Statistics and Its Interface Volume 10 (2017) 657–676

Estimation of directed subnetworks in ultra high dimensional data for gene network problems

Sung Won Han*, SungHwan Kim*, Junhee Seok,
Jeewhan Yoon†, and Hua Zhong†

The next generation sequencing technology generates ultra high dimensional data. However, it is computationally impractical to estimate an entire Directed Acyclic Graph (DAG) under such high dimensionality. In this paper, we discuss two different types of problems to estimate subnetworks in ultra high dimensional data. The first problem is to estimate DAGs of a subnetwork adjacent to a target gene, and the second problem is to estimate DAGs of multiple subnetworks without information about a target gene. To address each problem, we propose efficient methods to estimate subnetworks by using layer-dependent weights with BIC criteria or by using community detection approaches to identify clusters as subnetworks. We apply such approaches to the gene expression data of breast cancer in TCGA as a practical example.

Keywords and phrases: Bayesian network, Directed acyclic graph, Penalized likelihood, High dimension, Subnetworks.

1. INTRODUCTION

Pathway analysis and gene-gene interaction studies play important roles to reveal the underlying molecular mechanisms associated with cancer development. The directed acyclic graph (DAG) is a commonly used model to estimate a gene regulatory network. Since the cost of generating high-throughput data has been reduced, ultra high dimensional data sets are abundantly available [64]. However, the estimation of DAGs is an NP-hard problem, so it requires heavy computational time even for middle-sized data. Not surprisingly, it is challenging to estimate a whole gene network using large-scale gene expression data. For example, in large omics data for cancer such as the TCGA (The Cancer Genome Atlas) or the GTEx (Genotype-Tissue Expression) project, the number of gene expressions is huge, more than 10,000. However, there is no existing method that can estimate an entire network within reasonable computational time.

A directed subnetwork that contains a few clusters with significant genomic features is sufficient to account for the system of biological components related to diseases without estimating an entire network. Thus, we estimate a network focusing on only a few gene expressions that serve important roles in the model. According to Gene Ontology [25, 2], protein coding-genes characterized with the same genomic function leading to a certain disease are often mutually associated and biologically function together. In addition, the function of a gene can be revealed within a group of genes having known genetic functions [16], and thus identifying directed interactions within a group of similar functional genes is a reasonable approach. In this paper, we propose a method to estimate directed subnetworks in ultra high dimensional data.

Probabilistic graphical models have received a wide attention in terms of estimating networks. Conditional dependency among variables can be represented by undirected or directed graphs. Undirected graphical models are known as full conditional models or conditional independence graphs, and contain undirected edges. Undirected graphs are typically used to construct an undirected network with gene-gene interaction. If two variables are dependent given all the other variables, the corresponding two nodes are connected [38]. For instance, in the Gaussian graphical model, a non-zero element of the precision matrix (i.e., an inverse covariance matrix) indicates conditional dependency.

For directed acyclic graphs (DAGs), the directed edges indicate a causal relationship; it is useful to perform protein-protein interaction analysis [32] or gene expression data analysis [20]. In high dimensional data, especially when the sample size is smaller than the variable dimension, a lasso penalty is usually applied to estimate sparse networks. In Meinshausen and Bühlmann (2006), the estimation of a full conditional independence graph is shown by neighborhood selection from a linear model with a lasso penalty. Shojaie and Michailidis [61] converted the L_1-penalized likelihood to separable lasso problems to estimate DAGs under the known variable ordering. Fu and Zhou [22] and Han et al. [26] studied the L_1-penalized likelihood to estimate DAGs when the variable ordering is unknown.

Clustering analysis for gene networks has been studied for a few decades. There are many clustering approaches applied to gene expression data such as k-means [67], hierarchical clustering [16, 1, 73, 27, 41], self-organizing maps

[*] Sung Won Han and SungHwan Kim contribute equally to this paper.
[†] Corresponding authors.

(SOMs) [66, 69], Markov clustering algorithm [37, 21], simulated annealing [42], Nearest Neighbor Networks [30], Consense clustering [47, 75] and Spectral clustering algorithm [35, 70, 28]. In addition, several clustering methods are discussed for protein-protein interaction networks [65]; for instance, molecular complex detection algorithm [3], Markov cluster algorithm [17], and Clique Percolation Method [54].

A graph-based clustering is useful to circumvent the challenge of high dimensionality, and community detection in a social network is a popular example of the graph-based clustering. This approach can be used for clustering genes based on interactions and their strength. There are a bunch of methods for community detection methods in constructing graphs (refer to Fortunato [19] and Porter et al. [57]). Modularity maximization or spectral clustering is one of the well-known methods for detecting communities.

As aforementioned, it is computationally impractical to estimate DAGs under high dimensionality. In this paper, we discuss two different types of problems to estimate subnetworks in ultra high dimensional data. The first problem is to estimate DAGs of a subnetwork adjacent to a target gene, and the second problem is to estimate DAGs of multiple subnetworks without information about a target gene. To address each problem, we propose an efficient method to estimate subnetworks. This paper is organized as follows. Section 2 explains background knowledge and introduces problems of interest. Section 3 includes proposed algorithms for estimating directed acyclic graphs in high dimensional data, and Section 4 covers simulation studies performed under various simulation scenarios. In Section 5, we apply the proposed method to real example data to show the application of our approaches. Lastly we make a conclusion in Section 6.

2. PROBLEM FORMULATION

2.1 Graphical modeling and linear structure

Denote p variables of interest by $X_1, X_2, ..., X_p$, and let χ be a $n \times p$ data matrix, where n indicates the sample size or the number of the observations. A graph G can be represented by the variable set V with p nodes and edge set $E(V \times V)$. Each edge in E accounts for a relationship between two nodes, and if (i, j) is present in E, (j, i) should not be. We define $pa_i = \{\text{all } j | i \leftarrow j\}$, which is a parent set for i. To model a causal relationship, a structural equation model can be widely used [55]. Suppose that γ_i is an unexplained latent variable (e.g., noise effects), and it follows the independent normal distribution, $N(0, \sigma^2_{\gamma_i})$. The causal relationship can be represented by linear regression [61] as follows:

$$X_i = \sum_{j=1}^{p} a_{ij} X_j + \gamma_i, \tag{1}$$

where a_{ij} is a causal effect of a directional relationship from a parent j to a child i. Using a vector representation, let $\gamma = [\gamma_1, \gamma_2, ..., \gamma_p]^T$, and so $\gamma \sim MN(0, Q)$, where $Q = diag[\sigma^2_{\gamma_1}, \sigma^2_{\gamma_2}, ..., \sigma^2_{\gamma_p}]$. With $X = [X_1, X_2, ..., X_p]$, Equation (2) can be represented by

$$X = XA + \gamma, \tag{2}$$

where A is an adjacency matrix,

$$A = \begin{pmatrix} 0 & a_{1,2} & \cdots & a_{1,p-1} & a_{1,p} \\ a_{2,1} & 0 & \cdots & b_{2,p-1} & a_{2,p} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{p-1,1} & a_{p-1,2} & \cdots & 0 & a_{p-1,p} \\ a_{p,1} & a_{p,2} & \cdots & a_{p,p-1} & 0 \end{pmatrix}. \tag{3}$$

Note that X follows the multivariate normal distribution with $Var[X] = (I - A)^{-1} Q \big((I - A)^T\big)^{-1}$. We assume that x_i is standardized for $1 \le i \le p$, and so $\bar{x}_i = \sum_{k=1}^{n} x_{ik}/n = 0$, and $\sum_{k=1}^{n} (x_{ik} - \bar{x}_i)^2/n = 1$, where x_i is the i_{th} column vector of data matrix χ, and x_{ik} is the k_{th} component of x_i.

2.2 Penalized likelihood and Lasso framework

The log likelihood for the linear model is

$$\begin{aligned} -\frac{2}{n} \sum_{i=1}^{n} \log f(x_i) &\propto -\log |(I - A)^T Q^{-1} (I - A)| \\ &+ \operatorname{tr} \left[Q^{-1} (I - A) \hat{\Sigma} (I - A)^T \right], \end{aligned} \tag{4}$$

where x_i is the i_{th} column vector of matrix χ, and I is a $p \times p$ identity matrix. $\hat{\Sigma}$ is a sample covariance matrix defined by $\hat{\Sigma} = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T/n$, where $\bar{x} = \sum_{i=1}^{n} x_i/n$. According to Shojaie and Michailidis [61] and Han et al. [26], the log likelihood can be approximated as

$$\sum_{k=1}^{p} \left[(I - A) \hat{\Sigma} (I - A)^T \right]_{k,k} + \lambda \sum_{j=1}^{p} |a_{kj}|. \tag{5}$$

However, this approximation requires the condition of acyclicity, and the complete optimization problem was proposed by Han et al. [26] as below:

$$\min_{A,T} \sum_{k=1}^{p} \left[\frac{1}{n} \| \chi_k - \chi a_k \|_2^2 + \lambda \sum_{j=1}^{p} |a_{kj}| \right], \tag{6}$$

subject to $T_{ij} = I(a_{kj} \neq 0)$, and

$$\sum_{l=1}^{min(Card(T),p)} \sum_{m=1}^{p} \left[T^l \right]_{m,m} = 0. \tag{7}$$

When applying this algorithm to high-dimensional data, we necessarily encounter two challenges. The first is the expensive computational cost. To address this challenge, [26] proposed an approach based on a two-step procedure. In the first step, the algorithm purposely omits the acyclic constraint (7), and then solves the optimization problem (6). This procedure is identical to the estimation of a full conditional independence graph by Lasso (Meinshausen and Bühlmann, 2006), and hence this reduces searching space within the full conditional independence graph. In the second step, they fit an optimal graphical model using the objective function (6) together with an acyclic condition. This algorithm seems to be computationally efficient, but yet might not be applicable to ultra high dimensional data consisting of about 20,000 genomic features. To improve the approach, we propose an algorithm for the target driven estimation of DAGs in case of ultra high dimensional data, as well as for the estimation of subnetworks.

The main challenge is how to select λ_1. Meinshausen and Bühlmann (2006) mentioned that the selection of probabilistic neighbors is equivalent to variable selection by the lasso regression. In the two-stage approach, the first stage deals with the neighbor selection problem, which is related to selection of the penalty parameter in the lasso regression. For selecting the penalty parameter, there are two main objectives to be considered; one is consistency of prediction error, and the other is consistency of variable selection. To minimize average prediction errors, cross validation (CV) can be used, and is well known to be asymptotically optimal for estimating penalties. Nevertheless, CV does not ensure consistency in model selection for the Lasso penalty [72]. In addition, applying CV often incurs a high computational expense. To tackle this problem, Meinshausen and Buhlmann [46] proposed a method for the penalty choice under some regularization conditions in order to control the false discovery. In principle, the method is proposed under an assumption of asymptotic properties, which, in practice, often does not function properly, especially when applied to a finite sample size and the moderate number of true neighbors. For consistency of model selection, the BIC (Bayesian Information criteria) along with the best subset selection approach is suggested [59, 60]. However, the approach using BIC criteria suffers from heavy computational time for high dimensional data, so the combination of the lasso regression and the BIC criteria has been widely implemented. It was proved that the lasso approach obtains more stable and correct estimations than the subset variable selection method [68], and the BIC approach is known to satisfy consistent model selection. Taken together, in this paper, we will experimentally examine the estimation of directed networks by various criteria, and will show the superiority of the proposed method compared to the existing approaches.

3. ESTIMATION OF THE TARGET SUBNETWORK

In this section, we propose an algorithm to estimate a directed acyclic graph around a target response. We propose a two-step approach: (1) finding neighbors around a target gene, (2) estimating a directionality among variables within the set of neighbors.

3.1 Finding neighbors

Without loss of generality, let X_p be a target response variable. We apply a neighbor selection algorithm starting from X_p. The neighbors of X_p are the set of variables with non-zero coefficients from

$$\frac{1}{n}\|\chi_p - \chi a_p\|_2^2 + \lambda_1 \sum_{j=1}^{p} |a_{pj}|. \tag{8}$$

Let the level of χ_p be L_1, then we denote $L_1 = \{X_P\}$. The variables selected by Equation (8) in terms of X_p are included in the layer 2 (L_2). Let X^{L_m} be a variable in the set of layer m, L_m. The set of the $m+1$ level, L_{m+1}, includes a variable $X^{L_{m+1}}$, which is obtained by

$$\frac{1}{n}\|\chi_{i_m} - \chi a_{i_m}\|_2^2 + \lambda_m \sum_{j=1}^{p} |a_{i_m j}|, \tag{9}$$

where $X_{i_m} \in L_m$. We keep running the above neighbor selection algorithm until a stopping criterion is satisfied. We propose several rules for the searching level. The first approach is to use a fixed upperbound for the searching level, and so the algorithm searches neighbors until the level U. The second approach is to use the level-dependent penalty λ_m so that the penalty value λ increases as the level increases.

