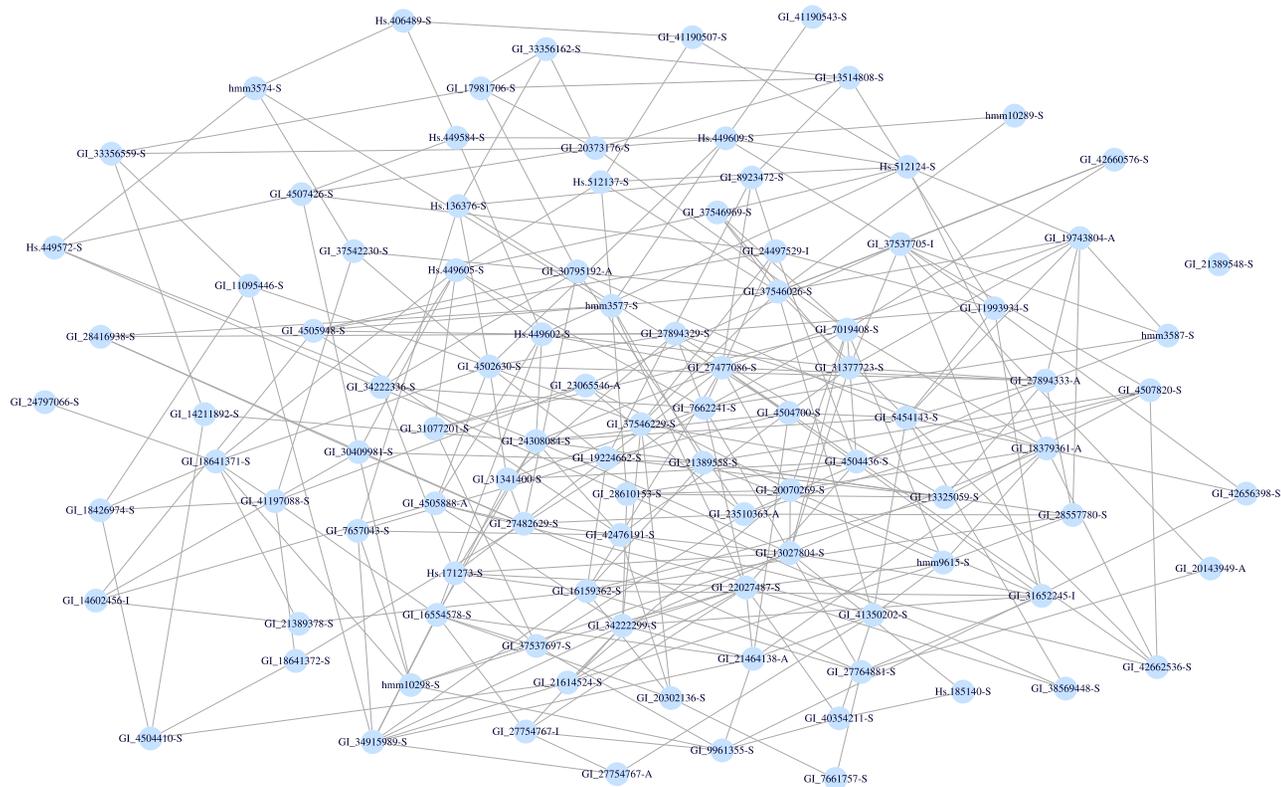
FIG 8. *The estimated networks of the human gene expression dataset. It includes all the 100 traits with 306 interactions. It includes almost all the interactions discovered by Bhadra and Mallick (2013) (40 out of 55), and it also includes the majority of the significant interactions identified by Mohammadi and Wit (2015) (70 out of 86).*

Recall that the matrix class $\mathcal{M}(\xi_{\max}, k)$ is defined as

$$\mathcal{M}(\xi_{\max}, k) \quad := \quad \left\{ \boldsymbol{\Theta} = \boldsymbol{\Theta}^T \in \mathbb{R}^{d \times d} : \boldsymbol{\Theta} \succ \mathbf{0}, \; \frac{\Lambda_{\max}(\boldsymbol{\Theta})}{\Lambda_{\min}(\boldsymbol{\Theta})} \leq \xi_{\max}, \right.$$
$$\left. \max_j \sum_k I(\boldsymbol{\Theta}_{jk} \neq 0) \leq k \right\},$$

let $\boldsymbol{\Theta} := \boldsymbol{\Sigma}^{-1} \in \mathcal{M}(\xi_{\max}, k)$, we have

$$\frac{\Lambda_{\max}(\boldsymbol{\Theta})}{\Lambda_{\min}(\boldsymbol{\Theta})} = \frac{\Lambda_{\max}(\boldsymbol{\Sigma}^{-1})}{\Lambda_{\min}(\boldsymbol{\Sigma}^{-1})} = \frac{\Lambda_{\max}(\boldsymbol{\Sigma})}{\Lambda_{\min}(\boldsymbol{\Sigma})} \leq \xi_{\max}.$$

We define the population correlation matrix $\mathbf{R}$ as

$$\mathbf{R} := [\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2} \boldsymbol{\Sigma} [\mathrm{diag}(\boldsymbol{\Sigma})]^{-1/2}.$$

We also define $\tau_j := \sigma_j \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}$. We recall that there are three assumptions:

(A1) $\boldsymbol{\Sigma}^{-1} \in \mathcal{M}(\xi_{\max}, k)$,
(A2) $k^2 \log d = o(n)$,
(A3) $\limsup_{n \to \infty} \max_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj}^2 \frac{\log d}{n} < \frac{1}{4}$.

We define

$$Q_j(\boldsymbol{\beta}_j) := \left\| \mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j} \boldsymbol{\beta}_j \right\|_2.$$

Therefore, $\hat{\boldsymbol{\beta}}_j$ from (3.5) can be written as

$$\hat{\boldsymbol{\beta}}_j = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{d-1}} \left\{ \frac{1}{\sqrt{n}} Q_j(\boldsymbol{\beta}_j) + \lambda \|\boldsymbol{\beta}\|_1 \right\}, \quad \text{for } j = 1, \ldots, d.$$

In the whole proof, for notational simplicity, we always denote the tuning parameter λ to be

$$\lambda = c \sqrt{\frac{2a \log d}{n}},$$

where $c > 1$ and $a > 2$. It is easy to see that $\lambda = \frac{\pi}{\sqrt{2}} \sqrt{\frac{\log d}{n}}$ is a special case of this setup.

A.2. Technical lemmas

Throughout this paper we often use one of the following tail bounds for central χ^2 random variables. We denote χ_d^2 to be a Chi-square variable with d degrees of freedom. Lemma A.1 presents some well known results of χ_d^2 and the proofs can be found in the original papers.

Lemma A.1 (Johnstone. (2000) and Laurent and Massart (1998))**.** *Let $X \sim \chi_d^2$. We have*

$$\max\left\{\mathbb{P}\left(X - d \geq 2\sqrt{dt} + 2t\right), \mathbb{P}\left(X - d \leq -2\sqrt{dt}\right)\right\} \leq \exp(-t)$$
$$\text{for all } t \geq 0, \tag{A.1}$$

$$\mathbb{P}\left(|X - d| > dt\right) \leq \exp\left(-\frac{3}{16}dt^2\right), \quad \text{for all } t \in \left[0, \frac{1}{2}\right), \tag{A.2}$$

$$\mathbb{P}\left(X \leq (1 - t)d\right) \leq \exp\left(-\frac{1}{4}dt^2\right), \quad \text{for all } t \in \left[0, \frac{1}{2}\right). \tag{A.3}$$

The next lemma bounds the tail of sample correlation for bivariate Gaussian random variables.

Lemma A.2. *Let $\boldsymbol{X} := (X_1, X_2)^T$ follows a bivariate normal distribution:*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N_2\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}\right).$$

Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^2$ be n independent data points from $\boldsymbol{X}$ and $\hat{\boldsymbol{\Sigma}} := \frac{1}{n}\sum_{i=1}^n \boldsymbol{X}_i \boldsymbol{X}_i^T$ be the sample covariance matrix. We define the sample and population correlations $\hat{\rho}$ and ρ as

$$\hat{\rho} := (\hat{\boldsymbol{\Sigma}}_{11})^{-1/2}\hat{\boldsymbol{\Sigma}}_{12}(\hat{\boldsymbol{\Sigma}}_{22})^{-1/2} \quad and \quad \rho := (\boldsymbol{\Sigma}_{11})^{-1/2}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22})^{-1/2}.$$

For any $t \in \left[0, \frac{1+|\rho|}{2}\right)$, we have

$$\mathbb{P}\left(|\hat{\rho} - \rho| > t\right) \leq 4\exp\left[-\frac{3nt^2}{64(1 + |\rho|)^2}\right].$$

Proof. Suppose that r is a small positive number such that $|\rho|r \leq t(1 - r)$, we have

$$\mathbb{P}\left(|\hat{\rho} - \rho| > t\right)$$
$$= \mathbb{P}\left(\left|(\hat{\boldsymbol{\Sigma}}_{11})^{-1/2}\hat{\boldsymbol{\Sigma}}_{12}(\hat{\boldsymbol{\Sigma}}_{22})^{-1/2} - \rho\right| > t\right)$$
$$= \mathbb{P}\left(\left|\hat{\boldsymbol{\Sigma}}_{12} - \rho(\hat{\boldsymbol{\Sigma}}_{11}\hat{\boldsymbol{\Sigma}}_{22})^{1/2}\right| > t(\hat{\boldsymbol{\Sigma}}_{11}\hat{\boldsymbol{\Sigma}}_{22})^{1/2}\right).$$

