Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data
AALTD 2015

Causality on Longitudinal Data: Stable Specification Search in Constrained Structural Equation Modeling

Ridho Rahmadi[1,2], Perry Groot[2], Marianne Heins[4],
Hans Knoop[3], and Tom Heskes[2]

[1] Department of Informatics, Universitas Islam Indonesia.
[2] Institute for Computing and Information Sciences, Radboud University Nijmegen.
[3] Expert Centre for Chronic Fatigue, Radboud University Medical Centre, Nijmegen.
[4] Netherlands Institute for Health Services Research, Utrecht.
r.rahmadi@cs.ru.nl

Abstract Developing causal models from observational longitudinal studies is an important, ubiquitous problem in many disciplines. A disadvantage of current causal discover algorithms, however, is the inherent instability in structure estimation. With finite data samples small changes in the data can lead to completely different optimal structures. The present work presents a new causal discovery algorithm for longitudinal data that is robust for finite data samples. We validate our approach on a simulated data set and real-world data on Chronic Fatigue Syndrome patients.

Keywords: Longitudinal data, Causal modeling, Structural equation model, Stability selection, Multi-objective evolutionary algorithm.

1 Introduction

Developing causal models from observational longitudinal studies is an important, ubiquitous problem in many disciplines, which has led to the development of a variety of causal discovery algorithms in the literature [1–5]. A disadvantage of current causal discovery algorithms, however, is the inherent instability in structure learning. With finite data samples small changes in the data can lead to completely different optimal structures, since errors made by the discovery algorithm may be propagated and lead to further errors [6]. In [7] we developed a robust causal discovery algorithm for cross-sectional data. The method performs structure search over Structural Equation Models (SEMs) by maximizing model scores in terms of data fit and complexity. The present work extends our causal discovery algorithm to longitudinal data. We describe how longitudinal causal relationships can be modelled for an arbitrary number of time slices. Furthermore, we show how a longitudinal causal model can easily be scored using standard SEM software by data reshaping. The algorithm produces accurate structure estimates and is shown to be robust for finite samples. We validate our approach on one simulated longitudinal data set and one real-world longitudinal data set for Chronic Fatigue Syndrome.

Copyright©2015 for this paper by its authors. Copying permitted for private and academic purposes.

2 Proposed method

We use a SEM for causal modeling. The general form of the equations is

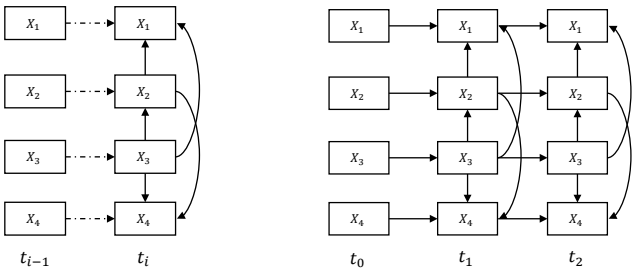
$$x_i = f_i(\mathrm{pa}_i, \varepsilon_i), \quad i = 1, \ldots, n. \tag{1}$$

where pa_i denotes the *parents* which represent the set of variables considered to be direct causes of X_i and ε_i represents errors on account of omitted factors that are assumed to be mutually independent [8]. In this study, we focus on causal models with no reciprocal relationships, and no latent variables. Thus the causal model can also be represented by a *Directed Acyclic Graph* (DAG). We score models using both the *chi-square* χ^2 (measuring the data fit) and the *model complexity* (measuring the number of parameters).

We use the method we developed in [7] to perform exploratory search over SEM models. Based on the idea of stability selection [9], the method subsamples the data D with size $\lfloor |D|/2 \rfloor$ without replacement and generates Pareto optimal models for each subset. After that, all Pareto optimal models are transformed into their corresponding model equivalent classes, called *Completed Partially Directed Acyclic Graph* (CPDAG) [10]. From these CPDAGs we compute the edge and causal path stability graph, such as Figure 3a, by grouping them according to model complexity and computing their *selection probability*, i.e., the number of occurrences divided by the total number of models for a certain level of model complexity. Stability selection is then performed by specifying two thresholds, π_{sel} (boundary of selection probability) and π_{bic} (boundary of complexity). For example, setting $\pi_{\mathrm{sel}} = 0.6$ means that all causal relationships with edge stability or causal path stability (Figure 3) above this threshold are considered *stable*. The second threshold π_{bic} is used to control overfitting. We set π_{bic} to the level of model complexity at which the minimum average *Bayesian Information Criterion* (BIC) score is found. For example, $\pi_{\mathrm{bic}} = 7$ means that all causal relationships with an edge stability or a causal path stability lower than this threshold (Figure 3) are considered *parsimonious*. Causal relationships that intersect with the top-left region are considered both stable and parsimonious and called *relevant*, from which we can derive a causal model.

