Journal of Machine Learning Research 14 (2013) 3365-3383 Submitted 7/12; Revised 6/13; Published 11/13

PC Algorithm for Nonparanormal Graphical Models

Naftali Harris NAFTALI@STANFORD.EDU
Department of Statistics
Stanford University
Stanford, CA, USA 94305

Mathias Drton MD5@UW.EDU
Department of Statistics
University of Washington
Seattle, WA, USA 98195

Editor: Chris Meek

Abstract

The PC algorithm uses conditional independence tests for model selection in graphical modeling with acyclic directed graphs. In Gaussian models, tests of conditional independence are typically based on Pearson correlations, and high-dimensional consistency results have been obtained for the PC algorithm in this setting. Analyzing the error propagation from marginal to partial correlations, we prove that high-dimensional consistency carries over to a broader class of Gaussian copula or *nonparanormal* models when using rank-based measures of correlation. For graph sequences with bounded degree, our consistency result is as strong as prior Gaussian results. In simulations, the 'Rank PC' algorithm works as well as the 'Pearson PC' algorithm for normal data and considerably better for non-normal data, all the while incurring a negligible increase of computation time. While our interest is in the PC algorithm, the presented analysis of error propagation could be applied to other algorithms that test the vanishing of low-order partial correlations.

Keywords: Gaussian copula, graphical model, model selection, multivariate normal distribution, nonparanormal distribution

1. Introduction

Let $G = (V,E)$ be an acyclic digraph with finite vertex set, and let $X = (X_v)_{v\in V}$ be a random vector whose entries are in correspondence with the graph's vertices. Then the graph G determines a statistical model for the joint distribution of X by imposing conditional independences that can be read off from G using the concept of d-separation. These independences are natural if the edges in E encode causal/functional relationships among the random variables X_v, and a distribution that satisfies them is said to be *Markov* with respect to G. Appendix B contains a brief review of these and other key notions that are relevant to this paper. More detailed introductions to statistical modeling with directed graphs can be found in Lauritzen (1996), Pearl (2009), Spirtes et al. (2000) or Drton et al. (2009, Chapter 3). As common in the field, we use the abbreviation DAG (for 'directed acyclic graph') when referring to acyclic digraphs.

We will be concerned with the consistency of the PC algorithm, which is named for its inventors, the first two authors of Spirtes et al. (2000). This algorithm uses conditional independence tests to infer a DAG from data. Alongside greedy-search techniques that optimize information criteria, the PC algorithm is one of the main methods for inference of directed graphs. Recent applications of the

PC algorithm can be found in Maathuis et al. (2010), Schmidberger et al. (2011), Le et al. (2013), and Verdugo et al. (2013).

Graph inference is complicated by the fact that two DAGs $G = (V,E)$ and $H = (V,F)$ with the same vertex set V may be *Markov equivalent*, that is, they may possess the same d-separation relations and, consequently, induce the same statistical model. Hence, the goal becomes estimation of the Markov equivalence class of an acyclic digraph G. For representation of the equivalence class, prior work considers a particular partially directed graph $C(G)$, for which it holds that $C(G) = C(H)$ if and only if the two DAGs G and H are Markov equivalent; see Andersson et al. (1997) and Chickering (2002). The graph $C(G)$ may contain both directed and undirected edges, and it is acyclic in the sense of its directed subgraph having no directed cycles. We will refer to $C(G)$ as the *completed partially directed acyclic graph* (CPDAG), but other terminology such as the *essential graph* is in use.

The PC algorithm uses conditional independence tests to infer a CPDAG from data (Spirtes et al., 2000). In its population version, the algorithm amounts to a clever scheme to reconstruct the CPDAG $C(G)$ from answers to queries about d-separation relations in the underlying DAG G. Theorem 1 summarizes the properties of the PC algorithm that are relevant for the present paper. For a proof of the theorem as well as a compact description of the PC algorithm we refer the reader to Kalisch and Bühlmann (2007). Recall that the degree of a node is the number of edges it is incident to, and that the degree of a DAG G is the maximum degree of any node, which we denote by $\deg(G)$.

Theorem 1 *Given only the ability to check d-separation relations in a DAG G, the PC algorithm finds the CPDAG $C(G)$ by checking whether pairs of distinct nodes are d-separated by sets S of cardinality $|S| \leq \deg(G)$.*

Let X_A denote the subvector $(X_v)_{v\in A}$. The joint distribution of a random vector $X = (X_v)_{v\in V}$ is *faithful* to a DAG G if, for any triple of pairwise disjoint subsets $A,B,S \subset V$, we have that S d-separates A and B in G if and only if X_A and X_B are conditionally independent given X_S; it is customary to denote this conditional independence by $X_A \perp\!\!\!\perp X_B \,|\, X_S$. Under faithfulness, statistical tests of conditional independence can be used to determine d-separation relations in a DAG and lead to a sample version of the PC algorithm that is applicable to data.

If X follows the multivariate normal distribution $N(\mu,\Sigma)$, with positive definite covariance matrix Σ, then

$$X_A \perp\!\!\!\perp X_B \,|\, X_S \quad \Longleftrightarrow \quad X_u \perp\!\!\!\perp X_v \,|\, X_S \quad \forall u \in A,\, v \in B.$$

Moreover, the pairwise conditional independence of X_u and X_v given X_S is equivalent to the vanishing of the *partial correlation* $\rho_{uv|S}$, that is, the correlation obtained from the bivariate normal conditional distribution of (X_u, X_v) given X_S. The iterations of the PC algorithm make use of the recursion

$$\rho_{uv|S} = \frac{\rho_{uv|S\setminus w} - \rho_{uw|S\setminus w}\rho_{vw|S\setminus w}}{\sqrt{\left(1-\rho^2_{uw|S\setminus w}\right)\left(1-\rho^2_{vw|S\setminus w}\right)}}, \tag{1}$$

for any $w \in S$, where $\rho_{uv|\emptyset} = \rho_{uv}$ is the correlation of u and v. Our later theoretical analysis will use the fact that

$$\rho_{uv|S} = -\frac{\Psi^{-1}_{uv}}{\sqrt{\Psi^{-1}_{uu}\Psi^{-1}_{vv}}}, \tag{2}$$

where $\Psi = \Sigma_{(u,v,S),(u,v,S)}$ is the concerned principal submatrix of Σ.

A natural estimate of $\rho_{uv|S}$ is the sample partial correlation obtained by replacing Σ with the empirical covariance matrix of available observations. Sample partial correlations derived from independent normal observations have favorable distributional properties (Anderson, 2003, Chapter 4), which form the basis for the work of Kalisch and Bühlmann (2007) who treat the PC algorithm in the Gaussian context with conditional independence tests based on sample partial correlations. The main results in Kalisch and Bühlmann (2007) show high-dimensional consistency of the PC algorithm, when the observations form a sample of independent normal random vectors that are faithful to a suitably sparse DAG.

The purpose of this paper is to show that the PC algorithm has high-dimensional consistency properties for a broader class of distributions, when standard Pearson-type empirical correlations are replaced by rank-based measures of correlations in tests of conditional independence. The broader class we consider includes continuous distributions with Gaussian copula. Phrased in the terminology of Liu et al. (2009), we consider *nonparanormal* distributions. Recall that a correlation matrix is a covariance matrix with all diagonal entries equal to one.

