Partial Correlation Estimation by Joint Sparse Regression Models

Jie PENG, Pei WANG, Nengfeng ZHOU, and Ji ZHU

In this article, we propose a computationally efficient approach—space (Sparse PArtial Correlation Estimation)—for selecting nonzero 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 nonzero partial correlation selection and the identification of hub variables, and also outperforms two existing methods. We then apply space to a microarray breast cancer dataset and identify a set of *hub genes* that 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.

KEY WORDS: Concentration network; Genetic regulatory network; High-dimension-low-sample-size; Lasso; Shooting.

1. INTRODUCTION

There has been a large amount of literature on *covariance selection*: the identification and estimation of nonzero entries in the inverse covariance matrix (a.k.a. *concentration matrix* or *precision matrix*) starting from the seminal paper by Dempster (1972). Covariance selection is very useful in elucidating associations among a set of random variables, as it is well known that nonzero entries of the concentration matrix correspond to nonzero partial correlations. Moreover, under Gaussianity, nonzero entries of the concentration matrix imply conditional dependency between corresponding variable pairs conditional on the rest of the variables (Edward 2000). Traditional methods do not work unless 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 datasets with $p > n$, for example, see Meinshausen and Buhlmann (2006), Yuan and Lin (2007), Li and Gui (2006), and Schafer and Strimmer (2005).

In this article, we propose a novel approach using sparse regression techniques for covariance selection. 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 $y_1, \ldots, y_p$. A *concentration network* 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 the partial correlation ρ^{ij} between y_i and y_j is nonzero. Note that, under the assumption that $y_1, \ldots, y_p$ are jointly normal, the partial correlation ρ^{ij} equals to $\mathrm{Corr}(y_i, y_j | y_{-(i,j)})$, where $y_{-(i,j)} = \{y_k : 1 \le k \ne i, j \le p\}$. Therefore, ρ^{ij} being nonzero is equivalent to y_i and y_j being conditionally dependent given all other variables $y_{-(i,j)}$. 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 (under normality assumption, this means that 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, di Bernardo, Lorenz, and Collins 2003; Jeong, Mason, Barabasi, and Oltvai 2001; Tegner, Yeung, Hasty, and Collins 2003). The proposed method is also particularly powerful in the identification of *hubs*: vertices (variables) that are connected to (have nonzero partial correlations 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 article is to propose a novel algorithm active-shooting for solving penalized optimization problems such as the lasso (Tibshirani 1996). This algorithm is computationally more efficient than the original shooting algorithm, which was first proposed by Fu (1998) and then extended by many others including Genkin, Lewis, and Madigan (2007) and Friedman, Hastie, Hofling, and Tibshirani (2007a). It enables us to implement the proposed procedure efficiently, such that we can conduct extensive simulation studies involving $\sim$1,000 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 article, they all assume sparsity 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. First, it does not take into account the intrinsic symmetry of

Jie Peng is Assistant Professor in Statistics, Department of Statistics, University of California, Davis, CA 95616 (E-mail: *jie@wald.ucdavis.edu*). Pei Wang is Assistant Member, Division of Public Health Science, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 (E-mail: *pwang@fhcrc.org*). Nengfeng Zhou is Ph.D. Candidate in Statistics, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *nfzhou@umich.edu*). Ji Zhu is Associate Professor in Statistics, Department of Statistics, University of Michigan, Ann Arbor, MI 48109 (E-mail: *jizhu@umich.edu*). Jie Peng and Pei Wang equally contributed to this article. Peng and Wang are partially supported by grant 1R01GM082802-01A1 from the National Institute of General Medical Sciences. Peng is also partially supported by grant DMS-0806128 from the National Science Foundation. Zhou and Zhu are partially supported by grants DMS-0705532 and DMS-0748389 from the National Science Foundation. We are grateful to two anonymous reviewers and an associate editor whose comments led to many improvements of the article.

Journal of the American Statistical Association
June 2009, Vol. 104, No. 486, Theory and Methods
DOI 10.1198/jasa.2009.0126

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 as suggested by their article, more or less equal effort is placed on building each neighborhood. This apparently is not the most efficient way to address the problem, unless the degree distribution of the network is nearly uniform. However, most real life networks have skewed degree distributions, such as the *power-law networks*. As observed by Schafer and Strimmer (2005), the neighborhood selection approach limits the number of edges connecting to each node. Therefore, it is not very effective in hub detection. On the contrary, 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 thus utilizes data more efficiently. We show by intensive simulation studies that our method performs better 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 previously, another class of methods employ the maximum likelihood framework. Yuan and Lin (2007) proposed a penalized maximum likelihood approach that performs model selection and estimation simultaneously and ensures the positive definiteness of the estimated concentration matrix. However, their algorithm cannot handle high dimensional data. The largest dimension considered by them is $p = 10$ in simulation and $p = 5$ in real data. Friedman, Hastie, and Tibshirani (2007b) proposed an efficient algorithm `glasso` to implement this method, such that it can be applied to problems with high dimensions. We show by simulation studies that, the proposed method performs better than `glasso` in both model selection and hub identification. Rothman, Bickel, Levina, and Zhu (2008) proposed another algorithm to implement the method of Yuan and Lin (2007). The computational cost is on the same order of `glasso`, but in general not as efficient as `glasso`. Li and Gui (2006) introduced a threshold gradient descent regularization procedure. Schafer and Strimmer (2005) proposed a shrinkage covariance estimation procedure to overcome the ill-conditioned problem of sample covariance matrix when $p > n$. There are also a large class of methods covering the situation where variables have a natural ordering (e.g., longitudinal data, time series, spatial data, or spectroscopy). See Wu and Pourahmadi (2003), Bickel and Levina (2008), Huang et al. (2006), and Levina et al. 2008, which are all based on the modified Cholesky decomposition of the concentration matrix. In this article, we, however, focus on the general case where an ordering of the variables is not available.

The rest of the article 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 likelihood based approach `glasso`. In Section 4, the proposed method is applied to a microarray expression dataset of $n = 244$ breast cancer tumor samples and $p = 1{,}217$ genes. In Section 5, we study the asymptotic properties of this procedure. A summary of the main results are given in Section 6. Technique details are provided in the Supplemental Material.

2. METHOD

2.1 Model

In this section, we describe a novel method for detecting pairs of variables having nonzero partial correlations among a large number of random variables based on iid samples. Suppose that, $(y_1, \ldots, y_p)^T$ has a joint distribution with mean 0 and covariance $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is a p by p positive definite matrix. Denote the partial correlation between y_i and y_j by ρ^{ij} $(1 \le i < j \le p)$. It is defined as $\mathrm{Corr}(\varepsilon_i, \varepsilon_j)$, where ε_i and ε_j are the prediction errors of the best linear predictors of y_i and y_j based on $y_{-(i,j)} = \{y_k : 1 \le k \ne i, j \le p\}$, respectively. Denote the *concentration matrix* $\mathbf{\Sigma}^{-1}$ by $(\sigma^{ij})_{p\times p}$. It is known that, $\rho^{ij} = -((\sigma^{ij})/(\sqrt{\sigma^{ii}\sigma^{jj}}))$. Let $y_{-i} = \{y_k{:}1 \le k \ne i \le p\}$. The following well-known result (Lemma 1) relates the estimation of partial correlations to a regression problem.

