International Journal of Approximate Reasoning ••• (••••) •••–•••

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

A constraint optimization approach to causal discovery from subsampled time series data ☆

Antti Hyttinen [a,*], Sergey Plis [b], Matti Järvisalo [a], Frederick Eberhardt [c], David Danks [d]

[a] HIIT, Department of Computer Science, University of Helsinki, Finland
[b] Mind Research Network and University of New Mexico, United States
[c] Humanities and Social Sciences, California Institute of Technology, United States
[d] Department of Philosophy, Carnegie Mellon University, United States

ARTICLE INFO

Article history:
Received 1 December 2016
Received in revised form 30 June 2017
Accepted 10 July 2017
Available online xxxx

Keywords:
Causality
Causal discovery
Graphical models
Time series
Constraint satisfaction
Constraint optimization

ABSTRACT

We consider causal structure estimation from time series data in which measurements are obtained at a coarser timescale than the causal timescale of the underlying system. Previous work has shown that such subsampling can lead to significant errors about the system's causal structure if not properly taken into account. In this paper, we first consider the search for system timescale causal structures that correspond to a given measurement timescale structure. We provide a constraint satisfaction procedure whose computational performance is several orders of magnitude better than previous approaches. We then consider finite-sample data as input, and propose the first constraint optimization approach for recovering system timescale causal structure. This algorithm optimally recovers from possible conflicts due to statistical errors. We then apply the method to real-world data, investigate the robustness and scalability of our method, consider further approaches to reduce underdetermination in the output, and perform an extensive comparison between different solvers on this inference problem. Overall, these advances build towards a full understanding of non-parametric estimation of system timescale causal structures from subsampled time series data.

1. Introduction

Time-series data has long constituted the basis for causal modeling in many fields of science [12,15,22]. These data often provide very precise measurements at regular time points, but the underlying causal interactions that give rise to those measurements can occur at a much faster timescale than the measurement frequency. As just one example: fMRI experiments measure neural activity (given various assumptions) roughly once per two seconds, but the underlying neural connections clearly operate much more quickly. Time order information can simplify causal analysis since it can provide

☆ This paper is part of the Virtual special issue on the Eighth International Conference on Probabilistic Graphical Models, Edited by Giorgio Corani, Alessandro Antonucci, Cassio De Campos.

* Corresponding author.
E-mail addresses: antti.hyttinen@helsinki.fi (A. Hyttinen), s.m.plis@gmail.com (S. Plis), matti.jarvisalo@helsinki.fi (M. Järvisalo), fde@caltech.edu (F. Eberhardt), ddanks@cmu.edu (D. Danks).

http://dx.doi.org/10.1016/j.ijar.2017.07.009
0888-613X/

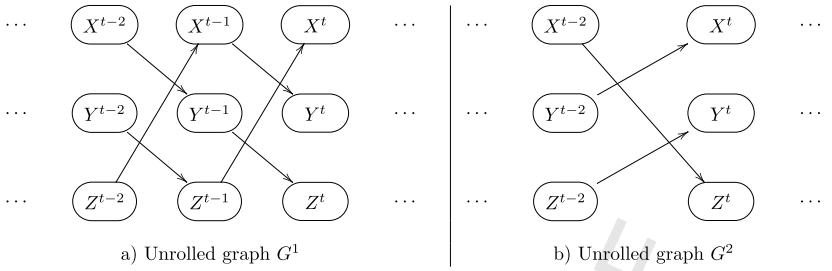

Fig. 1. (a) The structure of the causal system-scale time series. (b) The structure of the corresponding measurement scale time series if only every second sample is observed i.e. nodes at time slice $t-1$ are marginalized. If subsampling is ignored and (b) is thought to depict the true causal structure, all direct causal relationships among $\{X, Y, Z\}$ are misspecified.

directionality, but time series data that undersamples the generating process can be especially misleading about the true direct causal connections [7,19].

For example, Fig. 1a shows the causal structure of a process unrolled over discrete time steps, and Fig. 1b shows the corresponding structure of the same process, obtained by marginalizing every second time step. If we do not take into account the possibility of subsampling, then we would conclude that Fig. 1b gives the correct structure — and thus totally miss the presences of all true edges. This drastic structure misspecification may lead us to perform a possibly costly intervention on Z to control Y, when the influence of Z on Y is, in fact, completely mediated by X and so, intervening on X would be a more effective choice. Also, a (parametric) model with the structure in Fig. 1b gives inaccurate predictions when intervening on both X and Z: the value of Y would be predicted to depend on Z and not on X, when in reality Y depends on X and not on Z.

Standard methods for estimating causal structure from time series either focus exclusively on estimating a transition model at the measurement timescale (e.g., Granger causality [12,13]) or combine a model of measurement timescale transitions with so-called "instantaneous" or "contemporaneous" causal relations that aim to capture interactions that are faster than the measurement process (e.g., SVAR [22,15,18]), though only very specific types of interactions can be captured with these latter models. In contrast, we follow Plis et al. [30,31] and Gong et al. [11], and explore the possibility of identifying (features of) the causal process at the true timescale from data that subsample this process.

Plis et al. [30,31] developed algorithms that can learn the set of causal timescale structures that could yield a given measurement timescale graph, either at a known or unknown undersampling rate. While these algorithms show that the inference problem is solvable, they face a number of computational challenges that limit their use. They do, however, show the importance of constraints for this problem, and so suggest that a constraint satisfaction approach might be more effective and efficient. Gong et al. [11] consider finding a linear SVAR from subsampled data. They show that if the error variables are non-Gaussian, the true causal effects matrix can be discovered even from subsampled data. However, their method is highly restricted in terms of numbers of variables and parametric form.

In this paper, we provide an exact discovery algorithm based on using a general-purpose Boolean constraint solver [4,10], and demonstrate that it is orders of magnitudes faster than the current state-of-the-art method by Plis et al. [31]. At the same time, our approach is much simpler and, as we show, it allows inference in more general settings. We then develop the approach to integrate possibly conflicting constraints obtained from the data. In addition to an application of the method to the real-world data, we investigate the robustness and scalability of our method, consider further approaches to reduce underdetermination in the output, and perform an extensive comparison between different solvers on this inference problem. Moreover, unlike the method by Gong et al. [11], our approach does not depend on a particular parameterization of the underlying model and scales to a more reasonable number of variables.

The code implementing the approach presented in this article, including the answer set programming and Boolean satisfiability encodings, is available at

http://www.cs.helsinki.fi/group/coreo/subsampled/.

This article considerably extends a preliminary version presented at the International Conference on Probabilistic Graphical Models 2016 (PGM 2016) [17]. Most noticeably, Sections 6–9 of this article provide entirely new contents, including a real-world case study (Section 6), an evaluation of the impact of the choice of constraint satisfaction and optimization solvers on the efficiency of the approach (Section 7), and a discussion on learning from mixed frequency data (Section 8). Furthermore, new simulations on accuracy and robustness (Section 5, Figures 7–9) are now included.

2. Representation

We assume that the system of interest relates a set of variables $\mathbf{V}^t = \{X^t, Y^t, Z^t, \ldots\}$ defined at discrete time points $t \in \mathbb{Z}$ with continuous ($\in \mathbb{R}^n$) or discrete ($\in \mathbb{Z}^n$) values [9]. We distinguish the representation of the true causal process at the *system or causal timescale* from the time series data that are obtained at the *measurement timescale*. Following Plis et al.

Please cite this article in press as: A. Hyttinen et al., A constraint optimization approach to causal discovery from subsampled time series data, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.07.009