Sure Screening for Gaussian Graphical Models

Shikai Luo SLUO@NCSU.EDU
Chengchun Shi CSHI4@NCSU.EDU
Rui Song RSONG@NCSU.EDU
Department of Statistics
North Carolina State University
Raleigh, NC 27695, USA

Yuxiang Xie YXXIE@UW.EDU
Daniela Witten DWITTEN@UW.EDU
Departments of Statistics & Biostatistics
University of Washington
Seattle, WA 98195, USA

Editor:

Abstract

We propose graphical sure screening, a very simple and computationally efficient screening procedure for recovering the structure of a Gaussian graphical model in the high dimensional setting. The graphical sure screening estimator of the conditional dependence graph is obtained by thresholding the elements of the sample covariance matrix. The proposed approach possesses the sure screening property: with very high probability, the estimated edge set contains the true edge set. Furthermore, with high probability, the size of the estimated edge set is controlled. We provide a choice of threshold for graphical sure screening that can control the expected false positive rate. In a simulation study and on a gene expression data set, graphical sure screening performs quite competitively with more complex and computationally demanding techniques for graph estimation.

Keywords: conditional dependence; high dimensionality; sparsity; undirected graph

1. Introduction

In recent years, graphical modeling has been of widespread interest in a number of fields. In genomics, graphical models have been extensively used to model gene regulatory networks, composed of tens of thousands of genes. It is typically of interest to infer the structure of the graph based on hundreds, or at most thousands, of observations for which gene expression measurements are available. Consequently, the setting is high dimensional, in the sense that there are many more features than observations.

Consider the random vector $x = (x_1, \ldots, x_p)^T$, and the conditional dependence graph $\mathcal{G} = (\Gamma, \mathcal{E})$. Here $\Gamma = \{1, \ldots, p\}$ is the set of nodes, and $\mathcal{E}$ is the set of edges in $\Gamma \times \Gamma$. A pair (j, l) is contained in the edge set $\mathcal{E}$ if and only if x_j is conditionally dependent on x_l, given all remaining variables $x_{\Gamma\backslash(j,l)} = \{x_k : k \in \Gamma\backslash(j, l)\}$.

The conditional dependence graph takes a particularly simple form if we suppose that $x \sim N_p(0, \Sigma)$, where Σ is a non-singular covariance matrix. In this setting, a pair of variables is conditionally independent if and only if the corresponding entry of the precision matrix Σ^{-1} equals zero

(Lauritzen, 1996; Mardia et al., 1980). Consequently, in the Gaussian graphical model, recovering the edge set $\mathcal{E}$ is equivalent to recovering the sparsity pattern of the matrix Σ^{-1}.

Recently, a number of proposals have been made for estimating and recovering the sparsity pattern of a large sparse precision matrix, using penalized likelihood (Yuan and Lin, 2007; Friedman et al., 2008; Lam and Fan, 2009; Rothman et al., 2008; Ravikumar et al., 2011) and neighborhood selection (Meinshausen and Bühlmann, 2006; Yuan, 2010; Cai et al., 2011, 2012; Sun and Zhang, 2013) approaches. For many of the aforementioned approaches, statistical convergence results in terms of various matrix norms have been established for the high-dimensional setting.

Although various computationally efficient algorithms for estimating a sparse precision matrix have been proposed (e.g. Friedman et al., 2008; Witten et al., 2011), the required computations can be burdensome when the number of variables is in the tens of thousands, or even higher. For example, the precision matrix for a problem with $p = 25,000$, the number of genes in the human genome, involves upwards of $300,000,000$ parameters. In such a setting, existing algorithms are infeasible. We are thus motivated to consider a computationally-efficient screening approach for Gaussian graphical models that possesses desirable statistical properties.

In recent years, computationally simple variable screening approaches have gained popularity in the context of high dimensional regression and classification. Fan and Lv (2008) proposed sure independence screening for linear models. This approach possesses the sure screening property: with probability going to one, all important variables are selected. Fan et al. (2009) and Fan and Song (2010) extended this approach to generalized linear models. Other marginal screening methods include tilting methods (Hall et al., 2009), generalized correlation screening (Hall and Miller, 2009), nonparametric screening (Fan et al., 2011), partial likelihood screening (Zhao and Li, 2012), and robust rank correlation based screening (Li et al., 2012; Zhu et al., 2011). Most screening methods aim to select variables by ranking utilities such as the correlation between the marginal covariates and the response, where variables with strong marginal utilities are selected.

In this paper, we propose a novel screening procedure for recovering the structure of a Gaussian graphical model. Our approach is motivated by the fact that the jth column of the precision matrix Σ^{-1} can be obtained by regressing the jth feature onto the $p-1$ other features (Mardia et al., 1980). This suggests that in order to estimate $\mathcal{E}_j$, the neighborhood of the jth node, we can emulate the sure screening procedure of Fan and Lv (2008) for linear models: we simply threshold the sample correlations of the jth feature with the $p-1$ other features. We show that this approach is well-founded theoretically. Furthermore, it enjoys good empirical performance relative to existing approaches for estimating a sparse precision matrix, at a much lower computational cost of $\mathcal{O}(p^2)$ versus $\mathcal{O}(p^3)$ operations (Friedman et al., 2008). As far as we know, this is the first time that a sure screening procedure has been applied in an unsupervised context.

2. Graphical Sure Screening

2.1 Proposed Approach

Consider the random vector $x = (x_1, \ldots, x_p)^T \sim N_p(0, \Sigma)$, where Σ has unit diagonals, i.e. $E(x_j^2) = 1$ for all $j = 1, \ldots, p$. The $n \times p$ data matrix $X = (X_1, \ldots, X_p)$ contains n independent draws from x. Let $\gamma_n > 0$ be some pre-specified threshold.

We propose to obtain a candidate edge set, $\widehat{\mathcal{E}}_{\gamma_n}$, and a candidate neighborhood for the jth node, $\widehat{\mathcal{E}}_{j,\gamma_n}$, by thresholding the sample correlation matrix by γ_n. That is, we define

$$\widehat{\mathcal{E}}_{\gamma_n} = \{(j,l) : j < l, |X_j^T X_l|/n > \gamma_n\}, \tag{1}$$

$$\widehat{\mathcal{E}}_{j,\gamma_n} = \{l : l \neq j, |X_j^T X_l|/n > \gamma_n\}. \tag{2}$$

We refer to (1) and (2) as the graphical sure screening estimators.

We now briefly illustrate the graphical sure screening approach with a small example. We consider two simulation settings: one in which the precision matrix is block diagonal, and one in which it is tridiagonal. These two settings are discussed in greater detail in Section 5.1. We applied both the graphical lasso (Friedman et al., 2008) and graphical sure screening, with tuning parameters chosen to yield the same number of estimated edges. Results are shown in Fig. 1. The edge sets estimated by the graphical lasso and graphical sure screening are quite similar. This indicates that despite its simplicity, graphical sure screening can provide results that are competitive with a computationally-intensive state-of-the-art approach.

