

Estimating and Controlling the False Discovery Rate of the PC Algorithm Using Edge-Specific P-Values

Eric V. Strobl

*Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA 15206, USA*

EVS17@PITT.EDU

Peter L. Spirtes

*Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213, USA*

PS7Z@ANDREW.CMU.EDU

Shyam Visweswaran

*Department of Biomedical Informatics
University of Pittsburgh
Pittsburgh, PA 15206, USA*

SHV3@PITT.EDU

Editor: TBA

Abstract

The PC algorithm allows investigators to estimate a complete partially directed acyclic graph (CPDAG) from a finite dataset, but few groups have investigated strategies for estimating and controlling the false discovery rate (FDR) of the edges in the CPDAG. In this paper, we introduce PC with p-values (PC-p), a fast algorithm which robustly computes edge-specific p-values and then estimates and controls the FDR across the edges. PC-p specifically uses the p-values returned by many conditional independence (CI) tests to upper bound the p-values of more complex edge-specific hypothesis tests. The algorithm then estimates and controls the FDR using the bounded p-values and the Benjamini-Yekutieli FDR procedure. Modifications to the original PC algorithm also help PC-p accurately compute the upper bounds despite non-zero Type II error rates. Experiments show that PC-p yields more accurate FDR estimation and control across the edges in a variety of CPDAGs compared to alternative methods¹.

Keywords: PC Algorithm, Causal Inference, False Discovery Rate, Bayesian Network, Directed Acyclic Graph

1. Introduction

Discovering causal relationships is often much more important than discovering associational relationships in the sciences. As a result, the research community has been conducting extensive investigations into causal inference with the hope of developing practically useful algorithms to speed-up the scientific process. This research has resulted in a wide range of high performing algorithms over the years such as PC (Spirtes et al., 2000), FCI (Spirtes et al., 1995, 2000), and CCD (Richardson, 1996)

1. MATLAB implementation: <https://github.com/ericstrobl/PCp/>

The PC algorithm is currently one of the most popular methods for inferring causation from observational data. Given an observational dataset, the algorithm outputs a complete partially directed acyclic graph (CPDAG) which has helped some investigators elucidate important causal relationships in several domains. For example, the PC algorithm has been used to discover new causal relationships between genes and brain regions in biology (Wu and Ye, 2006; Li et al., 2008; Joshi et al., 2010; Sun et al., 2012; Harris and Drton, 2013; Iyer et al., 2013; Le et al., 2013; Teramoto et al., 2014; Ha et al., 2015). The algorithm has also been used to discover causal relations between corporate structures and strategies in economics (Chong et al., 2009) as well as academic and musical achievements in psychology (Mullensiefen et al., 2015).

The increased use of PC in recent years has nonetheless led to growing concern about algorithm’s confidence level in each edge of the CPDAG. For example, PC may have more confidence in the edge $A - B$ but have less confidence in the edge $B \rightarrow C$ in the subgraph $A - B \rightarrow C$ of the CPDAG. Currently, PC alone does not output any edge-specific measure of confidence, even though scientists often must report measures of confidence such as p-values or confidence intervals in their scientific articles. This incongruity has resulted in the relatively slow adoption or even avoidance of the PC algorithm in the sciences, despite the algorithm’s impressive capabilities in causal inference. Clearly then rectifying the problem by developing an edge-specific measure of confidence will increase the adoption of PC as well as hopefully ease the transition of ever-more complex causal inference algorithms into the scientific community.

We can of course consider multiple different ways of representing the confidence level in each edge of the CPDAG. However, we choose to pay special attention to the p-value, since it is by far the most popular notion of confidence in the sciences. Indeed, nearly all scientists report p-values in modern scientific reports because they rely on p-values to help justify their hypotheses. We therefore would ideally like to assign a p-value to each edge in the CPDAG as in Figure 1a in order to best integrate the algorithm within a well-known framework. In this paper, we propose such a “causal p-value” in detail.

We however also believe that assigning p-values to each edge is not enough to ease the transition of PC into mainstream science, since the CPDAG actually contains many edges and therefore also represents a complicated multiple hypothesis testing problem. Fortunately, the problem of multiple hypothesis testing has a long history, as scientists have often required the results of multiple hypothesis tests in order to answer complex scientific questions. Currently, a standard approach to tackling the multiple hypothesis testing problem involves controlling the proportion of false positives among the rejected null hypotheses, or the *false discovery rate (FDR)*, by using an *FDR controlling procedure* that takes a desired FDR level q and a set of p-values as input. The procedure then outputs a corresponding significance level α^* for the set of p-values. An investigator subsequently rejects the null hypotheses for those tests with p-values that fall below α^* in order to ensure that the expected FDR does not exceed q . For example, consider the set of p-values $\{0.02, 0.01, 0.03\}$ and suppose that the FDR controlling procedure with $q = 0.1$ outputs $\alpha^* = 0.019$. Then, rejecting the null hypotheses of the first and third hypothesis tests guarantees that the expected FDR does not exceed 10%. Several other FDR controlling strategies also exist, but the ease of use, speed and accuracy of the above method have made it the most widely adopted strategy in the last two decades.

We would therefore like to control the FDR in the edges of the CPDAG using an FDR controlling procedure like in Figure 1b. As a first idea, one may wonder whether an investigator can control the FDR in the CPDAG by simply feeding in all of the CI test p-values computed by PC into an FDR controlling procedure. Unfortunately, this approach fails for at least two reasons. First, it is unclear how to use the p-values which exceed the α^* cut-off to reject edges in the CPDAG, since one would first need to elucidate the correspondence between the p-values and edges. Second, even if one could solve this problem, the strategy may only loosely bound the expected FDR. An accurate FDR controlling procedure should instead take into account the specific computations executed by PC in order to identify a sharp bound. The p-value based approach therefore necessitates a more fine-grained strategy which has thus far remained undiscovered.

Several groups have nonetheless attempted to control the FDR in the CPDAG by avoiding the complicated nature of the above problem with a different, data re-sampling approach. For example, Friedman and colleagues proposed to estimate the FDR by using the parametric bootstrap (Friedman et al., 1999). This procedure involves first learning a causal graph with the PC algorithm. The procedure then generates data from the causal graph and re-applies the PC algorithm multiple times on each generated dataset to estimate the FDR using the learnt causal graphs. An investigator can subsequently control the FDR by repeating the above process with different α values until he or she reaches the desired FDR level q . However, notice that the method requires multiples calls to PC and can therefore require too much time with high dimensional data. The procedure also requires parametric knowledge about the underlying distribution which limits the applicability of the method to simple cases. Fortuitously, two groups later proposed a permutation-based method which drops the parametric assumption (Listgarten and Heckerman, 2007; Armen and Tsamardinos, 2014). The permutation method nevertheless also requires multiple calls

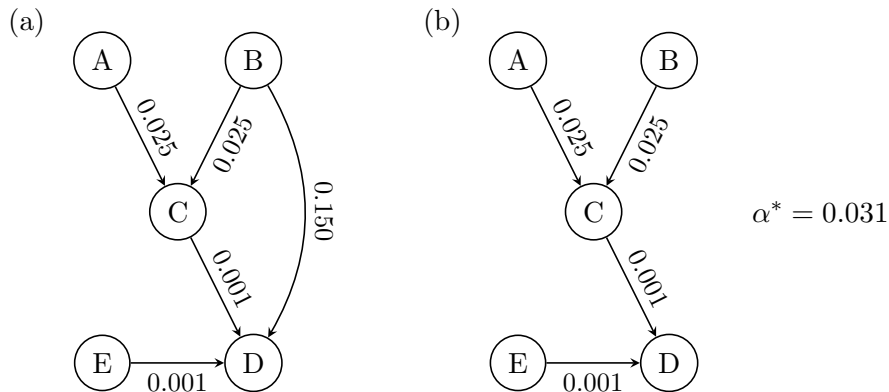


Figure 1: We seek to associate edge-specific p-values to the output of the PC algorithm such as in (a). The PC algorithm currently does not associate such p-values with its output. We would also like to control the FDR of the edges. In (b), we set the FDR to 0.1 and obtained a α^* cutoff of 0.031 for the output in (a), so we eliminated the edge between B and D because its p-value exceeds α^* .

to an algorithm and in fact only applies to the parts of PC which can be decomposed into independent searches for the parents of each vertex; this has thus far limited the applicability of the method to adjacency discovery with local to global discovery algorithms (e.g., MMHC) and incomplete edge orientation. We conclude that both the bootstrap and permutation approaches to FDR estimation and control are either incomplete or too time consuming.

Another class of methods fortunately attempts to control the FDR without resampling procedures by instead using a standard FDR controlling procedure with bounded p-values. For instance, one method proposed in (Tsamardinos and Brown, 2008) and then refined in (Armen and Tsamardinos, 2011, 2014) assigns a p-value to each adjacency by taking the maximum over all of the significant p-values from the associated CI tests executed by PC. The method then controls the FDR in the estimated adjacencies by applying an FDR controlling procedure, such as the one proposed by Benjamini and Yekutieli (BY) (Benjamini and Yekutieli, 2001), on the edge-specific p-values. Under faithfulness and a zero Type II error rate, the method controls the FDR across the estimated adjacencies, or the estimated *skeleton* (Armen and Tsamardinos, 2014). This two stage method also performs comparably with the one stage method proposed in (Li et al., 2008; Li and Wang, 2009), which controls the FDR during, as opposed to after, the execution of the skeleton discovery phase of the PC algorithm. Of course, the Type II error rate never reaches zero in practice but researchers have also investigated a strategy for reducing the realized Type II error rate by introducing a heuristic reliability criterion for CI tests when dealing with discrete data (Armen and Tsamardinos, 2014). Experiments have shown that these methods finish in a relatively short amount of time and perform well in practice. However, the methods are also incomplete because they only apply to the skeleton discovery phase of PC.

In this report, we build on the previous outstanding work for deriving p-values for adjacencies by contributing a sound, complete and fast algorithm called PC with p-values (PC-p) which appropriately combines the p-values of PC’s CI tests and then uses the BY FDR controlling procedure to accurately control the FDR in a CPDAG. The method relies on two upper bounds of the p-value that relate to logical conjunctions and disjunctions as described in Section 3. These upper bounds allow us to formulate several hypothesis tests for recovering the skeleton, discovering unshielded v-structures, and orienting additional edges as presented in Section 4. Accurately estimating the p-values of the hypothesis tests nonetheless requires a modified version of PC called PC-p which we propose in Section 5. Finally, we provide experimental results in Section 6 which show that PC-p’s p-value estimates yield accurate estimates of the FDR with the BY procedure and improve upon alternative methods.

2. Preliminaries

2.1 Causal graphs

A *causal graph* consists of vertices representing variables and edges representing causal relationships between any two variables. In this paper, we will use the terms “vertices” and “variables” interchangeably. *Directed graphs* are graphs where two distinct vertices can be connected by edges “ \rightarrow ” and “ \leftarrow .” We only consider *simple graphs* in this paper, or graphs

with no edges originating from and connecting to the same vertex. *Directed acyclic graphs* (DAGs) are directed graphs without directed cycles. We say that X and Y are *adjacent* if they are connected by an edge independent of the edge’s direction. A *path* p from X to Y is a set of consecutive edges (also independent of their direction) from X to Y such that no vertex is visited more than once. Given a path between two vertices X and Y with a middle vertex Z , the path is a *chain* if $X \rightarrow Y \rightarrow Z$, a *fork* if $X \leftarrow Y \rightarrow Z$, and a *v-structure* if $X \rightarrow Y \leftarrow Z$. We refer to Y as a *collider*, if it is the middle vertex in a v-structure. A v-structure is called an *unshielded v-structure* if $X \rightarrow Y \leftarrow Z$, but X and Z are non-adjacent. A *directed path* from X to Y is a set of consecutive edges with direction. We say that X is an *ancestor* of Y (and Y is a *descendant* of X), if there exists a directed path from X to Y .

If \mathbb{G} is a directed graph in which \mathbf{X} , \mathbf{Y} and \mathbf{Z} are disjoint sets of vertices, then \mathbf{X} and \mathbf{Y} are *d-connected* by \mathbf{Z} in \mathbb{G} if and only if there exists an undirected path p between some vertex in \mathbf{X} and some vertex in \mathbf{Y} such that, for every collider C on p , either C or a descendant of C is in \mathbf{Z} , and no non-collider on p is in \mathbf{Z} . On the other hand, \mathbf{X} and \mathbf{Y} are *d-separated* by \mathbf{Z} in \mathbb{G} if and only if they are not d-connected by \mathbf{Z} in \mathbb{G} . Next, the joint probability distribution \mathbb{P} over variables \mathbf{X} satisfies the *global directed Markov property* for a directed graph \mathbb{G} if and only if, for any three disjoint subsets of variables \mathbf{A} , \mathbf{B} and \mathbf{C} from \mathbf{X} , if \mathbf{A} and \mathbf{B} are d-separated given \mathbf{C} in \mathbb{G} , then \mathbf{A} and \mathbf{B} are conditionally independent given \mathbf{C} in \mathbb{P} . We refer to the converse of the global directed Markov property as *d-separation faithfulness*; that is, if \mathbf{A} and \mathbf{B} are conditionally independent given \mathbf{C} in \mathbb{P} , then \mathbf{A} and \mathbf{B} are d-separated given \mathbf{C} in \mathbb{G} .

A *Markov equivalence class* of DAGs refers to a set of DAGs which entail the same conditional independencies. A *complete partially directed acyclic graph* (CPDAG) is a partially directed acyclic graph with the following properties: (1) each directed edge exists in every DAG in the Markov equivalence class, and (2) there exists a DAG with $X \rightarrow Y$ and a DAG with $X \leftarrow Y$ in the Markov equivalence class for every undirected edge $X - Y$. A CPDAG \mathbb{G}^C represents a DAG \mathbb{G} , if \mathbb{G} belongs to the Markov equivalence class described by \mathbb{G}^C . We will occasionally use the meta-symbol “ \circ ” at the endpoint(s) of an edge to denote the presence or absence of an arrowhead. For example, the edge “ \neg ” may denote either “ $-$ ” or “ \rightarrow ”.

2.2 The PC Algorithm

The PC algorithm is comprised of three stages. We have summarized these stages as pseudocode in Algorithms 5, 6 and 7 in Section A.1 of the Appendix. The first stage estimates the adjacencies of \mathbb{G} , or the skeleton of \mathbb{G} . Starting with a fully connected skeleton, the algorithm attempts to eliminate the adjacency between any two variables, say A and B , by testing if A and B are conditionally independent given some subset of the neighbors of A or the neighbors of B . The search is performed progressively, whereby the algorithm increases the size of the conditioning set starting from zero using a step size of 1. The edge between A and B is removed, if A and B are rendered conditionally independent given some subset of the neighbors of A or the neighbors of B .

The PC algorithm orients unshielded colliders in its second stage. Specifically, PC finds triples A, B, C such that $A - B - C$, but A and C are non-adjacent. The algorithm

then determines whether B is contained in the set which rendered A and C conditionally independent in the first stage of PC. If not, $A - B - C$ is replaced with $A \rightarrow B \leftarrow C$.

The third and final stage of PC involves the repetitive application of three rules to orient as many of the remaining undirected edges as possible. The three rules include:

1. If $A - B$, $C \rightarrow A$ and C and B are non-adjacent, then replace $A - B$ with $A \rightarrow B$.
2. If $A - B$ and $A \rightarrow C \rightarrow B$, then replace $A - B$ with $A \rightarrow B$. (1)
3. If $A - B$, $A - C \rightarrow B$, $A - D \rightarrow B$, and C and D are non-adjacent, then replace $A - B$ with $A \rightarrow B$.

Overall, the PC algorithm has been shown to be complete in the sense that it finds and then orients edges up to \mathbb{G}^C , a CPDAG that represents \mathbb{G} (Meek, 1995).

2.3 Hypothesis Testing

A *hypothesis test* is a method of statistical inference usually composed of one *null* (H_0) and one *alternative* (H_1) *hypothesis* which are mutually exclusive; that is, if one occurs, then the other cannot occur. The null hypothesis refers to the default position which asserts that whatever one is trying to statistically infer actually did not happen. Note that the null and alternative do not necessarily need to be logical complements of each other. For example, one may be interested in determining whether the parameter μ is greater than zero. In this case, the null can be defined as $\mu = 0$ while the alternative can be defined as $\mu > 0$ instead of $\mu \neq 0$.

A *Type I error* is the incorrect rejection of a true null hypothesis, or a false positive. On the other hand, a *Type II error* is the failure to reject a false null hypothesis, or a false negative. The *p-value* (p) is the probability of the Type I error, or the Type I error rate. More specifically, the p-value is the probability of obtaining a result equal to or more extreme than the observed value under the assumption of the null hypothesis. The null hypothesis is thus rejected when the p-value is at or below a predefined α *threshold* (typically the α threshold is set to 0.05), because a low p-value demonstrates the improbability of the null hypothesis.

2.4 False Discovery Rate

Multiple comparisons or *multiple hypothesis testing* refers to the process of considering more than one statistical inference simultaneously. Failure to compensate for multiple comparisons can result in erroneous inferences. For example, if an investigator performs one hypothesis test with an α threshold of 0.05, then he or she has only a 5% chance of making a Type I error. However, if the investigator performs 100 independent tests with the same α threshold, then he or she has a $1 - (1 - 0.05)^{100} = 99.4\%$ chance of making a Type I error on at least one test.

In multiple hypothesis testing, the *false discovery rate* (FDR) at threshold α is the expected proportion of false positives among the rejected null hypotheses. Specifically, we

define the FDR at α as follows:

$$FDR(\alpha) \triangleq \mathbb{E} \left[\frac{V}{\max\{R, 1\}} \right],$$

where V is the number of false positives, R is the total number of null hypotheses rejected, and $\max\{R, 1\}$ ensures that $FDR(\alpha)$ is well-defined when $R = 0$. We define the *realized FDR* at α as $V/\max\{R, 1\}$.

FDR estimation, or *conservative point estimation of the FDR*, refers to the process of estimating $FDR(\alpha)$ in a conservative manner such that:

$$\mathbb{E}[\widehat{FDR}(\alpha)] \geq FDR(\alpha),$$

where $\widehat{FDR}(\alpha)$ represents an estimate of $FDR(\alpha)$. We denote $\mathbb{E}[\widehat{FDR}(\alpha)] - FDR(\alpha)$ as the *estimation bias*. Note that there are several ways of obtaining $\widehat{FDR}(\alpha)$. In 2001, Benjamini and Yekutieli proposed the following *FDR estimator* for m hypothesis tests:

$$\widehat{FDR}_{BY}(\alpha) \triangleq \frac{m\alpha \sum_{i=1}^m \frac{1}{i}}{\max\{R, 1\}}. \quad (2)$$

FDR estimators such as \widehat{FDR}_{BY} can be used to define *FDR controlling procedures*. These procedures determine the optimal threshold α^* which achieves *strong control*² of the FDR in the following sense:

$$\alpha^* \triangleq \arg \max_{\alpha} \{\widehat{FDR}(\alpha) \leq q\} \quad (3)$$

The FDR controlling procedure based on \widehat{FDR} involves the rejection of all null hypotheses with p-values below the α^* threshold. We refer to the quantity $FDR(\alpha^*) - q$ as the *control bias*. Benjamini and Yekutieli proved that the estimate \widehat{FDR}_{BY} in particular achieves strong control of the FDR with any form of dependence among the p-values of m hypothesis tests.

3. Upper Bounds on the P-Value

We present two upper bounds of the Type I error rate of hypothesis tests which can be constructed using a set of simpler hypothesis tests. These upper bounds will serve as useful tools in Section 4 for bounding the Type I error rate of the hypothesis tests which will be used to infer the presence or absence of edges in a CPDAG.

3.1 Union Bound

Consider the following hypothesis test for two random variables given a conditioning set:

$$\begin{aligned} H_0 &: \text{Conditionally independent,} \\ H_1 &: \text{Conditionally dependent.} \end{aligned}$$

2. *Strong control* of the FDR refers the process of controlling the FDR under any configuration of true and false null hypotheses; on the other hand, *weak control* refers to the process of controlling the FDR when all of the null hypotheses are true. Strong control is therefore preferable to weak control.

