

Learning Causal Bayesian Networks from Observations and Experiments: A Decision Theoretic Approach

Stijn Meganck¹, Philippe Leray², and Bernard Manderick¹

¹ Vrije Universiteit Brussel, Pleinlaan 2, 1050 Brussels, Belgium,
smeganck, bmanderi@vub.ac.be,

WWW home page: <http://como.vub.ac.be>

² INSA Rouen, Laboratoire PSI, BP 08 - Avenue de l'Université
76801 St-Etienne du Rouvray Cedex, France,

Philippe.Leray@insa-rouen.fr,

WWW home page: <http://psiserver.insa-rouen.fr/psi/>

Abstract. We discuss a decision theoretic approach to learn causal Bayesian networks from observational data and experiments. We use the information of observational data to learn a completed partially directed acyclic graph using a structure learning technique and try to discover the directions of the remaining edges by means of experiment. We will show that our approach allows to learn a causal Bayesian network optimally with relation to a number of decision criteria. Our method allows the possibility to assign costs to each experiment and each measurement. We introduce an algorithm that allows to actively add results of experiments so that arcs can be directed during learning. A numerical example is given as demonstration of the techniques.

1 Introduction

Bayesian networks (BNs), introduced by Pearl [1], have become well known tools for working in domains with uncertainty. They allow performant probabilistic inference and give an intuitive representation of the domain.

BNs can also represent causal information when the edges represent causal relations between the corresponding variables [2]. The causal relation between two variables, in the form of a directed edge from a cause variable C to an effect variable E , is understood as the effect a manipulation of variable C (the cause) would have on variable E (the effect).

Learning BNs can be done from observations alone, by first learning the completed partially directed acyclic graph (CPDAG) and then choosing a possible complete instantiation in the space of equivalent graphs defined by this CPDAG.

This is impossible for causal Bayesian networks (CBNs), because there is only one true causal network that represents the underlying mechanisms, so the remaining edges have to be directed to represent the correct causal influence.

We discuss a decision theoretic approach for learning CBNs from a mixture of observational and experimental data. We assume we learn a CPDAG using

a structure learning technique and then direct all the remaining arcs in the resulting CPDAG based on the results of experiments.

Algorithms exist to learn CBNs based on experiments [3, 4] and in [5] techniques have been developed to learn CBN from a mixture of experimental and observational data.

In [6, 7] it has been shown that in order to learn a complete structure at most $\log_2(N) + 1$, with N the number of variables, experiments are needed. This result is given as a theoretical bound for the worst case scenario.

The main difference with the Active Learning approaches in [3, 4] is that we assume that there is a number of observational data which we can use to form an initial CPDAG in which every directed edge is a representative of a causal mechanism. Next we introduce an experimentation phase in which we perform specific experiments in order to learn the completed CBN optimally based on some decision criterion. Our technique tries to find an optimal experimentation strategy in order to minimize the number of experiments, which should be lower than the bound derived in [7]. We allow the possibility to assign costs to an experiment, which might influence the decision of choice for performing an experiment. This type of setting is typical for medical applications where there is a lot of data from patients but it might be very costly to perform experiments.

The remainder of this paper is as follows, in the next section we introduce some notations and definitions. Then we will discuss several of the assumptions we make before introducing our decision theoretic approach and criteria used. We end with a conclusion and future work.

2 Notations and Definitions

In this work uppercase letters are used to represent variables or sets of variables, i.e. $V = \{X_1, \dots, X_n\}$, while corresponding lowercase letters are used to represent their instantiations, i.e. x_1, x_2 and v is an instantiation of all x_i . $P(X_i)$ is used to denote the probability distribution over all possible values of variable X_i , while $P(X_i = x_i)$ is used to denote the probability distribution over the instantiation of variable X_i to value x_i . Usually, $P(x_i)$ is used as an abbreviation of $P(X_i = x_i)$.

$Ch(X_i)$, $Pa(X_i)$, $Ne(X_i)$ respectively denote the children, parents and neighbors of variable X_i in a graph. Furthermore, $Pa(x_i)$ represents the values of the parents of X_i .

A *causal Bayesian network* (CBN) $\langle V, G, P(x_i|Pa(x_i)) \rangle$, with:

- $V = \{X_1, \dots, X_n\}$, a set of observable discrete random variables
- a directed acyclic graph (DAG) G , where each node represents a variable from V
- conditional probability distributions (CPD) $P(x_i|Pa(x_i))$ of each variable X_i from V conditional on its parents in the graph G .

is a Bayesian network in which the directed edges are viewed as representing *autonomous causal relations* among the corresponding variables, while in a BN

the directed edges only represent a probabilistic dependency, and not necessarily a causal one.