For the fixed upper bound approach of the searching level, λ_m at a level m can be decided by the formula suggested by Meinshausen and Bühlmann (2006):

$$\lambda_k(\alpha) = \frac{2\hat{\sigma}_{X_k}}{\sqrt{n}}\Big[1 - \Phi^{-1}\Big(\frac{\alpha}{2p^2}\Big)\Big]. \tag{10}$$

We call this approach Fix-alpha, which is suggested to use in [26]. The second fixed upper bound approach is using BIC. Let $\hat{a}_{i_m}$ be an estimate from Equation (9), then based on several values of λ, λ_m can be obtained by the BIC criterion, which is

$$\frac{1}{n}\|\chi_{i_m} - \chi \hat{a}_{i_m}\|_2^2 + Card(\hat{a}_{i_m})\frac{log(n)}{n}. \tag{11}$$

We call this approach Fix-BIC.

Besides the fixed upper bound approach using the layer-independent penalty, we can use the BIC with a level-dependent penalty:

$$\frac{1}{n}\|\chi_{i_m} - \chi \hat{a}_{i_m}\|_2^2 + Card(\hat{a}_{i_m})\frac{log(n)}{n} \times e^{h(m)}. \tag{12}$$

We propose two functions for $h(m)$; $h_1(m) = c_1(m-1)$ or $h_2(m) = c_2 \log(m)$. If $h_1(0) = h_2(0) = 1$, then Equation

(12) becomes the BIC approach in (11). We call the approach using the function $h_1(\cdot)$ as Var-BIC-h1, and we call the approach using the function $h_2(\cdot)$ as Var-BIC-h2. As the layer increases, a higher weight is assigned to the degree of freedom term, which also leads to an increase of λ. The Var-BIC-h1 approach gives a penalty, which increases exponentially in the level, whereas the Var-BIC-h2 approach increases polynomially. As the algorithm search in the higher layer, the penalty term becomes large, which automatically stops the search.

3.2 Estimation of DAGs

After we identify the set of neighbors around a target gene, we need to estimate the structure of a DAG. To estimate a DAG, we can apply a search algorithm to find the directionality within the neighborhood structure obtained from the previous step. Since the variable selection step is complete in the previous step, we no longer need a penalty term, and so can only estimate the directionality based on likelihood. The optimization problem is as below:

$$\min_{A,T} \sum_{k=1}^{p} \frac{1}{n} \|\chi_k - \chi a_k\|_2^2, \tag{13}$$

subject to

$$\sum_{l=1}^{min(Card(T),p)} \sum_{m=1}^{p} \left[T^l\right]_{m,m} = 0, \tag{14}$$

and

$$T_{kj} \leq N_{kj}, \tag{15}$$

where N is a neighborhood structure matrix. The description of how to find the solution in the optimization problem is in Appendix.

4. ESTIMATION OF THE MULTIPLE SUBNETWORKS

To estimate multiple subnetworks of DAGs in ultra high dimensional data, we propose a hybrid approach with community detection algorithms. We first estimate an undirected graph or neighbors based on the L_1-penalized likelihood in Equation (6). Based on the undirected graph, we apply a community detection algorithm to find graphs or clusters of the genes based on their estimated interaction. Given each cluster, we estimate directionality by incorporating Equation (6) with the acyclic constraint (7).

We use several well known community detection algorithms. For instance, one of the well-known metrics to gauge how effectively a network is divided into groups or communities is *modularity* [52, 51]. Modularity is defined by "the total number of edges among vertexes within the same community" minus "the expected number of edges among such vertexes, if they are randomly distributed". Here, we investigate three searching methods: fast greedy algorithm, walktrap algorithm, and leading eigenvector algorithm, which are implemented in "igraph" package in R. The fastgreedy algorithm by Clauset et al. [13] is a hierarchical approach, which improves a modularity function in a greedy manner. It starts its own separate community for each node, and communities are merged at each iteration in order to increase the modularity function value. The process stops when the modularity function value cannot increase any more.

The walktrap algorithm by Pons and Latapy [56] uses a stochastic approach based on a random walk. It generates random walks on the network. Since there are more edges between nodes within a community than those outside a community, the random walks tend to stay within the same community. The walktrap technique generates short random walks, and by using the results, it merges smaller communities into larger ones, which automatically generates a dendrogram. The modularity score is used to select the threshold for cutting a dendrogram.

The leading eigenvector algorithm by Newman [51] is the extension of the spectral partitioning method in a graph, but the former uses the modularity matrix instead of the Laplacian matrix used by the latter. It starts with the whole network at the initial step, and it splits the network into two parts which gives significant increase of the modularity. The leading eigenventor of the modularity matrix is used to split the networks.

5. SIMULATION STUDY

To experimentally compare performance of the proposed methods, we perform simulation studies. We consider various simulation scenarios in terms of the number of variables (p) and the number of parents per child (d). The parameter values we used are ($p = 200, 500,$ and 1000), ($n = 500$), and ($d = 2$, 4, or 6). The latent variables (γ_i) are generated from the standard normal distribution. The value a_{ij}, the causal relationship from X_i to X_j, is set at 0.8. For estimating target-driven subnetworks, we distribute edges from X_p, which is a child node for all other edges. For multiple subnetworks, we distribute edges randomly from each node with blocks of matrices, which indicate clusters.

Comparing performances of the methods is not a trivial task, since the methods detect subnetworks after defining clusters. Therefore, we consider estimating performance within the clusters. First, we define $p(C)$ by the number of nodes within clusters, and $p(A)$ by the number of nodes in an entire graph. We also define $e(C)$ by the number of true edges within clusters, and $p(A)$ by the number of nodes in an entire graph. Within the cluster, we define the number of true positives (TP), the number of false positives (FP), the number of true negatives (TN), and the number of false negatives (FN). Similarly, we also use directional true positive (dTP) defined as the number of correctly estimated directionality [22, 26]. In addition, we also consider directional false negative (dFN), which counts the number of wrongly estimated directions of edges as well as the number of estimated non-existing edges.

ROC curves based on TP and FP is not suitable to apply since they depend on the size of the clusters. Instead, we show similar curves based on overall performance such as MCC with $p(C)/p(A)$ or $e(C)/e(A)$. The Matthew's correlation coefficient (MCC) is calculated by

(16)
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]^{1/2}}.$$

Within a cluster, the higher value indicates the better fit. Similarly, dMCC can be defined based on dTP and dFN instead of TP and FP. We also use positive predictive value (PPV), which is defined by TP over the number of all estimated edges. Similarly, dPPV is defined by dTP over the number of all estimated edges. PPV does not count FN or TN. MCC(dMCC) and PPV(dPPV) focus on values within clusters.

5.1 Simulation results for selection of neighbors

Finding neighbors in subsection 3.1 is the iterative variable selection procedure. Before we perform full simulation studies, we thus first investigate performances of the criteria at one iteration, which is essentially single variable selection procedure. More precisely, we investigate performances in terms of the penalty parameter criteria under one child structure. For p variables, X_p is assumed to be a child node, and edges from parent nodes are randomly selected.

We come up with three simulation scenarios. Scenario 1 is that the true parent variables are independent of one another, and also independent of other non-parent variables. In Scenario 2, the true parent variables are independent of one another, but they are dependent on other non-parent variables. Especially, we assume that the non-parent variables are ancestors of the child variables. For example, if X_p is a child variable and X_i is the parent variable for X_p. We assign an edge between X_i and X_j, where X_j is not in the set of $\{X_k | X_k \text{ is a parent for } X_p\}$. In Scenario 3, we consider all parents are dependent on one another. Especially, we control each parent variable to be affected by only one other parent. The simulation result for d=2 and 3 is presented in Table 1.

Based on the simulation study, the approach based on the formula in (10) performs very well for all three scenarios when $n = 500$, as we expected. The formula is derived based on asymptotic property (i.e., n is large). However, if n is a small size (i.e., $n = 100$), performance become worse. For example, in Scenario 1, if $d = 2$, $p = 200$, and $n = 100$, the average MCC from α-method is 0.85, and the corresponding MCC if $d = 3$ is 0.56. The BIC approach shows good performance under most cases, which indicates robustness. As a data-driven approach, it adjusts the correlation structure of the data. Finally, we also investigate performance of GCV criteria. As we discuss in the introduction, CV satisfies asymptotic consistency in prediction, but not in variable selection. Taken together, in all cases, performance of GCV is the worst. We also investigate higher d such as d=4 and 5, and the simulation result is in Table 2. The patterns of performance among the three approaches are similar to those when d=2 and 3.

5.2 Simulation results for target-driven subnetworks

We compare four different techniques: the Fix-alpha method, the Fix-BIC method, the Var-BIC-h1 method, and the Var-BIC-h2 method. On the whole, using the BIC criteria gives better performance than using the alpha formula, the fixed-alpha method. Among the methods using BIC criteria, the techniques with layer-dependent weights such as the Var-BIC-h1 and Var-BIC-h2 methods show higher performance than the fixed layer technique such as the Fix-BIC method, especially for small sized clusters. Such performance patterns are clear when d=2, but when d=4 or 6, overall performances decrease for all methods since the signal-to-noise ratio becomes small. The detailed explanation of the performance comparisons especially for d=2 is as follows.

We first compare the PPV or MCC metric along with the different sizes of clusters in terms of the number of nodes, as shown in Figure 1. PPV considers only true positives or false positives based on estimated edges, and so it does not consider which part of the edges should be counted as true or false negatives. Thus, PPV is also a good metric for estimation of a subnetwork.

Based on PPV metric plots in Figure 1 (a), the Fix-alpha method is shown to have the worst performance. For example, when $p = 200$, the PPV of the Fix-alpha method is less than 0.4 in the range of p(C)/p(A) between 0 and 0.8. As p increases up to 1000, the PPV of the Fix-alpha method reduces and becomes less than 0.2. If we use the BIC criteria, say Fix-BIC, performance is significantly improved. For the range of p(C)/p(A) greater than 0.2, the PPV of the Fix-BIC method is over 0.4, and for the range of p(C)/p(A) greater than 0.4, it becomes over 0.6. The performance increases if we use the methods with layer dependent weights. The Var-BIC-h1 method shows the PPV over 0.5 even in the range of small p(C)/p(A). In the range greater than 0.2 of p(C)/p(A), the PPV of the Var-BIC-h1 method is over 0.6. Furthermore, the Var-BIC-h2 shows higher PPVs than 0.6 in the entire range of p(C)/p(A). Overall, the Var-BIC-h2 shows the best performance regardless of the size of the subnetwork, which indicates the robustness of the methods.

If we compare them based on dPPV, the values reduce. The performance patterns in terms of dPPV are similar to those based on PPV. The dPPV of the Fix-alpha method is less than 0.2 when p=200, and less than 0.1 when p=1000. However, the dPPVs of the Var-BIC-h1 and Var-BIC-h2 methods are around or over 0.4.