Lemma 1: For $1 \le i \le p$, y_i is expressed as $y_i = \sum_{j\ne i} \beta_{ij} y_j + \epsilon_i$, such that ε_i is uncorrelated with y_{-i} if and only if $\beta_{ij} = -(\sigma^{ij}/\sigma^{ii}) = \rho^{ij}\sqrt{(\sigma^{jj}/\sigma^{ii})}$. Moreover, for such defined $\beta_{ij}, \mathrm{var}(\epsilon_i) = (1/\sigma^{ii}), \mathrm{cov}(\epsilon_i, \epsilon_j) = \sigma^{ij}/(\sigma^{ii}\sigma^{jj})$.

Note that, under the normality assumption, $\rho^{ij} = \mathrm{Corr}(y_i, y_j|y_{-(i,j)})$ and in Lemma 1, we can replace "uncorrelated" with "independent." Because $\rho^{ij} = \mathrm{sign}(\beta_{ij})\sqrt{\beta_{ij}\beta_{ji}}$, the search for nonzero partial correlations can be viewed as a model selection problem under the regression setting. In this article, 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 datasets 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 $\mathbf{Y}^k = (y^k{}_1, \ldots, y^k{}_p)^T$ are iid observations from $(0, \mathbf{\Sigma})$, for $k = 1, \ldots, n$. Denote the sample of the ith variable as $\mathbf{Y}_i = (y^1{}_i, \ldots, y^n{}_i)^T$. Based on Lemma 1, we propose the following joint loss function

$$L_n(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{Y}) = \frac{1}{2}\left(\sum_{i=1}^{p} w_i \|\mathbf{Y}_i - \sum_{j\ne i} \beta_{ij}\mathbf{Y}_j\|^2\right) = \frac{1}{2}\left(\sum_{i=1}^{p} w_i \|\mathbf{Y}_i - \sum_{j\ne i} \rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\,\mathbf{Y}_j\|^2\right), \quad (1)$$

where $\boldsymbol{\theta} = (\rho^{12}, \cdots, \rho^{(p-1)p})^T$, $\boldsymbol{\sigma} = \{\sigma^{ii}\}_{i=1}^p$; $\mathbf{Y} = \{\mathbf{Y}^k\}_{k=1}^n$; and $\boldsymbol{w} = \{w_i\}_{i=1}^p$ are nonnegative weights. For example, we can choose $w_i = 1/\mathrm{var}\,(\varepsilon_i) = \sigma^{ii}$ to weigh individual regressions in the joint loss function according to their residual variances, as is done in regression with heteroscedastic noise. We propose to

estimate the partial correlations $\boldsymbol{\theta}$ by minimizing a penalized loss function

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

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

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

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

1. In `space`, sparsity is used for the partial correlations $\boldsymbol{\theta}$ as a whole view. However, in the neighborhood selection approach, sparsity is imposed on each neighborhood. The former treatment is more natural and utilizes the data more efficiently, especially for networks with hubs. A prominent example is the genetic regulatory network, where master regulators are believed to exist and are of great interest.
2. 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 $\text{sign}(\widehat{\beta}_{ji})$, which may lead to contradictory neighborhoods.
3. 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` versus $(p-1)^2$ for `MB`), and thus improves the efficiency.
4. 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 that 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 nonzero partial correlation selection and hub detection.

As compared with the penalized maximum likelihood approach `glasso` (Friedman et al. 2007b), the simulation study in Section 3 shows that `space` also outperforms `glasso` in both edge detection and hub identification under all settings that we have considered. In addition, `space` has the following advantages.

1. The complexity of `glasso` is $O(p^3)$, whereas 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.
2. 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 Smoothly Clipped Absolute Deviation (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.
3. In Section 5, we prove 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 penalized maximum likelihood approach.

Note that, in the penalized loss function (2), $\boldsymbol{\sigma}$ needs to be specified. We propose to estimate $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ by a two-step iterative procedure. Given an initial estimate $\boldsymbol{\sigma}^{(0)}$ of $\boldsymbol{\sigma}$, $\boldsymbol{\theta}$ is estimated by minimizing the penalized loss function (2), whose implementation is discussed in Section 2.2. Then given the current estimates $\boldsymbol{\theta}^{(c)}$ and $\boldsymbol{\sigma}^{(c)}$, $\boldsymbol{\sigma}$ is updated based on Lemma 1: $1/\widehat{\sigma}^{ii} = (1/n)\| \mathbf{Y}_i - \Sigma_{j \ne i} \widehat{\beta}_{ij}^{(c)} \mathbf{Y}_j \|^2$, where $\widehat{\beta}_{ij}^{(c)} = (\rho^{ij})^{(c)} \sqrt{((\sigma^{ij})^{(c)}/(\sigma^{ii})^{(c)})}$. We then iterate between these two steps until convergence. Because $1/\sigma^{ii} \le \text{var}(y_i) = \sigma_{ii}$, we can use $1/\widehat{\sigma}_{ii}$ as the initial estimate of σ^{ii}, where $\widehat{\sigma}_{ii} = 1/(n-1)\Sigma_{k=1}^n (y_i^k - \bar{y}_i)^2$ is the sample variance of y_i. Our simulation study shows that it usually takes no more than three iterations for this procedure to stabilize.

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 reformulate 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 efficiently.

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

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

where $\tilde{\sigma}^{ii} = \sigma^{ii}/w_i$ $(i = 1, \ldots, p)$. Then it is easy to see that the loss function (1) equals to $(1/2)\|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\boldsymbol{\theta}\|_2^2$, and the corresponding ℓ_1 minimization problem is equivalent to: $\min_{\boldsymbol{\theta}} \frac{1}{2}\|\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}\boldsymbol{\theta}\|_2^2 + \lambda\|\boldsymbol{\theta}\|_1$. 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 can be efficiently implemented by taking into consideration this structure (see the Supplemental Material 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. `Active-shooting` is a modification of the `shooting` algorithm, which was first proposed by Fu (1998) and then extended by many others including Genkin et al. (2007) and Friedman et al. (2007a). `Active-shooting` exploits the sparse nature of sparse penalization problems in a more efficient way, and is therefore computationally much faster. This is crucial for applying `space` for large p or n. It can be shown that the computational cost of `space` is min $(O(np^2), O(p^3))$, which is the same as 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 et al. 2007a). (Remark by the authors: after this article was submitted, recently the `active-shooting` idea was also proposed by Friedman, Hastie, and Tibshirani (2008).)

Finally, note that the concentration matrix should be positive definite. In principle, the proposed method (or more generally, the regression based methods) does not guarantee the positive definiteness of the resulting estimator, whereas the likelihood based method by Yuan and Lin (2007) and Friedman et al. (2007b) assures the positive definiteness. Whereas admitting that this is one limitation of the proposed method, we argue that, because we are more interested in model selection than parameter estimation in this article, we are less concerned with this issue. Moreover, in Section 5, we show that the proposed estimator is consistent under a set of suitable assumptions. Therefore, it is asymptotically positive definite. Indeed, the `space` estimators are rarely nonpositive-definite under the high dimensional sparse settings that we are interested in. More discussions on this issue can be found in Section 3.

