JMLR: Workshop and Conference Proceedings vol 52, 487-498, 2016 PGM 2016

# Computing Lower and Upper Bounds on the Probability of Causal Statements

**Elena Sokolova** ESOKOLOVA@CS.RU.NL
**Martine Hoogman** MARTINE.HOOGMAN@RADBOUDUMC.NL
**Perry Groot** PERRY.GROOT@SCIENCE.RU.NL
**Tom Claassen** TOMC@CS.RU.NL
**Tom Heskes** T.HESKES@SCIENCE.RU.NL
*Radboud University, Nijmegen, The Netherlands*

## Abstract

Causal discovery provides an opportunity to infer causal relationships from purely observational data and to predict the effect of interventions. Constraint-based methods for causal discovery exploit conditional (in)dependencies to infer the direction of causal relationships. They typically work through forward chaining: given some causal statements, others can be inferred by applying relatively straightforward causal logic such as transitivity and acyclicity. Starting from the premise that we can estimate reliabilities for base causal statements, we propose a novel approach to estimate the reliability of novel statements inferred by forward chaining. Since reliabilities for base statements are clearly dependent, if only because inferred from the same data, exact computation is infeasible. However, lending ideas from the area of imprecise probability theory, we can compute bounds on the reliabilities on inferred statements. Specifically, we make use of the good old Fréchet inequalities and discuss two different variants: greedy and delayed. In simulation experiments, we show that the delayed variant, at the expense of more bookkeeping and computation time, does provide slightly tighter intervals. We illustrate our method on a real-world data set about attention deficit/hyperactivity disorder.

**Keywords:** Causal discovery; Fréchet inequalities; constraint-based methods.

## 1. Introduction

The use of causal discovery algorithms has become increasingly popular in recent years. Causal discovery algorithms are able to predict the result of an intervention under some reasonable assumptions, purely from observational data. Causal relationships between variables are typically represented through a causal directed acyclic graph (DAG), where a directed path from variable $X$ to $Y$ represents a causal path from $X$ to $Y$.

One of the most common approaches to learn a causal DAG from data is the so-called constraint-based approach. This approach employs a statistical independence test and typically consists of two steps. In the first step, the skeleton of the graph is learned based on conditional independencies inferred from the data. In the second step, edges of the graph are oriented. This orientation is based on the presence of so-called V-structures (or colliders). A V-structure is a triple $(X, Y, Z)$, where $Z$ is a common child of $X$ and $Y$, variables $X$ and $Y$ are independent, but they become dependent when we condition on variable $Z$. Based on the V-structures, causal statements are inferred, which determine the orientation of the edges. By combining causal statements, new statements can be inferred to orient other edges in the graph. Examples of popular constraint-based approaches are IC (Pearl and Verma, 1991) and PC/FCI (Spirtes et al., 2000).

If a causal statement is estimated incorrectly, this mistake may propagate through the whole graph, leading to many erroneous orientations. A potential remedy, suggested by, e.g., Triantafilou et al. (2014); Claassen and Heskes (2012a), is to keep track of the reliability of the inferred causal statements. Standard constraint-based methods consecutively apply hypothesis tests for conditional independence that provide a p-value as output. Triantafilou et al. (2014) propose to translate these p-values into probabilities, following the generic approach of Sellke et al. (2001), and use these probabilities to estimate the reliability of the skeleton. Claassen and Heskes (2012a) estimate the reliability of conditional independence statements by computing Bayesian scores on DAGs over subsets of variables, and combine these into reliabilities of both edges and orientations.

In constraint-based methods causal statements are typically inferred using forward chaining: consecutively combining earlier inferred statements. The causal statements themselves are clearly dependent, since inferred from the same data, which makes combination of their reliabilities highly nontrivial. In this paper we propose to apply ideas from the theory of imprecise probabilities (Walley, 1991) or interval probabilities (Weichselberger, 2000) to keep track of reliability intervals instead of point estimates. We develop a new method that estimates these probability intervals using the so-called Fréchet inequalities (Fréchet, 1935) and prove that a particular form of the logical statement gives the best lower and upper bound using Fréchet inequalities. Although in this paper we focus on Bayesian Constraint-based Causal Discovery (BCCD) (Claassen and Heskes, 2012a), in principle our method can be applied to any constraint-based causal discovery algorithm that infers causal statements. We also propose an approximation of our algorithm that provides similar accuracy but requires lower computational complexity.

The rest of this paper is organized as follows. In Section 2, we give the background and derive our methods for computing reliability intervals. In Section 3, we apply these methods to simulated and real-world data. In Section 4 we provide our conclusions and discuss future work.

## 2. A Method for Computing Causal Reliability Intervals

### 2.1 Constrained-Based Causal Discovery

In order to infer a skeleton and causal statements from data, we use the BCCD algorithm (Claassen and Heskes, 2012a). BCCD is a state-of-the-art algorithm for estimating causal relationships between variables that also provides a reliability measure for inferred statements and can handle both potential confounding (i.e., does not assume causal sufficiency) and selection bias. By selection bias we mean a process of data selection that introduces dependencies between variables that are not representative of the population. Confounding refers to an unobserved common cause between several variables. The output of BCCD is a so-called partial ancestral graph (PAG) (Richardson and Spirtes, 2003) that is used to represent DAGs with latent variables. Due to space limitation we provide only a short description of BCCD, a more detailed description can be found in (Claassen and Heskes, 2012a).