Definition 2 *Let $f = (f_v)_{v\in V}$ be a collection of strictly increasing functions $f_v : \mathbb{R} \to \mathbb{R}$, and let $\Sigma \in \mathbb{R}^{V\times V}$ be a positive definite correlation matrix. The nonparanormal distribution $NPN(f,\Sigma)$ is the distribution of the random vector $(f_v(Z_v))_{v\in V}$ for $(Z_v)_{v\in V} \sim N(0,\Sigma)$.*

Taking the functions f_v to be affine shows that all multivariate normal distributions are also nonparanormal. If $X \sim NPN(f,\Sigma)$, then the univariate marginal distribution for a coordinate, say X_v, may have any continuous cumulative distribution function F, as we may take $f_v = F^- \circ \Phi$, where Φ is the standard normal distribution function and $F^-(u) = \inf\{x : F(x) \geq u\}$. Note that f_v need not be continuous.

Definition 3 *The nonparanormal graphical model $NPN(G)$ associated with a DAG G is the set of all distributions $NPN(f,\Sigma)$ that are Markov with respect to G.*

Since the marginal transformations f_v are deterministic, the dependence structure in a nonparanormal distribution corresponds to that in the underlying latent multivariate normal distribution. In other words, if $X \sim NPN(f,\Sigma)$ and $Z \sim N(0,\Sigma)$, then it holds for any triple of pairwise disjoint sets $A,B,S \subset V$ that

$$X_A \perp\!\!\!\perp X_B \,|\, X_S \iff Z_A \perp\!\!\!\perp Z_B \,|\, Z_S.$$

Hence, for two nodes u and v and a separating set $S \subset V \setminus \{u,v\}$, it holds that

$$X_u \perp\!\!\!\perp X_v \,|\, X_S \iff \rho_{uv|S} = 0, \tag{3}$$

with $\rho_{uv|S}$ calculated from Σ as in (1) or (2). In light of this equivalence, we will occasionally speak of a correlation matrix Σ being Markov or faithful to a DAG, meaning that the requirement holds for any distribution $NPN(f,\Sigma)$.

In the remainder of the paper we study the PC algorithm in the nonparanormal context, proposing the use of Spearman's rank correlation and Kendall's τ for estimation of the correlation matrix parameter of a nonparanormal distribution. In Section 2, we review how transformations of Spearman's rank correlation and Kendall's τ yield accurate estimators of the latent Gaussian correlations. In particular, we summarize tail bounds from Liu et al. (2012a). Theorem 8 in Section 4 gives our main result, an error bound for the output of the PC algorithm when correlations are used to determine nonparanormal conditional independence. In Corollary 9, we describe high-dimensional asymptotic scenarios and suitable conditions that lead to consistency of the PC algorithm. The proof of Theorem 8 is given in Section 3, which provides an analysis of error propagation from marginal to partial correlations. Our numerical work in Section 5 makes a strong case for the use of rank correlations in the PC algorithm. Some concluding remarks are given in Section 6.

2. Rank Correlations

Let (X,Y) be a pair of random variables, and let F and G be the cumulative distribution functions of X and Y, respectively. Spearman's ρ for the bivariate distribution of (X,Y) is defined as

$$\rho^S = \mathrm{Corr}\left(F(X), G(Y)\right),$$

that is, it is the ordinary Pearson correlation between the quantiles $F(X)$ and $G(Y)$. Another classical measure of correlation is Kendall's τ, defined as

$$\tau = \mathrm{Corr}\left(\mathrm{sign}\left(X - X'\right), \mathrm{sign}\left(Y - Y'\right)\right)$$

where (X',Y') is an independent copy of (X,Y), that is, (X',Y') and (X,Y) are independent and have the same distribution.

Suppose $(X_1,Y_1),\ldots(X_n,Y_n)$ are independent pairs of random variables, each pair distributed as (X,Y). Let $\mathrm{rank}(X_i)$ be the rank of X_i among $X_1,\ldots,X_n$. In the nonparanormal setting, the marginal distributions are continuous so that ties occur with probability zero, making ranks well-defined. The natural estimator of ρ^S is the sample correlation among ranks, that is,

$$\begin{aligned}\hat{\rho}^S &= \frac{\frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{rank}(X_i)}{n+1} - \frac{1}{2}\right)\left(\frac{\mathrm{rank}(Y_i)}{n+1} - \frac{1}{2}\right)}{\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{rank}(X_i)}{n+1} - \frac{1}{2}\right)^2}\sqrt{\frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{rank}(Y_i)}{n+1} - \frac{1}{2}\right)^2}} \\ &= 1 - \frac{6}{n\left(n^2-1\right)}\sum_{i=1}^n \left(\mathrm{rank}(X_i) - \mathrm{rank}(Y_i)\right)^2,\end{aligned}$$

which can be computed in $O(n\log n)$ time. Kendall's τ may be estimated by

$$\hat{\tau} = \frac{2}{n\left(n-1\right)}\sum_{1\le i<j\le n} \mathrm{sign}\left(X_i - X_j\right)\mathrm{sign}\left(Y_i - Y_j\right).$$

A clever algorithm using sorting and binary trees to compute $\hat{\tau}$ in time $O(n\log n)$ instead of the naive $O(n^2)$ time has been developed by Christensen (2005).

It turns out that simple trigonometric transformations of $\hat{\rho}^S$ and $\hat{\tau}$ are excellent estimators of the population Pearson correlation for multivariate normal data. In particular, Liu et al. (2012a) show that if (X,Y) are bivariate normal with $\mathrm{Corr}(X,Y) = \rho$, then

$$\mathbb{P}\left(\left|2\sin\left(\frac{\pi}{6}\hat{\rho}^S\right) - \rho\right| > \varepsilon\right) \;\le\; 2\exp\left(-\frac{2}{9\pi^2}n\varepsilon^2\right) \tag{4}$$

and

$$\mathbb{P}\left(\left|\sin\left(\frac{\pi}{2}\hat{\tau}\right) - \rho\right| > \varepsilon\right) \;\le\; 2\exp\left(-\frac{2}{\pi^2}n\varepsilon^2\right). \tag{5}$$

Clearly, $\hat{\rho}^S$ and $\hat{\tau}$ depend on the observations $(X_1,Y_1),\dots(X_n,Y_n)$ only through their ranks. Since ranks are preserved under strictly increasing functions, (4) and (5) still hold if $(X,Y) \sim NPN(f,\Sigma)$ with Pearson correlation $\rho = \Sigma_{xy}$ in the underlying latent bivariate normal distribution. Throughout the rest of this paper, we will assume that we have some estimator $\hat{\rho}$ of ρ which has the property that, for nonparanormal data,

$$\mathbb{P}(|\hat{\rho}-\rho| > \varepsilon) \ < \ A\exp\left(-Bn\varepsilon^2\right) \tag{6}$$

for fixed constants $0 < A,B < \infty$. As just argued, the estimators considered in (4) and (5) both have this property.

When presented with multivariate observations from a distribution $NPN(f,\Sigma)$, we apply the estimator from (6) to every pair of coordinates to obtain an estimator $\hat{\Sigma}$ of the correlation matrix parameter. Plugging $\hat{\Sigma}$ into (1) or equivalently into (2) gives partial correlation estimators that we denote $\hat{\rho}_{uv|S}$.

3. Error Propagation from Marginal to Partial Correlations

The PC algorithm leverages statistical decisions on conditional independence. An analysis of the algorithm in the context of nonparanormal distributions thus requires bounds on errors in partial correlations. The following Lemma 4 is our main tool. It provides a uniform bound on errors in partial correlations when a uniform bound on errors in marginal correlations is available. At times we will write such uniform bounds in terms of the l_∞ vector norm of a matrix. For matrix $A = (a_{ij}) \in \mathbb{R}^{q\times q}$ we denote this norm by

$$\|A\|_\infty = \max_{1\le i,j\le q} |a_{ij}|.$$

Some proofs involve the spectral norm $\|A\|$, that is, the square-root of the maximal eigenvalue of A^TA.

