A Simulation Study of Three Related Causal Data Mining Algorithms

Subramani Mani
(mani@cbmi.upmc.edu)
Center for Biomedical Informatics and
Intelligent Systems Program
University of Pittsburgh PA 15213

Gregory F. Cooper
(gfc@cbmi.upmc.edu)
Center for Biomedical Informatics and
Intelligent Systems Program
University of Pittsburgh PA 15213

Abstract

In all scientific domains causality plays a significant role. This study focused on evaluating and refining efficient algorithms to learn causal relationships from observational data. Evaluation of learned causal output is difficult, due to lack of a gold standard in real-world domains. Therefore, we used simulated data from a known causal network in a medical domain—the Alarm network. For causal discovery we used *three* variants of the Local Causal Discovery (LCD) algorithms, that are referred to as LCDa, LCDb and LCDc. These algorithms use the framework of causal Bayesian Networks to represent causal relationships among model variables. LCDa, LCDb and LCDc take as input a dataset and a partial node ordering, and output purported causes of the form *variable Y causally influences variable Z*. Using the simulated Alarm dataset as input, LCDa had a false positive rate of 0.09, LCDb 0.08, and LCDc 0.04. All the algorithms had a true positive rate of about 0.27. Most of the false positives occurred when a causal relationship was confounded. LCDc output as causal only those causally confounded pairs that had very weak confounding. We identify and discuss the causally confounded relationships that often seem to induce false positive results.

1 INTRODUCTION

Seeking causes for various phenomena is a significant part of human endeavor. Causal knowledge aids planning and decision making in almost all fields. For example, in the domain of medicine, determining the cause of a disease helps in prevention and treatment. Well designed experimental studies, such as randomized control trials, are typically employed in assessing causal relationships. Here the value of the variable postulated to be *causal* is set randomly and its effects measured. These studies are appropriate in certain situations, for example, animal studies and studies involving human subjects that have undergone a thorough procedural and ethical review. Experimental studies may not, however, be feasible in many contexts due to ethical, logistical, or cost considerations. These practical limitations of experimental studies heighten the importance of exploring, evaluating and refining techniques to learn more about causal relationships from observational data, for example, data routinely collected in astronomy, earth sciences or healthcare. The goal is not to replace experimental studies, which are extremely valuable in science, but rather to augment and guide experimental studies when feasible.

This paper introduces three algorithms called LCDa, LCDb and LCDc that are designed for efficient discovery of possible causal relationships from large observational databases. In this study we apply them to a simulated dataset obtained from a known causal network (Alarm) and evaluate the output. We have previously applied LCD (Cooper, 1997) to a population-based infant birth and death dataset of 41,000 instances and 87 attributes. We obtained nine relationships out of which eight seemed plausibly causal (Mani & Cooper, 1999). The present simulation study is a prelude to further work using large real-world medical datasets. By improving performance from insights gained through simulation experiments, such as the one reported here, we expect later to do better on real-world datasets.

2 METHODS

2.1 Assumptions for Causal Discovery

In the research reported here, we use causal Bayesian networks to represent causal relationships among

model variables. This section provides a brief introduction to causal Bayesian networks, as well as a description of the assumptions we used to apply these networks for causal discovery.

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network (Pearl, 1991). Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure, which contains five nodes. Due to limited space, the states of the nodes and the probabilities that are associated with this structure are not shown.

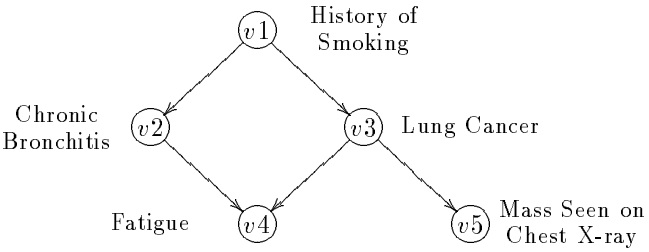

Figure 1: A hypothetical causal Bayesian network structure

The causal network structure in Figure 1 indicates, for example, that a *History of Smoking* can causally influence whether *Lung Cancer* is present, which in turn can causally influence whether a patient experiences *Fatigue* or presents with a *Mass Seen on Chest X-ray*.

