Biostatistics (2015), **16**, 4, *pp.* 670–685
doi:10.1093/biostatistics/kxv013
Advance Access publication on April 1, 2015

Statistical completion of a partially identified graph with applications for the estimation of gene regulatory networks

DONGHYEON YU

Department of Statistics, Keimyung University, Daegu, Korea

WON SON, JOHAN LIM

Department of Statistics, Seoul National University, Seoul, Korea

GUANGHUA XIAO*

Department of Clinical Sciences, University of Texas Southwestern Medical Center, TX 75390, USA
guanghua.xiao@utsouthwestern.edu

SUMMARY

We study the estimation of a Gaussian graphical model whose dependent structures are partially identified. In a Gaussian graphical model, an off-diagonal zero entry in the concentration matrix (the inverse covariance matrix) implies the conditional independence of two corresponding variables, given all other variables. A number of methods have been proposed to estimate a sparse large-scale Gaussian graphical model or, equivalently, a sparse large-scale concentration matrix. In practice, the graph structure to be estimated is often partially identified by other sources or a pre-screening. In this paper, we propose a simple modification of existing methods to take into account this information in the estimation. We show that the partially identified dependent structure reduces the error in estimating the dependent structure. We apply the proposed method to estimating the gene regulatory network from lung cancer data, where protein–protein interactions are partially identified from the human protein reference database. The application shows that proposed method identified many important cancer genes as hub genes in the constructed lung cancer network. In addition, we validated the prognostic importance of a newly identified cancer gene, PTPN13, in four independent lung cancer datasets. The results indicate that the proposed method could facilitate studying underlying lung cancer mechanisms and identifying reliable biomarkers for lung cancer prognosis.

Keywords: Concentration matrix; Gaussian graphical models; Gene regulatory network; Lung cancer; Partially identified graph; Protein–protein interaction.

1. INTRODUCTION

In recent years, statistical approaches have been developed to construct gene regulatory networks (GRNs) from mRNA expression data. A GRN describes the interactions among genes and how the genes work

*To whom correspondence should be addressed.

together to form modules of cell functions under specific contexts, such as disease status. It provides a systematic understanding of the molecular mechanisms underlying the biological processes (Friedman, 2004). In GRNs, highly connected genes are called hub genes. Because the hub genes are in key positions, their activities may affect many genes in the network and hence play an important role in biological processes. Recently, analysis of hub genes has shown to be a promising approach in identifying key disease driver genes (Akavia *and others*, 2010) and important biomarkers for predicting disease progression (Taylor *and others*, 2009; Tang *and others*, 2013). Currently, most of the existing computational methods use purely data-driven approaches to construct gene regulatory networks from gene expression data. These approaches do not rely on any prior knowledge about the network and are widely suited to many applications. However, for gene regulatory networks, information about many known connections (edges) between genes has been accumulated over decades of biological research, such as protein–protein interactions or transcriptional factor-binding sites. Using these known edges, we can turn a network construction problem into a statistical completion of a partially identified graph problem, which could lead to much better power in identifying the unknown edges. In this paper, we propose a statistical completion of a partially identified graph (SCPG) method, which is a modification of existing methods to incorporate the information about known edges. We show that the information on known edges reduces the error in identifying the unknown edges and improves the accuracy of the constructed networks.

Consider a p-dimensional random vector from a multivariate Gaussian distribution with mean 0 and covariance matrix Σ, or the concentration matrix $\Omega \equiv \Sigma^{-1} = (\sigma^{ij})_{1 \leqslant i,j \leqslant p}$. In the Gaussian graphical model, the dependent structure among p variables $X_1, X_2, \ldots, X_p$ can be expressed using a graph $G = (V, E)$, where a vertex set $V = \{i \mid i = 1, 2, \ldots, p\}$ represents p variables and an edge set $E = \{(i, j) \mid \sigma^{ij} \neq 0\}$ represents pairs of random variables that are dependent on each other, given all other variables. In other words, that σ^{ij}, the (i, j)th element of Ω, is equal to 0 implies that X_i and X_j are conditionally independent given all other variables X_k, $k \neq i, j$.

Covariance selection, first introduced by Dempster (1972), is a class of problems that estimate dependent structures among multivariate Gaussian variables by detecting non-zero elements of the concentration matrix Ω. Recently, researchers revisited the problem and studied the estimation of large-scale concentration matrices from a small number of observations. The ℓ_1-regularization on a concentration matrix or a partial correlation matrix is popularly used to obtain a sparse estimate of the dependent structure. Yuan and Lin (2007) propose the ℓ_1-regularized maximum likelihood estimator (MLE) of the concentration matrix and an algorithm to solve it using the determinant maximization problem (MAXDET), which has computational complexity of order $O(p^6)$ and becomes slower as p increases. Friedman *and others* (2007) propose a block coordinate descent procedure to solve the ℓ_1-regularized MLE, namely the graphical lasso. Meinshausen and Bühlmann (2006) formulate the covariance selection problem as a set of lasso regression problems and solve each of the lasso regression problems independently. Peng *and others* (2009) propose the sparse partial correlation estimation (SPACE) method, which solves the set of lasso regression problems jointly under the symmetry of the concentration matrix $\sigma^{ij} = \sigma^{ji}$. However, this symmetry is not guaranteed by Meinshausen and Bühlmann (2006). Recently, Cai *and others* (2011) propose the constrained ℓ_1-minimization for inverse matrix estimation (CLIME) that directly minimizes the ℓ_1-norm of the concentration matrix with a relaxed constraint for the condition $\Sigma\Omega = I$.

In this paper, we consider a simple modification of existing methods that incorporates "partially identified" dependent structures. The dependent structure to be estimated is often partially identified in practice. We also denote partially identified structures as "pre-identified" to emphasize that these structures are previously known. For example, GRNs and protein–protein interaction (PPI) networks are available in the public databases that were constructed for many previous laboratory experiments. In comparison, the pre-screening procedures recently proposed by many authors identify pairs of variables that are conditionally independent (Bair *and others*, 2006; Wasserman and Roeder, 2009). The existing methods do not

take into account these partially identified structures. They frequently estimate a known dependence as independence, or vice versa, due to a lack of data information from small samples.

The modification is done by simply redefining the existing ℓ_1-regularization. To be specific, σ^{ij} and σ^{ji} are not penalized in the objective function or its constraints if X_i and X_j are pre-identified as conditionally dependent. The modification can be applied to the ℓ_1-regularized MLE, the ℓ_1-minimization and the regression-based methods. However, in this paper, we restrict our discussion to the modification of the regression-based method, i.e., the SPACE method.

The paper is organized as follows. In Section 2, we briefly introduce the SPACE method and propose the SCPG method as a modification to incorporate pre-identified dependent structures. In Section 3, we analytically show that the SCPG method reduces the asymptotic probability of mistakenly identifying an independent pair of variables as dependent. In Section 4, we numerically investigate the gains in accuracy by estimating the network from assuming the pre-identified graph structure. In Section 5, we apply the SCPG method to estimating the gene regulatory network from lung cancer data. We conclude the paper in Section 6.

2. MODIFICATION FOR PARTIALLY IDENTIFIED GRAPH

Suppose $\mathbf{X}$ random vector with mean 0 and positive definite covariance matrix Σ. The partial correlation between X_i and X_j, denoted by ρ^{ij}, is the conditional correlation of X_i and X_j given $X_{-[i,j]} = \{X_k \mid k \neq i, j,\ 1 \leqslant k \leqslant p\}$. This partial correlation is closely related to the concentration matrix $\Omega = (\sigma^{ij})_{1 \leqslant i,j \leqslant p}$. It is known that $\rho^{ij} = -\sigma^{ij}/\sqrt{\sigma^{ii}\sigma^{jj}}$. Also, for every $i = 1, 2, \ldots, p$,

$$X_i = \sum_{j \neq i} \beta_{ij} X_j + \epsilon_i, \tag{2.1}$$

where $\beta_{ij} = -\sigma^{ij}/\sigma^{ii} = \rho^{ij}\sqrt{\sigma^{jj}/\sigma^{ii}}$, and ϵ_i is uncorrelated with $X_{-[i]} = \{X_j \mid j \neq i,\ 1 \leqslant j \leqslant p\}$ and has mean 0 and variance $1/\sigma^{ii}$.

