Estimating Causal Effects in Pragmatic Settings With Imperfect Information

The Harvard community has made this article openly available. Please share how this access benefits you. Your story matters

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Citable link</td>
<td><a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:41128471">http://nrs.harvard.edu/urn-3:HUL.InstRepos:41128471</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>This article was downloaded from Harvard University’s DASH repository, and is made available under the terms and conditions applicable to Other Posted Material, as set forth at <a href="http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA">http://nrs.harvard.edu/urn-3:HUL.InstRepos:dash.current.terms-of-use#LAA</a></td>
</tr>
</tbody>
</table>
Estimating Causal Effects in Pragmatic Settings with Imperfect Information

A dissertation presented

by

David Cheng

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

May 2018
Abstract

Precision medicine seeks to identify the optimal treatment for each individual based on his or her unique features. This invariably involves some form of estimation of causal effects for different patient subgroups to determine the treatment that leads to superior outcomes. Implementing methods to estimate causal effects in modern large and rich data sources such as electronic medical records (EMR), however, still faces challenges as information on patients is imperfectly captured in the observed data. In this work, we propose approaches to address some of the primary issues encountered in estimating causal effects in these pragmatic settings.

In Chapter 1, we consider estimating average treatment effects (ATE) in observational data where the number of covariates is not small relative to the sample size. We develop a double-index propensity score (DiPS) obtained by smoothing treatment over linear predictors for the covariates from initial working parametric propensity score (PS) and outcome models fit with regularization. We show that an inverse probability weighting (IPW) estimator based on DiPS maintains the doubly-robustness and local semiparametric efficiency properties of the usual doubly-robust estimator and achieves further gains in robustness and efficiency under model misspecification. Simulations demonstrate the benefit of the approach in finite samples, and the method is illustrated by applications estimating the effects of statins on colorectal cancer risk and smoking on C-reactive protein.

In Chapter 2, we extend the work from Chapter 1 to allow for incorporation of a large set of unlabeled data. This arises in EMR data when chart review is performed to ascertain gold-standard outcomes in case outcomes of interest are not directly observed. We frame the problem in a semi-supervised learning setting, where a small set of observations
are labeled and a large set of observations are unlabeled but includes features predictive of the outcome. We develop an imputation followed by IPW approach that is robust to misspecification of the imputation model. The estimator is also doubly-robust and efficient under an ideal semi-supervised model where the distribution of the unlabeled data is known. We demonstrate the robustness and efficiency of the approach through simulations and an application to compare rates of response to biologic therapies among inflammatory bowel disease patients.

In Chapter 3, we turn to the problem of identifying interpretable treatment subgroups. Although many statistical and machine learning approaches have been developed to discriminate patients exhibiting enhanced treatment effects, many produce output that are difficult to interpret for clinicians. Tree-based methods are a natural way of producing interpretable output but are typically not competitive in discriminative performance. We consider adapting the method of “born-again” trees (Breiman and Shang, 1996) for subgroup identification to balance interpretability and performance by re-approximating flexible initial estimators for the conditional average treatment effect (CATE). The approach is applied to data from two large phase 3 trials evaluating the effect of oral fumarate for preventing relapses among patients with multiple sclerosis.

In Chapter 4, we further consider estimating CATE when both randomized and observational data are simultaneously observed. Observational estimates could potentially be combined with randomized estimates to improve efficiency, but there may be concerns about whether confounding and treatment effect heterogeneity have been adequately addressed. We propose a combination approach that always yields an estimator consistent for a conditional causal effect. It weights heavily towards the randomized estimator in case bias in the OS estimator is detected or else combines the estimators for optimal efficiency. We show the weights can be estimated through a penalized least square criteria. The performance of the weights are evaluated through simulations, and we illustrate the method by estimating effects of hormone therapy on coronary heart disease in data from the Women’s Health Initiative.
Contents

Title page ......................................................... i
Abstract ......................................................... iii
Table of Contents .................................................. v

1 Estimating Average Treatment Effects with a Double-Index Propensity Score .................. 1
   1.1 Introduction ............................................... 2
   1.2 Notations and Problem Setup ................................ 5
   1.3 Parametric Models for Nuisance Functions .................... 6
   1.4 Double-Index Propensity Score and IPW Estimator ............... 8
   1.5 Asymptotic Robustness and Efficiency Properties ............... 9
       1.5.1 Robustness .................................... 10
       1.5.2 Efficiency .................................... 11
   1.6 Perturbation Resampling .................................. 13
   1.7 Numerical Studies ........................................ 14
       1.7.1 Simulation Study ................................ 14
   1.8 Discussion .............................................. 18
   1.9 Appendix .............................................. 20
       1.9.1 Appendix A: Regularity Conditions .................... 20
       1.9.2 Appendix B: Supporting Lemmas ....................... 21
       1.9.3 Appendix C: Expansion of Normalization Constant .......... 24
       1.9.4 Appendix D: Main Results ......................... 25

Acknowledgments .................................................. viii
2 Efficient and Robust Semi-Supervised Estimation of Average Treatment Effects in Electronic Medical Records Data

2.1 Introduction ................................................................. 38
2.2 Notations and Semi-Supervised Framework ................................... 40
  2.2.1 Target Parameter and Leveraging Unlabeled Data .................. 40
2.3 Robust Imputations ......................................................... 41
2.4 Doubly-Robust IPW Based on the Double-Index PS ...................... 43
2.5 Asymptotic Robustness and Efficiency Properties of \( \hat{\Delta} \) .......... 44
2.6 Perturbation Resampling .................................................. 46
2.7 Numerical Studies .......................................................... 48
  2.7.1 Simulations ............................................................. 48
  2.7.2 EMR Data Application ................................................. 52
2.8 Discussion ................................................................. 54
2.9 Appendix ................................................................. 55
  2.9.1 Appendix A: Supporting Lemmas ..................................... 56
  2.9.2 Appendix B: Consistency and Asymptotic Linearity of \( \hat{\Delta} \) ........ 60
  2.9.3 Appendix C: Semiparametric Efficiency .............................. 64

3 Identifying Treatment Subgroups by Tree Approximations of Conditional Treatment Effects 68

3.1 Introduction ................................................................. 69
3.2 Re-Approximation of Conditional Treatment Effects ...................... 71
  3.2.1 Tree Approximation and Subgroup Definition ..................... 71
  3.2.2 Estimation ............................................................. 72
3.3 Tuning and Validation ..................................................... 72
3.4 Application to MS Data .................................................... 74
  3.4.1 Results ................................................................. 76
3.5 Discussion ................................................................. 79

4 Adaptive Combination of Conditional Treatment Effect Estimators Based on
Acknowledgments

I am deeply grateful for my advisors, Professors Tianxi Cai and LJ Wei. Not only have their insight and experience been invaluable, their encouragement and support helped me persevere through all my struggles in conducting this research. I would also like to thank Professor Hajime Uno for serving on my committee and generously sharing with me his time and advice over the years. A special thanks goes out to my former supervisor James Signorovitch who largely introduced me to the field of biostatistics and has supported me ever since.

I would like to thank all my friends from the Department of Biostatistics. Thank you for all the stimulating conversations and wonderful memories over the years from which I’ve learned so much. Thanks especially to Matthew Goodman, who was always game to discuss any statistics problem I had.

Thank you to my family and also friends from Boston Chinese Evangelical Church. They always wanted the best for me and provided me with support in my pursuits even when they did not understand exactly what I was doing. Their care for me through their kind words, sincere advice, and persistent prayer has sustained me throughout my time in the program.

This dissertation is dedicated to my parents, Ann Cai and Yunhui Cheng. From a young age they kindled in me an enthusiasm for science. Their unconditional love and sacrifice made possible the opportunities I have today. It is also dedicated to my wife Glory Song, who has stood by my side through all the late nights, weekends spent working, and times of uncertainty. Thank you for your love and for believing in me. I am grateful to be able to join alongside you as heirs of the grace of life. Finally, I thank God for His loving grace offered to me in Christ. To Him be the glory.
Estimating Average Treatment Effects with a Double-Index Propensity Score

David Cheng
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Abhishek Chakrabortty
Department of Statistics
University of Pennsylvania

Ashwin N. Ananthakrishnan
Division of Gastroenterology
Massachusetts General Hospital

Tianxi Cai
Department of Biostatistics
Harvard T.H. Chan School of Public Health
1.1 Introduction

While randomized clinical trials (RCT) remain the gold standard, large-scale observational data such as electronic medical record (EMR), mobile health, and insurance claims data are playing an increasing role in evaluating treatments in biomedical studies. Such observational data are valuable because they can be collected in contexts where RCTs cannot be run due to cost, ethical, or other feasibility issues (Rosenbaum, 2002). In the absence of randomization, adjustment for covariates $X$ that satisfy “no unmeasured confounding” (or “ignorability” or “exchangeability”) assumptions is needed to avoid bias from confounding. Many methods based on propensity score (PS), outcome regression, or a combination have been developed to estimate treatment effects under these assumptions (for a review of the basic approaches see Lunceford and Davidian (2004) and Kang and Schafer (2007)). These methods were initially developed in settings where $p$, the dimension of $X$, was small relative to the sample size $n$. Modern observational data however tend to include a large number of variables with little prior knowledge known about them.

Regardless of the size of $p$, adjustment for different covariates among all observed $X$ yields different effects. Let $A_\pi$ index the subset of $X$ that contributes to the PS $\pi_1(x) = \mathbb{P}(T = 1 \mid X = x)$, where $T \in \{0, 1\}$ is the treatment. Let $A_\mu$ index the subset of $X$ contributing to either $\mu_1(x)$ or $\mu_0(x)$, where $\mu_k(x) = \mathbb{E}(Y \mid X = x, T = k)$ and $Y$ is an outcome. Beyond adjusting for covariates indexed in $A_\pi \cap A_\mu$ to mitigate confounding, it is well-known that adjusting for covariates in $A_\pi^c \cap A_\mu$ improves the efficiency of PS-based treatment effect estimators, whereas adjusting for covariates in $A_\pi \cap A_\mu^c$ decreases efficiency (Lunceford and Davidian, 2004; Hahn, 2004; Brookhart et al., 2006; Rotnitzky et al., 2010; De Luna et al., 2011). This parallels similar phenomena that occur when adjusting for covariates in regression (Tsiatis et al., 2008; De Stavola and Cox, 2008).

Early studies using PS-based approaches cautioned against excluding variables among $X$ to avoid incurring bias from excluding confounders despite potential efficiency loss (Perkins et al., 2000; Rubin and Thomas, 1996). VanderWeele and Shpitser (2011) proposed a simple criteria to adjust for covariates that are known to be either a cause of $T$
or $Y$. These strategies are not feasible when $p$ is large. Initial data-driven approaches based on screening marginal associations between covariates in $X$ with $T$ and $Y$ were considered in Schneeweiss et al. (2009) and Hirano and Imbens (2001). These approaches, however, can be misleading because marginal associations need not agree with conditional associations. Vansteelandt et al. (2012) and Gruber and van der Laan (2015) distinguish the problem of variable selection for estimating causal effects from variable selection for building predictive models and propose stepwise procedures focusing on optimizing treatment effect estimation.

More recently, a number of authors considered regularization methods for variable selection in treatment effect estimation. For example, Belloni et al. (2013) estimated the joint support $A_\pi \cup A_\mu$ through regularization and obtained treatment effects through a partially linear regression model. Belloni et al. (2017) then considered estimating treatment effects using orthogonal estimating equations (i.e. those of the doubly-robust (DR) estimator (Robins et al., 1994)) using regularization to estimate models for $\pi_1(x)$ and $\mu_k(x)$. Farrell (2015) similarly considers estimating treatment effects with the DR estimator using group LASSO to group main and interaction effects in the outcome model. These papers focus on developing theory for existing treatment effect estimators to allow for valid inferences following variable selection. Wilson and Reich (2014) considered estimating $A_\pi \cup A_\mu$ through a regularized loss function, which simplifies to an adaptive LASSO (Zou, 2006) problem with a modified penalty that selects out covariates not conditionally associated with either $T$ or $Y$. Treatment effects are estimated through an outcome model after selection. Shortreed and Ertefaie (2017) proposed an approach that also modifies the weights in an adaptive LASSO for a PS model to select out covariates in $A_x \cap A_\mu^c$, estimating the final treatment effect through an inverse probability weighting (IPW) estimator. These approaches modify existing regularization techniques to identify relevant covariates, but there is limited theory to support their performance compared to existing causal inference estimators used with regularization. Furthermore, by relying exclusively on PS- and outcome regression-based approaches to estimate the treatment effect in the end, the double-robustness property is often forfeited.

Alternatively, Koch et al. (2017) proposed an adaptive group LASSO to estimate mod-
els for $\pi_k(x)$ and $\mu_k(x)$ through simultaneously minimizing the sum of their loss functions with a group LASSO penalty grouping together coefficients between the models for the same covariate. The penalty is also weighted by the inverse of the association between each covariate and $Y$ from an initial outcome model to select out covariates belonging to $A_\pi \cap A_\mu^c$. The estimated nuisance functions are plugged into the standard doubly-robust estimator to estimate the treatment effect. However, selecting out covariates in $A_\pi \cap A_\mu^c$ may inadvertently induce a misspecified model for $\pi_1(x)$ when it is in fact correctly specified given covariates indexed in $A_\pi$. Moreover, the asymptotic distribution is not identified when the nuisance functions are estimated, making it difficult to compare its efficiency with other methods. Bayesian model averaging provides an alternative to regularization methods for variable selection in causal inference problems (Wang et al., 2012; Zigler and Dominici, 2014; Talbot et al., 2015). These methods, however, rely on strong parametric assumptions and encounters burdensome computations when $p$ is not small. Cefalu et al. (2017) applied Bayesian model averaging to doubly-robust estimators, averaging doubly-robust estimates over posterior model probabilities of a large collection of combinations of parametric models for the nuisance functions. Priors that prefer models not including covariates belonging to $A_\pi \cap A_\mu^c$ ease the computations. Despite this innovation, the computations could still be burdensome and possibly unfeasible for large $p$. Most of the aforementioned methods did not consider asymptotic properties allowing $p$ to diverge with $n$.

We consider an IPW-based approach to estimate treatment effects with possibly high-dimensional $X$. We first use regularized regression to estimate a parametric model for $\pi_1(x)$. Since this neglects associations between $X$ and $Y$, we also use regularized regression to estimate a model for $\mu_k(x)$, for $k = 0, 1$. We then calibrate the initial PS estimates by performing nonparametric smoothing of $T$ over the linear predictors for $X$ from both the initial PS and outcome models. Smoothing over the linear predictors from the outcome model, which can be viewed as a working prognostic score (Hansen, 2008a), uses the variation in $X$ predictive of $Y$ to inform estimation of the calibrated PS. We show that our proposed estimator is doubly-robust and locally semiparametric efficient for the ATE under a nonparametric model. Moreover, we show that it achieves potentially sub-
stantial gains in robustness and efficiency under misspecification of working models for $\pi_1(x)$ and $\mu_k(x)$. The results are shown to hold allowing $p$ to diverge with $n$ under sparsity assumptions with suitable regularization. The broad approach is similar to a method proposed for estimating mean outcomes in the presence of data missing at random (Hu et al., 2012), except we use the double-score to estimate a PS instead of an outcome model. In contrast to their results, we show that a higher-order kernel is required due to the two-dimensional smoothing, find explicit efficiency gains under misspecification of the outcome model, and consider asymptotics with diverging $p$. The combined use of PS and prognostic scores has also recently been suggested for matching and subclassification (Leacy and Stuart, 2014), but the theoretical properties have not been established.

The rest of this paper is organized as follows. The method is introduced from Section 2.2 to Section 1.4 and its asymptotic properties are discussed in Section 1.5. A perturbation-resampling method is proposed for inference in Section 1.6. Numerical studies including simulations and applications to estimating treatment effects in an EMR study and a cohort study with a large number of covariates is presented in Section 1.7. We conclude with some additional discussion in Section 1.8. Regularity conditions and proofs are relegated to the Appendix.

1.2 Notations and Problem Setup

Let $Z_i = (Y_i, T_i, X_i^T)^T$ be the observed data for the $i$th subject, where $Y_i$ is an outcome that could be modeled by a generalized linear model (GLM), $T_i \in \{0, 1\}$ a binary treatment, and $X_i$ is a $p$-dimensional vector of covariates with compact support $X \subseteq \mathbb{R}^p$. Here we allow $p$ to diverge with $n$ such that such that $\log(p)/\log(n) \to \nu$, for $\nu \in [0, 1)$. The observed data consists of $n$ independent and identically distributed (iid) observations $\mathcal{D} = \{Z_i : i = 1, \ldots, n\}$ drawn from the distribution $\mathbb{P}$. Let $Y_i^{(1)}$ and $Y_i^{(0)}$ denote the counterfactual outcomes (Rubin, 1974) had an individual been treated with treatment 1 or 0. Based on $\mathcal{D}$, we want to make inferences about the average treatment effect (ATE):

$$\Delta = \mathbb{E}\{Y^{(1)}\} - \mathbb{E}\{Y^{(0)}\} = \mu_1 - \mu_0. \quad (1.1)$$
For identifiability, we require the following standard causal inference assumptions:

\[ Y = TY^{(1)} + (1 - T)Y^{(0)} \text{ with probability 1} \]  
\[ \pi_1(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for some } \epsilon_\pi > 0, \text{ when } x \in \mathcal{X} \]  
\[ Y^{(1)} \perp\!\perp T \mid X \text{ and } Y^{(0)} \perp\!\perp T \mid X, \]

where \( \pi_k(x) = P(T = k \mid X = x) \), for \( k = 0, 1 \). Through the third condition we assume from the onset no unmeasured confounding holds given the entire \( X \), which could be plausible when a large set of baseline covariates are included in \( X \). Under these assumptions, it is well-known that \( \Delta \) can be identified from the observed data distribution \( \mathbb{P} \) through the g-formula (Robins, 1986):

\[ \Delta^* = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} = \mathbb{E}\left\{ \frac{I(T = 1)Y}{\pi_1(X)} - \frac{I(T = 0)Y}{\pi_0(X)} \right\}, \]  

where \( \mu_k(x) = \mathbb{E}(Y \mid X = x, T = k) \), for \( k = 0, 1 \). We will consider an estimator based on the IPW form that will nevertheless be doubly-robust so that it is consistent under parametric models where either \( \pi_k(x) \) or \( \mu_k(x) \) is correctly specified.

