Partial Correlation Estimation by Joint Sparse Regression Model

Jie Peng[†,*], Pei Wang[‡,*], Nengfeng Zhou[§], Ji Zhu[§]

[†] Department of Statistics, University of California, Davis, CA.

[‡] PHS, Fred Hutchinson Cancer Research Center, Seattle, WA.

[§] Department of Statistics, University of Michigan, Ann Arbor, MI.

Abstract

In this paper, we propose a computationally efficient approach —`space`(Sparse PArtial Correlation Estimation)— for selecting non-zero partial correlations under the high-dimension-low-sample-size setting. This method assumes the overall sparsity of the partial correlation matrix and employs sparse regression techniques for model fitting. We illustrate the performance of `space` by extensive simulation studies. It is shown that `space` performs well in both non-zero partial correlation selection and the identification of hub variables, and also outperforms two existing methods. We then apply `space` to a microarray breast cancer data set and identify a set of *hub genes* which may provide important insights on genetic regulatory networks. Finally, we prove that, under a set of suitable assumptions, the proposed procedure is asymptotically consistent in terms of model selection and parameter estimation (under ℓ_2 norm).

[*] equal contributors

key words: partial correlation, high-dimension-low-sample-size, sparse regression, lasso, concentration network

1 Introduction

There has been a large amount of literature on *covariance selection*: the identification and estimation of non-zero entries in the inverse covariance matrix (*concentration matrix*) starting from the seminal paper by Dempster (1972). Covariance selection is very useful in elucidating associations among a set of random variables, as under Gaussianity, non-zero entries of the concentration matrix imply conditional dependency (i.e., non-zero *partial correlation*) between corresponding variable pairs (Edward 2000). Traditional methods using greedy stepwise forward-selection or backward-elimination only work when the sample size (n) is larger than the number of variables (p) (Whittaker 1990; Edward 2000). Recently, a number of methods have been introduced to perform covariance selection for data sets with $p > n$, for example, see Meinshausen and Buhlmann (2006), Yuan and Lin (2007), Li and Gui (2006), Schafer and Strimmer (2007).

In this paper, we propose a novel approach using sparse regression techniques for covariance selection, or equivalently, the determination of non-zero partial correlations. Our work is partly motivated by the construction of *genetic regulatory networks (GRN)* based on high dimensional gene expression data. Denote the expression levels of p genes as $y_1, \cdots, y_p$. A *concentration network (or a Gaussian graphical model)* is defined as an undirected graph, in which the p vertices represent the p genes and an edge connects gene i and gene j if and only if y_i and y_j are conditionally dependent given all other variables $y_{-(i,j)} = \{y_k : 1 \leq k \neq i, j \leq p\}$. Under the assumption that $y_1, \cdots, y_p$ are from a multivariate normal distribution, y_i and y_j being conditionally dependent is equivalent to the corresponding partial correlation

$\rho^{ij} = \text{Corr}(y_i, y_j | y_{-(i,j)})$ being non-zero. The proposed method is specifically designed for the high-dimension-low-sample-size scenario. It relies on the assumption that the partial correlation matrix is sparse (i.e., most variable pairs are conditionally independent), which is reasonable for many real life problems. For instance, it has been shown that most genetic networks are intrinsically sparse (Gardner et al. 2003; Jeong et al. 2001; Tegner et al. 2003). The proposed method is also particularly powerful in the identification of *hubs*: vertices (variables) that are connected to (conditionally dependent with) many other vertices (variables). The existence of hubs is a well known phenomenon for many large networks, such as the internet, citation networks, and protein interaction networks (Newman 2003). In particular, it is widely believed that genetic pathways consist of many genes with few interactions and a few hub genes with many interactions (Barabasi and Oltvai 2004).

Another contribution of this paper is to propose a novel algorithm `active-shooting` for solving lasso type regression problems (Tibshirani 1996). This algorithm is computationally more efficient than the original `shooting` algorithm (Fu 1998; Friedman et al. 2007a). It enables us to implement the proposed procedure efficiently, such that we can conduct extensive simulation studies involving ~ 1000 variables and hundreds of samples. To our knowledge, this is the first set of intensive simulation studies for covariance selection with such high dimensions.

A few methods have also been proposed recently to perform covariance selection in the context of $p \gg n$. Similar to the method proposed in this paper, they all utilize the sparse property of the partial correlation matrix. Meinshausen and Buhlmann (2006) introduced a variable-by-variable approach for neighborhood selection via the lasso regression. They proved that neighborhoods can be consistently selected under a set of suitable assumptions. However, as regression models are fitted for each variable separately, this method has two major limitations. Firstly, it does not take

into account the intrinsic symmetry of the problem (i.e., $\rho^{ij} = \rho^{ji}$). This could result in loss of efficiency, as well as contradictory neighborhoods. Secondly, if the same penalty parameter is used for all p lasso regressions, more or less equal effort is placed on building each neighborhood. This does not seem to be the most efficient way to address the problem, unless the degree distribution of the network is nearly uniform. However, most real life networks have right skewed degree distributions, such as the *power law networks*. As observed by Schafer and Strimmer (2007), the neighborhood selection approach limits the number of edges connecting to each node. Therefore, it is not very effective in hub detection. The proposed method is based on a joint sparse regression model, which simultaneously performs neighborhood selection for all variables. It also preserves the symmetry of the problem and uses data more efficiently. We show by intensive simulation studies that our method outperforms in both model selection and hub identification. Moreover, as a joint model is used, it is easier to incorporate prior knowledge such as network topology into the model. This is discussed in Section 2.1.

Besides the regression approach mentioned above, another class of methods employ the maximum likelihood framework. Yuan and Lin (2007) proposed a penalized maximum likelihood approach which performs model selection and estimation simultaneously and ensures the positive definiteness of the estimated concentration matrix. However, their algorithm can not handle high dimensional data. The largest dimension considered by them is $p = 10$ in simulation and $p = 5$ in real data. Friedman et al. (2007b) proposed an efficient algorithm `glasso` to implement this method, such that it can be applied to problems with high dimensions. We show by extensive simulation studies that, the proposed method agian performs better than `glasso` in both model selection and hub identification. Other methods on similar topics include a threshold gradient descent (TGD) regularization procedure (Li and Gui 2006); a

shrinkage covariance estimation procedure to overcome the ill-conditioned problem of sample covariance matrix when $p > n$ (Schafer and Strimmer 2007); and a hard thresholding method to regularize the covariance matrix for families of covariance matrices satisfying suitable sparsity assumptions (Bickel and Levina 2008). The latter two methods are for estimating the covariance matrix, rather than the concentration matrix.

The rest of the paper is organized as follows. In Section 2, we describe the joint sparse regression model, its implementation and the `active-shooting` algorithm. In Section 3, the performance of the proposed method is illustrated through simulation studies and compared with that of the neighborhood selection approach and the MLE based approach `glasso`. In Section 4, the proposed method is applied to a microarray expression data set of $n = 244$ breast cancer tumor samples and $p = 1217$ genes. In Section 5, we study the asymptotic properties of this procedure. A discussion of future work is given in Section 6. Technique details are provided in the Appendices.

2 Method

2.1 Model

In this section, we describe a novel method for detecting conditionally dependent pairs among a large number of random variables based on i.i.d. samples. Suppose that, $Y = (y_1, \cdots, y_p)^T \sim \text{Normal}_\text{p}(0, \Sigma)$, where Σ is a p by p positive definite matrix. Denote the partial correlations as $\rho^{ij} = \text{Corr}(y_i, y_j | y_{-(i,j)})$ for $1 \leq i < j \leq p$ and $-(i,j) \equiv \{k : 1 \leq k \neq i, j \leq p\}$. Under the normality assumption, y_i and y_j are conditionally dependent if and only if $\rho^{ij} \neq 0$. Denote the *concentration matrix* Σ^{-1} by $(\sigma^{ij})_{p\times p}$. It is known that, $\rho^{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$. Denote $-i = \{k : 1 \leq k \neq i \leq p\}$. The following well-known result (Lemma 1) relates the estimation of partial correlations

to a regression problem.

Lemma 1 *: For $1 \leq i \leq p$, y_i is expressed as $y_i = \sum_{j \neq i} \beta_{ij} y_j + \epsilon_i$, such that ϵ_i is independent of y_{-i} if and only if $\beta_{ij} = -\frac{\sigma^{ij}}{\sigma^{ii}} = \rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}$. Moreover, for such defined β_{ij}, $\mathrm{Var}(\epsilon_i) = \frac{1}{\sigma^{ii}}, \mathrm{Cov}(\epsilon_i, \epsilon_j) = \frac{\sigma^{ij}}{\sigma^{ii}\sigma^{jj}}$.*

Remark 1 *: When we only assume that $Y = (y_1, \cdots, y_p)^T$ has a joint distribution with mean zero and covariance Σ (but not necessarily normal), partial correlation ρ^{ij} between two variables y_i and y_j are defined as the correlation between the prediction errors of the best linear predictors of y_i and y_j given $y_{-(i,j)}$. It can be shown that $\rho^{ij} = -\frac{\sigma^{ij}}{\sqrt{\sigma^{ii}\sigma^{jj}}}$, where σ^{ij} is the $(i,j)-$entry of the concentration matrix Σ^{-1}. Under normality assumption, such defined partial correlation ρ^{ij} equals to the conditional correlation $\mathrm{Corr}(y_i, y_j | y_{-(i,j)})$. Also, Lemma 1 holds even without normality assumption as long as we replace "independent" by "uncorrelated" in its statement. More details and a proof of Lemma 1 can be found in Appendix C.*

Since $\rho^{ij} = \mathrm{sign}(\beta_{ij})\sqrt{\beta_{ij}\beta_{ji}}$, the search for non-zero partial correlations can be viewed as a model selection problem under the regression setting. In this paper, we are mainly interested in the case where the dimension p is larger than the sample size n. This is a typical scenario for many real life problems. For example, high throughput genomic experiments usually result in data sets of thousands of genes for tens or at most hundreds of samples. However, many high-dimensional problems are intrinsically sparse. In the case of genetic regulatory networks, it is widely believed that most gene pairs are not directly interacting with each other. Sparsity suggests that even if the number of variables is much larger than the sample size, the effective dimensionality of the problem might still be within a tractable range. Therefore, we propose to employ sparse regression techniques by imposing the ℓ_1 penalty on a suitable loss function to tackle the high-dimension-low-sample-size problem.

Suppose $Y^k = (y_1^k, \cdots, y_p^k)^T$ are i.i.d. observations from $\text{Normal}_\text{p}(0, \Sigma)$, for $k = 1, \cdots, n$. Denote the sample of the ith variable as $Y_i = (y_i^1, \cdots, y_i^n)^T$. Based on Lemma 1, we propose the following joint loss function

$$\begin{aligned} L_n(\theta, \sigma, \mathbf{Y}) &= \frac{1}{2}\Big(\sum_{i=1}^{p} w_i ||Y_i - \sum_{j \neq i} \beta_{ij} Y_j||^2\Big) \\ &= \frac{1}{2}\Big(\sum_{i=1}^{p} w_i ||Y_i - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} Y_j||^2\Big), \end{aligned} \tag{1}$$

where $\theta = (\rho^{12}, \cdots, \rho^{(p-1)p})^T$, $\sigma = \{\sigma^{ii}\}_{i=1}^p$; $\mathbf{Y} = \{Y^k\}_{k=1}^n$; and $w = \{w_i\}_{i=1}^p$ are nonnegative weights. For example, we can choose $w_i = 1/\text{Var}(\epsilon_i) = \sigma^{ii}$ to weigh individual regressions in the joint loss function according to their residual variances, as is done in weighted regressions. We propose to estimate the partial correlations θ by minimizing a penalized loss function

$$\mathcal{L}_n(\theta, \sigma, \mathbf{Y}) = L_n(\theta, \sigma, \mathbf{Y}) + \mathcal{J}(\theta), \tag{2}$$

where the penalty term $\mathcal{J}(\theta)$ controls the overall sparsity of the final estimation of θ. In this paper, we focus on the ℓ_1 penalty (Tibshirani 1996):

$$\mathcal{J}(\theta) = \lambda ||\theta||_1 = \lambda \sum_{1 \leq i < j \leq p} |\rho^{ij}|. \tag{3}$$

The proposed joint method is referred as `space` (Sparse PArtial Correlation Estimation) hereafter. It is closely related to the *neighborhood selection approach* by Meinshausen and Buhlmann (2006) (referred as `MB` hereafter), where a lasso regression is performed separately for each variable. However, `space` has several important advantages.

(i) In `space`, sparsity is utilized for the partial correlations θ as a whole view.

However, in the neighborhood selection approach, sparsity is imposed on each neighborhood. The former treatment seems more natural, especially for networks with hubs. A prominent example is the genetic regulatory network, where master regulators are usually believed to exist and are of great interest.

(ii) According to Lemma 1, β_{ij} and β_{ji} have the same sign. The proposed method assures this sign consistency as it estimates $\{\rho^{ij}\}$ directly. However, when fitting p separate (lasso) regressions, it is possible that $\text{sign}(\widehat{\beta}_{ij})$ is different from $\text{sign}(\widehat{\beta}_{ji})$.

(iii) Furthermore, the utility of the symmetric nature of the problem allows us to reduce the number of unknown parameters in the model by almost half ($p(p+1)/2$ for `space` vs. $(p-1)^2$ for `MB`), and thus improves the estimation efficiency.

(iv) Finally, prior knowledge of the network structure are often available. The joint model is more flexible in incorporating such prior knowledge. For example, we may assign different weights w_i to different nodes according to their "importance". We have already discussed the residual variance weights, where $w_i = \sigma^{ii}$. We can also consider the weight that is proportional to the (estimated) degree of each variable, i.e., the estimated number of edges connecting with each node in the network. This would result in a preferential attachment effect which explains the cumulative advantage phenomena observed in many real life networks including GRNs (Barabasi and Albert 1999).

These advantages help enhance the performance of `space`. As illustrated by the simulation study in Section 3, the proposed joint method performs better than the neighborhood selection approach in both non-zero partial correlation detection and hub detection.

As compared to the MLE based approach `glasso` (Friedman, Hastie, and Tibshi-

rani 2007b), the simulation study in Section 3 shows that `space` always outperforms `glasso` in both edge detection and hub identification under all simulation settings that we have considered. In addition, `space` has the following advantages.

(i) The complexity of `glasso` is $O(p^3)$, while as discussed in Section 2.2, the `space` algorithm has the complexity of $\min(O(np^2), O(p^3))$, which is much faster than the algorithm of Yuan and Lin (2007) and in general should also be faster than `glasso` when $n < p$, which is the case in many real studies.

(ii) As discussed in Section 6, `space` allows for trivial generalizations to other penalties of the form of $|\rho^{ij}|^q$ rather than simply $|\rho^{ij}|$, which includes ridge and bridge (Fu 1998) or other more complicated penalties like SCAD (Fan and Li 2001). The `glasso` algorithm, on the other hand, is tied to the lasso formulation and cannot be extended to other penalties in a natural manner.

(iii) In Section 5, we have proved that our method consistently identifies the correct network neighborhood when *both* n *and* p go to ∞. As far as we are aware, no such theoretical results have been developed for the MLE approach.

(iv) As discussed earlier, `space` is more flexible in incorporating certain prior knowledge of the network structure.

Note that, in the penalized loss function (2), σ needs to be specified. We propose to estimate θ and σ by a two-step iterative procedure. Given an initial estimate $\sigma^{(0)}$ of σ, θ is estimated by minimizing the penalized loss function (2), whose implementation is discussed in Section 2.2. Then given the current estimates $\theta^{(c)}$ and $\sigma^{(c)}$, σ is updated based on Lemma 1: $1/\widehat{\sigma}^{ii} = \frac{1}{n}||Y_i - \sum_{j\neq i}\widehat{\beta}_{ij}^{(c)}Y_j||^2$, where $\widehat{\beta}_{ij}^{(c)} = (\rho^{ij})^{(c)}\sqrt{\frac{(\sigma^{jj})^{(c)}}{(\sigma^{ii})^{(c)}}}$. We then iterate between these two steps until convergence. Since $1/\sigma^{ii} = \mathrm{Var}(y_i|y_{-i}) \leq \mathrm{Var}(y_i) = \sigma_{ii}$, we can use $1/\widehat{\sigma}_{ii}$ as the initial estimate of σ^{ii}, where $\widehat{\sigma}_{ii} = \frac{1}{n-1}\sum_{k=1}^{n}(y_i^k -$

$\bar{y}_i)^2$ is the sample variance of y_i. Our simulation study shows that, it usually takes no more than three iterations for this procedure to converge.

2.2 Implementation

In this section, we discuss the implementation of the `space` procedure: that is, minimizing (2) under the ℓ_1 penalty (3). We first re-formulate the problem, such that the loss function (1) corresponds to the ℓ_2 loss of a "regression problem." We then use the `active-shooting` algorithm proposed in Section 2.3 to solve this lasso regression problem.

Given σ and positive weights w, let $\mathcal{Y} = (\tilde{Y}_1^T, ..., \tilde{Y}_p^T)^T$ be a $np \times 1$ column vector, where $\tilde{Y}_i = \sqrt{w_i} Y_i$ $(i = 1, \cdots, p)$; and let $\mathcal{X} = (\tilde{\mathcal{X}}_{(1,2)}, \cdots, \tilde{\mathcal{X}}_{(p-1,p)})$ be a np by $p(p-1)/2$ matrix, with

$$\tilde{\mathcal{X}}_{(i,j)} = (0, ..., 0, \underset{\substack{\uparrow \\ i^{th}\text{block}}}{\sqrt{\tfrac{\tilde{\sigma}^{jj}}{\tilde{\sigma}^{ii}}}\tilde{Y}_j^T}, 0, ..., 0, \underset{\substack{\uparrow \\ j^{th}\text{block}}}{\sqrt{\tfrac{\tilde{\sigma}^{ii}}{\tilde{\sigma}^{jj}}}\tilde{Y}_i^T}, 0, ..., 0)^T,$$

where $\tilde{\sigma}^{ii} = \sigma^{ii}/w_i$ $(i = 1, \cdots, p)$. Then it is easy to see that the loss function (1) equals to $\frac{1}{2}||\mathcal{Y} - \mathcal{X}\theta||_2^2$, and the corresponding ℓ_1 minimization problem is equivalent to: $\min_\theta \frac{1}{2}||\mathcal{Y} - \mathcal{X}\theta||_2^2 + \lambda||\theta||_1$. Note that, the current dimension $\tilde{n} = np$ and $\tilde{p} = p(p-1)/2$ are of a much higher order than the original n and p. This could cause serious computational problems. Fortunately, $\mathcal{X}$ is a block matrix with many zero blocks. Thus, algorithms for lasso regressions such as `lars` and `shooting` can be efficiently implemented by taking into consideration this structure (see Appendix B for the detailed implementation). To further decrease the computational cost, we develop a new algorithm `active-shooting` (Section 2.3) for the `space` model fitting. The `active-shooting` is a modification of the `shooting` algorithm (Fu 1998; Friedman

et al. 2007a), which exploits the sparse nature of the ℓ_1 penalization problem in a more efficient way, and is therefore computationally much faster. This is crucial for applying the proposed method for large p and/or n. It can be shown that the computational cost of `space` is $\min(O(np^2), O(p^3))$, which is no more than applying p individual lasso regressions as in the neighborhood selection approach. We want to point out that, the proposed method can also be implemented by `lars` (Efron, Hastie, Johnstone, and Tibshirani 2004). However, unless the exact whole solution path is needed, compared with `shooting` type algorithms, `lars` is computationally less appealing (Friedman, Hastie, Hofling, and Tibshirani 2007a).

