SCIENCE CHINA
Mathematics

• ARTICLES •

June 2020 Vol. 63 No. 6: 1203–1218
https://doi.org/10.1007/s11425-017-9438-5

A two-step method for estimating high-dimensional Gaussian graphical models

Yuehan Yang[1,*] & Ji Zhu[2]

[1]*School of Statistics and Mathematics, Central University of Finance and Economics, Beijing* 100081*, China;*
[2]*Department of Statistics, University of Michigan, Ann Arbor, MI* 48109*, USA*

Email: yyh@cufe.edu.cn, jizhu@umich.edu

Received December 20, 2017; accepted July 29, 2018; published online May 14, 2020

Abstract The problem of estimating high-dimensional Gaussian graphical models has gained much attention in recent years. Most existing methods can be considered as one-step approaches, being either regression-based or likelihood-based. In this paper, we propose a two-step method for estimating the high-dimensional Gaussian graphical model. Specifically, the first step serves as a screening step, in which many entries of the concentration matrix are identified as zeros and thus removed from further consideration. Then in the second step, we focus on the remaining entries of the concentration matrix and perform selection and estimation for nonzero entries of the concentration matrix. Since the dimension of the parameter space is effectively reduced by the screening step, the estimation accuracy of the estimated concentration matrix can be potentially improved. We show that the proposed method enjoys desirable asymptotic properties. Numerical comparisons of the proposed method with several existing methods indicate that the proposed method works well. We also apply the proposed method to a breast cancer microarray data set and obtain some biologically meaningful results.

Keywords covariance estimation, graphical model, penalized likelihood, sparse regression, two-step method

MSC(2010) 62-09, 62P10

Citation: Yang Y H, Zhu J. A two-step method for estimating high-dimensional Gaussian graphical models. Sci China Math, 2020, 63: 1203–1218, https://doi.org/10.1007/s11425-017-9438-5

1 Introduction

The problem of estimating high-dimensional Gaussian graphical models has gained much attention in recent years. The basic model assumes that the observations are independent and identically distributed samples from a multivariate Gaussian distribution, and the goal is to infer the conditional dependence structure among the variables, which can be implied from the inverse of the covariance matrix, also known as the concentration matrix, of the multivariate Gaussian distribution. Specifically, zero entries of the concentration matrix correspond to conditionally independent pairs of variables, and vice versa. This is also referred as the covariance selection problem in the literature [6].

Many methods have been proposed in the past 10–15 years to address the problem in the high-dimensional scenario, and most of them roughly fall into two categories. One type of methods is based on penalized likelihood. For example, Yuan and Lin [37] and Banerjee et al. [1] proposed to use the

* Corresponding author

© Science China Press and Springer-Verlag GmbH Germany, part of Springer Nature 2020

LASSO (least absolute shrinkage and selection operator) penalty on off-diagonal elements of the concentration matrix to regularize the Gaussian log-likelihood, while Fan et al. [7] and Lam and Fan [16] considered to use the smoothly clipped absolute deviation (SCAD) penalty [8]. Friedman et al. [11] and Yuan [36] developed efficient algorithms to maximize the LASSO penalized log-likelihood; the former is often referred as the GLASSO (graphical LASSO). Rothman et al. [29] and Ravikumar et al. [28] studied theoretical properties of the GLASSO, while Lam and Fan [16] did the same for SCAD penalized log-likelihood. Another type of methods is based on regularized regression. For example, Meinshausen and Bühlmann [19] proposed the neighborhood selection method, in which each variable is regressed on other variables using LASSO, and the zero/nonzero structure of the concentration matrix can be inferred by using the estimated regression coefficients. Peng et al. [25] proposed the sparse partial correlation estimation (SPACE) method by piling the separate regressions into one single regression and showed that the SPACE method performs well in both nonzero partial correlation selection and the identification of hub variables.

In this paper, we propose a two-step method, where the first step serves as a screening step, in which a significant amount of entries of the concentration matrix are identified as zeros, and then in the second step, we restrict covariance selection only on the reduced parameter space. Since the parameter space has been reduced, we anticipate the final estimate of the concentration matrix is improved over the above mentioned one-step approaches.

The rest of the paper is organized as follows. Section 2 presents the proposed method and Section 3 shows its theoretical properties. Section 4 contains numerical results, including both simulation studies and a data example. Section 5 summarizes the paper. All proofs are in Appendix A.

2 A two-step method for estimating high-dimensional Gaussian graphical models

We focus on the Gaussian graphical model. Specifically, we consider p random variables that follow a multivariate Gaussian distribution

$$X = (X_1, \ldots, X_p) \sim \mathcal{N}_p(0, \Sigma),$$

where Σ is a $p \times p$ positive-definite covariance matrix. We use Θ to denote the concentration matrix, i.e., $\Theta = (\theta_{jj'})_{j,j'=1,\ldots,p} = \Sigma^{-1}$. We note that the dimension p of the concentration matrix is allowed to grow as n increases. However, for notational simplicity, we do not index the parameters with n.

The structure of the concentration matrix Θ can be represented by using an undirected graph $G = (V, S)$, where $V = \{1, \ldots, p\}$ is the set of nodes corresponding to the random variables $X_1, \ldots, X_p$ and S is a set of undirected edges representing the conditional dependence relationships between the variables. Specifically, if variables X_j and $X_{j'}$ are conditionally independent given the rest of the variables, nodes j and j' are not linked in the graph; otherwise they are connected.

It is well known that the elements in Θ can be associated with a set of regression coefficients [13]. Specifically, under the Gaussianity assumption, the variable X_j can be represented as follows:

$$X_j = \sum_{j \neq j'} X_{j'} \beta_{jj'} + \epsilon_j, \tag{2.1}$$

where $\epsilon_j \sim N(0, \sigma_j^2)$ is independent of $X_{j'}$, $j' \neq j$, and $\sigma_j^2 = 1/\theta_{jj}$. Furthermore, it is true that $\beta_{jj'} = -\theta_{jj'}/\theta_{jj}$. Thus, for the Gaussian graphical model, we have

$$(j, j') \quad \text{and} \quad (j', j) \in S \Leftrightarrow \theta_{jj'}, \quad \theta_{jj'} \neq 0 \Leftrightarrow \beta_{jj'}, \quad \beta_{jj'} \neq 0. \tag{2.2}$$

One natural question is then how to efficiently and correctly recover the graphical structure by drawing upon the results in (2.2). In general, as mentioned in the introduction section, there are two types of methods. One type of methods is based on sparse regression techniques, utilizing the relationship in (2.1)

and the LASSO penalty, e.g., the neighborhood selection method [19] and the SPACE method [25]. The other type of methods is based on regularized likelihood, by using the Gaussianity assumption and certain sparse penalty, e.g., the GLASSO algorithm [11] and the SCAD method [7]. In this section, we propose a new two-step approach, which involves an initial screening step and a selection and estimation step. In the first step, we apply a graph learning algorithm to identify most unlinked variable pairs and set the corresponding $\hat{\theta}_{jj'}$'s to zeros; if Θ is sparse, this step can effectively reduce the dimension of the parameter space. In the second step, we restrict the estimation of Θ within the reduced space obtained from the first step. Since the parameter space is reduced for the second step, the statistical estimation accuracy of Θ is expected to be possibly better for the two-step estimate than for the one-step estimate. Similar ideas have been utilized in the regression setting, such as [9,18], but have not been explored with graphical models. Specifically, the proposed algorithm proceeds as follows:

A two-step algorithm for estimating the concentration matrix:

Step 1 (Screening)**.** Denote the j-th variable of the i-th observation as x_{ij}, $i = 1, \ldots, n$. Let

$$\{\hat{\beta}_{jj'}\} = \arg\min \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \sum_{j' \neq j} x_{ij'} \beta_{jj'})^2 + \lambda_1 \sum_{j \neq j'} |\beta_{jj'}|.$$

Let $\hat{A}$ be the nonzero index set of $\hat{\beta}_{jj'}$, i.e., $\hat{A} = \{(j, j') : \hat{\beta}_{jj'} \neq 0 \text{ or } \hat{\beta}_{j'j} \neq 0, j \neq j'\}$.

Step 2 (Selection and estimation)**.** Let $\widetilde{\Theta}$ be the maximum likelihood estimate of Θ based on $\hat{A}$. Furthermore, let $w_{jj'} = 1/|\widetilde{\theta}_{jj'}|$ if $(j, j') \in \hat{A}$. Let

$$\mathcal{M}_{\hat{A}} = \{\Theta \in \mathbb{R}^{p \times p}; \Theta \succ 0 \text{ and } \theta_{jj'} = 0 \text{ for all } (j, j') \notin \hat{A}, j \neq j'\}.$$

Set $\hat{\Theta} = \arg\min_{\Theta \in \mathcal{M}_{\hat{A}}} \{-\log|\Theta| + \mathrm{tr}(\Theta\hat{\Sigma}) + \lambda_2 \sum_{(j,j') \in \hat{A}} w_{jj'} |\theta_{jj'}|\}$, where $\hat{\Sigma}$ is the sample covariance matrix.