Lemma 4 (Errors in partial correlations) *Suppose $\Sigma \in \mathbb{R}^{q\times q}$ is a positive definite matrix with minimal eigenvalue $\lambda_{\min} > 0$. If $\hat{\Sigma} \in \mathbb{R}^{q\times q}$ satisfies*

$$\|\hat{\Sigma}-\Sigma\|_\infty < \frac{c\lambda_{\min}^2}{(2+c)q+\lambda_{\min}cq}$$

with $c > 0$, then all partial correlations are well-defined and their differences are bounded as

$$|\hat{\rho}_{uv|\setminus\{u,v\}} - \rho_{uv|\setminus\{u,v\}}| \ := \ \left| \frac{\Sigma^{-1}_{uv}}{\sqrt{\Sigma^{-1}_{uu}\Sigma^{-1}_{vv}}} - \frac{\hat{\Sigma}^{-1}_{uv}}{\sqrt{\hat{\Sigma}^{-1}_{uu}\hat{\Sigma}^{-1}_{vv}}} \right| < c, \quad 1 \le u < v \le q.$$

The proof of Lemma 4 follows by combining the conclusions of Lemmas 5, 6 and 7 from this section. The first of these, that is, Lemma 5, invokes classical results on error propagation in matrix inversion.

Lemma 5 (Matrix inversion) *Suppose $\Sigma \in \mathbb{R}^{q\times q}$ is a positive definite matrix with minimal eigenvalue $\lambda_{\min} > 0$. If $E \in \mathbb{R}^{q\times q}$ is a matrix of errors with $\|E\|_\infty < \varepsilon < \lambda_{\min}/q$, then $\Sigma + E$ is invertible and*

$$\|(\Sigma+E)^{-1} - \Sigma^{-1}\|_\infty \le \frac{q\varepsilon/\lambda_{\min}^2}{1-q\varepsilon/\lambda_{\min}}.$$

Proof First, note that

$$\|E\|_\infty \le \|E\| \le q\|E\|_\infty; \tag{7}$$

see entries $(2,6)$ and $(6,2)$ in the table on p. 314 in Horn and Johnson (1990). Using the sub-multiplicativity of a matrix norm, the second inequality in (7), and our assumption on ε, we find that

$$\|E\Sigma^{-1}\| \le \|\Sigma^{-1}\| \cdot \|E\| < \frac{q\varepsilon}{\lambda_{\min}} < 1. \tag{8}$$

As discussed in Horn and Johnson (1990, Section 5.8), this implies that $I + E\Sigma^{-1}$ and thus also $\Sigma + E$ is invertible. Moreover, by the first inequality in (7) and inequality (5.8.2) in Horn and Johnson (1990), we obtain that

$$\|(\Sigma+E)^{-1} - \Sigma^{-1}\|_\infty \le \|(\Sigma+E)^{-1} - \Sigma^{-1}\| \le \|\Sigma^{-1}\| \cdot \frac{\|E\Sigma^{-1}\|}{1-\|E\Sigma^{-1}\|}.$$

Since the function $x \mapsto x/(1-x)$ is increasing for $x < 1$, our claim follows from the fact that $\|\Sigma^{-1}\| = 1/\lambda_{\min}$ and the inequality $\|E\Sigma^{-1}\| < q\varepsilon/\lambda_{\min}$ from (8). ■

Lemma 6 (Diagonal of inverted correlation matrix) *If $\Sigma \in \mathbb{R}^{q\times q}$ is a positive definite correlation matrix, then the diagonal entries of $\Sigma^{-1} = (\sigma^{ij})$ satisfy $\sigma^{ii} \ge 1$.*

Proof The claim is trivial for $q = 1$. So assume $q \ge 2$. By symmetry, it suffices to consider the entry σ^{qq}, and we partition the matrix as

$$\Sigma = \begin{pmatrix} A & b \\ b^T & 1 \end{pmatrix}$$

with $A \in \mathbb{R}^{(q-1)\times(q-1)}$ and $b \in \mathbb{R}^{q-1}$. By the Schur complement formula for the inverse of a partitioned matrix,

$$\sigma^{qq} = \frac{1}{1 - b^T A^{-1} b};$$

compare Horn and Johnson (1990, Section 0.7.3). Since A is positive definite, so is A^{-1}. Hence, $b^T A^{-1} b \ge 0$. Since Σ^{-1} is positive definite, σ^{qq} cannot be negative, and so we deduce that $\sigma_{qq} \ge 1$, with equality if and only if $b = 0$. ■

The next lemma addresses the error propagation from the inverse of a correlation matrix to partial correlations.

Lemma 7 (Correlations) *Let $A = (a_{ij})$ and $B = (b_{ij})$ be symmetric 2×2 matrices. If A is positive definite with $a_{11}, a_{22} \ge 1$ and $\|A - B\|_\infty < \delta < 1$, then*

$$\left| \frac{a_{12}}{\sqrt{a_{11}a_{22}}} - \frac{b_{12}}{\sqrt{b_{11}b_{22}}} \right| < \frac{2\delta}{1-\delta}.$$

Proof Without loss of generality, suppose $a_{12} \geq 0$. Since $\|A-B\|_\infty < \delta$,

$$\frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} < \frac{a_{12}+\delta}{\sqrt{(a_{11}-\delta)(a_{22}-\delta)}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}}$$
$$= \frac{\delta}{\sqrt{(a_{11}-\delta)(a_{22}-\delta)}} + a_{12}\left(\frac{1}{\sqrt{(a_{11}-\delta)(a_{22}-\delta)}} - \frac{1}{\sqrt{a_{11}a_{22}}}\right).$$

Using that $a_{11}, a_{22} \geq 1$ to bound the first term and $a_{12}^2 < a_{11}a_{22}$ to bound the second term, we obtain that

$$\frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} < \frac{\delta}{1-\delta} + \sqrt{a_{11}a_{22}}\left(\frac{1}{\sqrt{(a_{11}-\delta)(a_{22}-\delta)}} - \frac{1}{\sqrt{a_{11}a_{22}}}\right)$$
$$= \frac{\delta}{1-\delta} + \left(\sqrt{\frac{a_{11}}{a_{11}-\delta}\cdot\frac{a_{22}}{a_{22}-\delta}} - 1\right).$$

Since the function $x \mapsto x/(x-\delta)$ is decreasing, we may use our assumption that $a_{11}, a_{22} \geq 1$ to get the bound

$$\frac{b_{12}}{\sqrt{b_{11}b_{22}}} - \frac{a_{12}}{\sqrt{a_{11}a_{22}}} < \frac{\delta}{1-\delta} + \left(\sqrt{\frac{1}{1-\delta}\cdot\frac{1}{1-\delta}} - 1\right) = \frac{2\delta}{1-\delta}$$

A similar argument yields that

$$\frac{a_{12}}{\sqrt{a_{11}a_{22}}} - \frac{b_{12}}{\sqrt{b_{11}b_{22}}} < \frac{2\delta}{1+\delta},$$

from which our claim follows. ∎

4. Rank PC Algorithm

Based on the equivalence (3), we may use the rank-based partial correlation estimates $\hat{\rho}_{uv|S}$ to test conditional independences. In other words, we conclude that

$$X_u \perp\!\!\!\perp X_v | X_S \quad \iff \quad \left|\hat{\rho}_{uv|S}\right| \leq \gamma, \tag{9}$$

where $\gamma \in [0,1]$ is a fixed threshold. We will refer to the PC algorithm that uses the conditional independence tests from (9) as the 'Rank PC' (RPC) algorithm. We write $\hat{C}_\gamma(G)$ for the output of the RPC algorithm with tuning parameter γ.