Finally, it is important to point out that there are natural constraints on the partial correlations $\{\rho^{ij}\}$, such as $-1 \leq \rho^{ij} \leq 1$. Also, the partial correlation matrix should be positive definite. In principle, the proposed method (or more generally, the regression based method) does not guarantee the positive definiteness of the resulting estimator, while the likelihood based method by Yuan and Lin (2007) assures the positive definiteness. While admitting that this is one limitation of the proposed method, we argue that, since we are more interested in model selection than parameter estimation, we are less concerned with this issue. Moreover, in Section 5, we show that the estimated partial correlation matrix $(\widehat{\rho}^{ij})_{p\times p}$ by `space` is consistent to the actual partial correlation matrix $(\rho^{ij})_{p\times p}$ under a set of suitable assumptions. Therefore, this estimate is asymptotically positive definite. Indeed, in our simulation studies, we observe that all the estimated partial correlation matrices (there are almost 3000 such estimates in our simulation studies) by the proposed method are legitimate (that is, they are positive definite as well as having all entries between -1 and 1).

2.3 Active Shooting

In this section, we propose a computationally very efficient algorithm `active-shooting` for solving general lasso regression problems. `Active-shooting` is motivated by the `shooting` algorithm (Fu 1998), which solves the lasso regression by updating each coordinate iteratively until convergence. `Shooting` is computationally very competitive compared with the well known `lars` procedure (Efron et al. 2004). Suppose that we want to minimize an ℓ_1 penalized loss function with respect to β

$$f(\beta) = \frac{1}{2}||Y - X\beta||_2^2 + \gamma \sum_j |\beta_j|,$$

where $Y = (y_1, \cdots, y_n)^T$, $X = (x_{ij})_{n\times p} = (X_1 : \cdots : X_p)$ and $\beta = (\beta_1, \cdots, \beta_p)^T$. The `shooting` algorithm proceeds as follows:

1. Initial step (univariate *soft-shrinkage*): for $j = 1, \cdots, p$,

$$\begin{aligned} \beta_j^{(0)} &= \arg\min_{\beta_j}\{\tfrac{1}{2}||Y - \beta_j X_j||^2 + \gamma|\beta_j|\} \\ &= \text{sign}(Y^T X_j)\tfrac{(|Y^T X_j| - \gamma)_+}{X_j^T X_j}, \end{aligned} \tag{4}$$

   where $(x)_+ = xI_{(x>0)}$.

2. For $j = 1, ..., p$, update $\beta^{(old)} \longrightarrow \beta^{(new)}$:

$$\begin{aligned} \beta_i^{(new)} &= \beta_i^{(old)}, i \neq j; \\ \beta_j^{(new)} &= \arg\min_{\beta_j} \tfrac{1}{2}\left\|Y - \textstyle\sum_{i\neq j}\beta_i^{(old)} X_i - \beta_j X_j\right\|^2 + \gamma|\beta_j| \\ &= \text{sign}\left(\tfrac{\epsilon^{(old)} X_j}{X_j^T X_j} + \beta_j^{(old)}\right)\left(\left|\tfrac{\epsilon^{(old)} X_j}{X_j^T X_j} + \beta_j^{(old)}\right| - \tfrac{\gamma}{X_j^T X_j}\right)_+, \end{aligned} \tag{5}$$

   where $\epsilon^{(old)} = Y - X\beta^{(old)}$.

3. Repeat step 2 until convergence.

At each updating step of the `shooting` algorithm, we define the set of currently non-zero coefficients as the *active set*. Since under the sparsity assumption, the active set should remain small, we propose to first update the coefficients within the active set until convergence is achieved before moving on to update other coefficients. The `active-shooting` algorithm proceeds as follows:

1. Initial step: same as the initial step of `shooting`.
2. Define the current active set $\Lambda = \{k : \text{current } \beta_k \neq 0\}$.
   (2.1) For each $k \in \Lambda$, update β_k with all other coefficients fixed at the current value as in equation (5);
   (2.2) Repeat (2.1) until convergence is achieved on the active set.
3. For $i = 1$ to p, update β_i with all other coefficients fixed at the current value as in equation (5). If no β_i changes during this process, return the current β as the final estimate. Otherwise, go back to step 2.

The idea of `active-shooting` is to focus on the set of variables that is more likely to be actually in the model, and thus it improves the computational efficiency by achieving a faster convergence. We illustrate the improvement of the `active-shooting` over the `shooting` algorithm by a small simulation study of the lasso regression (generated in the same way as in Section 5.1 of Friedman et al. (2007a)). The two algorithms result in exact same solutions. However, `active-shooting` takes much fewer iterations to converge (one iteration is counted whenever an update of a β_i occurs) (Table 1). The convergence of `active-shooting` is about 10 times faster than that of the `shooting` algorithm. In particular, it takes less than 30 seconds (on average) to fit the `space` model by `active-shooting` (implemented in `c` code) for cases with 1000 variables, 200 samples and around 1000 non-zero partial correlations on a server with Dual/Core, CPU 3 GHz and 4 GB RAM. This great computational advantage enables us to conduct large scale simulation studies to examine the model performance

(Section 3).

Remark 2 *: In the initial step, instead of using the univariate soft-shrinkage estimate, we can use a previous estimate as the initial estimate if such a thing is available. For example, when iterating between* $\{\rho^{ij}\}$ *and* $\{\sigma^{ii}\}$*, we can use the previous estimate of* $\{\rho^{ij}\}$ *in the current iteration as the initial value. This can further improve the computational cost of the proposed method, as a better initial value implies a faster convergence. Moreover, in practice, often estimates are desired for a series of tuning parameters* λ*, whether it is for data exploration or for the selection of* λ*. When this is the case, a decreasing-lambda approach can be used to facilitate computation. That is, we start with the largest* λ *(which results in the smallest model), then use the resulting estimate as the initial value when fitting the model under the second largest* λ *and continue in this manner until all estimates are obtained. We have incorporated these features in the R package* `space`*.*

2.4 Selection of Tuning Parameter

In this section, we discuss the selection of the tuning parameter λ. We propose to use BIC model selection. An alternative would be v-fold cross validation, which is however computationally very demanding. We want to point out that, if v-fold cross validation is used, we should use the ordinary-least-square estimates (OLS) based on the selected model (Efron et al. 2004). Otherwise, if the shrinked estimates are used, cross validation tends to select very small λ which usually corresponds to overly large models. This is because, when there are many noise variables (which is the case under sparsity assumption), in order to select the right model, sever shrinkage is necessary. However, at the same time, parameters are overly shrinked, and thus the corresponding prediction error will be large.

As shown by [Zou,...], in lasso regression, the degrees of freedom correspond to

each λ is simply the number of nonzero coefficients in the resulting model. Thus the BIC criterion for `space` is (under the normality assumption):

$$BIC(\lambda) = n \times \Big[-\log |\widehat{\Sigma}^{-1}| + \mathrm{trace}(\widehat{\Sigma}^{-1}\mathrm{S}) \Big] + \log n \times \#\{(i,j) : 1 \le i \le j \le p, \widehat{\rho}^{ij} \neq 0\}, \tag{6}$$

where S is the sample covariance matrix, and $\widehat{\Sigma}^{-1}$ and $\{\widehat{\rho}^{ij}\}$ are estimates under λ. The same criterion is proposed by Yuan and Lin (2007) for the penalized maximum likelihood estimator. We will compare the selection of the tuning parameter for all three methods in the next section.

3 Simulation

In this section, we conduct a series of simulation experiments to examine the performance of the proposed method `space` and compare it with the neighborhood selection approach `MB` as well as the likelihood based method `glasso`. For all three methods, variables are first standardized to have (sample) mean zero and (sample) standard deviation one before model fitting. For `space`, we consider three different types of weights: (1) uniform weights: $w_i = 1$; (2) residual variance based weights: $w_i = \widehat{\sigma}^{ii}$; and (3) degree based weights: w_i is proportional to the estimated degree of y_i, i.e., $|\{j : \widehat{\rho}^{ij} \neq 0, j \neq i\}|$. The corresponding methods are referred as `space`, `space.sw` and `space.dew`, respectively. For all three `space` methods, the initial value of σ^{ii} is set to be 1. Iterations are used for these `space` methods as discussed in Section 2.1. For `space.dew`, the initial weights are taken to be 1 (i.e., equal weights). In each subsequent iteration, new weights are calculated based on the estimated residual variances (for `space.sw`) or the estimated degrees (for `space.dew`) of the previous iteration. For all three `space` methods, three iterations (that is updating between $\{\sigma^{ii}\}$ and $\{\rho^{ij}\}$) are used since the procedure converges very fast and more iterations

result in essentially the same estimator (especially in terms of the resulting model). For `glasso`, the diagonal of the concentration matrix is not penalized.

We simulate networks consisting of disjointed modules. This is done because many real life large networks exhibit a modular structure comprised of many disjointed or loosely connected components of relatively small size. For example, experiments on model organisms like yeast or bacteria suggest that the transcriptional regulatory networks have modular structures (Lee et al. 2002). Each of our network modules is set to have 100 nodes and generated according to a given degree distribution, where the *degree* of a node is defined as the number of edges connecting to it. We consider several different types of degree distributions and denote their corresponding networks by `Hub Network`, `Power-Law Network`, `Empirical Network`, `Uniform Network` and `AR Network` (details are given later). Given an undirected network with p nodes, the "partial correlation matrix" $\tilde{\rho} = (\tilde{\rho}^{ij})_{p\times p}$ is generated by

$$\tilde{\rho}^{ij} = \begin{cases} 1, & i = j; \\ 0, & i \neq j \text{ and no edge between nodes } i \text{ and } j; \\ \sim Uniform([-1,-0.5]\cup[0.5,1]), & i \neq j \text{ and an edge connecting nodes } i \text{ and } j. \end{cases} \tag{7}$$

We then rescale the non-zero elements in the above matrix to assure positive definiteness. The resulting partial correlation matrix is denoted by $\rho = (\rho^{ij})_{p\times p}$. The covariance matrix Σ is then determined by

$$\Sigma(i,j) = \rho^{-1}(i,j)/\sqrt{\rho^{-1}(i,i)\rho^{-1}(j,j)},$$

where ρ^{-1} is the inverse of ρ. Finally, i.i.d. samples $\{Y^k\}_{k=1}^n$ are generated from Normal$(0,\Sigma)$. Note that, $\Sigma(i,i) = 1$, and $\Sigma^{-1}(i,i) = \sigma^{ii} \geq 1$.

Hub networks In the first set of simulations, module networks are generated by

inserting a few hub nodes into a very sparse graph. Specifically, each module consists of three hubs with degrees around 15, and the other 97 nodes with degrees of at most 4. This setting is designed to mimic the genetic regulatory networks where there exist a few hub genes, and most other genes have only a few edges. A network consisting of 5 such modules is shown in Figure 1(a). In this network, there are $p = 500$ nodes and 568 edges. The simulated non-zero partial correlations fall in $(-0.67, -0.1] \cup [0.1, 0.67)$, with two modes around -0.28 and 0.28. Based on this network and the partial correlation matrix, we generate 50 independent data sets each consisting of $n = 250$ i.i.d. samples.

We then evaluate each method at a series of different values of the ℓ_1 penalty parameter λ. The number of total detected edges (N_t) decreases as λ increases. Figure 2(a) shows the number of correctly detected edges (N_c) vs. the number of total detected edges (N_t) averaged across the 50 independent data sets for each method. We observe that all three `space` methods (`space`, `space.sw` and `space.dew`) consistently detect more correct edges than the neighborhood selection method `MB` (except for `space.sw` when $N_t < 470$) and the likelihood based method `glasso`. `MB` performs favorably over `glasso` when N_t is relatively small (say less than 530), but performs worse than `glasso` when N_t is large. Overall, `space.dew` is the best among all methods. Specifically, when $N_t = 568$ (which is the number of true edges), `space.dew` detects 501 correct edges on average with a standard deviation 4.5 edges. The corresponding sensitivity and specificity are both 88%. On the other hand, `MB` and `glasso` detect 472 and 480 correct edges on average, respectively, when the number of total detected edges N_t being 568.

In terms of hub detection, for a given N_t, a rank (averaged over the 50 replicates) is assigned to each variable y_i based on its estimated degree (the larger the estimated degree, the smaller the rank value). We then calculate the average rank of the 15

true hub nodes for each method. The results are shown in Figure 2(b). This average rank would achieve the minimum value 8 (indicated by the grey horizontal line), if the 15 true hubs have larger estimated degrees than all other non-hub nodes. As can be seen from the figure, the average rank curves (as a function of N_t) for the three `space` methods are very close to the optimal minimum value 8 when N_t is sufficiently large. This suggests that these methods can successfully identify most of the true hubs. Indeed, for `space.dew`, when N_t equals to the number of true edges (568), the top 15 nodes with the highest estimated degrees contain at least 14 out of the 15 true hub nodes in all replicates. On the other hand, both `MB` and `glasso` identify far fewer hub nodes, as their corresponding average rank curves are much higher than the grey horizontal line.

To investigate the impact of dimensionality p and sample size n, we perform simulation studies for a larger dimension with $p = 1000$ and various sample sizes with $n = 200, 300$ and 500. The simulated network includes 10 disjointed modules of size 100 and has 1163 edges in total. Non-zero partial correlations form a similar distribution as that of the $p = 500$ network discussed above. The ROC curves for `space.dew`, `MB` and `glasso` resulted from these simulations are shown in Figure 3. When FDR is controlled at 0.05, the powers (sensitivities) for detecting correct edges are given in Table 2. From the figure and the table, we observe that the sample size has a big impact on the performance of all methods. For $p = 1000$, when the sample size increases from 200 to 300, the power of `space.dew` increases more than 20%; when the sample size is 500, `space.dew` achieves an impressive power of 96%. On the other hand, the dimensionality seems to have relatively less influence. When the total number of variables is doubled from 500 to 1000, with only 20% more samples, all three methods achieve similar powers. This is presumably because the larger network ($p = 1000$) is much sparser than the smaller network ($p = 500$) and also the

complexity of the modules remains unchanged. Finally, it is obvious from Figure 3 that, `space.dew` again performs best among the three methods.

Power-law networks Many real world networks have a *power-law (a.k.a scale-free)* degree distribution with an estimated power parameter $\alpha = 2 \sim 3$ (Newman 2003). Thus, in the second set of simulations, the module networks are generated according to a power-law degree distribution with the power law parameter $\alpha = 2.3$, as this value is close to the estimated power parameters for biological networks (Newman 2003). Figure 1(b) illustrates a network formed by five such modules with each having 100 nodes. Figure 5(a) shows the degree distribution of this network. It can be seen that there are three obvious hub nodes in this network with degrees of at least 20. The simulated non-zero partial correlations fall in the range $(-0.51, -0.08] \cup [0.08, 0.51)$, with two modes around -0.22 and 0.22. Similar to the simulation done for `Hub networks`, we generate 50 independent data sets each consisting of $n = 250$ i.i.d. samples. We then compare the number of correctly detected edges by various methods. The result is shown in Figure 4 and Table 3. On average, when the number of total detected edges equals to the number of true edges which is 495, `space.dew` detects 406 correct edges, while `MB` detects only 378 and `glasso` detects only 381 edges. In terms of hub detection, all methods can correctly identify the three hub nodes for this network.

Empirical networks In this set of simulation, five module networks (each with 100 nodes) were simulated according to an empirical degree distribution of one genetic regulatory network (Schadt et al. 2005). The whole network structure, the empirical degree distribution and the none zero partial correlation distribution are illustrated in Figure 6. Similar as before, we generate 50 independent data sets each consisting of $n = 250$ i.i.d samples. The results of all methods on this network were summarized in Figure 7 and Table 4. `space.dew` appears to be the best method. When the

total detected edges are equal to the total true edges (656), on average, `space.dew` detects 25 more correct edges than `MB`. When $FDR = 0.05$ and 0.1, the sensitivity of `space.dew` is improved around 5% over the sensitivity of `MB`.

These simulation results suggest that when the (concentration) network is reasonably sparse, we should be able to characterize the conditional dependency relationships among thousands of variables with only a couple-of-hundreds of samples. In addition, `space.dew` outperforms `MB` by at least 6% on the power of edge detection under all simulation settings when FDR is controlled at 0.05, and the improvements are even larger when FDR is controlled at a higher level say 0.1 (see Figure 3). Also, compared to `glasso`, the improvement of `space.dew` is at least 15% when FDR is controlled at 0.05, and the improvements become smaller when FDR is controlled at a higher level (see Figure 3). Moreover, the `space` methods perform much better in hub identification than both `MB` and `glasso`.

In the following, we consider networks without obvious hubs. We apply `space`, `MB` and `glasso` on networks with nearly uniform degree distributions generated by following the simulation procedure in Meinshausen and Buhlmann (2006), Yuan and Lin (2007) and Friedman et al. (2007b). For these networks, `space` performs similarly to, if not better than, `glasso` and `MB` on identifying correct edges. We also found that `space` and `MB` give much more accurate estimation for $\{\sigma^{ii}\}$ than `glasso`. The results are summarized below.

