Biometrics 63, 1–??
March 2014
DOI: 10.1111/j.1541-0420.2005.00454.x

# PenPC: A Two-step Approach to Estimate the Skeletons of High Dimensional Directed Acyclic Graphs

**Min Jin Ha**

Department of Biostatistics, MD Anderson Cancer Center, Houston, Texas

*email: MJHa@mdanderson.org

**and**

**Wei Sun**

Department of Biostatistics, Department of Genetics, UNC Chapel Hill, North Carolina

*email: weisun@email.unc.edu

**and**

**Jichun Xie**

Department of Biostatistics & Bioinformatics, Duke University, Durham, North Carolina

*email: jichun.xie@duke.edu

Summary: Estimation of the skeleton of a directed acyclic graph (DAG) is of great importance for understanding the underlying DAG and causal effects can be assessed from the skeleton when the DAG is not identifiable. We propose a novel method named PenPC to estimate the skeleton of a high-dimensional DAG by a two-step approach. We first estimate the non-zero entries of a concentration matrix using penalized regression, and then fix the difference between the concentration matrix and the skeleton by evaluating a set of conditional independence hypotheses. For high dimensional problems where the number of vertices $p$ is in polynomial or exponential scale of sample size $n$, we study the asymptotic property of PenPC on two types of graphs: traditional random graphs where all the vertices have the same expected number of neighbors, and scale-free graphs where a few vertices may have a large number of neighbors. As illustrated by extensive simulations and applications on gene expression data of cancer patients, PenPC has higher sensitivity and specificity than the state-of-the-art method, the PC-stable algorithm.

Key words: DAG, Penalized regression, log penalty, PC-algorithm, skeleton, high dimensional

This paper has been submitted for consideration for publication in *Biometrics*

## 1. Introduction

Many statistical methods have been developed to identify the associations between genomic features and disease outcomes or cancer subtypes. However, such association results are descriptive in their nature, and they cannot deliver "actionable" conclusions for disease treatment. Many recently developed cancer drugs are so-called "targeted drugs" that target particular (mutated) proteins in cancer cells, and the mechanism of such drugs can be understood as direct interventions on tumor cells. To characterize or predict the consequences such drug interventions, statical methods that allow causal inference based on high dimensional genomic data are urgently needed.

One of the most commonly used tools for causal inference among a large number of random variables is the probabilistic directed acyclic graph (DAG) (also known as Bayesian Network) (Lauritzen, 1996; Pearl, 2009). In a DAG, all the edges are directed, and the direction of an edge implies a direct causal relation. There is no loop in a DAG. Such "acyclic" property is necessary to study causal relations (Spirtes et al., 2000). When we remove the directions of all the edges in a DAG, the resulting undirected graph is the *skeleton* of the DAG.

Estimation of the skeleton of a DAG is of great importance because it is a crucial step towards estimating the underlying DAG and skeleton itself may provide a limited amount of information for causal inference (Maathuis et al., 2009, 2010). Several methods have been developed to estimate DAGs or their skeletons from observational data (Heckerman et al., 1995; Spirtes et al., 2000; Chickering, 2003; Kalisch and Bühlmann, 2007), however most of them are not suitable for the high dimensional genomic problems that motivate our study. In this paper, we proposed a new method named `PenPC` to address this challenging problem. We proved the estimation consistency of `PenPC` for high dimensional settings of $p = O\left(\exp\{n^a\}\right)$ for $0 \leqslant a < 1$, and we also derived the conditions for estimation consistency for two types of graphs: random graphs where all the vertices have the same expected number of neighbors,

and scale-free graphs where a few vertices have much larger number of neighbors than other vertices. As verified by both simulation and real data analyses, `PenPC` provides more accurate estimates of DAG skeletons than existing methods. In addition to skeleton, `PenPC` can further estimate the complete partially directed acyclic graph (CPDAG), which can used to estimate causal effects (Maathuis et al., 2009).

The remaining parts of this paper are organized as follows. In Section 2, we give a brief review of DAG estimation methods and the conceptual advantages of our `PenPC` algorithm. We present the details of the `PenPC` algorithm and its theoretical properties in Sections 3 and 4, followed by simulations and real data analyses in Section 5 and Section 6, respectively. Finally, we conclude with some discussions in Section 7.

## 2. Review of DAG Estimation

### 2.1 *Directed Acyclic Graph (DAG)*

A DAG of random variables $X_1, ..., X_p$ can be denoted by $\mathcal{G} = (V, E)$, where $V$ contains $p$ vertices $1, 2, ...., p$ that correspond to $X_1, ..., X_p$, and $E$ contains all the directed edges. In a DAG, a *chain* of length $n$ from $i$ to $j$ is a sequence $i = i_0 - i_1 - \cdots - i_{n-1} - i_n = j$ of distinct vertices such that $i_{l-1} \to i_l \in E$ or $i_l \to i_{l-1} \in E$ for $l = 1, \ldots, n$; and a *path* of length $n$ from $i$ to $j$ is a sequence $i = i_0 \to i_1 \to \cdots \to i_n = j$ of distinct vertices such that $i_{l-1} \to i_l \in E$ for $l = 1, ..., n$. Given this path, $i_{l-1}$ is a *parent* of $i_l$, $i_l$ is a *child* of $i_{l-1}$, $i_0, i_1, ..., i_{l-1}$ are *ancestors* of $i_l$, and $i_{l+1}, ..., i_n$ are *descendants* of $i_l$.

Given a DAG $\mathcal{G}$ for random variables $X_1, \ldots, X_p$ and assume that $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \in \mathbb{R}^p \sim P_{\boldsymbol{X}}$ with density $f_{\boldsymbol{X}}$. Let $X_{\mathtt{pa}_i}$ be the parents of $X_i$. We say that the distribution $P_{\boldsymbol{X}}$ is *Markov* to $\mathcal{G}$ if the joint density $f_{\boldsymbol{X}}$ satisfies the *recursive factorization*: $f(X_1, \ldots, X_p) = \prod_{i=1}^{p} f(X_i | X_{\mathtt{pa}_i})$. The factorization naturally implies acyclic restriction of the graph struc-

ture. Equivalently $P_{\boldsymbol{X}}$ is Markov to $\mathcal{G}$ if every variable is conditionally independent of its non-descendants given its parents. A related concept is the so-called *faithfulness*:

DEFINITION 1: Let $P_{\boldsymbol{X}}$ be Markov to $\mathcal{G}$. $< \mathcal{G}, P_{\boldsymbol{X}} >$ satisfies the faithfulness condtiontion if and only if every conditional independence relation true in $P_{\boldsymbol{X}}$ is entailed by the Markov property applied to $\mathcal{G}$ (Spirtes et al., 2000).

This means that if a distribution $P_{\boldsymbol{X}}$ is faithful to a DAG $\mathcal{G}$, all conditional independences can be read off from the DAG $\mathcal{G}$ using d-separation defined in the following definition 2, and thus the faithfulness assumption requires stronger relationship between the distribution $P_{\boldsymbol{X}}$ and the DAG $\mathcal{G}$ than the Markov property.

DEFINITION 2: (d-separation). A vertex set $\boldsymbol{S}$ block a chain $\mathtt{p}$ if either (i) $\mathtt{p}$ contains at least one arrow-emitting vertex belonging to $\boldsymbol{S}$, or (ii) $\mathtt{p}$ contains at least one collision vertex ($j$ is a collision vertex if the chain includes $i \to j \leftarrow k$) that is outside $\boldsymbol{S}$ and no descendant of the collision vertex belongs to $\boldsymbol{S}$. If $\boldsymbol{S}$ blocks all the chains between two sets of random variables $X$ and $Y$, we say "$\boldsymbol{S}$ d-separates $X$ and $Y$" (Pearl, 2009).

Not all the distributions can be faithfully represented by a DAG. In this paper, we assume the random variables follow multivariate Gaussian distribution, then the faithfulness assumption can be justified by the fact that among all the multivariate Gaussian distributions associated with $\mathcal{G}$, the non-faithful ones form a Lebesgue null set (Meek, 1995).