3 Assumptions

In this section we will discuss some assumptions we make about the domain to apply our algorithm.

3.1 Faithful distribution

We assume the observed samples come from a distribution faithful to a CBN, i.e. there are no hidden variables or confounding factors. This means that for each two variables X_i and X_j connected in the graph either $X_i \rightarrow X_j$ or $X_i \leftarrow X_j$ must hold [8]. During the remainder of this paper we will always accept this assumption.

3.2 Correct CPDAG

We assume that after an initial learning phase we are given the correct CPDAG as a consequence of applying a learning algorithm such as PC [9], BLCD [10], etc. For the technique presented in this paper we will accept this assumption.

In general it is possible that these algorithms do not retrieve the correct CPDAG, for a discussion on the properties of these algorithms see [9]. If this assumption is false, there are a lot of problems that might occur and we will have to adapt our algorithm. We are currently studying learning CBNs in this setting, and this is a part of our future research.

3.3 Modular experiments

In order to find the causal relation between two variables X and Y we have to check whether randomizing X holding all other variables fixed at a certain value induces a variation in Y and/or vice versa. If we just randomized X and not put any constraints on the other variables then we can only detect which other variables covary with X but we cannot detect whether this relationship is direct or mediated by other variables.

If no variation is found by randomizing X this is possible due to the specific value assignment of all the other variables, and so it could be possible that we have to perform a randomization proces for all possible value assignments. Fortunately if we have the correct CPDAG, it is easier to find the causal relationships between all variables. The only possible unknown direct effects of a certain variable X are those that are connected to X by an undirected edge in the CPDAG. So performing an experiment at X , this is randomizing X , will give us direct information on the directionality of all undirected edges connected to X .

4 Decision theoretic approach

We use a decision theoretic approach to learn the CBN from a given CPDAG. In general a decision problem consists of three parts: values (symptoms, observables), actions and possible consequences. It is assumed that these are given in advance. It is possible to order the consequences by preference by using a utility function. Hence we can choose the action that will lead to our preferred result based on some decision criteria such as least risk or optimistic estimation.

Our decision problem is represented graphically in Figure 1, in which the possible actions are performing experiments, the values are the results of these, and the consequences are the relative utilities of the experiment. It is clear that we cannot construct the entire decision tree for this problem. Since the problem is iterative a decision can be dependent on the choice of a previous one, so we would have to construct a subtree for each possible sequence of actions and the size of the tree would explode.

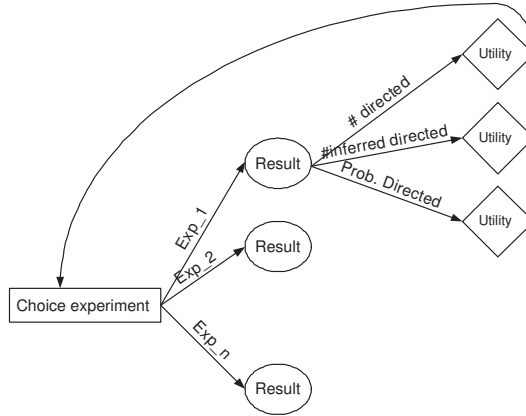


Fig. 1. Decision problem of learning a CBN from a given CPDAG.

4.1 Utility function

In general our utility function $U()$ will be a function of three variables: $gain(exp)$, $cost(exp)$, $cost(meas)$, respectively the gained information, the cost of performing an experiment and the cost of measuring other variables. If we denote performing an action (=experiment) at X_i by A_{X_i} , and measuring the neighboring variables by M_{X_i} then the utility function can be noted as:

$$U(A_{X_i}) = f(gain(A_{X_i}), cost(A_{X_i}), cost(M_{X_i})) \quad (1)$$

The only restriction that is placed on the utility function is that it is proportional to $gain(A_{X_i})$ and negative proportional to $cost(A_{X_i})$ and $cost(M_{X_i})$. In

this paper we assume the following utility function:

$$U(A_{X_i}) = \frac{\alpha \text{gain}(A_{X_i})}{\beta \text{cost}(A_{X_i}) + \gamma \text{cost}(M_{X_i})} \quad (2)$$

where α, β and γ are measures of importance for every part. We will assume $\alpha = \beta = \gamma$ unless stated otherwise, this allows to simplify the notation.

