Bayesian Probabilities for Constraint-based Causal Discovery*

Tom Claassen and Tom Heskes
Radboud University Nijmegen
Netherlands
{T.Claassen,T.Heskes}@science.ru.nl

Abstract

We target the problem of accuracy and robustness in causal inference from finite data sets. Our aim is to combine the inherent robustness of the Bayesian approach with the theoretical strength and clarity of constraint-based methods. We use a Bayesian score to obtain probability estimates on the input statements used in a constraint-based procedure. These are subsequently processed in decreasing order of reliability, letting more reliable decisions take precedence in case of conflicts, until a single output model is obtained. Tests show that a basic implementation of the resulting Bayesian Constraint-based Causal Discovery (BCCD) algorithm already outperforms established procedures such as FCI and Conservative PC. It indicates which causal decisions in the output have high reliability and which do not. The approach is easily adapted to other application areas such as complex independence tests.

1 Introduction: Robust Causal Discovery

In real-world systems interactions between a set of variables $\mathbf{V}$ are often modeled in the form of a *causal DAG* (directed ayclic graph) $\mathcal{G}_C$. A *directed path* from A to B in such a graph $\mathcal{G}_C$ indicates a **causal relation** $A \Rightarrow B$ in the system, whereas an edge $A \rightarrow B$ in $\mathcal{G}_C$ indicates a *direct* causal link.

The *causal Markov* and *faithfulness* assumptions link the structure of the graph $\mathcal{G}_C$ to observed probabilistic in/dependencies through d-separation [Pearl, 2000], which forms the basis behind most existing causal discovery procedures. Together, they imply that the causal DAG $\mathcal{G}_C$ is also *minimal*, in the sense that no proper subgraph can satisfy both assumptions *and* produce the same probability distribution [Zhang and Spirtes, 2008].

If some of the variables in the causal DAG are hidden then the independence relations between the observed variables may be represented in the form of an *ancestral graph* (AG) [Richardson and Spirtes, 2002]; intuitively similar to a DAG except that it can also contain *bi-directed arcs* $X \leftrightarrow Y$ (unobserved common cause between X and Y) and undirected edges $X - - Y$ (selection effects, ignored in this article).

Different graphs can produce the same set of independencies: the *equivalence class* $[\mathcal{G}]$ is the set of all graphs that are indistinguishable from $\mathcal{G}$ in terms of implied independencies. The invariant features, common to all members of $[\mathcal{G}]$, can be represented in the form of a *partial ancestral graph* (PAG) $\mathcal{P}$: an ancestral graph where circle marks '$\circ$' on edges indicate 'unknown tail or arrowhead'. A complete PAG encodes all identifiable, present or absent causal relations; see e.g. [Zhang, 2008] for details on how to read PAGs.

With this in mind, the task of a causal discovery algorithm is to find as many invariant features of the equivalence class over observed variables from a given data set as possible.

Causal discovery paradigms

The so-called **constraint-based** algorithms search for conditional independencies $X \perp\!\!\!\perp Y \mid \mathbf{Z}$ in order to eliminate direct causal links $X - Y$ from the causal PAG $\mathcal{P}$. An efficient search strategy can uncover the entire skeleton, after which a number of orientation rules are executed to find invariant tails '$-$' and arrowheads '$>$' on edges. Members of this group include the IC-algorithm [Pearl and Verma, 1991], PC/FCI [Spirtes *et al.*, 2000], and many others; see e.g. [Glymour *et al.*, 2004; Kalisch *et al.*, 2011]. Of these, the FCI algorithm in conjunction with the orientation rules in [Zhang, 2008] is sound and complete in the large-sample limit when hidden common causes and/or selection bias may be present.

They tend to output a crisp and clear causal model. The downside is that for realistic, finite data sets they give little indication of which parts of the network are stable (reliable), and which are not: if unchecked, even one erroneous borderline independence decision may lead to multiple incorrect orientations [Spirtes, 2010]. To tackle this lack of robustness, Ramsey *et al.* [2006] proposed a conservative approach involving explicit validation of certain orientation rules. The resulting Conservative PC algorithm is indeed more robust, but also significantly less informative than vanilla PC.