Table 1. Simulation results from independent or dependent variables under one child structure

d	p	n	Scenario 1			Scenario 2			Scenario 3		
			α	BIC_p	GCV	α	BIC_p	GCV	α	BIC_p	GCV
2	50	100	.96 (.10)	.96 (.09)	.18 (.03)	1.00 (.04)	.96 (.09)	.18 (.03)	1.00 (.03)	.99 (.03)	.18 (.03)
2	50	200	1.00 (.00)	.98 (.06)	.19 (.03)	1.00 (.00)	.97 (.07)	.19 (.03)	1.00 (.00)	1.00 (.02)	.19 (.03)
2	50	500	1.00 (.00)	.98 (.06)	.20 (.03)	1.00 (.00)	.99 (.04)	.18 (.02)	1.00 (.00)	1.00 (.00)	.20 (.03)
2	50	1000	1.00 (.00)	.99 (.04)	.22 (.03)	1.00 (.00)	.98 (.06)	.18 (.03)	1.00 (.00)	1.00 (.00)	.19 (.03)
2	100	100	.91 (.20)	.95 (.09)	.09 (.02)	.99 (.05)	.96 (.09)	.10 (.02)	.99 (.04)	1.00 (.03)	.10 (.02)
2	100	200	1.00 (.00)	.97 (.08)	.13 (.02)	1.00 (.00)	.97 (.08)	.13 (.02)	1.00 (.00)	1.00 (.03)	.13 (.02)
2	100	500	1.00 (.00)	.98 (.05)	.14 (.02)	1.00 (.00)	.98 (.06)	.13 (.01)	1.00 (.00)	1.00 (.00)	.14 (.01)
2	100	1000	1.00 (.00)	.99 (.05)	.15 (.01)	1.00 (.00)	.99 (.05)	.13 (.01)	1.00 (.00)	1.00 (.00)	.13 (.01)
2	200	100	.85 (.22)	.96 (.08)	.10 (.00)	.96 (.10)	.96 (.08)	.11 (.00)	.99 (.06)	.99 (.03)	.11 (.00)
2	200	200	1.00 (.00)	.96 (.08)	.07 (.01)	1.00 (.00)	.97 (.07)	.07 (.01)	1.00 (.00)	1.00 (.02)	.07 (.01)
2	200	500	1.00 (.00)	.99 (.04)	.09 (.01)	1.00 (.00)	.99 (.04)	.10 (.01)	1.00 (.00)	1.00 (.00)	.09 (.02)
2	200	1000	1.00 (.00)	1.00 (.03)	.10 (.02)	1.00 (.00)	.99 (.05)	.09 (.01)	1.00 (.00)	1.00 (.00)	.10 (.01)
2	500	100	.75 (.30)	.94 (.09)	.12 (.00)	.94 (.12)	.96 (.08)	.13 (.00)	.98 (.07)	.99 (.03)	.13 (.00)
2	500	200	1.00 (.00)	.97 (.07)	.08 (.00)	1.00 (.00)	.98 (.06)	.08 (.00)	1.00 (.00)	1.00 (.00)	.08 (.00)
2	500	500	1.00 (.00)	.99 (.04)	.04 (.01)	1.00 (.00)	.99 (.04)	.05 (.00)	1.00 (.00)	1.00 (.00)	.05 (.00)
2	500	1000	1.00 (.00)	.99 (.04)	.06 (.00)	1.00 (.00)	.99 (.03)	.06 (.00)	1.00 (.00)	1.00 (.00)	.07 (.00)
3	50	100	.84 (.16)	.82 (.15)	.21 (.03)	.95 (.08)	.78 (.15)	.21 (.04)	.94 (.09)	.91 (.11)	.22 (.04)
3	50	200	1.00 (.00)	.84 (.13)	.24 (.04)	1.00 (.00)	.83 (.14)	.23 (.04)	.84 (.07)	.90 (.11)	.28 (.06)
3	50	500	1.00 (.00)	.89 (.12)	.22 (.03)	1.00 (.00)	.87 (.12)	.25 (.04)	1.00 (.00)	.92 (.11)	.24 (.04)
3	50	1000	1.00 (.00)	.92 (.10)	.21 (.03)	1.00 (.00)	.88 (.11)	.24 (.04)	.93 (.06)	.94 (.09)	.32 (.05)
3	100	100	.80 (.18)	.81 (.15)	.12 (.03)	.87 (.14)	.80 (.17)	.12 (.03)	1.00 (.00)	.96 (.09)	.24 (.03)
3	100	200	1.00 (.00)	.85 (.12)	.16 (.02)	1.00 (.00)	.85 (.13)	.15 (.02)	1.00 (.00)	.96 (.07)	.26 (.04)
3	100	500	1.00 (.00)	.89 (.11)	.16 (.01)	1.00 (.00)	.88 (.12)	.17 (.03)	1.00 (.00)	.98 (.06)	.28 (.07)
3	100	1000	1.00 (.00)	.92 (.10)	.15 (.02)	1.00 (.00)	.90 (.11)	.17 (.02)	1.00 (.00)	.97 (.06)	.33 (.05)
3	200	100	.56 (.30)	.22 (.24)	.13 (.00)	.75 (.22)	.34 (.31)	.13 (.01)	.92 (.09)	.90 (.11)	.12 (.03)
3	200	200	1.00 (.03)	.85 (.14)	.08 (.02)	1.00 (.01)	.82 (.14)	.09 (.02)	.82 (.09)	.90 (.11)	.17 (.02)
3	200	500	1.00 (.00)	.91 (.10)	.12 (.01)	1.00 (.00)	.89 (.12)	.10 (.01)	1.00 (.00)	.92 (.11)	.16 (.03)
3	200	1000	1.00 (.00)	.93 (.10)	.11 (.01)	1.00 (.00)	.89 (.11)	.13 (.01)	.92 (.05)	.97 (.06)	.22 (.04)
3	500	100	.50 (.29)	.15 (.00)	.15 (.00)	.66 (.23)	.15 (.00)	.15 (.00)	1.00 (.00)	.94 (.09)	.17 (.02)
3	500	200	.99 (.04)	.87 (.12)	.10 (.00)	1.00 (.02)	.85 (.12)	.10 (.00)	1.00 (.00)	.96 (.07)	.19 (.02)
3	500	500	1.00 (.00)	.91 (.11)	.05 (.00)	1.00 (.00)	.89 (.11)	.06 (.00)	1.00 (.00)	.97 (.06)	.20 (.05)
3	500	1000	1.00 (.00)	.91 (.10)	.08 (.00)	1.00 (.00)	.89 (.11)	.08 (.01)	1.00 (.00)	.95 (.07)	.24 (.02)

We also compare the methods based on MCC criteria, which consider TP and FP as well as true and false negatives within clusters. The overall patterns of performance are similar to those based on PPV. In the small range of p(C)/p(A) between 0 and 0.2, the MCCs of the Fix-alpha method are less than 0.2 for all p cases. However, the MCC of the Var-BIC-h2 method is over 0.6 in the small range of p(C)/p(A) between 0 and 0.2 when p=200, and over 0.7 when p=1000. The MCCs of the Fix-BIC method are smaller than those of the variable weighted methods in the range of small p(C)/p(A), and the MCCs of the Fix-BIC method get close to those of the weighted methods as the ratio p(C)/p(A) increases. Based on the directional MCC (dMCC) metric, the values reduce, but performance patterns are similar. We also compare performance based on the ratios of edges in the clusters to those in the entire graph, e(C)/e(A), which is shown in Figure 2. The performance patterns are similar to those based on p(C)/p(A). The methods with layer-dependent weights based on BIC criteria (the Var-BIC-h1 and the Var-BIC-h2 methods) show better performance than others, and performances are robust against the e(C)/e(A). The Var-BIC-h2 method shows a slightly better performance than the Var-BIC-h1 method in the range of small e(C)/e(A), but their performances become similar in the range of large e(C)/e(A). In addition, the performance pattern when d=4 or 6 are similar to that when d=2 (Appendix A and B in Supplementary Materials http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0010/0004/s003). However, the performance difference is not shown clearly since overall performances of all methods decrease.

We also compare computational times especially when p=1000. In order to obtain computational times, we ran the simulations under the workstation with E5-2630 v3 CPU

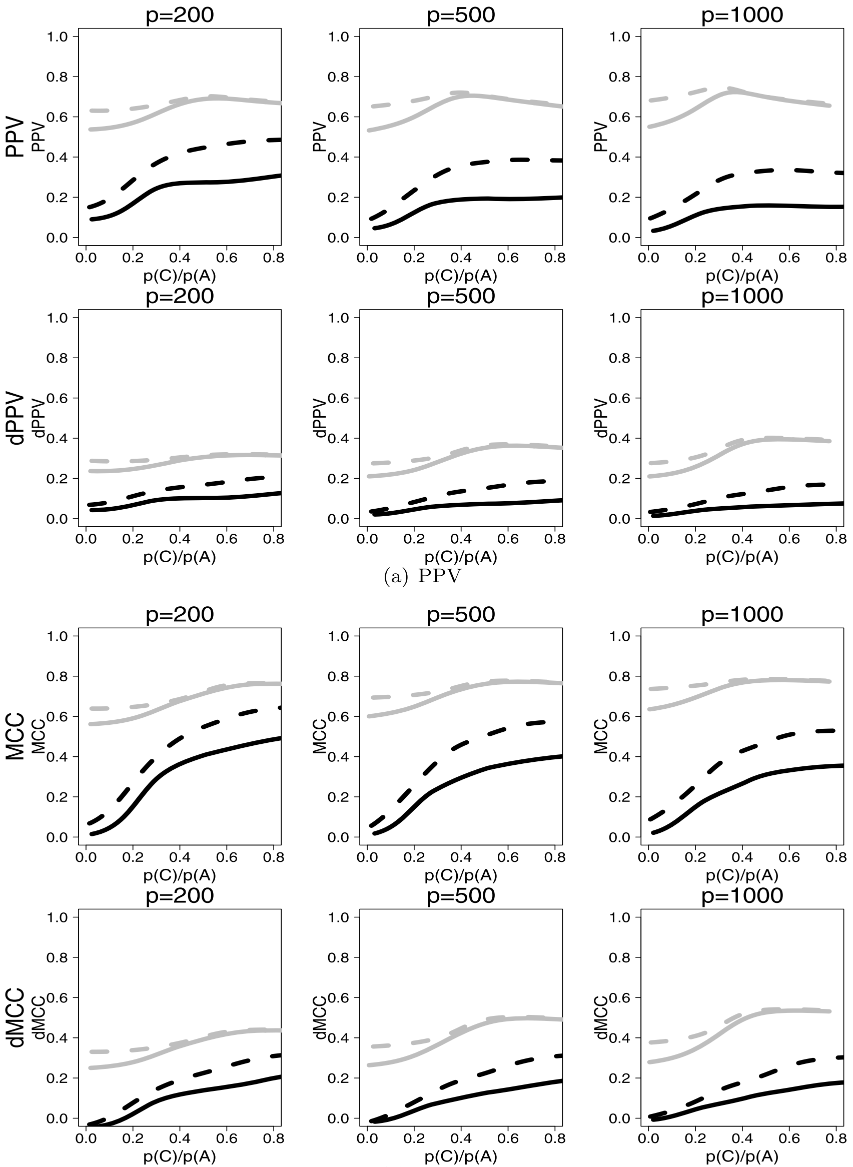
Table 2. Simulation results from independent or dependent variables under one child structure