Gain of an experiment In this section we describe the information that can be retrieved after performing an experiment. Since it is our goal to direct all remaining undirected edges of our current CPDAG, the amount of edges we can direct after having results from an experiment is the gain of our experiment.

Lets assume that we perform an experiment on X_i and that we can measure all neighboring variables $Ne(X_i)$. In this case we can direct all links connecting X_i and $Ne(X_i)$ as a result of the experiment. So in this case the $\text{gain}(A_{X_i})$, with $A_{X_i} = \text{experiment on } X_i$, is based entirely on the number of variables that are connected to X_i by an undirected arc.

However it is possible that directing one arc can infer direction of other arcs, see the final phase of the PC-algorithm [9]. It is possible to take into account the possibility of inferred edges in $\text{gain}(A_{X_i})$. Note that the amount of edges of which the direction can be inferred after performing an experiment is entirely based on the instantiation of the undirected edges connected to the one being experimented on. An instantiation of an undirected edge is assigning a direction to it, so for instance if we have an edge $X - Y$, then $X \rightarrow Y$ and $X \leftarrow Y$ are the two possible instantiations of that edge. We denote $\text{inst}(A_{X_i})$ as the set of instantiation of the undirected edges connected to X_i . The number of inferred edges based on $\text{inst}(A_{X_i})$ is noted as $\#\text{inferred}(\text{inst}(A_{X_i}))$.

Note that two parts of a graph that are not linked in any way by undirected edges can not be influenced by performing an experiment in the other part when the CPDAG is correct. Since we assume that all discovered arcs are correct no existing arcs can change based on inferration by experiments, and hence no new information can be inferred through a set of already directed arcs. So the calculation of the utility of an experiment is only based on that part of the graph that is connected to the variable by undirected links. The problem can hence be separated in sub-problems, each concerning a part of the graph linked by undirected edges. In the remainder of this paper we will introduce solutions for a single substructure that is entirely constituted of undirected links. This result can then be mimicked for the other undirected substructures.

Cost of experiment and measurement The cost of an experiment can be the time needed, the amount of space it takes or simply the amount of money it costs to perform an experiment. It is dependent on the situation in which every experiment takes place and will typically be given by experts.

It is important to note that there are certain experiments that can not be performed, either because of ethical reasons (e.g. infecting people with HIV) or

simply because it is impossible to do so (e.g. changing the season). These types of experiments will be assigned a cost value of infinity (∞) and thus the gain of performing such an experiment will be 0, and therefore it will not add any new information.

In order to gain anything from an experiment we have to perform measurements on the variables of interest. It is however important to note that measuring itself can be costly and can diminish the usefulness of an experiment although it does not directly concern the variable that is being altered. For instance injecting someone with a certain fluid might not cost that much, but when the only way to check for changes is performing a CT-scan, measuring the results might add a huge cost factor.

5 Decision criteria

In this section we will discuss a number of decision criteria for our learning problem. Our approach allows the possibility to maximize any of these criteria in order to converge optimally to the solution for these criteria. Depending on the type of situation in which to perform the experiments it might be advantageous to choose a specific criterion. We will show by example the working mechanism of the different criteria on the network in Figure 2, a comparative study is part of future research.

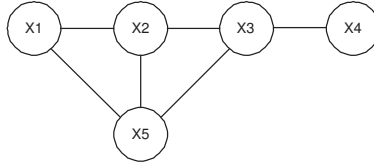


Fig. 2. Example CPDAG on which we show all optimization criteria.

5.1 Maximax

The *maximax* decision criterion is an optimistic one, which means that we choose the action that could give the best result, i.e. the one that might direct the most arrows. In our case this means that we perform an experiment on X_{best} with:

$$X_{best} = \underset{X_i}{argmax} \left(\frac{NeU(X_i) + \max_{inst(A_{X_i})} (\#inferred(inst(A_{X_i})))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (3)$$

This is the sum of the number of undirected edges connected to X_i and the maximum number of inferred edges by any of the instantiations of the directions of the undirected edges connected to X_i , divided by the cost.

In our example in Figure 2, if all costs are equal all variables except X_4 will have an equal maximal utility value, $U(X_i) = 6, i = 1, 2, 3, 5$.