The method in [7] only handles cross-sectional data. Based on the idea of "unrolling" the network in Dynamic Bayesian Networks [4, 5], we extended the method to handle longitudinal data. We model longitudinal causal relationships with a SEM model consisting of two time slices (Figure 1a) that can be "unrolled" into a network with an arbitrary number of time slices (Figure 1b). Time slice t_i represents the relationships *within* a time slice (intra-slice causal relationships, solid arcs in Figure 1a). Causal relationships *between* time slices (inter-slice causal relationships, dashed arcs in Figure 1a) always go forward in time, i.e., from time slice t_{i-1} to time slice t_i.

To score our models on longitudinal data we use data reshaping. In the reshaped data, the first n data points contain the relations that occur in the first two time slices t_0 and t_1. The next n data points contain the relations that occur in time slices t_1 and t_2. The i-th subset of n data points contain



(a) (b)

Figure 1: (a) The longitudinal causal model. (b) The "unrolled" causal graph used to generate longitudinal data. It contains four continuous variables $(X_1, \ldots, X_4)$ in three different time slices $t_0, \ldots, t_2$.

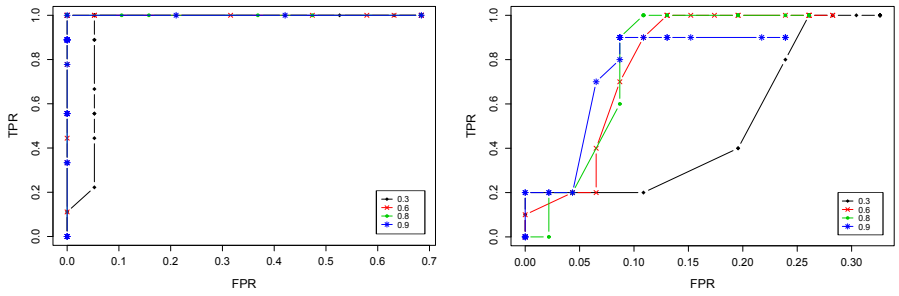
the relations in time slices t_{i-1} and t_i. The reshaped data then allows us to use standard SEM software to compute the scores.

3 Application to Simulated Data

For this experiment, we generated a longitudinal data set with 400 instances from a causal graph as depicted in Figure 1b. The data set consists of three time slices with four continuous variables for each time slice.[1] When we searched over SEM models we added prior knowledge that variables X_1 and X_2 do not cause variable X_3 directly. We performed the search over 200 subsets.

As the true model is known, we measure the performance of our method by means of the *Receiver Operating Characteristic* (ROC) [11] for both edges and causal paths. The threshold π_{sel} is fixed to a value ($\pi_{\text{sel}} \in \{0.3, 0.6, 0.8, 0.9\}$) while π_{bic} is varied. We compute the *True Positive Rate* (TPR) and the *False Positive Rate* (FPR) from the CPDAG of the true model. As for an example, in the case of edge stability, a true positive means that an edge that appears within the top-left region bounded by π_{sel} and π_{bic} also exists in the CPDAG of the true model. Figure 2 portrays the ROC curves for both edge and causal path stability. Generally we can see that higher values of π_{sel} tend to give better ROC curves. This suggests that our approach is able to find the underlying structure with high reliability scores. A notable point is that the ROC curves stop at a TPR and/or FPR value lower than 1. Since some of the edges and paths are disallowed (i.e., no edges in time-slice t_{i-1} and no paths from t_i to t_{i-1}) some of the edges and causal paths in the stability graphs end up with a selection probability of 0 and the result is that the ROC curves cannot reach the upper right corner with TPR = FPR = 1.