2.3 Active Shooting

In this section, we propose a computationally very efficient algorithm `active-shooting` for solving 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

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

where $\mathbf{Y} = (y_1, \ldots, y_n)^T$, $\mathbf{X} = (x_{ij})_{n\times p} = (\mathbf{X}_1 : \ldots : \mathbf{X}_p)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$. The `shooting` algorithm proceeds as follows:

1. Initial step: for $j = 1, \ldots, p$,

$$\begin{aligned}\beta_j^{(0)} &= \arg\min_{\beta_j}\left\{\tfrac{1}{2}\|\mathbf{Y} - \beta_j\mathbf{X}_j\|^2 + \gamma|\beta_j|\right\} \\ &= \operatorname{sign}(\mathbf{Y}^T\mathbf{X}_j)\,\frac{(|\mathbf{Y}^T\mathbf{X}_j| - \gamma)_+}{\mathbf{X}_j^T\mathbf{X}_j},\end{aligned} \tag{4}$$

   where $(x)_+ = xI_{(x>0)}$.
2. For $j = 1, \ldots, p$, update $\boldsymbol{\beta}^{(old)} \to \boldsymbol{\beta}^{(new)}$:

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

   where $\boldsymbol{\epsilon}^{(old)} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(old)}$.
3. Repeat step 2 Until convergence.

At each updating step of the `shooting` algorithm, we define the set of currently nonzero coefficients as the *active set*. Because under sparse models, 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 $\Lambda = \{k$: 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 $j = 1$ to p, update β_j with all other coefficients fixed at the current value as in Equation (5). If no β_j 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 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, as can be seen from Table 1, `active-shooting` takes much fewer iterations to converge (where one iteration is counted whenever an attempt to update a β_j is made). In particular, it takes less than 30 sec (on average) to fit the `space` model by `active-shooting` (implemented in c code) for cases with 1,000 variables, 200 samples and when the resulting model has around 1000 nonzero partial correlations on a server with dual core 3GHz processor and 4 GB RAM. This great computational advantage enables us to conduct large scale simulation studies to examine the performance of the proposed method (Section 3).

2.4 Tuning

The choice of the tuning parameter λ is of great importance. Because the `space` method uses a lasso criterion, methods that

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

p	coef. #	`shooting`	`active-shooting`
200	14	29,600	4,216
500	25	154,000	10,570
1,000	28	291,000	17,029

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, Hastie, and Tibshirani (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 (2005), the likelihood based method in Huang et al. (2006), and the cross-validation and bootstrap methods in Li and Gui (2006). In this article, we propose to use a "BIC-type" criterion for selecting the tuning parameter mainly due to its simplicity and computational easiness. For a given λ, denote the space estimator by $\widehat{\boldsymbol{\theta}}_\lambda = \{\widehat{\rho}_\lambda^{ij} : 1 \le i < j \le p\}$ and $\widehat{\boldsymbol{\sigma}}_\lambda = \{\widehat{\sigma}_\lambda^{ii} : 1 \le i \le p\}$. The corresponding residual sum of squares for the i-th regression: $y_i = \sum_{j \ne i} \beta_{ij} y_j + \epsilon_i$ is

$$RSS_i(\lambda) = \sum_{k=1}^{n} \left(y_i^k - \sum_{j \ne i} \widehat{\rho}_\lambda^{ij} \sqrt{\frac{\widehat{\sigma}_\lambda^{jj}}{\widehat{\sigma}_\lambda^{ii}}} y_j^k \right)^2.$$

We then define a "BIC-type" criterion for the i-th regression as

$$BIC_i(\lambda) = n \times \log(RSS_i(\lambda)) + \log n \times \#\{j : j \ne i, \widehat{\rho}_\lambda^{ij} \ne 0\}. \quad (6)$$

Finally, we define $BIC(\lambda) := \sum_{i=1}^{p} BIC_i(\lambda)$ and select λ by minimizing $BIC(\lambda)$. This method is referred to as space.joint hereafter.

In Yuan and Lin (2007), a BIC criterion is proposed for the penalized maximum likelihood approach. Namely

$$BIC(\lambda) : = n \times \left[-\log |\widehat{\boldsymbol{\Sigma}}_\lambda^{-1}| + \operatorname{trace}(\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} \mathbf{S}) \right] + \log n \times \#\{(i,j) : 1 \le i \le j \le p, \widehat{\sigma}_\lambda^{ij} \ne 0\}, \quad (7)$$

where $\mathbf{S}$ is the sample covariance matrix, and $\widehat{\boldsymbol{\Sigma}}_\lambda^{-1} = (\widehat{\sigma}_\lambda^{ij})$ is the estimator under λ. In this article, we refer to this method as glasso.like. For the purpose of comparison, we also consider the selection of the tuning parameter for MB. Because MB essentially performs p individual lasso regressions, the tuning parameter can be selected for each of them separately. Specifically, we use criterion (6) (evaluated at the corresponding MB estimators) to select the tuning parameter λ_i for the i-th regression. We denote this method as MB.sep. Alternatively, as suggested by Meinshausen and Buhlmann (2006), when all Y_i are standardized to have sample standard deviation one, the same $\lambda(\alpha) = \sqrt{n}\Phi^{-1}(1 - (\alpha/2p^2))$ is applied to all regressions. Here, Φ is the standard normal cumulative distribution function; α is used to control the false discovery rate and is usually taken as 0.05 or 0.1. We denote this method as MB.alpha. These methods are examined by the simulation studies 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 penalized likelihood 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., $\#\{j : \widehat{\rho}^{ij} \ne 0, j \ne i\}$). 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 one. Iterations are used for these space methods as discussed in Section 2.1. For space.dew and space.sw, the initial weights are taken to be one (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 because the procedure converges very fast and more iterations result in essentially the same estimator. 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 mainly consider two different types of degree distributions and denote their corresponding networks by Hub network and Power-law network (details are given later). Given an undirected network with p nodes, the initial "concentration matrix" $(\tilde{\sigma}^{ij})_{p \times p}$ is generated by

$$\tilde{\sigma}^{ij} = \begin{cases} 1, & i = j; \\ 0, & i \ne j, i \nsim j; \\ \sim U(\mathcal{D}) & i \ne j, i \sim j, \end{cases} \quad (8)$$

where $i \sim j$ means that there is an edge between nodes i and j, $i \nsim j$ means otherwise; and domain $\mathcal{D} = [-1, -0.5] \cup [0.5, 1]$. We then rescale the non-zero elements in the preceding matrix to assure positive definiteness. Specifically, for each row, we first sum the absolute values of the off-diagonal entries, and then divide each off-diagonal entry by 1.5 fold of the sum. We then average this rescaled matrix with its transpose to ensure symmetry. Finally the diagonal entries are all set to be one. This process results in diagonal dominance. Denote the final matrix as $\mathbf{A}$. The covariance matrix $\boldsymbol{\Sigma}$ is then determined by

$$\boldsymbol{\Sigma}(i,j) = \mathbf{A}^{-1}(i,j) / \sqrt{\mathbf{A}^{-1}(i,i)\mathbf{A}^{-1}(j,j)}.$$

Finally, iid samples $\{\mathbf{Y}^k\}_{k=1}^n$ are generated from Normal($\mathbf{0}, \boldsymbol{\Sigma}$). Note that, $\boldsymbol{\Sigma}(i,i) = 1$, and $\boldsymbol{\Sigma}^{-1}(i,i) = \sigma^{ii} \ge 1$.

