International Journal of Approximate Reasoning 55 (2014) 926–939

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Two optimal strategies for active learning of causal models from interventional data

Alain Hauser [a,*], Peter Bühlmann [b]

[a] Department of Biology, Bioinformatics, University of Bern, CH-3012 Bern, Switzerland
[b] Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland

ARTICLE INFO

Article history:
Available online 9 December 2013

Keywords:
Active learning
Graphical model
Causal inference
Bayesian network
Interventions

ABSTRACT

From observational data alone, a causal DAG is only identifiable up to Markov equivalence. Interventional data generally improves identifiability; however, the gain of an intervention strongly depends on the intervention target, that is, the intervened variables. We present active learning (that is, optimal experimental design) strategies calculating optimal interventions for two different learning goals. The first one is a greedy approach using single-vertex interventions that maximizes the number of edges that can be oriented after each intervention. The second one yields in polynomial time a minimum set of targets of arbitrary size that guarantees full identifiability. This second approach proves a conjecture of Eberhardt (2008) [1] indicating the number of unbounded intervention targets which is sufficient and in the worst case necessary for full identifiability. In a simulation study, we compare our two active learning approaches to random interventions and an existing approach, and analyze the influence of estimation errors on the overall performance of active learning.

1. Introduction

Causal relationships between random variables are usually modeled by directed acyclic graphs (DAGs), where an arrow between two random variables, $X \to Y$, reveals the former (X) as a *direct* cause of the latter (Y). From observational data alone (that is *passively* observed data from the undisturbed system), directed graphical models are only identifiable up to Markov equivalence, and arrow directions (which are crucial for the causal interpretation) are in general not identifiable. Without the assumption of specific functional model classes and error distributions [2], the only way to improve identifiability is to use interventional data for estimation, that is data produced under a perturbation of the system in which one or several random variables are forced to specific values, irrespective of the original causal parents. Examples of interventions include random assignment of treatments in a clinical trial, or gene knockdown or knockout experiments in systems biology.

The investigation of observational Markov equivalence classes has a long tradition in the literature [3–5]. Hauser and Bühlmann [6] extended the notion of Markov equivalence to the interventional case and presented a graph-theoretic characterization of corresponding Markov equivalence classes for a given set of interventions (possibly affecting several variables simultaneously). Recently, we presented strategies for actively learning causal models with respect to sequentially improving identifiability [7]. One of the strategies greedily optimizes the number of orientable edges with single-vertex interventions, and one that minimizes the number of interventions at arbitrarily many vertices to attain full identifiability. This paper is an extended version of our previous work: besides a more detailed presentation of the algorithms, we evaluate their

[*] Corresponding author.
E-mail addresses: alain.hauser@biology.unibe.ch (A. Hauser), buhlmann@stat.math.ethz.ch (P. Bühlmann).

0888-613X/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2013.11.007

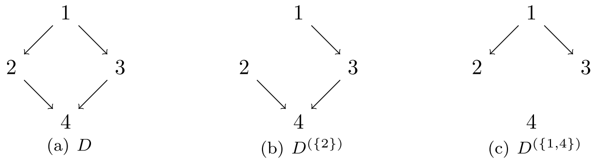

Fig. 1. A DAG D and the corresponding intervention graphs $D^{(\{2\})}$ and $D^{(\{1,4\})}$.

performance in the absence and presence of estimation errors and compare them to competing methods, and finally provide proofs for the correctness of the algorithms.