Let $r := \dfrac{t}{1 + 2|\rho| + t}$, we define the event

$$\mathcal{A} := \left\{\left|\hat{\boldsymbol{\Sigma}}_{11} - \boldsymbol{\Sigma}_{11}\right| \leq r\boldsymbol{\Sigma}_{11} \text{ and } \left|\hat{\boldsymbol{\Sigma}}_{22} - \boldsymbol{\Sigma}_{22}\right| \leq r\boldsymbol{\Sigma}_{22}\right\}.$$

On $\mathcal{A}$, we have

$$\left|(\hat{\boldsymbol{\Sigma}}_{11}\hat{\boldsymbol{\Sigma}}_{22})^{1/2} - (\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22})^{1/2}\right| \leq r(\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{22})^{1/2}.$$

Therefore, using (A.4), we have

$$
\begin{aligned}
&\mathbb{P}\Big(\big|\hat{\rho} - \rho\big| > t\Big)\\
&\leq \quad \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \rho(\hat{\mathbf{\Sigma}}_{11}\hat{\mathbf{\Sigma}}_{22})^{1/2}\big| > t(\hat{\mathbf{\Sigma}}_{11}\hat{\mathbf{\Sigma}}_{22})^{1/2}\,\big|\,\mathcal{A}\right) + \mathbb{P}(\mathcal{A}^c)\\
&\leq \quad \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \rho(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}\big| > [t(1-r) - |\rho|r](\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}\right) + \mathbb{P}(\mathcal{A}^c)\\
&= \quad \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \mathbf{\Sigma}_{12}\big| > [t(1-r) - |\rho|r](\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}\right) + \mathbb{P}(\mathcal{A}^c)\\
&\leq \quad \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \mathbf{\Sigma}_{12}\big| > [t(1-r) - |\rho|r](\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}\right) \qquad\qquad\text{(A.4)}\\
&\quad + \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{11} - \mathbf{\Sigma}_{11}\big| > r\mathbf{\Sigma}_{11}\right) + \mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{22} - \mathbf{\Sigma}_{22}\big| > r\mathbf{\Sigma}_{22}\right).
\end{aligned}
$$

Therefore, it suffices to analyze the terms $\mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \mathbf{\Sigma}_{12}\big| > \epsilon\right)$ and $\mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{11} - \mathbf{\Sigma}_{11}\big| > \epsilon\right)$. For this, we denote $\tilde{\boldsymbol{X}}_i := (\tilde{X}_{i1}, \tilde{X}_{i2})$, where

$$
\tilde{X}_{i1} := \frac{X_{i1}}{(\mathbf{\Sigma}_{11})^{1/2}} \sim N(0,1) \ \ \text{and} \ \ \tilde{X}_{i2} := \frac{X_{i2}}{(\mathbf{\Sigma}_{22})^{1/2}} \sim N(0,1).
$$

We have

$$
\sum_{i=1}^n \Big[\frac{(\tilde{X}_{i1} + \tilde{X}_{i2})^2}{2(1+\rho)}\Big] \sim \chi_n^2 \ \ \text{and} \ \ \sum_{i=1}^n \Big[\frac{(\tilde{X}_{i1} - \tilde{X}_{i2})^2}{2(1-\rho)}\Big] \sim \chi_n^2.
$$

Furthermore, since $\sum_{i=1}^n \tilde{X}_{i1}^2 \sim \chi_n^2$, we have

$$
\begin{aligned}
\mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{11} - \mathbf{\Sigma}_{11}\big| > \epsilon\right) &= \mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n \tilde{X}_{i1}^2 - 1\Big| > \frac{\epsilon}{\mathbf{\Sigma}_{11}}\right)\\
&= \mathbb{P}\left(\Big|\sum_{i=1}^n \tilde{X}_{i1}^2 - n\Big| > \frac{n\epsilon}{\mathbf{\Sigma}_{11}}\right)\\
&\leq \exp\left(-\frac{3n\epsilon^2}{16(\mathbf{\Sigma}_{11})^2}\right), \qquad\qquad\text{(A.5)}
\end{aligned}
$$

where the last inequality follows from (A.3).

Similarly,

$$
\begin{aligned}
&\mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \mathbf{\Sigma}_{12}\big| > \epsilon\right)\\
&= \mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n \tilde{X}_{i1}\tilde{X}_{i2} - \rho\Big| > \frac{\epsilon}{(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\right)\\
&= \mathbb{P}\left(\Big|\frac{1}{n}\sum_{i=1}^n \big[(\tilde{X}_{i1} + \tilde{X}_{i2})^2 - 2(1+\rho) - (\tilde{X}_{i1} - \tilde{X}_{i2})^2 + 2(1-\rho)\big]\Big|\right.\\
&\qquad\left. > \frac{4\epsilon}{(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\leq \quad \mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^{n}\big[(\tilde{X}_{i1}+\tilde{X}_{i2})^2 - 2(1+\rho)\big]\Big| > \frac{2\epsilon}{(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\bigg) \\
&\quad +\mathbb{P}\bigg(\Big|\frac{1}{n}\sum_{i=1}^{n}\big[(\tilde{X}_{i1}-\tilde{X}_{i2})^2 - 2(1-\rho)\big]\Big| > \frac{2\epsilon}{(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\bigg)
\end{aligned}
$$

$$
\leq \quad \mathbb{P}\bigg(\Big|\sum_{i=1}^{n}\big[\frac{(\tilde{X}_{i1}+\tilde{X}_{i2})^2}{2(1+\rho)}\big] - n\Big| > \frac{n\epsilon}{(1+\rho)(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\bigg) \tag{A.6}
$$

$$
+\mathbb{P}\bigg(\Big|\sum_{i=1}^{n}\big[\frac{(\tilde{X}_{i1}-\tilde{X}_{i2})^2}{2(1-\rho)}\big] - n\Big| > \frac{n\epsilon}{(1-\rho)(\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22})^{1/2}}\bigg). \tag{A.7}
$$

Applying (A.3) on (A.6) and (A.7), we have

$$
\begin{aligned}
\mathbb{P}\left(\big|\hat{\mathbf{\Sigma}}_{12} - \mathbf{\Sigma}_{12}\big| > \epsilon\right) &\leq \quad \exp\left(-\frac{3n\epsilon^2}{16(1+\rho)^2\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22}}\right) \\
&\quad + \exp\left(-\frac{3n\epsilon^2}{16(1-\rho)^2\mathbf{\Sigma}_{11}\mathbf{\Sigma}_{22}}\right). \tag{A.8}
\end{aligned}
$$

Combining (A.4) with (A.5) and (A.8), we get

$$
\mathbb{P}\left(\big|\hat{\rho} - \rho\big| > t\right) \leq 2\exp\left(-\frac{3n[t(1-r)-|\rho|r]^2}{16(1+|\rho|)^2}\right) + 2\exp\left(-\frac{3nr^2}{16}\right).
$$

Since

$$
r := \frac{t}{1+2|\rho|+t},
$$

we have

$$
\mathbb{P}\left(\big|\hat{\rho} - \rho\big| > t\right) \leq 4\exp\left(-\frac{3nt^2}{64(1+|\rho|)^2}\right).
$$

We thus complete the whole proof. □

The following lemma bounds the difference between the sample correlation matrix and the true correlation matrix in elementwise sup-nrom.

Lemma A.3. *Let $\hat{\mathbf{R}}$ and $\mathbf{R}$ be the sample and population correlation matrices. We define the event*

$$
\mathcal{A}_1 := \left\{\big\|\hat{\mathbf{R}} - \mathbf{R}\big\|_{\max} \leq 18\sqrt{\frac{\log d}{n}}\right\}.
$$

Then, we have $\mathbb{P}\left(\mathcal{A}_1\right) \geq 1 - 1/d$.

Proof. From Lemma A.2, we have

$$
\mathbb{P}\left(\big\|\hat{\mathbf{R}} - \mathbf{R}\big\|_{\max} > t\right) \leq 4d^2 \exp\left(-\frac{nt^2}{86}\right). \tag{A.9}
$$

The result follows by choosing $t = 18\sqrt{\dfrac{\log d}{n}}$. □

Lemma A.4. *Let $\boldsymbol{\epsilon}_j := (\epsilon_{j1}, \ldots, \epsilon_{jn})^T \in \mathbb{R}^n$ and $\boldsymbol{\epsilon}_j \sim N_n(\mathbf{0}, \sigma_j^2 \mathbf{I}_n)$. We define the event*

$$\mathcal{A}_2 := \left\{ \max_{1 \leq j \leq d} \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} \leq 1.4 \quad \text{and} \quad \max_{1 \leq j \leq d} \left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} - 1 \right| \leq 3.5 \sqrt{\frac{\log d}{n}} \right\}.$$

We have $\mathbb{P}(\mathcal{A}_2) \geq 1 - d\exp\left(-\dfrac{n}{100}\right) - \dfrac{1}{d}.$

Proof. Since

$$\frac{\|\boldsymbol{\epsilon}_j\|_2^2}{\sigma_j^2} = \sum_{i=1}^{n} \frac{\epsilon_{ij}^2}{\sigma_j^2} \sim \chi_n^2,$$

by Lemma A.1, it is easy to see that for any constant $1 \leq w < 1.5$,

$$\mathbb{P}\left(\frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} \leq w \right) \geq 1 - \exp\left(\frac{-3n(w-1)^2}{16} \right),$$

for all $j = 1, 2, \cdots, d$. By setting $w = 1.4$, we have

$$\mathbb{P}\left(\max_{1 \leq j \leq d} \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} \leq 1.4 \right) \geq 1 - d\exp\left(-\frac{n}{100} \right). \tag{A.10}$$

Similarly, for $t \in [0, 1/2)$, we have

$$\mathbb{P}\left(\left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n} - \sigma_j^2 \right| > t\sigma_j^2 \right) = \mathbb{P}\left(\left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{\sigma_j^2} - n \right| > nt \right) \leq \exp\left(-\frac{3}{16} nt^2 \right).$$

By setting $t = 3.5\sqrt{\dfrac{\log d}{n}}$, we have

$$\mathbb{P}\left(\max_{1 \leq j \leq d} \left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} - 1 \right| \leq 3.5\sqrt{\frac{\log d}{n}} \right) \geq 1 - \frac{1}{d}. \tag{A.11}$$

The desired result of this lemma follows from a union bound of (A.10) and (A.11). $\qquad\square$

The next lemma bounds the sample standard deviation of each marginal univariate Gaussian random variable.

