Causal Discovery from Databases with Discrete and Continuous Variables

Elena Sokolova, Perry Groot, Tom Claassen, and Tom Heskes

Radboud University, Faculty of Science,
Postbus 9010, 6500 GL Nijmegen, The Netherlands

Abstract. Bayesian Constraint-based Causal Discovery (BCCD) is a state-of-the-art method for robust causal discovery in the presence of latent variables. It combines probabilistic estimation of Bayesian networks over subsets of variables with a causal logic to infer causal statements. Currently BCCD is limited to discrete or Gaussian variables. Most of the real-world data, however, contain a mixture of discrete and continuous variables. We here extend BCCD to be able to handle combinations of discrete and continuous variables, under the assumption that the relations between the variables are monotonic. To this end, we propose a novel method for the efficient computation of BIC scores for hybrid Bayesian networks. We demonstrate the accuracy and efficiency of our approach for causal discovery on simulated data as well as on real-world data from the ADHD-200 competition.

Keywords: Causal discovery, hybrid data, structure learning.

1 Introduction

Causal discovery is widely used for analysis of experimental data focusing on the exploratory analysis and suggesting probable causal dependencies. There is a variety of causal discovery algorithms in the literature. Some of these algorithms rely on the assumption that there are no latent variables in the model; others do not provide a scoring metric to easily compare the reliability of two candidate models. Bayesian Constraint-based Causal Discovery (BCCD) [6] is a state-of-the-art-algorithm for causal discovery that tries to combine the strength of the best algorithms in the field. BCCD is able to detect latent variables in the model and determines the reliability of the edges between variables that makes it very easy to compare alternative models.

The idea of BCCD is to estimate the reliability of causal relations by scoring Directed Acyclic Graphs (DAGs) for a smaller subset of variables using a Bayesian score and then to combine these statements to infer a final causal model. The Bayesian score has a closed form solution for discrete variables that makes the scoring of causal relations fast and efficient. The BCCD algorithm is currently limited to discrete or Gaussian variables as there is no closed form solution for the Bayesian score for a mixture of discrete and continuous variables.

L.C. van der Gaag and A.J. Feelders (Eds.): PGM 2014, LNAI 8754, pp. 442–457, 2014.

To extend BCCD, we need a new scoring method to estimate the reliability of causal relations.

There are several scoring methods in the literature for mixtures of discrete and continuous variables. Most of these methods either rely on strict assumptions about the structure of the network that do not apply in practice, such as forbidding structures in the network with a continuous variable as a parent having a discrete variable as child, or are time consuming or/and memory inefficient. In this paper we propose a fast and memory efficient method to score DAGs with both discrete and continuous variables, under the assumption that the relationships between these variables are monotonic. This appears to be a reasonable assumption for many real-world data sets.

The scoring method proposed in this paper estimates the Bayesian information criterion (BIC) score by approximating the mutual information for a combination of discrete and continuous variables. Through simulations we will show that the BCCD algorithm with this scoring method can accurately estimate the structure of the Bayesian network for both simulated and real-world data.

The rest of the paper is organized as follows. Section 2 describes background information about causal discovery and graphical models. Section 3 describes algorithms for structure learning. Section 4 briefly summarizes the idea of the BCCD algorithm. Section 5 explains the scoring method for a mixture of discrete and continuous variables. Section 6 presents the results of the experiments of the BCCD algorithm on simulated data and real-world data. Section 7 provides our conclusion and future work.

2 Background

A Bayesian network is a pair $(\boldsymbol{\mathcal{G}}, \boldsymbol{\Theta})$ where $\boldsymbol{\mathcal{G}} = (\mathbf{X}, \mathbf{E})$ is a Directed Acyclic Graph (DAG) with a set of nodes $\mathbf{X}$ representing domain variables and a set of arcs $\mathbf{E}$; $\theta_{X_i} \subset \boldsymbol{\Theta}$ is a set of parameters representing the conditional probability of variable $X_i \subset \mathbf{X}$ given its parents Pa_i in a graph $\boldsymbol{\mathcal{G}}$. Using Bayesian networks we can model causal relationships between variables. In that case an edge $A \rightarrow B$ between variables represents a direct causal link from A to B. This means that A influences the values of B, but not the other way around.

Saying that two variables A and B are conditionally independent given C, means that if we know C, learning B would not change our belief in A. Two DAGs are called equivalent to one another, if they entail the same conditional (in)dependencies. All DAGs that are equivalent to a graph $\mathcal{G}$ form an equivalence class of a graph $\mathcal{G}$, where all members are indistinguishable in terms of implied independencies. To represent the members of this equivalence class, a different type of structure is used, known as a partially directed acyclic graph (PDAG).

The three main assumptions that are often used when learning the structure of causal networks are the following.

1. Causal Markov Condition: each variable is independent of its non-descendant conditioned on all its direct causes.

2. Faithfulness assumption: there are no independencies between variables that are not implied by the Causal Markov Condition.
3. Causal sufficiency assumption: there are no common confounders of the observed variables in $\mathcal{G}$.