Uniform networks In this set of simulation, we generate networks similarly as the ones used in Meinshausen and Buhlmann (2006). These networks have uniform degree distribution with degrees ranging from zero to four. Figure 8(a) illustrates a network formed by five such modules with each having 100 nodes. There are in total 447 edges. Figure 8 (b) illustrates the performance of `MB`, `space` and `glasso` over 50 independent data sets each having $n = 250$ i.i.d. samples. As can be been from

this figure, all three methods perform similarly. When the total number of detected edges equals to the total number of true edges (447), `space` detects 372 true edges, `MB` detects 369 true edges and `glasso` 371 true edges.

AR networks In this simulation, we consider the so called AR network used in Yuan and Lin (2007) and Friedman et al. (2007b). Specifically, we have $\sigma^{ii} = 1$ for $i = 1, \cdots p$ and $\sigma^{i-1,i} = \sigma^{i,i-1} = 0.25$ for $i = 2, \cdots, p$. Figure 9 (a) illustrates such a network with $p = 500$ nodes and thus 499 edges. Figure 9 (b) illustrates the performance of `MB`, `space` and `glasso` over 50 independent data sets each having $n = 250$ i.i.d. samples. As can be been from this figure, all three methods again perform similarly. When the total number of detected edges equals to the total number of true edges (499), `space` detects 416 true edges, `MB` detects 417 true edges and `glasso` 411 true edges. As a slight modification of the AR network, we also consider a big circle network with: $\sigma^{ii} = 1$ for $i = 1, \cdots p$; $\sigma^{i-1,i} = \sigma^{i,i-1} = 0.3$ for $i = 2, \cdots, p$ and $\sigma^{1,p} = \sigma^{p,1} = 0.3$. Figure 10 (a) illustrates such a network with $p = 500$ nodes and thus 500 edges. Figure 10 (b) compares the performance of the three methods. When the total number of detected edges equals to the total number of true edges (500), `space`, `MB` and `glasso` detect 478, 478 and 475 true edges, respectively.

We also compare the mean squared error (MSE) of estimation of $\{\sigma^{ii}\}$. For the uniform network, the median (across all samples and λ) of the square-root MSE is $0.108, 0.113, 0.178$ for `MB`, `space` and `glasso`. These numbers are $0.085, 0.089, 0.142$ for the AR network and $0.128, 0.138, 0.233$ for the circle network. It seems that `MB` and `space` work considerably better than `glasso` on this aspect.

Tuning In the following, we consider the selection of tuning parameter λ. As discussed in Section 2.4, we propose to use BIC criterion. Specifically, for `space.dew` and `glasso`, criterion (6) will be used. As for `MB`, we treat p individual regressions

separately, and use BIC criterion for each of them to select a variable specific λ. The reason is to allow greater flexibility in model building, as different λ can be used for different regressions now. That is, for the ith regression: $y_i = \sum_{j \neq i} \beta_{ij} y_j + \epsilon_i$, we use the criterion,

$$BIC_i(\lambda) = n \times \log(RSS_i) + \log n \times \#\{j : j \neq i, \widehat{\beta}_{ij} \neq 0\},$$

where $\{\widehat{\beta}_{ij}\}$ are the lasso estimates under tuning parameter λ for the ith regression, and RSS_i is the corresponding residual sum of squares. We consider the Hub network discussed above, with $p = 1000$ and in total 1163 edges. As before, three different sample sizes $n = 200, 300, 500$ are considered. The results are reported in Table 5. All results are averaged over 25 independent replicates. As can be seen from this table, BIC tends to select large models for all three methods. In particular, among the three methods, BIC selects the largest model for `glasso` and the smallest for `MB`. In terms of edge detection, `MB` and `space.dew` perform similarly under the selected model, and `glasso` works considerably worse than the other two. The overall performance of `MB` and `space.dew` (under selected λ) improves with sample size n. However, it seems that larger models are selected when sample size increases.

We then look at hub identification under the selected tuning parameter. As before, we consider the average rank of the 30 (three for each module) true hubs. If all 30 hubs are correctly identified, the average rank of their estimated degrees should achieve the minimum value 15.5. The smaller the average rank, the better the method is in hub identification. The results are reported in Table 6. As can be seen there, it is obvious that, `space.dew` performs best, as the mean average rank of the 30 true hubs are always the smallest (except when $n = 500$, which is slightly larger than that of `glasso`). Also, `MB` is the worst in this aspects among all three methods. All three methods improve in hub detection with sample size n increases and their performance

become similar when $n = 500$.

Overall, it seems that, when BIC is used for selecting λ, `space.dew` works best under small sample size, both in edge detection and hub identification. When sample size becomes larger, the performance of `MB` becomes similarly as, or even better than that of `space.dew`.

Computational cost

4 Application

More than 500,000 women die annually of breast cancer world wide. Great efforts are being made to improve the prevention, diagnosis and treatment for breast cancer. Specifically, in the past couple of years, molecular diagnostics of breast cancer have been revolutionized by high throughput genomics technologies. A large number of gene expression signatures have been identified (or even validated) to have potential clinical usage. However, since breast cancer is a complex disease, the tumor process cannot be understood by only analyzing individual genes. There is pressing need to study the interactions between genes, which may well lead to better understanding of the disease pathologies.

In a recent breast cancer study, microarray expression experiments were conducted for 295 primary invasive breast carcinoma samples (Chang et al. 2005; van de Vijver et al. 2002). Raw array results and patient clinical outcomes for 244 of these samples are available on-line and are used in this paper. Data can be downloaded at `http://microarray-pubs.stanford.edu/wound_NKI/explore.html`. To globally characterize the association among thousands of mRNA expression levels in this group of patients, we apply the proposed method on this data set as follows. First, for each expression array, we performed the global normalization by centering the mean to 0 and scaling the median absolute deviation to 1. Then we focused on

a subset of $p = 1217$ genes/clones whose expression levels are significantly associated with the tumor progression (p-values from univariate Cox models < 0.0008, corresponding FDR $= 0.01$). We estimated the partial correlation matrix of these 1217 genes with `space.dew` for a series of λ values. The degree distribution of the inferred network is heavy tailed to the right. Specifically, when 629 edges are detected, 598 out of the 1217 genes do not connect to any other genes, while 5 genes have degrees of at least 10. The power law parameter of this degree distribution is $\alpha = 2.56$ [Figure 5(b)], which is consistent with the findings in the literature for GRNs (Newman 2003). The topology of the inferred network is shown in Figure 11(a), which supports the statement that genetic pathways consist of many genes with few interactions and a few hub genes with many interactions.

We then search for potential hub genes by ranking nodes according to their degrees. There are 11 candidate hub genes whose degrees consistently rank the highest under various λ [see Figure 11(b)]. Among these 11 genes, five are important known regulators in breast cancer. For example, *HNF3A* (also known as *FOXA1*) is a transcription factor expressed predominantly in a subtype of breast cancer, which regulates the expression of the cell cycle inhibitor $p27kip1$ and the cell adhesion molecule E-cadherin. This gene is essential for the expression of approximately 50% of estrogene-regulated genes and has the potential to serve as a therapeutic target (Nakshatri and Badve 2007). Except for *HNF3A*, all the other 10 hub genes fall in the same big network component related to cell cycle/proliferation. This is not surprising as it is well-agreed that cell cycle/proliferation signature is prognostic for breast cancer. Specifically, *KNSL6, STK12, RAD54L* and *BUB1* have been previously reported to play a role in breast cancer: *KNSL6* (also known as *KIF2C*) is important for anaphase chromosome segregation and centromere separation, which is overexpressed in breast cancer cells but expressed undetectably in other human tissues except testis (Shimo et al. 2008);

STK12 (also known as *AURKB*) regulates chromosomal segregation during mitosis as well as meiosis, whose LOH contributes to an increased breast cancer risk and may influence the therapy outcome (Tchatchou et al. 2007); RAD54L is a recombinational repair protein associated with tumor suppressors BRCA1 and BRCA2, whose mutation leads to defect in repair processes involving homologous recombination and triggers the tumor development (Matsuda et al. 1999); in the end, BUB1 is a spindle checkpoint gene and belongs to the BML-1 oncogene-driven pathway, whose activation contributes to the survival life cycle of cancer stem cells and promotes tumor progression. The roles of the other six hub genes in breast cancer are worth of further investigation. The functions of all hub genes are briefly summarized in Table 7.

5 Asymptotics

In this section, we show that under appropriate conditions, the `space` procedure achieves both model selection consistency and estimation consistency when σ is fixed at the truth $\overline{\sigma}$ (thus sometimes σ is omitted in the notation). Use $\overline{\theta}$ and $\overline{\sigma}$ to denote the true parameters of θ and σ. As discussed in Section 2.1, when σ is given, θ is estimated by solving the following ℓ_1 penalization problem:

$$\widehat{\theta^{\lambda_n}} = \arg\min_{\theta} L_n(\theta, \mathbf{Y}) + \lambda_n ||\theta||_1. \tag{8}$$

In this section, the *loss function* $L_n(\theta, \sigma, \mathbf{Y}) := \frac{1}{n}\sum_{k=1}^{n} L(\theta, \sigma, Y^k)$, with

$$L(\theta, \sigma, Y) \;=\; \sum_{i=1}^{p} w_i (y_i - \sum_{j \neq i} \sqrt{\sigma^{jj}/\sigma ii}\, \rho^{ij} y_j)^2 = \sum_{i=1}^{p} \tilde{w}_i (\tilde{y}_i - \sum_{j \neq i} \rho^{ij} \tilde{y}_j)^2, \tag{9}$$

where $\tilde{y}_i = \sqrt{\sigma^{ii}} y_i$, $\tilde{w}_i = w_i/\sigma^{ii}$.

We first state regularity conditions that are needed for the proof. Define $\mathcal{A} =$

$\{(i,j) : \bar{\rho}^{ij} \neq 0\}$.

C0: The weights satisfy $0 < w_0 \leq \min_i\{w_i\} \leq \max_i\{w_i\} \leq w_\infty < \infty$

C1: There exist constants $0 < \Lambda_{\min}(\bar{\theta}) \leq \Lambda_{\max}(\bar{\theta}) < \infty$, such that the true covariance $\overline{\Sigma} = \overline{\Sigma}(\bar{\theta}, \bar{\sigma})$ satisfies: $0 < \Lambda_{\min}(\bar{\theta}) \leq \lambda_{min}(\overline{\Sigma}) \leq \lambda_{\max}(\overline{\Sigma}) \leq \Lambda_{\max}(\bar{\theta}) < \infty$.

C2: There exist a constant $\delta < 1$ such that for all $(i,j) \notin \mathcal{A}$

$$\left| \overline{L}''_{ij,\mathcal{A}}(\bar{\theta}, \bar{\sigma}) \left[\overline{L}''_{\mathcal{A},\mathcal{A}}(\bar{\theta}, \bar{\sigma}) \right]^{-1} \text{sign}(\bar{\theta}_{\mathcal{A}}) \right| \leq \delta (< 1).$$

Condition C0 says that the weights are bounded away from zero and infinity. Condition C1 assumes that the eigenvalues of the true covariance matrix $\overline{\Sigma}$ are bounded away from zero and infinity. Condition C2 corresponds to the *incoherence condition* in Meinshausen and Buhlmann (2006), which plays a crucial role in proving model selection consistency of l_1 penalization problems. An "almost sufficient" condition for C2 is the following *network stability* condition.

Proposition 1 *(network stability) If there exits $\eta > 0$, such that*

$$\widehat{\theta}^{\eta} = \arg\min_{\theta} E_{\bar{\theta}} \left(L(\theta, \bar{\sigma}, Y) \right) + \eta ||\theta||_1,$$

satisfying $\text{sign}(\widehat{\theta}^{\eta}) = \text{sign}(\bar{\theta})$, *then for all* $(i,j) \notin \mathcal{A}$

$$\left| \overline{L}''_{ij,\mathcal{A}}(\bar{\theta}, \bar{\sigma}) \left[\overline{L}''_{\mathcal{A},\mathcal{A}}(\bar{\theta}, \bar{\sigma}) \right]^{-1} \text{sign}(\bar{\theta}_{\mathcal{A}}) \right| \leq 1.$$

Its proof is similar as the proof of the *neighborhood stability* condition in Meinshausen and Buhlmann (2006) and is given in Appendix A.1.

We then state notations used in the main results. Let $q_n = |\mathcal{A}|$ denote the number of nonzero partial correlations and let $\{s_n\}$ be a positive sequence of real numbers such that for any $(i,j) \in \mathcal{A}$: $|\overline{\rho}^{ij}| \geq s_n$. Note that, s_n can be viewed as the signal size. We follow the similar strategy as in Meinshausen and Buhlmann (2006) and Massam et al. (2007) in deriving the asymptotic result: (i) First prove estimation consistency and sign consistency for the restricted penalization problem with $\theta_{\mathcal{A}^c} = 0$ (Theorem 1). We employ the method of the proof of Theorem 1 in Fan and Peng (2004); (ii) Then prove that with probability tending to one, no wrong edge is selected (Theorem 2); (iii) The final consistency result then follows (Theorem 3).

Theorem 1 *(consistency of the restricted problem) Suppose that the loss function $L(\cdot,\cdot)$ is as defined in (9) with σ fixed at $\overline{\sigma}$, and conditions C0-C1 are satisfied. Suppose further that $q_n \sim o(\sqrt{n/\log n})$, $\lambda_n n^{1/4} \to \infty$ and $\sqrt{q_n}\lambda_n \sim o(1)$, as $n \to \infty$. Then there exists a constant $C(\overline{\theta}) > 0$, such that for any $\eta > 0$, the following events hold with probability at least $1 - O(n^{-\eta})$:*

- *there exists a solution $\widehat{\theta}^{\mathcal{A},\lambda_n}$ of the restricted problem:*

$$\min_{\theta:\theta_{\mathcal{A}^c}=0} L_n(\theta, \overline{\sigma}, \mathbf{Y}) + \lambda_n ||\theta||_1. \tag{10}$$

- *(estimation consistency) any solution $\widehat{\theta}^{\mathcal{A},\lambda_n}$ of the restricted problem (10) satisfies:*

$$||\widehat{\theta}^{\mathcal{A},\lambda_n} - \overline{\theta}_{\mathcal{A}}||_2 \leq C(\overline{\theta})\sqrt{q_n}\lambda_n.$$

- *(sign consistency) if further suppose that the signal sequence satisfies: $\frac{s_n}{\sqrt{q_n}\lambda_n} \to \infty$, $n \to \infty$, then $\text{sign}(\widehat{\theta}_{ij}^{\mathcal{A},\lambda_n}) = \text{sign}(\overline{\theta}_{ij})$, for all $1 \leq i < j \leq p$.*

Theorem 2 *Suppose that the loss function $L(\cdot,\cdot)$ is as defined in (9) with σ fixed at $\overline{\sigma}$, and conditions C0-C2 are satisfied. Suppose further that $p = O(n^{\kappa})$ for some*

$\kappa \geq 0$; $q_n \sim o(\sqrt{n/\log n})$, $\lambda_n n^{1/4} \to \infty$ *and* $\sqrt{q_n}\lambda_n \sim o(1)$, *as* $n \to \infty$. *Then for any* $\eta > 0$, *for* n *sufficiently large*

$$P_{\overline{\theta}}\left(\max_{(i,j)\in\mathcal{A}^c} |L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \mathbf{Y})| < \lambda_n\right) \geq 1 - O(n^{-\eta}).$$

Theorem 3 *Assume the same conditions of Theorem 2. Then there exists a constant* $C(\overline{\theta}) > 0$, *such that for any* $\eta > 0$, *the following events hold with probability at least* $1 - O(n^{-\eta})$:

- *there exists a solution* $\widehat{\theta}^{\lambda_n}$ *of the* ℓ_1 *penalization problem (8).*
- *(estimation consistency): any solution* $\widehat{\theta}^{\lambda_n}$ *of (8) satisfies:*

$$||\widehat{\theta}^{\lambda_n} - \bar{\theta}||_2 \leq C(\bar{\theta})(\sqrt{q_n}\lambda_n).$$

- *Model selection consistency/sign consistency:*

$$\text{sign}(\widehat{\theta}_{ij}^{\lambda_n}) = \text{sign}(\bar{\theta}_{ij}), \text{ for all } 1 \leq i < j \leq p.$$

Proofs of these theorems are given in Appendix A.2.

6 Discussion

In this paper, we propose a joint regression model with sparse penalty for selecting non-zero partial correlations under the high-dimension-low-sample-size setting. By controlling the overall sparsity of the partial correlation matrix, the proposed `space` method is able to automatically adjust for different neighborhood sizes and thus to utilize data more effectively. It also explicitly employs the symmetry among the partial correlations, which helps to improve the estimation efficiency. Moreover, the

joint model makes it easy to incorporate prior knowledge. We develop a fast algorithm `active-shooting` to implement the proposed procedure. With extensive simulation studies, we demonstrate that `space` achieves good power in non-zero partial correlation selection as well as hub identification, and also performs favorably compared to two existing methods. The impact of the sample size and dimensionality on its performance has been examined on simulation examples as well. We then apply `space` on a microarray data set of 1217 genes from 244 breast cancer tumor samples, and find 11 candidate hubs, of which 5 are known breast cancer related regulators. Finally, we show consistency (in terms of model selection and estimation) of the proposed procedure under suitable regularity and sparsity conditions.

In the proposed method, the ℓ_1 penalty can be readily replaced with other sparse penalties to achieve additional benefits. One option is the elastic net penalty by Zou and Hasite (2005):

$$\mathcal{J}(\theta) = \lambda_1 \sum_{1 \le i < j \le p} |\rho^{ij}| + \lambda_2 \sum_{1 \le i < j \le p} (\rho^{ij})^2.$$

Here the ℓ_2 term forces a grouping effect such that variables that are highly correlated tend to be in and out of the model simultaneously. The lasso regression, in contrast, tends to select only one variable from a highly correlated group. Since co-regulated genes often co-express, elastic net could be especially preferred over lasso for hub identification. This has been supported by some preliminary simulation studies (results not shown). It is easy to see that the `active-shooting` algorithm described in this paper can be easily extended to solve elastic net and some other penalized regression problems such as SCAD (Fan and Li 2001).