The **causal Markov condition** gives the independence relationships[1] that are specified by a causal Bayesian network:

> A variable is independent of its non-descendants (i.e., non-effects) given just its parents (i.e., its direct causes).

According to the Markov condition, the causal network in Figure 1 is representing that the chance of a *Mass Seen on Chest X-ray* will be independent of a *History of Smoking*, given that we know whether *Lung Cancer* is present or not. While the causal Markov condition specifies independence relationships among variables, the **causal faithfulness condition** specifies *dependence* relationships:

> Variables are independent only if their independence is implied by the causal Markov condition.

[1] We use the terms *independence* and *dependence* in this section in the standard probabilistic sense.

For the causal network structure in Figure 1, three examples of the causal faithfulness condition are (1) *History of Smoking* and *Lung Cancer* are probabilistically dependent, (2) *History of Smoking* and *Mass Seen on Chest X-ray* are dependent, and (3) *Mass Seen on Chest X-ray* and *Fatigue* are dependent. The intuition behind that last example is as follows: a *Mass Seen on Chest X-ray* increases the chance of *Lung Cancer* which in turn increases the chance of *Fatigue*; thus, the variables *Mass Seen on Chest X-ray* and *Fatigue* are expected to be probabilistically dependent. In other words, the two variables are dependent because of a common cause (i.e., a confounder).

The causal Markov and faithfulness conditions describe *probabilistic* independence and dependence relationships, respectively, that are represented by a causal Bayesian network. In causal discovery, we do not know the probabilistic relationships among variables precisely, because we only have a finite amount of data. Thus, we make the following **statistical testing assumption**:

> A statistical test performed to determine independence (or alternatively dependence) given a finite dataset will be correct relative to independence (dependence) in the joint probability distribution that is defined by the causal process under study.

That is, we assume our statistical test gives valid independence and dependence results about the generating causal process. We are empirically investigating the dependence/independence hypotheses in context using varying sample sizes. In general, the greater the number of records in a dataset, the more likely it is that the statistical testing assumption will hold. But at very large sample sizes spurious correlations can also emerge eroding the validity of statistical tests. The reader is referred to chapters 8–11 of the book (Glymour & Cooper, 1999) for a detailed discussion of this and other related issues. Since a simulated dataset was used, the sample size could be varied easily in our experiments.

2.2 An Algorithm for Causal Discovery

In this section, we introduce the LCD algorithm on which we based several variant algorithms. LCD assumes the following:

> **Assumption 1:** The causal Markov condition
> **Assumption 2:** The causal faithfulness condition
> **Assumption 3:** The statistical testing assumption

In addition, LCD makes the following assumption:

Assumption 4: Given measured variables X, Y, and Z, if Y causes Z, and Y and Z are not confounded, then one of the causal networks in Figure 2 must hold.

Assumption 4 implicitly states that X is not causally influenced by Y or by Z. As we discuss in later sections, in our experiments we chose X so that this assumption is tenable.

Before introducing the LCD algorithm in more detail, we define some terms. Let Independent$_T(A,\ B)$ denote that A and B are independent according to test T applied to our dataset. Let Independent$_T(A,\ B$ given $C)$ denote that A and B are independent given C, according to T. Finally, let Dependent$_T(A,\ B)$ denote that A and B are dependent according to T[2]. These independence and dependence tests are labeled as given below for easy reference.

Test$_1$. Dependent$_T(X,\ Y)$
Test$_2$. Dependent$_T(Y,\ Z)$
Test$_3$. Dependent$_T(X,\ Z)$
Test$_4$. Independent$_T(X,\ Z$ given $Y)$

If all these four tests are satisfied then LCD outputs that Y causally influences Z. The first network in Figure 2 violates Test$_1$, and thus, LCD is unable to detect that Y causally influences Z in such situations. Under Assumptions 1 through 3, the other three networks in Figure 2 satisfy Test$_1$ through Test$_4$. In (Cooper, 1997), it is shown that if Y and Z are confounded, then one or more of the four tests will be violated. As an example, Figure 3 shows an important case in which Y and Z are confounded by a hidden variable H. For this causal network, it follows from Assumptions 1 and 2 that X and Z will be dependent given Y, and thus, Test$_4$ will fail.