In this paper we do not rely on the causal sufficiency assumption, i.e. we do allow for latent variables. One can represent the structure of a Bayesian network with latent variables using a so-called Maximal Ancestral Graph (MAG) on only the observed variables. In contrast to DAGs, MAGs can also contain bi-directed $X \leftrightarrow Y$ arcs (indicating that there is a common confounder) and undirected arcs $X - Y$. The equivalence class for MAGs is a partial ancestral graph (PAG). Edge directions are marked with " $-$ " and ">" if the direction is the same for all graphs belonging to the PAG and with "$\circ$" otherwise.

3 Structure Learning of Causal Networks

There is a variety of methods that can be used to learn the structure of a causal network. A broad description of methods can be found in [8]. In general, methods are divided into two approaches: constraint-based and score-based. The constraint-based approach works with statistical independence tests. Firstly, this approach finds a skeleton of a graph by starting from the complete graph and excluding edges between variables that are conditionally independent, given some other set of variables (possibly empty). Secondly, the edges are oriented to arrive at an output graph. The constraint-based approach learns the equivalence class of DAGs and outputs a PDAG. Examples of the constraint-based approach are the IC algorithm [23], PC-FCI [27], and TC [24].

A score-based approach uses a scoring metric. It measures the data goodness of fit given a particular graph structure and accounts for the complexity of the network. There are many different scoring metrics, where the Bayesian score [9] and the BIC score [26] are among the most common. The goal is to find the graph that has the highest score. Unfortunately, this optimization problem is NP-hard, so different heuristics are used in practice. These methods are divided in local search methods, such as greedy search [5], greedy equivalence search [4], and global search methods, such as simulated annealing [10] and genetic algorithms [17].

An advantage of the constraint-based approach is that it does not have to rely on the causal sufficiency assumption, which means that the algorithm can detect common causes of the observed variables. A disadvantage of the constraint-based approach is that it is sensitive to propagating mistakes in the resulting graph. A standard approach makes use of independence tests, whose results for borderline independencies/dependencies sometimes can be incorrect. The outcome of learning a network can be sensitive to such errors. In particular, one such error can produce multiple errors in the resulting graph. A set of conservative methods such as CPC [25] and CFCI [28] tackles the problem of lack of robustness, outperforming standard constraint-based methods such as PC.

An advantage of the score-based approach is that it indicates the measure of reliability of inferred causal relations. This makes the interpretation of the results easier and prevents incorrect categorical decisions. A main drawback of the approach is that it relies on the causal sufficiency assumption and as a result cannot detect latent confounders.

4 Bayesian Constraint-Based Causal Discovery

One of the state-of-the-art algorithms in causal discovery is Bayesian Constraint-based Causal Discovery (BCCD). Claassen and Heskes [6] showed that BCCD outperforms reference algorithms in the field, such as FCI and Conservative PC. Moreover, it provides an indication of the reliability of the causal links that makes it easier to interpret the results and compare alternative models. The advantage of the BCCD algorithm is that it combines the strength of constraint-based and score-based approaches. We here describe only the basic idea of the method. A more detailed description can be found in [6]. The main two steps of BCCD are the following:

Step 1. Start with a fully connected graph and perform adjacency search, estimating the reliability of causal relations, for example $X \to Y$. If a causal relation declares a variable conditionally independent with a reliability higher than a predefined threshold, delete an edge from the graph between these variables.

Step 2. Rank all causal relations in decreasing order of reliability and orient edges in the graph starting from the most reliable relations. If there is a conflict, pick the causal relation that has a higher reliability.

Based on the score of the causal relations, we can rank these relations and avoid propagating unreliable decisions giving preference to more confident ones. This can solve the drawback of a standard constraint-based method that can end up with an unreliable result. Moreover, using a Bayesian score we get a reliability measure of the final output, which makes it easier to interpret the results and compare with other alternative models. The BCCD algorithm does not rely on the causal sufficiency assumption, thus it can detect latent variables in the model.

The first step of the algorithm requires estimating the reliability of a causal relations $L : `X \to Y'$ given a data set $\mathbf{D}$, which is done using a Bayesian score:

$$p(L : `X \to Y' | \mathbf{D}) = \frac{\sum_{\mathcal{M} \in \mathbf{M}(L)} p(\mathbf{D}|\mathcal{M}) p(\mathcal{M})}{\sum_{\mathcal{M} \in \mathbf{M}} p(\mathbf{D}|\mathcal{M}) p(\mathcal{M})} \quad , \tag{1}$$

where $p(\mathbf{D}|\mathcal{M})$ denotes the probability of data $\mathbf{D}$ given structure $\mathcal{M}$, $p(\mathcal{M})$ represents the prior distribution over structures and $\mathbf{M}(L)$ is the set of structures containing the relation L. This reliability measure (1) gives a conservative estimate of the probability of a causal relation. Claassen and Heskes approximate the probability $p(\mathbf{D}|\mathcal{M})$ by $p(\mathbf{D}|\mathcal{G})$, the marginal likelihood of the data given

graph $\mathcal{G}$ that has a closed form solution for discrete variables, known as the Bayesian Dirichlet (BD) metric [16]. There is also a closed-form solution when all variables have a Gaussian distribution, called the BGe metric [16]. To estimate (1), the algorithm requires calculating the marginal likelihood over all possible graphs for each causal relation that we infer. For speed and efficiency of the algorithm, the set of possible graphs is limited to the graphs with at most five vertices, which gives a list of at most 29,281 DAGs. In theory, limiting the number of vertices to five may lead to loss of information. In practice, however, the accuracy of the BCCD algorithm is hardly affected and it still outperforms standard algorithms that perform conditional independence tests for more than five vertices [6].