### 1.3 Parametric Models for Nuisance Functions

As \( X \) is high-dimensional, we consider parametric modeling as a means to reduce the dimensions of \( X \) when estimating the nuisance functions \( \pi_k(x) \) and \( \mu_k(x) \). For reference, let \( \mathcal{M}_{np} \) be the nonparametric model for the distribution of \( Z, \mathbb{P} \), that has no restrictions on \( \mathbb{P} \) except requiring the second moment of \( Z \) to be finite. Let \( \mathcal{M}_\pi \subseteq \mathcal{M}_{np} \) and \( \mathcal{M}_\mu \subseteq \mathcal{M}_{np} \) respectively denote parametric working models under which:

\[ \pi_1(x) = g_\pi(\alpha_0 + \alpha^T x), \]  
\[ \text{and } \mu_k(x) = g_\mu(\beta_0 + \beta_1 k + \beta_k^T x), \text{ for } k = 0, 1, \]

where \( g_\pi(\cdot) \) and \( g_\mu(\cdot) \) are known link functions, and \( \bar{\alpha} = (\alpha_0, \alpha^T) \in \Theta_\alpha \subseteq \mathbb{R}^{p+1} \) and \( \bar{\beta} = (\beta_0, \beta_1, \beta_0^T, \beta_1^T) \in \Theta_\beta \subseteq \mathbb{R}^{2p+2} \) are unknown parameters. The specifications in (1.6) or (1.7) could be made more flexible by applying basis expansion functions, such as splines, to \( X \). In (1.7) slopes are allowed to differ by treatment arms to allow for heterogeneous
effects of \( X \) between treatments. In data where it is reasonable to assume heterogeneity is weak or nonexistent, it may be beneficial for efficiency to restrict \( \beta_0 = \beta_1 \), in which case (1.7) is simply a main effects model. We discuss concerns about ancillarity with the main effects model in the Section 1.8.

Regardless of the validity of either working model (i.e. whether \( P \) belongs in \( \mathcal{M}_\pi \cup \mathcal{M}_\mu \)), we first obtain estimates of \( \alpha \) and \( \beta_k \)’s through regularized estimation:

\[
(\hat{\alpha}_0, \hat{\alpha}^T) = \arg\min_{\bar{\alpha}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\pi(\bar{\alpha}; T_i, X_i) + p_\pi(\bar{\alpha}_{(-1)}; \lambda_n) \right\}
\]

(1.8)

\[
(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_0^T, \hat{\beta}_1^T) = \arg\min_{\bar{\beta}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\mu(\bar{\beta}; Z_i) + p_\mu(\bar{\beta}_{(-1)}; \lambda_n) \right\},
\]

(1.9)

where \( \ell_\pi(\bar{\alpha}; T_i, X_i) \) denotes the log-likelihood for \( \bar{\alpha} \) under \( \mathcal{M}_\pi \) given \( T_i, X_i \), and \( \ell_\mu(\bar{\beta}; Z_i) \) is a log-likelihood for \( \bar{\beta} \) from a GLM suitable for the outcome type of \( Y \) under \( \mathcal{M}_\mu \) given \( Z_i \), and, for any vector \( v \), \( v_{(-1)} \) denotes the subvector of \( v \) excluding the first element.

We require penalty functions \( p_\pi(u; \lambda) \) and \( p_\mu(u; \lambda) \) to be chosen such that the oracle properties (Fan and Li, 2001) hold. An example is the adaptive LASSO (Zou, 2006), where \( p_\pi(\bar{\alpha}_{(-1)}; \lambda_n) = \lambda_n \sum_{j=1}^{p} |\alpha_j| / |\tilde{w}_{\pi,j}| \) with initial weights \( \tilde{w}_{\pi,j} \) estimated from ridge regression, tuning parameter \( \lambda_n \) such that \( \lambda_n \sqrt{n} \to 0 \) and \( \lambda_n n^{(1-\nu)(1+\gamma)/2} \to \infty \), and \( \gamma > 2\nu/(1-\nu) \) (Zou and Zhang, 2009). When additional structure is known, other penalties yielding the oracle properties such as adaptive elastic net (Zou and Zhang, 2009) or Group LASSO (Wang and Leng, 2008) penalties can also be used. In principle other variable selection and estimation procedures (Wilson and Reich, 2014; Koch et al., 2017) can also be used, but the theoretical properties may be difficult to verify without the oracle properties.

We assume the true \( \alpha \) and \( \beta_k \)’s to be sparse when working models are correctly specified (i.e. when \( P \) belongs to \( \mathcal{M}_\pi \cap \mathcal{M}_\mu \)). For example, in EMRs, a large number of covariates extracted from codified or narrative data can be expected to actually be irrelevant to either the outcome or the treatment assignment processes. Even when \( \mathcal{M}_\pi \) or \( \mathcal{M}_\mu \) is not exactly correctly specified, we assume that the correlation between covariates are not so strong such that the limiting values of \( \hat{\alpha} \) and \( \hat{\beta}_k \), \( \bar{\alpha} \) and \( \bar{\beta}_k \), are still sparse. The oracle properties then ensure that the correct support can be selected in large samples. Let \( A_\alpha \) and \( A_\beta_k \) denote the respective support of \( \bar{\alpha} \) and \( \bar{\beta}_k \), regardless of the validity of either
working model. By selecting out irrelevant covariates belonging to \( A^c_\alpha \cap A^c_\beta \) when estimating the PS \( \pi_k(x) \), the efficiency of subsequent IPW estimators for \( \mu_k \) can be expected to improve substantially when \( p \) is not small. The regularization in (1.8) selects all variables belonging to \( A_\alpha \), which guards against misspecification of the model in (1.6) if variables in \( A_\alpha \) are selected out. But applying regularization to estimate the PS directly through (1.8) only may be inefficient because covariates belonging to \( A^c_\alpha \cap A^c_\beta \) would be selected out (Lunceford and Davidian, 2004; Brookhart et al., 2006). In the following we consider a calibrated PS based on both \( \hat{\alpha}^T X \) and \( \hat{\beta}^T_k X \) that addresses this shortcoming.

### 1.4 Double-Index Propensity Score and IPW Estimator

To mitigate the effects of misspecification of (1.6), one could calibrate an initial PS estimate \( g_\pi(\hat{\alpha}_0 + \hat{\alpha}^T X) \) by performing nonparametric smoothing of \( T \) over \( \hat{\alpha}^T X \). This adjusts the initial estimates \( g_\pi(\hat{\alpha}_0 + X^T \hat{\alpha}) \) closer to the true probability of receiving treatment 1 given \( \hat{\alpha}^T X \). However, we consider a smoothing over not only \( \hat{\alpha}^T X \) but also \( \hat{\beta}^T_k X \) as well to allow variation in covariates predictive of the outcome, i.e. covariates indexed in \( A^\beta_k \), to inform this calibration. In other words, this serves to as a means to include covariates indexed in \( A^\beta_k \) in the smoothing, except that such covariates are initially reduced into \( \hat{\beta}^T_k X \). The double-index PS (DiPS) estimator for each treatment is thus given by:

\[
\hat{\pi}_k(x; \hat{\theta}_k) = \frac{n^{-1} \sum_{j=1}^{n} K_h\{ (\hat{\alpha}, \hat{\beta}_k)^T (X_j - x) \} I(T_j = k)}{n^{-1} \sum_{j=1}^{n} K_h\{ (\hat{\alpha}, \hat{\beta}_k)^T (X_j - x) \}}, \text{ for } k = 0, 1,
\]  

(1.10)

where \( \hat{\theta}_k = (\hat{\alpha}^T, \hat{\beta}_k^T)^T \), \( K_h(u) = h^{-2} K(u/h) \), and \( K(u) \) is a bivariate \( q \)-th order kernel function with \( q > 2 \). A higher-order kernel is required here for the asymptotics to be well-behaved, which is the price for estimating the nuisance functions \( \pi_k(x) \) using two-dimensional smoothing. This allows for the possibility of negative values for \( \hat{\pi}_k(x; \hat{\theta}_k) \). Nevertheless, \( \hat{\pi}_k(x; \hat{\theta}_k) \) are nuisance estimates not of direct interest, and we find in numerical studies that negative values occur infrequently and do not compromise the performance of the final estimator. A monotone transformation of the input scores for each treatment \( \tilde{S}_k = (\hat{\alpha}, \hat{\beta}_k)^T X \) can be applied prior to smoothing to improve finite sample performance (Wand et al., 1991). In numerical studies, for instance, we applied a a prob-
ability integral transform based on the normal cumulative distribution function to the standardized scores to obtain approximately uniformly distributed inputs. We also scaled components of \( \hat{S}_k \) such that a common bandwidth \( h \) can be used for both components of the score.

With \( \pi(x) \) estimated by \( \hat{\pi}_k(x; \hat{\theta}_k) \), the estimator for \( \Delta \) is given by \( \hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0 \), where:

\[
\hat{\mu}_k = \left\{ \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \hat{\theta}_k)} \right\}^{-1} \left\{ \sum_{i=1}^{n} \frac{I(T_i = k)Y_i}{\hat{\pi}_k(X_i; \hat{\theta}_k)} \right\}^{-1}, \text{ for } k = 0, 1. \tag{1.11}
\]

This is the usual normalized IPW estimator, where, the PS is given by the double-index PS estimates. In the following, we show that this simple construction leads to an estimator that also possesses the robustness and efficiency properties of the doubly-robust estimator derived from semiparametric efficiency theory (Robins et al., 1994), and, in certain scenarios, achieves additional gains in robustness and efficiency.

### 1.5 Asymptotic Robustness and Efficiency Properties

We first directly present the influence function expansion of \( \hat{\Delta} \) in general. Robustness and efficiency results are subsequently derived based on the expansion. To present the influence function expansion, let \( \bar{\Delta} = \bar{\mu}_1 - \bar{\mu}_0 \) be the limiting estimand, with:

\[
\bar{\mu}_k = \mathbb{E} \left\{ \frac{I(T_i = k)Y_i}{\pi_k(X_i; \theta_k)} \right\}, \text{ for } k = 0, 1,
\]

\( \theta_k = (\alpha^T, \beta_k^T)^T \), and \( \pi_k(x; \theta_k) = \mathbb{P}(T_i = k \mid \alpha^T X_i = \alpha^T x, \beta_k^T X_i = \beta_k^T x) \). Moreover, for any vector \( v \) of length \( p \) and any index set \( A \subseteq \{1, 2, \ldots, p\} \), let \( v_A \) denote the subvector with elements indexed in \( A \). Let \( \hat{W}_k = n^{1/2}(\hat{\mu}_k - \bar{\mu}_k) \) for \( k = 0, 1 \) so that \( n^{1/2}(\hat{\Delta} - \bar{\Delta}) = \hat{W}_1 - \hat{W}_0 \).

We show in Appendix D the following result.

**Theorem 1.** Suppose that causal assumptions (1.2), (1.3), (1.4) and the regularity conditions in Appendix A hold. Let the sparsity of \( \alpha \) and \( \beta_k \), \( |A_\alpha| = s_\alpha \) and \( |A_{\beta_k}| = s_{\beta_k} \), be fixed. If \( \log(p)/\log(n) \rightarrow \nu \) for \( \nu \in [0, 1) \), then, with probability tending to 1, \( \hat{W}_k \) has the expansion:

\[
\hat{W}_k = n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)Y_i}{\pi_k(X_i; \theta_k)} - \left\{ \frac{I(T_i = k)}{\pi_k(X_i; \theta_k)} - 1 \right\} \mathbb{E}(Y_i \mid \alpha^T X_i, \beta_k^T X_i, T_i = k) - \bar{\mu}_k \tag{1.12}
\]

\[
+ u_{k,A_\alpha} n^{1/2}(\hat{\alpha} - \alpha)_{A_\alpha} + v_{k,A_{\beta_k}} T n^{1/2}(\hat{\beta}_k - \beta_k)_{A_{\beta_k}} + O_p(n^{1/2}h^q + n^{-1/2}h^{-2}), \tag{1.13}
\]

9
for \( k = 0, 1 \), where \( u_{k, A\alpha} \) and \( u_{k, A\beta_k} \) are deterministic vectors, \( n^{1/2}(\hat{\alpha} - \alpha)_{A\alpha} = O_p(1) \), and \( n^{1/2}(\hat{\beta}_k - \bar{\beta}_k)_{A\beta_k} = O_p(1) \). Under model \( M_\pi \nu_{k, A\beta_k} = 0 \), for \( k = 0, 1 \). Under \( M_\pi \cap M_\mu \), we additionally have that \( u_{k, A\alpha} = 0 \), for \( k = 0, 1 \).

The challenge in showing Theorem 1 is to obtain an influence function expansion when the nuisance functions \( \pi_k(x) \) are estimated with two-dimensional kernel-smoothing rather than finite-dimensional models. We show in Appendix D that a V-statistic projection lemma (Newey and McFadden, 1994) can be applied to obtain the expansion in this situation.

Let \( \hat{\Delta}_{dr} \) denote the usual doubly-robust estimator with \( \pi_k(x) \) and \( \mu_k(x) \) estimated in the same way through (1.8) and (1.9). The influence function expansion for \( \hat{\Delta} \) in Theorem 1 is nearly identical to that of \( \hat{\Delta}_{dr} \). The terms in (1.12) are the same except \( \pi_k(X_i; \bar{\theta}_k) \) and \( E(Y_i | \alpha^T X_i, \beta^T X_i, T_i = k) \) replaces parametric models at the limiting estimates. Terms in (1.13) analogously represent the additional contributions from estimating the nuisance parameters. The expansion also shows that asymptotically no contribution from smoothing is incurred. This similarity in the influence functions yields similar desirable robustness and efficiency properties, which are improved upon in some cases due to the calibration through smoothing.

1.5.1 Robustness

In terms of robustness, under \( M_\pi \), \( \pi_k(x; \bar{\theta}_k) = \pi_k(x) \) so the limiting estimands are:

\[
\bar{\mu}_k = E \left\{ \frac{I(T_i = k)Y_i}{\pi_k(X_i)} \right\} = E\{Y_i^{(k)}\}, \text{ for } k = 0, 1.
\]

On the other hand, under \( M_\mu \), \( E(Y_i | \alpha^T X_i, \bar{\theta}_k^T X_i = \beta^T X_i, T_i = k) = \mu_k(x) \) so that:

\[
\bar{\mu}_k = E \{ E(Y_i | \alpha^T X_i, \beta^T X_i, T_i = k) \} = E \{ \mu_k(X_i) \} = E\{Y_i^{(k)}\}, \text{ for } k = 0, 1.
\]

Thus by Theorem 1, under \( M_\pi \cup M_\mu \), \( \hat{\Delta} - \Delta = O_p(n^{-1/2}) \) provided that \( h = O(n^{-\alpha}) \) for \( \alpha \in (\frac{1}{2q}, \frac{1}{4}) \). That is, \( \hat{\Delta} \) is doubly-robust for \( \Delta \). Beyond this usual form of double-robustness, if the PS model specification (1.6) is incorrect, we expect the calibration step to at least partially correct for the misspecification since, in large samples, given \( x, \pi_k(x; \bar{\theta}_k) \) is closer
to the true $\pi_k(x)$ than the misspecified parametric model $g_\pi(\tilde{\alpha}_0 + \alpha^T x)$. In some specific scenarios, the calibration can completely overcome the misspecification of the PS model. For example, let $\tilde{M}_\pi$ denote a model for $P$ under which:

$$\pi_1(x) = \tilde{g}_\pi(\alpha^T x)$$

for some unknown link function $\tilde{g}_\pi(\cdot)$ and unknown $\alpha \in \mathbb{R}^p$, and $X$ are known to be elliptically distributed such that $E(a^T X \mid \alpha_s^T X)$ exists and is linear in $\alpha_s^T X$, where $\alpha_s$ denotes the true $\alpha$ (e.g. if $X$ is multivariate normal). Then, using the results of Li and Duan (1989), it can be shown that $\tilde{\alpha} = c\alpha_s$ for some scalar $c$. But since $\tilde{\pi}_k(x; \tilde{\theta}_k)$ recovers $\pi_k(x; \theta_k) = P(T = k \mid \tilde{\alpha}^T X = \tilde{\alpha}_k^T x, \beta_k^T X = \beta_k^T x)$ asymptotically, it also recovers $\pi_k(x)$ under $\tilde{M}_\pi$. Consequently, $\tilde{\Delta}$ is more than doubly-robust in that $\tilde{\Delta} - \Delta = O_p(n^{-1/2})$ under the larger model $M_\pi \cup \tilde{M}_\pi \cup M_\mu$. The same phenomenon also occurs when estimating $\beta_k$ under misspecification of the link in (1.7), if we do not assume $\beta_0 = \beta_1$ and use a common model to estimate the $\beta_k$'s. In this case, if $\tilde{M}_\mu$ is an analogous model under which:

$$\mu_1(x) = \tilde{g}_{\mu,1}(\beta_1^T x) \quad \text{and} \quad \mu_0(x) = \tilde{g}_{\mu,0}(\beta_0^T x)$$

for some unknown link functions $\tilde{g}_{\mu,0}(\cdot)$ and $\tilde{g}_{\mu,1}(\cdot)$ and $X$ are elliptically distributed, then $\tilde{\Delta} - \Delta = O_p(n^{-1/2})$ under the even larger model $M_\pi \cup \tilde{M}_\pi \cup M_\mu \cup \tilde{M}_\mu$. This does not hold when $\beta_0 = \beta_1$ is assumed, as $T$ is binary so $(T, X^T)^T$ is not exactly elliptically distributed. But the result may still be expected to hold approximately when $X$ is elliptically distributed.