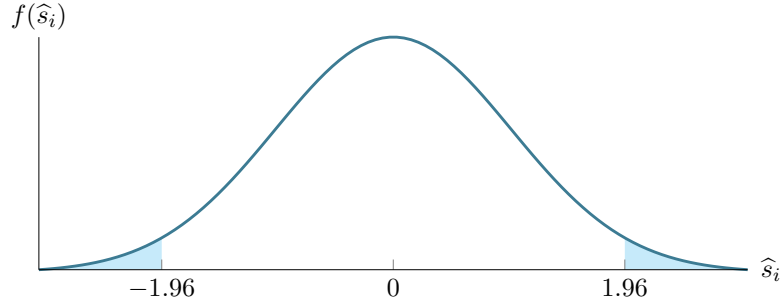


Figure 2: In the above standard normal case, we have $\Pr(|\hat{s}_i| \geq s_i^\alpha | s_i = 0) = 0.05$, where $s_i^\alpha = 1.96$. We reject the null hypothesis when $|\hat{s}_i|$ falls in the blue colored regions at the tails.

Trivially, we can rephrase the null and alternative in terms of a conditional independence (CI) oracle:

$$\begin{aligned} H_0 &: \text{The CI oracle outputs independent,} \\ H_1 &: \text{The CI oracle outputs dependent.} \end{aligned}$$

Now suppose we want to query m CI oracles about m CI relations. We can then consider the following null and alternative:

$$\begin{aligned} H_0 &: \text{All CI oracles output independent,} \\ H_1 &: \text{At least one CI oracle outputs dependent.} \end{aligned} \tag{4}$$

From here on, we write $\Pr(\text{CI test } i \text{ outputs dependent} \mid \text{CI oracle } i \text{ outputs independent})$ to denote $\Pr(|\hat{s}_i| \geq s_i^\alpha | s_i = 0)$, where s_i refers to a parameter of some standardized distribution used by CI test i , \hat{s}_i a random variable and the test statistic estimating s_i , and s_i^α a value of s_i determined by an α level. We provide an example in Figure 2, where s_i may correspond to Fisher's z -statistic in the case of Fisher's z -test for the mean parameter $s_i = 0$ of the standard normal distribution.

We now bound the Type I error rate of the hypothesis test (4) by using the new notation and the union bound:

$$\begin{aligned}
 & \Pr(\text{Type I error}) \\
 &= \Pr(\text{at least one CI test outputs dependent} | H_0) \\
 &= \Pr\left(\bigvee_{i=1}^m \text{CI test } i \text{ outputs dependent} | H_0\right) \\
 &\leq \sum_{i=1}^m \Pr(\text{CI test } i \text{ outputs dependent} | H_0) \\
 &= \sum_{i=1}^m \Pr(\text{CI test } i \text{ outputs dependent} | \text{CI oracle } i \text{ outputs independent}) \\
 &= \sum_{i=1}^m p_i,
 \end{aligned} \tag{5}$$

where p_i denotes the Type I error rate of CI test i . Thus, if the r.h.s. of (5) is less than the α threshold, then we can conclude that the Type I error rate of (4) is also below the threshold. In other words, (5) is a conservative p-value.

Note that the third equality in the derivation of (5) uses the simplifying assumption that the probability of the output of CI test i only depends on the output of CI oracle i when given the outputs of all CI oracles. Several papers have used this assumption implicitly in their proofs (Tsamardinos and Brown, 2008; Li and Wang, 2009), and we will also use it throughout this paper. We can justify the assumption based on three facts. First, most CI test statistics s_i have a limiting distribution which only depends on $s_i = 0$ under the null. For example, Fisher's z -statistic has a limiting standard normal distribution with mean parameter $z_i = 0$ and constant variance. Moreover, the G -statistic for the G -test has a limiting χ^2 -distribution with non-centrality parameter $g_i = 0$ and degrees of freedom determined by the number of cells in the contingency table. Second, existing methods which utilize bounds based on the assumption have strong empirical performance; loose-enough bounds therefore appear to accommodate the assumption well in most finite sample cases. Third, recall that simplifying assumptions are not new in the causality literature; indeed, many authors have made simplifying assumptions regarding parameter independence for Bayesian methods which similarly increase computational efficiency and achieve strong empirical performance (e.g., (Cooper and Yoo, 1999)).

We can now also generalize the bound in (5) to any hypothesis test consisting of a series of logical disjunctions in the alternative and a series of logical conjunctions in the null. Namely:

$$\begin{aligned}
 H_0 &: \bigwedge_{i=1}^m \text{oracle } i \text{ outputs } \neg P_i, \\
 H_1 &: \bigvee_{i=1}^m \text{oracle } i \text{ outputs } P_i,
 \end{aligned} \tag{6}$$

where P_i denotes an arbitrary output of oracle i . We now have:

$$\begin{aligned}
\Pr(\text{Type I error}) &= \Pr\left(\bigvee_{i=1}^m \text{test } i \text{ outputs } P_i \mid H_0\right) \\
&\leq \sum_{i=1}^m \Pr(\text{test } i \text{ outputs } P_i \mid H_0) \\
&= \sum_{i=1}^m \Pr(\text{test } i \text{ outputs } P_i \mid \text{oracle } i \text{ outputs } \neg P_i) \\
&= \sum_{i=1}^m h_i,
\end{aligned}$$

where h_i is the Type I error rate of test i , and the second equality uses the assumption that the probability of the output of test i only depends on the output of oracle i when given all oracles. We will use this generalization in Section 4.

3.2 Intersection Bound

Suppose we want to perform a hypothesis test with the following null and alternative which are different than the null and alternative in (4):

$$\begin{aligned}
H_0 &: \text{At least one CI oracle outputs independent,} \\
H_1 &: \text{All CI oracles output dependent.}
\end{aligned} \tag{7}$$

Now assume we know that the i^{th} CI oracle outputs independent. We can then bound the Type I error rate of (7) as follows with m queries to the CI oracle:

$$\begin{aligned}
\Pr(\text{Type I error}) &= \Pr(\text{all } m \text{ CI tests output dependent} \mid H_0) \\
&\leq \Pr(\text{CI test } i \text{ outputs dependent} \mid H_0) \\
&= \Pr(\text{CI test } i \text{ outputs dependent} \mid \text{CI oracle } i \text{ outputs independent} \\
&\quad \wedge \text{other CI oracles may output independent}) \\
&= \Pr(\text{CI test } i \text{ outputs dependent} \mid \text{CI oracle } i \text{ outputs independent}) \\
&= p_i,
\end{aligned} \tag{8}$$

where the second equality again holds under the assumption that the probability of the output of CI test i only depends on the output of CI oracle i when given the outputs of all CI oracles. We can therefore bound the Type I error rate of (7) using the p-value of a single CI test for which the CI oracle outputs independent. Nevertheless, in practice, we often do not know for which query the oracle outputs independent in the null. We do however know that at least one unknown CI oracle i outputs independent, so we can bound the Type I error rate of (7) using the maximum over all of the m CI test p-values:

$$\begin{aligned}
\Pr(\text{Type I error}) &= \Pr(\text{all } m \text{ CI tests output dependent} \mid H_0) \\
&\leq \Pr(\text{CI test } i \text{ outputs dependent} \mid \text{CI oracle } i \text{ outputs independent}) \\
&= p_i \leq \max_{j=1, \dots, m} p_j.
\end{aligned} \tag{9}$$

Note that we can generalize the above bound to any hypothesis test consisting of a series of logical conjunctions in the alternative and a series of logical disjunctions in the null. Namely:

$$\begin{aligned} H_0 : & \bigvee_{i=1}^m \text{oracle } i \text{ outputs } \neg P_i, \\ H_1 : & \bigwedge_{i=1}^m \text{oracle } i \text{ outputs } P_i. \end{aligned} \tag{10}$$

We therefore have:

$$\begin{aligned} \Pr(\text{Type I error}) &= \Pr\left(\bigwedge_{i=1}^m \text{test } i \text{ outputs } P_i \mid H_0\right) \\ &\leq \Pr(\text{test } i \text{ outputs } P_i \mid H_0) \\ &= \Pr(\text{test } i \text{ outputs } P_i \mid \text{oracle } i \text{ outputs } \neg P_i) \\ &= h_i \leq \max_{j=1, \dots, m} h_j, \end{aligned} \tag{11}$$

where the second equality again uses the assumption that the probability of the output of test i only depends on the output of oracle i when given all oracles.

4. Edge-Specific Hypothesis Tests

We now show how to apply the two upper bounds of the Type I error rate to derive p-value estimates for both the undirected and directed edges in the CPDAG as estimated by PC. Bounding the p-value for each edge therefore amounts to adding up and/or maximizing over the p-values returned from multiple CI tests.

Note that we will sometimes invoke a zero Type II error rate assumption in this section. This assumption is necessary to correctly upper bound the p-values of the edge-specific hypothesis tests of the CPDAG according to the CI tests executed by the PC algorithm. In fact, we can always correctly bound the p-values, if we perform all of the possible CI tests between the considered variables; however, this approach is impractical, since it ignores the efficiencies of the PC algorithm. A more interesting strategy involves designing the edge-specific hypothesis tests so that the p-value bounds are robust to Type II errors as well as redesigning the PC algorithm to catch many Type II errors. We will discuss these approaches in detail in Sections 4.4 and 5, so we encourage readers to accept the zero Type II error rate assumption for now.

4.1 Skeleton Discovery

We first consider the skeleton discovery phase of the PC algorithm. We wish to test whether each edge is absent in the true skeleton starting from a completely connected undirected graph. This problem has already been investigated in (Li et al., 2008; Tsamardinos and Brown, 2008; Li and Wang, 2009; Armen and Tsamardinos, 2011, 2014), but we review it here for completeness. We construct a hypothesis test with the following null and alterna-

tive:

$$\begin{aligned} H_0 : A - B \text{ is absent,} \\ H_1 : A - B \text{ is present.} \end{aligned} \tag{12}$$

Now consider the following proposition, where $\mathbf{Pa}(A)$ denotes the true parents of A :

Proposition 1 (*Spirites et al., 2000*) *Consider a DAG \mathbb{G} which satisfies the global directed Markov property. Moreover, assume that the probability distribution is d-separation faithful. Then, there is an edge between two vertices A and B if and only if A and B are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ and any subset of $\mathbf{Pa}(B) \setminus A$.*

We thus consider the following two scenarios for the undirected edge $A - B$:

1. If A and B are conditionally independent given some subset of $\mathbf{Pa}(A) \setminus B$ or some subset of $\mathbf{Pa}(B) \setminus A$, then $A - B$ is absent.
 2. If A and B are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ and any subset of $\mathbf{Pa}(B) \setminus A$, then $A - B$ is present.
- (13)

The following null and alternative are therefore equivalent to (12), where CI oracles are queried about A and B given all possible subsets of $\mathbf{Pa}(A) \setminus B$ and all possible subsets of $\mathbf{Pa}(B) \setminus A$:

$$\begin{aligned} H_0 : \text{At least one CI oracle outputs independent,} \\ H_1 : \text{All CI oracles output dependent.} \end{aligned} \tag{14}$$

Notice that the above hypothesis test is the same as the hypothesis test in (7). We can therefore bound the p-value of (12) using:

$$p'_{A-B} \triangleq \max_{i=1, \dots, q'} p_{A \perp\!\!\!\perp B | \mathbf{R}_i}, \tag{15}$$

where $\mathbf{R}_i \subseteq \{\mathbf{Pa}(A) \setminus B\}$ or $\mathbf{R}_i \subseteq \{\mathbf{Pa}(B) \setminus A\}$ and q' denotes the total number of such subsets.

Note that the skeleton discovery phase of the PC algorithm cannot differentiate between the parents and children of a particular vertex using its neighbors. However, we can further bound (15) using the following quantity:

$$p'_{A-B} \leq \max_{i=1, \dots, q} p_{A \perp\!\!\!\perp B | \mathbf{S}_i} \triangleq p_{A-B}, \tag{16}$$

where $\mathbf{S}_i \subseteq \{\mathbf{N}(A) \setminus B\}$ or $\mathbf{S}_i \subseteq \{\mathbf{N}(B) \setminus A\}$ and q denotes the total number of such subsets.

Now assume that the Type II error rate of all CI tests is zero. Then, if the alternative holds for the CI tests (conditional dependence), then the alternative is accepted. Hence, the PC algorithm will not remove any of the edges between $\mathbf{N}(A)$ and A as well as any of the edges between $\mathbf{N}(B)$ and B . PC therefore performs all necessary CI tests for computing (16), so upper bounding the Type I error rate for (12) reduces to taking the maximum of the p-values for all of the CI tests performed by PC regarding A and B . For example, suppose we measure three random variables A , B and C . Then we obtain p-values after the PC algorithm tests whether $A \perp\!\!\!\perp B$ and $A \perp\!\!\!\perp B | C$. Suppose these p-values are $(0.03, 0.04)$ so that the PC algorithm with an α threshold of 0.05 determines that $A - B$ is present. The p-value upper bound of (12) thus corresponds to $\max\{0.03, 0.04\} = 0.04$.

4.2 Detecting V-Structures

4.2.1 DETERMINISTIC SKELETON

The hypothesis testing procedure for directed edges is more complicated than the procedure for adjacencies. Edges can be oriented in the PC algorithm according to unshielded v-structures or the orientation rules as described in Section 2.2. Let us first focus on the former and, for further simplicity, let us also assume that 1) we have access to the ground truth skeleton and 2) no edge is involved in more than one unshielded v-structure (we will later drop these assumptions in Section 4.2.2). Our task then is to statistically infer the presence of an unshielded v-structure.

We now present the following null and alternative for each unshielded v-structure after finding a triple $A - C - B$ such that A and B are non-adjacent in the skeleton:

$$\begin{aligned} H_0 : \text{Unshielded } A \rightarrow C \leftarrow B \text{ is absent,} \\ H_1 : \text{Unshielded } A \rightarrow C \leftarrow B \text{ is present.} \end{aligned} \tag{17}$$

Next, consider the following proposition:

Proposition 2 (*Spirtes et al., 2000*) *Consider the same assumptions as Proposition 1. Further assume that A, C are adjacent and C, B are adjacent but A, B are non-adjacent. Then, A and B are conditionally independent given some subset of $\mathbf{Pa}(A) \setminus B$ which does not include C or some subset of $\mathbf{Pa}(B) \setminus A$ which does not include C if and only if $A \rightarrow C \leftarrow B$.*

The following null and alternative is therefore equivalent to (17):

$$\begin{aligned} H_0 : A \text{ and } B \text{ are conditionally dependent given any subset of } \mathbf{Pa}(A) \setminus B \\ \text{which does not include } C \text{ and any subset of } \mathbf{Pa}(B) \setminus A \text{ which} \\ \text{does not include } C, \\ H_1 : A \text{ and } B \text{ are conditionally independent given some subset of } \mathbf{Pa}(A) \setminus B \\ \text{which does not include } C \text{ or some subset of } \mathbf{Pa}(B) \setminus A \text{ which} \\ \text{does not include } C. \end{aligned} \tag{18}$$

The above alternative is reminiscent of the way in which PC determines the presence of an unshielded v-structure according to Algorithm 6 in the Appendix; specifically, if C is not in the set which renders A and B conditionally independent, then C in $A - C - B$ must be a collider. We however cannot bound the p-value of (18) using CI tests, because conditional dependence is in the null and conditional independence is in the alternative, as opposed to vice versa. As a result, we also consider the following proposition:

Proposition 3 *Consider the same assumptions as Proposition 2. Then, A and B are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ containing C and any subset of $\mathbf{Pa}(B) \setminus A$ containing C if and only if $A \rightarrow C \leftarrow B$.*

Proof First notice that $\mathbf{Pa}(A) = \{\mathbf{Pa}(A) \setminus B\}$ and $\mathbf{Pa}(B) = \{\mathbf{Pa}(B) \setminus A\}$, since A and B are non-adjacent. As a result, we can instead prove that the if and only if statement holds for $\mathbf{Pa}(A)$ and $\mathbf{Pa}(B)$ without loss of generality.

For the forward direction, suppose A and B are conditionally dependent given any subset of $\mathbf{Pa}(A)$ containing C and any subset of $\mathbf{Pa}(B)$ containing C . Then A and B are d-connected given any subset of $\mathbf{Pa}(A)$ containing C and any subset of $\mathbf{Pa}(B)$ containing C by the global directed Markov property. Clearly, $C \in \mathbf{N}(A)$ and $C \in \mathbf{N}(B)$, so C must either be a parent of A and a parent of B , a child of A and a parent of B , a parent of A and a child of B , or a child of A and a child of B . Note that A and B are non-adjacent, so A and B are d-separated given some subset of $\mathbf{Pa}(A)$ or some subset of $\mathbf{Pa}(B)$ by Proposition 1 and d-separation faithfulness. Moreover, the subset must include C if C is a parent of A and a parent of B , a child of A and a parent of B , or a parent of A and a child of B ; otherwise, A and B would be d-connected. As a result, in those three situations, we arrive at the contradiction that A and B are d-separated given some subset of $\mathbf{Pa}(A)$ containing C or some subset of $\mathbf{Pa}(B)$ containing C . We conclude that C must be a child of A and a child of B .

For the other direction, if $A \rightarrow C \leftarrow B$ holds, then A and B are d-connected given any subset of $\mathbf{Pa}(A)$ containing C and any subset of $\mathbf{Pa}(B)$ containing C . D-separation faithfulness then implies that A and B are conditionally dependent given any subset of $\mathbf{Pa}(A)$ containing C and any subset of $\mathbf{Pa}(B)$ containing C . \blacksquare

We can thus equivalently write (18) as:

$$\begin{aligned} H_0 : & A \text{ and } B \text{ are conditionally independent given some subset of } \mathbf{Pa}(A) \setminus B \\ & \text{containing } C \text{ or some subset of } \mathbf{Pa}(B) \setminus A \text{ containing } C, \\ H_1 : & A \text{ and } B \text{ are conditionally dependent given any subset of } \mathbf{Pa}(A) \setminus B \\ & \text{containing } C \text{ and any subset of } \mathbf{Pa}(B) \setminus A \text{ containing } C. \end{aligned} \quad (19)$$

We can bound the Type I error rate of the above hypothesis test by taking the maximum p-value over certain CI tests:

$$p'_{\gamma_{AB|C}} \triangleq \max_{i=1,\dots,m'} p_{A \perp\!\!\!\perp B | \mathbf{M}_i},$$

where \mathbf{M}_i denotes a subset of $\mathbf{Pa}(A) \setminus B$ containing C or a subset of $\mathbf{Pa}(B) \setminus A$ containing C , and m' is the total number of subsets \mathbf{M}_i . Of course, in practice, we do not know which vertices are the parents. However, we can also upper bound (19) as follows:

$$p'_{\gamma_{AB|C}} \leq \max_{i=1,\dots,m} p_{A \perp\!\!\!\perp B | \mathbf{T}_i} \triangleq p_{\gamma_{AB|C}}, \quad (20)$$

where \mathbf{T}_i denotes a subset of $\mathbf{N}(A) \setminus B$ containing C or a subset of $\mathbf{N}(B) \setminus A$ containing C , and m denotes the total number of subsets \mathbf{T}_i . Note that we do not need the zero Type II error rate assumption for computing (20), since we assume that the skeleton is provided.

4.2.2 INFERRED SKELETON

We have considered orienting the colliders, if we have access to the ground truth skeleton. We now consider the more complex problem of orienting the colliders, if we must also statistically infer the skeleton.