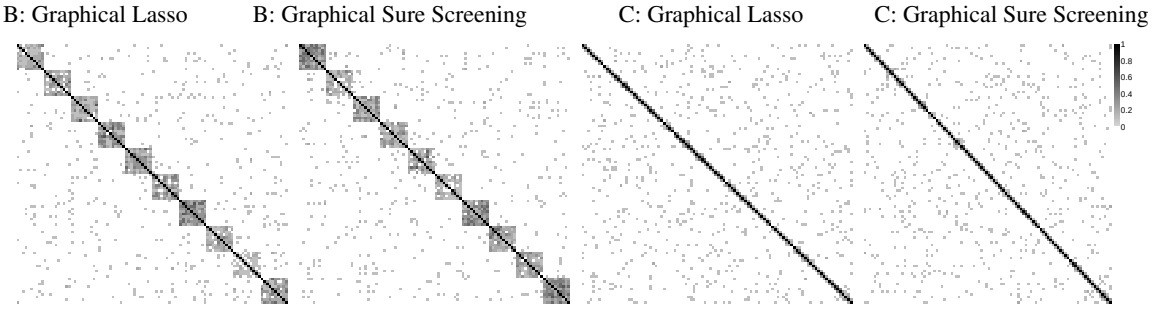

Figure 1: The adjacency matrices corresponding to the graphical lasso and graphical sure screening estimates, averaged over ten simulated data sets and displayed as heatmaps, for Simulations B and C with $p = 100$ and $n = 50$. Simulations B and C are described in detail in Section 5.1.

2.2 Theoretical Properties

In what follows, we use the notation $\sigma_{jl} \equiv E(x_j x_l)$. Let $\Lambda_{\max}(A)$ and $\Lambda_{\min}(A)$ denote the largest and smallest singular values of the matrix A, respectively. All proofs are provided in the supplementary article.

Our key assumption bounds the elements of Σ corresponding to an edge in $\mathcal{E}$.

Assumption 1 *For some constants $C_1 > 0$ and $0 < \kappa < 1/2$, $\min_{(j,l)\in\mathcal{E}} |\sigma_{jl}| \geq C_1 n^{-\kappa}$.*

We now establish the sure screening property of graphical sure screening.

Theorem 1 *Suppose that $\log(p) = C_3 n^{\xi}$ for some constants $C_3 > 0$ and $\xi \in (0, 1-2\kappa)$. Let $\gamma_n = 2C_1 n^{-\kappa}/3$. Then, given Assumption 1, there exist constants C_4 and C_5 such that*

$$\mathrm{pr}(\mathcal{E}_j \subseteq \widehat{\mathcal{E}}_{j,\gamma_n}) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa}), \quad \mathrm{pr}(\mathcal{E} \subseteq \widehat{\mathcal{E}}_{\gamma_n}) \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa}). \tag{3}$$

Conversely, if $\min_{(j,l)\in\mathcal{E}} |\sigma_{jl}| < C_1 n^{-\kappa}/3$*, then there exist constants* C_6 *and* C_7 *such that*

$$\mathrm{pr}(\mathcal{E} \not\subseteq \widehat{\mathcal{E}}_{\gamma_n}) \geq 1 - C_6 \exp(-C_7 n^{1-2\kappa}). \tag{4}$$

Theorem 1 indicates that Assumption 1 is necessary and sufficient for the sure screening property to hold. The following corollary further indicates that under Assumption 1, graphical sure screening can identify the connected components of $\mathcal{E}$ with probability converging to one.

Corollary 2 *Suppose that the graph* $\mathcal{E}$ *is composed of* g *connected components, and that the* k*th connected component contains the variables* $x_1^{(k)}, \ldots, x_{p_k}^{(k)}$*, where* $\sum_{k=1}^{g} p_k = p$*. That is,* $x_l^{(s)}$ *and* $x_j^{(t)}$ *are independent for* $s \neq t$*. Let* $\gamma_n = 2C_1 n^{-\kappa}/3$*. Under Assumption 1, if* $\log(p) = C_3 n^{\xi}$ *for* $\xi \in (0, 1-2\kappa)$*, then the connected components of* $\widehat{\mathcal{E}}_{\gamma_n}$ *are the same as the connected components of* $\mathcal{E}$ *with probability at least* $1 - C_4 \exp\left(-C_5 n^{1-2\kappa}\right)$.

Next, we will establish a bound on the size of $\widehat{\mathcal{E}}_{j,\gamma_n}$, the estimated neighborhood for the jth node. We first assume that the largest eigenvalue of Σ does not diverge too quickly as n grows. For example, this holds for the covariance matrix of a stationary time series (Fan and Lv, 2008).

Assumption 2 *There exist constants* $\tau \geq 0$ *and* $C_2 > 0$ *such that* $\Lambda_{\max}(\Sigma) \leq C_2 n^{\tau}$.

Theorem 3 *Let* $\gamma_n = 2C_1 n^{-\kappa}/3$*. Under Assumptions 1–2, if* $\log(p) = C_3 n^{\xi}$ *for* $\xi \in (0, 1-2\kappa)$*, then* $\mathrm{pr}\left\{|\widehat{\mathcal{E}}_{j,\gamma_n}| \leq O(n^{2\kappa+\tau})\right\} \geq 1 - C_4 \exp(-C_5 n^{1-2\kappa})$.

It is worthwhile to note that we have made no explicit assumption about the size of the edge set $\mathcal{E}$. Nonetheless, together, Theorems 1 and 3 guarantee that with high probability, the size of the true neighborhood of the jth node, $|\mathcal{E}_j|$, is no larger than a polynomial order of n.

The false positive rate is defined as $|\widehat{\mathcal{E}}_{\gamma_n} \cap \mathcal{E}^c|/|\mathcal{E}^c|$, where $\mathcal{E}^c = \{(j,l) : (j,l) \notin \mathcal{E}, j \neq l\}$. Along the lines of Zhao and Li (2012), next we propose a choice of γ_n that allows us to control the expected false positive rate at a pre-specified value. This requires another assumption.

Assumption 3 *For the same* ξ *as in Theorem 1,* $\max_{(j,l)\in\mathcal{E}^c} |\sigma_{jl}| = o\left\{n^{-(1-\xi)/2}\right\}$.

Theorem 4 *Under Assumptions 1–3, if* $\log(p) = C_3 n^{\xi}$ *for* ξ *as in Theorem 1, then for any positive* m*, we can control the asymptotic expected false positive rate at* $m/|\mathcal{E}^c|$ *by choosing* $\gamma_n = n^{-1/2}\Phi^{-1}\left[1 - m/\left\{p(p-1)\right\}\right]$*. Furthermore, with this threshold, the sure screening property of Theorem 1 holds.*

3. Connection with the Faithfulness Assumption

Assumption 1 puts conditions on the elements of Σ corresponding to elements in $\mathcal{E}$. However, it remains unclear what types of matrices Σ satisfy this assumption. In order to clarify this point, we present a proposition below which connects Assumption 1 with the faithfulness assumption (c.f, Definition 13.4, Bühlmann and van de Geer, 2011).