Several approaches for actively learning causal models have been proposed during the last decade, Bayesian as well as non-Bayesian ones, optimizing different utility functions. All these active learning strategies consider sequential improvement of identifiability, which is different from the more classical active learning setting where one aims for sequential optimization of estimation accuracy [8]. In the non-Bayesian setting, Eberhardt [1] and He and Geng [9] considered the problem of finding interventions that guarantee full identifiability of all representatives in a given (observational) Markov equivalence class which is assumed to be correctly learned. The approach of Eberhardt [1] works with intervention targets of unbounded size. We prove the conjecture of Eberhardt [1] on the number of intervention experiments sufficient and in the worst case necessary for fully identifying a causal model, and provide an algorithm that finds such a set of interventions in polynomial time (OptUnb, see Section 4.2). He and Geng [9] restrict the considerations to single-vertex interventions. They propose an iterative line of action for learning causal models: their method estimates the observational Markov equivalence class in a first step and then incorporates interventional data to decide about edge orientations in subsequent steps. This is not favorable from a statistical point of view since interventional data also yields information about parts of the graph that are not adjacent to the intervened variable. We will see in Section 5 that we indeed get smaller estimation errors in the finite sample case if we do not decouple the estimation of the observational Markov equivalence class and that of edge directions. Moreover, the approach of He and Geng [9] is not able to cope with a situation in which we have few or no observational data, in contrast to ours. Meganck et al. [10] compare different utility functions for single-vertex interventions, but do not address algorithmic questions of efficiently calculating optima of the utility functions. In the Bayesian setting, Tong and Koller [11] and Masegosa and Moral [12] use entropy-based utility functions. While the approach of Tong and Koller [11] only interacts with the system under investigation, the approach of Masegosa and Moral [12] uses (error-free) expert knowledge.

This paper is organized as follows: in Section 2, we specify our notation of causal models and formalize our learning goals. In Section 3, we summarize graph-theoretic background material upon which our active learning algorithms, presented in Section 4, are based. In Section 5, we evaluate our algorithms in a simulation study. The proofs of the theoretical results of Section 4 can be found in Appendix A.

2. Model

We consider a causal model on p random variables $(X_1, \ldots, X_p)$ described by a DAG D. Formally, a causal model is a pair (D, f), where D is a DAG on the vertex set $V = [p] := \{1, \ldots, p\}$ which encodes the **Markov property** of the (observational) density f: $f(x) = \prod_{i=1}^{p} f(x_i \mid x_{\mathrm{pa}_D(i)})$; $\mathrm{pa}_D(i)$ denotes the parent set of vertex i.

Our notation and definitions related to graphs are summarized in Section 3. Unless stated otherwise, all graphs in this paper are assumed to have the vertex set $[p]$.

2.1. Causal calculus

Beside the conditional independence relations of the observational density implied by the Markov property, a causal model also makes statements about effects of **interventions**. We consider **stochastic interventions** [13] modeling the effect of setting or forcing one or several random variables $X_I := (X_i)_{i \in I}$, where $I \subset [p]$ is called the **intervention target**, to the value of *independent* random variables U_I. Extending the do() operator [14] to stochastic interventions, we denote the **interventional density** of X under such an intervention by

$$f\bigl(x \mid \mathrm{do}_D(X_I = U_I)\bigr) := \prod_{i \notin I} f(x_i \mid x_{\mathrm{pa}_D(i)}) \prod_{i \in I} \tilde{f}(x_i),$$

where $\tilde{f}$ is the density of U_I on $\mathcal{X}_I$. We also encompass the observational case as an intervention target by using $I = \emptyset$ and the convention $f(x \mid \mathrm{do}(X_\emptyset = U_\emptyset)) = f(x)$. The interventional density $f(x \mid \mathrm{do}_D(X_I = U_I))$ has the Markov property of the **intervention graph** $D^{(I)}$, the DAG that we get from D by removing all arrows pointing to vertices in I. An illustration is given in Fig. 1.

We consider experiments based on data sets originating from *multiple* interventions. The **family of targets** $\mathcal{I} \subset \mathcal{P}([p])$, where $\mathcal{P}([p])$ denotes the power set of $[p]$, lists all (distinct) intervention targets used in an experiment. A family of targets $\mathcal{I} = \{\emptyset, \{2\}, \{1,4\}\}$ for example characterizes an experiment in which observational data as well as data originating from