The **score-based** algorithms build on the implied minimality of the causal graph. They define a scoring criterion $S(\mathcal{G}, \mathbf{D})$, often corresponding to a (Bayesian) likelihood, that measures how well a Bayesian network with structure $\mathcal{G}$ fits the observed data $\mathbf{D}$, while preferring simpler networks, with

*The paper on which this extended abstract is based was the recipient of the best paper award of the 2012 Conference on Uncertainty in Artificial Intelligence (UAI) [Claassen and Heskes, 2012].

fewer free parameters, over more complex ones. For DAG structures with either discrete or Gaussian variables closed form solutions exist that can be computed efficiently from data. These are employed in algorithms such as K2 [Cooper and Herskovits, 1992] and the Greedy Equivalence Search (GES) [Chickering, 2002] to search for an optimal structure.

Score-based procedures can output a range of high-scoring models. Multiple alternatives are arguably less straightforward to interpret, but it does allow for a measured interpretation of the reliability of inferred causal relations, and is less susceptible to incorrect categorical decisions [Heckerman *et al.*, 1999]. The main drawback is the need to rely on the *causal sufficiency* assumption (no latent confounders).

2 The Best of Both Worlds

The strength of constraint-based algorithms lies in the ability to handle data from arbitrary causal DAGs and turn it into clear, unambiguous causal output. The strength of Bayesian score-based approaches lies in the robustness and implicit confidence measure that a likelihood-weighted combination of multiple models can bring.

► Our idea is to improve on both methods by using a Bayesian approach to estimate the reliability of different constraints, and use this to decide if, when, and how that information should be used.

Instead of classifying pieces of information as 'true' or not, we want to rank and process independence constraints according to a principled confidence measure, and build up a global causal model starting from the most reliable information down to a discretionary minimum confidence level.

For that we use a recently developed method to break up the causal inference process into a series of modular steps that can be executed in arbitrary order. It works by translating observed in/dependence constraints into basic **logical causal statements** L via:

1. $X \perp\!\!\!\perp Y \,|\, [\mathbf{Z} \cup Z] \;\vdash\; (Z \Rightarrow X) \vee (Z \Rightarrow Y)$,
2. $X \not\perp\!\!\!\perp Y \,|\, \mathbf{Z} \cup [Z] \;\vdash\; (Z \not\Rightarrow X) \wedge (Z \not\Rightarrow Y) \wedge (Z \not\Rightarrow \mathbf{Z})$,

where square brackets indicate a *minimal* set of nodes.

The logical causal statement $L : (Z \Rightarrow X) \vee (Z \Rightarrow Y)$ states that there is either a causal relation from Z to X, or from Z to Y, or both. Similarly, the logical causal statement $L : (Z \not\Rightarrow X)$ states that there is *no* directed path from Z to X in the underlying causal DAG $\mathcal{G}_C$; see [Claassen and Heskes, 2011] for details. Subsequent statements follow from straightforward deduction on the causal properties *transitivity* and *acyclicity*.

Crucial in this connection is that these logical causal statements are implied directly by the structure of the underlying causal graph $\mathcal{G}_C$, and so in turn by the induced ancestral graph over the observed variables. As a result, we can obtain **probability estimates** on logical causal statements $L \in \mathcal{L}$ by summing the normalized posterior likelihoods of all structures $\mathcal{G}$ that entail L through d-separation:

$$p(L|\mathbf{D}) \propto \sum_{\mathcal{G} \in \mathbf{G}(\vdash L)} p(\mathbf{D}|\mathcal{G}) p(\mathcal{G}), \qquad (1)$$

For the likelihood estimates $p(\mathbf{D}|\mathcal{G})$ on possible DAG structures we employ the well-known *Bayesian Dirichlet* (BD) metric for discrete variables [Heckerman *et al.*, 1995].

$$p(\mathbf{D}|\mathcal{G}) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ijk})}, \quad (2)$$

with n the number of variables, r_i the multiplicity of variable X_i, q_i the number of possible instantiations of the parents of X_i in $\mathcal{G}$, N_{ijk} the number of cases in data set $\mathbf{D}$ in which variable X_i has the value $r_{i(k)}$ while its parents are instantiated as $q_{i(j)}$, and with $N'_{ij} = \sum_{k=1}^{r_i} N'_{ijk}$ pseudocounts for a Dirichlet prior over the free parameters.

