Causal discovery from databases with discrete and continuous variables[1]

Elena Sokolova [a] Perry Groot [a] Tom Claassen [a] Tom Heskes[a]

[a] *Radboud University, Faculty of Science,*
Postbus 9010, 6500 GL Nijmegen, The Netherlands

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) [1] 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. 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 and/or memory inefficient.

2 Methods

In this paper we propose a fast and memory efficient method to score DAGs with both discrete and continuous variables using BIC score, 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 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. To estimate BIC 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 [2]:

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

[1]The full paper has been published in *Proceedings of the Seventh European Workshop on Probabilistic Graphical Models.*, 2014.

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 parameters in the regression model.

In addition to (1) we propose substituting Pearson correlation with Spearman. We assume hereby explicitly that the variables obey a so-called non-paranormal distribution: a multivariate Gaussian distribution on latent variables, each of which is related to the observed variables through a monotonic transformation. As shown in [3], 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, this bias hardly affects the scoring of the network structures.

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.

3 Results

To test our algorithm on simulated data, we chose two widely used Bayesian networks: the Asia Network and the Waste Incinerator Network to test the algorithm for discrete variables and a mixture of discrete and continuous variables, respectively. We randomly generated data for four different sample sizes: 100, 500, 1000, and 1500 and repeated our experiments 20 times. Performance was 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. Simulation studies showed that the proposed scoring method appears to rely on a reasonable approximation of mutual information. The small bias introduced to it hardly influences the outcome of the BCCD algorithm. To test the BCCD algorithm on real-world data, we used the data set collected for ADHD-200 competition that contains a mixture of discrete and continuous variables. The resulting network inferred by the BCCD algorithm provided the causal relationships between different factors and symptoms. Several publication in medical journals have been found that confirm these relationships. Based on this results, we conclude that the BCCD algorithm with the proposed scoring method can accurately estimate the structure of the Bayesian network for both simulated and real-world data.

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

[2] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, 2006.

[3] N. Harris and M. Drton. PC algorithm for nonparanormal graphical models. *Journal of Machine Learning Research*, 14:3365–3383, 2013.