We again consider the following null and alternative:

$$\begin{aligned} H_0 : & \text{Unshielded } A \rightarrow C \leftarrow B \text{ is absent,} \\ H_1 : & \text{Unshielded } A \rightarrow C \leftarrow B \text{ is present.} \end{aligned} \quad (21)$$

Now, the PC algorithm determines that the alternative holds, if all of the following conditions are true:

1. A and C are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus C$ and any subset of $\mathbf{Pa}(C) \setminus A$.
 2. B and C are conditionally dependent given any subset of $\mathbf{Pa}(B) \setminus C$ and any subset of $\mathbf{Pa}(C) \setminus B$.
 3. A and B are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ containing C and any subset of $\mathbf{Pa}(B) \setminus A$ containing C .
- (22)

We therefore have the following equivalent form of the null and alternative as in (21), if we assume A and B are non-adjacent:

$$\begin{aligned} H_0 &: \text{At least one condition from (22) does not hold,} \\ H_1 &: \text{All conditions from (22) hold.} \end{aligned} \tag{23}$$

Note that the non-adjacency assumption is reasonable because we did not have enough statistical evidence to invalidate the assumption when we executed (12). Indeed, non-adjacencies are always assumed unless the data suggests that the null of (12) is unlikely. Now, the alternative of (23) is a series of three logical conjunctions, and the null is a series of three logical disjunctions as in (10), so the Type I error rate of (23) can be bounded using the intersection bound:

$$\begin{aligned} \Pr(\text{Conditions 1, 2, 3} | H_0) &\leq \Pr(\text{Any one condition} | H_0) \\ &\leq \max\{h_1, h_2, h_3\}. \end{aligned} \tag{24}$$

We will be using shorthand from here on. We write (23) equivalently as:

$$\begin{aligned} H_0 &: \neg(A - C) \vee \neg(B - C) \vee \neg\gamma_{AB|C}, \\ H_1 &: (A - C) \wedge (B - C) \wedge \gamma_{AB|C}, \end{aligned} \tag{25}$$

where $A - C$, $B - C$, and $\gamma_{AB|C}$ represent Condition 1, 2 and 3 from (22), respectively. We therefore have a p-value bound of (21) similar to (24):

$$\begin{aligned} &\Pr((A - C) \wedge (B - C) \wedge \gamma_{AB|C} | H_0) \\ &\leq \Pr(A - C | \neg(A - C)) \\ &\leq \max \left\{ \Pr(A - C | \neg(A - C)), \Pr(B - C | \neg(B - C)), \Pr(\gamma_{AB|C} | \neg\gamma_{AB|C}) \right\} \\ &\leq \max\{p_{A-C}, p_{B-C}, p_{\gamma_{AB|C}}\}. \end{aligned} \tag{26}$$

Notice that computing $p_{\gamma_{AB|C}}$ requires $\mathbf{N}(A)$ and $\mathbf{N}(B)$, not just their respective empirical estimates $\widehat{\mathbf{N}}(A)$ and $\widehat{\mathbf{N}}(B)$ which PC can discover. However, we can invoke a zero Type II error rate assumption in order to ensure that $\mathbf{N}(A) \subseteq \widehat{\mathbf{N}}(A)$ and $\mathbf{N}(B) \subseteq \widehat{\mathbf{N}}(B)$ as explained in detail in Section 4.4, so $p_{\gamma_{AB|C}}$ can still be upper bounded. The assumption

also ensures that we can upper bound p_{A-C} and p_{B-C} according to Section 4.1. We conclude that a zero Type II error rate ensures that (26) can be computed.

Next, consider the situation where PC can orient any one edge by using more than one unshielded v-structure. For example, consider the DAG in Figure 3. In this case, PC can orient $A - C$ by using either $B_1 \rightarrow C$ or $B_2 \rightarrow C$ (or both); we may therefore want to take both situations into account. Note that the original PC algorithm always orients an edge according to one v-structure which it picks arbitrarily according to the ordering of its computations. We thus only require the bound (26) in this case. However, we will propose a modified PC algorithm in Section 5 which takes into account all possible ways to orient one edge. Now, we can use the following null and alternative for Figure 3 when assuming that both A and B_1 and A and B_2 are non-adjacent:

$$\begin{aligned} H_0 &: \neg(A - C) \vee \left([\neg(B_1 - C) \vee \neg\gamma_{AB_1|C}] \wedge [\neg(B_2 - C) \vee \neg\gamma_{AB_2|C}] \right), \\ H_1 &: (A - C) \wedge \left([(B_1 - C) \wedge \gamma_{AB_1|C}] \vee [(B_2 - C) \wedge \gamma_{AB_2|C}] \right). \end{aligned} \quad (27)$$

We can therefore bound the Type I error rate of (27) as follows, where $\mathcal{G} = \neg(A - C)$ and $\mathcal{H} = [\neg(B_1 - C) \vee \neg\gamma_{AB_1|C}] \wedge [\neg(B_2 - C) \vee \neg\gamma_{AB_2|C}]$, $\mathcal{H}_1 = \neg(B_1 - C) \vee \neg\gamma_{AB_1|C}$, and $\mathcal{H}_2 = \neg(B_2 - C) \vee \neg\gamma_{AB_2|C}$:

$$\begin{aligned} & \Pr\left((A - C) \wedge \left([(B_1 - C) \wedge \gamma_{AB_1|C}] \vee [(B_2 - C) \wedge \gamma_{AB_2|C}] \right) \middle| H_0\right) \\ & \leq \max \left\{ \Pr(A - C | \mathcal{G}), \Pr\left([(B_1 - C) \wedge \gamma_{AB_1|C}] \vee [(B_2 - C) \wedge \gamma_{AB_2|C}] \middle| \mathcal{H} \right) \right\} \\ & \leq \max \left\{ \Pr(A - C | \mathcal{G}), \Pr\left((B_1 - C) \wedge \gamma_{AB_1|C} \middle| \mathcal{H} \right) + \Pr\left((B_2 - C) \wedge \gamma_{AB_2|C} \middle| \mathcal{H} \right) \right\} \\ & = \max \left\{ \Pr(A - C | \mathcal{G}), \Pr\left((B_1 - C) \wedge \gamma_{AB_1|C} \middle| \mathcal{H}_1 \right) + \Pr\left((B_2 - C) \wedge \gamma_{AB_2|C} \middle| \mathcal{H}_2 \right) \right\} \quad (28) \\ & \leq \max \left\{ \Pr(A - C | \mathcal{G}), \max \left\{ \Pr(B_1 - C | \neg(B_1 - C)), \Pr(\gamma_{AB_1|C} | \neg\gamma_{AB_1|C}) \right\} \right. \\ & \quad \left. + \max \left\{ \Pr(B_2 - C | \neg(B_2 - C)), \Pr(\gamma_{AB_2|C} | \neg\gamma_{AB_2|C}) \right\} \right\} \\ & \leq \max \left\{ p_{A-C}, \max\{p_{B_1-C}, p_{\gamma_{AB_1|C}}\} + \max\{p_{B_2-C}, p_{\gamma_{AB_2|C}}\} \right\}. \end{aligned}$$

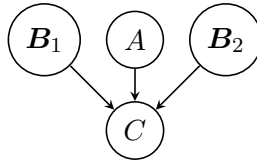


Figure 3: Here, one can orient the edge $A - C$ according to the two unshielded v-structures $A \rightarrow C \leftarrow B_1$ and $A \rightarrow C \leftarrow B_2$.

More generally, for an arbitrary number, say j , of multiple possible ways to orient $A - C$ by unshielded v-structures, we have:

$$\begin{aligned} & \Pr\left((A - C) \wedge \left([(\mathbf{B}_1 - C) \wedge \gamma_{A\mathbf{B}_1|C}] \vee \dots \vee [(\mathbf{B}_j - C) \wedge \gamma_{A\mathbf{B}_j|C}]\right) \middle| H_0\right) \\ & \leq \max\left\{p_{A-C}, \sum_{i=1}^j \max\{p_{\mathbf{B}_i-C}, p_{\gamma_{A\mathbf{B}_i|C}}\}\right\}, \end{aligned} \quad (29)$$

assuming that $\mathbf{B}_1, \dots, \mathbf{B}_j$ are all non-adjacent to A .

4.3 Orientation Rules

We now consider bounding the p-values of edges which are oriented using the orientation rules of the PC algorithm. Recall from (1) that the PC algorithm only requires the repeated application of three orientation rules to be complete. We analyze these three orientation rules in separate subsections.

4.3.1 FIRST ORIENTATION RULE

We can construct the hypothesis test for the first orientation rule as follows according to the sufficient conditions of the first rule in (1):

$$\begin{aligned} H_0 &: \neg(A - B) \vee \neg(C \rightarrow A), \\ H_1 &: (A - B) \wedge (C \rightarrow A). \end{aligned} \quad (30)$$

We again also assume that C and B are non-adjacent. Now the PC algorithm determines that the alternative holds, if all of the following conditions are true:

1. $A - B$: A and B are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ and any subset of $\mathbf{Pa}(B) \setminus A$.
2. $C \rightarrow A$: An edge is oriented from C to A under two scenarios. In the first, the edge is oriented because A is the collider in an unshielded v-structure. In the second, the edge is oriented due to the previous application of an orientation rule.

We thus have a logical conjunction and can bound the Type I error rate using the intersection bound:

$$\Pr((A - B) \wedge (C \rightarrow A) | H_0) \leq \max\{p_{A-B}, p_{C \rightarrow A}\},$$

where $p_{C \rightarrow A}$ refers to the p-value bound for the hypothesis test of an unshielded v-structure or a previously applied orientation rule. Of course, $p_{C \rightarrow A}$ will be the former when the PC algorithm begins to execute the orientation rules. More generally, for $\mathbf{C}_i \rightarrow A$ that can orient $A - B$ where $i = 1, \dots, j$, we have:

$$\begin{aligned} & \Pr\left((A - B) \wedge \left[(\mathbf{C}_1 \rightarrow B) \vee \dots \vee (\mathbf{C}_j \rightarrow A)\right] \middle| H_0\right) \\ & \leq \max\left\{p_{A-B}, \sum_{i=1}^j p_{\mathbf{C}_i \rightarrow A}\right\}, \end{aligned} \quad (31)$$

where we require that $\mathbf{C}_1, \dots, \mathbf{C}_j$ are all non-adjacent to B .

4.3.2 SECOND ORIENTATION RULE

We have the following hypothesis test according to the sufficient conditions of the second rule in (1):

$$\begin{aligned} H_0 &: \neg(A - B) \vee \neg(A \rightarrow C \rightarrow B), \\ H_1 &: (A - B) \wedge (A \rightarrow C \rightarrow B). \end{aligned} \quad (32)$$

Hence, by conjunction:

$$\Pr((A - B) \wedge (A \rightarrow C \rightarrow B) | H_0) \leq \max\{p_{A-B}, p_{A \rightarrow C \rightarrow B}\},$$

where $p_{A \rightarrow C \rightarrow B} \leq \max\{p_{A \rightarrow C}, p_{C \rightarrow B}\}$. The above Type I error rate can therefore be further upper bounded by $\max\{p_{A-B}, p_{A \rightarrow C}, p_{C \rightarrow B}\}$. More generally, we have:

$$\begin{aligned} & \Pr\left((A - B) \wedge \left[(A \rightarrow \mathbf{C}_1 \rightarrow B) \vee \dots \vee (A \rightarrow \mathbf{C}_j \rightarrow B)\right] | H_0\right) \\ & \leq \max\left\{p_{A-B}, \sum_{i=1}^j p_{A \rightarrow \mathbf{C}_i \rightarrow B}\right\} \leq \max\left\{p_{A-B}, \sum_{i=1}^j \max\{p_{A \rightarrow \mathbf{C}_i}, p_{\mathbf{C}_i \rightarrow B}\}\right\}. \end{aligned} \quad (33)$$

4.3.3 THIRD ORIENTATION RULE

We have the following null and alternative by the sufficient conditions of the third rule in (1), assuming that C and D are non-adjacent:

$$\begin{aligned} H_0 &: \neg(A - B) \vee \neg(A - C \rightarrow B) \vee \neg(A - D \rightarrow B), \\ H_1 &: (A - B) \wedge (A - C \rightarrow B) \wedge (A - D \rightarrow B). \end{aligned} \quad (34)$$

We can bound the Type I error rate of the above hypothesis test as follows:

$$\begin{aligned} & \Pr\left((A - B) \wedge (A - C \rightarrow B) \wedge (A - D \rightarrow B) | H_0\right) \\ & \leq \max\{p_{A-B}, p_{A-C \rightarrow B}, p_{A-D \rightarrow B}\} \\ & \leq \max\left\{p_{A-B}, \max\{p_{A-C}, p_{C \rightarrow B}\}, \max\{p_{A-D}, p_{D \rightarrow B}\}\right\}. \end{aligned}$$

The general case is slightly more complicated than the first and second orientation rules. In this case, we need to control the Type I error rate of accepting at least two paths as opposed to one. Let the set \mathbf{D} include all three-node paths from A to B with the first edge undirected from A to a middle vertex and the second edge directed from the middle vertex to B such that the i^{th} element of \mathbf{D} is:

$$\mathbf{D}_i \triangleq A - \mathbf{C}_i \rightarrow B.$$

Let us suppose \mathbf{D} has a total of n elements and assume that no middle vertex \mathbf{C}_i is adjacent to any other middle vertex. Now, let \mathbf{D}' be the set containing all of the n choose 2 elements of \mathbf{D} . The i^{th} element in \mathbf{D}' is therefore:

$$\mathbf{D}'_i \triangleq \{A - \mathbf{C}_k \rightarrow B, A - \mathbf{C}_l \rightarrow B\},$$

where k and l are the distinct indices represented the two chosen middle vertices. Let $\mathbf{D}'_{i,1}$ and $\mathbf{D}'_{i,2}$ be the first and second elements in \mathbf{D}'_i , respectively. Also let $r = \binom{n}{2}$. We then have:

$$\begin{aligned} & \Pr\left((A - B) \wedge \left[(\mathbf{D}'_{1,1} \wedge \mathbf{D}'_{1,2}) \vee \dots \vee (\mathbf{D}'_{r,1} \wedge \mathbf{D}'_{r,2})\right] \middle| H_0\right) \\ & \leq \max \left\{ p_{A-B}, \sum_{i=1}^r \Pr(\mathbf{D}'_i | H_0) \right\}, \end{aligned}$$

where $\Pr(\mathbf{D}'_i | H_0) \triangleq p_{\{A-C_k \rightarrow B, A-C_l \rightarrow B\}}$ and is bounded as follows:

$$\begin{aligned} \Pr(\mathbf{D}'_i | H_0) & \leq \max\{p_{A-C_k \rightarrow B}, p_{A-C_l \rightarrow B}\} \\ & \leq \max\{p_{A-C_k}, p_{C_k \rightarrow B}, p_{A-C_l}, p_{C_l \rightarrow B}\} \\ & \triangleq \ddot{\Pr}(\mathbf{D}'_i | H_0), \end{aligned}$$

We therefore have:

$$\begin{aligned} & \Pr\left((A - B) \wedge \left[(\mathbf{D}'_{1,1} \wedge \mathbf{D}'_{1,2}) \vee \dots \vee (\mathbf{D}'_{r,1} \wedge \mathbf{D}'_{r,2})\right] \middle| H_0\right) \\ & \leq \max \left\{ p_{A-B}, \sum_{i=1}^r \ddot{\Pr}(\mathbf{D}'_i | H_0) \right\}. \end{aligned} \tag{35}$$

4.4 Summary and Analysis of the Bounds

We derived several bounds for edge orientation as summarized in Table 1. We created the bounds by engineering specific hypothesis tests and successively applying the union and intersection bounds accordingly. Note that j and r are usually very small in sparse graphs.

One may now wonder whether PC can actually control the bounds listed in Table 1 (we say that a quantity can be *controlled*, if the quantity can be upper bounded). Recall that we

Table 1: P-value bounds for all of the edge types in a CPDAG. Note that $\mathbf{S}_i \subseteq \{\mathbf{N}(A) \setminus B\}$ or $\mathbf{S}_i \subseteq \{\mathbf{N}(B) \setminus A\}$.

Edge Type	P-Value Bound	Equation Num.
Undirected	$\max_i p_{A \perp\!\!\!\perp B \mathbf{S}_i}$	(16)
Unshielded v-structure	$\max \left\{ p_{A-C}, \sum_{i=1}^j \max\{p_{\mathbf{B}_i-C}, p_{\gamma_{AB_i C}}\} \right\}$	(29)
First orientation rule	$\max \left\{ p_{A-B}, \sum_{i=1}^j p_{C_i \rightarrow A} \right\}$	(31)
Second orientation rule	$\max \left\{ p_{A-B}, \sum_{i=1}^j \max\{p_{A \rightarrow C_i}, p_{C_i \rightarrow B}\} \right\}$	(33)
Third orientation rule	$\max \left\{ p_{A-B}, \sum_{i=1}^r \ddot{\Pr}(\mathbf{D}'_i H_0) \right\}$	(35)

provided a rough, affirmative answer to the question in Sections 4.1 and 4.2.2 by assuming a zero Type II error rate. We now spell out a more detailed answer via a theorem whose proof builds on the argument of Theorem 4 in (Armen and Tsamardinos, 2014).

Theorem 4 *Suppose that the PC algorithm is applied to a sample from \mathbb{P} represented by DAG \mathbb{G} . If we have:*

1. \mathbb{P} is d -separation faithful to \mathbb{G} ,
2. The Type II error rate is zero,
3. The PC algorithm also tests whether any two non-adjacent vertices A, B with common neighbor C are conditionally dependent given any subset of $\mathbf{Pa}(A) \setminus B$ containing C and any subset of $\mathbf{Pa}(B) \setminus A$ containing C ,

then all of the p-value bounds in Table 1 can be controlled using the p-values of the CI tests executed by PC.

Proof Consider any two vertices A and B . Algorithm 1 starts with a fully connected graph, so we have $B \in \widehat{\mathbf{N}}(A)$ and $A \in \widehat{\mathbf{N}}(B)$ in the beginning. Note that Algorithm 1 executes $\text{test}_{A \perp\!\!\!\perp B | \mathbf{S}}$ for all $\mathbf{S} \subseteq \widehat{\mathbf{N}}(A) \setminus B$ and for all $\mathbf{S} \subseteq \widehat{\mathbf{N}}(B) \setminus A$. The zero Type II error rate ensures the following: if the alternative holds, then the alternative is accepted. As a result, Algorithm 1 will not remove any vertices adjacent to A and any vertices adjacent to B with a zero Type II error rate. Hence, we always have $\{\mathbf{N}(A) \setminus B\} \subseteq \{\widehat{\mathbf{N}}(A) \setminus B\}$ and $\{\mathbf{N}(B) \setminus A\} \subseteq \{\widehat{\mathbf{N}}(B) \setminus A\}$. Algorithm 1 therefore must eventually execute $\text{test}_{A \perp\!\!\!\perp B | \mathbf{S}}$ for all $\mathbf{S} \subseteq \{\mathbf{N}(A) \setminus B\}$ and for all $\mathbf{S} \subseteq \{\mathbf{N}(B) \setminus A\}$, so (16) can be controlled.

For (29), the p-value bounds for undirected edges can already be controlled by the previous paragraph. We must now argue that $p_{\gamma_{AB|C}}$ can be controlled. Let C be a collider between non-adjacent vertices A and B . Now notice that $C \in \mathbf{N}(A) \subseteq \widehat{\mathbf{N}}(A)$ and $C \in \mathbf{N}(B) \subseteq \widehat{\mathbf{N}}(B)$, so Algorithm 2 must execute $\text{test}_{A \perp\!\!\!\perp B | \mathbf{S}}$ for all $\mathbf{S} \subseteq \{\mathbf{N}(A) \setminus B\}$ containing C and for all $\mathbf{S} \subseteq \{\mathbf{N}(B) \setminus A\}$ containing C . Hence $p_{\gamma_{AB|C}}$ can be controlled.

Now the p-value bounds (31), (33) and (35) can be controlled trivially because the p-value bounds for (15) and (29) can be controlled. \blacksquare

In other words, PC can control the bounds in Table 1 with some additional CI tests and a zero Type II error rate.

Of course, the Type II error rate is never zero in practice, but this becomes less of an issue as the sample size increases. We may also consider reducing the Type II error rate by simultaneously implementing three strategies:

1. Use a liberal (higher) α threshold. We for example often use an α threshold of 0.20 in the experiments. This is the simplest strategy which decreases the Type II error rate but also increases the Type I error rate. However, we can then control the Type I error rate post-hoc with an FDR controlling procedure. Of course, setting the α threshold too high will prevent the PC algorithm from terminating within a reasonable amount of time as well as loosen the p-value bounds, since the CI tests will fail to explain away many edges. We therefore cannot rely entirely on this first strategy.

2. Use hypothesis tests whose p-value bounds are robust to Type II errors. The hypothesis tests in Section 4.4 are in fact robust to such errors due to the intersection bound as explained in detail in Appendix A.2. Briefly, we can also reasonably consider modifying the null hypotheses of (25), (30), (32) and (34) to “no edges between any of the vertices.” This corresponds to converting the logical disjunctions in the null of (10) into conjunctions which in turns leads to a less robust p-value bound involving the minimum of a set of p-values instead of the maximum. As a result, under-estimating one p-value in the p-value set due to Type II error(s) can cause PC to also under-estimate the bound of (25), (30), (32) or (34).
3. Modify the PC algorithm to prevent and catch many Type II errors.

The last strategy is more complex, so we discuss it in detail in the next section.