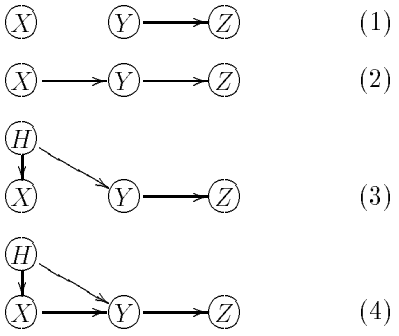

Figure 2: Causal models in which Y causes Z.

[2]Although the three tests in this paragraph should technically be distinguished from each other by using separate labels, such as *T1*, *T2*, and *T3*, for simplicity of notation we use a single label T.

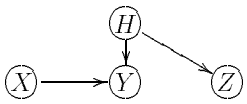

Figure 3: Causal model in which X causes Y, and Y and Z are dependent due to confounding by a hidden variable(s) represented by H.

To summarize, under Assumptions 1 through 4, when Y causally influences Z and these two variables are unconfounded, the four tests hold (unless X and Y are independent). Conversely, when Y and Z are confounded (or when X and Y are independent), one or more of the four tests will fail. From these propositions, we can conclude that if the four tests hold, then one of the three causal networks (2,3,4) in Figure 2 must hold, and thus, we can determine that Y causes Z and the two variables are unconfounded.

2.3 LCD variants—LCDa, LCDb and LCDc

The motivation for considering variants of LCD came from the observation that in all the false positive "causal" output from LCD based on Alarm data, the independence test (Test$_4$) was returned as positive when it should have failed. LCD output 45 true *causal* relationships, 21 *causally-confounded* and 21 confounded by a *common ancestor* (Section 2.5.1 contains a detailed description of these categories). The dependence tests (Test$_1$ through Test$_3$) did not fail. This led us to explore more stringent tests of independence. For example, performing an increased number of independence tests for the same YZ pair using different X nodes might improve independence testing resulting in a more accurate assessment of the causal influence of Y on Z. This was our working hypothesis in the design of LCD variants LCDa, LCDb, and LCDc.

We now describe these variants in greater detail. The LCDa, LCDb and LCDc algorithms apply Test$_1$ through Test$_4$ in exploring a database for possible causal relationships. These variants make an additional assumption apart from the *four* given earlier.

Assumption 5: The variables in the dataset are assigned a partial ordering.

The algorithms are given as input an ordered set of variables V. Let the total number of observed variables be n. A partial ordering will give k partitions of the variables, where k is less than n. If the variables are completely ordered, the number of partitions k will be equal to n. A partial ordering is a sufficient input to the algorithm. Even though a partial ordering with one partition containing all the variables

will satisfy this assumption, for our causal discovery framework to output purported causal relationships we should have at least three non-empty partitions. We obtained a partial ordering of the variables of the Alarm network by performing a modified topological sort on the Alarm network. In a real-world dataset, one could do a temporal ordering or ask an expert to provide an ordering of the variables. Table 1 gives the

Table 1: Partitioning (ordering) of the variables of the hypothetical network with five nodes

PARTITION	VARIABLES
p_0	$v1$ (HISTORY OF SMOKING)
p_1	$v2$,$v3$ (CHRONIC BRONCHITIS, LUNG CANCER)
p_2	$v4$,$v5$ (FATIGUE, MASS SEEN ON CHEST X-RAY)

partial ordering of the variables of the network given in Figure 1. In general, the first partition (p_0) contains the root node(s) and the last one (p_k) the leaf nodes (nodes having no descendents). The algorithm evaluates each triplet XYZ satisfying the following two constraints.

1. The nodes of the triplet belong to different partitions.
2. Let there be k partitions $p_0 \ldots p_k$. If X is in p_i, Y is in p_j, and Z is in p_k, then $i < j < k$.

