Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Estimating bounds on causal effects in high-dimensional and possibly confounded systems ☆

Daniel Malinsky *, Peter Spirtes

Carnegie Mellon University, Pittsburgh, PA USA

ARTICLE INFO

Article history:
Received 1 December 2016
Received in revised form 13 June 2017
Accepted 14 June 2017
Available online xxxx

Keywords:
Causal inference
Ancestral graphs
Latent confounding
Markov equivalence

ABSTRACT

We present an algorithm for estimating bounds on causal effects from observational data which combines graphical model search with simple linear regression. We assume that the underlying system can be represented by a linear structural equation model with no feedback, and we allow for the possibility of latent confounders. Under assumptions standard in the causal search literature, we use conditional independence constraints to search for an equivalence class of ancestral graphs. Then, for each model in the equivalence class, we perform the appropriate regression (using causal structure information to determine which covariates to adjust for) to estimate a set of possible causal effects. Our approach is based on the IDA procedure of Maathuis et al. [17], which assumes that all relevant variables have been measured (i.e., no latent confounders). We generalize their work by relaxing this assumption, which is often violated in applied contexts. We validate the performance of our algorithm in simulation experiments.

1. Introduction

It is well known that regression estimates for causal effects will be biased unless a variety of conditions on the data are satisfied; methods which correct for confounding by covariate adjustment rely on facts about the causal structure of the system under study (e.g., whether all the relevant variables have been measured and how the measured covariates are causally linked to the variables of interest). Maathuis et al. [17] provide a good overview and explanation of this idea; see also [7] for related analysis. Roughly speaking, regressing Y on X while controlling for additional covariates does not produce an unbiased estimate of the effect of intervening on X unless the additional covariates account for any possible confounding of X and Y. In the language of causal graphs, the covariates must block all causal pathways from variables (measured or not) which are causes of both X and Y and the covariates should not include effects of X. The conditions under which regression can produce an unbiased estimate of a causal effect can be readily translated into conditions on an appropriate causal graphical model [21].

The method proposed here combines techniques from automated causal search and regression to estimate causal effects (also called intervention effects) from observational data. In particular, the algorithms described in Section 4 estimate causal effects even when there are relevant unmeasured variables (i.e., "latent confounding" or "causal insufficiency"). The method is based on the one developed by Maathuis et al. [17], which has been fruitfully applied in the context of genetics research

☆ 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: malinsky@cmu.edu (D. Malinsky), ps7z@andrew.cmu.edu (P. Spirtes).

http://dx.doi.org/10.1016/j.ijar.2017.06.005
0888-613X/© 2017 Published by Elsevier Inc.

[16,32]. The IDA ("Intervention when the DAG is Absent") algorithm of Maathuis et al. is consistent under a set of assumptions which includes causal sufficiency: the assumption that no variables which are common direct causes of at least two measured variables are unmeasured. Importantly, IDA is feasible in high-dimensional settings, where sample sizes are small but the number of covariates is very large. In their genetics applications there are more than 4000 variables, and the goal is to find variables which are likely strong regulators (causes) of some chosen variable of interest in order to prioritize gene knock-out experiments. In the data which is typical in the social sciences and many areas of biomedical research, the assumption of causal sufficiency is often unwarranted. Even genome-wide expression data may be causally insufficient if there are unmeasured factors like proteins which act as common causes of multiple gene expressions. Our procedure is consistent in the presence of latent confounders and is feasible for large numbers of variables. Note that the procedure presented here can also be considered an alternative to causal estimation techniques based on propensity scores (e.g., [26,14]). While adapting propensity score techniques to high-dimensional settings is an active area of research (e.g., [2]), it is typical to assume unconfoundedness (a.k.a. "strong ignorability").[1] Our approach dispenses with this assumption, but as a consequence some effects will not be identifiable with our method, and in other cases we may produce bounds rather than a single point estimate. On the other hand, instrumental variables methods are a popular approach to estimating causal effects in possibly confounded settings. In addition to the various statistical difficulties with IV methods like two stage least squares (e.g., "weak instrument" issues), there is a more fundamental difficulty to data-driven IV estimation: instrumental variable analysis requires knowing that a potential instrument satisfies the exclusion restriction, which is not in general testable. IV methods are not, therefore, feasible to implement in data-driven, high-dimensional settings without substantial knowledge of causal mechanisms.