We introduce some notations. A path from j_1 to j_m where $j_1, j_m \in \mathcal{V}$ is sequence of vertices $\{j_1, j_2, \ldots, j_{m-1}, j_m\}$ such that for all $k = 1, \ldots, m-1$, $j_k \in \mathcal{V}$ and $(j_k, j_{k+1}) \in \mathcal{E}$. For any disjoint subsets $A, B, C \subseteq \mathcal{V}$, C separates A and B if every path from $j_1 \in A$ to $j_m \in B$ contains a

vertex in C. A and B is separated if there are no path from $j_1 \in A$ to $j_m \in B$. For any two random variables X and Y, $X \perp Y$ means X and Y are independent. Besides, we denoted by $P \iff Q$ if P and Q are logically equivalent. Our assumption 1 is implies by the following two conditions.
A1. For any $a, b \in \mathcal{V}$, $a \neq b$, we have

$$x_a \perp x_b \iff \{a\} \text{ and } \{b\} \text{ is separated.}$$

A2. For some constants $C_1 > 0$ and $0 < \kappa < 1/2$, $\min\{|\sigma_{a,b}| : \sigma_{a,b} \neq 0\} \geq C_1 n^{-\kappa}$.

For A1, the implication $\Longleftarrow$ is implied by global Markov property, which automatically satisfies for Gaussian graphical model. A2 requires the minimum nonzero covariance is greater than some thresholds which converges to 0 as sample size increases.

Proposition 5 *Assumption 1 holds if A1 and A2 are satisfied.*

The proof is easy by noting that under A1, $\mathcal{E} \subseteq \{(a, b) : \sigma_{a,b} \neq 0\}$. Assume there exists some $(a, b) \in \mathcal{E}$ and that $\sigma_{a,b} = 0$. By A1, this suggests $\{a\}$ and $\{b\}$ is separated, and therefore $\{a\}$ and $\{b\}$ is separated by $V - \{a, b\}$. By the global Markov property, this further implies $x_a \perp x_b | x_j, j \in V - \{a, b\}$, i.e, $(a, b) \subsetneq \mathcal{E}$, and we've reached a contraction.

We name A1 "marginal pairwise faithfulness" assumption. For fixed p, A1 automatically holds. Besides, it is implied by A1*, better known as the faithfulness assumption, introduced as follows.

A1*. For any disjoint sets $A, B, C \subseteq \mathcal{V}$,

$$x_A \perp x_B | x_C \iff C \text{ separates } A \text{ and } B.$$

By definition, our assumption A1 is in general weaker than A1* in that it restricts the class of sets (A, B, C) to the following special class $(\{a\}, \{b\}, \emptyset)$. Bühlmann and van de Geer (2011) gave two examples when A1* fails and commented that such assumption is unlikely to be violated. The graph generated in Setting E in Section 5 violates the faithfulness assumption and results are given therein to show numerical performance of our method.

4. Connection with the Graphical Lasso

We now consider the connections between graphical sure screening and the graphical lasso estimator (Friedman et al., 2008; Yuan and Lin, 2007), which solves the optimization problem

$$\text{maximize}_{\Theta} \left[\log\det\Theta - \text{tr}\left\{(X^T X/n)\Theta\right\} - \lambda \sum_{i \neq j} |\Theta_{ij}| \right]. \tag{5}$$

4.1 The Connected Components of the Graphical Lasso and Graphical Sure Screening

Recently, Witten et al. (2011) and Mazumder and Hastie (2012) established that the connected components of the graphical lasso estimator are exactly the same as the connected components that result from hard-thresholding the matrix $X^T X/n$ by λ. In other words, the connected components of the graphical lasso estimator are the same as the connected components of the graphical sure screening edge set estimator, $\widehat{\mathcal{E}}_{\gamma_n}$, when $\gamma_n = \lambda$.

However, the overall sparsity patterns of the graphical lasso and of graphical sure screening are typically not the same. The graphical lasso can be regarded as a two-stage procedure, in which we first perform graphical sure screening with $\gamma_n = \lambda$, and then perform a smaller graphical lasso problem on each connected component of $\widehat{\mathcal{E}}_{\gamma_n}$.

4.2 A Comparison of Assumption 1 and the Irrepresentability Condition

Ravikumar et al. (2011) showed that the the graphical lasso is model selection consistent under an irrepresentability condition, which bounds the Hessian of the log-determinant barrier, evaluated at the true precision matrix. We now show, by way of two simple examples, that our Assumption 1 is neither weaker nor stronger than the irrepresentability condition.

Example 1 *Consider the diamond-shaped graph illustrated in Fig. 2, with*

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & 2\rho^2 \\ \rho & 1 & 0 & \rho \\ \rho & 0 & 1 & \rho \\ 2\rho^2 & \rho & \rho & 1 \end{pmatrix}, \quad \Sigma^{-1} = \frac{1}{1-2\rho^2} \begin{pmatrix} 1 & -\rho & -\rho & 0 \\ -\rho & 1 & 2\rho^2 & -\rho \\ -\rho & 2\rho^2 & 1 & -\rho \\ 0 & -\rho & -\rho & 1 \end{pmatrix}.$$

As discussed in Ravikumar et al. (2011), the irrepresentability condition holds provided that $\rho \in (-0.2017, 0.2017)$. *However, Assumption 1 is violated, since* $(2,3) \in \mathcal{E}$ *and* $\sigma_{23} = 0$.

Example 2 *Consider the star-shaped graph illustrated in Fig. 2, with*

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho^2 & \rho^2 \\ \rho & \rho^2 & 1 & \rho^2 \\ \rho & \rho^2 & \rho^2 & 1 \end{pmatrix}, \quad \Sigma^{-1} = \frac{1}{1-\rho^2} \begin{pmatrix} 1+2\rho^2 & -\rho & -\rho & -\rho \\ -\rho & 1 & 0 & 0 \\ -\rho & 0 & 1 & 0 \\ -\rho & 0 & 0 & 1 \end{pmatrix}.$$

As discussed in Ravikumar et al. (2011), the irrepresentability condition fails if $|\rho| > 0.414$. *However, Assumption 1 holds for any value of* ρ, *since* $\min_{(j,l)\in\mathcal{E}} |\sigma_{jl}| = \rho$.

