Published in final edited form as:
J Multivar Anal. 2013 April 1; 116: . doi:10.1016/j.jmva.2013.01.005.

Adjusting for High-dimensional Covariates in Sparse Precision Matrix Estimation by ℓ_1-Penalization

Jianxin Yin and **Hongzhe Li**[1]
Center for Applied Statistics and School of Statistics, Renmin University of China, No. 59 Zhongguancun Street, Haidian District, Beijing 100872, China and Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104-6021, USA

Abstract

Motivated by the analysis of genetical genomic data, we consider the problem of estimating high-dimensional sparse precision matrix adjusting for possibly a large number of covariates, where the covariates can affect the mean value of the random vector. We develop a two-stage estimation procedure to first identify the relevant covariates that affect the means by a joint ℓ_1 penalization. The estimated regression coefficients are then used to estimate the mean values in a multivariate sub-Gaussian model in order to estimate the sparse precision matrix through a ℓ_1-penalized log-determinant Bregman divergence. Under the multivariate normal assumption, the precision matrix has the interpretation of a conditional Gaussian graphical model. We show that under some regularity conditions, the estimates of the regression coefficients are consistent in element-wise ℓ_∞ norm, Frobenius norm and also spectral norm even when $p \gg n$ and $q \gg n$. We also show that with probability converging to one, the estimate of the precision matrix correctly specifies the zero pattern of the true precision matrix. We illustrate our theoretical results via simulations and demonstrate that the method can lead to improved estimate of the precision matrix. We apply the method to an analysis of a yeast genetical genomic data.

Keywords

Estimation bounds; Graphical Model; Model selection consistency; Oracle property

1. Introduction

Estimation of high-dimensional covariance/precision matrix has attracted a great deal of interest in recent years [1, 2, 3, 4, 5, 6]. The problem is related to sparse Gaussian graphical modeling where the precision matrix provides information on the conditional independency among a large set of variables. Application of estimating the precision matrix includes analysis of gene expression data, spectroscopic imaging, FMRI data, numerical weather forecasting. Under the assumption of sparsity and some regularity conditions on the underlying precision matrix, regularization methods have been proposed to estimate such precision matrices. Some explicit rates of convergence of the resulting estimates have been

[1]Corresponding author. hongzhe@upenn.edu.

obtained [1, 2, 7, 6]. Furthermore, [4] and [3] have studied the optimal convergence rate of the estimates in Frobenius and operator norms, as well as the matrix ℓ_1 norm.

Almost all current methods for precision matrix estimation or Gaussian graphical model estimation assume that the random vector has zero or constant mean. However, in many real applications, it is often important to adjust for the covariate effects on the mean of the random vector in order to obtain more precise and interpretable estimate of the precision matrix. One such example is related to analysis of genetical genomic data, where we have both high dimensional genetic marker data and high dimensional gene expression data measured on the same set of samples in a segregation population. One important goal is to study the conditional independency structure among a set of genes at the expression level. This is related to estimating the precision matrix when the data are assumed to be normally distributed. However, it is now known that genetic marker data can affect the mean gene expression levels for many genes [8]. It is therefore important to adjust for the marker effects on gene expression when the conditional independency structure is studied.

In this paper, we consider the problem of adjusting for high-dimensional covariates in precision matrix estimation by ℓ_1-penalization. It can be formulated as the sparse multivariate regression with correlated errors. The model has both high dimensional regression coefficient matrix and high dimensional covariance matrix. Estimation of such multivariate regressions with correlated errors have been studied in literature. [9] focused on estimating the regression coefficient matrix and presented several algorithms based on ℓ_1 penalization. However, no theoretical results were provided. [10] developed an estimation procedure that iteratively estimates the regression coefficients and the precision matrix based on ℓ_1-penalization. They provided asymptotic results on estimation bounds and consistency. However, the computation is quite intensive.

We propose a two-stage ℓ_1 penalization procedure that first jointly estimates the multiple regression coefficients to obtain a sparse estimation of the regression coefficient matrix. We extend the results of [11] on sharp recovery and convergence rate for sparsity in single regression to multiple regression setting. The estimates of the regression coefficients are then used to adjust for the means in estimating the precision matrix. Under the assumption of a matrix version of the irrepresentable condition [12, 11] on the covariate matrix as well as a matrix version of the irrepresentable condition on the precision matrix, we obtain the consistency results. We additionally obtain the explicit convergence rates for both the estimates of the regression coefficient matrix and estimates of precision matrix in element-wise ℓ_∞ norm, hence also in spectral and Frobenius norms. The theoretical property of our estimates depends on the method of primal dual witness construction [11, 5]. If the primal-dual witness construction succeeds, it acts as a witness to the fact that the solution to the restricted problem is equivalent to the solution to the original problem. When further conditions on the minimum values of the true coefficient matrix and the precision matrix are assumed, we also establish the sign consistency results for the estimates.

2. Model and notation

Consider a random vector $Y \in \mathbb{R}^p$ and a deterministic covariate vector $X \in \mathbb{R}^q$, we assume that

$$Y = \Gamma X + \varepsilon, \quad (1)$$

where Γ is the $p \times q$ regression coefficient matrix, ε is a mean-zero error vector and is assumed to distribute as a sub-Gaussian vector with covariance matrix $\Sigma = \Theta^{-1}$ and precision matrix Θ. Specifically, we assume that for each ε^j in $\varepsilon = (\varepsilon^1, \ldots, \varepsilon^p)$, $\varepsilon^j / \sqrt{\Sigma_{jj}}$ is

sub-Gaussian with parameter σ. A zero-mean random variable Z is sub-Gaussian if there exists a constant $\sigma \in (0, \infty)$ such that $E[\exp(tZ)] \leq \exp(\sigma^2 t^2/2)$, for all $t \in \mathbb{R}$. By Chernoff bound, this upper bound on the moment generating function implies a two-sided tail bound of the form $\mathrm{pr}(|Z| > z) \leq 2\exp(-z^2/(2\sigma^2))$. If every element in the vector ε is sub-Gaussian, we call the vector ε sub-Gaussian.

Given n independent and identically distributed observations of a random vector $(Y|X)$, we propose to estimate the regression coefficient matrix Γ and precision matrix Θ in model (1) in a two-step ℓ_1 penalization procedure. To simplify the problem, we assume the X_i are fixed observations for $i = 1, \cdots, n$. Denote $X = (X_1, \cdots, X_n)^\top = (X^{(1)}, \cdots, X^{(q)})$ as the design matrix. Denote $W = (\varepsilon_1^\top, \cdots, \varepsilon_n^\top)^\top$ as the realized noise matrix and $Y = (Y_1, \cdots, Y_n)^\top$. We further denote $C_X = X^\top X/n = \sum_{i=1}^n X_i X_i^\top/n, C_{YX} = Y^\top X/n = \sum_{i=1}^n Y_i X_i^\top/n$ and $C_Y = Y^\top Y/n = \sum_{i=1}^n Y_i Y_i^\top/n$.

We first introduce notation related to vector and matrix norms. We use the notation $A \succ 0$ for the positive definiteness of matrix A. We denote $\bar{A} = \mathrm{vec}(A)$ as the vectorization of an arbitrary matrix A. Define $\|A\|_1 = \sum_{i,j} |A_{ij}|$ as the element-wise ℓ_1 norm for a matrix A and $\|A\|_{1,\mathrm{off}} = \sum_{i \neq j} |A_{ij}|$ as the off-diagonal ℓ_1 norm of matrix A. We denote $\|A\|_\infty = \max_{i,j} |A_{ij}|$ and $|||A|||_\infty = \max_{i=1,\cdots,p} \sum_{j=1}^p |A_{ij}|$ as the element-wise ℓ_∞ norm and the matrix ℓ_∞ norm of a matrix A, respectively. Furthermore, we use $\|A\|_F$ as the Frobenius norm, which is the square-root of the sum of the squares of the entries of A, and $\|A\|_2$ as the spectral norm, which is the largest singular value of A. Finally, we use Γ^*, Σ^* and Θ^* to denote the true matrix parameters in model (1), while $\hat{\Gamma}, \hat{\Sigma}$ and $\hat{\Theta}$ as their estimates.

As commonly used in Gaussian graphical model, we similarly relate the nonzero elements of the precision matrix Σ^* to the edges between two variables, and define the support of the precision matrix as

$$E(\Theta^*) := \{i, j \in (1, \cdots, p) | i \neq j, \Theta^*_{ij} \neq 0\},$$

and the maximum degree or row cardinality of Θ^* as

$$d_1 := \max_{i=1,\cdots,p} \left| \{j \in (1, \cdots, p) | \Theta^*_{ij} \neq 0\} \right|.$$

Similarly, for the regression coefficient matrix, let $T(\Gamma^*)$ be the support of a matrix Γ^*, defined as

$$T(\Gamma^*) := \left\{(i, j) : \Gamma^*_{i,j} \neq 0, \text{where} i \in (1, \cdots, p), j \in (1, \cdots, q)\right\}.$$

Also define $T(i) := \{j : \Gamma^*_{i,j} \neq 0, j = (1, \cdots, q)\}$, which is the support of the regression coefficients for the ith variable. We define the maximum degree or row cardinality of Γ^* as

$$d_2 := \max_{i=1,\cdots,p} \left| \left\{j \in \{1, \cdots, q\} | \Gamma^*_{ij} \neq 0\right\} \right|,$$

which corresponds to the maximum number of non-zeros in any row of Γ*. Denote the cardinality of $T(\Gamma^*)$ as $k_n = |T(\Gamma^*)|$. Finally, we define the extended sign matrix of Γ* as

$$S_{\pm}(\Gamma^*_{ij}) := \begin{cases} +1, & \text{if} \quad \Gamma^*_{ij}>0 \\ -1, & \text{if} \quad \Gamma^*_{ij}<0 \\ 0, & \text{if} \quad \Gamma^*_{ij}=0. \end{cases}$$

3. Two-stage Penalized log-Determinant Bregman Divergence Estimation

We propose to develop a two-stage penalized estimation procedure for estimating the regression coefficient matrix Γ and the precision matrix Θ, where in the first stage, we estimate Γ through a penalized joint least square estimation and in the second stage, we estimate Θ by minimizing a penalized log-determinant Bregman divergence after plugging in the regression coefficient estimates. This algorithm can be summarized as the following:

Step 1. Estimate Γ by minimizing a joint penalized residual sum of squares,

$$\hat{\Gamma}=\arg \min \left[\frac{1}{2n}\sum_{i=1}^{n} tr\{(Y_i-\Gamma X_i)(Y_i-\Gamma X_i)^{\top}\}+\rho_n\|\Gamma\|_1\right], \quad (2)$$

where ρ_n is a tuning parameter.

Step 2. Compute

$$\hat{\Sigma}_{\hat{\Gamma}}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{\Gamma}X_i)(Y_i-\hat{\Gamma}X_i)^{\top}. \quad (3)$$

Step 3. Solve the optimization problem,

$$\hat{\Theta}_{\hat{\Gamma}}=\arg\min_{\Theta\succ 0}\{\mathrm{tr}(\hat{\Sigma}_{\hat{\Gamma}}\Theta)-\log \det\Theta+\lambda_n\|\Theta\|_{1,\mathrm{off}}\}, \quad (4)$$

where λ_n is a tuning parameter.

