Causal and counterfactual views of missing data models

Razieh Nabi*[1], Rohit Bhattacharya[2], Ilya Shpitser[3], and James Robins[4]

[1]Emory University – United States
[2]Williams College – United States
[3]Johns Hopkins University – United States
[4]Harvard University – United States

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 work, 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. The validity of identification and estimation results using such techniques rely on the assumptions encoded by the graph holding true. Thus, we also provide new insights on the testable implications of a few common classes of missing data models, and design goodness-of-fit tests around them.

*Speaker