The RPC algorithm consist of two parts. The first part computes the correlation matrix $\hat{\Sigma} = (\hat{\rho}_{uv})$ in time $O(p^2 n \log n)$, where $p := |V|$. This computation takes $O(\log n)$ longer than its analogue under use of Pearson correlations. The second part of the algorithm is independent of the type of correlations involved. It determines partial correlations and performs graphical operations. For an accurate enough estimate of a correlation matrix Σ that is faithful to a DAG G, this second part takes $O(p^{\deg(G)})$ time in the worst case, but it is often much faster; compare Kalisch and Bühlmann (2007). For high-dimensional data with n smaller than p, the computation time for RPC is dominated by the second part, the PC-algorithm component. Moreover, in practice, one may wish to apply RPC for several different values of γ, in which case the estimate $\hat{\Sigma}$ needs to be calculated only once. As a result, Rank PC takes only marginally longer to compute than Pearson PC for high-dimensional data.

What follows is our main result about the correctness of RPC. For a correlation matrix $\Sigma \in \mathbb{R}^{V\times V}$, let

$$c_{\min}(\Sigma) := \min\left\{|\rho_{uv|S}| \,:\, u,v\in V,\, S\subseteq V\setminus\{u,v\},\, \rho_{uv|S}\neq 0\right\} \tag{10}$$

be the minimal magnitude of any non-zero partial correlation, and let $\lambda_{\min}(\Sigma)$ be the minimal eigenvalue. Then for any integer $q\geq 2$, let

$$c_{\min}(\Sigma,q) := \min\{\, c_{\min}(\Sigma_{I,I}) \,:\, I\subseteq V,\, |I|\leq q\,\}, \quad \text{and} \tag{11}$$

$$\lambda_{\min}(\Sigma,q) := \min\{\lambda_{\min}(\Sigma_{I,I}) \,:\, I\subseteq V,\, |I|\leq q\,\} \tag{12}$$

be the minimal magnitude of a non-zero partial correlation and, respectively, the minimal eigenvalue of any principal submatrix of order at most q.

Theorem 8 (Error bound for RPC-algorithm) *Let $X_1,\ldots,X_n$ be a sample of independent observations drawn from a nonparanormal distribution $NPN(f,\Sigma)$ that is faithful to a DAG G with p nodes. For $q := \deg(G)+2$, let $c := c_{\min}(\Sigma,q)$ and $\lambda := \lambda_{\min}(\Sigma,q)$. If $n>q$, then there exists a threshold $\gamma\in[0,1]$ for which*

$$\mathbb{P}\big(\hat{C}_\gamma(G)\neq C(G)\big) \;\leq\; \frac{A}{2}p^2\exp\left(-\frac{B\lambda^4 nc^2}{36q^2}\right),$$

where $0<A,B<\infty$ are the constants from (6).

We remark that while all subsets of size q appear in the definitions in (11) and (12), our proof of Theorem 8 only requires the corresponding minima over those principal submatrices that are actually inverted in the run of the PC-algorithm.

Proof (Theorem 8) We will show that our claimed probability bound for the event $\hat{C}_\gamma(G)\neq C(G)$ holds when the threshold in the RPC algorithm is $\gamma = c/2$. By Theorem 1, if all conditional independence tests for conditioning sets of size $|S|\leq\deg(G)$ make correct decisions, then the output of the RPC algorithm $\hat{C}_\gamma(G)$ is equal to the CPDAG $C(G)$. When $\gamma=c/2$, the conditional independence test accepts a hypothesis $X_u \perp\!\!\!\perp X_v|X_S$ if and only if $|\hat{\rho}_{uv|S}|<\gamma=c/2$. Hence, the test makes a correct decision if $|\hat{\rho}_{uv|S}-\rho_{uv|S}|<c/2$ because all non-zero partial correlations for $|S|\leq\deg(G)$ are bounded away from zero by c; recall (10) and (11). It remains to argue, using the error analysis from Lemma 4, that the event $|\hat{\rho}_{uv|S}-\rho_{uv|S}|\geq c/2$ occurs with small enough probability when $|S|\leq\deg(G)$.

Suppose our correlation matrix estimate $\hat{\Sigma}=(\hat{\rho}_{uv})$ satisfies $\|\hat{\Sigma}-\Sigma\|_\infty<\varepsilon$ for

$$\varepsilon = \frac{c\lambda^2}{(4+c)q+\lambda cq} = \frac{\lambda^2 c/2}{(2+c/2)q+\lambda qc/2} > 0. \tag{13}$$

Choose any two nodes $u,v\in V$ and a set $S\subseteq V\setminus\{u,v\}$ with $|S|\leq\deg(G)=q-2$. Let $I=\{u,v\}\cup S$. Applying Lemma 4 to the $I\times I$ submatrix of Σ and $\hat{\Sigma}$ yields

$$|\hat{\rho}_{uv|S}-\rho_{uv|S}| \;<\; \frac{c}{2}.$$

Therefore, $\|\hat{\Sigma}-\Sigma\|_\infty < \varepsilon$ implies that our tests decide all conditional independences correctly in the RPC algorithm.

Next, using (6) and a union bound, we find that

$$\mathbb{P}\left(\hat{C}_\gamma(G) \neq C(G)\right) \leq \mathbb{P}\left(|\hat{\Sigma}_{uv}-\Sigma_{uv}| \geq \varepsilon \text{ for some } u,v \in V\right) \\ \leq A\frac{p(p-1)}{2}\exp\left(-Bn\varepsilon^2\right).$$

Plugging in the definition of ε gives the claimed inequality

$$\mathbb{P}\left(\hat{C}_\gamma(G) \neq C(G)\right) \leq \frac{A}{2}p^2\exp\left(-\frac{B\lambda^4 nc^2}{((4+c)q+\lambda cq)^2}\right) \leq \frac{A}{2}p^2\exp\left(-\frac{B\lambda^4 nc^2}{36q^2}\right)$$

because $c \leq 1$ and $\lambda \leq 1$. The inequality $c \leq 1$ holds trivially because partial correlations are in $[-1,1]$. The inequality $\lambda \leq 1$ holds because a $q \times q$ correlation matrix has trace q, this trace is equal to the sum of the q eigenvalues, and λ is the minimal eigenvalue. ■

From the probability bound in Theorem 8, we may deduce high-dimensional consistency of RPC. For two positive sequences (s_n) and (t_n), we write $s_n = O(t_n)$ if $s_n \leq Mt_n$, and $s_n = \Omega(t_n)$ if $s_n \geq Mt_n$ for a constant $0 < M < \infty$.

Corollary 9 (Consistency of RPC-algorithm) *Let (G_n) be a sequence of DAGs. Let p_n be the number of nodes of G_n, and let $q_n = \deg(G_n)+2$. Suppose (Σ_n) is a sequence of $p_n \times p_n$ correlation matrices, with Σ_n faithful to G_n. Suppose further that there are constants $0 \leq a,b,d,f < 1$ that govern the growth of the graphs as*

$$\log p_n = O(n^a), \qquad q_n = O(n^b),$$

and minimal signal strengths and eigenvalues as

$$c_{\min}(\Sigma_n, q_n) = \Omega(n^{-d}), \qquad \lambda_{\min}(\Sigma_n, q_n) = \Omega(n^{-f}).$$

If $a+2b+2d+4f < 1$, then there exists a sequence of thresholds γ_n for which

$$\lim_{n\to\infty}\mathbb{P}\left(\hat{C}_{\gamma_n}(G_n) = C(G_n)\right) = 1,$$

where $\hat{C}_{\gamma_n}(G_n)$ is the output of the RPC algorithm for a sample of independent observations $X_1,\dots,X_n$ from a nonparanormal distribution $NPN(\cdot,\Sigma_n)$.