[1] Available at http://bit.ly/1L6dBOo



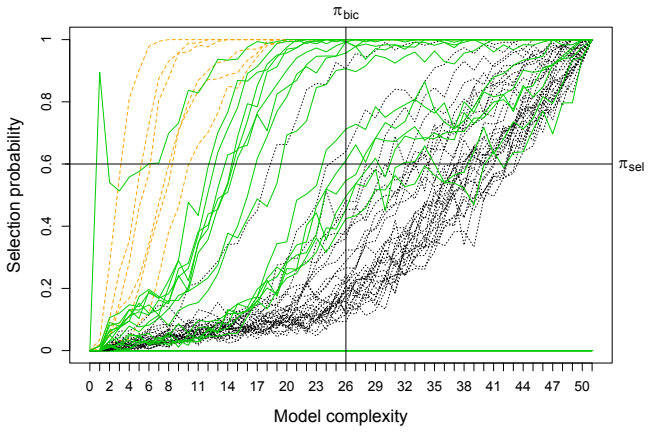
(a) (b)

Figure 2: A plot of ROC curves for (a) the edge stability and (b) the causal path stability, for different values of π_{sel}. A higher π_{sel} shows a better ROC curve. See the main text for an explanation why the ROC curves stop at some point.

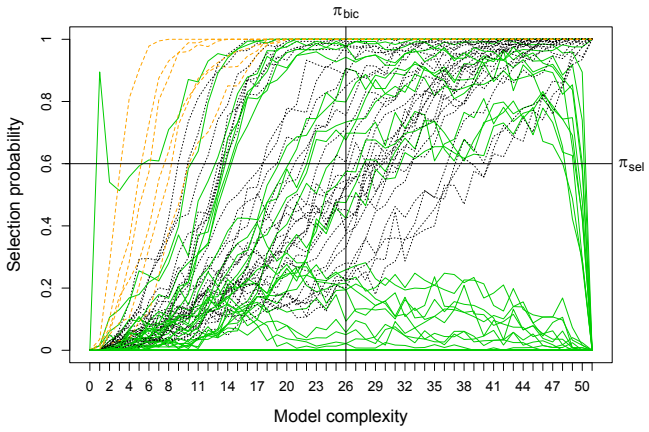
4 Application to Real-world Data

For an application to real-world data, we consider a data set about *Chronic Fatigue Syndrome* (CFS) which consists of 183 subjects and five time slices with six discrete variables [12]. The variables are, *fatigue* severity, the sense of *control* over fatigue, *focusing* on the symptoms, the objective activity of the patient (*oActivity*), the subject's perceived activity (*pActivity*), and the physical *functioning*. We use *Expectation Maximization* implemented in SPSS [13] to impute the missing values. As all of the variables have large scales, e.g., in the range between 0 to 155, we treat them as continuous variables. We added prior knowledge that the variable fatigue does not cause any of the other variables directly. We performed the search over 200 subsets.

Figure 3a shows that nineteen relevant edges were found, consisting of eleven intra-slice and eight inter-slice relationships which among of these, six are between the same variables and two are between different variables. Figure 3b shows that thirty-two relevant causal paths were found, consisting of twelve intra-slice and twenty inter-slice relationships which among of these, six are between the same variables and fourteen are between different variables. For a more intuitive representation, we combine the stability graphs into a model using the following procedure. First, the nodes are linked according to the nineteen relevant edges. Second, edges are oriented according to our background knowledge. Eight of the inter-slice relationships are oriented from time slice t_{i-1} to t_i and five of the intra-slice edges can be oriented since it is known that the variable fatigue does not directly cause any other variable. Third, the edges are oriented according to the relevant causal paths, which results in another twenty-eight directed edges. The inferred model is shown in Figure 4. Each edge is annotated with a reliability score which is the maximum score obtained in the top-left region of the edge stability graph.



(a)



(b)

Figure 3: The stability graphs for CFS together with π_{sel} and π_{bic}, yielding four regions. The top-left region contains the relevant causal relations. (a) The edge stability graph. (b) The causal path stability graph. Orange-dashed lines represent inter-slice relationships between the same variables, black-dotted lines represent inter-slice relationships between different variables, green-solid lines represent intra-slice relationships.

From the stability graphs we can see that the most stable causal relations are the inter-slice relations between the same variables followed by some of the intra-slice causal relations. Almost all of the inter-slice relations between different variables are not considered relevant. A directed edge $X \to Y$ in Figure 4 indicates

that a change in variable X causes a change in variable Y. In the intra-slice causal relationships, we found that all variables are direct causes for fatigue severity. We also found all variables, except *fatigue*, to be direct causes for the perceived activity. Furthermore, the variable *control* is a direct cause for both focusing on the symptoms and physical functioning. Generally the inter-slice relationships show direct causes between the same variables. In addition, the variables *pActivity* and *control* indicate a stronger direct cause for fatigue severity and focusing on symptoms, respectively, as they contribute a direct cause in both time slices. The inferred model is consistent with results reported in the medical literature [12, 14, 15].

