Causal Discovery Algorithms based on Y Structures

Subramani Mani SUBRAMANI.MANI@VANDERBILT.EDU
Department of Biomedical Informatics, Vanderbilt University, 400 Eskind Biomedical Library, 2209 Garland Avenue, Nashville, TN, 37232-8340, USA

Gregory F. Cooper GFC@CBMI.PITT.EDU
Department of Biomedical Informatics, University of Pittsburgh, Parkvale Building M-183, 200 Meyran Avenue, Pittsburgh, PA 15260, USA

Keywords: Causal discovery, Y structures, Bayesian networks, observational data

1. Introduction and Background

Discovering relationships of the form "A causally influences B" is valuable in different fields of study. These relationships are also referred to as "cause and effect" relationships where A represents the cause, and B denotes the effect. Generally, experimental studies are performed to ascertain causality where the value of a variable is set randomly and its effects measured under controlled experimental settings. However, such experiments may not be feasible due to logistical, ethical or cost considerations. We believe that discovery algorithms that can ascertain causality from observational (passively collected) data are valuable. The framework we use for causal discovery is founded on causal Bayesian networks. A causal Bayesian network (CBN) is a Bayesian network (BN) 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). For example, if there is a directed edge from A to B ($A \rightarrow B$), node A is said to exert a causal influence on node B.

We characterize the *causal influence* of a variable A on variable B using the *manipulation criterion* (Spirtes et al., 2000; Glymour & Cooper, 1999). The manipulation criterion states that if we had a way of setting just the values of A and then measuring B, the causal influence of A on B will be reflected as a change in the conditional distribution of B. That is, there exist values a_1 and a_2 of A such that $P(B|\text{ set } A = a_1) \neq P(B|\text{ set } A = a_2)$. In a CBN an arc between any pair of nodes represents a causal influence.

The two basic assumptions that are necessary for our causal discovery framework are the causal Markov condition and the causal faithfulness condition. We now introduce the concept of a Y structure. Let $W_1 \rightarrow X \leftarrow W_2$ be a V structure. If there is a node Z such that there is an arc from X to Z, then the nodes W_1, W_2, X and Z form a Y structure (see G_1 in Figure 1). If a Y structure is learned from data, the arc from X to Z represents an unconfounded[1] causal relationship. See (Mani et al., 2006) for a formal proof.

The CBN learning methods can be classified as global or local based on their search space and their output. If the goal is to learn a unified CBN over all the model variables, the search methodology is termed *global*. PC (Spirtes et al., 2000, page 84–85) and OR (Moore & Wong, 2003) are global BN search algorithms. If the goal of the learning procedure is to discover causal models on subsets of the model variables (for example, pairwise causal relationships), a *local* search methodology is employed such as in BLCD.

2. Y structure algorithms

In this section we introduce three algorithms that make use of Y structures to discover cause and effect relationships from observational data. The PC algorithm takes as input a dataset D over a set of random variables $\mathbf{V}$, a conditional independence test, and an α level of significance threshold for the test. It then outputs an essential graph that we define below. Markov equivalence (also known as independence equivalence) is a relationship based on independence that establishes an equivalence class of directed acyclic graphs over an observed set of variables $\mathbf{V}$. These DAGs are statistically indistinguishable based on independence relationships among $\mathbf{V}$. Let U be one equivalence class of DAGs over $\mathbf{V}$. An essential graph $\mathcal{E}$ of U over $\mathbf{V}$ will have directed and undirected edges such that each di-

[1] A pair of nodes A and B are said to be *unconfounded* iff there is no node C such that there is a directed path from C to A and a directed path from C to B that does not pass through A.

rected edge between a pair of nodes A and B will be represented in *all* the DAGs in U and each undirected edge between a pair A and B in $\mathcal{E}$ will be represented as either $A \rightarrow B$ or $A \leftarrow B$ in *all* the DAGs in U (Chickering, 1995) with both arc types represented. PC also makes an assumption of *causal sufficiency*. This means that all the variables of the causal network are measured and there are no latent or hidden variables. Hence PC is not designed to discover hidden variables that are common causes of any pair of observed variables. See (Spirtes et al., 2000, page 84–85) for more details on the PC algorithm. The PC algorithm outputs both directed and undirected edges. A post-processing step that we add is performed on the set of arcs to identify the Y structures. The PC algorithm with this additional post-processing step is referred to as the PC-Y algorithm.

