Adjusting for Many Covariates in Matching-Based Causal Inference

Adam Sales
Joint work with Ben Hansen

Researchers attempting observational causal inference or program evaluation frequently encounter more data—in particular, pre-treatment covariates—than traditional causal models can accommodate. Generally, some subset of available covariates is theoretically known, or thought, to pose a significant confounding threat, and these variables are included as covariates in a model. However, other variables in the dataset can potentially confound an estimate, as well. This paper suggests a way to supplement a traditional matching design—perhaps, but not necessarily, based on propensity scores—with high dimensional modeling to account for a large set of covariates. The method attempts to protect matching estimates against the bias that would result from omitting a measured covariate from a matching scheme. After constructing a match using a limited number of covariates, selected for their theoretical importance, a researcher would then use a high-dimensional outcome regression on the unmatched controls to estimate a predictive model of the outcome of interest as a function of the full covariate matrix. The resulting predictions, in the matched set, can be used to test balance, as well as to further adjust the causal estimate and reduce confounding bias. In some circumstances, they will decrease the variance of the estimate as well. We will present theoretical results to justify the method's bias-reducing properties, as well as a simulation study that demonstrates its bias-reducing properties. Additionally, we will illustrate the method in an evaluation of a school-level intervention in Texas.