The choice of the tuning parameter λ is of great importance. Since the `space` method uses a lasso criterion, methods that have been developed for selecting the tuning parameter for lasso can also be applied to `space`, such as the GCV in Tibshirani

(1996), the CV in Fan and Li (2001), the AIC in Buhlmann (2006) and the BIC in Zou et al. (2007). Several methods have also been proposed for selecting the tuning parameter in the setting of covariance estimation, for example, the MSE based criterion in Schafer and Strimmer (2007), the likelihood based method in Huang et al. (2006) and the cross-validation and bootstrap methods in Li and Gui (2006). We consider BIC criterions in this paper, and show by simulation studies that, the proposed method works reasonably well in edge detection and hub identification under the selected tuning parameter. It also works favorably over the other two methods (all using BIC criterion) on these two aspects, at least when sample size is relatively small. Our method can also be considered as an exploratory data analysis tool, hence we have presented numerical results on a wide range of λ values. According to the simulation results, the choice of λ does not seem to be crucial in terms of hub identification for the proposed `space` methods. As can be seen in Figure 2(b), hub identification is very consistent across a wide range of λ values. V-fold cross validation is an alternative we will explore in the future.

We want to point out that, although in the asymptotic analysis, we only derive consistency results when $\bar{\sigma}$ is known, the techniques used in that proof can also be applied to the cases in which $\bar{\sigma}$ is estimated, given that more stringent sparsity conditions are provided. In addition, due to exponential small tails of the probabilistic bounds, model selection consistency can be easily extended when the network consists of N disjointed components with $N = O(n^{\alpha})$ for some $\alpha \geq 0$, as long as the size and the number of edges of each component satisfy the corresponding conditions in Theorem 2.

Finally, the R package *space* is available on http://cran.r-project.org/.

References

Barabasi, A. L., and Albert, R. (1999), "Emergence of Scaling in Random Networks," *Science*, 286, 509–512.

Barabasi, A. L., and Oltvai, Z. N. (2004), "Network Biology: Understanding the Cells Functional Organization," *Nature Reviews Genetics*, 5, 101–113.

Bickel, P., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

Buhlmann, P. (2006), "Boosting for High-dimensional Linear Models," *The Annals of Statistics*, 34, 559–583.

Chang, H. Y., Nuyten, D. S. A., Sneddon, J. B., Hastie, T., Tibshirani, R., Sorlie, T., Dai, H., He,Y., van't Veer, L., Bartelink, H., and et al (2005), "Robustness, Scalability, and Integration of a Wound Response Gene Expression Signature in Predicting Survival of Human Breast Cancer Patients," *PNAS*, 8;102(10): 3738-43

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

Edward, D. (2000), *Introduction to Graphical Modelling* (2nd ed.), New York: Springer.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

Fan, J., and Peng, H. (2004), "Nonconcave Penalized Likelihood with a Diverging Number of Paramters," *Annals of statistics*, 32(3), 928–961.

Friedman, J., Hastie, T., Hofling, H., and Tibshirani, R. (2007a), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1(2), 302-332.

Friedman, J., Hastie, T., and Tibshirani, R. (2007b), "Sparse Inverse Covariance Estimation with the Graphical Lasso," *Biostatistics* doi:10.1093/biostatistics/kxm045.

Fu, W. (1998), "Penalized Regressions: the Bridge vs the Lasso," *Journal of Computational and Graphical Statistics*, 7(3), 397–416.

Gardner, T. S., Bernardo, D. di, Lorenz, D., and Collins, J. J. (2003), "Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling," *Science*, 301, 102–105.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006), "Covariance Matrix Selection and Estimation via Penalised Normal Likelihood," *Biometrika*, 93, 85–98.

Jeong, H., Mason, S. P., Barabasi, A. L., and Oltvai, Z. N. (2001), "Lethality and Centrality in Protein Networks," *Nature*, 411, 41–42.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., Zeitlinger, J., and et al. (2002), "Transcriptional Regulatory Networks in Saccharomyces Cerevisiae," *Science*, 298, 799–804.

Li, H., and Gui, J. (2006), "Gradient Directed Regularization for Sparse Gaussian Concentration Graphs, with Applications to Inference of Genetic Networks," *Biostatitics*, 7(2) ,302–317.

Massam, H., Paul, D., and Rajaratnam, B. (2007), "Penalized Empirical Risk Minimization Using a Convex Loss Function and ℓ_1 Penalty," *Unpublished Manuscript*.

Matsuda, M., Miyagawa, K., Takahashi, M., Fukuda, T., Kataoka, T., Asahara,

T., Inui, H., Watatani, M., Yasutomi, M., Kamada, N., Dohi, K., and Kamiya, K. (1999), "Mutations in the Rad54 Recombination Gene in Primary Cancers," *Oncogene*, 18, 3427–3430.

Meinshausen, N., and Buhlmann, P. (2006), "High Dimensional Graphs and Variable Selection with the Lasso, " *Annals of Statistics*, 34, 1436-1462.

Nakshatri, H., and Badve, S. (2007), "FOXA1 as a Therapeutic Target for Breast Cancer," *Expert Opinion on Therapeutic Targets*, 11, 507–514.

Newman, M. (2003), "The Structure and Function of Complex Networks," *Society for Industrial and Applied Mathematics*, 45(2), 167–256.

Schafer, J., and Strimmer, K. (2007), "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics," *Statistical Applications in Genetics and Molecular Biology*, 4(1), Article 32.

Shimo, A., Tanikawa, C., Nishidate, T., Lin, M., Matsuda, K., Park, J., Ueki, T., Ohta, T., Hirata, K., Fukuda, M., Nakamura, Y., and Katagiri, T. (2008), "Involvement of Kinesin Family Member 2C/Mitotic Centromere-Associated Kinesin Overexpression in Mammary Carcinogenesis," *Cancer Science*, 99(1), 62–70.

Tchatchou, S., Wirtenberger, M., Hemminki, K., Sutter, C., Meindl, A., Wappenschmidt, B., Kiechle, M., Bugert, P., Schmutzler, R., Bartram, C., and Burwinkel, B. (2007), "Aurora Kinases A and B and Familial Breast Cancer Risk," *Cancer Letters*, 247(2), 266–272.

Tegner, J., Yeung, M. K., Hasty, J., and Collins, J. J. (2003), "Reverse Engineering Gene Networks: Integrating Genetic Perturbations with Dynamical Modeling," *Proceedings of the National Academy of Sciences USA*, 100, 5944–5949.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal*

of the Royal Statistical Society, Series B, 58, 267–288.

van de Vijver, M. J., He, Y. D., van ′t Veer, L. J., Dai, H., Hart, A. A.M., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J., and et al. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *New England Journal of Medicine*, 347, 1999–2009.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics,* Wiley.

Yuan, M., and Lin, Y. (2007), "Model Selection and Estimation in the Gaussian Graphical Model," *Biometrika*, 94(1), 19–35.

Zou, H., and Hasite, T. (2005), "Regularization and Variable Selection via Elastic Net," *Journal of the Royal Statistical Society, Series B*, 67(2), 301–320.

Zou, H., Hasite, T., and Tibshirani, R. (2007), "On the Degrees of Freedom of the Lasso," *The Annals of Statistics*, 35, 2173–2192.

Appendix A.1

In this section, we list properties of the loss function (9) which are used for the proof of the main results.

A1: for all θ, σ and $Y \in \mathcal{R}^p$, $L(\theta, \sigma, Y) \geq 0$.

A2: for any $Y \in \mathcal{R}^p$ and any $\sigma > 0$, $L(\cdot, \sigma, Y)$ is convex in θ; and with probability one, $L(\cdot, \sigma, Y)$ is strictly convex.

A3: for $1 \leq i < j \leq p$

$$\overline{L}'_{ij}(\bar{\theta}, \bar{\sigma}) := E_{(\bar{\theta}, \bar{\sigma})} \left(\frac{\partial L(\theta, \sigma, Y)}{\partial \rho^{ij}} \Big|_{\theta=\bar{\theta}, \sigma=\bar{\sigma}} \right) = 0.$$

A4: for $1 \leq i < j \leq p$ and $1 \leq k < l \leq p$,

$$(\overline{L}''(\theta, \sigma))_{ij,kl} := E_{(\theta,\sigma)} \left(\frac{\partial^2 L(\theta, \sigma, Y)}{\partial \rho^{ij} \rho^{kl}} \right) = \frac{\partial}{\partial \rho^{kl}} \left[E_{(\theta,\sigma)} \left(\frac{\partial L(\theta, \sigma, Y)}{\partial \rho^{ij}} \right) \right],$$

and $\overline{L}''(\bar{\theta}, \bar{\sigma})$ is positive semi-definite.

If assuming C0-C1, then we have

B0 : There exist constants $0 < \bar{\sigma}_0 \leq \bar{\sigma}_\infty < \infty$ such that: $0 < \bar{\sigma}_0 \leq \min\{\bar{\sigma}^{ii} : 1 \leq i \leq p\} \leq \max\{\bar{\sigma}^{ii} : 1 \leq i \leq p\} \leq \bar{\sigma}_\infty$.

B1 : There exist constants $0 < \Lambda^L_{\min}(\bar{\theta}) \leq \Lambda^L_{\max}(\bar{\theta}) < \infty$, such that

$$0 < \Lambda^L_{\min}(\bar{\theta}) \leq \lambda_{min}(\overline{L}''(\bar{\theta})) \leq \lambda_{\max}(\overline{L}''(\bar{\theta})) \leq \Lambda^L_{\max}(\bar{\theta}) < \infty$$

B1.1 : There exists a constant $K(\bar{\theta}) < \infty$, such that for all $1 \leq i < j \leq p$, $\overline{L}''_{ij,ij}(\bar{\theta}) \leq K(\bar{\theta})$.

B1.2 : There exist constants $M_1(\bar{\theta}), M_2(\bar{\theta}) < \infty$, such that for any $1 \leq i < j \leq p$

$$\mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(L'_{n,ij}(\bar{\theta},\bar{\sigma},Y)) \leq M_1(\bar{\theta}),\ \mathrm{Var}_{(\bar{\theta},\bar{\sigma})}(L''_{n,ij,ij}(\bar{\theta},\bar{\sigma},Y)) \leq M_2(\bar{\theta}).$$

B1.3 : There exists a constant $0 < g(\bar{\theta}) < \infty$, such that for all $(i,j) \in \mathcal{A}$

$$\overline{L}''_{ij,ij}(\bar{\theta},\bar{\sigma}) - \overline{L}''_{ij,\mathcal{A}_{ij}}(\bar{\theta},\bar{\sigma}) \left[\overline{L}''_{\mathcal{A}_{ij},\mathcal{A}_{ij}}(\bar{\theta},\bar{\sigma})\right]^{-1} \overline{L}''_{\mathcal{A}_{ij},ij}(\bar{\theta},\bar{\sigma}) \geq g(\bar{\theta}),$$

where $\mathcal{A}_{ij} = \mathcal{A}/\{(i,j)\}$.

B1.4 : There exists a constant $M(\bar{\theta}) < \infty$, such that for any $(i,j) \in \mathcal{A}^c$

$$||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}||_2 \leq M(\bar{\theta}).$$

B2 There exists a constant $K_1(\bar{\theta}) < \infty$, such that for any $1 \leq i \leq j \leq p$, $\lambda_{\max}(E_{\bar{\theta}}(\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T)) \leq K_1(\bar{\theta})$, where $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_p)^T$.

B0 follows from C1 immediately. B1.1–B1.4 are direct consequences of B1. B2 follows from B1 and Gaussianity.

proof of A1: obvious.

proof of A2: obvious.

proof of A3: denote the residual for the ith term by

$$e_i(\theta,\sigma) = \tilde{y}_i - \sum_{j \neq i} \rho^{ij}\tilde{y}_j.$$

Then evaluated at the true parameter values $(\bar{\theta},\bar{\sigma})$, we have $e_i(\bar{\theta},\bar{\sigma})$ independent with $\tilde{y}_{(-i)}$ and $E_{(\bar{\theta},\bar{\sigma})}(e_i(\bar{\theta},\bar{\sigma})) = 0$. It is easy to show

$$\frac{\partial L(\theta,\sigma,Y)}{\partial \rho^{ij}} = -2\tilde{w}_i e_i(\theta,\sigma)\tilde{y}_j - 2\tilde{w}_j e_j(\theta,\sigma)\tilde{y}_i.$$

This proves A3.

proof of A4: see the proof of B1.

proof of B1: Denote $\tilde{y} = (\tilde{y}_1, \cdots, \tilde{y}_p)^T$, and $\tilde{x} = (\tilde{x}_{(1,2)}, \tilde{x}_{(1,3)}, \cdots, \tilde{x}_{(p-1,p)})$ with $\tilde{x}_{(i,j)} = (0, \cdots, 0, \tilde{y}_j, \cdots, \tilde{y}_i, 0, \cdots, 0)^T$. Then the loss function (9) can be written as $L(\theta, \sigma, Y) = ||\tilde{w}(\tilde{y} - \tilde{x}\theta)||_2^2$, with $\tilde{w} = diag(\sqrt{\tilde{w}}_1, \cdots, \sqrt{\tilde{w}}_p)$. Thus $\overline{L}''(\theta, \sigma) = E_{(\theta,\sigma)}\left[\tilde{x}^T\tilde{w}^2\tilde{x}\right]$ (this proves A4). Let $d = p(p-1)/2$, then $\tilde{x}$ is a p by d matrix. Denote its ith row by x_i^T $(1 \le i \le p)$. Then for any $a \in \mathcal{R}^d$, with $||a||_2 = 1$, we have

$$a^T\overline{L}''(\bar{\theta})a = E_{\bar{\theta}}(a^T\tilde{x}^T\tilde{w}^2\tilde{x}a) = E_{\bar{\theta}}\left(\sum_{i=1}^{p}\tilde{w}_i(x_i^Ta)^2\right).$$

Index the elements of a by $a = (a_{(1,2)}, a_{(1,3)}, \cdots, a_{(p-1,p)})^T$, and for each $1 \le i \le p$, define $a_i \in \mathcal{R}^p$ by $a_i = (a_{(1,i)}, \cdots, a_{(i-1,i)}, 0, a_{(i,i+1)}, \cdots, a_{(i,p)})^T$. Then by definition $x_i^Ta = \tilde{y}^Ta_i$. Also note that $\sum_{i=1}^{p}||a_i||_2^2 = 2||a||_2^2 = 2$. This is because, for $i \ne j$, the jth entry of a_i appears exactly twice in a. Therefore

$$a^T\overline{L}''(\bar{\theta})a = \sum_{i=1}^{p}\tilde{w}_iE_{\bar{\theta}}\left(a_i^T\tilde{y}\tilde{y}^Ta_i\right) = \sum_{i=1}^{p}\tilde{w}_ia_i^T\tilde{\Sigma}a_i \ge \sum_{i=1}^{p}\tilde{w}_i\lambda_{\min}(\tilde{\Sigma})||a_i||_2^2 \ge 2\tilde{w}_0\lambda_{\min}(\tilde{\Sigma}),$$

where $\tilde{\Sigma} = \mathrm{Var}(\tilde{y})$ and $\tilde{w}_0 = w_0/\bar{\sigma}_\infty$. Similarly $a^T\overline{L}''(\bar{\theta})a \le 2\tilde{w}_\infty\lambda_{\max}(\tilde{\Sigma})$, with $\tilde{w}_\infty = w_\infty/\bar{\sigma}_0$. By C1, $\tilde{\Sigma}$ has bounded eigenvalues, thus B1 is proved.

proof of B1.1: obvious.

proof of B1.2: note that $\mathrm{Var}_{(\bar\theta,\bar\sigma)}(e_i(\bar\theta,\bar\sigma)) = 1/\bar\sigma^{ii}$ and $\mathrm{Var}_{(\bar\theta,\bar\sigma)}(\tilde y_i) = \bar\sigma^{ii}$. Then for any $1 \le i < j \le p$, by Cauchy-Schwartz

$$\begin{aligned}
\mathrm{Var}_{(\bar\theta,\bar\sigma)}(L'_{n,ij}(\bar\theta,\bar\sigma,Y)) &= \mathrm{Var}_{(\bar\theta,\bar\sigma)}(-2\tilde w_i e_i(\bar\theta,\bar\sigma)\tilde y_j - 2\tilde w_j e_j(\bar\theta,\bar\sigma)\tilde y_i) \\
&\le 4\{E_{(\bar\theta,\bar\sigma)}(\tilde w_i^2 e_i^2(\bar\theta,\bar\sigma)\tilde y_j^2) + E_{(\bar\theta,\bar\sigma)}(\tilde w_j^2 e_j^2(\bar\theta,\bar\sigma)\tilde y_i^2) \\
&+ 2\sqrt{\tilde w_i^2\tilde w_j^2 E_{(\bar\theta,\bar\sigma)}(e_i^2(\bar\theta,\bar\sigma)\tilde y_j^2)E_{(\bar\theta,\bar\sigma)}(e_j^2(\bar\theta,\bar\sigma)\tilde y_i^2)}\} \\
&= 4\left\{\frac{w_i^2\bar\sigma^{jj}}{(\bar\sigma^{ii})^3} + \frac{w_j^2\bar\sigma^{ii}}{(\bar\sigma^{jj})^3} + 2\frac{w_i w_j}{\bar\sigma^{ii}\bar\sigma^{jj}}\right\}
\end{aligned}$$

The RHS is bounded because of C0 and B0.

proof of B1.3: for $(i,j) \in \mathcal{A}$, denote

$$D := \overline{L}''_{ij,ij}(\bar\theta,\bar\sigma) - \overline{L}''_{ij,\mathcal{A}_{ij}}(\bar\theta,\bar\sigma)\left[\overline{L}''_{\mathcal{A}_{ij},\mathcal{A}_{ij}}(\bar\theta,\bar\sigma)\right]^{-1}\overline{L}''_{\mathcal{A}_{ij},ij}(\bar\theta,\bar\sigma).$$

Then D^{-1} is the (ij,ij) entry in $\left[\overline{L}''_{\mathcal{A},\mathcal{A}}(\bar\theta)\right]^{-1}$. Thus by B1, D^{-1} is positive and bounded from above, so D is bounded away from zero.

proof of B1.4: note that $||\overline{L}''_{ij,\mathcal{A}}(\bar\theta)[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1}||_2^2 \le ||\overline{L}''_{ij,\mathcal{A}}(\bar\theta)||_2^2\lambda_{\max}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-2})$. By B1, $\lambda_{\max}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-2})$ is bounded from above, thus it suffices to show that $||\overline{L}''_{ij,\mathcal{A}}(\bar\theta)||_2^2$ is bounded. Since $(i,j) \in \mathcal{A}^c$, define $\mathcal{A}^+ := (i,j)\cup\mathcal{A}$. Then $\overline{L}''_{ij,ij}(\bar\theta) - \overline{L}''_{ij,\mathcal{A}}(\bar\theta)[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar\theta)$ is the inverse of the $(1,1)$ entry of $\overline{L}''_{\mathcal{A}^+,\mathcal{A}^+}(\bar\theta)$. Thus by B1, it is bounded away from zero. Therefore by B1.1, $\overline{L}''_{ij,\mathcal{A}}(\bar\theta)[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar\theta)$ is bounded from above. Since $\overline{L}''_{ij,\mathcal{A}}(\bar\theta)[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1}\overline{L}''_{\mathcal{A},ij}(\bar\theta) \ge ||\overline{L}''_{ij,\mathcal{A}}(\bar\theta)||_2^2\lambda_{\min}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1})$, and by B1, $\lambda_{\min}([\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar\theta)]^{-1})$ is bounded away from zero, we have $||\overline{L}''_{ij,\mathcal{A}}(\bar\theta)||_2^2$ bounded from above.

proof of B2: the (k,l)-th entry of the matrix $\tilde y_i\tilde y_j\tilde y\tilde y^T$ is $\tilde y_i\tilde y_j\tilde y_k\tilde y_l$, for $1 \le k < l \le p$. Thus, the (k,l)-th entry of the matrix $\mathbb{E}[\tilde y_i\tilde y_j\tilde y\tilde y^T]$ is $\mathbb{E}[\tilde y_i\tilde y_j\tilde y_k\tilde y_l] = \tilde\sigma_{ij}\tilde\sigma_{kl} + \tilde\sigma_{ik}\tilde\sigma_{jl} + \tilde\sigma_{il}\tilde\sigma_{jk}$.