Note $\lambda_1 > 0$ and $\lambda_2 > 0$ are tuning parameters controlling the amount of regularization. In the first step, we choose to use the regularized regression based neighborhood selection method, as it has been demonstrated to be effective at identifying zero elements in the concentration matrix [19,25]; furthermore, we use the same tuning parameter λ_1 for the p separate regressions, which reduces both the amount of tuning and the variability of the estimate. In general, due to the nature of the LASSO optimization, we have $\mathrm{Cardinality}(\hat{A}) \leqslant n \times p$. In the second step, we choose to use the weighted graphical LASSO method as the parameter space has been reduced, and the likelihood based method is expected to perform better than sparse regression based methods in the low-dimensional setting. Note for variable pairs that are identified as conditionally uncorrelated in the first step, i.e., $\hat{\beta}_{jj'} = \hat{\beta}_{j'j} = 0$, we set the corresponding $\hat{\theta}_{jj'} = \hat{\theta}_{j'j}$ to zeros. Furthermore, for other variable pairs, we assign different penalties on different entries of the concentration matrix when using the graphical LASSO, which is also expected to yield more accurate estimate.

It should also be noted that the proposed method is related to but different from the adaptive graphical LASSO method [7] and the Gelato (graph estimation with LASSO and thresholding) method [39], which are also two-step methods. In the adaptive graphical LASSO, different weights are also assigned to different entries of the concentration matrix; however, all weights are finite, and hence the dimension of the parameter space is not reduced in the second step, and consequently, the estimate does not benefit from the bias-variance tradeoff as much as the proposed method. In the Gelato method, the first step also reduces the parameter space by setting certain $\hat{\theta}_{jj'}$'s to zeros; however, in the second step, the method simply fits the maximum likelihood estimation to the rest entries of the concentration matrix, and thus the estimated graph structure is completely determined by the first step; in other words, the first step is responsible for selecting the graph structure, while the second step is responsible for estimating the value of nonzero entries of $\hat{\Theta}$. While in the proposed method, the first step serves for the purpose of screening, i.e., setting a significant amount of entries of the concentration matrix to zeros, and the second step does both selection and estimation, on a much reduced parameter space. As we will see in numerical studies, these differences turn out to be crucial and help the proposed method gain better performance than both the adaptive graphical LASSO and the Gelato methods.

3 Theoretical results

Suppose the variables are centered and normalized such that $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$. Let S_j be the set of indices corresponding to variables that are conditionally correlated with the j-th variable, and Θ_{S_j} denotes the $|S_j| \times |S_j|$ submatrix of Θ. To present our theoretical results, we need the following assumptions:

(C.1) For $j = 1, \ldots, p$, it holds that Cardinality$(S_j) \leqslant l_0$, where $l_0 = O(n^{\gamma})$ for some $0 < \gamma < 1$.

(C.2) The following condition holds for a positive constant κ_1:

$$\kappa_1 \triangleq \min_{J \subseteq \{1,\ldots,p\}, |J| \leqslant l_0} \min_{\|v_{J^c}\|_1 \leqslant 3\|v_J\|_1} \frac{\|\Theta^{1/2} v\|_2}{8\|v_J\|_2} > 0,$$

where $|J|$ denotes the cardinality of J.

Note that (C.1) limits the maximal possible rate of growth for the number of variables that are conditionally correlated with each variable. (C.2) is the restricted eigenvalue condition, which is weaker than requiring the minimum eigenvalue of Θ to be bounded away from 0. Note that both (C.1) and (C.2) are regular conditions and are commonly used in the literature, e.g., [19, 22, 25–28, 39]. We use them to show that with high probability the first step does not shrink the nonzero entries of Θ into zeros. In addition, we use the following assumptions for the second step of the algorithm:

(C.3) For any index set A of variable pairs and Cardinality$(A) \leqslant n \times p$, there exists some constant $\alpha \in (0, 1]$ such that $\max_{e \in A/S} \|\Gamma_{e,S}(\Gamma_{SS})^{-1}\|_1 \leqslant 1 - \alpha$, where $\|\cdot\|_1$ denotes the usual ℓ_1 norm of a vector and $\Gamma = \Theta^{-1} \otimes \Theta^{-1}$, with $\otimes$ denoting the Kronecker matrix product.

(C.4) The maximum eigenvalue of Σ is bounded, i.e., for some constant κ_2, we have $\Lambda_{\max}(\Sigma) \leqslant \kappa_2 < \infty$.

Condition (C.3) limits the influence that the non-edge terms, indexed by A/S, can have on the edge-based terms, indexed by S. We note that it can be considered as an alternative version of [28, Assumption 1]. (C.4) assumes that the eigenvalues of the true covariance matrix are bounded away from infinity, and it is used to assure the tail bound of the associated sample covariance of Σ.

Now we state our major theoretical results.

Theorem 3.1. *Suppose* (C.1) *and* (C.2) *hold. Set* $\lambda_1 = K_1\sqrt{\log p/n}$, *where* $K_1 > 4$. *Let* β_{min} *be the minimum absolute value of nonzero elements of* $\beta_{jj'}$, $j \neq j'$. *Assume* $\beta_{min} > \frac{3}{\kappa_1} K_1 \sqrt{l_0 \log p/n}$, *where* κ_1 *is defined in* (C.2). *Suppose the sample size satisfies* $n > K_2 l_0 \log p$ *for some constant* K_2. *We have* $\mathrm{P}(S \subseteq \hat{A}) \geqslant 1 - 1/p$.

This implies that after the screening step, we obtain an index set $\hat{A}$ that contains all nonzero elements of Θ with high probability. It is worth noting that one may further apply a hard-thresholding step to the estimated coefficients, and under certain conditions, one could show such hard-thresholding may achieve sign consistency [20]. However, this involves a thresholding parameter and also once a coefficient is hard-thresholded to zero, it will not be estimated as nonzero anymore in the second step, which is restrictive. Thus we choose not to utilize hard-thresholding and instead use the estimated coefficients for screening, and perform the final selection and estimation in the second step of the algorithm. This turns out to work better than hard-thresholding in practice as illustrated in numerical studies in Section 4.

The following result shows sign consistency for the proposed two-step algorithm.

Theorem 3.2. *Suppose* (C.1)–(C.4) *hold. Set* $\lambda_2 = 4\frac{M}{\alpha}(\frac{\log(np)^{\tau}}{n})^{1/2}$, *where* $\tau > 1$, $M > 0$ *and* α *is defined in* (C.3). *Let* $\theta_{\min}$ *be the minimum absolute value of nonzero entries of* Θ. *Assume* $\theta_{\min} > 4K(\Gamma)(1 + \frac{4M}{\alpha})(\frac{\tau \log(np)}{n})^{1/2}$, *where* $K(\Gamma) = \|\Gamma_{SS}^{-1}\|_{\infty}$ *with* $\|\cdot\|_{\infty}$ *denoting the usual* ℓ_{∞} *norm of a matrix, e.g.,* $\|\Sigma\|_{\infty} = \max_{j=1,\ldots,p} \sum_{j'=1}^{p} |\Sigma_{jj'}|$. *Furthermore, suppose the sample size satisfies*

$$\left(\frac{\tau \log(np)}{n}\right)^{1/2} \leqslant \left\{3\left(1 + \frac{4M}{\alpha}\right) K(\Gamma) \|\Sigma\|_{\infty}^3 l_0\right\}^{-1}.$$

Then the following result holds with probability at least $1 - \frac{1}{p^{\tau-1}} - \frac{1}{p}$:

$$\mathrm{sign}(\hat{\theta}_{jj'}) = \mathrm{sign}(\theta_{jj'}), \quad \textit{for all} \quad 1 \leqslant j, j' \leqslant p.$$

Let q be the number of nonzero off-diagonal elements in Θ. Next, we obtain the rate of convergence for $\hat{\Theta}$ in the Frobenius norm.

Theorem 3.3. *Suppose* (C.1)–(C.4) *hold. Under the same setting of Theorem* 3.2*, the following event holds with probability at least* $1-\frac{1}{p^{\tau-1}}-\frac{1}{p}$:

$$\|\hat{\Theta}-\Theta\|_F \leqslant 2K(\Gamma)\left(1+\frac{4M}{\alpha}\right)\left(\frac{\tau(p+q)\log(np)}{n}\right)^{1/2},$$

where $\|\cdot\|_F$ *denotes the usual Frobenius norm of a matrix.*

4 Numerical results

4.1 Simulation studies

In this section, we conduct simulation experiments to examine the finite sample performance of the proposed method. We also compare the proposed method with two other two-step methods, the adaptive GLASSO and Gelato [7,39], as well as two other one-step methods, the GLASSO and SCAD [7,37].