The BCCD algorithm consists of two main stages.
1. **Inference of the skeleton and base causal statements.** BCCD considers subsets of maximum $K$ variables. For each subset, it computes a Bayesian score (Dawid, 1984) for every directed acyclic

graph. Each directed acyclic graph implies particular causal statements of the form

$$\begin{array}{rl} \text{no collider:} & (Z \Rightarrow X) \vee (Z \Rightarrow Y) \vee (Z \Rightarrow \mathbf{S}) \\ \text{no causal path:} & (Z \not\Rightarrow X) \wedge (Z \not\Rightarrow \mathbf{S}) \\ \text{causal path:} & (Z \Rightarrow X) \wedge (Z \not\Rightarrow \mathbf{S}) \end{array} \quad (1)$$

where statement $(Z \Rightarrow \mathbf{S})$ indicates a selection bias $\mathbf{S}$ on variable $Z$. The first line follows from a so-called minimal conditional independence: if conditioning upon a variable $Z$ breaks the conditional dependence between two variables $X$ and $Y$, $Z$ must have a causal path to $X$, $Y$, or both (more details can be found in (Claassen and Heskes, 2012b)). The second line follows from a minimal conditional dependence: if adding $Z$ to the conditioning set makes a variable $X$ dependent of another variable, say $Y$, there cannot be a causal path from $Z$ to $X$ (nor from $Z$ to $Y$) and there can be no selection bias ($\mathbf{S}$) on $Z$. This corresponds to the V-structure mentioned before. The first line is in a sense the negation of the second line: it states that $Z$ must be on a path between $X$ and $Y$, but cannot lead to a V-structure, so cannot be a collider on this path. BCCD infers the reliability for each causal statement by combining the Bayesian scores for all DAGs that match this statement. This reliability gives a conservative estimate of the probability of a causal relation (Claassen and Heskes, 2012a). We interpret it as an estimate for the probability that the causal statement is true.
2. **Combination of causal statements.** In the second stage, BCCD infers new causal statements by combining causal statements and applying rules from standard causal logic:

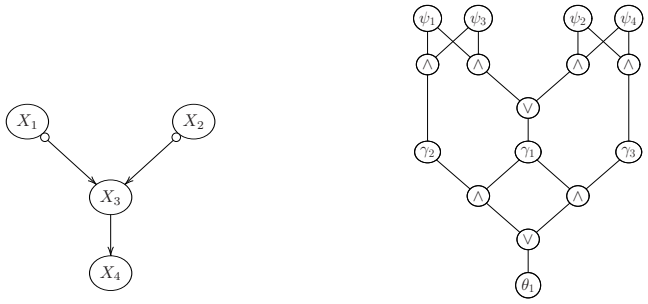
$$\begin{array}{rl} \text{irreflexive:} & (X \Rightarrow X) \vdash \text{false} \\ \text{acyclic:} & (X \Rightarrow Y) \vdash (Y \not\Rightarrow X) \\ \text{transitive:} & (X \Rightarrow Y) \wedge (Y \Rightarrow Z) \vdash (X \Rightarrow Z) \end{array} \quad (2)$$

The system of causal statements is closed, in the sense that all newly inferred causal statements can also be written in the form (1), with special cases $(Z \Rightarrow X) \vee (Z \Rightarrow \mathbf{S})$ (causal path or selection bias; first line, when $Y$ equals $X$) and $(Z \not\Rightarrow \mathbf{S})$ (no selection bias; second line, when $X$ equals $Z$).

The output of the BCCD algorithm is a list of statements of the form $(X \Rightarrow Y), (Y \not\Rightarrow X)$, or $(X \Rightarrow S)$. Given a skeleton, these statements can be used to determine the directions of edges, e.g., a combination of two statements $(X \Rightarrow Y), (Y \not\Rightarrow X)$ suggests a causal effect of $X$ on $Y$ that is represented as $X \rightarrow Y$ in a PAG. Statements $(X \Rightarrow Y)$, and $(Y \Rightarrow X)$ indicate a selection bias between $X$ and $Y$ and is represented with $X - Y$ in a PAG. If a list of statements contains $(X \not\Rightarrow Y), (Y \not\Rightarrow X)$ then there is a latent confounding between two variables that is represented as $X \leftrightarrow Y$ in a PAG. Circle marks are used to mark edges which directions cannot be fully determined, e.g., if only statement $(X \Rightarrow Y)$ was inferred, it is either a causal effect $X \rightarrow Y$ or a selection bias $X - Y$, which will be represented as $X -\!\circ Y$ in a PAG.

The parameter $K$, which determines the subsets of variables considered by BCCD to infer causal statements, plays an important role in the BCCD algorithm. The higher $K$, the more causal statements can be inferred directly, without the need to combine statements using causal logic. On the other hand the complexity of the algorithm grows exponentially with $K$. For example the number of possible causal models for which likelihoods should be estimated is 25 for $K = 3$ and 29.281 for $K = 5$. The default value of $K$ is five variables, a fine compromise between complexity and accuracy. In this paper, we will also consider $K = 3$, to better demonstrate the effect of different strategies for combining causal statements.

**Example 1** *As a running example, we will consider a so-called Y-structure (Mani et al., 2012). This structure, sketched in Figure 1(a), consists of four variables, where $X_1$, $X_2$, and $X_3$ form a*




(a) (b)

Figure 1: (a) Example of a Y-structure. (b) Different levels of representing the inference of the statement $\theta$ that encodes an arrow between variable $X_4$ and variable $X_3$ in Figure 1(a).