Step 4. Output the solution $(\hat{\Gamma}, \hat{\Theta}_{\hat{\Gamma}})$.

Note that in Step 1, we ignore the correlation among the *Y* variables when estimating the multiple regression coefficients. [9] showed that only when the correlation of the errors is high, incorporation of of such an dependency can lead to increased efficiency in estimating Γ. Theorem 1 in the next section shows that the maximum estimation error is controlled in a certain rate. The estimate $\hat{\Sigma}_{\hat{\Gamma}}$ in Step 2 is a plug-in estimate based on the estimated residuals. This leads to our two-stage estimate of the precision matrix $\hat{\Theta}_{\hat{\Gamma}}$ in Step 3, formulated as the ℓ_1-penalized log-determinant divergence problem [5]. Efficient coordinate descent algorithm can be applied to implement the optimization problems in Step 1 and Step 3 [13]. The tuning parameters can be chosen based on the BIC. The convergent rate in element-wise ℓ_∞ norm of this estimate is established in Theorem 2, followed by rates in other norms.

4. Theoretical properties

4.1. Estimation bound and sign consistency of $\hat{\Gamma}$

Let $T \equiv T(\Gamma^*)$, $T(i)$ and C_X be defined as above and denote I_p as the identity matrix of dimension *p*. In addition, for any matrix *A*, let $A_{S,T}$ be the submatrix with the row indices

given in set S and column indices given in set T. We first make several assumptions on the covariate matrix X.

Assumption 1. *There exists a* $\gamma \in (0, 1]$, *such that*

$$\left\|\left|(C_X \otimes I_P)_{T^c,T}[(C_X \otimes I_P)_{T,T}]^{-1}\right|\right\|_\infty \leq 1 - \gamma. \quad (5)$$

This is the matrix extension of the irrepresentable condition used in ℓ_1 penalized regression setting [12]. This assumption is equivalent to the irrepresentable assumption for Lasso for each of the p components of the response, i.e., $\|\|(C_X)_{T(i)^c,T(i)}[(C_X)_{T(i),T(i)}]^{-1}\|\|_\infty \leq 1 - \gamma$, for $i = 1 \cdots, p$. This can be further written as

$$\sup_i \left\|\left|(C_X)_{T(i)^c,T(i)}[(C_X)_{T(i),T(i)}]^{-1}\right|\right\|_\infty \leq 1 - \gamma.$$

The assumption implies that the number of non-zero elements in each row of Γ should be less than n.

Assumption 2. *There exists a constant* C_{max}, *such that the largest eigenvalue*

$$\lambda_{max}\left([(C_X \otimes I_P)_{T,T}]^{-1}(C_X \otimes \Sigma)_{T,T}[(C_X \otimes I_P)_{T,T}]^{-1}\right) \leq C_{max}. \quad (6)$$

This condition assumes an upper bound on the operator norm of the matrix $[(C_X \otimes I_p)_{T,T}]^{-1}(C_X \otimes \Sigma)_{T,T}[(C_X \otimes I_p)_{T,T}]^{-1}$, which is a combination of the assumptions (26b) and (26c) in [11]. It is easy to check that this assumption holds if

$$\frac{\lambda_{\max}\left((C_X \otimes \Sigma)_{T,T}\right)}{\lambda^2_{\min}\left((C_X \otimes I_p)_{T,T}\right)} \leq C_{\max}.$$

Since $C_X \otimes \Sigma$ is no longer a block diagonal matrix, we cannot obtain an equivalent assumption for each of the p components of the response and then take the supreme over all p components.

Assumption 3. *For all* $n > 0$, *the largest eigenvalue of* C_X *has a common upper bound* $\Lambda_{\max}$, *that is*

$$\lambda_{\max}(C_X) \leq \Lambda_{\max}.$$

This is also commonly used assumption in sparse high dimensional regression analysis ([12] and [11]).

Theorem 1. *Suppose that the design matrix X satisfies the Assumptions* (1) *and* (2) *and X is column-standardized such that*

$$n^{-1/2} \max_{i \in \{1,\cdots,p\}} \max_{j \in \{T(i)\}^c} \|X^{(j)}\|_2 \leq 1. \quad (7)$$

If the sequence of regularization parameters $\{\rho_n\}$ *satisfies*

$$\rho_n > \frac{2}{\gamma}\sqrt{\frac{2\max_i \Sigma^*_{ii}\{\log(p_n)+\log(q_n)\}}{n}}, \quad (8)$$

then for some constant $C_1 > 0$, *the following properties hold with probability greater than* $1 - 4\exp(-C_1 n\rho_n^2) \to 1$,

1. *The minimization of Step 1 of the algorithm has a unique solution* $\hat{\Gamma} \in \mathbb{R}^{p\times q}$ *with its support contained within the true support, i.e.* $T(\hat{\Gamma}) \subseteq T(\Gamma^*)$. *In addition, the element-wise* ℓ_∞ *norm and the Frobenius norm have the following bounds*

$$\|\hat{\Gamma}_T - \Gamma^*_T\|_\infty \le \rho_n \left\{ \left| \left| \left| \{(C_X \otimes I_p)_{T,T}\}^{-1} \right| \right| \right|_\infty + \frac{\gamma}{2}\sqrt{\frac{C_{max}}{\max_i\{\Sigma^*_{ii}\}}} \right\} := \rho_n M_n(X,T,\Sigma^*), \|\hat{\Gamma} - \Gamma^*\|_F \le \sqrt{k_n}\rho_n M_n(X,T,\Sigma^*).$$

2. *If the minimum absolute value of the regression coefficient matrix* Γ^* on its support is bounded below as $|\Gamma^*|_{\min} > \rho_n M_n(X, T, \Sigma^*)$, *then* $\hat{\Gamma}$ *has the correct signed support, i.e.* $S_\pm(\hat{\Gamma}) = S_\pm(\Gamma^*)$.

Theorem 1 is an extension of the results for single regression of [11] to multiple regressions when we simultaneously estimate the regression coefficients of multiple regressions. A lower bound on the minimum absolute value of elements of Γ^* is required for sign consistency. Such an estimation bound on the regression coefficient matrix is required to establish the theoretical property of $\hat{\Theta}_{\hat{\Gamma}}$.

4.2. Estimation bound and sign consistency of $\hat{\Theta}$

We next present results on the estimate of the precision matrix $\hat{\Theta} = \hat{\Theta}_{\hat{\Gamma}}$. Define $\Omega^* = \Theta^{*-1} \otimes \Theta^{*-1}$, which is the Hessian of the log-determinant objective function respect to Θ^* [5]. Since $\Omega^*_{(j,k),(l,m)} = \mathrm{cov}\{\varepsilon_j\varepsilon_k, \varepsilon_l\varepsilon_m\}$, it can be viewed as an edge-based counterpart to the usual covariance matrix Σ^* [5]. Let $S(\Theta^*) = \{E(\Theta^*) \cup \{(1, 1), \cdots, (p, p)\}\}$ be the augmented set including the diagonals. With slight abuse of notation, we also use S and S^c to denote $S(\Theta^*)$ and its complement. We further define

$$K_{\Sigma^*} := \left| \left| \left| \Sigma^* \right| \right| \right|_\infty = (\max_i \sum_{j=1}^{p} |\Sigma^*_{ij}|),$$

as the matrix ℓ_∞ norm of the true covariance matrix Σ^*, and

$$K_{\Omega^*} := \left| \left| \left| (\Omega^*_{SS})^{-1} \right| \right| \right|_\infty = \left| \left| \left| ([\Theta^{*-1} \otimes \Theta^{*-1}]_{SS})^{-1} \right| \right| \right|_\infty.$$

Before we present the theorem on $\hat{\Theta}_{\hat{\Gamma}}$, we need one assumption on the Heissian matrix Ω^*,

Assumption 4. *There exists an* $\alpha \in (0, 1]$, *such that*

$$\left| \left| \left| \Omega^*_{S^cS}(\Omega^*_{SS})^{-1} \right| \right| \right|_\infty \le 1 - \alpha.$$

This assumption is the mutual incoherence or irrepresentable condition introduced in [5], which controls the influence of the non-edge terms on the edge-based terms.

Define $\bar{\delta}_f(n,p_n^\tau) := \sqrt{(\log 4+\tau\log p_n)/(C_* n)}$ for some $\tau > 2$, where $C_*=[128(1+4\sigma^2)^2\max_i\{\Sigma^*_{ii}\}^2]^{-1}$. We then have the following main theorem on the estimation error bound and edge selection.

Theorem 2. *Under the model of Theorem 1 and additional Assumptions* (3) *and* (4), *assume that* ε *is a sub-Gaussian random vector with parameter* σ^2. Let $\hat{\Gamma}$ *be the estimate of* Γ *from Step 1 of two-stage procedure and* $\hat{\Theta_{\hat{\Gamma}}}$ *be the unique solution in Step 3 of the procedure, that is*

$$\hat{\Theta}_{\hat{\Gamma}} := \underset{\Theta\succ 0}{\operatorname{argmin}}\left\{\operatorname{tr}(\Theta\hat{\Sigma}_{\hat{\Gamma}}) - \log\ \det\Theta+\lambda_n\|\Theta\|_{1,\text{off}}\right\},$$

where $\hat{\Sigma_{\hat{\Gamma}}}$ *is defined in* (3). *Suppose that* d_2 *in* Γ^* *satisfies the following upper bound*

$$d_2<\frac{\gamma}{2M_n(X,T,\Sigma^*)\sqrt{\Lambda_{\max}}}\sqrt{\frac{\log q_n}{\log p_n+\log q_n}}\times\left[\sqrt{\{\frac{C^*}{\log q_n}\sqrt{2n(\log 4+\tau\log p_n)}+1\}-1}\right],$$

where $C^*=4(1+4\sigma^2)(1-\sqrt{2/\tau})$, *and tuning parameter* ρ_n *satisfies*

$$\frac{2}{\gamma}\sqrt{\frac{2\max_i\{\Sigma^*_{ii}\}\log(p_n q_n)}{n}}<\rho_n^2<\frac{1-\sqrt{2/\tau}}{C_2M_2(X,T,\Sigma^*)^2\Lambda_{\max}d_2^2}\sqrt{\frac{\log 4+\tau\log p_n}{C_* n}},$$

where C^2 *is some constant. Choosing the regularization parameter*

$$\lambda_n=(8/\alpha)\bar{\delta}_f(n,p_n^\tau).$$

If the sample size exceeds the lower bound

$$n>2(\log 4+\tau\log p_n)\max\left\{C_*^{*2}d_1^2(1+\frac{8}{\alpha})^2,1\right\}, \quad (9)$$

where $C_*^*=48(1+4\sigma^2)\max_i\{\Sigma^*_{ii}\}\max\{K_{\Sigma^*}K_{\Omega^*},K_{\Sigma^*}^3K_{\Omega^*}^2\}$, *then with probability greater than*

$$1-\frac{4}{p_n^{\tau^*-2}}-8\exp(-C_1n\rho_n^2)-2\exp(-C_2n\rho_n^2)\to 1,$$

where

$$\tau^*=\tau\left(1-\frac{C_2M_n(X,T,\Sigma^*)^2\Lambda_{\max}\rho_n^2d_2^2\sqrt{C_*n}}{\sqrt{\log 4+\tau\log p_n}}\right)^2>2,$$

we have:

1. *The estimate* $\hat{\Theta_{\hat{\Gamma}}}$ *satisfies the element-wise* ℓ_∞*-bound:*

$$\|\hat{\Theta}_{\hat{\Gamma}} - \Theta^*\|_\infty \leq \left\{16\sqrt{2}(1+4\sigma^2)(1+8\alpha^{-1})\max_i\{\Sigma^*_{ii}\}K_{\Omega^*}\right\}\sqrt{\frac{\log 4+\tau\log p_n}{n}}.$$

2. *The edge set $E(\hat{\Theta})$ is a subset of the true edge set $E(\Theta^*)$ and includes all edges (i, j) with*

$$|\Theta^*_{ij}| > \{16\sqrt{2}(1+4\sigma^2)(1+8\alpha^{-1})\max_i\{\Sigma^*_{ii}\}K_{\Omega^*}\}\sqrt{(\log 4+\tau\log p_n)/n}.$$

The proof of this theorem is based on the primal-dual witness method used in [5]. The key difference between our approach and that of [5] is the result on controlling the sampling noise. Define $U := \Sigma_{\hat{\Gamma}} - \Sigma^*$, where $\hat{\Sigma}_{\hat{\Gamma}} = \sum_{i=1}^n (Y_i - \hat{\Gamma}X_i)(Y_i - \hat{\Gamma}X_i)^\top/n$. Our proof is mainly on the control of $\|U\|_\infty$. As part of the proof of this theorem, a new result on controlling the sampling noise in our setting is given as Lemma 2 in the Appendix, taking into account that Γ has to be estimated. [5] on the other hand considered the model with zero mean and only has to consider the noise control for $\sum_{i=1}^n Y_i Y_i^\top/n - \Sigma^*$. Theorem 2 indicates that we have the same bound on the element-wise ℓ_∞ norm of the discrepancy between the estimate and the truth as that in [5], but with a slower convergence probability, which is the price we pay for estimating Γ.

Based on the result on of the element-wise ℓ_∞ norm bound, we can get the results on Frobenius and spectral norm bounds. Denote $s_n = |E(\Theta^*)|$ as the total number of off-diagonal non-zeros in Θ^*. We have following corollary:

Corollary 1 (Rates in Frobenius and spectral norm). *Under the same assumptions as Theorem 2, with probability at least* $1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2)$, *the estimator* $\hat{\Theta}_{\hat{\Gamma}}$ *satisfies*

$$\|\hat{\Theta}_{\hat{\Gamma}} - \Theta^*\|_F \leq \{2K_{\Omega^*}(1+\frac{8}{\alpha})\}\sqrt{\frac{(s_n+p_n)(\log 4+\tau\log p_n)}{C_* n}},$$

$$\|\hat{\Theta}_{\hat{\Gamma}} - \Theta^*\|_2 \leq \{2K_{\Omega^*}(1+\frac{8}{\alpha})\}\min(\sqrt{s_n+p_n}, d_1)\sqrt{\frac{\log 4+\tau\log p_n}{C_* n}},$$

where $C_* = [128(1+4\sigma^2)^2\max_i\{\Sigma^*_{ii}\}^2]^{-1}$.

Our final theoretical result is on sign consistency, which requires a lower bound on the minimum value of Θ^*. Define $\theta_{\min} := \min_{(i,j)\in E(\Theta^*)}|\Theta^*_{ij}|$ and the sign recovery event $\mathscr{M}(\hat{\Theta}, \Theta^*) := \{\text{sign}(\hat{\Theta}_{ij}) = \text{sign}(\Theta^*_{ij}), \forall (i, j) \in E(\Theta^*)\}$. We have the following theorem on sign consistency:

Theorem 3. *Under the same conditions as in Theorem 2, suppose that the sample size satisfies the lower bound*

$$n > 2(\log 4+\tau\log p_n)\max\left\{2K^2_{\Omega^*}(1+\frac{8}{\alpha})^2\theta^{-2}_{\min}, C_*{}^2 d_1^2(1+\frac{8}{\alpha})^2, 1\right\},$$

then the estimator is model selection sign consistent with high probability,

$$pr(\mathscr{M}(\hat{\Theta}_{\hat{\Gamma}}, \Theta^*)) \geq 1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2) \to 1.$$

5. Monte Carlo simulations

5.1. Models for comparisons and generation of data

We present results from Monte Carlo simulations to examine the performance of the proposed two-stage estimates. We simulated data to mimic genetical genomic data, where both binary genetic marker data and continuous gene expression data are simulated. We compare our estimate with several other procedures in terms of estimating the precision matrix and neighborhood selection, including the standard Gaussian graphical model implemented as GLASSO [13] using only the gene expression data, a procedure that iteratively updates the regression coefficient matrix and the precision matrix [9, 10] and a neighbor-based graphical model selection procedure of [14], where each gene is regressed on other genes and also the genetic markers using the ℓ_1 regularized regression, and a link is defined between gene *i* and *j* if gene *i* is selected for gene *j* and gene *j* is also selected by gene *i*. Note that in our setting, the neighbor-based procedure does not provide an estimate of the precision matrix. For each simulated data set, we chose the tuning parameters ρ and λ based on the BIC.

To compare the performance of different estimators for the precision matrix, we use the quadratic loss function $\text{LOSS}(\Theta, \hat{\Theta}) = tr(\Theta^{-1}\hat{\Theta} - I)^2$, where $\hat{\Theta}$ is an estimate of the true precision matrix Θ. We also compare $\|\Delta\|_\infty$, $|||\Delta|||_\infty$, $\|\Delta\|_2$ and $\|\Delta\|_F$, where $\Delta = \Theta - \hat{\Theta}$ is the difference between the true precision matrix and its estimate. In order to compare how different methods recover the true graphical structures, we consider the specificity (SPE), sensitivity (SEN) and Matthews correlation coefficient (MCC) scores, which are defined as

$$\text{SPE}=\text{TN}/(\text{TN}+\text{FP}), \text{SEN}=\text{TP}/(\text{TP}+\text{FN}),$$

and

$$\text{MCC}=(\text{TP}\times\text{TN}-\text{FP}\times\text{FN})/\{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})\},$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives in identifying the non-zero elements in the precision matrix. Here we consider the non-zero entry in a sparse precision matrix as “positive.”

In the following simulations, we consider a general sparse precision matrix, where we randomly generate a link (i.e., non-zero elements in the precision matrix, indicated by δ_{ij}) between variables *i* and *j* with a success probability proportional to 1/*p*. Similar to the simulation setup of Li and Gui [15], Fan et al. [16] and Peng et al. [17], for each link, the corresponding entry in the precision matrix is generated uniformly over $[-1, -0.5]\cup[0.5, 1]$. Then for each row, every entry except the diagonal one is divided by the sum of the absolute value of the off-diagonal entries multiplied by 1.5. Finally the matrix is symmetrized and the diagonal entries are fixed at 1. To generate the $p \times q$ coefficient matrix $\Gamma = (\gamma_{ij})$, we first generated a $p \times q$ sparse indicator matrix (δ_{ij}), where $\delta_{ij} = 1$ with a probability proportional to 1/*q*. If $\delta_{ij} = 1$, we generated γ_{ij} from Unif $([\upsilon_m, 1] \cup [-1, -\upsilon_m])$, where υ_m is the minimum absolute non-zero value of Θ generated.

After Γ and Θ were generated, we generated the marker genotypes $X = (X_1, \cdots, X_q)$ by assuming $X_i \sim$ Bernoulli(1, 1/2), for $i = 1, \cdots, q$. Finally, given X, we generated Y the multivariate normal distribution $Y|X \sim \mathcal{N}\Gamma X, \Sigma)$. For a given model and a given simulation, we generated a data set of n independent and identically distributed random vectors (X, Y). The simulations were repeated 50 times.

5.2. Simulation results

We first consider the setting when the sample size n is larger than the number of genes p and the number of genetic markers q. We simulated data from three models with different values of p, q (See Table 1 Model 1 – Model 3) and present the simulation results in Table 2. We observe that the two-stage procedure performs very similarly to the iterative procedure. Clearly, the two-stage procedure and the iterative procedure provide much improved estimates of the precision matrix over the Gaussian graphical model for all three models considered in all measurements. This is expected since the Gaussian graphical model assumes a constant mean of the multivariate vector, which is a misspecified model. In addition, the two-stage procedure resulted in higher sensitivities, specificities and MCC than the Gaussian graphical model and the neighbor-based method. We observed that the Gaussian graphical model often resulted in much denser graphs than the real graphs. This is partially due to the fact that some of the links identified by Gaussian graphical model can be explained by shared common genetic variants. By assuming constant means, in order to compensate for the model misspecification, the Gaussian graphical tends to identify many non-zero elements in the precision matrix. The results indicate that by adjusting the effects of the covariates on the means, we can reduce both false positives and false negatives in identifying the non-zero elements of the precision matrix. The neighbor-based selection procedure using multiple LASSO accounts for the genetic effects in modeling the relationship among the genes. It performed better than the Gaussian graphical in graph structure selection, but worse than the two-stage procedure. This procedure, however, did not provide an estimate of the precision matrix.

We next consider the setting when $p > n$ and simulated data from three models with different values of n, p and q (see Table 1 Model 4 – Model 6). Note that for all three models, the graph structure is very sparse due to the large number of genes considered. The performances over 50 replications are reported in Table 2 for the optimal tuning parameters chosen by the BIC. For all three models, we observed much improved estimates of the precision matrix from the proposed two-stage procedure as reflected by smaller norms of the difference between the true and estimated precision matrices. In terms of graph structure selection, in general, we observe that when p is larger than the sample size, the sensitivities from all four procedures are much lower than the settings when the sample size is larger. This indicates that recovering the graph structure in a high-dimensional setting is statistically difficult. However, the specificities are in general very high, agreeing with our theoretical result of the estimates.

Finally, Table 3 presents the comparison of the estimates of Γ of three different procedures. Overall, we observe no differences in estimates of Γ from the two-stage and the iterative procedures, both perform better than the neighbor-based procedure.

6. Real data analysis

To demonstrate the proposed method, we present results from the analysis of a data set generated by [18], where 112 yeast segregants, one from each tetrad, were grown from a cross involving parental strains BY4716 and wild isolate RM11-1A and gene expression levels of 6,216 genes were measured. These 112 segregants were individually genotyped at 2,956 marker positions throughout the genome. Since many of these markers are in high

linkage disequilibrium, we combined the markers into 585 blocks where the markers within a block differed by at most one sample. For each block, we chose the marker that had the least number of missing values as the representative marker.

To demonstrate our methods, we focused our analysis on a set of genes of the protein-protein interaction (PPI) network obtained from a previously compiled set by [19] combined with protein physical interactions deposited in Munich information center for protein sequences. We further selected 1,207 genes with variance greater than 0.05. Based on the most recent yeast protein-protein interaction database BioGRID [20], there are a total of 7,619 links among these 1,207 genes. Our goal is to construct a conditional independent network among these genes based on the sparse Gaussian graphical model adjusting for possible genetic effects on gene expression levels.