### 1.5.2 Efficiency

In terms efficiency, let the terms contributed to the influence function for $\tilde{\Delta}$ when $\alpha$ and $\beta_k$ are known be:

$$\varphi_{i,k} = \frac{I(T_i = k)Y_i}{\pi_k(X_i; \tilde{\theta}_k)} - \left\{ \frac{I(T_i = k)}{\pi_k(X_i; \theta_k)} - 1 \right\} E(Y_i \mid \tilde{\alpha}^T X_i, \tilde{\beta}_k^T X_i, T_i = k) - \bar{\mu}_k. \tag{1.14}$$

Under $M_\pi \cap M_\mu$, $\varphi_{i,k}$ is the full influence function for $\tilde{\Delta}$. This influence function is the efficient influence function for $\Delta^*$ under $M_{np}$ since $E(Y_i \mid \tilde{\alpha}^T X_i = \tilde{\alpha}_k^T x, \tilde{\beta}_k^T X_i = \beta_k^T x, T_i = k) = \mu_k(x)$ and $\pi_k(x; \tilde{\theta}_k) = \pi_k(x)$ for all $x \in X$ at distributions for $P$ belonging to $M_\pi \cap$
\( \mathcal{M}_\mu \). An important consequence of this is that \( \hat{\Delta} \) reaches the semiparametric efficiency bound under \( \mathcal{M}_{np} \), at distributions belonging to \( \mathcal{M}_\pi \cap \mathcal{M}_\mu \). That is, \( \hat{\Delta} \) is also a \textit{locally semiparametric efficient} estimator under \( \mathcal{M}_{np} \). In addition, based on the same arguments as in the preceding subsection, the local efficiency property could potentially be broadened so that \( \hat{\Delta} \) is locally efficient at distributions for \( \mathbb{P} \) belonging to \( (\mathcal{M}_\pi \cup \tilde{\mathcal{M}}_\pi) \cap (\mathcal{M}_\mu \cup \tilde{\mathcal{M}}_\mu) \).

Beyond this characterization of efficiency similar to that of \( \hat{\Delta}_{dr} \), there are additional benefits of \( \hat{\Delta} \) under model \( \mathcal{M}_\pi \cap \mathcal{M}_\mu^c \). In this case, akin to \( \hat{\Delta}_{dr} \), estimating \( \beta_k \) does not contribute to the asymptotic variance since \( v_{k,\mathcal{A}_\beta} = 0 \), and a similar \( n^{1/2}u_{k,\mathcal{A}_\alpha}(\hat{\alpha} - \alpha)_{A\alpha} \) term is contributed from estimating \( \alpha \). The analogous term in the expansion for \( \hat{\Delta}_{dr} \) contributes the negative of a projection of the preceding terms onto the space of the linear span of the scores for \( \alpha \) (restricted to \( \mathcal{A}_\alpha \)) to its influence function. We show in Appendix D the same interpretation can be adopted for \( \hat{\Delta} \).

**Theorem 2.** Let \( U_\alpha \) be the score for \( \alpha \) under \( \mathcal{M}_\pi \) and let \( [U_{\alpha,\mathcal{A}_\alpha}] \) denote the linear span of its components indexed in \( \mathcal{A}_\alpha \). In the Hilbert space of random variables with mean 0 and finite variance \( \mathcal{L}_2^0 \) with inner product given by the covariance, let \( \Pi\{V \mid S\} \) denote the projection of some \( V \in \mathcal{L}_2^0 \) into a subspace \( S \subseteq \mathcal{L}_2^0 \). If the assumptions required for Theorem 1 hold, under \( \mathcal{M}_\pi \), \( u_{k,\mathcal{A}_\alpha} n^{1/2}(\hat{\alpha} - \alpha)_{A\alpha} = -n^{-1/2} \sum_{i=1}^n \Pi\{\varphi_{i,k} \mid [U_{\alpha,\mathcal{A}_\alpha}]\} + o_p(1) \).

This result implies through the Pythagorean theorem the familiar efficiency paradox that under \( \mathcal{M}_\pi \cap \mathcal{M}_\mu^c \), \( \hat{\Delta} \) would be more efficient if \( \alpha \) were estimated, even if the true \( \alpha \) were known (Lunceford and Davidian, 2004). Moreover, under \( \mathcal{M}_\pi \cap \mathcal{M}_\mu^c \), the influence function of \( \hat{\Delta} \) involves projecting \( \varphi_{i,k} \) rather than:

\[
\phi_{i,k} = \frac{I(T_i = k)Y_i}{\pi_k(X_i)} - \left\{ \frac{I(T_i = k)}{\pi_k(X_i)} - 1 \right\} g_\mu(\hat{\beta}_0 + \hat{\beta}_1 k + \hat{\beta}_k^T X_i) - \hat{\mu}_k,
\]

which are the preceding influence function terms for \( \hat{\Delta}_{dr} \). But since \( \mathbb{E}(Y_i \mid \alpha^T X_i = \alpha^T x, \beta_k^T X_i = \beta_k^T x, T_i = k) \) better approximates \( \mu_k(x) \) than the limiting parametric model \( g_\mu(\beta_0 + \beta_1 k + \beta_k^T x) \), it can be shown that \( \mathbb{E}(\phi_{i,k}^2) > \mathbb{E}(\varphi_{i,k}^2) \) for \( k = 0, 1 \). Given that under \( \mathcal{M}_\pi \cap \mathcal{M}_\mu^c \) the influence function for both \( \hat{\Delta} \) and \( \hat{\Delta}_{dr} \) are then in the form of a residual after projecting \( \varphi_{i,k} \) and \( \phi_{i,k} \) onto the same space, the asymptotic variance of \( \hat{\Delta} \) can be seen to be less than \( \hat{\Delta}_{dr} \). That is, \( \hat{\Delta} \) is more efficient than \( \hat{\Delta}_{dr} \) under \( \mathcal{M}_\pi \cap \mathcal{M}_\mu^c \). We show
in the simulation studies that this improvement can lead to substantial efficiency gains under $M_\pi \cap M_\mu$ in finite samples. These robustness and efficiency properties distinguish $\hat{\Delta}$ from the usual doubly-robust estimators and their variants. Moreover, despite being motivated by data with high-dimensional $p$, these properties still hold if $p$ is small relative to $n$, which makes $\hat{\Delta}$ effective in low-dimensional settings as well. We next consider a perturbation scheme to estimate standard errors (SE) and confidence intervals (CI) for $\hat{\Delta}$.

1.6 Perturbation Resampling

Although the asymptotic variance of $\hat{\Delta}$ can be determined through its influence function specified in Theorem (1), a direct empirical estimate based on the influence function is difficult because $u_k, A_\alpha$ and $v_k, A_\beta$ involve complicated functionals of $P$ that are difficult to estimate. Instead we propose a simple perturbation-resampling procedure. Let $G = \{G_i : i = 1, \ldots, n\}$ be a set of non-negative iid random variables with unit mean and variance that are independent of $\mathcal{D}$. The procedure then perturbs each “layer” of the estimation of $\hat{\Delta}$. Let the perturbed estimates of $\tilde{\alpha}$ and $\tilde{\beta}$ be:

$$(\hat{\alpha}_0, \hat{\alpha}^* T)^T = \arg \min_{\tilde{\alpha}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\pi(\tilde{\alpha}; T_i, X_i)G_i + p_\pi(\tilde{\alpha}_{-1}; \lambda_n) \right\}$$

$$(\hat{\beta}_0, \hat{\beta}^*_T , \hat{\beta}_0^*, \hat{\beta}^* T)^T = \arg \min_{\tilde{\beta}} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\mu(\tilde{\beta}; Z_i)G_i + p_\mu(\tilde{\beta}_{-1}; \lambda_n) \right\}.$$  

The perturbed DiPS estimates are calculated by:

$$\hat{\pi}_k(x; \tilde{\theta}_k^*) = \frac{\sum_{j=1}^{n} K_h \{(\hat{\alpha}^*, \hat{\beta}_k^*)(X_j - x)\} I(T_j = k)G_j}{\sum_{j=1}^{n} K_h \{(\hat{\alpha}^*, \hat{\beta}_k^*)(X_j - x)\} G_j}, \text{ for } k = 0, 1. \tag{1.15}$$

Lastly the perturbed estimator is given by $\hat{\Delta}^* = \hat{\mu}_1^* - \hat{\mu}_0^*$ where:

$$\hat{\mu}_k = \left\{ \sum_{i=1}^{n} I(T_i = k) \hat{\pi}_k \left( X_i; \tilde{\theta}_k^* \right) G_i \right\}^{-1} \left\{ \sum_{i=1}^{n} I(T_i = k) Y_i \hat{\pi}_k \left( X_i; \tilde{\theta}_k^* \right) G_i \right\}^{-1}, \text{ for } k = 0, 1.$$  

It can be shown based on arguments similar to those in Tian et al. (2007) that the asymptotically distribution of $n^{1/2}(\hat{\Delta} - \bar{\Delta})$ coincides with that of $n^{1/2}(\hat{\Delta}^* - \bar{\Delta}) | \mathcal{D}$. We can thus
approximate the SE of $\hat{\Delta}$ based on the empirical standard deviation or, as a robust alternative, the empirical mean absolute deviations (MAD) of a large number of resamples $\hat{\Delta}^*$ and construct CI’s using empirical percentiles of resamples of $\hat{\Delta}^*$.

### 1.7 Numerical Studies

#### 1.7.1 Simulation Study

We performed extensive simulations to assess the finite sample bias and relative efficiency (RE) of $\hat{\Delta}$ (DiPS) compared to alternative estimators. We also assessed in a separate set of simulations the performance of the perturbation procedure. Adaptive LASSO was used to estimate $\alpha$ and $\beta_k$, and a Gaussian product kernel of order $q = 4$ with a plug-in bandwidth at the optimal order (see Section 1.8) was used for smoothing. Alternative estimators include an IPW with $\pi_k(x)$ estimated by adaptive LASSO (ALAS), $\hat{\Delta}_{dr}$ (DR-ALAS) with nuisances estimated by adaptive LASSO, the DR estimator with nuisance functions estimated by “rigorous” LASSO (DR-rLAS) of Belloni et al. (2017), outcome-adaptive LASSO (ALS) of Shortreed and Ertefaie (2017), Group Lasso and Doubly Robust Estimation (GLiDeR) of Koch et al. (2017), and model average double robust (MADR) of Cefalu et al. (2017). ALS and GLiDeR were implemented with default settings from code provided in the Supplementary Materials of the respective papers. DR-rLAS was implemented using the hdm R package (Chernozhukov et al., 2016) with default settings, and MADR was implemented using the madr package with $M = 500$ MCMC iterations to reduce the computations. Throughout the numerical studies, unless noted otherwise, we postulated $g_\pi(u) = 1/(1 + e^{-u})$ for $M_\pi$ and $g_\mu(u) = u$ with $\beta_0 = \beta_1$ for $M_\mu$ as the working models. For adaptive LASSO, the tuning parameter was chosen by an extended regularized information criteria (Hui et al., 2015), which showed good performance for variable selection. We re-fitted models with selected covariates as suggested in Hui et al. (2015) to reduce bias.

We focused on a continuous outcome in the simulations, generating the data according to:

$$X \sim N(0, \Sigma), \quad T \mid X \sim Ber\{\pi_1(X)\}, \quad \text{and} \quad Y \mid X, T \sim N\{\mu_T(X), 10^2\}.$$
where \(\Sigma = (\sigma_{ij}^2)\) with \(\sigma_{ij}^2 = 1\) if \(i = j\) and \(\sigma_{ij}^2 = A(.5)^{|i-j|/3}I(|i - j| \leq 15)\) if \(i \neq j\). The simulations were varied over scenarios where working models were correct or misspecified:

Both correct: \(\pi_1(x) = g_\pi(\alpha^T x), \quad \mu_k(x) = k + \beta_k^T x\)

Misspecified \(\mu_k(x)\): \(\pi_1(x) = g_\pi(\alpha^T x), \quad \mu_k(x) = k + 3 \{\beta_k^T x(\beta_k^T x + 3)\}^{1/3}\)

Misspecified \(\pi_k(x)\): \(\pi_1(x) = g_\pi \{-1 + \alpha_1^T x(0.5\alpha_2^T x + .5)\}, \quad \mu_k(x) = k + \beta_k^T x\)

Both misspecified: \(\pi_1(x) = g_\pi \{-1 + \alpha_1^T x(0.5\alpha_2^T x + .5)\}, \mu_k(x) = k + 3 \{\beta_k^T x(\beta_k^T x + 3)\}^{1/3}\).

The parameters are given by:

\[
\alpha = (4, -3, .4, .32, .4, .3, .3, .0_{p-10})^T, \quad \beta_0 = \beta_1 = (-1, 1, -1, 1, 2, -1, -1, 2, -1, 0_{p-10})^T, \\
\alpha_1 = (.9, 0, -9, 0, .9, 0, .9, 0, -9, 0, 0_{p-10})^T, \quad \alpha_2 = (0, -6, 0, .6, 0, .6, 0, -6, 0, -6, 0_{p-10})^T,
\]

with \(a_n\) denoting a \(1 \times m\) vector that has all its elements as \(a\). In the misspecified \(\mu_k(x)\) scenario, \(\mu_k(x)\) is actually a single-index model among subjects with either \(T = 1\) or \(T = 0\), which allows for complex nonlinearities and interactions. Nevertheless, this is a genuine misspecification of \(\mu_k(x)\) (i.e. \(\mathbb{P}\) would not belong to \(\mathcal{M}_\mu \cup \mathcal{M}_\pi\)) because we assumed \(\beta_0 = \beta_1\) for \(\mathcal{M}_\mu\). A double-index model is used for misspecifying \(\pi_k(x)\) so that we consider when \(\mathbb{P}\) would not belong to \(\mathcal{M}_\pi \cup \mathcal{M}_\mu\). The simulations were run over \(R = 2,000\) repetitions.

Table 2.1 presents the bias and root mean square error (RMSE) for \(n = 1,000,5,000\) under the different specifications when \(p = 15\). Among the four scenarios considered, the bias for DiPS is small relative to the RMSE and generally diminishes to zero as \(n\) increases, verifying its double-robustness. In contrast, IPW-ALAS and OAL rely on consistent estimates of the PS and show non-negligible bias under the “misspecified \(\pi_1(x)\)” scenario. The extra robustness of DiPS is evident under the “both misspecified” scenario, where most other estimators incur non-negligible bias. As the true outcome model is close to being a single-index model with elliptical covariates, DiPS still estimates the average treatment effect \(\Delta\) with little bias. In results not shown, we also checked the bias for \(p\) up to \(p = 75\), where we generally found similar patterns. DiPS incurred around 1-2% more bias than other estimators when \(n\) was small relative to \(p\), which may be a consequence of smoothing in finite samples.
Table 1.1: Bias and RMSE of estimators by \( n \) and model specification scenario.

<table>
<thead>
<tr>
<th>Size</th>
<th>Estimator</th>
<th>Both Correct ( \mu_k(x) ) Bias</th>
<th>Both Correct ( \pi_1(x) ) RMSE</th>
<th>Misspecified ( \mu_k(x) ) Bias</th>
<th>Misspecified ( \pi_1(x) ) RMSE</th>
<th>Both misspecified Bias</th>
<th>Both misspecified RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=1,000</td>
<td>IPW-ALAS</td>
<td>0.009</td>
<td>0.279</td>
<td>0.001</td>
<td>0.414</td>
<td>-0.117</td>
<td>0.312</td>
</tr>
<tr>
<td></td>
<td>DR-ALAS</td>
<td>-0.002</td>
<td>0.252</td>
<td>-0.014</td>
<td>0.401</td>
<td>0.003</td>
<td>0.244</td>
</tr>
<tr>
<td></td>
<td>DR-rLAS</td>
<td>-0.057</td>
<td>0.275</td>
<td>-0.092</td>
<td>0.426</td>
<td>-0.022</td>
<td>0.284</td>
</tr>
<tr>
<td></td>
<td>OAL</td>
<td>0.002</td>
<td>0.263</td>
<td>-0.013</td>
<td>0.403</td>
<td>0.013</td>
<td>0.283</td>
</tr>
<tr>
<td></td>
<td>GLiDeR</td>
<td>-0.002</td>
<td>0.244</td>
<td>-0.016</td>
<td>0.379</td>
<td>0.004</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>MADR</td>
<td>-0.002</td>
<td>0.249</td>
<td>-0.015</td>
<td>0.402</td>
<td>0.002</td>
<td>0.239</td>
</tr>
<tr>
<td></td>
<td>DiPS</td>
<td>0.015</td>
<td>0.252</td>
<td>0.003</td>
<td>0.293</td>
<td>0.017</td>
<td>0.243</td>
</tr>
<tr>
<td>n=5,000</td>
<td>IPW-ALAS</td>
<td>-0.003</td>
<td>0.116</td>
<td>0.001</td>
<td>0.186</td>
<td>-0.127</td>
<td>0.181</td>
</tr>
<tr>
<td></td>
<td>DR-ALAS</td>
<td>-0.003</td>
<td>0.110</td>
<td>0.002</td>
<td>0.184</td>
<td>0.000</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>DR-rLAS</td>
<td>-0.003</td>
<td>0.110</td>
<td>0.002</td>
<td>0.184</td>
<td>0.037</td>
<td>0.122</td>
</tr>
<tr>
<td></td>
<td>OAL</td>
<td>-0.002</td>
<td>0.115</td>
<td>0.002</td>
<td>0.186</td>
<td>-0.098</td>
<td>0.155</td>
</tr>
<tr>
<td></td>
<td>GLiDeR</td>
<td>-0.003</td>
<td>0.108</td>
<td>0.001</td>
<td>0.180</td>
<td>0.000</td>
<td>0.108</td>
</tr>
<tr>
<td></td>
<td>MADR</td>
<td>-0.003</td>
<td>0.110</td>
<td>0.002</td>
<td>0.184</td>
<td>0.000</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>DiPS</td>
<td>0.005</td>
<td>0.109</td>
<td>0.008</td>
<td>0.117</td>
<td>0.005</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Figure 1.1 presents the RE under the different scenarios for \( n = 1,000, 5,000 \) and \( p = 15, 30, 75 \). RE was defined as the mean square error (MSE) of DR-ALAS relative to that of each estimator, with RE > 1 indicating greater efficiency compared to DR-ALS. Under the “both correct” scenario many of the estimators have similar efficiency since they are variants of the doubly-robust estimator and are locally semiparametric efficient. DiPS is also no less efficient than other estimators under the “misspecified \( \pi_1(x) \)” scenario. On the other hand, efficiency gains of more than 100% are achieved relative to other estimators when \( n \) is large relative to \( p \), under the “misspecified \( \mu_k(x) \)” scenario. This demonstrates that the efficiency gain resulting from Theorem 2 can be substantial. The gains diminish with larger \( p \) for a fixed \( n \) as expected but remain substantial even when \( p = 75 \). Similar results occur under the “both misspecified” scenario, except that the efficiency gains for DiPS are even larger due its robustness to misspecification in terms of the bias.
(a) Both correct

(b) Misspecified $\mu_k(x)$

(c) Misspecified $\pi_1(x)$

(d) Misspecified both

Figure 1.1: RE relative to DR-ALAS by $n, p$, and specification scenario.
Table 1.2 presents the performance of perturbation for DiPS when \( p = 15 \). SEs for DiPS were estimated using the MAD. The empirical SEs (Emp SE), calculated from the sample standard deviations of \( \hat{\Delta} \) over the simulation repetitions, were generally similar to the average of the SE estimates over the repetitions (ASE), despite some under slight underestimation. The coverage of the percentile CI’s (Cover) were generally close to nominal 95%, with slight over-coverage in small samples that diminished with larger \( n \).