For the prior over structures $p(\mathcal{G})$ we can opt for a uniform distribution over all possible graphs, or even include additional background information. To obtain proper probability estimates we still need to normalize eq.(1) by dividing through the sum of all contributions from all graphs. It is well-known that the number of possible graphs increases super-exponentially with the number of nodes. However, eq.(1) equally applies to graphs over arbitrary subsets $\mathbf{X} \subset \mathbf{V}$ of the observed variables. For sparse graphs we can limit the search to (small) subsets of max. size $K \ll |\mathbf{V}|$ in order to keep the computations feasible, without losing much information on the possible logical causal statements that can be inferred. It implies that we can employ an efficient search strategy over **increasing subsets** $\mathbf{X}$ of variables, similar to PC/FCI. It does mean that there are now multiple ways (different subsets of variables) to obtain a given logical causal statement L, but it can be shown that it is sufficient to only keep track of the *maximum* probability estimates obtained so far. The resulting $p(L|\mathbf{D}_{(\mathbf{X})})$ form a conservative estimate of the theoretical optimal probability that could be obtained from the entire data set $\mathbf{D}$.

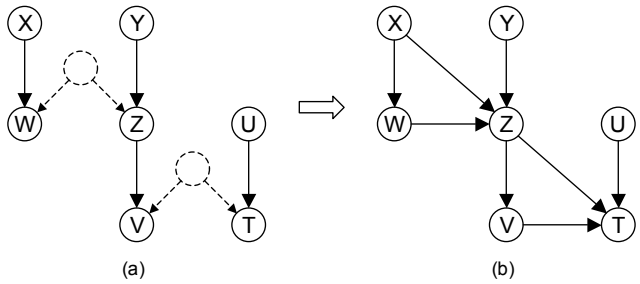

Figure 1: (a) causal DAG with hidden nodes, (b) minimal uDAG over observed variables

When considering graphs over subsets of variables, we still need to account for the fact that the minimal DAG over an arbitrary subset $\mathbf{X} \subset \mathbf{V}$ may be **unfaithful** (a 'uDAG') to the underlying causal structure. For example in Figure 1(a) the corresponding ancestral graph contains invariant bi-directed edges $W \leftrightarrow Z$ and $V \leftrightarrow T$ from hidden common causes that cannot be accommodated in a DAG. As a result, apparent direct causal links such as $X \rightarrow Z$ and $V \rightarrow T$ appear in the minimal uDAG in (b).

To avoid drawing incorrect conclusions we need to rely on a modified d-separation inference rule:

Lemma 1. Let $\mathcal{G}$ be a uDAG for some distribution $p(\mathbf{X})$. Let $\mathcal{G}_{X\|Y}$ be the graph obtained by eliminating the edge $X - Y$ from $\mathcal{G}$ (if present). Then, if $X \perp\!\!\!\perp_{\mathcal{G}_{X\|Y}} Y \mid \mathbf{Z}$ then:

$$(X \perp\!\!\!\perp_{\mathcal{G}} Y \mid \mathbf{Z}) \Leftrightarrow (X \perp\!\!\!\perp_{p} Y \mid \mathbf{Z}).$$

In words: independence from d-separation remains valid, but the identifiable dependencies are restricted. The rule can be extended to *indirect* dependencies in a uDAG $\mathcal{G}$ by showing that if $\pi = \langle X, .., Y \rangle$ is the *only* unblocked path from X to Y given $\mathbf{Z}$ in $\mathcal{G}$, then $X \not\perp\!\!\!\perp_{p} Y \mid \mathbf{Z}$; see [Bouckaert, 1995; Claassen and Heskes, 2012] for details. From this we build a (pre-computed) **mapping** $\mathcal{G} \rightarrow \mathcal{L}$ from each possibly unfaithful uDAG $\mathcal{G}$ to all valid logical causal statements $\mathcal{L}$.

Finally, as we now consider graphs over subsets of different sizes, it becomes essential to ensure that we also have a **consistent prior** $p(\mathcal{G})$ for structures of different sizes. Perhaps surprisingly, this is *not* obtained by applying the same strategy at different levels: a uniform distribution over DAGs over $\{X, Y, Z\}$ implies $p(\text{“}X \perp\!\!\!\perp Y\text{”}) = 6/25$, whereas a uniform distribution over two-node DAGs implies $p(\text{“}X \perp\!\!\!\perp Y\text{”}) = 1/3$. We obtain a consistent multi-level prior by starting from a preselected level K, and then extend to different sized structures through marginalization.

3 Implementation and results

We can now turn the results from the previous section into a working algorithm, called the **Bayesian Constraint-based Causal Discovery** (BCCD) algorithm, depicted below.

It starts from a data set $\mathbf{D}$ and available background information $\mathcal{I}$, and outputs a matrix of identified causal relations $\mathbf{M}_C$, indicating explicitly for each pair of variables whether there is a causal relation $X \Rightarrow Y$, absence of causal relation $X \not\Rightarrow Y$, or ‘unknown’. It also produces a graphical causal model in the form of a PAG $\mathcal{P}$.

