Essays on Individualized Treatment Rules

Daido Kido
Graduate School of Economics, Kyoto University

Abstract

The effects of interventions (i.e., treatments) are often heterogeneous, based on individual characteristics (i.e., covariates). For example, an unemployment policy may differently impact the future incomes of unemployed individuals depending on their levels of education and previous earnings. Meanwhile, responses to an online advertisement may depend on demographics (e.g., age and sex) and browsing history. Along these same lines, the efficacy of a therapy may differ according to patients' physical constitutions. Policymakers who wish to maximize the welfare of the population of interest (typically, the population's mean outcome) should take advantage of such heterogeneity and optimize treatment allocations using covariate values. Such personalized treatment assignment is often called an *individualized treatment rule (ITR)*. The importance of ITRs is widely acknowledged in various fields, including healthcare, public policy, and business.

When data on the relationship between the treatment response and covariates are available, an optimal ITR must be learned from the data. Three main estimation approaches have been proposed for optimal ITRs. The first is to estimate the mean of the potential outcome conditioned on the covariates and use it to optimize the ITRs (e.g., Qian & Murphy, 2011). The second is to construct an unbiased estimator of the welfare of each ITR using the inverse propensity score method and optimize it directly (e.g., Kitagawa & Tetenov, 2018; Swaminathan & Joachims, 2015a, 2015b). The third approach replaces the inverse propensity score method with a doubly robust method in the second approach (e.g., Athey & Wager, 2021; Dudík et al., 2011; Zhou et al., 2023).

These existing approaches implicitly rely on the assumptions that (i) the population from which data are generated and the population at which the estimated ITR is deployed are essentially identical and (ii) the welfare of each ITR in the target population can be identified. However, these assumptions often fail. Frequently, policymakers must decide which ITR to apply in the future based on data obtained elsewhere. Alternatively, owing to limited budgets, policymakers often rely on observational data that do not allow for the point identification of causal effects. When these assumptions do not hold, the theoretical guarantees of existing approaches also no longer hold.

This dissertation primarily examines situations in which such existing methods cannot be applied. Specifically, this dissertation consists of three chapters and one appendix. Chapter 1 studies a situation when the population for which a policymaker wants to deploy an ITR and the population from which the available data is generated are different

in an unknown way. Existing estimation methods often assume that these two populations are the same; however, this assumption often fails in practice because of limited data. Unfortunately, existing estimation methods do not work as expected in the new setting; in the first place, the conventional learning goal (i.e., the optimal ITR in the target population) is not identified. This study examines the application of *distributionally robust optimization (DRO)*. The DRO formalizes the ambiguity about the target population and adapts to the worst-case scenario in the set. Compared to the DRO with Kullback-Leibler divergence-based ambiguity studied by Mo et al. (2021) and Si et al. (2020, forthcoming), this study shows that the DRO with Wasserstein distance-based ambiguity provides simple intuitions and a simple estimation method. Next, this study develops an estimator for the distributionally robust ITR and evaluates its theoretical performance. An empirical application shows that the proposed approach outperforms the naive approach in the target population.

Chapter 2 studies the statistical treatment assignment problem when the (conditional) average treatment effect (ATE) is partially identified. The ATE characterizes the optimal ITR for the target population. In practice, the ATE is typically only partially identified. Presuming partial identification of the ATE, several studies (Ishihara & Kitagawa, 2021; Manski, 2007; Stoye, 2012; Yata, 2021) solved finite-sample statistical decision problems by restricting the class of data generating processes or identifying assumptions. However, these exact approaches are not feasible for general classes of data generating processes. Instead, this study conducts a local asymptotic analysis and develops a locally asymptotically minimax statistical treatment rule (LAM STR). Importantly, this study does not assume full differentiability but rather directional differentiability of the boundary functions of the identification region of the ATE. The results show that the LAM STR differed from the plug-in STR. A simulation study also demonstrated that the LAM STR outperformed the plug-in STR.

Chapter 3 investigates the problem of individualizing treatment allocations using stated preferences for treatments. Treatment preference is a significant factor of heterogeneous effects. In the presence of such heterogeneity, it is tempting to individualize assignments using stated preferences. However, such individualization is often problematic; if individuals know in advance how the assignment will be individualized based on their stated preferences, they may state false preferences. Such strategic responses induce differences between the target population and the population that generates data. This study derives the optimal ITR when individuals strategically state their preferences. Further, this study shows that the optimal ITR is strategy-proof; that is, individuals do not have a strong incentive to lie even if they know the optimal ITR a priori. Constructing the optimal ITR requires information on the distribution of true preferences and the average treatment effect conditioned on the true preferences. In practice, information must be identified and estimated from the data. Because true preferences are hidden information,

identification is not straightforward. This study discusses two experimental designs that allow the identification: strictly strategy-proof randomized controlled trials and doubly randomized preference trials. Under the assumption that data come from one of these experiments, this study develops data-dependent procedures for determining ITR. The maximum regret of the proposed procedure converges to zero at the rate of the square root of the sample size. An empirical application demonstrates our proposed STRs.

The research presented in Appendix A is a joint study by Takanori Ida, Takunori Ishihara, Koichiro Ito, Toru Kitagawa, Shosei Sakaguchi, and Shusaku Sasaki. Appendix A presents the benefits of combining compulsion and self-selection to allocate treatments. The effects of treatments depend not only on observable covariates but also on unobservable covariates. Previous studies on ITRs focused on treatment assignments using only observable covariates, and did not consider using unobservable covariates. This study argues that viewing compulsory and self-selection treatment assignments as distinct options and optimizing the ITR allows for the indirect use of unobservable covariates and improves welfare over conventional ITRs. Using the framework of the *local average treatment effect (LATE)* (Imbens & Angrist, 1994), the analysis relates the improvement to the LATEs for those who would and would not self-select into treatments. This argument implies that a combination of compulsion and self-selection significantly improves welfare when considerable heterogeneity exists between LATEs and the ATE. An empirical study based on a randomized controlled trial concerning energy-saving programs demonstrates the advantage of the combined use of compulsion and self-selection.