The identity (2.1) introduces a regression-based method to estimate Σ or Ω. Let $\mathbf{X}^k = (X_1^k, X_2^k, \ldots, X_p^k)^{\mathrm{T}}$ be the kth observation of the random vector $\mathbf{X}$ for $k = 1, 2, \ldots, n$. Meinshausen and Bühlmann (2006) propose the neighborhood selection method to solve a set of lasso regression problems with respect to β_{ij}s;that is, for $i = 1, 2, \ldots, p$,

$$\min_{\beta_{ij}, j \neq i} \frac{1}{2} \sum_{k=1}^{n} \left(X_i^k - \sum_{j \neq i} \beta_{ij} X_j^k \right)^2 + \lambda \sum_{j \neq i} |\beta_{ij}|. \tag{2.2}$$

Later, Peng *and others* (2009) propose the SPACE method, which minimizes the weighted sum of p squared loss functions in (2.2) with a penalty term on the ℓ_1-norm of the partial correlation ρ^{ij}s:

$$\begin{aligned} \min_{\rho^{ij}, 1 \leqslant i < j \leqslant p} \quad & \frac{1}{2} \sum_{i=1}^{p} \left\{ w_i \sum_{k=1}^{n} \left(X_i^k - \sum_{j \neq i} \rho^{ij} \sqrt{\sigma^{jj}/\sigma^{ii}}\, X_j^k \right)^2 \right\} + \lambda \sum_{1 \leqslant i < j \leqslant p} |\rho^{ij}|, \\ \text{subject to} \quad & \rho^{ij} = \rho^{ji}, \quad 1 \leqslant i < j \leqslant p, \end{aligned} \tag{2.3}$$

where σ^{ii} is the ith diagonal element of the concentration matrix and w_i is a nonnegative weight for the ith squared loss function.

The SPACE method has several advantages over the neighborhood selection method by Meinshausen and Bühlmann (2006) in estimating the concentration matrix and the graph structure. First,

the SPACE method estimates the partial correlations and the diagonal elements of the concentration matrix. Thus, the estimation of the concentration matrix can easily be calculated by the relationship $\sigma^{ij} = -\rho^{ij}\sqrt{\sigma^{ii}\sigma^{jj}}$ in the SPACE method, while the neighborhood selection method only obtains information about whether or not each off-diagonal element of the concentration matrix is zero. Second, the estimated edges from the SPACE method are symmetrical in the sense that $\hat{\rho}^{ij} = \hat{\rho}^{ji}$; thus, if $\hat{\rho}^{ij} = 0$ (or $\hat{\rho}^{ij} \neq 0$), then $\hat{\rho}^{ji} = 0$ (or $\hat{\rho}^{ji} \neq 0$). Conversely, the neighborhood selection method separately solves p problems in (2.2) and may obtain the contradictory edges (i.e., $\hat{\beta}_{ij} \neq 0$ and $\hat{\beta}_{ji} = 0$). Finally, the SPACE method outperforms the neighborhood selection method in estimating graph structure and finding hubs in practice. This comparison study is reported in Peng *and others* (2009).

We now introduce the SCPG method as a modification of the SPACE method to take into account the pre-identified graph structure. The same modification can be applied to the ℓ_1-regularized MLEs and the ℓ_1-minimization, but those examples are omitted here. We consider the concentration matrix Ω and its induced graph $G = (V, E)$. Let $\mathcal{K}$ be a set of pre-identified edges in E. In this paper, we propose to solve

$$\begin{aligned} \min_{\rho^{ij}, 1 \leqslant i < j \leqslant p} \quad & \frac{1}{2} \sum_{i=1}^{p} \left\{ w_i \sum_{k=1}^{n} \left(X_i^k - \sum_{j \neq i} \rho^{ij} \sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}} X_k^j \right)^2 \right\} + \lambda \sum_{1 \leqslant i < j \leqslant p, (i,j) \notin \mathcal{K}} |\rho^{ij}|, \\ \text{subject to} \quad & \rho^{ij} = \rho^{ji}, \quad 1 \leqslant i < j \leqslant p. \end{aligned} \tag{2.4}$$

The modification in (2.4) only removes the penalties on the partial correlations corresponding to the pre-identified edges from (2.3). Thus, we can directly apply the active shooting algorithm, proposed by Peng *and others* (2009), to solve the modified problem. To be specific, we first rewrite the main problem (2.4) using matrix notation. The problem (2.4) without symmetry constraints becomes

$$\min_{\rho^{ij}, 1 \leqslant i < j \leqslant p} \frac{1}{2} \sum_{k=1}^{n} \sum_{i=1}^{p} \left(v_{ii} X_i^k - \sum_{j<i} \rho^{ji} v_{ij} X_j^k - \sum_{j>i} \rho^{ij} v_{ij} X_j^k \right)^2 + \lambda \sum_{1 \leqslant i < j \leqslant p, (i,j) \notin \mathcal{K}} |\rho^{ij}|,$$

where $v_{ij} = \sqrt{w_i \sigma^{jj} / \sigma^{ii}}$ for $1 \leqslant i, j \leqslant p$.

Let $\mathcal{G}_0$ denote a set of pairs such that $\rho^{ij} = 0$ (i.e., $(i, j) \in \mathcal{G}_0 \Leftrightarrow (i, j) \notin E$) and $\mathcal{G}_1$ denote a set of edges such that $(i, j) \in E$ and $(i, j) \notin \mathcal{K}$ (i.e., $\mathcal{G}_1 \equiv E \setminus \mathcal{K}$). Let $\alpha = (\alpha_1, \ldots, \alpha_{|\mathcal{G}_1|})^{\mathrm{T}}$ be a $|\mathcal{G}_1|$-dimensional vector of ρ^{ij}s for $(i, j) \in \mathcal{G}_1$; let $\gamma = (\gamma_1, \ldots, \gamma_{|\mathcal{G}_0|})^{\mathrm{T}}$ be a $|\mathcal{G}_0|$-dimensional vector of ρ^{ij}s for $(i, j) \in \mathcal{G}_0$; and let $\eta = (\eta_1, \ldots, \eta_{|\mathcal{K}|})^{\mathrm{T}}$ be a $|\mathcal{K}|$-dimensional vector of ρ^{ij}s for $(i, j) \in \mathcal{K}$. For $k = 1, 2, \ldots, n$, let

$$\mathrm{Y}^k = \begin{pmatrix} Y_1^k \\ \vdots \\ Y_p^k \end{pmatrix} = \begin{pmatrix} v_{11} X_1^k \\ \vdots \\ v_{pp} X_p^k \end{pmatrix} \quad \text{and} \quad \mathbf{Y} = \begin{pmatrix} \mathrm{Y}^1 \\ \vdots \\ \mathrm{Y}^n \end{pmatrix}.$$

We define a covariate matrix A^k of α, for Y^k, as a matrix with a size of $p \times |\mathcal{G}_1|$ and, if $\alpha_l = \rho^{ij}$, its lth column vector $\mathbf{a}^{k,l} = (a_1^{k,l}, \ldots, a_p^{k,l})^{\mathrm{T}}$ is defined as

$$a_m^{k,l} = \begin{cases} v_{ij} X_j^k & \text{if } m = i \\ v_{ji} X_i^k & \text{if } m = j \\ 0 & \text{otherwise} \end{cases}, \tag{2.5}$$

where $v_{ij} = \sqrt{w_i \sigma^{jj} / \sigma^{ii}}$ for $1 \leqslant i, j \leqslant p$.

