# Causal Discovery from Medical Data: Dealing with Missing Values and a Mixture of Discrete and Continuous Data

Elena Sokolova[1(✉)], Perry Groot[1], Tom Claassen[1], and Daniel von Rhein[2], and Jan Buitelaar[2] and Tom Heskes[1]

[1] Faculty of Science, Radboud University, Nijmegen, The Netherlands
e.sokolova@cs.ru.nl
[2] Donders Institute for Brain, Cognition and Behaviour, Radboud University Medical Center, Nijmegen, The Netherlands

**Abstract.** Causal discovery is an increasingly popular method for data analysis in the field of medical research. In this paper we consider two challenges in causal discovery that occur very often when working with medical data: a mixture of discrete and continuous variables and a substantial amount of missing values. To the best of our knowledge there are no methods that can handle both challenges at the same time. In this paper we develop a new method that can handle these challenges based on the assumption that data is missing completely at random and that variables obey a non-paranormal distribution. We demonstrate the validity of our approach for causal discovery for empiric data from a monetary incentive delay task. Our results may help to better understand the etiology of attention deficit-hyperactivity disorder (ADHD).

**Keywords:** Causal discovery · Missing data · Mixture of discrete and continuous data · ADHD

## 1 Introduction

In recent years, the use of causal discovery in the field of medical research has become increasingly popular. Existing algorithms deal reasonably well with models that contain only discrete variables or only Gaussian variables, while real-world data often contains mixture variables, where continuous variables are not Gaussian. To tackle the problem of mixture variables, several approaches have been developed, including partial correlation tests [7], Mercer kernels [2], and neural networks [8]. Existing algorithms for causal discovery mainly start from the assumption that the data is complete, whereas in practice, medical data sets often have missing values. For example, some tests are often performed only for part of the patients, the quality of some data is poor, participants drop out etc. Several methods have been proposed to deal with missing values for causal discovery, including imputation methods, expectation maximization algorithms, and importance sampling [6,9]. However, these methods usually rely on the assumption that data is either discrete or continuous Gaussian. To the best of our


J.H. Holmes et al. (Eds.): AIME 2015, LNAI 9105, pp. 177–181, 2015.
DOI: 10.1007/978-3-319-19551-3_23

knowledge there are no methods that can handle both a mixture of discrete and continuous variables as well as missing values for directed graphical models. An ad hoc solution would be to apply standard methods and ignore all missing values. However, when the percentage of missing values is significant one can end up with the insufficient data to learn a causal structure.

We propose a method that can handle missing values and mixture variables based on the ideas for undirected graphical models presented in [12,1]. As a prototypical example we apply our algorithm to a data set of patients with Attention-deficit/hyperactivity disorder (ADHD). The data set is ideal for our purposes because it provides all characteristics of a typical medical data set: it describes relationships between various possible factors of the disease such as genes, age, gender, and different types of symptoms and behavioral characteristics. Moreover, it has a mixture of discrete and continuous variables and approximately 10% of missing data.

The rest of the paper is organized as follows. Section 2 explains the proposed method. Section 3 presents the results for ADHD data. Section 4 provides our conclusion and future work.

## 2 Proposed Method

In this section we propose a causal discovery algorithm that can deal with both a mixture of discrete and continuous variables and missing data. In the first two steps of this algorithm we estimate the correlation matrix for mixed data with missing values, based on the ideas proposed in [12,1]. In the third step, we use this correlation matrix as an input into a causal discovery algorithm to infer the causal structure. We use the BCCD algorithm for this purpose, one of the state-of-the-art algorithms in causal discovery. Claassen and Heskes [3] showed that BCCD outperforms reference algorithms in the field, such as FCI and Conservative PC. Moreover, BCCD provides an indication of the reliability of the causal links that makes it easier to interpret the results and compare alternative models. We rely on the assumption that data is missing completely at random and that variables obey a non-paranormal distribution.

**Step 1 Mixture of discrete and continuous variables**
To deal with mixed data we propose to use a Gaussian copula. For each variable $X_i$ we estimate the rescaled empirical distribution

$$\hat{F}_i(x) = \frac{1}{n+1} \sum_{i=1}^{n} \mathcal{I}\{X_i < x\}, \tag{1}$$

and then transform the data into Gaussian normal scores $\hat{X}_i = \hat{\Phi}_i^{-1}(\hat{F}(X_i))$. In this step missing values are ignored.