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, at most, four. This setting is designed to mimic the genetic regulatory networks, which usually contains a few hub genes plus

many other genes with only a few edges. A network consisting of five such modules is shown in Figure 1(a). In this network, there are $p = 500$ nodes and 568 edges. The simulated nonzero 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 datasets each consisting of $n = 250$ iid samples.

We then evaluate each method at a series of different values of the tuning 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) versus the number of total detected edges (N_t) averaged across the 50 independent datasets 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%. Here, sensitivity is defined as the ratio of the number of correctly detected edges to the total number of true edges; and specificity is defined as the ratio of the number of correctly detected edges to the number of total detected edges. 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 is 568.

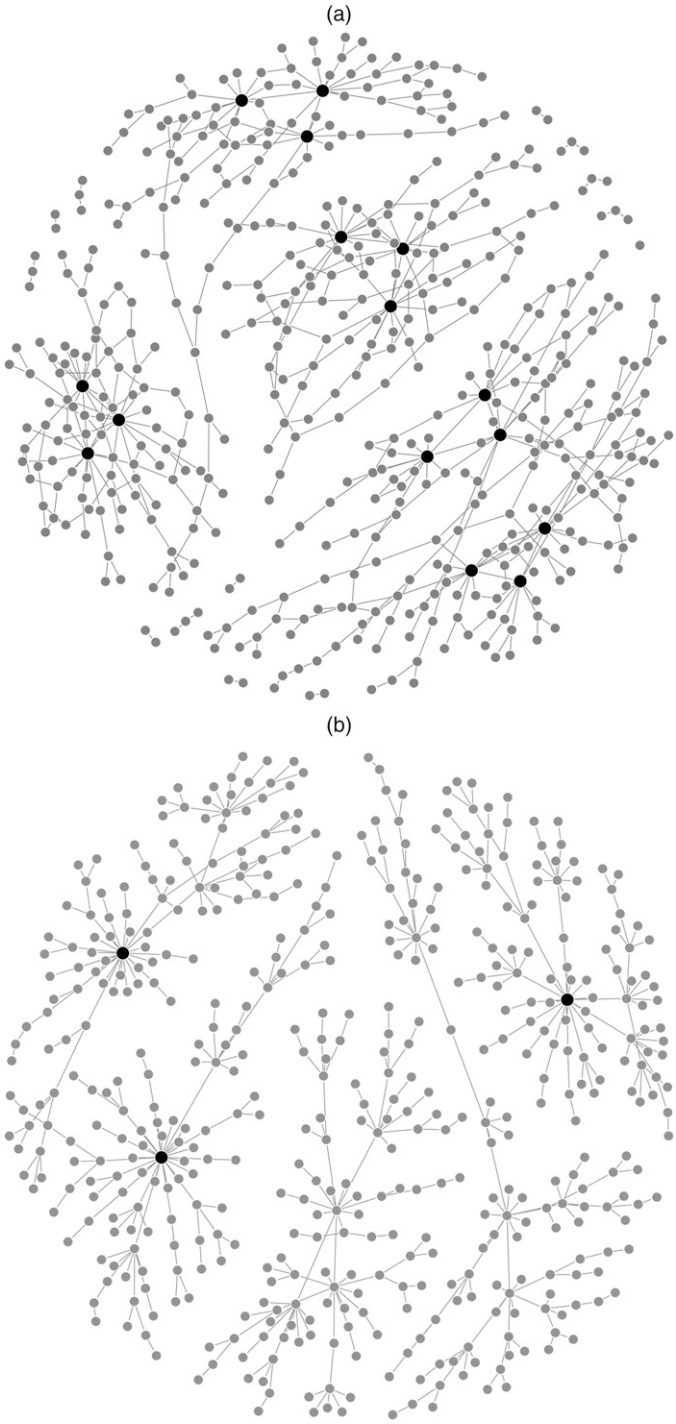

Figure 1. Topology of simulated networks. (a) Hub network: 500 nodes and 568 edges. 15 nodes (in black) have degrees of around 15. (b) Power-law network: 500 nodes and 495 edges. 3 nodes (in black) have degrees at least 20.

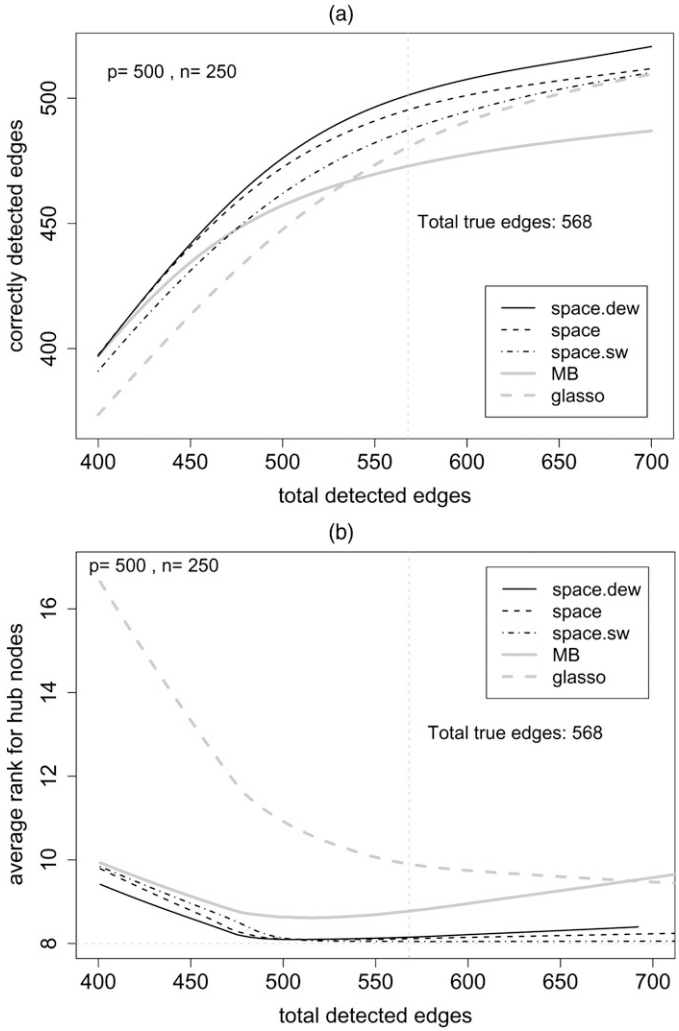

Figure 2. Simulation results for Hub network. (a) *x-axis*: the number of total detected edges (i.e., the total number of pairs (i, j) with $\widehat{\rho}^{ij} \neq 0$; *y-axis*: the number of correctly identified edges. The vertical grey line corresponds to the number of true edges. (b) *x-axis*: the number of total detected edges; *y-axis*: the average rank of the estimated degrees of the 15 true hub nodes.