The use of BD/BGe metrics to score DAGs estimating the marginal likelihood, limits the BCCD algorithm to work only with discrete variables or only with Gaussian variables. However, many real-world data sets contain a mixture of discrete and continuous variables. One of the possible solutions to this problem would be to discretize continuous variables [12,21]. This, however, can lead to a loss of information comprised in continuous variables, spurious dependencies and misleading results of the method [27]. Another solution to this problem is to define a new scoring method that can estimate the marginal likelihood when the data is a mixture of discrete and continuous variables. Since the estimation of the marginal likelihood is repeated many times for a large number of possible graphs, such a new scoring method must be fast and memory efficient.

5 The Scoring Metric for Discrete and Continuous Variables

In this section we discuss methods that can estimate the Bayesian score (1) for a mixture of discrete and continuous variables. In Sect. 5.1 we provide an overview of existing scoring algorithms and in Sect. 5.2 we propose our scoring method.

5.1 Alternative Approaches

There are few score-based methods that can work with a mixture of discrete and continuous variables. Geiger and Heckerman [13] proposed a closed-form solution for the Bayesian score of a mixture of discrete and continuous variables, but this solution only works in case a number of assumptions are met. These assumptions imply that the data are drawn from a conditional Gaussian distribution and forbid structures in the network with a continuous variable having a discrete variable as a child.

An alternative method is described in [11] which uses a multiple regression framework for scoring structures. However, the method is applicable only for time-series data. Bach and Jordan [1] use Mercer kernels to estimate the structure of causal models, but calculation of a Gramm matrix requires significant computational costs ($O(N^3)$, where N is the sample size) and may be inefficient

for data sets with large sample sizes. Monti and Cooper [20] use neural networks to represent the density function for a mixture of discrete and continuous variables. Estimation of the neural network parameters to calculate (1) requires significant computational costs which should be repeated multiple times and becomes too slow to be applicable in the BCCD algorithm.

5.2 Proposed Scoring Method

Existing scoring methods for a mixture of discrete and continuous variables are either computationally or/and memory expensive or rely on strict assumptions that limit the types of data that can be used considerably. In this section we propose a fast and memory efficient method for scoring DAGs that relies only on the assumption that relationships between variables are monotonic. This is a reasonable assumption for many domains.

We consider the BIC score, which is an approximation of the Bayesian score, to estimate the marginal likelihood $p(\mathbf{D}|\mathcal{G})$. The BIC score can be decomposed into the sum of two components, the mutual information $I(X_i, Pa_i)$ with Pa_i the parents of node X_i and $Dim[\mathcal{G}]$ the number of parameters necessary to estimate the model. The first component measures the goodness of fit, and the second penalizes the complexity of the model:

$$BICscore(\mathbf{D}|\mathcal{G}) = M \sum_{i=1}^{n} I(X_i, Pa_i) - \frac{\log M}{2} Dim[\mathcal{G}] \ , \tag{2}$$

where n is the number of variables and M is a sample size.

To estimate (2) for a mixture of discrete and continuous variables we need to estimate the mutual information and the complexity penalty. We propose to approximate the mutual information based on the formula for continuous variables drawn from a Gaussian distribution [7]:

$$I(X_i, Pa_i) = -\frac{1}{2} \log \frac{|R|}{\mid R_{Pa_i} \mid} \ , \tag{3}$$

where R is a correlation matrix between all variables and R_{Pa_i} is a correlation matrix between the parents of variable X_i.

Estimation of the complexity penalty for the model containing a mixture of variables is reduced to calculation of the number of parameters in the model:

$$Dim[\mathcal{G}] = \sum_{i=1}^{n} d_i d_{Pa_i}, \text{with } d_{Pa_i} = \sum_{j \in Pa_i(\mathcal{G})} d_j, \text{ and}$$
$$d_j = \begin{cases} 1, & \text{if } X_j \text{ is continuous} \\ s-1, & \text{if } X_j \text{ is discrete} \end{cases} , \tag{4}$$

where s is the number of states of the discrete variable X_j. In case the variable X_i does not have parents, we assign (3) and (4) a zero value. Due to that assignment,

(4) represents the difference in the number of parameters, when there is an edge between X_i and Pa_i and when there are no edges between them.