Proof By Theorem 8, for large enough n, we can pick a threshold γ_n such that

$$\mathbb{P}(\hat{C}_{\gamma_n}(G_n) \neq C(G_n) \leq A' \exp\left(2n^a - B' n^{1-2b-2d-4f}\right)$$

for constants $0 < A', B' < \infty$. The bound goes to zero if $1-2b-2d-4f > a$. ■

As previously mentioned, Kalisch and Bühlmann (2007) prove a similar consistency result in the Gaussian case. Whereas our proof consists of propagation of errors from correlation to partial correlation estimates, their proof appeals to Fisher's result that under Gaussianity, sample partial correlations follow the same type of distribution as sample correlations when the sample size is adjusted by subtracting the cardinality of the conditioning set (Anderson, 2003, Chapter 4). It is then natural to work with a bound on the partial correlations associated with small conditioning sets. More precisely, Kalisch and Bühlmann (2007) assume that there is a constant $0 \leq M < 1$ such that for any n, the partial correlations $\rho_{uv|S}$ of the matrix Σ_n satisfy

$$|\rho_{uv|S}| \leq M \qquad \forall\, u,v \in V,\ S \subseteq V \setminus \{u,v\},\ |S| \leq q_n. \tag{14}$$

It is then no longer necessary to involve the minimal eigenvalues from (12). The work in Kalisch and Bühlmann (2007) is thus free of an analogue to our constant f. Stated for the case of polynomial growth of p_n (with $a = 0$), their result gives consistency when $b + 2d < 1$; our constant b corresponds to $1 - b$ in Kalisch and Bühlmann (2007). The condition from Corollary 9, on the other hand, requires $2b + 2d < 1$ even if $f = 0$. This is more restrictive as larger b allows for faster growth in the degree of the graphs and larger d allows for faster decay of the minimal signal strength.

In the important special case of bounded degree, however, our nonparanormal result is just as strong as the previously established Gaussian consistency guarantee. Staying with polynomial growth of p_n, that is, $a = 0$, suppose the sequence of graph degrees $\deg(G_n)$ is indeed bounded by a fixed constant, say $q_0 - 2$. Then clearly, $b = 0$. Moreover, the set of correlation matrices of size q_0 satisfying (14) with $q_n = q_0$ is compact. Since the smallest eigenvalue is a continuous function, the infimum of all eigenvalues of such matrices is achieved for some invertible matrix. Hence, the smallest eigenvalue is bounded away from zero, and we conclude that $f = 0$. Corollary 9 thus implies consistency if $2d < 1$, or if $d < \frac{1}{2} = \frac{1-b}{2}$, precisely as in Kalisch and Bühlmann (2007). (No generality is lost by assuming $a = 0$; in either one of the compared results this constant is involved solely in a union bound over order p^2 events.)

5. Numerical Experiments

In this section we evaluate the finite-sample properties of the RPC algorithm in simulations and in an application to gene expression data. In implementations of the PC algorithm in the `pcalg` package for R (Kalisch et al., 2012) and other software such as `Tetrad IV`,[1] the Gaussian conditional independence tests use a fixed level $\alpha \in [0,1]$ and decide that

$$X_u \perp\!\!\!\perp X_v | X_S \iff \sqrt{n - |S| - 3}\left|\frac{1}{2}\log\left(\frac{1+\hat{\rho}_{uv|S}}{1-\hat{\rho}_{uv|S}}\right)\right| \leq \Phi^{-1}(1 - \alpha/2)\,. \tag{15}$$

If the observations are multivariate normal and $\hat{\rho}_{uv|S}$ are sample partial correlations then α is an asymptotic significance level for the test. The sample size adjustment from n to $n - |S| - 3$ achieves a bias-correction (Anderson, 2003).

Suppose for a moment that in (15) the square root of $n - |S| - 3$ was simply $\sqrt{n}$. Then, for fixed n and α, the acceptance region in (15) could be translated into a corresponding fixed value for γ in (9). Hence, our Theorem 8 would apply directly when plugging rank correlations into the mentioned software implementations of the PC algorithm. With the sample size adjustment from n to $n - |S| - 3$, however, the value of γ depends on $|S|$ and further arguments are needed. We postpone these to Appendix A, where we show that the sample size adjustment has indeed no effect on the consistency result in Corollary 9.

1. Tetrad IV can be found at `http://www.phil.cmu.edu/projects/tetrad`.

5.1 Simulations

We compare RPC to two other versions of the PC-algorithm: (i) 'Pearson-PC', by which we mean the standard approach of using sample partial correlations to test Gaussian conditional independences, and (ii) 'Q_n-PC', which is based on a robust estimator of the covariance matrix and was considered in Kalisch and Bühlmann (2008). All our computations are done with the `pcalg` package for R.

Following Kalisch and Bühlmann (2007), we simulate random DAGs and sample from probability distributions faithful to them. Fix a sparsity parameter $s \in [0,1]$ and enumerate the vertices as $V = \{1,\dots,p\}$. Then we generate a DAG by including the edge $u \to v$ with probability s, independently for each pair (u,v) with $1 \le u < v \le p$. In this scheme, each node has the same expected degree $(p-1)s$.

Given a DAG $G = (V,E)$, let $\Lambda = (\lambda_{uv})$ be a $p \times p$ matrix with $\lambda_{uv} = 0$ if $u \to v \notin E$. Furthermore, let $\varepsilon = (\varepsilon_1,\dots,\varepsilon_p)$ be a vector of independent random variables. Then the random vector X solving the equation system

$$X = \Lambda X + \varepsilon \tag{16}$$

is well-known to be Markov with respect to G. Here, we draw the edge coefficients λ_{uv}, $u \to v \in E$, independently from a uniform distribution on the interval $(0.1,1)$. For such a random choice, with probability one, the vector X solving (16) is faithful with respect to G. We consider three different types of data:

(i) multivariate normal observations, which we generate by taking ε in (16) to have independent standard normal entries;

(ii) observations with Gaussian copula obtained by transforming the marginals of the normal random vectors from (i) to an $F_{1,1}$-distribution;

(iii) contaminated data, for which we generate the entries of ε in (16) as independent draws from a 80-20 mixture between a standard normal and a standard Cauchy distribution.

The contaminated distributions in (iii) do not belong to the nonparanormal class.

For the simulations we sample from two graph distributions: A small graph on ten vertices with an expected vertex degree of three, and a larger graph on one hundred vertices with an expected vertex degree of six. For each $n \in \{50, 1000\}$ and each of the three types of data listed above, we sample 201 random graphs from both the small and large graph distributions, and then sample n observations from the graph with the given data distribution.

For each resulting combination, we run each of the three considered versions of the PC algorithm on a grid of α's ranging from 10^{-100} to 0.8. We consider the RPC algorithm in the version that uses Spearman correlations as in (4); the results for Kendall's τ were similar. For each estimated skeleton, we compute the proportions of true and of false positives by comparing the estimated skeleton to the true skeleton. The skeleton of a graph G is the undirected graph with edges between nodes that are adjacent in G. Finally, we compute the area under the receiver operating characteristic curve (AUC) for each of the 201 repetitions. Mean areas with standard deviation in parenthesis are listed in Tables 1- 3.