Table 1.2: Perturbation performance under correctly specified models. Emp SE: empirical standard error over simulations, ASE: average of standard error estimates based on MAD over perturbations, Cover: Coverage of 95% percentile intervals.

<table>
<thead>
<tr>
<th>( p )</th>
<th>( n )</th>
<th>Emp SE</th>
<th>ASE</th>
<th>Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>500</td>
<td>0.350</td>
<td>0.326</td>
<td>0.971</td>
</tr>
<tr>
<td>15</td>
<td>1000</td>
<td>0.248</td>
<td>0.228</td>
<td>0.956</td>
</tr>
<tr>
<td>15</td>
<td>2000</td>
<td>0.170</td>
<td>0.161</td>
<td>0.954</td>
</tr>
<tr>
<td>30</td>
<td>500</td>
<td>0.351</td>
<td>0.330</td>
<td>0.973</td>
</tr>
<tr>
<td>30</td>
<td>1000</td>
<td>0.248</td>
<td>0.230</td>
<td>0.963</td>
</tr>
<tr>
<td>30</td>
<td>2000</td>
<td>0.174</td>
<td>0.161</td>
<td>0.950</td>
</tr>
</tbody>
</table>

### 1.8 Discussion

In this paper we developed a novel IPW-based approach to estimate the ATE that accommodates settings with high-dimensional covariates. Under sparsity assumptions and using appropriate regularization, the estimator achieves double-robustness and local semiparametric efficiency when adjusting for many covariates. By calibrating the initial PS through a smoothing, we showed that additional gains in robustness and efficiency are guaranteed in large samples under misspecification of working models. Simulation results demonstrate that DiPS exhibits comparative performance to existing approaches under correctly specified models but achieves potentially substantial gains in efficiency under model misspecification.

In numerical studies, we used the extended regularized information criterion (Hui et al., 2015) to tune adaptive LASSO, which is shown to maintain selection consistency for the diverging \( p \) case when \( \log(p)/\log(n) = \nu \), for \( \nu \in [0, 1) \). Other criteria such as cross-validation can also be used and may exhibit better performance in some cases. To obtain
a suitable bandwidth \( h \), the bandwidth must be selected such that the dominating errors in the influence function, which are of order \( O_p(n^{1/2}h^q + n^{-1/2}h^{-2}) \), converges to 0. This is satisfied for \( h = O(n^{-\alpha}) \) for \( \alpha \in \left( \frac{1}{2q}, \frac{1}{4} \right) \). The optimal bandwidth \( h^* \) is one that balances these bias and variance terms and is of order \( h^* = O(n^{-1/(q+2)}) \). In practice we use a plug-in estimator \( \hat{h}^* = \hat{\sigma}n^{-1/(q+2)} \), where \( \hat{\sigma} \) is the sample standard deviation of either \( \hat{\alpha}^T X_i \) or \( \hat{\beta}_k^T X_i \), possibly after applying a monotonic transformation. Similarly, cross-validation can also be used to obtain an optimal bandwidth for smoothing itself and re-scaled to be of the optimal order.

The second estimated direction \( \hat{\beta}_k^T X \) can be considered a working prognostic score (Hansen, 2008a). In the case where \( \beta_0 = \beta_1 \) is postulated in the working outcome model (1.7) but, in actuality \( \beta_0 \neq \beta_1 \), \( \hat{\beta}_k^T X \) could be a mixture of the true prognostic and propensity scores and may not be ancillary for the ATE. This could bias estimates of ATE when adjusting only for \( \hat{\beta}_k^T X \). DiPS avoids this source of bias by adjusting for both the working PS \( \hat{\alpha}^T X \) and \( \hat{\beta}_k^T X \) so that consistency is still maintained in this case, provided the working PS model (1.6) is correct (i.e. under \( M_\pi \cap M_{\mu}^c \)). Otherwise, if the PS model is also incorrect, no method is guaranteed to be consistent. If adaptive LASSO is used to estimate \( \alpha \) and \( \beta_k \) and the true \( \alpha \) and \( \beta_k \) are of order \( O(n^{-1/2}) \), \( \hat{\alpha} \) and \( \hat{\beta}_k \) are not \( n^{1/2} \)-consistent when the penalty is tuned to achieve consistent model selection (Pötscher and Schneider, 2009). This is a limitation of relying on procedures that satisfy oracle properties only under fixed-parameter asymptotics. DiPS is thus preferred in situations where \( n \) is large and signals are not extremely weak. Moreover, if adaptive LASSO is used, when \( \nu \) is large, a large power parameter \( \gamma \) would be required to maintain the oracle properties, leading to an unstable penalty and potentially poor performance in finite samples. It would be of interest to consider other approaches to estimate \( \alpha \) and \( \beta_k \) that have good performance in broader settings, such as settings allowing for larger \( p \) and more general sparsity assumptions. It would also be of interest to extend the approach to accommodate survival data.
1.9 Appendix

This Appendix identifies the requisite regularity conditions (Appendix A), establishes supporting lemmas (Appendix B and Appendix C), and derives two theorems in the text (Appendix D). The following notations will facilitate the derivations. Throughout this Appendix, we suppress the $k$ in $\beta_k, \bar{\beta}_k, \tilde{\beta}_k, \theta_k, \bar{\theta}_k$, and $\hat{\theta}_k$ for ease of notation but implicitly understand these quantities to be defined with respect to treatment $k = 0, 1$ in general. Let $\Sigma = (\alpha^T X, \beta^T X)^T$ be $X$ in the directions of $\alpha$ and $\beta$, regardless of the adequacy of the working models. Let the true density of $\pi$ at $s$ be $f(s)$, the propensity score given $\bar{S} = s$ for $k = 0, 1$ be $\pi_k(s) = P(T = k \mid \bar{S} = s)$, and $l_k(s) = \pi_k(s)f(s)$. Given a $x \in \mathbb{R}^p$, $\alpha, \beta \in \mathbb{R}^p$, for $\theta = (\alpha^T, \beta^T)^T$, let:

$$\tilde{\pi}_k(x; \theta) = \pi_k(x; \alpha, \beta) = \frac{\tilde{l}_k(x; \theta)}{\tilde{f}(x; \theta)} = \frac{\sum_{j=1} I_k \{ (\alpha, \beta)^T (X_j - x) \} I(T_i = k)}{\sum_{j=1} I_k \{ (\alpha, \beta)^T (X_j - x) \}}.$$ (1.16)

For a $p$ length random vector $V$, let $V^\dagger = (V, 0_p)$ be the $p \times 2$ matrix of the vector augmented by column of zeros on the right and $V^\ddagger = (0_p, V)$ similarly by a column of zeros on the left. For any two vectors $V_i$ and $V_j$, let $V_{ji} = V_j - V_i$. Let $K(u)$ be a bivariate symmetric kernel function of order $q > 2$, with a finite $q$-th moment. Let $K'(u) = \partial K(u)/\partial u$ and $K''(u) = h^{-3}K'(u/h)$. For any vector $V$ of length $p$ and $A \subseteq \{1, 2, \ldots, p\}$, with $|A| = p_0$, let $V_A$ denote a $p_0$-length vector that is $V$ restricted to coordinates indexed in $A$. Similarly, let $V_A^\dagger$ denote $V^\dagger$ restricted to coordinates indexed in $A$.

1.9.1 Appendix A: Regularity Conditions

(i) $K(u)$ is a bivariate kernel function of order $q > 2$, with a finite $q$-th moment. (ii) $K(u)$ is bounded and continuously differentiable with a compact support. (iii) $K'(u)$ is bounded, integrable, and Lipshitz continuous. (iv) $X$ is compact. (v) $f(s)$ is bounded and bounded away from 0 over its support. (vi) $f(s)$, $\pi_k(s)$, and $E(Y \mid \bar{S} = s, T = k)$ for $k = 0, 1$ are $q$-times continuously differentiable. (vii) $E(X \mid \bar{S} = s), E(X \mid \bar{S} = s, T = k)$, and $E(XY \mid \bar{S} = s, T = k)$ are continuously differentiable for $k = 0, 1$. (viii) There exists $0 < k_1 < k_2 < \infty$ such that the minimum and maximum eigenvalues of $\frac{1}{n} \sum_{i=1}^n X_iX_i^T$ around bounded below by $k_1$ and above by $k_2$. (ix) $\Theta_\alpha$ and $\Theta_\beta$ are compact. (x) For all
\[ u \in \mathbb{R}, 1/M \leq g'_\mu(u) \leq M \text{ and } |g''_\mu(u)| \leq M \text{ and for some } 0 < M < \infty. \]

1.9.2 Appendix B: Supporting Lemmas

Lemma 1 identifies the stochastic order of a standardized mean when the variance of the observations is of a known order. It will be useful for controlling certain terms that will emerge in the expansion. Lemma 2 shows the uniform convergence rate for kernel smoothing when \( \alpha \) and \( \beta \) are fixed, which is a fundamental result used in our approach. Lemma 3 simplifies the average of the gradients of the average of terms that are inversely weighted by the calibrated PS evaluated at the least false parameters. These terms appear repeatedly in subsequent derivations.

**Lemma 1.** Let \( \{X_{i,n}\} \) be a triangular array such that \( X_{1,n}, \ldots, X_{n,n} \) are iid for each \( n \in \mathbb{N} \). Suppose that \( \sigma^2_n = \text{Var}(X_{i,n}) = O(c^2_n) \), where \( c_n \) is some positive sequence. Then:

\[
n^{1/2} |\bar{X}_n - \mu_n| \leq O_p(c_n),
\]

where \( \bar{X}_n = n^{-1} \sum_{i=1}^n X_{i,n} \) and \( \mu_n = E(X_{i,n}) \).

**Proof.** By Chebyshev’s inequality, for any \( k > 0 \):

\[
\mathbb{P} \left( n^{1/2} |\bar{X}_n - \mu_n| / c_n \geq k \right) \leq \frac{\sigma^2_n}{c^2_n k^2}.
\]

Let \( M = \sup_{n \in \mathbb{N}} \frac{\sigma^2_n}{c^2_n} \). For any \( \epsilon > 0 \), the desired result is obtained by taking \( k = (M/\epsilon)^{1/2} \).

**Lemma 2.** The uniform convergence rate for two-dimensional smoothing over \( X \) in the directions \( \bar{\alpha} \) and \( \bar{\beta} \) is given by:

\[
\sup_x \| \hat{\pi}_k(x; \bar{\theta}) - \pi_k(x; \bar{\theta}) \| = O_p(a_n),
\]

where \( \bar{\theta} = (\bar{\alpha}^T, \bar{\beta}^T)^T, \pi_k(x; \bar{\theta}) = P(T = k \mid \bar{\alpha}^T X = \bar{\alpha}^T x, \bar{\beta}^T X = \bar{\beta}^T x), \) and:

\[
a_n = h^q + \{\log(n)/(nh^2)\}^{1/2}.
\]

**Proof.** Smoothing over \( X \) in the directions of \( \bar{\alpha} \) and \( \bar{\beta} \) is the same as a two-dimensional kernel smoothing since \( \bar{\alpha} \) and \( \bar{\beta} \) are fixed. See, for example, Hansen (2008b) for the derivation of uniform convergence rates for \( d \)-dimensional smoothing. \( \square \)
Lemma 3. Let \( g(Z) \) denote a real-valued square-integrable transformation of the data \( Z = (X, T, Y)^T \). Under the above regularity conditions and that \( E\{g(Z)|\bar{S} = s\} \) and \( E\{Xg(Z)|\bar{S} = s\} \) are continuous in \( s \):

\[
\begin{align*}
\frac{\partial}{\partial \alpha} \frac{g(Z_i)}{\pi_k(X_i; \theta)} &= \mathbb{E} \left[ \hat{K}_h(S_{ji})^T \left\{ \pi_k(S_i) - I(T_j = k) \right\} \frac{g(Z_i)}{\pi_k(S_i)l_k(S_i)} X_{ji} \right] + O_p(b_n) \\
\frac{\partial}{\partial \beta} \frac{g(Z_i)}{\pi_k(X_i; \theta)} &= \mathbb{E} \left[ \hat{K}_h(S_{ji})^T \left\{ \pi_k(S_i) - I(T_j = k) \right\} \frac{g(Z_i)}{\pi_k(S_i)l_k(S_i)} X_{ji} \right] + O_p(b_n),
\end{align*}
\]

where \( b_n = n^{-1/2}h^{-1} + n^{-1}h^{-3}, \) for \( k = 0, 1 \).

**Proof.** We will show the first equality for the gradient with respect to \( \alpha \), with the second equality being analogous. First note each of the gradients can be written:

\[
\frac{\partial}{\partial \alpha} \frac{1}{\pi_k(X_i; \theta)} = \frac{\partial}{\partial \alpha} \frac{\hat{f}(X_i; \theta)l_k(X_i; \theta)}{\hat{l}_k(X_i; \theta)^2} - \frac{\partial}{\partial \alpha} \frac{\hat{f}(X_i; \theta)}{\hat{l}_k(X_i; \theta)} \frac{\hat{l}_k(X_i; \theta)}{\hat{l}_k(X_i; \theta)^2}
\]

Consequently, the average of the gradients can be written:

\[
\begin{align*}
n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha} \frac{g(Z_i)}{\pi_k(X_i; \theta)} &= n^{-1} \sum_{i,j} \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(X_i; \theta) - I(T_j = k)\hat{f}(X_i; \theta)}{\hat{l}_k(X_i; \theta)^2} X_{ji} g(Z_i) \\
&= n^{-1} \sum_{i,j} \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(S_i) - I(T_j = k)\hat{f}(S_i)}{\hat{l}_k(S_i)^2} X_{ji} g(Z_i) + O_p(a_n e_n) \\
&= n^{-1} \sum_{i,j} \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(S_i)}{\hat{l}_k(S_i)^2} X_{ji} g(Z_i) + O_p(a_n e_n),
\end{align*}
\]

where we make repeated use of the uniform convergence of \( \hat{l}_k(X_i; \theta) \) and \( \hat{f}(X_i; \theta) \) to \( l_k(S_i) \) and \( f(S_i) \), \( e_n \) is a term of the same order as the main term so that \( O_p(a_n e_n) \) will be a negligible lower-order term, and use that \( l_k(s) \) is bounded over its support in the last equality. To facilitate application of the V-statistic projection lemma, define:

\[
\begin{align*}
m_{1,k}(Z_j) &= \mathbb{E}_{Z_i} \left\{ \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(S_i) - I(T_j = k)\hat{f}(S_i)}{\hat{l}_k(S_i)^2} X_{ji}^T g(Z_i) \right\} \\
m_{2,k}(Z_i) &= \mathbb{E}_{Z_i} \left\{ \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(S_i) - I(T_j = k)\hat{f}(S_i)}{\hat{l}_k(S_i)^2} X_{ji}^T g(Z_i) \right\} \\
m_k &= \mathbb{E} \left\{ \hat{K}_h(S_{ji})^T \frac{\hat{l}_k(S_i) - I(T_j = k)\hat{f}(S_i)}{\hat{l}_k(S_i)^2} X_{ji}^T g(Z_i) \right\}
\end{align*}
\]
\[ \varepsilon_{1,k} = n^{-1} \mathbb{E} \left\| \frac{\hat{K}_h(\bar{S}_i)^T l_k(\bar{S}_i)}{l_k(\bar{S}_i)^2} - I(T_i = k) f(\bar{S}_i) X_{ji}^T g(Z_i) \right\| = 0 \]

\[ \varepsilon_{2,k} = n^{-1} \left( \mathbb{E} \left\{ \left( \frac{\hat{K}_h(\bar{S}_i)^T l_k(\bar{S}_i)}{l_k(\bar{S}_i)^2} - I(T_j = k) f(\bar{S}_i) X_{ji}^T g(Z_i) \right)^2 \right\} \right)^{1/2} \]

We now further evaluate each term. The first term can be simplified through a change-of-variables:

\[ m_{1,k}(Z_j) = \mathbb{E}_{\bar{S}_i} \left[ \hat{K}_h(\bar{S}_{ji})^T \left\{ 1 - \frac{I(T_j = k)}{\pi_k(\bar{S}_i)} \right\} \frac{1}{l_k(\bar{S}_i)} \mathbb{E} \left\{ X_{ji}^T g(Z_i) \mid \bar{S}_i \right\} \right] \]

\[ = \int \hat{K}_h(\bar{S}_j - s_1)^T \left\{ 1 - \frac{I(T_j = k)}{\pi_k(s_1)} \right\} \frac{1}{\pi_k(s_1)} \mathbb{E} \left\{ X_{ji}^T g(Z_i) \mid \bar{S}_i = s_1 \right\} ds_1 \]

\[ = h^{-1} \int \hat{K}(\psi_j)^T \left\{ 1 - \frac{I(T_j = k)}{\pi_k(h\psi_j + \bar{S}_j)} \right\} \frac{1}{\pi_k(h\psi_j + \bar{S}_j)} \mathbb{E} \left\{ X_{ji}^T g(Z_i) \mid \bar{S}_i = h\psi_j + \bar{S}_j \right\} d\psi_j \]

\[ = O_p(h^{-1}), \]

where the last step follows from bounding the integrand. Similarly for the second term:

\[ m_{2,k}(Z_i) = \mathbb{E}_{\bar{S}_j} \left[ \hat{K}_h(\bar{S}_{ji})^T \mathbb{E} \left\{ \left( 1 - \frac{I(T_j = k)}{\pi_k(\bar{S}_i)} \right) X_{ji}^T \mid \bar{S}_j \right\} \frac{g(Z_i)}{l_k(\bar{S}_i)} \right] \]

\[ = \int \hat{K}_h(s_2 - \bar{S}_i)^T \mathbb{E} \left\{ \left( 1 - \frac{I(T_j = k)}{\pi_k(s_2)} \right) X_{ji}^T \mid \bar{S}_j = s_2 \right\} \frac{g(Z_i)}{l_k(\bar{S}_i)} ds_2 \]

\[ = h^{-1} \int \hat{K}(\psi_i)^T \mathbb{E} \left\{ \left( 1 - \frac{I(T_j = k)}{\pi_k(s_2)} \right) X_{ji}^T \mid \bar{S}_j = h\psi_i + \bar{S}_i \right\} \frac{g(Z_i)}{l_k(\bar{S}_i)} f(h\psi_i + \bar{S}_i) d\psi_i \]

\[ = O_p(h^{-1}), \]

where again the last step follows from bounding the integrand. Now, \( \varepsilon_{2,k} = O_p(n^{-1}h^{-3}) \) from bounding the terms in the expectation, except for \( g(Z_i) \). The projection lemma thus yields:

\[ n^{-2} \sum_{i,j} \hat{K}_h(\bar{S}_{ji}) \frac{l_k(\bar{S}_i) - I(T_j = k) f(\bar{S}_i) X_{ji}^T g(Z_i)}{l_k(\bar{S}_i)^2} \]

\[ = m_k + n^{-1} \sum_{j=1}^n m_{1,k}(Z_j) - m_k + n^{-1} \sum_{i=1}^n m_{2,k}(Z_i) - m_k + O_p(\varepsilon_1 + \varepsilon_2) \]

\[ = m_k + O_p(n^{-1/2}h^{-1}) + O_p(h^{-1}h^{-3}), \]

for \( k = 0, 1 \), where the last line follows from application of Lemma 1. Re-arrangement of terms and collecting the dominant errors yield the desired result.
1.9.3 Appendix C: Expansion of Normalization Constant