Given multivariate Gaussian distribution assumption, a commonly used graphical model is Gaussian Graphic Model (GGM), where two vertices are connected if the corresponding two variables are dependent, given all the other variables. A GGM can be constructed by a *concentration matrix* (i.e., precision matrix or inverse of covariance matrix) in that two vertices are connected if the corresponding elements in the concentration matrix is non-zero. The skeleton of a DAG is different from its GGM because of v-structures. In a *v-structure*

$X \to W \leftarrow Z$, *co-parent* $X$ and $Z$ are marginally independent or conditionally independent given their parents, but given every vertex set that contains $W$ (a collision vertex) or any descendant of $W$, $X$ and $Z$ are dependent with each other. Note that by the definition of v-structure, the co-parents $X$ and $Z$ are not connected. A few examples are shown in Figure 1, and instances of the covariance and concentration matrices of the GGM in Figure 1(a) are shown in the Supplementary Materials, Section 1.

[Figure 1 about here.]

### 2.2 *DAG estimation using observational data*

In this paper we focus on DAG skeleton estimation using observational data instead of interventional data. When the $p$ variables have a nature ordering (i.e., all the parents or ancestors of $X_i$ are among the vertices $X_1$, ..., $X_{i-1}$, and all the children or descendants of $X_i$ are among vertices $X_{i+1}$, ..., $X_p$), the problem of skeleton estimation is greatly simplified because a regression of $X_i$ versus $X_1$, ..., $X_{i-1}$ can be used to identify the true skeleton (Shojaie and Michailidis, 2010). However, in many high-dimensional problems, such a nature ordering is not available. Throughout this paper, we assume no knowledge of nature ordering. Then the underlying DAG is not identifiable from observational data, because conditional dependencies implied by the Markov property on the observational distribution $P_{\boldsymbol{X}}$ only determine the *skeleton* and *v-structures* of the graph (Pearl, 2009). All the DAGs with the same skeleton and v-structures correspond to the same probability distribution and they form a *Markov equivalence class*. After estimating skeleton, the v-structures can be identified by a set of deterministic rules, and thus we do not distinguish the estimation of a DAG skeleton and a Markov equivalence class.

In general, there are two approaches for DAG or DAG skeleton estimation. The first one is the search-and-score approach that searches for the DAG that maximizes or minimizes a pre-defined score, such as BIC (Bayesian Information Criterion) or $L_0$-penalized maximum

likelihood estimates (van de Geer and Bühlmann, 2013). Instead of searching across all the DAGs, which is often computationally infeasible, elegant methods have been developed to search across Markov equivalence classes (Chickering, 2003) or the nature orderings of the variables (Teyssier and Koller, 2005). However, these methods are still computationally very challenging for genomic applications with thousands of variables.

The second approach for DAG (skeleton) estimation is constraint-based approach that constructs DAGs by assessing conditional independence of random variables. One representative method is the PC algorithm (named after its authors, Peter Sprites and Clark Glymour) (Spirtes et al., 2000). Starting with a complete undirected graph where any two vertices are connected with each other, the PC algorithm first thins the complete graph by removing edges between vertices that are marginally independent. Then it removes edges by assessing conditional independence given one vertex, two vertices, and so on. Kalisch and Bühlmann (2007) proved the consistency of the PC-algorithm in high-dimensional settings where $p = O(n^a)$ for $a > 0$. The results of the PC algorithm depend on the order of the edges to be assessed. Colombo and Maathuis (2012) proposed PC-stable algorithm, which modified the PC algorithm to remove such order dependency and substantially improve the performance of the PC algorithm. We consider the PC-stable algorithm as the state-of-the-art method and compare our method with the PC-stable algorithm.

The Independence Graph (IG) algorithm (Chapter 5.4.3 of Spirtes et al. (2000)) modifies the PC algorithm by using a different initial graph: an undirected independence graph where two vertices are connected if the corresponding two variables are conditionally dependent given all the other variables, i.e., a GGM under multivariate Gaussian distribution assumption. In such an independence graph, the neighbors of a vertex $Y_j$ include its parents, children, and co-parents of v-structures in the underlying DAG, which constitute the so-called *Markov blanket* of $Y_j$ such that $Y_j$ is independent of all the other vertices given its Markov blanket.

The Max-Min Hill-Climbing (MMHC) algorithm is a popular hybrid method that first estimates DAG skeleton using a constraint-based method (the Max-Min part of the algorithm), and then orient the edges using a search-and-score technique (the Hill-Climbing part of the algorithm) (Tsamardinos et al., 2006). Schmidt et al. (2007) proposed to replace the Max-Min part of the MMHC algorithm by a penalized regression with Lasso ($l_1$) penalty, i.e., neighborhood selection (Meinshausen and Bühlmann, 2006). Variable selection consistency of Lasso requires the irrepresentable condition (Zhao and Yu, 2006): there is weak correlation between the variables within and outside a Markov blanket. This is a strong condition and it generally does not hold for the genomic problems that motivate this study.

We propose a `PenPC` algorithm for DAG skeleton estimation in two steps. It first adapts neighborhood selection method to select Markov blanket of each vertex, and then it applies a modified PC-stable algorithm to remove false positive edges between co-parents of v-structures. Although the two-step approach of the `PenPC` algorithm shares similar spirit to the IG algorithm (Spirtes et al., 2000) and the modified MMHC algorithm (Schmidt et al., 2007), we have made the following novel contributions. First, we employ the log penalty $p(|b|; \lambda, \tau) = \lambda \log(|b|+\tau)$ (Mazumder et al., 2011) for neighborhood selection, which significantly improves the accuracy of Markov blanket search for higher dimensional problems, e.g., $n = 30$ and $p = 100$, or $n = 300$ and $p = 1000$. In contrast, Schmidt et al. (2007) explicitly assume $n \gg p$ in their paper. The resulting `PenPC` algorithm outperforms the state-of-the-art PC-stable algorithm and also enjoys some advantage in terms of computational efficiency in high dimensional settings. Second, we provide theoretical justifications of the estimation consistency of the `PenPC` algorithm in high dimensional settings where $p = O\left(\exp\{n^a\}\right)$ for $0 \leqslant a < 1$. We also discuss the implications for estimation consistency for two types of graphs: traditional random graph where all the vertexes have the same expected number of connections, and scale-free graph where a few vertices can have much larger number of

neighbors than the other vertices. Whereas non-scale-free graph is often assumed in previous studies (Kalisch and Bühlmann, 2007), scale-free graph is more frequently observed in gene networks as well as many other applications (Barabási and Albert, 1999).

## 3. Methods

We adopt a multivariate Gaussian distribution assumption: $\boldsymbol{X} = (X_1, \ldots, X_p)^{\mathrm{T}} \sim N(0, \Sigma)$. Let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_p)$ be the $n \times p$ observed data matrix. Our `PenPC` algorithm proceeds in two steps: (1) neighborhood selection, and (2) application of a modified PC-stable algorithm to remove false connections. Theoretical justification of our algorithm is presented in Section 4.

**Step 1. (Neighborhood Selection)** We first select the neighborhood of vertex $i$ by a penalized regression with $X_i$ as response variable and all the other variables corresponding to vertices $V \setminus \{i\}$ as covariates:

$$\hat{\mathbf{b}}_i = \underset{\mathbf{b}_i \in \mathbb{R}^{p-1}}{\arg\min} \frac{1}{2}(\mathbf{x}_i - \mathbf{X}_{-i}\mathbf{b}_i)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{X}_{-i}\mathbf{b}_i) + n\sum_{j \neq i} p(|b_{i,j}|; \boldsymbol{\varpi}_i), \tag{1}$$

where $\mathbf{X}_{-i}$ is an $n \times (p-1)$ matrix for $n$ measurements of the remaining $p-1$ covariates, $\mathbf{b}_i = (b_{i,1}, ..., b_{i,i-1}, b_{i,i+1}, ..., b_{i,p})^{\mathrm{T}}$, and $p(|b_{i,j}|; \boldsymbol{\varpi}_i)$ denotes a penalty function with one or more tuning parameters, denoted by $\boldsymbol{\varpi}_i$. We consider a class of folded concave penalty functions satisfying the following condition:

**Condition 1:** The penalty function $p(\beta; \boldsymbol{\varpi})$ is concave in $\beta \in [0, \infty)$, with continuous derivative $p'(\beta; \boldsymbol{\varpi}) \geqslant 0$, and $p'(0+; \boldsymbol{\varpi}) > 0$.

