

# Asymptotically Normal and Efficient Estimation of Covariate-Adjusted Gaussian Graphical Model

Mengjie Chen,

Program of Computational Biology and Bioinformatics,

Zhao Ren,

Department of Statistics,

Hongyu Zhao,

Department of Biostatistics, School of Public Health,

Harrison Zhou

Department of Statistics,

Yale University, New Haven, CT 06520, USA.

Email: [harrison.zhou@yale.edu](mailto:harrison.zhou@yale.edu).

September 23, 2013

## **Abstract**

A tuning-free procedure is proposed to estimate the covariate-adjusted Gaussian graphical model. For each finite subgraph, this estimator is asymptotically normal and efficient. As a consequence, a confidence interval can be obtained for each edge. The procedure enjoys easy implementation and efficient computation through parallel estimation on subgraphs or edges. We further apply the asymptotic normality result to perform support recovery through edge-wise adaptive thresholding. This support recovery procedure is called ANTAC, standing for Asymptotically Normal estimation with Thresholding after Adjusting Covariates. ANTAC outperforms other methodologies in the literature in a range of simulation studies. We apply ANTAC to identify gene-gene interactions using an eQTL dataset. Our result achieves better interpretability and accuracy in comparison with CAMPE.

**KEYWORDS:** Sparsity, Precision matrix estimation, Support recovery, High-dimensional statistics, Gene regulatory network, eQTL

## 1. INTRODUCTION

Graphical models have been successfully applied to a broad range of studies that investigate the relationships among variables in a complex system. With the advancement of high-throughput technologies, an unprecedented amount of features can be collected for a given system. Therefore, the inference with graphical models has become more challenging. To better understand the complex system, novel methods under high dimensional setting are extremely needed. Among graphical models, Gaussian graphical models have recently received considerable attention for their applications in the analysis of gene expression data. It provides an approach to discover and analyze gene relationships, which offers insights into gene regulatory mechanism. However gene expression data alone are not enough to fully capture the complexity of gene regulation. Genome-wide expression quantitative trait loci (eQTL) studies, which simultaneously measure genetic variation and gene expression levels, reveal that genetic variants account for a large proportion of the variability of gene expression across different individuals (Rockman & Kruglyak 2006). Some genetic variants may confound the genetic network analysis, thus ignoring the influence of them may lead to false discoveries. Adjusting the effect of genetic variants is of importance for the accurate inference of genetic network at the expression level. A few papers in the literature have considered to accommodate covariates in graphical models. See, for example, Li, Chun & Zhao (2012), Yin & Li (2013) and Cai, Li, Liu & Xie (2013) introduced Gaussian graphical model with adjusted covariates, and Cheng, Levina, Wang & Zhu (2012) introduced additional covariates to Ising models.

This problem has been naturally formulated as joint estimation of the multiple regression coefficients and the precision matrix in Gaussian settings. Since it is widely believed that genes operate in biological pathways, the graph for gene expression data is expected to be sparse. Many regularization-based approaches have been proposed in the literature. Some use a joint regularization penalty for both the multiple regression coefficients and the precision matrix and solve iteratively (Obozinski, Wainwright & Jordan 2011; Yin & Li 2011; Peng, Zhu, Bergamaschi, Han, Noh, Pollack & Wang 2010). Others apply a two-stage strategy: estimating the regression coef-

ficients in the first stage and then estimating the precision matrix based on the residuals from the first stage. For all these methods, the thresholding level for support recovery depends on the unknown matrix  $l_1$  norm of the precision matrix or an irrepresentable condition on the Hessian tensor operator, thus those theoretically justified procedures can not be implemented practically. In practice, the thresholding level is often selected through cross-validation. When the dimension  $p$  of the precision matrix is relatively high, cross-validation is computationally intensive, with a jeopardy that the selected thresholding level is very different from the optimal one. As we show in the simulation studies presented in Section 5, the thresholding levels selected by the cross-validation tend to be too small, leading to an undesired denser graph estimation in practice. In addition, for current methods in the literature, the thresholding level for support recovery is set to be the same for all entries of the precision matrix, which makes the procedure non-adaptive.

In this paper, we propose a tuning free methodology for the joint estimation of the regression coefficients and the precision matrix. The estimator for each entry of the precision matrix or each partial correlation is asymptotically normal and efficient. Thus a P-value can be obtained for each edge to reflect the statistical significance of each entry. In the gene expression analysis, the P-value can be interpreted as the significance of the regulatory relationships among genes. This method is easy to implement and is attractive in two aspects. First, it has the scalability to handle large datasets. Estimation on each entry is independent and thus can be parallelly computed. As long as the capacity of instrumentation is adequate, those steps can be distributed to accommodate the analysis of high dimensional data. Second, it has the modulability to estimate any subgraph with special interests. For example, biologists may be interested in the interaction of genes play essential roles in certain biological processes. This method allows them to specifically target the estimation on those genes. An R package implementing our method has been developed and is available on the CRAN website.

We apply the asymptotic normality and efficiency result to do support recovery by edge-wise adaptive thresholding. This rate-optimal support recovery procedure is called ANTAC, standing for Asymptotically Normal estimation with Thresholding after Adjusting Covariates. This work is closely connected to a growing literature on optimal estimation of large covariance and precision

matrices. Many regularization methods have been proposed and studied. For example, Bickel and Levina (Bickel & Levina 2008a; Bickel & Levina 2008b) proposed banding and thresholding estimators for estimating bandable and sparse covariance matrices respectively and obtained rate of convergence for the two estimators. See also El Karoui (2008) and Lam & Fan (2009). Cai, Zhang & Zhou (2010) established the optimal rates of convergence for estimating bandable covariance matrices. Cai & Zhou (2012) and Cai, Liu & Zhou (2012) obtained the minimax rate of convergence for estimating sparse covariance and precision matrices under a range of losses including the spectral norm loss. Most closely related to this paper is the work in Ren, Sun, Zhang & Zhou (2013) where fundamental limits were given for asymptotically normal and efficient estimation of sparse precision matrices. Due to the complication of the covariates, the analysis in this paper is more involved.

We organize the rest of the paper as follows. Section 2 describes the covariate-adjusted Gaussian graphical model and introduces our novel two-step procedure. Corresponding theoretical studies on asymptotic normal distribution and adaptive support recovery are presented in Sections 3-4. Simulation studies are carried out in Section 5. Section 6 presents the analysis of eQTL data. Proofs for theoretical results are collected in Section 7. We collect a key lemma and auxiliary results for proving the main results in Section 8 and Appendix 9.

## 2. COVARIATE-ADJUSTED GAUSSIAN GRAPHICAL MODEL AND METHODOLOGY

In this section we first formally introduce the covariate-adjusted Gaussian graphical model, and then propose a two-step procedure for estimation of the model.

### 2.1 Covariate-adjusted Gaussian Graphical Model

Let  $(X^{(i)}, Y^{(i)})$ ,  $i = 1, \dots, n$ , be i.i.d. with

$$Y^{(i)} = \Gamma_{p \times q} X^{(i)} + Z^{(i)}, \quad (1)$$

where  $\Gamma$  is a  $p \times q$  unknown coefficient matrix, and  $Z^{(i)}$  is a  $p$ -dimensional random vector following a multivariate Gaussian distribution  $N(0, \Omega^{-1})$  and is independent of  $X^{(i)}$ . For the genome-wide expression quantitative trait studies,  $Y^{(i)}$  is the observed expression levels for  $p$  genes of the

$i$ -th subject and  $X^{(i)}$  is the corresponding values of  $q$  genetic markers. We will assume that  $\Omega$  and  $\Gamma_{p \times q}$  are sparse. The precision matrix  $\Omega$  is assumed to be sparse partly due to the belief that genes operate in biological pathways, and the sparseness structure of  $\Gamma$  reflects the sensitivity of confounding of genetic variants in the genetic network analysis.

We are particularly interested in the graph structure of random vector  $Z^{(i)}$ , which represents the genetic networks after removing the effect of genetic markers. Let  $G = (V, E)$  be an undirected graph representing the conditional independence relations between the components of a random vector  $Z^{(1)} = (Z_{11}, \dots, Z_{1p})^T$ . The vertex set  $V = \{V_1, \dots, V_p\}$  represents the components of  $Z$ . The edge set  $E$  consists of pairs  $(i, j)$  indicating the conditional dependence between  $Z_{1i}$  and  $Z_{1j}$  given all other components. In the genetic network analysis, the following question is fundamental: Is there an edge between  $V_i$  and  $V_j$ ? It is well known that recovering the structure of an undirected Gaussian graph  $G = (V, E)$  is equivalent to recovering the support of the population precision matrix  $\Omega = (\omega_{ij})$  of the data in the Gaussian graphical model. There is an edge between  $V_i$  and  $V_j$ , i.e.,  $(i, j) \in E$ , if and only if  $\omega_{ij} \neq 0$ . See, for example, Lauritzen (1996). Consequently, the support recovery of the precision matrix  $\Omega$  yields the recovery of the structure of the graph  $G$ .

Motivated by biological applications, we consider the high-dimensional case in this paper, allowing the dimension to exceed or even be far greater than the sample size,  $\min\{p, q\} \geq n$ . The main goal of this work is not only to provide a fully data driven and easily implementable procedure to estimate the network for the covariate-adjusted Gaussian graphical model, but also to provide a confidence interval for estimation of each entry of the precision matrix  $\Omega$ .

## 2.2 A Two-step Procedure

In this section, we propose a two-step procedure to estimate  $\Omega$ . In the first step of the two-step procedure, we apply a scaled lasso method to obtain an estimator  $\hat{\Gamma} = (\hat{\gamma}_1, \dots, \hat{\gamma}_p)^T$  of  $\Gamma$ . This procedure is tuning free. This is different from other procedures in the literature for the sparse linear regression, such as standard lasso and Dantzig selector which select tuning parameters by cross-validation and can be computationally very intensive for high dimensional data. In the second step, we approximate each  $Z^{(i)}$  by  $\hat{Z}^{(i)} = Y^{(i)} - \hat{\Gamma}_{p \times q} X^{(i)}$ , then apply the tuning-free methodology

proposed in Ren et al. (2013) for the standard Gaussian graphical model to estimate each entry  $\omega_{ij}$  of  $\Omega$ , pretending that each  $\hat{Z}^{(i)}$  was  $Z^{(i)}$ . As a by-product, we have an estimator  $\hat{\Gamma}$  of  $\Gamma$ , which is shown to be rate optimal under different matrix norms, however our main goal is not to estimate  $\Gamma$ , but to make inference on  $\Omega$ .

**Step 1** Denote the  $n$  by  $q$  dimensional explanatory matrix by  $\mathbf{X} = (X^{(1)}, \dots, X^{(n)})^T$ , where the  $i$ th row of matrix is from the  $i$ -th sample  $X^{(i)}$ . Similarly denote the  $n$  by  $p$  dimensional response matrix by  $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(n)})^T$  and the noise matrix by  $\mathbf{Z} = (Z^{(1)}, \dots, Z^{(n)})^T$ . Let  $\mathbf{Y}_j$  and  $\mathbf{Z}_j$  be the  $j$ -th column of  $\mathbf{Y}$  and  $\mathbf{Z}$  respectively. For each  $j = 1, \dots, p$ , we apply a scaled lasso penalization to the univariate linear regression of  $\mathbf{Y}_j$  against  $\mathbf{X}$  as follows,

$$\textbf{Step 1} : \left\{ \hat{\gamma}_j, \hat{\sigma}_{jj}^{1/2} \right\} = \arg \min_{b \in \mathbb{R}^q, \theta \in \mathbb{R}^+} \left\{ \frac{\|\mathbf{Y}_j - \mathbf{X}b\|^2}{2n\theta} + \frac{\theta}{2} + \lambda_1 \sum_{k=1}^q \frac{\|\mathbf{X}_k\|}{\sqrt{n}} |b_k| \right\}, \quad (2)$$

where the weighted penalties are chosen to be adaptive to each variance  $\text{Var}(X_{1k})$  such that an explicit value can be given for the parameter  $\lambda_1$ , for example, one of the theoretically justified choices is  $\lambda_1 = \sqrt{2(1 + \frac{\log p}{\log q})/n}$ . The scaled lasso (2) is jointly convex in  $b$  and  $\theta$ . The global optimum can be obtained through alternatively updating between  $b$  and  $\theta$ . The computational cost is nearly the same as that of the standard lasso. For more details about its algorithm, please refer to Sun & Zhang (2012).

Define the estimate of “noise”  $\mathbf{Z}_j$  as the residue of the scaled lasso regression by

$$\hat{\mathbf{Z}}_j = \mathbf{Y}_j - \mathbf{X}\hat{\gamma}_j, \quad (3)$$

which will be used in the second step to make inference for  $\Omega$ .

**Step 2** In the second step, we propose a tuning-free regression approach to estimate  $\Omega$  based on  $\hat{\mathbf{Z}}$  defined in Equation (3), which is different from other methods proposed in the literature, including Cai et al. (2013) or Yin & Li (2013). An advantage of our approach is the ability to provide an asymptotically normal and efficient estimation of each entry of the precision matrix  $\Omega$ .

We first introduce some convenient notations for a subvector or a submatrix. For any index subset  $A$  of  $\{1, 2, \dots, p\}$  and a vector  $W$  of length  $p$ , we use  $W_A$  to denote a vector of length

$|A|$  with elements indexed by  $A$ . Similarly for a matrix  $U$  and two index subsets  $A$  and  $B$  of  $\{1, 2, \dots, p\}$ , we can define a submatrix  $U_{A,B}$  of size  $|A| \times |B|$  with rows and columns of  $U$  indexed by  $A$  and  $B$ , respectively. Let  $W = (W_1, \dots, W_p)^T$ , representing each  $Z^{(i)}$ , follow a Gaussian distribution  $N(0, \Omega^{-1})$ . It is well known that

$$W_A | W_{A^c} = N(-\Omega_{A,A}^{-1} \Omega_{A,A^c} W_{A^c}, \Omega_{A,A}^{-1}). \quad (4)$$

For  $A = \{i, j\}$ , equivalently we may write

$$(W_i, W_j) = W_{\{i,j\}^c}^T \beta + (\eta_i, \eta_j), \quad (5)$$

where the coefficients and error distributions are

$$\beta = -\Omega_{A^c,A} \Omega_{A,A}^{-1}, \quad (\eta_i, \eta_j)^T \sim N(0, \Omega_{A,A}^{-1}). \quad (6)$$

Based on the regression interpretation (5), we have the following data version of the multivariate regression model

$$\mathbf{Z}_A = \mathbf{Z}_{A^c} \beta + \epsilon_A, \quad (7)$$

where  $\beta$  is a  $(p-2)$  by 2 dimensional coefficient matrix. If we know  $\mathbf{Z}_A$  and  $\beta$ , an asymptotically normal and efficient estimator of  $\Omega_{A,A}$  is  $(\epsilon_A^T \epsilon_A / n)^{-1}$ .

But of course  $\beta$  is unknown and we only have access to the estimated observations  $\hat{\mathbf{Z}}$  from Equation (3). We replace  $\mathbf{Z}_A$  and  $\mathbf{Z}_{A^c}$  by  $\hat{\mathbf{Z}}_A$  and  $\hat{\mathbf{Z}}_{A^c}$  respectively in the regression (7) to estimate  $\beta$  as follows. For each  $m \in A = \{i, j\}$ , we apply a scaled lasso penalization to the univariate linear regression of  $\hat{\mathbf{Z}}_m$  against  $\hat{\mathbf{Z}}_{A^c}$ ,

$$\text{Step 2 : } \left\{ \hat{\beta}_m, \hat{\psi}_{mm}^{1/2} \right\} = \arg \min_{b \in \mathbb{R}^{p-2}, \sigma \in \mathbb{R}^+} \left\{ \frac{\|\hat{\mathbf{Z}}_m - \hat{\mathbf{Z}}_{A^c} b\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda_2 \sum_{k \in A^c} \frac{\|\hat{\mathbf{Z}}_k\|}{\sqrt{n}} |b_k| \right\}, \quad (8)$$

where the vector  $b$  is indexed by  $A^c$ , and one of the theoretically justified choices of  $\lambda_2$  is  $\lambda_2 = \sqrt{\frac{2 \log p}{n}}$ . Denote the residuals of the scaled lasso regression by

$$\hat{\epsilon}_A = \hat{\mathbf{Z}}_A - \hat{\mathbf{Z}}_{A^c} \hat{\beta}, \quad (9)$$

and then define

$$\hat{\Omega}_{A,A} = (\hat{\epsilon}_A^T \hat{\epsilon}_A / n)^{-1}. \quad (10)$$

This extends the methodology proposed in Ren et al. (2013) for Gaussian graphical model to corrupted observations. The approximation error  $\hat{\mathbf{Z}} - \mathbf{Z}$  affects inference for  $\Omega$ . Later we show if  $\Gamma$  is sufficient sparse,  $\Gamma_{p \times q} X$  can be well estimated so that the approximation error is negligible. When both  $\Omega$  and  $\Gamma$  are sufficiently sparse,  $\hat{\Omega}_{A,A}$  can be shown to be asymptotically normal and efficient. An immediate application of the asymptotic normality result is to perform adaptive graphical model selection by explicit entry-wise thresholding, which yields a rate-optimal adaptive estimation of the precision matrix  $\Omega$  under various matrix  $l_w$  norms. See Theorems 2, 3 and 4 in Section 3 and 4 for more details.

### 3. ASYMPTOTIC NORMALITY DISTRIBUTION OF THE ESTIMATOR

In this section we first give theoretical properties of the estimator  $\hat{\mathbf{Z}}$  as well as  $\hat{\Gamma}$ , then present the asymptotic normality and efficiency result for estimation of  $\Omega$ .

We assume the coefficient matrix  $\Gamma$  is sparse, and entries of  $X$  with mean zero are bounded since the gene marker is usually bounded.

1. The coefficient matrix  $\Gamma$  satisfies the following sparsity condition,

$$\max_i \sum_{j \neq i} \min \left\{ 1, \frac{|\gamma_{ij}|}{\lambda_1} \right\} = s_1, \quad (11)$$

where in this paper  $\lambda_1$  is at an order of  $\sqrt{\frac{\log q}{n}}$ . Note that  $s_1 \leq \max_i \sum_{j \neq i} I \{ \gamma_{ij} \neq 0 \}$ , the maximum of the exact row sparseness among all rows of  $\Gamma$ .

2. There exist positive constants  $M_1$  and  $M_2$  such that  $1/M_1 \leq \lambda_{\min}(\text{Cov}(X^{(1)}))$  and  $1/M_2 \leq \lambda_{\min}(\Omega) \leq \lambda_{\max}(\Omega) \leq M_2$ .
3. There is a constant  $B > 0$  such that

$$|X_{ij}| \leq B \text{ for all } i \text{ and } j. \quad (12)$$

It is worthwhile to note that the boundedness assumption (12) does not imply the  $X^{(1)}$  is jointly sub-gaussian, i.e.,  $X^{(1)}$  is allowed to be not jointly sub-gaussian as long as above conditions are satisfied. In the high dimensional regression literature, it is common to assume the joint sub-gaussian condition on the design matrix as follows,

3'. We shall assume that the distribution of  $X^{(1)}$  is jointly sub-gaussian with parameter  $(M_1)^{1/2} > 0$  in the sense that

$$\mathbb{P}\{|v^T X^{(1)}| > t\} \leq e^{-t^2/2M_1} \text{ for all } t > 0 \text{ and } \|v\|_2 = 1. \quad (13)$$

We analyze the Step 1 of the procedure in Equation (2) under Conditions 1-3 as well as Conditions 1-2 and 3'. The optimal rates of convergence are obtained under the matrix  $l_\infty$  norm and Frobenius norm for estimation of  $\Gamma$ , which yield a rate of convergence for estimation of each  $\mathbf{Z}_j$  under the  $l_2$  norm.