The optimal reinsertion (OR) algorithm is an algorithm for learning Bayesian networks using a score-based approach developed by Moore and Wong (Moore & Wong, 2003). The algorithm introduces a new search operator called optimal reinsertion. On each step a node is labeled as the target node. All incoming and outgoing arcs of this target node are removed and the node is reinserted with the "optimal" combination of incoming and outgoing arcs. The process is repeated with all nodes taking turns as the target node multiple times until no step changes the Bayesian network structure. Note that the OR algorithm outputs a global Bayesian network consisting of all the variables of the input dataset. We perform a post-processing step on the output (Bayesian network) to identify the Y structures. The OR algorithm with this additional post-processing step is referred to as the OR-Y algorithm.

The Bayesian local causal discovery algorithm (BLCD) conjectures unconfounded causal relationships between pairs of variables. Figure 1 shows models G_1, G_2, G_3, and G_4 from a four variable domain. The Score function of BLCD assigns a score to a model that represents the probability of the model given data and prior knowledge. For scoring the DAGs, we use the Bayesian likelihood equivalent (BDeu) scoring measure (Heckerman et al., 1995). Note that for a Y structure, the causal claim is valid for only the arc from X to Z. We represent the Y structure using the notation $X \Rightarrow Z$. In the large sample limit under the causal Markov and causal Faithfulness assumptions, $\mathrm{P}(X \Rightarrow Z|\mathrm{D})$ (D denotes the dataset) can be estimated using Equation 1:

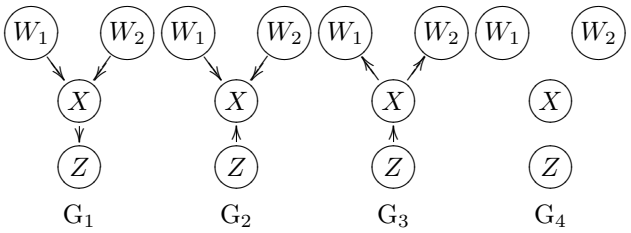

Figure 1: Several CBN models that contain four nodes, out of the possible 543 models. G_1 is a Y structure.

$$\frac{\mathrm{Score}(\mathrm{G}_1|\mathrm{D})}{\sum_{i=1}^{543} \mathrm{Score}(\mathrm{G}_i|\mathrm{D})} \quad (1)$$

where G_i represents one of the 543 CBNs over $\mathbf{V} = \{W_1, W_2, X, Z\}$. Interpreting the score as a probability is a heuristic approximation borne out by simulation studies (Mani, 2005).

The following are the steps of the BLCD algorithm. For each node $X \in \mathbf{X}$ ($\mathbf{X}$ denotes the set of all random observed variables in the dataset) perform the following:

1. **Estimate the Markov Blanket.** Estimate the Markov Blanket of X using the *Procedure MB* (Mani, 2005). Let $\mathbf{B}$ denote the estimated MB of X.
2. **Update B.** Apply the MB *update* rule (Mani, 2005).
3. **Pick W_1, W_2, and Z.** Obtain all possible distinct triplets (sets of three nodes) from $\mathbf{B}$. Add X to each triplet to get sets of four variables. We refer to each set of four variables as a tetraset $\mathbf{T}$. Since we are focusing on the MB of X, X is an essential element of $\mathbf{T}$. Note that each tetraset can give rise to 3 "Y" patterns where the X variable is a cause and each of the other three variables are potential effects.
4. **Derive $\mathbf{P}(X \rightarrow Z \,|\, \mathbf{D})$**: For each tetraset $\mathbf{T}$ and for each of the three "Y" structures defined by the variables in $\mathbf{T}$, derive the posterior probability of $X \Rightarrow Z$ using Equation 1.
5. **Generate output**: If $P(X \rightarrow Z \,|\, \mathrm{D}) > t$, where t is a user-set threshold, then output $X \rightarrow Z$ as a purported, unconfounded causal relationship.