*V-structure and $X_4$ is a child of $X_3$. Given enough data generated from such a Y-structure, BCCD with $K = 3$ would infer the base causal statements.*

$$\psi_1 : (X_3 \not\Rightarrow X_1) \wedge (X_3 \not\Rightarrow \mathbf{S})$$
$$\psi_2 : (X_3 \not\Rightarrow X_2) \wedge (X_3 \not\Rightarrow \mathbf{S})$$
$$\psi_3 : (X_3 \Rightarrow X_1) \vee (X_3 \Rightarrow X_4) \vee (X_3 \Rightarrow \mathbf{S})$$
$$\psi_4 : (X_3 \Rightarrow X_2) \vee (X_3 \Rightarrow X_4) \vee (X_3 \Rightarrow \mathbf{S})$$

*$\psi_1$ and $\psi_2$ follow from the minimal conditional dependence between $X_1$ and $X_2$ given $X_3$, $\psi_3/\psi_4$ from the minimal conditional independencies between $X_1/X_2$ and $X_4$ given $X_3$. Applying the causal rules (2), we can infer various new statements:*

$$(\psi_1 \wedge \psi_3) \vdash \gamma_1 \;\; \textit{with} \;\; \gamma_1 : (X_3 \Rightarrow X_4) \wedge (X_3 \not\Rightarrow \mathbf{S})$$
$$(\psi_2 \wedge \psi_4) \vdash \gamma_1$$
$$(\psi_1 \wedge \psi_3) \vdash \gamma_2 \;\; \textit{with} \;\; \gamma_2 : (X_4 \not\Rightarrow X_1) \wedge (X_4 \not\Rightarrow \mathbf{S})$$
$$(\psi_2 \wedge \psi_4) \vdash \gamma_3 \;\; \textit{with} \;\; \gamma_3 : (X_4 \not\Rightarrow X_2) \wedge (X_4 \not\Rightarrow \mathbf{S})$$

*The derivation of $\gamma_1$ is relatively straightforward; $\gamma_2$ and $\gamma_3$ are most easily proven by contradiction. Note further that two combinations here lead to the same statement. Given these new statements $\gamma_1$ through $\gamma_3$, we can then also infer*

$$\gamma_1 \wedge \gamma_2 \vdash \theta_1 \;\; \textit{with} \;\; \theta_1 : (X_4 \not\Rightarrow X_3) \wedge (X_4 \not\Rightarrow \mathbf{S})$$
$$\gamma_1 \wedge \gamma_3 \vdash \theta_1$$

*Figure 1(b) gives the logic tree for deriving the causal statement $\theta_1$. The question we would like to answer is: given probabilities of the base causal statements $\psi_1$ through $\psi_4$, what we can say about the probability of the inferred statement $\theta_1$? Note that in this example, BCCD with $K = 5$ would already give $\theta_1$ as a base causal statement with a corresponding reliability.*

### 2.2 Estimation of Probability Intervals

The native version of BCCD estimates reliabilities from inferred statements by taking the product when statements are combined with an AND (as if the underlying statements are independent) and

taking the maximum when statements are combined with an OR (giving a conservative estimate). Empirically, this appears to work fine in practice. Here, we propose to give up on estimating reliabilities on inferred statements, but instead derive an algorithm to compute reliability intervals making use of the well-known Fréchet inequalities (Fréchet, 1935).

As should be clear from Figure 1(b), newly inferred statements are derived from a (potentially complicated) mixture of conjunctions (ANDs) and disjunctions (ORs) of the base statements. Suppose we have the conjunction $\psi_{\text{conjunct}} \dashv \psi_1 \wedge \psi_2 \wedge \ldots \wedge \psi_n$, then the Fréchet inequalities give

$$\max\left(0, \sum_{i=1}^{n} P(\psi_i) - (n-1)\right) \leq P(\psi_{\text{conjunct}}) \leq \min_i P(\psi_i)\,,$$

with $P(\psi_i)$ the probability of the causal statement $\psi_i$. Similarly, applying Fréchet inequalities to a disjunction $\psi_{\text{disjunct}} \dashv \psi_1 \vee \psi_2 \vee \ldots \vee \psi_n$ gives

$$\max_i P(\psi_i) \leq P(\psi_{\text{disjunct}}) \leq \min\left(1, \sum_{i=1}^{n} P(\psi_i)\right)\,.$$

Using these inequalities, we can keep track of lower bounds and upper bounds on the probability of inferred causal statements, indicated by $\underline{P}$ and $\bar{P}$, respectively. We will use shorthand notation $I(\psi) = [\underline{P}(\psi), \bar{P}(\psi)]$ to refer to the probability interval for causal statement $\psi$. We will use lower case Greek letters $\psi$, $\gamma$, $\theta$ to refer to causal statements and upper case Greek letters $\Psi$, $\Gamma$, $\Theta$ for the corresponding formulae that have been used to derive the causal statements. We will write both $\bar{P}(\Phi)$ and $\bar{P}(\phi)$, the interpretation of which should be clear from the context.

Specifically, we have the following rules for combining intervals of two causal statements:

$$\begin{array}{ll} \underline{P}(\psi_1 \wedge \psi_2) = \max\left(0, \underline{P}(\psi_1) + \underline{P}(\psi_2) - 1\right), & \bar{P}(\psi_1 \wedge \psi_2) = \min\left(\bar{P}(\psi_1), \bar{P}(\psi_2)\right) \\ \underline{P}(\psi_1 \vee \psi_2) = \max\left(\underline{P}(\psi_1), \underline{P}(\psi_2)\right), & \bar{P}(\psi_1 \vee \psi_2) = \min\left(1, \bar{P}(\psi_1) + \bar{P}(\psi_2)\right) \end{array} \tag{3}$$

We now propose two different algorithms: *greedy* and *delayed*. When two causal statements are combined to derive a new one, the greedy algorithm immediately applies the rules (3) to compute a new interval for the newly inferred statement. The delayed algorithm, on the other hand, delays the computation of the intervals as much as possible. It keeps track of the propositional formula that led to the causal statement of interest, attempts to simplify it, and only then computes the interval.