**Theorem 1** *Let  $\lambda_1 = (1 + \varepsilon_1) \sqrt{\frac{2\delta_1 \log q}{n}}$  for any  $\delta_1 \geq 1$  and  $\varepsilon_1 > 0$  in Equation (2). Assume that*

$$s_1 = o\left(\min\left\{\frac{n}{\log^3 n \log q}, \sqrt{\frac{n}{\log q}}\right\}\right).$$

*Under Conditions 1-3 we have*

$$\mathbb{P}\left\{\frac{1}{n} \|\hat{\mathbf{Z}}_j - \mathbf{Z}_j\|^2 > C_1 s_1 \frac{\log q}{n}\right\} \leq o(q^{-\delta_1+1}) \text{ for each } j, \quad (14)$$

$$\mathbb{P}\left\{\|\hat{\Gamma} - \Gamma\|_{l_\infty} > C_2 s_1 \sqrt{\frac{\log q}{n}}\right\} \leq o(p \cdot q^{-\delta_1+1}), \quad (15)$$

$$\mathbb{P}\left\{\frac{1}{p} \|\hat{\Gamma} - \Gamma\|_F^2 > C_3 s_1 \frac{\log q}{n}\right\} \leq o(p \cdot q^{-\delta_1+1}). \quad (16)$$

*Moreover, if we replace Condition 3 by the weaker version Condition 3', all results above still hold under a weaker assumption on  $s_1$ ,*

$$s_1 = o\left(\frac{n}{\log q}\right). \quad (17)$$

The proof of Theorem 1 is provided in the Section 7.1.

**Remark 1** Under the assumption that the  $l_r$  norm of each row of  $\Gamma$  is bounded by  $k_{n,q}^{1/r}$ , an immediate application of Theorem 1 yields corresponding results for  $l_r$  ball sparseness. For example,

$$\mathbb{P} \left\{ \frac{1}{p} \left\| \hat{\Gamma} - \Gamma \right\|_F^2 > C_4 k_{n,q} \left( \frac{\log q}{n} \right)^{1-r/2} \right\} \leq o(p \cdot q^{-\delta_1+1}),$$

provided that  $k_{n,q} = o\left(\frac{n}{\log q}\right)^{1-r/2}$  and Conditions 1-2, 3' hold.

**Remark 2** Cai et al. (2013) assumes that the matrix  $l_1$  norm of  $(\text{Cov}(X^{(1)}))^{-1}$ , the inverse of the covariance matrix of  $X^{(1)}$ , is bounded, and their tuning parameter depends on the unknown  $l_1$  norm. In Theorem 1 we don't need the assumption on the  $l_1$  norm of  $(\text{Cov}(X^{(1)}))^{-1}$  and the tuning parameter  $\lambda_1$  is given explicitly.

To analyze the Step 2 of the procedure in Equation (8), we need the following assumptions for  $\Omega$ .

4. The precision matrix  $\Omega = (\omega_{ij})_{p \times p}$  has the following sparsity condition

$$\max_i \sum_{j \neq i} \min \left\{ 1, \frac{|\omega_{ij}|}{\lambda_2} \right\} = s_2, \quad (18)$$

where  $\lambda_2$  is at an order of  $\sqrt{\frac{\log p}{n}}$ .

5. There exists a positive constant  $M_2$  such that  $\|\Omega\|_{l_\infty} \leq M_2$ .

It is convenient to introduce a notation for the covariance matrix of  $(\eta_i, \eta_j)^T$  in Equation (5).

Let

$$\Psi_{A,A} = \Omega_{A,A}^{-1} = \begin{pmatrix} \psi_{ii} & \psi_{ij} \\ \psi_{ji} & \psi_{jj} \end{pmatrix}.$$

We will estimate  $\Psi_{A,A}$  first and show that an efficient estimator of  $\Psi_{A,A}$  yields an efficient estimation of entries of  $\Omega_{A,A}$  by inverting the estimator of  $\Psi_{A,A}$ . Denote a sample version of  $\Psi_{A,A}$  by

$$\Psi_{A,A}^{ora} = (\psi_{kl}^{ora})_{k \in A, l \in A} = \epsilon_A^T \epsilon_A / n, \quad (19)$$

which is an oracle MLE of  $\Psi_{A,A}$ , assuming that we know  $\beta$ , and

$$\Omega_{A,A}^{ora} = (\omega_{kl}^{ora})_{k \in A, l \in A} = (\Psi_{A,A}^{ora})^{-1}. \quad (20)$$

Let

$$\hat{\Psi}_{A,A} = \hat{\epsilon}_A^T \hat{\epsilon}_A / n, \quad (21)$$

where  $\hat{\epsilon}_A$  is defined in Equation (9). Note that  $\hat{\Omega}_{A,A}$  defined in Equation (10) is simply the inverse of the estimator  $\hat{\Psi}_{A,A}$ . The following result shows that  $\hat{\Omega}_{A,A}$  is asymptotically normal and efficient when both  $\Gamma$  and  $\Omega$  are sufficient sparse.

**Theorem 2** Let  $\lambda_1$  be defined as in Theorem 1 with  $\delta_1 \geq 1 + \frac{\log p}{\log q}$  and  $\lambda_2 = (1 + \varepsilon_2) \sqrt{\frac{2\delta_2 \log p}{n}}$  for any  $\delta_2 \geq 1$  and  $\varepsilon_2 > 0$  in Equation (8). Assume that

$$s_1 = o\left(\sqrt{\frac{n}{\log q}}\right) \text{ and } s_2 = o\left(\sqrt{\frac{n}{\log p}}\right). \quad (22)$$

Under Conditions 1-2 and 4-5, and Condition 3 or 3', we have

$$\mathbb{P} \left\{ \left\| \hat{\Psi}_{A,A} - \hat{\Psi}_{A,A}^{ora} \right\|_{\infty} > C_5 \left( s_2 \frac{\log p}{n} + s_1 \frac{\log q}{n} \right) \right\} \leq o(p^{-\delta_2+1} + pq^{-\delta_1+1}), \quad (23)$$

$$\mathbb{P} \left\{ \left\| \hat{\Omega}_{A,A} - \Omega_{A,A}^{ora} \right\|_{\infty} > C_6 \left( s_2 \frac{\log p}{n} + s_1 \frac{\log q}{n} \right) \right\} \leq o(p^{-\delta_2+1} + pq^{-\delta_1+1}), \quad (24)$$

for some positive constants  $C_5$  and  $C_6$ . Furthermore,  $\hat{\omega}_{ij}$  is asymptotically efficient

$$\sqrt{n F_{ij}} (\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} N(0, 1), \quad (25)$$

when  $s_2 = o\left(\frac{\sqrt{n}}{\log p}\right)$  and  $s_1 = o\left(\frac{\sqrt{n}}{\log q}\right)$ , where

$$F_{ij}^{-1} = \omega_{ii}\omega_{jj} + \omega_{ij}^2.$$

**Remark 3** The asymptotic normality result can be obtained for estimation of the partial correlation. Let  $r_{ij} = -\omega_{ij}/(\omega_{ii}\omega_{jj})^{1/2}$  be the partial correlation between  $Z_i$  and  $Z_j$ . Define  $\hat{r}_{ij} = -\hat{\omega}_{ij}/(\hat{\omega}_{ii}\hat{\omega}_{jj})^{1/2}$ . Under the same assumptions in Theorem 2, the estimator  $\hat{r}_{ij}$  is asymptotically efficient, i.e.,  $\sqrt{n(1-r_{ij}^2)^{-2}}(\hat{r}_{ij} - r_{ij}) \xrightarrow{D} N(0, 1)$ , when  $s_2 = o(\sqrt{n}/\log p)$  and  $s_1 = o(\sqrt{n}/\log q)$ .

**Remark 4** In Equations (4) and (7), we can replace  $A = \{i, j\}$  by a bounded size subset  $B \subset [1 : p]$  with cardinality more than 2. Similar to the analysis of Theorem 2, we can show the estimator for any smooth functional of  $\Omega_{B,B}^{-1}$  is asymptotic normality as shown in Ren et al. (2013) for Gaussian graphical model.

**Remark 5** A stronger result can be obtained for the choice of  $\lambda_1$  and  $\lambda_2$ . Theorems 1 and 2 still hold, when  $\lambda_1 = (1 + \varepsilon_1) \sqrt{\frac{2\delta_1 \log(q/s_{\max,1})}{n}}$  and  $\lambda_2 = (1 + \varepsilon_2) \sqrt{\frac{2\delta_2 \log(p/s_{\max,2})}{n}}$ , where  $s_{\max,1} = o\left(\frac{\sqrt{n}}{\log q}\right)$  and  $s_{\max,2} = o\left(\frac{\sqrt{n}}{\log p}\right)$ . Another alternative choice of  $\lambda_1$  and  $\lambda_2$  will be introduced in Section 5.

#### 4. ADAPTIVE SUPPORT RECOVERY AND ESTIMATION OF $\Omega$ UNDER MATRIX NORMS

In this section, the asymptotic normality result obtained in Theorem 2 is applied to perform adaptive support recovery and to obtain rate-optimal estimation of the precision matrix under various matrix  $l_w$  norms. The two-step procedure for support recovery is first removing the effect of the covariate  $X$ , then applying ANT (Asymptotically Normal estimation with Thresholding) procedure. We thus call it ANTAC, which stands for ANT after Adjusting Covariates.

##### 4.1 ANTAC for Support Recovery of $\Omega$

The support recovery on covariate-adjusted Gaussian graphical model has been studied by several papers, for example, Yin & Li (2013) and Cai et al. (2013). Denote the support of  $\Omega$  by  $Supp(\Omega)$ . In these literature, the theoretical properties on the support recovery were obtained but they all assumed that  $\min_{(i,j) \in Supp(\Omega)} |\omega_{ij}| \geq CM_{n,p}^2 \sqrt{\frac{\log p}{n}}$ , where  $M_{n,p}$  is either the matrix  $l_\infty$  norm or related to the irrerepresentable condition on  $\Omega$ , which is unknown. The ANTAC procedure, based on the asymptotic normality estimation in Equation (25), performs entry-wise thresholding adaptively to recover the graph with explicit thresholding levels.

Recall that in Theorem 2 we have

$$\sqrt{nF_{ij}} (\hat{\omega}_{ij} - \omega_{ij}) \xrightarrow{D} N(0, 1),$$

where  $F_{ij} = (\omega_{ii}\omega_{jj} + \omega_{ij}^2)^{-1}$  is the Fisher information of estimating  $\omega_{ij}$ . Suppose we know this Fisher information, we can apply a thresholding level  $\sqrt{\frac{2\xi(\omega_{ii}\omega_{jj} + \omega_{ij}^2) \log p}{n}}$  with any  $\xi \geq 2$  for  $\hat{\omega}_{ij}$  to correctly distinguish zero and nonzero entries, noting the total number of edges is  $p(p-1)/2$ . However, when the variance  $\omega_{ii}\omega_{jj} + \omega_{ij}^2$  is unknown, all we need is to plug in a consistent estimator. The ANTAC procedure is defined as follows

$$\hat{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr})_{p \times p}, \quad \text{where } \hat{\omega}_{ii}^{thr} = \hat{\omega}_{ii}, \text{ and } \hat{\omega}_{ij}^{thr} = \hat{\omega}_{ij} 1\{|\hat{\omega}_{ij}| \geq \tau_{ij}\} \quad (26)$$

$$\text{with } \tau_{ij} = \sqrt{\frac{2\xi_0 (\hat{\omega}_{ii}\hat{\omega}_{jj} + \hat{\omega}_{ij}^2) \log p}{n}} \text{ for } i \neq j, \quad (27)$$

where  $\hat{\omega}_{kl}$  is the consistent estimator of  $\omega_{kl}$  defined in (10) and  $\xi_0$  is a tuning parameter which can be taken as fixed at any  $\xi_0 > 2$ .

The following sufficient condition for support recovery is assumed in Theorem 3 below. Define the sign of  $\Omega$  by  $\mathcal{S}(\Omega) = \{sgn(\omega_{ij}), 1 \leq i, j \leq p\}$ . Assume that

$$|\omega_{ij}| \geq 2\sqrt{\frac{2\xi_0 (\omega_{ii}\omega_{jj} + \omega_{ij}^2) \log p}{n}}, \quad \forall \omega_{ij} \in \text{Supp}(\Omega). \quad (28)$$

The following result shows that not only the support of  $\Omega$  but also the signs of the nonzero entries can be recovered exactly by  $\hat{\Omega}_{thr}$ .

**Theorem 3** *Assume that Conditions 1-2 and 4-5, and Condition 3 or 3' hold. Let  $\lambda_1$  be defined as in Theorem 1 with  $\delta_1 \geq 1 + \frac{\log p}{\log q}$  and  $\lambda_2 = (1 + \varepsilon_2) \sqrt{\frac{2\delta_2 \log p}{n}}$  with any  $\delta_2 \geq 3$  and  $\varepsilon_2 > 0$  in Equation (8). Also let  $\xi_0 > 2$  in the thresholding level (27). Under the assumptions (22) and (28), we have that the ANTAC defined in (26) recovers the support  $\mathcal{S}(\Omega)$  consistently, i.e.,*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \mathcal{S}(\hat{\Omega}_{thr}) = \mathcal{S}(\Omega) \right) = 1. \quad (29)$$

**Remark 6** *If the assumption (28) does not hold, the procedure recovers part of the true graph with high partial correlation.*

The proof of Theorem 3 depends on the oracle inequality (24) in Theorem 2, a moderate deviation result of the oracle  $\hat{\omega}_{ij}$  and a union bound. The detail of the proof is in spirit the same as that of Theorem 6 in Ren et al. (2013), and thus will be omitted due to the limit of space.

## 4.2 ANTAC for Estimation under the Matrix $l_w$ Norm

In this section, we consider the rate of convergence of a thresholding estimator of  $\Omega$  under the matrix  $l_w$  norm, including the spectral norm. The convergence under the spectral norm leads to the consistency of eigenvalues and eigenvectors estimation. Define  $\check{\Omega}_{thr}$ , a modification of  $\hat{\Omega}_{thr}$  defined in (26), as follows

$$\check{\Omega}_{thr} = (\hat{\omega}_{ij}^{thr} 1\{|\hat{\omega}_{ij}| \leq \log p\})_{p \times p}. \quad (30)$$

From the idea of the proof of Theorem 3 (see also the proof of Theorem 6 in Ren et al. (2013)), we see that with high probability  $\|\hat{\Omega} - \Omega\|_\infty$  is dominated by  $\|\Omega^{ora} - \Omega\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}}\right)$  under the sparsity assumptions (22). The key of the proof in Theorem 4 is to derive the upper bound under matrix  $l_1$  norm based on the entry-wise supnorm  $\|\check{\Omega}_{thr} - \Omega\|_\infty$ . Then the theorem follows immediately from the inequality  $\|M\|_{l_w} \leq \|M\|_{l_1}$  for any symmetric matrix  $M$  and  $1 \leq w \leq \infty$ , which can be proved by applying the Riesz-Thorin interpolation theorem. The proof follows similarly from that of Theorem 3 in Cai & Zhou (2012). We omit the proof due to the limit of space.

**Theorem 4** *Assume that Conditions 1-2 and 4-5, and Condition 3 or 3' hold. Under the assumptions (22) and  $n = \max\{O(p^{\xi_1}), O(q^{\xi_2})\}$  with some  $\xi_1, \xi_2 > 0$ , the  $\check{\Omega}_{thr}$  defined in (30) with sufficiently large  $\delta_1$  and  $\delta_2$  satisfies, for all  $1 \leq w \leq \infty$ ,*

$$\mathbb{E}\|\check{\Omega}_{thr} - \Omega\|_{l_w}^2 \leq C s_2^2 \frac{\log p}{n}. \quad (31)$$

**Remark 7** *The rate of convergence result in Theorem 4 also can be easily extended to the parameter space in which each row of  $\Omega$  is in a  $l_r$  ball of radius  $k_{n,q}^{1/r}$ . See, e.g., Theorem 3 in Cai & Zhou (2012). Under the same assumptions of Theorem 4 except replacing  $s_2 = o\left(\sqrt{n/\log p}\right)$  by  $k_{n,p}^2 = o\left((n/\log p)^{1-r}\right)$ , we have*

$$\mathbb{E}\|\check{\Omega}_{thr} - \Omega\|_{l_w}^2 \leq C k_{n,p}^2 \left(\frac{\log p}{n}\right)^{1-r}. \quad (32)$$

**Remark 8** *For the Gaussian graphical model without covariate variables, Cai et al. (2012) showed the rates obtained in Equations (31) and (32) are optimal when  $p \geq cn^\gamma$  for some  $\gamma > 1$  and*

$k_{n,p} = o\left(n^{1/2}(\log p)^{-3/2}\right)$  for the corresponding parameter spaces of  $\Omega$ . This implies that our estimator is rate optimal.

## 5. SIMULATION STUDIES

### 5.1 Asymptotic Normal Estimation

In this section, we compare the sample distribution of the proposed estimator for each edge  $\omega_{ij}$  with the normal distribution in Equation (25). Three models are considered with corresponding  $\{p, q, n\}$  listed in Table 1. Based on 200 replicates, the distributions of the estimators match the asymptotic distributions very well.

Three sparse models are generated in a similar way to those in Cai et al. (2013). The  $p \times q$  coefficient matrix  $\Gamma$  is generated as following for all three models,

$$\Gamma_{ij} \stackrel{i.i.d.}{\sim} N(0, 1) \cdot \text{Bernoulli}(0.025),$$

where the Bernoulli random variable is independent with the standard normal variable, taking one with probability 0.025 and zero otherwise. We then generate the  $p \times p$  precision matrix  $\Omega$  with identical diagonal entries  $\omega_{ii} = 4$  for the model of  $p = 200$  or 400 and  $\omega_{ii} = 5$  for the model of  $p = 1000$ , respectively. The off-diagonal entries of  $\Omega$  are generated i.i.d. as follows for each model,

$$\omega_{ij} = \begin{cases} 0.3 & \text{with probability } \frac{\pi}{3} \\ 0.6 & \text{with probability } \frac{\pi}{3} \\ 1 & \text{with probability } \frac{\pi}{3} \\ 0 & \text{otherwise} \end{cases}, \text{ for } i \neq j$$

where the probability of being nonzero  $\pi = P(\omega_{ij} \neq 0)$  for three models is shown in Table 1. Once both  $\Gamma$  and  $\Omega$  are chosen for each model, the  $n \times p$  outcome matrix  $Y$  is simulated from  $Y = X\Gamma^T + Z$  where rows of  $Z$  are i.i.d.  $N(0, \Omega^{-1})$  and rows of  $X$  are i.i.d.  $N(0, I_{q \times q})$ . We generate 200 replicates of  $X$  and  $Y$  for each model.