We will first show the normalization constant is 1 up to some lower order terms, which will allow us to account for the normalization in the expansion. The approach for the analysis parallels that of the main expansion. First note that:

$$n^{-1} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} = n^{-1} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} + n^{-1} \sum_{i=1}^{n} \left\{ \frac{1}{\pi_k(X_i; \theta)} - \frac{1}{\pi_k(X_i; \theta)} \right\} I(T_i = k)$$

$$= \hat{V}_{1,k} + \hat{V}_{2,k} + \hat{V}_{3,k},$$

where:

$$\hat{V}_{1,k} = n^{-1} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)}, \quad \hat{V}_{2,k} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{1}{\pi_k(X_i; \theta)} - \frac{1}{\pi_k(X_i; \theta)} \right\} I(T_i = k),$$

$$\hat{V}_{3,k} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{1}{\pi_k(X_i; \theta)} - \frac{1}{\pi_k(X_i; \theta)} \right\} I(T_i = k).$$

The second term is of order:

$$\left| \hat{V}_{2,k} \right| = n^{-1} \left| \sum_{i=1}^{n} \frac{\pi_k(X_i; \theta) - \pi_k(X_i; \theta)}{\pi_k(X_i; \theta)\pi_k(X_i; \theta)} I(T_i = k) \right|$$

$$\leq \sup_{X_i} \left| \pi_k(X_i; \theta) - \pi_k(X_i; \theta) \right| n^{-1} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)\pi_k(X_i; \theta)}$$

$$= O_p(a_n),$$

where the last step follows from uniform convergence of $$\pi_k(X_i; \beta)$$ to $$\pi_k(X_i; \theta)$$ and noting the remaining sum is $$O_p(1)$$ plus some lower-order term. The third term can be written:

$$\hat{V}_{3,k} = n^{-1} \sum_{i=1}^{n} \left\{ \frac{1}{\pi_k(X_i; \alpha, \beta)} - \frac{1}{\pi_k(X_i; \alpha, \beta)} + \frac{1}{\pi_k(X_i; \alpha, \beta)} - \frac{1}{\pi_k(X_i; \alpha, \beta)} \right\} I(T_i = k)$$

$$= n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \alpha^T} \frac{\partial}{\partial \beta} - \frac{\partial}{\partial \alpha^T} \frac{\partial}{\partial \beta} \right\} I(T_i = k) + O_p(n^{-1}E_{\alpha,n} + n^{-1}E_{\beta,n})$$

$$= n^{-1} \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \alpha^T} \frac{\partial}{\partial \beta} - \frac{\partial}{\partial \alpha^T} \frac{\partial}{\partial \beta} \right\} I(T_i = k)$$

$$+ O_p(n^{-1}E_{\alpha,n} + n^{-1}E_{\beta,n} + n^{-1}E_{\alpha\beta,n}),$$
where the last equality uses that that \( \dot{K}(u) \) is Lipshitz continuous and that \( \mathcal{E}_{\alpha,n}, \mathcal{E}_{\beta,n}, \) and \( \mathcal{E}_{\alpha\beta,n} \) are terms of the same order as \( n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha^T} \hat{\pi}_k(X_i; \alpha; \beta)^{-1} \) so that the error terms will be negligible lower-order terms. Applying Lemma 3, we can simplify:

\[
 n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \alpha^T} I(T_i = k) = \mathbb{E} \left[ \dot{K}_h(\bar{S}_{ji})^{T} \left\{ \pi_k(S_i) - I(T_j = k) \right\} \frac{I(T_i = k)}{\pi_k(S_i)l_k(S_i)} X_{ji}^{\text{\textup{T}}} \right] + O_p(b_n).
\]

Further simplifying the expectation we have:

\[
 \mathbb{E} \left[ \dot{K}_h(\bar{S}_{ji})^{T} \left\{ \pi_k(S_i) - I(T_j = k) \right\} \frac{I(T_i = k)}{\pi_k(S_i)l_k(S_i)} X_{ji}^{\text{\textup{T}}} \right] \\
= \mathbb{E} \left( \dot{K}_h(\bar{S}_{ji})^{T} \left[ \pi_k(S_i) \left\{ \mathbb{E}(X_j^{\text{\textup{T}}} | S_j) - \mathbb{E}(X_i^{\text{\textup{T}}} | \bar{S}_i, T_i = k) \right\} \right] \\
- \pi_k(\bar{S}_j) \left\{ \mathbb{E}(X_j^{\text{\textup{T}}} | \bar{S}_j, T_j = k) - \mathbb{E}(X_i^{\text{\textup{T}}} | \bar{S}_i, T_i = k) \right\} \right) \\
= h^{-1} \int \int \dot{K}(\psi_1) \frac{f(h\psi_1 + s_1)}{\pi_k(s_1)} \left[ \pi_k(s_1) \left\{ \mathbb{E}(X_j | S_j = h\psi_1 + s_1) - \mathbb{E}(X_i^{\text{\textup{T}}} | \bar{S}_i = s_1, T_i = k) \right\} \\
- \pi_k(h\psi_1 + s_1) \left\{ \mathbb{E}(X_j^{\text{\textup{T}}} | \bar{S}_j = h\psi_1 + s_1, T_j = k) - \mathbb{E}(X_i^{\text{\textup{T}}} | \bar{S}_i = s_1, T_i = k) \right\} \right] d\psi_1 ds_1 \\
= O(h^{-1}),
\]

where the last step follows from bounding terms in the integrand. Similarly:

\[
 n^{-1} \sum_{i=1}^{n} \frac{\partial}{\partial \beta^T} I(T_i = k) = O_p(h^{-1}) + O_p(b_n).
\]

Collecting all the results:

\[
 n^{-1} \sum_{i=1}^{n} I(T_i = k) = 1 + O_p(n^{-1/2}) + O_p(a_n) + O_p(n^{-1/2}h^{-1}) + O_p(n^{-1/2}b_n) \\
= 1 + O_p(a_n).
\]

### 1.9.4 Appendix D: Main Results

The approach for the expansion will be to decompose \( \hat{W}_k \) into terms representing the variability contributed from smoothing, with known \( \theta \), and from estimating \( \theta \). The term contributed from smoothing is written in terms of a V-statistic and analyzed using a V-statistic projection lemma (Newey and McFadden, 1994). The term contributed from estimating \( \theta \) is analyzed applying results from literature on adaptive LASSO in the diverging
parameters setting (Zou and Zhang, 2009; Hui et al., 2015). First note that:

\[
\tilde{W}_k = \left\{ n^{-1} \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \theta)} \right\}^{-1} \left\{ n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \theta)} (Y_i - \bar{\mu}_k) \right\}
\]

\[
= n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \theta)} (Y_i - \bar{\mu}_k) + \left\{ (1 + O_p(a_n)) \right\}^{-1} - 1 \right\} n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \theta)} (Y_i - \bar{\mu}_k)
\]

\[
= n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \theta)} (Y_i - \bar{\mu}_k) \{ 1 + O_p(a_n) \},
\]

where the second step follows from the result in Appendix C. Define:

\[
\tilde{W}_k = n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} (Y_i - \bar{\mu}_k) = \tilde{W}_{1,k} + \tilde{W}_{2,k} + \tilde{W}_{3,k},
\]

where:

\[
\tilde{W}_{1,k} = n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} (Y_i - \bar{\mu}_k)
\]

\[
\tilde{W}_{2,k} = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{1}{\hat{\pi}_k(X_i; \theta)} - \frac{1}{\pi_k(X_i; \theta)} \right\} I(T_i = k) (Y_i - \bar{\mu}_k)
\]

\[
\tilde{W}_{3,k} = n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{1}{\hat{\pi}_k(X_i; \theta)} - \frac{1}{\pi_k(X_i; \theta)} \right\} I(T_i = k) (Y_i - \bar{\mu}_k).
\]

We now proceed to further expand the second and third terms. For the second term:

\[
\tilde{W}_{2,k} = -n^{-1/2} \sum_{i=1}^{n} \frac{\hat{l}(X_i; \theta) - \hat{f}(X_i; \theta) \pi_k(S_i)}{l_k(X_i; \theta)} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} (Y_i - \bar{\mu}_k)
\]

\[
+ n^{-1/2} \sum_{i=1}^{n} \left\{ \frac{1}{l_k(X_i; \theta)} - \frac{1}{l_k(X_i; \theta)} \right\} \left\{ \hat{l}(X_i; \theta) - \hat{f}(X_i; \theta) \pi_k(S_i) \right\} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} (Y_i - \bar{\mu}_k)
\]

\[
= n^{-1/2} \sum_{i=1}^{n} \frac{\hat{l}(X_i; \theta) - \hat{f}(X_i; \theta) \pi_k(S_i)}{l_k(X_i; \theta)} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} (Y_i - \bar{\mu}_k) + O_p(n^{1/2}a_n^2),
\]

where the last equality follows from repeated use of uniform convergence of \( \hat{l}(X_i; \theta) \) and \( \hat{f}(X_i; \theta) \) to \( l(X_i; \theta) \) and \( f(X_i; \theta) \) and that \( n^{-1/2} \sum_{i=1}^{n} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} \frac{Y_i - \bar{\mu}_k}{l_k(X_i; \theta)} = O_p(n^{1/2}) \). Thus:

\[
\tilde{W}_{2,k} = -n^{1/2} n^{-2} \sum_{i,j} K_h(S_{ji}) \left\{ I(T_j = k) - \pi_k(S_i) \right\} \frac{I(T_i = k)}{\pi_k(X_i; \theta)} \frac{Y_i - \bar{\mu}_k}{l_k(X_i; \theta)} + O_p(n^{1/2}a_n^2)
\]

\[
= \tilde{W}_{ct,2,k} + \tilde{W}_{nc,2,k} + O_p(n^{1/2}a_n^2),
\]
with a centered and a non-centered V-statistic:

\[
\tilde{W}_{ct,2,k} = -n^{1/2}n^{-2} \sum_{i,j} K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)}
\]

\[
\tilde{W}_{nc,2,k} = -n^{1/2}n^{-2} \sum_{i,j} K_h(\tilde{S}_{ji}) \left\{ \pi_k(\tilde{S}_i) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)}
\]

To facilitate application of the projection lemma, let:

\[
m_{1,ct,2,k}(Z_j) = \mathbb{E}_{Z_j} \left[ K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)} \right]
\]

\[
m_{2,ct,2,k}(Z_j) = \mathbb{E}_{Z_j} \left[ K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)} \right]
\]

\[
m_{ct,2,k} = \mathbb{E} \left[ K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)} \right]
\]

\[
\varepsilon_{1,ct,2,k} = n^{-1} \mathbb{E} \left[ K_h(\tilde{S}_{ji}) \left\{ I(T_i = k) - \pi_k(\tilde{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)} \right]
\]

\[
\varepsilon_{2,ct,2,k} = n^{-1} \mathbb{E} \left[ \left( K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) \ l_k(S_i)} \right)^2 \right]^{1/2}
\]

We now evaluate each term. The first term can be simplified through change-of-variables:

\[
m_{1,ct,2,k}(Z_j) = \mathbb{E}_{S_j} \left[ K_h(\tilde{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \frac{\mathbb{E}(Y_i - \bar{\mu}_k \mid \tilde{S}_i = \tilde{S}, T_i = k)}{l_k(S_i)} \right]
\]

\[
= \int K_h(\tilde{S}_j - s_1) \xi_k(s_1) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \ ds_1
\]

\[
= \int K(\psi_j) \xi_k(h \psi_j + \tilde{S}_j) \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \ d\psi_j
\]

\[
= \int K(\psi_j) \left\{ \xi_k(\tilde{S}_j) + h \psi_j^T \frac{\partial}{\partial s} \xi_k(\tilde{S}_j) + \ldots + \frac{h^q}{q!} \psi_j^\otimes q \otimes \frac{\partial}{\partial s^\otimes q} \xi_k(\tilde{S}_j) \right\} \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \ d\psi_j
\]

\[
= \left\{ \xi_k(\tilde{S}_j) + \frac{h^q}{q!} \int K(\psi_j) \psi_j^\otimes q \otimes \frac{\partial}{\partial s^\otimes q} \xi_k(\tilde{S}_j) \ d\psi_j \right\} \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\}
\]

\[
= \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\} \mathbb{E}(Y_i - \bar{\mu}_k \mid \tilde{S}_i = \tilde{S}, T_i = k)
\]

\[
+ \frac{h^q}{q!} \int K(\psi_j) \psi_j^\otimes q \otimes \frac{\partial}{\partial s^\otimes q} \xi_k(\tilde{S}_j) \ d\psi_j \left\{ I(T_j = k) - \pi_k(\tilde{S}_j) \right\},
\]

where \( \| \tilde{S}_j - S_j \| \leq h \| \psi_j \| \) and:

\[
\xi_k(s) = \frac{\mathbb{E}(Y_i - \bar{\mu}_k \mid S_i = s, T_i = k)}{\pi_k(s)}.
\]
For the second term, due to the centering:

\[
m_{2,ct,2,k}(Z_i) = \mathbb{E}_{\xi_i} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right] = 0
\]

\[
m_{ct,2,k}(Z_i) = \mathbb{E}(m_{2,ct,2,k}(Z_i)) = 0.
\]

For the remaining terms:

\[
\varepsilon_{1,ct,2,k} = n^{-1} h^{-2} K(0) \mathbb{E} \left[ \{I(T_i = k) - \pi_k(S_i)\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right] = O(n^{-1} h^{-2})
\]

\[
\varepsilon_{2,ct,2,k} = n^{-1} \mathbb{E} \left( \left[ K_h(\bar{S}_{ji}) \left\{ I(T_j = k) - \pi_k(\bar{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]^2 \right)^{1/2} = O(n^{-1} h^{-2}),
\]

where the order of the second error can be obtained from bounding terms inside the expectation. Now, we apply the projection lemma to find that:

\[
\tilde{W}_{ct,2,k} = -n^{-1/2} \left[ n^{-1} \sum_{j=1}^n m_{1,ct,2,k}(Z_j) - m_{ct,2,k} + n^{-1} \sum_{j=1}^n m_{1,ct,2,k}(Z_j) - m_{ct,2,k} \right.
\]
\[
+ m_{ct,2,k} + O_p(\varepsilon_{1,ct,2,k} + \varepsilon_{2,ct,2,k})
\]
\[
= n^{-1/2} \sum_{j=1}^n \left\{ \frac{I(T_j = k)}{\pi_k(S_j)} - 1 \right\} \mathbb{E}(Y_j - \bar{\mu}_k | \bar{S}_j, T_j = k) + O_p(h^n) + O_p(n^{-1/2} h^{-2}).
\]

We used that \( \pi(s) \) and \( E(Y|S = s, T = k) \) are \( q \)-times continuously differentiable to bound the remainder error term from \( m_{1,ct,2,k}(Z_j) \).