**Example 2** *We consider the logic tree of Figure 1(b). Suppose that the probabilities of the base causal statements are* $P(\psi_1) = 0.8$, $P(\psi_2) = 0.85$, $P(\psi_3) = 0.9$, *and* $P(\psi_4) = 0.95$. *The greedy algorithm sequentially applies (3) to yield* $I(\gamma_2) = [\max(0, P(\psi_1) + P(\psi_3) - 1), \min(P(\psi_1), P(\psi_3))] = [\max(0, 0.8 + 0.9 - 1), \min(0.8, 0.9)] = [0.7, 0.8]$, $I(\gamma_3) = [0.8, 0.85]$, $I(\gamma_1) = [0.8, 1]$, $I(\gamma_1 \wedge \gamma_2) = [0.5, 0.8]$, $I(\gamma_1 \wedge \gamma_3) = [0.6, 0.85]$, *and, finally,* $I(\theta_1) = [0.6, 1]$. *The delayed algorithm, on the other hand, first expresses each of the inferred causal statements as a formula in terms of the base statements* $\psi_1$ *through* $\psi_4$. *For* $\gamma_1$ *and* $\theta_1$, *we arrive at*

$$\theta_1 \dashv \Theta_1, \gamma_1 \dashv \Gamma_1 \;\; \textit{with} \;\; \Theta_1 \equiv \Gamma_1 = (\psi_1 \wedge \psi_3) \vee (\psi_2 \wedge \psi_4)\,. \tag{4}$$

*Now applying the Fréchet inequalities, first on the conjunctions and then on the disjunction, we obtain* $I(\theta_1) = I(\gamma_1) = [0.8, 1]$*: a tighter bound than for the greedy algorithm.*

A causal statement combining more than two base causal statements can have various equivalent formulae, each suggesting a different ordering in the application of the Fréchet inequalities. In the above example, the formula (4) happens to be in disjunctive normal form (DNF): a disjunction of clauses, each of which is a conjunction of literals. Instead, we could also write it in conjunctive normal form (CNF), as a conjunction of disjunctive terms:

$$\Theta_1 = (\psi_1 \vee \psi_2) \wedge (\psi_1 \vee \psi_4) \wedge (\psi_3 \vee \psi_2) \wedge (\psi_3 \vee \psi_4)\,. \tag{5}$$

Given this formula, we would first apply the Fréchet inequalities to the disjunctive clauses and only then to their conjunction. This would give $I(\theta_1) = [0.65, 1]$.

Each formula is monotone, i.e., only contains positive statements. Note that here we treat $(Z \not\Rightarrow X) \wedge (Z \not\Rightarrow \mathbf{S})$ and $(Z \Rightarrow X) \vee (Z \Rightarrow \mathbf{S})$ as two separate (positive) causal statements, where one follows from the other because of the acyclicity condition. The minimal DNF and CNF representations corresponding to a monotone formula are unique (Goldsmith et al., 2005). Given a formula $\Gamma$, we now consider two ways of computing bounds: following the natural ordering corresponding to its minimal DNF representation $\Gamma_{\text{minDNF}}$ or according to its minimal CNF representation $\Gamma_{\text{minCNF}}$. In the Appendix, we show that, when using Fréchet inequalities, the minimal DNF representation is better for computing the lower bound, i.e., $\underline{P}(\Gamma_{\text{minDNF}}) \geq \underline{P}(\Gamma)$, whereas the minimal CNF representation is better for computing the upper bound, i.e., $\bar{P}(\Gamma_{\text{minCNF}}) \leq \bar{P}(\Gamma)$.

This then suggests the following approach for the delayed algorithm. For each statement, we keep track of its minimal DNF and CNF. Whenever we combine two statements $\gamma_1$ and $\gamma_2$ with corresponding formulae $\Gamma_1$ and $\Gamma_2$ to derive a novel statement $\theta_1$ using a conjunction, we combine the minimal CNFs of $\Gamma_1$ and $\Gamma_2$ into a CNF through $\Theta_1 = (\Gamma_{\text{minCNF},1} \wedge \Gamma_{\text{minCNF},2})$, simplify that to its minimal CNF, and convert this to a (minimal) DNF using Quine's algorithm (Quine, 1955). When it so happens that the novel statement $\theta_1$ coincides with an earlier derived statement $\theta_2$, we keep only one statement and replace $\Theta_1$ and $\Theta_2$ by their disjunction $(\Theta_1 \vee \Theta_2)$ simplified into its minimal DNF.

In practice, the total number of literals in the formulae may grow very fast. Since translating between CNF and DNF can produce expressions of exponential size, the delayed algorithm is practically infeasible. In practice, we therefore restrict the number of literals in the minimal DNF representation to a prespecified maximum $M$. That is, if combining two causal statements $\gamma_1$ and $\gamma_2$ with corresponding formulae $\Gamma_1$ and $\Gamma_2$ leads to a new formula $\Theta_{\text{minDNF}}$ with in total more than $M$ literals, we choose to ignore how $\gamma_1$ and $\gamma_2$ were derived and switch to a greedy approximation at this level, cutting the tree and treating $\gamma_1$ and $\gamma_2$ as new base statements with their respective lower and upper bounds. In the experiments described in the next section, we set $M = 13$.