Results from several different procedures are summarized in Table 4. We observe that the neighbor-based method resulted in sparsest graph and the standard Gaussian graphical model without adjusting for the effects of genetic markers resulted in the densest graph, and the two-stage procedure was in between. A summary of the degrees of the graphs estimated by these three procedures is given in Table 4. We observe that the standard Gaussian graphical model gave a much denser graph than the other two procedures, agreeing with what we observed in simulation studies. The Gaussian graphical selected a lot more links than the other two methods, among the links that were identified by the Gaussian graphical model only, 476 pairs are associated with at least one common genetic marker based on the two-stage procedure, further explaining that some of the links identified by gene expression data alone can be due to shared comment genetic variants. The neighbor-based selection procedure identified only 1,917 edges, out of which 1880 were identified by the two-stage procedure and 1,916 were identified by the graphical model. There was a common set of 1749 links that were identified by all three procedures.

If we treat the PPI of the BioGRID database as the true network among these genes, the true positive rate from the two-stage procedure, the Gaussian graphical model and the neighbor-based selection procedure was 0.068, 0.071 and 0.019, respectively, and the false positive rate was 0.018, 0.026 and 0.0025, respectively. The reason for having low true positive rates is that many of the protein-protein interactions cannot be reflected at the gene expression level. Figure 1 (a) shows the histogram of the correlations of genes that are linked on the BioGRID PPI network, indicating that many linked gene pairs have very small marginal correlations. The Gaussian graphical models are not able to recover these links. Figure 1 plots (b) – (d) show the marginal correlations of the gene pairs that were identified by the two-stage procedure, the Gaussian graphical model and the neighbor-based procedure, clearly indicating that the linked genes identified by the two-stage procedure have higher marginal correlations. In contrast, some linked genes identified by the Gaussian graphical model have quite small marginal correlations.

7. Discussion

The proposed two-stage procedure is computationally efficient through coordinate descent algorithm and can be applied to high dimensional settings. Our simulation results show that this two-stage procedure performs very similarly to the iterative procedure of [9, 10]. To ensure model selection consistency and to derive the estimation bounds, our main theoretical assumption is an irrepresentable or mutual incoherence condition on both the covariates matrix and the true precision matrix. These conditions are similar to those required for model selection consistency of the LASSO or precision matrix estimation. Compared to the asymptotic results in [10], the results in this paper provide more explicit bounds in different matrix norms and present conditions for correct sign support. Our theoretical results on the

estimate of the precision matrix parallel to those in [5]. However, the proofs are more difficult since the estimation biases of the regression coefficients have to be accounted for when studying the properties of the estimate of the precision matrix. This is achieved by proving an important lemma on control of sampling noise.

Partially due to computational consideration, we used the ℓ_1-penalization to obtain sparse results for both regression coefficient matrix and the precision matrix. However, other non-convex penalty functions can be applied in our two-stage algorithm, although computationally the optimizations are more challenging. Alternatively, one can extend the Dantzig selector [21] to estimate the regression coefficient matrix and the constrained ℓ_1 minimization [22] to estimate the precision matrix. It would be interesting to compare the performances of these different approaches. Finally, we can also consider to impose low-rank sparsity in stage 1 of the estimation using a penalty proportional to the rank of Γ [23]. This approach yields a closed form solution and different rates of convergence. It is interesting to compare these alternatives with the proposed approach in this paper.

Acknowledgments

This research is supported by NIH grants R01CA127334 and R01GM097505 and National Natural Science Foundation of China (grant No. 11201479).

Appendix

We present the proofs of the theorems in this Appendix. Proof of Theorem 1 extends that of [11] to multiple regressions and coefficient matrix settings. The key of the proof of Theorem 2 is a lemma on control of sampling noise, which we present detailed proof. Using this lemma, the proof of Theorem 2 is mainly based on the technique of primal-dual witness method [5].

Proof of Theorem 1.

From equation (2) and the model $Y_i = \Gamma^* X_i + \varepsilon_i$, the estimation equation becomes

$$(\Gamma - \Gamma^*)C_X - \frac{1}{n}W^\top X + \rho_n B = 0,$$

where B is the sub-differential of $\|\Gamma\|_1$, defined as $B_{ij} = \text{sign}(\Gamma_{ij})$ if $\Gamma_{ij} \neq 0$ and $\in [-1, 1]$, if $\Gamma_{ij} = 0$. With these definitions, we have the following lemma

Lemma 1. *(a) A matrix $\hat{\Gamma} \in \mathbb{R}^{p\times q}$ is optimal to the ℓ_1 penalization problem* (2) *if and only if there exists an element $\hat{B}$ of the sub-differential $\partial\|\Gamma\|_1$ such that*

$$(\hat{\Gamma} - \Gamma^*)C_X - \frac{1}{n}W^\top X + \rho_n \hat{B} = 0.$$

(b) Suppose that the sub-different matrix satisfies the strict dual feasibility condition $|\hat{B}_{ij}| < 1$ for all $(i, j) \notin T(\hat{\Gamma})$. Then any optimal solutione $\tilde{\Gamma}$ to the ℓ_1 penalization problem (2) *satisfiese $\tilde{\Gamma}_{ij} = 0$ for all $(i, j) \notin T(\hat{\Gamma})$.*

(c) Under the condition of part (b), if the $|T(\hat{\Gamma})|\times|T(\hat{\Gamma})|$ matrix $(C_X \otimes I_p)_{T(\hat{\Gamma}),T(\hat{\Gamma})}$ is invertible, then $\hat{\Gamma}$ is the unique optimal solution of the ℓ_1 penalization problem (2).

Similar technique as in [11] can be used to prove this Lemma. From Lemma 1, we know that strict dual feasibility conditions are sufficient to ensure the uniqueness of $\hat{\Gamma}$. We construct the primal-dual witness solution $(\tilde{\Gamma}, \tilde{B})$ as follows:

a. First, we determine the matrix $\tilde{\Gamma}$ by solving the restricted LASSO problem

$$\tilde{\Gamma}=\arg\min_{\Gamma_{T^c}=0}\{\frac{1}{2n}\sum_{i=1}^{n}\mathrm{tr}\{(Y_i-\Gamma X_i)(Y_i-\Gamma X_i)^\top\}+\rho_n\|\Gamma\|_1\}. \quad (.1)$$

b. Second, we choose $\tilde{B}_T$ as an element of the sub-differential of the regularizer $\|\cdot\|_1$, evaluated at $\tilde{\Gamma}$.

c. Third, we set e $\tilde{B}_{T^c}$ to satisfy the zero sub-differential condition (.1), and check whether or not the dual feasibility condition $\tilde{B}_{ij} \le 1$ for all $(i, j) \in T^c$ is satisfied. To ensure the uniqueness, we check for strict dual feasibility $\tilde{B}_{ij} < 1$ for all $(i, j) \in T^c$.

d. Fourth, we check whether the sign consistency condition $\tilde{\mathrm{B}}_T=\mathrm{sign}(\Gamma^*_T)$ is satisfied.

PROOF OF THEOREM 1. From the primal-dual witness construction, denote $\Lambda = \tilde{\Gamma} - \Gamma^*$, where $\tilde{\Gamma}$ is the solution to (.1) and Γ^* is the true parameter. The equation (.1) can be rewritten as:

$$(C_X\otimes I_p)_{T,T}\bar{\Lambda}_T-\frac{1}{n}(X^\top\otimes I_p)_{T,\cdot}\overline{W^\top}+\rho_n\overline{\tilde{B}}_T=0, \quad (.2)$$

$$(C_X\otimes I_p)_{T^c,T}\bar{\Lambda}_T-\frac{1}{n}(X^\top\otimes I_p)_{T^c,\cdot}\overline{W^\top}+\rho_n\overline{\tilde{B}}_{T^c}=0. \quad (.3)$$

Since $\bar{\Lambda}_{T^c} = 0$, in order to establish strict dual feasibility, we need to check whether $\|\overline{\tilde{B}}_{T^c}\|_\infty<1$. From (.2), we have

$$\bar{\Lambda}_T=[(C_X\otimes I_p)_{T,T}]^{-1}[\frac{1}{n}(X^\top\otimes I_p)_{T,\cdot}\overline{W^\top}-\rho_n\overline{\tilde{B}}_T],$$

substituting this into (.3) leads to

$$\begin{aligned}\overline{\tilde{B}}_{T^c}=&-\frac{1}{\rho_n}(C_X\otimes I_p)_{T^c,T}[(C_X\otimes I_p)_{T,T}]^{-1}[\frac{1}{n}(X^\top\times I_p)_{T,\cdot}\overline{W^\top}-\rho_n\overline{\tilde{B}}_T]\\&+(C_X\otimes I_p)_{T^c,T}[(C_X\otimes I_p)_{T,T}]^{-1}\overline{\tilde{B}}_T\\&\cdot=\frac{1}{n\rho_n}(X^\top\otimes I_p)_{T^c,\cdot}\left\{I_{np}-(\frac{1}{n}X\otimes I_p)_{\cdot,T}[(C_X\otimes I_p)_{T,T}]^{-1}(X^\top\otimes I_p)_{T,\cdot}\right\}\\&\times\overline{W^\top}+(C_X\otimes I_p)_{T^c,T}[(C_X\otimes I_p)_{T,T}]^{-1}\overline{\tilde{B}}_T=(\mathrm{I})+(\mathrm{II}).\end{aligned} \quad (.4)$$

For the second term (II) of (.4), from the Assumption (1) and $\|\overline{\tilde{B}}_{T^c}\|_\infty \le 1$, we have

$$\|(C_X\otimes I_p)_{T^c,T}[(C_X\otimes I_p)_{T,T}]^{-1}\overline{\tilde{B}}_T\|_\infty<1-\gamma.$$

From the sub-Gaussian (sG for short) distribution assumption on ε, $\overline{W^\top}\sim sG(0, I_n\otimes\Sigma^*)$. Denote the projection matrix as

$$\Pi_{(C_X \otimes I_p)_{T,T}} := I_{np} - (\frac{1}{n} X \otimes I_p)_{\cdot,T} [(C_X \otimes I_p)_{T,T}]^{-1} (X^\top \otimes I_p)_{T,\cdot} := I_{np} - A.$$

Choosing a particular element (*j*, *i*) in the first term (I) of (.4),

$$\frac{1}{n\rho_n}(X^{(j)T} \otimes \mathbf{e}_i^\top)(I_{np} - A)\overline{W^\top}$$

with $\overline{W^\top} \sim sG(0, I_n \otimes \Sigma^*)$ and $\mathbf{e}_i$ is *i*th row of identity matrix I_p, then

$$\frac{1}{n\rho_n}(X^{(j)T} \otimes \mathbf{e}_i^\top)(I_{np} - A)\overline{W^\top} \sim sG(0, \sigma^2_{(j,i)}),$$

where using the fact that $I_{np} - A$ is a projection matrix and the condition (7) in the theorem, we have

$$\begin{aligned}\sigma^2_{(j,i)} = & \frac{1}{n^2\rho_n^2}(X^{(j)T} \otimes \mathbf{e}_i^\top)(I_{np} \\ & - A)(I_n \otimes \Sigma^*)(I_{np} \\ & - A)(X^{(j)} \otimes \mathbf{e}_i) \le \frac{1}{n^2\rho_n^2}(X^{(j)T}X^{(j)}) \otimes (\mathbf{e}_i^\top \Sigma^* \mathbf{e}_i) \le \frac{1}{n\rho_n^2}\Sigma^*_{ii} \le \frac{1}{n\rho_n^2}\max_i(\Sigma^*_{ii}).\end{aligned}$$

By applying the Chernoff bound, we have,

$$\mathrm{pr}(\max_i \max_{j \in (T(i))^c} |(\mathrm{I})_{(j,i)}| \ge t) \le 2(p_n q_n - k_n)\exp\{-\frac{n\rho_n^2 t^2}{2\max_i(\Sigma^*_{ii})}\},$$

where $(\mathrm{I})_{(j,i)}$ is the (*j*, *i*)th element in the first term (I) in (.4) and k_n is the number of nonzero elements in the true parameter Γ*. Setting $t = \gamma/2$ yields

$$\mathrm{pr}(\max_i \max_{j \in (T(i))^c} |(\mathrm{I})_{(j,i)}| \ge \frac{\gamma}{2}) \le 2\exp\{\frac{n\rho_n^2\gamma^2}{8\max_i(\Sigma^*_{ii})} + \log(p_n q_n - k_n)\}.$$

Putting together the pieces and using our choice (8) of ρ_n, we have

$$\mathrm{pr}(\|\overline{\tilde{B}}_{T^c}\|_\infty > 1 - \frac{\gamma}{2}) \le 2\exp(-c_1 n\rho_n^2) \to 0,$$

for some constant c_1. So from Lemma 1, the estimated support $T(\hat{\Gamma})$ is contained in the support T hence in the true support $T^*(\Gamma^*)$ with probability at least $1 - 2\exp(-c_1 n\rho_n^2)$.