5. The PC Algorithm with P-Values

We now propose a modified PC algorithm called PC with p-values (PC-p) that reduces the influence of Type II errors by preventing and catching potential Type II errors. At the same time, PC-p is correct - the algorithm operates differently than PC, but it maintains PC’s desirable soundness and completeness properties.

The PC-p algorithm involves two ideas. First, PC-p performs skeleton discovery with the same skeleton discovery procedure used in the PC-stable algorithm (Colombo and Maathuis, 2014). This procedure ensures that the algorithm does not skip some CI tests due to Type II errors and variable ordering. The second idea behind PC-p involves a modification to the procedure for propagating edge orientations. Specifically, if two edge orientations conflict, PC-p admits bidirected edges instead of over-writing previous orientations like PC. PC-p then unorients the bidirected edges as well as the directed edges which were directly used to infer the presence of the bidirected edges. The algorithm subsequently labels the resulting undirected edges as “ambiguous” which ensures that PC-p does not orient additional edges using the ambiguous edges. Indeed, the PC-p algorithm uses conflicts in edge orientation to detect potential Type II errors and prevent the propagation of the errors throughout the graph. In practice, we find that these two modifications to the PC algorithm help PC-p with the BY estimator achieve more accurate strong estimation and control of the FDR than PC, as we will see in Section 6.

We now describe the PC-p algorithm in detail; however, we will not describe the computation of the p-value upper bounds until Section 5.5 in order to keep the presentation clear. We have divided the PC-p algorithm into Algorithms 1, 2, 3 and 4, where the first three procedures correspond to Algorithms 5, 6 and 7 of the original PC algorithm.

5.1 Skeleton Discovery

We first consider skeleton discovery. The original PC algorithm uses Algorithm 5 to discover the skeleton. However, Algorithm 5 can cause the sample version of the PC algorithm to skip some CI tests due to variable ordering and Type II errors. For example, consider the causal graph in Figure 4a as first presented in (Colombo and Maathuis, 2014). In this example, suppose the CI tests correctly determine that $A \perp\!\!\!\perp B$ and $B \perp\!\!\!\perp D|\{A, C\}$ but incorrectly determine that $C \perp\!\!\!\perp D|\{A, E\}$. The incorrect inference is a Type II error, since C and D

are adjacent in the true graph. Now consider the following ordering of variables for the PC algorithm: $order_1(\mathbf{X}) = (A, D, B, C, E)$. In this case, the ordered pair (D, B) is considered before (D, C) in Algorithm 5, since (D, B) comes earlier in $order_1(\mathbf{X})$. The PC algorithm removes $D - B$ because a CI test determines that $D \perp\!\!\!\perp B | \{A, C\}$ and $\{A, C\}$ is a subset of $\mathbf{N}(D) = \{A, B, C, E\}$. Next, $D - C$ is considered and erroneously removed because a CI test determines that $D \perp\!\!\!\perp C | \{A, E\}$ and $\{A, E\}$ is a subset of $\mathbf{N}(D) = \{A, C, E\}$. We thus ultimately obtain the skeleton in Figure 4b with $order_1(\mathbf{X})$.

Now consider an alternative ordering of the variables: $order_2(\mathbf{X}) = (A, C, D, B, E)$. In this case, (C, D) is considered before (D, B) in Algorithm 5, and the algorithm erroneously removes $C - D$. Next, the algorithm considers $D - B$ but $\{A, C\}$ is not a subset of $\mathbf{N}(D) = \{A, B, E\}$, so $D - B$ remains. Even when the PC algorithm eventually also considers the same undirected edge as $B - D$, $\{A, C\}$ is again not a subset of $\mathbf{N}(B) = \{C, D, E\}$, so $B - D$ remains. In other words, (C, D) is considered first in $order_2(\mathbf{X})$ which causes C to be removed from $\mathbf{N}(D)$. Algorithm 5 therefore never executes $test_{B \perp\!\!\!\perp D | \{A, C\}}$. We thus ultimately obtain the skeleton in Figure 4c with $order_2(\mathbf{X})$.

The previous two examples show that the Type II error of incorrectly determining that $C \perp\!\!\!\perp D | \{A, E\}$ leads PC to infer two different skeletons due to differences in variable ordering. Clearly, we would like to eliminate the dependency of skeleton discovery on variable ordering and also reduce its dependency on Type II errors at the same time. Fortunately, Colombo and Maathuis proposed such a modification of Algorithm 5 as outlined in Algorithm 1. The key difference between Algorithm 5 and 1 involves the for loop in steps 5-7 of Algorithm 1 which computes and stores the adjacency sets after each new conditioning set size. As a result, an incorrect edge deletion due to a Type II error on line 16 of Algorithm 1 no longer effects which CI tests are performed for other pairs of variables with conditioning set size l . Indeed, the algorithm only modifies the adjacency sets when it increases the conditioning set size. Colombo and Maathuis proved that Algorithm 1 is order-independent. We review the proof here, since it is informative:

Proposition 5 (Colombo and Maathuis, 2014). *The skeleton resulting from Algorithm 1 is order-independent.*

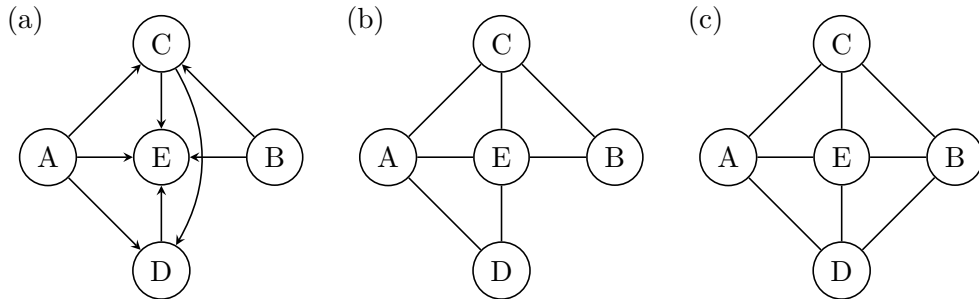


Figure 4: An example of a situation when PC infers different skeletons due to a Type II error and two variable orderings. (a) The true causal graph, (b) the skeleton inferred by PC from $order_1(\mathbf{X})$, (c) the skeleton inferred by PC from $order_2(\mathbf{X})$.

Proof Consider the removal or retention of some undirected edge $A - B$ at some conditioning set size l . The ordering of the variables determines the order in which the edges (line 9) and subsets $\mathbf{S} \subseteq \mathbf{a}(A)$ and $\mathbf{S} \subseteq \mathbf{a}(B)$ (line 11) are considered. However, by construction, the order in which the edges are considered does not affect the sets $\mathbf{a}(A)$ and $\mathbf{a}(B)$.

If there is at least one subset \mathbf{S} of $\mathbf{a}(A)$ or $\mathbf{a}(B)$ such that $A \perp\!\!\!\perp B | \mathbf{S}$, then any ordering of the variables will find a separating set for A and B (but different orderings may lead to different separating sets as illustrated in Example 2 of (Colombo and Maathuis, 2014)). Conversely, if there is no subset \mathbf{S}' of $\mathbf{a}(A)$ or $\mathbf{a}(B)$ such that $A \perp\!\!\!\perp B | \mathbf{S}'$, then no ordering will find a separating set.

Hence, any ordering of the variables leads to the same edge deletions and therefore to the same skeleton. \blacksquare

In other words, modifying the adjacency sets only when changing the conditioning set size prevents PC-p from skipping some CI tests during skeleton discovery because of Type II errors and variable ordering. As a result, Algorithm 1 enables PC-p to perform more of the required CI tests than Algorithm 5 in order to correctly upper bound the p-value of (12). However, notice that Algorithm 1 does not prevent all Type II errors from effecting the skeleton. The edge $C - D$ is for example eliminated in Figure 4 regardless of the ordering because of the erroneous conclusion that $C \perp\!\!\!\perp D | \{A, E\}$. As a result, we have $C \notin \widehat{\mathbf{N}}(D)$ which may lead to under-estimation of the p-value bounds for undirected edges connected to D . We will nonetheless see in Section 6 that Algorithm 1 does help PC-p achieve tighter estimation and control of the FDR than the original skeleton discovery procedure, since Algorithm 1 eliminates the influence of at least some Type II errors.

5.2 Unshielded V-Structures

We now describe Algorithm 2, where we use the circle edge endpoint “o” as a meta-symbol representing either a tail or an arrowhead. In Algorithm 2, PC-p orients edges according to all unshielded v-structures in line 3, even if two v-structures conflict with each other in

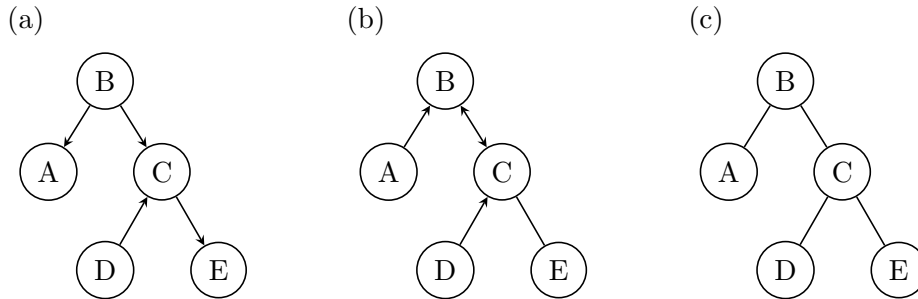


Figure 5: Example of how Algorithm 2 deals with conflicting edge orientations. a) The ground truth, b) the inferred graph with two v-structures $A \rightarrow B \leftarrow C$ and $D \rightarrow C \leftarrow B$ that lead to the bi-directed edge $B \leftrightarrow C$, and c) the final graph after unorienting both v-structures.

the direction of a particular edge. In the case of conflict, PC-p admits a bidirected edge instead of favoring one particular direction over the other. The algorithm then unorients all v-structures involving the bidirected edges and labels the unoriented edges as “ambiguous” in line 23 because bidirected edges may result from a Type II error. For example, consider the ground truth in Figure 5a and assume that Algorithm 1 correctly discovers all of the undirected edges. Moreover, assume Algorithm 1 correctly finds a separating set of B and D that does not contain C but incorrectly finds a separating set of A and C that does not contain B . The latter is a Type II error, since the alternative should have been accepted rather than rejected when conditioning on a subset not containing B . In this case, PC-p first orients the edges according to Figure 5b. However, notice that the two unshielded v-structures conflict with each other due to the bidirected edge $B \leftrightarrow C$, and PC-p cannot determine which v-structure admitted the Type II error. As a result, the algorithm unorients all of the edges in both v-structures as in Figure 5c. PC-p then labels the three unoriented edges as “ambiguous” so that the algorithm does not orient any other undirected edges based on these three edges using the orientation rules. The labeling thus prevents the algorithm from propagating Type II errors by orienting additional edges based on the erroneous directions.

5.3 Orientation Rules

Notice that Algorithm 7 uses “else if” statements instead of all “if” statements. The “else if” approach is of course faster, but it also causes PC to ignore any interactions between the orientation rules in the sense that, if one rule orients an edge, then no other rule can orient an edge. PC-p performs the orientation rules according Algorithm 3 which uses the “if” approach to attempt to apply all three orientation rules to each non-ambiguous undirected edge. Now, if bidirected edges exist after the rules are applied, then Algorithm 3 unorients the edge as well as all edges involved in the sufficient conditions of the associated orientation rules in lines 16-18. The algorithm then labels the unoriented edges as “ambiguous” in line 19 similar to unshielded v-structure orientation in Section 5.2. For example, in Figure 6, rule 1 of PC-p induces a bidirected edge between $A - B$, so PC-p unorients and labels all directed edges which satisfy the sufficient conditions of rule 1 as ambiguous; these include $D \rightarrow A, C \rightarrow A, E \rightarrow B$, and $F \rightarrow B$.

5.4 Analysis of PC-p

We now have the following analysis of the PC-p algorithm:

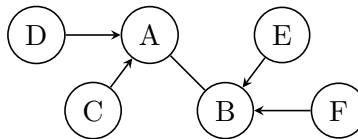


Figure 6: Here, a bidirected edge between A and B results from the application of rule 1. PC-p therefore unorients and labels all edges in the above graph as “ambiguous” according to the sufficient conditions of rule 1.

Data: \mathbf{X}^n, α
Result: $\widehat{\mathbb{G}}, \mathcal{P}^1, \mathcal{S}, \mathcal{I}$

```

1 Form a completely connected undirected graph  $\widehat{\mathbb{G}}$  on the variable set in  $\mathbf{X}^n$ 
2  $l = -1$ 
3 repeat
4    $l = l + 1$ 
5   for each variable  $A$  in  $\widehat{\mathbb{G}}$  do
6      $\mathbf{a}(A) \leftarrow \widehat{\mathbf{N}}(A)$ 
7   end
8   repeat
9     Select a new ordered pair of variables  $(A, B)$  that are adjacent in  $\widehat{\mathbb{G}}$  and
       satisfy  $|\mathbf{a}(A) \setminus B| \geq l$ 
10    repeat
11      Choose a new set  $\mathbf{S} \subseteq \{\mathbf{a}(A) \setminus B\}$  with  $|\mathbf{S}| = l$ 
12       $p \leftarrow$  p-value from  $\text{test}_{A \perp\!\!\!\perp B | \mathbf{S}}$ 
13      if  $p \leq \alpha$  then
14        Insert  $p$  into  $\mathcal{P}_{AB}^1$  and  $\mathcal{P}_{BA}^1$ 
15      else
16        Delete  $A - B$  from  $\widehat{\mathbb{G}}$ 
17        Empty  $\mathcal{P}_{AB}^1$  and  $\mathcal{P}_{BA}^1$ 
18        Insert  $\mathbf{S}$  into  $\mathcal{S}_{AB}$  and  $\mathcal{S}_{BA}$ 
19      end
20    until  $A - B$  is deleted from  $\widehat{\mathbb{G}}$  or all  $\mathbf{S} \subseteq \{\mathbf{a}(A) \setminus B\}$  with  $|\mathbf{S}| = l$  have been
       considered;
21  until all ordered pairs of adjacent variables  $(A, B)$  in  $\widehat{\mathbb{G}}$  with  $|\mathbf{a}(A) \setminus B| \geq l$  have
     been considered;
22 until all ordered pairs of adjacent variables  $(A, B)$  in  $\widehat{\mathbb{G}}$  satisfy  $|\mathbf{a}(A) \setminus B| \leq l$ ;
23 for each nonempty  $\mathcal{P}_{AB}^1$  in  $\mathcal{P}^1$  do
24    $\mathcal{P}_{AB}^1 \leftarrow \max\{\mathcal{P}_{AB}^1\}$ 
25   Place the same unique identifier for  $A - B$  into  $\mathcal{I}_{AB}$  and  $\mathcal{I}_{BA}$ 
26 end

```

Algorithm 1: Skeleton Discovery

Data: $\mathbf{X}^n, \widehat{\mathbb{G}}, \mathcal{P}^1, \mathcal{S}, \mathcal{I}$

Result: $\widehat{\mathbb{G}}, \mathcal{P}^2, \mathcal{I}$

```

1 for all ordered pairs of non-adjacent variables  $(A, B)$  with common neighbor  $C$  do
2   if  $C \notin$  any set in  $\mathcal{S}_{AB}$  then
3     Replace  $A \circ\circ C \circ\circ B$  with  $A \circ\rightarrow C \leftarrow\circ B$ 
4      $l = 0$ 
5     repeat
6        $l = l + 1$ 
7       repeat
8         Choose a new set  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(A)$  including  $C$  or  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(B)$  including  $C$ 
          with  $|\mathcal{S}| = l$ 
9          $p \leftarrow$  p-value from  $\text{test}_{A \perp\!\!\!\perp B | \mathcal{S}}$ 
10        Insert  $p$  into  $\mathcal{P}''$ 
11      until all  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(A)$  including  $C$  and all  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(B)$  including  $C$  with
           $|\mathcal{S}| = l$  have been considered;
12    until all  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(A)$  including  $C$  and all  $\mathcal{S} \subseteq \widehat{\mathcal{N}}(B)$  including  $C$  satisfy
           $|\mathcal{S}| \leq l$ ;
13    Insert  $\max\{\mathcal{P}_{AC}^1, \mathcal{P}''\}$  into  $\mathcal{P}'_{BC}$ 
14    Insert  $\max\{\mathcal{P}_{BC}^1, \mathcal{P}''\}$  into  $\mathcal{P}'_{AC}$ 
15    Empty  $\mathcal{P}''$ 
16  end
17 end
18  $\widehat{\mathbb{G}}' \leftarrow \widehat{\mathbb{G}}$ 
19 for each  $A \leftrightarrow C$  in  $\widehat{\mathbb{G}}$  do
20   Unorient the edge to  $A - C$  in  $\widehat{\mathbb{G}}'$ 
21   For each additional edge directed to  $A$  in  $\widehat{\mathbb{G}}$ , unorient the edge in  $\widehat{\mathbb{G}}'$ 
22   For each additional edge directed to  $C$  in  $\widehat{\mathbb{G}}$ , unorient the edge in  $\widehat{\mathbb{G}}'$ 
23   Label the unoriented edges as “ambiguous” in  $\widehat{\mathbb{G}}'$ 
24 end
25  $\widehat{\mathbb{G}} \leftarrow \widehat{\mathbb{G}}'$ 
26 for each  $A \rightarrow C$  in  $\widehat{\mathbb{G}}$  do
27    $\mathcal{P}_{AC}^2 \leftarrow \max\{\mathcal{P}_{AC}^1, \text{sum}[\mathcal{P}'_{AC}]\}$ 
28   if one p-value in  $\mathcal{P}'_{AC}$  then
29     Place the same unique identifier into  $\mathcal{I}_{AC}$  and  $\mathcal{I}_{BC}$  for unshielded collider
       $A \rightarrow C \leftarrow B$ 
30   else if more than one p-value in  $\mathcal{P}'_{AC}$  then
31     Place a unique identifier into  $\mathcal{I}_{AC}$ 
32   end
33 end

```

Algorithm 2: Unshielded V-structures

Data: $\widehat{\mathbb{G}}, \mathcal{P}^1, \mathcal{P}^2, \mathcal{I}$

Result: $\widehat{\mathbb{G}}, \mathcal{P}^2, \mathcal{I}$

```

1 repeat
2    $\widehat{\mathbb{G}}' \leftarrow \widehat{\mathbb{G}}$ 
3   if  $A - B$  non-ambiguous and  $\exists i$  s.t.  $C_i \rightarrow A$  with  $C_i$  and  $B$  non-adjacent in  $\widehat{\mathbb{G}}$ 
      then
4     Replace  $A \circ \circ B$  in  $\widehat{\mathbb{G}}'$  with  $A \circ \rightarrow B$ 
5      $\mathcal{P}'_{A,B} \leftarrow \text{sum}\{\mathcal{P}'_{A,B}, \text{sum}\{\mathcal{P}^2_{C_i A}, \forall i \text{ s.t. } C_i \rightarrow A \text{ with } C_i, B \text{ non-adjacent}\}\}$ 
6   end
7   if  $A - B$  non-ambiguous and  $\exists i$  s.t.  $A \rightarrow C_i \rightarrow B$  in  $\widehat{\mathbb{G}}$  then
8     Replace  $A \circ \circ B$  in  $\widehat{\mathbb{G}}'$  with  $A \circ \rightarrow B$ 
9      $\mathcal{P}'_{AB} \leftarrow \text{sum}\{\mathcal{P}'_{AB}, \text{sum}[\max\{\mathcal{P}^2_{AC_i}, \mathcal{P}^2_{C_i B}\}, \forall i \text{ s.t. } A \rightarrow C_i \rightarrow B]\}$ 
10  end
11  if  $A - B$  non-ambiguous and  $\exists i, j$  s.t.  $A - C_i \rightarrow B, A - C_j \rightarrow B$  with  $A - C_i$ 
      and  $A - C_j$  non-ambiguous, and  $C_i$  and  $C_j$  non-adjacent in  $\widehat{\mathbb{G}}$  then
12    Replace  $A \circ \circ B$  in  $\widehat{\mathbb{G}}'$  with  $A \circ \rightarrow B$ 
13     $\mathcal{P}'_{AB} \leftarrow \text{sum}\{\mathcal{P}'_{AB}, \text{sum}[\max\{\mathcal{P}^1_{AC_i}, \mathcal{P}^2_{C_i B}, \mathcal{P}^1_{AC_j}, \mathcal{P}^2_{C_j B}\}, \forall i, j \text{ s.t. } A - C_i \rightarrow$ 
         $B, A - C_j \rightarrow B \text{ with } A - C_i \text{ and } A - C_j \text{ non-ambiguous, and } C_i \text{ and } C_j$ 
        non-adjacent  $]\}$ 
14  end
15  for each  $A \leftrightarrow B$  in  $\widehat{\mathbb{G}}'$  do
16    Unorient to  $A - B$  in  $\widehat{\mathbb{G}}'$ 
17    For each edge in the sufficient conditions of each orientation rule that led to
        the creation of a directed edge to  $A$  in  $\widehat{\mathbb{G}}$ , unorient the edge in  $\widehat{\mathbb{G}}'$ 
18    For each edge in the sufficient conditions of each orientation rule that led to
        the creation of a directed edge to  $B$  in  $\widehat{\mathbb{G}}$ , unorient the edge in  $\widehat{\mathbb{G}}'$ 
19    Label the unoriented edges as “ambiguous” in  $\widehat{\mathbb{G}}'$ 
20  end
21   $\widehat{\mathbb{G}} \leftarrow \widehat{\mathbb{G}}'$ 
22  for each  $A \rightarrow B$  in  $\widehat{\mathbb{G}}$  do
23    Place a unique identifier into  $\mathcal{I}_{AB}$ 
24     $\mathcal{P}^2_{AB} \leftarrow \max\{\mathcal{P}^1_{AB}, \mathcal{P}'_{AB}\}$ 
25  end
26  Empty  $\mathcal{P}'$ 
27 until there are no more edges to orient;
28 for each non-empty cell  $\mathcal{P}^1_{AB}$  s.t.  $\mathcal{P}^2_{AB}$  and  $\mathcal{P}^2_{BA}$  are empty do
29    $\mathcal{P}^2_{AB}, \mathcal{P}^2_{BA} \leftarrow \mathcal{P}^1_{AB}$ 
30 end
    
```

