4th Workshop on Data Mining for Medical Informatics: Causal Inference for Health Data Analytics
Washington D.C., November 4th, 2017

An Assessment of the Calibration of Causal Relationships Learned Using RFCI and Bootstrapping

Mahdi Pakdaman Naeini[1], Fattaneh Jabbari[2], and Gregory F. Cooper[2,3]
[1]Department of Biomedical Informatics, Harvard Medical School, MA, USA
[2]Intelligent Systems Program, University of Pittsburgh, PA, USA
[3]Department of Biomedical Informatics, University of Pittsburgh, PA, USA

Abstract *The discovery of causal relationships from data is an important problem in many fields, including biomedical science. Algorithms have been developed that learn causal Bayesian networks from observational data, even when latent confounders are possible. Causal structure discovery from observational data is generally made under uncertainty. Thus, it is helpful for scientists to know how likely a structure or substructure is to be correct. Previous researchers have proposed using bootstrap probabilities in Bayesian structure learning; however, the accuracy of this approach for causal discovery with latent variables has not been studied, to our knowledge. This paper reports extensive simulation studies to evaluate the accuracy (calibration) of such bootstrap probabilities in causal discovery. Using a state-of-the-art causal discovery method (RFCI with typical parameter settings), our results show that bootstrap probabilities are usually well-calibrated for directed edge types in the discovered causal network, especially for datasets that contain thousands of variables, which is an increasingly common situation in biomedicine and other fields. Furthermore, the results indicate that the bootstrap RFCI (BRFCI) consistently yields a significantly higher precision than RFCI in discovering directed causal edges over a wide range of experimental setups, yet it obtains lower recall. While its recall is lower, BRFCI nonetheless outputs many correct causal relationships, which might serve as hypotheses that drive experimentation. This suggests that BRFCI may often be a more suitable method for guiding causal structure discovery in real applications compared to RFCI.*

1 Introduction

Discovering and modeling causal relationships is an important and challenging problem in many areas of science. Increasingly, scientists have available multiple complex data types and a large number of samples, each of which has a large number of measurements recorded, thanks to rapid advancements in sophisticated measurement technology. Some such data may result from experiments in which one or more variables were experimentally manipulated. Often, however, this data are purely observational. In the past 25 years, there has been significant progress in developing general computational methods for representing and discovering causal knowledge from observational data[1–5]. A primary use of such methods is to analyze observational scientific data to generate causal networks. Such networks can provide novel causal relationships to test experimentally; if such proposed relationships prove to be often correct, this approach may significantly increase the efficiency of causal discovery in science.

To make informed decisions about which proposed causal relationships to investigate experimentally, scientists need to know how likely the relationships are to be true. Such information is provided by having calibrated posterior probabilities of the relationships. To our knowledge, however, none of the existing causal discovery methods that model the possibility of latent confounders, such as RFCI[6], assign probabilities to the causal relationships that are output. Consequently, it would be difficult for a scientist to decide which of the output, proposed causal relationships, to test experimentally.

In this paper, we use a bootstrapping method[7] for generating probability estimates of edge types. This method resamples a dataset n times with replacement and learns a model for each dataset. In particular, for each dataset we used the RFCI algorithm[6] to derive a causal network representation known as a Partial Ancestral Graph (PAG), which can model latent common causes (confounders) [a] of measured variables, which are common in scientific data. For any given pair of nodes (A, B), the probability of a given edge type (e.g., $A \rightarrow B$) is estimated as the fraction of that edge type for (A, B) in the n PAGs.

Previously, bootstrapping has been used in the applications of evolutionary biology[8,9] to find the uncertainty of esti-

[a] A latent (hidden) confounder of the variables A and B is a hidden variable(s) that causes both A and B. In such a case, A and B will generally be statistically dependent, even if A does not cause B and B does not cause A.

mated clades of evolutionary trees and the intensity of edges in gene networks, respectively . In addition, researchers have successfully applied this approach for estimating the probabilities of edge types in Bayesian networks[10–12]. However, to our knowledge, the accuracy (calibration) of this approach has not been studied for causal discovery when latent confounders are being modeled from observational data, which is an important problem.