Next we establish the ℓ_∞ bounds, from (.4) we know

$$\overline{\Lambda^*}_T := (\hat{\Gamma} - \Gamma^*)_T = [(C_X \otimes I_p)_{T,T}]^{-1}[\frac{1}{n}(X^\top \otimes I_p)_{T,\cdot}\,\overline{W^\top} - \rho_n \overline{\hat{B}}_T],$$

where $\hat{B_T}$ is in the sub-differential of $\|\hat{\Gamma}\|_1$. So

$$\|\overline{\Lambda^*}_T\|_\infty \le \|[(C_X \otimes I_p)_{T,T}]^{-1}(X^\top \otimes I_p)_{T,\cdot}\frac{1}{n}\overline{W^\top}\|_\infty + \rho_n \|\,|[(C_X \otimes I_p)_{T,T}]^{-1}|\,\|_\infty$$

Note that the second term in (.4) is a fixed term. Since $\overline{W^\top} \sim sG(0, I_n \otimes \Sigma^*)$, then

$$[(C_X \otimes I_p)_{T,T}]^{-1}(X^\top \otimes I_p)_{T,\cdot}\frac{1}{n}\overline{W^\top} \sim sG(0, \Omega),$$

where

$$\begin{aligned}\Omega &= \frac{1}{n^2}[(C_X \otimes I_p)_{T,T}]^{-1}(X^\top \otimes I_p)_{T,\cdot}(I_n \otimes \Sigma^*)(X \otimes I_p)_{\cdot,T}[(C_X \otimes I_p)_{T,T}]^{-1} \\ &= \frac{1}{n}[(C_X \otimes I_p)_{T,T}]^{-1}(C_X \otimes \Sigma^*)_{T,T}[(C_X \otimes I_p)_{T,T}]^{-1}.\end{aligned}$$

Define the first term in (.4) as ξ, then the (j, i)-th element of ξ where $j \in T(i)$ is distributed as sub-Gaussian with parameter σ^2, that is $\xi_{(j,i)} \sim sG(0, \sigma^2)$, with $\sigma^2 \le 1/nC_{\max}$, where $C_{\max}$ is defined in the Assumption 2. Again from Chernoff bound,

$$\mathrm{pr}(\max_i \max_{j \in T(i)} |\xi_{(j,i)}| > t) \le 2\exp(-\frac{nt^2}{2C_{\max}} + \log k_n).$$

Setting $t = \rho_n \gamma/2\sqrt{C_{\max}/\max_i\{\Sigma^*_{ii}\}}$, then $nt^2/2C_{\max}) = n\rho_n^2\gamma^2/(8\max_i\{\Sigma^*_{ii}\})$. Since ρ_n satisfies (8), $n\rho_n^2\gamma^2/(8\max_i\{\Sigma^*_{ii}\}) > \log(p_n q_n) > \log(k_n)$. So $\mathrm{pr}(\max_i \max_{j \in T(i)} |\xi_{(j,i)}| > \rho_n\gamma/2\sqrt{C_{\max}/\max_i\{\Sigma^*_{ii}\}})$ vanishes at the rate at least $2\exp(-c_2 n\rho_n^2)$, where c_2 is a constant. Overall, we conclude that

$$\|\hat{\Gamma} - \Gamma^*\|_\infty \le \rho_n \left\{\frac{\gamma}{2}\sqrt{\frac{C_{\max}}{\max_i\{\Sigma^*_{ii}\}}} + \|\,|[(C_X \otimes I_p)_{T,T}]^{-1}|\,\|_\infty\right\}$$

with probability greater than $1 - 4\exp(-C_1 n\rho_n^2)$ where C_1 is a constant (for example C_1 can be chosen as $\min\{c_1, c_2\}$). Thus assertion (1) in Theorem 1 is proved and the (2) directly follows when (1) is proved. Thus complete the proof of Theorem 1.

Proof of Theorem 2:

We define

$$M_n(X,T,\Sigma^*)=\left|\left|\left|\left[(C_X\otimes I_p)_{T,T}\right]^{-1}\right|\right|\right|_\infty+\frac{\gamma}{2}\sqrt{\frac{C_{\max}}{\max_i\{\Sigma^*_{ii}\}}},$$

and $C_{\max}$ is the constant in Assumption 2. Define $U:=\hat{\Sigma}_{\hat{\Gamma}}-\Sigma^*$, where $\hat{\Sigma}_{\hat{\Gamma}}=\sum_{i=1}^n(Y_i-\hat{\Gamma}X_i)(Y_i-\hat{\Gamma}X_i)^\top/n$. Our proof is mainly on the control of $\|U\|_\infty$, which is the major difference between our Theorem 2 and Theorem 1 in [5]. We state this noise control result in the following lemma:

Lemma 2 (Control of Sampling Noise). *Under the assumptions that log p_n = $o(n)$, $\log q_n=o(\sqrt{n})$, d_2 = $o(q_n)$ and furthermore, for some real number $\tau>2$,*

$$d_2<\frac{\gamma}{2M_n(X,T,\Sigma^*)\sqrt{\Lambda_{\max}}}\sqrt{\frac{\log q_n}{\log p_n+\log q_n}}\times\left\{\sqrt{\left[\frac{C^*}{\log q_n}\sqrt{2n(\log 4+\tau\log p_n)}+1\right]}-1\right\},$$

where $C^*=4(1+4\sigma^2)(1-\sqrt{2/\tau})$ and $\Lambda_{\max}$ is the constant in Assumption 3 and σ is the parameter in the tail condition on ε_i. Choose a constant $C_2>1$, such that

$$C_2>1+\frac{\gamma}{M_n(X,T,\Sigma^*)\sqrt{\Lambda_{\max}}d_2}\sqrt{\frac{\log q_n}{\log p_n+\log q_n}}. \quad (.5)$$

Assume the conditions in Theorem 1 are satisfied and in addition to the tuning parameter ρ_n satisfying condition (8), *ρ_n also satisfies*

$$\rho_n^2<\frac{1-\sqrt{2/\tau}}{C_2M_n(X,T,\Sigma^*)^2\Lambda_{\max}d_2^2}\sqrt{\frac{\log 4+\tau\log p_n}{C_*n}},$$

where $C_*=\left[128(1+4\sigma^2)^2\max_i\{\Sigma^*_{ii}\}^2\right]^{-1}$. Under this condition, denote

$$\tau^*=\tau\left[1-\frac{C_2M_n(X,T,\Sigma^*)^2\Lambda_{\max}\rho_n^2d_2^2\sqrt{C_*n}}{\sqrt{\log 4+\tau\log p_n}}\right]^2>2, \quad (.6)$$

then

$$pr\left(\|U\|_\infty\geq\sqrt{\frac{\log 4+\tau\log p_n}{C_*n}}\right)\leq\frac{4}{p_n^{\tau^*-2}}+8\exp\{-C_1n\rho_n^2\}+2\exp\{-C_2n\rho_n^2\},$$

where C_1 and C_2 are constants.

PROOF OF LEMMA 2. From the definition of U, we have

$$U=\frac{1}{n}W^\top W-(\Theta^*)^{-1}+(\hat{\Gamma}-\Gamma^*)C_X(\hat{\Gamma}-\Gamma^*)^\top-(\hat{\Gamma}-\Gamma^*)(\frac{1}{n}X^\top W)-(\frac{1}{n}W^\top X)(\hat{\Gamma}-\Gamma^*)^\top.$$

We want to bound the element-wise ℓ_∞ norm $\|U\|_\infty$,

$$\begin{aligned}\|U\|_\infty \le & \|\frac{1}{n}W^\top W-(\Theta^*)^{-1}\|_\infty \\ & +\|(\hat\Gamma-\Gamma^*)C_X(\hat\Gamma-\Gamma^*)^\top\|_\infty \\ & +\|(\hat\Gamma-\Gamma^*)(\frac{1}{n}X^\top W)\|_\infty \\ & +\|(\frac{1}{n}W^\top X)(\hat\Gamma-\Gamma^*)^\top\|_\infty . \le \|\frac{1}{n}W^\top W-(\Theta^*)^{-1}\|_\infty \\ & +|||(\hat\Gamma-\Gamma^*)\otimes(\hat\Gamma-\Gamma^*)|||_\infty\|C_X\|_\infty+2\|(\hat\Gamma-\Gamma^*)(\frac{1}{n}X^\top W)\|_\infty .\end{aligned}$$

Let $\mathscr{A}$ be the event that $T(\hat\Gamma)\subseteq T(\Gamma^*)$ and $\|(\hat\Gamma_T-\Gamma^*_T)\|_\infty \le \rho_n M_n(X,T,\Sigma^*)$. Then $\mathrm{pr}(\mathscr{A})\ge 1-4\exp(-C_1 n\rho_n^2)$. Under event $\mathscr{A}$, $\overline{(\hat\Gamma-\Gamma^*)(X^\top W/n)}=\{(W^\top X/n)\otimes I_p\}_{\cdot,T}\overline{(\hat\Gamma-\Gamma^*)}_T$. So

$$\|(\hat\Gamma-\Gamma^*)(\frac{1}{n}X^\top W)\|_\infty=\|[(\frac{1}{n}W^\top X)\otimes I_p]_{\cdot,T}\overline{(\hat\Gamma-\Gamma^*)}_T\|_\infty \le |||[(\frac{1}{n}W^\top X)\otimes I_p]_{\cdot,T}|||_\infty\|(\hat\Gamma-\Gamma^*)_T\|_\infty .$$

So under event $\mathscr{A}$

$$\|U\|_\infty \le \mathrm{I}+\mathrm{II}+\mathrm{III},$$

where

$$\mathrm{I}=\|\frac{1}{n}W^\top W-(\Theta^*)^{-1}\|,\ \mathrm{II}=|||(\hat\Gamma-\Gamma^*)\otimes(\hat\Gamma-\Gamma^*)|||_\infty\|C_X\|_\infty,\ \mathrm{III}=2|||[(\frac{1}{n}W^\top X)\otimes I_p]_{\cdot,T}|||_\infty\|(\hat\Gamma-\Gamma^*)_T\|_\infty .$$