This is a generalization of the Condition 1 in Fan and Lv (2011). In this study, we employ the log penalty $p(|b|; \lambda, \tau) = \lambda \log(|b| + \tau)$, which has been demonstrated to have good performance in high-dimensional genetic studies (Sun et al., 2010). We solve penalized regression with log penalty using a coordinate descent algorithm (Sun et al., 2010), and the two tuning parameters $\lambda$ and $\tau$ are selected by two-grid search to minimize extended

BIC (Chen and Chen, 2008). After $p$ penalized regressions for each of the $p$ variables, we construct the GGM by adding an edge between vertices $i$ and $j$ if $\hat{b}_{ij} \neq 0$ or $\hat{b}_{ji} \neq 0$.

**Step 2. (Modified PC-stable algorithm)** We apply a modified PC-stable algorithm to remove the false edges between parents of v-structures. For each edge $i - j$, we first assess marginal association between vertices $i$ and $j$. If they remain dependent, we test whether they are conditionally dependent. The conditional set should be selected from the Markov blanket of $i$ and $j$, after excluding $i$ and $j$'s common children or descendants. Specifically. we use the following strategy to search for candidate separation sets. Let $\boldsymbol{A}_{i,j}$ be the Markov blanket of $i$ and $j$, and let $\boldsymbol{C}_{i,j}$ be the set of vertices that could be common children or descendants of $i$ and $j$. Then the candidate conditional sets are

$$\boldsymbol{\Pi}_{i,j} = \{\boldsymbol{A}_{i,j} \setminus \boldsymbol{D}_{i,j}, \boldsymbol{D}_{i,j} \subseteq \boldsymbol{C}_{i,j}\}. \tag{2}$$

Each element of $\boldsymbol{\Pi}_{i,j}$ is a set $\boldsymbol{A}_{i,j} \setminus \boldsymbol{D}_{i,j}$, where $\boldsymbol{D}_{i,j}$ is exhaustively searched across all subsets of $\boldsymbol{C}_{i,j}$. More details are described in the Supplementary Materials, Section 2.

We test the conditional independence of $X_i$ and $X_j$ given $\mathcal{K} \in \boldsymbol{\Pi}_{i,j}$ using Fisher transformation of partial correlation. Specifically, denote the partial correlation between $X_i$ and $X_j$ given $\mathcal{K} \in \boldsymbol{\Pi}_{i,j}$ by $\rho_{i,j|\mathcal{K}}$. With the significance level $\alpha$, we reject the null hypothesis $H_0 : \rho_{i,j|\mathcal{K}} = 0$ against the alternative hypothesis $H_a : \rho_{i,j|\mathcal{K}} \neq 0$ if $\sqrt{n - |\mathcal{K}| - 3}\hat{z}_{i,j|\mathcal{K}} > \Phi^{-1}(1-\alpha/2)$, where $\hat{z}_{i,j|\mathcal{K}} = 0.5\log((1 + \hat{\rho}_{i,j|\mathcal{K}})/(1 - \hat{\rho}_{i,j|\mathcal{K}}))$ and $\Phi(\cdot)$ is the cdf of $N(0, 1)$.

The final output of `PenPC` algorithm is the estimated skeleton and separation sets $S(i, j)$ for all $(i, j)$. If vertices $i$ and $j$ are connected in the skeleton, the separate set is an empty set, otherwise $X_i$ and $X_j$ are independent given $S(i, j)$, hence the name separation set. Given the skeleton and the separation sets, one can estimate CPDAG (Complete Partially Directed Acyclic Graphs) (Supplementary Materials Section 3) and then apply the `idaFast` or `ida` functions of R package pcalg (Kalisch et al., 2012) to estimate multi-set of possible causal effects.

## 4. Theoretical Properties

### 4.1 *Fixed Graphs*

We first introduce the following notations. For an $m \times n$ matrix $\mathbf{A}$, denote the matrix $L_b$ norm of $\mathbf{A}$ by $\|\mathbf{A}\|_b = \sup_{\mathbf{x}\neq 0}(\|\mathbf{Ax}\|_b/\|\mathbf{x}\|_b)$, where $\mathbf{x}$ is a vector of length $n$, and $\|\mathbf{x}\|_b = \left(\sum_{i=1}^n x_i^b\right)^{1/b}$. In particular, $\|\mathbf{A}\|_2$ is the spectral norm of $\mathbf{A}$, $\|\mathbf{A}\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^m |a_{ij}|$ is the maximum absolute column summation, and $\|\mathbf{A}\|_\infty = \max_{1\leqslant i\leqslant m}\sum_{j=1}^n |a_{ij}|$ is the maximum absolute row summation. Denote the vector $L_b$ norm of $\mathbf{A}$ by $|\mathbf{A}|_b$. In particular, $|\mathbf{A}|_2 = \left(\sum_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n} a_{ij}^2\right)^{1/2}$, and $|\mathbf{A}|_\infty = \max_{1\leqslant i\leqslant m, 1\leqslant j\leqslant n}|a_{ij}|$. Denote by $\mathbf{A}_{i,-i}$ the submatrix of $\mathbf{A}$ that includes the $i$-th row and excludes the $i$-th column of $\mathbf{A}$. $\mathbf{A}_{i,i}$, $\mathbf{A}_{-i,i}$ and $\mathbf{A}_{-i,-i}$ are defined similarly. For any compatible subsets $\mathcal{S}_1$ and $\mathcal{S}_2$, $\mathbf{A}_{\mathcal{S}_1,\mathcal{S}_2}$ is the submatrix that contains all the rows with indices in $\mathcal{S}_1$ and all the columns with indices in $\mathcal{S}_2$.

We denote $p$ as $p_n$ to emphasize it is a function of sample size $n$. Let a DAG and the corresponding GGM be $\mathcal{G}_n = (V_n, E_n)$ and $\mathcal{C}_{\mathcal{G}_n} = (V_n, F_n)$, respectively. We further denote the skeleton of $\mathcal{G}_n$ by $\mathcal{G}_n^u = (V_n, E_n^u)$ where $a - b \in E_n^u \Leftrightarrow a \to b \in E_n$ or $b \to a \in E_n$. For any vertex $i$, denote the observed centralized data of the variables within and outside of the neighbors of $i$ in $\mathcal{C}_{\mathcal{G}_n}$ (denoted by $\texttt{adj}(i, \mathcal{C}_{\mathcal{G}_n})$), but not including $X_i$, by $\mathscr{X}_{i1}$ and $\mathscr{X}_{i2}$, respectively, *i.e.*, $\mathscr{X}_{i1} = \mathbf{X}_{\mathcal{S}_0,\mathcal{S}_i}$ and $\mathscr{X}_{i2} = \mathbf{X}_{\mathcal{S}_0,-\{i,\mathcal{S}_i\}}$ where $\mathcal{S}_0 = \{1, 2, ..., n\}$ and $\mathcal{S}_i = \{j : j \in \texttt{adj}(i, \mathcal{C}_{\mathcal{G}_n})\}$.

The following Lemma 1 is a well-known conclusion that gives the relation between concentrate matrix of multivariate Gaussian distribution and the regression coefficients when we regress one variable versus all the other variables (Anderson, 2003).

LEMMA 1: *Suppose* $\boldsymbol{X} = (X_1, ..., X_p)^{\mathrm{T}} \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ *and* $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} = (\omega_{ij})_{p\times p}$. *Then* $X_i = \boldsymbol{X}_{-i}^{\mathrm{T}}\mathbf{b}_i + \epsilon_i$ *where* $\mathbf{b}_i = -\boldsymbol{\Omega}_{-i,i}/\omega_{ii}$, *and* $\epsilon_i \sim \mathrm{N}(0, 1/\omega_{ii})$ *independent of* $\boldsymbol{X}_{-i}$.

With the aforementioned notations and definitions, we can state the following conditions that are needed for the consistency of the `PenPC` algorithm.

(A1) Dimensionality of the problem. $p_n = O\left(\exp\{n^a\}\right)$ with $a \in [0, 1)$.

(A2) Sparseness assumption. Let $q_n = \max_{1 \leqslant j \leqslant p_n} |\texttt{adj}(j, \mathcal{C}_{\mathcal{G}_n})|$, i.e., the maximum degree of the Gaussian graphical model $\mathcal{C}_{\mathcal{G}_n}$. $q_n = O(n^b)$ for some $0 \leqslant b < (1-a)/2$. Let $M_n = \max_{1 \leqslant j \leqslant p_n} |\texttt{adj}(j, \mathcal{G}_n)|$. By the following Lemma 2, $M_n \leqslant q_n = O(n^b)$.