In terms of hub detection, for a given N_t, a rank 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 nonhub 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 for a large range of N_t. 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 $p =$ 1,000 and various sample sizes with $n = 200$, 300, and 500. The simulated network includes 10 disjointed modules of size 100 each and has 1,163 edges in total. Nonzero partial correlations form a similar distribution as that of the $p = 500$ network discussed previously. The ROC curves for `space.dew`, `MB`, and `glasso` resulted from these simulations are shown in Figure 3. When false discovery rate (FDR, defined as 1-specificity) is controlled at 0.05, the power (=sensitivity) for detecting correct edges is 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 $p = 1{,}000$, 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 1,000, with only 20% more samples (that is $p = 500$, $n = 250$ versus $p =$ 1,000, $n = 300$), all three methods achieve similar powers. This is presumably because the larger network ($p = 1{,}000$) is 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` performs best among the three methods.

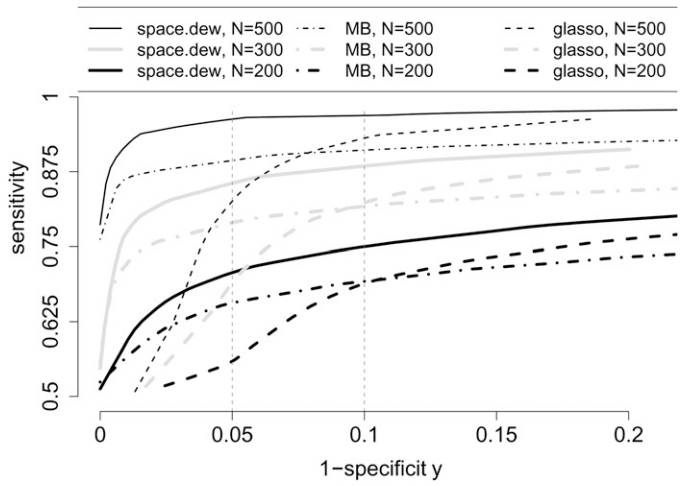

Figure 3. Hub network: ROC curves for different samples sizes ($p =$ 1,000).

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

Network	p	n	`space.dew`	`MB`	`glasso`
Hub-network	500	250	0.844	0.784	0.655
		200	0.707	0.656	0.559
Hub-network	1,000	300	0.856	0.790	0.690
		500	0.963	0.894	0.826
Power-law network	500	250	0.704	0.667	0.580

We then investigate the performance of these methods at the selected tuning parameters (see Section 2.4 for details). For the preceding Hub network with $p = 1{,}000$ nodes and $n = 200$, 300, 500, the results are reported in Table 3. As can be seen from the table, BIC based approaches tend to select large models (compared with the true model that has 1,163 edges). `Space.joint` and `MB.sep` perform similarly in terms of specificity, and `glasso.like` works considerably worse than the other two in this regard. On the other hand, `space.joint` and `glasso.like` performs similarly and are better than `MB.sep` in terms of sensitivity. In contrast, `MB.alpha` selects very small models and thus results in very high specificity, but very low sensitivity. In terms of hub identification, `space.joint` apparently performs better than other methods (indicated by a smaller average rank over 30 true hub nodes). Moreover, the performances of all methods improve with sample size.

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. It can be seen that there

Table 3. Edge detection under the selected tuning parameter λ. *For average rank*, the optimal value is 15.5. For MB.alpha, $\alpha = 0.05$ is used.

Sample size	Method	Total edge detected	Sensitivity	Specificity	Average rank
$n = 200$	`space.joint`	1,357	0.821	0.703	28.6
	`MB.sep`	1,240	0.751	0.703	57.5
	`MB.alpha`	404	0.347	1.00	175.8
	`glasso.like`	1,542	0.821	0.619	35.4
$n = 300$	`space.joint`	1,481	0.921	0.724	18.2
	`MB.sep`	1,456	0.867	0.692	30.4
	`MB.alpha`	562	0.483	1.00	128.9
	`glasso.like`	1,743	0.920	0.614	21
$n = 500$	`space.joint`	1,525	0.980	0.747	16.0
	`MB.sep`	1,555	0.940	0.706	16.9
	`MB.alpha`	788	0.678	1.00	52.1
	`glasso.like`	1,942	0.978	0.586	16.5

are three obvious hub nodes in this network with degrees of at least 20. The simulated nonzero 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 datasets each consisting of $n = 250$ iid samples. We then compare the number of correctly detected edges by various methods. The result is shown in Figure 4. 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, whereas `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.

Comments. These simulation results suggest that when the (concentration) networks are reasonably sparse, we should be able to characterize their structures with only a couple-of-hundreds of samples when there are a couple of thousands of nodes. In addition, `space.dew` outperforms `MB` by at least 6% on the power of edge detection under all simulation settings above 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 with `glasso`, the improvement of `space.dew` is at least 15% when FDR is controlled at 0.05, and the advantages 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`. We have also applied `space` methods, `MB` and `glasso` on networks with nearly uniform degree distributions generated by following the simulation procedures in Meinshausen and Buhlmann (2006), as well as the autoregressive network discussed in Yuan and Lin (2007) and Friedman et al. (2007b). For these cases, the `space` methods perform comparably, if not better than, the other two methods. However, for these networks without hubs, the advantages of `space` become smaller compared with the results on the networks with hubs. Due to space limitation, detailed results are not reported here.

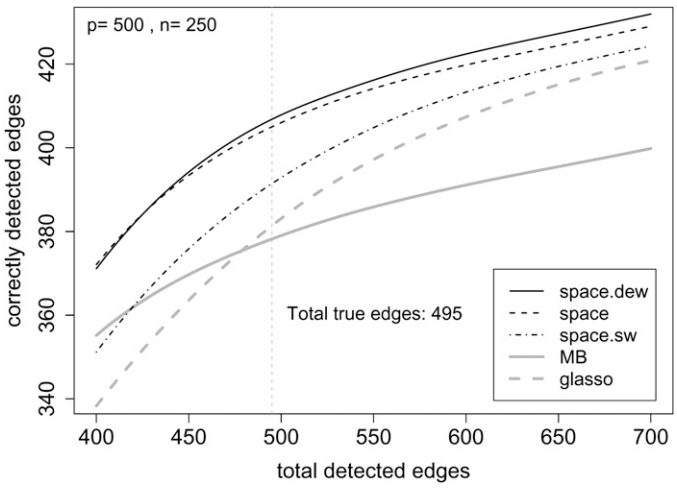

Figure 4. Simulation results for Power-law network. *x-axis*: the number of total detected edges; *y-axis*: the number of correctly identified edges. The vertical grey line corresponds to the number of true edges.