We randomly select four entries of  $\Omega$  with values  $v_\omega$  of 0, 0.3, 0.6 and 1 in each model and draw histograms of our estimators for those four entries based on the 200 replicates. The penalty

Table 1: Model parameters and simulation results: mean and standard deviation (in parentheses) of the proposed estimator for the randomly selected entry with value  $v_\omega$  based on 200 replicates.

$(p, q, n)$	$\pi = P(\omega_{ij} \neq 0)$	$v_\omega = 0$	$v_\omega = 0.3$	$v_\omega = 0.6$	$v_\omega = 1$
(200, 100, 400)	0.025	-0.015 (0.168)	0.289 (0.184)	0.574 (0.165)	0.986 (0.182)
(400, 100, 400)	0.010	-0.003 (0.24)	0.268 (0.23)	0.606 (0.23)	0.954 (0.244)
(1000, 100, 400)	0.005	0.011 (0.21)	0.292 (0.26)	0.507 (0.232)	0.862 (0.236)

parameter  $\lambda_1$ , which controls the weight of penalty in the regression of the first step (2), is set to be  $B_1/\sqrt{n-1+B_1^2}$ , where  $B_1 = q_t(1 - \frac{1}{2}(s_{\max,1}/q)^{1+\frac{\log p}{\log q}}, n-1)$ , and  $q_t(\cdot, n)$  is the quantile function of  $t$  distribution with degrees of freedom  $n$ . This parameter  $\lambda_1$  is a finite sample version of the asymptotic level  $\sqrt{2(1 + \frac{\log p}{\log q}) \log(q/s_{\max,1})/n}$  we proposed in Theorem 2 and Remark 5. Here we pick  $s_{\max,1} = \sqrt{n}/\log q$ . The penalty parameter  $\lambda_2$ , which controls the weight of penalty in the second step (8), is set to be  $B_2/\sqrt{n-1+B_2^2}$  where  $B_2 = q_t(1 - s_{\max,2}/(2p), n-1)$ , which is asymptotically equivalent to  $\sqrt{2 \log(p/s_{\max,2})/n}$ . The  $s_{\max,2}$  is set to be  $\sqrt{n}/\log p$ .

In Figure 1, we show the histograms of the estimators with the theoretical normal density super-imposed for those randomly selected four entries with values  $v_\omega$  of 0, 0.3, 0.6 and 1 in each of the three models. The distributions of our estimators match well with the theoretical normal distributions.

## 5.2 Support recovery

In this section, we evaluate the performance of the proposed ANTAC method and competing methods in support recovery with different simulation settings. ANTAC always performs among the best under all model settings. Under the Heterogeneous Model setting, the ANTAC achieves superior precision and recall rates and performs significantly better than others. Besides, ANTAC is computationally more efficient compared to a state-of-art method CAPME due to its tuning free property.

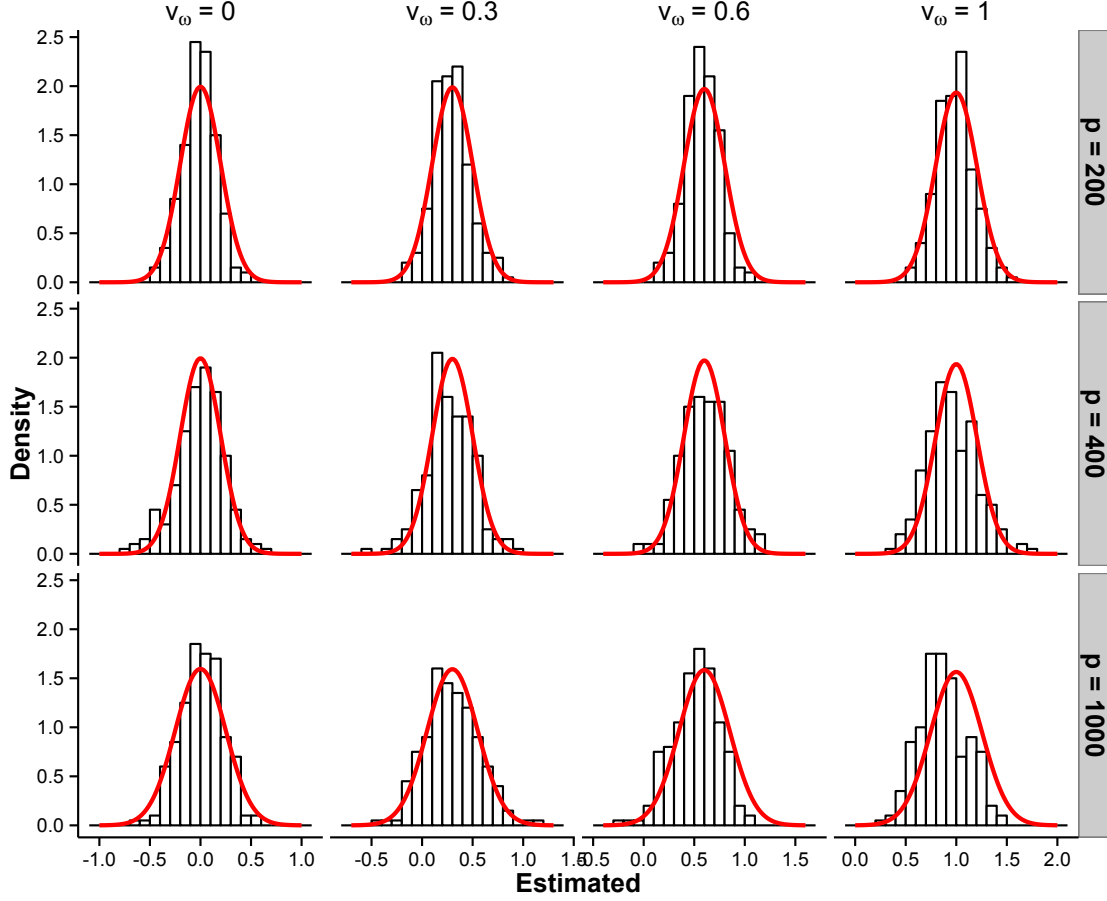


Figure 1: The histograms of the estimators for randomly selected entries with values  $v_\omega = 0, 0.3, 0.6$  and  $1$  in three models listed in Table 1. The theoretical normal density curves are shown as solid curves. The variance for each curve is  $(\omega_{ii}\omega_{jj} + \omega_{ij}^2)/n$ , the inverse of the Fisher information.

**Homogeneous Model** We consider three models with corresponding  $\{p, q, n\}$  listed in Table 2, which are similar to the models listed in Table 2 and used in (Cai et al. 2013). Since every model has identical values along the diagonal, we call them “Homogeneous Model”. In terms of the support recovery, ANTAC performs among the best in all three models, although the performance from all procedures is not satisfactory due to the intrinsic difficulty of support recovery problem for models considered.

We generate the  $p \times q$  coefficient matrix  $\Gamma$  in the same way as Section 5.1,

$$\Gamma_{ij} \stackrel{i.i.d.}{\sim} N(0, 1) \cdot \text{Bernoulli}(0.025).$$

The off-diagonal entries of the  $p \times p$  precision matrix  $\Omega$  are generated as follows,

$$\omega_{ij} \stackrel{i.i.d.}{\sim} N(0, 1) \cdot \text{Bernoulli}(\pi),$$

where the probability of being nonzero  $\pi = P(\omega_{ij} \neq 0)$  is shown in Table 2 for three models respectively. We generate 50 replicates of  $X$  and  $Y$  for each of the three models.

Table 2: Model parameters used in the simulation of support recovery.

	$(p, q, n)$	$P(\Gamma_{ij} \neq 0)$	$\pi = P(\omega_{ij} \neq 0), i \neq j$
Model 1	(200, 200, 200)	0.025	0.025
Model 2	(200, 100, 300)	0.025	0.025
Model 3	(800, 200, 200)	0.025	0.010

We compare our method with graphical Lasso (GLASSO) (Friedman, Hastie & Tibshirani 2008), a state-of-art method — CAPME (Cai et al. 2013) and a conditional GLASSO procedure (short as cGLASSO), where we apply the same scaled lasso procedure as the first stage of the proposed method and then estimate the precision matrix by GLASSO. This cGLASSO procedure is similar to that considered in Yin & Li (2013) except that in the first stage Yin & Li (2013) applies ordinary lasso rather than the scaled lasso, which requires another cross-validation for this step. For GLASSO, the precision matrix is estimated directly from the sample covariance matrix

without taking into account the effects from  $X$ . The tuning parameter for the  $l_1$  penalty is selected using five-fold cross validation by maximizing the log-likelihood function. For CAPME, the tuning parameters  $\lambda_1$  and  $\lambda_2$ , which control the penalty in the two stages of regression, are chosen using five-fold cross validation by maximizing the log-likelihood function. The optimum is achieved via a grid search on  $\{(\lambda_1, \lambda_2)\}$ . For Models 1 and 2,  $10 \times 10$  grid is used and for Model 3,  $5 \times 5$  grid is used because of the computational burden. Specifically, we use the CAPME package implemented by the authors of (Cai et al. 2013). For Model 3, each run with  $5 \times 5$  grid search and five-fold cross validation takes 160 CPU hours using one core from PowerEdge M600 nodes 2.33 GHz and 16–48 GB RAM, whereas ANTAC takes 46 CPU hours. For ANTAC, the parameter  $\lambda_1$  is set to be  $B_1/\sqrt{n-1+B_1^2}$ , where  $B_1 = q_t(1 - \frac{1}{2}(s_{\max,1}/q)^{1+\frac{\log p}{\log q}}, n-1)$ ,  $q_t(\cdot, n)$  is the quantile function of  $t$  distribution with degrees of freedom  $n$  and  $s_{\max,1} = \sqrt{n}/\log q$ . The parameter  $\lambda_2$ , is set to be  $B_2/\sqrt{n-1+B_2^2}$  where  $B_2 = q_t(1 - (s_{\max,2}/p)^3/2, n-1)$ . For cGLASSO, the first step is the same as ANTAC. In the second step, the precision matrix is estimated by applying GLASSO to the estimated  $Z$ , where the tuning parameter is selected using five-fold cross validation by maximizing the log-likelihood function.

We evaluate the performance of the estimators for support recovery problem in terms of the misspecification rate, specificity, sensitivity, precision and Matthews correlation coefficient, which are defined as,

$$\begin{aligned} \text{MISR}(\hat{\Omega}, \Omega) &= \frac{\text{FN} + \text{FP}}{p(p-1)}, \text{SPE} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \text{SEN} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{PRE} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{[(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})]^{1/2}}. \end{aligned}$$

Here, TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives respectively. True positives are defined as the correctly identified nonzero entries of the off-diagonal entries of  $\Omega$ . For GLASSO and CAPME, nonzero entries of  $\hat{\Omega}$  are selected as edges with no extra thresholding applied. For ANTAC, edges are selected by the theoretical bound with  $\xi_0 = 2$ . The results are summarized in Table 3. It can be seen that ANTAC achieves superior specificity and precision. Besides, ANTAC has the best overall performance in terms of the Matthews correlation coefficient.

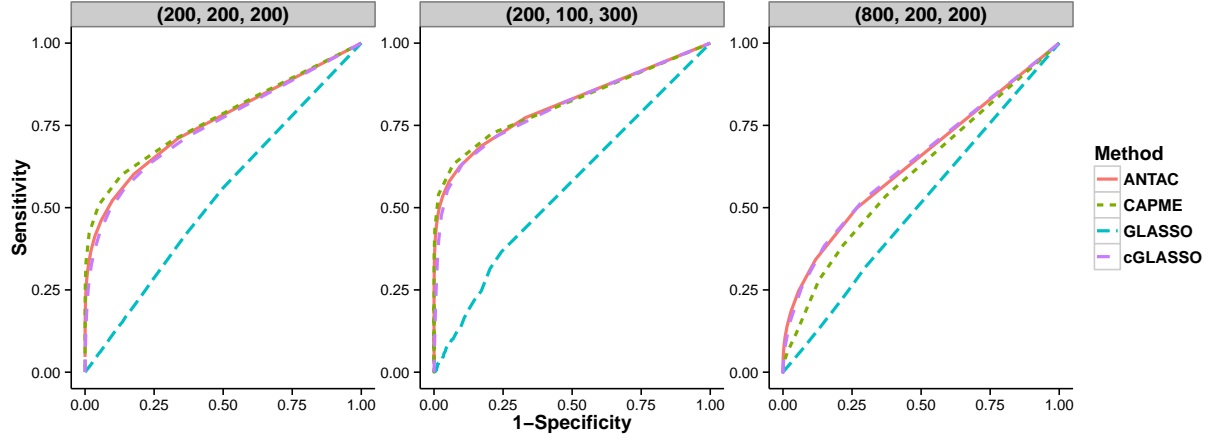


Figure 2: The ROC curves for different methods. For GLASSO or cGLASSO, the ROC curve is obtained by varying its tuning parameter. For CAPME,  $\lambda_1$  is fixed as the value selected by the cross validation and the ROC curve is obtained by varying  $\lambda_2$ . For ANTAC, the ROC curve is obtained by varying the cut-off on P-values.

We further construct ROC curves to check how this result would vary by changing the tuning parameters. For GLASSO or cGLASSO, the ROC curve is obtained by varying the tuning parameter. For CAPME,  $\lambda_1$  is fixed as the value selected by the cross validation and the ROC curve is obtained by varying  $\lambda_2$ . For proposed ANTAC method, the ROC curve is obtained by varying the thresholding level  $\xi_0$ . When  $p$  is small, CAPME, cGLASSO and ANTAC have comparable performance. As  $p$  grows, both ANTAC and cGLASSO outperform CAPME.

The purpose of simulating “Homogeneous Model” is to compare the performance of ANTAC and other procedures under models with similar settings used in Cai et al. (2013). Overall the performance from all procedures is not satisfactory due to the difficulty of support recovery problem. All nonzero entries are sampled from a standard normal. Hence, most signals are very weak and hard to be recovered by any method, although ANTAC performs among the best in all three models.

**Heterogeneous Model** We consider some models where the diagonal entries of the precision matrix have different values. These models are different from “Homogeneous Model” and we call

Table 3: Simulation results of the support recovery for homogeneous models based on 50 replications. Specifically, the performance is measured by misspecification rate, specificity, sensitivity (recall rate), precision and the Matthews correlation coefficient with all the values multiplied by 100. Numbers in parentheses are the simulation standard deviations.

	$(p, q, n)$	Method	MISR	SPE	SEN	PRE	MCC
Model 1	(200, 200, 200)	GLASSO	35(1)	65(1)	37(2)	2(0)	1(1)
		cGLASSO	25(6)	76(6)	64(7)	6(1)	8(1)
		CAPME	2(0)	100(0)	4(1)	96(1)	21(1)
		ANTAC	2(0)	100(0)	4(0)	88(8)	18(2)
Model 2	(200, 100, 300)	GLASSO	43(0)	57(0)	51(2)	3(0)	3(1)
		cGLASSO	5(0)	97(0)	47(1)	25(1)	32(1)
		CAPME	4(0)	97(0)	56(1)	29(1)	39(1)
		ANTAC	2(0)	100(0)	22(1)	97(2)	46(1)
Model 3	(800, 200, 200)	GLASSO	19(1)	81(1)	19(1)	1(0)	0(0)
		cGLASSO	1(0)	100(0)	0(0)	100(0)	2(0)
		CAPME	1(0)	100(0)	0(0)	0(0)	0(0)
		ANTAC	1(0)	100(0)	7(0)	71(2)	22(1)

them “Heterogeneous Model”. The performance of ANTAC and other procedures are explored under “Magnified Block” model and “Heterogeneous Product” model, respectively. The ANTAC performs significantly better than GLASSO, cGLASSO and CAMPE in both settings.

In “Magnified Block” model, we apply the following randomized procedure to choose  $\Omega$  and  $\Gamma$ . We first simulate a  $50 \times 50$  matrix  $\Omega_B$  with diagonal entries being 1 and each non-diagonal entry i.i.d. being nonzero with  $P(\omega_{ij} \neq 0) = 0.02$ . If  $\omega_{ij} \neq 0$ , we sample  $\omega_{ij}$  from  $\{0.4, 0.5\}$ . Then we generate two matrices by multiplying  $\Omega_B$  by 5 and 10, respectively. Then we align three matrices along the diagonal, resulting in a block diagonal matrix  $\Omega$  with sequentially magnified signals. A visualization of the simulated precision matrix is shown in Figure 3. The  $150 \times 100$  matrix  $\Gamma$  is simulated with each entry being nonzero i.i.d. follows  $N(0, 1)$  with  $P(\Gamma_{ij} \neq 0) = 0.05$ . Once the matrices  $\Omega$  and  $\Gamma$  are chosen, 50 replicates of  $X$  and  $Y$  are generated.

In “Heterogeneous Product” model, the matrices  $\Omega$  and  $\Gamma$  are chosen in the following ran-

domized way. We first simulate a  $200 \times 200$  matrix  $\Omega$  with diagonal entries being 1 and each non-diagonal entry i.i.d. being nonzero with  $P(\omega_{ij} \neq 0) = 0.005$ . If  $\omega_{ij} \neq 0$ , we sample  $\omega_{ij}$  from  $\{0.4, 0.5\}$ . Then we replace the  $100 \times 100$  submatrix at bottom-right corner by multiplying the  $100 \times 100$  submatrix at up-left corner by 2, which results in a precision matrix with possibly many different product values  $\omega_{ii}\sigma_{jj}$  over all  $i, j$  pairs, where  $\sigma_{jj}$  is the  $j$ th diagonal entry of the covariance matrix  $\Sigma = (\sigma_{kl})_{p \times p} = \Omega^{-1}$ . Thus we call it “Heterogeneous Product” model. A visualization of the simulated precision matrix is shown in Figure 4. The  $200 \times 100$  matrix  $\Gamma$  is simulated with each entry being nonzero i.i.d. follows  $N(0, 1)$  with  $P(\Gamma_{ij} \neq 0) = 0.05$ . Once  $\Omega$  and  $\Gamma$  are chosen, 50 replicates of  $X$  and  $Y$  are generated.

We compare our method with GLASSO, CAPME and cGLASSO procedures in “Heterogeneous Model”. We first compare the performance of support recovery when a single procedure from each method is applied. The tuning parameters for each procedure are set in the same way as in “Homogeneous Model” except that for CAPME, the optimal tuning parameter is achieved via a  $10 \times 10$  grid search on  $\{(\lambda_1, \lambda_2)\}$  by five-fold cross validation. We summarize the support recovery results in Table 4. A visualization of the support recovery result for a replicate of “Magnified Block” model and a replicate of “Heterogeneous Product” model are shown in Figure 3 and 4 respectively. In both models, ANTAC significantly outperforms others and achieves high precision and recall rate. Specifically, ANTAC has precision of 0.99 for two models respectively while no other procedure achieves precision rate higher than 0.21 in either model. Besides, ANTAC returns true sparse graph structure while others report much denser results.