Probabilistic causal relationships are most useful when the probabilities assigned to them are accurate, that is, well-calibrated. Informally, we say that probabilities are well-calibrated if events predicted to occur with probability p do occur about p fraction of the time, for all p[13]. Obtaining well-calibrated probability estimates can be important in many practical causal network discovery problems. For example, an important area of cancer research is to discover gene alterations (G) that drive (i.e., cause) cancer (C). Algorithms exist for analyzing observational data to find gene alterations that appear to be cancer drivers[14], which is a hypothesis of the form $G \rightarrow C$. Such hypotheses generally require experimental validation in order to be accepted by the cancer biology community as drivers. If a cancer biologist knows that a given alteration has a 0.98 probability of being a driver, and if that probability is well-calibrated, it means that a wet-lab experiment is very likely to confirm the hypothesis, which is quite helpful to know. If the probability is not well-calibrated, it may be difficult for the biologist to decide whether to perform an experiment to evaluate the causal relationship.

In this paper, we report the results of an extensive set of experiments using simulated data [b] to investigate the following important question: *How calibrated are the bootstrap-derived probabilities of directed causal relationships when using typical algorithm parameter settings?* A directed causal relationship is an arc $A \rightarrow B$ in a PAG, which indicates that A is a cause of B. If the bootstrap probabilities are found to be well-calibrated, it supports that scientists can use them with confidence in deciding which experiments to perform or not perform. Such prioritization can in turn save them significant amounts of time and monetary costs. Although we have focused on scientific decision making as a use case for wanting calibrated probabilities, the possible applications are much broader, including for example robot decision making as well. Indeed, in any domain in which decision theory is being applied, it is important to have calibrated probabilities.

2 Method

In this section we briefly describe the RFCI causal network discovery algorithm that we applied in our experiments. We then introduce the bootstrap method that we used to compute bootstrap probabilities for edge types using RFCI.

2.1 Overview of RFCI

Colombo et al.[6] developed an algorithm called Really Fast Causal Inference (RFCI), which identifies the causal structure of the data-generating process in the presence of latent variables using PAGs as a representation. A PAG represents a Markov equivalence class of causal Bayesian network structures (possibly with latent variables) that have the same conditional independence relationships among the measured variables. RFCI is a two-phase, constraint-based algorithm that includes an adjacency search followed by an orientation phase. The first phase of RFCI starts with a fully connected undirected graph, $\mathcal{G}$. For each adjacency A — B in $\mathcal{G}$, if it finds a subset of nodes C that are adjacent to A and B and that make A and B independent conditioned on C (i.e., $A \perp\!\!\!\perp B|C$), it deletes the edge and stores C. The conditional independence tests at this step starts from empty conditioning sets and continues checking all adjacent nodes in $\mathcal{G}$ by increasing the size of the conditioning sets up to a specified set size. In the second phase, it applies the orientation rules to orient the endpoints and may perform additional independence tests. For more information about this method see Colombo et al.[6].

As is typical of constraint-based causal discovery algorithms, RFCI outputs a single causal graph structure (PAG) and does not provide any information about the uncertainty of the edges between the nodes in the structure. In the following section, we apply bootstrapping to obtain probabilities for causal edge types among the measured nodes.

[b] It is difficult to obtain a gold-standard set of causal relationships among the variables in large observational datasets. As a result, using simulated data is an important and commonly used approach for evaluating causal discovery methods.