(A3) Minimum effect size for neighborhood selection.

$$\delta_n \equiv (1/2) \inf_{i,j} \left\{ \left| \frac{\omega_{ij}}{\omega_{ii}} \right| : \omega_{ij} \neq 0 \right\} \geqslant O(n^{-d_1}), \quad 0 < d_1 < (1-a-b)/2.$$

(A4) Conditions for the population covariance matrix $\mathbf{\Sigma} = \{\sigma_{ij}\}$. Let $\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}})$ be the minimum eigen-value of a sub matrix $\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}$. For any $\mathcal{S}$ with $|\mathcal{S}| \leqslant q_n$, $\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{S},\mathcal{S}}) > C_1$. We also assume $\max_i \sigma_{ii} < C_2$. Here $C_1$ and $C_2$ are two positive constants. Consequently $C_2$ is also an upper bound of all the off-diagonal elements of $\mathbf{\Sigma}$ because $\sigma_{ij} \leqslant (\sigma_{ii}\sigma_{jj})^{1/2}$.

(A5) Conditions for penalty function. Let $\kappa(\boldsymbol{v}; \boldsymbol{\varpi}) = \lim_{\epsilon \to 0+} \sup_{t_1 < t_2 \in (|v_j|-\epsilon, |v_j|+\epsilon)} \frac{p'(t_2;\boldsymbol{\varpi}) - p'(t_1;\boldsymbol{\varpi})}{t_2 - t_1}$ for $\boldsymbol{v} = (v_1, ..., v_r)^{\mathrm{T}} \in \mathbb{R}^r$ and $v_j \neq 0$. Thus $\kappa(\boldsymbol{v}; \boldsymbol{\varpi}) = \max_{1 \leqslant j \leqslant r} -p''(|v_j|; \boldsymbol{\varpi})$ if the penalty function has continuous second derivative. Let $\kappa_{0,i} = \max_{\boldsymbol{\beta}_1 \in \mathcal{N}_i} \kappa(\boldsymbol{\beta}_1; \boldsymbol{\varpi}_i)$ and $\mathcal{N}_i$ is a hypercube around the vector $\mathbf{b}_{i1} = (\omega_{ij}/\omega_{ii} : \ \omega_{ij} \neq 0)$ such that $\mathcal{N}_i = \{\boldsymbol{\beta}_1 : |\boldsymbol{\beta}_1 - \mathbf{b}_{i1}|_\infty \leqslant Cn^{-d_1}\}$. We assume $\max_i \kappa_{0,i} \leqslant C_1$, where $C_1$ is defined in (A4) , $p'(\delta_n; \boldsymbol{\varpi}_i) = o(q_n^{-1/2} n^{-d_1})$, and $q_n^{1/2} \max\{(\log p/n)^{1/2}, p'(\delta_n; \boldsymbol{\varpi}_i)\} = o(p'(0+; \boldsymbol{\varpi}_i))$.

(A6) Restriction on the size of conditional partial correlation. Denote the partial correlations between $X_i$ and $X_j$ given a set of variables $\{X_r : r \in \mathcal{K}\}$ for $\mathcal{K} \subseteq V_n \setminus \{i, j\}$ by $\varrho_{i,j|\mathcal{K}}$. For $\mathcal{K} \in \mathbf{\Pi}_{ij}$ ($\mathbf{\Pi}_{i,j}$ was defined in equation (2)), the absolute values of $\varrho_{i,j|\mathcal{K}}$'s are bounded:

$$\inf_{i,j,\mathcal{K}} \left\{ \left| \varrho_{i,j|\mathcal{K}} \right| : \rho_{i,j|\mathcal{K}} \neq 0, \mathcal{K} \in \mathbf{\Pi}_{i,j} \right\} \geqslant c_n, \text{ and } \sup_{i,j,\mathcal{K}} \left\{ \left| \varrho_{i,j|\mathcal{K}} \right| : \mathcal{K} \in \mathbf{\Pi}_{ij} \right\} \leqslant M < 1,$$

where $c_n = O(n^{-d_2})$ for some $0 < d_2 < \min\{(1-a)/2, (1-b)/2\}$.

The sparseness assumption (A2) will be replaced by tighter assumptions for two specific random graph models later. Assumptions (A3)-(A5) ensure that the step 1 of `PenPC` can recover the partial correlation graph. There are fairly reasonable conditions to ensure the identifiability of the problem. Assumption (A3) requires the minimum effect size is larger

than noise level (e.g., larger than $O(n^{-1/2})$ when $p = O(1)$). Assumption (A4) requires the covariance matrix for those important covariates in a neighborhood selection problem is not singular. Assumption (A5) are conditions for the penalty function, which can be easily satisfied by adjusting the two tuning parameters of the Log penalty (Chen et al., 2014). Assumption (A6) ensures the summation of the mistaken probabilities of the step 2 of the PenPC algorithm goes to 0 asymptotically. The condition $q_n^{1/2} p'(\delta_n; \boldsymbol{\varpi}_i) = o(p'(0+; \boldsymbol{\varpi}))$ of assumption (A5) deserves more discussions because it corresponds to the irrepresentable condition that limits the performance of Lasso regression. Specifically, in Supplementary Materials, we show that $\left\| \mathscr{X}_{i2}^{\mathrm{T}} \mathscr{X}_{i1} (\mathscr{X}_{i1}^{\mathrm{T}} \mathscr{X}_{i1})^{-1} \right\|_\infty = O(q_n^{1/2})$ with probability approaching to 1. The assumption $q_n^{1/2} p'(\delta_n; \boldsymbol{\varpi}_i) = o(p'(0+; \boldsymbol{\varpi}))$ is needed so that $\left\| \mathscr{X}_{i2}^{\mathrm{T}} \mathscr{X}_{i1} (\mathscr{X}_{i1}^{\mathrm{T}} \mathscr{X}_{i1})^{-1} \right\|_\infty = o(p'(0+; \boldsymbol{\varpi})/p'(\delta_n; \boldsymbol{\varpi}))$. For the Lasso, $p'(0+; \boldsymbol{\varpi})/p'(\delta_n; \boldsymbol{\varpi}) = 1$, and thus this is a very strong assumption for the size of $\left\| \mathscr{X}_{i2}^{\mathrm{T}} \mathscr{X}_{i1} (\mathscr{X}_{i1}^{\mathrm{T}} \mathscr{X}_{i1})^{-1} \right\|$. In contrast, for the log penalty, $p'(t; \lambda_i, \tau_i) = \lambda_i \mathtt{sgn}(t)/(|t| + \tau_i)$, and thus $p'(0+; \tau_i)/p'(\delta_n; \tau_i) \to (\delta_n + \tau_i)/\tau_i$, which can goes to infinity if $\tau_i = o(\delta_n)$. We can show that the log penalty satisfies other assumptions and refer the readers to Chen et al. (2014) for details.

Consider the neighborhood selection problem for the $i$-th variable versus all the other variables. Recall that $\mathcal{S}_i = \{j : j \in \mathtt{adj}(i, \mathcal{C}_{\mathcal{G}_n})\} = \mathrm{supp}(\mathbf{b}_i)$ is the support of the true regression coefficient $\mathbf{b}_i$ with size $|\mathcal{S}_i| = s_i$. Let $\mathbf{b}_{i1}$ and $\hat{\mathbf{b}}_{i1}$ be respectively the sub-vectors of $\mathbf{b}_i$ and $\hat{\mathbf{b}}_i$ corresponding to $\mathcal{S}_i$.

THEOREM 1: *Given Assumptions (A1) - (A5), with probability at least* $1 - C\exp\{-n^a\}$ *for a constant* $0 < C < \infty$, *there exists a local minimizer* $\hat{\mathbf{b}}_i = (\hat{\mathbf{b}}_{i1}, \hat{\mathbf{b}}_{i2})^{\mathrm{T}}$ *that satisfies the following conditions for any* $i = 1, \ldots, p_n$,

*(a) Sparsity:* $\hat{\mathbf{b}}_{i2} = \mathbf{0}$,

*(b)* $L_\infty$ *loss:* $\|\hat{\mathbf{b}}_{i1} - \mathbf{b}_{i1}\|_\infty = o(n^{-d_1})$, *where* $d_1$ *is defined in (A3).*

*Therefore, if we denote the estimate of* $\mathcal{C}_{\mathcal{G}_n}$ *by the neighborhood selection as* $\hat{\mathcal{C}}_{\mathcal{G}_n}(\boldsymbol{\varpi})$*, where* $\boldsymbol{\varpi}$ *are tuning parameters of the penalty function,* $\mathbb{P}(\hat{\mathcal{C}}_{\mathcal{G}_n}(\boldsymbol{\varpi}) = \mathcal{C}_{\mathcal{G}_n}) \geqslant 1 - C\exp\{-n^a\}$.