## 3. Results

### 3.1 Simulated Data

Through a simulation study, we aim to investigate to what extent the type of algorithm (greedy and delayed) affects the tightness of the bounds. Furthermore, we will check whether any improvement in the bounds then also leads to an improvement in the accuracy of the causal statements derived.

We generated data with sample sizes 500, 1500, and 3000 from random graphs with linear interactions between 6, 9, and 12 Gaussian random variables and some other predefined properties (adapted from (Melancon et al., 2000)). Each experiment is repeated 20 times. For the delayed

algorithm, we set the maximum number of literals in any formula to $M = 13$, to obtain a compromise between accuracy and computational complexity. Higher values do not lead to significantly different results, but considerably increase the computational costs. We considered BCCD with two different values, $K = 3$ and $K = 5$ (the default), for the parameter $K$ that specifies the maximum number of variables used to infer the base causal statements. Since with higher $K$, BCCD will already find many causal statements without the need to explicitly combine statements using causal logic, we expect the difference between the greedy and the delayed algorithm to be more distinct for the smaller value of $K$. As in native BCCD, we process causal statements sequentially, going from the most to least reliable (in terms of the lower bound reliability) and ignore any causal statements with a (lower bound) reliability below 0.5.

We first focus on the lower bounds obtained by both variants for causal statements involving only pairs of variables (and possibly selection bias), i.e., statements of the form $Z \Rightarrow X$ (with and without potential selection bias on $Z$) and $Z \not\Rightarrow X$. We will refer to these as pairwise causal statements. Since for many such pairwise causal statements the greedy and delayed variants give the exact same lower bound, we only consider those cases where the lower bounds are indeed different. Because both variants ignore any causal statements with a lower bound reliability below 0.5, the greedy variant will typically end up with less pairwise causal statements than the delayed variant. When comparing the lower bound reliability for causal statements that are only inferred by the delayed variant, we set the lower bound for the greedy algorithm to 0.5.

The plots on the top of Figure 2 give the mean difference between the lower bounds inferred by both variants for different sample sizes, numbers of variables, and $K = 3$ (green, solid) or $K = 5$ (blue, dashed), averaged over 20 experiments. The errorbars give the 95% confidence interval of the mean. With $K = 3$, the delayed algorithm indeed improves the lower bound, in particular for larger sample sizes where base causal statements tend to have a higher reliability and more causal statements can be combined before the reliability drops below the threshold of 0.5. The size of the graph does not appear to have a serious effect. For $K = 5$, the improvement in the tightness of the lower bound is smaller: most causal statements are directly derived as base causal statements.

Having access to the ground truth, we can compare the inferred causal structures, in PAG form, with the true underlying PAG. Here, for each output PAG, we evaluate how many edges are correctly oriented by each of the algorithms. The plots on the bottom of Figure 2 display the mean difference between the PAG accuracy of the delayed and the greedy variant, over the same 20 experiments. Errorbars again give the standard error of the mean. It can be seen that, for larger sample sizes, the better bounds indeed appear to lead to a small improvement for the delayed over the greedy algorithm. In conclusion, the more expensive delayed algorithm leads to better bounds: the greedy algorithm is to be preferred when computational complexity is really an issue and then the greedy algorithm still leads to acceptable bounds and PAG accuracies.

### 3.2 Real-World Data

To illustrate the estimate of the lower and upper bound by the delayed algorithm ($K = 5$) on real-world data, we use the data described in (Hoogman et al., 2013, 2011) describing patients with attention deficit-hyperactivity disorder (ADHD). The study included 164 participants, 87 patients, and 77 control subjects from the Dutch chapter of the International Multicentre persistent ADHD CollaboraTion (IMpACT). The goal of the study was to investigate the connection between candidate gene DAT1, brain functioning, and behavior characteristics associated with reward-related

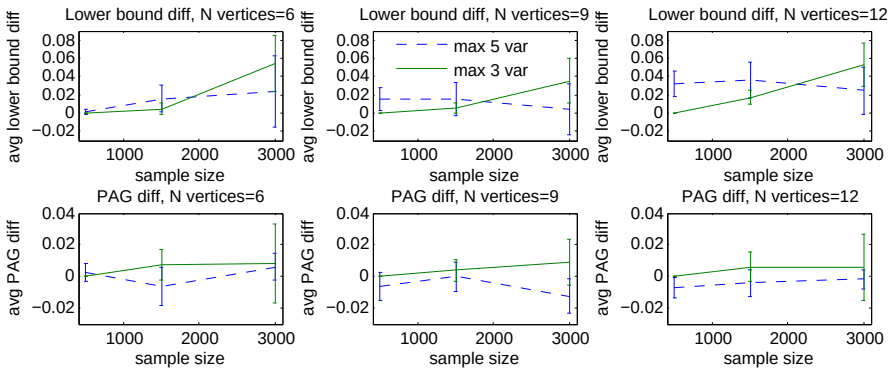



Figure 2: Results on simulated data. Top row: mean difference between the lower bound reliabilities for pairwise causal statements computed using the delayed and the greedy variants as a function of sample size. Bottom row: difference in PAG accuracy (% of edge marks in PAG) for both variants as a function of sample size. Results for graphs with 6, 9, and 12 variables (from left to right) and $K = 3$ (green, solid) and $K = 5$ (blue, dashed). See the main text for further explanation.

problems in ADHD. Two experiments were performed to learn these relationships. In the first experiment brain activity during the reward anticipation phase of the Monetary Incentive Delay (MID) task in a functional MRI paradigm was assessed. The MID task activates the ventral striatum, where DAT1 is most highly expressed. In the second experiment a delay discount task (Dom et al., 2006) was performed that aims to evaluate reward-related impulsivity.