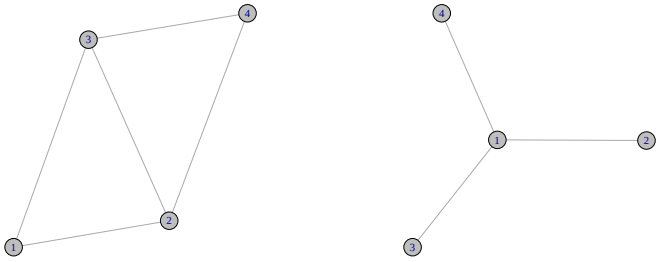

Figure 2: *Left:* The diamond graph described in Example 1. *Right:* The star-shaped graph described in Example 2.

The theoretical results for graphical sure screening are somewhat milder than those for the graphical lasso: while the graphical lasso estimates exactly the correct edge set with high probability, graphical sure screening instead yields no false negatives and a controlled expected false positive rate. The price that the graphical lasso pays for this stronger theoretical result is relatively high, in terms of complexity of both assumptions and computations. We also note that Assumptions 1 and 3 for graphical sure screening are simple to explain to a practitioner. In fact, even without Assumption 3, both Theorems 1 and 2 and Corollary 1 hold.

5. Simulation Studies

In this section, we explore the empirical performance of the threshold proposed in Theorem 4 for expected false positive rate control, and compare the performance of graphical sure screening to competitors from the literature. In the supplementary article, we explore the extent to which $\mathcal{E} \subseteq \widehat{\mathcal{E}}_{\gamma_n}$ holds for finite samples.

5.1 Data Generation

We considered four simulation set-ups. The covariance matrices were generated as follows.

Simulation A: For all $j < l$, we set $(j, l) \in \mathcal{E}$ with probability 0.01. We then generated a $p \times p$ matrix A, where

$$A_{jl} = A_{lj} = \begin{cases} 1, & j = l, \\ \text{Unif}(-0.3, 0.7), & (j, l) \in \mathcal{E}, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Let I denote the $p \times p$ identity matrix. We then created a positive definite matrix

$$\Sigma^{-1} = A + \{0.1 - \Lambda_{\min}(A)\} I. \tag{7}$$

Simulation B: We partitioned the p features into ten equally-sized and non-overlapping sets, i.e.

$$C_k = \{(k-1)p/10 + 1, \ldots, kp/10\}, \quad (k = 1, \ldots, 10).$$

For all $l \in C_k, j \in C_k, l < j$, we set $(l, j) \in \mathcal{E}$. We then generated A according to (6), and Σ^{-1} according to (7).

Simulation C: For $1 \leq j, l \leq p$, we set $\Sigma_{jl} = 0.3^{|j-l|}$.

Simulation D: We partitioned the features into $p/10$ equally-sized and non-overlapping sets, i.e.

$$C_k = \{10(k-1) + 1, \ldots, 10k\}, \quad k = 1, \ldots, p/10.$$

For $j \in C_k$ and $l \in C_k$, we set $(\Sigma^{-1})_{jl} = 0.9^{|j-l|}$. We set $(\Sigma^{-1})_{jl} = 0$ otherwise.

In Simulations A–D, we then rescaled the covariance matrix Σ to have diagonal elements equal to 1. Finally, we generated n observations independently from a $N_p(0, \Sigma)$ distribution.

The covariance and precision matrices corresponding to Simulations A–D are displayed in Fig. 3. These four simulation set-ups allow us to explore the performance of graphical sure screening under various violations to the assumptions in Section 2.2. Roughly speaking, Assumption 1 in Section 2.2 states that any element in $\mathcal{E}$ — that is, any non-zero element of the precision matrix — should correspond to a sufficiently large value of the covariance matrix. We see from Fig. 3 that in Simulations A, B, and C, all non-zero elements of the precision matrix are also non-zero in the covariance matrix. However, this is not the case in Simulation D.

In contrast, Assumption 3 in Section 2.2 states that any element not in the edge set — that is, any zero element of the precision matrix — should correspond to a sufficiently small value of the covariance matrix. We see from Fig. 3 that in Simulations B and D, all zero elements of the precision matrix are also zero in the covariance matrix. However, this is not the case in Simulations A and C.

The constants -0.3, 0.7, and 0.1 in (6) and (7) were chosen in order to obtain a suitable signal-to-noise ratio and to guarantee a well-conditioned covariance matrix. Broadly similar results to those in Sections 5.2 and 5.3 are obtained if these values are modified.

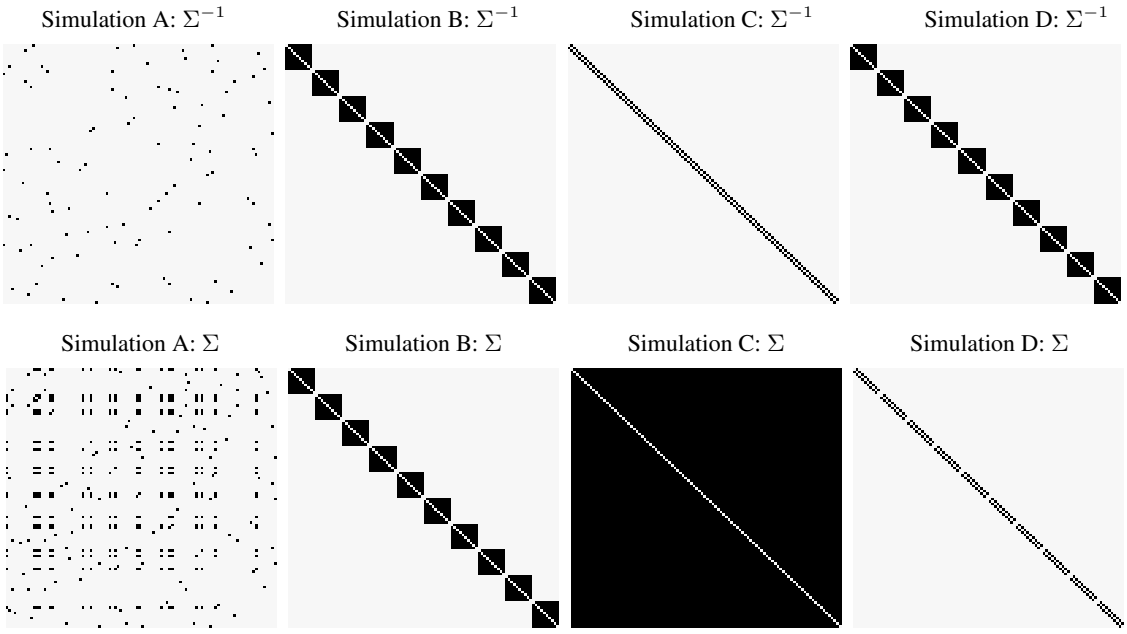

Figure 3: The sparsity patterns of Σ^{-1} and Σ, with $p = 100$.