We propose a scoring method that approximates the BIC score for a mixture of discrete and continuous variables using (3) and (4) and substituting Pearson correlation with Spearman in (3). We assume hereby explicitly that the variables obey a so-called non-paranormal distribution. For univariate monotone functions $f_1, ..., f_d$ and a positive-definite correlation matrix $\Sigma^0 \in \mathbb{R}^{d \times d}$ we say that a d-dimensional random variable $X = (X_1, ..., X_d)^T$ has a non-paranormal distribution $X \backsim \text{NPN}_d(f, \Sigma^0)$, if $f(X) = (f_1(X_1), ..., f_d(X_d)) \backsim N_d(0, \Sigma^0)$. As shown in [15], conditional independence tests for non-paranormal data based on Spearman correlations are more accurate than those based on Pearson correlations for non-Gaussian continuous data. Ignoring the discreteness of the discrete variables does introduce some bias in the approximation of the BIC score, however, as we will argue and show in the next section, this bias hardly affects the scoring of the network structures. In case of categorical variables we propose to perform so-called dummy coding, binarising the variable into several variables when calculating the correlation matrix and then adjusting for the correlation between these variables when calculating the mutual information.

The most computationally expensive part of the proposed scoring method is the calculation of the correlation matrix. However, one can compute the full correlation matrix once beforehand, which can then be stored and used to efficiently construct the correlation matrices for any subset of variables. The proposed method is thus computationally and memory efficient.

In this section we described the scoring method for a mixture of discrete and continuous variables. Now we can use this scoring method to estimate marginal likelihood in the loop of the BCCD algorithm.

6 Experiments

In this section we describe the results of our experiments. Section 6.1 describes the accuracy of estimating mutual information using (3). In Sect. 6.2 we describe the results of the experiments on simulated data, where the ground truth about the structure of the network is known. In Sect. 6.3 we describe the results of the experiments on real-world data, where the ground truth is unknown.

6.1 Testing Mutual Information

In the previous section we proposed a method to score DAGs for a mixture of discrete and continuous variables that approximates mutual information in the BIC score using (3). To estimate the accuracy of this approximation, we performed a series of simulations. In these simulations we randomly generated two variables with particular distributions and a sample size of 1000 and then compared the exact value of mutual information with the approximated value. Since the strength of association between variables can influence the accuracy of the estimation we performed the simulations changing the correlation between

the variables from 0 to 1. Simulations were performed for three types of data: continuous, discrete, and a mixture of discrete and continuous.

Figure 1 (a) shows the difference between the exact mutual information and mutual information estimated using (3) between two normally distributed random variables. There is a slight difference between the exact and approximated mutual information, since we use Spearman correlation instead of Pearson in (3) in order to capture monotonic relations between variables that may be non-normally distributed.

Figure 1 (b) shows the difference between the exact mutual information and mutual information estimated using (3) between two discrete binary variables. The difference between the estimates of mutual information arises when the correlation between two variables is close to one. This happens because our mutual information estimate is based upon the incorrect assumption that discrete variables can have an infinite amount of values.

In practice, however, the approximated mutual information for discrete variables will not strongly affect the selection of the correct structure. For the range of correlations when the mutual information is off, the probability of the edge between two variables is always close to one. Thus, higher levels of mutual information cannot overestimate the probability of an edge, since it has already achieved its maximum. As a result, overestimation of the mutual information for discrete variables hardly affects the scoring of the structures.

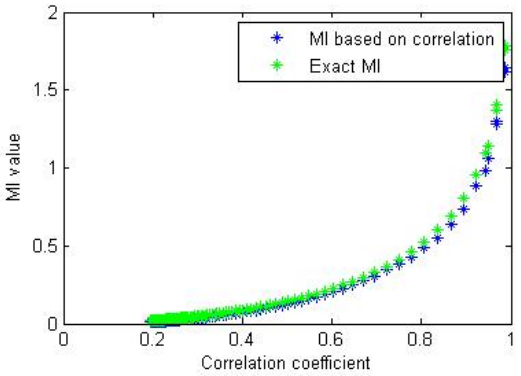
Figure 1 (c) shows the difference between the exact mutual information and mutual information estimated using (3) between one binary parent and one normally distributed child with its mean depending on the value of the binary parent. Figure 1 (c) shows that the approximated value of mutual information stays very close to the actual value.

6.2 Application on Simulated Data

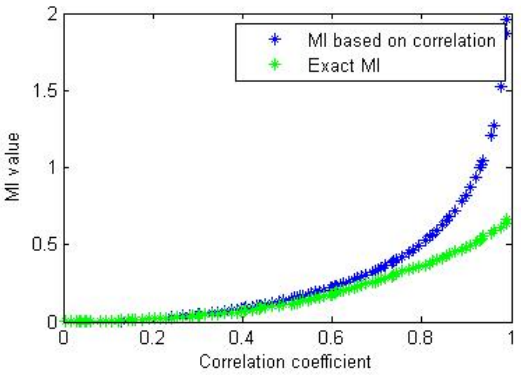
To test our algorithm on simulated data we chose two widely used Bayesian networks: the Asia Network [18] and the Waste Incinerator Network [19] to test the algorithm for discrete variables and a mixture of discrete and continuous variables, respectively. The case of only continuous variables was already discussed in [15]. We randomly generated data for four different sample sizes: 100, 500, 1000, and 1500 and repeated our experiments 20 times. Performance is measured by PAG accuracy measure, that evaluates how many edges were oriented correctly in the output PAG. We also estimated the correctness of the skeleton by calculating the amount of correct, missing, and spurious edges of the resulting graph.