We consider three simulation settings:

Model 1. Diagonal but heterogeneous model with $\Sigma=\mathrm{diag}(\sigma_j)$, where $\sigma_j=\min(j,200)$ and $j=1,\ldots,p$.

Model 2. Tri-diagonal model with $\theta_{jj}=1$ and $\theta_{j,j-1}=\theta_{j-1,j}=0.3$ for $j=1,\ldots,p$.

Model 3. Set $\Theta=B+\delta I$, where the diagonal entries of B are all equal to zeros, and each off-diagonal entry is generated independently and equals 0.5 with probability ρ and 0 with probability $1-\rho$. We set $\rho=0.02$ and choose δ such that the conditional number of Θ is close to $\min(p,200)$. The matrix is then standardized to have unit diagonals.

To demonstrate the strength of the two-step approach in general, we first compare the performance of the proposed two-step method with that of the GLASSO. To make the comparison fair, we set $\lambda_1=\lambda_2=\lambda$ for the two-step method so that there is only one tuning parameter for both the proposed two-step method and the GLASSO. We fix the sample size $n=100$ and vary the dimensionality $p=50$, $p=80$, $p=120$ and $p=200$. We use the Frobenius norm error $\|\hat{\Theta}-\Theta\|_F$ to evaluate the performance of the estimate. The results are shown in Figure 1. As one can see, across different simulation settings, the proposed two-step method in general outperforms the one-step method GLASSO.

Next, we add SCAD and two other two-step methods, i.e., the adaptive GLASSO and Gelato for comparison. For Model 3, in addition to $\rho=0.02$ (Model 3a), we also consider a denser setting with $\rho=0.1$ (Model 3b). For the adaptive GLASSO, the regularization weights are defined as follows: Let $\widetilde{\Theta}$ be the GLASSO estimate. Then we set $w_{jj'}=\max\{\sqrt{n},1/\min_{(j,j')\in\widetilde{S}}|\widetilde{\theta}_{jj'}|\}$ if $\widetilde{\theta}_{jj'}=0$, and $w_{jj'}=1/|\widetilde{\theta}_{jj'}|$ otherwise. All tuning parameters are selected by using five-fold cross-validation as done in [7,29], and all simulations are replicated 100 times. The results are shown in Table 1. As one can see, in both high- and low-dimensional settings, the proposed two-step method nearly uniformly outperforms other methods.

Furthermore, we compare different methods in terms of identifying zero/nonzero elements of the concentration matrix using the receiver operating characteristic (ROC) curve. Each point on the ROC curve corresponds to a false positive rate (FPR) and a true positive rate (TPR), obtained by a particular value of the tuning parameter. The FPR and the TPR are respectively defined as

$$\mathrm{FPR}=\frac{\#\text{ of incorrectly identified nonzero elements}}{p(p-1)-q},$$
$$\mathrm{TPR}=\frac{\#\text{ of correctly identified nonzero elements}}{q},$$

where p is the number of variables and q is the number of nonzero off-diagonal elements in Θ. For the purpose of comparison, we consider the following model: Set the diagonal elements of Θ to be 1, and as for the off-diagonal elements, we set $\theta_{jj'}=\theta_{j'j}=\eta$ if $|j-j'|\leqslant k$ and 0 otherwise, where η is generated

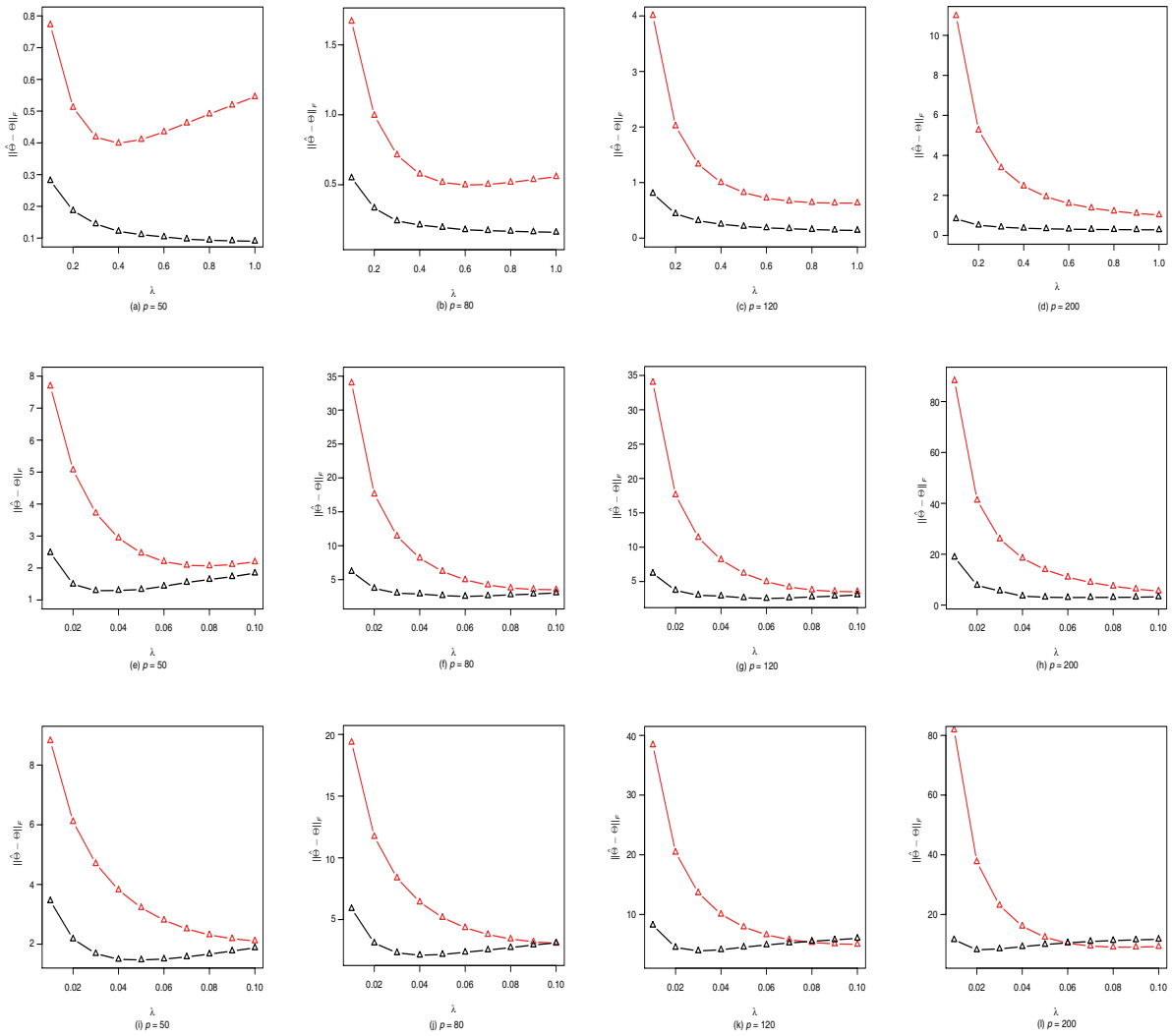

Figure 1 (Color online) Results for three simulation models ($n = 100$, $\rho = 0.02$ for Model 3), corresponding to the three rows. The red line shows the results for the GLASSO and the black line for the proposed two-step method. The vertical axis is the Frobenius norm error $\|\hat{\Theta} - \Theta\|_F$ and the horizontal axis corresponds to the tuning parameter λ

from the uniform distribution Unif$[0, 0.5]$, and we use k to control the sparsity of Θ. Specifically, we have considered four settings: (1) $n = 100$, $p = 50$, $q = 100$; (2) $n = 100$, $p = 50$, $q = 200$; (3) $n = 50$, $p = 100$, $q = 600$; (4) $n = 50$, $p = 100$, $q = 800$; and they correspond to the low/high-dimensional and sparse/non-sparse scenarios. Both the adaptive GLASSO and our method have two tuning parameters and we set $\lambda_1 = \lambda_2 = \lambda$. The results are shown in Figure 2. As one can see, all methods perform better in the sparse settings in comparison with the non-sparse settings, and the two-step methods in general perform better than the one-step methods, in this case, the GLASSO and SCAD. Furthermore, the proposed two-step method performs the best among the five methods in terms of identifying zero/nonzero elements of the concentration matrix.