Algorithm 3: Orientation Rules

Data: $\widehat{\mathbb{G}}, \mathcal{P}^2, \mathcal{I}, \alpha, q$

Result: $\widehat{FDR}_{BY}(\alpha), \widehat{\mathbb{G}}^*$

// Estimation

- 1 $\widehat{FDR}_{BY}(\alpha) \leftarrow$ Solution of 2 using threshold α and m corresponding to the number of unique identifiers in \mathcal{I}

// Control

- 2 $\alpha^* \leftarrow$ Solution of 3 using FDR level q and m corresponding to the number of unique identifiers in \mathcal{I}
- 3 $\widehat{\mathbb{G}}^* \leftarrow \widehat{\mathbb{G}}$ with edges associated with p-values above α^* eliminated

Algorithm 4: FDR Estimation and Control

Theorem 6 *The PC-p algorithm with a CI oracle is sound and complete.*

Proof PC is sound and complete, so it is enough to prove that PC-p and PC will perform the exact same edge deletion and edge orientation operations with a CI oracle. Note that Algorithm 1 has already been shown to be sound and complete up to skeleton discovery (Colombo & Maathius 2014). Algorithm 1 will therefore perform the exact same edge deletions as Algorithm 5 with a CI oracle. Now, Algorithm 2 will also perform the same edge orientations as Algorithm 6 with a CI oracle, since there will never be conflicting edge orientations. Lastly, for Algorithm 3, if there exists an edge that can be oriented by more than rule, then the edge must be oriented in the same direction by the other two rules. Algorithm 3 therefore returns the same edge orientations as Algorithm 7. We have proved equivalence in outputs of Algorithms 1, 2 and 3 of PC-p to Algorithms 5, 6 and 7 of PC, respectively. Algorithm 4 is not involved in graph structure discovery. ■

The output of PC-p is therefore equivalent to the output of PC in the large sample limit with a consistent CI test, even though PC-p performs more operations than PC.

5.5 Computation of the P-Values

We now address the issue of computing the upper bounds of the p-values. Let us first consider Algorithm 1. Algorithm 1 takes as input the dataset \mathbf{X}^n and the significance threshold α . The algorithm then stores the p-values of all significant CI tests in cell \mathcal{P}^1

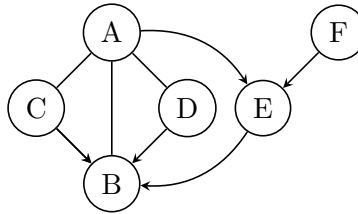


Figure 7: An example of a situation where two of PC’s orientation rules, specifically rules 2 and 3, can orient one undirected edge ($A - B$) in the same direction. In this case, PC oriented all of the currently directed edges using unshielded v-structures.

when it reaches line 14. Notice that the algorithm stores the p-values of all significant tests involving A and B in both \mathcal{P}_{AB}^1 and \mathcal{P}_{BA}^1 . Algorithm 1 next computes the maximum over the p-values for all surviving edges in line 24 as in (16).

Algorithm 2 takes \mathcal{P}^1 from Algorithm 1 as input. Moreover, unlike Algorithm 6 of PC, Algorithm 2 also takes as input the dataset \mathbf{X}^n , since PC-p must apply (29) in order to obtain the upper bounds of the p-values for oriented unshielded v-structures. Indeed, Algorithm 2 executes $test_{A \perp\!\!\!\perp B | \mathbf{S}}$ for all $\mathbf{S} \subseteq \widehat{\mathcal{N}}(A)$ containing C and all $\mathbf{S} \subseteq \widehat{\mathcal{N}}(B)$ containing C in steps 4-12 for each $A - C - B$ such that A and B are non-adjacent and $C \notin \mathcal{S}_{AB}$. Now, Algorithm 2 ultimately stores all of the p-values needed to compute $p_{\gamma_{AB|C}}$ as in (20) in \mathcal{P}'' via line 10. Algorithm 2 then stores the maximum over p_{A-C} and $p_{\gamma_{AB|C}}$ in \mathcal{P}'_{BC} instead of \mathcal{P}'_{AC} in line 13. A similar set of operations eventually stores the maximum over p_{B-C} and $p_{\gamma_{AB|C}}$ into \mathcal{P}'_{AC} in line 14. Note that multiple elements can enter into \mathcal{P}'_{BC} and \mathcal{P}'_{AC} when multiple v-structures can orient one edge. Finally, in line 27, Algorithm 2 takes the maximum over \mathcal{P}_{AC}^1 as returned from Algorithm 1 and the sum of \mathcal{P}'_{AC} to obtain $p_{A \rightarrow C}$ in \mathcal{P}^2 according to (29) and similarly takes the maximum over \mathcal{P}_{BC}^1 and the sum of \mathcal{P}'_{BC} to obtain $p_{B \rightarrow C}$ in \mathcal{P}^2 .

Algorithm 3 takes \mathcal{P}^2 from Algorithm 2 as input. Next, in rule 1, Algorithm 3 adds up the p-values associated with $C_i \rightarrow A, \forall i$ and places the result in \mathcal{P}'_{AB} in line 5 for computing (31). Then, Algorithm 3 sums over the maxima of $p_{A \rightarrow C_i}$ and $p_{C_i \rightarrow B}$ in rule 2 $\forall i$ s.t. $A \rightarrow C_i \rightarrow B$ in line 9 for ultimately computing (33). Subsequently, in rule 3, Algorithm 3 finds all n edges such that $A - C_i \rightarrow B$. The algorithm then finds all of the n choose 2 pairs, say r of them. For each pair, say $A - C_1 \rightarrow B$ and $A - C_2 \rightarrow B$, Algorithm 3 computes $p_{A-C_1 \rightarrow B}$ and $p_{A-C_2 \rightarrow B}$ as the maximum over p_{A-C_1} and $p_{C_1 \rightarrow B}$ and the maximum over p_{A-C_2} and $p_{C_2 \rightarrow B}$, respectively. Algorithm 3 next sums the p-values over all r pairs in line 13 for computing (35). Note that Algorithm 3 also takes an outer-sum involving \mathcal{P}'_{AB} in lines 5, 9 and 13 of rules 1, 2 and 3, respectively; these summations correspond to logical disjunctions when multiple orientation rules can orient one edge in the same direction. For example, rules 2 and 3 can orient $A - B$ in the same direction in Figure 7. Two applications of rule 1 can also orient $A - B$ in the same direction in Figure 6, if we remove one of the unshielded v-structures from the graph. Now, for all non-ambiguous edges, Algorithm 3 then stores the maximum over the p-values from Algorithm 1 and \mathcal{P}' into \mathcal{P}^2 in line 24. This process is repeated until no more edges can be oriented. Algorithm 3 finally transfers the p-values of all of the remaining undirected edges in $\widehat{\mathcal{G}}$ from \mathcal{P}^1 to \mathcal{P}^2 in lines 28- 30. The algorithm therefore eventually outputs all of the final p-values in \mathcal{P}^2 as desired.

5.6 Controlling the False Discovery Rate

PC-p controls the FDR per hypothesis test as opposed to per edge, since the algorithm can sometimes orient two edges according to the same hypothesis test during unshielded v-structure discovery. Indeed, controlling the p-values per edge as opposed to per hypothesis test can result in overly conservative FDR estimation or control because an FDR estimator or controlling procedure may count the p-value of one hypothesis test multiple times.

PC-p keeps track of each distinct hypothesis test in Algorithms 1, 2 and 3 by using indexing cell \mathcal{I} as follows. First, Algorithm 1 assigns the same, unique identifier to the p-

value bounds in both \mathcal{P}_{AB} and \mathcal{P}_{BA} in line 25. Next, if we have one v-structure $A \rightarrow C \leftarrow B$ that orients $A - C$ and $C - B$, then Algorithm 2 associates both $A \rightarrow C$ and $C \leftarrow B$ with the same hypothesis test and therefore the same identifier in line 29. On the other hand, if multiple unshielded v-structures can orient one edge, then Algorithm 2 assigns the edge a unique identifier in line 31, since a unique hypothesis test exists per edge in this case. Algorithm 3 finally assigns a unique identifier to each newly oriented edge in line 23 because each newly oriented edge also corresponds to a distinct hypothesis test.

We can now use Algorithm 4 to estimate and control the FDR using the identifiers in \mathcal{I} and the p-value bounds in \mathcal{P}^2 as returned from Algorithm 3. Algorithm 4 estimates the FDR by solving 2 to obtain \widehat{FDR}_{BY} , where m corresponds to the number of unique identifiers in \mathcal{I} . The algorithm subsequently controls the FDR by solving 3 to obtain α^* . Algorithm 4 then eliminates all edges with p-values below α^* in \mathcal{P}^2 in order to obtain $\widehat{\mathbb{G}}^*$; this process ensures that the expected FDR does not exceed q in $\widehat{\mathbb{G}}^*$.

5.7 Conclusion

We wrap-up this section with the following theorem:

Theorem 7 *Consider the same assumptions as Theorem 4. Then PC-p achieves conservative point estimation and strong control of the FDR across the edges in $\widehat{\mathbb{G}}$.*

Proof We have already shown that PC-p can control the p-values of all of the edges in $\widehat{\mathbb{G}}$ from Theorem 4. Estimation follows because the solution of 2 achieves conservative point estimation of the FDR at threshold α when the p-values are controlled (Benjamini and Yekutieli, 2001). Similarly, control follows because eliminating the edges associated with p-values above α^* as obtained from 3 achieves strong control of the FDR at level q when the p-values are in turn controlled (Benjamini and Yekutieli, 2001). ■

The PC-p algorithm thus corresponds to a valid method for estimating and controlling the FDR in the estimated CPDAG.

Note finally that PC-p takes slightly longer than original PC to complete because it performs extra computations. However, PC-p runs at approximately the same speed as PC-stable, since v-structure detection and orientation rule application take an infinitesimal amount of time compared to skeleton discovery.

6. Experiments

6.1 Algorithms and Metrics

We evaluated six algorithms:

1. PC-p,
2. PC-p without stabilization in the skeleton discovery procedure,
3. PC-p without ambiguous labelings during v-structure orientation and orientation rule application (PC-p without ambiguity),
4. PC-p without both stabilization and ambiguity,

5. PC-p without hypothesis tests with robust p-value bounds – we chose the null hypotheses to be a series of logical conjunctions so that no edges are present between any of the variables. As a result, the p-values take on minimal values as described in Appendix A.2. We call this procedure PC-p without robust p-values.
6. The original PC algorithm with p-value computation – that is, we do not incorporate stabilization, and the algorithm arbitrarily over-writes edge orientations. We compute p-values according to the v-structure or rule which ultimately orients each edge in the CPDAG. The algorithm also performs some additional CI tests in order to compute 20 as described in Section 4.1.

We ran these six algorithms because they are the only algorithms that allow us to compute the FDR across the entire CPDAG from the estimated p-values.

We assessed the FDR of the above six algorithms in detail using control and estimation bias³. An algorithm exhibits low control bias at FDR level q when an FDR controlling procedure can accurately eliminate edges in the CPDAG using the p-values so that the FDR is in fact q . On the other hand, an algorithm exhibits low estimation bias when an FDR estimate closely matches the true FDR of the CPDAG. Notice that both control and estimation bias are important and can serve different purposes. As a result, we prefer an algorithm that exhibits both low control and estimation bias.

We used the mean of the following quantities to assess control bias:

$$\begin{aligned} uc(\widehat{FDR}_{BY}, q) &:= \max\{FDR(\alpha^*) - q, 0\}, \\ oc(\widehat{FDR}_{BY}, q) &:= \max\{q - FDR(\alpha^*), 0\}, \end{aligned} \tag{36}$$

where $uc(\widehat{FDR}_{BY}, q)$ denotes under-control at FDR level q with the BY FDR estimate, and $oc(\widehat{FDR}_{BY}, q)$ similarly denotes over-control. In the experiments, we varied q from $[0.001, 0.1]$ using 100 equispaced intervals. Note that we compute both under-control and over-control per CPDAG. A method achieves strong control when the mean under-control taken across the hypothesis tests is zero (Armen and Tsamardinos, 2014). Moreover, the less the mean over-control, the tighter the strong control. As a result, achieving a lower mean under-control is more important than achieving a lower mean over-control. We therefore say that one method outperforms another if the method achieves a lower mean under-control while also maintaining a reasonably low mean over-control.

We used the mean of the following similar quantities for estimation bias:

$$\begin{aligned} ue(\widehat{FDR}_{BY}, \alpha) &:= \max\{FDR(\alpha) - \widehat{FDR}_{BY}(\alpha), 0\}, \\ oe(\widehat{FDR}_{BY}, \alpha) &:= \max\{\widehat{FDR}_{BY}(\alpha) - FDR(\alpha), 0\}, \end{aligned} \tag{37}$$

where $ue(\widehat{FDR}_{BY}, \alpha)$ denotes under-estimation at threshold level α with the BY FDR estimate, and $oe(\widehat{FDR}_{BY}, \alpha)$ similar denotes over-estimation. We varied the α threshold from $[1E-10, 0.1]$ with 100 equispaced intervals in the experiments. Now, we say that estimation is conservative in a α threshold region when the underestimation is zero. Moreover, the

3. We also measured the false negative rate using the structural Hamming distance as a metric in Figure 19 of the Appendix.

greater the over-estimation in a p-value threshold region, the more conservative the estimate. A method should conservatively estimate the FDR but not do so over-conservatively. As a result, achieving lower under-estimation is more important than achieving lower over-estimation, and one method outperforms another if the method achieves a lower mean under-estimation while maintaining a reasonably low mean over-estimation.

Below, we report the relative performance differences of the six algorithms in recovering the CPDAG at a liberal α threshold of 0.20, since this threshold consistently provided a nice tradeoff between p-value bound looseness and low Type II error rates. We have reported the results using other α thresholds of 0.01, 0.05, 0.10, and 0.15 or 0.50 in Figures 12-15 of Appendix A.3, with similar relative performance differences between the algorithms. Figures 16-18 in the Appendix also contain results for skeleton discovery, where we compared the original skeleton discovery procedure of PC against the same procedure with stabilization. As expected, the stabilization procedure improved performance. We finally provide results with the more commonly used structural Hamming distance in 19 of the Appendix; here, PC-p achieved superior performance by conservatively estimating the graph.

Note that for the simulations in Sections 6.2 and 6.3, we generated the DAGs using the TETRAD V package (version 5.2.1) by drawing uniformly over all DAGs with a maximum in-degree of 2 and a maximum out-degree of 2. We then converted each of the DAGs to linear non-recursive SEM-IEs by 1) drawing the linear coefficients from independent standard normal distributions, and 2) setting independent Gaussian distributions over the error terms with standard deviations also drawn from the standard normal. Each linear SEM-IE with the error distributions therefore induced a multivariate Gaussian distribution across the observed variables. We finally ran all of the six algorithms using Fisher’s z-test with a liberal α threshold of 0.20 and a maximum conditioning set size of 2.

6.2 Low Dimensional Inference

We generated 30 DAGs by drawing uniformly over all DAGs with 20 vertices. We converted each of the DAGs to 5 linear non-recursive SEM-IEs. We subsequently created 5 datasets using each linear SEM-IE with sample sizes of 100, 500, 1000, 5000, and 10000. We therefore created a total of $30 \times 5 \times 5 = 750$ datasets.

We analyzed the ability of the algorithms in correctly estimating the CPDAG in terms of the four metrics proposed in Section 6.1 as well as the FDR values. Results as averaged over DAGs, parameters and sample sizes are summarized in Figure 8. We assessed the significance of all inter-algorithm differences using paired Wilcoxon signed rank tests. PC-p obtained lower mean FDR values than PC-p without robust p-values (Figures 8a and 8b; $z = -4.782, p = 1.734E-6$), PC-p without stabilization ($z = -4.371, p = 1.238E-5$), PC-p without ambiguation ($z = -4.782, p = 1.734E-6$), PC-p without both stabilization and ambiguation ($z = -4.782, p = 1.734E-6$), and PC original ($z = -4.433, p = 9.316E-6$). Moreover, PC-p achieved significantly lower mean under-control than the competing methods (Figures 8c and 8d; vs. no robust: $z = -4.782, p = 1.734E-6$; vs. no stable: $z = -3.898, p = 9.711E-5$; vs. no ambig: $z = -4.782, p = 1.734E-6$; vs. no stable & no ambig: $z = -4.782, p = 1.734E-6$; vs. PC original: $z = -4.700, p = 2.603E-6$); meanwhile, PC-p kept the mean over-control small at 3.865% (SD: 0.795%).

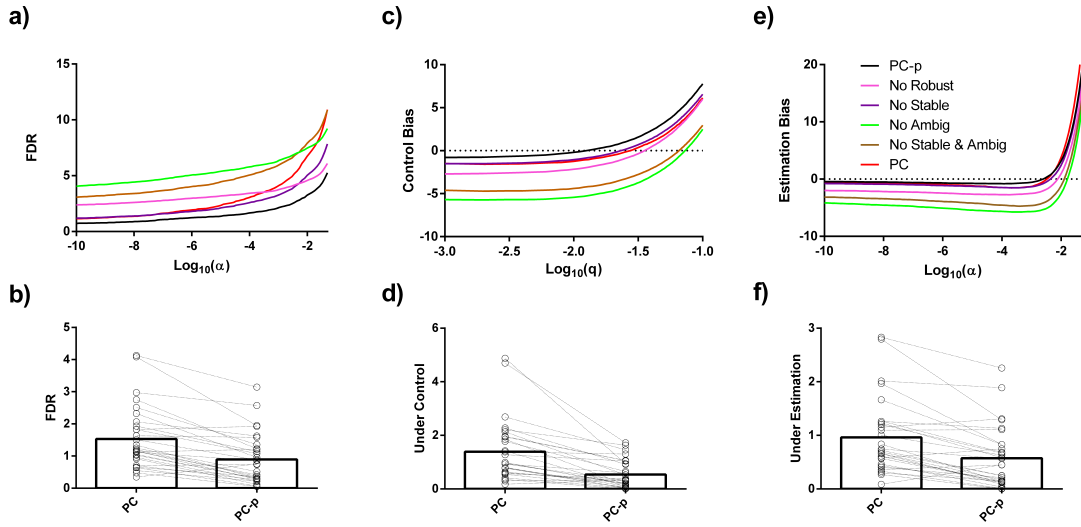


Figure 8: Performances of PC-p, PC-p without robust p-values (no robust), PC-p without ambiguation (no ambig), PC-p without stabilization (no stable), PC-p without ambiguation and stabilization (no stable & no ambig), and the original PC algorithm (PC) as assessed by (a,b) the FDR, (c,d) control bias, and (e,f) estimation bias in units of percent. PC-p achieved significantly lower FDR, under-estimation and under-control than the other five methods suggesting that robust p-values, stabilization and ambiguation are all important components of PC-p.

