DSSV 2023
Data Science, Statistics & Visualisation
ECDA 2023
European Conference on Data Analysis

Which variables should you impute in treatment effects?

J. Berrevoets[1], F. Imrie[2], T. Kyono[3], J. Jordon[4], and M. van der Schaar[1, 4]

[1]University of Cambridge, DAMTP
[2]UCLA, EEC
[3]Meta
[4]Alan Turing Institute

Missing data is a systemic problem in practical scenarios that causes noise and bias when estimating treatment effects. This makes treatment effect estimation from data with missingness a particularly tricky endeavour. A key reason for this is that standard assumptions on missingness are rendered insufficient due to the presence of an additional variable, treatment, besides the input (e.g. an individual) and the label (e.g. an outcome). The treatment variable introduces additional complexity with respect to *why* some variables are missing that is not fully explored by previous work. In our work we introduce *mixed confounded missingness* (MCM), a new missingness mechanism where some missingness *determines* treatment selection and other missingness *is determined by* treatment selection. Given MCM, we show that naively imputing all data leads to poor performing treatment effects models, as the act of imputation effectively *removes* information necessary to provide unbiased estimates. However, no imputation at all also leads to biased estimates, as missingness determined by treatment introduces bias in covariates. Our solution is *selective* imputation, where we use insights from MCM to inform precisely which variables should be imputed and which should not. We empirically demonstrate how various learners benefit from selective imputation compared to other solutions for missing data. We highlight that our experiments encompass both average treatment effects *and* conditional average treatment effects. This work was published at AISTATS 2023 [1]

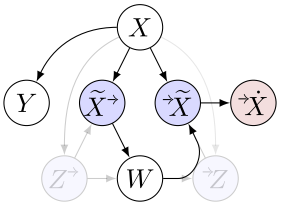
X
Y
$\widetilde{X}^{\rightarrow}$
${}^{\rightarrow}\widetilde{X}$
${}^{\rightarrow}\dot{X}$
$Z^{\rightarrow}$
W
${}^{\rightarrow}Z$

Figure 1: **Selective imputation.** Above DAG depicts MCM, the missingness mechanism we introduce in our paper and will discuss in our talk. Beyond MCM, we also show our solution (in red).

[1] J. Berrevoets, F. Imrie, T. Kyono, J. Jordon, and M. van der Schaar, "To impute or not to impute? missing data in treatment effect estimation," in *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics* (F. Ruiz, J. Dy, and J.-W. van de Meent, eds.), vol. 206 of *Proceedings of Machine Learning Research*, pp. 3568–3590, PMLR, 25–27 Apr 2023.