2.2 Bootstrap RFCI (BRFCI)

For a pair of nodes (A, B), a PAG may contain one of the following seven different edge types, which we describe here assuming no selection bias: (1) $A \cdots B$: the no-edge type; (2) $A \rightarrow B$: a directed-edge from A to B means that A is a cause of B; (3) $B \rightarrow A$: this is similar to (2); (4) $A \circ\!\!\rightarrow B$: this edge type indicates either A is a cause of B, there is an unmeasured confounder of A and B, or both; (5) $B \circ\!\!\rightarrow A$: this is similar to (4); (6) $A \circ\!\!-\!\!\circ B$: this edge type expresses that A is a cause of B, B is a cause of A, there is an unmeasured confounder of A and B, or that there is an unmeasured confounder and one of those two causal relationships holds; (7) $A \leftrightarrow B$: the bi-directed edge between A and B represents the presence of an unmeasured confounder of A and B.

To obtain edge type probabilities for a pair (A, B), we used the bootstrapping method. The Bootstrap RFCI (BRFCI) method that we apply has three main steps. In the first step, it performs bootstrap sampling over the training data n times ($n = 400$ in our experiments) to create n different bootstrap datasets. In the second step, it runs RFCI on those n datasets to obtain n PAGs. Finally, for every pair of nodes, it uses the frequency counts of each edge type for that pair over the generated PAGs to determine a bootstrap probability distribution for the seven possible edge types. Although bootstrap probabilities need not be calibrated, this paper investigates the extent to which they are calibrated for directed edge types when using a state-of-the-art algorithm (RFCI) that applies typical parameter settings ($\alpha = 0.01$ and $\alpha = 0.001$) over a range of datasets generated from simulated data.

3 Experimental Method

This section describes the experimental methods that we used to evaluate the performance of BRFCI in obtaining bootstrap probabilities for predicted edge types. The evaluation involves the following steps:

1. Create a random causal Bayesian network, CBN, with V real-valued nodes and E edges. We set V to be 500 and 1000. We also set the average number of edges per node to be 1 and 2 (i.e., $E = V$ and $E = 2V$). To construct the CBN, we first ordered the nodes. Then, we randomly added edges in a forward direction until obtaining the specified mean graph density. This process generates a graph with a power-law-type distribution over the number of parents, with some nodes having many more than the average number of parents (Figure 1). In each CBN, the nodes correspond to continuous random variables where for every pair of nodes, (A, B), we parametrize a relation $A \rightarrow B$ in the CBN as a structural equation model (SEM): $A = \epsilon_A$ and $B = A\beta + \epsilon_B$, where ϵ_A and ϵ_B are zero-mean Gaussian noise terms and β is the linear coefficient. In our experiments, similar to Ramsey[15,16], variances of ϵ_A and ϵ_B are uniformly randomly chosen from the interval $[1.0, 3.0]$ and β is drawn uniformly randomly from the interval $[-1.5, 0.2] \cup [0.2, 1.5]$. This choice of parameter values for the simulations implies that, on average, around half of the variance of the variables is due to the error term, which makes structure learning more difficult[15,16].
2. Simulate a dataset, D, from the CBN. To evaluate the effect of sample size, C, we simulated datasets of size 2000 and 4000 instances (samples), subject to the structure and parameter constraints of the CBN from step 1.
3. Set a percentage of variables, H, to be unobserved (i.e., latent). These latent variables are randomly chosen from confounder variables (i.e., common causes) in a given data-generating CBN. We set H to be either 5% or 20%.
4. Generate 400 bootstrap datasets, $DB[1..400]$ from D.
5. For each of the bootstrap datasets, learn a PAG using the RFCI method; let $PAG[1..400]$ designate these PAGs. RFCI uses Fisher's Z test to check conditional independence of variables in the dataset. We set the significance level at which independence judgments were made to be $\alpha = 0.01$ and $\alpha = 0.001$. The smaller values of α result in the more stringent independence test in the RFCI method (i.e., less rejection of the independence null hypotheses) and a sparser PAG as a result.
6. For each observed node pair (A, B), calculate the bootstrap probability distribution $P_e(A, B)$ of the seven possible edge types of (A, B) using maximum likelihood estimates on the counts in $PAG[1..400]$.
7. Evaluate the performance of P_e in correctly predicting the directed edge type ($A \rightarrow B$ or $B \rightarrow A$) versus all other edge types ($A \cdots B$, $A \circ\!\!\rightarrow B$, $B \circ\!\!\rightarrow A$, $A \circ\!\!-\!\!\circ B$, $A \leftrightarrow B$) for every pair of nodes in the CBN. The edge type predictions were compared to the gold-standard PAG, which is obtained by running RFCI on the observed variables using d-separation properties that hold among those variables in the data-generating CBN.