The proof is in the Supplementary Materials. The following Lemma 2 and 3 provide the theoretical justifications for using GGM as a starting point of our modified PC-algorithm.

LEMMA 2: *If the distribution* $P_{\boldsymbol{X}}$ *is* Markov *to* $\mathcal{G}$*, i.e., if the joint density* $f_X$ *satisfies the recursive factorization, the set of edges* $F_n$ *of* $\mathcal{C}_{\mathcal{G}_n}$ *includes all edges* $E_n^u$ *of* $\mathcal{G}_n^u$ *plus the edges between co-parents of v-structures in* $\mathcal{G}_n$*.*

LEMMA 3: *Assume (A1). If* $(i,j) \in F_n$ *of* $\mathcal{C}_{\mathcal{G}_n}$ *but* $(i,j) \notin E_n^u$ *of* $\mathcal{G}_n^u$*, the conditioning set* $\boldsymbol{\Pi}_{i,j}$ *in (2) includes at least one set which d-separates vertices* $i$ *and* $j$ *in* $\mathcal{G}$*.*

Lemma 2 has been proved in Lemma 3.21 of Lauritzen (1996). The proof of Lemma 3 is presented in the Supplementary Materials. Lemma 2 shows that the concentration matrix recovers all the edges in the skeleton with no false negatives, but some false positives between the co-parents of v-structures. Lemma 3 shows that we can remove such false positives by examining partial correlation conditioning on some set in $\boldsymbol{\Pi}_{i,j}$.

Next we discuss the theoretical property of the modified PC-stable algorithm given a perfect estimation of GGM.

THEOREM 2: *Let* $\alpha_n$ *be the p-value threshold for testing whether a partial correlation is 0. Let* $\hat{\mathcal{G}}_n^u(\alpha_n)$ *be the estimates of* $\mathcal{G}_n^u$ *from the second step of the* `PenPC` *algorithm given a perfect estimation of GGM from the first step of the* `PenPC` *algorithm. Assume (A1), (A2) and (A6), then there exists* $\alpha_n \to 0$*, such that* $\mathbb{P}\left[\hat{\mathcal{G}}_n^u(\alpha_n) = \mathcal{G}_n^u\right] = 1 - O\left(\exp\{-Cn^{1-2d_2}\}\right) \to 1$*, where* $0 < C < \infty$ *is a constant.*

The proof is in the Supplementary Materials. Similar theorem has been proved in Kalisch and Bühlmann (2007) with $p_n$ at polynomial order of $n$. By starting with GGM, we extend

the theorem to $p_n = O\left(\exp\{n^a\}\right)$ case. Combining the results of Theorem 1 and Theorem 2, corollary 1 show that the summation of mistaken probabilities of GGM estimation and skeleton estimation given GGM goes to 0 as $n \to \infty$.

COROLLARY 1: *Let $\hat{\mathcal{G}}_n^u(\boldsymbol{\theta}, \alpha_n)$ be the estimates of $\mathcal{G}_n^u$ from the two-step approach* PenPC *algorithm. Assume (A1)-(A6), then there exists an $\alpha_n \to 0$, such that $\mathbb{P}\left[\hat{\mathcal{G}}_n^u(\boldsymbol{\theta}, \alpha_n) = \mathcal{G}_n^u\right] = 1 - O\left(\exp\{-n^a\}\right) \to 1$, where $0 < C < \infty$ is a constant.*

4.2 *Random Graphs*

Next we extend our theoretical results to two commonly used models for random graphs: Erdős and Rényi (ER) Model (Erdős and Rényi, 1960) and Barabási and Albert (BA) Model (Barabási and Albert, 1999). Let $q_n = \max_{1 \leqslant j \leqslant p_n} |\texttt{adj}(j, \mathcal{C}_{\mathcal{G}_n})|$ and $M_n = \max_{1 \leqslant j \leqslant p_n} |\texttt{adj}(j, \mathcal{G}_n)|$. In general, assumption (A2) no longer holds for random graphs. It is easy to see that assumption (A2) can be relaxed to (A2') and we introduced an additional assumption (A7)

(A2') $\mathbb{P}\{q_n \leqslant O(n^b)\} = 1$, for some $0 \leqslant b < 1$.

(A7) $M_n \geqslant C q_n$, where $0 < C < 1$ is a constant.

Assumption (A7) says that the maximal degree in the GGM is not dominated by the edges induced by co-parents of v-structures as $n$ goes to infinity, which is a reasonable assumption. Given this assumption, $M_n$ and $q_n$ are on the same scale.

4.2.1 *Erdős and Rényi (ER) Model.* The ER model constructs a graph $G(p_n, p_E)$ of $p_n$ vertices by connecting vertices randomly. Each edge is included in the graph with probability $p_E$ independent from all other edges. By law of large numbers, such vertex is almost surely connected to $(p_n - 1)p_E$ edges. Erdős and Rényi (1960) proved the following results about $M_n$, the maximal degree of the graph.

LEMMA 4: *In the graph $G(p_n, p_E)$ following the ER model, the maximal degree $M_n$ almost*

*surely converges to* $m_n$*, where* $m_n = O(\log p_n)$ *if* $p_n p_E < 1$*,* $m_n = p_n^{2/3}$ *if* $p_n p_E = 1$*, and* $m_n = O(p_n)$ *if* $\lim_{p_n \to \infty} p_n p_E = c > 1$*.*

When $p_n = O\{\exp(n^a)\}$, by Lemma 4 and assumption (A7), assumption (A2') holds if $p_n p_E < 1$ and $b \geqslant a$. When $p_n p_E \geqslant 1$, our proof cannot handle the general case $p_n = O\{\exp(n^a)\}$. However, when the number of vertices is of the polynomial order of $n$, assumption (A2') may still hold. In particular, suppose $p_n = O(n^r)$. When $p_n p_E < 1$, assumption (A2') holds for any $b \in [0, \infty)$. When $p_n p_E = 1$, assumption (A2') holds if $b \geqslant 2r/3$. When $p_n p_E \to c > 1$, assumption (A2') holds if $r < 1$ and $b \geqslant r$.

4.2.2 *Barabási and Albert (BA) Model.* The BA model is used to generate scale free graphs whose degree distribution follows a power law: $\mathbb{P}(\nu) = \gamma_0 \nu^{-\gamma_1}$, with a normalizing constant $\gamma_0$ and a exponent $\gamma_1$. Specifically, BA model generates a graph by adding vertices into the graph over time and when each new vertex is introduced into the graph, it is connected with larger probability to the existing vertices with larger number of connections. Since the distribution does not depend on the size of the network (or time), the graph organizes itself into a scale free state (Barabási and Albert, 1999). Móri (2005) showed that $M_n$ (the maximal degree of the graph) almost surely converges to $O(p^{1/2})$. Thus, assumption (A2') holds for the case $p_n = O(n^r)$ with $b \leqslant r/2$.

## 5. Simulation Studies

We evaluated the performance of the `PenPC` algorithm and the PC-stable algorithm in terms of sensitivity and specificity of skeleton estimation using DAGs simulated by the ER model or the BA model. In both simulations and real data analysis, we used the implementation of the PC-stable algorithm by function `skeleton` in `R` package `pcalg` (version 1.1-6), and we have implemented `PenPC` algorithm in `R` package `PenPC`.

Following Kalisch and Bühlmann (2007), we simulated DAGs of $p$ vertices by the ER model

as follows. For any vertex pair $(i, j)$ where $i < j$, we added an edge $i \to j$ with probability $p_E$. For the BA model, the DAGs were simulated following Barabási and Albert (1999). The initial graph had one vertex and no edge. In the $(t+1)$-th step, $e$ edges were proposed. For each edge, the new vertex was connected to the $i$-th $(1 \leqslant i \leqslant t)$ existing vertex with probability $\nu_i^{(t)} / \sum_j \nu_j^{(t)}$, where $\nu_i^{(t)} = |\mathtt{adj}(i, \mathcal{G}^{(t)})|$, and $\mathcal{G}^{(t)}$ was the DAG at the $t$-th step. The distribution of the degrees $\nu$ from simulated DAGs under ER model ($p = 1000$ and $p_E = 2/p$) and BA model ($p = 1000$ and $e = 1$) are shown in Figure 2 and similar graph for BA model ($p = 1000$ and $e = 2$) is shown in Figure S2 of Supplementary Materials.