d	p	n	Scenario 1			Scenario 2			Scenario 3		
			α	BIC_p	GCV	α	BIC_p	GCV	α	BIC_p	GCV
4	50	100	.69 (.17)	.94 (.08)	.25 (.04)	.77 (.14)	.95 (.07)	.24 (.04)	.88 (.07)	1.00 (.01)	.25 (.05)
4	50	200	1.00 (.00)	.97 (.06)	.26 (.04)	1.00 (.00)	.96 (.06)	.25 (.04)	.98 (.05)	1.00 (.00)	.25 (.04)
4	50	500	1.00 (.00)	.99 (.03)	.27 (.04)	1.00 (.00)	.98 (.05)	.25 (.04)	1.00 (.00)	1.00 (.00)	.33 (.05)
4	50	1000	1.00 (.00)	.99 (.03)	.25 (.03)	1.00 (.00)	.99 (.03)	.30 (.04)	1.00 (.00)	1.00 (.00)	.25 (.04)
4	100	100	.59 (.19)	.95 (.07)	.14 (.03)	.66 (.18)	.95 (.07)	.14 (.03)	.87 (.07)	1.00 (.00)	.14 (.02)
4	100	200	.99 (.04)	.96 (.06)	.18 (.02)	1.00 (.01)	.96 (.06)	.18 (.02)	.97 (.06)	1.00 (.00)	.19 (.02)
4	100	500	1.00 (.00)	.98 (.04)	.20 (.02)	1.00 (.00)	.98 (.04)	.17 (.02)	1.00 (.00)	1.00 (.00)	.21 (.05)
4	100	1000	1.00 (.00)	.99 (.04)	.19 (.02)	1.00 (.00)	.98 (.05)	.22 (.02)	1.00 (.00)	1.00 (.00)	.18 (.02)
4	200	100	.48 (.24)	.94 (.07)	.15 (.01)	.56 (.21)	.94 (.08)	.15 (.01)	.85 (.06)	1.00 (.00)	.17 (.01)
4	200	200	.96 (.07)	.98 (.05)	.10 (.02)	1.00 (.02)	.97 (.05)	.10 (.02)	.96 (.06)	1.00 (.00)	.11 (.01)
4	200	500	1.00 (.00)	.99 (.04)	.13 (.02)	1.00 (.00)	.99 (.03)	.13 (.01)	1.00 (.00)	1.00 (.00)	.11 (.02)
4	200	1000	1.00 (.00)	.99 (.03)	.14 (.01)	1.00 (.00)	.98 (.04)	.16 (.01)	1.00 (.00)	1.00 (.00)	.13 (.01)
4	500	100	.30 (.27)	.93 (.08)	.18 (.00)	.44 (.24)	.93 (.08)	.18 (.00)	.83 (.08)	1.00 (.02)	.19 (.01)
4	500	200	.93 (.09)	.97 (.05)	.12 (.00)	.98 (.05)	.97 (.05)	.12 (.00)	.93 (.07)	1.00 (.00)	.13 (.00)
4	500	500	1.00 (.00)	.99 (.03)	.06 (.00)	1.00 (.00)	.99 (.04)	.07 (.00)	1.00 (.00)	1.00 (.00)	.09 (.00)
4	500	1000	1.00 (.00)	.99 (.04)	.10 (.00)	1.00 (.00)	.99 (.04)	.07 (.02)	1.00 (.00)	1.00 (.00)	.09 (.00)
5	50	100	.55 (.20)	.77 (.12)	.27 (.05)	.65 (.13)	.76 (.13)	.28 (.05)	.91 (.09)	.87 (.17)	.14 (.01)
5	50	200	.98 (.05)	.82 (.11)	.28 (.04)	.99 (.03)	.81 (.13)	.28 (.04)	.82 (.08)	.94 (.08)	.19 (.01)
5	50	500	1.00 (.00)	.86 (.13)	.30 (.05)	1.00 (.00)	.84 (.11)	.28 (.04)	1.00 (.02)	.94 (.09)	.09 (.01)
5	50	1000	1.00 (.00)	.88 (.09)	.30 (.05)	1.00 (.00)	.85 (.10)	.33 (.08)	.91 (.04)	.96 (.07)	.14 (.01)
5	100	100	.44 (.23)	.76 (.15)	.15 (.03)	.51 (.16)	.76 (.11)	.16 (.03)	1.00 (.00)	.94 (.09)	.13 (.01)
5	100	200	.94 (.09)	.83 (.12)	.19 (.03)	.98 (.05)	.81 (.12)	.20 (.03)	1.00 (.00)	.95 (.07)	.15 (.01)
5	100	500	1.00 (.00)	.87 (.10)	.22 (.03)	1.00 (.00)	.85 (.10)	.20 (.02)	1.00 (.00)	.98 (.05)	.11 (.03)
5	100	1000	1.00 (.00)	.88 (.08)	.22 (.02)	1.00 (.00)	.86 (.10)	.26 (.04)	1.00 (.00)	.97 (.05)	.18 (.01)
5	200	100	.31 (.26)	.27 (.22)	.17 (.01)	.38 (.25)	.42 (.31)	.18 (.01)	.88 (.09)	.22 (.21)	.16 (.00)
5	200	200	.90 (.11)	.83 (.12)	.11 (.02)	.94 (.08)	.80 (.12)	.12 (.02)	.78 (.09)	.82 (.26)	.21 (.01)
5	200	500	1.00 (.00)	.87 (.10)	.14 (.02)	1.00 (.00)	.84 (.10)	.15 (.01)	.99 (.04)	.94 (.10)	.11 (.00)
5	200	1000	1.00 (.00)	.89 (.09)	.16 (.01)	1.00 (.00)	.86 (.09)	.18 (.03)	.90 (.03)	.94 (.09)	.15 (.01)
5	500	100	.16 (.23)	.20 (.00)	.20 (.00)	.22 (.24)	.20 (.01)	.20 (.01)	1.00 (.00)	.94 (.09)	.07 (.00)
5	500	200	.82 (.12)	.83 (.12)	.13 (.00)	.90 (.09)	.82 (.11)	.14 (.00)	1.00 (.00)	.95 (.06)	.11 (.00)
5	500	500	1.00 (.00)	.87 (.09)	.08 (.00)	1.00 (.00)	.85 (.09)	.08 (.00)	1.00 (.00)	.97 (.08)	.07 (.00)
5	500	1000	1.00 (.00)	.88 (.10)	.11 (.02)	1.00 (.00)	.85 (.10)	.08 (.00)	1.00 (.00)	.96 (.06)	.12 (.01)

and linux OS with R. The curves of computational times based on p(C)/p(A) are shown in Figure 3. Overall, the Fix-BIC-method shows smaller computational times than other methods except for the range after p(C)/p(A) when d=2. It was expected since the Fix-alpha method uses only one penalty parameter based on Equation 10. The computational time of the Fix-BIC method as well as variable BIC approaches is higher than that of the Fix-alpha method. However, all of the approaches show reasonable computational times if we focus on the estimation of networks around the target genes. For example, when d=2, the computational time is less than 2 hours if the subnetwork is 50% of the whole one (i.e., p(C)/p(A)<0.5). If d=4 or 6, the computation time is less than 2 hours if the subnetwork is 60%∼70% of the whole one (i.e., p(C)/p(A)<0.6∼0.7). Thus, the Var-BIC-h2 method shows the better performance than or as good as those of other approaches with comparative computational time.

5.3 Simulation results for multiple subnetworks

We also compare three different community detection algorithms to estimate multiple subnetworks by simulation study: the walktrap, fastgreedy, and leading eigenvector algorithms. We first estimate the entire undirected graph with the lasso linear regression in (8) with the penalty parameter (10). Then, we apply the community detection algorithm to find the cluster. We examine whether different algorithms affect performance of estimation in subnetworks. Figure 4 shows PPV (dPPV) or MCC (dMCC) plots based on p(C)/p(A) when d=2. Given p(C)/p(A), PPV (dPPV) and MCC (dMCC) are almost similar. Since the community detection algorithm defines clusters in terms of connection by edges, the ratio of the number of nodes within clusters is irrelevant to the detection performance. On the other hand, the ratio of e(C)/e(A) is related to performance



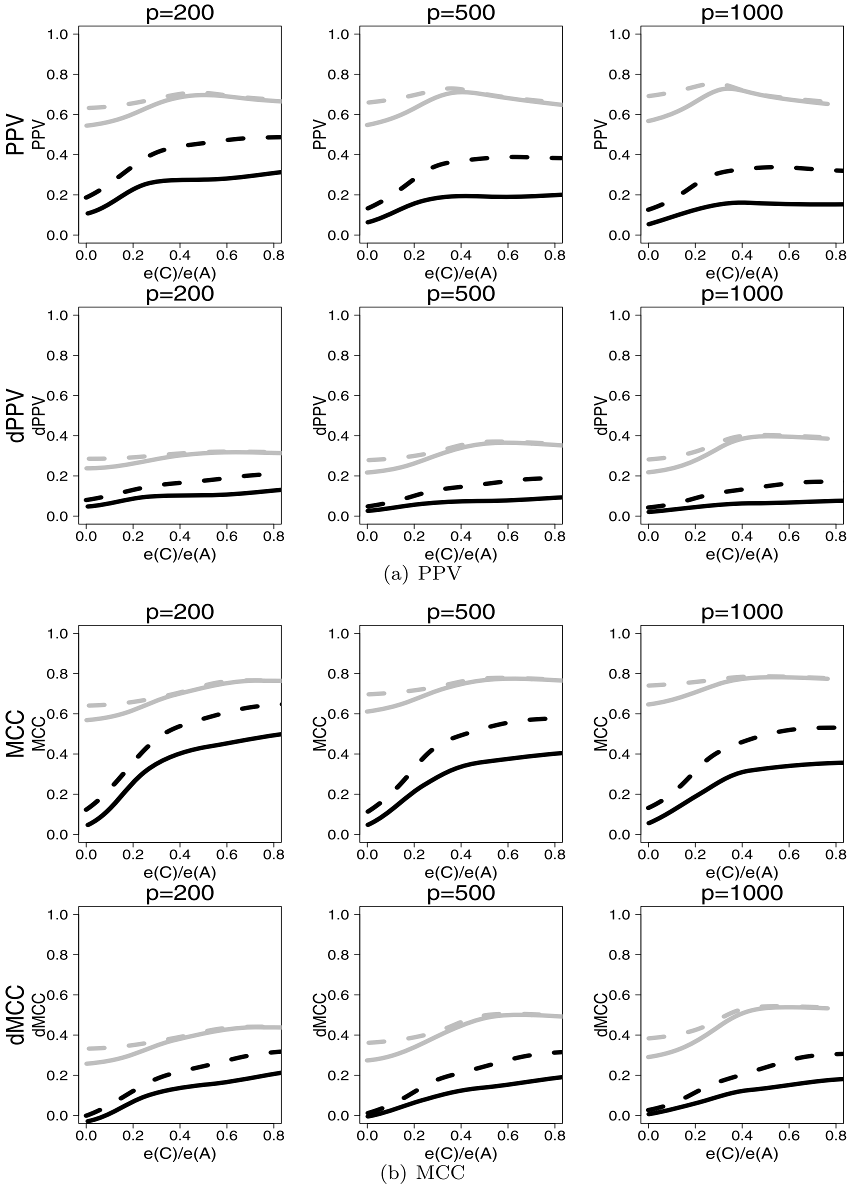
(a) PPV

(b) MCC

Figure 1. PPV or MCC plots based on p(C)/p(A) when d=2. The solid black line indicates the Fix-alpha method, and the dashed black line indicates the Fix-BIC method. The solid gray line indicates the Var-BIC-h1 method, and the dashed gray line indicates the Var-BIC-h2 method.

in terms of PPV or MCC. Figure 5 shows the PPV (dPPV) or MCC (dMCC) plots based on e(C)/e(A). The fastgreedy algorithm is the heuristic search algorithm. PPV or MCC from the fast greedy algorithm are high at small e(C)/e(A), and they become reduced as e(C)/e(A) increases. For example, when $p = 500$, PPV is close to 0.9 at e(C)/e(A)=0.2, and it is close to 0.7 at e(C)/e(A)=1.0.

For the walktrap algorithm, similar patterns are observed. However, the PPV or MCC of the walktrap algorithm are generally smaller than those of the fastgreedy algorithm given the same e(C)/e(A). Furthermore, when $p = 1000$, the PPV (dPPV) or MCC (dMCC) decreases sharply as the e(C)/e(A) increases until e(C)/e(A)=0.6. After e(C)/e(A)=0.6, they increase smoothly. This down-and-up pattern is clearly shown as p is high. The leading eigenvector algorithm shows a more severe down-and-up pattern than the walktrap algorithm. When $p = 500$ or $p = 1000$, the PPV decreases as e(C)/e(A) increases until e(C)/e(A)=0.35. After e(C)/e(A)=0.35, the PPV or MCC increases, and after e(C)/e(A)=0.6, they become stable. Overall, the fastgreedy algorithm generally gives a slightly better and more consistent performance in terms of the ratio e(C)/e(A) than the other two algorithms' performance. In addition, we also compare the performances of three approaches when d=4



(a) PPV

(b) MCC

Figure 2. PPV or MCC plots based on e(C)/e(A) when d=2. The solid black line indicates the Fix-alpha method, and the dashed black line indicates the Fix-BIC method. The solid gray line indicates the Var-BIC-h1 method, and the dashed gray line indicates the Var-BIC-h2 method.