4.2 The data example

In this subsection, we apply the proposed two-step method to a breast cancer microarray data set that consists of 1,217 genes from 244 breast cancer tumor samples [4, 33, 35]. Breast cancer is a complex disease, where molecular characterization is needed to improve diagnostic and therapeutic strategies, and the tumor process cannot be understood by only analyzing individual genes. Over the past 15 years,

Table 1 Comparison of Frobenius norm errors for four simulation models

		$n = 100$	$p = 50$	
	Two-step method	Adaptive GLASSO	Gelato	SCAD
Model 1	0.16 (0.06)	0.17 (0.07)	0.17 (0.09)	0.38 (0.05)
Model 2	2.23 (0.12)	2.32 (0.16)	2.91 (0.11)	2.29 (0.07)
Model 3a	1.41 (0.14)	2.63 (0.11)	1.68 (0.31)	2.16 (0.12)
Model 3b	2.10 (0.14)	2.79 (0.11)	2.65 (0.41)	2.54 (0.09)
		$n = 50$	$p = 100$	
	Two-step method	Adaptive GLASSO	Gelato	SCAD
Model 1	0.19 (0.07)	0.26 (0.13)	0.24 (0.14)	0.82 (0.02)
Model 2	4.36 (0.16)	4.61 (0.07)	4.86 (0.04)	5.81 (0.02)
Model 3a	4.08 (0.25)	4.43 (0.25)	4.22 (0.36)	5.65 (0.35)
Model 3b	5.69 (0.33)	6.19 (0.17)	5.86 (0.88)	6.98 (0.15)
		$n = 200$	$p = 400$	
	Two-step method	Adaptive GLASSO	Gelato	SCAD
Model 1	0.15 (0.02)	0.82 (0.03)	0.61 (0.02)	0.60 (0.02)
Model 2	4.25 (0.07)	5.19 (0.08)	4.83 (0.06)	5.23 (0.05)
Model 3a	6.55 (0.05)	6.70 (0.09)	5.41 (0.08)	6.56 (0.06)
Model 3b	7.16 (0.19)	10.09 (0.04)	9.17 (0.05)	9.83 (0.05)
		$n = 400$	$p = 800$	
	Two-step method	Adaptive GLASSO	Gelato	SCAD
Model 1	0.08 (0.04)	0.78 (0.02)	0.25 (0.02)	1.03 (0.01)
Model 2	4.07 (0.04)	6.40 (0.03)	4.13 (0.02)	6.45 (0.04)
Model 3a	9.28 (0.20)	14.77 (0.36)	15.99 (0.12)	9.93 (0.08)
Model 3b	10.51 (0.04)	16.57 (0.17)	18.21 (0.10)	14.39 (0.03)

a large amount of gene expression signatures have been identified to have potential clinical usage, and in this section, we focus on $p = 1{,}217$ genes whose expression levels are significantly associated with the breast cancer.

We apply the proposed two-step method to estimate the concentration graph and then use two ways to search for potential interesting genes: (1) we obtain genes by ranking nodes according to their estimated degrees; and (2) we look for gene pairs that are connected over a range of values of the tuning parameters. By using the first way, the top seven genes are: STK15, AARS, VACM-1, SMARCA2, CUL5, PSMB2, and XBP1 (see Figure 3), which are all important known regulators in the tumor process. For example, STK15, the serine/threoninkinase 15 gene, is named the breast tumor amplified kinase (BTAK) since it is strongly correlated with the amplification of breast cancer organizations [21]. AARS, the alanyl-tRNA synthetase, is a protein coding gene which is associated with several diseases. VACM-1, a cul-5 gene, has been shown to have an inhibition effect on the T47D breast cancer cell growth [3].

Using the second way, we find several gene pairs that are always connected when the tuning parameter is varied over a range of values. For example, EXO1 is always connected with SOX2 and NESP55, while recent literature in biology [34] has studied the association between the polymorphisms of the EXO1 gene and the risk of breast cancer and has provided evidence that EXO1 may be associated with the development of breast cancer and may be a useful biomarker for breast cancer detection and primary prevention. It is also found that PTTG is always connected with GNAS1, and it is known that GNAS is associated with survival of patients with invasive breast carcinoma [24].

We have also applied the adaptive GLASSO and Gelato to the same dataset. All three methods always find STK15 and SMARCA2 among the top ranking nodes (in terms of estimated degrees); however, the proposed method also identifies AARS and VACM-1 as important genes while the other two methods do not, and both genes have been shown in the literature that they have specific influences on breast cancer [3, 21]. In terms of gene pairs, over a range of values of the tuning parameter, the Gelato selects a

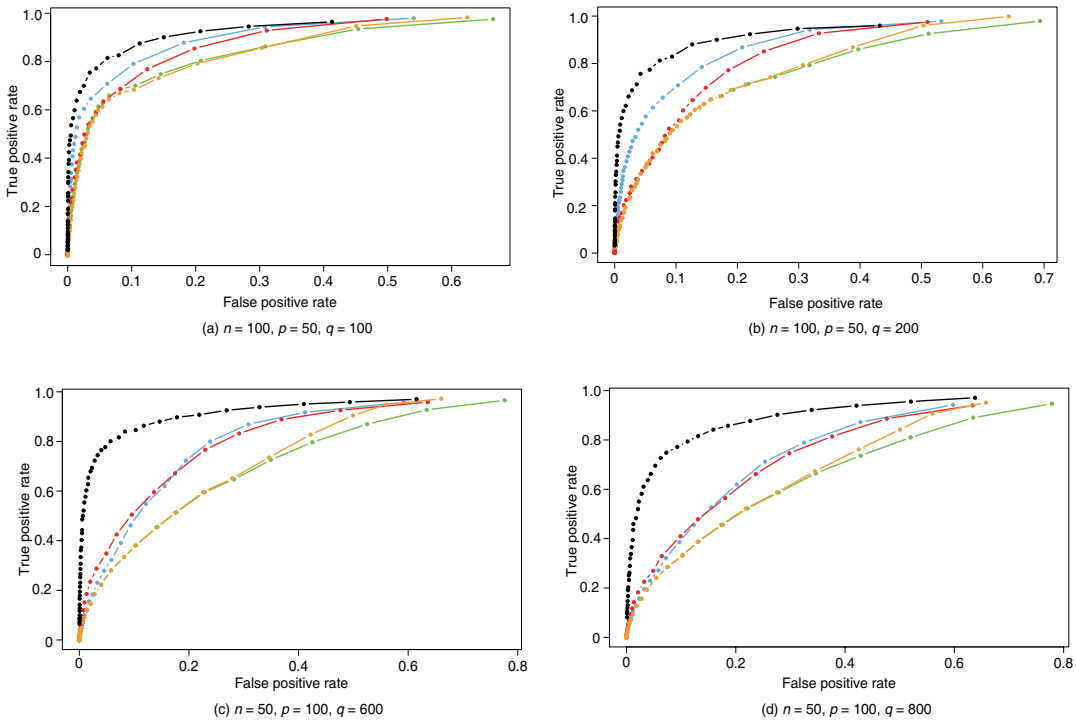

Figure 2 (Color online) Results of ROC curves under the four simulation scenarios. The black, red, blue, orange and green lines correspond to the proposed two-step method, adaptive GLASSO, Gelato, SCAD and GLASSO methods, respectively

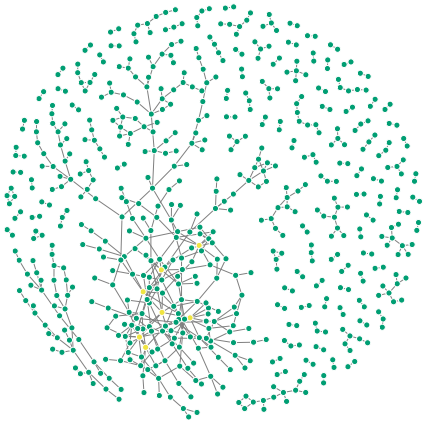

Figure 3 (Color online) The inferred graph from the breast cancer microarray data set by the proposed method. The yellow nodes correspond to the seven genes that have the highest estimated degrees

set of gene pairs similar to that of the proposed method, while the adaptive GLASSO tends to select a smaller set and many of the gene pairs in the selected set are different from those selected by the other two methods. Nevertheless, four gene pairs have always been chosen by all three methods: PTTG2-PTTG1, C4B-C4A, GAPD-ADAMTS7 and GNAS1-NESP55. The first two pairs may be less interesting as PTTG1 and PTTG2 share greater than 80% identity at the amino acid level [12], and C4A and C4B are two isoforms encoded from C4 [17], while for GNAS1-NESP55, it is known that the transcripts from NESP55 splice onto GNAS1 [14].

5 Summary

In this paper, we have proposed a two-step method for estimating high-dimensional Gaussian graphical models. The first step serves as a screening step, in which many entries of the concentration matrix are identified as zeros and thus removed from further consideration. Then in the second step, we focus on the remaining entries of the concentration matrix and apply a weighted graphical LASSO method to select and estimate nonzero entries of the concentration matrix. Since the dimension of the parameter space is effectively reduced by the screening step, the estimation accuracy of the estimated concentration matrix can be potentially improved. We have shown the proposed method enjoys desirable asymptotic properties. Comparisons of the proposed method with several existing methods for Gaussian graphical models on simulation studies indicate that the proposed method works well. We have also applied the proposed method to a breast cancer microarray data set and obtained some biologically meaningful results.

