Title:
Learning Surrogate Indices from Historical A/B Tests

Abstract:

Experimentation on digital platforms often faces a dilemma: we want to experiment rapidly at scale but we also want to make decisions based on long-term impact. Usually one resorts to looking at indices (i.e., scalar-valued functions) that combine multiple short-term surrogate outcomes. Constructing indices by regressing long-term metrics on short-term ones can suffer bias from confounding between the long and short terms as well as from direct (i.e., unmediated) effects. I will discuss how to instead leverage past experiments as instrumental variables (IVs) and some surrogates as negative-control outcomes, with real-world examples from Netflix. There are two key challenges to surmount to make this possible. First, past experiments characterize the right surrogate index as a solution to an ill-posed inverse problem: it does not uniquely identify an index, and nearly solving it does not translate to being near a solution. I tackle this via a novel automatic debiasing for inference on linear functionals of solutions to inverse problems (as average long-term effects are such functionals of the index). Second, this approach suffers from a many-weak-IVs phenomenon where even as we observe more past experiments we have non-vanishing bias in estimating the implied moment conditions. We tackle this by incorporating an instrument-splitting technique into our nonparametric nuisance estimators leading to a machine-learning version of the classic (linear) jackknife IV estimator (JIVE). Taken together the methods and results enable reliable and fast long-term causal inference in domains that experiment rapidly at scale.

Papers:

https://arxiv.org/abs/2208.08291

https://arxiv.org/abs/2307.13793

https://arxiv.org/abs/2402.17637

https://arxiv.org/abs/2406.14140