The algorithms LCDa, LCDb and LCDc perform Test_1 through Test_3 for all such triplets XYZ in the database. Test_1 through Test_3 output many triplets XYZ such that for the same pair YZ there often is more than one X. In such situations Test_4 could be taken as positive if it is satisfied for any *one* such triplet (LCDa), satisfied by any *two* such triplets (LCDb) or satisfied by *all* such triplets (LCDc). All three of these algorithms were used in the causal discovery study reported here. Note that if there is only one triplet for a pair YZ, LCDa, LCDb and LCDc perform Test_4 on just that one triplet XYZ. Simple variations of the Independence and Dependence tests described in (Cooper, 1997) were used. Both tests have $O(m)$ time complexity, where m is the number of records (cases) in the database. If all four tests are passed, LCDa, LCDb and LCDc output that Y causally influences Z and the two variables are unconfounded (under Assumptions 1–5), and the probability distribution of Z given Y is displayed.

2.4 Related work

Traditional statistical approaches using χ^2 tests or logistic regression can establish dependence between variables. Likewise, machine learning algorithms such as decision tree learners (e.g., C4.5 and CART), rule inducers (e.g., C4.5Rules and FOCL) and neural networks can build useful domain models from data and capture the inter-dependence among the variables. But none of these techniques is intended to establish causal relationships of the form Y causally influences Z.

Structural equation models (SEMs) (Bollen, 1989), represent causal relationships, thus going beyond correlation and dependence. The emphasis in SEM research is on hypothesis testing of manually specified models, rather than on automated search over the space of models. Typically the SEM assumes linear relationships (with statistical noise) among the model variables than modeling with discrete variables, although more recently non-linear relationships are modeled.

A review of the philosophical literature on causality is beyond the scope of this paper. For a detailed discussion of the relationship between statistical association and causation, including philosophical issues, see for example (Fetzer, 1989) and (Salmon, 1997).

Constraint-based approaches to causal discovery were put forward by Pearl and Verma (Pearl & Verma, 1991) and by Spirtes, Glymour, and Scheines (Spirtes et al., 1991). The PC and FCI algorithms, for instance, take a global approach to causal discovery and output a graph with different types of edges between all the variables to represent for example that X causes Y, X does not cause Y, or the causal direction is undetermined (Spirtes et al., 1993). The FCI can also model latent variables.

Earlier research on learning Bayesian networks from data using a Bayesian approach (Cooper & Herskovits, 1992; Heckerman et al., 1995) has simultaneously modeled all the causal relationships among the model variables. These global approaches have worst-case search time complexities that are exponential in the number of measured variables V. LCD constrains the search space to triplets of variables.

LCD and its variants (LCDa, LCDb and LCDc) output only causes of the form Y causes Z and take a local approach to causal discovery (evaluate only triplets of the form XYZ). By searching only for pairwise causal relationships, they trade off completeness for efficiency.

Recently, Silverstein and others have used a variant of LCD to perform *market basket analysis* to discover causal association rules (Silverstein et al., 2000). Their algorithm uses in addition patterns such as $A \rightarrow C \leftarrow B$ to infer that A and B cause C, assuming no hidden variables and confounding.

2.4.1 Time Complexity of LCDa, LCDb and LCDc

We assume here that the number of levels (states) of any of the variables in V is bounded by some constant. We have used a statistical test based on a Bayesian scoring metric for establishing dependence and independence. The time complexity is $O(m)$ where m is the size of the dataset (number of instances). If there are n variables in the database, the time complexity of the LCD variants is $O(mn^2r)$, where m is the number of records in the database, n is the number of variables and r is the number of variables of type X. If we restrict the number of variables of type X in the search, so that r is bounded above by some constant, then the time complexity is $O(mn^2)$. Likewise, if we focus on a bounded number of effects of interest (variable Z), the time complexity becomes $O(mn)$. The space complexity of these algorithms is also $O(mn)$, which is the size of the database.

The efficiency can be further improved if we are interested in just answering the question whether some particular Y_i causes a particular Z_j. Based on the ordering (partitioning) only a limited number of X variables will be required to determine the causal influence of Y_i on Z_j . If we restrict our choice to a constant number k of such variables, the time complexity and space complexity can be reduced to just $O(km)$ and hence $O(m)$. Using this framework we can assess the causal influence of any one arbitrary node Y on Z in $O(m)$ time and space requirements. Note though that the LCD variants are *incomplete* algorithms i.e. in general they cannot find all the unconfounded causal influences. Hence the absence of an output of the form Y_i causally influences Z_j, does not guarantee the absence of such a relationship between them.