[Figure 2 about here.]

The probability of finding a highly connected vertex decreases exponentially with $\nu$ for the graphs generated by the ER model (Figure 2(a)). However, for the graphs generated by the BA model, there is a linear relation between degree and degree probability in log-log scale, confirming its scale-free property (Figure 2(b)).

After constructing the DAGs, the observed data were simulated by structure equations under multivariate Gaussian assumption. For example, denote the parents of $X_j$ by $\mathtt{pa}_j$, then $\mathbf{x}_j = \sum_{k \in \mathtt{pa}_j} b_{jk} \mathbf{x}_k + \epsilon_j$, where $\epsilon_j \sim \mathcal{N}(0, \sigma^2 I_{n \times n})$. In our simulations, all $b_{jk}$'s and $\sigma^2$ were set to be 1. For either ER or BA model, we considered low dimension setting where $p = 11, n = 100$ and high-dimension settings where $p = 100, n = 30$ and $p = 1000, n = 300$ with various sparsity levels determined by $p_E$ for ER model and $e$ for BA model (Table 1).

[Table 1 about here.]

Due to limited space, here we only show the results for the simulation setups using ER model with $p$=1000, $n$=300, and $p_E$=0.005; and BA model with $p$=1000, $n$=300, and $e$=1. The remaining results are presented in Figure S3 - S14 of the Supplementary Materials.

[Figure 3 about here.]

[Figure 4 about here.]

First consider the results from the ER model when $\alpha = 0.01$. Recall that $\alpha$ is the p-value threshold for conditional independence testing. The penalized regression (step 1 of the `PenPC` algorithm) identifies more true positives than the PC-stable algorithm, but also introduce more false positives (Figure 3 (a-b)), while `PenPC` algorithm significantly reduces the number of false positives, though some true positives are also removed. At the end, the `PenPC` has the lowest number of false positives plus false negatives, as measured by Hamming distance (HD) (Figure 3 (c)). Figures 3(d-f) show that across various values of $\alpha$, `PenPC` consistently has better performance than the PC-stable algorithm. Finally, Figure 3(g) shows the ROC curves for the `PenPC` and the PC-stable algorithms, which illustrate that `PenPC` has better sensitivity and specificity than the PC-stable algorithm regardless of the cutoff $\alpha$. Similar conclusions can be drawn for the simulation results shown in Figure 4, where the DAGs are simulated by the BA model. We note that although `PenPC` performs well for the BA model, further improvement is possible by incorporating special consideration for the scale-free structure of BA graphs (Liu and Ihler, 2011).

## 6. Application

We applied the PC-stable algorithm and the `PenPC` algorithm to study gene-gene network using gene expression data from tumor tissue of 550 TCGA (The Cancer Genome Atlas) breast cancer patients (Cancer Genome Atlas Network, 2012). Gene expression were measured by RNA-seq. We quantified the expression of each gene within each sample by log(total read count) (logTReC). After removing genes with low expression across most samples, we ended up with 18,827 genes. We first removed the effects of several covariates by taking residuals of logTReC for each gene using a linear regression with the following covariates: 75 percentile

of logTReC per sample, which captures read depth, plate, institution, age, and six PCs from the corresponding germline genotype data.

The computational cost of the PC-stable algorithm increases quickly as the number of vertices or the p-value cutoff increases. When we study 410 genes, the computational time of the PC-stable and `PenPC` algorithms are both within an hour. When we expand the number of genes to 8,261. The step 1 in `PenPC` algorithm took 3 hours in total while searching for 1000 combinations of tuning parameters for each gene. Given the GGM, the 2nd step of the `PenPC` is computationally much more efficient than the PC-stable algorithm (Figure 5 (a)). For example, with p-value threshold varies from $10^{-7}$ to $10^{-5}$, the computational time of the PC-stable algorithm increases from 20 to 50 hours. In contrast, the computation time of the `PenPC` remains below 10 hours even for p-value cutoff $5 \times 10^{-3}$. All the computation are done in Linux server with an 2.93 GHz Intel processor and 48GB RAM.

[Figure 5 about here.]

Since PC-stable algorithm is not computationally feasible for larger gene set, we first discussed the results on 410 genes from the cancer Gene Census in `http://cancer.sanger.ac.uk/cancergenome/projects/census/`. For $\alpha = 0.0001$ to 0.05, we estimated the skeleton by the PC-stable and `PenPC` algorithms. The estimated skeletons were evaluated by comparing the estimated edge sets with protein-protein interaction (PPI) database at `http://www.pathwaycommons.org/pc2/downloads.html`. PPI is a reasonable resource to evaluate graphical model estimates because genes with PPI tend to co-express (Rhodes et al., 2005). There were 3315 PPIs where both proteins belong to the 410 genes. Figure 5(a) shows the total number of detected edges versus the number of edges in PPI data. For both methods, the total number of detected edges increase monotonically as $\alpha$ increases. The `PenPC` results have higher sensitivity to detect PPI given the number of edges discovered.

Next we applied PenPC to 8,261 genes with PPI annotation. Using $\alpha = 0.001$, we detected

12,150 edges, that is 0.03% of the total number of edges. We arbitrarily define the genes with more than 7 neighbors as hub genes and there are 46 hub genes (Supplementary Table 1). Interestingly, many of the hub genes are cancer-related. For example, all the hub genes with more than 9 neighbors, MYC, ELF3, and RAB15, are cancer-related. MYC encodes Myc proto-oncogene protein, which are associated with multiple types of human cancers including breast cancer. ELF3 is one of the ETS transcription factor and it modulates breast cancer-associated gene expression. RAB15 is a member RAS oncogene family.

## 7. Discussions

The seminal works of Kalisch and Bühlmann (2007) have greatly advanced our understating of the PC-algorithm and provided well-designed and user-friendly software packages (Kalisch et al., 2012). Our `PenPC` algorithm provides some helpful improvements, especially in high dimensional settings. The `PenPC` algorithm has three tuning parameters, two for step 1 (tuning parameters of the Log penalty for neighborhood selection) and one for step 2 (p-value cutoff) of the `PenPC` algorithm. The selection of tuning parameters for step 1 and step 2 of `PenPC` are two independent procedures. For step 1, it is a classical problem of tuning parameter selection for penalized regression. We chose to use extended BIC as it delivers the best performance and it has sound theoretical justifications (Chen and Chen, 2008). Choosing the best combination of the two tuning parameters for log penalty using extended BIC does not induce heavy computational cost. For example, we use 1,000 combinations of $\lambda$ and $\tau$ for each of the penalized regressions with Log penalty, and for our real data analysis with $n = 550$ and $p = 8,261$, it takes about 3 hours for all the $p = 8,261$ penalized regressions. In the second step of `PenPC`, we need to choose a p-value threshold for conditional independence tests, similar to the tuning parameter for the PC algorithm. How to choose this p-value cutoff is an open problem that warrants further research.

We have compared PenPC with the approach of replacing the Log penalty with the Lasso penalty. As expected, the Lasso penalty leads to much worse performance in high dimensional settings (Figures S15-S16 in Supplementary Materials). Following Kalisch and Bühlmann (2007), we assume a multivariate Gaussian assumption so that we may test conditional independence by assessing conditional correlation. The first step of the PenPC algorithm (neighborhood selection) does not require this assumption. Our method is robust to this multivariate Gaussian assumption. For example, Figures S15-S16 show that the performance of PenPC algorithm is comparable when the data are simulated from multivariate Gaussian and multivariate t-distribution (df=5). Assuming the observed data are generated by linear, potentially non-Gaussian structural equation model (SEM), Loh and Bühlmann (2014) proved that the moral graph of a DAG (i.e., the DAG skeleton plus the edges that connect the co-parents of v-structures) can be estimated by the support of the inverse covariance matrix and they estimated inverse covariance matrix using graphical Lasso. Borrowing their theoretical justifications, we can extend PenPC to non-Gaussian cases. The first step of PenPC is similar to graphical Lasso, but with log penalty instead of Lasso penalty. The second step of PenPC can be modified by using a conditional independent test that does not rely on Gaussian assumption. Finally, our work assumes no hidden confounders or latent variables, which may be justified by the fact that we examine the expression of all the genes and the effects of confounders or latent variables may be manifested by the expression of certain genes.

