Towards A Complete Identification Algorithm for Missing Data Problems

Ilya Shpitser and James M. Robins

Abstract

The fundamental problem of causal inference is a missing data problem – the comparison of responses to two hypothetical treatment assignments is made difficult because for every experimental unit, only one treatment assignment is actually observed. Simple identification results in causal inference that link observed and counterfactual quantities using the stable unit treatment value assumption and ignorability has been extended to a complete general theory using graphical causal models [9, 8, 7], and a rich estimation theory for resulting functionals of observed data has been developed.

We consider the implications of the converse view: that missing data problems are a form of causal inference. We consider the classical missing data problem of identifying the full data law from the observed data law as a problem of inferring a joint law over counterfactual variables from a joint law over factual variables. We encode the relationship between the factual and counterfactual variables in graphical models, in an approach closely related to similar modeling approaches in causal inference, review recent identification results developed in this framework, and develop a new algorithm for identifying the full data law in settings with both missing data and hidden variables. Our algorithm can be viewed as a version of the **ID** algorithm for identifying causal effects adapted to peculiarities of the missing data setting. Completeness of our algorithm is currently an open problem.

1 Introduction

Missing data is a common difficulty in the analysis of survey, experimental and observational data, both for the purpose of creating classifiers in machine learning, and for drawing causal inferences. Existing approaches to missing data include inference on a parametric model of missing variables, such as the expectation minimization algorithm [1], and multiple imputation [5], matrix completion methods that use ideas from the sparsity literature [4], and methods closely related to causal inference that exploit information about the mechanism that drives missingness [3, 2, 6].

Simple versions of the missing mechanism based methods use assumptions known as Missing Completely At Random (MCAR) and Missing At Random (MAR) to recover functions of the underlying data distribution exactly, without the need for strong parametric assumptions on missing variables. A recent strand of work has used graphical models to show cases where it is possible to recover functions of the underlying data law under assumptions weaker than MAR, in other words when data is Missing Not At Random (MNAR). In particular [6] has given a general algorithm for recovering the full data distribution given a set of constraints, possibly weaker than MAR, represented by a graphical model.

We review these results, and show that the existing algorithm is not complete for the problem of identification of the full data distribution. We provide a repaired algorithm which recursively decomposes the identification problem into subproblems that are processed either in parallel or sequentially, and generalize it to settings with completely hidden variables. Our algorithm can be viewed as a version of the **ID** algorithm for identification of causal effects, extended to missing data settings

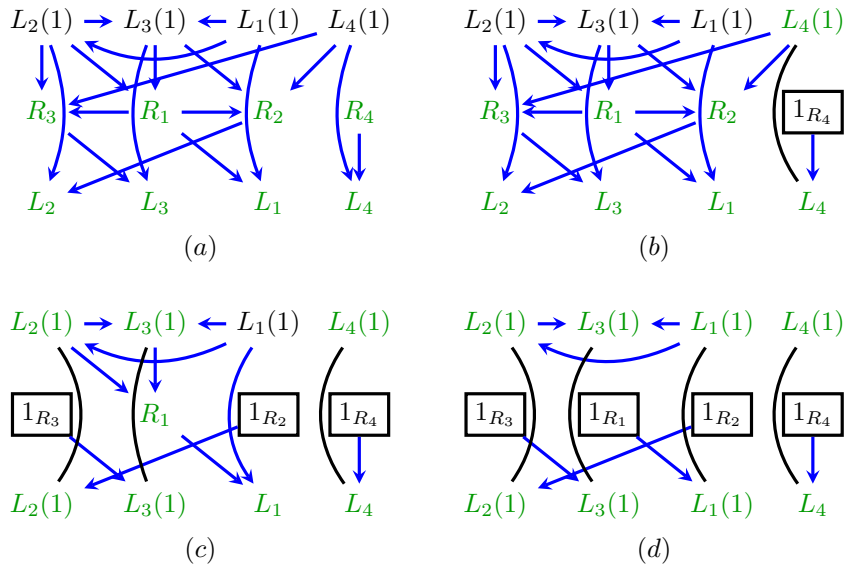

Figure 1: An example where the **MID** algorithm in [6] and g-computation inspired algorithms for missing data problems fail, but our new algorithm succeeds.

where constraints on the missingness model can be expressed as a graph. The purpose of our work is to build towards a characterization of missing data problems where the full data law is identifiable from the observed data law, and only equality constraints between variables and missingness indicators are permitted. Here equality constraints are taken to mean conditional independence constraints, and generalized independence constraints, also known as Verma constraints [10].