Thus, we can write

$$\mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T] = \tilde{\sigma}_{ij}\tilde{\Sigma} + \tilde{\sigma}_{i\cdot}\tilde{\sigma}_{j\cdot}^T + \tilde{\sigma}_{j\cdot}\tilde{\sigma}_{i\cdot}^T, \tag{11}$$

where $\tilde{\sigma}_{i\cdot}$ is the $p \times 1$ vector $(\tilde{\sigma}_{ik})_{k=1}^p$. From (11), we have

$$\| \mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T] \| \quad \leq \quad |\tilde{\sigma}_{ij}| \, \| \tilde{\Sigma} \| + 2 \, \| \tilde{\sigma}_{i\cdot} \|_2 \| \tilde{\sigma}_{j\cdot} \|_2, \tag{12}$$

where $|| \cdot ||$ is the operator norm. By C0-C1, the first term on the RHS is uniformly bounded. Now, we also have,

$$\tilde{\sigma}_{ii} - \tilde{\sigma}_{i\cdot}^T \tilde{\Sigma}_{(-i)}^{-1} \tilde{\sigma}_{i\cdot} > 0 \tag{13}$$

where $\tilde{\Sigma}_{(-i)}$ is the submatrix of $\tilde{\Sigma}$ removing i-th row and column. From this, it follows that

$$\begin{aligned} \| \tilde{\sigma}_{i\cdot} \|_2 &= \| \tilde{\Sigma}_{(-i)}^{1/2} \tilde{\Sigma}_{(-i)}^{-1/2} \tilde{\sigma}_{i\cdot} \|_2 \\ &\leq \| \tilde{\Sigma}_{(-i)}^{1/2} \| \quad \| \tilde{\Sigma}_{(-i)}^{-1/2} \tilde{\sigma}_{i\cdot} \|_2 \\ &\leq \sqrt{\| \tilde{\Sigma} \|} \sqrt{\tilde{\sigma}_{ii}}, \end{aligned} \tag{14}$$

where the last inequality follows from (13), and the fact that $\tilde{\Sigma}_{(-i)}$ is a principal submatrix of $\tilde{\Sigma}$. Thus the result follows by applying (14) to bound the last term in (12).

proof of Proposition 1: In the following fix σ at the true value $\overline{\sigma}$ and ignore it in the notation. Without loss of generality, assume $w_i \equiv 1$. Consider the loss function: $\overline{L}(\theta) = E_{\overline{\theta}}(||\tilde{y} - \tilde{x}\theta||_2^2)$. Then for $1 \leq i < j \leq p$, $\overline{L}'_{ij}(\theta) = -2E_{\overline{\theta}}(\tilde{x}^T(\tilde{y} - \tilde{x}\theta))$. For $(i,j) \in \mathcal{A}$, since $\text{sign}(\widehat{\theta}_{ij}^{\eta}) = \text{sign}(\overline{\theta}_{ij}) \neq 0$, by Karush-Kuhn-Tucker condition (Lemma

2, Appendix A.2)

$$-2E_{\overline{\theta}}(\tilde{x}_{(i,j)}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))=-\eta\text{sign}(\overline{\theta}_{ij}). \tag{15}$$

For $(i,j)\notin\mathcal{A}$, since $\text{sign}(\widehat{\theta}_{ij}^{\eta})=\text{sign}(\overline{\theta}_{ij})=0$, by Karush-Kuhn-Tucker condition

$$|-2E_{\overline{\theta}}(\tilde{x}_{(i,j)}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))|\leq\eta. \tag{16}$$

Write

$$\tilde{x}_{(i,j)}=\tilde{x}_{\mathcal{A}}\theta_{(i,j)}+\epsilon_{(i,j)}, \tag{17}$$

with $\theta_{(i,j)}=(\overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}))^{-1}\overline{L}''_{\mathcal{A},ij}(\overline{\theta})$. Then it is easy to see, $E_{\overline{\theta}}\left(\epsilon_{(i,j)}^T\tilde{x}_{\mathcal{A}}\right)=0$. Plug (17) into the LHS of (16), we get

$$|2E_{\overline{\theta}}(\tilde{x}_{(i,j)}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))|=|2\theta_{(i,j)}^TE_{\overline{\theta}}(\tilde{x}_{\mathcal{A}}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))+2E_{\overline{\theta}}(\epsilon_{(i,j)}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))|.$$

Since $\text{sign}(\widehat{\theta}^{\eta})=\text{sign}(\overline{\theta})$,

$$\begin{aligned}E_{\overline{\theta}}(\epsilon_{(i,j)}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))&=E_{\overline{\theta}}(\epsilon_{(i,j)}^T(\tilde{y}-\tilde{x}_{\mathcal{A}}\widehat{\theta}_{\mathcal{A}}^{\eta}))\\&=E_{\overline{\theta}}(\epsilon_{(i,j)}^T\tilde{y})=0.\end{aligned}$$

The last equation is because, $\tilde{y}=\tilde{x}\overline{\theta}+\epsilon=\tilde{x}_{\mathcal{A}}\overline{\theta}_{\mathcal{A}}+\epsilon$, with ϵ_i independent of $\tilde{y}_{(-i)}$ for $i=1,\cdots,p$. Thus by definition $E_{\overline{\theta}}(\epsilon^T\tilde{x}_{\mathcal{A}})=0$ and $E_{\overline{\theta}}(\epsilon^T\tilde{x}_{(i,j)})=0$. Therefore $E_{\overline{\theta}}(\epsilon^T\epsilon_{(i,j)})=0$. Also by (15), $2E_{\overline{\theta}}(\tilde{x}_{\mathcal{A}}^T(\tilde{y}-\tilde{x}\widehat{\theta}^{\eta}))=\eta\text{sign}(\overline{\theta}_{\mathcal{A}})$, thus together with (16)

$$\left|\eta\overline{L}''_{ij,\mathcal{A}}(\bar{\theta},\bar{\sigma})\left[\overline{L}''_{\mathcal{A},\mathcal{A}}(\bar{\theta},\bar{\sigma})\right]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}})\right|\leq\eta.$$

This proves proposition 1.

Appendix A.2

In this section, we proof the main results (Theorems 1–3). We first give a few lemmas.

Lemma 2 *(Karush-Kuhn-Tucker condition)* $\widehat{\theta}$ *is a solution of the optimization problem (8), if and only if*

$$\begin{aligned} L'_{n,ij}(\widehat{\theta}, \mathbf{Y}) &= \lambda_n \text{sign}(\widehat{\theta}_{ij}), \quad \text{if} \quad \widehat{\theta}_{ij} \neq 0 \\ |L'_{n,ij}(\widehat{\theta}, \mathbf{Y})| &\leq \lambda_n, \qquad \text{if} \quad \widehat{\theta}_{ij} = 0, \end{aligned}$$

where $\theta_{ij} = \rho^{ij}$ *and* $1 \leq i < j \leq p$*. Moreover, if the solution of (8) is not unique,* $|L'_{n,ij}(\tilde{\theta}, \mathbf{Y})| < \lambda_n$ *for some specific solution* $\tilde{\theta}$ *and* $L'_{n,ij}(\theta, \mathbf{Y})$ *being continuous in* θ *imply that* $\widehat{\theta}_{ij} = 0$ *for all solutions* $\widehat{\theta}$*.*

Lemma 3 *For the loss function (9), if conditions C0-C1 hold, then for any* $\eta > 0$*, there exist constants* $c_{1,\eta}, c_{2,\eta}, c_{3,\eta} > 0$*, such that for any* $u \in R^{q_n}$ *the following hold with probability as least* $1 - O(n^{-\eta})$ *for sufficiently large* n*:*

$$\begin{aligned} &|u^T L'_{n,\mathcal{A}}(\overline{\theta}, \mathbf{Y})| \leq c_{1,\eta} ||u||_2 (\sqrt{\frac{q_n \log n}{n}}) \\ &|u^T L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \mathbf{Y}) u - u^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u| \leq c_{2,\eta} ||u||_2 (q_n \sqrt{\frac{\log n}{n}}) \\ &||L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \mathbf{Y}) u - \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u||_2 \leq c_{3,\eta} ||u||_2 (q_n \sqrt{\frac{\log n}{n}}) \end{aligned}$$

proof of Lemma 3: The proof uses Cauchy-Schwartz and Bernstein's inequalities, and B1.2. Details are omitted.

The following two lemmas are used for proving Theorem 1.

Lemma 4 *Assuming the same conditions of Theorem 1. Then there exists a constant* $C_1(\overline{\theta}) > 0$*, such that for any* $\eta > 0$*, the probability that there exists a local minima of the restricted problem (10) within the disc:*

$$\{\theta : ||\theta - \overline{\theta}||_2 \leq C_1(\overline{\theta})\sqrt{q_n}\lambda_n\}.$$

is at least $1 - O(n^{-\eta})$ *for sufficiently large* n*.*

proof of Lemma 4: Let $\alpha_n = \sqrt{q_n}\lambda_n$, and $Q_n(\theta, \mathbf{Y}, \lambda_n) = L_n(\theta, \mathbf{Y}) + \lambda_n||\theta||_1$. Then for any given constant $C > 0$ and any vector $u \in R^p$ such that $u_{\mathcal{A}^c} = 0$ and $||u||_2 = C$, by the triangle inequality and Cauchy-Schwartz inequality, we have

$$||\overline{\theta}||_1 - ||\overline{\theta} + \alpha_n u||_1 \leq \alpha_n||u||_1 \leq C\alpha_n\sqrt{q_n}.$$

Thus

$$\begin{aligned}
& Q_n(\overline{\theta} + \alpha_n u, \mathbf{Y}, \lambda_n) - Q_n(\overline{\theta}, \mathbf{Y}, \lambda_n) \\
= \; & \{L_n(\overline{\theta} + \alpha_n u, \mathbf{Y}) - L_n(\overline{\theta}, \mathbf{Y})\} - \lambda_n\{||\overline{\theta}||_1 - ||\overline{\theta} + \alpha_n u||_1\} \\
\geq \; & \{L_n(\overline{\theta} + \alpha_n u, \mathbf{Y}) - L_n(\overline{\theta}, \mathbf{Y})\} - C\alpha_n\sqrt{q_n}\lambda_n \\
= \; & \{L_n(\overline{\theta} + \alpha_n u, \mathbf{Y}) - L_n(\overline{\theta}, \mathbf{Y})\} - C\alpha_n^2.
\end{aligned}$$

Thus for any $\eta > 0$, there exists $c_{1,\eta}, c_{2,\eta} > 0$, such that, with probability at least $1 - O(n^{-\eta})$

$$\begin{aligned}
& L_n(\overline{\theta} + \alpha_n u, \mathbf{Y}) - L_n(\overline{\theta}, \mathbf{Y}) = \alpha_n u_{\mathcal{A}}^T L'_{n,\mathcal{A}}(\overline{\theta}, \mathbf{Y}) + \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \mathbf{Y}) u_{\mathcal{A}} \\
\geq \; & \frac{1}{2}\alpha_n^2 u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} - c_{1,\eta}(\alpha_n q_n^{1/2} n^{-1/2}\sqrt{\log n}) - c_{2,\eta}(\alpha_n^2 q_n n^{-1/2}\sqrt{\log n}).
\end{aligned}$$

In the above, the first equation is because the loss function $L(\theta, Y)$ is quadratic in

θ and $u_{\mathcal{A}^c} = 0$. The inequality is due to Lemma 3 and the union bound. Since $\lambda_n n^{1/4} \to \infty$, thus $q_n^{1/2} n^{-1/2} \sqrt{\log n} = o(\sqrt{q_n}\lambda_n) = o(\alpha_n)$. Also by the assumption that $q_n \sim o(\sqrt{n/\log n})$, with n sufficiently large

$$Q_n(\overline{\theta} + \alpha_n u, \mathbf{Y}, \lambda_n) - Q_n(\overline{\theta}, \mathbf{Y}, \lambda_n) \geq \frac{1}{4}\alpha_n^2 u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} - C\alpha_n^2$$

with probability at least $1 - O(n^{-\eta})$. By B1, $u_{\mathcal{A}}^T \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta}) u_{\mathcal{A}} \geq \Lambda_{\min}^L(\bar{\theta})||u_{\mathcal{A}}||_2^2 = \Lambda_{\min}^L(\bar{\theta})C^2$. Thus, if we choose $C = 4/\Lambda_{\min}^L(\bar{\theta}) + \epsilon$, then for any $\eta > 0$, for sufficiently large n, the following holds

$$\inf_{u: u_{\mathcal{A}^c}=0, ||u||_2 = C} Q_n(\overline{\theta} + \alpha_n u, \mathbf{Y}, \lambda_n) > Q_n(\overline{\theta}, \mathbf{Y}, \lambda_n),$$

with probability at least $1 - O(n^{-\eta})$. This means that a local minima exists within the disc $\{\theta : ||\theta - \overline{\theta}||_2 \leq C\alpha_n = C\sqrt{q_n}\lambda_n\}$ with probability at least $1 - O(n^{-\eta})$.

Lemma 5 *Assuming the same conditions of Theorem 1. Then there exists a constant $C_2(\overline{\theta}) > 0$, such that for any $\eta > 0$, for sufficiently large n, the following holds with probability at least $1 - O(n^{-\eta})$: for any θ belongs to the set $S = \{\theta : ||\theta - \overline{\theta}||_2 \geq C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \theta_{\mathcal{A}^c} = 0\}$, it has $||L'_{n,\mathcal{A}}(\theta, \mathbf{Y})||_2 > \sqrt{q_n}\lambda_n$.*

proof of Lemma 5: Let $\alpha_n = \sqrt{q_n}\lambda_n$. Any θ belongs to S can be written as: $\theta = \overline{\theta} + \alpha_n u$, with $u_{\mathcal{A}^c} = 0$ and $||u||_2 \geq C_2(\bar{\theta})$. Note that

$$\begin{aligned} L'_{n,\mathcal{A}}(\theta, \mathbf{Y}) &= L'_{n,\mathcal{A}}(\overline{\theta}, \mathbf{Y}) + \alpha_n L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \mathbf{Y}) u \\ &= L'_{n,\mathcal{A}}(\overline{\theta}, \mathbf{Y}) + \alpha_n (L''_{n,\mathcal{A}\mathcal{A}}(\overline{\theta}, \mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta})) u + \alpha_n \overline{L}''_{\mathcal{A}\mathcal{A}}(\overline{\theta})) u. \end{aligned}$$

By the triangle inequality and Lemma 3, for any $\eta > 0$, there exists constants

$c_{1,\eta}, c_{3,\eta} > 0$, such that

$$||L'_{n,\mathcal{A}}(\theta, \mathbf{Y})||_2 \geq \alpha_n ||\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta}))u||_2 - c_{1,\eta}(q_n^{1/2} n^{-1/2}\sqrt{\log n}) - c_{3,\eta}||u||_2(\alpha_n q_n n^{-1/2}\sqrt{\log n})$$

with probability at least $1 - O(n^{-\eta})$. Thus, similar as in Lemma 4, for n sufficiently large, $||L'_{n,\mathcal{A}}(\theta, \mathbf{Y})||_2 \geq \frac{1}{2}\alpha_n ||\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta}))u||_2$ with probability at least $1 - O(n^{-\eta})$. By B1, $||\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta}))u||_2 \geq \Lambda^L_{\min}(\bar{\theta})||u||_2$. Therefore $C_2(\bar{\theta})$ can be taken as $2/\Lambda^L_{\min}(\bar{\theta}) + \epsilon$.

The following lemma is used in proving Theorem 2.