Note that in the first step of the algorithm, we have used a regularized regression based method to screen zero entries of the concentration matrix. We do not use penalized likelihood based methods as it has been demonstrated that the former tends to be more effective at identifying zero elements in the concentration matrix than the latter [19, 25]; our own experiences have also confirmed the same (results not shown). A potential alternative is to use (corrected) p-values and a pre-specified FDR threshold, e.g., [5, 15, 23, 30, 32, 38], and we will explore that in our future work.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant No. 11671059).

References

1 Banerjee O, Ghaoui L-E, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. J Mach Learn Res, 2008, 9: 485–516
2 Bickel P-J, Levina E. Regularized estimation of large covariance matrices. Ann Statist, 2008, 36: 199–227
3 Burnatowska-Hledin M-A, Kossoris J-B, Van Dort C-J, et al. T47D breast cancer cell growth is inhibited by expression of VACM-1, a cul-5 gene. Biochem Bioph Res Co, 2004, 319: 817–825
4 Chang H-Y, Nuyten D-S, Sneddon J-B, et al. Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. Proc Natl Acad Sci USA, 2001, 102: 3738–343
5 Chen M-J, Ren Z, Zhao H-Y, et al. Asymptotically normal and efficient estimation of covariate-adjusted Gaussian graphical model. J Amer Statist Assoc, 2016, 111: 394–406
6 Dempster A-P. Covariance selection. Biometrics, 1972, 28: 157–175
7 Fan J-Q, Feng Y, Wu Y-C. Network exploration via the adaptive LASSO and scad penalties. Ann Appl Stat, 2009, 1: 521–541
8 Fan J-Q, Li R-Z. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360
9 Fan J-Q, Lv L-C. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B Stat Methodol, 2008, 70: 849–911
10 Ferguson T-S. An inconsistent maximum likelihood estimate. J Amer Statist Assoc, 1982, 77: 831–834
11 Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical LASSO. Biostatistics, 2008, 9: 432–441
12 Han X, Poon R. Critical differences between isoforms of securin reveal mechanisms of separase regulation. Mol Cell Biol, 2013, 33: 3400–3415
13 Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009
14 Hayward B, Moran V, Strain L, et al. Bidirectional imprinting of a single gene: GNAS1 encodes maternally, paternally, and biallelically derived proteins. Proc Natl Acad Sci USA, 1998, 95: 15475–15480
15 Jankova J, van de Geer S. Confidence intervals for high-dimensional inverse covariance estimation. Electron J Stat, 2015, 9: 1205–1229
16 Lam C, Fan J-Q. Sparsistency and rates of convergence in large covariance matrix estimation. Ann Statist, 2009, 37: 42–54
17 Li N, Zhang J, Liao D, et al. Association between C4, C4A, and C4B copy number variations and susceptibility to autoimmune diseases: A meta-analysis. Sci Rep, 2017, 7: 42628
18 Meinshausen N. Relaxed LASSO. Comput Statist Data Anal, 2007, 52: 374–393

19 Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the LASSO. Ann Statist, 2006, 34: 1436–1462
20 Meinshausen N, Yu B. Lasso-type recovery of sparse representations for high-dimensional data. Ann Statist, 2009, 37: 246–270
21 Miyoshi Y, Iwao K, Egawa C, et al. Association of centrosomal kinase STK15/BTAK mRNA expression with chromosomal instability in human breast cancers. Int J Cancer, 2001, 92: 370–373
22 Negahban S, Ravikumar P, Wainwright M-J, et al. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Statist Sci, 2012, 27: 1348–1356
23 Ning Y, Liu H. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. Ann Statist, 2017, 45: 158–195
24 Otterbach F, Callies R, Frey U-H, et al. The T393C polymorphism in the gene GNAS1 of G protein is associated with survival of patients with invasive breast carcinoma. Breast Cancer Res Treat, 2007, 105: 311–317
25 Peng J, Wang P, Zhou N, et al. Partial correlation estimation by joint sparse regression models. J Amer Statist Assoc, 2009, 104: 735–746
26 Raskutti G, Wainwright M-J, Yu B. Restricted eigenvalue properties for correlated Gaussian designs. J Mach Learn Res, 2010, 11: 2241–2259
27 Raskutti G, Wainwright M-J, Yu B. Minimax rates of estimation for high-dimensional linear regression over-balls. IEEE Trans Inform Theory, 2011, 57: 6976–6994
28 Ravikumar P, Wainwright M-J, Raskutti G, et al. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electron J Stat, 2011, 5: 935–980
29 Rothman A-J, Bickel P-J, Levina E, et al. Sparse permutation invariant covariance estimation. Electron J Stat, 2008, 2: 494–515
30 Taylor J, Tibshirani R. Post-selection inference for-penalized likelihood models. Canad J Statist, 2018, 46: 41–61
31 Uhler C, Raskutti G, Bühlmann P, et al. Geometry of the faithfulness assumption in causal inference. Ann Statist, 2013, 41: 436–463
32 van de Geer S, Bühlmann P, Ritov Y, et al. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Statist, 2014, 42: 1166–1202
33 van de Vijver M-J, He Y-D, van't Veer L-J, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med, 2002, 37: 1999–2009
34 Wang H-C, Chiu C-F, Tsai R-Y, et al. Association of genetic polymorphisms of EXO1 gene with risk of breast cancer in taiwan. Anticancer Res, 2009, 29: 3897–3901
35 West M, Blanchette C, Dressman H, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci USA, 2001, 98: 11462–11467
36 Yuan M. Efficient computation of l_1 regularized estimates in Gaussian graphical models. J Comput Graph Statist, 2008, 17: 809–826
37 Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94: 19–35
38 Zhang C-H, Zhang S-S. Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B Stat Methodol, 2014, 76: 217–242
39 Zhou S-H, Rütimann P, Xu M, et al. High-dimensional covariance estimation based on Gaussian graphical models. J Mach Learn Res, 2011, 12: 2975–3026

Appendix A

Lemma A.1. *Let $\hat{\beta}$ be the solution to*

$$\{\hat{\beta}_{jj'}\} = \arg\min \frac{1}{2n} \sum_{j=1}^{p} \sum_{i=1}^{n} \left(x_{ij} - \sum_{j' \neq j} x_{ij'} \beta_{jj'} \right)^2 + \lambda_1 \sum_{j \neq j'} |\beta_{jj'}|. \tag{A.1}$$

The optimality conditions for (A.1) *can be written as*

$$\frac{1}{n} \sum_{i=1}^{n} x_{ij'} \cdot \left(x_{ij} - \sum_{j' \neq j} x_{ij'} \hat{\beta}_{jj'} \right) = \lambda_1 \gamma_{jj'}, \tag{A.2}$$

where

$$\gamma_{jj'} = \begin{cases} \operatorname{sign}(\hat{\beta}_{jj'}), & \text{if } j \neq j' \text{ and } \hat{\beta}_{jj'} \neq 0, \\ \text{a real number } \in [-1, 1], & \text{if } j \neq j' \text{ and } \hat{\beta}_{jj'} = 0. \end{cases} \tag{A.3}$$

Note that given j, (A.2) can be seen as an LASSO solution and hence has at most $\min\{n,p\}$ nonzero components.

Proof of Lemma A.1. The partial derivative of the first part of (A.1) with respect of $\beta_{jj'}$ can be calculated as follows:

$$\begin{aligned}&\partial\frac{1}{2n}\Big\{\sum_{j=1}^{p}\cdot\sum_{i=1}^{n}\Big(x_{ij}-\sum_{j'\neq j}x_{ij'}\beta_{jj'}\Big)^2\Big\}\Big/\partial\beta_{jj'}\\&=\frac{1}{2n}\sum_{i=1}^{n}\partial\Big\{\sum_{j=1}^{p}\Big(x_{ij}-\sum_{j'\neq j}x_{ij'}\beta_{jj'}\Big)^2\Big\}\Big/\partial\beta_{jj'}\\&=-\frac{1}{n}\sum_{i=1}^{n}x_{ij'}\cdot\Big(x_{ij}-\sum_{j'\neq j}x_{ij'}\beta_{jj'}\Big).\end{aligned}$$

Note that $\gamma_{jj'}$ of (A.3) is the subgradient of the function $f(x)=\|x\|_1$ evaluated at $x=\hat{\beta}$. Therefore, by the Karush-Kuhn-Tucker (KKT) conditions, we have $\{\hat{\beta}_{jj'}\}$ is a solution of (A.1) if and only if $\hat{\beta}_{jj'}$ satisfies (A.2) and (A.3). □

The following notation will be used throughout the proof of Theorem 3.1. With a bit abuse of notation, we let $\hat{u}=\sqrt{n}(\hat{\beta}-\beta)$ and W be $(p-1)\times p$ matrices with elements $\sqrt{n}(\hat{\beta}_{jj'}-\beta_{jj'})$ and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_{ij'}\epsilon_{ij}$, $j\neq j'$, respectively, and C be the $p\times p$ matrix with elements $\frac{1}{n}\sum_{i=1}^{n}x_{ij}x_{ij'}$, where $j,j'=1,\ldots,p$. Let $\beta_{\cdot j}$, $\hat{u}_{\cdot j}$ and $W_{\cdot j}$ denote the j-th column of β, $\hat{u}$ and W, respectively, e.g.,

$$\hat{u}_{\cdot j}=(\hat{u}_{1j},\ldots,\hat{u}_{(j-1)j},\hat{u}_{(j+1)j},\ldots,\hat{u}_{pj})^{\mathrm{T}}.$$

Similarly, W_{S_j} and $\hat{u}_{S_j}$ denote the sub-vector of $W_{\cdot j}$ and $\hat{u}_{\cdot j}$ corresponding to S_j, respectively.