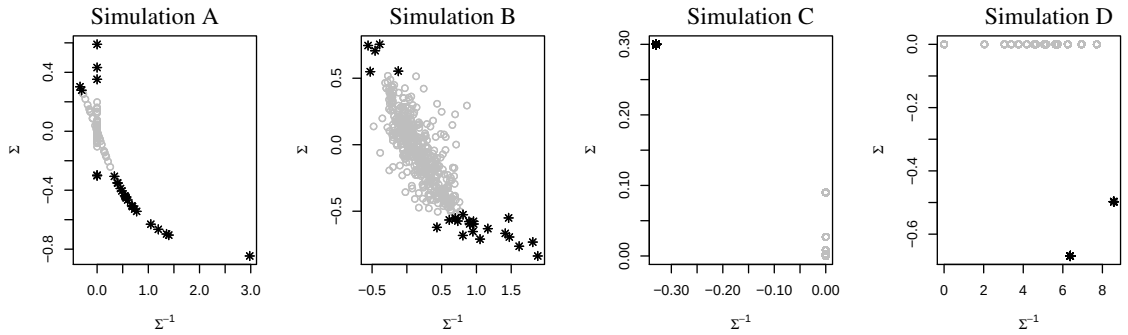

Figure 4: The off-diagonal elements of Σ^{-1} and Σ, with $p = 100$. The 0.5% of largest elements of Σ are displayed as black stars ($*$). The pronounced relationship in the left-most panel is due to the fact that in Simulation A, with high probability, a given column of Σ^{-1} contains no more than one non-zero off-diagonal element. Consequently, Σ^{-1} is approximately a block-diagonal matrix with blocks containing no more than two features, and there is a simple monotone relationship between the off-diagonal elements of a 2×2 symmetric matrix and its inverse.

5.2 Control of False Positive Rate

Theorem 4 states that under certain conditions, graphical sure screening with a particular choice of threshold controls the asymptotic expected false positive rate. We now investigate the extent to which the false positive rate is controlled in finite samples. Results for Simulations A–D, with $n = 100$ and $p = 1000$, are shown in Table 1.

In Simulation B, the sparsity patterns of the precision and covariance matrices are identical. Thus, it is not surprising that the false positive rate is controlled successfully. However, in Simulations A, C, and D, the sparsity patterns of Σ and Σ^{-1} are quite different, as revealed in Fig. 3. Nonetheless, Table 1 indicates that the false positive rate is very well-controlled. This is due to the fact that Theorem 4 simply requires that the elements in $\mathcal{E}$ correspond to large elements of Σ, and that the elements of $\mathcal{E}^c$ correspond to small, though not necessarily zero, elements of Σ. Indeed, even though these assumptions do not hold exactly in Simulations A, C, and D, they do hold for most of the elements in $\mathcal{E}$ and $\mathcal{E}^c$, as can be seen in Fig. 4.

Table 1: For various values of the desired false positive rate q, the false positive and negative rates and $|\widehat{\mathcal{E}}_{\gamma_n}|$ are reported for each of Simulations A–D with $n = 100$ and $p = 1000$. False positive and negative rates and q are multiplied by 100 for convenience of display. Results are averaged over 250 simulated data sets.

	$100 \times q$	$\lvert\widehat{\mathcal{E}}_{\gamma_n}\rvert$	$100\times$ False positive rate	$100\times$ False negative rate
Simulation A	0.01	552	0.03	92.0
	0.1	1614	0.18	85.6
	1	7736	1.30	74.2
	10	55712	10.8	52.8
	20	106094	20.9	42.5
	30	155747	30.8	35.0
	50	254296	50.6	23.3
Simulation B	0.01	1595	0.02	96.9
	0.1	3521	0.13	94.1
	1	11301	1.11	87.2
	10	61400	10.2	68.2
	20	111797	20.1	56.7
	30	161065	30.0	47.6
	50	258327	50.0	32.2
Simulation C	0.01	279	0.02	80.4
	0.1	1041	0.13	61.8
	1	6222	1.12	34.9
	10	51727	10.2	9.3
	20	101312	20.1	4.5
	30	150882	30.1	2.6
	50	250228	50.0	0.98
Simulation D	0.01	866	0.02	82.6
	0.1	1515	0.13	80.8
	1	6436	1.11	79.3
	10	51612	10.2	71.9
	20	101132	20.1	63.9
	30	150674	30.0	56.0
	50	250050	50.0	40.1

5.3 Comparison to Existing Approaches

We now compare the performances of the graphical lasso (Friedman et al., 2008), neighborhood selection (Meinshausen and Bühlmann, 2006), and graphical sure screening on Simulations A–D, with $n = 50$ and $p = 750$. Results are displayed in Fig. 5.

In Simulation B, the sparsity patterns of Σ and Σ^{-1} are identical. Graphical sure screening outperforms the competitors: it pays no price by thresholding the sample covariance matrix rather than estimating the precision matrix, and in fact it benefits from a reduction in variance relative to the graphical lasso and neighborhood selection. In Simulations A and C, graphical sure screening again performs quite well. This is because Assumption 1 holds approximately: elements in $\mathcal{E}$ tend to correspond to large elements of Σ, as is shown in Fig. 4. In a sense, Simulation D is a worst-case scenario for graphical sure screening: recall from Fig. 4 that many of the elements in $\mathcal{E}$ correspond to zero elements in Σ. Such edges are not detected by graphical sure screening. But nevertheless, Fig. 5 indicates that graphical lasso and neighborhood selection perform only slightly better than graphical sure screening in Simulation D: the former two approaches have trouble detecting many of the very small non-zero elements in Σ^{-1}.

Overall, the results in Fig. 5 indicate that even though graphical sure screening is extremely simple, in practice it performs quite competitively with specialized and computationally-intensive procedures for estimating a precision matrix, across a range of simulation settings.

5.4 Violations of Assumption 1

In order to obtain greater insight into violations of Assumption 1, we consider the following simulation set-up:

Simulation E: We partitioned the features into $p/3$ sets of the form

$$C_k = \{3(k-1)+1, 3(k-1)+2, 3k\}, \quad k = 1, \ldots, p/3.$$

For $j \in C_k$ and $l \in C_{k'}$ where $k \neq k'$, we set $\Sigma_{jl} = 0$. Thus, Σ and Σ^{-1} are block diagonal with $p/3$ blocks, each of dimension 3×3. The blocks of Σ take the form $\begin{pmatrix} 1 & 1 & 0 \\ 1 & 3 & 1 \\ 0 & 1 & 1 \end{pmatrix}$, and the blocks of Σ^{-1} take the form $\begin{pmatrix} 2 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 2 \end{pmatrix}$.

Figure 6 displays the marginal and conditional dependence graphs corresponding to each triple of features in Simulation E. Let $\mathcal{E}_M$ denote the marginal dependence graph, and let $\mathcal{E}_{NM} \equiv \mathcal{E} \backslash \mathcal{E}_M$ denote the set of edges in $\mathcal{E}$ that are not in $\mathcal{E}_M$. We expect graphical sure screening to successfully detect the edges in $\mathcal{E}_M$, but not the edges in $\mathcal{E}_{NM}$.