A clear message emerges from the tables. First, Table 1 shows that for normal data, RPC performs only marginally worse than Pearson-PC. The Q_n-PC algorithm does well on larger sample sizes, but it not as good on smaller sample sizes. Second, Table 2 shows a dramatic relative gain

	Pearson-PC	Q_n-PC	RPC
Small graph, $n = 50$	0.824 (0.065)	0.734 (0.102)	0.809 (0.072)
Small graph, $n = 1000$	0.938 (0.050)	0.930 (0.053)	0.936 (0.050)
Large graph, $n = 50$	0.721 (0.016)	0.584 (0.022)	0.706 (0.016)
Large graph, $n = 1000$	0.837 (0.023)	0.830 (0.023)	0.835 (0.023)

Table 1: Mean AUC for Normal data

	Pearson-PC	Q_n-PC	RPC
Small graph, $n = 50$	0.668 (0.079)	0.506 (0.062)	0.813 (0.067)
Small graph, $n = 1000$	0.774 (0.068)	0.566 (0.082)	0.930 (0.054)
Large graph, $n = 50$	0.587 (0.012)	0.502 (0.004)	0.704 (0.016)
Large graph, $n = 1000$	0.678 (0.021)	0.525 (0.011)	0.833 (0.024)

Table 2: Mean AUC for Nonparanormal data

in performance for RPC for the Gaussian copula data with $F_{1,1}$ marginals. As expected, the performance of RPC on nonparanormal data is the same as on normal data, while that of Pearson-PC and Q_n-PC deteriorate. Finally, Table 3 shows that RPC continues to do well in the presence of contaminated data, the mean AUC for the other two algorithms is significantly lower. Curiously, despite using a robust covariance matrix estimator, the Q_n-PC performs substantially worse than Pearson-PC on this data.

5.2 Gene Expression Data

While Kendall's τ and Spearman's rank correlation give similar results for continuous observations from a distribution with Gaussian copula, the two measures of correlation can give quite different results in applications. We illustrate this for data on gene expression in yeast from Brem and Kruglyak (2005), where we focus on $p = 54$ genes from the MAPK signaling pathway as was done in Sun and Li (2012). The sample size is $n = 112$.

When plotting histograms of the expression measurements for each of the 54 genes, the majority of the plots do not show any obvious deviation from normality but, as one might suspect, there are several with signs of skewness as well as some outliers. Moreover, for five genes, the marginal distribution appears to be bimodal; see Figure 1 for an example. Multimodal marginals can arise under nonparanormal distributions, which thus have the potential to alleviate the effects of such

	Pearson-PC	Q_n-PC	RPC
Small graph, $n = 50$	0.781 (0.075)	0.656 (0.102)	0.819 (0.073)
Small graph, $n = 1000$	0.905 (0.078)	0.859 (0.110)	0.939 (0.053)
Large graph, $n = 50$	0.646 (0.023)	0.518 (0.008)	0.690 (0.017)
Large graph, $n = 1000$	0.738 (0.039)	0.616 (0.044)	0.832 (0.024)

Table 3: Mean AUC for Contaminated data

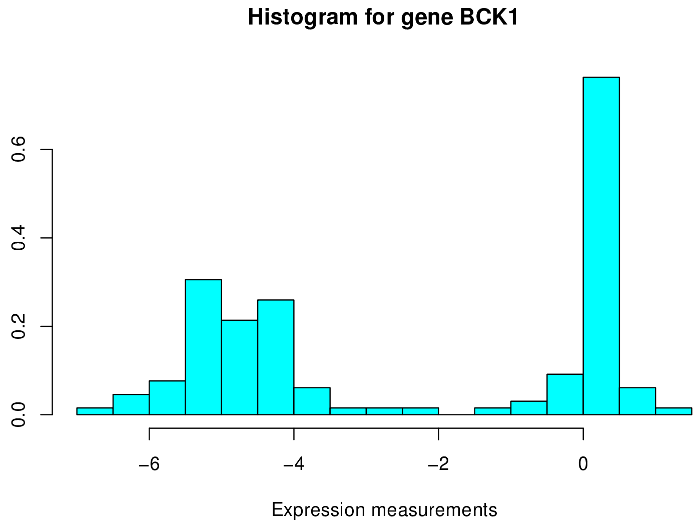

Figure 1: A histogram suggesting a bimodal distribution for the expression values of gene BCK1.

obvious departures from multivariate normality. This said, a Gaussian copula remains of course a strong assumption about the joint distribution.

We ran the PC algorithm using Pearson correlations, Spearman correlations as well as Kendall's τ. We considered a grid of values for α from 10^{-8} to 0.5 and selected α by optimizing the Bayesian information criterion (BIC) of Schwarz (1978). (Extensions in the spirit of Chen and Chen, 2008 and Foygel and Drton, 2010 could be attractive for this tuning problem but have yet to be adapted and studied for directed graphs.) The computations were done using routines from the aforementioned R package `pcalg` as well as the package `ggm` (Sadeghi and Marchetti, 2012). The former package offers, in particular, routines to create DAGs from the PC output and the latter package contains a routine to fit a DAG model by maximum likelihood.

For the case of Pearson correlations, tuning with BIC gave $\alpha = 0.5$ and a graph with 178 edges. Spearman correlations behaved similarly. No true optimum arose during the BIC tuning, which again suggested $\alpha = 0.5$ and led to a graph with 171 edges. For Kendall's τ on the other hand, the BIC was minimal for $\alpha = 0.1$ and only values in the range $[0.05, 0.1]$ gave comparable BIC values. The graph inferred for $\alpha = 0.1$ has 74 edges. We display its largest connected component in Figure 2.

Figure 2 was produced using output from TETRAD IV and features directed, undirected and bidirected edges. While the former two arise in CPDAGs, the latter type of edge indicates inconsistencies that the PC algorithm encountered. Briefly put, a bidirected edge arises when this edge appears in the skeleton inferred in the first stage of the PC algorithm but the edge orientation rules in the second stage of the algorithm yield arrowheads at both tails of the edge.

As mentioned in Sun and Li (2012), some prior biological knowledge about the pathway is available but not in a form that can be translated into a statistical model as considered here. Nevertheless, in this example, the use of Kendall's τ seems preferable to that of Pearson and also Spearman correlations. Both the sparsity of the inferred graph as well as the more favorable behavior in the likelihood computations underlying the BIC search speak for Kendall's τ.

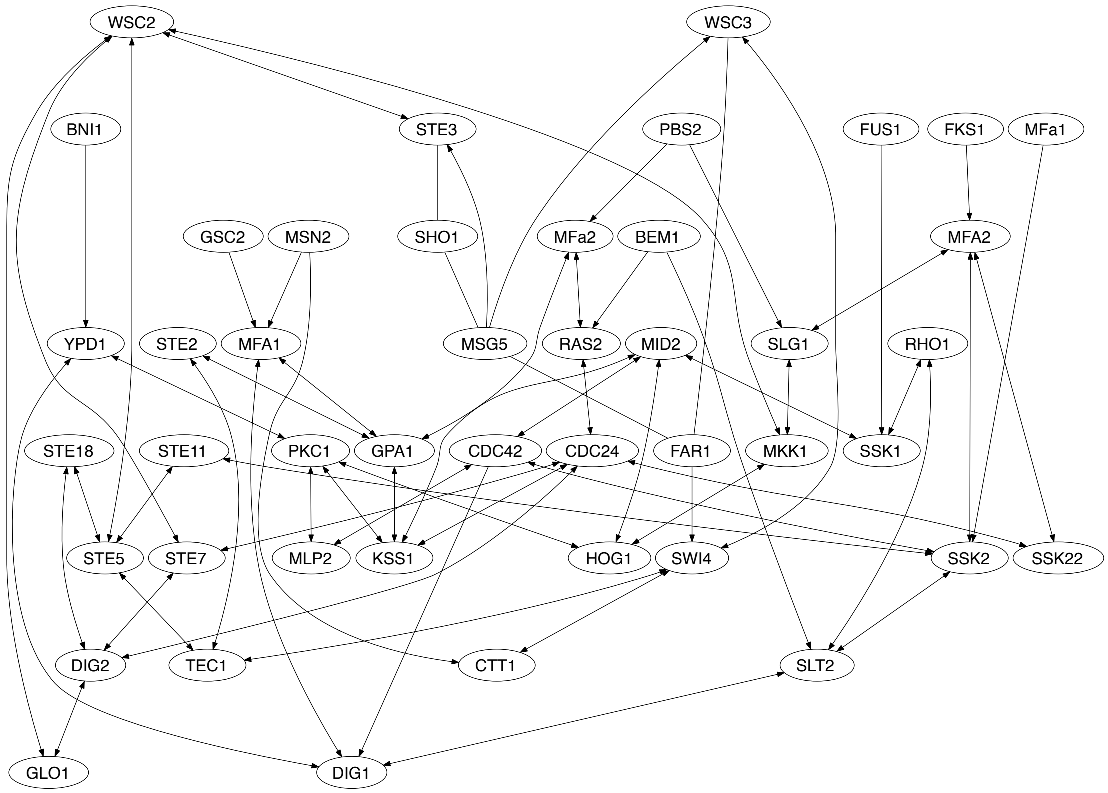

Figure 2: Largest connected component in the output of the Kendall RPC algorithm applied to expression data for genes in the MAPK pathway in yeast.