The covariate matrices B^k and C^k with sizes of $p \times |\mathcal{G}_0|$ and $p \times |\mathcal{K}|$, respectively, are defined similarly for coefficient vectors γ and η. The whole group of covariate matrices **A**, **B,** and **C** are then defined as

$$\mathbf{A} = \begin{pmatrix} \mathrm{A}^1 \\ \vdots \\ \mathrm{A}^n \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathrm{B}^1 \\ \vdots \\ \mathrm{B}^n \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} \mathrm{C}^1 \\ \vdots \\ \mathrm{C}^n \end{pmatrix}.$$

The first part of the objective function in (2.3) is read as the least square error of the linear model

$$\mathbf{Y} = \mathbf{A}\alpha + \mathbf{B}\gamma + \mathbf{C}\eta + \mathrm{E} \equiv \tilde{\mathbf{X}}\rho + \mathrm{E}, \tag{2.6}$$

where $\rho = (\rho^{12}, \rho^{13}, \ldots, \rho^{(p-1)p})^{\mathrm{T}}$ is a $(p(p-1)/2)$-dimensional vector, $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}^{1,2}, \tilde{\mathbf{X}}^{1,3}, \ldots, \tilde{\mathbf{X}}^{(p-1),p})$ is a design matrix with a size of $np \times (p(p-1)/2)$, E is from the (np)-dimensional multivariate normal distribution with mean 0 and covariance matrix $(I_n \otimes D_p)$, where $D_p = \mathrm{diag}(1/\sigma^{11}, \ldots, 1/\sigma^{pp})$ and an operator $\otimes$ denotes the Kronecker product. Thus, we can represent the problem (2.4) as

$$\min_{\rho^{ij}, 1 \leqslant i < j \leqslant p} \frac{1}{2}\|\mathbf{Y} - \tilde{\mathbf{X}}\rho\|_2^2 + \lambda \sum_{1 \leqslant i < j \leqslant p, (i,j) \notin \mathcal{K}} |\rho^{ij}|.$$

Note that we set weights w_is to one in this paper since we do not assume any strengths for nodes. We can only assume that we have partial information about true edges. If there is information about strength or importance for specific nodes, that information can be incorporated by changing the weights.

Now we briefly describe the proposed algorithm, which depends largely on the algorithm in Peng *and others* (2009). We first set initial values $\hat{\sigma}^{ii} = 1$ for $i = 1, 2, \ldots, p$. The proposed algorithm alternately updates the estimates $\hat{\rho}^{ij}$s and $\hat{\sigma}^{ii}$s by the following steps:

- Step 1: For a given $\hat{\sigma}^{ii}$ for $i = 1, 2, \ldots, p$,

$$\hat{\rho} = \mathrm{argmin}_{\rho} \frac{1}{2}\|\mathbf{Y} - \tilde{\mathbf{X}}\rho\|_2^2 + \lambda \sum_{1 \leqslant i < j \leqslant p, (i,j) \notin \mathcal{K}} |\rho^{ij}|,$$

  where $\tilde{\mathbf{X}}$ is defined by (2.5) and (2.6) with $\sigma^{ii} = \hat{\sigma}^{ii}$ for $i = 1, 2, \ldots, p$.

- Step 2: Based on the identity (2.1), for a given $\hat{\rho}$ and $\hat{\sigma}^{ii,(\mathrm{old})}$ for $i = 1, 2, \ldots, p$,

$$\hat{\sigma}^{ii} = n \cdot \left(\sum_{k=1}^{n} \left(X_i^k - \sum_{j \neq i} \hat{\rho}^{ij} \sqrt{\frac{\hat{\sigma}^{jj,(\mathrm{old})}}{\hat{\sigma}^{ii,(\mathrm{old})}}} X_k^j \right)^2 \right)^{-1},$$

  where $\hat{\sigma}^{ii,(\mathrm{old})}$ is the estimate of σ^{ii} from the previous iteration.

- Step 3: Repeat Steps 1 and 2 until the convergence occurs.

In Step 1, we apply the modified active shooting algorithm to incorporate the pre-identified edges $\mathcal{K}$. Details on the active shooting algorithm we propose are given in Appendix A of Supplementary material available at *Biostatistics* online.

3. Asymptotic error probabilities

In this section, we analytically find the changes in asymptotic error probability by using the information on pre-identified edges. As shown in Section 2, our main problem can be rewritten as the estimation of the sparse linear model (lasso regression), and we are able to compute the asymptotic true negative/positive probabilities of the model both with and without the information on dependent pairs of variables. The computation shows that the pre-identified dependent information asymptotically increases the true negative probability (the probability of identifying independent pairs as independent) while the true positive probability (the probability of identifying dependent pairs as dependent) of both methods converge to 1. Thus, the SCPG method reduces the error probability asymptotically. Our analysis of this section relies heavily on the results of Knight and Fu (2000) and Anderson (1955), which are reviewed in Appendix B of Supplementary material available at *Biostatistics* online.

For the asymptotic true negative probability, we consider the simplified model

$$\mathbf{Y} = \tilde{\mathbf{X}}\rho + \mathrm{E} \equiv \mathbf{B}\gamma + \mathbf{C}\eta + \mathrm{E}, \tag{3.1}$$

where $\gamma = 0$ and $\eta \neq 0$. Under this model, we compare the true negative probabilities of the SCPG and SPACE methods.

Theorem 3.1 Suppose we have knowledge about $\eta \neq 0$. Let $\hat{\rho} = (\hat{\gamma}^T, \hat{\eta}^T)^{\mathrm{T}}$ and $\hat{\rho}_{\mathcal{K}} = (\hat{\gamma}_{\mathcal{K}}^{\mathrm{T}}, \hat{\eta}_{\mathcal{K}}^T)^{\mathrm{T}}$ be the solutions of the SPACE and SCPG methods, respectively. Then, the asymptotic true negative probabilities $\mathrm{P}(\hat{\gamma} = 0)$ and $\mathrm{P}(\hat{\gamma}_{\mathcal{K}} = 0)$ satisfy the inequality

$$\mathrm{P}(\hat{\gamma} = 0) \leqslant \mathrm{P}(\hat{\gamma}_{\mathcal{K}} = 0).$$

Proof. See Appendix C.1 of Supplementary material available at *Biostatistics* online. □

We next compare the asymptotic true positive probabilities of the SCPG and SPACE methods. Here, we consider the simplified model

$$\mathbf{Y} = \mathbf{X}\rho + \mathrm{E} \equiv \mathbf{A}\alpha + \mathbf{C}\eta + \mathrm{E},$$

where both $\alpha \neq 0$ and $\eta \neq 0$. Suppose we have knowledge on $\eta \neq 0$. Let $\hat{\rho} = (\hat{\alpha}^T, \hat{\eta}^T)^{\mathrm{T}}$ and $\hat{\rho}_{\mathcal{K}} = (\hat{\alpha}_{\mathcal{K}}^{\mathrm{T}}, \hat{\eta}_{\mathcal{K}}^T)^{\mathrm{T}}$ be the solutions of the SPACE and SCPG methods, respectively. This section shows the following theorem:

Theorem 3.2 The asymptotic true positive probabilities of both the SPACE and SCPG models (which are $\mathrm{P}(\hat{\alpha} \neq 0)$ and $\mathrm{P}(\hat{\alpha}_{\mathcal{K}} \neq 0)$, respectively) converge to one as $n \to \infty$.