5.2 Maximin

The *maximin* decision criterion is a pessimistic one, which means that we assume that for each experiment at a variable X_i the least number of possible inferred edges can be found. This means the minimum amount of edges oriented by any instantiation of all edges connected to X_i . In our case this means that we perform an experiment on X_{best} with:

$$X_{best} = \underset{X_i}{\operatorname{argmax}} \left(\frac{Ne_U(X_i) + \min_{inst(A_{X_i})} (\#inferred(inst(A_{X_i})))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (4)$$

The instantiation of edges that would induce the least inferred edges would be the one where all arrows are pointing at X_i , but this might create new v -structures and thus is not always possible. So if two neighbors of X_i are not directly connected, one of the links has to be out of X_i and hence leads to inferred edges.

In the example Figure 2 the optimal choices are variable X_2, X_3 and X_5 , since they all orient minimal 3 edges. For example, if we perform an experiment at X_2 , a minimal instantiation would be: $X_1 \rightarrow X_2, X_3 \rightarrow X_2, X_5 \rightarrow X_2$.

5.3 Laplace

Using the *Laplace* criterion means that we assume that all directionalities of edges are equally probable, for example for any two connected variables $P(X_i \rightarrow X_j) = P(X_i \leftarrow X_j) = 0.5$, for all non-directed edges $X_i - X_j$.

Every instantiation has thus a probability of $\frac{1}{\#inst(A_{X_i})}$ and for every possible instantiation we can calculate the number of inferred edges that will be directed.

In this case it would mean that we perform the experiment on X_{best} with:

$$\underset{X_i}{\operatorname{argmax}} \left(\frac{Ne_U(X_i) + \frac{\sum_{inst(A_{X_i})} \#inferred(inst(A_{X_i}))}{\#inst(A_{X_i})}}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (5)$$

In the example Figure 2 this is the variable X_3 with $U(X_3) = 3 + \frac{10}{5} = 5$. The derivation of this result is based entirely on the number of inferred edges for each instantiation. The results for X_3 are:

instantiation	#inferred	inferred
$\rightarrow X_2, \rightarrow X_4, \rightarrow X_5$	2	$X_2 \rightarrow X_1, X_5 \rightarrow X_1$
$\leftarrow X_2, \rightarrow X_4, \rightarrow X_5$	3	$X_2 \rightarrow X_1, X_5 \rightarrow X_1, X_2 \rightarrow X_5$
$\rightarrow X_2, \leftarrow X_4, \rightarrow X_5$	2	$X_2 \rightarrow X_1, X_5 \rightarrow X_1$
$\rightarrow X_2, \rightarrow X_4, \leftarrow X_5$	3	$X_2 \rightarrow X_1, X_5 \rightarrow X_1, X_2 \rightarrow X_5$
$\leftarrow X_2, \rightarrow X_4, \leftarrow X_5$	0	none

in which $\rightarrow X_i$ indicates an arrow from X_3 to X_i and vice versa.

5.4 Expected utility

The expected utility is based on a distribution of the directions of the links. Based on this distribution it is possible to calculate the probability of any instantiation of directions that might occur. We will discuss several ways to give distributions for the directionalities.

Probabilities based on equivalence class for general graph structure

Instead of just assuming a uniform distribution of the edges we can look at all possible dags in the equivalence class of the discovered CPDAG and count for each pair $X_i - X_j$, the number of times $X_i \rightarrow X_j$ and $X_i \leftarrow X_j$ appears and hence we can assume that:

$$P_{eq}(X_i \rightarrow X_j) = \frac{\#(X_i \rightarrow X_j)}{\#\text{members of eq. class}} \quad (6)$$

$$P_{eq}(X_i \leftarrow X_j) = \frac{\#(X_i \leftarrow X_j)}{\#\text{members of eq. class}} \quad (7)$$

Note that in future steps in the learning phase we no longer have a CPDAG, because some arcs may be directed based on knowledge from experiments. We should then take into account all members of the original equivalence class that share the exact same directed edges, for convenience we will still refer to this set of dags as the members of the equivalence class of the current PDAG.

Using this approach it would mean that we perform the experiment on the variable X_{best} with:

$$\underset{X_i}{\operatorname{argmax}} \left(\frac{Ne_U(X_i) + \sum_{inst(A_{X_i})} \#inferred(inst(A_{X_i}))P_{eq}(inst(A_{X_i}))}{cost(A_{X_i}) + cost(M_{X_i})} \right) \quad (8)$$

with $P_{eq}(inst(A_{X_i}))$ meaning the number of times a certain instantiation is present in the equivalence class divided by the number of members in that class.