**Step 2 Correlation matrix with missing data**
New variables now have a Gaussian distribution, so we can use Pearson correlation to estimate dependencies between variables. Since our data has

missing values, we propose to first use the expectation maximization algorithm to estimate the correlation matrix, since this algorithm provides an unbiased estimate of parameters and their standard error [4].

**Step 3 Apply BCCD**

Having the correlation matrix we can use it in the BCCD algorithm to estimate the causal structure of the graph. A more detailed description can be found in [3,10]. The core part of the algorithm BCCD is the estimation of the reliability of the causal relations, based on the marginal likelihood of data given graphical structure $p(\mathbf{D}|\mathcal{G})$. We approximate the logarithm of the marginal likelihood $\log(p(\mathbf{D}|\mathcal{G}))$ using the Bayesian Information Criterion (BIC), which depends on the correlation matrix computed in Step 2:

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

where $n$ is the number of variables, $M$ is the sample size, $\mathrm{Dim}[\mathcal{G}]$ is the number of parameters in the model corresponding to graph $\mathcal{G}$, $Pa_i$ are the parents of node $X_i$, and $\Sigma_{Pa_i}$ is a correlation matrix between the parents of variable $X_i$. Having all the causal relations, BCCD ranks them in decreasing order of reliability and uses logical deduction with transitivity and acyclicity to derive additional causal statements. If there is a conflict, it picks the causal statement that has a higher reliability.

## 3 ADHD Data

In a future paper we will analyze the performance of our method on simulated data, where we know the ground truth. Here we describe an application of our algorithm to the ADHD data set that was collected as a part of the NeuroIMAGE study [11]. This study investigated the brain response during reward anticipation and receipt with a monetary incentive delay (MID) task in a large sample of adolescents and young adults with ADHD, their unaffected siblings and healthy controls. The brain activation was measured in ventral striatum (VS) and orbitalfrontal cortex (OFC) brain areas during the anticipation of the reward cue and feedback after reward cue. During the experiments the difference in reaction time was measured when there was and there was no reward cue (Reaction time difference). The data set contained 189 probands with ADHD, 104 unaffected siblings, and 116 age-matched controls. Since the presence of the unaffected siblings can blur the effect of the genes, we did not include them in our study and consider only ADHD patients and healthy controls.

The goal of our study is to identify the endophenotypic model [5] that explains the relationships between genes, brain functioning, behaviors, and disease symptoms. To apply causal discovery to this data set, we selected 12 variables that represent genes, brain functioning in different regions of the brain, symptoms, and general factors. We included the prior knowledge that no variable in the network can cause gender, and the endophenotypic assumption from [5] that symptoms are the consequence of the brain functioning problems.

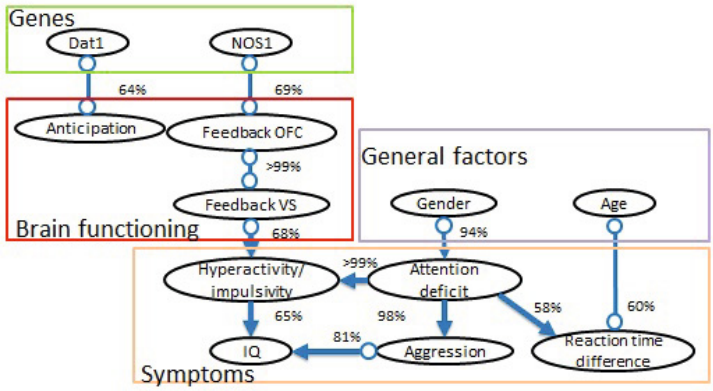


**Fig. 1.** The causal graph representing causal relationships between variables for the ADHD data set. The graph represents a PAG, where edge directions are marked with " – " and " >" for invariant edge directions and with "○" for non-invariant edge directions. The reliability of an edge between two variables is depicted with a percentage value near each edge.