Lemma 6 *Assuming conditions C0-C1. Let* $D_{\mathcal{A}\mathcal{A}}(\bar{\theta}, Y) = L''_{1,\mathcal{A}\mathcal{A}}(\bar{\theta}, Y) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})$. *Then there exists a constant* $K_2(\bar{\theta}) < \infty$, *such that for any* $(k, l) \in \mathcal{A}$, $\lambda_{\max}(\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, Y))) \leq K_2(\bar{\theta})$.

proof of Lemma 6: $\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, Y)) = E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T) - \overline{L}''_{\mathcal{A},kl}(\bar{\theta})\overline{L}''_{\mathcal{A},kl}(\bar{\theta})^T$. Thus it suffices to show that, there exists a constant $K_2(\bar{\theta}) > 0$, such that for all (k, l)

$$\lambda_{\max}(E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T)) \leq K_2(\bar{\theta}).$$

Use the same notations as in the proof of B1. Note that $L''_{1,\mathcal{A},kl}(\bar{\theta}, Y) = \tilde{x}^T\tilde{w}^2\tilde{x}_{(k,l)} = \tilde{w}_k\tilde{y}_l x_k + \tilde{w}_l\tilde{y}_k x_l$. Thus

$$E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T) = \tilde{w}_k^2\mathbb{E}[\tilde{y}_l^2 x_k x_k^T] + \tilde{w}_l^2\mathbb{E}[\tilde{y}_k^2 x_l x_l^T] + \tilde{w}_k\tilde{w}_l\mathbb{E}[\tilde{y}_k\tilde{y}_l(x_k x_l^T + x_l x_k^T)],$$

and for $a \in \mathcal{R}^{p(p-1)/2}$

$$\begin{aligned} & a^T E_{\bar{\theta}}(L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)L''_{1,\mathcal{A},kl}(\bar{\theta}, Y)^T)a \\ = \;& \tilde{w}_k^2 a_k^T\mathbb{E}[\tilde{y}_l^2\tilde{y}\tilde{y}^T]a_k + \tilde{w}_l^2 a_l^T\mathbb{E}[\tilde{y}_k^2\tilde{y}\tilde{y}^T]a_l + 2\tilde{w}_k\tilde{w}_l a_k^T\mathbb{E}[\tilde{y}_k\tilde{y}_l\tilde{y}\tilde{y}^T]a_l. \end{aligned}$$

Since $\sum_{k=1}^p ||a_k||_2^2 = 2||a||_2^2$, and by B2: $\lambda_{\max}(\mathbb{E}[\tilde{y}_i\tilde{y}_j\tilde{y}\tilde{y}^T]) \leq K_1(\bar{\theta})$ for any $1 \leq i \leq$

$j \leq p$, the conclusion follows.

proof of Theorem 1: The existence of a solution of (10) follows from Lemma 4. By the Karush-Kuhn-Tucker condition (Lemma 2), for any solution $\widehat{\theta}$ of (10), it has $||L'_{n,\mathcal{A}}(\widehat{\theta}, \mathbf{Y})||_\infty \leq \lambda_n$. Thus $||L'_{n,\mathcal{A}}(\widehat{\theta}, \mathbf{Y})||_2 \leq \sqrt{q_n}||L'_{n,\mathcal{A}}(\widehat{\theta}, \mathbf{Y})||_\infty \leq \sqrt{q_n}\lambda_n$. Thus by Lemma 5, for any $\eta > 0$, for n sufficiently large with probability at least $1-O(n^{-\eta})$, all solutions of (10) are inside the disc $\{\theta : ||\theta - \overline{\theta}||_2 \leq C_2(\overline{\theta})\sqrt{q_n}\lambda_n\}$. Since $\frac{s_n}{\sqrt{q_n}\lambda_n} \to \infty$, for sufficiently large n and $(i,j) \in \mathcal{A}$: $\overline{\theta}_{ij} \geq s_n > 2C_2(\overline{\theta})\sqrt{q_n}\lambda_n$. Thus

$$\begin{aligned} 1 - O(n^{-\eta}) &\leq P_{\overline{\theta}}\left(||\widehat{\theta}^{\mathcal{A},\lambda_n} - \overline{\theta}_{\mathcal{A}}||_2 \leq C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \bar{\theta}_{ij} > 2C_2(\overline{\theta})\sqrt{q_n}\lambda_n, \text{ for all}(i,j) \in \mathcal{A}\right) \\ &\leq P_{\bar{\theta}}\left(\text{sign}(\widehat{\theta}_{ij}^{\mathcal{A},\lambda_n}) = \text{sign}(\overline{\theta}_{ij}), \text{ for all}(i,j) \in \mathcal{A}\right). \end{aligned}$$

proof of Theorem 2: For any given $\eta > 0$, let $\eta' = \eta + \kappa$. Let $\mathcal{E}_n = \{\text{sign}(\widehat{\theta}^{\mathcal{A},\lambda_n}) = \text{sign}(\bar{\theta})\}$. Then by Theorem 1, $P_{\bar{\theta}}(\mathcal{E}_n) \geq 1 - O(n^{-\eta'})$ for sufficiently large n. On $\mathcal{E}_n$, by the Karush-Kuhn-Tucker condition and the expansion of $L'_{n,\mathcal{A}}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \mathbf{Y})$ at $\bar{\theta}$

$$-\lambda_n \text{sign}(\bar{\theta}_{\mathcal{A}}) = L'_{n,\mathcal{A}}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \mathbf{Y}) = L'_{n,\mathcal{A}}(\bar{\theta}, \mathbf{Y}) + L''_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \mathbf{Y})\nu_n,$$

where $\nu_n := \widehat{\theta}_{\mathcal{A}}^{\mathcal{A},\lambda_n} - \bar{\theta}_{\mathcal{A}}$. By the above expression

$$\nu_n = -\lambda_n[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}}) - [\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}[L'_{n,\mathcal{A}}(\bar{\theta}, \mathbf{Y}) + D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \mathbf{Y})\nu_n], \qquad (18)$$

where $D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \mathbf{Y}) = L''_{n,\mathcal{A}\mathcal{A}}(\bar{\theta}, \mathbf{Y}) - \overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})$. Fix $(i,j) \in \mathcal{A}^c$, and consider the expansion of $L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \mathbf{Y})$ around $\bar{\theta}$:

$$L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n}, \mathbf{Y}) = L'_{n,ij}(\bar{\theta}, \mathbf{Y}) + L''_{n,ij,\mathcal{A}}(\bar{\theta}, \mathbf{Y})\nu_n. \qquad (19)$$

Then plug in (18) into (19), we get

$$L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\mathbf{Y}) = -\lambda_n \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}}) - \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}L'_{n,\mathcal{A}}(\bar{\theta},\mathbf{Y})$$
$$+ \quad L'_{n,ij}(\bar{\theta},\mathbf{Y}) + \left[D_{n,ij,\mathcal{A}}(\bar{\theta},\mathbf{Y}) - \overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\mathbf{Y})\right]\nu_n. \tag{20}$$

By condition C2, for any $(i,j) \in \mathcal{A}^c$: $|\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}\text{sign}(\bar{\theta}_{\mathcal{A}})| \leq \delta < 1$. Thus it suffices to prove that the remaining terms in (20) are all $o(\lambda_n)$ with probability at least $1 - O(n^{-\eta'})$ (uniformly for all $(i,j) \in \mathcal{A}^c$). Then since $|\mathcal{A}^c| \leq p \sim O(n^\kappa)$, by the union bound, the event $\max_{(i,j)\in\mathcal{A}^c} |L'_{n,ij}(\widehat{\theta}^{\mathcal{A},\lambda_n},\mathbf{Y})| < \lambda_n$ holds with probability at least $1 - O(n^{\kappa-\eta'}) = 1 - O(n^{-\eta})$, when n is sufficiently large.

By B1.4, for any $(i,j) \in \mathcal{A}^c$: $||\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}||_2 \leq M(\bar{\theta})$. Therefore by Lemma 3, for any $\eta > 0$, there exists a constant $C_{1,\eta} > 0$ (not dependent on (i,j)), such that $|\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}L'_{n,\mathcal{A}}(\bar{\theta},\mathbf{Y})| \leq C_{1,\eta}(\sqrt{\frac{q_n}{n}\log n})$ with probability at least $1-O(n^{-\eta})$. Since $q_n \sim o(\sqrt{n/\log n})$ and $\lambda_n n^{1/4} \to \infty$, $O(\sqrt{q_n \log n/n}) = o(\lambda_n)$.

By B1.2, $||\text{Var}_{\bar{\theta}}(L'_{n,ij}(\bar{\theta},\mathbf{Y}))||_2 \leq M_1(\bar{\theta})$. Then similarly as above, for any $\eta > 0$, there exists a constant $C_{2,\eta} > 0$, such that $|L'_{n,ij}(\bar{\theta},\mathbf{Y})| \leq C_{2,\eta}(\sqrt{\frac{\log n}{n}}) = o(\lambda_n)$, with probability at least $1 - O(n^{-\eta})$.

Note that by Theorem 1, for any $\eta > 0$, $||\nu_n||_2 \leq C(\bar{\theta})\sqrt{q_n}\lambda_n$ with probability at least $1 - O(n^{-\eta})$ for large enough n. Thus, similarly as in Lemma 3, for any $\eta > 0$, there exists a constant $C_{3,\eta}$, such $|D_{n,ij,\mathcal{A}}(\bar{\theta},\mathbf{Y})\nu_n| \leq C_{3,\eta}(\sqrt{\frac{q_n}{n}\log n}\sqrt{q_n}\lambda_n) = o(\lambda_n)$, with probability at least $1 - O(n^{-\eta})$.

Finally, let $b^T = |\overline{L}''_{ij,\mathcal{A}}(\bar{\theta})[\overline{L}''_{\mathcal{A}\mathcal{A}}(\bar{\theta})]^{-1}$. By Cauchy-Schwartz inequality

$$|b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\mathbf{Y})\nu_n|^2 \leq ||b^T D_{n,\mathcal{A}\mathcal{A}}(\bar{\theta},\mathbf{Y})||_2^2||\nu_n||_2^2 \leq q_n^2\lambda_n^2 \max_{(k,l)\in\mathcal{A}} |b^T D_{n,\mathcal{A},kl}(\bar{\theta},\mathbf{Y})|^2.$$

In order to show the RHS is $o(\lambda_n^2)$ with probability at least $1 - O(n^{-\eta})$, it suffices to show $\max_{(k,l)\in\mathcal{A}} |b^T D_{n,\mathcal{A},kl}(\bar{\theta},\mathbf{Y})|^2 = O(\log n/n)$ with probability at least $1 - O(n^{-\eta})$,

which in turn is implied by

$$E_{\bar{\theta}}(|b^T D_{\mathcal{A},kl}(\bar{\theta}, Y)|^2) \leq ||b||_2^2 \lambda_{\max}(\mathrm{Var}_{\bar{\theta}}(D_{\mathcal{A},kl}(\bar{\theta}, Y)))$$

being bounded. This follows immediately from B1.4 and Lemma 6.

proof of Theorem 3: By Theorems 1 and 2 and the Karush-Kuhn-Tucker condition, for any $\eta > 0$, with probability at least $1 - O(n^{-\eta})$, a solution of the restricted problem is also a solution of the original problem. On the other hand, by Theorem 2 and the Karush-Kuhn-Tucker condition, with high probability, any solution of the original problem is a solution of the restricted problem. Therefore, by Theorem 1, the conclusion follows.

Appendix B

In this section, we provide details for the implementation of `space` which take into account of the special structure of $\mathcal{X}$. Denote the target loss function as

$$f(\theta) = \frac{1}{2} \|\mathcal{Y} - \mathcal{X}\theta\|^2 + \lambda_1 ||\theta||_{l_1} + \lambda_2 ||\theta||_{l_2}^2, \tag{21}$$

where $\theta = \{\rho^{ij}\}_{1 \leq i < j \leq p}$. Our goal is to find $\widehat{\theta} = \mathrm{argmin}_\theta f(\theta)$ for given λ_1 and λ_2. We will employ `active-shooting` algorithm (section 2.3) to solve this optimization problem.

Without loss of generality, we assume $\mathrm{mean}(Y_i) = 1/n \sum_{k=1}^n y_i^k = 0$ for $i = 1, .., p$. Denote $\xi_i = Y_i^T Y_i$. We have

$$\mathcal{X}_{(i,j)}^T \mathcal{X}_{(i,j)} = \xi_j \frac{\sigma^{jj}}{\sigma^{ii}} + \xi_i \frac{\sigma^{ii}}{\sigma^{jj}};$$

$$\mathcal{Y}^T\mathcal{X}_{(i,j)} = \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_i^TY_j + \sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}Y_j^TY_i.$$

Further denote $\rho_{(i,j)} = \rho^{ij}$. We give details of the initiation step and the updating steps in the `active-shooting` algorithm.

I. Initiation

Let

$$\begin{aligned}\rho^0_{(i,j)} &= \frac{1}{1+\lambda_2}\frac{\left(|\mathcal{Y}^T\mathcal{X}_{(i,j)}|-\lambda_1\right)_+\cdot\text{sign}(\mathcal{Y}^T\mathcal{X}_{(i,j)})}{\mathcal{X}_{(i,j)}^T\mathcal{X}_{(i,j)}} \\ &= \frac{1}{1+\lambda_2}\frac{\left|\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_i^TY_j+\sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}Y_j^TY_i\ -\lambda_1\right|_+\cdot\text{sign}(Y_i^TY_j)}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}.\end{aligned} \tag{22}$$

For $j = 1, ..., p$, compute

$$\widehat{Y_j}^{(0)} = \left(\sqrt{\frac{\sigma^{11}}{\sigma^{jj}}}Y_1, ..., \sqrt{\frac{\sigma^{pp}}{\sigma^{jj}}}Y_p\right)\cdot\begin{pmatrix}\rho^{(0)}_{(1,j)}\\ \vdots \\ \rho^{(0)}_{(p,j)}\end{pmatrix}, \tag{23}$$

and

$$E^{(0)} = \mathcal{Y} - \widehat{\mathcal{Y}}^{(0)} = \left((E_1^{(0)})^T, ..., (E_p^{(0)})^T\right), \tag{24}$$

where $E_j^{(0)} = Y_j - \widehat{Y_j}^{(0)}$ for $1 \le j \le p$.

II. Update $\rho^{(0)}_{(i,j)} \longrightarrow \rho^{(1)}_{(i,j)}$

Let

$$A_{(i,j)} = (E_j^{(0)})^T\cdot\sqrt{\frac{\sigma^{ii}}{\sigma^{jj}}}Y_i; \tag{25}$$

$$A_{(j,i)} = (E_i^{(0)})^T\cdot\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}Y_j. \tag{26}$$

We have

$$\begin{aligned}(E^{(0)})^T\mathcal{X}_{(i,j)} &= (E_i^{(0)})^T\cdot\sqrt{\tfrac{\sigma^{jj}}{\sigma^{ii}}}Y_j + (E_j^{(0)})^T\cdot\sqrt{\tfrac{\sigma^{ii}}{\sigma^{jj}}}Y_i \\ &= A_{(j,i)} + A_{(i,j)}.\end{aligned} \tag{27}$$

It follows

$$\begin{aligned}\rho^{(1)}_{(i,j)} &= \tfrac{1}{1+\lambda_2}\text{sign}\left(\frac{(E^{(0)})^T\mathcal{X}_{(i,j)}}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}}+\rho^{(0)}_{(i,j)}\right)\left(\left|\frac{(E^{(0)})^T\mathcal{X}_{(i,j)}}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}}+\rho^{(0)}_{(i,j)}\right|-\frac{\lambda_1}{\mathcal{X}^T_{(i,j)}\mathcal{X}_{(i,j)}}\right)_+ \\ &= \tfrac{1}{1+\lambda_2}\text{sign}\left(\frac{A_{(j,i)}+A_{(i,j)}}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}+\rho^{(0)}_{(i,j)}\right)\left(\left|\frac{A_{(j,i)}+A_{(i,j)}}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}+\rho^{(0)}_{(i,j)}\right|-\frac{\lambda_1}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}\right)_+ .\end{aligned} \tag{28}$$

III. Update $\rho^{(t)} \longrightarrow \rho^{(t+1)}$

From the previous iteration, we have

- $E^{(t-1)}$: residue in the previous iteration ($np \times 1$ vector).
- (i_0, j_0): index of coefficient that is updated in the previous iteration.
- $\rho^{(t)}_{(i,j)} = \begin{cases} \rho^{(t-1)}_{(i,j)} & \text{if } (i,j) \neq (i_0,j_0),\ \text{nor } (j_0,i_0) \\ \rho^{(t-1)}_{(i,j)} - \Delta & \text{if } (i,j) = (i_0,j_0),\ \text{or } (j_0,i_0) \end{cases}$

Then,

$$\begin{aligned} E^{(t)}_k &= E^{(t-1)}_k \text{ for } k \neq i_0, j_0; \\ E^{(t)}_{j_0} &= E^{(t-1)}_{j_0} + \widehat{Y}^{(t-1)}_{j_0} - \widehat{Y}^{(t)}_{j_0} \\ &= E^{(t-1)}_{j_0} + \textstyle\sum_{i=1}^P \sqrt{\frac{\sigma^{ii}}{\sigma^{j_0j_0}}} Y_i(\rho^{(t-1)}_{(i,j_0)} - \rho^{(t)}_{(i,j_0)}) \\ &= E^{(t-1)}_{j_0} + \sqrt{\frac{\sigma^{i_0i_0}}{\sigma^{j_0j_0}}} Y_{i_0}\cdot\Delta; \\ E^{(t)}_{i_0} &= E^{(t-1)}_{i_0} + \sqrt{\frac{\sigma^{j_0j_0}}{\sigma^{i_0i_0}}} Y_{j_0}\cdot\Delta. \end{aligned} \tag{29}$$

Suppose the index of the coefficient we want to update in this iteration is (i_1, j_1), then let

$$A_{(i_1,j_1)} = (E^{(t)}_{j_1})^T \cdot \sqrt{\frac{\sigma^{i_1i_1}}{\sigma^{j_1j_1}}} Y_{i_1},$$

$$A_{(j_1,i_1)} = (E^{(t)}_{i_1})^T \cdot \sqrt{\frac{\sigma^{j_1j_1}}{\sigma^{i_1i_1}}} Y_{j_1}.$$

We have

$$\begin{aligned}\rho_{(i,j)}^{(t+1)} = & \frac{1}{1+\lambda_2}\text{sign}\left(\frac{A_{(j_1,i_1)}+A_{(i_1,j_1)}}{\xi_j\frac{\sigma^{j_1j_1}}{\sigma^{i_1i_1}}+\xi_{i_1}\frac{\sigma^{i_1i_1}}{\sigma^{j_1j_1}}}+\rho_{(i_1,j_1)}^{(t)}\right) \\ & \times\left(\left|\frac{A_{(j_1,i_1)}+A_{(i_1,j_1)}}{\xi_j\frac{\sigma^{j_1j_1}}{\sigma^{i_1i_1}}+\xi_{i_1}\frac{\sigma^{i_1i_1}}{\sigma^{j_1j_1}}}+\rho_{(i_1,j_1)}^{(t)}\right|-\frac{\lambda_1}{\xi_j\frac{\sigma^{jj}}{\sigma^{ii}}+\xi_i\frac{\sigma^{ii}}{\sigma^{jj}}}\right)_+ .\end{aligned} \tag{30}$$

With above **I-III** steps, it is easy to implement the `active-shooting` algorithm to solve the optimization problem in `space`. It is also obvious that the computational cost is $O(np^2)$. It is implemented in `c` and available in the `R` package `space`.