Lemma A.5. *Let $\hat{\boldsymbol{\Sigma}}$ be the sample covariance matrix. Under the assumption that*

$$\limsup_{n \to \infty} \max_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj} \sqrt{\frac{\log d}{n}} < \frac{1}{2}, \tag{A.12}$$

we define the event

$$\mathcal{A}_3 := \left\{ \frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}) \leq \min_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq \max_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq \frac{3}{2} \Lambda_{\max}(\boldsymbol{\Sigma}) \right\}.$$

We have $\mathbb{P}(\mathcal{A}_3) \geq 1 - 1/d.$

Proof. By the definitions of $\Lambda_{\max}(\boldsymbol{\Sigma})$ and $\Lambda_{\min}(\boldsymbol{\Sigma})$, we have

$$\Lambda_{\max}(\boldsymbol{\Sigma}) \geq \max_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj} \geq \min_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj} \geq \Lambda_{\min}(\boldsymbol{\Sigma}).$$

By (A.5), we know that for any $j \in \{1, \ldots, d\}$ and $0 \leq \epsilon < 1/2$,

$$\mathbb{P}\left(\left|\hat{\boldsymbol{\Sigma}}_{jj} - \boldsymbol{\Sigma}_{jj}\right| > \epsilon\right) \leq \exp\left(-\frac{3n\epsilon^2}{16(\boldsymbol{\Sigma}_{jj})^2}\right). \tag{A.13}$$

We choose $\epsilon = t\boldsymbol{\Sigma}_{jj}$, then

$$\mathbb{P}\left(\min_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj}(1-t) \leq \hat{\boldsymbol{\Sigma}}_{jj} \leq (1+t) \max_{1 \leq j \leq d} \boldsymbol{\Sigma}_{jj}\right) \leq \exp\left(-\frac{3nt^2}{16}\right),$$

By the union bound, we have

$$\mathbb{P}\left((1-t)\Lambda_{\min}(\boldsymbol{\Sigma}) \leq \min_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq \max_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq (1+t)\Lambda_{\max}(\boldsymbol{\Sigma})\right)$$
$$\leq d \exp\left(-\frac{3nt^2}{16}\right).$$

Under the assumption (A.12), we know that, for large enough n, there must be $\epsilon = t\boldsymbol{\Sigma}_{jj} < 1/2$. The desired result of the lemma follows by setting $t = 3.5\sqrt{\dfrac{\log d}{n}}$. $\qquad\square$

Recall that we define

$$Q_j(\boldsymbol{\beta}_j) := \left\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\right\|_2,$$

the next lemma provides theoretical justification to the choice of the tuning parameter λ.

Lemma A.6. *Let* $\lambda = c\sqrt{\dfrac{2a \log d}{n}}$ *with* $c > 1$ *and* $a > 2$, *we define an event*

$$\mathcal{A}_4 := \left\{\max_{1 \leq j \leq d} c\left\|\nabla Q_j(\boldsymbol{\beta}_j)\right\|_\infty \leq \lambda\sqrt{n}\right\}.$$

Then

$$\mathbb{P}(\mathcal{A}_4) \geq 1 - \sqrt{\frac{2}{\pi a \log d}} \cdot d^{2-a\left(1-2\sqrt{\frac{(a-1)\log d}{n}}\right)} - \frac{1}{d^{a-2}}.$$

Proof. Using the fact that $\mathbf{Z}_{*j} = \mathbf{X}_{*j}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}$ and $\mathbf{Z}_{*\backslash j} = \mathbf{X}_{*\backslash j}\hat{\boldsymbol{\Gamma}}_{\backslash j, \backslash j}^{-1/2}$, we have

$$\mathbf{Z}_{*j} = \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j,$$

where $\boldsymbol{\epsilon}_j \sim N_n(\mathbf{0}, \sigma_j^2\mathbf{I}_n)$. We then have

$$\left\|\nabla Q_j(\boldsymbol{\beta}_j)\right\|_\infty = \frac{\left\|\mathbf{Z}_{*\backslash j}^T(\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j)\right\|_\infty}{\left\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\right\|_2}$$

$$= \frac{\left\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\right\|_\infty}{\left\|\boldsymbol{\epsilon}_j\right\|_2}$$

From the properties of multivariate Gaussian, we know that $\mathbf{Z}_{*\backslash j}$ and $\boldsymbol{\epsilon}_j$ are independent. For any $\ell \neq j$, $\mathbf{Z}_{*\ell}^T \boldsymbol{\epsilon}_j$ follows $N(0, n\sigma_j^2)$ distribution when conditioning on $\mathbf{Z}_{*\backslash j}$. In the following argument, we suppose everything is conditioning on $\mathbf{Z}_{*\backslash j}$. Since $\|\boldsymbol{\epsilon}_j\|_2^2/\sigma_j^2 \sim \chi_n^2$, by Lemma A.1, we know that, for any $0 \leq r_n < 1/2$,

$$\mathbb{P}\Big(\max_{1\leq j\leq d} \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} \leq 1 - r_n\Big) \leq d\exp\left(-\frac{nr_n^2}{4}\right).$$

Let $\Phi(\cdot)$ and $\phi(\cdot)$ be the cumulative and density functions of a standard Gaussian random variable. For any $0 \leq r_n < 1/2$, we have

$$\mathbb{P}\left(\max_{1\leq j\leq d} \frac{\left\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\right\|_\infty}{\left\|\boldsymbol{\epsilon}_j\right\|_2} > \sqrt{2a\log d}\right)$$

$$\leq d\mathbb{P}\left(\left\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\right\|_\infty > \sqrt{1-r_n}\sqrt{2an\sigma_j^2\log d}\right) + \mathbb{P}\Big(\max_{1\leq j\leq d} \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n\sigma_j^2} \leq 1-r_n\Big)$$

$$\leq d\sum_{\ell\neq j}\mathbb{P}\left(|\mathbf{Z}_{*\ell}^T\boldsymbol{\epsilon}_j| > \sqrt{1-r_n}\sqrt{2an\sigma_j^2\log d}\right) + d\exp\left(-\frac{nr_n^2}{4}\right)$$

$$\leq 2d^2\big(1-\Phi(\sqrt{1-r_n}\sqrt{2a\log d})\big) + d\exp\left(-\frac{nr_n^2}{4}\right)$$

$$\leq 2d^2 \frac{d^{-a(1-r_n)}}{\sqrt{2\pi}\sqrt{1-r_n}\sqrt{2a\log d}} + d\exp\left(-\frac{nr_n^2}{4}\right)$$

$$= \frac{d^{2-a(1-r_n)}}{\sqrt{\pi(1-r_n)a\log d}} + d\exp\left(-\frac{nr_n^2}{4}\right).$$

where the second to last inequality follows from the fact that

$$1 - \Phi(t) \leq \frac{\phi(t)}{t} = \frac{1}{\sqrt{2\pi}t}\exp\left(-\frac{t^2}{2}\right),$$

whenever $t \geq 1$. Now let

$$r_n = 2\sqrt{\frac{(a-1)\log d}{n}},$$

it can be seen that, when n is large enough,

$$\mathbb{P}\left(\max_{1\leq j\leq d} \frac{\left\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\right\|_\infty}{\left\|\boldsymbol{\epsilon}_j\right\|_2} \leq \sqrt{2a\log d}\right)$$

$$\geq 1 - \sqrt{\frac{2}{\pi a\log d}}\cdot d^{2-a\left(1-2\sqrt{\frac{(a-1)\log d}{n}}\right)} - \frac{1}{d^{a-2}}.$$

We finish the proof of this lemma.   □

The next lemma shows that the error vector falls in a restricted set: $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j \in \boldsymbol{\Delta}_{\bar{c}}^{d-1}(k)$.

Lemma A.7. *Let* $\hat{\beta}_j$ *be defined as in* (3.5) *and* $\bar{c} = \frac{c+1}{c-1}$. *Then, on the event* $\mathcal{A}_4$, *we have* $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j \in \boldsymbol{\Delta}_{\bar{c}}^{d-1}(k)$ *for all* $j = 1, \ldots, d$.