Judea Pearl and his collaborators provide techniques for calculating the outcomes of interventions when the true causal structure (i.e., true causal graph) is known (e.g., [34,28]). These results relate to the general conditions for "back-door adjustment" and "front-door adjustment" described in [21]. The back-door criterion is a graphical criterion that is sufficient for adjustment in the following sense: if a set of variables satisfies the back-door criterion for a given graph, then conditioning on that set is sufficient for estimating intervention effects from observed distributions. Maathuis and Colombo [15] generalize the back-door criterion to different types of graphical objects, and their result will play an instrumental role in the algorithms we propose. In order to estimate intervention effects via (generalized) back-door adjustment from data, the researcher must be able to identify the set of covariates which satisfy the (generalized) back-door criterion. To determine which variables satisfy this condition without substantial background causal knowledge, we use (variations of) an automated causal search algorithm called FCI [31,38]. Our procedure is closely related to the work of Hyttinen et al. [11], and we discuss that method in Section 4.

One alternative approach to estimating causal effects is worth mentioning here. Algorithms which learn latent variable LiNGAM models [10,13,8,33] allow for the possibility of unmeasured variables. These algorithms exploit assumptions about the causal structure (assumed to be structural equation models which are acyclic, linear, and which have non-Gaussian error terms) to estimate graphical structure and some estimate causal strength parameters simultaneously. See also [9,27] for related Bayesian procedures. One substantial benefit to these algorithms is that they can often identify a unique model or a smaller equivalence class of models than the FCI algorithm can. Unfortunately, computational complexity makes these algorithms mostly infeasible in applied contexts when there are more than a few variables and the sample sizes required are unrealistic for many applications. Furthermore, these algorithms generally require that the researcher stipulates the number of (possible) latent variables explicitly; the approach proposed here is more general in that it does not make any assumptions about the number of (possible) unmeasured variables.

Though our procedure cannot always pin down a unique causal graphical model, from an equivalence class of graphs we can estimate bounds on causal effects. That is, for a given variable pair (X, Y) we can calculate a set of estimates for the causal effect of X on Y. Each estimate corresponds to some model in the equivalence class. Some effects, for some or all models in the equivalence class, will not be identified because possible confounding cannot be blocked. Otherwise, the minimum and maximum estimates in the estimated set are bounds on the true causal effect, and these bounds can be used to prioritize follow-up experiments by, for example, concentrating on experimental manipulations of variables with effects bounded away from zero.

2. Definitions and background

It is assumed here that the causal structure of the system under study can be represented by a Directed Acyclic Graph (a DAG). A graph $\mathcal{G}$ is a pair $(\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is a set of vertices corresponding to random variables $\mathbf{V} = \{X_1, \ldots, X_p\}$ and $\mathbf{E}$ is a set of edges. A DAG contains only directed edges ($\rightarrow$) and has no cycles (no sequence of directed edges from any variable to itself). If $X_i \rightarrow X_j$ then X_i is called a parent of X_j, and X_j is a child of X_i. Two variables are adjacent if there is some edge between them, and a path is a sequence of distinct adjacent vertices (e.g., $X_i \leftarrow X_j \leftarrow X_k \rightarrow X_l$). A directed path from X_i to X_j is a path which contains only directed edges away from X_i and toward X_j. When there is a directed path from X_i to X_j we call X_i an ancestor of X_j, and X_j is a descendent of X_i. Denote the set of parents of a vertex X in $\mathcal{G}$ by $pa(X, \mathcal{G})$, and the sets of ancestors of X and descendents of X by $An(X, \mathcal{G})$ and $De(X, \mathcal{G})$ respectively. The adjacency set of X is $adj(X, \mathcal{G})$.

[1] See [29] for a simple example where ignorability fails, and propensity score estimation produces an incorrect conclusion.

Please cite this article in press as: D. Malinsky, P. Spirtes, Estimating bounds on causal effects in high-dimensional and possibly confounded systems, Int. J. Approx. Reason. (2017), http://dx.doi.org/10.1016/j.ijar.2017.06.005