Proof. See Appendix C.2 of Supplementary material available at *Biostatistics* online. □

In comparing the asymptotic error probabilities between the SPACE and SCPG models, we show that the SCPG model asymptotically improves the true negative probability with the same performance as the true positive probability. Note that there is a difference between the asymptotic biases for $\hat{\alpha}$ and $\hat{\alpha}_{\mathcal{K}}$. However, a direct comparison of these asymptotic biases is difficult since the difference varies with the signs and structures of partial correlations in the model.

4. Numerical study

In the previous section, we showed that the SCPG method improves the asymptotic true negative probability and has the same performance for the asymptotic true positive probability in estimating graph structure

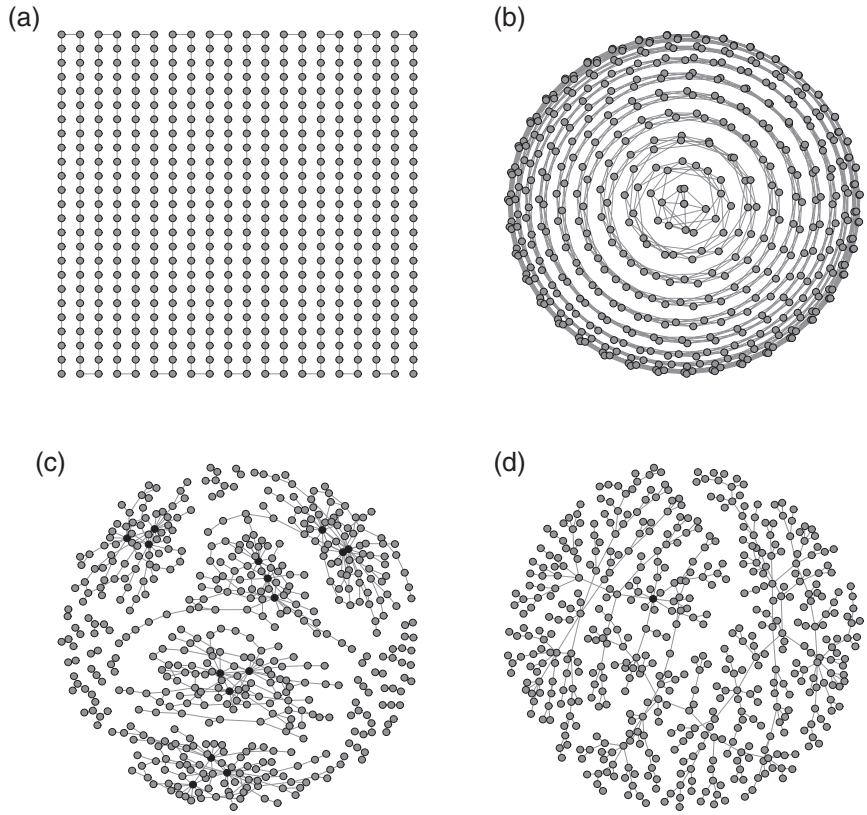

Fig. 1. Graphs of four networks used in simulation. Black nodes (•) denote nodes whose degrees are >9. (a) **(N1)** AR(1), (b) **(N2)** AR(2), (c) **(N3)** hub network, and (d) **(N4)** scale-free.

for a given λ_0. This is mainly due to the bias reduction by using the prior information about partially identified edges. In practice, however, we generally encounter datasets with finite samples and choose a tuning parameter that minimizes an information criterion, such as the Bayesian information criterion (BIC) (Schwarz, 1978). Thus, we additionally investigate the performance of the SCPG method with finite samples for several graph structures and also compare the SCPG method with the SPACE method to confirm the improvements of the SCPG method in estimating a graph structure.

We first consider the Gaussian graphical model accompanied by the following AR(1), AR(2), hub and scale-free networks in simulation. The AR(1) and the AR(2) networks are from the time series model and the hub and scale-free networks reflect real biological networks. These four networks are illustrated in Figure 1. The details of the four networks, including how they are generated, are given in Appendix D of Supplementary material available at *Biostatistics* online.

We consider moderate-sized networks with 500 nodes and sample sizes of 100, 250, and 500. To apply the SCPG method, we define two pre-identified edge sets $\mathcal{K}_{0.1}$ and $\mathcal{K}_{0.3}$ by randomly selecting 10 and 30% of the true edges, respectively. These two pre-identified edge sets are also used to find the effects of the amount of information on estimating a graph structure. In each network, we generate 50 datasets from a Gaussian distribution with mean 0 and covariance matrix Σ defined with the (i, j)th element $\sigma_{ij} = (\Omega^{-1})_{ij}/\sqrt{(\Omega^{-1})_{ii}(\Omega^{-1})_{jj}}$. Note that, for the hub and scale-free networks, we make the network have five exclusive sub-networks, each of which has 100 nodes; the nodes in one sub-network are not connected to

those in other sub-networks. This procedure is applied in Peng *and others* (2009) to describe the module-based networks frequently observed in real networks.

Let $\rho = (\rho^{ij})_{1 \leqslant i < j \leqslant p}$ and $\hat{\rho}_\lambda = (\hat{\rho}_\lambda^{ij})_{1 \leqslant i < j \leqslant p}$ be a $(p(p-1)/2)$-dimensional vector of the true partial correlations and the estimates of partial correlations at λ, respectively. To investigate the theoretical properties of the SCPG method with finite samples, we introduce the true positive rate (TPR) and the true negative rate (TNR) defined as follows:

$$\mathrm{TPR}(\hat{\rho}_\lambda, \rho) \equiv \frac{\mathrm{TP}(\hat{\rho}_\lambda, \rho)}{\mathrm{P}(\rho)} \quad \text{and} \quad \mathrm{TNR}(\hat{\rho}_\lambda, \rho) \equiv \frac{\mathrm{TN}((\hat{\rho}_\lambda, \rho)}{\mathrm{N}(\rho)},$$

where $\mathrm{TP}(\hat{\rho}_\lambda, \rho) = \sum_{i<j} I(\rho^{ij} \neq 0) I(\hat{\rho}_\lambda^{ij} \neq 0)$, $\mathrm{P}(\rho) = \sum_{i<j} I(\rho^{ij} \neq 0)$, $\mathrm{TN}(\hat{\rho}_\lambda, \rho) = \sum_{i<j} I(\rho^{ij} = 0) I(\hat{\rho}_\lambda^{ij} = 0)$, $\mathrm{N}(\rho) = \sum_{i<j} I(\rho^{ij} = 0)$, and $I(\cdot)$ denotes an indicator function. We additionally define the false discovery rate (FDR) as

$$\mathrm{FDR}(\hat{\rho}_\lambda, \rho) \equiv \frac{\mathrm{FP}(\hat{\rho}_\lambda, \rho)}{\mathrm{P}(\hat{\rho})} = \frac{\mathrm{FP}(\hat{\rho}_\lambda, \rho)}{\mathrm{TP}(\hat{\rho}_\lambda, \rho) + \mathrm{FP}(\hat{\rho}_\lambda, \rho)},$$

where $\mathrm{FP}(\hat{\rho}_\lambda, \rho) = \sum_{i<j} I(\rho^{ij} = 0) I(\hat{\rho}_\lambda^{ij} \neq 0)$ and $\mathrm{P}(\hat{\rho}) = \sum_{i<j} I(\hat{\rho}_\lambda^{ij} \neq 0)$. Note that the FDR is not defined if $\sum_{i<j} I(\hat{\rho}_\lambda^{ij} \neq 0) = 0$. In this case, we consider the FDR value to be 0 to summarize results with all datasets.

Figure 2 plots the average of TPRs, TNRs, and FDRs for various λs in the aforementioned four networks and shows several interesting features containing the result that are related to the theoretical properties in the previous section. Compared with SPACE, the SCPG method improves TPRs in all networks considered except the AR(1) network for a given λ.

