Randomization inference with machine learning

Gary Chan

Department of Biostatistics, University of Washington, U.S.A.

Abstract

Incorporating covariate information via machine learning (ML) is a highly effective way to improve precision in causal inference. However, integrating ML into design-based randomization inference, particularly for the weak null hypothesis of zero average treatment effect, poses significant theoretical and computational challenges. We introduce a novel framework for the studentized Fisher Randomization Test using ML-estimated pseudo-outcomes. We explore three key facets of this approach. First, we characterize the efficiency gains associated with different choices of pseudo-outcomes. Second, we analyze the behavior of various studentization methods, revealing how the choice of variance estimator affects statistical power. Finally, we provide a rigorous rationale for evaluating permutations using frozen ML models rather than refitting the algorithms for each permutation. We show that holding the ML models fixed across the permutation distribution maintains asymptotic validity for the weak null while rendering the procedure computationally manageable.