We conjecture that, under the sparse and high dimensional setting, the superior performance in model selection of the regression based method `space` over the penalized likelihood method is partly due to its simpler quadratic loss function. Moreover, because `space` ignores the correlation structure of the regression residuals, it amounts to a greater degree of regularization, which may render additional benefits under the sparse and high dimensional setting.

In terms of parameter estimation, we compare the entropy loss of the three methods. We find that, they perform similarly when the estimated models are of small or moderate size. When the estimated models are large, `glasso` generally performs better in this regard than the other two methods. Because the interest of this article lies in model selection, detailed results of parameter estimation are not reported here.

As discussed earlier, one limitation of `space` is its lack of assurance of positive definiteness. However, for simulations reported previously, the corresponding estimators we have examined (over 3,000 in total) are all positive definite. To further investigate this issue, we design a few additional simulations. We first consider a case with a similar network structure as the Hub network, however having a nearly singular concentration matrix (the condition number is 16,240; as a comparison, the condition number for the original Hub network is 62). For this case, the estimate of `space` remains positive definite until the number of total detected edges increases to 50,000; whereas the estimate of `MB` remains positive definite until the number of total detected edges is more than 23,000. Note that, the total number of true edges of this model is only 568, and the model selected by `space.joint` has 791 edges. In the second simulation, we consider a denser network ($p = 500$ and the number of true edges is 6,188) with a nearly singular concentration matrix (condition number is 3,669). Again, we observe that, the `space` estimate only becomes nonpositive-definite when the estimated models are huge (the number of detected edges is more than 45,000). This suggests that, for the regime we are interested in in this article (the sparse and high dimensional setting), nonpositive-definiteness does not seem to be a big issue for the proposed method, as it only occurs when the resulting model is huge and thus very far away from the true model. As long as the estimated models are reasonably sparse, the corresponding estimators by `space` remain positive definite. We believe that this is partly due to the heavy shrinkage imposed on the off-diagonal entries to ensure sparsity.

Finally, we investigate the performance of these methods when the observations come from a nonnormal distribution. Particularly, we consider the multivariate t_{df} distribution with $df = 3, 6, 10$. The performances of all three methods deteriorate compared with the normal case, however the overall picture in terms of relative performance among these methods remains essentially unchanged (detailed results not shown).

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 of 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, because breast cancer is a complex disease, the tumor process cannot be understood by only analyzing individual genes. There is a 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 data and patient clinical outcomes for 244 of these samples are available online and are used in this article. 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 `space` method on this dataset as follows. First, for each expression array, we perform the global normalization by centering the mean to zero and scaling the median absolute deviation to one. Then we focus on a subset of $p = 1{,}217$ genes/clones whose expression levels are significantly associated with tumor progression (p values from univariate Cox models < 0.0008, corresponding FDR = 0.01). We estimate the partial correlation matrix of these 1,217 genes with `space.dew` for a series of λ values. The degree distribution of the inferred network is heavily skewed to the right. Specifically, when 629 edges are detected, 598 out of the 1,217 genes do not connect to any other genes, whereas five genes have degrees of at least 10. The power-law parameter of this degree distribution is $\alpha = 2.56$, which is consistent with the findings in the literature for GRNs (Newman 2003). The topology of the inferred network is shown in Figure 5(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 5(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 *p*27*kip*1 and the cell adhesion molecule E-cadherin. This gene is essential for the expression of approximately 50% of estrogen-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 loss of heterozygosity 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

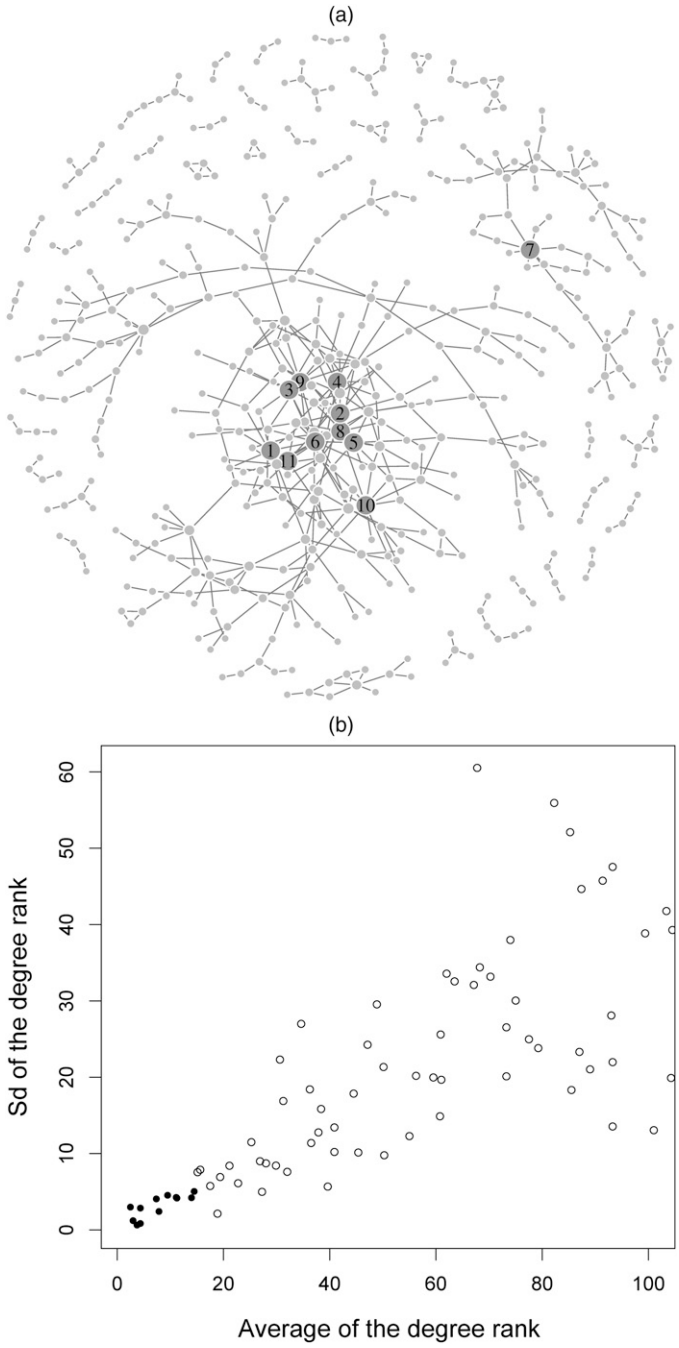

Figure 5. Results for the breast cancer expression dataset. (a) Network inferred from the real data (only showing components with at least three nodes). The gene annotation of the hub nodes (numbered) are given in Table 4. (b) Degree ranks (for the 100 genes with highest degrees). Different circles represent 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 1,217 genes), and thus are identified as hub nodes.

Table 4. Annotation of hub genes

Index	Gene Symbol	Summary Function (GO)
1	CENPA	Encodes a centromere protein (nucleosome assembly)
2	*NA*	*Annotation not available*
3	KNSL6	Anaphase chromosome segregation (cell proliferation)
4	STK12	Regulation of chromosomal segregation (cell cycle)
5	*NA*	*Annotation not available*
6	URLC9	*Annotation not available* (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)

breast cancer are worth of further investigation. The functions of all hub genes are briefly summarized in Table 4.