In view of TNRs, however, the SCPG method improves on the SPACE method for any given λs in all networks we consider and also increases TNRs as the amount of pre-identified information increases. This result shows that the theoretical result for the true negative probability described in the previous section still holds with finite samples. Moreover, the SCPG method decreases FDRs for any given λs in all networks compared with the SPACE method. Interestingly, the SCPG method decreases FDRs while the TPRs decrease as the amount of pre-identified information increases in the AR(1) network.

The tuning parameter λ in both the SCPG and SPACE methods plays an important role in estimating the network, where a large (or a small) value of λ results in a sparse (or a dense) estimate of the network with low false positives (or low false negatives). Several information criteria, such as the Akaike information criterion and Bayesian information criterion (BIC), are heuristically used for the network model of these papers (Danaher *and others*, 2014; Yuan and Lin, 2007; Peng *and others*, 2009). They are originally designed for the linear regression model and some of them are theoretically shown to select the correct model (Wang *and others*, 2009; Fan and Tang, 2013). However, these are limited to the linear regression model, and there is no optimal rule for choosing λ in the network model. In this paper, we adopt the generalized information criterion (GIC) proposed by Fan and Tang (2013), which is shown to outperform the BIC in identifying the correct model in the linear regression. The "GIC-type" criterion used in this paper is defined like the "BIC type" criterion in Peng *and others* (2009) as

$$\mathrm{GIC}(\lambda) = \sum_{k=1}^{p} n \log \left(\sum_{i=1}^{n} \left(X_k^i - \sum_{j \neq k} \rho_\lambda^{jk} \sqrt{\frac{\sigma_\lambda^{jj}}{\sigma_\lambda^{kk}}} X_j^i \right)^2 \right) + \log(\log n) \log(p-1) \sum_{k=1}^{p} \mathrm{df}_k(\hat{\rho}),$$

where $\mathrm{df}_k(\hat{\rho}) = |\{j \mid \hat{\rho}_\lambda^{jk} \neq 0,\ j \neq k\}|$ and $|A|$ is a cardinality of a set A. For each dataset, we evaluate the $\mathrm{GIC}(\lambda)$ on a grid of $(20, \lambda^{\max})$ and choose a tuning parameter λ^* such that $\lambda^* = \mathrm{argmin}_{\lambda \in (20, \lambda^{\max})} \mathrm{GIC}(\lambda)$,

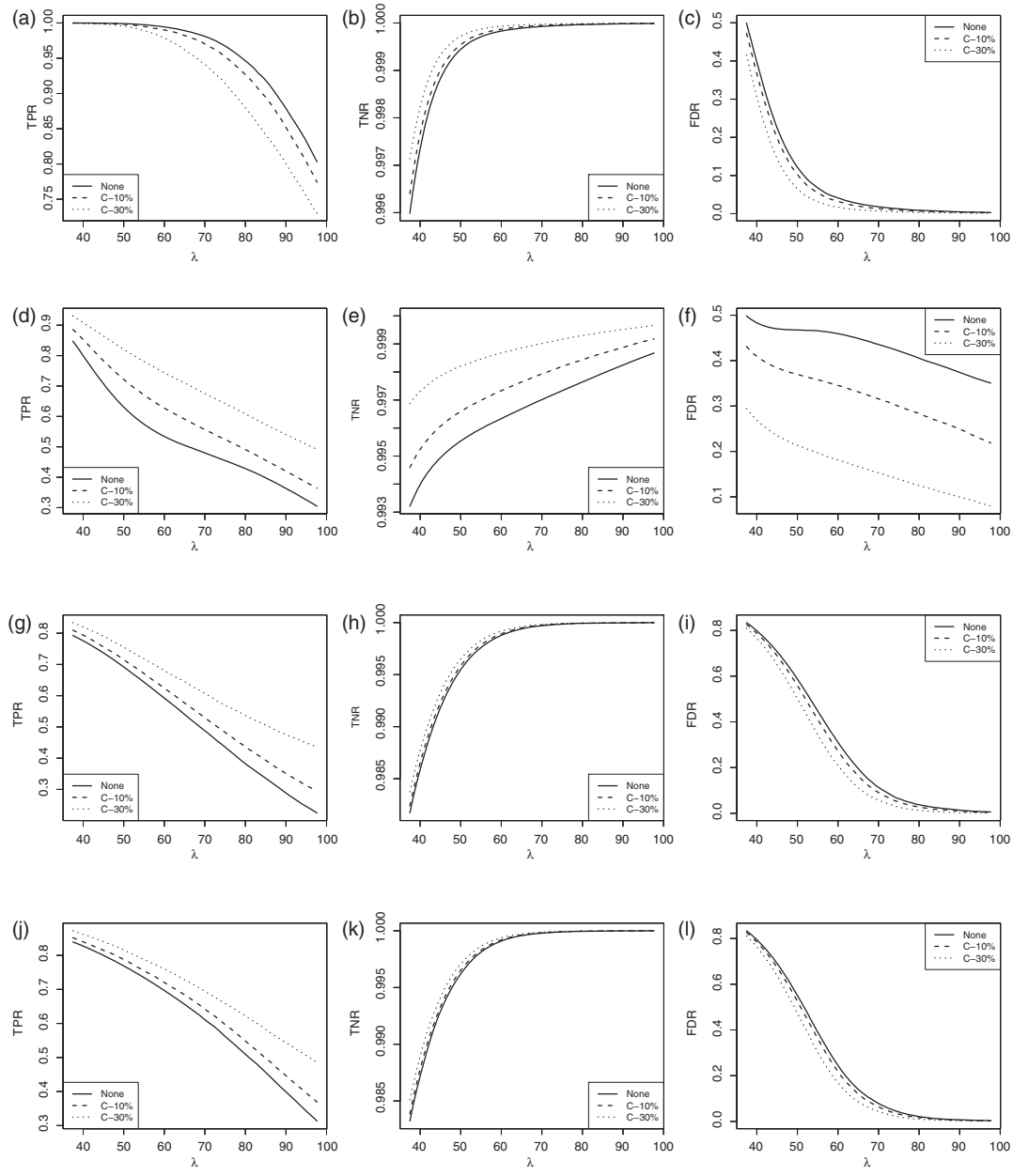

Fig. 2. The averages of TPR, TNR, and FDR for **(N1)**–**(N4)** networks in Figure 1 with $p = 500$ and $n = 100$. "None", "C-10%", and "C-30%" denote the SPACE method (solid), the SCPG method with 10% (dashed) and 30% (dotted) partially identified edges, respectively. (a) **(N1)** TPR($\hat{\rho}_\lambda, \rho$), (b) **(N1)** TNR($\hat{\rho}_\lambda, \rho$), (c) **(N1)** FDR($\hat{\rho}_\lambda, \rho$), (d) **(N2)** TPR($\hat{\rho}_\lambda, \rho$), (e) **(N2)** TNR($\hat{\rho}_\lambda, \rho$), (f) **(N2)** FDR($\hat{\rho}_\lambda, \rho$), (g) **(N3)** TPR($\hat{\rho}_\lambda, \rho$), (h) **(N3)** TNR($\hat{\rho}_\lambda, \rho$), (i) **(N3)** FDR($\hat{\rho}_\lambda, \rho$), (j) **(N4)** TPR($\hat{\rho}_\lambda, \rho$), (k) **(N4)** TNR($\hat{\rho}_\lambda, \rho$), and (l) **(N4)** FDR($\hat{\rho}_\lambda, \rho$).