The reader is referred to (Mani & Cooper, 2004; Mani, 2005) for additional details on BLCD.

3. Experimental methods

By using expert-defined CBNs we can (1) generate data from those models, (2) apply CBN discovery algorithms to the data, and (3) evaluate the causal relationships output by the algorithms relative to the data generating CBNs that serve as gold standards. The

output of the algorithms were compared with the data generating structure and scored as explained below.

BLCD was implemented in the C programming language. We obtained the Tetrad program that implements the PC algorithm from Professor Peter Spirtes. The OR implementation was obtained from Professor Andrew Moore (http://www.autonlab.org/autonweb/software/10530.html). For PC-Y, OR-Y and BLCD the Y structures output by the algorithms were compared with the Y structures of the data generating networks. In short, we examine the intersection set of Y structures output by the algorithms and the Y structures present in the data generating model. Precision and recall were computed for the three algorithms.

Five Bayesian networks built by domain experts in such varied fields as medicine, atmospheric sciences and agriculture were identified. These networks are Alarm (Beinlich et al., 1990), Hailfinder (Abramson et al., 1996), Barley (Kristensen & Rasmussen, 2002), Pathfinder (Heckerman et al., 1992), and Munin (Andreassen et al., 1987). For causal discovery, we generated simulated training instances by stochastic sampling (Henrion, 1986). Table 1 gives the distribution of the nodes, arcs and Y structures for the various networks used in our study.

Table 1: Nodes, arcs and Y structures in the Alarm, Hailfinder, Barley, Pathfinder, and Munin networks

Category	Alarm	HF	Barley	PF	Munin
Nodes	37	56	48	131	189
Arcs	46	66	84	195	282
Y structures	13	20	44	5	147

HF: Hailfinder; PF: Pathfinder

4. Results

The results presented below are based on a sample size of 20,000 instances for each of the five domain datasets that were generated. We present a summary performance of the causal discovery algorithms based on the aggregate number of Y arcs (YA)[2] present in all the data generating networks (see Table 2). For all the three algorithms, default thresholds were used to generate these results. Altogether there were 229 YA in the five domain CBNs. BLCD had the highest precision (0.827) followed by PC-Y (0.548). The best recall was achieved by OR-Y (0.314). We used a Z test statistic (two sided) to test the difference between the two proportions across all the algorithms pairwise for both precision and recall (algorithm A precision versus algorithm B precision, and algorithm A recall versus algorithm B recall). Standard errors were estimated and 95% confidence intervals (CI) were computed after pooling the two proportions (Daniel, 1991, pages 152 and 225). The null hypothesis of no difference in the two proportions was rejected if the p value was < 0.017 as we did multiple comparisons (3) of precision and recall proportions. Based on this analysis, there is a significant pairwise difference in the proportions of all the three precision comparisons ($p < 0.0001$). However, there is no significant difference in the recall proportions. Figure 2 presents the precision recall-curve

Table 2: Precision and recall based on 229 Y arcs from all datasets (20,000 samples).

Algorithm	Total	YA in both	YA_P	YA_R
OR-Y	283	72	0.254	**0.314**
BLCD	75	62	**0.827**	0.271
PC -Y	93	51	0.548	0.223

Total: Total number of arcs output as causal by the algorithm; YA in both: Y arcs output by the algorithm and present in the generating network; YA_P: Y arc Precision; YA_R: Y arc Recall

obtained by varying the threshold for BLCD. PC went out of memory with significance levels greater than 0.05 (default) and the OR algorithm does not have a parameter that can be varied to generate precision and recall curves. Hence for PC-Y and OR-Y, we provide the precision and recall values using the default thresholds as a comparison. Based on the precision recall-curve, the best precision and recall values for BLCD are 0.9 and 0.389 respectively.