To apply causal discovery using the BCCD algorithm, we selected ten variables from this data set. The first seven variables (disease status, smoking behavior, hyperactivity/impulsivity symptom score, attention-deficit symptoms score, medication status, presence of the DAT1 risk haplotype, ventral-striatal brain activation) were previously described in (Sokolova et al., 2014). The extra three variables that were added based on a delay discount task are: reward-related impulsivity behavior; IQ level; education level. As prior information we incorporated the assumption that the DAT1 risk haplotype cannot be influenced by any other factor in the model, and that diagnosis is present downstream of symptoms, i.e., that symptoms cannot be caused by diagnosis.

The resulting causal graph is presented in Figure 3. This figure includes only edges with a reliability of a direct causal link higher than 50%. The edges inferred in the graph are in line with several literature studies. The link between DAT1 risk haplotype and ADHD symptoms is also found in (Gizer et al., 2009). The effect of the ADHD on brain functioning was shown in (Scheres et al., 2007). The association between ADHD and smoking was described in (Milberger et al., 1997). Correlation between ADHD and reward-related impulsivity was discussed in (Paloyelis et al., 2009). Thus, we can conclude that BCCD has inferred a reliable skeleton from the data.

BCCD was also able to infer the directions of some edges in Figure 3. Here we used a lower bound threshold of 30% to get a broader overview of possible edge directions. Combining causal statements, BCCD was able to infer the lower and upper bounds for two arrows and two tails in the graph. Other edge directions were directly inferred from the conditional independencies observed in the data. The lower and upper bounds for the edge directions suggest that there is strong evidence

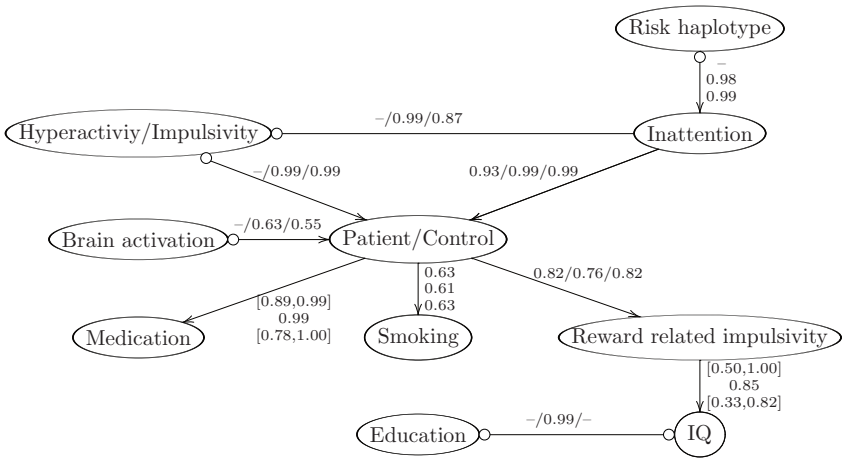



Figure 3: The causal graph representing causal relationships between variables for the ADHD data set. The graph represents a PAG, where edge directions are marked with "−"and ">" for invariant edge directions and with "∘" for non-invariant edge directions. The reliability of an edge and its direction are depicted with a reliability score in the interval [0,1] near each edge, in the following format: "the arrow or tail on the left/edge/arrow or tail on the right". In case the annotation is vertical then top row is the reliability of the top arrow or tail and bottom row is the reliability of the bottom arrow or tail. Edge directions where the reliability of the arrow or tail is lower than 30% are marked with "-".

that ADHD status (patient/control) influences the medication status (reliability tail [0.78,1.0], arrow [0.89,0.99]). On the other hand there is vague evidence that the association between reward-related impulsivity and IQ goes from former to latter (reliability tail [0.5,1.0]). We compared these bounds with bounds obtained by the default BCCD algorithm that estimates the probability of a causal statement by taking the product of probabilities in case of a conjunction and maximum in case of a disjunction. Using the default algorithm the probability of a tail between reward-related impulsivity and IQ was 0.56 and the probability of an arrow was 0.46. Thus, both probabilities were slightly higher than the lower bound obtained with the delayed algorithm. The probability of an arrow and a tail between ADHD status (patient/control) and the medication status obtained with the default BCCD algorithm coincided with the lower bound obtained with the delayed algorithm.

## 4. Conclusion and Discussion

In this paper we provided a method to estimate lower and upper bounds for the reliability of a causal statement. Such bounds are valuable to convey the uncertainties involved in causal discovery to practitioners. They help to provide guidelines for setting up new (intervention) experiments to test causal hypotheses inferred from observational data.

We demonstrated how our approach can be integrated in the BCCD algorithm. However, any other constraint-based algorithm that infers causal statements can be used potentially. We showed that a full, "delayed" version of our algorithm gives the best bound and PAG accuracies. However, it can be expensive for larger graphs, which is why we came up with a cheaper, "greedy" variant. In

this paper we considered Fréchet inequalities to estimate the lower and upper bounds of causal statements, but other approaches such as linear programming can also be used to compute the bounds. For example, using SAT solvers in some cases it is possible to provide more accurate lower and upper bound estimates than using Fréchet inequalities, as shown in (Hailperin, 1965).