Further, all these algorithms can be implemented in an *anytime* framework, to output the causes as they are discovered. The time complexity of LCDa, LCDb and LCDc makes them appropriate for exploring possible causal relationships in databases that contain a very large number of records (on the order of hundreds of thousands) and a moderately large number of measured variables per record (on the order of hundreds).

2.5 Experimental Methods

In the evaluation of causal output, we have to consider two dimensions—qualitative or structural and quantitative or parameterization. The output of the algorithm can be evaluated for causal influences—both in terms of structure and parameterization of the variables of interest. These are compared with the true structure and scored as explained below.

2.5.1 The Alarm network and dataset

In this study, we used as gold standard a causal model that was constructed by an expert—the Alarm causal Bayesian network, which contains 37 nodes and 46 causal arcs. Each node can have two to four possible states. Beinlich developed Alarm to model potential interactions in the operating room while providing anesthesia to the patient (Beinlich et al., 1990). His expertise as an anesthesiologist and medical knowledge from literature went into the development of the Alarm network. Alarm has been used extensively in evaluations of Bayesian network induction. We believe it remains a useful standard benchmark. The total number of possible distinct pairs in the Alarm network is 666. Each pair (Y,Z) is categorized as follows:

Causal and not confounded(C) There is a directed path from Y to Z, and there is no common ancestor X that has a directed path to Y and a directed path to Z that does not traverse Y. The nodes *Lung Cancer* and *Mass seen on Chest X-ray* in Figure 1 are *causal*.

Causally-confounded (CC) There is a directed path from Y to Z, and there is a common ancestor X that has a directed path to Y, and a directed path to Z that does not traverse Y. The nodes *Chronic Bronchitis* and *Fatigue* in Figure 1 are *causally-confounded* by *History of Smoking* and *Lung Cancer*.

Confounded-only (CO) There is no directed path between Y and Z, and there is a common ancestor X that has a directed path to Y, and a directed path to Z that does not traverse Y. The nodes *Chronic Bronchitis* and *Lung Cancer* in Figure 1 have the *confounded-only* relationship.

Independent (I) There is no d-connecting path (Pearl, 1991) between Y and Z. There are no *independent* node pairs in Figure 1.

Table 2: Categories of node pairs in the Alarm network

Description	Abbreviation	Number
Causal	C	167
Causally-confounded	CC	56
Confounded-only	CO	78
Independent	I	365
Total		666

Table 2 gives the distribution of these different categories for the actual ALARM network. Note that for *causal* and *causally-confounded* pairs, directionality is also to be considered while evaluating the output of an algorithm. When categorizing a pair (Y,Z) as *causal* or *causally-confounded*, the direction of the arc between Y and Z is important. If the direction is incorrect, two types of mis-categorization can occur: *causally-reversed* and *causally-confounded-reversed*. But since in the study reported here, we assume an ordering of

the variables, the *causally-reversed* and the *causally-confounded-reversed* are not possible. For causal discovery we generated a set of 5000 instances by simulation from the Alarm network. The number of these instances that we applied in our experiments varied from 50 to 5000.

2.5.2 Evaluation Metrics

Error metrics have been proposed to predict the distribution of Z given that Y is observed and also to predict the distribution of Z given that Y is manipulated (Cooper & Yoo, 1999). We adapted these metrics. In particular, since our study focused on causal discovery from observational data, we derive the following metric (see Equation 1) where we manipulate the nodes in the true Alarm network but use the simulated data as observational. In Equation 1, y refers to an arbitrary state of Y and z an arbitrary state of Z. The notation $manip(Y = y)$ means that Y is manipulated to the state y. $P_A(Z = z|manip(Y = y))$ is the conditional probability inferred from Alarm when observing Z is z while Y is manipulated to be y. If Y causes Z and the two variables are not confounded, then in the large sample limit $P_A(Z = z|manip(Y = y))$ will equal $P_E(Z = z|(Y = y))$, where $P_E(Z = z|(Y = y))$ is an estimate from the dataset of the conditional probability of Z given Y. r_Y and r_Z denote the number of states of variable Y and Z respectively. The manipulation observation prediction error for a pair of nodes Y, Z is computed as follows:

$$MOPErr_{Y,Z}(D) = \sum_y \frac{1}{r_Y}[\frac{1}{r_Z} \cdot \sum_z |P_A(Z = z|manip(Y = y)) - P_E(Z = z|(Y = y))|] \quad (1)$$