The Asia data set describes the effect of visiting Asia and smoking behavior on the probability of contracting tuberculosis, cancer or bronchitis. The network contains eight binary variables that are connected by eight arcs as can be seen in Figure 2. For this network we compared the results of the BCCD using BD as a scoring method, following [16] and using our approximated BIC score.

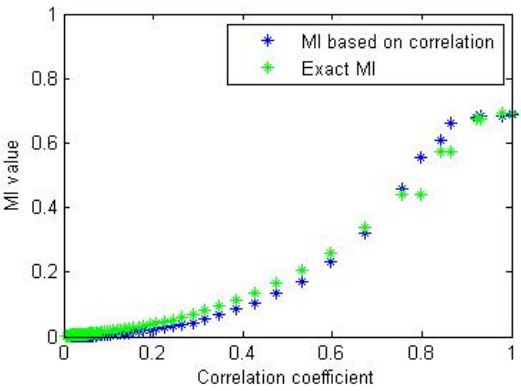
Figure 3 compares the accuracy of the BCCD algorithm for the Asia Network when the BIC score is estimated using BD metric (BIC original) or correlation



(a) two normally distributed variables



(b) two binary variables



(c) one binary and one normally distributed variable

Fig. 1. The exact mutual information (MI) and mutual information calculated using Spearman correlation for different values of the correlation coefficient between: (a) two normally distributed variables, (b) two binary variables, (c) one binary and one normally distributed variable.

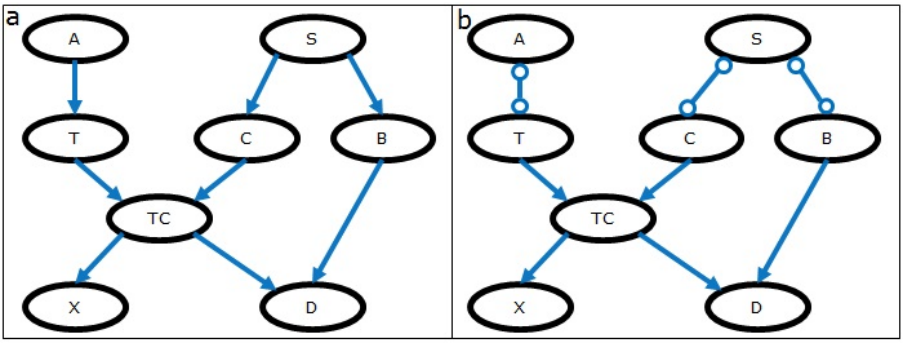


Fig. 2. Asia Network represented as (a) DAG and (b) PAG. The node names are abbreviated as follows: Visit to Asia? (A) Smoker? (S) Has tuberculosis? (T) Has lung cancer? (C) Has bronchitis? (B) Tuberculosis or cancer (TC) Positive X-ray? (X) Dyspnoea? (D)

matrix (BIC Corr). Figure 3 suggests that there is no significant difference in PAG accuracy and in the skeleton of the graph between the two methods which is also confirmed by a Students t-test. The larger the sample size, the higher is the PAG accuracy and the more accurate the skeleton. Based on these results, we suggest that estimating mutual information using the correlation matrix for discrete variables gives an accurate estimation of the causal structure.

The Waste Incinerator Network describes the emission from a waste incinerator depending on filter efficiency, waste type, burning regime, and other factors. The network contains nine variables that are connected by ten arcs as can be seen in Fig. 4.

The network contains three discrete (B, F, W) and six continuous variables (C, E, MW, ME, D, L). The original parameters of Waste Incinerator Network have a very low variance that results in an almost one-to-one deterministic relationship between the variables, which violates the faithfulness assumption. To avoid this problem, we increased the variance between variables E, F, and V to 2.5 and between other variables to 0.2.

Figure 5 represents the accuracy of the BCCD algorithm when the BIC score is estimated using the correlation matrix for the Waste Incinerator Network. Figure 5 shows that the PAG and skeleton accuracy increases with the increase of sample size and becomes close to one. That suggests that the BCCD algorithm is able to estimate the structure of a Bayesian network for a mixture of discrete and continuous variables. Based on this conclusion we now can apply the BCCD algorithm on real-world data.