Moreover, we construct the precision-recall curve to compare a sequence of procedures from different methods. In terms of precision-recall curve, CAPME has closer performance as the proposed method in “Magnified Block” model whereas cGLASSO has closer performance as the proposed method in “Heterogeneous Product” model, which indicates the proposed method performs comparable to the better of CAPME and cGLASSO. Another implication indicates from the precision-recall curve is that while the tuning free ANTAC method is always close to the best point along the curve created by using different values of threshold  $\xi_0$ , CAPME and cGLASSO via cross validation cannot select the optimal parameters on their corresponding precision-recall

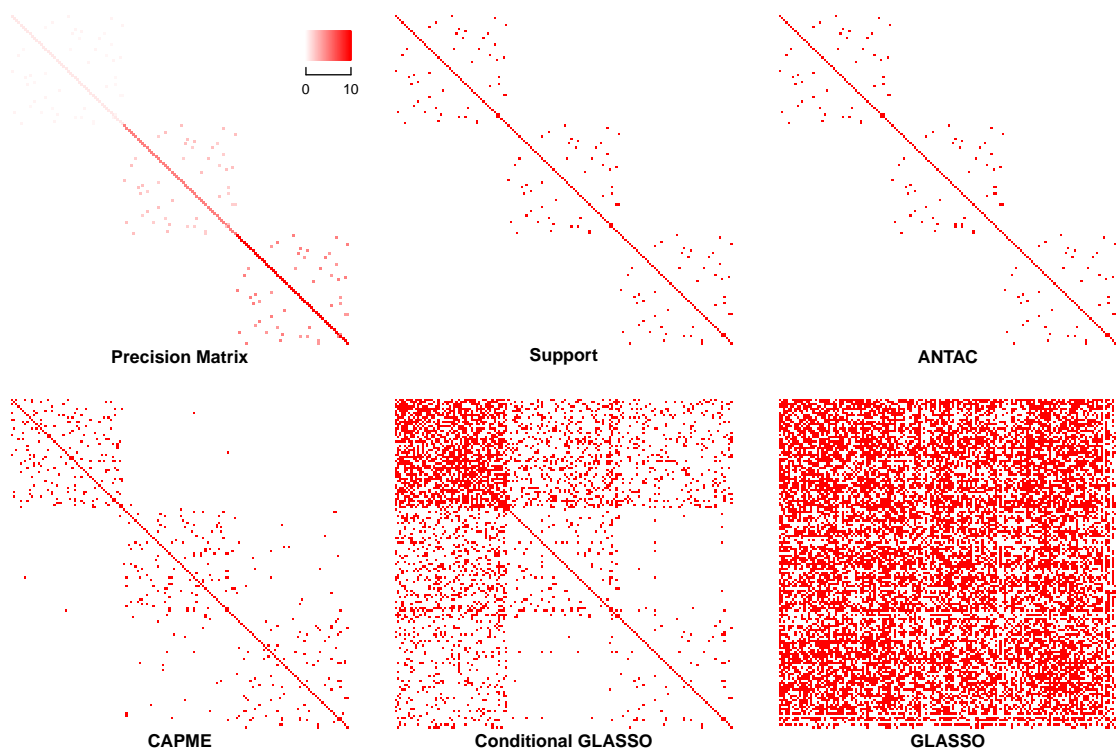


Figure 3: Heatmap of support recovery using different methods for a "Magnified Block" model.

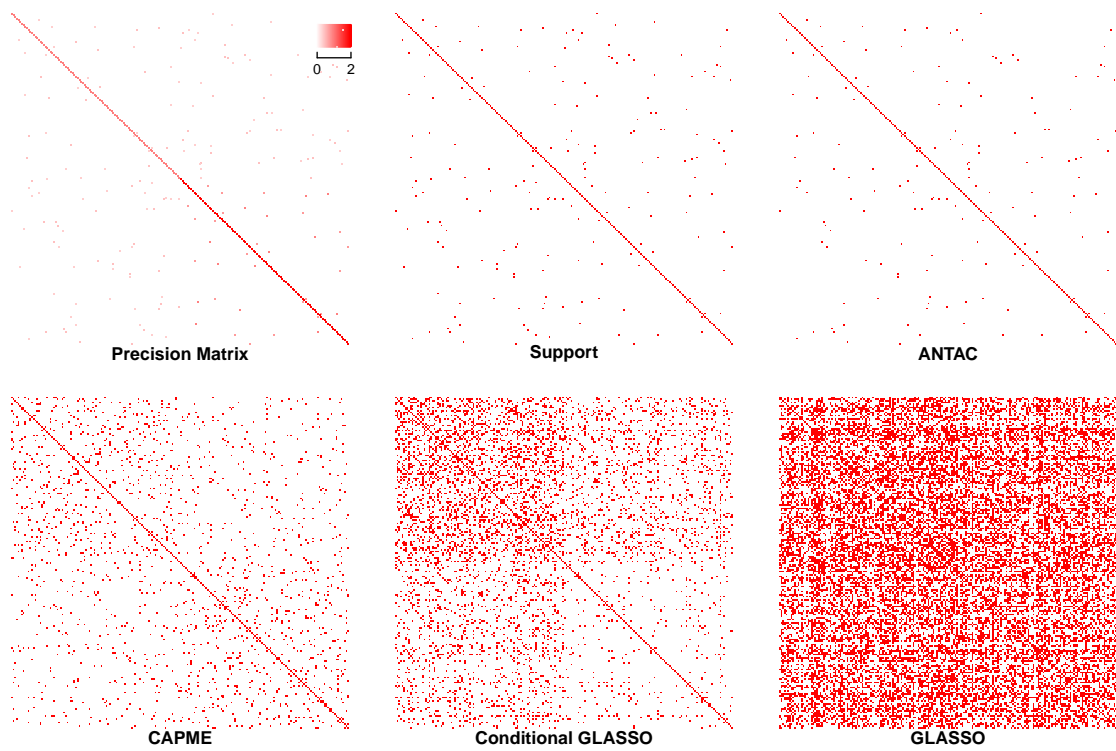


Figure 4: Heatmap of support recovery using different methods for a “Heterogeneous Product” model.

Table 4: Simulation results of the support recovery for heterogeneous models based on 50 replications. The performance is measured by overall error rate, specificity, sensitivity (recall rate), precision and the Matthews correlation coefficient with all the values multiplied by 100. Numbers in parentheses are the simulation standard deviations.

	$(p, q, n)$	Method	MISR	SPE	SEN	PRE	MCC
Magnified Block	(150, 100, 300)	GLASSO	54(0)	46(0)	80(4)	1(0)	4(1)
		cGLASSO	14(0)	86(0)	99(1)	4(0)	19(0)
		CAPME	1(0)	99(0)	99(0)	24(2)	58(1)
		ANTAC	0(0)	100(0)	98(1)	99(1)	99(1)
Heterogeneous Product	(200, 100, 300)	GLASSO	42(0)	58(0)	85(3)	1(0)	6(0)
		cGLASSO	12(0)	88(0)	100(0)	4(0)	18(0)
		CAPME	4(0)	96(0)	97(0)	15(0)	33(0)
		ANTAC	0(0)	100(0)	80(3)	99(1)	89(2)

curves, even though one of them has potentially good performance when using appropriate tuning parameters. Here is an explanation why the ANTAC procedure is better than CAPME and cGLASSO in “Magnified Block” model and “Heterogeneous Product” model settings. Recall that in the second stage, CAPME applies the same penalty level  $\lambda$  for each entry of the difference  $\Omega\hat{\Sigma} - I$ , where  $\hat{\Sigma}$  denotes the sample covariance matrix, but the  $i, j$  entry has variance  $\omega_{ii}\sigma_{jj}$  after scaling. Thus CAPME may not recover the support well in the “Heterogeneous Product” model settings, where the variances of different entries may be very different. As for the cGLASSO, we notice that essentially the same level of penalty is put on each entry  $\omega_{ij}$  while the variance of each entry in the  $i$ th row  $\Omega_{i\cdot}$  depends on  $\omega_{ii}$ . Hence we cannot expect cGLASSO performs very well in the “Magnified Block” model settings, where the diagonals  $\omega_{ii}$  vary a lot. In contrast, the ANTAC method adaptively puts the right penalty level (asymptotic variance) for each estimate of  $\omega_{ij}$ , therefore it works well in either setting.

Overall, the simulation results on heterogeneous models reveal the appealing practical properties of the ANTAC procedure. Our procedure enjoys tuning free property and has superior performance. In contrast, it achieves better precision and recall rate than the results from CAPME

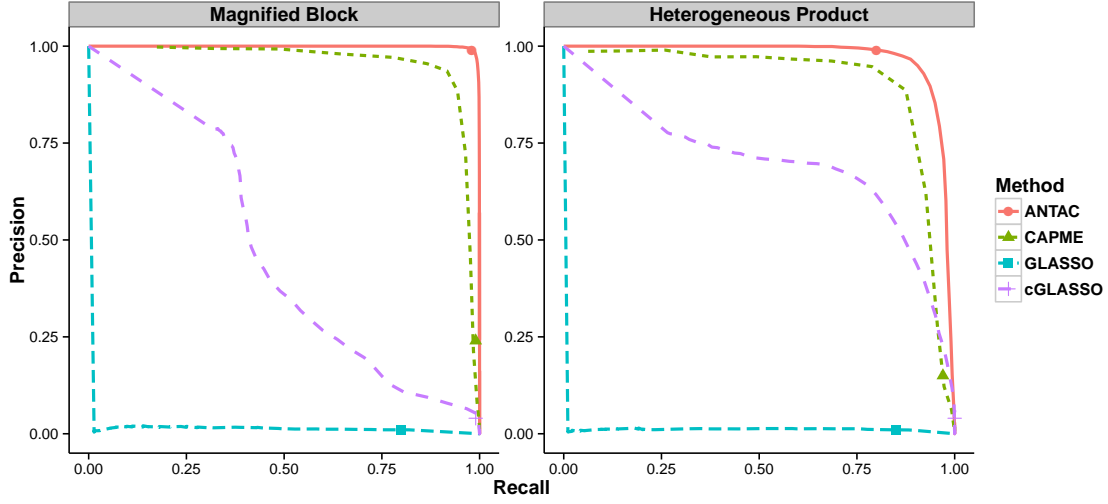


Figure 5: The precision-recall curves for “Magnified Block” model and “Heterogeneous Product” model using different methods. For GLASSO or cGLASSO, the curve is obtained by varying its tuning parameter. For CAPME,  $\lambda_1$  is fixed as the value selected by the cross validation and the curve is obtained by varying  $\lambda_2$ . For ANTAC, the precision-recall curve is obtained by varying threshold level  $\xi_0$ . The points on the curves correspond to the results obtained by cross-validation for GLASS, cGLASS and CAPME and by using theoretical threshold level  $\xi_0 = 2$  for tuning free ANTAC.

and cGLASSO using cross validation. Although in terms of precision-recall curve, the better of CAPME and cGLASSO is comparable with our procedure, generally the optimal sensitivity and specificity could not be obtained through cross-validation.

## 6. APPLICATION TO AN EQTL STUDY

We apply the ANTAC procedure to a yeast dataset from Smith & Kruglyak (2008) (GEO accession number GSE9376), which consists of 5,493 gene expression probes and 2,956 genotyped markers measured in 109 segregants derived from a cross between BY and RM treated with glucose. We find the proposed method achieves both better interpretability and accuracy in this example.

There are many mechanisms leading to the dependency of genes at the expression level. Among those, the dependency between transcription factors (TFs) and their regulated genes has been intensively investigated. Thus the gene-TF binding information could be utilized as an external biological evidence to validate and interpret the estimation results. Specifically, we used the high-confidence TF binding site (TFBS) profiles from m:Explorer, a database recently compiled using perturbation microarray data, TF-DNA binding profiles and nucleosome positioning measurements (Reimand, Aun, Vilo, Vaquerizas, Sedman & Luscombe 2012).

We first focus our analysis on a medium size dataset that consists of 121 genes on the yeast cell cycle signaling pathway (from the Kyoto Encyclopedia of Genes and Genomes database (Kanehisa & Goto 2000)). There are 119 markers marginally associated with at least 3 of those 121 genes with a Bonferroni corrected P-value less than 0.01. The parameters  $\lambda_1$  and  $\lambda_2$  for the ANTAC method are set as described in Theorem 3. 55 edges are identified using a cutoff of 0.01 on the FDR controlled P-values and 200 edges with a cutoff of 0.05. The number of edges goes to 375 using a cutoff of 0.1. For the purpose of visualization and interpretation, we further focus on 55 edges resulted from the cutoff of 0.01.

We then check how many of these edges involve at least one TF and how many TF-gene pairs are documented in the m:Explorer database. In 55 detected edges, 12 edges involve at least one TF and 2 edges are documented. In addition, we obtain the estimation of precision matrix from CAPME, where the tuning parameters  $\lambda_1 = 0.356$  and  $\lambda_2 = 0.5$  are chosen by five-fold cross val-

idation. To compare with the results from ANTAC, we select top 55 edges from CAPME solution based on the magnitude of partial correlation. Within these 55 edges, 13 edges involve at least one TF and 2 edges are documented. As shown in Figure 6, 22 edges are detected by both methods. Our method identifies a promising cell cycle related subnetwork featured by CDC14, PDS1, ESP1 and DUN1, connecting through GIN4, CLB3 and MPS1. In the budding yeast, CDC14 is a phosphatase functions essentially in late mitosis. It enables cells to exit mitosis through dephosphorylation and activation of the enemies of CDKs (Wurzenberger & Gerlich 2011). Throughout G1, S/G2 and early mitosis, CDC14 is inactive. The inactivation is partially achieved by PDS1 via its inhibition on an activator ESP1 (Stegmeier, Visintin, Amon et al. 2002). Moreover, DUN1 is required for the nucleolar localization of CDC14 in DNA damage-arrested yeast cells (Liang & Wang 2007).

We then extend the analysis to a larger dataset constructed from GSE9376. For 285 TFs documented in m:Explorer database, expression levels of 20 TFs are measured in GSE9376 with variances greater than 0.25. For these 20 TFs, 875 TF-gene interactions with 377 genes with variances greater than 0.25 are documented in m:Explorer. Applying the screening strategy as the previous example, we select 644 genetic markers marginally associated with at least 5 of the 377 genes with a Bonferroni corrected P-value less than 0.01. We apply the proposed ANTAC method and CAPME to this new dataset. For ANTAC, the parameters  $\lambda_1$  and  $\lambda_2$  are set as described in Theorem 3. For CAPME, the tuning parameters  $\lambda_1 = 0.089$  and  $\lambda_2 = 0.281$  are chosen by five-fold cross validation. We use TF-gene interactions documented in m:Explorer as an external biological evidence to validate the results. The results are summarized in Table 5. For ANTAC, 540 edges are identified using a cutoff of 0.05 on the FDR controlled P-values. Within these edges, 67 edges are TF-gene interactions and 44 out of 67 are documented in m:Explorer. In comparison, 8499 nonzero edges are detected by CAPME, where 915 edges are TF-gene interactions and 503 out of 915 are documented. This result is hard to interpret biologically. We further ask if identifying the same number of TF-gene interactions, which method achieves higher accuracy according to the concordance with m:Explorer. Based on the magnitude of partial correlation, we select top 771 edges from the CAPME solution, which capture 67 TF-gene interactions. Within these inter-

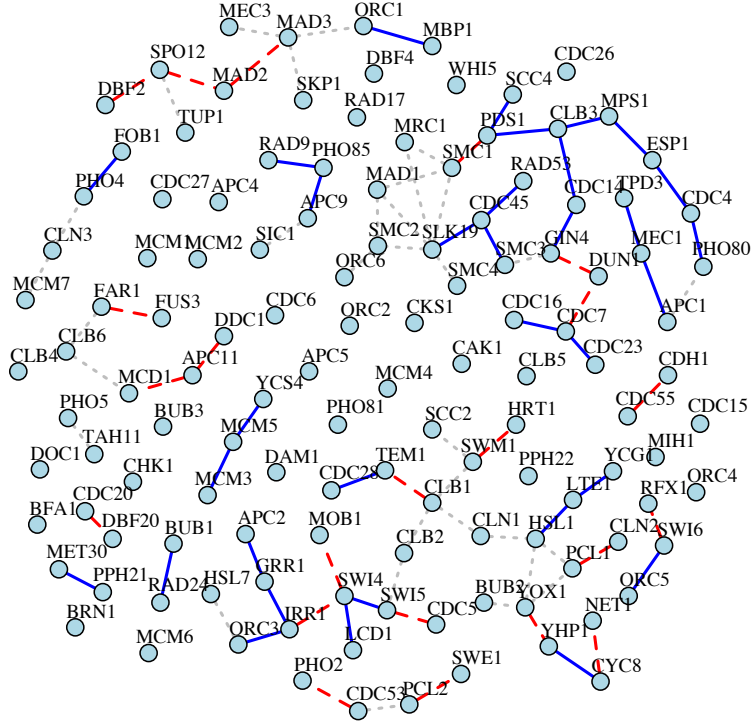


Figure 6: Visualization of the network constructed from yeast cell cycle expression data by CAPME and the proposed ANTAC method. For ANTAC, 55 edges are identified using a cut-off of 0.01 on the FDR controlled p-values. For CAPME, top 55 edges are selected based on the magnitude of partial correlation. 22 common edges detected by both methods are shown in dashed lines. Edges only detected by the proposed method are shown in solid lines. CAPME-specific edges are shown in dotted lines.

actions, 38 are documented in m:Explorer. Thus in this example, the proposed ANTAC method achieves both better interpretability and accuracy.

Table 5: Results for a dataset consists of 644 markers and 377 genes, which was constructed from GSE9376.

Method	Solution Criteria	Total # of gene-gene interactions	# of TF-gene interactions	# of TF-gene interactions documented	Documented Proportion
ANTAC	FDR controlled P-values $\leq 0.05$	540	67	44	65%
CAPME	Magnitude of partial correlation	771	67	38	57%
CAPME	Nonzero entries	8499	915	503	55%

## 7. PROOF OF MAIN THEOREMS

In this section, we will prove the main results Theorems 1 and 2.

### 7.1 Proof of Theorem 1

This proof is based on the key lemma, Lemma 5, which is deterministic in nature. We apply Lemma 5 with  $\mathbf{R} = \mathbf{D}b^{true} + \mathbf{E}$  replaced by  $\mathbf{Y}_j = \mathbf{X}\gamma_j + \mathbf{Z}_j$ ,  $\lambda$  replaced by  $\lambda_1$  and sparsity  $s$  replaced by  $s_1$ . The following lemma is the key to the proof.

**Lemma 1** *There exist some constants  $C'_k$ ,  $1 \leq k \leq 3$  such that for each  $1 \leq j \leq p$ ,*

$$\|\gamma_j - \hat{\gamma}_j\|_1 \leq C'_1 \lambda_1 s_1, \quad (33)$$

$$\|\gamma_j - \hat{\gamma}_j\| \leq C'_2 \lambda_1 \sqrt{s_1}, \quad (34)$$

$$\|\mathbf{X}(\gamma_j - \hat{\gamma}_j)\|^2 / n \leq C'_3 \lambda_1^2 s_1, \quad (35)$$

with probability  $1 - o(q^{-\delta_1+1})$ .