6. Conclusion

The PC algorithm of Spirtes et al. (2000) addresses the problem of model selection in graphical modelling with directed graphs via a clever scheme of testing conditional independences. For multivariate normal observations, the algorithm is known to have high-dimensional consistency properties when conditional independence is tested using sample partial correlations (Kalisch and Bühlmann, 2007). We showed that the PC algorithm retains these consistency properties when observations follow a Gaussian copula model and rank-based measures of correlation are used to assess conditional independence. The assumptions needed in our analysis are no stronger than those in prior Gaussian work when the considered sequence of DAGs has bounded degree. When the degree grows our assumptions are slightly more restrictive as our proof requires control of the conditioning of principal submatrices of correlation matrices that are inverted to estimate partial correlations in the rank-based PC (RPC) algorithm.

In our simulations, the use of the RPC algorithm led to negligible differences in statistical efficiency when data were indeed normal. For nonnormal data, RPC clearly outperformed the other considered versions of the algorithm. Since rank correlations take only marginally longer to compute than sample correlations, the simulations suggest that there are hardly any downsides associated with making RPC the standard version of the PC algorithm for continuous data.

Consistency results assume the data-generating distribution to be faithful to an underlying DAG. In fact, our results make the stronger assumption that non-zero partial correlations are sufficiently far from zero. As shown in Uhler et al. (2013), this can be a restrictive assumption, which provides an explanation for why consistency does not 'kick-in' quicker in simulation studies such as the one in Kalisch and Bühlmann (2007) and also ours.

Our analysis of the PC algorithm made use of two main arguments. First, for graphs with suitably bounded degree the population version of the PC algorithm only needs to check conditional independences with small conditioning sets. Second, the low-order partial correlations whose vanishing corresponds to these conditional independence can be estimated accurately. Lemma 4, which provides the error propagation from marginal to partial correlations, could similarly be used to analyze other algorithms that test the vanishing of low-order partial correlations. One example is the FCI algorithm that infers a more complex graphical object to deal with situations in which some relevant variables remain unobserved (Spirtes et al., 2000; Colombo et al., 2012).

Recent work shows that Kendall's τ can be used to obtain accurate estimates of the dispersion parameters in a more general setting of elliptical (rather than nonparanormal) distributions. Our analysis would again carry over to this case as an analogue to (5) is available in this setting. However, in the elliptical family zeros in the dispersion matrix do not correspond to independences and would have to be interpreted in terms of a latent normal random vector (Liu et al., 2012b).

Acknowledgments

We would like to thank three anonymous reviewers for carefully reading a draft manuscript and offering insightful suggestions which greatly improved our paper. Mathias Drton was supported by the NSF under Grant No. DMS-0746265 and by an Alfred P. Sloan Fellowship.

Appendix A. Sample Size Adjustment

We now show that the consistency result in Corollary 9 still holds when using the conditional independence tests from (15). In these tests, the sample size is adjusted from n to $n-|S|-3$.

Proof The test in (15) accepts a conditional independence hypothesis if and only if

$$|\hat{\rho}_{uv|S}| \;\leq\; \gamma(n,|S|,z), \tag{17}$$

where

$$\gamma(n,|S|,z) = \frac{\exp\left(z/\sqrt{n-|S|-3}\right)-1}{\exp\left(z/\sqrt{n-|S|-3}\right)+1}$$

and $z = z(\alpha) = 2\Phi^{-1}(1-\alpha/2)$. We need to find a sequence (α_n) of values for α such that consistency holds under the scaling assumptions made in Corollary 9. We will do this by specifying a sequence (z_n) for values for the (doubled) quantiles z.

We claim that the RPC algorithm using the tests from (17) is consistent when choosing the quantile sequence

$$z_n = \sqrt{n-3}\cdot\log\left(\frac{1+c_n/3}{1-c_n/3}\right), \tag{18}$$

where we use the abbreviation

$$c_n := c_{\min}(\Sigma_n, q_n).$$

We will show that as the sample size n tends to infinity, with probability tending to one, $|\hat{\rho}_{uv|S} - \rho_{uv|S}| < c_n/3$ for every $u,v \in V$ and $|S| \leq q_n$. Furthermore, we will show that for the above choice of z_n and all sufficiently large n, we have $c_n/3 \leq \gamma(n,|S|,z_n) \leq 2c_n/3$ for each relevant set S with $0 \leq |S| \leq q_n$. These facts imply that, with asymptotic probability one, every conditional independence test is correct, and the RPC algorithm succeeds.

First, we slightly adapt the proof of Theorem 8. Choosing the uniform error threshold for the correlation estimates as

$$\varepsilon = \frac{c\lambda^2}{(6+c)q+\lambda cq} > 0$$

in place of (13) yields that, with probability at least

$$1 - \frac{A}{2}p^2 \exp\left(-\frac{B\lambda^4 nc^2}{64q^2}\right), \tag{19}$$

we have that $|\hat{\rho}_{uv|S} - \rho_{uv|S}| < c/3$ for every $u,v \in V$ and $|S| \leq q$. When substituting p_n, q_n, c_n and $\lambda_{\min}(\Sigma_n, q_n)$ for p, q, c and λ, respectively, the scaling assumptions in Corollary 9 imply that the probability bound in (19) tends to one as $n \to \infty$, and we obtain the first part of our claim.

For the second part of our claim, note that our choice of z_n in (18) gives $\gamma(n,0,z_n) = c_n/3$. Since $\gamma(n,|S|,z)$ is monotonically increasing in $|S|$, we need only show that for sufficiently large n,

$$\gamma(n,q_n,z_n) - \gamma(n,0,z_n) \leq c_n/3.$$

For $x \geq 0$, the function

$$f(x) = \frac{\exp(x)-1}{\exp(x)+1}$$

is concave and, thus, for any $q_n \geq 0$,

$$\begin{aligned}\gamma(n,q_n,z_n) - \gamma(n,0,z_n) &= f\left(\frac{z}{\sqrt{n-q_n-3}}\right) - f\left(\frac{z}{\sqrt{n-3}}\right) \\ &\leq f'\left(\frac{z}{\sqrt{n-3}}\right)\left(\frac{z}{\sqrt{n-q_n-3}} - \frac{z}{\sqrt{n-3}}\right).\end{aligned} \tag{20}$$

The derivative of f is

$$f'(x) = \frac{2\exp(x)}{(\exp(x)+1)^2}.$$

Evaluating the right hand side of (20), we obtain that

$$\begin{aligned}\gamma(n,q_n,z_n) - \gamma(n,0,z_n) &\leq \frac{1}{2}\left(1-\frac{c_n^2}{9}\right)\log\left(\frac{1+c_n/3}{1-c_n/3}\right)\left(\frac{\sqrt{n-3}}{\sqrt{n-q_n-3}}-1\right) \\ &\leq \frac{1}{2}\log\left(\frac{1+c_n/3}{1-c_n/3}\right)\left(\frac{\sqrt{n-3}}{\sqrt{n-q_n-3}}-1\right).\end{aligned} \tag{21}$$