## 8. Supplementary Materials

The Supplementary Figures and Results referenced in Sections 2-7 are available with this paper at the Biometrics website on Wiley Online Library http://www.biometrics.tibs.org, along with our method in an R package named PenPC.

Acknowledgements

This research is supported in part by NIH R01 GM105785-01 and HG006292-03. We are grateful for the constructive comments from two anonymous reviewers and the editors.

References

Anderson, T. (2003). *An Introduction to Multivariate Statistical Analysis. 2003*. Wiley, New York.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks. *science* **286,** 509–512.

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

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika* **95,** 759–771.

Chen, T., Sun, W., and Fine, J. (2014). Designing penalty functions in high dimensional problems: The role of tuning parameters. Technical report, University of North Carolina, Chapel Hill.

Chickering, D. M. (2003). Optimal structure identification with greedy search. *The Journal of Machine Learning Research* **3,** 507–554.

Colombo, D. and Maathuis, M. (2012). A modification of the pc algorithm yielding order-independent skeletons. *arXiv preprint arXiv:1211.3295* .

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5,** 17–61.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with np-dimensionality. *Information Theory, IEEE Transactions on* **57,** 5467–5484.

Heckerman, D., Geiger, D., and Chickering, D. (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning* **20,** 197–243.

Kalisch, M. and Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *The Journal of Machine Learning Research* **8,** 613–636.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* **47,** 1–26.

Lauritzen, S. (1996). *Graphical models*, volume 17. Oxford University Press, USA.

Liu, Q. and Ihler, A. T. (2011). Learning scale free networks by reweighted l1 regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 40–48.

Loh, P.-L. and Bühlmann, P. (2014). High-dimensional learning of linear causal networks via inverse covariance estimation. *Journal of Machine Learning Research* **15,** 3065–3105.

Maathuis, M., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods* **7,** 247–248.

Maathuis, M., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* **37,** 3133–3164.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). Sparsenet: Coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106,**.

Meek, C. (1995). Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 411–418. Morgan Kaufmann Publishers Inc.

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

Móri, T. (2005). The maximum degree of the barabási-albert random tree. *Combinatorics Probability and Computing* **14,** 339–348.

Pearl, J. (2009). *Causality: models, reasoning and inference*. Cambridge Univ Press.

Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-

Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A. M. (2005). Probabilistic model of the human protein-protein interaction network. *Nature biotechnology* **23,** 951–959.

Schmidt, M., Niculescu-Mizil, A., and Murphy, K. (2007). Learning graphical model structure using l1-regularization paths. In *AAAI*, volume 7, pages 1278–1283.

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

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, prediction and search*, volume 81. The MIT Press.

Sun, W., Ibrahim, J. G., and Zou, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185,** 349–359.

Teyssier, M. and Koller, D. (2005). Ordering-based search: A simple and effective algorithm for learning bayesian networks. In *In UAI*, pages 584–590.

Tsamardinos, I., Brown, L. E., and Aliferis, C. F. (2006). The max-min hill-climbing bayesian network structure learning algorithm. *Machine learning* **65,** 31–78.

van de Geer, S. and Bühlmann, P. (2013). l0-penalized maximum likelihood for sparse directed acyclic graphs. *The Annals of Statistics* **41,** 536–567.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7,** 2541–2563.

*Received April* 2014. *Revised xxx* 2014. *Accepted xxx* 2014.

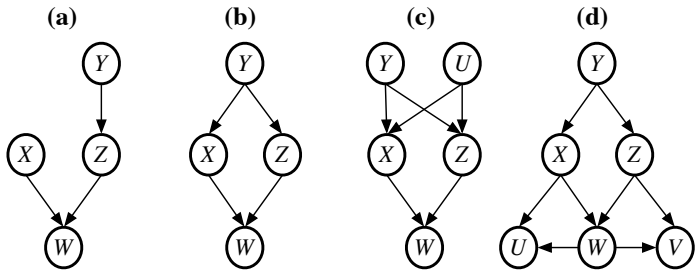


**Figure 1.** Four DAGs where $X$ and $Z$ are not connected in the skeleton, but are connected in the corresponding GGMs. The true relation between $X$ and $Z$ can be revealed by appropriate conditional independence testing. For example, $X \perp Z$ in Figure 1(a), $X \perp Z|Y$ in Figure 1(b), $X \perp Z|(Y, U)$ in Figure 1(c), and $X \perp Z|Y$ in Figure 1(d).

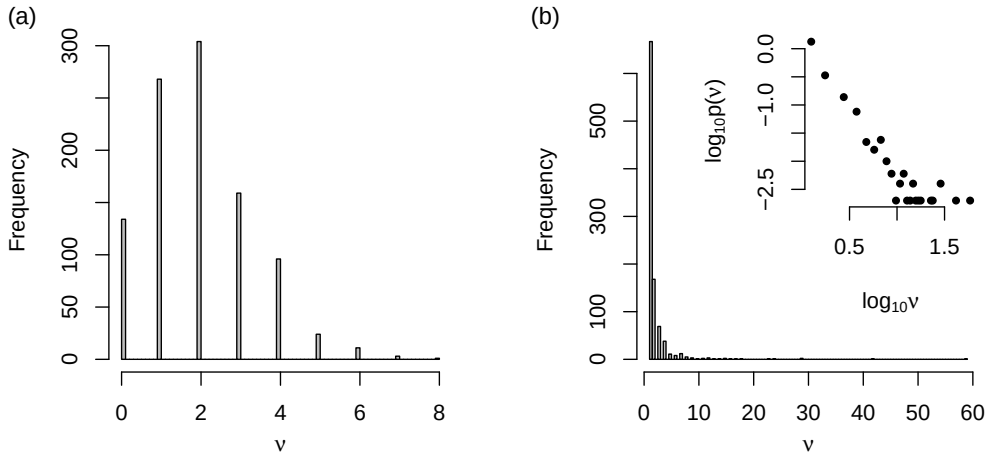


**Figure 2.** Histograms of the degree $\nu$. (a) ER model with $p = 1000$ and $p_E = 2/p$. (b) BA model with $p = 1000$ and $e = 1$ and the $\log_{10}$ scale density of $\log_{10} \nu$ in its subplot.