Equation 1 derives the average absolute error in the predicted probabilities for the states of Z given a uniform random manipulation of Y. The metric was computed for 534 pairs[3] out of the total possible 666 pairs (Y,Z). $P_E(Z = z|(Y = y))$ was estimated from each of the datasets and error metrics were computed for the 12 different dataset sizes.

2.5.3 Experimental Runs

LCDa, LCDb and LCDc were run as follows. We used 12 different dataset sizes, varying from 50 to 5000 instances. The causal output was categorized as described in section 2.5.1. Due to the ordering schema used, the categories of *causally-reversed* and *causally-confounded-reversed* are not output by LCDa, LCDb, and LCDc. This reduces the false positive base (denominator) to 499[4], which is used to calculate the false positive rate (FPR). We used the total causal pairs in Alarm (167) as the base for computing the true positive rate (TPR). For each causal output, error rates were derived using Equation 1. Mean errors were computed separately for each category of output (causal, causally-confounded, and confounded-only) for each training set size.

[3]The number of pairs are lower due to the ordering of the nodes.

[4]This was obtained as follows. Total number of pairs in Alarm (666) minus the unconfounded causal pairs (167).

3 RESULTS AND DISCUSSION

Table 3: LCDa output at different dataset sizes. Integers indicate instance counts and reals denote mean error rates.

Inst	C	C_err	CC	CC_err	CO	CO_err
50	15	0.0530	7	0.0732	0	0
100	24	0.0501	9	0.0727	4	0.2688
200	29	0.0374	14	0.0666	4	0.2282
300	32	0.0290	16	0.0699	8	0.2007
400	32	0.0260	20	0.0591	7	0.1912
500	33	0.0251	21	0.0611	6	0.1878
750	32	0.0202	24	0.0509	5	0.1774
1000	36	0.0185	25	0.0435	6	0.1944
2000	41	0.0115	38	0.0469	2	0.1809
3000	43	0.0108	42	0.0457	2	0.1379
4000	44	0.0097	41	0.0430	2	0.1375
5000	45	0.0102	43	0.0378	1	0.0759

Inst–Instances, C–Causal, CC–Causally-confounded, CO–Confounded-only, err–error

Table 4: LCDb output at different dataset sizes

Inst	C	C_err	CC	CC_err	CO	CO_err
50	15	0.0530	6	0.0745	0	0
100	20	0.0444	9	0.0727	1	0.3069
200	27	0.0374	11	0.0568	2	0.2545
300	31	0.0286	16	0.0699	2	0.2059
400	32	0.0260	17	0.0633	1	0.2281
500	33	0.0251	18	0.0608	1	0.2331
750	32	0.0202	20	0.0485	2	0.1988
1000	35	0.0183	25	0.0435	1	0.2426
2000	41	0.0115	34	0.0359	0	0
3000	43	0.0108	38	0.0347	0	0
4000	44	0.0097	38	0.0340	0	0
5000	45	0.0102	39	0.0316	0	0

Tables 3, 4, and 5 summarize the respective performance of LCDa, LCDb and LCDc. Figure 4 gives the true positive rates (TPR) and false positive rates (FPR) for these LCD variants. Regarding TPR, LCDa and LCDb converge at the dataset size of 2000, and LCDc approaches LCDa and LCDb curves at the training set size of 3000. On the other hand, the FPR is consistently different across the datasets, with FPR.c < FPR.b < FPR.a. With LCDc only one false positive category (causally-confounded) is output. This is also true of LCDb with a training set size