The problem with this approach is that we need to know the exact number of elements in the equivalence class. As far as we know there is no exact way of calculating the number of elements in the equivalence class of a certain DAG. In theory it is possible to construct all equivalent graphs, but this is very time consuming, certainly for large graphs. Hence, we can not calculate $P_{eq}(X_i \rightarrow X_j)$ in practice for a general graph.

However we can solve the problem for tree structures, since the number of elements in the equivalence class of a tree is equal to the number of nodes (= number of edges + 1). We will use this property to construct an approximation for general structures.

Approximate probabilities for general graph structure Since the number of members in an equivalence class (of a CPDAG or PDAG as introduced earlier)

is generally unknown and hard to compute we will need an approximation to solve the problem.

Checking the utility of performing an experiment at X_i means running over all possible instantiations of $X_i - Ne(X_i)$, so in our approximate technique we need to have information on all these directions. We have seen that for tree structures this problem can be solved, so instead of working with the original structure we will use a Minimum Weight Spanning Tree (MWST) algorithm in which we force that all edges $X_i - Ne(X_i)$ are present and no edges except those in the original structure are present to approximate the original structure. The weights on all the edges are given by the mutual information between the variables based on the observational data. We then use the technique for trees and use the results as approximations for the general structure.

For example if we want to check for the expected utility based on the equivalence class of performing an experiment at X_1 in Figure 2, a possible MWST is given in Figure 3.

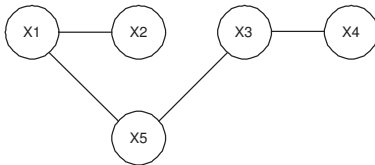


Fig. 3. Possible MWST for structure given in Figure 2.

Keep in mind that we will look at all possible instantiations of $X_i - Ne(X_i)$ as they occur in the original graph. For example in Figure 3 we will also look at the instantiation $X_2 \rightarrow X_1 \leftarrow X_3$, although this is a v -structure in the tree.

Expert knowledge It is also possible that an expert can give a certain probability to the direction of edges in the CPDAG. This is the least costly procedure since at the time of performing experiments an expert is present and it is no added cost to obtain the extra knowledge. So in this case the probabilities are based on the belief of an expert and is noted as:

$$P_{exp}(X_i \rightarrow X_j) = \text{belief of expert} \quad (9)$$

The utility function would be the same as given in equation 8 but with $P_{exp}()$ instead of $P_{eq}()$.

6 Learning algorithm

In this section we propose our learning algorithm. We introduce an adaptive algorithm in which it is assumed that experiments are performed during learning.

The first phase of the algorithm consists of applying a learning algorithm to obtain the correct CPDAG representing the equivalence class of all BNs faithful the distribution.

As stated, we allow the possibility to add newly discovered knowledge due to the experiments during the learning phase. Since experiments are performed we gain information on the direction of certain links, these may remove the need to perform certain other experiments. Remember that parts of the graph that are not connected by undirected links can be treated separately, so multiple instances of the algorithm can be applied in parallel to the substructures.

The complete algorithm is given in Algorithm 1.

Algorithm 1 Adaptive learning of CBN.

Require: Observational data set.

Ensure: A CBN.

1. Apply a learning algorithm on the data-set to obtain CPDAG G .
 2. Compute for each node X_i for which $\#Ne_U(X_i) > 0$ in G $U(A_{X_i})$ with equation 3, 4, 5 or 8.
 3. Perform an experiment at the node with the optimal $U(A_{X_i})$ value in relation to the decision criterion, X_{best} .
 4. For all $X_j \in Ne_U(X_{best})$
 - If distribution of X_j changed because of experiment,
 - then orient $X_{best} - X_j$ as $X_{best} \rightarrow X_j$
 - else orient $X_{best} - X_j$ as $X_{best} \leftarrow X_j$
 - end
 5. repeat
 - if $X_i \rightarrow X_j$ and X_j and X_k are adjacent, X_k and X_i are not and there is no arrow into X_j then orient $X_j - X_k$ as $X_j \rightarrow X_k$.
 - if there is a directed path from X_i to X_j and an edge between X_i and X_j then orient $X_i - X_j$ as $X_i \rightarrow X_j$.
 - until no more edges can be oriented.
 6. Return to Step (2) until all links are directed.
 7. Return CBN G .
-

7 Example

For example lets assume that the correct causal network of Figure 2 is as given in Figure 4.

Run with Maximin We have seen in Section 5.2 that for the *maximin* criterion the optimal choices were X_2 , X_3 and X_5 . Suppose we choose to perform an experiment at X_2 , this would result immediatly in the following instantiation: $X_2 \rightarrow X_1$, $X_2 \leftarrow X_3$ and $X_2 \rightarrow X_5$.