To begin, we generated $n = 50$ observations independently from a $N_p(0, \Sigma)$ distribution, with $p = 750$. We applied graphical sure screening and the graphical lasso to the resulting data, over a range of values of the tuning parameter γ_n. We let $\bar{\mathcal{E}}_{\gamma_n}$ denote the edge set estimated by the graphical lasso when applied with tuning parameter γ_n; that is, it is the set of non-zero elements in the solution to (5) when $\lambda = \gamma_n$. We choose to apply graphical lasso and graphical sure screening using the same value of the tuning parameter because, as pointed out in Section 3.1, the resulting

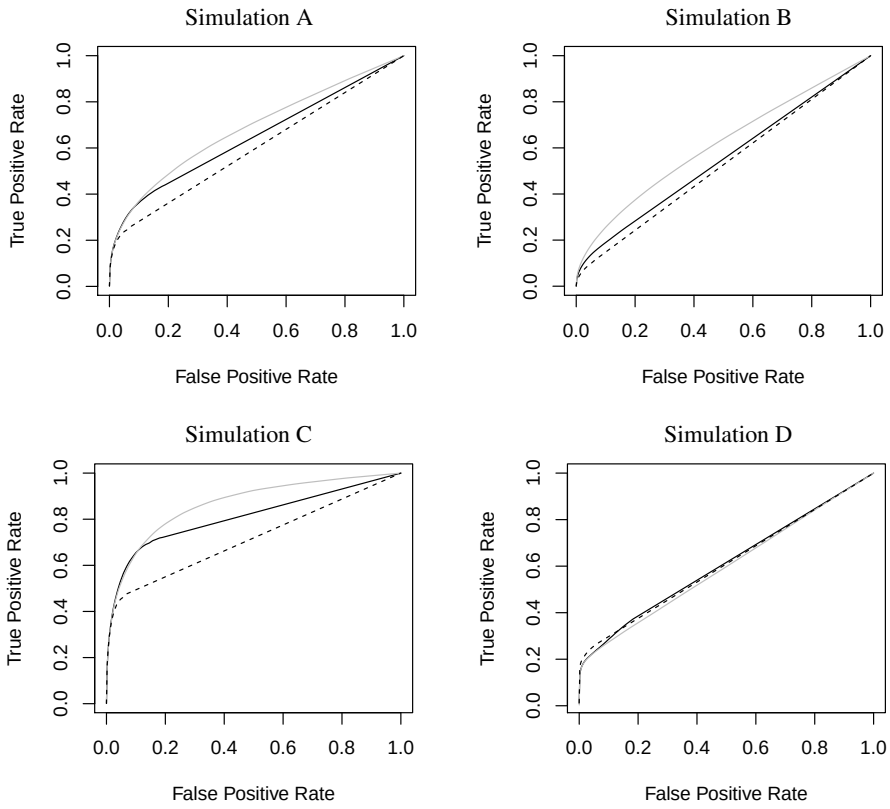

Figure 5: True positive and false positive rates, for $p = 750$ and $n = 50$. Curves are obtained by varying the tuning parameter for each method. Results for graphical lasso (—), neighborhood selection (- - -), and graphical sure screening (—) are averaged over 20 simulated data sets.

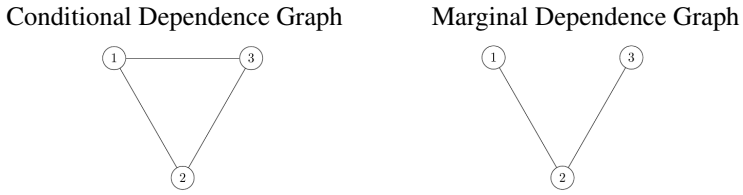

Figure 6: Conditional ($\mathcal{E}$) and marginal ($\mathcal{E}_M$) dependence graphs for each triple of features in Simulation E.

estimators have the same connected components. The results, averaged over 200 simulated data sets, are displayed in Fig. 7.

Figure 7 indicates that for any fixed value of the tuning paramer γ_n, the graphical lasso and graphical sure screening estimators are almost identical. For instance, for $\gamma_n = 0.5$, the graphical sure screening estimator averages 452.9 edges and the graphical lasso estimator averages 452.6 edges; furthermore, the former is a strict superset of the latter. We see from Fig. 7(a) that for a fixed value of γ_n, the edges estimated by the two methods to fall within $\mathcal{E}_M$ are identical. A similar result holds in Fig. 7(c) for the edges estimated to fall within $\mathcal{E}^c$.

Fig. 7(b) indicates that graphical lasso outperforms graphical sure screening at detecting edges in the set $\mathcal{E}\backslash\mathcal{E}_M$, as expected. However, this difference may not be of practical interest: a close inspection of Fig. 7(b) reveals that the graphical lasso correctly detects only five edges in $\mathcal{E}\backslash\mathcal{E}_M$ when $\gamma_n = 0.33$, corresponding to more than 5400 edges in the graphical lasso estimator. Recall that $|\mathcal{E}_M| = 500$. Thus, on average, graphical lasso detects almost no edges in $\mathcal{E}\backslash\mathcal{E}_M$ unless a network is estimated with a false discovery rate of at least $4900/5400 \approx 91\%$. Furthermore, graphical lasso detects all of the edges in $\mathcal{E}_M$ before detecting any edges in $\mathcal{E}\backslash\mathcal{E}_M$; this means that graphical lasso detects edges in $\mathcal{E}\backslash\mathcal{E}_M$ at a false discovery rate that exceeds $4895/4900 \approx 99.9\%$. Such a network estimate is unlikely to be of practical interest. For tuning parameter values that yield more reasonable false discovery rates, the recovery of true edges by graphical lasso and graphical sure screening is identical.

We now consider the results obtained with a much larger sample size of $n = 500$. In this case, Fig. 8 indicates that for a fixed value of γ_n, $\widehat{\mathcal{E}}_{\gamma_n}$ and $\bar{\mathcal{E}}_{\gamma_n}$ detect the exact same set of edges in $\mathcal{E}_M$, and almost exactly the same set of false positives in $\mathcal{E}^c$. However, not surprisingly, $\bar{\mathcal{E}}_{\gamma_n}$ detects more elements in $\mathcal{E}\backslash\mathcal{E}_M$. But inspection of Fig. 8(b) reveals that the estimators $\bar{\mathcal{E}}_{\gamma_n}$ and $\widehat{\mathcal{E}}_{\gamma_n}$ are in fact identical for all values of γ_n that yield network estimates with fewer than 500 edges. This is because the 500 edges in $\mathcal{E}_M$ are easier to detect by either method than the 250 edges in $\mathcal{E}\backslash\mathcal{E}_M$. We concede that graphical lasso outperforms graphical sure screening when estimates with greater than 500 edges are obtained. Indeed, in Fig. 8, $n \approx p$; in contrast, graphical sure screening is intended for the high-dimensional setting in which $p \gg n$.