From Lemma 1 of [5] on sub-Gaussian tail condition, we have

$$\mathrm{pr}(\mathrm{I}>\delta)\le 4p^2\exp\{-\frac{n\delta^2}{128(1+4\sigma^2)^2\max_i\{\Sigma^*_{ii}\}^2}\}. \quad (7)$$

Since $\lambda_{\max}(C_X)\le\Lambda_{\max}$, then $\|C_X\|_\infty\le\Lambda_{\max}|||\hat\Gamma-\Gamma^*|||^2_\infty$ because of $|||A\otimes B|||_\infty\le|||A|||_\infty |||B|||_\infty$. Then $\mathrm{II}\le\Lambda_{\max}(d_2\|\Gamma-\Gamma^*\|_\infty)^2$. Under event $\mathscr{A}$ $T(\hat\Gamma\subseteq T(\Gamma^*))$ and $\|\hat\Gamma-\Gamma^*\|_\infty=\|\hat\Gamma_T-\Gamma^*_T\|_\infty\le M_n(X,T,\Sigma^*)\rho_n$. So

$$\mathrm{pr}(\mathrm{II}>\Lambda_{\max}d_2^2g^2(\rho_n))\le 1-\mathrm{pr}(\mathscr{A})\le 4\exp\{-C_1n\rho_n^2\}.$$

Next we bound (III).We know under event $\mathscr{A}$ $\|(\hat\Gamma-\Gamma^*)^T\|_\infty\le\rho_n M_n(X,T,\Sigma^*)$. We need further bound each row's ℓ_1 norm in $[(W^\top X/n)\otimes I_p]_{\cdot,T}$. Since $\overline{W^\top X/n}$ is a $p_nq_n\times 1$ random vector with mean zero and covariance matrix $C_X\otimes\Sigma^*/n$, for certain index (i,j), the (i,j)-th row in $[(W^\top X/n)\otimes I_p]_{\cdot,T}$ is $[e_i^\top(W^\top X/n)]\otimes e_j^\top$, and e_i, $e_j\in\mathbb{R}^p$ are the simple base functions for $i,j=1,\cdots,p$. Since $\overline{e_i^\top(W^\top X/n)}=(I_q\otimes e_i^\top)\overline{(W^\top X/n)}$, $\overline{e_i^\top(W^\top X/n)}$ is with mean zero and covariance matrix

$$(I_q \otimes e_i^\top)(\frac{1}{n}C_X \otimes \Sigma^*)(I_q \otimes e_i)=\frac{1}{n}C_X \otimes \Sigma^*_{ii}=\frac{1}{n}\Sigma^*_{ii}C_X.$$

The non-zero elements in $[e_i^\top(W^\top X/n)] \otimes e_j^\top$ is $[e_i^\top(W^\top X/n)]_{T(j)} \otimes 1$, so $\overline{[e_i^\top(W^\top X/n)]_{T(j)}}$ is mean zero with covariance matrix $\frac{1}{n}\Sigma^*_{ii}(C_X)_{T(j),T(j)}$, and $|||[(W^\top X/n) \otimes I_p]_{\cdot,T}|||_\infty$ equals the maximum value for all (i, j) pair, the ℓ_1 norm of $\overline{[e_i^\top(W^\top X/n)]_{T(j)}}$. Obviously variables in vector $\overline{[e_i^\top(W^\top X/n)]_{T(j)}}$ are sub-Gaussian. In next lemma, we bound the ℓ_1 norm of such type of sub-Gaussian vectors.

Lemma 3. *For any $j \in \{1, \cdots, p\}$, let $T(j)$ be defined as before. Suppose that $|T(j)| \geq 1$. If $y \in \mathbb{R}^{|T(j)|}$ is a random vector with mean zero and covariance matrix $\Sigma^*_{ii}(C_X)_{T(j),T(j)}/n$, and every variable in y is sub-Gaussian. Then*

$$pr(\|y\|_1>t) \leq 2|T(j)|\exp\{-\frac{nt^2}{2|T(j)|^2\max_i\{\Sigma^*_{ii}\}\Lambda_{\max}}\}.$$

PROOF OF LEMMA 3: First we have:

$$\mathrm{pr}(\|y\|_1>t)=\mathrm{pr}\left(|y_1|+\cdots+|y_{|T(j)|}|>t\right) \leq \mathrm{pr}\left(|y_1|>\frac{t}{|T(j)|}\right)+\cdots+\mathrm{pr}\left(|y_{|T(j)|}|>\frac{t}{|T(j)|}\right).$$

Note that y_k is sub-Gaussian with parameter $\Sigma^*_{ii}(C_X)_{kk}/n \leq \max_i\{\Sigma^*_{ii}\}\Lambda_{\max}/n \forall k \in \{1, \cdots, |T(j)|\}$ and $i \in \{1, \cdots, |T(j)|\}$. From Chernoff bound,

$$\mathrm{pr}(\|y\|_1>t) \leq 2|T(j)|\exp\{-\frac{nt^2}{2|T(j)|^2\max_i\{\Sigma^*_{ii}\}\Lambda_{\max}}\},$$

which completes the proof.

Since $f(x) = x\exp\{-a/x^2\}$ for some $a > 0$ is an increasing function of x and $\forall_j \in \{1, \cdots, p\}$, $|T(j)| \leq d_2$, we have

$$\mathrm{pr}\left(|||\left[(\frac{1}{n}W^\top X) \otimes I_p\right]_{\cdot,T}|||_\infty>t\right) \leq 2d_2\exp\{-\frac{nt^2}{2d_2^2\max_i\{\Sigma^*_{ii}\}\Lambda_{\max}}\}. \quad (.8)$$

If we choose $t^2=\frac{1}{4}\Lambda_{\max}d_2^2\rho_n^2\gamma^2\log q_n/(\log p_n+\log q_n)$, we have

$$\frac{nt^2}{2d_2^2\max_i\{\Sigma^*_{ii}\}\Lambda_{\max}}=\frac{n\rho_n^2\gamma^2}{8\max_i\{\Sigma^*_{ii}\}}\frac{\log q_n}{\log p_n+\log q_n}.$$

From the choice of ρ_n in (8), we can see

$$n\rho_n^2\gamma^2/[8\max_i\{\Sigma_{ii}^*\}(\log p_n+\log q_n)]>1,$$

so

$$\frac{nt^2}{2d_2^2\max_i\{\Sigma_{ii}^*\}\Lambda_{\max}}>\log q_n$$

and from the condition $d_2 = o(q_n)$, we know in (.8), the exponential part dominates and converges to zero at some exponential rate. On the other hand the term on the exponential shoulder is bounded by $n\rho_n^2\gamma^2/[8\max_i\{\Sigma_{ii}^*\}]=C_2n\rho_n^2$ for some constant $C_2>0(C_2=\gamma^2/[8\max_i\{\Sigma_{ii}^*\}])$. Denote for any event B, $\mathrm{pr}_{\mathscr{A}}^*(B)=\mathrm{pr}(B\cap\mathscr{A})$, then

$$\mathrm{pr}_{\mathscr{A}}\left(|||[(\frac{1}{n}W^\top X)\otimes I_p]_{\cdot,T}|||_\infty>\frac{1}{2}\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\gamma d_2\rho_n\right)\le 2\exp\{-C_2n\rho_n^2\}.$$

So

$$\mathrm{pr}\left(\mathrm{III}>\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\gamma d_2\rho_n^2M_n(X,T,\Sigma^*)\right)\le \mathrm{pr}_{\mathscr{A}}\left(\mathrm{III}>\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\gamma d_2\rho_n^2M_n(X,T,\Sigma^*)\right)$$
$$+\mathrm{pr}(\mathscr{A}^c)\le \mathrm{pr}_{\mathscr{A}}\left(|||[(\frac{1}{n}W^\top X)\otimes I_p]_{\cdot,T}|||_\infty>\frac{1}{2}\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\gamma d_2\rho_n\right)+\mathrm{pr}(\mathscr{A}^c)$$

That is

$$\mathrm{pr}\left(\mathrm{III}>\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\gamma d_2\rho_n^2M_n(X,T,\Sigma^*)\right)\le 2\exp\{-C_2n\rho_n^2\}+4\exp\{-C_1n\rho_n^2\},\quad (.9)$$

where C_2 is defined above and C_1 is defined in Theorem 1.

Denote

$$\bar{\delta}_f(n,p_n^\tau)=\sqrt{(\log 4+\tau\log p_n)/(C_*n)},$$

where $C_*=[128(1+4\sigma^2)^2\max_i\{\Sigma_{ii}^*\}^2]^{-1}$. Define

$$\alpha=\sqrt{\frac{\Lambda_{\max}\log q_n}{\log p_n+\log q_n}}\frac{\gamma d_2\rho_n^2M_n(X,T,\Sigma^*)}{\bar{\delta}_f(n,p_n^\tau)}$$

and

$$\beta=\frac{\Lambda_{\max}d_2^2M_n(X,T,\Sigma^*)^2\rho_n^2}{\bar{\delta}_f(n,p_n^\tau)}.$$

So

$$\alpha+\beta=\frac{M_n(X,T,\Sigma^*)^2\Lambda_{\max}\rho_n^2d_2^2}{\bar{\delta}_f(n,p_n^\tau)}$$
$$\times\left(1+\frac{\gamma}{M_n(X,T,\Sigma^*)\sqrt{\Lambda_{\max}d_2}}\sqrt{\frac{\log q_n}{\log p_n+\log q_n}}\right)<\frac{C_2M_n(X,T,\Sigma^*)^2\Lambda_{\max}\rho_n^2d_2^2}{\bar{\delta}_f(n,p_n^\tau)}<1-\sqrt{\frac{2}{\tau}},$$

from the choice of C_2 in (.5) and inequality (.6). We have

$$\begin{aligned}\mathrm{pr}\left(\|U\|_\infty\geq\bar{\delta}_f(n,p_n^\tau)\right)&\leq\mathrm{pr}\,(\mathrm{I}\\&+\mathrm{II}+\mathrm{III}\geq\bar{\delta}_f(n,p_n^\tau)\leq\mathrm{pr}(\mathrm{I}\geq(1\\&-\alpha-\beta)\bar{\delta}_f(n,p_n^\tau))\\&+\mathrm{pr}(\mathrm{II}\geq\beta\bar{\delta}_f(n,p_n^\tau))\\&+\mathrm{pr}(\mathrm{III}\geq\alpha\bar{\delta}_f(n,p_n^\tau)).\end{aligned}$$

Choosing the parameter $\delta=(1-\alpha-\beta)\bar{\delta}_f(n,p_n^\tau)$ in (.7), so from (.7),

$$\mathrm{pr}\left(\mathrm{I}\geq(1-\alpha-\beta)\bar{\delta}_f(n,p_n^\tau)\right)\leq4p_n^2\exp\left\{-(1-\alpha-\beta)^2[\log4+\tau\log p_n]\right\}=\frac{4^{1-(1-\alpha-\beta)^2}}{p_n^{\tau(1-\alpha-\beta)^2-2}}.$$

So

$$\mathrm{pr}\left(\mathrm{I}\geq(1-\alpha-\beta)\bar{\delta}_f(n,p_n^\tau)\right)\leq\frac{4}{p_n^{\tau^*-2}}$$

Note that

$$\beta\bar{\delta}_f(n,p_n^\tau)=\Lambda_{\max}d_2^2(\rho_nM_n(X,T,\Sigma^*))^2$$

and

$$\alpha\bar{\delta}_f(n,p_n^\tau)=\gamma d_2\rho_n^2M_n(X,T,\Sigma^*)\sqrt{\Lambda_{\max}\log q_n/[\log p_n+\log q_n]},$$

further with (.8) and (.9),

$$\mathrm{pr}\left(\|U\|_\infty\geq\bar{\delta}_f(n,p_n^\tau)\right)\leq\frac{4}{p_n^{\tau^*-2}}+8\exp\{-C_1np_n^2\}+2\exp\{-C_2n\rho_n^2\}.$$

Thus we proved Lemma 2.