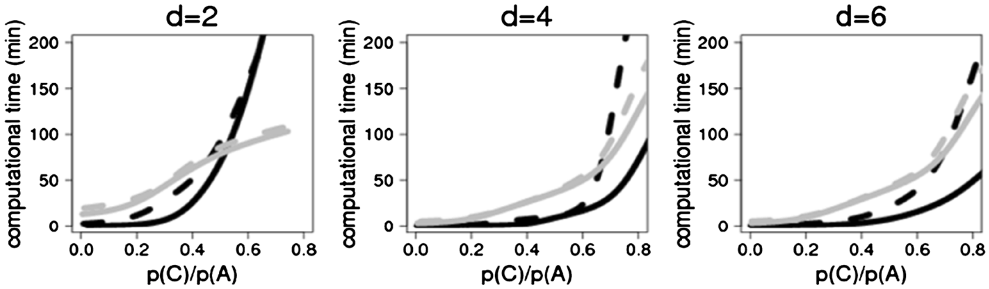
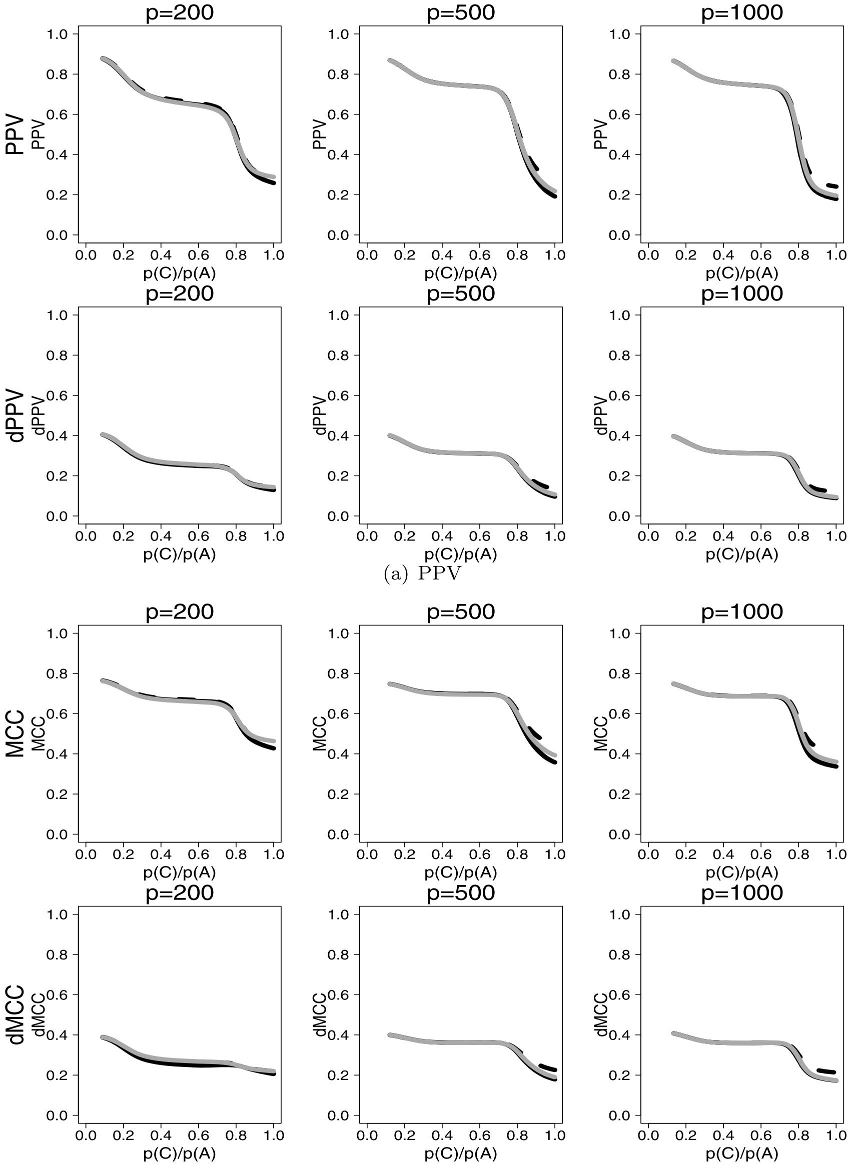


Figure 3. Computational time of the combined steps when p=1000. The solid black line indicates the Fix-alpha method, and the dashed black line indicates the Fix-BIC method. The solid gray line indicates the Var-BIC-h1 method, and the dashed gray line indicates the Var-BIC-h2 method.



(a) PPV

(b) MCC

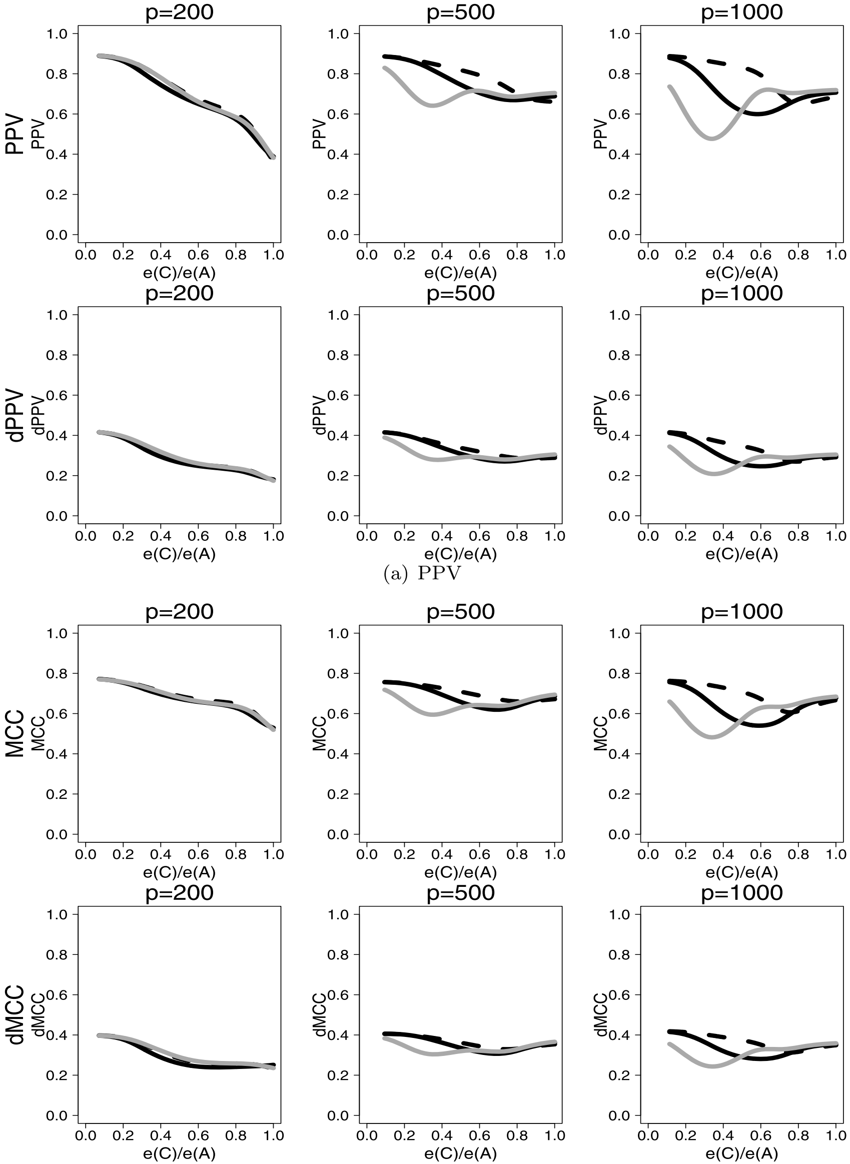
Figure 4. PPV or MCC plots based on p(C)/p(A) when d=2. The solid black line indicates the Walktrap algorithm, and the dashed black line indicates the Fastgreedy algorithm. The solid gray line indicates the Leading eigenvector algorithm.

or 6, and their performances are very similar, and the performance difference is not shown clearly since overall performances of all methods decrease (Appendix C and D in Supplementary Materials).

We also compare computational times when p=1000 under the environment of workstation with E5-2530 v3 CPU and Linux OS with R. The curves of computational times based on p(C)/p(A) are shown in Figure 6. Overall, the computational times of three approaches are very similar, especially if p(C)/p(A) is less than 0.8. If p(C)/p(A) is greater than 0.8, the fast greedy algorithm shows the shortest computational time when d=2, but the leading eigenvector algorithm shows the shortest computational time when d=4 or 6. However, the computational times of all approaches are less than 30 minutes under the entire range of p(C)/p(A), which indicates the computational efficiency for the estimation of networks when p=1000.

6. APPLICATION

In this section, we illustrate how to apply our proposed approach to the gene expression dataset, for example, of breast cancer. It is reported that the incidence of breast cancer is about 1.3 million per year over the world, and the number of deaths is about 450 thousand [12]. Recently, the National Cancer Institute (NCI) and National Human



(a) PPV

(b) MCC

Figure 5. PPV or MCC plots based on e(C)/e(A) when d=2. The solid black line indicates the Walktrap algorithm, and the dashed black line indicates the Fastgreedy algorithm. The solid gray line indicates the Leading eigenvector algorithm.

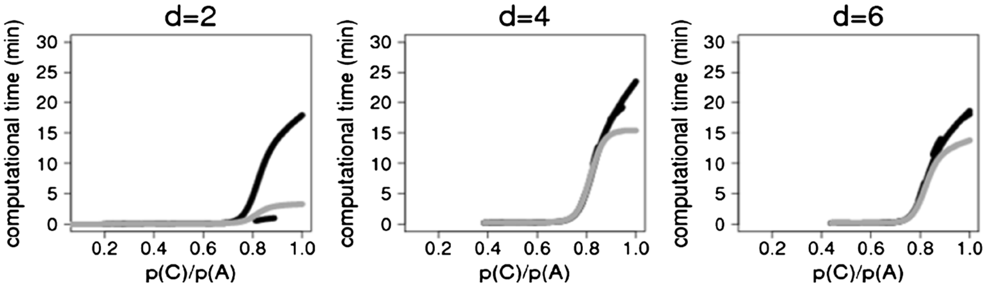


Figure 6. Computational time of the combined steps when p=1000. The solid black line indicates the Walktrap algorithm, and the dashed black line indicates the Fastgreedy algorithm. The solid gray line indicates the Leading eigenvector algorithm.

Table 3. Three most significant gene sets found in each cluster

Cluster number	Source gene	Found gene set names	the number of overlapped genes	p-value	FDR q-value
1	CDKN1B	CREIGHTON ENDOCRINE THERAPY RESISTANCE 2	14	8.14E-12	2.81E-08
		JOHNSTONE PARVB TARGETS 2 DN	9	1.20E-07	0.000156
		MASSARWEH RESPONSE TO ESTRADIOL	5	3.66E-07	0.000316
2	CDH1	CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL UP	24	3.66E-21	1.26E-17
		ROYLANCE BREAST CANCER 16Q COPY NUMBER UP	12	6.46E-18	1.67E-14
		NIKOLSKY BREAST CANCER 16Q24 AMPLICON	8	1.71E-11	1.27E-08
3	RB1	FARMER BREAST CANCER BASAL VS LULMINAL	53	4.15E-64	4.29E-60
		SMID BREAST CANCER BASAL DN	53	2.45E-46	1.27E-42
		VANTVEER BREAST CANCER ESR1 UP	34	7.86E-45	2.71E-41
4	AKT1	GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	40	7.85E-16	1.16E-12
		GRAESSMANN RESPONSE TO MC AND DOXORUBICIN DN	24	9.15E-13	6.31E-10
		ELVIDGE HYPOXIA DN	8	6.88E-07	8.20E-05
5	MLL3	GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	50	1.24E-27	4.27E-24
	PIK3CA	JOHNSTONE PARVB TARGETS 3 DN	38	8.02E-27	2.08E-23
		JOHNSTONE PARVB TARGETS 2 DN	15	1.29E-11	5.12E-09
6	MAP3K1	SMID BREAST CANCER BASAL DN	43	1.23E-31	1.28E-27
		MASSARWEH TAMOXIFEN RESISTANCE DN	18	9.09E-15	1.34E-11
		RAF UP.V1 DN	16	2.13E-14	2.45E-11
7	BRCA1	GRAESSMANN APOPTOSIS BY DOXORUBICIN DN	31	1.27E-12	2.62E-09
		GRAESSMANN RESPONSE TO MC AND DOXORUBICIN DN	15	2.87E-07	9.20E-05
		MCBRYAN PUBERTAL BREAST 5 6WK DN	7	1.00E-06	0.00024

Genome Research Institute (NHGRI) collaborated on the Cancer Genome Atlas (TCGA) project to accumulate gene expression data for breast cancer [12]. The data for breast invasive carcinoma was downloaded from TCGA website (https://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). We applied the proposed method to derive subnetworks from 1,045 primary tumor samples based on mRNA expression data of the level 3 RNA SeqV2 platform. From previous literature, important cancer genes implicated in breast cancer include AKT1 (v-akt murine thymoma viral oncogene homolog 1), BRCA1 (breast cancer 1, early onset), CDH1 (cadherin 1, type 1, E-cadherin (epithelial)), CDKN1B (cyclin-dependent kinase inhibitor 1B (p27, Kip1)), GATA3 (GATA binding protein 3), MAP3K1 (mitogen-activated protein kinase kinase kinase 1), MLL3 (myeloid/lymphoid or mixed-lineage leukemia 3), PIK3CA (phosphoinositide-3-kinase, catalytic, alpha polypeptide), PTEN (phosphatase and tensin homolog), RB1 (retinoblastoma 1), and TP53 (tumor protein p53), which are illustrated in [12] and [62].

In what follows, we aim at finding subnetworks among genes to determine regulatory encoding genes and regulated target encoding genes in each subnetwork. We combine the procedures used in this paper. From the simulation study of subsection 5.2, the Var-BIC-h2 method shows the best performance. In step 1, starting from the important cancer genes mentioned above, we estimate probabilistic neighbors with the Var-BIC-h2 of C_1=1. For the purpose of illustration, we search the neighbors up to the network size of 3,000. In step 2, we detect clusters and estimate subnetworks in each cluster. We use the fast greedy algorithm to identify the clusters. After that, we investigate how many genes in each cluster account for similar genetic functions. To obtain the information of gene functions, we use Gene Set Enrichment Analysis (GSEA, http://www.broadinstitute.org/gsea/index.jsp). Based on the Molecular Signatures Database (MSigDB) in the GSEA software, annotated gene sets are collected. The GSEA software first identifies overlapping between the gene set we provide and the gene set extracted from MSigDB, and then categorizes genes by biological functions. It also measures statistical significance of the overlapping gene sets via p-values based on hypergeometric test [7] and FDR q-values [8, 9]. Table 3 and 4 show the most significant three gene sets found in each cluster related to breast cancer and the detailed descriptions of the sets, respectively.