With the help of the lemma above, it's trivial to finish our proof. In fact,  $\|\hat{\mathbf{Z}}_j - \mathbf{Z}_j\|^2 = \|\mathbf{X}(\gamma_j - \hat{\gamma}_j)\|^2$  and hence Equation (14) immediately follows from result (35). Equations (15)

and (16) are obtained by the union bound and Equations (33) and (34) because of the following relationship,

$$\begin{aligned}\left\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\right\|_{l_\infty} &= \max_j \|\gamma_j - \hat{\gamma}_j\|_1, \\ \frac{1}{p} \left\|\hat{\mathbf{\Gamma}} - \mathbf{\Gamma}\right\|_F^2 &= \frac{1}{p} \sum_{j=1}^p \|\gamma_j - \hat{\gamma}_j\|^2.\end{aligned}$$

It is then enough to prove Lemma 1 to complete the proof of Theorem 1.

**Proof of Lemma 1** The Lemma is an immediate consequence of Lemma 5 applied to  $\mathbf{Y}_j = \mathbf{X}\gamma_j + \mathbf{Z}_j$  with tuning parameter  $\lambda_1$  and sparsity  $s_1$ . To show the union of Equations (33)-(35) holds with high probability  $1 - o(q^{-\delta_1+1})$ , we only need to check the following conditions of the Lemma 5 hold with probability at least  $1 - o(q^{-\delta_1+1})$ ,

$$\begin{aligned}I_1 &= \left\{ \nu \leq \theta^{ora} \lambda_1 \frac{\xi - 1}{\xi + 1} (1 - \tau) \text{ for some } \xi > 1 \right\}, \\ I_2 &= \left\{ \frac{\|\mathbf{X}_k\|}{\sqrt{n}} \in [1/A_1, A_1] \text{ for all } k \right\}, \\ I_3 &= \left\{ \theta^{ora} \in [1/A_2, A_2], \text{ where } \theta^{ora} = \frac{\|\mathbf{Z}_j\|}{\sqrt{n}} \right\}, \\ I_4 &= \left\{ \begin{array}{l} \frac{\mathbf{W}^T \mathbf{W}}{n} \text{ satisfies lower-RE with } (\alpha_1, \zeta(n, q)) \text{ s.t.} \\ s_1 \zeta(n, q) 8(1 + \xi)^2 A_1^2 \leq \min\{\frac{\alpha_1}{2}, 1\} \end{array} \right\},\end{aligned}$$

where we set  $\xi = 3/\varepsilon_1 + 1$  for the current setting,  $A_1 = C_{A_1} \max\{B, \sqrt{M_1}\}$  under Condition 3 and  $A_1 = C_{A_1} \sqrt{M_1}$  under Condition 3' for some universal constant  $C_{A_1} > 0$ ,  $A_2 = \sqrt{2M_2}$ ,  $\alpha_1 = \frac{1}{2M_1}$  and  $\zeta(n, q) = o(1/s_1)$ . Let us still define  $\mathbf{W} = \mathbf{X} \cdot \text{diag}\left(\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right)$  as the standardized  $\mathbf{X}$  in Lemma 5 of Section 8.

We will show that

$$\begin{aligned}\mathbb{P}\{I_1^c\} &\leq O\left(q^{-\delta_1+1}/\sqrt{\log q}\right), \\ \mathbb{P}\{I_i^c\} &\leq o(q^{-\delta_1}) \text{ for } i = 2, 3 \text{ and } 4,\end{aligned}$$

which implies

$$\mathbb{P}\{E_j\} \geq 1 - o(q^{-\delta_1+1}),$$

where  $E_j$  is the union of  $I_1$  to  $I_4$ . We will first consider  $\mathbb{P}\{I_2^c\}$  and  $\mathbb{P}\{I_3^c\}$ , then  $\mathbb{P}\{I_4^c\}$ , and leave  $\mathbb{P}\{I_1^c\}$  to the last, which relies on the bounds for  $\mathbb{P}\{I_i^c\}$ ,  $2 \leq i \leq 4$ .

(1). To study  $\mathbb{P}\{I_2^c\}$  and  $\mathbb{P}\{I_3^c\}$ , we need the following Bernstein-type inequality (See e.g. (Vershynin 2010), Section 5.2.4 for the inequality and definitions of norms  $\|\cdot\|_{\varphi_2}$  and  $\|\cdot\|_{\varphi_1}$ ) to control the tail bound for sum of i.i.d. sub-exponential variables,

$$\mathbb{P}\left\{\left|\sum_{i=1}^n U_i\right| \geq tn\right\} \leq 2 \exp\left(-cn \min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\}\right), \quad (36)$$

for  $t > 0$ , where  $U_i$  are i.i.d. centered sub-exponential variables with parameter  $\|U_i\|_{\varphi_1} \leq K$  and  $c$  is some universal constant. Notice that  $\theta^{ora} = \|\mathbf{Z}_j\|/\sqrt{n}$  with  $\mathbf{Z}_j^{(1)} \sim N(0, \sigma_{jj})$ , and  $\mathbf{X}_k^{(1)}$  is sub-gaussian with parameter  $\|\mathbf{X}_k^{(1)}\|_{\varphi_2} \in [c_{\varphi_2} M_1^{-1/2}, C_{\varphi_2} B]$  by Conditions 2 and 3 with some universal constants  $c_{\varphi_2}$  and  $C_{\varphi_2}$  (all formulas involving  $\varphi_1$  or  $\varphi_2$  parameter of  $\mathbf{X}_k^{(1)}$  replace “ $B$ ” by “ $\sqrt{M_1}$ ” under Condition 3’ hereafter). The fact that sub-exponential is sub-gaussian squared implies that there exists some universal constant  $C_1 > 0$  such that

$$\begin{aligned} n((\theta^{ora})^2 - \sigma_{jj}) & \text{ is sum of i.i.d. sub-exponential with } \varphi_1 \text{ parameter } C_1 \sigma_{jj}, \\ \|\mathbf{X}_k\|^2 - n\mathbb{E}x_k^2 & \text{ is sum of i.i.d. sub-exponential with } \varphi_1 \text{ parameter } \left\|\left(\mathbf{X}_k^{(1)}\right)^2\right\|_{\varphi_1}. \end{aligned}$$

Note that  $\sigma_{jj} \in [1/M_2, M_2]$  by Condition 2 and  $\left\|\left(\mathbf{X}_k^{(1)}\right)^2\right\|_{\varphi_1} \in [c'_{\varphi_1} M_1^{-1}, C'_{\varphi_1} B^2]$  with some universal constants  $c'_{\varphi_1}$  and  $C'_{\varphi_1}$ . Let  $A_1 = C_{A_1} \max\{B, \sqrt{M_1}\}$  and  $A_2 = \sqrt{2M_2}$  for some large constant  $C_{A_1}$ . Equation (36) with a sufficiently small constant  $c_0$  implies

$$\begin{aligned} \mathbb{P}\{I_3^c\} &= \mathbb{P}\left\{(\theta^{ora})^2 \notin [1/A_2^2, A_2^2]\right\} \leq \mathbb{P}\left\{n|(\theta^{ora})^2 - \sigma_{jj}| \geq \frac{\sigma_{jj}}{2}n\right\} \\ &\leq 2 \exp(-C'n) \leq o(q^{-\delta_1}), \end{aligned} \quad (37)$$

and

$$\begin{aligned} \mathbb{P}\{I_2^c\} &= \mathbb{P}\left\{\frac{\|\mathbf{X}_k\|^2}{n} \notin [1/A_1^2, A_1^2] \text{ for some } k\right\} \\ &\leq q \mathbb{P}\left\{|\|\mathbf{X}_k\|^2 - n\mathbb{E}x_k^2| > c_0 \left\|\left(\mathbf{X}_k^{(1)}\right)^2\right\|_{\varphi_1} n\right\} \leq q 2 \exp(-C'n) \leq o(q^{-\delta_1}). \end{aligned} \quad (38)$$

(2). To study the lower-RE condition of  $\frac{\mathbf{W}^T \mathbf{W}}{n}$ , we essentially need to study  $\frac{\mathbf{X}^T \mathbf{X}}{n}$ . On the event  $I_2$ , it's easy to see that if  $\frac{\mathbf{X}^T \mathbf{X}}{n}$  satisfies lower-RE with  $(\alpha_x, \zeta_x(n, q))$ , then  $\frac{\mathbf{W}^T \mathbf{W}}{n}$  satisfies lower-RE with  $(\alpha_1, \zeta) = (\alpha_x A_1^{-2}, \zeta_x A_1^2)$ , since  $\mathbf{W} = \mathbf{X} \text{diag}\left(\frac{\sqrt{n}}{\|\mathbf{X}_k\|}\right)$  and  $\frac{\|\mathbf{X}_k\|}{\sqrt{n}} \in [1/A_1, A_1]$  on event  $I_2$ . Moreover, to study the lower-RE condition, the following lemma implies that we only need to consider the behavior of  $\frac{\mathbf{X}^T \mathbf{X}}{n}$  on sparse vectors.

**Lemma 2** *For any symmetric matrix  $\Delta_{q \times q}$ , suppose  $|v^T \Delta v| \leq \delta$  for any unit  $2s$  sparse vector  $v \in \mathbb{R}^q$ , i.e.  $\|v\| = 1$  and  $v \in \mathbb{B}_0(2s) = \{a : \sum_{i=1}^q 1\{a_i \neq 0\} \leq 2s\}$ , then we have*

$$|v^T \Delta v| \leq 27\delta \left( \|v\|^2 + \frac{1}{s} \|v\|_1^2 \right) \text{ for any } v \in \mathbb{R}^q.$$

See Supplementary Lemma 12 in (Loh & Wainwright 2012) for the proof. With a slight abuse of notation, we define  $\Sigma_x = \text{Cov}(X^{(1)})$ . By applying Lemma 2 on  $\frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x$  and Condition 2,  $v^T \Sigma_x v \geq \frac{1}{M_1} \|v\|^2$ , we know that  $\frac{\mathbf{X}^T \mathbf{X}}{n}$  satisfies lower-RE  $(\alpha_x, \zeta_x(n, q))$  with

$$\alpha_x \geq \frac{1}{M_1} \left( 1 - \frac{27}{L} \right) \text{ and } \zeta_x(n, q) \leq \frac{27}{s_1 M_1 L}, \quad (39)$$

provided

$$\left| v^T \left( \frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x \right) v \right| \leq \frac{1}{L M_1} \text{ for all } v \in \mathbb{B}_0(2s_1) \text{ with } \|v\| = 1, \quad (40)$$

which implies that the population covariance matrix  $\Sigma_x$  and its sample version  $\frac{\mathbf{X}^T \mathbf{X}}{n}$  behave similarly on all  $2s_1$  sparse vectors.

Now we show Equation (40) holds under Conditions 3 and 3' respectively for a sufficiently large constant  $L$  such that the inequality in the event  $I_4$  holds. Under Condition 3',  $X^{(1)}$  is jointly sub-gaussian with parameter  $(M_1)^{1/2}$ . A routine one-step chaining (or  $\delta$ -net) argument implies that there exists some constant  $c_{sg} > 0$  such that

$$\mathbb{P} \left\{ \sup_{v \in \mathbb{B}_0(2s_1)} \left| v^T \left( \frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x \right) v \right| > t \|v\|^2 \right\} \leq 2 \exp \left( -c_{sg} n \min \left\{ \frac{t^2}{M_1^2}, \frac{t}{M_1} \right\} + 2s_1 \log q \right). \quad (41)$$

See e.g. Supplementary Lemma 15 in (Loh & Wainwright 2012) for the proof. Hence by picking small  $t = \frac{1}{L M_1}$  with any fixed but arbitrary large  $L$ , the sparsity condition (17)  $s_1 = o\left(\frac{n}{\log q}\right)$  and Equation (41) imply that Equation (40) holds with probability  $1 - o(p^{-\delta_1})$ .

Under Condition 3, if  $s_1 = o\left(\sqrt{\frac{n}{\log q}}\right)$ , Hoeffding's inequality and a union bound imply that

$$\mathbb{P}\left\{\left\|\frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x\right\|_\infty > 2B\sqrt{\frac{(1 + \delta_1) \log q}{n}}\right\} \leq 2q^{-2\delta_1},$$

where the norm  $\|\cdot\|_\infty$  denotes the entry-wise supnorm and  $\|X^{(1)}\|_\infty \leq B$  (see, e.g. (Massart 2007) Proposition 2.7). Thus with probability  $1 - o(q^{-\delta_1})$ , we have for any  $v \in \mathbb{B}_0(2s_1)$  with  $\|v\| = 1$ ,

$$\left|v^T \left(\frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x\right) v\right| \leq \left\|\frac{\mathbf{X}^T \mathbf{X}}{n} - \Sigma_x\right\|_\infty \|v\|_1^2 \leq 2B\sqrt{\frac{(1 + \delta_1) \log q}{n}} \frac{1}{2s_1} = o(1),$$

where the last inequality follows from  $\|v\|_1 \leq \sqrt{2s_1} \|v\|$  for any  $v \in \mathbb{B}_0(2s_1)$ . Therefore we have Equation (40) holds with probability  $1 - o(q^{-\delta_1})$  for any arbitrary large  $L$ . If  $s_1 = o\left(\frac{n}{\log^3 n \log q}\right)$ , an involved argument using Dudley's entropy integral and Talagrand's concentration theorem for empirical processes implies (see, (Rudelson & Zhou 2013) Theorem 23 and its proof) Equation (40) holds with probability  $1 - o(q^{-\delta_1})$  for any fixed but arbitrary large  $L$ .

Therefore we showed that under Condition 3 or 3', Equation (40) holds with probability  $1 - o(q^{-\delta_1})$  for any arbitrary large  $L$ . Consequently, Equation (39) with  $(\alpha_1, \zeta) = (\alpha_x A_1^{-2}, \zeta_x A_1^2)$  on event  $I_2$  implies that we can pick  $\alpha_1 = \frac{1}{2M_1}$  and sufficiently small  $\zeta$  such that event  $I_4$  holds with probability  $1 - o(q^{-\delta_1})$ .

(3). Finally we study the probability of event  $I_1$ . The following tail probability of  $t$  distribution is helpful in the analysis.

**Proposition 1** *Let  $T_n$  follows a  $t$  distribution with  $n$  degrees of freedom. Then there exists  $\epsilon_n \rightarrow 0$  as  $n \rightarrow \infty$  such that  $\forall t > 0$*

$$\mathbb{P}\left\{T_n^2 > n\left(e^{2t^2/(n-1)} - 1\right)\right\} \leq (1 + \epsilon_n) e^{-t^2} / (\pi^{1/2} t).$$

Please refer to (Sun & Zhang 2012) Lemma 1 for the proof. Recall that  $I_1 = \left\{\frac{\nu}{\theta^{ora}} \leq \lambda_1 \frac{\xi-1}{\xi+1} (1 - \tau)\right\}$ , where  $\tau$  defined in Equation (67) satisfies that  $\tau = O(s_1 \lambda_1^2) = o(1)$  on  $\cap_{i=2}^4 I_i$ . From the definition of  $\nu$  in Equation (66) we have

$$\frac{\nu}{\theta^{ora}} = \max_k |h_k|, \text{ with } h_k = \frac{\mathbf{W}_k^T \mathbf{Z}_j}{n\theta^{ora}},$$

where each column of  $\mathbf{W}$  has norm  $\|\mathbf{W}_k\| = \sqrt{n}$ . Given  $\mathbf{X}$ , equivalently  $\mathbf{W}$ , it's not hard to check that we have  $\frac{\sqrt{n-1}h_k}{\sqrt{1-h_k^2}} \sim t_{(n-1)}$  by the normality of  $\mathbf{Z}_j$ , where  $t_{(n-1)}$  is the  $t$  distribution with  $(n-1)$  degrees of freedom. From Proposition 1 we have

$$\begin{aligned} & \mathbb{P} \left\{ |h_k| > \sqrt{\frac{2t^2}{n}} \right\} \\ = & \mathbb{P} \left\{ \frac{(n-1)h_k^2}{1-h_k^2} > \frac{2(n-1)t^2/n}{1-2t^2/n} \right\} \leq \mathbb{P} \left\{ \frac{(n-1)h_k^2}{1-h_k^2} > \frac{2(n-1)t^2/(n-2)}{1-t^2/(n-2)} \right\} \\ \leq & \mathbb{P} \left\{ \frac{(n-1)h_k^2}{1-h_k^2} > (n-1) \left( e^{2t^2/(n-2)} - 1 \right) \right\} \leq (1 + \epsilon_{n-1}) e^{-t^2} / (\pi^{1/2} t), \end{aligned}$$

where the first inequality holds when  $t^2 \geq 2$ , and the second inequality follows from the fact  $e^x - 1 \leq x/(1 - \frac{x}{2})$  for  $0 < x < 2$ . Now let  $t^2 = \delta_1 \log q > 2$ , and  $\lambda_1 = (1 + \epsilon_1) \sqrt{\frac{2\delta_1 \log q}{n}}$  with  $\xi = 3/\epsilon_1 + 1$ , then we have  $\lambda_1 \frac{\xi-1}{\xi+1} (1 - \tau) > \sqrt{\frac{2\delta_1 \log q}{n}}$  and

$$\begin{aligned} \mathbb{P} \{ \cap_{i=1}^4 I_i \} & \geq \mathbb{P} \left\{ \frac{\nu}{\theta^{ora}} \leq \sqrt{\frac{2\delta_1 \log q}{n}} \right\} - \mathbb{P} \{ (\cap_{i=2}^4 I_i)^c \} \\ & \geq 1 - q \cdot \mathbb{P} \left\{ |h_k| > \sqrt{\frac{2\delta_1 \log q}{n}} \right\} - \mathbb{P} \{ (\cap_{i=2}^4 I_i)^c \} \\ & \geq 1 - \left( \frac{1}{\sqrt{\pi\delta_1}} + o(1) \right) \frac{q^{-\delta_1+1}}{\sqrt{\log q}}, \end{aligned}$$

which immediately implies  $\mathbb{P} \{ I_1^c \} \leq O(q^{-\delta_1+1}/\sqrt{\log q})$ .

## 7.2 Proof of Theorem 2

The whole proof is based on the results in Theorem 1. In particular with probability  $1 - o(p \cdot q^{-\delta_1+1})$ , the following events hold,

$$\frac{1}{n} \left\| \hat{\mathbf{Z}}_j - \mathbf{Z}_j \right\|^2 \leq C_1 s_1 \frac{\log q}{n} \text{ for all } j, \quad (42)$$

$$\left\| \hat{\gamma}_j - \gamma_j \right\|_1 \leq C_2 s_1 \sqrt{\frac{\log q}{n}} \text{ for all } j. \quad (43)$$

From now on the analysis is conditioned on the two events above. This proof is also based on the key lemma, Lemma 5. We apply Lemma 5 with  $\mathbf{R} = \mathbf{D}b^{true} + \mathbf{E}$  replaced by  $\hat{\mathbf{Z}}_m = \hat{\mathbf{Z}}_{A^c} \beta_m + \mathbf{E}_m$  for each  $m \in A = \{i, j\}$ ,  $\lambda$  replaced by  $\lambda_2$  and sparsity  $s$  replaced by  $C_\beta s_2$ , where  $\mathbf{E}_m$  is defined

by the regression model Equation (7) as

$$\begin{aligned}\mathbf{E}_m &\equiv \boldsymbol{\epsilon}_m + \Delta_{Z_m} \\ &= \boldsymbol{\epsilon}_m + \left(\hat{\mathbf{Z}}_m - \mathbf{Z}_m\right) + \left(\mathbf{Z}_{A^c} - \hat{\mathbf{Z}}_{A^c}\right) \beta_m,\end{aligned}\tag{44}$$

and  $C_\beta = 2M_2$  because the definition of  $\beta_m$  in Equation (6) implies it is a weighted sum of two columns of  $\Omega$  with weight bounded by  $M_2$ .