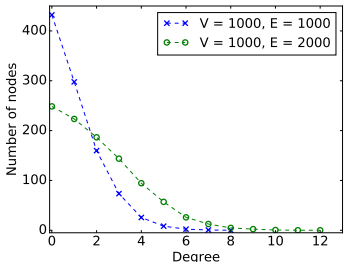

Figure 1: Parent size distribution for simulated CBNs with $V = 1000$ nodes and $E = 1000$ or 2000 edges.

Table 1: Summary Results. V and E indicate the number of nodes and edges. Avg_P, Avg_R, and Avg_NP indicate the mean precision, recall, and number of positive predictions, respectively. BRFCI-8 and BRFCI-9 represent BRFCI results using 0.8 and 0.9 threshold values.

V, E	Method	$\alpha = 0.01$			$\alpha = 0.001$		
		Avg precision	Avg recall	Avg NP	Avg precision	Avg recall	Avg NP
500, 500	BRFCI-9	0.94	0.13	9.9	0.88	0.33	24.3
	BRFCI-8	0.91	0.24	17.9	0.82	0.44	35.3
	RFCI	0.52	0.65	82.9	0.50	0.64	80.2
500, 1000	BRFCI-9	0.89	0.15	70.2	0.81	0.19	99.4
	BRFCI-8	0.83	0.22	107.8	0.74	0.24	135.7
	RFCI	0.59	0.37	262.9	0.60	0.34	241.3
1000, 1000	BRFCI-9	0.98	0.09	12.1	0.90	0.28	40.9
	BRFCI-8	0.95	0.17	23.5	0.85	0.38	60.2
	RFCI	0.54	0.66	155.0	0.49	0.64	167.6
1000, 2000	BRFCI-9	0.92	0.13	111.7	0.84	0.21	195.9
	BRFCI-8	0.88	0.21	183.0	0.77	0.27	269.2
	RFCI	0.59	0.42	558.6	0.59	0.39	518.6

8. Steps 1 through 7 above were repeated for 20 randomly generated CBNs and the performance results were averaged. For all simulations, we used Tetrad[c], which is an open-source, freely available software application that is coded in Java.

As described in step 7, we evaluated the performance of the bootstrap probabilities to measure how well they predict the directed edge type compared to all other (non-directed) edge types. A directed edge type is arguably the most important edge type because it is the one most likely to drive experimentation. In particular, directed edges that a scientist considers to be highly probable, as well as novel and important, are prime candidates for experimental validation. Furthermore, we are interested to see how well the method discovers highly probable (e.g., above 0.9) directed edges since a high probability region is the most critical one for making decisions about which directed arcs to investigate, such that false positive experimental investigations are minimized. Avoiding false positive experiments is important because typically they are expensive in terms of a scientist's time, financial cost, and other resources.

We used *precision* and *recall* as evaluation measures to compute the performance of the method on each of these two classes of edge types. For the directed edge type, for example, a pair of nodes (A, B) is considered to be classified as a directed edge type if the probability of having an arc $A \to B$ (or $B \to A$) is above a specified threshold. Such an edge is considered a true positive (TP) if the underlying true edge type is $A \to B$ (or $B \to A$). Otherwise, it is considered a false positive (FP). The precision is the ratio of TP to all instances predicted as directed with high probability (e.g., above 0.9). The recall is the ratio of TP to the total number of directed edges in the gold-standard PAG.