Appendix C

Table 1: The numbers of iterations required by the `shooting` algorithm and the proposed `active-shooting` algorithm to achieve convergence ($n = 100$, $\lambda = 2$). "coef. #" is the number of non-zero coefficients

p	coef. #	`Shooting`	`active-shooting`
200	14	29600	4216
500	25	154000	10570
1000	28	291000	17029

Table 2: Power (sensitivity) of `MB`, `glasso` and `space.dew` in identifying correct edges when FDR is controlled at 0.05.

Network	p	n	`MB`	`glasso`	`space.dew`
Hub network	500	250	0.784	0.655	0.844
Hub network	1000	200	0.656	0.559	0.707
		300	0.790	0.690	0.856
		500	0.894	0.826	0.963
Power-law network	500	250	0.667	0.580	0.704

Table 3: Power Law Network (.tr means using true $\{\sigma^{ii}\}$)

Method	MB	space.sw	space	space.dew	space.tr.sw	space.tr	space.tr.dew
Correct No.	378	398	405	406	411	406	415
$FDR = 0.05$	0.662	0.691	0.730	0.736	0.665	0.675	0.717
$FDR = 0.1$	0.720	0.765	0.787	0.797	0.791	0.778	0.808

Table 4: Empirical Network (.tr means using true $\{\sigma^{ii}\}$)

Method	MB	space.sw	space	space.dew	space.tr.sw	space.tr	space.tr.dew
Correct No.	456	467	476	481	489	481	510
$FDR = 0.05$	0.593	0.587	0.616	0.628	0.643	0.635	0.681
$FDR = 0.1$	0.633	0.639	0.661	0.670	0.689	0.676	0.730

Table 5: Edge detection under the selected tuning parameter λ by BIC

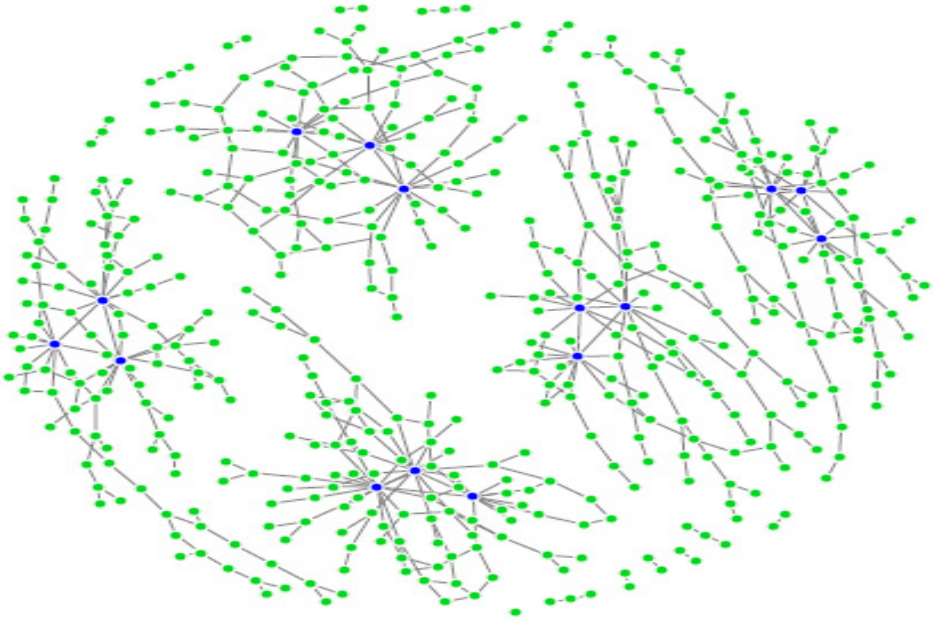
Sample size	Method	Total edge	Total true	False positive	False negative	Total False
$n = 200$	`space.dew`	1390.16	957.88	432.28	205.12	637.40
	`MB`	1240.36	872.92	367.44	290.08	657.52
	`glasso`	1542.32	954.56	587.76	208.44	796.20
$n = 300$	`space.dew`	1638.48	1079.88	558.60	83.12	641.72
	`MB`	1455.96	1008.68	447.28	154.32	601.60
	`glasso`	1743.16	1069.88	673.28	93.12	766.40
$n = 500$	`space.dew`	1701.92	1143.16	558.76	19.84	578.60
	`MB`	1555.48	1098.48	457.00	64.52	521.52
	`glasso`	1942.20	1137.88	804.32	25.12	829.44

Table 6: Hub detection under the selected tuning parameter λ by BIC. (The optimal average rank would be 15.5)

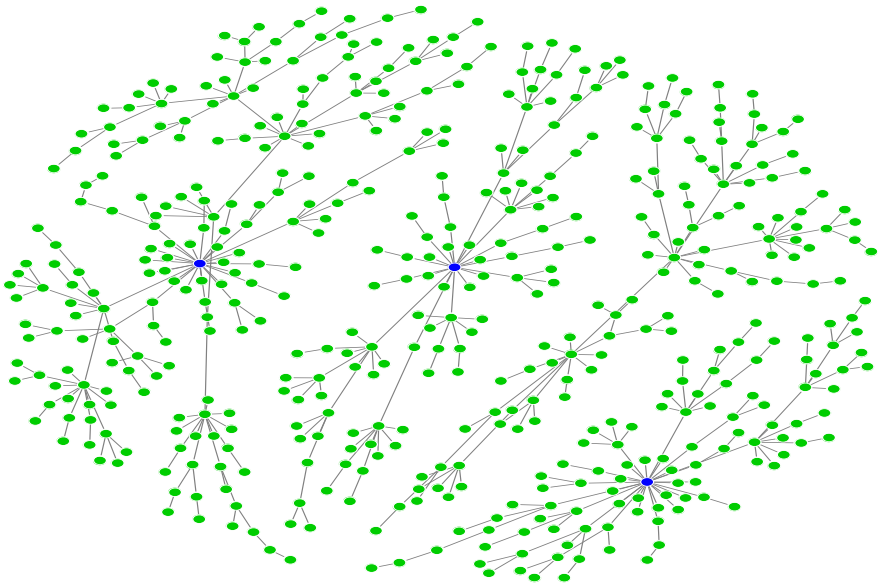
Sample size	Method	Mean of average rank	Sd of average rank
$n = 200$	`space.dew`	29.2	8.6
	`MB`	57.5	15.8
	`glasso`	35.42	11.1
$n = 300$	`space.dew`	19.5	2.13
	`MB`	30.4	12.6
	`glasso`	21	5.02
$n = 500$	`space.dew`	16.7	1.01
	`MB`	16.9	1.67
	`glasso`	16.5	0.6

Table 7: Annotation of hub genes

Index	Gene Symbol	Summary Function (GO)
1	CENPA	Encodes a centromere protein (nucleosome assembly)
2	*NA*	*NA*
3	KNSL6	Anaphase chromosome segregation (cell proliferation)
4	STK12	Regulation of chromosomal segregation (cell cycle)
5	*NA*	*NA*
6	URLC9	*NA* (up-regulated in lung cancer)
7	HNF3A	Transcriptional factor activity (epithelial cell differentiation)
8	TPX2	Spindle formation (cell proliferation)
9	RAD54L	Homologous recombination related DNA repair (meiosis)
10	ID-GAP	Stimulate GTP hydrolysis (cell cycle)
11	BUB1	Spindle checkpoint (cell cycle)

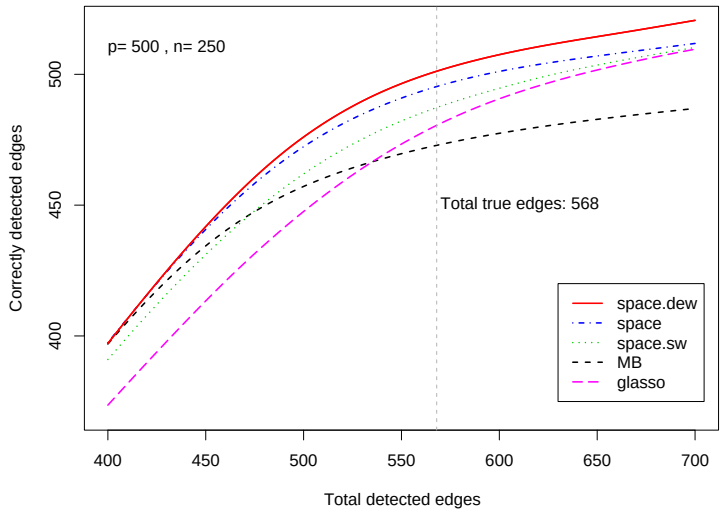


(a) Hub network: 500 nodes and 568 edges. 15 nodes (in blue) have degrees of around 15.

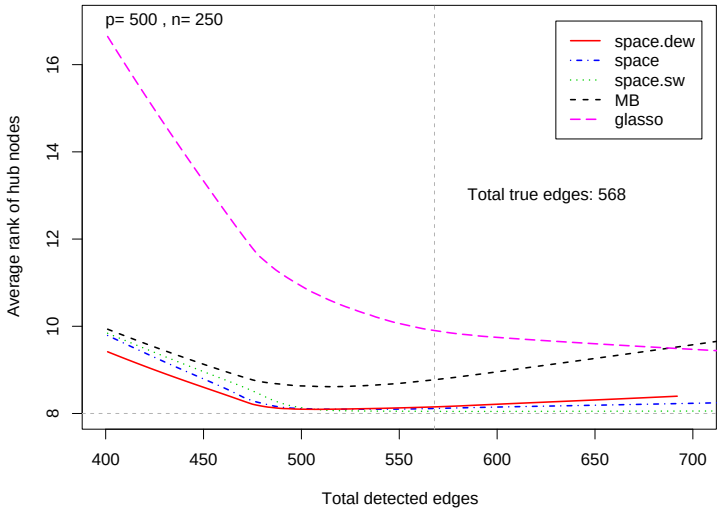


(b) Power-law network: 500 nodes and 495 edges. 3 nodes (in blue) have degrees at least 20.

Figure 1: Topology of simulated networks.



(a) *x-axis*: the total number of edges detected (i.e., the total number of pairs (i, j) with $\widehat{\rho}^{ij} \neq 0$); *y-axis*: the total number of correctly identified edges. The vertical grey line corresponds to the number of true edges.



(b) *x-axis*: the total number of edges detected; *y-axis*: the average rank of the estimated degrees of the 15 true hub nodes.

Figure 2: Simulation results for Hub network.

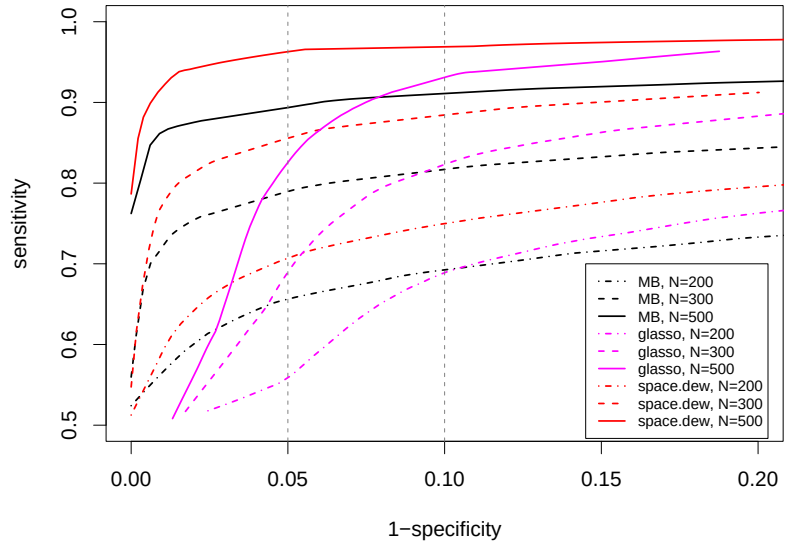


Figure 3: Hub network: ROC curves for different samples sizes ($p = 1000$).

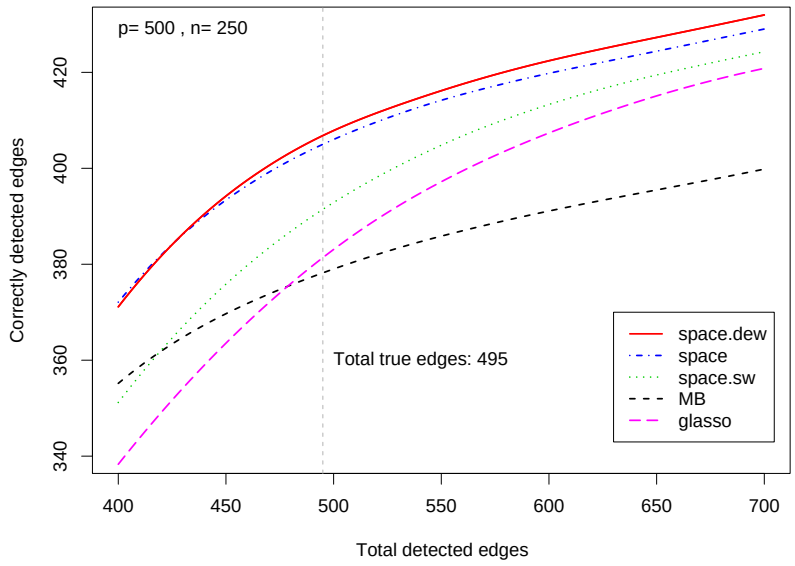
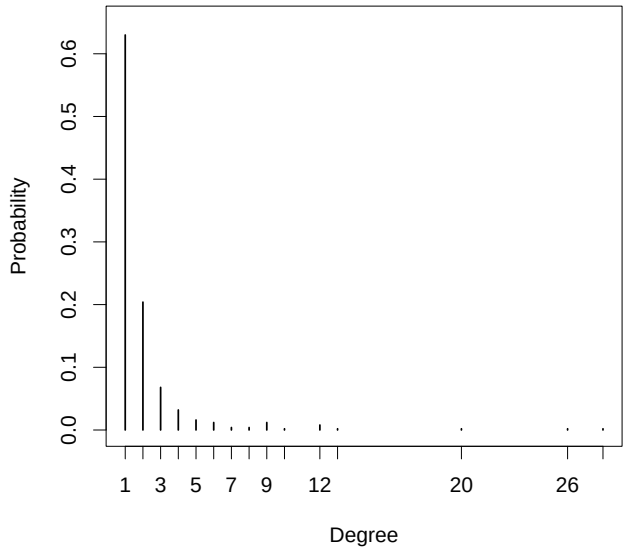
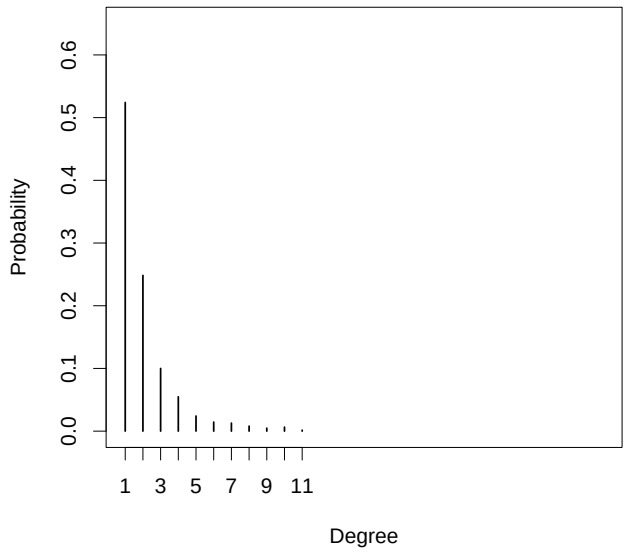


Figure 4: Simulation results for Power-law network. *x-axis*: the total number of edges detected; *y-axis*: the total number of correctly identified edges. The vertical grey line corresponds to the number of true edges.



(a) Power-law network. Power law parameter $\alpha = 2.3$.



(b) Inferred real network. Power law parameter $\alpha = 2.56$.

Figure 5: The degree distributions of the Power-law network used in the simulation study and the inferred network based on the breast cancer expression data set.