To obtain our desired result  $\left|\hat{\psi}_{kl} - \psi_{kl}^{ora}\right| = \left|\boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_l / n - \hat{\boldsymbol{\epsilon}}_k^T \hat{\boldsymbol{\epsilon}}_l / n\right|$  for each pair  $k, l \in A = \{i, j\}$ , it's sufficient for us to bound  $\left|\hat{\boldsymbol{\epsilon}}_k^T \hat{\boldsymbol{\epsilon}}_l / n - \mathbf{E}_k^T \mathbf{E}_l / n\right|$  and  $\left|\boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_l / n - \mathbf{E}_k^T \mathbf{E}_l / n\right|$  separately and then to apply the triangle inequality. The following two lemmas are useful to establish those two bounds.

**Lemma 3** *There exists some constant  $C_{in} > 0$  such that  $\left|\boldsymbol{\epsilon}_k^T \boldsymbol{\epsilon}_l / n - \mathbf{E}_k^T \mathbf{E}_l / n\right| \leq C_{in} \lambda_1^2 s_1$  with probability  $1 - o(q^{-\delta_1})$ .*

**Lemma 4** *There exist some constants  $C'_k$ ,  $1 \leq k \leq 3$  such that for each  $m \in A = \{i, j\}$ ,*

$$\begin{aligned}\left\|\beta_m - \hat{\beta}_m\right\|_1 &\leq C'_1 \lambda_2 s_2, \\ \left\|\hat{\mathbf{Z}}_{A^c} \left(\beta_m - \hat{\beta}_m\right)\right\|^2 / n &\leq C'_2 \lambda_2^2 s_2, \\ \left\|\hat{\mathbf{Z}}_{A^c}^T \mathbf{E}_m / n\right\|_\infty &\leq C'_3 \lambda_2,\end{aligned}$$

with probability  $1 - o(p^{-\delta_2+1})$ .

Before moving on, we point out a fact we will use several times in the proof,

$$\|\beta_m\|_1 \leq 2M_2^2,\tag{45}$$

which follows from Equation (6) and Condition 5. Hence we have

$$\|\Delta_{Z_m}\| / \sqrt{n} \leq \frac{\left\|\hat{\mathbf{Z}}_m - \mathbf{Z}_m\right\|}{\sqrt{n}} + \frac{\max_k \left\|\hat{\mathbf{Z}}_k - \mathbf{Z}_k\right\|}{\sqrt{n}} \|\beta_m\|_1 \leq \sqrt{C_3 s_1 \lambda_1^2},\tag{46}$$

for some constant  $C_3 > 0$ .

To bound the term  $|\hat{\epsilon}_k^T \hat{\epsilon}_l/n - \mathbf{E}_k^T \mathbf{E}_l/n|$ , we note that for any  $k, l \in A = \{i, j\}$ ,

$$\begin{aligned}
|\hat{\epsilon}_k^T \hat{\epsilon}_l/n - \mathbf{E}_k^T \mathbf{E}_l/n| &= \left| \left( \mathbf{E}_k + \hat{\mathbf{Z}}_{A^c} \left( \beta_k - \hat{\beta}_k \right) \right)^T \left( \mathbf{E}_l + \hat{\mathbf{Z}}_{A^c} \left( \beta_l - \hat{\beta}_l \right) \right) /n - \mathbf{E}_k^T \mathbf{E}_l/n \right| \\
&\leq \left\| \hat{\mathbf{Z}}_{A^c}^T \mathbf{E}_k/n \right\|_\infty \left\| \beta_k - \hat{\beta}_k \right\|_1 + \left\| \hat{\mathbf{Z}}_{A^c}^T \mathbf{E}_l/n \right\|_\infty \left\| \beta_l - \hat{\beta}_l \right\|_1 \\
&\quad + \left\| \hat{\mathbf{Z}}_{A^c} \left( \beta_k - \hat{\beta}_k \right) \right\| \cdot \left\| \hat{\mathbf{Z}}_{A^c} \left( \beta_l - \hat{\beta}_l \right) \right\| /n \\
&\leq (2C'_1 C'_3 + C'_2) \lambda_2^2 s_2,
\end{aligned} \tag{47}$$

where we applied Lemma 4 in the last inequality.

Lemma 3, together with Equation (47), immediately implies the desired result (23),

$$|\hat{\psi}_{kl} - \psi_{kl}^{ora}| \leq C'_4 \left( s_2 \frac{\log p}{n} + s_1 \frac{\log q}{n} \right),$$

for some constant  $C'_4$  with probability  $1 - o(p \cdot q^{-\delta_1+1} + p^{-\delta_2+1})$ . Since the spectrum of  $\Psi_{A,A}$  is bounded below by  $M_2^{-1}$  and above by  $M_2$  and the functional  $(\Omega_{A,A})_{kl} = (\Psi_{A,A}^{-1})_{kl}$  is Lipschitz in a neighborhood of  $\Psi_{A,A}$  for  $k, l \in A$ , we obtain that Equation (24) is an immediate consequence of Equation (23). Note that  $\omega_{ij}^{ora}$  is the MLE of  $\omega_{ij}$  in the model  $(\eta_i, \eta_j)^T \sim N(0, \Omega_{A,A}^{-1})$  with three parameters given  $n$  samples. Whenever  $s_2 = o\left(\frac{\sqrt{n}}{\log p}\right)$  and  $s_1 = o\left(\frac{\sqrt{n}}{\log q}\right)$ , we have  $s_2 \frac{\log p}{n} + s_1 \frac{\log q}{n} = o\left(\frac{1}{\sqrt{n}}\right)$ . Therefore we have  $\omega_{ij}^{ora} - \hat{\omega}_{ij} = o_p(\omega_{ij}^{ora} - \omega_{ij})$ , which immediately implies Equation (25) in Theorem 2,

$$\sqrt{nF_{ij}}(\hat{\omega}_{ij} - \omega_{ij}) \stackrel{D}{\sim} \sqrt{nF_{ij}}(\omega_{ij}^{ora} - \omega_{ij}) \xrightarrow{D} N(0, 1),$$

where  $F_{ij}$  is the Fisher information of  $\omega_{ij}$ .

It is then enough to prove Lemma 3 and Lemma 4 to complete the proof of Theorem 2.

**Proof of Lemma 3** We show that  $|\epsilon_k^T \epsilon_l/n - \mathbf{E}_k^T \mathbf{E}_l/n| \leq C_{in} \lambda_1^2 s_1$  with probability  $1 - o(q^{-\delta_1})$  in this section. By Equation (46), we have

$$|\epsilon_k^T \epsilon_l/n - \mathbf{E}_k^T \mathbf{E}_l/n| \leq \left| \frac{\epsilon_k^T \Delta_{Z_l} + \epsilon_l^T \Delta_{Z_k} + \Delta_{Z_l}^T \Delta_{Z_k}}{n} \right| \leq C_3 s_1 \lambda_1^2 + \frac{|\epsilon_k^T \Delta_{Z_l}| + |\epsilon_l^T \Delta_{Z_k}|}{n}. \tag{48}$$

To bound the term  $|\epsilon_k^T \Delta_{Z_l}|$ , we note that by the definition of  $\Delta_{Z_l}$  in Equation (44) there exists some constant  $C_4$  such that,

$$\begin{aligned}
\frac{|\epsilon_k^T \Delta_{Z_l}|}{n} &= \left| \epsilon_k^T \left[ (\hat{\mathbf{Z}}_l - \mathbf{Z}_l) + (\mathbf{Z}_{A^c} - \hat{\mathbf{Z}}_{A^c}) \beta_m \right] \right| / n \\
&= \left| \epsilon_k^T \mathbf{X} [(\gamma_l - \hat{\gamma}_l) + (\hat{\gamma}_{A^c} - \gamma_{A^c}) \beta_m] \right| / n \\
&\leq \left\| \epsilon_k^T \mathbf{X} / n \right\|_\infty \|(\gamma_l - \hat{\gamma}_l) + (\hat{\gamma}_{A^c} - \gamma_{A^c}) \beta_m\|_1 \\
&\leq \left\| \epsilon_k^T \mathbf{X} / n \right\|_\infty \max_i \|\gamma_i - \hat{\gamma}_i\|_1 (1 + \|\beta_m\|_1) \leq \left\| \epsilon_k^T \mathbf{X} / n \right\|_\infty C_4 s_1 \lambda_1, \quad (49)
\end{aligned}$$

where the last inequality follows from Equations (43) and (45). Since  $\epsilon_k \stackrel{i.i.d.}{\sim} N(0, \psi_{kk})$  and  $\mathbf{X}$  are independent, it can be seen that each coordinate of  $\epsilon_k^T \mathbf{X}$  is a sum of  $n$  i.i.d. sub-exponential variables with bounded parameter under either Condition 3 or 3'. A union bound with  $q$  coordinates and another application of Bernstein inequality in Equation (36) with  $t = C_5 \sqrt{\frac{\log q}{n}}$  imply that  $\left\| \epsilon_k^T \mathbf{X} / n \right\|_\infty \leq C_5 \sqrt{\frac{\log q}{n}}$  with probability  $1 - o(q^{-\delta_1})$  for some large constant  $C_5 > 0$ . This fact, together with Equation (49), implies that  $|\epsilon_k^T \Delta_{Z_l}| / n = O(s_1 \lambda_1^2)$  with probability  $1 - o(q^{-\delta_1})$ . Similar result holds for  $|\epsilon_l^T \Delta_{Z_k} / n|$ . Together with Equation (48), this result completes our claim on  $|\epsilon_k^T \epsilon_l / n - \mathbf{E}_k^T \mathbf{E}_l / n|$  and finishes the proof of Lemma 3.

**Proof of Lemma 4** The Lemma 4 is an immediate consequence of Lemma 5 applied to  $\hat{\mathbf{Z}}_m = \hat{\mathbf{Z}}_{A^c} \beta_m + \mathbf{E}_m$  for each  $m \in A = \{i, j\}$  with parameter  $\lambda_2$  and sparsity  $C_\beta s_2$ . We check the following conditions  $I_1 - I_4$  in the Lemma 5 hold with probability  $1 - o(p^{-\delta_2+1})$  to finish our proof. This part of the proof is similar to the proof of Lemma 1. We thus directly apply those facts already shown in the proof of Lemma 1 whenever possible. Let

$$\begin{aligned}
I_1 &= \left\{ \nu \leq \theta^{ora} \lambda_2 \frac{\xi - 1}{\xi + 1} (1 - \tau_2) \text{ for some } \xi > 1 \right\}, \\
I_2 &= \left\{ \frac{\|\hat{\mathbf{Z}}_k\|}{\sqrt{n}} \in [1/A'_1, A'_1] \text{ for all } k \in A^c \right\}, \\
I_3 &= \left\{ \theta^{ora} \in [1/A'_2, A'_2], \text{ where } \theta^{ora} = \frac{\|\mathbf{E}_m\|}{\sqrt{n}} \right\}, \\
I_4 &= \left\{ \begin{array}{l} \frac{\mathbf{W}^T \mathbf{W}}{n} \text{ satisfies lower-RE with } (\alpha_1, \zeta(n, p)) \text{ s.t.} \\ s_2 \zeta(n, p) 8(1 + \xi)^2 A_1'^2 \leq \min\{\frac{\alpha_1}{2}, 1\} \end{array} \right\},
\end{aligned}$$

where we can set  $\xi = 3/\varepsilon_2 + 1$  for the current setting,  $A'_1 = A'_2 = \sqrt{3M_2}$ ,  $\alpha_1 = \frac{1}{2M_2}$  and  $\zeta(n, p) = o(1/s_2)$ . Let us still define  $\mathbf{W} = \hat{\mathbf{Z}}_{A^c} \cdot \text{diag}\left(\frac{\sqrt{n}}{\|\hat{\mathbf{Z}}_k\|}\right)$  as the standardized  $\hat{\mathbf{Z}}_{A^c}$  in the Lemma 5 of Section 8. The strategy is to show that  $\mathbb{P}\{I_1^c\} \leq O(p^{-\delta_2+1}/\sqrt{\log p})$  and  $\mathbb{P}\{I_i^c\} \leq o(p^{-\delta_2})$  for  $i = 2, 3$  and 4, which completes our proof.

(1). To study  $\mathbb{P}\{I_2^c\}$  and  $\mathbb{P}\{I_3^c\}$ , we note that  $\frac{\|\hat{\mathbf{Z}}_k\|}{\sqrt{n}} \leq \frac{\|\mathbf{Z}_k\|}{\sqrt{n}} + \frac{\|\hat{\mathbf{Z}}_k - \mathbf{Z}_k\|}{\sqrt{n}}$ , where  $\mathbf{Z}_k \sim N(0, \sigma_{kk}I)$  and  $\frac{\|\hat{\mathbf{Z}}_k - \mathbf{Z}_k\|}{\sqrt{n}} = O\left(\sqrt{s_1 \lambda_1^2}\right) = o(1)$  according to Equation (42). Similarly  $\frac{\|\mathbf{E}_m\|}{\sqrt{n}} \leq \frac{\|\epsilon_m\|}{\sqrt{n}} + \frac{\|\Delta_{Z_m}\|}{\sqrt{n}}$  where  $\epsilon_m \sim N(0, \psi_{mm}I)$  and  $\frac{\|\Delta_{Z_m}\|}{\sqrt{n}} = o(1)$  from Equation (46). Noting that  $\psi_{mm}, \sigma_{kk} \in [M_2^{-1}, M_2]$ , we use the same argument as that for  $\mathbb{P}\{I_3^c\}$  in the proof of Lemma 1 to obtain

$$\begin{aligned}
\mathbb{P}\{I_2^c\} &\leq \mathbb{P}\left\{ \frac{\|\mathbf{Z}_k\|}{\sqrt{n}} \notin [1/\sqrt{2M_2}, \sqrt{2M_2}] \text{ for some } k \right\} \leq o(p^{-\delta_2}), \\
\mathbb{P}\{I_3^c\} &\leq \mathbb{P}\left\{ \frac{\|\epsilon_m\|}{\sqrt{n}} \notin [1/\sqrt{2M_2}, \sqrt{2M_2}] \right\} \leq o(p^{-\delta_2}).
\end{aligned}$$

(2). To study the lower-RE condition of  $\frac{\mathbf{W}^T \mathbf{W}}{n}$ , as what we did in the proof of Lemma 1 and Lemma 2, we essentially need to study  $\frac{\hat{\mathbf{Z}}_{A^c}^T \hat{\mathbf{Z}}_{A^c}}{n}$  and to show the following fact

$$\left| v^T \left( \frac{\hat{\mathbf{Z}}_{A^c}^T \hat{\mathbf{Z}}_{A^c}}{n} - \Sigma_{A^c, A^c} \right) v \right| \leq \frac{1}{LM_2} \text{ for all } v \in \mathbb{B}_0(2C_\beta s_2) \text{ with } \|v\| = 1,$$

where  $\lambda_{\min}(\Sigma_{A^c, A^c}) \geq 1/M_2$ . Following the same line of the proof in Lemma 1 for the lower-RE condition of  $\frac{\mathbf{X}^T \mathbf{X}}{n}$  with normality assumption on  $\mathbf{Z}$  and sparsity assumption  $s_2 = o\left(\sqrt{\frac{n}{\log p}}\right)$ , we can obtain that with probability  $1 - o(p^{-\delta_2})$ ,

$$\sup_{v \in \mathbb{B}_0(2C_\beta s_2)} \left| v^T \left( \frac{\mathbf{Z}_{A^c}^T \mathbf{Z}_{A^c}}{n} - \Sigma_{A^c, A^c} \right) v \right| \leq \frac{\|v\|^2}{2LM_2}. \quad (50)$$

Therefore all we need to show in the current setting is that with probability  $1 - o(p^{-\delta_2})$ ,

$$\left| v^T \left( \frac{\hat{\mathbf{Z}}_{A^c}^T \hat{\mathbf{Z}}_{A^c}}{n} - \frac{\mathbf{Z}_{A^c}^T \mathbf{Z}_{A^c}}{n} \right) v \right| \leq \frac{1}{2LM_2} \text{ for all } v \in \mathbb{B}_0(2C_\beta s_2) \text{ with } \|v\| = 1. \quad (51)$$

To show Equation (51), we notice

$$\left| v^T \left( \frac{\hat{\mathbf{Z}}_{A^c}^T \hat{\mathbf{Z}}_{A^c}}{n} - \frac{\mathbf{Z}_{A^c}^T \mathbf{Z}_{A^c}}{n} \right) v \right| \leq 2 \left| v^T \left( \frac{\mathbf{Z}_{A^c}^T (\hat{\mathbf{Z}}_{A^c} - \mathbf{Z}_{A^c})}{n} \right) v \right| + \frac{\|(\hat{\mathbf{Z}}_{A^c} - \mathbf{Z}_{A^c}) v\|^2}{n} \equiv D_1 + D_2.$$

To control  $D_2$ , we find

$$\sqrt{D_2} \leq \max_k \frac{\|\hat{\mathbf{Z}}_k - \mathbf{Z}_k\|}{\sqrt{n}} \|v\|_1 \leq \sqrt{s_1} \lambda_1 \sqrt{2C_\beta s_2} = o(1), \quad (52)$$

by Equation (42),  $\|v\|_1 \leq \sqrt{2C_\beta s_2} \|v\|$  and sparsity assumptions (22).