We also reported the **n**umber of **p**redicted directed edges with probability more than a certain threshold (the **NP** column in Tables 1 and 2). This number is helpful in determining whether the method identifies a sufficient number of candidate directed edges to plausibly support experimental investigation.

4 Experimental Results

This section presents the results of our experiments in evaluating the performance of the generated bootstrap probabilities for the directed edge type [d]. For each set of configurations (e.g., $H = 0.2$, $V = 1000$, $E = 2000$), we report the mean results over 20 randomly simulated CBNs, as well as the standard deviation of each mean, which is shown in parentheses following the reported means. Tables 2a and 2c report the results of our experiments when 2000 randomly generated instances are used for learning the CBN structure with the RFCI causal discovery method. Tables 2b and 2d report the results for 4000 randomly generated instances.

These results show that when the bootstrap probabilities indicate with high probability (e.g., greater than 0.9) that there is a directed edge between a pair of variables, then that prediction is often correct (i.e., it is typically correct more than 0.9 fraction of the time). Thus, the predictions are generally quite discriminative and well calibrated. Therefore, the predictions provide confidence in prioritizing the experimental investigation of such node pairs for direct causal relationships. As expected, the results show that by increasing the percentage of hidden nodes, we often have lower

[c] https://github.com/cmu-phil/tetrad

[d] The precision and recall for the non-directed edge type is close to 1.0 since most of them are no-edge; therefore, they are not reported.

Table 2: The average results of experiments performed using 20 randomly generated CBNs. V, E, C, and H indicate the number of nodes, edges, data samples, and the percentage of hidden variables, respectively. Tables 2a and 2b show results when $\alpha = 0.01$, where α is the confidence level that is used in the RFCI method for the independence tests (i.e., null hypothesis is independence); Tables 2c and 2d show results when $\alpha = 0.001$. BRFCI-8 and BRFCI-9 represent BRFCI results using 0.8 and 0.9 threshold values.

(a) α=0.01, $C = 2000$

H	V, E	Method	Precision	Recall	NP
0.05	500, 500	BRFCI-9	0.97(0.01)	0.16(0.01)	14.2(0.94)
		BRFCI-8	0.96(0.01)	0.28(0.01)	24.7(1.01)
		RFCI	0.59(0.02)	0.7(0.02)	103.9(3.94)
	500, 1000	BRFCI-9	0.92(0.01)	0.16(0.01)	74.2(2.53)
		BRFCI-8	0.86(0.01)	0.24(0.01)	118.6(3.4)
		RFCI	0.62(0.01)	0.42(0.01)	290.9(3.49)
	1K, 1K	BRFCI-9	0.99(0.01)	0.09(0.01)	16.1(2.08)
		BRFCI-8	0.96(0.01)	0.19(0.01)	32.1(2.95)
		RFCI	0.61(0.02)	0.71(0.01)	190.4(5.17)
	1K, 2K	BRFCI-9	0.93(0.01)	0.14(0.0)	120.8(2.8)
		BRFCI-8	0.89(0.01)	0.22(0.0)	200.8(4.33)
		RFCI	0.61(0.01)	0.46(0.01)	613.3(5.26)
0.2	500, 500	BRFCI-9	0.91(0.05)	0.06(0.01)	3.0(0.44)
		BRFCI-8	0.89(0.04)	0.14(0.02)	7.3(0.76)
		RFCI	0.42(0.02)	0.52(0.02)	59.6(2.65)
	500, 1000	BRFCI-9	0.86(0.02)	0.09(0.0)	43.2(1.97)
		BRFCI-8	0.8(0.01)	0.14(0.0)	70.6(1.91)
		RFCI	0.55(0.01)	0.28(0.01)	205.9(2.51)
	1K, 1K	BRFCI-9	0.97(0.02)	0.06(0.01)	5.3(0.64)
		BRFCI-8	0.95(0.02)	0.11(0.01)	11.1(0.99)
		RFCI	0.46(0.02)	0.55(0.02)	110.1(4.17)
	1K, 2K	BRFCI-9	0.93(0.01)	0.08(0.0)	67.5(2.22)
		BRFCI-8	0.9(0.01)	0.14(0.0)	119.7(3.41)
		RFCI	0.58(0.01)	0.33(0.0)	423.7(6.57)