Proof. Since $\hat{\boldsymbol{\beta}}_j$ is the empirical minimizer of the objective function in (3.5), we have

$$\frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j\big\|_2 + \lambda\big\|\hat{\boldsymbol{\beta}}_j\big\|_1 \leq \frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\big\|_2 + \lambda\big\|\boldsymbol{\beta}_j\big\|_1. \quad (A.14)$$

Let $S := \{\ell : \beta_{j\ell} \neq 0\}$, it is obvious that $|S| \leq k$. From (A.14), we get

$$\frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j\big\|_2 - \frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\big\|_2$$
$$\leq \quad \lambda\big\|\boldsymbol{\beta}_j\big\|_1 - \lambda\big\|\hat{\boldsymbol{\beta}}_j\big\|_1$$
$$\leq \quad \lambda\big\|(\boldsymbol{\beta}_j)_S\big\|_1 - \lambda\big\|(\hat{\boldsymbol{\beta}}_j)_S\big\|_1 - \lambda\big\|(\hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1$$
$$\leq \quad \lambda\Big(\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1 - \big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1\Big).$$

On the event $\mathcal{A}_4$, we have $c\big\|\nabla Q_j(\boldsymbol{\beta}_j)\big\|_\infty \leq \lambda\sqrt{n}$, where

$$Q_j(\boldsymbol{\beta}_j) := \big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\big\|_2.$$

Therefore, using the fact that $Q_j(\cdot)$ is a convex function

$$-\lambda\Big(\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1 + \big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1\Big)$$
$$= \quad -\lambda\big\|\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\big\|_1$$
$$\leq \quad -\frac{c}{\sqrt{n}}\big\|\nabla Q_j(\boldsymbol{\beta}_j)\big\|_\infty\big\|\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j\big\|_1$$
$$\leq \quad -\frac{c}{\sqrt{n}}(\nabla Q_j(\boldsymbol{\beta}_j))^T(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)$$
$$\leq \quad \frac{c}{\sqrt{n}}\big(Q_j(\hat{\boldsymbol{\beta}}_j) - Q_j(\boldsymbol{\beta}_j)\big)$$
$$= \quad c\Big(\frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j\big\|_2 - \frac{1}{\sqrt{n}}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j\big\|_2\Big).$$

Combining the above analysis, we have

$$-\Big(\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1 + \big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1\Big) \leq c\Big(\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1 - \big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1\Big).$$

Therefore,

$$\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_{S^c}\big\|_1 \leq \frac{c+1}{c-1}\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1 = \bar{c}\big\|(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)_S\big\|_1.$$

We finish the proof of this lemma. $\qquad\qquad\square$

Appendix B: Proof of matrix restricted eigenvalue conditions

The following lemma bounds the restricted eigenvalue of the sample correlation matrix $\hat{\mathbf{R}}$.

Lemma B.1. *Let the event $\mathcal{A}_3$ be defined as in Lemma A.5. We assume $k \log d = o(n)$ and define an event*

$$\mathcal{B}_1 := \left\{ \inf_{\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k \boldsymbol{\beta}^T \hat{\mathbf{R}} \boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_1} \geq \frac{1}{5(1 + \bar{c}) \xi_{\max}^{1/2}} \right\}.$$

Then, there exist constants c_1 and c_2, such that $\mathbb{P}\left(\mathcal{B}_1 \,|\, \mathcal{A}_3\right) \geq 1 - c_1 \exp\left(-c_2 n\right)$.

Proof. For any $S \subset \{1, 2, \cdots, d\}$ with $|S| \leq k$, we have, for any $\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^d(S)$,

$$\|\boldsymbol{\beta}\|_1 \leq (1 + \bar{c}) \|\boldsymbol{\beta}_S\|_1 \leq (1 + \bar{c}) \sqrt{k} \|\boldsymbol{\beta}_S\|_2 \leq (1 + \bar{c}) \sqrt{k} \|\boldsymbol{\beta}\|_2, \tag{B.1}$$

and

$$\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta} \geq \Lambda_{\min}(\boldsymbol{\Sigma}) \|\boldsymbol{\beta}\|_2^2 \geq \Lambda_{\min}(\boldsymbol{\Sigma}) \|\boldsymbol{\beta}_S\|_2^2 \geq \Lambda_{\min}(\boldsymbol{\Sigma}) \frac{\|\boldsymbol{\beta}\|_1^2}{k(1 + \bar{c})^2},$$

where the last inequality uses the fact that $\|\boldsymbol{\beta}\|_1 \leq (1 + \bar{c}) \sqrt{k} \|\boldsymbol{\beta}_S\|_2$. This result is obtained from (B.1).

Recall that $\mathbf{X} \in \mathbb{R}^{n \times d}$ is a matrix and the rows of $\mathbf{X}$ are independent $N(0, \boldsymbol{\Sigma})$ Gaussian random vectors, where $\boldsymbol{\Sigma}$ is the $d \times d$ covariance matrix. Let $\hat{\boldsymbol{\Sigma}}$ be the sample covariance matrix of $\mathbf{X}$. From Theorem 1 of Raskutti et al. (2010), we know that there exist two positive constants c_1 and c_2 such that

$$\mathbb{P}\left(\sqrt{\boldsymbol{\beta}^T \hat{\boldsymbol{\Sigma}} \boldsymbol{\beta}} \geq \frac{1}{4} \sqrt{\boldsymbol{\beta}^T \boldsymbol{\Sigma} \boldsymbol{\beta}} - 9 \max_{1 \leq j \leq d} \sqrt{\boldsymbol{\Sigma}_{jj}} \sqrt{\frac{\log d}{n}} \|\boldsymbol{\beta}\|_1, \forall \boldsymbol{\beta} \in \mathbb{R}^d \right)$$
$$\geq 1 - c_1 \exp\left(-c_2 n\right).$$

Let $\hat{\boldsymbol{\Gamma}} := \mathrm{diag}\left(\hat{\boldsymbol{\Sigma}}\right)$, we have, for any $\boldsymbol{\beta} \in \mathbb{R}^d$,

$$\mathbb{P}\Big(\sqrt{(\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta})^T \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta})}$$
$$\geq \frac{1}{4} \sqrt{(\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta})^T \boldsymbol{\Sigma} (\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta})} - 9 \max_{1 \leq j \leq d} \sqrt{\boldsymbol{\Sigma}_{jj}} \sqrt{\frac{\log d}{n}} \|\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta}\|_1 \Big)$$
$$\geq 1 - c_1 \exp\left(-c_2 n\right). \tag{B.2}$$

Since $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Gamma}}^{1/2} \hat{\mathbf{R}} \hat{\boldsymbol{\Gamma}}^{1/2}$, we have

$$(\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta})^T \hat{\boldsymbol{\Sigma}} (\hat{\boldsymbol{\Gamma}}^{-1/2} \boldsymbol{\beta}) = \boldsymbol{\beta}^T \hat{\mathbf{R}} \boldsymbol{\beta}. \tag{B.3}$$

On the event $\mathcal{A}_3$ as defined in Lemma A.5, we have

$$\frac{1}{2} \Lambda_{\min}(\boldsymbol{\Sigma}) \leq \min_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq \max_{1 \leq j \leq d} \hat{\boldsymbol{\Sigma}}_{jj} \leq \frac{3}{2} \Lambda_{\max}(\boldsymbol{\Sigma}).$$

It is easy to see that

$$\|\hat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\beta}\|_1 \leq \|\boldsymbol{\beta}\|_1 \min_{1\leq j\leq d}(\hat{\boldsymbol{\Sigma}}_{jj})^{-1/2} \leq \sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}}\|\boldsymbol{\beta}\|_1, \qquad \text{(B.4)}$$

$$\|\hat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\beta}\|_2 \geq \|\boldsymbol{\beta}\|_2 \max_{1\leq j\leq d}(\hat{\boldsymbol{\Sigma}}_{jj})^{-1/2} \geq \sqrt{\frac{2}{3\Lambda_{\max}(\boldsymbol{\Sigma})}}\|\boldsymbol{\beta}\|_2.$$

Therefore,

$$\sqrt{(\hat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\beta})^T\boldsymbol{\Sigma}(\hat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\beta})} \geq \sqrt{\Lambda_{\min}(\boldsymbol{\Sigma})}\|\hat{\boldsymbol{\Gamma}}^{-1/2}\boldsymbol{\beta}\|_2 \geq \sqrt{\frac{2\Lambda_{\min}(\boldsymbol{\Sigma})}{3\Lambda_{\max}(\boldsymbol{\Sigma})}}\|\boldsymbol{\beta}\|_2. \quad \text{(B.5)}$$

Plugging (B.3), (B.4) and (B.5) into (B.2), we have

$$\mathbb{P}\left(\sqrt{\boldsymbol{\beta}^T\hat{\mathbf{R}}\boldsymbol{\beta}} \geq \frac{1}{4}\sqrt{\frac{2\Lambda_{\min}(\boldsymbol{\Sigma})}{3\Lambda_{\max}(\boldsymbol{\Sigma})}}\|\boldsymbol{\beta}\|_2 - 9\sqrt{\frac{\max_{1\leq j\leq d}\boldsymbol{\Sigma}_{jj}}{\Lambda_{\min}(\boldsymbol{\Sigma})}}\sqrt{\frac{2\log d}{n}}\|\boldsymbol{\beta}\|_1,\right.$$

$$\left.\forall\boldsymbol{\beta}\in\mathbb{R}^d\right) \geq 1 - c_1\exp(-c_2 n).$$

By (B.1), for any $\boldsymbol{\beta}\in\boldsymbol{\Delta}_{\bar{c}}^d(k)$, we have

$$\|\boldsymbol{\beta}\|_2 \geq \frac{1}{(1+\bar{c})\sqrt{k}}\|\boldsymbol{\beta}\|_1.$$

Therefore,

$$\mathbb{P}\left(\inf_{\boldsymbol{\beta}\in\boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k\boldsymbol{\beta}^T\hat{\mathbf{R}}\boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_1} \geq \frac{1}{4(1+\bar{c})}\sqrt{\frac{2\Lambda_{\min}(\boldsymbol{\Sigma})}{3\Lambda_{\max}(\boldsymbol{\Sigma})}} - 9\xi_{\max}^{1/2}\sqrt{\frac{2k\log d}{n}}\right)$$

$$\geq 1 - c_1\exp(-c_2 n).$$

Since we assume $k\log d = o(n)$, for n large enough, we have

$$\frac{1}{4(1+\bar{c})}\sqrt{\frac{2\Lambda_{\min}(\boldsymbol{\Sigma})}{3\Lambda_{\max}(\boldsymbol{\Sigma})}} - 9\xi_{\max}^{1/2}\sqrt{\frac{2k\log d}{n}} \geq \frac{1}{5(1+\bar{c})\xi_{\max}^{1/2}}.$$

We finish the proof of this lemma. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

We can see that when $k\log d = o(n)$ and n large enough, with high probability,

$$\eta(\hat{\mathbf{R}}) := \inf_{\boldsymbol{\beta}\in\boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k\boldsymbol{\beta}^T\hat{\mathbf{R}}\boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_1} \geq \frac{1}{5(1+\bar{c})\xi_{\max}^{1/2}}$$

is bounded away from zero by a positive constant. This implies the restricted eigenvalue condition required by the SQRT-Lasso method, which is provided in the next lemma.