A crucial step in the algorithm is the mapping $\mathcal{G} \times \mathcal{L}$ from (possibly unfaithful) DAG structures to logical causal statements in line 9, which converts posterior likelihoods for structures into probability estimates for causal relations. This mapping is the same for each run, so it can be precomputed once from the rules such as in Lemma 1, and stored for use afterwards (l.1) The uDAGs $\mathcal{G}$ are represented as adjacency matrices. For speed and efficiency purposes, we choose to limit the structures to size $K \leq 5$, which gives a list of 29,281 uDAGs at the highest level. For details about representation and rules, see [Claassen and Heskes, 2012].

The adjacency search (l.3-15), loops over subsets from neigbouring nodes, looking for identifiable causal information, while keeping track of adjacencies that can be eliminated (l.11). As the set $\mathbf{W} = \{X, Y\} \cup \mathbf{Z}$ can be encountered in different ways, line (7) checks if the test on that set has been performed already. A list of probability estimates $p(L|\mathbf{D})$ for each logical causal statement is built up (l.10), until no more information is found.

The inference stage (l.16-21) then processes the list $\mathcal{L}$ in decreasing order of reliability, until the threshold is reached.

Algorithm 1 Bayesian Constraint-based Causal Discovery

In : database $\mathbf{D}$ over variables $\mathbf{V}$, backgr.info $\mathcal{I}$
Out: causal relations matrix $\mathbf{M}_C$, causal PAG $\mathcal{P}$
Stage 0 - Mapping
1: $\mathcal{G} \times \mathcal{L} \leftarrow Get\_uDAG\_Mapping(\mathbf{V}, K_{max} = 5)$
2: $p(\mathcal{G}) \leftarrow Get\_Prior(\mathcal{I})$
Stage 1 - Search
3: fully connected $\mathcal{P}$, empty list $\mathcal{L}$, $K = 0$, $\theta = 0.5$
4: **while** $K \leq K_{max}$ **do**
5: **for all** $X \in \mathbf{V}, Y \in Adj(X)$ in $\mathcal{P}$ **do**
6: **for all** $\mathbf{Z} \subseteq Adj(X)_{\setminus Y}$, $|\mathbf{Z}| = K$ **do**
7: $\mathbf{W} \leftarrow Check\_Unprocessed(X, Y, \mathbf{Z})$
8: $\forall \mathcal{G} \in \mathcal{G}_{\mathbf{W}}$: compute $p(\mathcal{G}|\mathbf{D}_{\mathbf{W}})$
9: $\forall L : p(L_{\mathbf{W}}|\mathbf{D}_{\mathbf{W}}) \leftarrow \sum_{\mathcal{G} \rightarrow L_{\mathbf{W}}} p(\mathcal{G}|\mathbf{D}_{\mathbf{W}})$
10: $\forall L : p(L) \leftarrow \max(p(L), p(L_{\mathbf{W}}|\mathbf{D}_{\mathbf{W}}))$
11: $\mathcal{P} \leftarrow p(\text{“}W_i \not\leftrightarrow W_j\text{”}|\mathbf{D}_{\mathbf{W}}) > \theta$
12: **end for**
13: **end for**
14: $K = K + 1$
15: **end while**
Stage 2 - Inference
16: $\mathbf{L}_C =$ empty 3D-matrix size $|\mathbf{V}|^3$, $i = 1$
17: Ł $\leftarrow Sort\_Descending\,(\text{Ł}, p(L))$
18: **while** $p(L_i) > \theta$ **do**
19: $\mathbf{L}_C \leftarrow Run\_Causal\_Logic(\mathbf{L}_C, L_i)$
20: $i \leftarrow i + 1$
21: **end while**
22: $\mathbf{M}_C \leftarrow Get\_Causal\_Matrix(\mathbf{L}_C)$
23: $\mathcal{P} \leftarrow Map\_To\_PAG(\mathcal{P}, \mathbf{M}_C)$

Statements in $\mathcal{L}$ are added one-by-one to the matrix of logical causal statements $\mathbf{L}_C$ (encoding identical to $\mathcal{L}$, see [Claassen and Heskes, 2011]), with additional information inferred from the causal logic rules. Basic conflict resolution is achieved by not overriding information already derived from more reliable statements. The final step (l.22,23) retrieves all explicit causal relations in the form of a causal matrix $\mathbf{M}_C$, and maps this onto the skeleton $\mathcal{P}$ obtained from Stage 1 to return a graphical PAG representation.