(b) $\alpha = 0.01$, $C = 4000$

H	V, E	Method	Precision	Recall	NP
0.05	500, 500	BRFCI-9	0.98(0.01)	0.18(0.01)	15.7(1.0)
		BRFCI-8	0.96(0.01)	0.31(0.01)	27.7(1.3)
		RFCI	0.65(0.02)	0.77(0.01)	100.9(2.5)
	500, 1000	BRFCI-9	0.91(0.01)	0.21(0.0)	103.9(2.43)
		BRFCI-8	0.85(0.01)	0.29(0.01)	152.3(3.63)
		RFCI	0.62(0.01)	0.47(0.01)	330.7(4.94)
	1K, 1K	BRFCI-9	0.99(0.01)	0.12(0.01)	20.3(1.15)
		BRFCI-8	0.98(0.01)	0.22(0.01)	37.3(1.9)
		RFCI	0.67(0.01)	0.76(0.01)	193.0(5.6)
	1K, 2K	BRFCI-9	0.93(0.01)	0.19(0.0)	164.3(4.24)
		BRFCI-8	0.89(0.01)	0.28(0.0)	254.2(4.35)
		RFCI	0.6(0.01)	0.52(0.01)	698.1(7.9)
0.2	500, 500	BRFCI-9	0.89(0.05)	0.13(0.02)	6.6(0.78)
		BRFCI-8	0.84(0.04)	0.22(0.02)	12.0(1.22)
		RFCI	0.41(0.02)	0.61(0.02)	67.3(2.71)
	500, 1000	BRFCI-9	0.87(0.01)	0.13(0.01)	59.4(2.58)
		BRFCI-8	0.81(0.01)	0.19(0.01)	89.5(2.81)
		RFCI	0.56(0.01)	0.32(0.01)	223.9(3.48)
	1K, 1K	BRFCI-9	0.96(0.02)	0.07(0.01)	6.5(0.77)
		BRFCI-8	0.89(0.03)	0.14(0.01)	13.4(0.81)
		RFCI	0.43(0.01)	0.61(0.01)	126.6(4.42)
	1K, 2K	BRFCI-9	0.9(0.01)	0.12(0.0)	94.3(2.71)
		BRFCI-8	0.85(0.01)	0.18(0.0)	157.3(4.02)
		RFCI	0.55(0.01)	0.37(0.01)	499.5(5.89)