We now repeat a similar analysis for \( \tilde{W}_{nc,2,k} \). Let:

\[
m_{1,nc,2,k}(Z_i) = \mathbb{E}_{\xi_i} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]
\]

\[
m_{2,nc,2,k}(Z_i) = \mathbb{E}_{\xi_i} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]
\]

\[
m_{nc,2,k} = \mathbb{E} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]
\]

\[
\varepsilon_{1,nc,2,k} = n^{-1} \mathbb{E} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]
\]

\[
\varepsilon_{2,nc,2,k} = n^{-1} \mathbb{E} \left( \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) \pi_k(\bar{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]^2 \right)^{1/2}
\]

The first term is:

\[
m_{1,nc,2,k}(Z_j) = \mathbb{E}_{\xi_i} \left[ K_h(\bar{S}_{ji}) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_i) \right\} \frac{E(Y_i - \bar{\mu}_k | \bar{S}_i, T_i = k)}{l_k(S_i)} \right]
\]
where $\bar{S}_j^*$ is such that $\|\bar{S}_j^* - \bar{S}_j\| \leq h \|\psi_j\|$ and the last equality can be obtained through bounding $\frac{\partial}{\partial s^{\otimes q}} \pi_k(\bar{S}_j^*)$ and $\xi_k(h\psi_j + \bar{S}_j)$. Similarly, for the second term:

$$m_{2, nc, 2, k}(Z_i) = \mathbb{E}_S \left[ K_h(\bar{S}_j) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right]$$

$$= \int K_h(s_2 - \bar{S}_i) \left\{ \pi_k(s_2) - \pi_k(\bar{S}_i) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} f(s_2) ds_2$$

$$= \int K(\psi_i) \left\{ \pi_k(h\psi_i + \bar{S}_i) - \pi_k(\bar{S}_i) \right\} f(h\psi_i + \bar{S}_i) d\psi_i \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)}$$

$$= \int K(\psi_i) \left\{ h\psi_i^{\otimes q} \frac{\partial}{\partial s} \pi_k(s_i) + \ldots + \frac{h q^q}{q!} \psi_i^{\otimes q} \otimes \frac{\partial}{\partial s^{\otimes q}} \pi_k(s_i) \right\} f(h\psi_i + \bar{S}_i) d\psi_i \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)}$$

$$= O_p(h^q),$$

where $\bar{S}_i^*$ is such that $\|\bar{S}_i^* - \bar{S}_i\| \leq h \|\psi_i\|$ and the last equality could be obtained through bounding $\frac{\partial}{\partial s^{\otimes q}} \pi_k(\bar{S}_i^*)$ and $f(h\psi_i + \bar{S}_i)$. The errors are:

$$\varepsilon_{1, nc, 2, k} = n^{-1} \mathbb{E} \left[ K_h(0) \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right] = 0$$

$$\varepsilon_{2, nc, 2, k} = n^{-1} \mathbb{E} \left[ \left( \left[ K_h(\bar{S}_j) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_j) \right\} \frac{I(T_i = k) Y_i - \bar{\mu}_k}{\pi_k(X_i; \theta) l_k(S_i)} \right\right] \right)^2 \right]^{1/2} = O(n^{-1}h^{-2}),$$

where the order of the second error can be obtained from bounding terms inside the expectation. Application of the projection lemma now yields:

$$\tilde{W}_{nc, 2, k} = -n^{-1/2} \left[ n^{-1} \sum_{i=1}^{n} \left\{ m_{1, nc, 2, k}(Z_j) - m_{nc, 2, k} \right\} + n^{-1} \sum_{i=1}^{n} \left\{ m_{2, nc, 2, k}(Z_j) - m_{nc, 2, k} \right\} \right]$$

$$+ m_{nc, 2, k} + O_p(\varepsilon_{1, nc, 2, k} + \varepsilon_{2, nc, 2, k})$$

$$= O_p(h^q) - n^{1/2} m_{nc, 2, k} + O_p(n^{-1/2}h^{-2}),$$

where we use that $\text{Var}\{m_{1, nc, 2, k}(Z_j)\} = O(h^{2q})$ and $\text{Var}\{m_{2, nc, 2, k}(Z_i)\} = O(h^{2q})$ and apply Lemma 1. We now evaluate $m_{nc, 2, k}$:

$$m_{nc, 2, k} = \mathbb{E} \left[ K_h(\bar{S}_j) \left\{ \pi_k(\bar{S}_j) - \pi_k(\bar{S}_j) \right\} \frac{E(Y_i | S_i, T_i = k) - \bar{\mu}_k}{l_k(S_i)} \right]$$
truly sparse, similar arguments can be used to show the result by viewing under model $P$. Consequently, on a set with probability tending to $Zou and Zhang (2009). Under misspecified working models, provided that models are correctly specified, this is shown by arguments from Hui et al. (2015) and using that

\[ \frac{\partial}{\partial s} \hat{\beta}_k(s_1) = h(y), \]

where $s^*$ is such that $||s^* - s_1|| \leq h \| \psi_1 \|$, and the last equality follows from bounding $\frac{\partial}{\partial s} \hat{\beta}_k(s_1)$ and $f(h \psi_1 + s_1)$. We have now have that:

\[ \hat{W}_{n,2,k} = O_p(n^{1/2}h^q) + O_p(n^{-1/2}h^{-2}). \]

We now proceed to expand $\hat{W}_{3,k}$. We then then first analyze the gradients in general, under model $\mathcal{M}_\pi$, and under model $\mathcal{M}_\pi \cap \mathcal{M}_\mu$, using Lemma 3. First note that:

\[ \hat{W}_{3,k} = n^{-1/2} \sum_{i=1}^n \left\{ \frac{\partial}{\partial \alpha^T} \hat{\pi}_k(X_i; \alpha, \beta)(\hat{\alpha} - \alpha) + \frac{\partial}{\partial \beta^T} \hat{\pi}_k(X_i; \alpha, \beta)(\hat{\beta} - \beta) \right\} I(T_i = k)(Y_i - \hat{\mu}_k) \]

\[ + O_p \left\{ n^{1/2} \left( \| \hat{\alpha} - \alpha \|^2 + \| \hat{\beta} - \beta \|^2 + \| \hat{\alpha} - \alpha \| \| \hat{\beta} - \beta \| \right) \right\}, \]

using that $\frac{\partial}{\partial \alpha^T} \hat{\pi}_k(X_i; \theta)^{-1}$ and $\frac{\partial}{\partial \beta^T} \hat{\pi}_k(X_i; \theta)^{-1}$ are Lipschitz continuous in $\theta$. Moreover, we have that $\mathbb{P}\{supp(\hat{\alpha}) = A_\alpha\} \to 1$ and $\mathbb{P}\{supp(\hat{\beta}) = A_\beta\} \to 1$. In the case when working models are correctly specified, this is shown by arguments from Hui et al. (2015) and Zou and Zhang (2009). Under misspecified working models, provided that $\hat{\alpha}$ and $\hat{\beta}$ are truly sparse, similar arguments can be used to show the result by viewing $(\hat{\alpha}_0, \hat{\alpha}_1)^T$ and $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2^T, \hat{\beta}_1^T)^T$ as maximizers of postulated penalized likelihood functions (Lu et al., 2012). Consequently, on a set with probability tending to 1:

\[ \hat{W}_{3,k} = n^{-1} \sum_{i=1}^n \left\{ \frac{\partial}{\partial \alpha^T} \hat{\pi}_k(X_i; \alpha, \beta) \right\}_{A_\alpha} n^{1/2}(\hat{\alpha} - \alpha)_{A_\alpha} \]

\[ + \left\{ \frac{\partial}{\partial \beta^T} \hat{\pi}_k(X_i; \alpha, \beta) \right\}_{A_\beta} n^{1/2}(\hat{\beta} - \beta)_{A_\beta} I(T_i = k)(Y_i - \hat{\mu}_k) \]

\[ + O_p \left\{ n^{1/2} \left( \| (\hat{\alpha} - \alpha)_{A_\alpha} \|^2 + \| (\hat{\beta} - \beta)_{A_\beta} \|^2 + \| (\hat{\alpha} - \alpha)_{A_\alpha} \| \| (\hat{\beta} - \beta)_{A_\beta} \| \right) \right\}. \]

Applying Lemma 3 to the gradient restricted to the respective active sets:

\[ n^{-1} \sum_{i=1}^n \left\{ \frac{\partial}{\partial \alpha^T} \hat{\pi}_k(X_i; \alpha, \beta) \right\}_{A_\alpha} I(T_i = k)(Y_i - \hat{\mu}_k) \]
Each of the four terms in \( u \) where:

\[
\sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \beta} I(T_i = k)(Y_i - \bar{\mu}_k) \right\}_{\alpha} = E \quad (Y_i - \bar{\mu}_k) X_j^T \bigg| S_i, T_i = k \bigg) \bigg)_{\alpha}
\]

We now verify that \( u \) and \( v \) are \( O_p(1) \) in general. First note that:

\[
u^T \sum_{i=1}^{n} \left\{ \frac{\partial}{\partial \alpha} I(T_i = k)(Y_i - \bar{\mu}_k) \right\}_{\beta} = E \quad (Y_i - \bar{\mu}_k) X_j^T \bigg| S_i, T_i = k \bigg) \bigg)_{\alpha}
\]

where:

\[
\eta_{1,1,k}(s) = E(Y_i - \mu_k \big| S_i = s, T_i = k) \quad \eta_{2,1,k}(s) = E(X_j^T \big| S_j = s) f(s)
\]

\[
\eta_{1,2,k}(s) = E \left\{ (Y_i - \mu_k) X_j^T \big| S_i = s, T_i = k \right\} \quad \eta_{2,2,k}(s) = f(s)
\]

\[
\eta_{1,3,k}(s) = \frac{E(Y_i - \mu_k \big| S_i = s, T_i = k)}{\pi_k(s)} \quad \eta_{2,3,k}(s) = l_k(s) E(X_j^T \big| S_j = s, T_j = k)
\]

\[
\eta_{1,4,k}(s) = \frac{E(Y_i - \mu_k \big| S_i = s, T_i = k)}{\pi_k(s)} \quad \eta_{2,4,k}(s) = l_k(s)
\]

Each of the four terms in \( u \) can be simplified through change-of-variables:

\[
\int \int \hat{K}_h(s_21)^T \eta_{1,u,k}(s_2) \eta_{2,u,k}(s_2) ds_1 ds_2 = \int \int \hat{K}_h(s_21)^T \eta_{1,u,k}(s_2) \eta_{2,u,k}(s_2) ds_1 ds_2
\]

For some vector \( u \), let \( \hat{K}(u) = \hat{K}(u)_{11}^T \), \( \hat{K}(u)_{21}^T \) be the partial derivatives of \( K(u) \) with respect to the first and second components of \( u \), evaluated at \( u \). Similarly, for some \( s_1 \) and
s₂, let \( \{ \eta(s_1)_1,u,k \eta(s_2)_1,u,k \}_{(i,j)} \) denote the \((i,j)\)-th element of \( \eta(s_1)_1,u,k \eta(s_2)_1,u,k \) evaluated at \( s_1 \) and \( s_2 \), for \( i = 1, 2 \) and \( j = 1, \ldots, p + 1 \). Applying integration by parts, the \( j \)-th element of the above expectation is:

\[
\sum_{i=1}^{2} \int \int h^{-1} K(\psi_2)_{ \{ i \} } \{ \eta_{1,u,k}(h\psi_2 + s_2) \eta_{2,u,k}(s_2) \}_{(i,j)} d\psi_2 ds_2
\]

\[
= \sum_{i=1}^{2} \int h^{-1} K(\psi_2) \{ \eta_{1,u,k}(h\psi_2 + s_2) \eta_{2,u,k}(s_2) \}_{(i,j)} \Bigg|_{\psi_2} ds_2
\]

\[
- \int \int K(\psi_2) \frac{\partial}{\partial \psi_{2i}} \{ \eta_{1,u,k}(h\psi_2 + s_2) \eta_{2,u,k}(s_2) \}_{(i,j)} d\psi_2 ds_2
\]

\[
= - \sum_{i=1}^{2} \int \int K(\psi_2) \frac{\partial}{\partial \psi_{2i}} \{ \eta_{1,u,k}(h\psi_2 + s_2) \eta_{2,u,k}(s_2) \}_{(i,j)} d\psi_2 ds_2
\]

\[
= O(1),
\]

where the second to last and last equalities can be shown by bounding terms using that \( \mathbb{E}(Y_i \mid \bar{S}_i = s, T_i = k), \pi_k(s), f(s), \mathbb{E}(Y_iX_i \mid \bar{S}_i = s, T_i = k), \mathbb{E}(X_i \mid \bar{S}_i = s, T_i = k) \) are differentiable in \( s \) for \( k = 0, 1 \), \( \mathbb{E}(X_i \mid \bar{S}_i = s) \) is continuous in \( s \), \( \mathcal{X} \) is compact, and \( K(u) \) is a kernel function. Consequently:

\[
u^T_{k,\alpha} = \sum_{u=1}^{4} \mathbb{E} \left\{ \hat{K}_h(\bar{S}_{ji}) \frac{\eta_{1,u,k}(\bar{S}_i) \eta_{2,u,k}(\bar{S}_j)}{f(S_i) \cdot f(S_j)} \right\} \mathcal{A}_\alpha = O(1).
\]

Applying the same argument it can be shown that \( v^T_{k,\beta} = O(1) \) for \( k = 0, 1 \) as well.

We now consider simplifying \( v^T_{k,\beta} \) under \( \mathcal{M}_\pi \). First we note that under \( \mathcal{M}_\pi \), \( T \perp X \cdot \alpha^T X \). This implies that \( T \perp X \mid \bar{S} \). Applying this and similar calculations used above for \( u^T_{k,\alpha} \):

\[
v^T_{k,\beta} = \mathbb{E} \left[ \int \int \hat{K}_h(s_21) \frac{\pi_k(s_1) - \pi_k(s_2)}{\pi_k(s_1)} f(s_2) \left\{ \mathbb{E}(Y_i \mid \bar{S}_i = s_1, T_i = k) \mathbb{E}(X^T_j \mid \bar{S}_j) - \mathbb{E}(Y_iX^T_j \mid \bar{S}_i, T_i = k) \right\} \right]_{\mathcal{A}_\beta}.
\]

We now evaluate:

\[
v^T_{k,\beta} = \left[ \int \int \hat{K}_h(s_21) \frac{\pi_k(s_1) - \pi_k(s_2)}{\pi_k(s_1)} f(s_2) \left\{ \mathbb{E}(Y_i \mid \bar{S}_i = s_1, T_i = k) \mathbb{E}(X^T_j \mid \bar{S}_j = s_2) \right. \right.
\]

\[
- \mathbb{E}(Y_iX^T_j \mid \bar{S}_i = s_1, T_i = k) \right\} ds_2 ds_1 \right]_{\mathcal{A}_\beta}
\]

32
= \left[ \int h^{-1} K(\psi_1)^T \frac{\pi_k(s)}{\pi_k(s_1)} \frac{\partial \pi_k(s)}{\partial s} f(h \psi_1 + s_1) \left\{ \mathbb{E}(Y_i | \bar{S}_i = s_1, T_i = k) \right\} \right]_{A_\beta} \\
E(X_j^{T^T} | \bar{S}_j = h \psi_1 + s_1) - \mathbb{E} \left( Y_i X_i^{T^T} | \bar{S}_i = s_1, T_i = k \right) \left\{ \mathbb{E}(Y_i | \bar{S}_i = s_1, T_i = k) \right\} \right] \right\} \right\} + O(h)

where \( s^* \) is such that \( \|s^* - s_1\| \leq h \|\psi_1\| \) and we use that \( f(s) \) and \( \mathbb{E}(X | \bar{S}) \) are continuously differentiable and that \( \pi_k(s) \) is twice continuously differentiable, \( \mathcal{X} \) is compact, \( K'(u) \) is bounded and integrable, to bound terms in the remainder. After some re-arrangement, this can be further simplified:

\[
\mathbf{v}_{k,A_\beta}^T = \left\{ - \int \frac{\partial \pi_k(s_1)}{\partial s} f(s_1) \left[ \mathbb{E}(Y_i | \bar{S}_i = s_1, T_i = k) \right] \left\{ \mathbb{E}(X_j^{T^T} | \bar{S}_j = s_1) - \mathbb{E} \left( Y_i X_i^{T^T} | \bar{S}_i = s_1, T_i = k \right) \right\} \right\} + O(h)

= \mathbb{E} \left\{ \frac{\partial \pi_k(s_i)}{\partial s} \mathbb{E}(X_i^{T^T} | \bar{S}_i, T_i = k) - \mathbb{E} \left( Y_i X_i^{T^T} | \bar{S}_i, T_i = k \right) \right\} + O(h)

= 0 + O(h),

(1.18)

where the second equality follows from that \( \int \psi_1 K(\psi_1)^T d\psi_1 = -I_{2 \times 2} \) by integration by parts. Let the partial derivatives of \( \pi_k(s) \) with respect to \( s \), evaluated at \( s \), be denoted by \( \partial \pi_k(s)/\partial s^T = (\partial \pi_k(s)/\partial s_1, \partial \pi_k(s)/\partial s_2) \). Under \( \mathcal{M}_\pi \) when the PS model is correct, \( \partial \pi_k(s)/\partial s_2 = 0 \) since \( \pi_k(s) \) would depend only on the first argument. The last equality follows from noting this and that the first row of \( X_i^{T^T} \) is 0T.

Finally, we consider the case under \( \mathcal{M}_T \cap \mathcal{M}_t \). In this case we have not only that \( T \perp X | S \) but also \( \mathbb{E}(Y | S, T = k, X) = g_\mu(\beta_0 + \beta_1 k + \beta^T X) = \mathbb{E}(Y | S, T = k) \). Thus in

33
this case:

\[
\mathbb{E}(Y_i X_i^T | \bar{S}_i, T_i = k) = \mathbb{E}(Y_i | \bar{S}_i, T_i = k) \mathbb{E}(X_i^T | \bar{S}_i).
\]

Consequently, continuing from an analogous expression for \(u_{k,A_\alpha}\) from (1.17):

\[
u_{k,A_\alpha}^T = \mathbb{E} \left[ \hat{K}_h(\bar{S}_i) \frac{\pi_k(\bar{S}_i)}{l_k(\bar{S}_i)} \mathbb{E}(Y_i - \bar{\mu}_k | \bar{S}_i, T_i = k) \left\{ \mathbb{E}(X_j^T | \bar{S}_j) - \mathbb{E}(X_i^T | \bar{S}_i) \right\} \right]_{A_\alpha}.
\]

Evaluating the expression, we obtain that:

\[
u_{k,A_\alpha}^T = \left\{ \int \int h^{-1} \hat{K}(\psi_1) \frac{\pi_k(s_1) - \pi_k(h\psi_1 + s_1)}{\pi_k(s_1)} \mathbb{E}(Y_i - \bar{\mu}_k | \bar{S}_i = s_1, T_i = k) \right\} f(h\psi_1 + s_1) d\psi_1 ds_1 \right\}_{A_\alpha}
\]

\[
= \left\{ -h \int \int h^{-1} \hat{K}(\psi_1) \frac{\partial}{\partial s} \frac{\pi_k(s_1)}{\pi_k(s_1)} \mathbb{E}(Y_i - \bar{\mu}_k | \bar{S}_i = s_1, T_i = k) \right\} f(h\psi_1 + s_1) d\psi_1 ds_1 \right\}_{A_\alpha}
\]

\[
= O(h),
\]

where \(s_1^*\) and \(s_1^{**}\) are values such that \(\|s_1^* - s_1\| \leq h \|\psi_1\|\) and \(\|s_1^{**} - s_1\| \leq h \|\psi_1\|\). The last equality can be shown by bounding terms inside the integral by using that \(\pi_k(s)\) is continuously differentiable and bounded away from 0, \(\mathbb{E}(Y - \bar{\mu}_k | \bar{S} = s, T = k)\) is continuous, \(\mathbb{E}(X | \bar{S} = s)\) is continuously differentiable, \(f(s)\) is continuous, and \(\mathcal{X}\) is compact. The same argument can be applied to show that \(v_{k,A_\beta}^T = O(h)\) for \(k = 0, 1\), under \(\mathcal{M}_\pi \cap \mathcal{M}_\mu\).

