Statistica Sinica **36** (2026), 1-39
doi:https://doi.org/10.5705/ss.202023.0382

CAUSAL AND COUNTERFACTUAL VIEWS OF MISSING DATA MODELS

Razieh Nabi*, Rohit Bhattacharya, Ilya Shpitser and James M. Robins

Emory University, Williams College,
Johns Hopkins University and Harvard University

Abstract: It is often said that 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 potential response is observed. In this paper, we consider the implications of the converse view: that missing data problems are a form of causal inference. We make explicit how the missing data problem of recovering the complete data law from the observed law can be viewed as identification of a joint distribution over counterfactual variables corresponding to values had we (possibly contrary to fact) been able to observe them. Drawing analogies with causal inference, we show how identification assumptions in missing data can be encoded in terms of graphical models defined over counterfactual and observed variables. We review recent results in missing data identification from this viewpoint. In doing so, we note interesting similarities and differences between missing data and causal identification theories.

Key words and phrases: Causal graphs, causal inference, missing not at random.

1. Introduction

Missing data is a common challenge in the analysis of survey, experimental, and observational data, both for the purpose of prediction and for drawing causal conclusions. Complete-case analysis is a popular and simple approach to handling missing data, but it is generally only justified when data entries are missing-completely-at-random (MCAR) (Rubin, 1976). When data entries are missing in a way that only depends on observed data values, the data are said to be missing-at-random (MAR) (Rubin, 1976). Under MAR assumptions, it is possible to identify target parameters of the underlying data distribution without the need for further parametric assumptions. Moreover, we can estimate parameters identified under MAR via likelihood-based methods such as expectation maximization (Dempster, Laird and Rubin, 1977; Horton and Laird, 1999; Little and Rubin, 2002), multiple imputation (Rubin, 1987; Schafer, 1999), inverse probability weighting (Robins, Rotnitzky and Zhao, 1994; Li et al., 2013), or semiparametric methods that exploit information about mechanisms determining missingness and are closely related to methods for estimating causal parameters

*Corresponding author. E-mail: razieh.nabi@emory.edu