Based on Lemma 2, the rest of the proof follows closely to the proof to Theorem 1 in [5]. We only outline the proof here.

Lemma 4. *For any* $\lambda_n > 0$ *and sample covariance of* ε_i *based on the estimate* $\hat{\Gamma}$, $\hat{\Sigma}_{\hat{\Gamma}}$ *with strictly positive diagonal, the* ℓ_1*-penalized log-determinant problem* (4) *has a unique solution* $\hat{\Theta}_{\hat{\Gamma}} \succ 0$ *characterized by*

$$\hat{\Sigma}_{\hat{\Gamma}} - \hat{\Theta}_{\hat{\Gamma}}^{-1} + \lambda_n \hat{Z} = 0, \quad (.10)$$

where $\hat{Z}$ *is an element of the sub-differential* $\partial \|\hat{\Theta}_{\hat{\Gamma}}\|_{1,\text{off}}$.

This lemma is a slightly revised version of Lemma 3 in [5] and hence we omit the proof here. Based on this lemma, we construct the primal-dual witness solution $(\tilde{\Theta}, \tilde{Z})$ as follows:

a. Determine the matrix $\tilde{\Theta}$ by solving the restricted log-determinant problem

$$\tilde{\Theta} := \arg_{\Theta \succ 0, \Theta_{S^c}=0} \min \{ \text{tr}(\hat{\Sigma}_{\hat{\Gamma}} \Theta) - \log \det \Theta + \lambda_n \|\Theta\|_{1,\text{off}} \}. \quad (.11)$$

Note that by construction, we have $\tilde{\Theta} \succ 0$ and $\tilde{\Theta}_{S^c} = 0$.

b. We choose $\tilde{Z}_S$ as a member of the sub-differential of the regularizer $\|\cdot\|_{1,\text{off}}$, evaluated at $\tilde{\Theta}$.

c. Set $\tilde{Z}_{S^c}$ as

$$\tilde{Z}_{S^c} = \frac{1}{\lambda_n} \{ -\hat{\Sigma}_{S^c} + [\tilde{\Theta}^{-1}]_{S^c} \},$$

where $\hat{\Sigma}$ is short for $\hat{\Sigma}_{\hat{\Gamma}}$ and the constructed $(\tilde{\Theta}, \tilde{Z})$ satisfy the optimality condition (.10).

d. We verify the strict dual feasibility condition

$$|\tilde{Z}_{ij}| < 1 \text{for all} (i,j) \in S^c.$$

If the primal-dual witness construction succeeds, then it acts as a witness to the fact that the solution $\tilde{\Theta}$ to the restricted problem (.11) is equivalent to the solution $\hat{\Theta}$ to the original unrestricted problem (4) [5]. The proof proceeds as this: we first show that the primal-dual witness technique succeeds with high probability, hence the support of the optimal solution $\hat{\Theta}$ is contained within the support of the true Θ^*. In addition, the characterization of $\hat{\Theta}$ provided by the primal-dual witness construction can establish the element-wise ℓ_∞ bounds claimed in Theorem 2. Note we define the "effective noise" in the sample covariance matrix $\hat{\Sigma}_{\hat{\Gamma}}$ in the appendix as $U := \hat{\Sigma}_{\hat{\Gamma}} - (\Theta^*)^{-1}$ and we use $\Delta := \tilde{\Theta} - \Theta^*$ to measure the discrepancy between the restricted estimate Θ in (.11) and the truth Θ^*. We define $R(\Delta) := \tilde{\Theta}^{-1} - \Theta^{*-1} + \Theta^{*-1} \Delta \Theta^{*-1}$.

PROOF OF THEOREM 2. We first show that with high probability the witness matrix $\tilde{\Theta}$ is equal to the solution $\hat{\Theta}$ to the original log-determinant problem (4), by showing that the primal-dual witness construction succeeds with high probability. Let $\mathcal{B}$ denote the event that $\|U\|_\infty \leq \bar{\delta}_f(n, p_n^\tau) = \sqrt{(\log 4 + \tau \log p_n)/(C_* n)}$ where $C_* = [128(1+4\sigma^2)^2 \max_i \{\Sigma_{ii}^*\}^2]^{-1}$.

Condition (9) on sample size *n* implies $\bar{\delta}_f(n, p_n^\tau) \leq 8(1+4\sigma^2)\max_i\{\Sigma^*_{ii}\}$, which indicates that sub-Gaussian tail condition can be used in our control of sampling noise. Lemma 2 can guarantee $\mathrm{pr}(\mathscr{B}) \geq 1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2)$.

Conditioning on event $\mathscr{B}$, the following analysis follows as that of [5]. The choice of regularization penalty $\lambda_n = (8/\alpha)\bar{\delta}_f(n, p_n^\tau)$ implies $\|U\|_\infty \leq (\alpha/8)\lambda_n$. Following the same steps as [5], we can show that $\|R(\Delta)\|_\infty \leq \alpha\lambda_n/8$. We can then show that the matrix $Z_{\tilde{S}^c}$ constructed in step (c) satisfies $\|Z_{\tilde{S}^c}\|_\infty < 1$ and therefore $\tilde{\Theta} = \hat{\Theta}$. The estimator $\hat{\Theta}$ then satisfies the ℓ_∞ bound as claimed in Theorem 2 (1), and moreover, $\hat{\Theta}_{\tilde{S}^c} = \tilde{\Theta}_{\tilde{S}^c} = 0$, as claimed in the first part Theorem 2 (2). Second part of Theorem 2 (2) follows directly after (1). Since the above is conditioned on the event $\mathscr{B}$, these statements hold with probability

$$\mathrm{pr}(\mathscr{B}) \geq 1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2).$$

Hence we proved Theorem 2.

Proof of Theorem 3:

The proof of Theorem 3 depends on the following lemma.

Lemma 5 (Sign Consistency). *Suppose the minimum absolute value* $\theta_{\min}$ *of nonzero entries in the true precision matrix* Θ^* *is bounded from below by*

$$\theta_{\min} \geq 2\|\tilde{\Theta} - \Theta^*\|_\infty, \quad (.12)$$

then $\mathrm{sign}(\tilde{\Theta}_S) = \mathrm{sign}(\Theta^*_S)$ *holds*.

Proof of Lemma 5. This claim follows from the bound (.12), which guarantees for all $(i, j) \in S$, the estimate $\tilde{\Theta}_{ij}$ cannot differ enough from Θ^*_{ij} to change sign.

Proof of Theorem 3. Using the notation $\bar{\delta}_f(n, p_n^\tau) = \sqrt{(\log 4 + \tau \log p_n)/(C_* n)}$, where $C_* = [128(1+4\sigma^2)^2 \max_i\{\Sigma^*_{ii}\}^2]^{-1}$, the lower bound on *n* implies

$$\theta_{\min} > 4K_{\Omega^*}(1+\frac{8}{\alpha})\bar{\delta}_f(n, p_n^\tau).$$

As in the proof of Theorem 2, with probability greater than

$$1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2),$$

we have $\tilde{\Theta_\Gamma} = \hat{\Theta_\Gamma}$ and $\|\tilde{\Theta_\Gamma} - \Theta^*\|_\infty \leq \theta_{\min}/2$. Consequently, Lemma 5 implies that $\mathrm{sign}(\tilde{\Theta}_{ij}) = \mathrm{sign}(\Theta^*_{ij})$ for all $(i, j) \in E(\Theta^*)$. Overall, we can conclude that with probability greater than

$$1 - 4/p_n^{\tau^*-2} - 8\exp(-C_1 n\rho_n^2) - 2\exp(-C_2 n\rho_n^2),$$

the sign consistency condition $\mathrm{sign}(\hat{\Theta}_{ij})=\mathrm{sign}(\Theta^*_{ij})$ holds for all $(i, j) \in E(\Theta^*)$. This proves the theorem.

References

1. Bickel P, Levina E. Regularized estimation of large covariance matrices. Annals of Statistics. 2008a; 36(1):199–227.
2. Bickel P, Levina E. Covariance regularization by thresholding. Annals of Statistics. 2008b; 36(6): 2577–2604.
3. Cai T, Zhou H. Minimax estimation of large covariance matrices under ℓ_1 norm. Technical Report. 2010
4. Cai T, Zhang C-H, Zhou H. Optimal rates of convergence for covariance matrix estimation. The Annals of Statistics. 2010; 38:2118–2144.
5. Ravikumar P, Wainwright M, Raskutti G, Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Electronic Journal of Statistics. 2011; 5:935–980.
6. Lam C, Fan J. Sparsistency and rates of convergence in large covariance matrices estimation. The Annals of Statistics. 2009; 37:4254–4278.
7. El Karoui N. Operator norm consistent estimation of large dimensional sparse covariance matrices. The Annals of Statistics. 2008; 36:2717–2756.
8. Cheung V, Spielman R. The genetics of variation in gene expression. Nature Genetics. 2002:522–525. [PubMed: 12454648]
9. Rothman A, Levina E, Zhu J. Sparse multivariate regression with covariate estimation. Journal of Computational and Graphical Statistics. 2010; 19(4):947–962.
10. Yin J, Li H. A sparse conditional gaussian graphical model for analysis of genetical genomics data. Annals of Applied Statistics. 2011; 5:2630–2650. [PubMed: 22905077]
11. Wainwright MJ. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). IEEE Transactions on Information Theory. 2009; 55:2183–2202.
12. Zhao P, Yu B. On model selection consistency of lasso. Journal of Machine Learning Research. 2006; 7:2541–2567.
13. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008; 9:432–441. [PubMed: 18079126]
14. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the lasso. Annals of Statistics. 2006; 34
15. Li H, Gui J. Gradient directed regularization for sparse gaussian concentration graphs, with applications to inference of genetic networks. Biostatistics. 2006; 7:302–317. [PubMed: 16326758]
16. Fan J, Feng Y, Wu Y. Network exploration via the adaptive lasso and scad penalties. The Annals of Applied Statistics. 2009; 3:521–541. [PubMed: 21643444]
17. Peng J, Wang P, Zhou N, Zhu J. Partial correlation estimation by joint sparse regression models. Journal of American Statistical Association. 2009; 104:735–746.
18. Brem R, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proceedings of National Academy of Sciences. 2005; 102:1572–1577.
19. Steffen M, Petti A, Aach J, D'Haeseleer P, Church G. Automated modelling of signal transduction networks. BMC Bioinformatics. 2002; 3:34. [PubMed: 12413400]
20. Stark C, Breitkreutz B, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone M, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust J, Winter A, Dolinski K, Tyers M. The biogrid interaction database: 2011 update. Nucleic Acids Research. 2011; 39:D698–D704. [PubMed: 21071413]
21. Candes E, Tao T. The dantzig selector: Statistical estimation when p is much larger than n. Annals of Statistics. 2007; 35:2313–2351.