## Appendix A.

First we recall some well-known concepts from Boolean logic. *Literals* are variables and negated variables. A conjunction of literals is a *term*, sometimes represented as a set of literals. A disjunction of literals is a *clause*. Every Boolean function can be represented as a conjunction of clauses, referred to as *conjunctive normal form* (CNF), as well as a disjunction of terms (DNF), referred to as *disjunctive normal form* (DNF). A *monotone* Boolean function is one without any negated variables. A term $\phi$ *subsumes* a term $\psi$ iff $\phi \subset \psi$.

**Lemma 1 (Lower bound)** *Given any monotone formula* $\Gamma$ *and its corresponding (unique) minimal DNF representation* $\Gamma_{\text{minDNF}}$. *When using Fréchet inequalities, the minimal DNF representation gives the best possible lower bound, i.e.,* $\underline{P}(\Gamma_{\text{minDNF}}) \geq \underline{P}(\Gamma)$.

**Proof** The proof is by induction on the number $n$ of operators ($\vee$,$\wedge$) in $\Gamma$. The base case $n = 0$ clearly holds. If $n > 0$ then $\Gamma$ can be rewritten as a disjunction, $\Gamma = \Gamma_1 \vee \Gamma_2$ or a conjunction, $\Gamma = \Gamma_1 \wedge \Gamma_2$ of two formulae $\Gamma_1$ and $\Gamma_2$, with $(n_1, n_2) < n$. We assume that the lemma holds for $\Gamma_1$ and $\Gamma_2$, and prove that it then also holds for $\Gamma$. Let us first consider the disjunction, i.e., we suppose that $\Gamma = \Gamma_1 \vee \Gamma_2$. Then, by definition of how we apply the Fréchet inequalities,

$$\underline{P}(\Gamma) \quad = \underline{P}(\Gamma_1 \vee \Gamma_2) = \max\left(\underline{P}(\Gamma_1), \underline{P}(\Gamma_2)\right) \leq \max\left(\underline{P}(\Gamma_{\text{minDNF},1}), \underline{P}(\Gamma_{\text{minDNF},2})\right) \ ,$$

where the last step follows from the induction assumption. Now,

$$\underline{P}(\Gamma) \quad \leq \max\left(\underline{P}(\Gamma_{\text{minDNF},1}), \underline{P}(\Gamma_{\text{minDNF},2})\right) = \underline{P}(\Gamma_{\text{minDNF},1} \vee \Gamma_{\text{minDNF},2}) = \underline{P}(\Gamma_{\text{minDNF}}) \ .$$

Since all formulae are monotone, $\Gamma_{\text{minDNF}}$ is unique and can be obtained by removing terms from the disjunction of $\Gamma_{\text{minDNF},1}$ and $\Gamma_{\text{minDNF},2}$ such that it does not contain any subsumed terms (Quine, 1955). Clearly, removing subsumed terms does not change the lower bound, which gives the last step in the disjunctive part of the proof. For the conjunction, we have

$$\begin{aligned} \underline{P}(\Gamma) \quad &= \underline{P}(\Gamma_1 \wedge \Gamma_2) = \max\left(0, \underline{P}(\Gamma_1) + \underline{P}(\Gamma_2) - 1\right) \\ &\leq \max\left(0, \underline{P}(\Gamma_{\text{minDNF},1}) + \underline{P}(\Gamma_{\text{minDNF},2}) - 1\right) \ , \end{aligned}$$

again with the first step by definition of how we compute a lower bound using the Fréchet inequalities and the last step from the induction assumption. Now, since for probability values $(a, b) \in [0, 1]$, we have $a + b - 1 \leq \max(a, b)$, we get

$$\begin{aligned} \underline{P}(\Gamma) \quad &\leq \max\left(0, \underline{P}(\Gamma_{\text{minDNF},1}) + \underline{P}(\Gamma_{\text{minDNF},2}) - 1\right) \\ &\leq \max\left(\underline{P}(\Gamma_{\text{minDNF},1}), \underline{P}(\Gamma_{\text{minDNF},2})\right) = \underline{P}(\Gamma_{\text{minDNF},1} \vee \Gamma_{\text{minDNF},2}) = \underline{P}(\Gamma_{\text{minDNF}}) \ . \end{aligned}$$

■

**Lemma 2 (Upper bound)** *Given any monotone formula $\Gamma$ and its corresponding (unique) minimal CNF representation $\Gamma_{\text{minCNF}}$. When using Fréchet inequalities, the minimal CNF representation gives the best possible upper bound, i.e., $\bar{P}(\Gamma_{\text{minCNF}}) \leq \bar{P}(\Gamma)$.*

**Proof** We follow exactly the same reasoning as in the proof of Lemma 1, but consider the upper bound instead of the lower bound and minCNF instead of minDNF. For $\Gamma = \Gamma_1 \wedge \Gamma_2$ we now have

$$\bar{P}(\Gamma) \quad = \bar{P}(\Gamma_1 \wedge \Gamma_2) = \min\left(\bar{P}(\Gamma_1), \bar{P}(\Gamma_2)\right) \geq \min\left(\bar{P}(\Gamma_{\text{minCNF},1}), \bar{P}(\Gamma_{\text{minCNF},2})\right) ,$$

and

$$\begin{aligned} \bar{P}(\Gamma) \quad &\geq \min\left(\bar{P}(\Gamma_{\text{minCNF},1}), \bar{P}(\Gamma_{\text{minCNF},2})\right) \\ &= \bar{P}(\Gamma_{\text{minCNF},1} \wedge \Gamma_{\text{minCNF},2}) = \bar{P}(\Gamma_{\text{minCNF}}) . \end{aligned}$$