(c) $\alpha = 0.001$, $C = 2000$

H	V, E	Method	Precision	Recall	NP
0.05	500, 500	BRFCI-9	0.9(0.01)	0.35(0.02)	31.6(1.66)
		BRFCI-8	0.82(0.02)	0.46(0.02)	45.7(1.69)
		RFCI	0.55(0.02)	0.64(0.02)	96.1(3.09)
	500, 1000	BRFCI-9	0.84(0.01)	0.2(0.01)	103.5(3.48)
		BRFCI-8	0.78(0.01)	0.26(0.01)	143.1(3.44)
		RFCI	0.61(0.01)	0.38(0.01)	264.1(3.96)
	1K, 1K	BRFCI-9	0.96(0.01)	0.31(0.01)	51.4(2.12)
		BRFCI-8	0.89(0.01)	0.42(0.01)	75.7(3.05)
		RFCI	0.57(0.01)	0.67(0.01)	188.7(5.73)
	1K, 2K	BRFCI-9	0.88(0.01)	0.21(0.0)	194.9(5.55)
		BRFCI-8	0.81(0.01)	0.27(0.01)	271.8(6.42)
		RFCI	0.63(0.01)	0.41(0.0)	537.6(10.03)
0.2	500, 500	BRFCI-9	0.87(0.04)	0.21(0.02)	11.1(0.84)
		BRFCI-8	0.82(0.03)	0.31(0.02)	17.4(1.13)
		RFCI	0.43(0.02)	0.53(0.02)	57.4(2.05)
	500, 1000	BRFCI-9	0.83(0.01)	0.12(0.01)	60.9(2.68)
		BRFCI-8	0.76(0.01)	0.17(0.01)	91.2(3.03)
		RFCI	0.6(0.01)	0.27(0.01)	184.9(3.46)
	1K, 1K	BRFCI-9	0.82(0.05)	0.16(0.01)	17.8(1.4)
		BRFCI-8	0.76(0.04)	0.25(0.02)	30.1(1.78)
		RFCI	0.39(0.01)	0.5(0.02)	119.6(3.82)
	1K, 2K	BRFCI-9	0.82(0.03)	0.14(0.0)	118.4(6.41)
		BRFCI-8	0.76(0.03)	0.19(0.01)	173.9(9.31)
		RFCI	0.55(0.03)	0.31(0.01)	392.9(18.9)

(d) $\alpha = 0.001$, $C = 4000$

H	V, E	Method	Precision	Recall	NP
0.05	500, 500	BRFCI-9	0.93(0.01)	0.45(0.02)	38.1(1.62)
		BRFCI-8	0.86(0.01)	0.58(0.02)	53.3(1.82)
		RFCI	0.6(0.02)	0.76(0.01)	100.9(2.45)
	500, 1000	BRFCI-9	0.79(0.02)	0.26(0.01)	141.1(3.66)
		BRFCI-8	0.73(0.01)	0.32(0.01)	184.6(3.97)
		RFCI	0.61(0.01)	0.42(0.01)	298.7(5.17)
	1K, 1K	BRFCI-9	0.95(0.01)	0.39(0.01)	69.2(2.19)
		BRFCI-8	0.91(0.01)	0.52(0.01)	96.7(2.97)
		RFCI	0.61(0.01)	0.77(0.01)	215.7(5.51)
	1K, 2K	BRFCI-9	0.83(0.01)	0.3(0.01)	297.9(4.8)
		BRFCI-8	0.77(0.01)	0.36(0.01)	388.4(5.67)
		RFCI	0.6(0.01)	0.48(0.01)	664.3(5.18)
0.2	500, 500	BRFCI-9	0.83(0.03)	0.3(0.02)	16.3(1.1)
		BRFCI-8	0.78(0.02)	0.42(0.02)	25.0(1.52)
		RFCI	0.42(0.02)	0.61(0.02)	66.3(3.18)
	500, 1000	BRFCI-9	0.77(0.01)	0.17(0.01)	92.0(3.34)
		BRFCI-8	0.7(0.01)	0.22(0.01)	123.8(4.21)
		RFCI	0.56(0.01)	0.3(0.01)	217.4(4.53)
	1K, 1K	BRFCI-9	0.88(0.02)	0.24(0.01)	25.2(1.26)
		BRFCI-8	0.82(0.02)	0.34(0.01)	38.3(1.68)
		RFCI	0.4(0.01)	0.63(0.01)	146.3(3.73)
	1K, 2K	BRFCI-9	0.81(0.01)	0.18(0.0)	172.5(3.84)
		BRFCI-8	0.74(0.01)	0.24(0.0)	242.8(3.63)
		RFCI	0.57(0.01)	0.36(0.0)	479.6(5.66)

precision. Also, recall and the number of positive predictions (NP) decreases when increasing the percentage of hidden nodes in the CBN. Increasing the number of cases that are used in the RFCI method led to better recall and usually improved the precision.