where $\lambda^{\max} = \inf_\lambda\{\lambda \mid \hat{\rho}_\lambda^{ij} = 0 \text{ for } 1 \leqslant i < j \leqslant p\}$. In addition, the selected models of the SPACE and SCPG methods by the GIC are evaluated by the TPR, TNR, FDR, the mis-specification rate (MISR), and the Matthews correlation coefficient (MCC). The first three measures have been defined already. Here, we introduce the MISR and MCC, defined as

$$\text{MISR}(\hat{\rho}, \rho) = \frac{\text{FN} + \text{FP}}{p(p-1)/2} \quad \text{and} \quad \text{MCC}(\hat{\rho}, \rho) = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

where $\text{TP} = \sum_{i<j} I(\rho^{ij} \neq 0) I(\hat{\rho}_\lambda^{ij} \neq 0)$, $\text{FP} = \sum_{i<j} I(\rho^{ij} = 0) I(\hat{\rho}_\lambda^{ij} \neq 0)$, $\text{TN} = \sum_{i<j} I(\rho^{ij} = 0) I(\hat{\rho}_\lambda^{ij} = 0)$, and $\text{FN} = \sum_{i<j} I(\rho^{ij} \neq 0) I(\hat{\rho}_\lambda^{ij} = 0)$. Here, the MISR corresponds to the total error rate of a classifier and the MCC, with a value between -1 and 1, measures the accuracy of a classifier, where $+1$, 0, and -1, respectively, denote a perfect classification, a random classification, and a total discordance of classification.

Tables 1 and 2 report the average of these five measures over 50 datasets, which reveals some interesting features of the proposed SCPG method. First, compared with the SPACE method, the TNRs and the MCCs of the SCPG method increase as the amount of pre-identified information increases, for all the cases we consider. Second, the SCPG method has smaller error rates than the SPACE method in terms of the FDRs and the MISRs. Finally, the TPRs of the SCPG method are approximately equal to or higher than those of the SPACE method (without pre-identified information) in all the cases we consider except the AR(1) model. In summary, these features indicate that the SCPG method's performance is superior to the SPACE method in all aspects.

Before we end this section, we implement three additional numerical studies. First, we compare the performances of the SCPG and naive methods in estimating the structure of the network. Here, the naive method implies the direct addition of pre-identified edges to the estimated network by the SPACE. The results show that the SCPG method outperforms the naive method in all cases considered. The details of this comparison are detailed in Appendix H of Supplementary material available at *Biostatistics* online and the results are summarized in Table H.1 of Supplementary material available at *Biostatistics* online. It indicates that incorporating information of pre-identified edges help the estimation of network structure. In the second study, we investigate how the SCPG method is sensitive to the misspecification rates (the ratio of false positives in the pre-identified edges). Both details of the second study and results are reported in Appendix I of of Supplementary material available at *Biostatistics* online. The results show that the SCPG method still works better than the SPACE method in terms of error rates unless the misspecification rate is low (not $>15\%$ in the study). However, we recommend readers choose the pre-identified edges in a conservative way. Finally, to understand how the SCPG performs in a large-scale network, we repeat the same numerical study as above for the hub and scale-free networks with 1000 nodes; these two networks are the most common assumptions for a large-scale network. The results are similar to what we had in Table 2. They are reported in Table J.1 in Appendix J of of Supplementary material available at *Biostatistics* online.

5. Applications with lung cancer adenocarcinoma

Two recent studies have shown that the hub genes in lung cancer gene regulatory networks may be potential robust biomarkers for lung cancer progression. To study whether our proposed method could discover novel gene biomarkers for cancer progression, we applied the proposed method to construct a network based on a microarray dataset from the Lung Cancer Consortium dataset (Shedden *and others*, 2008). This dataset measures the gene levels in 442 lung cancer adenocarcinoma patients. We identified 794 genes whose expression levels are significantly associated with patients' survival time, after adjusting for clinical variables based on a univariate Cox regression (See Appendix E of of Supplementary material

Table 1. *The averages of* $|\hat{E}|$*, TPR, TNR, FDR, MISR, and MCC for AR*(1) *and AR*(2) *networks over* 50 *datasets*

Network	n	Info.	$\lvert\hat{E}\rvert$	TPR	TNR	FDR	MISR	MCC
AR(1) ($\lvert E\rvert = 499$)	100	None	627.64	99.97	99.9	20.45	0.1	89.12
			(2.55)	(0.01)	(0)	(0.32)	(0)	(0.18)
		10%	613.7	99.92	99.91	18.7	0.09	90.08
			(2.22)	(0.02)	(0)	(0.29)	(0)	(0.16)
		30%	584.38	99.82	99.93	14.68	0.07	92.24
			(2.51)	(0.03)	(0)	(0.36)	(0)	(0.19)
	250	None	609.58	100	99.91	18.09	0.09	90.46
			(2.24)	(0)	(0)	(0.3)	(0)	(0.17)
		10%	596.62	100	99.92	16.3	0.08	91.44
			(2.4)	(0)	(0)	(0.33)	(0)	(0.18)
		30%	570.74	100	99.94	12.52	0.06	93.5
			(1.87)	(0)	(0)	(0.28)	(0)	(0.15)
	500	None	600.84	100	99.92	16.9	0.08	91.12
			(2.18)	(0)	(0)	(0.3)	(0)	(0.17)
		10%	584.68	100	99.93	14.61	0.07	92.37
			(1.83)	(0)	(0)	(0.26)	(0)	(0.14)
		30%	566.2	100	99.95	11.83	0.05	93.87
			(1.62)	(0)	(0)	(0.25)	(0)	(0.13)
AR(2) ($\lvert E\rvert = 997$)	100	None	1431.8	75.21	99.45	47.59	0.74	62.4
			(14.03)	(0.63)	(0.01)	(0.14)	(0)	(0.24)
		10%	1509.6	87.28	99.48	42.29	0.61	70.68
			(10.81)	(0.37)	(0.01)	(0.2)	(0)	(0.11)
		30%	1403.8	94.87	99.63	32.56	0.41	79.8
			(7.23)	(0.19)	(0)	(0.25)	(0)	(0.12)
	250	None	1873.38	100	99.29	46.75	0.7	72.71
			(6.57)	(0)	(0.01)	(0.19)	(0.01)	(0.13)
		10%	1724.14	100	99.41	42.14	0.58	75.84
			(6.22)	(0)	(0.01)	(0.21)	(0)	(0.14)
		30%	1455.84	100	99.63	31.49	0.37	82.61
			(4)	(0)	(0)	(0.19)	(0)	(0.11)
	500	None	1801.44	100	99.35	44.61	0.64	74.17
			(7.39)	(0)	(0.01)	(0.23)	(0.01)	(0.15)
		10%	1665.48	100	99.46	40.09	0.54	77.18
			(6.52)	(0)	(0.01)	(0.23)	(0.01)	(0.15)
		30%	1418.54	100	99.66	29.69	0.34	83.7
			(3.67)	(0)	(0)	(0.18)	(0)	(0.11)

$|\hat{E}|$ denotes the number of estimated edges. All values except for $|\hat{E}|$ are multiplied by 100. The numbers in parentheses denote the standard errors of measures.

available at *Biostatistics* online). In addition, we used a list of PPIs from the human protein reference database (HPRD), which provided 39 240 pairs of PPIs for 9617 genes. Only 222 pairs of PPIs for 211 genes were matched to 794 genes in the lung cancer dataset. We used these 222 pairs as the pre-identified information.