Lemma B.2. *We define the event*

$$\mathcal{B}_2 := \left\{ \max_{1 \le j \le d} \frac{\left\| \mathbf{Z}_{*\setminus j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j) \right\|_2}{\tau_j} \le C \cdot \sqrt{k \log d} \right\}.$$

Then $\mathbb{P}(\mathcal{B}_2 \,|\, \mathcal{B}_1) \ge 1 - 1/d.$

Proof. We only need to verify that the L_1-restricted eigenvalue condition of the SQRT-Lasso is satisfied. More specifically, we need to verify the conditions in Theorem 1 of Belloni et al. (2012). Since $\hat{\mathbf{R}} = \dfrac{1}{n}\mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z}$ is defined as in Section 3, we have proved that, on $\mathcal{B}_1$,

$$\inf_{\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k}\|\mathbf{Z}\boldsymbol{\beta}\|_2}{\sqrt{n}\|\boldsymbol{\beta}\|_1} = \inf_{\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k\boldsymbol{\beta}^T\hat{\mathbf{R}}\boldsymbol{\beta}}}{\|\boldsymbol{\beta}\|_1} \ge \frac{1}{5(1+\bar{c})\xi_{\max}^{1/2}}.$$

Recall that $\mathbf{Z}_{*\setminus j}$ is the $n \times (d-1)$ submatrix of $\mathbf{Z}$ with the jth column removed, then

$$\min_{1 \le j \le d} \inf_{\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^{d-1}(k)} \frac{\sqrt{k}\|\mathbf{Z}_{*\setminus j}\boldsymbol{\beta}\|_2}{\sqrt{n}\|\boldsymbol{\beta}\|_1} \ge \inf_{\boldsymbol{\beta} \in \boldsymbol{\Delta}_{\bar{c}}^d(k)} \frac{\sqrt{k}\|\mathbf{Z}\boldsymbol{\beta}\|_2}{\sqrt{n}\|\boldsymbol{\beta}\|_1} \ge \frac{1}{5(1+\bar{c})\xi_{\max}^{1/2}}.$$

Therefore, on the event $\mathcal{B}_1$, the L_1-restricted eigenvalue condition holds for all the SQRT-Lasso subproblem defined in (3.5). The desired result follows from Theorem 1 of Belloni et al. (2012). □

Appendix C: Main lemmas

We define the event $\mathcal{E} := \mathcal{A}_1 \bigcap \mathcal{A}_2 \bigcap \mathcal{A}_3 \bigcap \mathcal{A}_4 \bigcap \mathcal{B}_1 \bigcap \mathcal{B}_2$. It is easy to see that $\mathbb{P}(\mathcal{E}) \ge 1 - o(1)$. To prove the main results, we separately analyze the diagonal and off-diagonal elements of $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$. In the next subsection, we first control the diagonal elements.

C.1. Analyzing the diagonal elements

Lemma C.1. *On the event $\mathcal{E}$, we have*

$$\max_{1 \le j \le d} \bigl|\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}\bigr| \le C \cdot \|\boldsymbol{\Theta}\|_2 \sqrt{\frac{\log d}{n}},$$

for large enough n.

Proof. Recall that $\mathbf{Z}_{*j} = \mathbf{Z}_{*\setminus j}\boldsymbol{\beta}_j + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j$, we have

$$\left| (\hat{\boldsymbol{\Theta}}_{jj})^{-1} - (\boldsymbol{\Theta}_{jj})^{-1} \right|$$
$$= \left| \hat{\boldsymbol{\Gamma}}_{jj}\hat{\tau}_j^2 - \hat{\boldsymbol{\Gamma}}_{jj}\tau_j^2 \right|$$

$$
= \left| \hat{\boldsymbol{\Gamma}}_{jj} \frac{\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j\|_2^2}{n} - \hat{\boldsymbol{\Gamma}}_{jj}\tau_j^2 \right|
$$

$$
= \left| \hat{\boldsymbol{\Gamma}}_{jj} \frac{\|\mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j) + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j\|_2^2}{n} - \sigma_j^2 \right|
$$

$$
\leq \left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n} - \sigma_j^2 \right| + \hat{\boldsymbol{\Gamma}}_{jj} \frac{\|\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2^2}{n} + 2\hat{\boldsymbol{\Gamma}}_{jj}^{1/2} \frac{|(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^T \mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j|}{n}.
$$

On the event $\mathcal{E}$, we have

$$
\frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n} \leq 1.4\sigma_j^2 \quad \text{and} \quad \left| \frac{\|\boldsymbol{\epsilon}_j\|_2^2}{n} - \sigma_j^2 \right| \leq 3.5\sigma_j^2 \sqrt{\frac{\log d}{n}},
$$

$$
\|\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2 \leq C \cdot \tau_j \sqrt{k \log d}, \tag{C.1}
$$

$$
\hat{\boldsymbol{\Gamma}}_{jj}^{1/2} \leq \max_{1 \leq j \leq d} (\hat{\boldsymbol{\Sigma}}_{jj})^{1/2} \leq \sqrt{\frac{3\Lambda_{\max}(\boldsymbol{\Sigma})}{2}},
$$

$$
\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} = (\hat{\boldsymbol{\Sigma}}_{jj})^{-1/2} \leq \sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}}.
$$

Futhermore, since $\tau_j^2 = \sigma_j^2 \hat{\boldsymbol{\Gamma}}_{jj}^{-1}$, we have

$$
\hat{\boldsymbol{\Gamma}}_{jj}\|\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2^2 \leq C^2 \cdot \sigma_j^2 \cdot k \log d.
$$

We also have

$$
2\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}|(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^T \mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j| \leq 2\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1 \cdot \|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\|_\infty.
$$

Since $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j \in \boldsymbol{\Delta}_{\bar{c}}^{d-1}(k)$, by the definition of the L_1-restricted eigenvalue, we know that

$$
\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\|_1
$$
$$
\leq \hat{\boldsymbol{\Gamma}}_{jj}^{1/2} \cdot 5(1+\bar{c})\xi_{\max}^{1/2} \cdot \sqrt{k(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^T \hat{\mathbf{R}}_{\backslash j,\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)}
$$
$$
\leq \hat{\boldsymbol{\Gamma}}_{jj}^{1/2} \cdot 5(1+\bar{c})\xi_{\max}^{1/2} \cdot \sqrt{\frac{k}{n}} \cdot \|\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2
$$
$$
\leq \hat{\boldsymbol{\Gamma}}_{jj}^{1/2} \cdot 5(1+\bar{c})\xi_{\max}^{1/2} \cdot \sqrt{\frac{k}{n}} \cdot C \cdot \sigma_j \Gamma_{jj}^{-1/2}\sqrt{k \log d}
$$
$$
\leq 5(1+\bar{c}) \cdot C \cdot \xi_{\max}^{1/2} \cdot \sigma_j \cdot k\sqrt{\frac{\log d}{n}}.
$$

Similarly, on the event $\mathcal{E}$, we have

$$
\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\|_\infty \leq \frac{1}{c}\|\boldsymbol{\epsilon}_j\|_2 \lambda \sqrt{n} = \sigma_j \sqrt{2.8 n \log d}.
$$

Therefore, on $\mathcal{E}$,

$$
\left| (\hat{\boldsymbol{\Theta}}_{jj})^{-1} - (\boldsymbol{\Theta}_{jj})^{-1} \right|
$$

$$\leq \quad 3.5\sigma_j^2\sqrt{\frac{\log d}{n}} + C^2\sigma_j^2 \cdot \frac{k\log d}{n} + 10\sqrt{2.8} \cdot (1+\bar{c}) \cdot C\xi_{\max}^{1/2}\sigma_j^2\frac{k\log d}{n}.$$

Since $k\sqrt{\dfrac{\log d}{n}} = o(1)$, there exists a constant C such that, for large enough n,

$$\left|(\hat{\boldsymbol{\Theta}}_{jj})^{-1} - (\boldsymbol{\Theta}_{jj})^{-1}\right| \leq C\sigma_j^2 \cdot \sqrt{\frac{\log d}{n}}.$$

Multiplying $\boldsymbol{\Theta}_{jj}$ on both sides and using the fact that $\boldsymbol{\Theta}_{jj} = \sigma_j^{-2}$, we have

$$\left|\boldsymbol{\Theta}_{jj}(\hat{\boldsymbol{\Theta}}_{jj})^{-1} - 1\right| \leq C\sqrt{\frac{\log d}{n}}. \tag{C.2}$$

This implies that

$$\left(1 + C\sqrt{\frac{\log d}{n}}\right)^{-1} \leq \frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} \leq \left(1 - C\sqrt{\frac{\log d}{n}}\right)^{-1}$$

Since, for large enough n,

$$1 - C\sqrt{\frac{\log d}{n}} \leq \left(1 + C\sqrt{\frac{\log d}{n}}\right)^{-1} \quad \text{and} \quad \left(1 - C\sqrt{\frac{\log d}{n}}\right)^{-1} \leq 1 + 2C\sqrt{\frac{\log d}{n}},$$

we have

$$\left(1 - C\sqrt{\frac{\log d}{n}}\right) \leq \frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} \leq \left(1 + C\sqrt{\frac{\log d}{n}}\right).$$

This implies that

$$\max_{1\leq j\leq d}\left|\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}\right| \leq C \cdot \|\boldsymbol{\Theta}\|_2\sqrt{\frac{\log d}{n}}.$$

The last inequality follows from the fact that $\max_{1\leq j\leq d}\boldsymbol{\Theta}_{jj} \leq \|\boldsymbol{\Theta}\|_2$. $\qquad\qquad\square$

C.2. Analyzing the off-diagonal elements in L_1-norm error

We first bound the L_1-norm of each column of the off-diagonal elements of $\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}$.