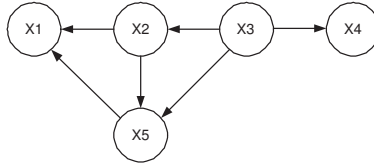


Fig. 4. The correct underlying CBN for the undirected structure in Figure 2.

This instantiation then triggers the inferration of other directions in the next phase of the algorithm, $X_3 \rightarrow X_5$ and $X_5 \rightarrow X_1$.

The only remaining undirected arc is $X_3 - X_4$, so an experiment on either one of these will lead to the final result. In practice the experiment with highest utility will be chosen.

Run with tree approximation In Figure 5 all MWSTs are given for each possible experiment, i.e. (a) is the tree for performing experiment at X_1 , (b) for X_2 , etc. Only arcs that were present in the original structure are used and all arcs connected to the variable being experimented on remain, as discussed previously.

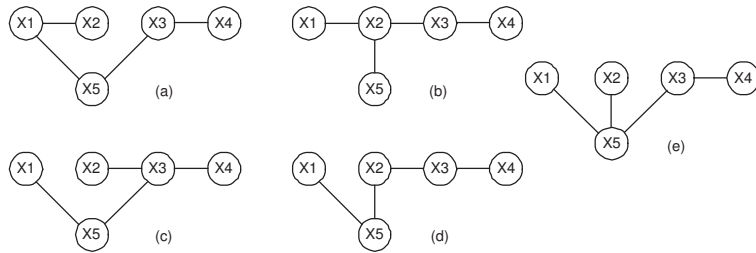


Fig. 5. All MWSTs for each possible experiment.

The optimal choice (again assuming all costs equal) is variable X_3 (since not all arcs can be directed inwards in the original graph, there are less instantiations without inferred edges in the tree), and this would result immediatly in the following instantiation: $X_3 \rightarrow X_2$, $X_3 \rightarrow X_4$, $X_3 \rightarrow X_5$.

This instantiation then triggers the inferration of $X_2 \rightarrow X_1$ and $X_5 \rightarrow X_1$.

The only remaining undirected arc now is $X_2 - X_5$, thus again an experiment on any of these nodes will solve the problem.

8 Conclusion and future work

We discussed a decision theoretic approach to learn causal Bayesian networks from a mixture of experiments and observational data.

We used the information of observational data to learn a completed partially directed graph and tried to discover the directions of the remaining edges by means of experiment. Our method allows the possibility to assign costs to each experiment and each measurement.

We demonstrated that our approach allows to learn a causal Bayesian network optimally with relation to a number of decision criteria. For the expected utility for a general structure we gave an approximation based on the solution for tree structures.

We introduced an algorithm that is adaptive, since it allows to actively add results of experiments. The algorithm is a general description and can be used with any decision criteria or utility function.

A first part of future work is a comparative study between the different decision criteria for learning CBN structure. We would also like to extend this approach to take into account that the CPDAG learned in the first phase is not correct and allow hidden variables.

Acknowledgments This work was partially funded by a IWT-scholarship. This work was partially supported by the IST Programme of the European Community, under the PASCAL network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. Pearl, J.: Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann (1988)
2. Pearl, J.: Causality: Models, Reasoning and Inference. MIT Press (2000)
3. Murphy, K.P.: Active learning of causal bayes net structure. Technical report, Department of Computer Science, UC Berkeley (2001)
4. Tong, S., Koller, D.: Active learning for structure in bayesian networks. In: Seventeenth International Joint Conference on Artificial Intelligence. (2001)
5. Cooper, G.F., Yoo, C.: Causal discovery from a mixture of experimental and observational data. In: Proceedings of Uncertainty in Artificial Intelligence. (1999) 116–125
6. Eberhardt, F., Glymour, C., Scheines, R.: N-1 experiments suffice to determine the causal relations among n variables. Technical report, Carnegie Mellon University (2005)
7. Eberhardt, F., Glymour, C., Scheines, R.: On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In: Uncertainty in Artificial Intelligence Conference (UAI). (2005)
8. Shipley, B.: Cause and Correlation in Biology. Cambridge University Press (2000)
9. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction and Search. MIT Press (2000)
10. Mani, S., Cooper, G.F.: Causal discovery using a bayesian local causal discovery algorithm. In: MEDINFO 2004, IOS Press (2004) 731–735