A causal network learned from the data is presented in Figure 1. This figure includes only edges with a reliability of a direct causal link higher than 50%. The graph built by BCCD shows an effect of genes on brain functioning, the effect of brain functioning and general factors on disease symptoms, and an interaction between these symptoms. The relationships between variables found by BCCD are in line with several other studies in ADHD that considered mainly pairs of variables [13]. An additional advantage of this study is that interactions are visualized in a single graph that makes it easy to interpret the results.

## 4 Conclusions and Discussion

The contribution of this paper is the presentation of an algorithm for causal discovery and application of it to real-world data describing ADHD. The results of the algorithm were corroborated by medical experts and literature. As any statistical approach, methods for causal discovery have to rely on assumptions. In this paper we relaxed standard assumptions on, for example, Gaussianity and the absence of missing values. By doing this we open up the application of causal discovery to a much wider class of data sets. However, we obviously rely on some other assumptions such as data being missing completely at random, monotonic interactions between variables, and absence no cycles. As future work we would like to consider data sets with more complex interactions between variables and relax in particular the assumption that data is missing completely at random.

**Acknowledgments.** This work was supported by NIH Grant R01MH62873 (to Stephen V. Faraone), NWO Large Investment Grant (1750102007010 to Jan Buitelaar), NWO Brain & Cognition grants (056-13-015 and 433-09-242 to Jan Buitelaar), and grants from

Radboud University Nijmegen Medical Center, University Medical Center Groningen and Accare and VU University Amsterdam. The research leading to these results also received support from the EU FP7 project TACTICS under grant agreement n°278948, the NWO grants MoCoCaDi (612.001.202) and CHILL (617.001.451).

## References

1. Abegaz, F., Wit, E.: Penalized EM algorithm and copula skeptic graphical models for inferring networks for mixed variables. Statistics in Medicine (2014)
2. Bach, F.R., Jordan, M.I.: Learning graphical models with Mercer kernels. In: Proceedings of the NIPS Conference, pp. 1009–1016 (2002)
3. Claassen, T., Heskes, T.: A Bayesian approach to constraint based causal inference. In: Proceedings of the UAI Conference, pp. 207–216 (2012)
4. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), pp. 1–38 (1977)
5. Franke, B., Neale, B.M., Faraone, S.V.: Genome-wide association studies in ADHD. Human Genetics 126(1), 13–50 (2009)
6. Friedman, N.: The bayesian structural EM algorithm. In: Proceedings of the Fourteenth
Conference on Uncertainty in Artificial Intelligence, pp. 129–138. Morgan Kaufmann Publishers Inc. (1998)
7. Harris, N., Drton, M.: PC algorithm for nonparanormal graphical models. Journal of Machine Learning Research 14, 3365–3383 (2013)
8. Monti, S., Cooper, G.F.: Learning hybrid Bayesian networks from data. Technical Report ISSP-97-01, Intelligent Systems Program, University of Pittsburgh (1997)
9. Riggelsen, C., Feelders, A.: Learning bayesian network models from incomplete data using importance sampling. In: Proc. of Artificial Intelligence and Statistics, pp. 301–308 (2005)
10. Sokolova, E., Groot, P., Claassen, T., Heskes, T.: Causal discovery from databases with discrete and continuous variables. In: van der Gaag, L.C., Feelders, A.J. (eds.) PGM 2014. LNCS, vol. 8754, pp. 442–457. Springer, Heidelberg (2014)
11. von Rhein, D., Mennes, M., van Ewijk, H., Groenman, A.P., Zwiers, M.P., Oosterlaan, J., Heslenfeld, D., Franke, B., Hoekstra, P.J., Faraone, S.V.: et al. The NeuroIMAGE study: a prospective phenotypic, cognitive, genetic and MRI study in children with attention-deficit/hyperactivity disorder. Design and descriptives. European Child & Adolescent Psychiatry, 1–17 (2014)
12. Wang, H., Fazayeli, F., Chatterjee, S., Banerjee, A., Steinhauser, K., Ganguly, A., Bhattacharjee, K., Konar, A., Nagar, A.: Gaussian copula precision estimation with missing values. Biotechnology Journal 4(9) (2009)
13. Willcutt, E.G., Pennington, B.F., DeFries, J.C.: Etiology of inattention and hyperactivity/impulsivity in a community sample of twins with learning difficulties. J. Abnorm. Child Psychol. 28(2), 149–159 (2000)