Lemma C.2. *On the event $\mathcal{E}$, we have*

$$\max_{1\leq j\leq d}\left\|\hat{\boldsymbol{\Theta}}_{\backslash j,j} - \boldsymbol{\Theta}_{\backslash j,j}\right\|_1 \leq C\left(\|\boldsymbol{\Theta}\|_2 k + \|\boldsymbol{\Theta}\|_1\right)\sqrt{\frac{\log d}{n}},$$

$$\max_{1\leq j\leq d}\left\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right\|_1 \leq 5\sqrt{2}C(1+\bar{c}).\xi_{\max}k\sqrt{\frac{\log d}{n}}, \tag{C.3}$$

for large enough n.

Proof. We recall that

$$\boldsymbol{\Theta}_{\backslash j,j} = -\boldsymbol{\Theta}_{jj}\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\boldsymbol{\beta}_j.$$

Since $\tau_j^{-2} := \sigma_j^{-2}\hat{\boldsymbol{\Gamma}}_{jj} = \boldsymbol{\Theta}_{jj}\hat{\boldsymbol{\Gamma}}_{jj}$, we have

$$
\begin{aligned}
&\left\|\hat{\boldsymbol{\Theta}}_{\backslash j,j} - \boldsymbol{\Theta}_{\backslash j,j}\right\|_1 \\
&= \left\|\hat{\tau}_j^{-2}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\hat{\boldsymbol{\beta}}_j - \tau_j^{-2}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\boldsymbol{\beta}_j\right\|_1 \\
&= \left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}(\hat{\boldsymbol{\Theta}}_{jj}\hat{\boldsymbol{\beta}}_j - \boldsymbol{\Theta}_{jj}\boldsymbol{\beta}_j)\right\|_1 \\
&\leq \left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\right\|_1 |\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}|\left\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right\|_1 + \hat{\boldsymbol{\Gamma}}_{jj}^{1/2}|\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}|\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\boldsymbol{\beta}_j\right\|_1 \\
&= \left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\right\|_1 |\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}|\left\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right\|_1 + |\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}|\left\|\boldsymbol{\Theta}_{\backslash j,j}\boldsymbol{\Theta}_{jj}^{-1}\right\|_1 \\
&\leq \left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\right\|_1 |\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}|\left\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right\|_1 + \left|\frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} - 1\right|\left\|\boldsymbol{\Theta}_{\backslash j,j}\right\|_1.
\end{aligned}
$$

On the event $\mathcal{E}$, using (C.2) and the analysis of Lemma C.1, we have, for large enough n,

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\right\|_1 &\leq \max_{1\leq j\leq d}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} = \max_{1\leq j\leq d}(\hat{\boldsymbol{\Sigma}}_{jj})^{-1/2} \leq \sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}}, \\
\hat{\boldsymbol{\Gamma}}_{jj}^{1/2} &\leq \sqrt{\frac{3}{2}\Lambda_{\max}(\boldsymbol{\Sigma})}, \\
\hat{\boldsymbol{\Theta}}_{jj} &\leq \left(1 + C\sqrt{\frac{\log d}{n}}\right)\boldsymbol{\Theta}_{jj} \leq 2\|\boldsymbol{\Theta}\|_2, \\
\left|\frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} - 1\right| &\leq C\sqrt{\frac{\log d}{n}}.
\end{aligned}
$$

By Lemma A.7, $\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j \in \boldsymbol{\Delta}_{\bar{c}}^{d-1}(k)$. By the definition of the L_1-restricted eigenvalue, we know that, on the event $\mathcal{B}_2$,

$$
\begin{aligned}
\left\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\right\|_1 &\leq 5(1+\bar{c})\xi_{\max}^{1/2}\cdot\sqrt{k(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)^T\hat{\mathbf{R}}_{\backslash j,\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)} \\
&\leq 5(1+\bar{c})\xi_{\max}^{1/2}\cdot\sqrt{\frac{k}{n}}\cdot\|\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\|_2 \\
&\leq 5(1+\bar{c})\xi_{\max}^{1/2}\cdot\sqrt{\frac{k}{n}}\cdot C\cdot\sigma_j\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\sqrt{k\log d} \\
&= 5(1+\bar{c})\cdot C\cdot\xi_{\max}^{1/2}\cdot\sigma_j\cdot\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\cdot k\sqrt{\frac{\log d}{n}} \\
&\leq 5\sqrt{2}C(1+\bar{c}).\xi_{\max}k\sqrt{\frac{\log d}{n}},
\end{aligned}
$$

where the last inequality follows from the fact that

$$\sigma_j \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2} \leq \sqrt{\boldsymbol{\Sigma}_{jj}} \cdot \sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}} \leq \sqrt{\frac{2\Lambda_{\max}(\boldsymbol{\Sigma})}{\Lambda_{\min}(\boldsymbol{\Sigma})}} \leq \sqrt{2}\xi_{\max}^{1/2}.$$

Therefore

$$\big\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\big\|_1 \big|\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}\big|\big\|\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j\big\|_1 \leq 10(1+\bar{c})\sqrt{3}\,\xi_{\max}\cdot C\cdot\sigma_j\big\|\boldsymbol{\Theta}\big\|_2\cdot\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\cdot k\sqrt{\frac{\log d}{n}}$$

and

$$\Big|\frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} - 1\Big|\big\|\boldsymbol{\Theta}_{\backslash j,j}\big\|_1 \leq C\cdot\big\|\boldsymbol{\Theta}\big\|_1\sqrt{\frac{\log d}{n}}.$$

Combining all the above analysis, we have

$$\big\|\hat{\boldsymbol{\Theta}}_{\backslash j,j} - \boldsymbol{\Theta}_{\backslash j,j}\big\|_1 \quad \leq \quad C\big\|\boldsymbol{\Theta}\big\|_2 k\sqrt{\frac{\log d}{n}} + C\cdot\big\|\boldsymbol{\Theta}\big\|_1\sqrt{\frac{\log d}{n}}.$$

We thus prove the desired result of the lemma.   □

C.3. Analyzing the off-diagonal elements in sup-norm error

To conduct sup-norm analysis of each of the column of the off-diagonal elements, we first present a technical lemma:

Lemma C.3. *On the event $\mathcal{E}$, there exists a constant C such that*

$$\Big\|\frac{\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}}{n}\mathbf{Z}_{*\backslash j}^T(\mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j)\Big\|_\infty \leq C\sigma_j\sqrt{\frac{\log d}{n}} \quad \text{for all } j = 1,\ldots,d.$$

Proof. Recall that we define

$$Q_j(\boldsymbol{\gamma}) = \big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\boldsymbol{\gamma}\big\|_2,$$

let $\lambda = c\sqrt{\frac{2a\log d}{n}}$ with $c > 1$ and $a > 2$, since $\hat{\boldsymbol{\beta}}_j$ is defined as:

$$\hat{\boldsymbol{\beta}}_j := \underset{\boldsymbol{\gamma}\in\mathbb{R}^{d-1}}{\operatorname{argmin}}\Big\{\frac{1}{\sqrt{n}}Q_j(\boldsymbol{\gamma}) + \lambda\|\boldsymbol{\gamma}\|_1\Big\},$$

we know that $c\|\nabla Q_j(\hat{\boldsymbol{\beta}}_j)\|_\infty \leq \sqrt{n}\lambda$. This means

$$\big\|\mathbf{Z}_{*,\backslash j}^T(\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j)\big\|_\infty \leq \sqrt{2a\log d}\big\|\mathbf{Z}_{*j} - \mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}\big\|_2.$$

Using the fact that $\mathbf{Z}_{*j} = \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j$, we have

$$\big\|\mathbf{Z}_{*\backslash j}^T\big(\mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j) + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j\big)\big\|_\infty \leq \sqrt{2a\log d}\big\|\mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j) + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\boldsymbol{\epsilon}_j\big\|_2.$$

Therefore

$$\left\|\mathbf{Z}_{*\backslash j}^T \mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)\right\|_\infty$$
$$\leq \quad \sqrt{2a\log d}\left\|\mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)\right\|_2 + \sqrt{2a\log d}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\left\|\boldsymbol{\epsilon}_j\right\|_2 + \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\left\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\right\|_\infty.$$

On the event $\mathcal{E}$, we have

$$\|\boldsymbol{\epsilon}_j\|_2 \leq \sigma_j \cdot \sqrt{1.4n},$$
$$\|\mathbf{Z}_{*\backslash j}^T \boldsymbol{\epsilon}_j\|_\infty \leq \sqrt{2a\log d}\|\boldsymbol{\epsilon}_j\|_2 \leq \sigma_j \cdot \sqrt{2.8an\log d},$$
$$\left\|\mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)\right\|_2 \leq C\tau_j\sqrt{k\log d} \leq C \cdot \sigma_j \cdot \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\sqrt{k\log d}.$$

By piecing all these terms together, we have

$$\left\|\mathbf{Z}_{*\backslash j}^T \mathbf{Z}_{*\backslash j}(\boldsymbol{\beta}_j - \hat{\boldsymbol{\beta}}_j)\right\|_\infty$$
$$\leq \quad C \cdot \sigma_j \cdot \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\sqrt{2ak} \cdot \log d + 2\sigma_j \cdot \hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\sqrt{2.8an\log d}.$$

Therefore

$$\left\|\frac{\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}}{n}\mathbf{Z}_{*\backslash j}^T(\mathbf{Z}_{*\backslash j}\hat{\boldsymbol{\beta}}_j - \mathbf{Z}_{*\backslash j}\boldsymbol{\beta}_j)\right\|_\infty \leq C \cdot \sigma_j \cdot \sqrt{2ak} \cdot \frac{\log d}{n} + 2\sigma_j\sqrt{\frac{2.8a\log d}{n}}.$$

The result follows from the fact that

$$\frac{\sqrt{k}\log d}{n} \leq \sqrt{\frac{\log d}{n}}$$

for large enough n. We finish the proof of this lemma. $\quad\square$

Lemma C.4. *On the event $\mathcal{E}$, we have*

$$\max_{1 \leq j \leq d}\left\|\hat{\boldsymbol{\Theta}}_{\backslash j,j} - \boldsymbol{\Theta}_{\backslash j,j}\right\|_\infty \leq C\|\boldsymbol{\Theta}\|_1\sqrt{\frac{\log d}{n}},$$

for large enough n.