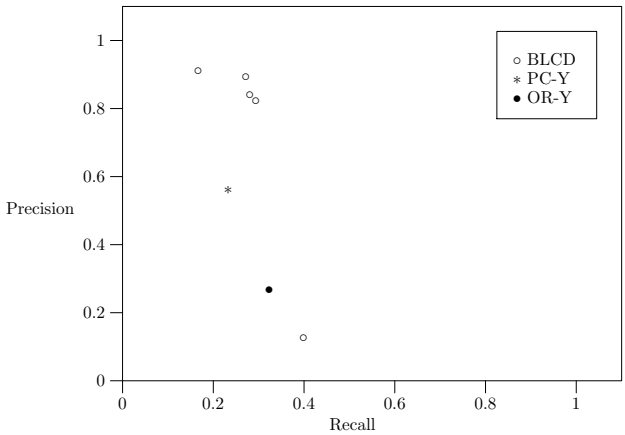

Figure 2: Precision versus recall plot. (20,000 samples)

[2]The arc from X to Z in a Y structure is referred to as a Y arc (YA).

5. Discussion

In this section we discuss the results and present the implications of our research for discovering causal relationships from observational data. This research has highlighted the role of Y structures for causal discovery from observational data using global and local causal Bayesian network learning algorithms. BLCD had statistically significantly better precision than PC-Y and OR-Y. This implies less number of false positives for BLCD when compared to PC-Y and OR-Y. There were no statistically significant differences in recall values among the three algorithms. The precision-recall curve of BLCD shows that varying the threshold can increase the precision to 0.9. Note that if the number of variables are considerably more than that used in this study (for example, gene expression datasets that have tens of thousands of variables) only BLCD can readily scale up among the three algorithms described in this paper. By design BLCD has access to only a small subset of variables at a time (4 variables) in the model evaluation stage. Using these four variables, BLCD ascertains the causal influence of a variable A on variable B.

The causal discovery framework that we presented for identifying direct causal relationships is dependent on the presence of Y structures in the data generating process. The three medical (Alarm, Pathfinder, Munin) and two non-medical (Hailfinder, Barley) networks that were used to generate data had varying numbers of Y structures. Since these networks were created by domain experts capturing the probabilistic dependencies and independencies in the domain, it is plausible to assume the occurrence of Y structures as components of the data generating process in many real-world domains.

References

Abramson, B., Brown, J., Edwards, W., Murphy, A., & Winkler, R. L. (1996). Hailfinder: A Bayesian System for Forecasting Severe Weather. *International Journal of Forecasting*, *12*, 57–71.

Andreassen, S., Woldbye, M., Falck, B., & Andersen, S. K. (1987). MUNIN — A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence* (pp. 366–372). San Mateo, CA: Morgan Kaufmann.

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.

Chickering, D. (1995). A transformational characterization of equivalent bayesian network structures. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (pp. 87–98). San Francisco: Morgan Kaufmann.

Daniel, W. W. (1991). *Biostatistics: A foundation for analysis in the health sciences*. John Wiley and Sons, Inc. 5 edition.

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.

Heckerman, D., Horvitz, E., & Nathwani, B. (1992). Towards normative expert systems: Part I The Pathfinder Project. *Methods of Information in Medicine*, *31*, 90–105.

Henrion, M. (1986). Propagating uncertainty in bayesian networks by probabilistic logic sampling. *Proceedings of the 2nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-86)* (pp. 0–0). New York, NY: Elsevier Science Publishing Company, Inc.

Kristensen, K., & Rasmussen, I. (2002). The use of a Bayesian network in the design of a decision support system for growing malting barley without use of pesticides. *Computers and Electronics in Agriculture*, *33*, 197–217.

Mani, S. (2005). *A Bayesian Local Causal Discovery Framework*. Doctoral dissertation, University of Pittsburgh.

Mani, S., & Cooper, G. F. (2004). Causal discovery using a Bayesian local causal discovery algorithm. *Proceedings of MedInfo* (pp. 731–735). IOS Press.

Mani, S., Spirtes, P., & Cooper, G. F. (2006). A theoretical study of Y structures for causal discovery. *Proceedings of the Conference on Uncertainty in Artificial Intelligence* (pp. 314–323). Corvallis, OR: AUAI Press.

Moore, A., & Wong, W.-K. (2003). Optimal reinsertion: A new search operator for accelerated and more accurate bayesian network structure learning. *Proceedings of the 20th International Conference on Machine Learning (ICML '03)* (pp. 552–559). Menlo Park, California: AAAI Press.

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

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT Press. 2nd edition.