Lemma A.2. *Conditional on* $\{2\|W_{\cdot j}\|_\infty\leqslant\sqrt{n}\lambda_1\}$, *we have*

$$\|\hat{u}_{S_j^c}\|_1\leqslant 3\|\hat{u}_{S_j}\|_1,\quad \textit{for}\quad j=1,\ldots,p.$$

Proof. Given $j=1,\ldots,p$, we have

$$\begin{aligned}&\frac{1}{2n}\sum_{i=1}^{n}\Big(x_{ij}-\sum_{j'\neq j}x_{ij'}\hat{\beta}_{jj'}\Big)^2+\lambda_1\sum_{j'\neq j}|\hat{\beta}_{jj'}|_1\\&\leqslant\frac{1}{2n}\sum_{i=1}^{n}\Big(x_{ij}-\sum_{j'\neq j}x_{ij'}\beta_{jj'}\Big)^2+\lambda_1\sum_{j'\neq j}|\beta_{jj'}|_1.\end{aligned}$$

Since $\sum_{j'\neq j}|\hat{\beta}_{jj'}|_1=\|\hat{\beta}_{S_j}\|_1+\|\hat{\beta}_{S_j^c}\|_1$ and $\|\hat{\beta}_{S_j}\|_1\geqslant-\|\hat{\beta}_{S_j}-\beta_{S_j}\|_1+\|\beta_{S_j}\|_1$, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\sum_{j'\neq j}x_{ij'}^2(\hat{\beta}_{jj'}-\beta_{jj'})^2+2\lambda_1\|\hat{\beta}_{S_j^c}\|_1\\&\leqslant 2\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\sum_{j'\neq j}x_{ij'}(\hat{\beta}_{jj'}-\beta_{jj'})+2\lambda_1\|\hat{\beta}_{S_j}-\beta_{S_j}\|_1.\end{aligned}\tag{A.4}$$

Conditional on $\{2\|W_{\cdot j}\|_\infty\leqslant\sqrt{n}\lambda_1\}$, we also have

$$\begin{aligned}2\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}\sum_{j'\neq j}x_{ij'}(\hat{\beta}_{jj'}-\beta_{jj'})&\leqslant\lambda_1\sum_{j'\neq j}|\hat{\beta}_{jj'}-\beta_{jj'}|\\&\leqslant\lambda_1\|\hat{\beta}_{S_j}-\beta_{S_j}\|_1+\lambda_1\|\hat{\beta}_{S_j^c}\|_1.\end{aligned}\tag{A.5}$$

Combining (A.4) and (A.5), we have

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j'\neq j}x_{ij'}^2(\hat{\beta}_{jj'}-\beta_{jj'})^2+2\lambda_1\|\hat{\beta}_{S_j^c}\|_1\leqslant 3\lambda_1\|\hat{\beta}_{S_j}-\beta_{S_j}\|_1+\lambda_1\|\hat{\beta}_{S_j^c}\|_1.$$

Therefore, $\|\hat{u}_{S_j^c}\|_1\leqslant 3\|\hat{u}_{S_j}\|_1$. □

Lemma A.3. *Given $j=1,\dots,p$, according to* (C.1) *and* (C.2)*, conditional on*

$$\{2\|W_{S_j}\|_2 \leqslant \sqrt{l_0}\sqrt{n}\lambda_1\} \cap \{2\|W_{\cdot j}\|_\infty \leqslant \sqrt{n}\lambda_1\},$$

we have

$$\|\hat{u}_{S_j}\|_2 \leqslant \frac{3}{\kappa_1}\sqrt{l_0}\sqrt{n}\lambda_1.$$

Proof. It holds that

$$\text{Set } F(\beta) = \frac{1}{2n}\sum_{j=1}^{p}\sum_{i=1}^{n}\left(x_{ij} - \sum_{j'\neq j} x_{ij'}\beta_{jj'}\right)^2 + \lambda_1 \sum_{j\neq j'}|\beta_{jj'}|.$$

Define $V(u) = F(\hat{\beta}) - F(\beta)$:

$$\begin{aligned} V(u) &= \sum_{j=1}^{p}\left\{\frac{1}{2n}u_{\cdot j}^{\mathrm{T}}C_{-j}u_{\cdot j} - \frac{1}{n}u_{\cdot j}^{\mathrm{T}}W_{\cdot j}\right\} + \lambda_1 \sum_{j\neq j'}\left(\left|\beta_{jj'} + \frac{u_{jj'}}{\sqrt{n}}\right| - |\beta_{jj'}|\right) \\ &= \sum_{j=1}^{p}\left\{\frac{1}{2n}u_{\cdot j}^{\mathrm{T}}C_{-j}u_{\cdot j} - \frac{1}{n}u_{\cdot j}^{\mathrm{T}}W_{\cdot j} + \lambda_1\left(\left\|\beta_{\cdot j} + \frac{u_{\cdot j}}{\sqrt{n}}\right\|_1 - \|\beta_{\cdot j}\|_1\right)\right\} \\ &\triangleq \sum_{j=1}^{p}\{G_j(u)\}, \end{aligned}$$

where C_{-j} denotes the $(p-1)\times(p-1)$ matrix without the j-th column and the j-th row, $j=1,\dots,p$. We have $\hat{u} = \arg\min_u V(u)$. Note that according to [26, Corollary 1] there are universal positive constants c' and c such that with probability at least $1-c'\exp(-cn)$, we have for all $v\in\mathbb{R}^p$,

$$\|C^{1/2}v\|_2 \geqslant \frac{1}{8}\|\Sigma^{1/2}v\|_2.$$

By (C.2) and $\|\hat{u}_{S_j^c}\|_1 \leqslant 3\|\hat{u}_{S_j}\|_1$, we have $\hat{u}_{\cdot j}^{\mathrm{T}}C_{-j}\hat{u}_{\cdot j} \geqslant \kappa_1\|\hat{u}_{S_j}\|_2^2$. Then for $j=1,\dots,p$, we have, for $u_{\cdot j}$ satisfying (C.2), that

$$\begin{aligned} G_j(u) &= \frac{1}{2n}u_{\cdot j}^{\mathrm{T}}C_{-j}u_{\cdot j} - \frac{1}{n}u_{\cdot j}^{\mathrm{T}}W_{\cdot j} + \lambda_1\left(\left\|\beta_{\cdot j} + \frac{u_{\cdot j}}{\sqrt{n}}\right\|_1 - \|\beta_{\cdot j}\|_1\right) \\ &\geqslant \frac{1}{n}\left[\frac{\kappa_1}{2}\|u_{S_j}\|_2^2 - |u_{S_j}^{\mathrm{T}}W_{S_j}| - \sqrt{n}\lambda_1\|u_{S_j}\|_1\right] + \left[\frac{\lambda_1}{\sqrt{n}}\|u_{S_j^c}\|_1 - \frac{1}{n}|u_{S_j^c}^{\mathrm{T}}W_{S_j^c}|\right]. \end{aligned} \tag{A.6}$$

Conditional on $\{2\|W_{\cdot j}\|_\infty \leqslant \sqrt{n}\lambda_1\}$, the second part of (A.6) can be calculated as

$$\frac{\lambda_1}{\sqrt{n}}\|u_{S_j^c}\|_1 - \frac{1}{n}|u_{S_j^c}^{\mathrm{T}}W_{S_j^c}| \geqslant \sum_{j'\notin S_j^c}|u_{jj'}|\left[\frac{\lambda_1}{\sqrt{n}} - \frac{\|W_{S_j^c}\|_\infty}{n}\right] > 0. \tag{A.7}$$

By (C.1), the first term of the right-hand side of (A.6) is bounded as

$$\begin{aligned} &\frac{1}{n}\left[\frac{\kappa_1}{2}\|u_{S_j}\|_2^2 - |u_{S_j}^{\mathrm{T}}W_{S_j}| - \sqrt{n}\lambda_1\|u_{S_j}\|_1\right] \\ &\quad\geqslant \frac{1}{n}\|u_{S_j}\|_2\left[\frac{\kappa_1}{2}\|u_{S_j}\|_2 - \|W_{S_j}\|_2 - \sqrt{n}\lambda_1\cdot\sqrt{l_0}\right]. \end{aligned}$$