Proof. We have the following decomposition

$$\left\|\hat{\boldsymbol{\Theta}}_{\backslash j,j} - \boldsymbol{\Theta}_{\backslash j,j}\right\|_\infty$$
$$= \quad \left\|\hat{\tau}_j^{-2}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\hat{\boldsymbol{\beta}}_j - \tau_j^{-2}\hat{\boldsymbol{\Gamma}}_{jj}^{-1/2}\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\boldsymbol{\beta}_j\right\|_\infty$$
$$= \quad \left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}(\hat{\boldsymbol{\Theta}}_{jj}\hat{\boldsymbol{\beta}}_j - \boldsymbol{\Theta}_{jj}\boldsymbol{\beta}_j)\right\|_\infty$$
$$\leq \quad \left|\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}\right|\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right\|_\infty + \hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\left|\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}\right|\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\boldsymbol{\beta}_j\right\|_\infty$$
$$= \quad \left|\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}\right|\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right\|_\infty + \left|\hat{\boldsymbol{\Theta}}_{jj} - \boldsymbol{\Theta}_{jj}\right|\left\|\boldsymbol{\Theta}_{\backslash j,j}\boldsymbol{\Theta}_{jj}^{-1}\right\|_\infty$$
$$\leq \quad \left|\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}\right|\left\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}(\hat{\boldsymbol{\beta}}_j - \boldsymbol{\beta}_j)\right\|_\infty + \left|\frac{\hat{\boldsymbol{\Theta}}_{jj}}{\boldsymbol{\Theta}_{jj}} - 1\right|\left\|\boldsymbol{\Theta}_{\backslash j,j}\right\|_\infty. \qquad (C.4)$$

Following a similar argument as in Lemma C.2, we have, on the event $\mathcal{E}$,

$$\Big|\frac{\hat{\Theta}_{jj}}{\Theta_{jj}} - 1\Big|\|\Theta_{\backslash j,j}\|_\infty \leq C \cdot \|\Theta\|_1\sqrt{\frac{\log d}{n}}. \tag{C.5}$$

Now for the first term in (C.4), we have

$$|\hat{\Gamma}_{jj}^{1/2}|\|\hat{\Gamma}_{\backslash j,\backslash j}^{-1/2}(\hat{\beta}_j - \beta_j)\|_\infty$$

$$= \|\hat{\Gamma}_{\backslash j,\backslash j}^{-1/2}\mathbf{R}_{\backslash j,\backslash j}^{-1}\mathbf{R}_{\backslash j,\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty \tag{C.6}$$

$$\leq \|\hat{\Gamma}_{\backslash j,\backslash j}^{-1/2}\mathbf{R}_{\backslash j,\backslash j}^{-1}\|_\infty\|\mathbf{R}_{\backslash j,\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty \tag{C.7}$$

$$\leq \|\hat{\Gamma}_{\backslash j,\backslash j}^{-1/2}\|_\infty\|\mathbf{R}_{\backslash j,\backslash j}^{-1}\|_\infty\|\mathbf{R}_{\backslash j,\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty. \tag{C.8}$$

To analyze the term $\|\mathbf{R}_{\backslash j,\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty$, we have

$$\|\mathbf{R}_{\backslash j,\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty$$

$$\leq \|(\mathbf{R}_{\backslash j,\backslash j} - \frac{1}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j})\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty + \|\frac{1}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j}\hat{\Gamma}_{jj}^{1/2}(\hat{\beta}_j - \beta_j)\|_\infty$$

$$\leq \|\mathbf{R}_{\backslash j,\backslash j} - \frac{1}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j}\|_{\max}|\hat{\Gamma}_{jj}^{1/2}|\|\hat{\beta}_j - \beta_j\|_1$$

$$+\|\frac{\hat{\Gamma}_{jj}^{1/2}}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j}(\hat{\beta}_j - \beta_j)\|_\infty. \tag{C.9}$$

From (C.3), we have, on the event $\mathcal{E}$:

$$\|\hat{\beta}_j - \beta_j\|_1 \leq 5\sqrt{2}C(1 + \bar{c}).\xi_{\max}k\sqrt{\frac{\log d}{n}}.$$

By Lemma A.3, we have, on the event $\mathcal{E}$:

$$\|\mathbf{R}_{\backslash j,\backslash j} - \frac{1}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j}\|_{\max} \leq \|\hat{\mathbf{R}} - \mathbf{R}\|_{\max} \leq 18\sqrt{\frac{\log d}{n}}.$$

On the other hand, let $\mathbf{A}$ be an invertible matrix, we have

$$1 = \|\mathbf{A}\mathbf{A}^{-1}\|_\infty \leq \|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_\infty,$$

which implies that

$$\frac{1}{\|\mathbf{A}\|_\infty} \leq \|\mathbf{A}^{-1}\|_\infty.$$

Therefore,

$$\|\mathbf{R}_{\backslash j,\backslash j}^{-1}\|_\infty = \min_{\beta \neq 0, \beta \in \mathbb{R}^{d-1}} \frac{\|\beta\|_\infty}{\|\mathbf{R}_{\backslash j,\backslash j}\beta\|_\infty}$$

$$
\begin{aligned}
&\leq && \min_{\boldsymbol{\beta}\neq\mathbf{0},\boldsymbol{\beta}\in\mathbb{R}^d}\frac{\|\boldsymbol{\beta}\|_\infty}{\|\mathbf{R}\boldsymbol{\beta}\|_\infty} \\
&\leq && \|\mathbf{R}^{-1}\|_\infty \\
&\leq && \max_{1\leq j\leq d}\boldsymbol{\Sigma}_{jj}\|\boldsymbol{\Theta}\|_1 \\
&\leq && \Lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{\Theta}\|_1.
\end{aligned}
\tag{C.10}
$$

From Lemma C.3, we have

$$
\Big\|\frac{\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}}{n}\mathbf{Z}_{*\backslash j}^T\mathbf{Z}_{*\backslash j}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\Big\|_\infty \leq C\sigma_j\sqrt{\frac{\log d}{n}}.
$$

Therefore, by piecing all these terms together:

$$
\begin{aligned}
&\big\|\mathbf{R}_{\backslash j,\backslash j}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\big\|_\infty \\
&\leq \hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\cdot 18\sqrt{\frac{\log d}{n}}\cdot 5\sqrt{2}C(1+\bar{c}).\xi_{\max}k\sqrt{\frac{\log d}{n}}+C\sigma_j\sqrt{\frac{\log d}{n}} \\
&\leq Ck\frac{\log d}{n}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}+C\sigma_j\sqrt{\frac{\log d}{n}}.
\end{aligned}
\tag{C.11}
$$

Again, on the event $\mathcal{E}$, we have

$$
\hat{\boldsymbol{\Gamma}}_{jj}\leq\frac{3}{2}\Lambda_{\max}(\boldsymbol{\Sigma}),
$$

$$
\big\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}\big\|_\infty\leq\frac{1}{\min_{1\leq j\leq d}(\hat{\boldsymbol{\Sigma}}_{jj})^{1/2}}\leq\sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}},
\tag{C.12}
$$

$$
\hat{\boldsymbol{\Theta}}_{jj}\leq\left(1+C\sqrt{\frac{\log d}{n}}\right)\boldsymbol{\Theta}_{jj}\leq 2\|\boldsymbol{\Theta}\|_2=\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})},
\tag{C.13}
$$

$$
\sigma_j=\frac{1}{\sqrt{\boldsymbol{\Theta}_{jj}}}\leq\frac{1}{\sqrt{\Lambda_{\min}(\boldsymbol{\Theta})}}=\sqrt{\Lambda_{\max}(\boldsymbol{\Sigma})}.
$$

Putting all the terms in (C.10), (C.11), (C.12), (C.13) together, we get, for large enough n:

$$
\begin{aligned}
&|\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}\hat{\boldsymbol{\Theta}}_{jj}|\big\|\hat{\boldsymbol{\Gamma}}_{\backslash j,\backslash j}^{-1/2}(\hat{\boldsymbol{\beta}}_j-\boldsymbol{\beta}_j)\big\|_\infty \\
&\leq \frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}\sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}}\cdot\Lambda_{\max}(\boldsymbol{\Sigma})\|\boldsymbol{\Theta}\|_1\cdot\left(Ck\frac{\log d}{n}\hat{\boldsymbol{\Gamma}}_{jj}^{1/2}+C\sigma_j\sqrt{\frac{\log d}{n}}\right) \\
&\leq 2\xi_{\max}\sqrt{\frac{2}{\Lambda_{\min}(\boldsymbol{\Sigma})}}\|\boldsymbol{\Theta}\|_1\cdot\left(Ck\frac{\log d}{n}\cdot\sqrt{\frac{3}{2}\Lambda_{\max}(\boldsymbol{\Sigma})}+C\sqrt{\Lambda_{\max}(\boldsymbol{\Sigma})}\sqrt{\frac{\log d}{n}}\right) \\
&\leq 2\sqrt{3}\xi_{\max}^{3/2}\cdot C\cdot\|\boldsymbol{\Theta}\|_1\cdot k\frac{\log d}{n}+2\sqrt{2}\xi_{\max}^{3/2}\cdot C\cdot\|\boldsymbol{\Theta}\|_1\cdot\sqrt{\frac{\log d}{n}}
\end{aligned}
$$

$$\leq \quad 2\big(\sqrt{3} + \sqrt{2}\big)\xi_{\max}^{3/2} \cdot C \cdot \big\|\mathbf{\Theta}\big\|_1 \cdot \sqrt{\frac{\log d}{n}},$$

where the last inequality follows from the fact that $k\sqrt{\log d} \leq \sqrt{n}$ for large enough n. Therefore, the lemma is proved. $\quad\square$

Appendix D: Proof of the main theorems

In this section, we prove the main results based on the previous technical lemmas.