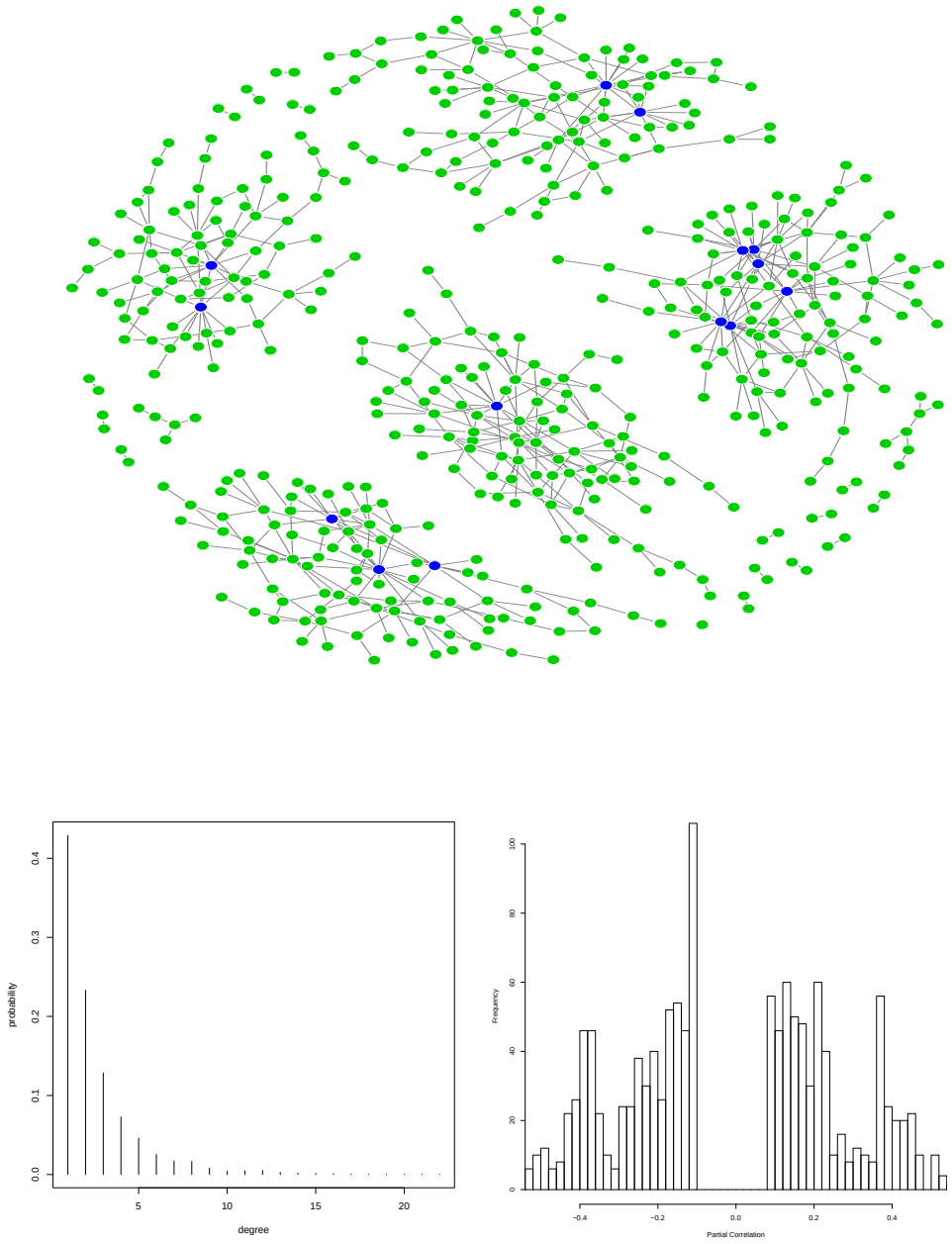


Figure 6: Empirical Network

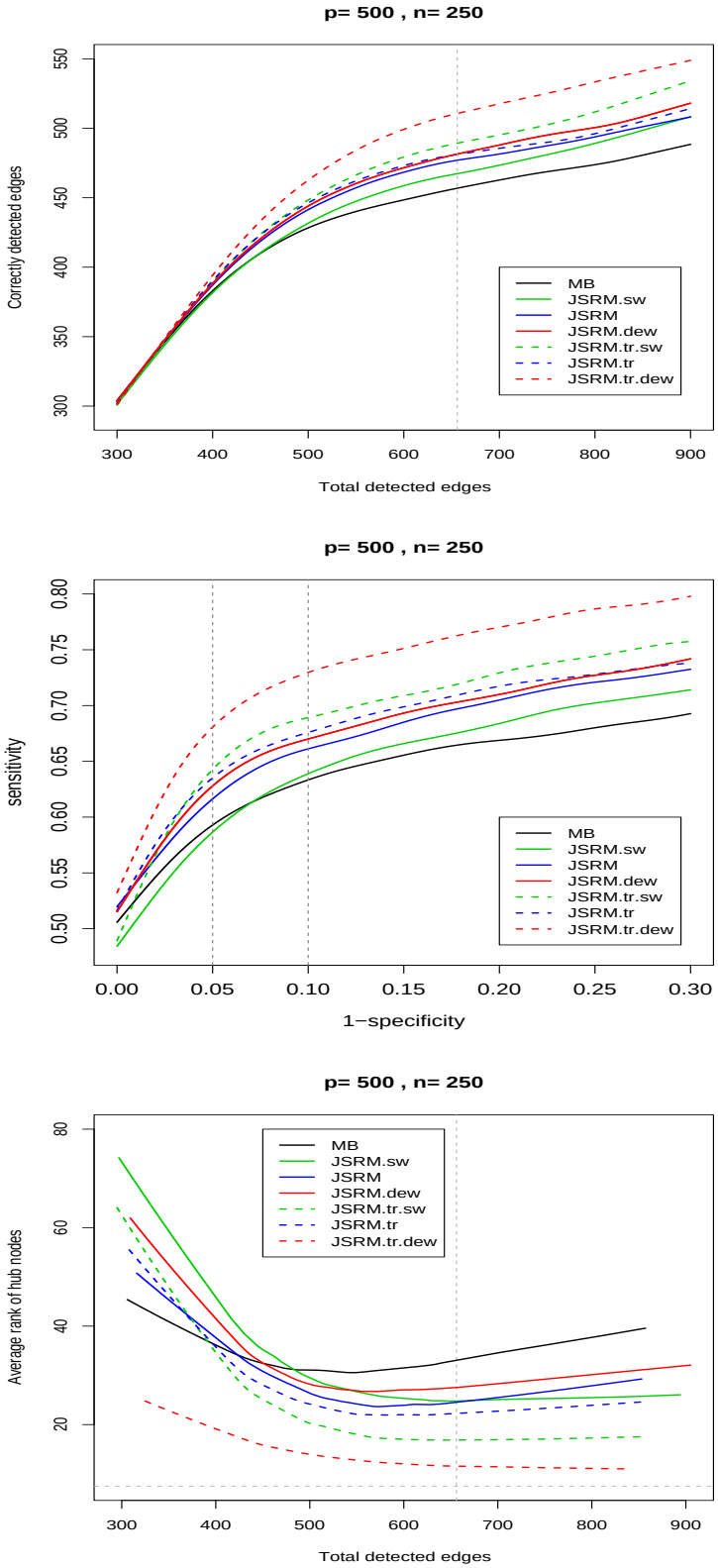
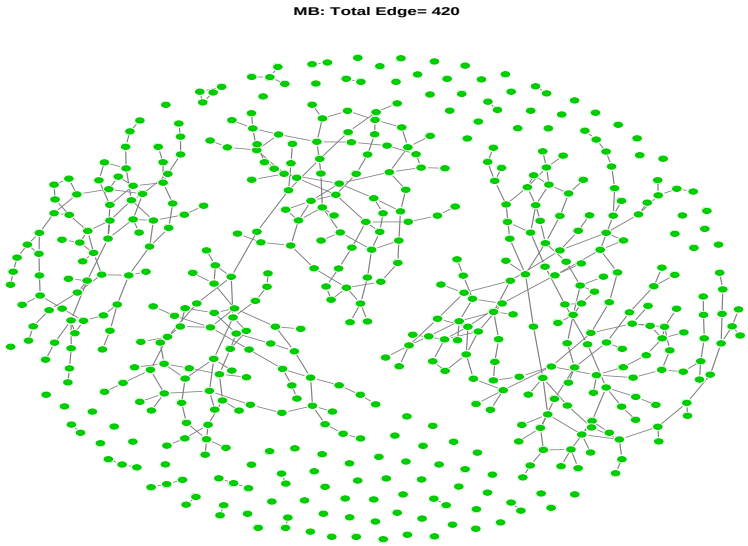
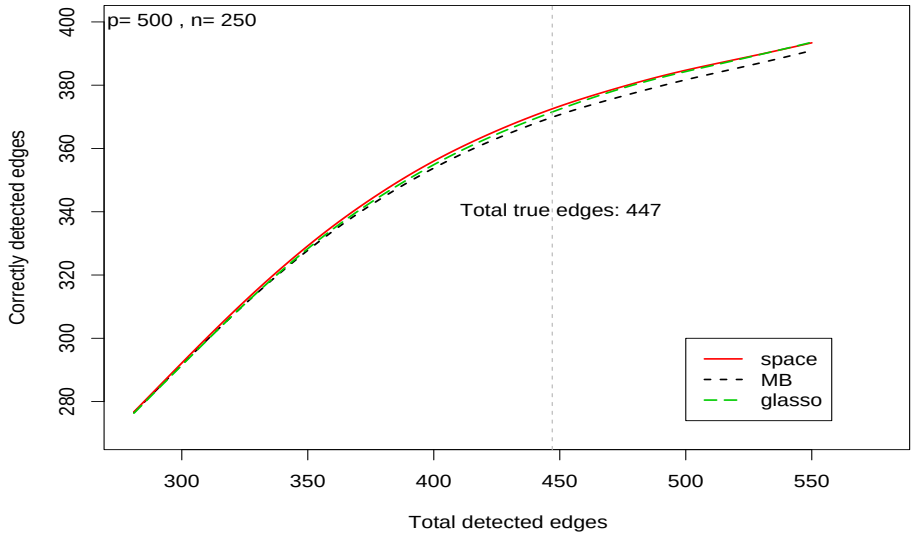


Figure 7: Empirical network: p=500, n=250 (.tr means using true $\{\sigma^{ii}\}$)

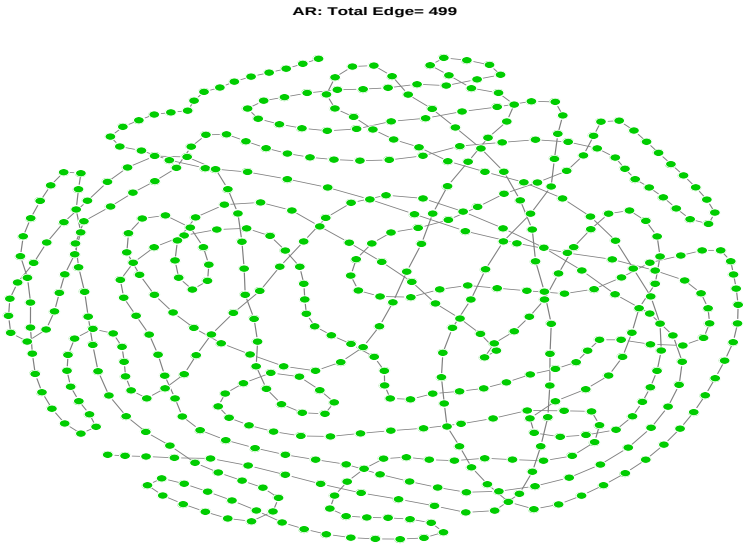


(a) Uniform network: 500 nodes and 447 edges.

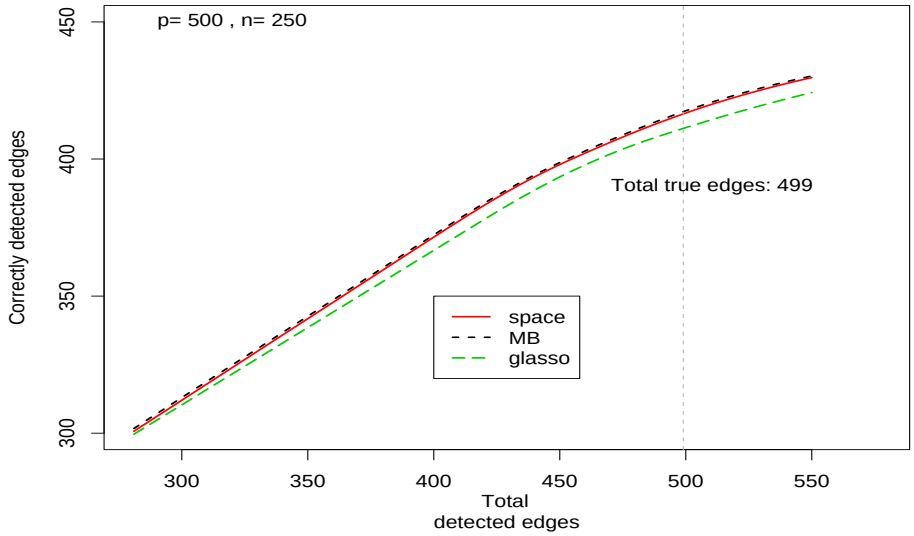


(b) Simulation results for Uniform network. *x-axis*: the total number of edges detected; *y-axis*: the total number of correctly identified edges. The vertical grey line corresponds to the number of true edges.

Figure 8: Simulation results for Uniform networks.

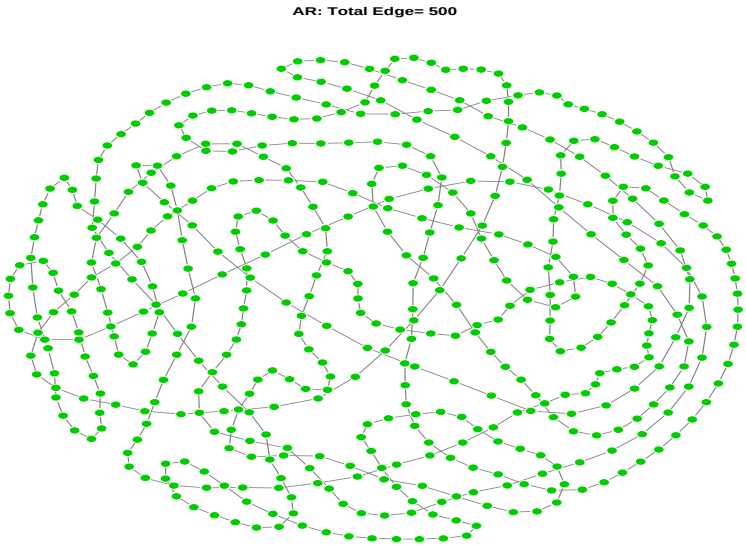


(a) AR network: 500 nodes and 499 edges.

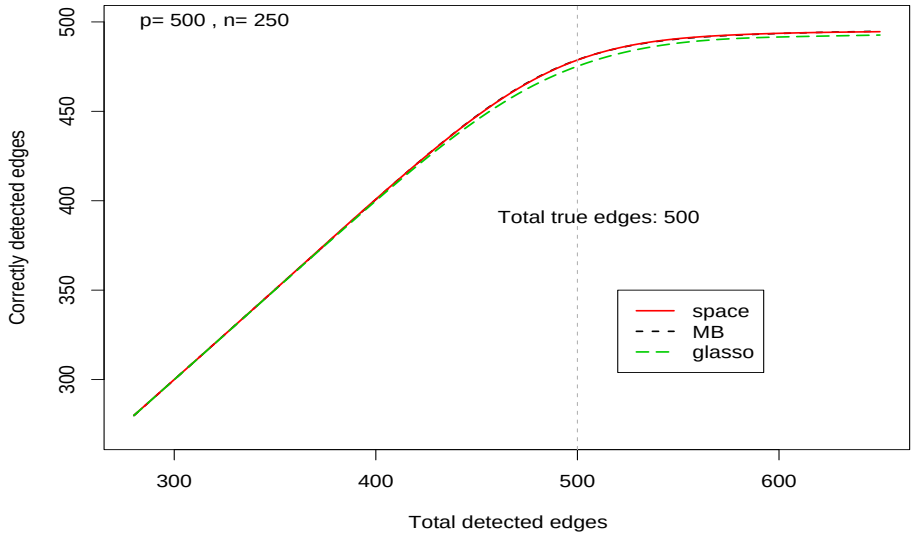


(b) Simulation results for AR network. *x-axis*: the total number of edges detected; *y-axis*: the total number of correctly identified edges. The vertical grey line corresponds to the number of true edges.

Figure 9: Simulation results for AR networks.

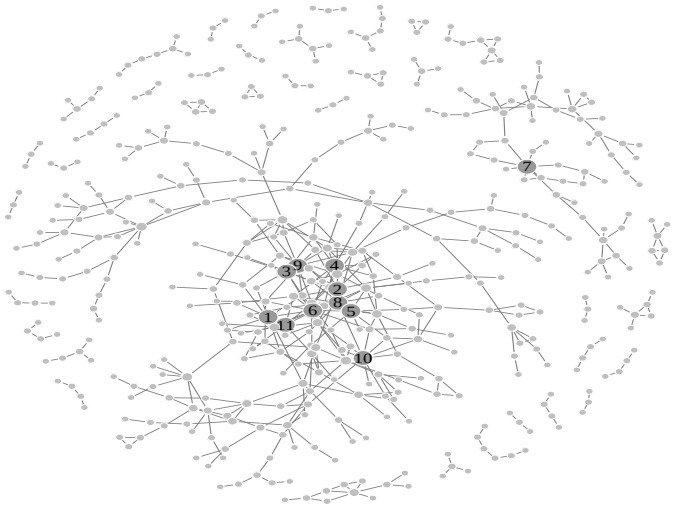


(a) Big-circle network: 500 nodes and 500 edges.

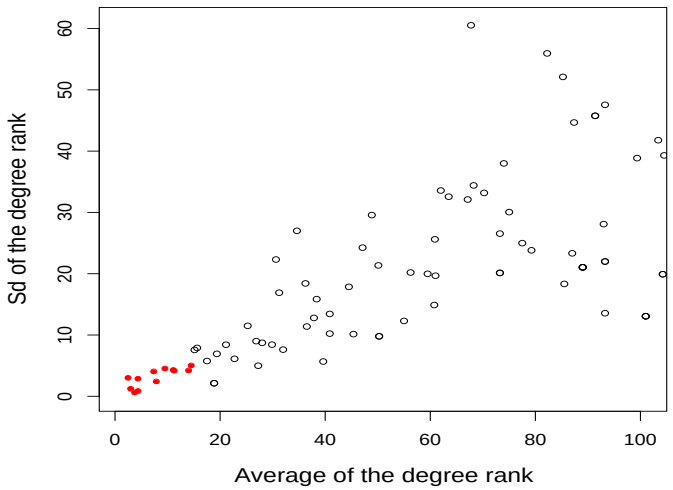


(b) Simulation results for Circle network. *x-axis*: the total number of edges detected; *y-axis*: the total number of correctly identified edges. The vertical grey line corresponds to the number of true edges.

Figure 10: Simulation results for Circle networks.



(a) Empirical network inferred from the real data (only showing components with at least three nodes). The gene annotation of the hub nodes (numbered) are given in Table 3.



(b) Degree ranks (for the 100 genes with highest degrees). Different symbols represent different genes. *Solid circles*: the 11 genes with highest degrees. *Circles*: the other genes. The sd(rank) of the top 11 genes are all smaller than 4.62 (4.62 is the 1% quantile of sd(rank) among all the 1217 genes), and thus are identified as hub nodes.

Figure 11: Result for the breast cancer expression data set.