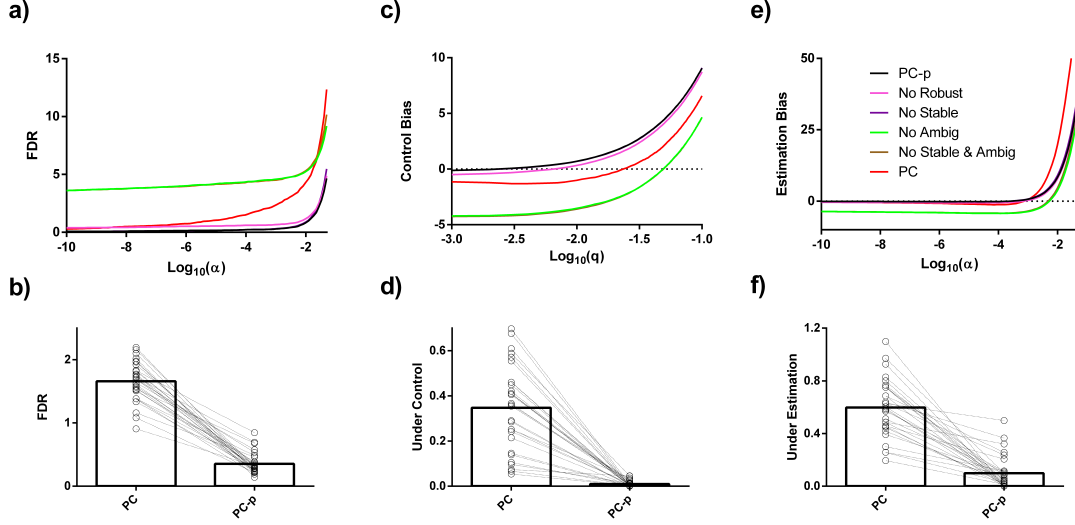


Figure 9: Same setup as Figure 8 except with high dimensional data. PC-p significantly outperformed all other methods except PC-p without stabilization in terms of under-control and under-estimation.

Results for estimation were similar. PC-p achieved significantly lower mean under-estimation than the competing methods (Figures 8e and 8f; vs. no robust: $z = -4.782, p = 1.734\text{E-}6$; vs. no stable: $z = -4.206, p = 2.597\text{E-}5$; vs. no ambig: $z = -4.782, p = 1.734\text{E-}6$; vs. no stable & no ambig: $z = -4.782, p = 1.734\text{E-}6$; vs. PC original: $z = -4.186, p = 2.843\text{E-}5$). PC-p also achieved a small degree of mean over-estimation (8.222%, SD: 0.400%). We conclude that robust p-values, stabilization, and ambiguation all help PC-p achieve the lowest under-control and under-estimation.

6.3 High Dimensional Inference

We next tested PC-p and the other five algorithms on high dimensional graph estimation. To do this, we generated thirty 100, thirty 200 and thirty 300 variable DAGs. We subsequently converted each of the DAGs to one linear non-recursive SEM-IE. Finally, we generated 1000 samples from each SEM-IE in order to obtain sample size to variable ratios of 10, 5 and 3.333.

Results are summarized using the FDR, control bias, and estimation bias metrics as averaged over the DAGs and their parameters in Figure 9. PC-p achieved similar results in low dimensions as it did for high dimensions. Specifically, PC-p obtained lower mean FDR values across the same α thresholds than PC-p without robust p-values (Figures 9a and 9b; $z = -4.703, p = 2.563\text{E-}6$), PC-p without stabilization ($z = -2.232, p = 0.026$), PC-p without ambiguation ($z = -4.782, p = 1.734\text{E-}6$), PC-p without both stabilization and ambiguation ($z = -4.782, p = 1.734\text{E-}6$), and PC original ($z = -4.782, p = 1.734\text{E-}6$). Moreover, PC-p achieved significantly lower mean under-control than four of the five competing methods

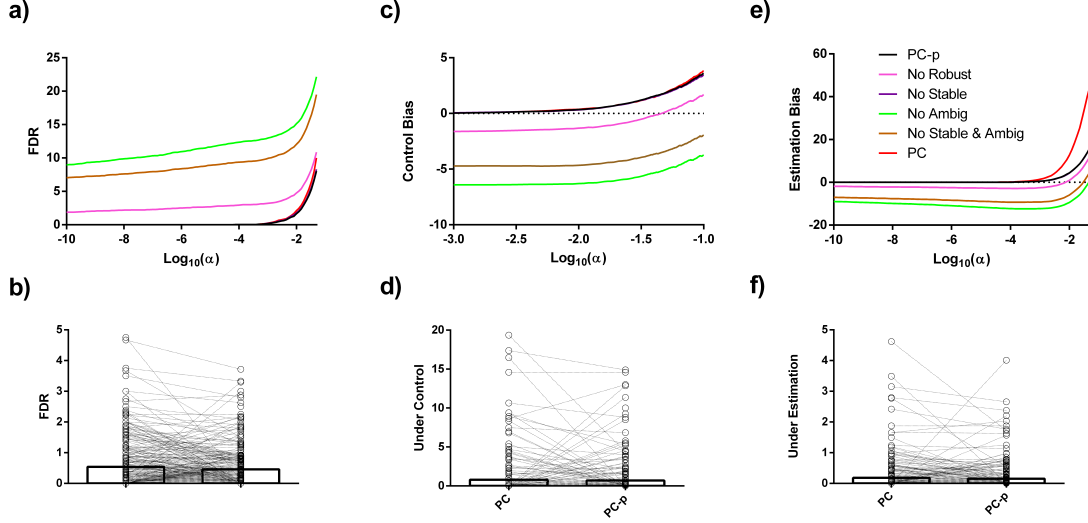


Figure 10: Same setup as Figure 8 except with the CYTO dataset. PC-p significantly outperformed all methods across all metrics except the original PC algorithm in under-control.

(Figures 9c and 9d; vs. no robust: $z = -4.623, p = 3.790E-6$; vs. no ambig: $z = -4.782, p = 1.734E-6$; vs. no stable & no ambig: $z = -4.782, p = 1.734E-6$; vs. PC original: $z = -4.782, p = 1.734E-6$). PC-p did not outperform PC-p without stabilization at four of the five threshold α thresholds tested (0.05 : $z = -1.121, p = 0.262$; 0.10 : $z = 0.504, p = 0.614$; 0.20 : $z = 0.985, p = 0.324$; 0.50 : $z = 0.760, p = 0.447$); however, PC-p did outperform PC-p without stabilization at an α threshold of 0.01 ($z = -3.692, p = 2.225E-4$). Meanwhile, PC-p kept the mean over-control small at 4.507% (SD: 0.240%).

Results for estimation were again similar. PC-p achieved significantly lower mean under-estimation than the competing methods except PC-p without stabilization (Figures 9e and 9f; vs. all methods except no stable: $z = -4.782, p = 1.734E-6$). PC-p did not outperform PC-p without stabilization at four of the five threshold α thresholds tested (0.05 : $z = -1.820, p = 0.069$; 0.10 : $z = 0.175, p = 0.861$; 0.20 : $z = 0.625, p = 0.532$; 0.50 : $z = 0.608, p = 0.543$); however, PC-p did outperform PC-p without stabilization at an α threshold of 0.01 ($z = -2.293, p = 0.028$). PC-p also achieved a small degree of mean over-control (2.129%, SD: 0.084%).

We conclude that the results for control and estimation bias for high dimensional graph estimation are similar to the low dimensional case. However, stabilization only increased performance at lower α thresholds in the high dimensional scenario; we may nonetheless view this as a desirable property, since a lower α threshold helps the algorithm complete more quickly.

6.4 Real Data: CYTO

We evaluated the six algorithms on the CYTO dataset which contains single cell recordings of the abundance of 11 phosphoproteins and phospholipids in human primary naive CD4+ T cells using flow cytometry (Sachs et al., 2005). The variables in the dataset and their causal relationships can be represented as a DAG, where vertices are proteins or lipids and edges are phosphorylation interactions between the proteins and lipids. We used the general perturbation samples (i.e., CD3-CD28 and CD3-CD28-ICAM2) as our observational data; these perturbations are required to activate the phosphorylation pathways. Note that algorithms typically cannot accurately infer the gold standard solution set using the observational data alone, as noted by the original authors⁴. As a result, we created a silver standard DAG by running LiNGAM as implemented in TETRAD V using default parameters on the full dataset of 1,755 samples; recall that LiNGAM is a method within a different class of causal discovery algorithms based on functional causal models. We then ran the six algorithms described in Section 6.1 on 1000 bootstrapped datasets of sample size 100 using Spearman’s rho to handle the class of non-paranormal distributions.

We have summarized the results in Figure 10. PC-p obtained lower mean FDR across the same α thresholds than PC-p without robust p-values ($z = -12.774, p = 2.282\text{E-}37$), PC-p without stabilization ($z = -8.402, p = 4.378\text{E-}17$), PC-p without ambiguity ($z = -24.598, p = 1.343\text{E-}133$), PC-p without both stabilization and ambiguity ($z = -23.714, p = 2.616\text{E-}124$), and original PC ($z = -4.924, p = 8.469\text{E-}7$). Moreover, PC-p achieved significantly lower mean under-control than four of the five competing methods (vs. no robust: $z = -12.601, p = 2.081\text{E-}36$; vs no stable: $z = -4.310, p = 1.631\text{E-}5$; vs. no ambig: $z = -24.339, p = 7.559\text{E-}131$; vs. no stable & no ambig: $z = -22.893, p = 5.515\text{E-}116$). PC-p did not outperform the original PC algorithm in mean under-control ($z = 0.827, p = 0.408$); however, PC-p did outperform the original PC algorithm in mean under-estimation ($z = -2.662, p = 0.008$). Meanwhile, PC-p kept the mean over-control small at 4.232% (SD: 1.635%). PC-p also outperformed the other four methods in mean under-estimation (vs. no robust: $z = -12.684, p = 7.289\text{E-}37$; vs. no stable: $z = -4.893, p = 9.917\text{E-}7$; vs. no ambig: $z = -24.411, p = 1.322\text{E-}131$; vs. no stable & no ambig: $z = -23.301, p = 4.333\text{E-}120$) while maintaining low mean over-estimation at 1.200% (SD: 0.559%). We conclude that PC-p outperforms the other methods similar to the results with synthetic data. PC-p only outperformed PC in 2 of the 3 metrics, however, probably because the LiNGAM solution is only an estimate of the ground truth.

6.5 Real Data: GDP Dynamics

One way of approximating the underlying DAG involves learning the graph with a large number of samples. Another way uses time series data, where we know a priori that we must have contemporaneous causal relations or causal relations directed forward in time. In this experiment, we strip the time information from the six algorithms, and then identify the false discoveries when algorithms mistakenly detect a causal relation directed backwards in time. We used a time series dataset downloaded from the Economic Research Service of the United States Department of Agriculture containing ten economic indicators per year

4. In general, we do not have gold standard causal graphs for real data, so we must approximate the solution in some manner.

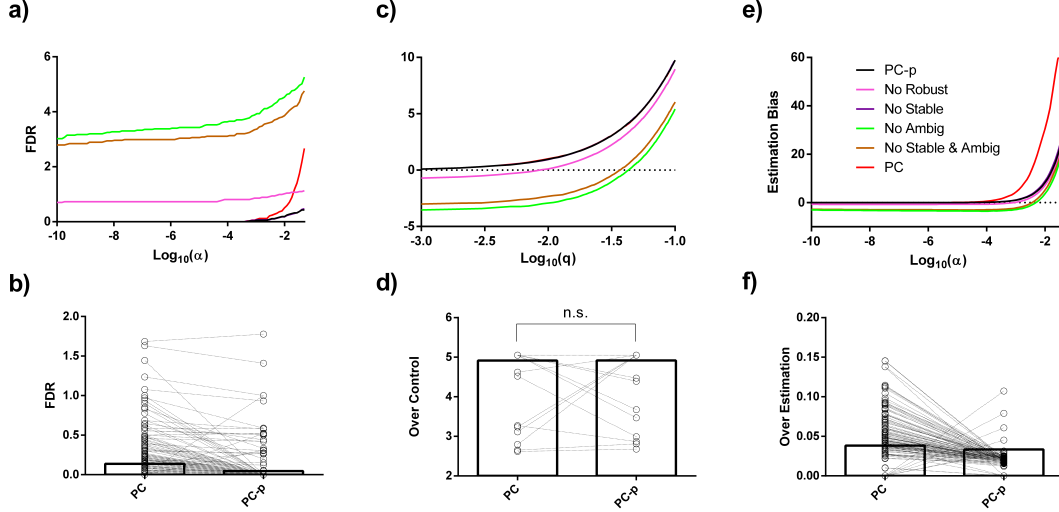


Figure 11: Similar to Figure 8 except with the GDP dataset as well as over-control and over-estimation bias values instead of under. PC-p did not achieve lower under-control and under-estimation than PC, but it did achieve significantly lower mean FDR and over-estimation than PC.

related to GDP among 192 countries⁵. We specifically evaluated the algorithms on their ability to discover causal relations among the indicators within and between 1987, 1988 and 1989, where we treated each country as an i.i.d. sample and used 100 bootstrapped datasets.

We have summarized the results in Figure 11. PC-p again obtained lower mean FDR values across the α thresholds than PC-p without robust p-values ($z = -4.623, p = 3.784E-6$), PC-p without ambiguity ($z = -8.054, p = 4.128E-16$), PC-p without both stabilization and ambiguity ($z = -8.135, p = 4.128E-16$), and original PC ($z = -6.624, p = 3.500E-11$). However, PC-p did not obtain significantly lower mean FDR values than PC-p without stabilization (signed-rank = 70, $p = 0.600$). Next, PC-p achieved significantly lower mean under-control than three of the five competing methods including PC-p without robust p-values ($z = -4.374, p = 1.218E-5$), PC-p without ambiguity ($z = -7.867, p = 3.647E-15$), and PC-p without both stabilization and ambiguity ($z = -7.819, p = 5.306E-15$). PC-p did not outperform PC-p without stabilization (signed-rank = 3, $p = 1$) as well as the original PC algorithm (signed-rank = 7, $p = 0.625$) in mean under-control; nevertheless, PC-p did outperform the former in over-control (signed-rank = 3, $p = 0.020$) while keeping its own mean over-control low at 4.920% (SD: 0.483%). PC-p also outperformed the same three methods in mean under-estimation (vs. no robust: $z = -4.372, p = 1.229E-5$; vs. no ambig: $z = -7.818, p = 5.363E-15$; vs. no stable & no ambig: $z = -7.770, p = 7.850E-15$) while maintaining low mean over-estimation at 2.032% (SD: 0.281%). PC-p again did not outperform PC-p without stabilization (signed-rank = 2, $p = 0.750$) and original

5. Web link: <http://www.ers.usda.gov/data/macroeconomics/Data/HistoricalRealGDPValues.xls>

PC (signed-rank = 7, $p = 0.625$) in under-estimation, but it outperformed both in over-estimation (no stable: $z = -7.386$, $p = 1.519\text{E-}13$; PC original: $z = -8.682$, $p = 3.897\text{E-}18$). We conclude that PC-p outperforms most methods in either the under or over-metrics. The results however are not as clean as the results with the synthetic data because we only have access to a portion of the ground truth.

7. Conclusion

We developed a new algorithm called PC-p which outputs a causal DAG with p-value bounds associated with each edge. One can then use the bounds with the BY procedure to achieve almost strong control and estimation of the FDR. The PC-p algorithm specifically integrates the skeleton discovery procedure of PC-stable, edge orientation with ambiguation, and robust hypothesis tests in order to accurately estimate p-values bounds while maintaining computational efficiency.

The PC-p algorithm represents the first global constraint-based method which can recover p-value estimates for every edge of a CPDAG. In our opinion, the algorithm is a significant advancement over previous methods which can only achieve strong control of the FDR under special conditions. Moreover, PC-p lays a foundation for developing similar methods which can also recover edge-specific p-values and achieve strong control of the FDR for graphs recovered by algorithms such as FCI and CCD. In particular, we suspect that a combination of the max and union bounds will also be sufficient for deriving upper bounds of the edge-specific p-values for more sophisticated constraint-based methods. The proposed approach may therefore represent one the earliest forms of a “causal p-value.”

Now readers may wonder whether PC-p can also use the p-values to control the family-wise error rate (FWER). The answer is yes, and we recommend using the Benjamini-Holm step-down procedure as opposed to Hochberg’s step-up procedure to control the FWER (Hochberg, 1988), since the latter assumes positive dependency among the test statistics. However, application of an FWER controlling procedure to constraint-based causal discovery requires additional justification, since most investigators do not use constraint-based methods to definitively conclude causal relationships but rather to screen for potential causal variables. With the screening goal in mind, the FWER may be too conservative in practice, since it controls the rate of making a single Type I error across all of the hypothesis tests as opposed to controlling the proportion of Type I errors.

In summary, we introduced an algorithm called PC-p which outputs a causal DAG along with edge-specific p-value bounds. One can then use the BY procedure with the bounds to achieve almost strong control or estimation of the FDR and therefore assess the algorithm’s confidence in each edge in a principled manner. We ultimately hope that this work will encourage more applications of constraint-based causal discovery to important problems in science.

Acknowledgments

Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-

NIH Big Data to Knowledge initiative. The research was also supported by the National Library of Medicine of the National Institutes of Health under award numbers T15LM007059 and R01LM012095. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Appendix

A.1 PC Algorithm Pseudocode

We provide pseudocode for the original PC algorithm. We summarize skeleton discovery in Algorithm 5, unshielded v-structure discovery in Algorithm 6, and orientation rule application in Algorithm 7.

Data: \mathbf{X}^n, α
Result: $\hat{\mathbb{G}}, \mathcal{S}$

```

1 Form a completely connected undirected graph  $\hat{\mathbb{G}}$  on the vertex set  $\mathbf{X}$ 
2  $l = -1$ 
3 repeat
4    $l = l + 1$ 
5   repeat
6     Select a new ordered pair of variables  $(A, B)$  that are adjacent in  $\hat{\mathbb{G}}$  s.t.
        $|\mathbf{N}(A) \setminus B| \geq l$ 
7     repeat
8       Choose a new set  $\mathbf{S} \subseteq \{\mathbf{N}(A) \setminus B\}$  with  $|\mathbf{S}| = l$  using  $order(\mathbf{X})$ 
9        $p \leftarrow$  p-value from  $test_{A \perp\!\!\!\perp B | \mathbf{S}}$ 
10      if  $p > \alpha$  then
11        Delete  $A - B$  from  $\hat{\mathbb{G}}$ 
12        Insert  $\mathbf{S}$  into  $\mathcal{S}_{A,B}$  and  $\mathcal{S}_{B,A}$ 
13      end
14    until  $A - B$  is deleted from  $\hat{\mathbb{G}}$  or all  $\mathbf{S} \subseteq \{\mathbf{N}(A) \setminus B\}$  with  $|\mathbf{S}| = l$  have been
      considered;
15  until all ordered pairs of adjacent variables  $(A, B)$  in  $\hat{\mathbb{G}}$  with  $|\hat{\mathbf{N}}(A) \setminus B| \geq l$  have
    been considered;
16 until all ordered pairs of adjacent variables  $(A, B)$  in  $\hat{\mathbb{G}}$  satisfy  $|\hat{\mathbf{N}}(A) \setminus B| \leq l$ ;

```

Algorithm 5: Skeleton Discovery

Data: $\hat{\mathbb{G}}, \mathcal{S}$
Result: $\hat{\mathbb{G}}$

```

1 for all ordered pairs of non-adjacent variables  $(A, B)$  with common neighbor  $C$  do
2   if  $C \notin$  any set in  $\mathcal{S}_{i,j}$  then
3     Replace  $A - C - B$  with  $A \rightarrow C \leftarrow B$ 
4   end
5 end

```

Algorithm 6: Unshielded V-structures

Data: $\hat{\mathbb{G}}$
Result: $\hat{\mathbb{G}}$

```

1 repeat
2   if  $A - B$  and  $\exists C$  s.t.  $C \rightarrow A$ , and  $C$  and  $B$  are non-adjacent then
3     | Replace  $A - B$  with  $A \rightarrow B$ 
4   else if  $A - B$  and  $\exists C$  s.t.  $A \rightarrow C \rightarrow B$  then
5     | Replace  $A - B$  with  $A \rightarrow B$ 
6   else if  $A - B$  and  $\exists B, D$  s.t.  $A - C \rightarrow B$ ,  $A - D \rightarrow B$ , and  $C$  and  $D$  are
      non-adjacent then
7     | Replace  $A - B$  with  $A \rightarrow B$ 
8   end
9 until there are no more edges to orient;

```

Algorithm 7: Orientation Rules

A.2 Hypothesis Tests with Less Robust Bounds

We claimed to propose edge-specific hypothesis tests whose bounds are robust to Type II errors in Section 4.4. We now explain our rationale.