22. Cai T, Liu W, Luo X. A constrained l1 minimization approach to sparse precision matrix estimation. Journal of American Statistical Association. 2011; 106:594–607.

23. Bunea F, She Y, Wegkamp M. Optimal selection of reduced rank estimators of high-dimensional matrices. Annals of Statistics. 2011; 39(2):1282–1309.

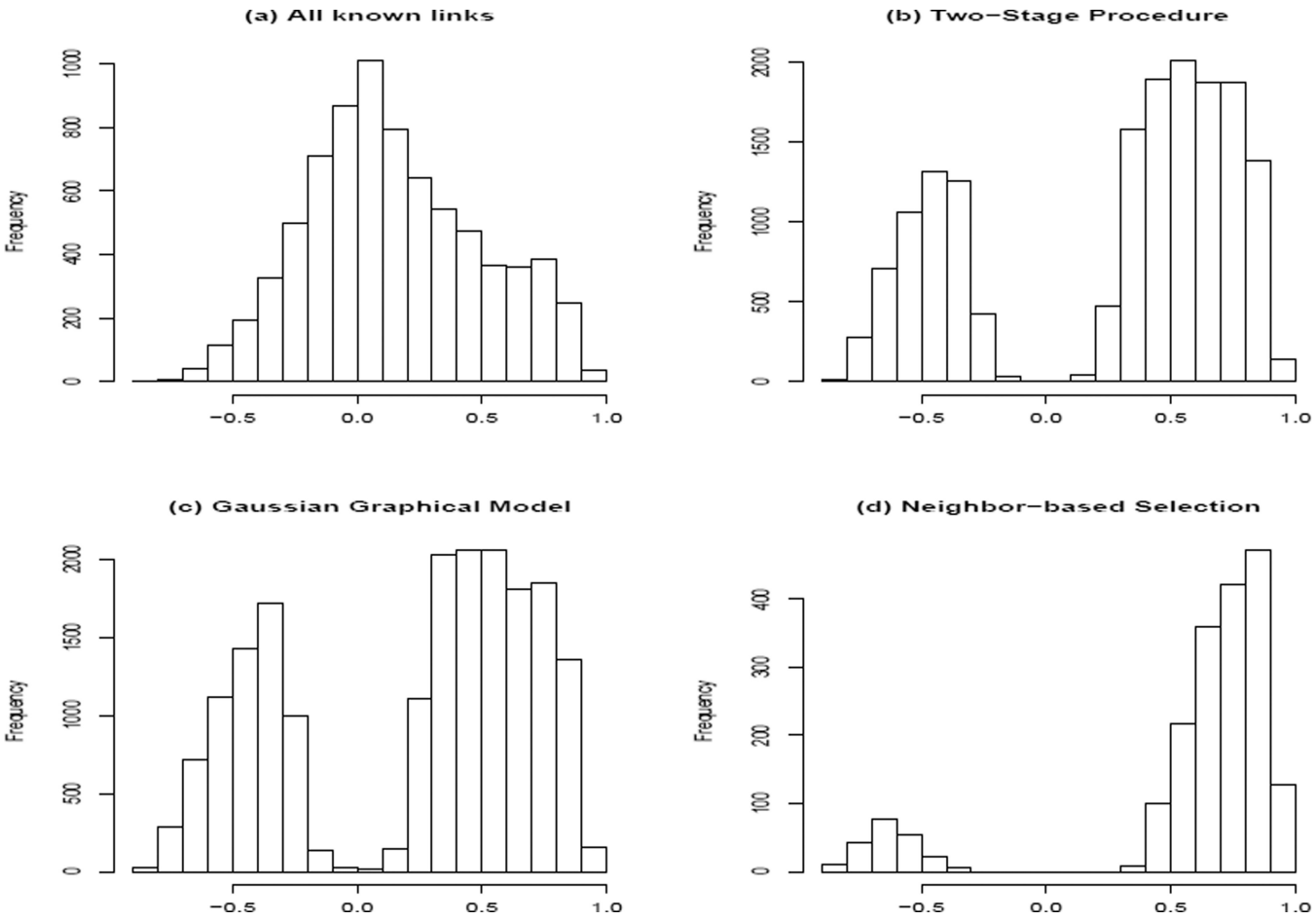

Figure 1.
Histograms of marginal correlations for pairs of linked genes based on BioGRID (a) and linked genes identified by the two-stage procedure (b), the Gaussian graphical model (c) and a neighbor-based selection procedure (d).

Table 1

Six models considered in simulations, where p is the number of the variables, q is the number of covariates and n is the sample size. $\mathrm{pr}(\Theta_{ij} \neq 0)$ and $\mathrm{pr}(\Gamma_{ij} \neq 0)$ specify the sparsity of the model.

Model	(p, q, n)	$\mathrm{pr}(\Theta_{ij} \neq 0)$	$\mathrm{pr}(\Gamma_{ij} \neq 0)$
1	(100, 100, 250)	$2/p$	$3/q$
2	(50, 50, 250)	$2/p$	$4/q$
3	(25, 10, 250)	$2/p$	$3.5/q$
4	(1000, 200, 250)	$1.5/p$	$20/q$
5	(800, 200, 250)	$1.5/p$	$25/q$
6	(400, 200, 250)	$2.5/p$	$20/q$

Table 2

Comparison of the performances on estimating the precision matrix Θ by the two-stage procedure, the iterative selection procedure of [10], a neighbor-based selection procedure [14] and the Gaussian graphical model using GLASSO [13], where $\Delta = \Theta - \hat{\Theta}$.

Method	AUC	SPE	SEN	MCC	$\|\Delta\|_\infty$	$\|\|\|\Delta\|\|\|_\infty$	$\|\Delta\|_2$	$\|\Delta\|_F$
Model 1: (p, q, n)=(100, 100, 250)								
Two-stage	0.91	0.99	0.49	0.56	0.32	1.18	0.68	3.24
Iterative	0.91	0.99	0.48	0.56	0.33	1.17	0.67	3.18
GLASSO	0.81	0.97	0.24	0.21	0.69	1.89	1.12	5.19
Neighbor	0.86	0.99	0.38	0.48				
Model 2: (p, q, n)=(50, 50, 250)								
Two-stage	0.91	0.97	0.69	0.65	0.35	1.31	0.73	2.43
Iterative	0.92	0.98	0.69	0.66	0.37	1.30	0.72	2.36
GLASSO	0.74	0.87	0.37	0.18	0.75	2.12	1.20	4.57
Neighbor	0.88	0.95	0.60	0.48				
Model 3: (p, q, n)= (25, 10, 250)								
Two-stage	0.89	0.91	0.76	0.62	0.23	0.90	0.51	1.20
Iterative	0.89	0.91	0.76	0.62	0.24	0.90	0.52	1.21
GLASSO	0.57	0.43	0.73	0.12	0.65	1.99	1.12	2.77
Neighbor	0.85	0.84	0.68	0.44				
Model 4: (p, q, n)=(1000, 200, 250)								
Two-stage	0.93	1	0.32	0.51	0.46	1.77	0.91	13.42
Iterative	0.90	1	0.31	0.47	0.59	1.81	0.97	13.48
GLASSO	0.88	0.98	0.08	0.02	0.71	2.86	1.31	19.82
Neighbor	0.87	1	0.12	0.16				
Model 5: (p, q, n)=(800, 200, 250)								
Two-stage	0.93	1	0.21	0.45	0.48	1.80	0.97	12.58
Iterative	0.89	1	0.21	0.34	0.75	2.30	1.20	12.82
GLASSO	0.87	0.97	0.07	0.02	0.76	2.97	1.40	18.39
Neighbor	0.87	0.96	0.61	0.19				
Model 6: (p, q, n)=(400, 200, 250)								
Two-stage	0.79	1	0.05	0.20	0.39	1.56	0.79	7.13
Iterative	0.75	1	0.05	0.21	0.44	1.55	0.77	6.86

J Multivar Anal. Author manuscript; available in PMC 2014 April 01.

Method	AUC	SPE	SEN	MCC	$\|\Delta\|_\infty$	$\|\|\|\Delta\|\|\|_\infty$	$\|\Delta\|_2$	$\|\Delta\|_F$
GLASSO	0.71	0.95	0.03	−0.01	0.69	2.72	1.22	11.01
Neighbor	0.73	0.99	0.08	0.10				

Table 3

Comparison of the performances on estimating the regression coefficient matrix Γ from the two-stage procedure, an iterative selection procedure of [10] and a neighbor-based procedure [14], where $\Delta = \Gamma - \hat{\Gamma}$.

Algorithm	AUC	SPE	SEN	MCC	$\|\Delta\|_\infty$	$\|\|\|\Delta\|\|\|_\infty$	$\|\Delta\|_F$
Model 1: (p, q, n)=(100,100,250)							
Two-stage	0.98	0.99	0.87	0.77	0.38	1.03	2.39
Iterative	0.98	0.98	0.90	0.64	0.36	1.01	2.16
Neighbor	0.97	0.99	0.87	0.78	0.38	1.06	2.39
Model 2: (p, q, n)=(50,50,250)							
Two-stage	0.98	0.99	0.89	0.84	0.37	1.65	2.48
Iterative	0.98	0.98	0.90	0.81	0.36	1.70	2.32
Neighbor	0.97	0.96	0.91	0.75	0.35	1.48	2.21
Model 3: (p, q, n)=(25,10,250)							
Two-stage	0.98	0.75	0.98	0.68	0.24	0.74	0.97
Iterative	0.97	0.81	0.98	0.74	0.25	0.75	1
Neighbor	0.98	0.90	0.95	0.81	0.31	1.02	1.30
Model 4: (p, q, n)=(1000,200,250)							
Two-stage	0.96	1	0.82	0.82	0.48	1.90	11.86
Iterative	0.96	1	0.83	0.79	0.62	2.98	11.98
Neighbor	0.83	1	0.65	0.80	0.81	3.51	18.75
Model 5: (p, q, n)=(800,200,250)							
Two-stage	0.97	1	0.83	0.82	0.48	2.49	11.69
Iterative	0.96	1	0.81	0.79	0.89	6.52	12.75
Neighbor	0.79	0.97	0.77	0.46	0.76	4.21	14.48
Model 6: (p, q, n)=(400,200,250)							
Two-stage	0.96	1	0.82	0.82	0.45	2.03	7.29
Iterative	0.96	0.99	0.86	0.65	0.44	2.27	6.40
Neighbor	0.86	1	0.78	0.83	0.56	2.64	8.35

Table 4

Comparison of the results of the two-stage procedure, the neighbor-based procedure [14] and the Gaussian graphical model using GLASSO [13] for the yeast protein-protein interaction data where $n = 112$, $p = 1207$, $q = 578$.

	Two-stage	Neighbor	Gaussian graph
No. of edges in $\hat{\Theta}$	13522	7518	18987
No. of links in $\hat{\Gamma}$	1030	330	NA
Tuning parameter	(0.326, 0.362)	0.324	0.224
Mean degree	27.16	3.18	31.5
Max degree	53	12	60