Hence $G_j(u)$ is greater than zero as well when

$$\|u_{S_j}\|_2 > \frac{2}{\kappa_1}\{\|W_{S_j}\|_2 + \sqrt{l_0}\sqrt{n}\lambda_1\}.$$

Conditional on $\{2\|W_{S_j}\|_2 \leqslant \sqrt{l_0}\sqrt{n}\lambda_1\}$, define

$$M_0 \equiv \frac{3}{\kappa_1}\sqrt{l_0}\sqrt{n}\lambda_1.$$

Since we have $G_j(0)=0$, it follows that the minimum of $G_j(u)$ cannot be attained at any u satisfying $\|\hat{u}_{S_j}\|_2 > M_0$. Thus, conditional on $\{2\|W_{S_j}\|_2 \leqslant \sqrt{l_0}\sqrt{n}\lambda_1\}$ and (C.2), we have $\|\hat{u}_{S_j}\|_2 \leqslant M_0$. □

Proof of Theorem 3.1. According to Lemma A.3, (C.1) and (C.2), conditional on

$$\{2\|W_{S_j}\|_2 \leqslant \sqrt{l_0}\sqrt{n}\lambda_1\} \cap \{2\|W_{\cdot j}\|_\infty \leqslant \sqrt{n}\lambda_1\},$$

we have for $j = 1, \ldots, p$,

$$\|\hat{u}_{S_j}\|_2 \leqslant \frac{3}{\kappa_1}\sqrt{l_0}\sqrt{n}\lambda_1$$

and

$$\sqrt{n}\|\hat{\beta}_{S_j} - \beta_{S_j}\|_\infty \leqslant \|\hat{u}_{S_j}\|_2 \leqslant \frac{3}{\kappa_1}\sqrt{l_0}\sqrt{n}\lambda_1.$$

Since $\lambda_1 = K_1\sqrt{\log p/n}$, we have for every $(j, j') \in S$, $\hat{\beta}_{jj'} \neq 0$.

For the event in Lemma A.3, by $n > K_2 l_0 \log p$, we have

$$\mathrm{P}\bigg(\|W_{\cdot j}\|_\infty > \frac{K_1}{2}\sqrt{\log p}\bigg) \leqslant p \cdot \bigg(\frac{K_1}{2}\sqrt{\log p/n}\bigg)^{-1} \exp\bigg(-\frac{K_1}{4}\log p\bigg) < \frac{1}{p},$$

$$\mathrm{P}\bigg(\|W_{S_j}\|_2 > \frac{K_1}{2}\sqrt{l_0}\sqrt{\log p}\bigg) \leqslant l_0 \cdot \bigg(\frac{K_1}{2}\sqrt{\log p/n}\bigg)^{-1} \exp\bigg(-\frac{K_1}{4}\log p\bigg) < \frac{1}{p}$$

and for all $v \in \mathbb{R}^p$,

$$\mathrm{P}\bigg(\|C^{1/2}v\|_2 \geqslant \frac{1}{8}\|\Sigma^{1/2}v\|_2\bigg) < \frac{1}{p}.$$

Thus, we have

$$\mathrm{P}(S \nsubseteq \hat{A}) \leqslant \frac{1}{p} \to 0, \quad \text{as} \quad n \to \infty.$$

This completes the proof. □

Now we prove Theorem 3.2. We change the definition of S a little. Specifically, let S be the set containing both diagonal and off-diagonal nonzero entries of Θ, i.e.,

$$S = \{(j, j') : \theta_{jj'} \neq 0\}, \tag{A.8}$$

and Cardinality$(S) = q + p$. Then we introduce the following Lemma, where more details can be found in [2, Lemma 3] and [29, Lemma 1].

Lemma A.4. *By* (C.4)*, let Z_i's be independent identically distributed from $\mathcal{N}(0, \Sigma)$ and $\Lambda_{\max}(\Sigma) < \kappa_2 < \infty$. Then for $\Sigma = \{\sigma_{jj'}\}$, the associated sample covariance $\hat{\Sigma} = \{\hat{\sigma}_{jj'}\}$ satisfies the tail bound*

$$\mathrm{P}(|\hat{\sigma}_{jj'} - \sigma_{jj'}| > \delta) \leqslant K_3 \exp(-K_4 n\delta^2),$$

where K_3 and K_4 are positive constants depending on the maximum eigenvalue of Σ, and $|\delta| \leqslant \kappa_3$, where κ_3 depends on κ_2 as well.

Proof. We omit the proof and refer the readers to [2, 29]. □

Note that given the property of the solution of LASSO, we have Cardinality$(\hat{A}) < n \cdot p$, and thus $\hat{A}$ satisfies (C.3). Furthermore, according to Theorem 3.1, we have $S \subseteq \hat{A}$ with high probability. For notational simplicity, we denote it as A for the rest of the proof instead of $\hat{A}$.

Lemma A.5. *For any $\lambda_2 > 0$ and sample covariance $\hat{\Sigma}$ with strictly positive diagonal elements, the restricted log-determinant problem has a unique solution $\hat{\Theta} \succ 0$ characterized by*

$$-\hat{\Theta}^{-1} + \hat{\Sigma} + \lambda_2\hat{Z} = 0, \tag{A.9}$$

where $\hat{Z}_A$ is an element of the subdifferential $\partial(\sum_{(j,j')\in A}|w_{jj'}\theta_{jj'}|)$ such that

$$\hat{Z}_{jj'} = \begin{cases} \mathrm{sign}(\hat{\theta}_{jj'}) \cdot \hat{w}_{jj'}, & \text{if } \hat{\theta}_{jj'} \neq 0, \\ \text{a real number } \in [-\hat{w}_{jj'}, \hat{w}_{jj'}], & \text{if } \hat{\theta}_{jj'} = 0. \end{cases}$$

This result is similar to [28, Lemma 3], and hence we omit the proof here.

Lemma A.6. *Suppose* (C.1)–(C.3) *hold. Conditional on the first step solution, with appropriately chosen* λ_2 *and the sample size, the following event holds with probability at least* $1-\frac{1}{p^{\tau-1}}$ $(\tau>1)$,

$$\operatorname{sign}(\hat{\Theta}_{S^c})=\operatorname{sign}(\Theta_{S^c})=0.$$

Proof. We assume there exists a solution $\bar{\Theta}$ for the following restricted log-determinant problem:

$$\bar{\Theta}=\arg\min_{\Theta\in\mathcal{M}_S}\{-\log|\Theta|+\operatorname{tr}(\Theta\hat{\Sigma})+\lambda_2\widetilde{Z}\},\tag{A.10}$$

where $\mathcal{M}_S=\{\Theta\in\mathbb{R}^{p\times p};\Theta\succ 0$ and $\theta_{jj'}=0$ for all $(j,j')\notin S, j\neq j'\}$.

By construction, set $\widetilde{Z}_{jj'}=\frac{1}{\lambda_2}(-\hat{\Sigma}_{jj'}+[\bar{\Theta}^{-1}]_{jj'})$, which makes $\widetilde{Z}$ satisfy (A.9). If we are able to show $|\widetilde{Z}_{jj'}|\leqslant w_{jj'}$, where $(j,j')\in A/S$, then it implies $\bar{\Theta}$ is equal to the solution $\hat{\Theta}$ of the original criterion. The argument is along the lines of a result due to [28]. We apply it to the second step of the proposed method, which has shrunk many edges to zeros.