6.3 Application on Real-World Data: ADHD

To test the BCCD algorithm on real-world data, we use the data set collected by [3] that contains a mixture of discrete and continuous variables. This data set

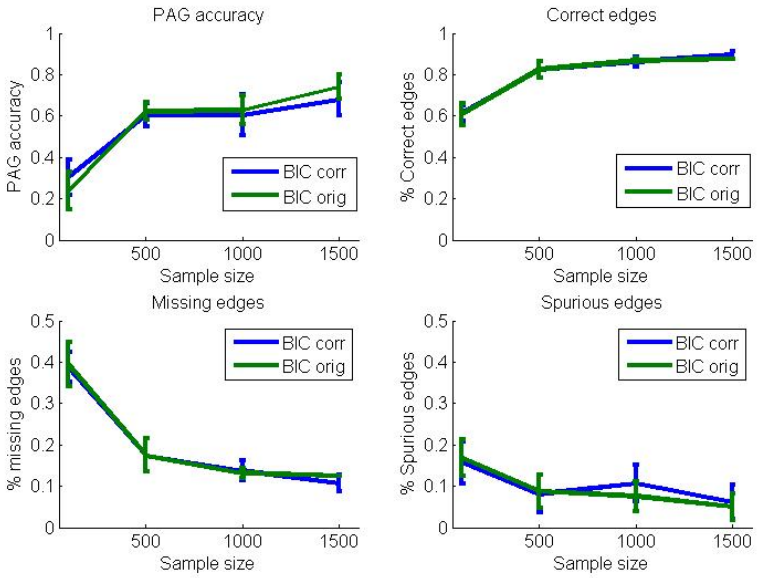


Fig. 3. The accuracy of the BCCD algorithm for the Asia Network in which mutual information for the BIC score is estimated using: frequency counts (BIC orig) and correlation matrices (BIC corr). (a) PAG accuracy. (b) Percentage of correct edges. (c) Percentage of missing edges. (d) Percentage of spurious edges.

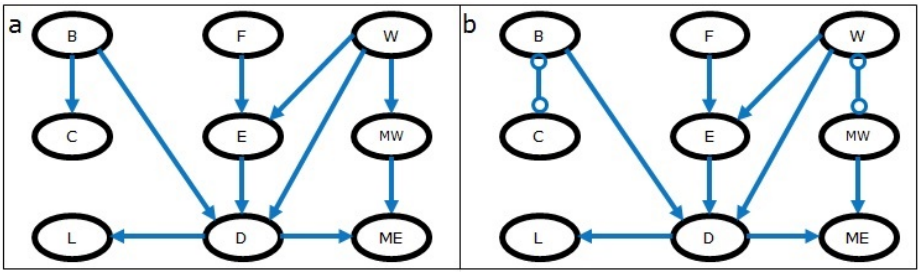


Fig. 4. Waste Incinerator Network represented as (a) DAG and (b) PAG. The node names are abbreviated as follows: Burning regime (B) Filter state (F) Waste type(W) CO2 concentration (C) Filter efficiency (E) Metal in waste (MW) Light penetrability (L) Dust emission (D) Metals emission (ME)

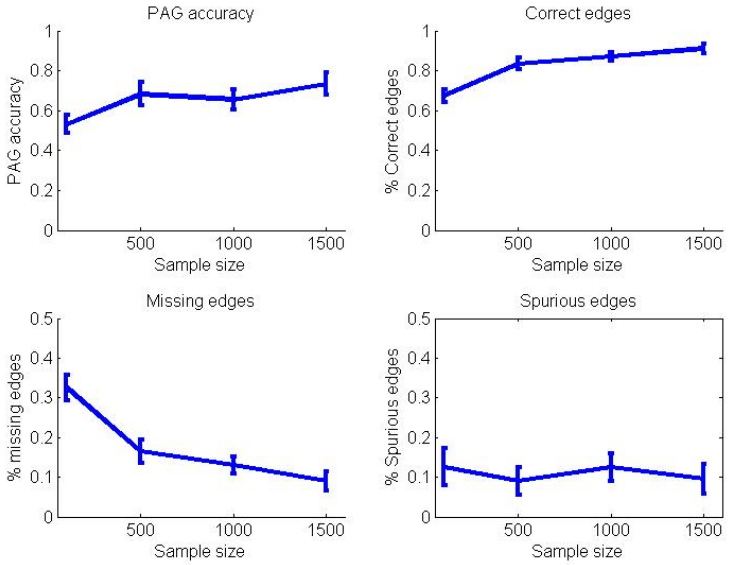


Fig. 5. The accuracy of the BCCD algorithm for the Waste Incinerator Network (a) PAG accuracy. (b) Percentage of correct edges. (c) Percentage of missing edges. (d) Percentage of spurious edges.

describes phenotypic information about children with Attention Deficit Hyperactivity Disorder (ADHD) and is publicly available as a part of the ADHD-200 competition. The ADHD data set contains 23 variables for 245 subjects. We excluded subjects that had missing information, so our final data set contained 223 subjects. We also excluded eleven variables that either did not have enough data or were considered irrelevant, leaving only the following nine variables:

1. Gender (binary: male/female)
2. Attention deficit score (continuous)
3. Hyperactivity/impulsivity score (continuous)
4. Verbal IQ (continuous)
5. Full IQ (continuous)
6. Performance IQ (continuous)
7. Aggressive behavior (binary: yes/no)
8. Medication status (binary: naïve/not naïve)
9. Handedness (binary: right/left)

The BCCD algorithm using Spearman correlation was applied to the ADHD data set. Due to the small sample size the BCCD algorithm inferred only the skeleton of the network, but not the direction of the edges for the resulting network. However, including prior knowledge about the domain that no variable in the network can cause gender, BCCD inferred the direction of several edges.