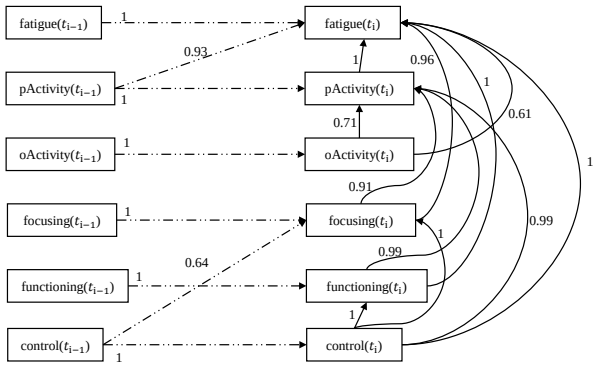


Figure 4: The inferred model of CFS by combining the edge stability and causal path stability graphs.

5 Conclusion

Causal discovery from longitudinal data is an important, ubiquitous problem in science. Current causal discovery algorithms, however, have difficulty dealing with the inherent instability in structure estimation. The present work introduces a new discovery algorithm for longitudinal data that is robust for finite samples. Experimental results on both artificial and real-world data sets show that the method results in reliable structure estimates. Future research will aim to estimate the size of causal effects.

Acknowledgments

The research leading to these results has received funding from the DGHE of Indonesia and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 305697.

References

1. Riva, A., Bellazzi, R.: Learning temporal probabilistic causal models from longitudinal data. Artificial Intelligence in Medicine **8**(3) (1996) 217–234
2. Parner, J., Arjas, E.: Causal reasoning from longitudinal data. Rolf Nevanlinna Inst., University of Helsinki (1999)
3. Marsh, H.W., Yeung, A.S.: Causal effects of academic self-concept on academic achievement: Structural equation models of longitudinal data. Journal of educational psychology **89**(1) (1997) 41–54
4. Friedman, N., Murphy, K., Russell, S.: Learning the structure of dynamic probabilistic networks. In: Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence, Morgan Kaufmann Publishers Inc. (1998) 139–147
5. Murphy, K., Mian, S., et al.: Modelling gene expression data using dynamic Bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA (1999)
6. Spirtes, P.: Introduction to causal inference. The Journal of Machine Learning Research **11** (2010) 1643–1662
7. Rahmadi, R., Groot, P., Heins, M., Knoop, H., Heskes, T., The OPTIMISTIC consortium: Causality on cross-sectional data: Stable specification search in constrained structural equation modeling. arXiv:1506.05600 [stat.ML] (2015)
8. Pearl, J.: Causality: models, reasoning and inference. Cambridge Univ Press (2000)
9. Meinshausen, N., Bühlmann, P.: Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **72**(4) (2010) 417–473
10. Chickering, D.M.: Learning equivalence classes of Bayesian-network structures. The Journal of Machine Learning Research **2** (2002) 445–498
11. Fawcett, T.: ROC graphs: Notes and practical considerations for researchers. Machine learning **31** (2004) 1–38
12. Heins, M.J., Knoop, H., Burk, W.J., Bleijenberg, G.: The process of cognitive behaviour therapy for chronic fatigue syndrome: Which changes in perpetuating cognitions and behaviour are related to a reduction in fatigue? Journal of psychosomatic research **75**(3) (2013) 235–241
13. IBM Corp. Armonk, NY: IBM SPSS Statistics for Windows, Version 19.0. (2010)
14. Vercoulen, J., Swanink, C., Galama, J., Fennis, J., Jongen, P., Hommes, O., Van der Meer, J., Bleijenberg, G.: The persistence of fatigue in chronic fatigue syndrome and multiple sclerosis: development of a model. Journal of psychosomatic research **45**(6) (1998) 507–517
15. Wiborg, J.F., Knoop, H., Frank, L.E., Bleijenberg, G.: Towards an evidence-based treatment model for cognitive behavioral interventions focusing on chronic fatigue syndrome. Journal of psychosomatic research **72**(5) (2012) 399–404