Table 2. *The averages of* $|\hat{E}|$, *TPR*, *TNR*, *FDR*, *MISR*, *and MCC for hub and scale-free networks over* 50 *datasets*

Network	n	Info.	$\lvert\hat{E}\rvert$	TPR	TNR	FDR	MISR	MCC
Hub ($\lvert E\rvert = 569$)	100	None	318	48.94	99.97	12.15	0.26	65.38
			(4.5)	(0.5)	(0)	(0.49)	(0)	(0.26)
		10%	343.1	53.79	99.97	10.56	0.24	69.2
			(4.17)	(0.44)	(0)	(0.44)	(0)	(0.21)
		30%	363.4	60.18	99.98	5.61	0.2	75.24
			(3.69)	(0.42)	(0)	(0.35)	(0)	(0.18)
	250	None	586.38	87.05	99.93	15.43	0.13	85.71
			(3.34)	(0.18)	(0)	(0.37)	(0)	(0.17)
		10%	574.96	87.45	99.94	13.39	0.12	86.96
			(2.74)	(0.18)	(0)	(0.31)	(0)	(0.14)
		30%	564.5	88.56	99.95	10.68	0.1	88.88
			(2.54)	(0.17)	(0)	(0.29)	(0)	(0.13)
	500	None	654.32	97.01	99.92	15.57	0.1	90.44
			(2.78)	(0.09)	(0)	(0.33)	(0)	(0.17)
		10%	644.34	96.99	99.93	14.31	0.09	91.12
			(2.29)	(0.08)	(0)	(0.28)	(0)	(0.14)
		30%	619.42	96.92	99.95	10.92	0.07	92.88
			(2.21)	(0.08)	(0)	(0.29)	(0)	(0.15)
Scale-free ($\lvert E\rvert = 495$)	100	None	396.6	66.73	99.95	16.58	0.19	74.49
			(3.02)	(0.26)	(0)	(0.42)	(0)	(0.18)
		10%	399.04	69.17	99.95	14.04	0.17	77
			(3.11)	(0.24)	(0)	(0.45)	(0)	(0.16)
		30%	402.3	73.28	99.97	9.72	0.14	81.25
			(2.74)	(0.24)	(0)	(0.38)	(0)	(0.15)
	250	None	526.1	89.28	99.93	15.87	0.11	86.59
			(3.3)	(0.18)	(0)	(0.42)	(0)	(0.2)
		10%	518.12	89.57	99.94	14.31	0.1	87.54
			(3.12)	(0.17)	(0)	(0.41)	(0)	(0.19)
		30%	500.22	90.47	99.96	10.37	0.08	89.99
			(2.9)	(0.18)	(0)	(0.39)	(0)	(0.17)
	500	None	561.52	96.79	99.93	14.61	0.08	90.86
			(2.26)	(0.09)	(0)	(0.34)	(0)	(0.19)
		10%	551.34	96.63	99.94	13.18	0.07	91.55
			(2.2)	(0.09)	(0)	(0.34)	(0)	(0.19)
		30%	536.6	97.14	99.96	10.33	0.06	93.29
			(2.09)	(0.07)	(0)	(0.33)	(0)	(0.17)

$|\hat{E}|$ denotes the number of estimated edges. All values except for $|\hat{E}|$ are multiplied by 100. The numbers in parentheses denote the standard errors of measures.

In this study, we compared performances in constructing the gene regulatory network using (i) the SPACE method and (ii) the proposed SCPG method, with λ determined by the GIC. An overview of the networks constructed using the SPACE method and the proposed method is shown in Figure 3. The SPACE method estimated 297 edges for 135 genes of 794 genes (659 genes had no connection). The SCPG method estimated 455 edges for 299 genes (495 genes had no connection). To identify hub genes in the estimated

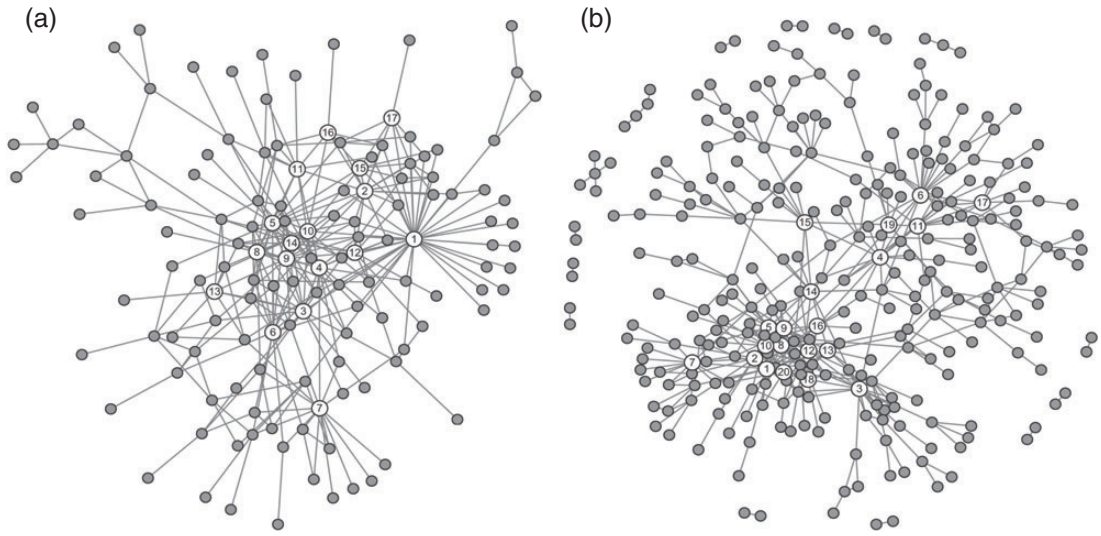

Fig. 3. Estimated graph structures for the SPACE and SCPG methods. The nodes with numbers denote the detected hub genes reported in Table 3. (a) SPACE and (b) SCPG with PPI networks.

graph, we applied a procedure similar to that described in Peng *and others* (2009). From the estimated graph structures, we first selected genes whose degrees lie over 0.95 quantiles of degree distribution. Then, we calculated the ranks of degrees of selected genes for various λ values. We selected potential hub genes such that the averages of the ranks of degrees were $\leqslant 20$, and the standard deviations were $\leqslant 2$. Following this procedure, we identified 17 hub genes from the SPACE method and 20 hub genes from the proposed method by incorporating the PPI network information. The identified hub genes are summarized in Table 3. There were 11 genes (highlighted in bold in Table 3) identified by both approaches, among which there were several key lung cancer genes, NKX2-1, HOP, and SFTPB (Further information is given in Appendix F of Supplementary material available at *Biostatistics* online). In comparing the two methods, we noted that the SCPG method identified nine genes that were missed by the SPACE method, including CTNNB1, CSNK2A1, ESR1, NEDD9, FYN, BRCA1, PTPN13, PIK3R1, and SLC34A2. Seven of these nine genes (identified only by SCPG) had been reported to play important roles in lung cancer, while two (UBE2C and TYMS) of six genes identified only by SPACE method are, based on our literature search, associated with lung cancer. (Further details are given in Appendix G of Supplementary material available at *Biostatistics* online.)

In addition, the SCPG method identified the PTPN13 gene, which had not been previously reported as a lung cancer related gene. To further study this gene, we have downloaded the mRNA expression together with the clinical annotation from four public lung cancer datasets, including (1) Tomida *and others* (2009) ($n = 117$), (2) Bhattacharjee *and others* (2001) ($n = 203$), (3) Raponi *and others* (2006) ($n = 129$), and (4) Jones *and others* (2004) ($n = 80$). These four datasets were selected because they were published in high-profile journals, contained relatively large sample sizes (at least 80 samples), and were measured from different microarray platforms. Interestingly, the under-expression of the PTPN13 gene is consistently associated with the poor prognosis of lung cancer patients in the four independent datasets, which were measured using different platforms (see Fig. G.1 of Supplementary material available at *Biostatistics* online). The results show that the mRNA expression of the PTPN13 gene is a novel and robust prognostic biomarker of potential clinical importance.