To control  $D_1$ , we find  $D_1 \leq \sqrt{D_2} \cdot \frac{\|\mathbf{Z}_{A^c} v\|}{\sqrt{n}}$ , which is  $o(1)$  with probability  $1 - o(p^{-\delta_2})$  by Equation (52) and the following result  $\frac{\|\mathbf{Z}_{A^c} v\|}{\sqrt{n}} = O(1)$ . Equation (50) implies that with probability  $1 - o(p^{-\delta_2})$ ,

$$\frac{\|\mathbf{Z}_{A^c} v\|}{\sqrt{n}} \leq \frac{\|v\|}{\sqrt{2LM_2}} + \left\| \Sigma_{A^c, A^c}^{1/2} v \right\| = O(1) \text{ for all } \|v\| = 1.$$

(3). Finally we study the probability of event  $I_1$ . In the current setting,

$$\begin{aligned} \nu &= \left\| \mathbf{W}^T \mathbf{E}_m / n \right\|_\infty = \left\| \mathbf{W}^T \boldsymbol{\epsilon}_m / n \right\|_\infty + O\left(\left\| \mathbf{W}^T \Delta_{Z_m} / n \right\|_\infty\right), \\ \theta^{ora} &= \frac{\|\mathbf{E}_m\|}{\sqrt{n}} = \frac{\|\boldsymbol{\epsilon}_m\|}{\sqrt{n}} + O\left(\frac{\|\Delta_{Z_m}\|}{\sqrt{n}}\right). \end{aligned}$$

Following the same line of the proof for event  $I_1$  in Lemma 1, we obtain that with probability  $1 - o(p^{-\delta_2+1})$ ,

$$\left\| \mathbf{W}^T \boldsymbol{\epsilon}_m / n \right\|_\infty < \frac{\|\boldsymbol{\epsilon}_m\|}{\sqrt{n}} \lambda_2 \frac{\xi - 1}{\xi + 1} (1 - \tau_2).$$

Thus to prove event  $I_1 = \left\{ \nu \leq \theta^{ora} \lambda_2 \frac{\xi - 1}{\xi + 1} (1 - \tau_2) \right\}$  holds with desired probability, we only need to show with probability  $1 - o(p^{-\delta_2+1})$ ,

$$\left\| \mathbf{W}^T \Delta_{Z_m} / n \right\|_\infty = o(\lambda_2) \text{ and } \frac{\|\Delta_{Z_m}\|}{\sqrt{n}} = o(1). \quad (53)$$

Equation (46) immediately implies  $\frac{\|\Delta_{Z_m}\|}{\sqrt{n}} = O\left(\sqrt{s_1\lambda_1^2}\right) = o(1)$  by the sparsity assumptions (22). To study  $\|\mathbf{W}^T \Delta_{Z_m}/n\|_\infty$ , we obtain that there exists some constant  $C_6 > 0$  such that,

$$\begin{aligned}
\|\mathbf{W}^T \Delta_{Z_m}/n\|_\infty &= \left\| \text{diag} \left( \frac{\sqrt{n}}{\|\hat{\mathbf{Z}}_k\|} \right) \hat{\mathbf{Z}}_{A^c}^T \Delta_{Z_m}/n \right\|_\infty \\
&\leq A'_1 \left\| \mathbf{Z}_{A^c}^T + \left( \hat{\mathbf{Z}}_{A^c}^T - \mathbf{Z}_{A^c}^T \right) \Delta_{Z_m}/n \right\|_\infty \\
&\leq A'_1 \left( \|\mathbf{Z}_{A^c}^T \Delta_{Z_m}/n\|_\infty + \max_i \left\| \left( \hat{\mathbf{Z}}_i - \mathbf{Z}_i \right) / \sqrt{n} \right\| \|\Delta_{Z_m}/\sqrt{n}\| \right) \\
&\leq A'_1 \|\mathbf{Z}_{A^c}^T \Delta_{Z_m}/n\|_\infty + C_6 s_1 \lambda_1^2,
\end{aligned} \tag{54}$$

where we used  $\frac{\|\hat{\mathbf{Z}}_k\|}{\sqrt{n}} \in [1/A'_1, A'_1]$  for all  $k$  on  $I_2$  in the first inequality and Equations (42) and (46) in the last inequality. By the sparsity assumptions (22), we have  $s_1 \lambda_1^2 = o(\lambda_2)$ . Thus it's sufficient to show  $\|\mathbf{Z}_{A^c}^T \Delta_{Z_m}/n\|_\infty = o(\lambda_2)$  with probability  $1 - o(p^{-\delta_2+1})$ . In fact,

$$\begin{aligned}
\|\mathbf{Z}_{A^c}^T \Delta_{Z_m}/n\|_\infty &\leq \|\mathbf{Z}_{A^c}^T \mathbf{X} [(\gamma_m - \hat{\gamma}_m) + (\hat{\gamma}_{A^c} - \gamma_{A^c}) \beta_m] / n\|_\infty \\
&\leq \|\mathbf{Z}_{A^c}^T \mathbf{X}/n\|_\infty \|[(\gamma_m - \hat{\gamma}_m) + (\hat{\gamma}_{A^c} - \gamma_{A^c}) \beta_m]\|_1 \\
&\leq \|\mathbf{Z}_{A^c}^T \mathbf{X}/n\|_\infty \max_i \|\hat{\gamma}_i - \gamma_i\|_1 (1 + \|\beta_m\|_1) \\
&\leq \|\mathbf{Z}_{A^c}^T \mathbf{X}/n\|_\infty C s_1 \lambda_1 = o(\|\mathbf{Z}_{A^c}^T \mathbf{X}/n\|_\infty),
\end{aligned} \tag{55}$$

where the last inequality follows from Equations (43) and (45).

Since each  $\mathbf{Z}_k^T \stackrel{i.i.d.}{\sim} N(0, \sigma_{kk})$  and is independent of  $\mathbf{X}$ , it can be seen that each entry of  $\mathbf{Z}_{A^c}^T \mathbf{X}$  is a sum of  $n$  i.i.d. sub-exponential variables with finite parameter under either Condition 3 or 3'. A union bound with  $pq$  entries and an application of Bernstein inequality in Equation (36) with  $t = C_7 \sqrt{\frac{\log(pq)}{n}}$  imply that  $\|\mathbf{Z}_{A^c}^T \mathbf{X}/n\|_\infty \leq C_7 \sqrt{\frac{\log(pq)}{n}}$  with probability  $1 - o((pq)^{-\delta_1})$  for some large constant  $C_7 > 0$ . This result, together with Equation (55), implies that  $\|\mathbf{Z}_{A^c}^T \Delta_{Z_m}/n\|_\infty = o(\lambda_2)$  with probability  $1 - o(p^{-\delta_2+1})$ . Now we finish the proof of  $\|\mathbf{W}^T \Delta_{Z_m}/n\|_\infty = o(\lambda_2)$  by Equation (54) and hence the proof of the Equation (53) with probability  $1 - o(p^{-\delta_2+1})$ . This completes our proof of Lemma 4.

## 8. A KEY LEMMA

The lemma of scaled lasso introduced in this section is deterministic in nature. It could be applied in different settings in which the assumptions of design matrix, response variables and noise distribution may vary, as long as the conditions of the lemma are satisfied. Therefore it's a very useful building block in the analysis of many different problems. In particular, the main theorems of both steps are based on this key lemma. The proof of this lemma is similar as that in (Sun & Zhang 2012) and (Ren et al. 2013), but we use the restrict eigenvalue condition for the gram matrix instead of CIF condition to easily adapt to different settings of design matrix for our probabilistic analysis.

Consider the following general scaled  $l_1$  penalized regression problem. Denote the  $n$  by  $p_0$  dimensional design matrix by  $\mathbf{D} = (\mathbf{D}_1, \dots, \mathbf{D}_{p_0})$ , the  $n$  dimensional response variable  $\mathbf{R} = (R_1, \dots, R_n)^T$  and the noise variable  $\mathbf{E} = (E_1, \dots, E_n)^T$ . The scaled lasso estimator with tuning parameter  $\lambda$  of the regression

$$\mathbf{R} = \mathbf{D}b^{true} + \mathbf{E}, \quad (56)$$

is defined as

$$\{\hat{b}, \hat{\theta}\} = \arg \min_{b \in \mathbb{R}^{p_0}, \theta \in \mathbb{R}^+} \left\{ \frac{\|\mathbf{R} - \mathbf{D}b\|^2}{2n\theta} + \frac{\theta}{2} + \lambda \sum_{k=1}^{p_0} \frac{\|\mathbf{D}_k\|}{\sqrt{n}} |b_k| \right\}, \quad (57)$$

where the sparsity  $s$  of the true coefficient  $b^{true}$  is defined as follows,

$$s = \sum_{j=1}^{p_0} \min \{1, |b_j^{true}| / \lambda\}, \quad (58)$$

which is a generalization of exact sparseness (the number of nonzero entries).

We first normalize each column of the design matrix  $\mathbf{D}$  to make the analysis cleaner by setting

$$d_k = \frac{\|\mathbf{D}_k\|}{\sqrt{n}} b_k, \text{ and } \mathbf{W} = \mathbf{D} \cdot \text{diag} \left( \frac{\sqrt{n}}{\|\mathbf{D}_k\|} \right), \quad (59)$$

and then rewrite the model (56) and the penalized procedure (57) as follows,

$$\mathbf{R} = \mathbf{W}d^{true} + \mathbf{E}_j, \quad (60)$$

and

$$L_\lambda(d, \theta) = \frac{\|\mathbf{R} - \mathbf{W}d\|^2}{2n\theta} + \frac{\theta}{2} + \lambda \|d\|_1, \quad (61)$$

$$\{\hat{d}, \hat{\theta}\} = \arg \min_{d \in \mathbb{R}^{p_0}, \theta \in \mathbb{R}^+} L_\lambda(d, \theta), \quad (62)$$

where the true coefficients and the estimator of the standardized scaled lasso regression (61) are  $d_k^{true} = \frac{\|\mathbf{D}_k\|}{\sqrt{n}} b_k^{true}$  and  $\hat{d}_k = \hat{b}_k \frac{\|\mathbf{D}_k\|}{\sqrt{n}}$  respectively.

For this standardized scaled lasso regression, we introduce some important notation, including the lower-RE condition on the gram matrix  $\frac{\mathbf{W}^T \mathbf{W}}{n}$ . The oracle estimator  $\theta^{ora}$  of the noise level can be defined as

$$\theta^{ora} = \frac{\|\mathbf{R} - \mathbf{W}d^{true}\|}{\sqrt{n}} = \frac{\|\mathbf{E}\|}{\sqrt{n}}. \quad (63)$$

Let  $|K|$  be the cardinality of an index set  $K$ . Define  $T$  as the index set of those large coefficients of  $d^{true}$ ,

$$T = \{k : |d_k^{true}| \geq \lambda\}. \quad (64)$$

We say the gram matrix  $\frac{\mathbf{W}^T \mathbf{W}}{n}$  satisfies a **lower-RE condition** with curvature  $\alpha_1 > 0$  and tolerance  $\zeta(n, p_0) > 0$  if

$$\mu^T \frac{\mathbf{W}^T \mathbf{W}}{n} \mu \geq \alpha_1 \|\mu\|^2 - \zeta(n, p_0) \|\mu\|_1^2 \text{ for all } \mu \in \mathbb{R}. \quad (65)$$

Moreover, we define

$$\nu = \|\mathbf{W}^T (\mathbf{R} - \mathbf{W}d^{true}) / n\|_\infty = \|\mathbf{W}^T \mathbf{E} / n\|_\infty, \quad (66)$$

$$\tau = s\lambda^2 \cdot A_2 \frac{2\xi}{(1+\xi)} C_1 \left( \frac{3\theta^{ora}}{2}, 4\xi A_1, \alpha_1 \right), \quad (67)$$

with constants  $\xi > 1$ ,  $A_1$  and  $A_2$  introduced in Lemma 5.  $C_1 \left( \frac{3\theta^{ora}}{2}, 4\xi A_1, \alpha_1 \right)$  is a constant depending on  $\frac{3\theta^{ora}}{2}$ ,  $\xi A_1$  and  $\alpha_1$  with its definition in Equation (76). It is bounded above if  $\frac{3\theta^{ora}}{2}$  and  $4\xi A_1$  are bounded above and  $\alpha_1$  is bounded below by some universal constants, respectively.

With the notation we can state the key lemma as follows.

**Lemma 5** *Consider the scaled  $l_1$  penalized regression procedure (57). Whenever there exist constants  $\xi > 1$ ,  $A_1$  and  $A_2$  such that the following conditions are satisfied*

1.  $\nu \leq \theta^{ora} \lambda \frac{\xi-1}{\xi+1} (1-\tau)$  for some  $\xi > 1$  and  $\tau \leq 1/2$  defined in (67);
2.  $\frac{\|\mathbf{D}_k\|}{\sqrt{n}} \in [1/A_1, A_1]$  for all  $k$ ;
3.  $\theta^{ora} \in [1/A_2, A_2]$ ;
4.  $\frac{\mathbf{W}^T \mathbf{W}}{n}$  satisfies the lower-RE condition with  $\alpha_1$  and  $\zeta(n, p_0)$  such that

$$s\zeta(n, p_0) 8(1+\xi)^2 A_1^2 \leq \min\{\frac{\alpha_1}{2}, 1\}, \quad (68)$$

we have the following deterministic bounds

$$|\hat{\theta} - \theta^{ora}| \leq C_1 \lambda^2 s, \quad (69)$$

$$\|\hat{b} - b^{true}\|^2 \leq C_2 \lambda^2 s, \quad (70)$$

$$\|\hat{b} - b^{true}\|_1 \leq C_3 \lambda s, \quad (71)$$

$$\|\mathbf{D}(\hat{b} - b^{true})\|^2 / n \leq C_4 \lambda^2 s, \quad (72)$$

$$\|\mathbf{D}^T \mathbf{E} / n\|_\infty \leq C_5 \lambda, \quad (73)$$

where constants  $C_i$  ( $i = 1, \dots, 5$ ) only depend on  $A_1$ ,  $A_2$ ,  $\alpha_1$  and  $\xi$ .

## 9. APPENDIX

### 9.1 Proof of Lemma 5

The function  $L_\lambda(d, \theta)$  in Equation (61) is jointly convex in  $(d, \theta)$ . For fixed  $\theta > 0$ , denote the minimizer of  $L_\lambda(d, \theta)$  over all  $d \in \mathbb{R}^{p_0}$  by  $\hat{d}(\theta\lambda)$ , a function of  $\theta\lambda$ , i.e.,

$$\hat{d}(\theta\lambda) = \arg \min_{d \in \mathbb{R}^{p_0}} L_\lambda(d, \theta) = \arg \min_{d \in \mathbb{R}^{p_0}} \left\{ \frac{\|\mathbf{R} - \mathbf{W}d\|^2}{2n} + \lambda\theta \|d\|_1 \right\}, \quad (74)$$

then if we knew  $\hat{\theta}$  in the solution of Equation (62), the solution for the equation is  $\{\hat{d}(\hat{\theta}\lambda), \hat{\theta}\}$ . We recognize that  $\hat{d}(\hat{\theta}\lambda)$  is just the standard lasso with the penalty  $\hat{\theta}\lambda$ , however we don't know the estimator  $\hat{\theta}$ . The strategy of our analysis is that we first show that  $\hat{\theta}$  is very close to its oracle estimator  $\theta^{ora}$ , then the standard lasso analysis would imply the desired result Equations (70)-(73) under the assumption that  $\hat{\theta}/\theta^{ora} = 1 + O(\lambda^2 s)$ . For the standard lasso analysis, some kind of

regularity condition is assumed on the design matrix  $\mathbf{W}^T \mathbf{W}/n$  in the regression literature. In this paper we use the lower-RE condition, which is one of the most general conditions.

Let  $\mu = \lambda\theta$ . From the Karush-Kuhn-Tucker condition,  $\hat{d}(\mu)$  is the solution to the Equation (74) if and only if

$$\begin{aligned} \mathbf{W}_k^T (\mathbf{R} - \mathbf{W} \hat{d}(\mu)) / n &= \mu \cdot \text{sgn}(\hat{d}_k(\mu)), \text{ if } \hat{d}_k(\mu) \neq 0, \\ \mathbf{W}_k^T (\mathbf{R} - \mathbf{W} \hat{d}(\mu)) / n &\in [-\mu, \mu], \text{ if } \hat{d}_k(\mu) = 0. \end{aligned} \quad (75)$$

Let  $C_2(a_1, a_2)$  and  $C_1(a_{11}, a_{12}, a_2)$  be constants depending on  $a_1, a_2$  and  $a_{11}, a_{12}, a_2$ , respectively. The constant  $C_2$  is bounded above if  $a_1$  is bounded above and  $a_2$  is bounded below by constants, respectively. The constant  $C_1$  is bounded above whenever  $a_{11}$  and  $a_{12}$  are bounded above and  $a_2$  is bounded below by constants. The explicit formulas of  $C_1$  and  $C_2$  are given as follows,

$$\begin{aligned} C_2(a_1, a_2) &= \frac{a_1}{a_2} + \left[ \left( \frac{a_1}{a_2} \right)^2 + \frac{2(a_1 + 1)}{a_2} \right]^{1/2}, \\ C_1(a_{11}, a_{12}, a_2) &= a_{12} (1 + C_2(a_{11} \times a_{12}, a_2)). \end{aligned} \quad (76)$$

The following propositions are helpful to establish our result. The proof is given in Sections 9.2 and 9.3.

**Proposition 2** *The sparsity  $s$  is defined in Equation (58). For any  $\xi > 1$ , assuming  $\nu \leq \mu \frac{\xi-1}{\xi+1}$  and conditions 2, 4 in Lemma 5 hold, we have*

$$\left\| \hat{d}(\mu) - d^{true} \right\|_1 \leq C_1 \left( \frac{\mu}{\lambda}, 4\xi A_1, \alpha_1 \right) \lambda s, \quad (77)$$

$$\left\| \hat{d}(\mu) - d^{true} \right\| \leq C_2 \left( 4\xi A_1 \frac{\mu}{\lambda}, \alpha_1 \right) \lambda \sqrt{s}, \quad (78)$$

$$\frac{1}{n} \left\| \mathbf{W} (d^{true} - \hat{d}(\mu)) \right\|^2 \leq (\nu + \mu) \left\| \hat{d}(\mu) - d^{true} \right\|_1. \quad (79)$$

**Proposition 3** *Let  $\{\hat{d}, \hat{\theta}\}$  be the solution of the scaled lasso (62). For any  $\xi > 1$ , assuming conditions 1 – 4 in Lemma 5 hold, then we have*

$$\left| \frac{\hat{\theta}}{\theta^{ora}} - 1 \right| \leq \tau. \quad (80)$$

Now we finish our proof with these two propositions. According to Conditions 1 – 4 in Lemma 5, Proposition 3 implies  $\nu \leq \mu \frac{\xi-1}{\xi+1}$  with  $\mu = \lambda \hat{\theta}$  and Proposition 2 further implies that there exist some constants  $c_1, c_2$  and  $c_3$  such that

$$\begin{aligned} \left\| \hat{d}(\lambda \hat{\theta}) - d^{true} \right\|_1 &\leq c_1 \lambda s, \\ \left\| \hat{d}(\lambda \hat{\theta}) - d^{true} \right\| &\leq c_2 \lambda \sqrt{s}, \\ \frac{\left\| \mathbf{W} \left( d^{true} - \hat{d}(\lambda \hat{\theta}) \right) \right\|^2}{n} &\leq c_3 \lambda^2 s. \end{aligned}$$

Note that  $\frac{\mu}{\lambda} = \hat{\theta} \in [\theta^{ora}(1-\tau), \theta^{ora}(1+\tau)] \subset \left[ \frac{1}{2A_2}, \frac{3A_2}{2} \right]$ . Thus the constants  $c_1, c_2$  and  $c_3$  only depends on  $A_1, A_2, \alpha_1$  and  $\xi$ . Now we transfer the results above on standardized scaled lasso (62) back to the general scaled lasso (57) through the bounded scaling constants  $\left\{ \frac{\sqrt{n}}{\|\mathbf{D}_k\|} \right\}$  and immediately have the desired results (70)-(72). Result (69) is an immediate consequence of Proposition 3 and Result (73) is an immediate consequence of assumptions 1 – 3.