To summarize, the results in this section indicate that even under this worst-case scenario for graphical sure screening, provided that the sample size is small relative to the number of features, then graphical sure screening and graphical lasso perform almost identically (Fig. 7) in the tuning parameter regime that is likely to be of scientific interest. When the sample size is very large, graphical lasso outperforms graphical sure screening by detecting the edges that are present in the conditional dependence graph and absent in the marginal dependence graph. However, in the high-dimensional regime for which graphical sure screening is intended, it performs competitively with graphical lasso even in this worst-case scenario. Furthermore, it does so at a small fraction of the computational cost.

6. Analysis of Gene Expression Data

We examined a gene expression data set from Spira et al. (2007), previously studied in Danaher et al. (2013), and available from the Gene Expression Omnibus (Barrett et al., 2007) at accession number GDS2771. The data consist of 22,283 microarray-derived gene expression measurements from large airway epithelial cells sampled from 97 patients with lung cancer and 90 controls. We limited our analysis to the 90 control samples, and to the 1,778 genes with the highest marginal variance. Each feature was standardized in order to have mean zero and standard deviation one.

Our goal is to compare the performances of graphical sure screening and the graphical lasso in terms of edge set recovery. Unfortunately, the underlying conditional dependence relationships in this data are unknown, so no gold standard is available.

Given the absence of a gold standard, we split the samples into two equally-sized sets, Set 1 and Set 2. We applied both the graphical lasso and graphical sure screening to each set. We refer to the resulting estimated edge sets as $\widehat{\mathcal{E}}_1^{\mathrm{GL}}$, $\widehat{\mathcal{E}}_1^{\mathrm{GRASS}}$, $\widehat{\mathcal{E}}_2^{\mathrm{GL}}$, and $\widehat{\mathcal{E}}_2^{\mathrm{GRASS}}$; tuning parameters were chosen such that $|\widehat{\mathcal{E}}_1^{\mathrm{GL}}| = |\widehat{\mathcal{E}}_2^{\mathrm{GL}}| = |\widehat{\mathcal{E}}_1^{\mathrm{GRASS}}| = |\widehat{\mathcal{E}}_2^{\mathrm{GRASS}}|$. In order to quantify the accuracy of

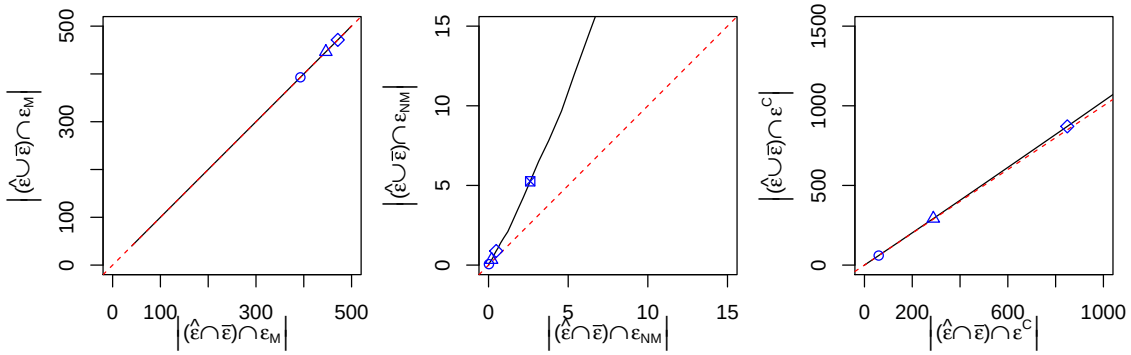

Figure 7: Results for Simulation E with $n = 50$ and $p = 750$. In each panel, the 45° line is shown in red. The symbols represent tuning parameter values of $\gamma_n = 0.5$ (○), resulting in an average of 452.9 edges for graphical sure screening and an average of 452.6 edges for graphical lasso, of which 452.6 are in the intersection; $\gamma_n = 0.45$ (△), resulting in 738.8 edges for graphical sure screening and 734.6 edges for graphical lasso, of which 734.6 are in the intersection; $\gamma_n = 0.41$ (◇), resulting in 1342.8 edges for graphical sure screening and 1320.5 edges for graphical lasso, of which 1320.5 are in the intersection; and $\gamma_n = 0.33$ (⊠), resulting in 5896.3 edges for graphical sure screening and 5404.1 edges for graphical lasso, of which 5402.0 are in the intersection.

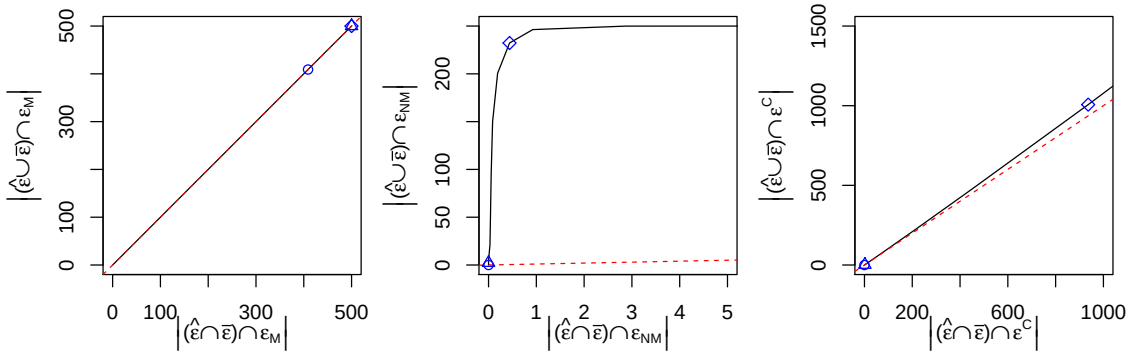

Figure 8: Results for Simulation E with $n = 500$ and $p = 750$. In each panel, the 45° line is shown in red. The symbols represent tuning parameter values of $\gamma_n = 0.55$ (○), resulting in an average of 409 edges for both graphical sure screening and the graphical lasso, of which 409 are in the intersection; $\gamma_n = 0.2$ (△), resulting in 501.8 edges for graphical sure screening and 504.1 edges for graphical lasso, of which 501.8 are in the intersection; and $\gamma_n = 0.13$ (◇), resulting in 1506.3 edges for graphical sure screening and 1668.7 edges for graphical lasso, of which 1436.8 are in the intersection.

the edges estimated on Set 2 by graphical sure screening and the graphical lasso, we first treated the edges estimated by the graphical lasso on Set 1 as the gold standard, and then we treated the edges estimated by graphical sure screening on Set 1 as the gold standard. In greater detail, we calculated (i) the accuracy of graphical sure screening on Set 2 when graphical lasso on Set 1 is treated as the gold standard, $\left|\widehat{\mathcal{E}}_1^{\mathrm{GL}} \cap \widehat{\mathcal{E}}_2^{\mathrm{GRASS}} \cap \left(\widehat{\mathcal{E}}_2^{\mathrm{GL}}\right)^c\right|$; (ii) the accuracy of graphical lasso on