Table 5: LCDc output at different dataset sizes

Inst	C	C_err	CC	CC_err	CO	CO_err
50	15	0.0530	5	0.0302	0	0
100	19	0.0444	2	0.0341	0	0
200	25	0.0388	6	0.0286	0	0
300	24	0.0283	5	0.0534	0	0
400	25	0.0273	6	0.0431	0	0
500	26	0.0270	6	0.0450	0	0
750	24	0.0213	9	0.0295	0	0
1000	30	0.0184	10	0.0203	0	0
2000	38	0.0112	19	0.0164	0	0
3000	42	0.0109	19	0.0138	0	0
4000	42	0.0097	18	0.0116	0	0
5000	44	0.0104	19	0.0126	0	0

of 2000 and above. The *confounded-only* pattern output of LCDa reduces to 1 at the training set size of 5000.

The TPR for LCDa and LCDb is 0.27 (45 causal pairs out of the total possible 167) when all the 5000 instances are used. The TPR for LCDc is 0.26 (44 causal pairs). These three LCD variants can in theory detect only 53 out of this 167. This is because the algorithms cannot detect causes which originate from the first partition. Twelve root nodes in Alarm formed the first partition (p_0), and there are 114 causal relationships originating from any one of these root nodes. This shows that the algorithms actually find more than 80 percent of the true causal relationships they can *possibly* find.

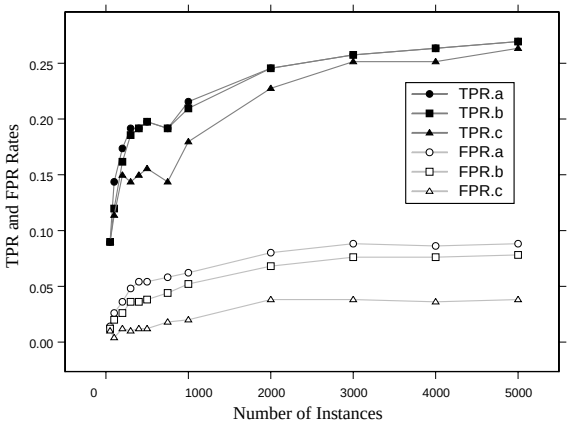

Figure 4: TPR and FPR of structural error for the LCDa, LCDb and LCDc algorithms

The manipulation-observation error for the relationships that are output by the three algorithms are given in Tables 3, 4, and 5 (see the C_err, CC_err, and CO_err columns). Figure 5 graphically displays these errors. In Figure 5, the dark lines show the error in predicting the probability distribution of Z, given manipulation of Y, even when an algorithm correctly concludes that Y causes Z (C_err). These errors are due to estimating a conditional probability using a finite sample of instances. The lighter lines (CC_err and CO_err) indicate errors due to sample size *and* to incorrectly assuming that Y and Z are unconfounded. Remarkably, LCDc has an error rate that reduces almost to sample size induced error. This result suggests that any confounded relationships being output by LCDc as causal and unconfounded are indeed only very weakly confounded.

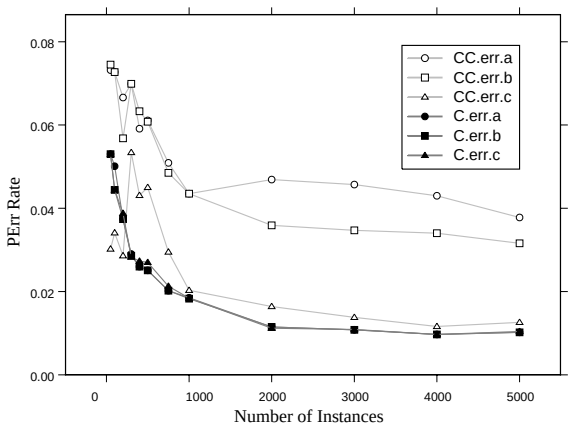

Figure 5: Causal and causally-confounded error rates of LCDa, LCDb and LCDc

LCDb and LCDc make use of more independence tests when there is more than one X variable for a pair Y, Z. This results in elimination of pairs with relatively higher manipulation-observation error which is an index of the parameters of the X, Y pair. (See CC-error plots in Figure 5) Qualitatively we identified the causally-confounded patterns which were output by LCDa but not by LCDb and LCDc. Figure 6 shows a representative example.

