Nonparametric estimation of nonlinear causal effects from incomplete observational data

Juha Karvanen

University of Jyväskylä, Finland

The estimation of causal effects is a challenging problem if only incomplete observational data are available and the causal effects are strongly nonlinear. In this work, the problem is approached with the tools of modern statistics. As a starting point, it is assumed that the true or hypothesized causal structure and missingness mechanism are qualitatively known, i.e. the directions of the causal effects are known but their functional form or magnitude are unknown.

First, the situation is described using a causal model with design (Karvanen, 2013), which systematically presents the causal assumptions, the study design and the missingness mechanism in the same graph. Causal calculus (Pearl, 2009) is then applied to express the causal effects of interest in terms of observational probabilities. Missing data are handled with multiple imputation (Van Buuren, 2012). Generalized additive models (GAM) (Hastie and Tibshirani, 1990) with smoothing splines are applied to estimate the observational probabilities needed to calculate the causal effects.

The causal model with design for the problem considered in this work is presented in Figure 1. The objective is to estimate the average causal effect $\mathrm{E}(Y \mid \mathrm{do}(X = x))$ from the observational data on X, Z and Y shown in Figure 2 (left panel).

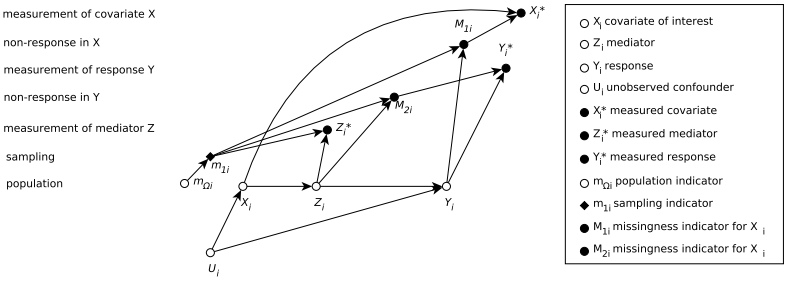

Figure 1: Causal model with design for the example studied.

It can be seen from Figure 1 that the causal effect of X to Y is confounded by unobserved variable U. Therefore the causal effect cannot be directly estimated from observational data. Fortunately, there is a mediator Z, which is independent on U on the condition X. Using the rules of causal calculus

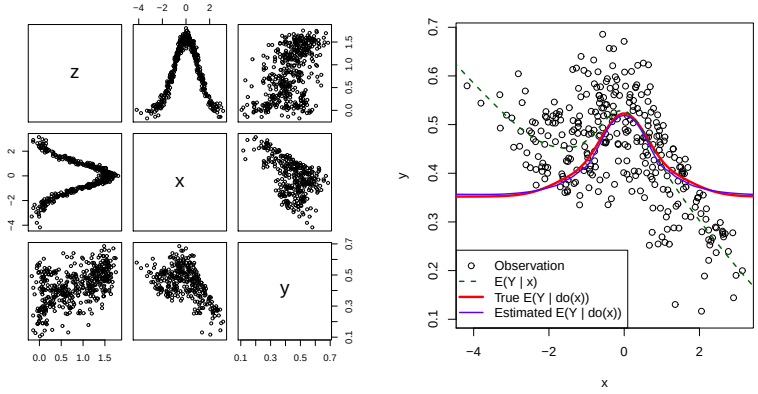

Figure 2: Left panel: Scatterplot of the observed data (a subsample of 500 out of 20000). Right panel: True and estimated average causal effects (solid lines) together with observations and their conditional average (dotted line).

(Pearl, 2009), the average causal effect can be expressed in terms of observed probabilities as follows (frontdoor adjustment)

$$\mathrm{E}(Y \mid \mathrm{do}(X = x)) = \int p(z \mid X = x) \int \mathrm{E}(Y \mid X = x', Z = z) p(X = x') \mathrm{d}x' \mathrm{d}z. \tag{1}$$

There are missing data in variables X and Y. According to the causal model with design in Figure 1, the missingness of Y depends on Z and the missingness of X depends on Y. The missing data are handled with multiple imputation applying MICE algorithm (Van Buuren, 2012). For each imputed dataset, the components of formula (1) are estimated from the data using GAM with smoothing splines. The right panel of Figure 2 shows that the average causal effect $\mathrm{E}(Y \mid \mathrm{do}(X = x))$ substantially differs from the observed conditional mean $\mathrm{E}(Y \mid X = x)$. Despite of this, the applied procedure correctly estimates the average causal effect.

Keywords: causal estimation, data analysis, missing data, structural equation model

References:

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* CRC Press.

Karvanen, J. (2013). Study design in causal models. Submitted, arXiv:1211.2958.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference.* Cambridge University Press, second edition.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data.* CRC Press.