Consider the following modification to (10), where we have replaced the null with a series of logical conjunctions:

$$\begin{aligned}
 H_0 &: \bigwedge_{i=1}^m \text{oracle } i \text{ outputs } \neg P_i, \\
 H_1 &: \bigwedge_{i=1}^m \text{oracle } i \text{ outputs } P_i.
 \end{aligned} \tag{38}$$

We can bound the Type I error rate of the above hypothesis test as follows:

$$\begin{aligned}
 \Pr(\text{Type I error}) &= \Pr\left(\bigwedge_{i=1}^m \text{test } i \text{ outputs } P_i \mid H_0\right) \\
 &\leq \min_{i=1, \dots, m} \Pr(\text{test } i \text{ outputs } P_i \mid H_0) \\
 &= \min_{i=1, \dots, m} \Pr(\text{test } i \text{ outputs } P_i \mid \text{oracle } i \text{ outputs } \neg P_i) \\
 &= \min_{i=1, \dots, m} g_i,
 \end{aligned} \tag{39}$$

We can use the above bound with the following variant of (26) for unshielded v-structures:

$$\begin{aligned}
 H_0 &: \neg(A - C) \wedge \neg(B - C) \wedge \neg\gamma_{AB|C}, \\
 H_1 &: (A - C) \wedge (B - C) \wedge \gamma_{AB|C},
 \end{aligned} \tag{40}$$

The above hypothesis test follows from the following natural hypothesis test:

$$\begin{aligned}
 H_0 &: \text{All edges between } A, C, B \text{ are absent,} \\
 H_1 &: (A - C) \wedge (B - C) \wedge \gamma_{AB|C},
 \end{aligned} \tag{41}$$

since the null of (41) implies the null in (40). From (39), the Type I error rate of (40) is bounded by $\min\{p_{A-C}, p_{B-C}, p_{\gamma_{AB|C}}\}$. This is a less robust bound than (26) in terms of the Type II error rate, since failing to control one p-value can cause an algorithm to under-estimate the bound. For example, suppose the underlying truth corresponds to $p_{A-C} = 0.01$, $p_{B-C} = 0.03$, $p_{\gamma_{AB|C}} = 0.02$. Thus, the Type I error rate of (41) is truly bounded by 0.01. However, suppose a Type II error causes PC-p to skip CI tests and therefore compute $p_{A-C} = 0.01$, $p_{B-C} = 0.03$, $p_{\gamma_{AB|C}} = 0.003$, where the third term is under-estimated. Then, PC-p will under-estimate the bound at 0.003 instead of the true 0.01.

Note that generalizing (41) to account for multiple possible ways of orienting a v-structure does not robustify the bound either, since we have:

$$\begin{aligned} H_0 : & \text{All edges between } A, C, \mathbf{B}_1 \text{ are absent, and all edges between} \\ & A, C, \mathbf{B}_2 \text{ are absent,} \\ H_1 : & (A - C) \wedge \left([(B_1 - C) \wedge \gamma_{AB_1|C}] \vee [(B_2 - C) \wedge \gamma_{AB_2|C}] \right). \end{aligned} \quad (42)$$

whose Type I error rate is bounded by $\min\{p_{A-C}, \min\{p_{B_1-C}, p_{\gamma_{AB_1|C}}\} + \min\{p_{B_2-C}, p_{\gamma_{AB_2|C}}\}\}$.

More broadly, we can consider the following hypothesis test:

$$\begin{aligned} H_0 : & \bigvee_{i=1}^m \left(\bigwedge_{j=1}^{n_i} \text{oracle } i, j \text{ outputs } \neg P_{i,j} \right), \\ H_1 : & \bigwedge_{i=1}^m \left(\bigwedge_{j=1}^{n_i} \text{oracle } i, j \text{ outputs } P_{i,j} \right). \end{aligned} \quad (43)$$

We bound its Type I error rate as follows:

$$\begin{aligned} \Pr(\text{Type I error}) &= \Pr\left(\bigvee_{i=1}^m \bigwedge_{j=1}^{n_i} \text{test } i, j \text{ outputs } P_{i,j} \mid H_0\right) \\ &\leq \max_{i=1, \dots, m} \min_{j=1, \dots, n_i} \Pr(\text{test } i, j \text{ outputs } P_{i,j} \mid H_0) \\ &= \max_i \min_j \Pr(\text{test } i, j \text{ outputs } P_{i,j} \mid \text{oracle } i, j \text{ outputs } \neg P_{i,j}) \\ &= \max_i \min_j g_{i,j}, \end{aligned} \quad (44)$$

The above bound is less robust to Type II errors than (11), since under-estimating one term in each group i composed of n_i terms can cause PC-p to also under-estimate (44).

A.3 Other Experimental Results

We have summarized the results for the low dimensional, high dimensional and real datasets across multiple α thresholds in Figures 12, 13, 14 and 15 respectively. Relative differences in performance largely remained consistent across the thresholds, since PC-p usually achieved the lowest mean FDR, under-control and under-estimation values with minimal increases in mean over-control and over-estimation.

We have also summarized the results for adjacency discovery in Figures 16, 17, and 18, where we tested whether the skeleton discovery procedure of PC with stabilization could

improve the estimation of the p-value bounds relative to the procedure without stabilization. Results show that stabilization improves performance across the three metrics particularly with the low α threshold values of 0.01 and 0.05. Note that we cannot compute the same figures for the GDP dataset, since we can only evaluate relative performance levels based on edge direction in this case.

We have finally summarized the results using the structural Hamming distance in Figure 19. Notice that ambiguation helps PC-p achieve significantly lower Hamming distances across multiple α thresholds by forcing the algorithm to conservatively orient the edges. Again, we cannot compute the structural Hamming distances for the GDP dataset for the aforementioned reason.

References

- A. P. Armen and I. Tsamardinos. A unified approach to estimation and control of the false discovery rate in bayesian network skeleton identification. In *In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2011.
- A. P. Armen and I. Tsamardinos. Estimation and control of the false discovery rate of bayesian network skeleton identification. Technical Report TR-441, University of Crete, 2014.
- Y. Benjamini and D. Yekutieli. Investigating the importance of self-theories of intelligence and musicality for students’ academic and musical achievement. *Annals of Statistics*, 29 (4):1165–1188, 2001.
- H. Chong, M. Zey, and D. A. Bessler. On corporate structure, strategy, and performance: a study with directed acyclic graphs and pc algorithm. *Managerial and Decision Informatics*, 31:47–62, 2009.
- D. Colombo and M. H. Maathuis. Order-independent constraint-based causal structure learning. *J. Mach. Learn. Res.*, 15(1):3741–3782, January 2014. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2627435.2750365>.
- G. F. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI’99, pages 116–125, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9. URL <http://dl.acm.org/citation.cfm?id=2073796.2073810>.
- N. Friedman, M. Goldszmidt, and A. Wyner. Data analysis with bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI’99, pages 196–205, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc. ISBN 1-55860-614-9. URL <http://dl.acm.org/citation.cfm?id=2073796.2073819>.

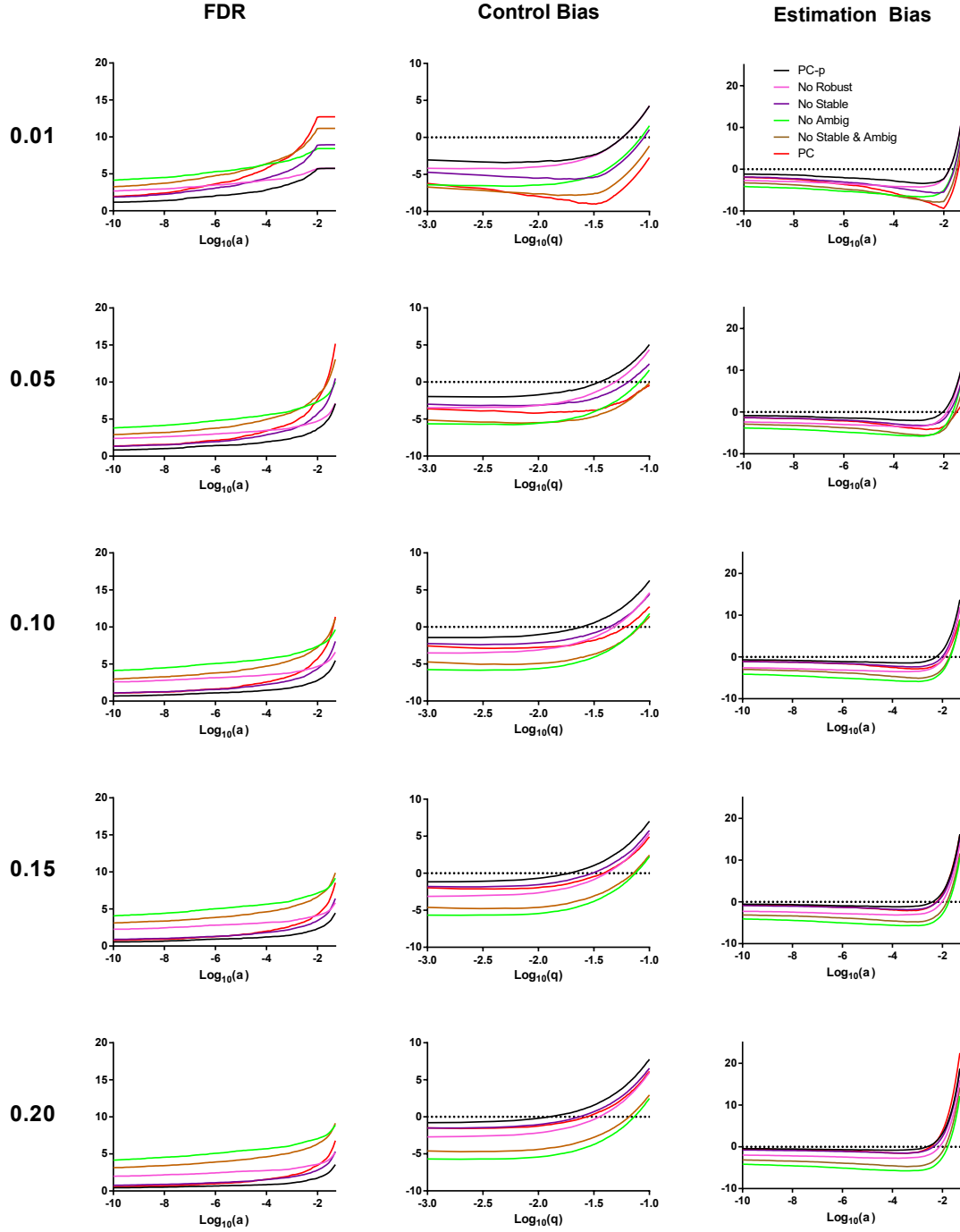


Figure 12: Mean FDR, control bias, and estimation bias values across multiple α thresholds for the low dimensional datasets.

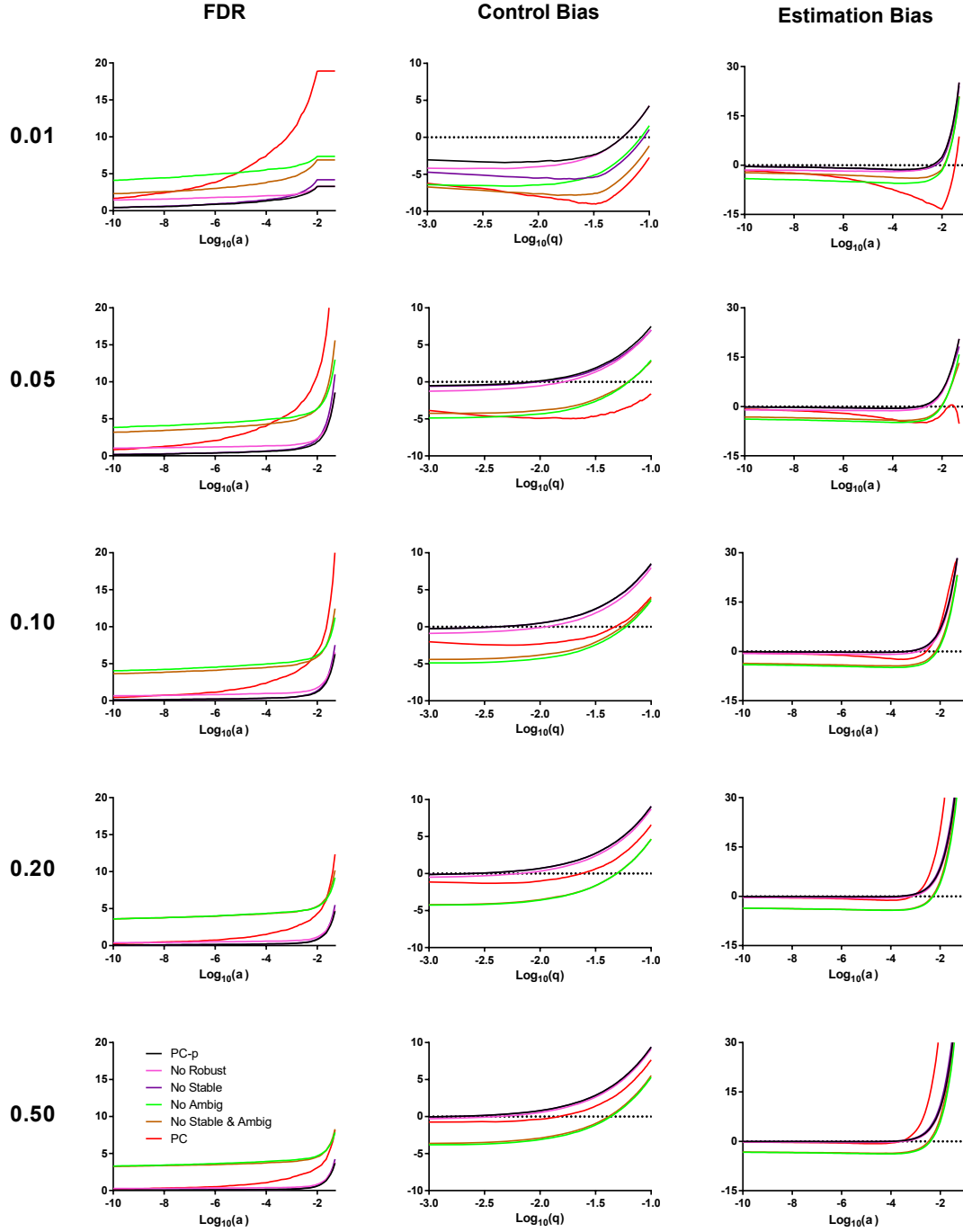


Figure 13: Same as Figure 12 except with high dimensional datasets.

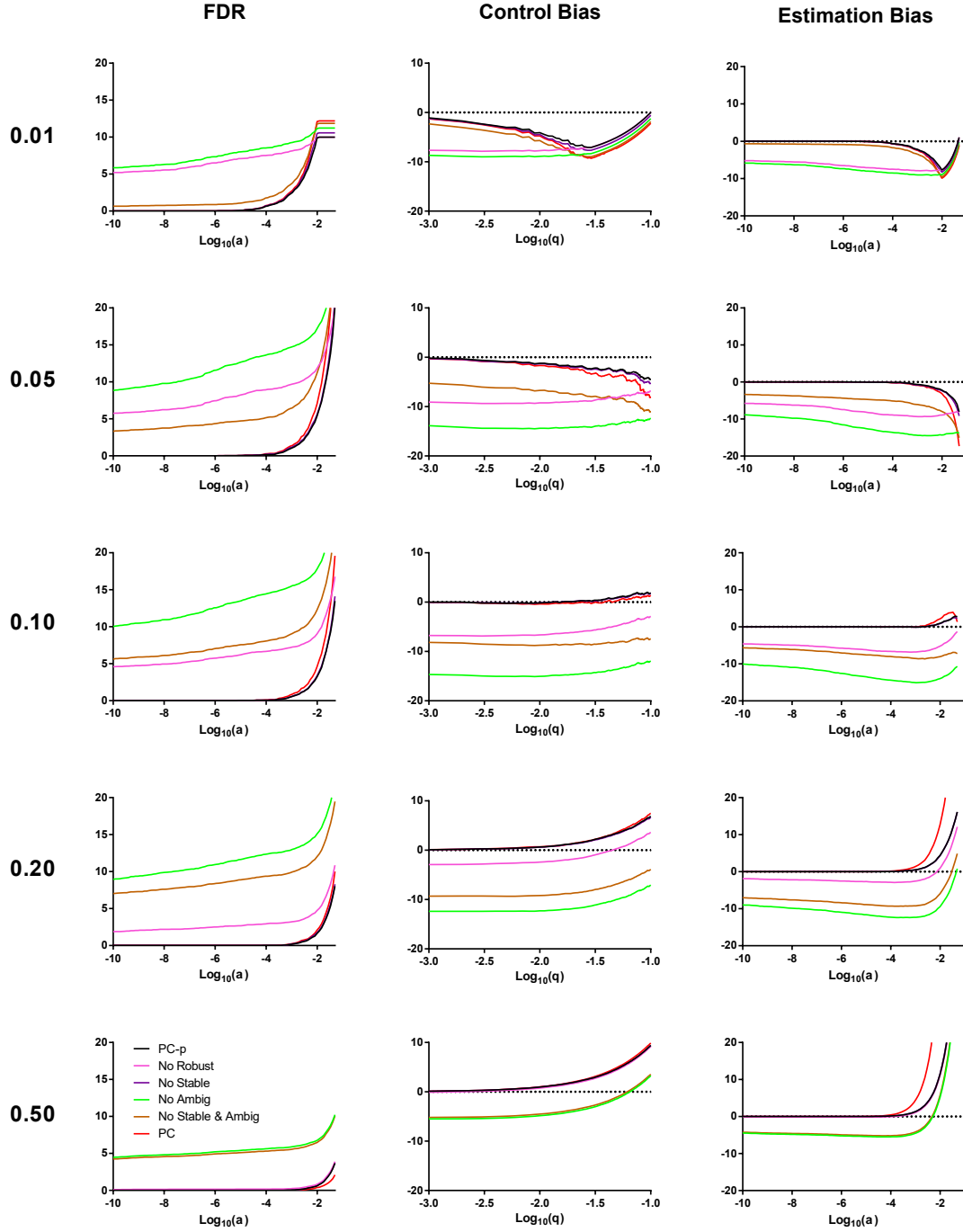


Figure 14: Same as Figure 12 except with bootstrapped CYTO datasets.