Let $\Delta=\bar{\Theta}-\Theta$, and $R(\Delta)$ be the difference between the gradient $\nabla(\log|\bar{\Theta}|)=\bar{\Theta}^{-1}$ and its first-order Taylor expansion around Θ by using the first and second derivatives of the log-determinant function, which is

$$R(\Delta)=\bar{\Theta}^{-1}-\Theta^{-1}+\Theta^{-1}\Delta\Theta^{-1}.$$

Then the solution of (A.10) can be rewritten as

$$\Theta^{-1}\Delta\Theta^{-1}+H-R(\Delta)+\lambda_2\widetilde{Z}=0,$$

where $H\triangleq\hat{\Sigma}-\Theta^{-1}=\hat{\Sigma}-\Sigma$. Let $\Gamma=\Theta^{-1}\otimes\Theta^{-1}$, where $\otimes$ denotes the Kronecker matrix product. Consider the sub-block of Γ, i.e.,

$$\Gamma_{SS}:=[\Theta^{-1}\otimes\Theta^{-1}]_{SS}\in R^{(q+p)\times(q+p)},$$

where q is the number of nonzero off-diagonal entries of the concentration matrix (divided by 2) and p is the number of diagonal elements of the concentration matrix. Let $\bar{\Delta}$ be the vector version of Δ. The above equality can be rewritten into two blocks,

$$\begin{aligned}&\Gamma_{SS}\bar{\Delta}+\bar{H}_S-\bar{R}_S+\lambda_2\bar{\bar{Z}}_S=0,\\&\Gamma_{(A/S)S}\bar{\Delta}+\bar{H}_{A/S}-\bar{R}_{A/S}+\lambda_2\bar{\bar{Z}}_{A/S}=0.\end{aligned}$$

It follows that

$$\bar{\bar{Z}}_{A/S}=-\frac{1}{\lambda_2}(\Gamma_{(A/S)S}\bar{\Delta}+\bar{H}_{(A/S)}-\bar{R}_{(A/S)}).\tag{A.11}$$

We now consider the upper bounds of $\|\bar{H}_A\|_\infty$ and $\|\bar{R}_A\|_\infty$. First, consider a multivariate Gaussian vector; the deviation of the sample covariance matrix $\hat{\Sigma}$ has an exponential type tail bound. Let $\delta=(\frac{\tau\log(np)}{n})^{1/2}$. We have

$$\begin{aligned}\mathrm{P}(\|\bar{H}_A\|_\infty>\delta)&<n\cdot p\cdot\mathrm{P}(|\hat{\sigma}_{jj'}-\sigma_{jj'}|>\delta)\\&\leqslant n\cdot p\cdot K_3\exp(-K_4n\delta^2)\\&\leqslant\frac{1}{p^{\tau-1}}.\end{aligned}$$

Second, use [28, Lemma 5], and set $G_0=\frac{3}{2}l_0\|\Sigma\|_\infty^3$, where $\|\Sigma\|_\infty=\max_{j=1,\ldots,p}\sum_{j'=1}^p|\Sigma_{jj'}|$. Then conditional on the first step, we have

$$\|\bar{R}_A\|_\infty\leqslant G_0\|\hat{\Theta}_A-\Theta_A\|_\infty.$$

Note that the above inequality holds according to the result

$$\|\hat{\Theta}_A - \Theta_A\|_\infty < (3l_0 \cdot \|\Sigma\|_\infty)^{-1}.$$

Given j, since Cardinality$(A_j) \leqslant n$, the maximum likelihood estimate of Θ based on A exists and is unique [31]. By [10,39], $\widetilde{\Theta}$ converges to a certain matrix with the rate at least $1/p^{\tau-1}$.

Define $\mathcal{B}$ as

$$\mathcal{B} := \{\Delta_S : \|\Delta_S\|_\infty \leqslant 2K(\Gamma)(\|H_S\|_\infty + \lambda_2 M_1\},$$

where $K(\Gamma) = \|\Gamma_{SS}^{-1}\|_\infty$ and there exists a constant M_1 such that $\|\hat{w}_S\|_\infty < M_1$. By [28, Lemma 6], the solution $\Delta_S = \bar{\Theta}_S - \Theta_S$ is contained inside the l_∞ ball. Conditional on the event $\mathcal{B}$ and choosing a suitable n, we have $\|\bar{R}_S\|_\infty \leqslant \delta$.

Now we prove the upper bound of $|\bar{\widetilde{Z}}_{A/S}|$ in (A.11). By setting $\widetilde{Z}_{jj'}$, we have $-\bar{\Theta}_S^{-1} + \hat{\Sigma}_S + \lambda_2 \widetilde{Z}_S = 0$. Thus the following equality holds:

$$\bar{\Delta}_S - (\Gamma_{SS})^{-1}\mathrm{vec}(-[(\Theta + \Delta)^{-1}]_S + \hat{\Sigma}_S + \lambda_2 \widetilde{Z}_S) = \bar{\Delta}_S. \tag{A.12}$$

By a direct calculation, we have

$$\begin{aligned}&\bar{\Delta}_S - (\Gamma_{SS})^{-1}\mathrm{vec}(-[(\Theta + \Delta)^{-1}]_S + \hat{\Sigma}_S + \lambda_2 \widetilde{Z}_S)\\ &\quad = (\Gamma_{SS})^{-1} R_S - (\Gamma_{SS})^{-1}(\bar{H}_S + \lambda_2 \bar{\widetilde{Z}}_S).\end{aligned} \tag{A.13}$$

By letting $\lambda_2 = \frac{4M_2}{\alpha}\delta$ and setting $\max_{(j,j')\in A/S} |\widetilde{\theta}_{jj'}| \leqslant M_2$, from (A.12) and (A.13), with $e \in A/S$, it follows

$$\begin{aligned}|e\bar{\widetilde{Z}}_{A/S}| &= e\left|\frac{1}{\lambda_2}(\Gamma_{(A/S)S}\bar{\Delta} + \bar{H}_{(A/S)} - \bar{R}_{(A/S)})\right|\\ &= \left| - \frac{1}{\lambda_2} e\Gamma_{(A/S)S}(\Gamma_{SS})^{-1}(\bar{H}_S - \bar{R}_S) + e\Gamma_{(A/S)S}(\Gamma_{SS})^{-1}\bar{\hat{Z}}_S\right.\\ &\quad \left. - e\frac{1}{\lambda_2}(\bar{H}_{(A/S)} - \bar{R}_{(A/S)})\right|\\ &\leqslant \frac{1}{\lambda_2}|e\Gamma_{(A/S)S}(\Gamma_{SS})^{-1}|(\|\bar{H}_S\|_\infty + \|\bar{R}_S\|_\infty) + |e\Gamma_{(A/S)S}(\Gamma_{SS})^{-1}\bar{\widetilde{Z}}_S|\\ &\quad + \frac{1}{\lambda_2}(\|\bar{H}_S\|_\infty + \|\bar{R}_S\|_\infty)\\ &\leqslant \frac{\alpha}{2M_2}(1-\alpha) + (1-\alpha)M_1 + \frac{\alpha}{2M_2}\\ &< \frac{q}{M_2}.\end{aligned}$$

Then we have $|\widetilde{Z}_{jj'}| \leqslant w_{jj'}$ where $(j,j') \in A/S$, which implies $\bar{\Theta}$ is equal to the solution $\hat{\Theta}$ and $\mathrm{sign}(\hat{\Theta}_{S^c}) = 0$. □

Proof of Theorem 3.2. We now have all necessary ingredients to prove Theorem 3.2. We show that with high probability the nonzero edge set of $\bar{\Theta}$ is equal to the solution of the proposed two-step method $\hat{\Theta}$. Thus we have

$$\|\hat{\Theta}_S - \Theta_S\|_\infty \leqslant 2K(\Gamma)(\|H_S\|_\infty + \lambda_2 M_1).$$

Let $M = M_1 \times M_2$. With probability at least $1 - \frac{1}{p^{\tau-1}}$, we have

$$\|\hat{\Theta}_S - \Theta_S\|_\infty \leqslant 2K(\Gamma)\left(1 + \frac{4M}{\alpha}\right)\left(\frac{\tau\log(np)}{n}\right)^{1/2}.$$

The minimum absolute value $\theta_{\min}$ of nonzero entries of Θ satisfies

$$\theta_{\min} > 4K(\Gamma)\left(1 + \frac{4M}{\alpha}\right)\left(\frac{\tau\log(np)}{n}\right)^{1/2}.$$

Hence it equals to saying $\text{sign}(\hat{\Theta}_S) = \text{sign}(\Theta_S)$.

In addition, according to the result of Lemma A.6, with probability at least $1 - \frac{1}{p^{\tau-1}}$, we have

$$\text{sign}(\hat{\Theta}_{S^c}) = \text{sign}(\Theta_{S^c}) = 0. \tag{A.14}$$

Since (A.14) is built on the first step of the proposed method, overall, we have

$$\begin{aligned}\text{P}(\text{sign}(\hat{\Theta}) \neq \text{sign}(\Theta)) &\leqslant \text{P}(\text{sign}(\hat{\Theta}_A) \neq \text{sign}(\Theta_A), S \subseteq A) + \text{P}(S \nsubseteq A)\\ &\leqslant \frac{1}{p^{\tau-1}} + \exp(-n^{\eta}).\end{aligned}$$

This completes the proof. □

Proof of Theorem 3.3. Note we have

$$\|\hat{\Theta}_S - \Theta_S\|_{\infty} \leqslant 2K(\Gamma)\left(1 + \frac{4M}{\alpha}\right)\left(\frac{\tau \log(np)}{n}\right)^{1/2},$$

with probability at least $1 - \frac{1}{p^{\tau-1}} - \exp(-n^{\eta})$. By the setting of S in (A.8), we have $\text{Cardinality}(S) = p+q$. It is straightforward to calculate that

$$\|\hat{\Theta} - \Theta\|_F \leqslant 2K(\Gamma)\left(1 + \frac{4M}{\alpha}\right)\left(\frac{\tau(p+q) \log(np)}{n}\right)^{1/2}.$$

This completes the proof. □