Set 2 when graphical lasso on Set 1 is treated as the gold standard, $\left|\widehat{\mathcal{E}}_1^{\mathrm{GL}} \cap \widehat{\mathcal{E}}_2^{\mathrm{GL}} \cap \left(\widehat{\mathcal{E}}_2^{\mathrm{GRASS}}\right)^c\right|$; the accuracy of graphical sure screening on Set 2 when graphical sure screening on Set 1 is treated as the gold standard, $\left|\widehat{\mathcal{E}}_1^{\mathrm{GRASS}} \cap \widehat{\mathcal{E}}_2^{\mathrm{GRASS}} \cap \left(\widehat{\mathcal{E}}_2^{\mathrm{GL}}\right)^c\right|$; and the accuracy of graphical lasso on Set 2 when graphical sure screening on Set 1 is treated as the gold standard, $\left|\widehat{\mathcal{E}}_1^{\mathrm{GRASS}} \cap \widehat{\mathcal{E}}_2^{\mathrm{GL}} \cap \left(\widehat{\mathcal{E}}_2^{\mathrm{GRASS}}\right)^c\right|$. In calculating these accuracies, we only considered feature pairs for which the two estimators disagree over whether an edge is present in Set 2.

The results, averaged over 20 splits of the data, are summarized in Table 2. Regardless of whether graphical lasso or graphical sure screening on Set 1 is treated as the gold standard, graphical sure screening on Set 2 agrees better with the gold standard than does the graphical lasso on Set 2. In other words, independent data provide greater evidence for edges estimated by graphical sure screening than for edges estimated by the graphical lasso, regardless of how the independent data are evaluated. These results are due to the fact that in this high-dimensional setting, the graphical lasso estimator is high-variance and yields quite different results across random splits of the data. In contrast, the graphical sure screening estimator is lower-variance and yields similar, and likely more accurate, results across random splits of the data.

Table 2: Mean (and standard error) of accuracy of graphical lasso (GL) and graphical sure screening (GRASS) on the gene expression data, over 20 splits of the observations into Set 1 and Set 2. $|\widehat{\mathcal{E}}|$, the size of the estimated edge set, is also reported.

	GRASS as Gold Standard		GL as Gold Standard	
$\lvert\widehat{\mathcal{E}}\rvert$	GRASS Accuracy	GL Accuracy	GRASS Accuracy	GL Accuracy
47371.8 (2387.7)	3002.4 (122.4)	1428.3 (45.6)	2248.7 (85.9)	1824 (64.5)
40781.3 (2346.6)	1991.6 (90.3)	1002.9 (34.9)	1562.5 (68)	1207 (45.7)
33555.9 (2229.2)	1186.6 (60.9)	625.7 (26.9)	968.5 (48.5)	714.5 (31.4)
25942.8 (2012.8)	603.2 (33.3)	341.9 (16.2)	523.4 (27.3)	373.1 (17.4)
18540.5 (1688.2)	256.6 (16.6)	144.4 (8.1)	232.6 (14.9)	155.4 (8.7)
11903.3 (1276.8)	86.7 (5.7)	56.5 (3.5)	81.6 (5.5)	58.2 (3.7)
6647.9 (834.5)	24.2 (2.1)	12.4 (1.2)	23.6 (2)	12.7 (1.3)
3095.2 (447.3)	2.9 (0.4)	2.1 (0.3)	2.9 (0.4)	2.1 (0.3)
1142 (184.7)	0.5 (0.2)	0.1 (0.1)	0.5 (0.2)	0.1 (0.1)

Acknowledgments

This work was supported by grants from the National Science Foundation to RS and DW, a grant from the National Institutes of Health to DW, and a Sloan Research Fellowship to DW.

References

Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, and Ron Edgar. NCBI GEO:

mining tens of millions of expression profiles – database and tools update. *Nucleic Acids Research*, 35(suppl 1):D760–D765, 2007.

Peter Bühlmann and Sara van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. ISBN 978-3-642-20191-2. doi: 10.1007/978-3-642-20192-9. URL http://dx.doi.org/10.1007/978-3-642-20192-9. Methods, theory and applications.

T Tony Cai, Weidong Liu, and Harrison H Zhou. Estimating sparse precision matrix: Optimal rates of convergence and adaptive estimation. *Annals of Statistics*, 2012.

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

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:373–397, 2013.

Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

Jianqing Fan and Rui Song. Sure independence screening in generalized linear models with np-dimensionality. *The Annals of Statistics*, 38(6):3567–3604, 2010.

Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of Machine Learning Research*, 10:2013–2038, 2009.

Jianqing Fan, Yang Feng, and Rui Song. Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association*, 106(494): 544–557, 2011.

Jerome Friedman, Trevor J. Hastie, and Robert J. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Peter Hall and Hugh Miller. Using generalized correlation to effect variable selection in very high dimensional problems. *Journal of Computational and Graphical Statistics*, 18(3):533–550, 2009.

Peter Hall, DM Titterington, and Jing-Hao Xue. Tilting methods for assessing the influence of components in a classifier. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):783–803, 2009.

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

Steffen L Lauritzen. *Graphical Models*. Oxford University Press, 1996.

Runze Li, Wei Zhong, and Liping Zhu. Feature screening via distance correlation learning. *Journal of the American Statistical Association*, 107(499):1129–1139, 2012.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate Analysis*. Academic Press, 1980.

Rahul Mazumder and Trevor J. Hastie. Exact covariance thresholding into connected components for large-scale graphical lasso. *Journal of Machine Learning Research*, 13(1):781–794, 2012.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

Pradeep Ravikumar, Martin J Wainwright, Garvesh Raskutti, and Bin Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

Adam J Rothman, Peter J Bickel, Elizaveta Levina, and Ji Zhu. Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics*, 2:494–515, 2008.

Avrum Spira, Jennifer E Beane, Vishal Shah, Katrina Steiling, Gang Liu, Frank Schembri, Sean Gilman, Yves-Martine Dumas, Paul Calner, Paola Sebastiani, et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nature Medicine*, 13 (3):361–366, 2007.

Tingni Sun and Cun-Hui Zhang. Sparse matrix inversion with scaled lasso. *Journal of Machine Learning Research*, 14:3385–3418, 2013.

Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

Ming Yuan. High dimensional inverse covariance matrix estimation via linear programming. *Journal of Machine Learning Research*, 99:2261–2286, 2010.

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

Sihai Dave Zhao and Yi Li. Principled sure independence screening for Cox models with ultra-high-dimensional covariates. *Journal of Multivariate Analysis*, 105(1):397–411, 2012.

Li-Ping Zhu, Lexin Li, Runze Li, and Li-Xing Zhu. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 106(496):1464–1475, 2011.