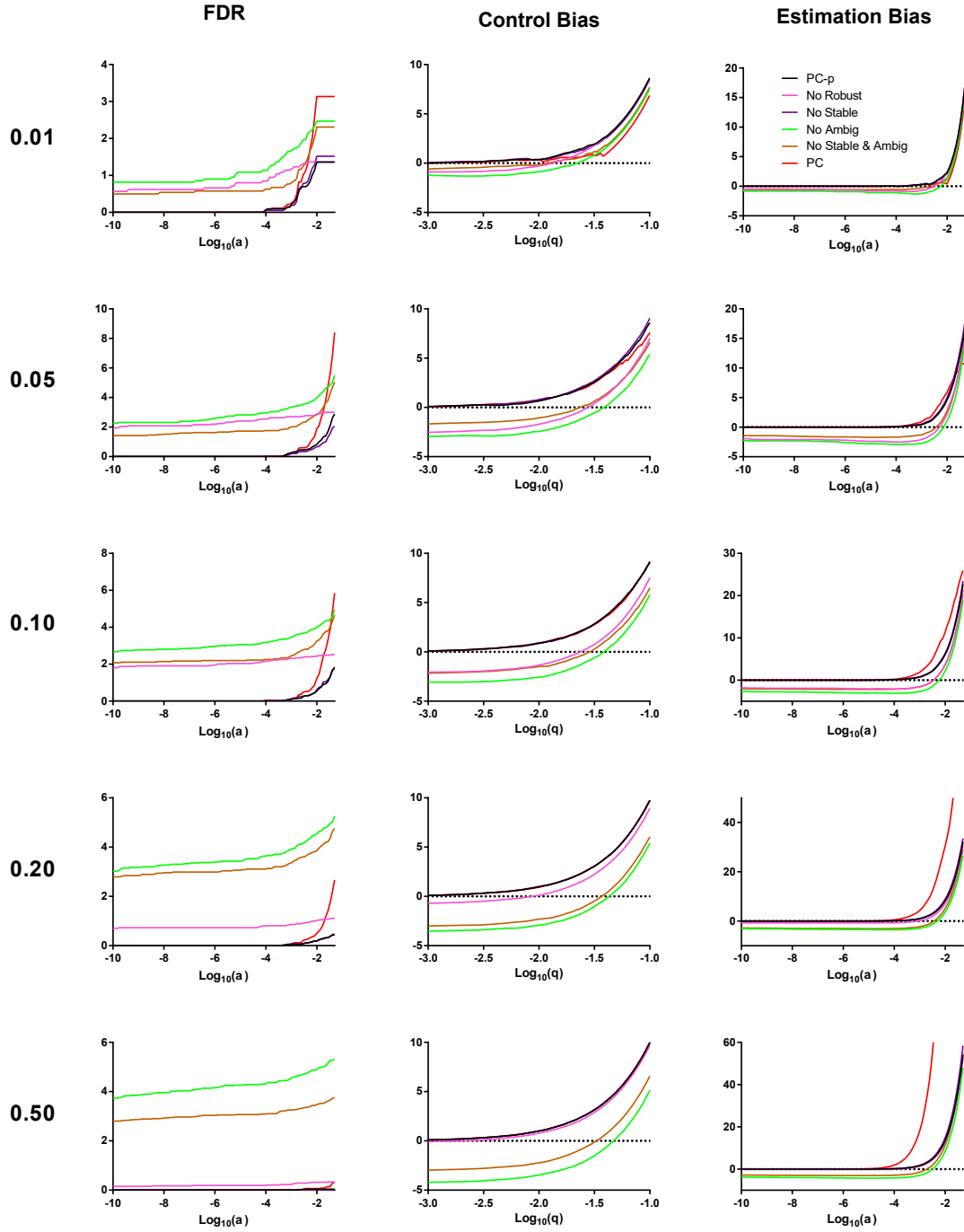


Figure 15: Same as Figure 12 except with bootstrapped GDP datasets.

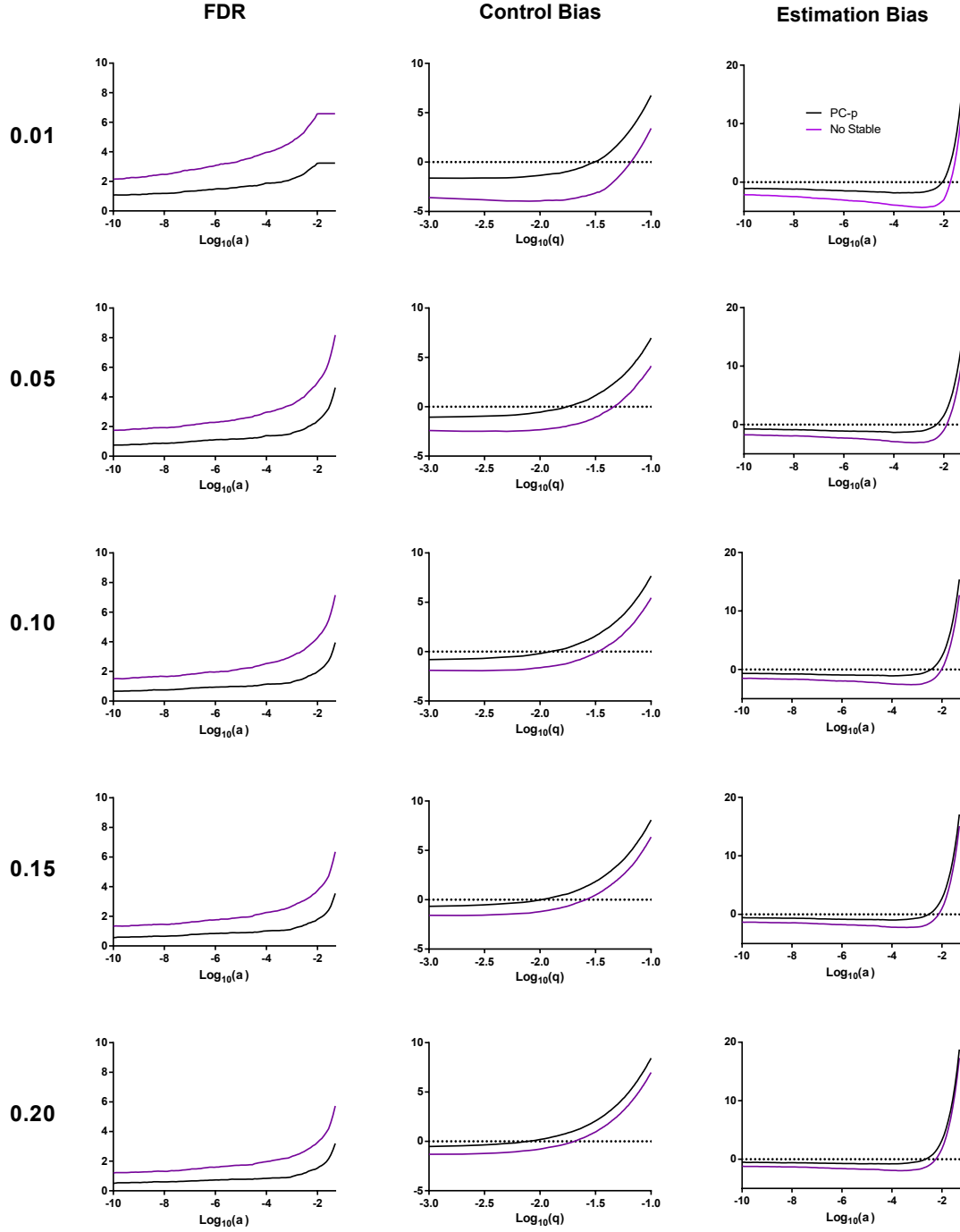


Figure 16: Mean FDR, control bias, and estimation bias values across multiple α thresholds for the low dimensional datasets.

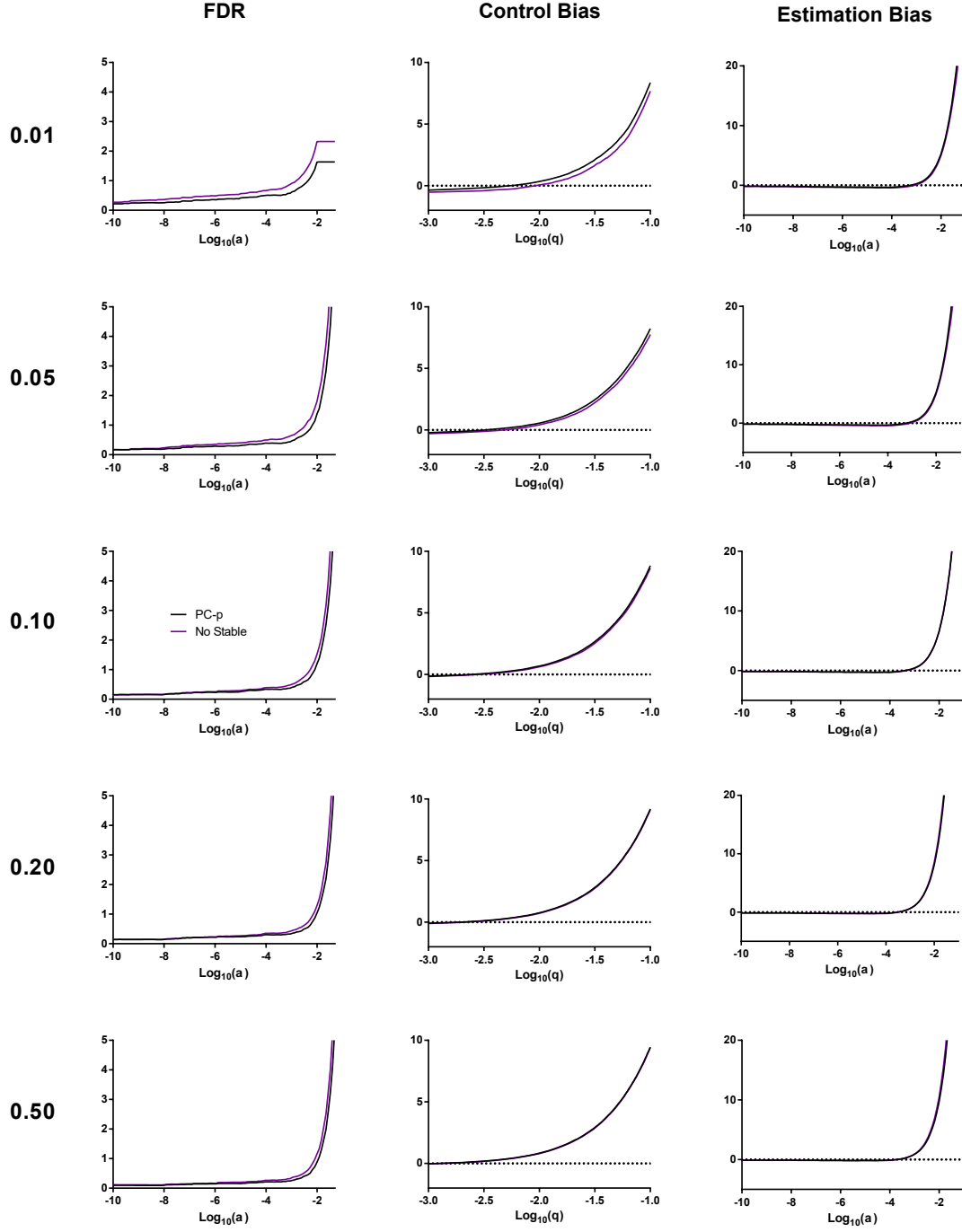


Figure 17: Same as Figure 16 except with high dimensional datasets.

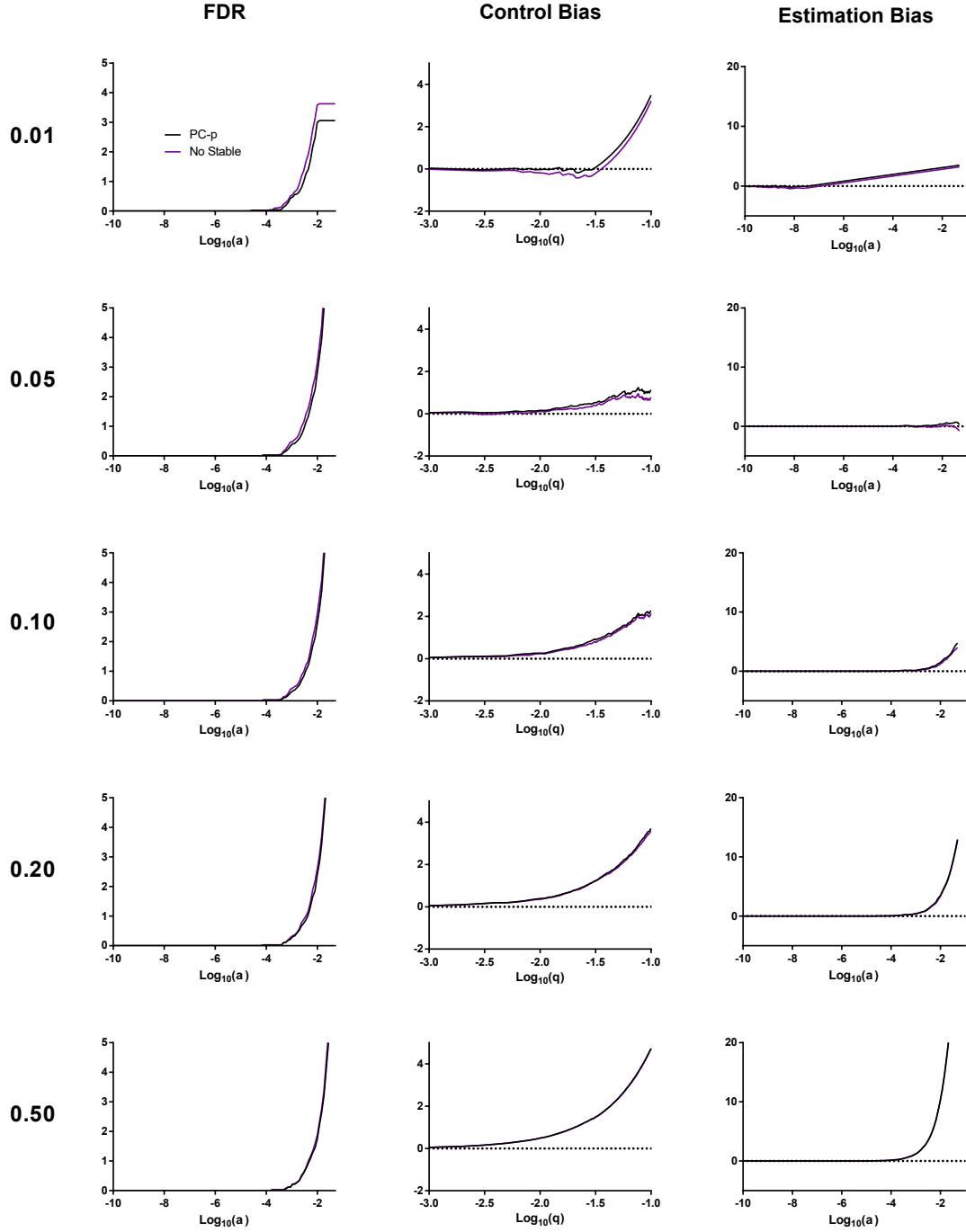


Figure 18: Same as Figure 16 except with bootstrapped CYTO datasets.

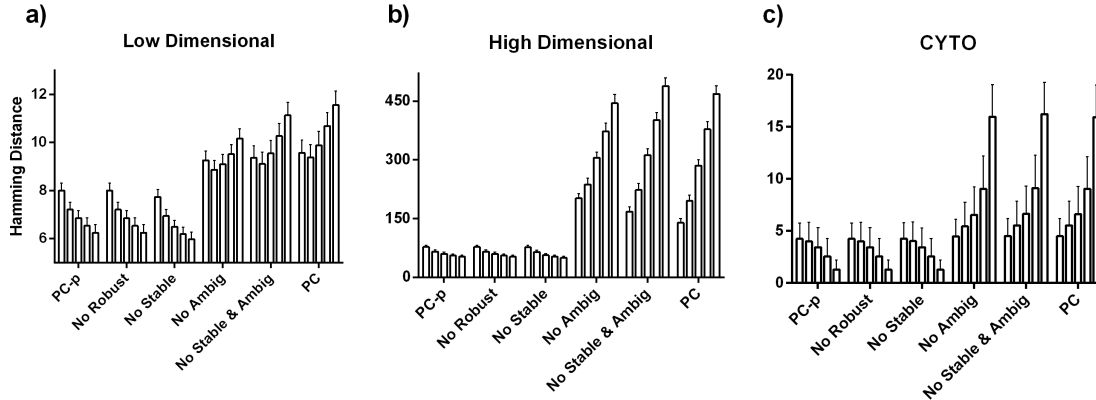


Figure 19: Structural Hamming distances for the a) low dimensional, b) high dimensional and c) CYTO datasets. The three sub-figures associate 5 bars with each algorithm; these bars correspond to α thresholds of 0.01, 0.05, 0.1, 0.20 and 0.50 for the low dimensional and CYTO datasets, and α thresholds of 0.01, 0.05, 0.1, 0.15 and 0.20 for the high dimensional datasets. Error bars represent standard errors for a) and standard deviations otherwise.

M. J. Ha, V. Baladandayuthapani, and K. A. Do. Prognostic gene signature identification using causal structure learning: applications in kidney cancer. *Cancer Inform*, 14(Suppl 1):23–35, 2015.

N. Harris and M. Drton. Pc algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, 14(1):3365–3383, January 2013. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2567709.2567770>.

Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, December 1988. ISSN 1464-3510. doi: 10.1093/biomet/75.4.800. URL <http://dx.doi.org/10.1093/biomet/75.4.800>.

S. P. Iyer, I. Shafran, D. Grayson, K. Gates, J. T. Nigg, and D. A. Fair. Inferring functional connectivity in MRI using Bayesian network structure learning with a modified PC algorithm. *Neuroimage*, 75:165–175, Jul 2013.

A. A. Joshi, S. H. Joshi, R. M. Leahy, D. W. Shattuck, I. Dinov, and A. W. Toga. *Bayesian approach for network modeling of brain structural features*, volume 7626. 2010. ISBN 9780819480279. doi: 10.1117/12.844548.

T. D. Le, L. Liu, A. Tsykin, G. J. Goodall, B. Liu, B. Y. Sun, and J. Li. Inferring microRNA-mRNA causal regulatory relationships from expression data. *Bioinformatics*, 29(6):765–771, Mar 2013.

- J. Li and Z. J. Wang. Controlling the false discovery rate of the association/causality structure learned with the pc algorithm. *J. Mach. Learn. Res.*, 10:475–514, June 2009. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=1577069.1577086>.
- J. Li, Z. Wang, and M. J. McKeown. Learning brain connectivity with the false-discovery-rate-controlled PC-algorithm. *Conf Proc IEEE Eng Med Biol Soc*, 2008:4617–4620, 2008.
- J. Listgarten and D. Heckerman. Determining the number of non-spurious arcs in a learned dag model: Investigation of a bayesian and a frequentist approach. In Ronald Parr and Linda C. van der Gaag, editors, *UAI*, pages 251–258. AUAI Press, 2007. ISBN 0-9749039-3-0. URL <http://dblp.uni-trier.de/db/conf/uai/uai2007.html#ListgartenH07>.
- C. Meek. Strong completeness and faithfulness in bayesian networks. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, pages 411–418, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9. URL <http://dl.acm.org/citation.cfm?id=2074158.2074205>.
- D. Mullensiefen, P. Harrison, F. Caprini, and A. Fancourt. Investigating the importance of self-theories of intelligence and musicality for students’ academic and musical achievement. *Front Psychol*, 6:1702, 2015.
- T. Richardson. A discovery algorithm for directed cyclic graphs. In *Proceedings of the Twelfth International Conference on Uncertainty in Artificial Intelligence*, UAI’96, pages 454–461, San Francisco, CA, USA, 1996. Morgan Kaufmann Publishers Inc. ISBN 1-55860-412-X. URL <http://dl.acm.org/citation.cfm?id=2074284.2074338>.
- K. Sachs, O. Perez, D. Pe’er, D. A. Lauffenburger, and G. P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, Apr 2005.
- P. Spirtes, C. Meek, and T. Richardson. Causal inference in the presence of latent variables and selection bias. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, UAI’95, pages 499–506, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc. ISBN 1-55860-385-9. URL <http://dl.acm.org/citation.cfm?id=2074158.2074215>.
- P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2nd edition, 2000.
- J. Sun, X. Hu, X. Huang, Y. Liu, K. Li, X. Li, J. Han, L. Guo, T. Liu, and J. Zhang. Inferring consistent functional interaction patterns from natural stimulus fMRI data. *Neuroimage*, 61(4):987–999, Jul 2012.
- R. Teramoto, C. Saito, and S. Funahashi. Estimating causal effects with a non-paranormal method for the design of efficient intervention experiments. *BMC Bioinformatics*, 15:228, 2014.
- I. Tsamardinos and L. E. Brown. Bounding the false discovery rate in local bayesian network learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*,

- AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, pages 1100–1105, 2008. URL <http://www.aaai.org/Library/AAAI/2008/aaai08-174.php>.
- X. Wu and Y. Ye. Exploring gene causal interactions using an enhanced constraint-based method. *Pattern Recogn.*, 39(12):2439–2449, December 2006. ISSN 0031-3203. doi: 10.1016/j.patcog.2006.05.003. URL <http://dx.doi.org/10.1016/j.patcog.2006.05.003>.