## 9.2 Proof of Proposition 2

Notice that

$$\begin{aligned} &\frac{1}{n} \left\| \mathbf{W} \left( d^{true} - \hat{d}(\mu) \right) \right\|^2 \\ &= \frac{\left( d^{true} - \hat{d}(\mu) \right)^T \left( \mathbf{W}^T (\mathbf{R} - \mathbf{W} \hat{d}(\mu)) - \mathbf{W}^T (\mathbf{R} - \mathbf{W} d^{true}) \right)}{n} \\ &\leq \mu \left( \|d^{true}\|_1 - \|\hat{d}(\mu)\|_1 \right) + \nu \|d^{true} - \hat{d}(\mu)\|_1 \end{aligned} \quad (81)$$

$$\leq (\mu + \nu) \left\| \left( d^{true} - \hat{d}(\mu) \right)_T \right\|_1 + 2\mu \left\| (d^{true})_{T^c} \right\|_1 - (\mu - \nu) \left\| \left( d^{true} - \hat{d}(\mu) \right)_{T^c} \right\|_1, \quad (82)$$

where the first inequality follows from the KKT conditions (75).

Now define  $\Delta = \hat{d}(\mu) - d^{true}$ . Equation (81) also implies the desired inequality (79)

$$\Delta^T \frac{\mathbf{W}^T \mathbf{W}}{n} \Delta \leq (\mu + \nu) \|\Delta\|_1 \quad (83)$$

We will first show that

$$\|\Delta_{T^c}\|_1 \leq \max \left\{ 2(1 + \xi) \left\| (d^{true})_{T^c} \right\|_1, 2\xi \|\Delta_T\|_1 \right\}, \quad (84)$$

then we are able to apply the lower-RE condition (65) to derive the desired results.

To show Equation (84), we note that our assumption  $\nu \leq \mu \frac{\xi-1}{\xi+1}$  with Equation (82) implies

$$\frac{1}{n} \|\mathbf{W}\Delta\|^2 \leq 2\mu \left( \frac{\xi \|\Delta_T\|_1}{\xi+1} + \|(d^{true})_{T^c}\|_1 - \frac{\|\Delta_{T^c}\|_1}{\xi+1} \right). \quad (85)$$

Suppose that

$$\|\Delta_{T^c}\|_1 \geq 2(1+\xi) \|(d^{true})_{T^c}\|_1, \quad (86)$$

then the inequality (85) becomes

$$0 \leq \frac{1}{n} \|\mathbf{W}\Delta\|^2 \leq \frac{\mu}{\xi+1} (2\xi \|\Delta_T\|_1 - \|\Delta_{T^c}\|_1),$$

which implies

$$\|\Delta_{T^c}\|_1 \leq 2\xi \|\Delta_T\|_1. \quad (87)$$

Therefore the complement of inequality (86) and Equation (87) together finish our proof of Equation (84).

Before proceeding, we point out two facts which will be used below several times. Note the sparseness  $s$  is defined in terms of the true coefficients  $b^{true}$  in Equation (58) before standardization but the index set  $T$  is defined in term of  $d^{true}$  in Equation (64) after standardization. Condition 2 implies that this standardization step doesn't change the sparseness up to a factor  $A_1$ . Hence it's not hard to see that  $|T| \leq A_1 s$  and  $\|(d^{true})_{T^c}\|_1 \leq A_1 \lambda s$ .

Now we are able to apply the lower-RE condition of  $\frac{\mathbf{W}^T \mathbf{W}}{n}$  to Equation (83) and obtain that

$$\begin{aligned} \Delta^T \frac{\mathbf{W}^T \mathbf{W}}{n} \Delta &\geq \alpha_1 \|\Delta\|^2 - \zeta(n, p_0) \|\Delta\|_1^2 \\ &\geq \alpha_1 \|\Delta\|^2 - \zeta(n, p_0) 8(1+\xi)^2 \left( \|(d^{true})_{T^c}\|_1^2 + \|\Delta_T\|_1^2 \right) \\ &\geq (\alpha_1 - |T| \zeta(n, p_0) 8(1+\xi)^2) \|\Delta\|^2 - \zeta(n, p_0) 8(1+\xi)^2 A_1^2 \lambda^2 s^2 \\ &\geq \frac{\alpha_1}{2} \|\Delta\|^2 - \lambda^2 s, \end{aligned}$$

where in the second, third and last inequalities we applied the facts (84),  $\|\Delta_T\|_1^2 \leq |T| \|\Delta_T\|^2$ ,  $|T| \leq A_1 s$ ,  $\|(d^{true})_{T^c}\|_1 \leq A_1 \lambda s$  and  $s \zeta(n, p_0) 8(1+\xi)^2 A_1^2 \leq \min\{\frac{\alpha_1}{2}, 1\}$ . Moreover, by apply-

ing those facts used in last equation again, we have

$$\begin{aligned}
(\mu + \nu) \|\Delta\|_1 &\leq 4\xi\mu \left( \|(d^{true})_{T^c}\|_1 + \|\Delta_T\|_1 \right) \\
&\leq 4\xi\mu \left( A_1\lambda s + \sqrt{|T|} \|\Delta\| \right) \\
&\leq 4\xi A_1 \frac{\mu}{\lambda} (\lambda^2 s + \sqrt{s}\lambda \|\Delta\|).
\end{aligned}$$

The above two inequalities together with Equation (83) imply that

$$4\xi A_1 \frac{\mu}{\lambda} (\lambda^2 s + \sqrt{s}\lambda \|\Delta\|) \geq \frac{\alpha_1}{2} \|\Delta\|^2 - \lambda^2 s.$$

Define  $S_u = 4\xi A_1 \frac{\mu}{\lambda}$ . Some algebra about this quadratic inequality implies the bound of  $\Delta$  under  $l_2$  norm

$$\begin{aligned}
\|\Delta\| &\leq \left( \frac{S_u}{\alpha_1} + \left[ \left( \frac{S_u}{\alpha_1} \right)^2 + \frac{2(S_u + 1)}{\alpha_1} \right]^{1/2} \right) \lambda \sqrt{s} \\
&\equiv C_2 \left( 4\xi A_1 \frac{\mu}{\lambda}, \alpha_1 \right) \lambda \sqrt{s}.
\end{aligned}$$

Combining this fact with Equation (84), we finally obtain the bound under  $l_1$  norm (77)

$$\begin{aligned}
\|\Delta\|_1 &\leq \|\Delta_T\|_1 + \|\Delta_{T^c}\|_1 \leq 2(1 + \xi) \left( \|(d^{true})_{T^c}\|_1 + \|\Delta_T\|_1 \right) \\
&\leq 2(1 + \xi) \left( A_1 s \lambda + \sqrt{|T|} \|\Delta\| \right) \\
&\leq 4\xi A_1 \left( 1 + C_2 \left( 4\xi A_1 \frac{\mu}{\lambda}, \alpha_1 \right) \right) s \lambda \\
&\equiv C_1 \left( \frac{\mu}{\lambda}, 4\xi A_1, \alpha_1 \right) s \lambda.
\end{aligned}$$

### 9.3 Proof of Proposition 3

For  $\tau$  defined in Equation (67), we need to show that  $\hat{\theta} \geq \theta^{ora} (1 - \tau)$  and  $\hat{\theta} \leq \theta^{ora} (1 + \tau)$  on the event  $\left\{ \nu \leq \theta^{ora} \lambda \frac{\xi-1}{\xi+1} (1 - \tau) \right\}$ . Let  $\hat{d}(\theta\lambda)$  be the solution of (74) as a function of  $\theta$ , then

$$\frac{\partial}{\partial \theta} L_\lambda \left( \hat{d}(\theta\lambda), \theta \right) = \frac{1}{2} - \frac{\left\| \mathbf{R} - \mathbf{W} \hat{d}(\theta\lambda) \right\|^2}{2n\theta^2}, \quad (88)$$

since  $\left\{ \frac{\partial}{\partial d} L_\lambda(d, \theta) \big|_{d=\hat{d}(\theta\lambda)} \right\}_k = 0$  for all  $\hat{d}_k(\theta\lambda) \neq 0$ , and  $\left\{ \frac{\partial}{\partial \theta} \hat{d}(\theta\lambda) \right\}_k = 0$  for all  $\hat{d}_k(\theta\lambda) = 0$  which follows from the fact that  $\left\{ k : \hat{d}_k(\theta\lambda) = 0 \right\}$  is unchanged in a neighborhood of  $\theta$  for almost all  $\theta$ . Equation (88) plays a key in the proof.

(1). To show that  $\hat{\theta} \geq \theta^{ora} (1 - \tau)$  it's enough to show

$$\frac{\partial}{\partial \theta} L_{\lambda} \left( \hat{d}(\theta \lambda), \theta \right) |_{\theta=t_1} < 0,$$

where  $t_1 = \theta^{ora} (1 - \tau)$ , due to the strict convexity of the objective function  $L_{\lambda}(d, \theta)$  in  $\theta$ . Equation (88) implies that

$$\begin{aligned} 2t_1^2 \frac{\partial}{\partial \theta} L_{\lambda} \left( \hat{d}(\theta \lambda), \theta \right) |_{\theta=t_1} &= t_1^2 - \frac{\left\| \mathbf{R} - \mathbf{W} \hat{d}(t_1 \lambda) \right\|^2}{n} \\ &\leq t_1^2 - \frac{\left\| \mathbf{R} - \mathbf{W} d^{true} + \mathbf{W} \left( \hat{d}(t_1 \lambda) - d^{true} \right) \right\|^2}{n} \\ &\leq t_1^2 - (\theta^{ora})^2 + 2 \left( d^{true} - \hat{d}(t_1 \lambda) \right)^T \frac{\mathbf{W}^T (\mathbf{R} - \mathbf{W} d^{true})}{n} \\ &\leq 2t_1 (t_1 - \theta^{ora}) + 2\nu \left\| d^{true} - \hat{d}(t_1 \lambda) \right\|_1. \end{aligned} \quad (89)$$

On the event  $\left\{ \nu \leq t_1 \lambda \frac{\xi-1}{\xi+1} \right\} = \left\{ \nu / \theta^{ora} < \lambda \frac{\xi-1}{\xi+1} (1 - \tau) \right\}$  we have

$$\begin{aligned} &2t_1^2 \frac{\partial}{\partial \theta} L_{\lambda} \left( \hat{d}(\theta \lambda), \theta \right) |_{\theta=t_1} \\ &\leq 2t_1 (t_1 - \theta^{ora}) + 2t_1 \lambda \frac{\xi-1}{\xi+1} \left\| d^{true} - \hat{d}(t_1 \lambda) \right\|_1 \\ &\leq 2t_1 \left[ -\tau \theta^{ora} + \lambda \frac{\xi-1}{\xi+1} \left\| d^{true} - \hat{d}(t_1 \lambda) \right\|_1 \right] < 0. \end{aligned}$$

The last inequality follows from the definition of  $\tau$  and the  $l_1$  error bound in Equation (77) of Proposition 2. Note that for  $\lambda^2 s$  sufficiently small, we have small  $\tau < 1/2$ . In fact, although  $\left\| d^{true} - \hat{d}(t_1 \lambda) \right\|_1$  also depends on  $\tau$ , our choice of  $\tau$  is well-defined and is larger than  $\frac{\lambda}{\theta^{ora}} \frac{3(\xi-1)}{2(\xi+1)} \left\| d^{true} - \hat{d}(t_1 \lambda) \right\|_1$ .

(2). Let  $t_2 = \theta^{ora} (1 + \tau)$ . To show the other side  $\hat{\theta} \leq \theta^{ora} (1 + \tau)$  it is enough to show

$$\frac{\partial}{\partial \theta} L_{\lambda} \left( \hat{d}(\theta \lambda), \theta \right) |_{\theta=t_2} > 0.$$

Equation (88) implies that on the event  $\left\{ \nu \leq t_2 \lambda^{\frac{\xi-1}{\xi+1}} \right\} = \left\{ \nu / \theta^{ora} < \lambda^{\frac{\xi-1}{\xi+1}} (1 + \tau) \right\}$  we have

$$\begin{aligned}
2t_2^2 \frac{\partial}{\partial \theta} L_\lambda \left( \hat{d}(\theta \lambda), \theta \right) \big|_{\theta=t_2} &= t_2^2 - \frac{\left\| \mathbf{R} - \mathbf{W} \hat{d}(t_2 \lambda) \right\|^2}{n} \\
&= t_2^2 - (\theta^{ora})^2 + (\theta^{ora})^2 - \frac{\left\| \mathbf{R} - \mathbf{W} \hat{d}(t_2 \lambda) \right\|^2}{n} \\
&= t_2^2 - (\theta^{ora})^2 + \frac{\left\| \mathbf{R} - \mathbf{W} d^{true} \right\|^2 - \left\| \mathbf{R} - \mathbf{W} \hat{d}(t_2 \lambda) \right\|^2}{n} \\
&= t_2^2 - (\theta^{ora})^2 + \frac{\left( \hat{d}(t_2 \lambda) - d^{true} \right)^T \mathbf{W}^T \left( \mathbf{R} - \mathbf{W} d^{true} + \mathbf{R} - \mathbf{W} \hat{d}(t_2 \lambda) \right)}{n} \\
&\geq t_2^2 - (\theta^{ora})^2 - \left\| \hat{d}(t_2 \lambda) - d^{true} \right\|_1 (\nu + t_2 \lambda) \\
&\geq (t_2 + \theta^{ora}) \theta^{ora} \tau - \frac{2\xi}{\xi+1} t_2 \lambda \left\| \hat{d}(t_2 \lambda) - d^{true} \right\|_1 \\
&\geq 2\theta^{ora} \left( \tau \theta^{ora} - \frac{3\xi}{2(\xi+1)} \lambda \left\| \hat{d}(t_2 \lambda) - d^{true} \right\|_1 \right) > 0,
\end{aligned}$$

where the second last inequality is due to the fact  $\tau \leq 1/2$  and the last inequality follows from the definition of  $\tau$  and the  $l_1$  error bound in Equation (77) of Proposition 2. Still, our choice of  $\tau$  is well-defined and is larger than  $\frac{\lambda}{\theta^{ora}} \frac{4\xi}{2(\xi+1)} \left\| d^{true} - \hat{d}(t_2 \lambda) \right\|_1$ .

## ACKNOWLEDGEMENTS

We thank Yale University Biomedical High Performance Computing Center for computing resources, and NIH grant RR19895 and RR029676-01, which funded the instrumentation. The research of Zhao Ren and Harrison Zhou was supported in part by NSF Career Award DMS-0645676, NSF FRG Grant DMS-0854975 and NSF Grant DMS-1209191.

## REFERENCES

- Bickel, P. J., & Levina, E. (2008a), “Regularized estimation of large covariance matrices.,” *Ann. Statist.*, 36(1), 199–227.
- Bickel, P. J., & Levina, E. (2008b), “Covariance regularization by thresholding.,” *Ann. Statist.*, 36(6), 2577–2604.

- Cai, T. T., Li, H., Liu, W., & Xie, J. (2013), “Covariate-adjusted precision matrix estimation with an application in genetical genomics,” *Biometrika*, 100(1), 139–156.
- Cai, T. T., Liu, W., & Zhou, H. H. (2012), “Estimating Sparse Precision Matrix: Optimal Rates of Convergence and Adaptive Estimation.,” *Manuscript*, .
- Cai, T. T., Zhang, C. H., & Zhou, H. H. (2010), “Optimal rates of convergence for covariance matrix estimation.,” *Ann. Statist.*, 38(4), 2118–2144.
- Cai, T. T., & Zhou, H. H. (2012), “Optimal rates of convergence for sparse covariance matrix estimation,” *Ann. Statist.*, 40(5), 2389–2420.
- Cheng, J., Levina, E., Wang, P., & Zhu, J. (2012), “Sparse ising models with covariates,” *arXiv preprint arXiv:1209.6342*, .
- El Karoui, N. (2008), “Operator norm consistent estimation of large dimensional sparse covariance matrices.,” *Ann. Statist.*, 36(6), 2717–2756.
- Friedman, J., Hastie, T., & Tibshirani, R. (2008), “Sparse inverse covariance estimation with the graphical lasso,” *Biostatistics*, 9(3), 432–441.
- Kanehisa, M., & Goto, S. (2000), “KEGG: kyoto encyclopedia of genes and genomes,” *Nucleic acids research*, 28(1), 27–30.
- Lam, C., & Fan, J. (2009), “Sparsistency and rates of convergence in large covariance matrix estimation,” *Ann. Statist.*, 37(6B), 4254–4278.
- Lauritzen, S. L. (1996), *Graphical Models* Oxford University Press.
- Li, B., Chun, H., & Zhao, H. (2012), “Sparse estimation of conditional graphical models with application to gene networks,” *Journal of the American Statistical Association*, 107(497), 152–167.
- Liang, F., & Wang, Y. (2007), “DNA damage checkpoints inhibit mitotic exit by two different mechanisms,” *Molecular and cellular biology*, 27(14), 5067–5078.

- Loh, P., & Wainwright, M. (2012), “High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity,” *Ann. Statist.*, 40(3), 1637–1664.
- Massart, P. (2007), *Concentration inequalities and model selection* Springer Verlag.
- Obozinski, G., Wainwright, M. J., & Jordan, M. I. (2011), “Support union recovery in high-dimensional multivariate regression,” *The Annals of Statistics*, 39(1), 1–47.
- Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J. R., & Wang, P. (2010), “Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer,” *The Annals of Applied Statistics*, 4(1), 53–77.
- Reimand, J., Aun, A., Vilo, J., Vaquerizas, J. M., Sedman, J., & Luscombe, N. M. (2012), “m: Explorer: multinomial regression models reveal positive and negative regulators of longevity in yeast quiescence,” *Genome biology*, 13(6), R55.
- Ren, Z., Sun, T., Zhang, C. H., & Zhou, H. H. (2013), “Asymptotic Normality and Optimalities in Estimation of Large Gaussian Graphical Model,” *Manuscript*, .
- Rockman, M. V., & Kruglyak, L. (2006), “Genetics of global gene expression,” *Nature Reviews Genetics*, 7(11), 862–872.
- Rudelson, M., & Zhou, S. (2013), “Reconstruction from anisotropic random measurements,” *IEEE Trans Inf Theory*, 59(6), 3434–3447.
- Smith, E. N., & Kruglyak, L. (2008), “Gene–environment interaction in yeast gene expression,” *PLoS biology*, 6(4), e83.
- Stegmeier, F., Visintin, R., Amon, A. et al. (2002), “Separase, polo kinase, the kinetochore protein Slk19, and Spo12 function in a network that controls Cdc14 localization during early anaphase,” *Cell*, 108(2), 207.
- Sun, T., & Zhang, C. H. (2012), “Scaled Sparse Linear Regression,” *Manuscript*, .

- Vershynin, R. (2010), “Introduction to the non-asymptotic analysis of random matrices,” *arXiv preprint arXiv:1011.3027*, .
- Wurzenberger, C., & Gerlich, D. W. (2011), “Phosphatases: providing safe passage through mitotic exit,” *Nature Reviews Molecular Cell Biology*, 12(8), 469–482.
- Yin, J., & Li, H. (2011), “A sparse conditional Gaussian graphical model for analysis of genetical genomics data,” *The annals of applied statistics*, 5(4), 2630.
- Yin, J., & Li, H. (2013), “Adjusting for high-dimensional covariates in sparse precision matrix estimation by  $l_1$ -penalization,” *Journal of multivariate analysis*, 116, 365–381.