Our framework uses *missingness graphs*, which display relationships between underlying missing variables (represented as counterfactual outcomes in the causal inference sense), missingness indicators (represented as treatments in the causal inference sense), and observed proxies (represented as observed data variables in the causal inference sense) as a graphical model. For example, in Fig. 1, the variables $L_i(1)$ represent underlying variables in the full data law that we do not always get to observe, L_i are observed proxies that either assume values of $L_i(1)$ (if $R_i = 1$) or assume values "**missing**" (if $R_i = 0$) and R_i are missingness indicators. The relationships between these variables are displayed according to a standard graphical model, which implies certain independence restrictions on the full data law. The algorithm succeeds if it is able to set, by intervention, all R_i indicators to 1, thereby forcing observed status on all variables. Unlike classical causal inference settings, such as the g-computation algorithm or the **ID** algorithm where interventions can be viewed as applied sequentially, in our algorithm interventions are sometimes applied sequentially and sometimes in parallel. Moreover, a single step of the algorithm (where we potentially intervene on multiple variables simultaneously) which in the **ID** algorithm involves a single application of the g-formula, in our case potentially involves an entire recursively solvable subproblem to identify the relevant propensity score distribution $p(R_i \mid \text{parents of } R_i)$.

As an example, we show that our algorithm is able to identify the full data law $p(L_1(1), L_2(1), L_3(1), L_4(1))$ in the model in Fig. 1 (a). To identify $p(L_1(1), L_2(1), L_3(1), L_4(1))$ in Fig. 1 (a), the algorithm first identifies $p(R_4)$, and intervenes on R_4, resulting in $L_4(1)$ becoming observed, and the subproblem shown in Fig. 1 (b). Next, the algorithm identifies $p(R_3 \mid L_2(1), L_4(1))$ and $p(R_2 \mid L_3(1), L_4(1))$ in the subproblem, and intervenes on both R_2 and R_3, resulting $L_2(1)$ and $L_3(1)$ becoming observed, and the subproblem shown in Fig. 1 (c). Finally, the algorithm identifies $p(R_1 \mid L_2(1), L_3(1))$ in the new subproblem, and intervenes on R_1, resulting in an expression that is a function of the observed data distribution and which is equal to the underlying full data law.

While our algorithm is very general and subsumes all previously known graph-based non-parametric identification approaches for missing data, it is not currently known whether it is complete.

References

[1] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38, 1977.

[2] Karthika Mohan and Judea Pearl. Graphical models for recovering probabilistic and causal queries from missing data. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1520–1528. Curran Associates, Inc., 2014.

[3] Karthika Mohan, Judea Pearl, and Jin Tian. Graphical models for inference with missing data. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1277–1285. Curran Associates, Inc., 2013.

[4] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[5] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley & Sons, 1987.

[6] Ilya Shpitser, Karthika Mohan, and Judea Pearl. Missing data as a causal and probabilistic problem. In *Proceedings of the Thirty First Conference on Uncertainty in Artificial Intelligence (UAI-15)*, pages 802–811. AUAI Press, 2015.

[7] Ilya Shpitser and Judea Pearl. Identification of conditional interventional distributions. In *Proceedings of the Twenty Second Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 437–444. AUAI Press, Corvallis, Oregon, 2006.

[8] Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI-06)*. AAAI Press, Palo Alto, 2006.

[9] Jin Tian and Judea Pearl. On the testable implications of causal models with hidden variables. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence (UAI-02)*, volume 18, pages 519–527. AUAI Press, Corvallis, Oregon, 2002.

[10] Thomas S. Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Department of Computer Science, University of California, Los Angeles, 1990.