4 Experimental Evaluation

We have tested various aspects of the BCCD algorithm in many different circumstances, and against various other methods. The principal aim in this paper is to verify the viability of the Bayesian approach. We compare our results and that of other methods from data against known ground-truth causal models. For that, we generate random causal graphs with certain predefined properties (adapted from Melancon *et al.* [2000]; Chung and Lu [2002]), generate random data from this model, and marginalize out one or more hidden confounders. We looked at the impact of the number of data points, size of the models, sparseness, choices for parameter settings etc. on the performance to get a good feel for expected strengths and weaknesses in real-world situations.

It is well-known that the relative performance of different causal discovery methods can depend strongly on the performance metric and/or specific test problems used in the eval-

uation. Therefore, we will not claim that our method is inherently better than others based on the experimental results below, but simply note that the fact that in nearly all test cases the BCCD algorithm performed as good or better than other methods, is a clear indication of its viability and potential.

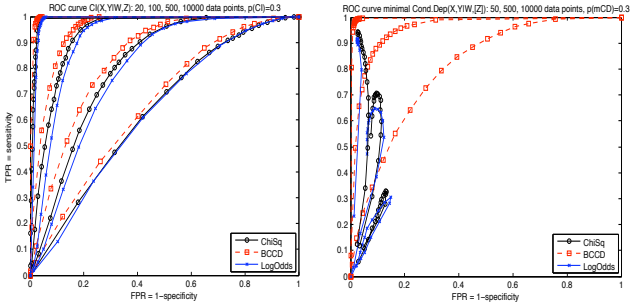

Figure 2: BCCD approach to (complex) independence test; (a) conditional independence $X \perp\!\!\!\perp Y \mid W, Z$, (b) *minimal* conditional dependence $X \not\perp\!\!\!\perp Y \mid W \cup [Z]$

First we implemented the BCCD approach as a simple independence test through a modified mapping $\mathcal{G} \rightarrow \mathbf{I}$ of structures $\mathcal{G}$ to implied in/dependence statements $\mathbf{I}$. Figure 2 shows a typical example in the form of ROC-curves for different sized data sets, compared against a chi-squared test and a Bayesian log-odds test from [Margaritis and Bromberg, 2009], with the prior on independence as the tuning parameter for BCCD. For 'regular' conditional independence there was no significant difference (BCCD slightly ahead, more as conditioning set increases). But for minimal independencies other methods reject for both high and low decision thresholds, resulting in the looped curves in (b); BCCD has no such problem.

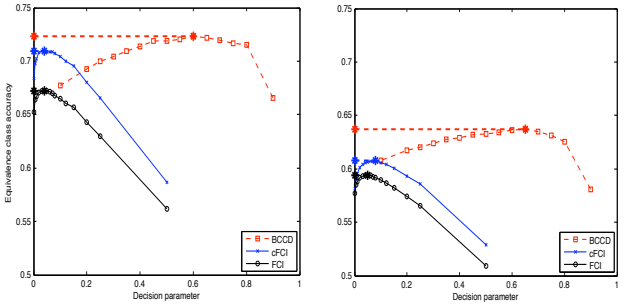

Figure 3: Equivalence class accuracy (% of edge marks in PAG) vs. decision parameter; for BCCD and (conservative) FCI, from 1000 random models; (a) 6 observed nodes, 1-2 hidden, 1000 points, (b) idem, 12 observed nodes

Figure 3 shows typical results for the BCCD algorithm itself: for a data set of 1000 records the *PAG accuracy* for both FCI and conservative FCI peaks around a threshold $\alpha \approx 0.05$ - lower for more records, higher for less - with conservative FCI consistently outperforming standard FCI. The BCCD algorithm peaks at a cut-off value $\theta \in [0.5, 0.7]$ with an accuracy that is slightly higher than the maximum for conservative FCI. The PAG accuracy tends not to vary much over this interval, making the default choice $\theta = 0.5$ fairly safe, even though the number of invariant edge marks does increase significantly (more decisions).