Figure 6 shows the network inferred by the BCCD algorithm. This figure includes the edges with a reliability of a direct causal link higher than 50%, that are calculated based on (1). The resulting network suggests that there is a strong effect of gender on the level of attention deficit and consequently hyperactivity/impulsivity symptoms. This statement is confirmed by different studies in the field of ADHD [2]. The effect of attention deficit on hyperactivity/impulsivity was also found in [30]. From the network we can see that the level of aggression is associated with both: attention deficit and hyperactivity/impulsivity. Moreover, the network suggests that left-handedness is associated with a higher risk of aggressive behavior. The BCCD algorithm inferred that prescription of medication is mainly associated with a high level of inattention and aggression. The network suggests that the association between level of performance IQ, verbal IQ, and full IQ is explained by a latent common cause. Thus, an IQ-related latent variable can be introduced to model the association between different types of IQ measurements. On the other hand, only the performance IQ has a causal link with attention deficit symptoms either direct or through a latent common cause between these two variables. The presence of a relation between attention deficit and intelligence is confirmed in several medical studies [22,29].

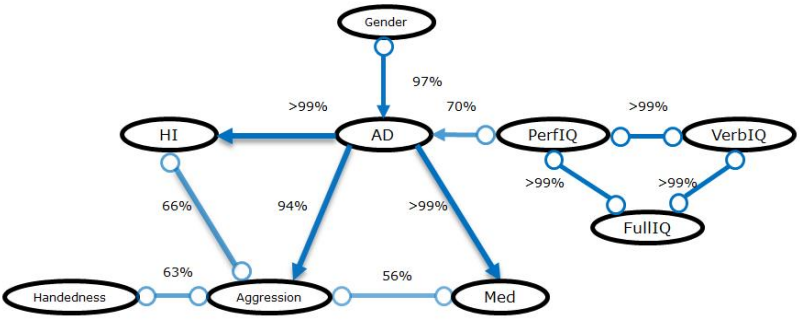


Fig. 6. The causal graph representing causal relationships between variables for the ADHD data set. The graph represents a PAG, where edge directions are marked with " – " and " >" for invariant edge directions and with "o" for non-invariant edge directions. The reliability of an edge between two variables is depicted with a percentage value near each edge.

7 Conclusion and Future Work

In this paper we extended the state-of-the-art BCCD algorithm for causal discovery to mixtures of discrete and continuous variables. In order to extend BCCD, we developed a method to estimate the score of the DAG for data sets with a mixture of discrete and continuous variables, under the assumption that relationships between variables are monotonic. The developed method approximates

mutual information using Spearman correlation to calculate the BIC score. Simulation studies on different types of data show that this appears to be a reasonable approximation of mutual information. The small bias introduced to it hardly influences the outcome of the BCCD algorithm.

The most computationally expensive part of the proposed scoring method is the calculation of the correlation matrix. However, to estimate the BIC score for all possible subsets of variables it is not necessary to recalculate correlation matrix for each subset. The correlation matrix between all variables can be calculated beforehand and the required elements of the matrix can be selected depending on a subset of variables used to calculate the BIC score. As a result, the proposed scoring method does not require expensive calculations and requires to store only the correlation matrix, which makes it fast and memory efficient.

Our extended version of BCCD is now able to learn causal relationships between a mixture of discrete and continuous variables, assuming that the relations between them are monotonic. For future work we will relax this assumption for discrete variables and learn non-monotonic relations, by introducing a preliminary transformation to combine the parents of the variables. Another possible extension of the method is to handle missing values. This is a quite common problem in many fields, while standard solutions such as data deletion or data imputation lead to loss of information or introduction of bias into the results.

An alternative approach for structure learning is through conditional independence tests. Several methods [31,14] employing kernels have been proposed to measure conditional independence without assuming a functional form of dependency between the variables as well as the data distributions. It would be interesting to consider how we can use these tests to obtain alternative scoring methods for the BCCD algorithms.

Acknowledgments. The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 278948.

References

1. Bach, F.R., Jordan, M.I.: Learning graphical models with Mercer kernels. In: Proceedings of the NIPS Conference, pp. 1009–1016 (2002)
2. Bauermeister, J.J., Shrout, P.E., Chávez, L., Rubio-Stipec, M., Ramírez, R., Padilla, L., Anderson, A., García, P., Canino, G.: ADHD and Gender: Are risks and sequela of ADHD the same for boys and girls? Journal of Child Psychology and Psychiatry 48(8), 831–839 (2007)
3. Cao, Q., Zang, Y., Sun, L., Sui, M., Long, X., Zou, Q., Wang, Y.: Abnormal neural activity in children with attention deficit hyperactivity disorder: a resting-state functional magnetic resonance imaging study. Neuroreport 17(10), 1033–1036 (2006)
4. Chickering, D.M.: Optimal structure identification with greedy search. Journal of Machine Learning Research 3, 507–554 (2002)