For the next step, we estimate the directed acyclic graph within the clusters. We compare the estimated edges with known interactions introduced in previous literatures. Such known gene-gene interactions provided in NetBox (http://cbio.mskcc.org/tools/netbox/index.html), which extracts the information from four data sources: NCI-Nature Pathway Interaction Database [58], Human Protein Reference Database [34], MSKCC Cancer Cell Map (http://www.mskcc.org/), and Reactome [33, 45].

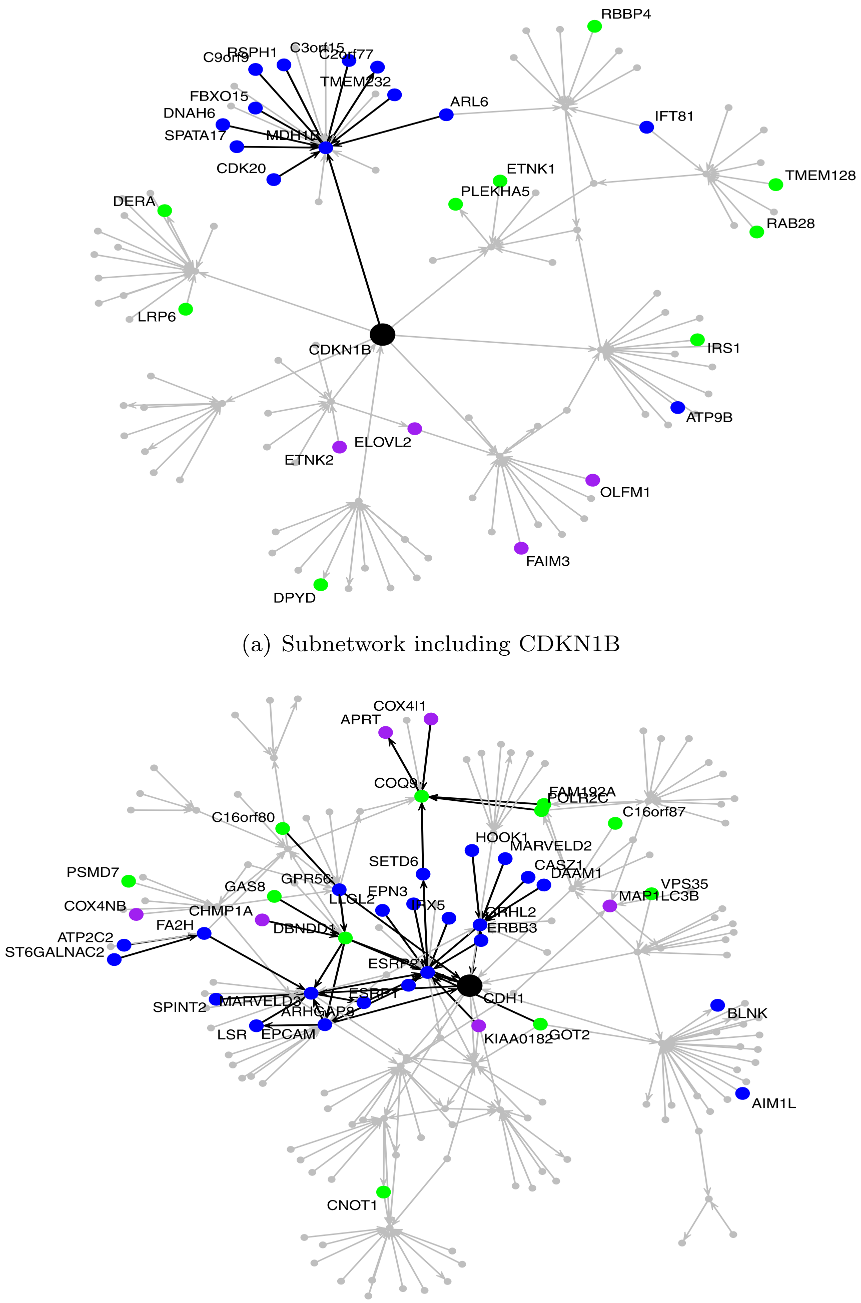
Figures 7, 8, 9, and 10 show the estimated directed networks in each cluster containing biologically important genes. The nodes corresponding to significant gene sets are colored by blue (the first significant gene set), green (the second significant gene set), and purple (the third signif-

Table 4. Explanation of three significant functional sets in each cluster

- Found gene set names in Cluster 1
  - CREIGHTON ENDOCRINE THERAPY RESISTANCE 2: The ‘group 2 set’ of genes associated with acquired endocrine therapy resistance in breast tumors expressing ESR1 and ERBB2 [GeneID=2099;2064].
  - JOHNSTONE PARVB TARGETS 2 DN: Genes down-regulated upon overexpression of PARVB [GeneID=29780] in MDA-MB-231 cells (breast cancer) cultured in 3D collagen I and 3D Matrigel only.
  - MASSARWEH RESPONSE TO ESTRADIOL: Genes rapidly up-regulated in breast cancer cell cultures by estradiol [PubChem=5757].
- Found gene set names in Cluster 2
  - CHARAFE BREAST CANCER LUMINAL VS MESENCHYMAL UP: Genes up-regulated in luminal-like breast cancer cell lines compared to the mesenchymal-like ones.
  - ROYLANCE BREAST CANCER 16Q COPY NUMBER UP: Genes in discrete regions of gain within 16q region detected in individual invasive breast cancer tumors.
  - NIKOLSKY BREAST CANCER 16Q24 AMPLICON: Genes within amplicon 16q24 identified in a copy number alterations study of 191 breast tumor samples.
- Found gene set names in Cluster 3
  - FARMER BREAST CANCER BASAL VS LULMINAL: Genes which best discriminated between two groups of breast cancer according to the status of ESR1 and AR [GeneID=2099;367]: basal (ESR1- AR-) and luminal (ESR1+ AR+).
  - SMID BREAST CANCER BASAL DN: Genes down-regulated in basal subtype of breast cancer samples.
  - VANTVEER BREAST CANCER ESR1 UP: Up-regulated genes from the optimal set of 550 markers discriminating breast cancer samples by ESR1 [GeneID=2099] expression: ER(+) vs ER(-) tumors.
- Found gene set names in Cluster 4
  - GRAESSMANN APOPTOSIS BY DOXORUBICIN DN: Genes down-regulated in ME-A cells (breast cancer) undergoing apoptosis in response to doxorubicin [PubChem=31703].
  - GRAESSMANN RESPONSE TO MC AND DOXORUBICIN DN: Genes down-regulated in ME-A cells (breast cancer; sensitive to apoptotic stimuli) exposed to doxorubicin [PubChem=31703] in the presence of medium concentrate (MC) from ME-C cells (breast cancer; resistant to apoptotic stimuli).
  - ELVIDGE HYPOXIA DN: Genes down-regulated in MCF7 cells (breast cancer) under hypoxia conditions.
- Found gene set names in Cluster 5
  - GRAESSMANN APOPTOSIS BY DOXORUBICIN DN: Genes down-regulated in ME-A cells (breast cancer) undergoing apoptosis in response to doxorubicin [PubChem=31703].
  - JOHNSTONE PARVB TARGETS 3 DN: Genes down-regulated upon overexpression of PARVB [GeneID=29780] in MDA-MB-231 cells (breast cancer) cultured in 3D Matrigel only.
  - JOHNSTONE PARVB TARGETS 2 DN: Genes down-regulated upon overexpression of PARVB [GeneID=29780] in MDA-MB-231 cells (breast cancer) cultured in 3D collagen I and 3D Matrigel only.
- Found gene set names in Cluster 6
  - SMID BREAST CANCER BASAL DN: Genes down-regulated in basal subtype of breast cancer samples.
  - MASSARWEH TAMOXIFEN RESISTANCE DN: Genes down-regulated in breast cancer tumors (formed by MCF-7 xenografts) resistant to tamoxifen [PubChem=5376].
  - RAF UP.V1 DN: Genes down-regulated in MCF-7 cells (breast cancer) positive for ESR1 [Gene ID=2099] MCF-7 cells (breast cancer) stably over-expressing constitutively active RAF1 [Gene ID=5894] gene.
- Found gene set names in Cluster 7
  - GRAESSMANN APOPTOSIS BY DOXORUBICIN DN: Genes down-regulated in ME-A cells (breast cancer) undergoing apoptosis in response to doxorubicin [PubChem=31703].
  - GRAESSMANN RESPONSE TO MC AND DOXORUBICIN DN: Genes down-regulated in ME-A cells (breast cancer; sensitive to apoptotic stimuli) exposed to doxorubicin [PubChem=31703] in the presence of medium concentrate (MC) from ME-C cells (breast cancer; resistant to apoptotic stimuli).
  - MCBRYAN PUBERTAL BREAST 5 6WK DN: Genes down-regulated during pubertal mammary gland development between week 5 and 6.

icant gene set). The estimated edges among nodes from significant gene sets are colored by black, otherwise colored by gray. The overlapping edges between the known interactions and estimated ones are colored by red. Interestingly, several of the edges are consistent with a priori biological knowledge, and so the identified sub-networks efficiently capture biological significance related to breast cancer. BRCA1 and ACACA in Figure 10 are good examples. It is widely known that breast cancer-associated mutations, which are associated with the BRCT domains of the tumor suppressor gene BRCA1, alter the function of BRCA1 to interact with acetyl coenzyme A carboxylase alpha (a.k.a ACACA or ACCA), the rate-limiting enzyme catalyzing de novo fatty acid biogenesis [10]. Precisely, the formation of the BRCA1/P-ACCA (i.e., inactive form of ACCA) complex controls ACCA activity by refraining P-ACCA dephosphorylation. In addition, [48] also experimentally verified that RNA inhibition-mediated down-regulation of BRCA1 expression was molecularly consistent with germ line BRCA1 mutations, while in the absence of the BRCA1ACCA interaction. Importantly note that such underexpression of BRCA1 is sporadically observed in breast and ovarian cancers. Related to CCND1 and RB1 in Figure 8 (a), the previous literature highlighted mutations, amplification and overexpression of this gene, which distort cell cycle progression, contribute to tumorigenesis in a range of cancers [44]. The pathogenic relationship between CCND1 and RB1 has been extensively studied across diverse species. For instance, the presence of down-regulated expression of CCND1 and RB1 appeared in human breast and pancreatic cancers [29, 43] and transgenic mice bladder tumors [24]. Taken together, we believe that our proposed method performs effective network inference in pursuit of detecting true molecular mechanisms.