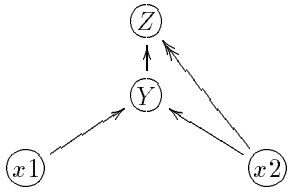

Figure 6: A causally-confounded pattern output by LCDa, but not by LCDb or LCDc. A double arrow denotes a path length greater than one.

In this example LCDa output Y causally influences Z, while LCDb and LCDc did not. The independence test—(IND $(x1, Z|Y)$) was positive while (IND $(x2, Z|Y)$) was negative. Since LCDa requires only one positive independence test, it output $Y \rightarrow Z$. This can be explained by the fact that with $x2$ confounding is more direct and local. The $x1$ confounding path $(x1 \rightarrow Y \leftarrow x2 \rightarrow Z)$ is longer than the $x2$ confounding path $(x2 \rightarrow Z)$.

4 CONCLUSION AND FUTURE WORK

LCDa, LCDb, and LCDc are efficient algorithms which use the local causal discovery framework. By making use of more independence tests LCDb and LCDc were able to reduce the FPR and causally-confounded error rates monotonically while at sufficient sample size obtaining the TPR of LCDa. All these LCD variants (in particular LCDb and LCDc) appear to be good candidate algorithms for efficient causal datamining. Since the motivation for this work arose from analyzing the false positive ouput of LCD on Alarm, testing LCD variants on a different network(s) will be important.

We plan to explore two future algorithmic directions. One is to develop more sensitive independence tests to reduce further the FPR. The other is to try to develop new search techniques to improve the accuracy of the causal output, yet retain computational efficiency. We also plan to use additional causal networks to evaluate the algorithms.

Acknowledgements

We thank Changwon Yoo for providing the formatted version of the Alarm dataset used in this research, and for helpful discussions. This work was supported by the National Library of Medicine (NLM) training grant LM07059 to Subramani Mani, by NLM research grants R01-LM06696 and R01-06759, and by National Science Foundation grant IIS-9812021 to Greg Cooper.

References

Beinlich, I. A., Suermondt, H., Chavez, R. M., & Cooper, G. F. (1990). The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. *Proceedings of the Second European Conference on Artificial Intelligence in Medicine* (pp. 247–256). London: Chapman and Hall.

Bollen, K. A. (1989). *Structural Equations With Latent Variables.* New York: Wiley.

Cooper, G., & Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning, 9*, 309–347.

Cooper, G. F. (1997). A simple constraint-based algorithm for efficiently mining observational databases for causal relationships. *Data Mining and Knowledge Discovery, 1*, 203–224.

Cooper, G. F., & Yoo, C. (1999). Causal discovery from a mixture of experimental and observational data. *Uncertainty in Artificial Intelligence 99* (pp. 116–125). San Francisco, California: Morgan Kaufmann Publishers.

Fetzer, J. (Ed.). (1989). *Probability and Causality.* Boston: D. Reidel Publishing Company.

Glymour, C., & Cooper, G. F. (Eds.). (1999). *Computation, Causation, and Discovery.* Cambridge, MA: MIT Press.

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

Mani, S., & Cooper, G. F. (1999). A study in causal discovery from population-based infant birth and death records. *Proceedings of the AMIA Annual Fall Symposium* (pp. 315–319). Philadelphia, PA: Hanley & Belfus.

Pearl, J. (1991). *Probabilistic Reasoning in Intelligent Systems.* San Francisco, California: Morgan Kaufmann. 2 edition.

Pearl, J., & Verma, T. (1991). A theory of inferred causation. *Principles of Knowledge Prepresentation and Reasoning: Proceedings of the Second International Conference* (pp. 441–452). San Francisco, California: Morgan Kaufmann Publishers.

Salmon, W. C. (1997). *Causality and Explanation.* New York: Oxford University Press.

Silverstein, C., Brin, S., Motwani, R., & Ullman, J. (2000). Scalable techniques for mining causal structures. *Data Mining and Knowledge Discovery, 4*, 163–192.

Spirtes, P., Glymour, C., & Scheines, R. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review, 9*, 62–72.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, Prediction, and Search.* New York: Springer-Verlag.