Table 1 shows the summary results for precision, recall, and the number of positive predictions (NP) for $\alpha = 0.01$ and $\alpha = 0.001$ averaged over all other settings of CBNs in our experiments. The results show that the BRFCI usually generates accurate and well-calibrated estimates for directed edge type probabilities; however, the predicted probabilities for $\alpha = 0.01$ tends to be more accurate compared to $\alpha = 0.001$. The results show that in all the settings there are often enough positive predictions made by the BRFCI to plausibly support meaningful scientific investigation.

These results are particularly true for datasets with larger number of nodes, which correspond to the types of big data that are often now available for data mining in science, including biomedical science.

5 Conclusions

In causal discovery, scientists generally want to make informed decisions about which novel causal hypotheses they prioritize investigating experimentally. Thus, it is useful for a causal discovery method to output an estimate of the probability of causal relationships. To our knowledge, of the constraint-based causal discovery methods that model latent variables, none provide posterior probabilities of causal relationships. In this paper, we introduced a simple approach to address this problem by using bootstrapping with the RFCI method to estimate the probabilities of the causal relationships between each pair of measured variables in the output model. We performed an extensive set of experiments on moderately large-scale simulation data to evaluate the calibration of the bootstrap probabilities. The results of these experiments provide support that bootstrap probabilities are generally accurate for directed types. These results provide support that bootstrap probabilities can help prioritize which experimental investigations to perform. Similarly, the results provide support for not investigating edges for which the assigned bootstrap probability of the non-directed type is high. This information could potentially save considerable time and expense in real applications in science and other domains that involve causal discovery and modeling. In the future, we would like to evaluate the bootstrap probabilities on real observational scientific data for which we know some of the directed edge types and can evaluate calibration using them. In addition, it would be useful to investigate a wider variety of simulated data, including data generated from denser CBNs.

Acknowledgements 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 (BD2K) initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Clark N Glymour and Gregory F Cooper. *Computation, Causation, and Discovery*. MIT Press, 1999.

[2] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

[3] Judea Pearl. *Economet. Theory*, 19:675–685, 2003.

[4] Peter Spirtes. Introduction to causal inference. *The Journal of Machine Learning Research*, 11:1643–1662, 2010.

[5] Phyllis Illari, Federica Russo, and Jon Williamson. *Causality in the Sciences*. OUP Oxford, 2011.

[6] Diego Colombo, Marloes H Maathuis, Markus Kalisch, Thomas S Richardson, et al. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321, 2012.

[7] Bradley Efron and Robert J Tibshirani. *An Introduction to the Bootstrap*. CRC press, 1994.

[8] Joseph Felsenstein. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, pages 783–791, 1985.

[9] Seiya Imoto, Sun Yong Kim, Hidetoshi Shimodaira, Sachiyo Aburatani, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Bootstrap analysis of gene networks based on Bayesian networks and nonparametric regression. *Genome Informatics*, 2002.

[10] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with Bayesian networks: A bootstrap approach. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, pages 196–205, 1999.

[11] Nir Friedman, Moises Goldszmidt, and Abraham J Wyner. On the application of the bootstrap for computing confidence measures on features of induced Bayesian networks. In *AISTATS*, 1999.

[12] Anna Roumpelaki, Giorgos Borboudakis, Sofia Triantafillou, and Ioannis Tsamardinos. Marginal causal consistency in constraint-based causal learning. In *CFA@ UAI*, pages 39–47, 2016.

[13] Morris H DeGroot and Stephen E Fienberg. The comparison and evaluation of forecasters. *The Statistician*, 1983.

[14] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6), 2010.

[15] Joseph D Ramsey. Scaling up greedy equivalence search for continuous variables. *arXiv preprint arXiv:1507.07749*, 2015.

[16] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter Spirtes. Learning the structure of linear latent variable models. *The Journal of Machine Learning Research*, 7:191–246, 2006.