5. ASYMPTOTICS

In this section, we show that under appropriate conditions, the `space` procedure achieves both model selection consistency and estimation consistency. Use $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\sigma}}$ to denote the true parameters of $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. As discussed in Section 2.1, when $\boldsymbol{\sigma}$ is given, $\boldsymbol{\theta}$ is estimated by solving the following ℓ_1 penalization problem:

$$\widehat{\boldsymbol{\theta}}^{\lambda_n}(\boldsymbol{\sigma}) = \arg\min_{\boldsymbol{\theta}} \; L_n(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{Y}) + \lambda_n \|\boldsymbol{\theta}\|_1, \tag{9}$$

where the *loss function* $L_n(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{Y}) := \frac{1}{n}\sum_{k=1}^{n} L(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{Y}^k)$, with, for $k = 1, \ldots, n$

$$L(\boldsymbol{\theta}, \boldsymbol{\sigma}, \mathbf{Y}^k) := \frac{1}{2}\sum_{i=1}^{p} w_i (y_i^k - \sum_{j \neq i} \sqrt{\sigma^{jj}/\sigma^{ii}}\rho^{ij} y_j^k)^2. \tag{10}$$

Throughout this section, we assume $\mathbf{Y}^1, \ldots, \mathbf{Y}^n$ are iid samples from $N_p(\mathbf{0}, \boldsymbol{\Sigma})$. The Gaussianity assumption here can be relaxed by assuming appropriate tail behaviors of the observations. The assumption of zero mean is simply for exposition simplicity. In practice, in the loss function (9), $\mathbf{Y}^k$ can be replaced by $\mathbf{Y}^k - \bar{\mathbf{Y}}$ where $\bar{\mathbf{Y}} = \frac{1}{n}\sum_{k=1}^{n} \mathbf{Y}^k a$ is the sample mean. All results stated in this section still hold under that case.

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 $0 < \Lambda_{\min}(\bar{\boldsymbol{\theta}}) \leq \Lambda_{\max}(\bar{\boldsymbol{\theta}}) < \infty$, such that the true covariance $\bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\sigma}})$ satisfies: $0 < \Lambda_{\min}(\bar{\boldsymbol{\theta}}) \leq \lambda_{\min}(\bar{\boldsymbol{\Sigma}}) \leq \lambda_{\max}(\bar{\boldsymbol{\Sigma}}) \leq \Lambda_{\max}(\bar{\boldsymbol{\theta}}) < \infty$, where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of a matrix, respectively.

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

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

where for $1 \leq i < j \leq p$, $1 \leq t < s \leq p$,

$$\overline{L}''_{ij,ts}(\bar{\boldsymbol{\theta}}, \bar{\boldsymbol{\sigma}}) := E_{(\bar{\theta}, \bar{\sigma})} \left(\frac{\partial^2 L(\boldsymbol{\theta}, \boldsymbol{\sigma}, Y)}{\partial \rho^{ij} \partial \rho^{ts}} |_{\theta = \bar{\theta}, \sigma = \bar{\sigma}} \right).$$

Condition C0 says that the weights are bounded away from zero and infinity. Condition C1 assumes that the eigenvalues of the true covariance matrix $\overline{\boldsymbol{\Sigma}}$ 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.

Furthermore, because $\bar{\boldsymbol{\sigma}}$ is usually unknown, it needs to be estimated. Use $\hat{\boldsymbol{\sigma}} = \hat{\boldsymbol{\sigma}}_n = \{\hat{\sigma}^{ii}\}_{i=1}^{p}$ to denote one estimator. The following condition says **D**: For any $\eta > 0$, there exists a constant $C > 0$, such that for sufficiently large n, $\max_{1 \leq i \leq p} |\hat{\sigma}^{ii} - \bar{\sigma}^{ii}| \leq C(\sqrt{\frac{\log n}{n}})$ holds with probability at least $1 - O(n^{-\eta})$.

Note that, the theorems following hold even when $\hat{\boldsymbol{\sigma}}$ is obtained based on the same dataset from which $\boldsymbol{\theta}$ is estimated as long as condition D is satisfied. The following proposition says that, when $p < n$, we can get an estimator of $\bar{\boldsymbol{\sigma}}$ satisfying condition D by simply using the residuals of the ordinary least square fitting.

Proposition 1: Suppose $\mathbf{Y} = [\mathbf{Y}^1 : \ldots : \mathbf{Y}^n]$ is a $p \times n$ data matrix with iid columns $\mathbf{Y}^i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$. Further suppose that $p = p_n$ such that $p/n \leq 1 - \delta$ for some $\delta > 0$; and $\boldsymbol{\Sigma}$ has a bounded condition number (that is assuming condition C1). Let $\bar{\sigma}^{ii}$ denote the (i,i)-th element of $\boldsymbol{\Sigma}^{-1}$; and let $\mathbf{e}_i$ denote the residual from regressing $\mathbf{Y}^i$ on to $\mathbf{Y}_{(-i)} := [\mathbf{Y}^1 : \ldots : \mathbf{Y}^{i-1} : \mathbf{Y}^{i+1} : \ldots : \mathbf{Y}^n]$, that is

$$\mathbf{e}_i = \mathbf{Y}^i - \mathbf{Y}_{(-i)}^T \left(\mathbf{Y}_{(-i)} \mathbf{Y}_{(-i)}^T \right)^{-1} \mathbf{Y}_{(-i)} \mathbf{Y}^i.$$

Define $\hat{\sigma}^{ii} = 1/\hat{\sigma}_{ii,-(i)}$, where

$$\hat{\sigma}_{ii,-(i)} = \frac{1}{n - p - 1} \mathbf{e}_i^T \mathbf{e}_i,$$

then condition D holds for $\{\hat{\sigma}^{ii}\}_{i=1}^{p}$.

The proof of this proposition is omitted due to space limitation.

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

Theorem 1: Suppose that conditions C0-C1 and D are satisfied. Suppose further that $q_n \sim o(\sqrt{(n/\log n)})$, $\lambda_n \sqrt{(n/\log n)} \to \infty$ and $\sqrt{q_n}\lambda_n \sim o(1)$, as $n \to \infty$. Then there

exists a constant $C(\bar{\boldsymbol{\theta}})>0$, such that for any $\eta>0$, the following events hold with probability at least $1-O(n^{-\eta})$:

(i) there exists a solution $\hat{\boldsymbol{\theta}}^{\mathcal{A},\lambda_n}=\hat{\boldsymbol{\theta}}^{\mathcal{A},\lambda_n}(\hat{\boldsymbol{\sigma}})$ of the restricted problem:

$$\min_{\boldsymbol{\theta}:\boldsymbol{\theta}_{\mathcal{A}^c}=0} L_n(\boldsymbol{\theta},\hat{\boldsymbol{\sigma}},\mathbf{Y})+\lambda_n\|\boldsymbol{\theta}\|_1, \tag{11}$$

where the loss function L_n is defined via *(10)*.
(ii) (estimation consistency) any solution $\hat{\boldsymbol{\theta}}^{\mathcal{A},\lambda_n}$ of the restricted problem *(11)* satisfies: $\|\hat{\boldsymbol{\theta}}^{\mathcal{A},\lambda_n}-\bar{\boldsymbol{\theta}}_A\|_2 \le C(\bar{\boldsymbol{\theta}})\sqrt{q_n}\lambda_n$.
(iii) (sign consistency) if further assume that the signal sequence satisfies: $\frac{s_n}{\sqrt{q_n}\lambda_n}\to\infty, n\to\infty$, then $\operatorname{sign}(\bar{\theta}_{ij}^{\mathcal{A},\lambda_n})=\operatorname{sign}(\bar{\theta}_{ij})$, for all $1\le i<j\le p$.