Taken together, this example of the estimation of subnetworks clearly confirms the biological applicability of the proposed algorithm to real data, and demonstrates how to extract the network information from ultra high dimensional data. We also apply other cluster approaches, walk-



(a) Subnetwork including CDKN1B

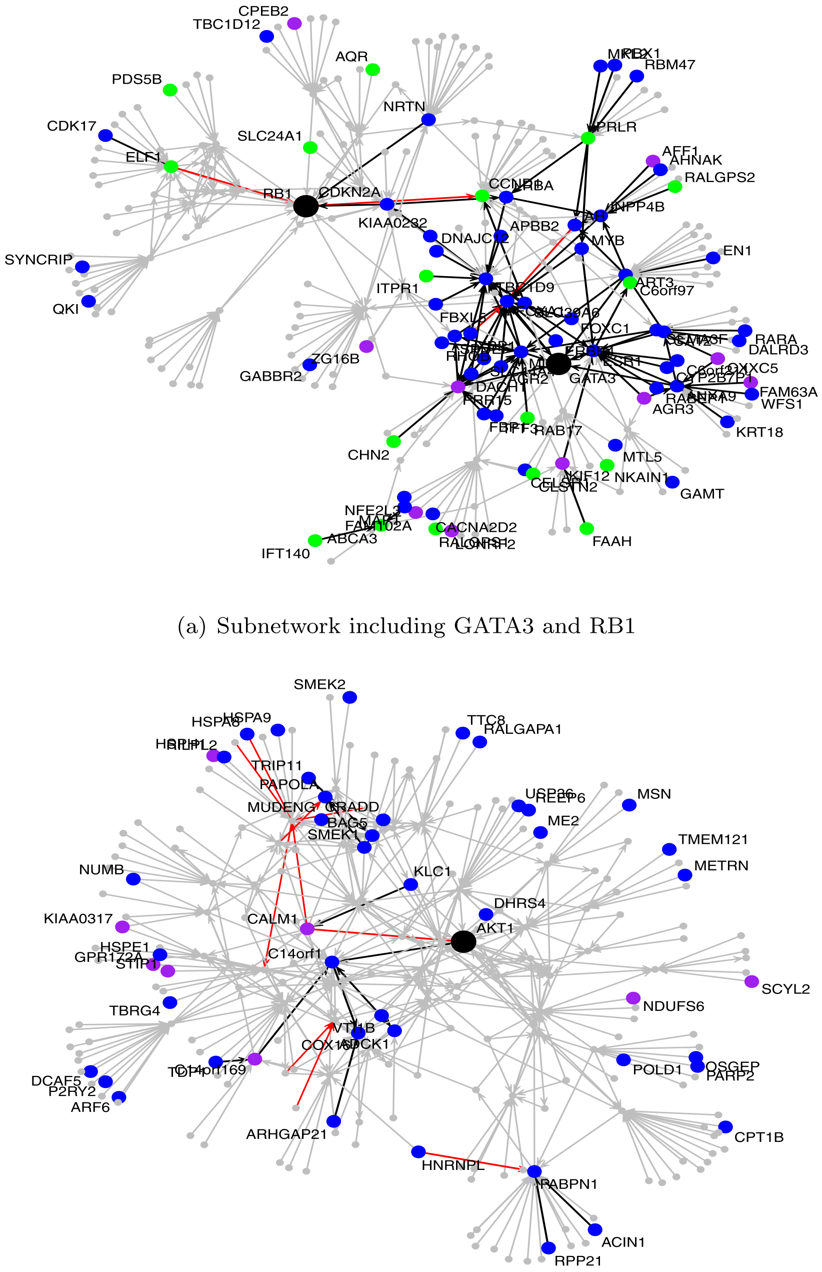
(b) Subnetwork including CDH1

Figure 7. Sub-network plots including CDKN1B or CDH1.

trap and leading eigenvector algorithms, to identify clusters from the probabilistic neighbors estimated by the Var-BIC-h2 of c_1=1. Subnetwork in each cluster by those algorithms are shown in Appendix E and F of Supplementary Materials. Different algorithms identify subnetworks with different size, and many parts of the subnetworks detected by those algorithm are overlapped with the subnetworks identified by the fast greedy algorithm. Computational times for the subnetwork estimation based on three cluster approach are 39.67 minutes (fastgreedy), 39.43 minutes (walktrap), 40.54 minutes (leading eigenvector), respectively.

7. CONCLUSION

In this paper, we discuss two different types of problems to estimate subnetworks in ultra high dimensional data. The first problem is to estimate DAGs of a subnetwork adjacent to a target gene, and the second problem is to estimate



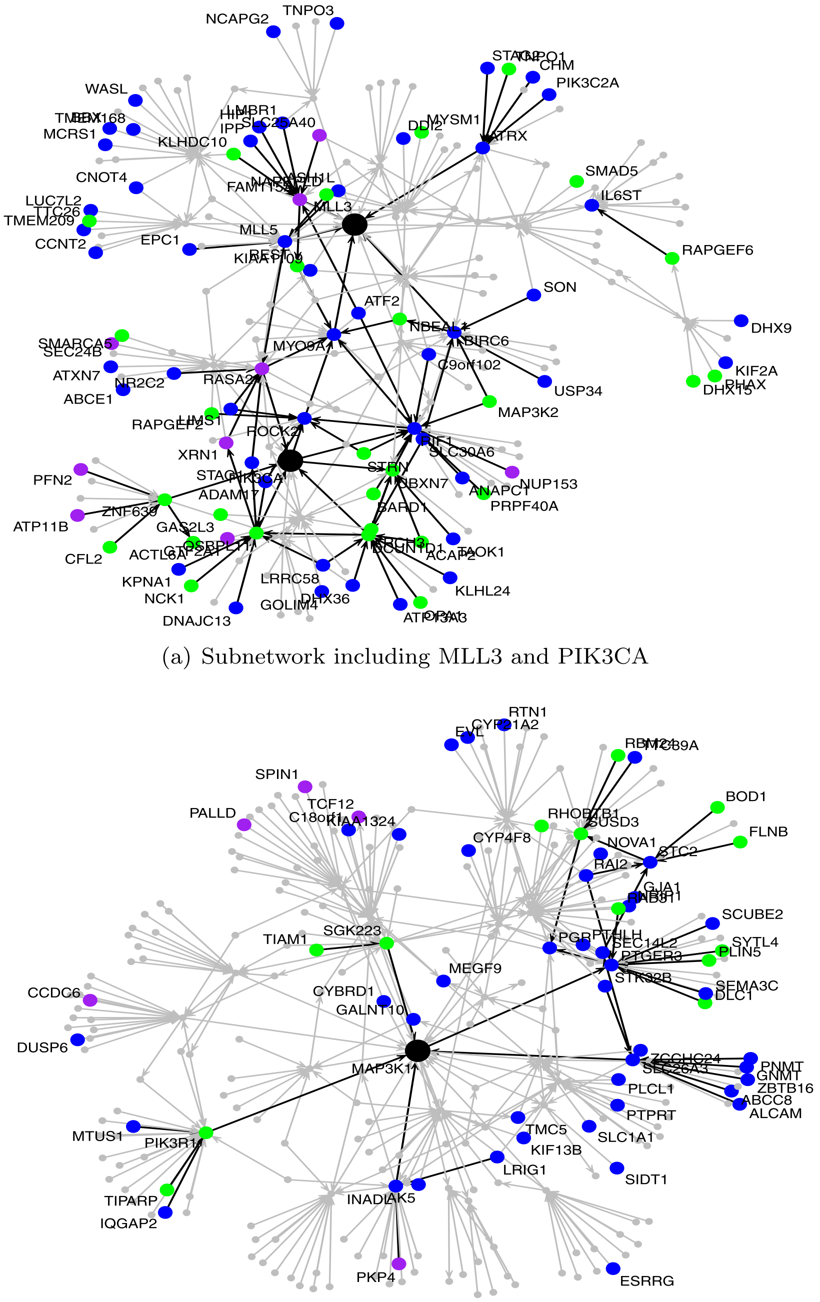
(a) Subnetwork including GATA3 and RB1

(b) Subnetwork including AKT1

Figure 8. Sub-network plots including GATA3 and RB1; or AKT1.

DAGs of multiple subnetworks without information about a target gene. To address each problem, we propose efficient methods to estimate subnetworks, which are to use layer-dependent weights with BIC criteria and to use community detection approaches to identify clusters as subnetworks. In order to estimate a target sub-network, we show that the BIC criteria gives better performance than the technique using the alpha formula, and that the approach with layer dependent weights in BIC outperforms the approach with fixed weights. In addition, to estimate multiple subnetworks without knowing target genes, a certain type of community detection algorithms such as the fast greedy algorithm generally gives a slightly better than and more stable performance than the other approaches.

[61] showed consistency in variable selection to estimate an entire network of the DAG under the L1-likelihood when variable order is known. [22] discussed the consistency in model selection for estimating an entire network



(a) Subnetwork including MLL3 and PIK3CA

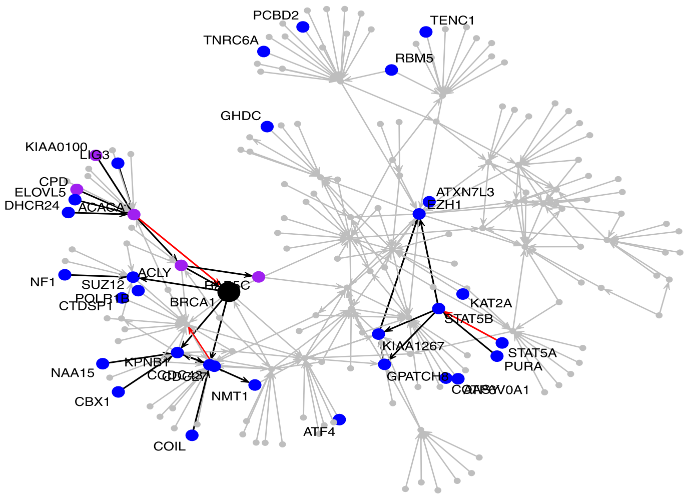
(b) Subnetwork including MAP3K1

Figure 9. Sub-network plots including CDKN1B or CDH1.

of the DAG when the variable order is unknown. BIC criteria in variable selection are known to satisfy the consistency. However, it is not clear whether the BIC with a lasso approach satisfies the consistency in estimating an entire network, even directed subnetworks. Instead, we show the performance of simulations for our proposed method such as MCC or PPV based on the ratio of the number of true edges (or nodes) within clusters to that in an entire graph.

ACKNOWLEDGEMENTS

We would like to give thank to James Wilson for discussing community detection algorithms. Research is supported by a Korea University Grant (K1607901). Research is also supported by NIH-1-R21 GM110450-01. Research is also supported by grants from National Research Foundation of Korea (NRF-2014R1A1A2A16050527, NRF-2016R1D1A1B03931077).



(a) Subnetwork including BRCA1

Figure 10. Sub-network plots including BRCA1.

Please send any inquiry to the corresponding authors, Jeewhan Yoon (Department of Management of Technology, Korea University Graduate School of Management of Technology, 145 Anam-ro, Seongbuk-gu, Seoul, 02841, South Korea. Email: towny@korea.ac.kr), and Hua Zhong (Division of Biostatistics, Department of Population Health, New York University School of Medicine, 650 First Avenue, Room 540, New York, NY 10016. Email: judy.zhong@nyumc.org).

APPENDIX

We provide a brief description of the solution search algorithm for the optimization problem in terms of of a_k or A in Equation (13) with constraints (14) and (15). The solution search algorithm is based on the discrete improving search technique, used in [26]. Equation (13) is convex, but with constraint, the optimization problem is a NP-Hard, mixed integer problem with acyclic constraints. This search algorithm tries to find the sequence of solutions which decrease the objective function values in (13) in addition to satisfying the acyclic constraint (14). There are three steps in the algorithm. The first step is finding an entering edge among unselected edges to improve the objective function in (13). If the entering edge causes cycles, the second step is finding a leaving edge among previously selected edges to eliminate cycles. A leaving edge is selected, which breaks cycles by giving the least increment in the objective function values in (13). The third step is updating the A and T matrix and the objective function value. This search procedure is performed within neighborhood structure matrix, N.

Received 1 December 2015

REFERENCES

[1] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745–6750.

[2] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). The Gene Ontology Consortium Gene ontology: tool for the unification of biology. *Nature genetics* **25**(1), 25–29.

[3] Bader G. D. and Hogue C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 1–17.

[4] Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehar, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jane-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K., Yu, J., Aspesi, P., de Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., MacConaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R., and Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia Enables Predictive Modelling of Anticancer Drug Sensitivity. *Nature* **483**, 603–607.

[5] Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2005). A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. *Bioinformatics* **21**, 349–356.

[6] Bedikian, A. Y., Millward, M., Pehamberger, H., Conry, R., Gore, M., Trefzer, U., Pavlick, A. C., DeConti, R., Hersh, E. M., Hersey, P., Kirkwood, J. M., and Haluska, F. G. (2006). Bcl-2 antisense (oblimersen sodium) plus dacarbazine in patients with advanced melanoma: the Oblimersen Melanoma Study Group. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* **24**, 4738–4745.

[7] Berkopec, A. (2007). HyperQuick algorithm for discrete hypergeometric distribution. *Journal of Discrete Algorithms* **5**, 341–347. MR2316556

[8] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* **57**, 289–300.

[9] Benjamini, Y. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 405–416.

[10] Brunet, J., Vazquez-Martin, A., Colomer, R., Graña-Suarez, B., Martin-Castillo, B., Menendez, J. A. (2008). BRCA1 and Acetyl-CoA Carboxylase: The Metabolic Syndrome of Breast Cancer. *Molecular Carcinogenesis* **47**, 157–163.

[11] Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R., and Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *Proc. Natl. Acad. Sci. USA* **97**, 12182–12186.

[12] The Cancer Genome Atlas Network. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61–70.

[13] Clauset, A., Newman, M. E. J., and Moore, C. (2004). Finding community structure in very large networks. *Physical review E* **70**, 066111.

[14] Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., Viale, G., Delorenzi, M., Zhang, Y., d'Assignies, M. S., Bergh, J., Lidereau, R., Ellis, P., Harris, A. L., Klijn, J. G., Foekens, J. A., Cardoso, F., Piccart, M. J., Buyse, M., and Sotiriou, C. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical cancer research* **13**, 3207–3214.