5. Chickering, D.M., Geiger, D., Heckerman, D.: Learning Bayesian networks: Search methods and experimental results. In: Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics, pp. 112–128 (January 1995)
6. Claassen, T., Heskes, T.: A Bayesian approach to constraint based causal inference. In: Proceedings of the UAI Conference, pp. 207–216. AUAI Press (2012)
7. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley Series in Telecommunications and Signal Processing. Wiley-Interscience (2006)
8. Daly, R., Shen, Q., Aitken, J.S.: Learning Bayesian networks: approaches and issues. Knowledge Eng. Review 26(2), 99–157 (2011)
9. Dawid, A.P.: Statistical theory: the prequential approach (with discussion). J. R. Statist. Soc. A 147, 278–292 (1984)
10. de Campos, L.M., Huete, J.F.: Approximating causal orderings for Bayesian networks using genetic algorithms and simulated annealing. In: Proceedings of the Eighth IPMU Conference, pp. 333–340 (2000)
11. de Santana, Á.L., Francês, C.R.L., Costa, J.C.W.: Algorithm for graphical Bayesian modeling based on multiple regressions. In: Gelbukh, A., Kuri Morales, Á.F. (eds.) MICAI 2007. LNCS (LNAI), vol. 4827, pp. 496–506. Springer, Heidelberg (2007)
12. Friedman, N., Goldszmidt, M.: Discretizing continuous attributes while learning Bayesian networks. In: Proceedings of the ICML Conference, pp. 157–165 (1996)
13. Geiger, D., Heckerman, D.: Learning Gaussian networks. In: Proceedings of the UAI Conference, pp. 235–243. Morgan Kaufmann (1994)
14. Gretton, A., Fukumizu, K., Teo, C.H., Song, L., Schölkopf, B., Smola, A.J.: A kernel statistical test of independence. In: NIPS. Curran Associates, Inc. (2007)
15. Harris, N., Drton, M.: PC algorithm for nonparanormal graphical models. Journal of Machine Learning Research 14, 3365–3383 (2013)
16. Heckerman, D., Geiger, D., Chickering, D.M.: Learning Bayesian networks: The combination of knowledge and statistical data. In: Machine Learning, pp. 197–243 (1995)
17. Larrañaga, P., Kuijpers, C.M.H., Murga, R.H., Yurramendi, Y.: Learning Bayesian network structures by searching for the best ordering with genetic algorithms. IEEE Transactions on Systems, Man and Cybernetics 26, 487–493 (1996)
18. Lauritzen, S., Spiegelhalter, D.J.: Local computations with probabilities on graphical structures and their application to expert systems (with discussion). Journal of the Royal Statistical Society Series B 50, 157–224 (1988)
19. Lauritzen, S.L., Lauritzen, S.L.: Propagation of probabilities, means and variances in mixed graphical association models. Journal of the American Statistical Association 87, 1098–1108 (1992)
20. Monti, S., Cooper, G.F.: Learning hybrid Bayesian networks from data. Technical Report ISSP-97-01, Intelligent Systems Program, University of Pittsburgh (1997)
21. Monti, S., Cooper, G.F.: A multivariate discretization method for learning Bayesian networks from mixed data. In: Cooper, G.F., Moral, S. (eds.) Proceedings of the UAI Conference, pp. 404–413. Morgan Kaufmann (1998)
22. Paloyelis, Y., Rijsdijk, F., Wood, A., Asherson, P., Kuntsi, J.: The genetic association between adhd symptoms and reading difficulties: the role of inattentiveness and IQ. J. Abnorm. Child Psychol. 38, 1083–1095 (2010)
23. Pearl, J., Verma, T.: A theory of inferred causation. In: Proceedings of the KR Conference, pp. 441–452. Morgan Kaufmann (1991)
24. Pellet, J.P., Elisseeff, A.: Using Markov blankets for causal structure learning. J. Mach. Learn. Res. 9, 1295–1342 (2008)
25. Ramsey, J., Zhang, J., Spirtes, P.: Adjacency-faithfulness and conservative causal inference. In: Proceedings of the UAI Conference, pp. 401–408. AUAI Press (2006)

26. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics 6, 461–464 (1978)
27. Spirtes, P., Glymour, C., Scheines, R.: Causation, Prediction, and Search, 2nd edn. MIT Press (2000)
28. Spirtes, P., Glymour, C., Scheines, R.: The TETRAD Project: Causal Models and Statistical Data. Carnegie Mellon University Department of Philosophy, Pittsburgh (2004)
29. Vaida, N., Mattoo, N.H., Wood, A., Madhosh, A.: Intelligence among attention deficit hyperactivity disordered (adhd) children (aged 5-9). J. Psychology 4(1), 9–12 (2013)
30. Willcutt, E., Pennington, B., DeFries, J.: Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. J. Abnorm. Child Psychol. 28(2), 149–159 (2000)
31. Zhang, K., Peters, J., Janzing, D., Schölkopf, B.: Kernel-based conditional independence test and application in causal discovery. CoRR 1202.3775 (2012)