Theorem 2: Suppose that conditions C0–C2 and D are satisfied. Suppose further that $p=O(n^{\kappa})$ for some $\kappa\ge 0$; $q_n\sim o(\sqrt{(n/\log n)})$, $\sqrt{(q_n\log n/n)}=o(\lambda_n), \lambda_n\sqrt{(n/\log n)}\to\infty$ and $\sqrt{q_n}\lambda_n\sim o(1)$, as $n\to\infty$. Then for any $\eta>0$, for n sufficiently large, the solution of (11) satisfies

$$P_{(\bar{\boldsymbol{\theta}},\bar{\boldsymbol{\sigma}})}\left(\max_{(i,j)\in\mathcal{A}^c}|L'_{n,ij}(\hat{\boldsymbol{\theta}}^{\mathcal{A},\lambda_n},\hat{\boldsymbol{\sigma}},\mathbf{Y})|<\lambda_n\right)\ge 1-O(n^{-\eta}),$$

where $L'_{n,ij}:=(\partial L_n/\partial\rho^{ij})$.

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

(i) there exists a solution $\hat{\boldsymbol{\theta}}^{\lambda_n}=\hat{\boldsymbol{\theta}}^{\lambda_n}(\hat{\boldsymbol{\sigma}})$ of the ℓ_1 penalization problem

$$\min_{\theta} L_n(\boldsymbol{\theta},\hat{\boldsymbol{\sigma}},\mathbf{Y})+\lambda_n\|\boldsymbol{\theta}\|_1, \tag{12}$$

where the loss function L_n is defined via *(10)*.
(ii) (estimation consistency): any solution $\hat{\boldsymbol{\theta}}^{\lambda_n}$ of *(12)* satisfies:

$$\|\hat{\boldsymbol{\theta}}^{\lambda_n}-\bar{\boldsymbol{\theta}}\|_2\le C(\bar{\boldsymbol{\theta}})(\sqrt{q_n}\lambda_n).$$

(iii) (sign consistency):

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

Proofs of these theorems are given in the Supplemental Material. Finally, 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 $\alpha\ge 0$, as long as the size and the number of true edges of each component satisfy the corresponding conditions in Theorem 2.

Remark 1: The condition $\lambda_n\sqrt{(n/\log n)}\to\infty$ is indeed implied by the condition $\sqrt{(q_n\log n/n)}=o(\lambda_n)$ as long as q_n does not go to zero. Moreover, under the ''worst case'' scenario, that is when q_n is almost in the order of $\sqrt{(n/\log n)}$, λ_n needs to be nearly in the order of $n^{-1/4}$. On the other hand, for the ''best case'' scenario, that is when $q_n=O(1)$ (for example, when the dimension p is fixed), the order of λ_n can be nearly as small as $n^{-1/2}$ (within a factor of $\log n$). Consequently, the ℓ_2-norm distance of the estimator from the true parameter is in the order of $\sqrt{\log n/n}$, with probability tending to one.

6. SUMMARY

In this article, we propose a joint sparse regression model—`space`—for selecting nonzero partial correlations under the high-dimension-low-sample-size setting. By controlling the overall sparsity of the partial correlation matrix, `space` is able to automatically adjust for different neighborhood sizes and thus to use data more effectively. The proposed method also explicitly employs the symmetry among the partial correlations, which also helps to improve efficiency. Moreover, this joint model makes it easy to incorporate prior knowledge about network structure. We develop a fast algorithm `active-shooting` to implement the proposed procedure, which can be readily extended to solve some other penalized optimization problems. We also propose a ''BIC-type'' criterion for the selection of the tuning parameter. With extensive simulation studies, we demonstrate that this method achieves good power in nonzero partial correlation selection as well as hub identification, and also performs favorably compared with two existing methods. The impact of the sample size and dimensionality has been examined on simulation examples as well. We then apply this method on a microarray dataset of 1,217 genes from 244 breast cancer tumor samples, and find 11 candidate hubs, of which five are known breast cancer related regulators. In the end, we show consistency (in terms of model selection and estimation) of the proposed procedure under suitable regularity and sparsity conditions.

The R package *space*—Sparse PArtial Correlation Estimation—is available on *http://cran.r-project.org*.

[Received March 2008. Revised November 2008.]

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. J., and Levina, E. (2008), ''Regularized Estimation of Large Covariance Matrices,'' *Annals of Statistics,* 36, 199–227.

Buhlmann, P. (2006), ''Boosting for High-Dimensional Linear Models,'' *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. J, Bartelink, H., van de Rijn M., Brown P. O., van de Vijver, M. J. (2005), ''Robustness, Scalability, and Integration of a Wound Response Gene Expression Signature in Predicting Survival of Human Breast Cancer Patients,'' *Proceedings of the National Academy of Sciences*, 8;102, 3738–3743.

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 Parameters,'' *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 (Oxford, England),* doi: 10.1093/biostatistics/kxm045.

———(2008), ''Regularization Paths for Generalized Linear Models via Coordinate Descent,'' *Technical Report*, Available at http://www-stat.stanford.edu/jhf/ftp/glmnet.pdf.

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

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

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics,* 49, 291–304.

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., Jennings, E. G., Murray, H. L., Gordon, D. B., Ren, B., Wyrick, J. J., Tagne, J. B., Volkert, T. L., Fraenkel, E., Gifford, D. K., and Young, R. A. (2002), "Transcriptional Regulatory Networks in Saccharomyces Cerevisiae," *Science,* 298, 799–804.

Levina, E., Rothman, A. J., and Zhu, J. (2008), "Sparse Estimation of Large Covariance Matrices via a Nested Lasso Penalty," *Annals of Applied Statistics,* 2, 245–263.

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

Rothman, A. J., Bickel, P. J., Levina, E., and Zhu, J. (2008), "Sparse Permutation Invariant Covariance Estimation," *Electronic Journal of Statistics,* 2, 494–515.

Schafer, J., and Strimmer, K. (2005), "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 of the United States of America,* 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., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E. T., Friend, S. H., and Bernards, R. (2002), "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer," *The New England Journal of Medicine,* 347, 1999–2009.

Whittaker, J. (1990), *Graphical Models in Applied Mathematical Multivariate Statistics,* Chichester: John Wiley & Sons Ltd.

Wu, W. B., and Pourahmadi, M. (2003), "Nonparametric Estimation of Large Covariance Matrices of Longitudinal Data," *Biometrika,* 90 (4), 831–844.

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

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