We now collect all the results in the main expansion. With probability tending to 1:

\[
\tilde{W}_k = n^{-1/2} \sum_{i=1}^n \frac{I(T_i = k)}{\pi_k(S_i)} (Y_i - \bar{\mu}_k) - \left\{ \frac{I(T_i = k)}{\pi_k(S_i)} - 1 \right\} \mathbb{E}(Y_i - \bar{\mu}_k | \bar{S}_i, T_i = k)
\]

34
\[ + \mathbf{u}_{k, \alpha}^T A_{\alpha} n^{1/2} (\hat{\alpha} - \bar{\alpha})_{A_{\alpha}} + \mathbf{v}_{k, \beta}^T A_{\beta} n^{1/2} (\hat{\beta} - \bar{\beta})_{A_{\beta}} \\
+ O_p(n^{1/2} h^2 + n^{-1/2} h^{-2}) + O_p(h^2 + n^{-1/2} h^{-2}) + O_p(n^{1/2} o_n^2) \]

\[ = n^{-1/2} \sum_{i=1}^{n} \left( I(T_i = k) Y_i \right) - \left( \frac{I(T_i = k)}{\pi_k(X_i; \theta)} - 1 \right) \mathbb{E}(Y_i \mid \hat{\alpha}^T X_i, \hat{\beta}^T X_i, T_i = k) - \mu_k \]

\[ + \mathbf{u}_{k, \alpha}^T A_{\alpha} n^{1/2} (\hat{\alpha} - \bar{\alpha})_{A_{\alpha}} + \mathbf{v}_{k, \beta}^T A_{\beta} n^{1/2} (\hat{\beta} - \bar{\beta})_{A_{\beta}} + O_p(n^{1/2} h^2 + n^{-1/2} h^{-2}) \]

where \( \mathbf{u}_{k, \alpha} \) and \( \mathbf{v}_{k, \beta} \) are deterministic vectors such that, for \( k = 0, 1 \), \( \mathbf{v}_{k, \beta} = 0 \) under \( \mathcal{M}_\pi \) and \( \mathbf{u}_{k, \alpha} = \mathbf{v}_{k, \beta} = 0 \) under \( \mathcal{M}_\pi \cap \mathcal{M}_\mu \).

We next verify Theorem 2 to characterize the influence function contribution from estimating the PS. Since \([U_{\alpha, A_{\alpha}}] \) is a finite dimensional subspace of \( \mathcal{L}_2 \) spanned by the components of \( U_{\alpha, A_{\alpha}} \), the projection of \( \varphi_{i,k} \) onto it is given by population least squares:

\[ \Pi \{ \varphi_{i,k} \mid [U_{\alpha, A_{\alpha}}] \} = \mathbb{E}(\varphi_{i,k} U_{\alpha, A_{\alpha}}^T) \mathbb{E} \left( U_{\alpha, A_{\alpha}} U_{\alpha, A_{\alpha}}^T \right)^{-1} U_{\alpha, A_{\alpha}}. \]

If \( \hat{\alpha} \) satisfies the oracle properties, then \( n^{1/2} (\hat{\alpha} - \bar{\alpha})_{A_{\alpha}} = \mathbb{E}(U_{\alpha, A_{\alpha}} U_{\alpha, A_{\alpha}}^T)^{-1} n^{-1/2} \sum_{i=1}^{n} U_{\alpha, i} A_{\alpha} + o_p(1) \), where \( U_{\alpha, i} = X_i \{ T_i - \pi_1(X_i; \bar{\alpha}_0, \bar{\alpha}) \} \) and \( \pi_1(X_i; \bar{\alpha}_0, \bar{\alpha}) = g_\pi(\bar{\alpha}_0 + \alpha^T X_i) \). It thus suffices to show that \( \mathbf{u}_{k, \alpha}^T = -\mathbb{E}(\varphi_{i,k} U_{\alpha, A_{\alpha}}^T) + o(1) \).

We proceed by simplifying the covariance term:

\[ \mathbb{E}(\varphi_{i,k} U_{\alpha, A_{\alpha}}^T) = \mathbb{E} \left( \left[ \frac{I(T_i = k) Y_i}{\pi_k(X_i; \theta)} - \left( \frac{I(T_i = k)}{\pi_k(X_i; \theta)} - 1 \right) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = k) - \mu_k \right] U_{\alpha, A_{\alpha}}^T \right) \]

\[ = \mathbb{E} \left( \left[ \frac{I(T_i = k) Y_i}{\pi_k(X_i; \theta)} - \left( \frac{I(T_i = k)}{\pi_k(X_i; \theta)} - 1 \right) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = k) \right] X_{i, A_{\alpha}}^T \{ T_i - \pi_1(X_i; \bar{\alpha}_0, \bar{\alpha}) \} \right). \]

First consider the \( k = 1 \) case. Using that \( \pi_1(X; \bar{\alpha}_0, \bar{\alpha}) = \pi_1(\bar{S}) \) and \( T \perp X \mid \bar{S} \) under \( \mathcal{M}_\pi \):

\[ \mathbb{E}(\varphi_{i,1} U_{\alpha, A_{\alpha}}^T) = \mathbb{E} \left( \left[ \frac{I(T_i = 1) Y_i}{\pi_1(X_i; \theta)} - \left( \frac{I(T_i = 1)}{\pi_1(X_i; \theta)} - 1 \right) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = 1) \right] X_{i, A_{\alpha}}^T \{ T_i \} \right) \]

\[ - \mathbb{E} \left( \left[ \frac{I(T_i = 1) Y_i}{\pi_1(X_i; \theta)} - \left( \frac{I(T_i = 1)}{\pi_1(X_i; \theta)} - 1 \right) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = 1) \right] X_{i, A_{\alpha}}^T \pi_1(X_i; \bar{\alpha}_0, \bar{\alpha}) \right) \]

\[ = \mathbb{E} \left( \mathbb{E}(Y_i X_{i, A_{\alpha}}^T \mid \bar{S}_i, T_i = 1) - \{ 1 - \pi_1(\bar{S}_i) \} \mathbb{E}(X_{i, A_{\alpha}}^T \mid \bar{S}_i, T_i = 1) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = 1) \right) \]

\[ - \mathbb{E} \left( \mathbb{E}(Y_i X_{i, A_{\alpha}}^T \mid \bar{S}_i, T_i = 1) \pi_1(\bar{S}_i) - \{ \pi_1(\bar{S}_i) - \pi_1(\bar{S}_i) \} \mathbb{E}(X_{i, A_{\alpha}}^T \mid \bar{S}_i, T_i = 1) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = 1) \right) \]

\[ = \{ 1 - \pi_1(\bar{S}_i) \} \{ \mathbb{E}(Y_i X_{i, A_{\alpha}}^T \mid \bar{S}_i, T_i = 1) - \mathbb{E}(X_{i, A_{\alpha}}^T \mid \bar{S}_i) \mathbb{E}(Y_i \mid \bar{S}_i, T_i = 1) \} \]

\[ = -\mathbf{u}_{1, \alpha}^T + O(h), \]
where $u_{k,\alpha}^T$ has same form as derived in (1.18) for $v_{k,\beta}$, except that $X_i^{\dagger T}$ is replaced by $X_i^{\dagger T}$.

In the $k = 0$ case, again using $\pi_1(X; \bar{\alpha}_0, \bar{\alpha}) = \pi_1(\bar{S})$ and $T \perp X \mid \bar{S}$ under $\mathcal{M}_z$:

$$E(\varphi_{i,0} U_{\alpha,\alpha}^T) = E\left(\left[\frac{I(T_i = 0) Y_i}{\pi_0(X_i; \theta)} - \left\{\frac{I(T_i = 0)}{\pi_0(X_i; \theta)} - 1\right\} E(Y_i \mid \bar{S}_i, T_i = 0)\right] X_{i,\alpha}^{T,\alpha} T_i\right)$$

$$- E\left(\left[\frac{I(T_i = 0) Y_i}{\pi_0(X_i; \theta)} - \left\{\frac{I(T_i = 0)}{\pi_0(X_i; \theta)} - 1\right\} E(Y_i \mid \bar{S}_i, T_i = 0)\right] X_{i,\alpha}^{T,\alpha} \pi_1(X_i; \bar{\alpha}_0, \bar{\alpha})\right)$$

$$= E\left\{E(Y_i \mid \bar{S}_i, T_i = 0)E(X_{i,\alpha}^{T,\alpha} \mid \bar{S}_i)\pi_1(\bar{S}_i)\right\} - E\left\{E(Y_i X_{i,\alpha}^{T,\alpha} \mid \bar{S}_i, T_i = 0)\pi_1(\bar{S}_i)\right\}$$

$$= E\left\{\pi_1(\bar{S}_i)E(Y_i \mid \bar{S}_i, T_i = 0)E(X_{i,\alpha}^{T,\alpha} \mid \bar{S}_i) - E(Y_i X_{i,\alpha}^{T,\alpha} \mid \bar{S}_i, T_i = 0)\right\}$$

$$= -u_{0,\alpha}^T + O(h).$$
Efficient and Robust Semi-Supervised Estimation of Average Treatment Effects in Electronic Medical Records Data

David Cheng
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Ashwin N. Ananthakrishnan
Division of Gastroenterology
Massachusetts General Hospital

Tianxi Cai
Department of Biostatistics
Harvard T.H. Chan School of Public Health
2.1 Introduction

Electronic medical records (EMR) aggregate rich patient-level data that are routinely collected during patient care. Since they include large samples in broad populations, EMRs have become a valuable data source for conducting comparative effectiveness research (CER) and identifying optimal treatment strategies among real-world patients (Fiks et al., 2012; Manion et al., 2012). However, many challenges arise when performing CER using EMR data (Hersh et al., 2013). Beyond the usual issue of confounding in observational data, a primary challenge is the lack of direct observation on a true clinical outcome of interest $Y$. In contrast with clinical trials or traditional observational studies, EMR data are not collected to evaluate any specific pre-specified outcome. This is frequently ignored, whereby researchers often explicitly or implicitly rely on “surrogate variables” $W$ abstracted from codified (e.g. billing codes) or narrative (e.g. physician notes) data to approximate $Y$ with some imputed outcome $Y^\dagger = g(W)$. But the accuracy of $Y^\dagger$ in approximating $Y$ is usually unclear, and reliance on naive imputations $Y^\dagger$ can lead to biased estimates of the average treatment effect (ATE) on the actual outcome of interest $Y$.

To address this issue, EMR studies often undertake a manual chart review process by domain experts to label records with a “gold-standard” $Y$. But labeling is a costly and time-consuming process, which is effectively unfeasible when scaling a study up to a large sample. An alternative approach used in practice is to label a relatively small number of records and use the labeled data $L$ to build an imputation model for $Y$ based on features extracted automatically from the EMR, including $W$ (Ananthakrishnan et al., 2016). This approach is attractive because it uses $Y$ to build the imputation model and avoids comprehensive labeling of all records. But it remains unclear when a resulting estimator of the ATE will be valid or efficient, especially under possible misspecification of the imputation model.

To the best of our knowledge, there have been no methods developed for this problem in the context of EMR data. This problem can be cast in terms of surrogate outcomes data, where $W$ can be regarded as surrogates for $Y$ and $L$ can be regarded as a validation sample. A variety of methods have been developed in this data setting for estimating re-
gression parameters (Pepe, 1992; Reilly and Pepe, 1995; Chen, 2000; Chen and Chen, 2000) and solutions to estimating equations (Chen et al., 2003, 2008). These methods tend to assume a univariate surrogate with low dimensional baseline covariates $X$ and have not been adapted to estimate causal effects in observational data. More generally, this problem can be regarded as a missing data problem in which $Y$ is missing in the large set of unlabeled data $\mathcal{U}$, for which semiparametric efficiency theory can potentially be applied (Robins et al., 1994, 1995; Robins and Rotnitzky, 1995). This approach has been used to develop robust and efficient estimators of the ATE in clinical trial settings with availability of surrogates (Davidian et al., 2005) and observational data with missingness in variables besides $Y$ (Williamson et al., 2012; Zhang et al., 2016a). But these methods assume the proportion of missingness is bounded away from 0, whereas in the EMR setting, $\mathcal{L}$ is always so much smaller than $\mathcal{U}$ that the proportion of missingness should be regarded as tending to 0 asymptotically. This in turn changes the semiparametric efficiency considerations. Moreover, these methods involve parametric modeling of propensity score (PS) and outcome models, and typical implementations based on logistic and linear regression models have poor performance if they are misspecified or if the number of covariates is not small. Other methods for handling missing data in non-EMR settings based on imputation models (Little and An, 2004; An and Little, 2008) are related to our proposed approach but do not address estimation of causal treatment effects.

In this paper, we propose a semi-supervised (SS) estimator for the ATE (SS$_{DR}$) based on an imputation followed by inverse probability weighting (IPW) that is doubly-robust and semiparametric efficient. The imputations are constructed such that the resulting estimator is robust to misspecification of the imputation model, enabling $\mathcal{U}$ to be safely used in improving estimation. We further employ the double-index propensity score from Chapter 1 for additional robustness and efficiency gain. The remainder of the paper is organized as follows. We formalize the SS estimation problem in Section 2.2 and develop the estimator from Section 2.3 to Section 2.5. A perturbation resampling procedure is proposed in Section 2.6 for inference. Section 2.7.1 presents simulations showing the robustness and efficiency of the proposed estimator, and Section 2.7.2 applies the method to compare two biologic therapies for treating inflammatory bowel disease (IBD) in EMR.
data from Partner’s Healthcare. We conclude with some remarks in Section 5. Proofs are deferred to the Appendix.

2.2 Notations and Semi-Supervised Framework

Let $Y$ denote an outcome that could be modeled by a generalized linear model (GLM), such as a binary, ordinal, or continuous response, $T \in \{0, 1\}$ a binary treatment, $X$ a $p_x$-dimensional vector of pre-treatment baseline covariates, $W$ a $p_w$-dimensional vector of post-treatment surrogate variables that are potentially predictive of $Y$, and $V = (W^T, X^T)^T$. The labeled data consists of $n$ independent and identically distributed (iid) observations $\mathcal{L} = \{(Y_i, T_i, V_i^T) : i = 1, \ldots, n\}$, while the unlabeled data consists of $N$ iid observations without $Y$, $\mathcal{U} = \{(T_i, V_i^T) : i = n + 1, \ldots, N\}$, with $\mathcal{U} \perp \perp \mathcal{L}$. In the SS setting $N \gg n$ so that $\nu_n = n/N \to 0$ as $n \to \infty$. We assume that the labeled observations were randomly selected so that $Y$ is missing completely at random (MCAR) from observations in $\mathcal{U}$.

2.2.1 Target Parameter and Leveraging Unlabeled Data

Let $Y^{(1)}$ and $Y^{(0)}$ denote the counterfactual outcomes had an individual received treatment or control. Based on the observed data $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ we want to estimate the ATE:

$$\Delta = \mathbb{E}\{Y^{(1)}\} - \mathbb{E}\{Y^{(0)}\} = \mu_1 - \mu_0. \quad (2.1)$$

We require the following standard assumptions to identify $\Delta$:

$$Y = TY^{(1)} + (1 - T)Y^{(0)} \quad (2.2)$$

$$(Y^{(1)}, Y^{(0)}) \perp \perp T \mid X \quad (2.3)$$

$$\pi(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for some } \epsilon_\pi > 0 \text{ when } f(x) > 0, \quad (2.4)$$

where $\pi(x) = \mathbb{P}(T = 1 \mid X = x)$ is the PS and $f(x)$ is the joint density for the covariates. In the typical setting where the outcome is fully observed, the ATE can be identified through the g-formula (Robins, 1986):

$$\Delta = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} = \mathbb{E}\left\{\frac{I(T = 1)Y}{\pi(X)} - \frac{I(T = 0)Y}{1 - \pi(X)}\right\}, \quad (2.5)$$
where \( \mu_k(x) = \mathbb{E}(Y \mid X = x, T = k) \) for \( k = 0, 1 \). This suggests the usual estimators based on averaging the outcome weighted by IPW weights or averaging estimated outcome models. When the outcome is scarce but surrogates \( W \) are available, a further decomposition can potentially be helpful:

\[
\Delta = \mathbb{E} \left\{ \mathbb{E} \left\{ \xi_1(V) \mid X, T = 1 \right\} - \mathbb{E} \left\{ \xi_0(V) \mid X, T = 0 \right\} \right\}
\]

where \( \xi_k(v) = \mathbb{E}(Y \mid V = v, T = k) \) for \( k = 0, 1 \). This form of the g-formula suggests that, if a consistent estimator for \( \xi_k(v) \) is available, then \( \Delta \) can be estimated by first imputing \( Y \) through the \( \xi_k(v) \) estimator and then applying standard IPW or outcome regression methods to the imputed outcome. However, obtaining a consistent estimator for \( \xi_k(v) \) may not be feasible without strong modeling assumptions due to the potential high dimensionality of \( v \) and complexity of the functional form of \( \xi_k(v) \). In the following we show that even with incorrectly specified models for \( \xi_k(v) \), it is still possible to leverage \( \mathcal{U} \) in estimating \( \Delta \) without introducing bias from their misspecification.