For $\Gamma = \Gamma_1 \vee \Gamma_2$ we have

$$\begin{aligned} \bar{P}(\Gamma) \quad &= \bar{P}(\Gamma_1 \vee \Gamma_2) = \min\left(1, \bar{P}(\Gamma_1) + \bar{P}(\Gamma_2)\right) \\ &\geq \min\left(1, \bar{P}(\Gamma_{\text{minCNF},1}) + \bar{P}(\Gamma_{\text{minCNF},2})\right) , \end{aligned}$$

and

$$\begin{aligned} \bar{P}(\Gamma) \quad &\geq \min\left(1, \bar{P}(\Gamma_{\text{minCNF},1}) + \bar{P}(\Gamma_{\text{minCNF},2})\right) \geq \min\left(\bar{P}(\Gamma_{\text{minCNF},1}), \bar{P}(\Gamma_{\text{minCNF},2})\right) \\ &= \bar{P}(\Gamma_{\text{minCNF},1} \wedge \Gamma_{\text{minCNF},2}) = \bar{P}(\Gamma_{\text{minCNF}}) . \end{aligned}$$

■

## References


T. Claassen and T. Heskes. A Bayesian approach to constraint based causal inference. In *Proceedings of the UAI Conference*, pages 207–216. AUAI Press, 2012a.

T. Claassen and T. Heskes. A logical characterization of constraint-based causal discovery. *arXiv preprint arXiv:1202.3711*, 2012b.

A. P. Dawid. Statistical theory: the prequential approach (with discussion). *J. R. Statist. Soc. A*, 147:278–292, 1984.

G. Dom, B. De Wilde, W. Hulstijn, W. van den Brink, and B. Sabbe. Behavioural aspects of impulsivity in alcoholics with and without a cluster-B personality disorder. *Alcohol and Alcoholism*, 41(4):412–420, 2006.

M. Fréchet. Généralisation du théoreme des probabilités totales. *Fund. Math.*, 1(25):379–387, 1935.

I. Gizer, C. Ficks, and I. Waldman. Candidate gene studies of ADHD: a meta-analytic review. *Human genetics*, 126(1):51–90, 2009.

J. Goldsmith, M. Hagen, and M. Mundhenk. Complexity of DNF and isomorphism of monotone formulas. In *Mathematical Foundations of Computer Science*, pages 410–421. Springer, 2005.

T. Hailperin. Best possible inequalities for the probability of a logical function of events. *Amer. Math. Monthly*, 72(4):343–359, 1965.

M. Hoogman, E. Aarts, M. Zwiers, D. Slaats-Willemse, M. Naber, M. Onnink, R. Cools, C. Kan, J. Buitelaar, and B. Franke. Nitric oxide synthase genotype modulation of impulsivity and ventral striatal activity in adult ADHD patients and healthy comparison subjects. *Am. J. Psychiatry*, 2011.

M. Hoogman, M. Onnink, R. Cools, E. Aarts, C. Kan, A. Arias Vasquez, J. Buitelaar, and B. Franke. The dopamine transporter haplotype and reward-related striatal responses in adult ADHD. *European Neuropsychopharmacology*, 23(6):469–478, 2013.

S. Mani, P. Spirtes, and G. Cooper. A theoretical study of Y structures for causal discovery. *arXiv preprint arXiv:1206.6853*, 2012.

G. Melancon, I. Dutour, and M. Bousquet-Melou. Random generation of dags for graph drawing. Technical report, Amsterdam, The Netherlands, The Netherlands, 2000.

S. Milberger, J. Biederman, S. Faraone, L. Chen, and J. Jones. ADHD is associated with early initiation of cigarette smoking in children and adolescents. *Journal of the American Academy of Child & Adolescent Psychiatry*, 36(1):37–44, 1997.

Y. Paloyelis, P. Asherson, and J. Kuntsi. Are ADHD symptoms associated with delay aversion or choice impulsivity? a general population study. *Journal of the American Academy of Child & Adolescent Psychiatry*, 48(8):837–846, 2009.

J. Pearl and T. Verma. A theory of inferred causation. In *Proceedings of the KR Conference*, pages 441–452. Morgan Kaufmann, 1991. ISBN 1-55860-165-1.

W. Quine. A way to simplify truth functions. *Amer. Math. Monthly*, 62(9):627–631, 1955.

T. Richardson and P. Spirtes. Causal inference via ancestral graph models. *Highly Structured Stochastic Systems*, 27:83, 2003.

A. Scheres, M. Milham, B. Knutson, and F. Castellanos. Ventral striatal hyporesponsiveness during reward anticipation in attention-deficit/hyperactivity disorder. *Biological psychiatry*, 61(5):720–724, 2007.

T. Sellke, M. Bayarri, and J. Berger. Calibration of $p$ values for testing precise null hypotheses. *The American Statistician*, 55(1):62–71, 2001.

E. Sokolova, P. Groot, T. Claassen, and T. Heskes. Causal discovery from databases with discrete and continuous variables. In *Probabilistic Graphical Models*, pages 442–457. Springer, 2014.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT press, 2000.

S. Triantafilou, I. Tsamardinos, and A. Roumpelaki. Learning neighborhoods of high confidence in constraint-based causal discovery. In *Probabilistic Graphical Models*, pages 487–502, 2014.

P. Walley. Statistical reasoning with imprecise probabilities. 1991.

K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, 24(2):149–170, 2000.