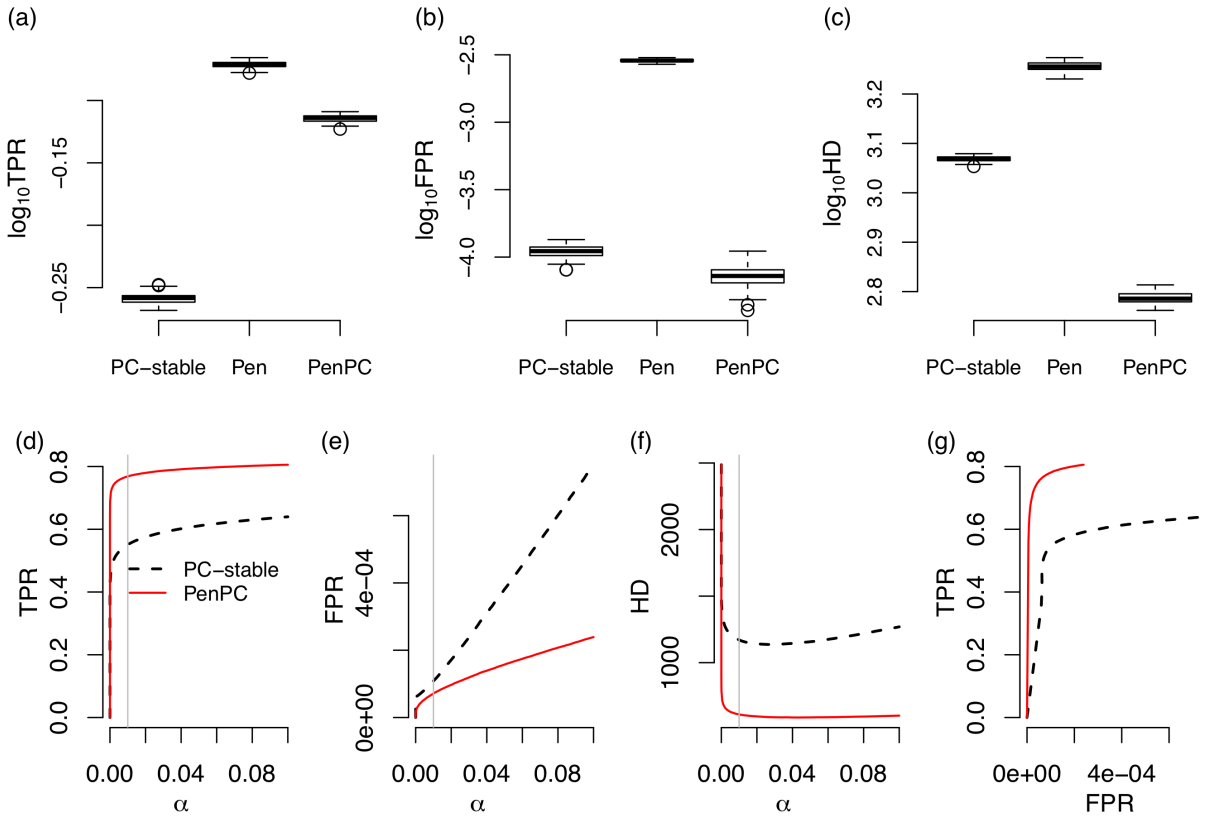


**Figure 3.** Performance of ER model ($p = 1000, n = 300, p_E = 0.005$). The upper panels are box plots (in $\log_{10}$ scale) of true positive rate (TPR) (a), false positive rate (FPR) (b) and hamming distance (HD) (c) from 100 replications at $\alpha = 0.01$. The lower panels are average true positive rate (d), false positive rate (e), and Hamming distance (f) from 100 replications when the tuning parameter $\alpha$ is changed from 0 to 0.1 (the grey vertical line are at $\alpha = 0.01$). ROC curves are shown in panel (g). This figure appears in color in the electronic version of this article.

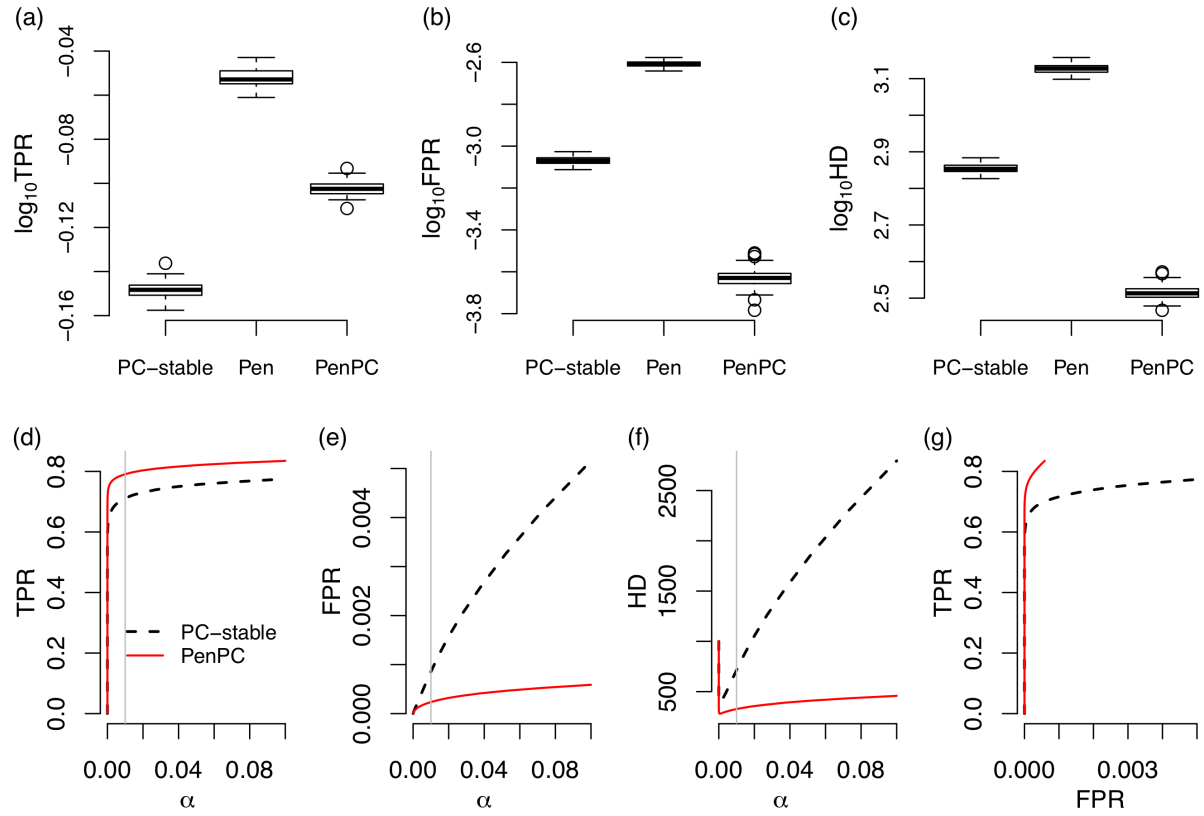


**Figure 4.** Performance of BA model (p=1000,n=300,e=1). The upper panels are box plots (in $\log_{10}$ scale) of true positive rate (TPR) (a), false positive rate (FPR) (b) and hamming distance (HD) (c) from 100 replications at $\alpha = 0.01$. The lower panels are average true positive rate (d), false positive rate (e), and Hamming distance (f) from 100 replications when the tuning parameter $\alpha$ is changed from 0 to 0.1 (the grey vertical line are at $\alpha = 0.01$). ROC curves are shown in panel (g). This figure appears in color in the electronic version of this article.

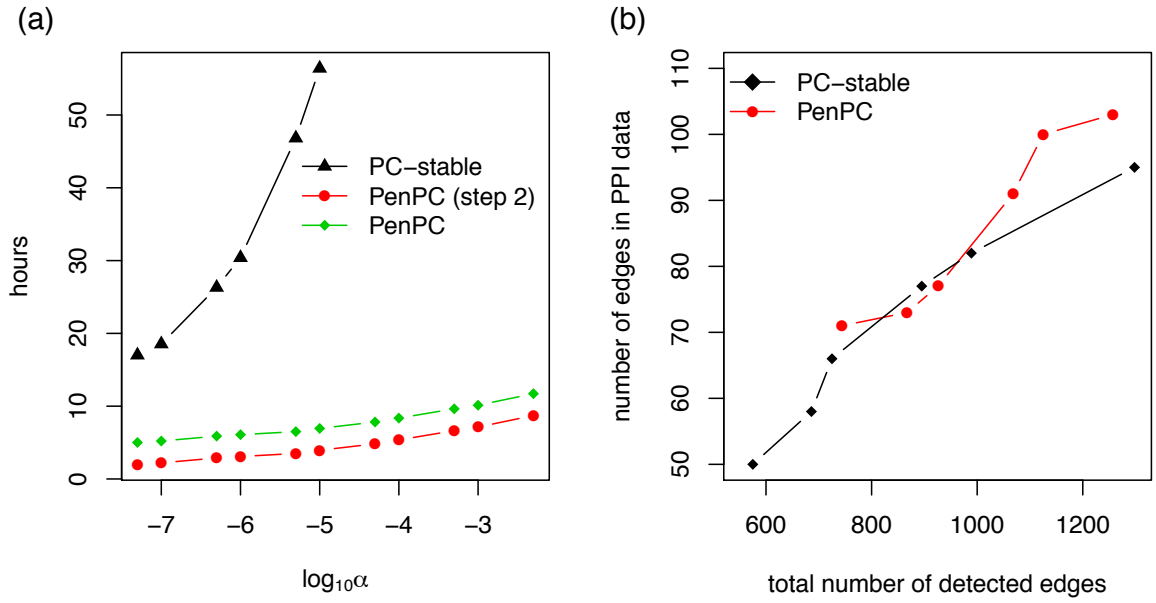


**Figure 5.** Comparing `PenPC` algorithm with PC-stable algorithm in terms of skeleton estimation by changing the significance levels for partial correlation testings: $\alpha$=0.0001, 0.0005, 0.001, 0.005, 0.01 and 0.05. (a) The computational time at different $\alpha$ values when we consider 8,261 genes. (b) The number of detected edges vs. the number of edges in the PPI (protein-protein interaction) data when we consider 410 genes. This figure appears in color in the electronic version of this article.

**Table 1**
*Simulation Setting*

| $p$ | $n$ | $p_E$ (ER) | $e$ (BA) |
|---|---|---|---|
| 11 | 100 | 0.2 | 1,2 |
| 100 | 30 | 0.02, 0.03, 0.04, 0.05 | 1,2 |
| 1000 | 300 | 0.002, 0.005, 0.01 | 1,2 |