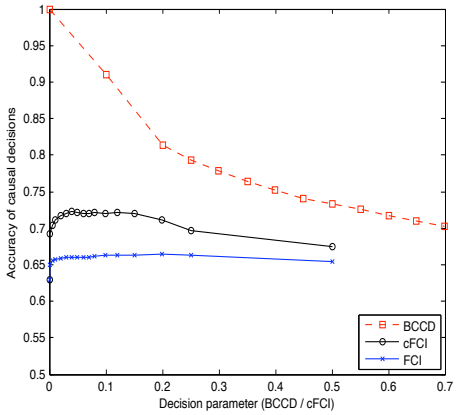

Figure 4: Accuracy of causal decisions as a function of the decision parameter

Figure 4 depicts the *causal accuracy* as a function of the tuning parameter for the three methods. The BCCD dependency is set against $(1-\theta)$ so that going from $0 \rightarrow 1$ matches processing the list of statements in decreasing order of reliability. As hoped/expected: changing the decision parameter θ allows to access a range of accuracies, from a few very reliable causal relations to more but less certain indications. In contrast, the accuracy of the two FCI algorithms cannot be tuned effectively through the decision parameter α. The reason behind this is apparent from Figure 2(b): changing the decision threshold in an independence test shifts the balance between dependence and independence decisions, but it cannot identify or alter the balance in favor of more reliable decisions. We consider the fact that the BCCD *can* do exactly that as the most promising aspect of the Bayesian approach.

5 Discussion

The experimental results confirm that the Bayesian approach is both viable and promising: even in a basic implementation the BCCD algorithm already outperforms other state-of-the-art causal discovery algorithms. It yields slightly better accuracy, and comes with an easy tuning parameter that can be used to vary from making just a few but very reliable causal statements to many possibly less certain decisions.

An interesting question is how far off from the theoretical optimum we are: at the moment it is not clear whether we are fighting for the last few percent or if sizeable gains can still be made. Obvious improvements include scoring equivalence classes (instead of all DAGs), handle larger substructures through sampling, and to score ADMGs directly [Silva and Ghahramani, 2009; Evans and Richardson, 2010]. Finally we want to extend the method to handle continuous/mixed data without the need for discretization.

References

R. Bouckaert. *Bayesian Belief Networks: From Construction to Inference*. PhD thesis, University of Utrecht, 1995.

D. Chickering. Optimal structure identification with greedy search. *Journal of Machine Learning Research*, 3(3):507–554, 2002.

F. Chung and L. Lu. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics*, 6(2):125–145, 2002.

T. Claassen and T. Heskes. A logical characterization of constraint-based causal discovery. In *Proc. of the 27th Conference on Uncertainty in Artificial Intelligence*, 2011.

T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *Proc. of the 28th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 207 – 216, 2012.

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

R. Evans and T. Richardson. Maximum likelihood fitting of acyclic directed mixed graphs to binary data. In *Proc. of the 26th Conference on Uncertainty in Artificial Intelligence*, 2010.

C. Glymour, R. Scheines, P. Spirtes, and J. Ramsey. The TETRAD project: Causal models and statistical data. www.phil.cmu.edu/projects/tetrad/current, 2004.

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

D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. In *Computation, Causation, and Discovery*, pages 141–166. 1999.

M. Kalisch, M. Mächler, D. Colombo, M. Maathuis, and P. Bühlmann. Causal inference using graphical models with the R package pcalg. http://cran.r-project.org/web/packages/ pcalg/vignettes/pcalgDoc.pdf, 2011.

D. Margaritis and F. Bromberg. Efficient Markov network discovery using particle filters. *Computational Intelligence*, 25(4):367–394, 2009.

G. Melancon, I. Dutour, and M. Bousquet-M'elou. Random generation of DAGs for graph drawing. Technical Report INS-R0005, Centre for Mathematics and Computer Sciences, 2000.

J. Pearl and T. Verma. A theory of inferred causation. In *Knowledge Representation and Reasoning: Proc. of the Second Int. Conf.*, pages 441–452, 1991.

J. Pearl. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

J. Ramsey, J. Zhang, and P. Spirtes. Adjacency-faithfulness and conservative causal inference. In *Proc. of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 401–408, 2006.

T. Richardson and P. Spirtes. Ancestral graph Markov models. *Ann. Stat.*, 30(4):962–1030, 2002.

R. Silva and Z. Ghahramani. The hidden life of latent variables: Bayesian learning with mixed graph models. *Journal of Machine Learning Research*, 10:1187–1238, 2009.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. The MIT Press, Cambridge, Massachusetts, 2nd edition, 2000.

P. Spirtes. Introduction to causal inference. *Journal of Machine Learning Research*, 11:1643–1662, 2010.

J. Zhang and P. Spirtes. Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 2(18):239–271, 2008.

J. Zhang. On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873 – 1896, 2008.