Being derived from absolute values of partial correlations, the sequence c_n is in $[0,1]$. Now, $\log[(1+x)/(1-x)]$ is a convex function of $x \geq 0$ that is zero at $x=0$ and equal to $\log(2)$ for $x = 1/3$. Therefore,

$$\frac{1}{2}\log\left(\frac{1+c_n/3}{1-c_n/3}\right) \leq \frac{1}{2}\log(2)\cdot c_n, \qquad c_n \in [0,1].$$

This shows that the bound in (21) is $o(c_n)$ because, by assumption, $q_n = o(\sqrt{n})$. In particular, the bound in (21) is less than $c_n/3$ for sufficiently large n, proving the claimed consistency result. ■

Appendix B. Background on Graphical Models

Let $G=(V,E)$ be an acyclic digraph with finite vertex set. We write $v \to w \in E$ to indicate that (v,w) is an edge in E. As mentioned in the introduction, the conditional independences associated with the graph G may be determined using d-separation; compare, for example, page 48 in Lauritzen (1996). We briefly review the concept.

Since a DAG contains at most one edge between any two nodes, we may define a path from a node u to a node v to be a sequence of distinct nodes $(v_0, v_1, \ldots, v_n)$ such that $v_0 = u$, $v_n = v$ and for all $1 \leq k \leq n$, either $v_{k-1} \to v_k \in E$ or $v_{k-1} \leftarrow v_k \in E$. Two distinct nodes u and v are then said to be *d-separated* by a set $S \subset V \setminus \{v,u\}$ if every path from u to v contains three consecutive nodes (v_{k-1}, v_k, v_{k+1}) for which one of the following is true:

(i) The three nodes form a chain $v_{k-1} \to v_k \to v_{k+1}$, a chain $v_{k-1} \leftarrow v_k \leftarrow v_{k+1}$, or a fork $v_{k-1} \leftarrow v_k \to v_{k+1}$, and the middle node v_k is in S.

(ii) The three nodes form a collider $v_{k-1} \to v_k \leftarrow v_{k+1}$, and neither v_k nor any of its descendants is in S.

Suppose A,B,S are pairwise disjoint subsets of V. Then S d-separates A and B if S d-separates any pair of nodes a and b with $a \in A$ and $b \in B$.

Two DAGs $G=(V,E)$ and $H=(V,F)$ with the same vertex set V are *Markov equivalent* if they may possess the same d-separation relations, that is, two sets A and B are d-separated given a third set C in the graph G if and only if the same holds in H. To give an example, the graphs $u \to v \to w$ and $u \leftarrow v \leftarrow w$ are Markov equivalent, but $u \to v \to w$ and $u \to v \leftarrow w$ are not. As first shown in Verma and Pearl (1991), two DAGs G and H are Markov equivalent if and only if they have the same skeleton and the same unshielded colliders. The *skeleton* of a digraph G is the undirected graph obtained by converting each directed edge into an undirected edge. An *unshielded collider* is a triple of nodes (u,v,w) that induces the subgraph $u \to v \leftarrow w$, that is, there is no edge between u and w.

Let $[G]$ be the Markov equivalence class of an acyclic digraph $G=(V,E)$. Write $E(H)$ for the edge set of a DAG H, and define the edge set

$$[E] = \bigcup_{H \in [G]} E(H).$$

That is, $(v,w) \in [E]$ if there exists a DAG $H \in [G]$ with the edge $v \to w$ in its edge set. We interpret the presence of both (v,w) and (w,v) in $[E]$ as an undirected edge between v and w. The graph

$C(G) = (V, [E])$ is known as the *completed partially directed acyclic graph* (CPDAG) for G or also as the *essential graph*. Two DAGs G and H satisfy $C(G) = C(H)$ if and only if $[G] = [H]$, making the CPDAG a useful graphical representation of a Markov equivalence class; see Andersson et al. (1997) and Chickering (2002).

References

Theodore. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, third edition, 2003.

Steen A. Andersson, David Madigan, and Michael D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *Ann. Statist.*, 25(2):505–541, 1997.

Rachel B. Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of National Academy of Sciences*, 102:1572–1577, 2005.

Jiahua Chen and Zehua Chen. Extended Bayesian information criterion for model selection with large model sp ace. *Biometrika*, 95:759–771, 2008.

David Maxwell Chickering. Learning equivalence classes of Bayesian-network structures. *J. Mach. Learn. Res.*, 2(3):445–498, 2002.

David Christensen. Fast algorithms for the calculation of Kendall's τ. *Comput. Statist.*, 20(1):51–62, 2005.

Diego Colombo, Marloes H. Maathuis, Markus Kalisch, and Thomas S. Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, 40(1): 294–321, 2012.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on Algebraic Statistics*, volume 39 of *Oberwolfach Seminars*. Birkhäuser Verlag, Basel, 2009.

Rina Foygel and Mathias Drton. Extended Bayesian information criteria for Gaussian graphical models. *Adv. Neural Inf. Process. Syst.*, 23:2020–2028, 2010.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original.

Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, 8:613–636, May 2007.

Markus Kalisch and Peter Bühlmann. Robustification of the PC-algorithm for directed acyclic graphs. *J. Comput. Graph. Statist.*, 17(4):773–789, 2008.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H. Maathuis, and Peter Bühlmann. Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11):1–26, 5 2012.

Steffen L. Lauritzen. *Graphical Models*, volume 17 of *Oxford Statistical Science Series*. The Clarendon Press Oxford University Press, New York, 1996.

Thuc Duy Le, Lin Liu, Anna Tsykin, Gregory J. Goodall, Bing Liu, Bing-Yu Sun, and Jiuyong Li. Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics*, 29(6):765–771, 2013.

Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: semiparametric estimation of high dimensional undirected graphs. *J. Mach. Learn. Res.*, 10:2295–2328, 2009.

Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *Ann. Statist.*, 40(4):2293–2326, 2012a.

Han Liu, Fang Han, and Cun-hui Zhang. Transelliptical graphical models. *Adv. Neural Inf. Process. Syst.*, 25:800–808, 2012b.

Marloes H. Maathuis, Diego Colombo, Markus Kalisch, and Peter Bühlmann. Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248, 2010.

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, second edition, 2009. Models, reasoning, and inference.

Kayvan Sadeghi and Giovanni M. Marchetti. Graphical Markov models with mixed graphs in R. *The R Journal*, 4(2):65–73, December 2012.

Markus Schmidberger, Sabine Lennert, and Ulrich Mansmann. Conceptual aspects of large meta-analyses with publicly available microarray data: A case study in oncology. *Bioinformatics and Biology Insights*, 5:13–39, 2011.

Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, second edition, 2000. With additional material by David Heckerman, Christopher Meek, Gregory F. Cooper and Thomas Richardson, A Bradford Book.

Hokeun Sun and Hongzhe Li. Robust Gaussian graphical modeling via l_1 penalization. *Biometrics*, 68:1197–1206, 2012.

Caroline Uhler, Garvesh Raskutti, Bin Yu, and Peter Bühlmann. Geometry of faithfulness assumption in causal inference. *Ann. Statist.*, 41(2):436–463, 2013.

Ricardo A. Verdugo, Tanja Zeller, Maxime Rotival, Philipp S. Wild, Thomas Münzel, Karl J. Lackner, Henri Weidmann, Ewa Ninio, David-Alexandre Trégouët, François Cambien, Stefan Blankenberg, and Laurence Tiret. Graphical modeling of gene expression in monocytes suggests molecular mechanisms explaining increased atherosclerosis in smokers. *PLoS ONE*, 8(1): e50888, 2013.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, UCLA, 1991.