Table 3. *List of potential hub genes that identified by the SPACE and SCPG methods*

SPACE				SCPG			
No.	Gene symbol	Degree	CR	No.	Gene symbol	Degree	CR
1	PRC1	39		1	**GPR116**	18	
2	**RRM2**	18		2	**NKX2-1**	18	•
3	**CYP2B7P1**	17		3	**RRM2**	18	
4	**GPR116**	17		4	CTNNB1	17	•
5	**SFTPB**	17	•	5	**CYP2B7P1**	17	
6	**NKX2-1**	16	•	6	CSNK2A1	16	•
7	**TFF1**	16		7	**TFF1**	15	
8	**HOP**	15	•	8	**C1orf116**	14	
9	**C1orf116**	14		9	**HOP**	14	•
10	**FMO5**	14		10	**SFTPB**	14	•
11	**CD302**	12		11	ESR1	13	•
12	**HSD17B6**	12		12	**FMO5**	12	
13	HOXD1	9		13	**CD302**	11	
14	TMPRSS2	9		14	NEDD9	11	•
15	TPX2	9		15	FYN	10	•
16	UBE2C	8	•	16	PTPN13	10	
17	TYMS	7	•	17	BRCA1	9	•
				18	**HSD17B6**	9	
				19	PIK3R1	9	•
				20	SLC34A2	9	

Bold font highlights the genes identified by both methods. "CR" denotes cancer-related genes identified by previous studies.

6. Conclusion

Recently, reconstructions of GRNs based on genome-wide mRNA expression data have been widely used to study biological mechanisms and identify novel biomarkers. Learning the gene network structures from gene expression data is a challenge because of the extremely large number of possible network edges and the small number of sample sizes in gene expression data to infer the true edges. However, for GRN, there are many previously identified edges (i.e., gene regulations) from pathway information, protein–protein interaction databases, and transcriptional factor binding databases. So instead of learning the structure of GRN from scratch, we can incorporate the known edges to mitigate the daunting task of network reconstruction. In this study, we proposed the SCPG method, a simple but effective modification of the SPACE method, to incorporate partially identified edges in estimating graph structure with a Gaussian graphical model. The SCPG method asymptotically increases the true negative probability and obtains the same performance in terms of the true positive probability compared with the SPACE method. Moreover, we numerically show that the SCPG method not only increases the true negative rate but also reduces the false discovery rate. The SCPG method was applied here to estimate the gene regulatory network of lung cancer data with pre-identified edges from the HPRD database, and it identified more cancer-related hub genes than the SPACE method. More importantly, the SCPG method identified a novel prognostic biomarker, the PTPN13 gene. We validated the prognostic performance of PTPN13 gene expression using four independent lung cancer mRNA expression datasets across different experimental platforms. The results indicate that the proposed SCPG method performs well in reconstructing a gene regulatory network and could be used to identify novel biomarkers for predicting disease outcomes.

In this study, we demonstrated that inferring gene network structures can be improved by incorporating information about previously identified edges from other resources. However, we need to be cautious because gene regulation could vary among different tissues or biological conditions, while most information available about previously identified edges (gene–gene interactions) is not condition specific. As a result, some edges reported in existing databases may not really be edges in the specific conditions under study, which may lead to false-positive edges. A reasonable way to avoid this is to select only the reported edges with high expression correlations for the corresponding gene pairs in the expression data to be used for constructing the network (Ahn *and others*, 2011). This step helps to identify the gene-gene interactions that are appropriate for the specific conditions under study. In addition, we used GIC to select the tuning parameter, which produced satisfactory results in the real data application. However, it is possible that there exist other examples where the GIC performs poorly. It is also possible that there are other methods for selecting the tuning parameter that could be superior to the GIC. In summary, methodology for objectively selecting tuning parameters is an interesting area for future research.

SUPPLEMENTARY MATERIAL

Supplementary material is available online at http://biostatistics.oxfordjournals.org.

FUNDING

This work was supported by the National Institutes of Health (R01CA172211 to Guanghua Xiao) and National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2011-0029104 to Johan Lim).

REFERENCES

AHN, J., YOON, Y., PARK, C., SHIN, E. AND PARK, S. (2011). Integrative gene network construction for predicting a set of complementary prostate cancer genes. *Bioinformatics* **27**(13), 1846–1853.

AKAVIA, U. D., LITVIN, O., KIM, J., SANCHEZ-GARCIA, F., KOTLIAR, D., CAUSTON, H. C., POCHANARD, P., MOZES, E. *and others* (2010). An integrated approach to uncover drivers of cancer. *Cell* **143**(6), 1005–1017.

ANDERSON, T. W. (1955). The integral of a symmetric unimodal function over a symmetric convex set and some probability inequalities. *Proceedings of the American Mathematical Society* **6**(2), 170–176.

BAIR, E., HASTIE, T., PAUL, D. AND TIBSHIRANI, R. (2006). Prediction by supervised principal components. *Journal of the American Statistical Association* **101**, 119–137.

BHATTACHARJEE, A., RICHARDS, W. G., STAUNTON, J., LI, C., MONTI, S., VASA, P., LADD, C. AND BEHESHTI, J. *and others* (2001). Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of the National Academy of Sciences of the United States of America* **98**(24), 13790–13795.

CAI, T., LIU, W. D. AND LUO, X. (2011). A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association* **106**(494), 594–607.

DANAHER, P., WANG, P. AND WITTEN, D. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B* **76**(2), 373–397.

DEMPSTER, A. (1972). Covariance selection. *Biometrics* **28**, 157–175.

FAN, Y. AND TANG, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society: Series B* **75**(3), 531–552.

Friedman, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **303**(5659), 799–805.

Friedman, J., Hastie, T. and Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9**(3), 432–441.

Jones, M. H., Virtanen, C., Honjoh, D., Miyoshi, T., Satoh, Y., Okumura, S., Nakagawa, K., Nomura, H. and Ishikawa, Y. (2004). Two prognostically significant subtypes of high-grade lung neuroendocrine tumours independent of small-cell and large-cell neuroendocrine carcinomas identified by gene expression profiles. *The Lancet* **363**(9411), 775–781.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics* **28**(5), 1356–1378.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graph and variable selection with the lasso. *Annals of Statistics* **34**(3), 1436–1462.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association* **104**, 735–746.

Raponi, M., Zhang, Y., Yu, J., Chen, G., Lee, G., Taylor, J. M., Macdonald, J. and Thomas, D. *and others* (2006). Gene expression signatures for predicting prognosis of squamous cell and adenocarcinomas of the lung. *Cancer Research* **66**(15), 7466–7472.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* **6**(2), 461–464.

Shedden, K., Taylor, J. M., Enkemann, S. A., Tsao, M. S., Yeatman, T. J., Gerald, W. L., Eschrich, S. and Jurisica, I. *and others* (2008). Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine* **14**(8), 822–827.

Tang, H., Xiao, G., Behrens, C., Schiller, J., Allen, J., Chow, C. W., Suraokar, M. and Corvalan, A. *and others* (2013). A 12-gene set predicts survival benefits from adjuvant chemotherapy in non-small cell lung cancer patients. *Clinical Cancer Research* **19**(6), 1577–1586.

Taylor, I. W., Linding, R., Warde-Farley, D., Liu, Y., Pesquita, C., Faria, D., Bull, S. and Pawson, T. *and others* (2009). Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology* **27**(2), 199–204.

Tomida, S., Takeuchi, T., Shimada, Y., Arima, C., Matsuo, K., Mitsudomi, T., Yatabe, Y. and Takahashi, T. (2009). Relapse-related molecular signature in lung adenocarcinomas identifies patients with dismal prognosis. *Journal of Clinical Oncology* **27**(17), 2793–2799.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society Series B* **75**(3), 531–552.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics* **37**(5), 2178–2201.

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

[*Received April 29, 2014; revised February 26, 2015; accepted for publication March 3, 2015*]