D.1. Proof of Theorem 4.1

Proof. By piecing together the results of Lemma C.1 and Lemma C.2, we have

$$
\begin{aligned}
\big\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\big\|_1 &= \max_{1 \leq j \leq d} \|\hat{\mathbf{\Theta}}_{*j} - \mathbf{\Theta}_{*j}\|_1 \\
&\leq \max_{1 \leq j \leq d}\big|\hat{\mathbf{\Theta}}_{jj} - \mathbf{\Theta}_{jj}\big| + \max_{1 \leq j \leq d}\big\|\hat{\mathbf{\Theta}}_{\backslash j,j} - \mathbf{\Theta}_{\backslash j,j}\big\|_1 \\
&\leq C \cdot \|\mathbf{\Theta}\|_2\sqrt{\frac{\log d}{n}} + C\big(\|\mathbf{\Theta}\|_2 k + \|\mathbf{\Theta}\|_1\big)\sqrt{\frac{\log d}{n}} \\
&\leq C\big(\|\mathbf{\Theta}\|_2 k + \|\mathbf{\Theta}\|_1\big)\sqrt{\frac{\log d}{n}} \\
&\leq C\Big(k\|\mathbf{\Theta}\|_2\sqrt{\frac{\log d}{n}}\Big).
\end{aligned}
$$

The last inequality follows from the fact that

$$\big\|\mathbf{\Theta}\big\|_1 \leq k\big\|\mathbf{\Theta}\big\|_{\max} \leq k\big\|\mathbf{\Theta}\big\|_2.$$

We complete the proof of this theorem. $\quad\square$

D.2. Proof of Theorem 4.2

Proof. By piecing together the results of Lemma C.1 and Lemma C.4, we have

$$\big\|\hat{\mathbf{\Theta}} - \mathbf{\Theta}\big\|_{\max} = \max_{1 \leq j \leq d}\big|\hat{\mathbf{\Theta}}_{jj} - \mathbf{\Theta}_{jj}\big| + \max_{1 \leq j \leq d}\big\|\hat{\mathbf{\Theta}}_{\backslash j,j} - \mathbf{\Theta}_{\backslash j,j}\big\|_\infty \leq C\big\|\mathbf{\Theta}\big\|_1\sqrt{\frac{\log d}{n}}.$$

We complete the proof of this theorem. $\quad\square$

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* 267–281. MR0483125

Bach, F. R. (2008). Bolasso: model consistent lasso estimation through the bootstrap. In *In Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*.

Banerjee, O., Ghaoui, L. E. and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research* **9** 485–516. MR2417243

Belloni, A., Chernozhukov, V. and Wang, L. (2012). Square-root lasso: Pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791–806. MR2860324

Ben-david, S., Luxburg, U. V. and Pal, D. (2006). A sober look at clustering stability. In *Proceedings of the Annual Conference of Learning Theory*. Springer.

Bhadra, A. and Mallick, B. K. (2013). Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. *Biometrics* **69** 447–457. MR3071063

Bishop, C., Spiegelhalter, D. and Winn, J. (2003). VIBES: A variational inference engine for Bayesian networks. In *Advances in Neural Information Processing Systems 15* (S. Becker, S. Thrun and K. Obermayer, eds.). MIT Press, Cambridge, MA, 777–784.

Boos, D. D., Stefanski, L. A. and Wu, Y. (2009). Fast fsr variable selection with applications to clinical trials. *Biometrics* **65** 692–700. MR2649841

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

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

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95** 759–771.

Chen, J. and Chen, Z. (2012). Extended bic for small-n-large-p sparse glm. *Statistica Sinica* **22** 555–574.

Chen, L., Tong, T. and Zhao, H. (2008). Considering dependence among genes and markers for false discovery control in eqtl mapping. *Bioinformatics* **24** 2015–2022.

Chen, S., Donoho, D. and Saunders, M. (1998). Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing* **20** 33–61. MR1639094

Dempster, A. (1972). Covariance selection. *Biometrics* **28** 157–175.

Drton, M. and Perlman, M. (2007). Multiple testing and error control in Gaussian graphical model selection. *Statistical Science* **22** 430–449.

Drton, M. and Perlman, M. (2008). A SINful approach to Gaussian graphical model selection. *Journal of Statistical Planning and Inference* **138** 1179–1200.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans.* SIAM [Society for Industrial and Applied Mathematics].

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association* **99** 619–632.

Foster, D. P. and George, E. I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics* **22** 1947–1975.

Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in Neural Information Processing Systems 23.*

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

Friedman, J., T. Hastie, H. H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics* **1** 302–332. MR2415737

Jalali, A., Johnson, C. and Ravikumar, P. (2012). High-dimensional sparse inverse covariance estimation using greedy methods. *International Conference on Artificial Intelligence and Statistics.*

Johnstone., I. (2000). *Chi-square oracle inequalities.* Lecture Notes-Monograph Series.

Lafferty, J., Liu, H. and Wasserman, L. (2012). Sparse Nonparametric Graphical Models. *Statistical Science* **27** 519–537.

Lam, C. and Fan, J. (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of Statistics* **37** 42–54.

Lange, T., Roth, V., Braun, M. L. and Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation* **16** 1299–1323.

Laurent, B. and Massart, P. (1998). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* **28** 1303–1338.

Li, X., Haupt, J., Arora, R., Liu, H., Hong, M. and Zhao, T. (2016). A first order free lunch for sqrt-lasso. *Technical Report.*

Liu, H., Chen, X., Lafferty, J. and Wasserman, L. (2010a). Graph-valued regression. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS).*

Liu, H., Han, F., Yuan, M., Lafferty, J. and Wasserman, L. (2012). High dimensional semiparametric gaussian copula graphical models. *The Annals of Statistics* **40** 2293–2326.

Liu, H., Roeder, K. and Wasserman, L. (2010b). Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS).*

LIU, H., XU, M., GU, H., GUPTA, A., LAFFERTY, J. and WASSERMAN, L. (2011). Forest density estimation. *Journal of Machine Learning Research* **12** 907–951. A short version has appeared in the 23rd Annual Conference on Learning Theory (COLT).

LIU, W. and LUO, X. (2015). High-dimensional sparse precision matrix estimation via sparse column inverse operator. *Journal of Multivariate Analysis* **135** 153–162.

LYSEN, S. (2009). Permuted inclusion criterion: A variable selection technique. *Publicly accessible Penn Dissertations. Paper 28.*

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MEINSHAUSEN, N. and BÜHLMANN, P. (2006). High dimensional graphs and variable selection with the lasso. *Annals of Statistics* **34** 1436–1462.

MEINSHAUSEN, N. and BÜHLMANN, P. (2010). Stability selection. *Journal of the Royal Statistical Society, Series B, Methodological* **72** 417–473.

MOHAMMADI, A. and WIT, E. C. (2015). Bayesian Structure Learning in Sparse Gaussian Graphical Models. *Bayesian Analysis* **10** 109–138.

RASKUTTI, G., WAINWRIGHT, M. J. and YU, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research* **99** 2241–2259.

RAVIKUMAR, P., WAINWRIGHT, M., 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.

ROTHMAN, A., BICKEL, P., LEVINA, E. and ZHU, J. (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics* **2** 494–515.

SCHWARZ, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.

SHEN, X., PAN, W. and ZHU, Y. (2012).). likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107** 223–232. MR2949354

STRANGER, B., NICA, A., FORREST, M., DIMAS, A., BEAZLEY, C. B. C., INGLE, C., DUNNING, M., FLICEK, P., KOLLER, D., MONTGOMERY, S., TAVARÉ, S., DELOUKAS, P. and DERMITZAKIS, E. (2007). Population genomics of human gene expression. *Nature Genetics* **39** 1217–24.

SUN, T. and ZHANG, C.-H. (2013). Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research* **14** 3358–3418.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* **58** 267–288.

WILLE, A., ZIMMERMANN, P., VRANOVA, E., FÜRHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIC, A., VON ROHR, P., THIELE, L., ZITZLER, E., GRUISSEM, W. and BÜHLMANN, P. (2004). Sparse graphical gaussian modeling of the isoprenoid gene network in arabidopsis thaliana. *Genome Biology* **5** R92.

WU, Y., BOOS, D. D. and STEFANSKI, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102** 235–243.

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

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

Zhao, T., Liu, H., Roeder, K., Lafferty, J. and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research* **13** 1059–1062. MR2930633