[15] Diepgen, T. L. and Mahler, V. (2002). The epidemiology of skin cancer. *The British journal of dermatology* **146**(Suppl 61), 1–6.

[16] Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95**, 14863–14868.

[17] Enright, A. J., Van Dongen, S., and Ouzounis C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584.

[18] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. Ann. Statist., **32**, 928–961. MR2065194

[19] Fortunato, S. (2010). Community detection in graphs. *Physics Reports* **486**, 75–174.

[20] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of Computational Biology* **7**, 601–620.

[21] Freeman, T. C., Goldovsky, L., Brosch, M., van Dongen, S., Mazière, P., Grocock, R. J., Freilich, S., Thornton, J., and Enright, A. J. (2007). Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Comput Biol.* **3**, e206.

[22] Fu, F. and Zhou, Q. (2013). Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of the American Statistical Association* **108**, 288–300.

[23] Garbe, C., McLeod, G. R., and Buettner, P. G. (2000). Time trends of cutaneous melanoma in Queensland, Australia and Central Europe. *Cancer* **89**, 1269–1278.

[24] Garcia-España, A., Salazar, E., Sun, T.-T., Wu, X.-R., and Pellicer, A. (2005). Differential expression of cell cycle regulators in phenotypic variants of transgenically induced bladder tumors: implications for tumor behavior. *Cancer Research* **65**, 1150–1157.

[25] Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M., and Barabàsi, A.-L. (2007). The human disease network. *PNAS* **104**, 8685–8690.

[26] Han, S. W., Chen, G., Cheon, M.-S., and Zhong, H. (2016). Estimation of Directed Acyclic Graphs Through Two-stage Adaptive Lasso for Gene Network Inference. Journal of the American Statistical Association, doi:10.1080/01621459.2016.1142880. MR3561925

[27] Hanisch, D., Zien, A., Zimmer, R., and Lengauer, T. (2002). Co-clustering of biological networks and gene expression data. *Bioinformatics* **18**(Suppl 1), S145–S154.

[28] Higham, D., Kalna, G., and Kibble, M. (2007). Spectral clustering and its use in bioinformatics. *Journal of computational and applied mathematics* **204**(1), 25–37.

[29] Huang, L., Lang, D., Geradts, J., Obara, T., Klein-Szanto, A. J. P., Lynch H. T., and Ruggeri, B. A. (1996). Molecular and immunochemical analyses of RB1 and cyclin D1 in human ductal pancreatic carcinomas and cell lines. *Molecular Carcinogenesis* **15**, 85–95.

[30] Huttenhower, C., Flamholz, A. I., Landis, J. N., Sahi, S., Myers, C. L., Olszewski, K. L., Hibbs, M. A., Siemers, N. O., Troyanskaya, O. G., and Coller, H. A. (2007). Nearest neighbor networks: clustering expression data based on gene neighborhoods. *BMC Bioinformatics* **8**, 250 doi:10.1186/1471-2105-8-250.

[31] Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., Smeds, J., Lindahl, T., Pawitan, Y., Hall, P., Nordgren, H., Wong, J. E., Liu, E. T., Bergh, J., Kuznetsov, V. A., and Miller, L. D. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer research* **66**(21), 10292–10301.

[32] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F., and Gerstein, M. (2003). A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. *Science* **302**, 449–453.

[33] Joshi-Tope, G., Gillespie, M., Vastrik, I., D'Eustachio, P., Schmidt, E., de Bono, B., Jassal, B., Gopinath, G. R., Wu, G. R., Matthews, L., Lewis, S., Birney, E., and Stein, L. (2005). Reactome: A knowledgebase of biological pathways. *Nucleic acids research* **33**, 428–432.

[34] Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., and Pandey, A. (2009). Human protein reference database–2009 update. *Nucleic acids research* **37**(Database issue), 767–772.

[35] Kluger, Y., Basri, R., Chang, J., and Gerstein, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research* **13**(4), 703–716.

[36] Kunz, M. (2013). MicroRNAs in Melanoma Biology. *MicroRNA Cancer Regulation, Advances in Experimental Medicine and Biology*, Schmitz U., Wolkenhauer O., Vera J. (eds). **774**, 103–120.

[37] Samuel Lattimore, B., van Dongen, S., and Crabbe, M. J. (2005). GeneMCL in microarray analysis. *Computational biology and chemistry* **29**(5), 354–359.

[38] Lauritzen, S. L. (1996). *Graphical Models*. Oxford: Clarendon Press.

[39] Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica* **16**, 1273–1284.

[40] Li, Y., Zou, L., Li, Q., Haibe-Kains, B., Tian, R., Li, Y., Desmedt, C., Sotiriou, C., Szallasi, Z., Iglehart, J. D., Richardson, A. L., and Wang, Z. C. (2010). Amplification of LAPTM4B and YWHAZ contributes to chemotherapy resistance and recurrence of breast cancer. *Nature Medicine* **16**, 214–218.

[41] Liu, Y., Gu, Q., Hou, J. P., Han, J., and Ma, J. (2014). A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics* **15**, 37.

[42] Lukashin, A. V. and Fuchs, R. (2001). Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters. *Bioinformatics* **17**, 405–414.

[43] Lung, J.-C., Chu, J.-S., Yu, J.-C., Yue, C.-T., Lo, Y.-L., Shen, C.-Y., and Wu, C.-W. (2002). Aberrant expression of cell-cycle regulator cyclin D1 in breast cancer is related to chromosomal genomic instability. *Genes, Chromosomes and Cancer* **34**, 276–284.

[44] Mao, X., Orchard, G., Vonderheid, E. C., Nowell, P. C., Bagot, M., Bensussan, A., Russell-Jones, R., Young, B. D., and Whittaker, S. J. (2006). Heterogeneous abnormalities of CCND1 and RB1 in primary cutaneous T-Cell lymphomas suggesting impaired cell cycle control in disease pathogenesis. *Journal of Investigative Dermatology* **126**, 1388–1395.

[45] Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., Kanapin, A., Lewis, S., Mahajan, S., May, B., Schmidt, E., Vastrik, I., Wu, G., Birney, E., Stein, L., and D'Eustachio, P. (2009). Reactome knowledgebase of human biological pathways and processes. *Nucleic acids research* **37**(Database issue), 619–622.

[46] Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics* **34**, 1436–1462.

[47] Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* **52**(1), 91–118.

[48] Moreau, K., Dizin, E., Ray, H., Luquain, C., Lefai, E., Foufelle, F., Billaud, M., Lenoir, G. M., and Venezia, N. D. (2006). BRCA1 affects lipid synthesis through its interaction with acetyl-CoA carboxylase. *The Journal of Biological Chemistry* **281**, 3172–3181.

[49] National Cancer Institute (2013). SEER Stat Fact Sheets: Melanoma of the Skin. URL http://seer.cancer.gov/statfacts/html/melan.html. Accessed 1/29/13.

[50] Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review* **45**, 167–256.

[51] Newman, M. E. J. (2006). Finding community structure in networks using the eigenvectors of matrices. *Physical Review E* **74**, 036104.

[52] Newman, M. E. J. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E* **69**, 026113.

[53] Oldham, M., Horvath, S., and Geschwind, D. (2006). Conservation and evolution of gene coexpression networks in human and chimpanzee brains. *Proc. Natl. Acad. Sci. USA*, 17973–17978.

[54] Palla, G., Derenyi, I., Farkas, I., and Vicsek, T. (2005). Uncovering the over-lapping community structure of complex networks in nature and society. *Nature* **435**, 814–818.

[55] Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*, Cambridge: Cambridge University Press.

[56] Pons, P. and Latapy, M. (2006). Computing Communities in Large Networks Using Random Walks. *Journal of Graph Algorithms and Applications* **10**, 191–218.

[57] Porter, M. A., Onnela, J.-P., and Mucha, P. J. (2009). Communities in networks. *Notices of the AMS* **56**, 1082–1097.

[58] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K. H. (2009). PID: The pathway interaction database. *Nucleic acids research* **37**(Database issue), 674–679.

[59] Shao, J. (1997). An asymptotic theory for linear model selection. *Statist. Sin.* **7**, 221–264.

[60] Shi, P. and Tsai, C. L. (2002) Regression model selection – a residual likelihood approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 237–252.

[61] Shojaie, A. and Michailidis, G. (2010). Penalized likelihood methods for estimation of sparse high-dimensional directed acyclic graphs. *Biometrika* **97**, 519–538. MR2672481

[62] Stephens, P. J., Tarpey, P. S., Davies, H., Loo, P. V., Greenman, C., Wedge, D. C., Nik-Zainal, S., Martin, S., Varela, I., Bignell, G. R., Yates, L. R., Papaemmanuil, E., Beare, D., Butler, A., Cheverton, A., Gamble, J., Hinton, J., Jia, M., Jayakumar, A., Jones, D., Latimer, C., Lau, K. W., McLaren, S., McBride, D. J., Menzies, A., Mudie, L., Raine, K., Rad, R., Chapman, M. S., Teague, J., Easton, D., Langerød, A., The Oslo Breast Cancer Consortium (OSBREAC), Lee, M. T. M., Shen, C.-Y., Tee, B. T.K, Huimin, B. W., Broeks, A., Vargas, A. C., Turashvili, G., Martens, J., Fatima, A., Miron, P., Chin, S.-F., Thomas, G., Boyault, S., Mariani, O., Lakhani, S. R., van de Vijver, M., van 't Veer, L., Foekens, J., Desmedt, C., Sotiriou, C., Tutt, A., Caldas, C., Reis-Filho, J. S., Aparicio, S. A. J. R., Salomon, A. V., Børresen-Dale, A.-L., Richardson, A. L., Campbell, P. J., Futreal, P. A., and Stratton, M. R. (2012). The landscape of cancer genes and mutational processes in breast cancer. *Nature* **486**, 400–404.

[63] Steuer, R. (2006). On the analysis and interpretation of correlations in metabolomic data. *Brief Bioinform* **151**, 151–158.

[64] Stone, E. A. and Ayroles, J. F. (2009). Modulated Modularity Clustering as an Exploratory Tool for Functional Genomic Inference. *PLoS Genetics* **5**(5), e1000479. doi:10.1371/journal.pgen.1000479.

[65] Sun, P. G., Gao, L., and Han, S. (2011). Prediction of Human Disease-Related Gene Clusters by Clustering Analysis. *International Journal of Biological Sciences* **7**(1), 61–73.

[66] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA* **96**, 2907–2912.

[67] Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

[68] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**, 267–288.

[69] Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999). Analysis of gene expression data using self-organizing maps. *FEBS Letters* **451**, 142–146.

[70] Tritchler, D., Fallah, S., and Beyene, J. (2005). A Spectral clustering method for microarray data. *Computational Statistics and Data Analysis* **49**(1), 63–76.

[71] Vassilev, L. T., Vu, B. T., Graves, B., Carvajal, D., Podlaski, F., Filipovic, Z., Kong, N., Kammlott, U., Lukacs, C., Klein, C., Fotouhi, N., and Liu, E. A. (2004). In vivo activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848.

[72] 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 (Statistical Methodology)* **71**, 671–683.

[73] Wen, X., Fuhrman, S., Michaels, G. S., Carr, D. B., Smith, S., Barker, J. L., and Somogyi, R. (1998). Large-scale temporal gene expression mapping of central nervous system development. *PNAS* **95**, 334–339.

[74] Werhli, A. V., Grzegorczyk, M., and Husmeier, D. (2006). Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, **22**, 2523–2531.

[75] Wilkerson, M. D. and Hayes, D. N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics* **26**(12), 1572–1573.

[76] Zou, M. and Conzen, S. D. (2005). A new dynamic Bayesian network (dbn) approach for identifying gene regulatory networks from time course microarray data. *Bioinformatics*, **21**, 71–79.

[77] Zou, H. (2006). The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Sung Won Han
School of Industrial Management Engineering
Korea University
145 Anam-Ro, Seongbuk-Gu
Seoul 02841
Republic of Korea
E-mail address: swhan@korea.ac.kr

SungHwan Kim
Department of Statistics
Keimyung University
Daegu
South Korea
E-mail address: swiss747@gmail.com

Junhee Seok
School of Electrical Engineering
Korea University
145 Anam-ro, Seongbuk-gu
Seoul 136-713
South Korea
E-mail address: jseok14@korea.ac.kr

Jeewhan Yoon
Graduate School of Management of Technology
Korea University
145 Anam-ro, Seongbuk-gu
Seoul 136-713
South Korea
E-mail address: towny@korea.ac.kr

Hua Zhong
Division of Biostatistics
Department of Population Health
New York University School of Medicine
650 First Avenue, Room 540
New York, NY 10016
USA
E-mail address: judy.zhong@nyumc.org