### 2.3 Robust Imputations

Let \( U_\pi = I(T = 1)/\pi(X) - I(T = 0)/(1 - \pi(X)) \) denote a utility covariate given \( \pi(x) \), assumed momentarily to be known. Suppose we postulate a parametric working model, possibly misspecified, for \( \xi_k(v) \):

\[
\xi_T(V) = g_\xi(\gamma_0 + \gamma_1^T h(V) + \gamma_2 T + \gamma_3 U_\pi) = g_\xi(\gamma^T Z_\pi),
\]

where \( \gamma = (\gamma_0, \gamma_1^T, \gamma_2, \gamma_3)^T \), \( Z_\pi = (1, V^T, T, U_\pi) \), \( g_\xi(\cdot) \) is a specified link function, and \( h(\cdot) \) is a vector of fixed basis expansion functions that can incorporate nonlinear effects. We estimate \( \gamma \) as \( \hat{\gamma} \), the solution to a penalized estimating equation with ridge regularization:

\[
n^{-1} \sum_{i=1}^n Z_{\pi,i} \left\{ Y_i - g_\xi(\gamma^T Z_{\pi,i}) \right\} + \lambda_n \gamma_{(-1)} = 0,
\]

where \( \gamma_{(-1)} \) denotes the vector \( \gamma \) excluding the first element \( \gamma_0 \) and \( \lambda_n = o(n^{-1/2}) \) is a tuning parameter chosen such that \( \hat{\gamma} \) has \( n^{-1/2} \) convergence rate. In particular, this
class of estimators includes ridge estimators for GLMs based on exponential families with canonical link functions. Other regularization penalties besides the ridge penalty can also be used, as long as \( \hat{\gamma} \) maintains a \( n^{-1/2} \) convergence rate. Using the fact that \( Y \) are MCAR, standard arguments can be used to show that \( \hat{\gamma} \) follows:

\[
E \left[ Z_\pi \left\{ Y - g_\xi (\gamma^T Z_\pi) \right\} \right] = 0,
\]

with the expectation being taken over the entire population and not restricted only to the labeled population. Specifically, for \( Y^\dagger = g_\xi (\hat{\gamma}^T Z_\pi) \), since \( Z_\pi \) includes \( U_\pi \) this implies that:

\[
E \left\{ \frac{I(T = 1)Y}{\pi(X)} - \frac{I(T = 0)Y}{1 - \pi(X)} \right\} = E \left\{ \frac{I(T = 1)Y^\dagger}{\pi(X)} - \frac{I(T = 0)Y^\dagger}{1 - \pi(X)} \right\}.
\]

This suggests that a standard IPW estimator based on the imputed outcome \( Y^\dagger \) has the same limit asymptotically as if the true outcomes were used, even if imputation model (2.6) is misspecified. Consequently the surrogate data from \( \mathcal{U} \) could be safely used to impute the outcome using a consistent estimator of \( Y^\dagger \).

In practice \( \pi(x) \) also needs to be estimated, which is typically done through parametric modeling such as logistic regression. When \( \pi(x) \) is estimated by an estimator \( \hat{\pi}(x) \), the IPW estimator discussed above will be consistent for \( \Delta \) if \( \hat{\pi}(x) \) is consistent for \( \pi(x) \) but otherwise could be biased if the parametric model for \( \pi(x) \) is misspecified. Alternatively, similar arguments can be used to construct imputations \( Y^\dagger \) that could be substituted for \( Y \) in an outcome regression estimator and still be robust to misspecification of the imputation model. However, such an approach would then require correct specification of an outcome regression model given baseline covariates for \( \mu_k(x) \) to be consistent for \( \Delta \).

In the following we propose an IPW approach but weighting with the double-index PS (DiPS) of Chapter 1. The resulting IPW estimator will be doubly-robust in that it will be consistent for \( \Delta \) when a model for either \( \pi(x) \) or \( \mu_k(x) \) is correctly specified. Whereas the IPW estimator in Chapter 1 considered only the scenario where \( Y \) is fully observed, the present paper uses the double-index PS to develop a novel ATE estimator in the SS setting. Using the double-index PS is not essential in that an augmented IPW estimator (Robins et al., 1994, 1995) can potentially be used to achieve double-robustness as well.
We use the double-index PS for the clarity of the construction, which directly follows the line of reasoning of the robust imputations.

## 2.4 Doubly-Robust IPW Based on the Double-Index PS

Suppose we postulate the following working parametric models for $\pi(x)$ and $\mu_k(x)$:

$$
\pi(X) = g_\pi(\alpha_0 + \alpha_1^T X) = \pi(X; \alpha) \quad (2.9)
$$

$$
\mu_T(X) = g_\mu(\beta_0 + \beta_1^T X + \beta_2 T) = \mu_T(X; \beta), \quad (2.10)
$$

where $\alpha = (\alpha_0, \alpha_1^T)^T$, $\beta = (\beta_0, \beta_1^T, \beta_2)^T$, and $g_\pi(\cdot)$ and $g_\mu(\cdot)$ are specified link functions. We estimate $\alpha$ and $\beta$ using regularized estimators:

$$
\hat{\alpha} = \arg\min_{\alpha} \left\{ -N^{-1} \sum_{i=1}^N \ell_\pi(\alpha; X_i, T_i) + p_{\lambda_N}(\alpha_1) \right\}, \quad (2.11)
$$

$$
\hat{\beta} = \arg\min_{\beta} \left\{ -n^{-1} \sum_{i=1}^n \ell_\mu(\beta; Y_i, X_i, T_i) + p_{\lambda_n}(\beta_{-1}) \right\}, \quad (2.12)
$$

where $\ell_\pi(\alpha; X_i, T_i)$ and $\ell_\mu(\beta; Y_i, X_i, T_i)$ are the log-likelihood contributions for the $i$-th observation, and $p_{\lambda_N}(\cdot)$ and $p_{\lambda_n}(\cdot)$ are penalty functions chosen such that the oracle properties (Fan and Li, 2001) hold. Examples of such estimators include the adaptive least absolute shrinkage and selection operator (ALASSO) (Zou, 2006) where $p_{\lambda}(u) = \lambda \sum_{j=1}^p |u_j| / |\tilde{w}_j|$ with initial weights $\tilde{w}_j$ estimated from ridge regression and tuning parameters are such that $N\lambda_N \to \infty$, $\sqrt{N}\lambda_N \to 0$, $n\lambda_n \to \infty$, and $\sqrt{n}\lambda_n \to 0$. We then calibrate the initial PS estimate $\pi(x; \hat{\alpha})$ by the following kernel smoothing estimator:

$$
\tilde{\pi}(x; \hat{\alpha}_1, \hat{\beta}_1) = \frac{N^{-1} \sum_{j=1}^N K_h(\hat{S}_j - \hat{s})I(T_j = 1)}{N^{-1} \sum_{j=1}^N K_h(\hat{S}_j - \hat{s})},
$$

where $\hat{S}_j = (\hat{\alpha}_1, \hat{\beta}_1)^T X_j$ and $\hat{s} = (\hat{\alpha}_1, \hat{\beta}_1)^T x$ are bivariate scores that represent the covariate in the directions of $\hat{\alpha}_1$ and $\hat{\beta}_1$, $K_h(\cdot) = h^{-2} K(\cdot/h)$, with $K(\cdot)$ being a bivariate $q$-th order kernel with $q > 2$ and $h = O(N^{-\alpha})$ being a bandwidth for which a suitable choice of $\alpha$ is discussed below. Finally, we define the proposed $\text{SS}_{\text{DR}}$ estimator as $\hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_0$ where:

$$
\hat{\mu}_1 = \left\{ \sum_{i=1}^N I(T_i = 1) \tilde{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1) \right\}^{-1} \left\{ \sum_{i=1}^N I(T_i = 1) \hat{Y}_i^\dagger \tilde{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1) \right\}, \quad (2.13)
$$

43
and $\hat{\mu}_0 = \left\{ \sum_{i=1}^{N} \frac{I(T_i = 0)}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}^{-1} \left\{ \sum_{i=1}^{N} \frac{I(T_i = 0)\hat{Y}_i^\dagger}{1 - \hat{\pi}(X_i; \hat{\alpha}_1, \hat{\beta}_1)} \right\}$, where $\hat{Y}_i^\dagger = g_\xi(\hat{\gamma}^TZ_{\pi, i})$. Here $\hat{\Delta}$ substitutes the robust imputations based on the PS estimated by the double-index PS $\hat{Y}_i^\dagger$ into an IPW estimator weighted also with the double-index PS. Had $Y$ been fully observed such that $\hat{Y}_i^\dagger = Y_i$ for $i = 1, \ldots, N$, Chapter 1 showed that an IPW estimator based on the double-index PS is doubly-robust in that it is consistent when a working model for either $\pi(x)$ in (4.10) or for $\mu_k(x)$ in (2.10) is correctly specified. We will show that in the SS setting, using the double-index PS in this robust imputation approach maintains this double-robustness property.

### 2.5 Asymptotic Robustness and Efficiency Properties of $\hat{\Delta}$

We show in Appendix B that, under the causal identification assumptions (2.2)-(2.4) and mild regularity conditions, given that $h = O(N^{-\alpha})$ with $\alpha \in \left(\frac{1 - \beta}{2q}, \frac{\beta}{2} \wedge \frac{1}{q}\right)$ and $n = O(N^{1 - \beta})$ with $\beta \in \left(\frac{1}{q + 1}, 1\right)$, $\hat{\Delta}$ is doubly-robust so that:

$$\hat{\Delta} - \Delta = O_p(n^{-1/2})$$

(2.15)

when either the PS model $\pi(x; \alpha)$ in (4.10) or the baseline outcome model $\mu_k(x; \beta)$ in (2.10) is correctly specified. To characterize the large sample variability of $\hat{\Delta}$, we next show it is asymptotically linear and identify its influence function. First define $\tilde{\Delta} = \bar{\mu}_1 - \bar{\mu}_0$, where:

$$\bar{\mu}_1 = \mathbb{E} \left\{ \frac{I(T = 1)Y^\dagger}{\pi(X; \alpha_1, \beta_1)} \right\} \quad \text{and} \quad \bar{\mu}_0 = \mathbb{E} \left\{ \frac{I(T = 0)Y^\dagger}{1 - \pi(X; \alpha_1, \beta_1)} \right\},$$

with $\pi(x; \alpha_1, \beta_1) = \mathbb{P}(T = 1 \mid \alpha_1^T X = \alpha_1^T x, \beta_1^T X = \beta_1^T x)$, $\alpha_1$ and $\beta_1$ as the probability limits of $\hat{\alpha}_1$ and $\hat{\beta}_1$ regardless of model adequacy, and $Y^\dagger$ being defined as in (2.8) except that $\pi(x)$ is replaced by $\pi(x; \alpha_1, \beta_1)$. We show in Appendix B that, under the same requirements for $\alpha$ and $\beta$, the influence function for $\tilde{\Delta}$ is given by $n^{1/2}(\tilde{\Delta}_k - \bar{\Delta}_k) = \hat{W}_1 - \hat{W}_0$, where $\hat{W}_k = n^{1/2}(\hat{\mu}_k - \bar{\mu}_k)$ for $k = 0, 1$ and:

$$\hat{W}_k = n^{-1/2} \sum_{i=1}^{n} (\nu_{\beta_1,k} + u_{\pi,\alpha,k}) \varphi_{\beta_1,i} + u_{\gamma,k} \varphi_{\gamma,i} + o_p(1),$$

(2.16)

with $v_{\beta_1,k} = 0$ when the PS model $\pi(x; \alpha)$ is correctly specified and $u_{\pi,\alpha,k} = 0$ when either the PS model $\pi(x; \alpha)$ or imputation model $g_\xi(\gamma^T z_\pi)$ without the utility covariate
is correctly specified. Here \( \varphi_{\beta_1,i} \) and \( \varphi_{\gamma,i} \) are influence functions for \( \hat{\beta}_1 \) and \( \hat{\gamma} \) such that

\[
\begin{align*}
    n^{1/2}(\hat{\beta}_1 - \beta_1) &= n^{-1/2} \sum_{i=1}^{n} \varphi_{\beta_1,i} + o_p(1) \\
    n^{1/2}(\hat{\gamma} - \gamma) &= n^{-1/2} \sum_{i=1}^{n} \varphi_{\gamma,i} + o_p(1).
\end{align*}
\]

Accordingly, the first term in (2.16) represents the contribution from estimating \( \beta_1 \) in the baseline outcome model \( \mu_k(x; \beta) \) for the double-index PS appearing in the IPW weight and the utility covariate. The remaining term represents the contribution from estimating \( \gamma \) in the imputation model \( g_\xi(\gamma^T z_\pi) \). The influence function does not include terms associated with the variability in estimating \( \alpha \) in the parametric PS or for smoothing in the double-index PS, as such contributions to the expansion are of higher order when \( N \gg n \) in the SS setting.

In terms of efficiency, when the PS model \( \pi(x; \alpha) \) is correctly specified, the influence function in (2.16) simplifies so that:

\[
    n^{1/2}(\hat{\Delta} - \Delta) = n^{-1/2} \sum_{i=1}^{n} (u_{\gamma,1} - u_{\gamma,0})^T \varphi_{\gamma,i} + o_p(1),
\]

where:

\[
    \varphi_{\gamma,i} = \left[ \mathbb{E} \left\{ Z_{\pi,i} Z_{\pi,i} \hat{g}_\xi(\gamma^T Z_{\pi,i}) \right\} \right]^{-1} Z_{\pi,i} \left\{ Y_i - g_\xi(\gamma^T Z_{\pi,i}) \right\},
\]

for \( \hat{g}_\xi(u) = \frac{\partial u}{\partial u} g_\xi(u) \bigg|_{u=\hat{u}} \). The centering of \( Y_i \) around a model approximation of \( \xi_T(V) \) suggests that \( \hat{\Delta} \) should achieve efficiency gain over complete-case (CC) estimators, which neglect the surrogates \( W \). Let \( \hat{\Delta}^* = \mathbb{E}\{\mu_1(X) - \mu_0(X)\} \) be strictly a functional of the observed data distribution not depending on identification assumptions (2.2)-(2.4) as in \( \Delta \).

To characterize the efficiency of \( \hat{\Delta} \) in a more full context, we show in Appendix C that the semiparametric efficiency bound for \( \hat{\Delta}^* \) under an ideal SS model where the distribution of \( (V^T, T)^T \) is known but the conditional distribution of \( Y \) given \( (V^T, T)^T \) is unrestricted, with respect to a class of regular parametric submodels subject to mild regularity conditions, is \( \mathbb{E}(\varphi_{eff}^2) \), where:

\[
    \varphi_{eff} = U_{\pi}\{Y - \xi_T(V)\}
\]

is the efficient influence function. This efficiency bound is lower than or equal to the efficiency bound in the fully nonparametric model where the distribution of \( (V^T, T)^T \) is unknown. Furthermore, we show in Appendix C that \( \hat{\Delta} \) indeed achieves the SS efficiency.
bound when both the PS and imputation model, \( \pi(x; \alpha) \) and \( g_\xi(\gamma^T z_\pi) \), are correctly specified so that \( \hat{\Delta} \) is locally semiparametric efficient. In this case, even though the distribution of \((V^T, T)^T\) is actually not known in our setup, the bound under the ideal SS model can still be achieved because \( N \gg n \). The correct specification of \( \mu_k(x; \beta) \) is not required for attaining the efficiency bound, as the bound does not involve \( \mu_k(x) \), but its specification is still important for double-robustness in case \( \pi(x; \alpha) \) is misspecified. The local efficiency of \( \hat{\Delta} \) prompts favorable efficiency compared to CC and other SS estimators that traditionally have sought to be efficient under non-SS nonparametric models where the distribution of \((V^T, T)^T\) is assumed to be unknown. In our SS setting, \( \hat{\Delta} \) gains efficiency over these approaches by taking advantage of the additional information from the large set of unlabeled data \( \mathcal{U} \). Even when the working models are not exactly correct as in practice, we find that \( \hat{\Delta} \) still achieves substantial efficiency gains if the models are adequate approximations. In particular, \( \mu_k(x; \beta) \) and \( g_\xi(\gamma^T z_\pi) \) may not be compatible with one another if non-linear models such as logistic regression are used for either working model. Nevertheless, we find in such cases that \( \hat{\Delta} \) still attains large efficiency gains over existing estimators when flexible basis functions are used in \( g_\xi(\gamma^T z_\pi) \) to more closely approximate \( \xi_k(v) \). We offer some more discussion on the efficiency gains of \( \hat{\Delta} \) under model misspecification in the Section 2.8. We next consider inference about \( \Delta \) based on \( \hat{\Delta} \) through a perturbation resampling procedure.

### 2.6 Perturbation Resampling

Although the asymptotic variance for \( \hat{\Delta} \) is specified through the influence function in (2.16), a direct estimate is difficult because the influence functions involves complicated functionals of the data distribution. We instead propose a simple perturbation resampling procedure for inference. Let \( \mathcal{G} = \{G_i : i = 1, \ldots, N\} \) be non-negative iid random variables with unit mean and variance that are independent of the observed data \( \mathcal{D} \). We first obtained perturbed estimators of \( \alpha \) and \( \beta \):

\[
\hat{\alpha}^* = \arg \min_{\alpha} \left\{ -n^{-1} \sum_{i=1}^{n} \ell_\pi(\alpha; X_i, T_i)G_i + p_{\lambda_N}^*(\alpha_1) \right\}
\]
\[ \hat{\beta}^* = \arg \min_\beta \left\{ -n^{-1} \sum_{i=1}^{n} \ell_{\mu}(\beta; Y_i, X_i, T_i) G_i + p_{\lambda_n}(\theta_{(\cdot)}) \right\}, \]

where \( p_{\lambda_n}(\cdot) \) and \( p_{\lambda_n}^*(\cdot) \) are the corresponding penalties based on the perturbed data if data-adaptive weights are used, such as for adaptive LASSO. This leads to the perturbed double-index PS:

\[ \hat{\pi}^* (x; \hat{\alpha}_1^*, \hat{\beta}_1^*) = \frac{\sum_{j=1}^{N} K_h(\hat{S}_j^* - \hat{s}^*) I(T_j = 1) G_j}{\sum_{j=1}^{N} K_h(\hat{S}_j^* - \hat{s}^*) G_j}, \]

where \( \hat{S}_j = (\hat{\alpha}_1^*, \hat{\beta}_1^*)^T X_j \) and \( \hat{s}^* = (\hat{\alpha}_1^*, \hat{\beta}_1^*)^T x \) are the perturbed bivariate scores. We then obtain the perturbed estimator for \( \gamma, \hat{\gamma}^* \), as the solution to:

\[ n^{-1} \sum_{i=1}^{n} Z_{\hat{\pi}^*,i} \left\{ Y_i - g_\xi(\gamma Z_{\hat{\pi}^*,i}) \right\} G_i + \lambda_n \gamma_{(\cdot)} = 0, \]

where \( \hat{\pi}^* \) specifies that the imputations use utility covariates that plug in \( \hat{\pi}^*(x; \hat{\alpha}_1^*, \hat{\beta}_1^*) \).

Finally, we calculate the perturbed SS_{or} estimator as \( \hat{\Delta}^* = \hat{\mu}_1^* - \hat{\mu}_0^* \), where:

\[ \hat{\mu}_1^* = \left\{ \frac{\sum_{i=1}^{N} I(T_i = 1) G_i}{\hat{\pi}(X_i; \hat{\alpha}_1^*, \hat{\beta}_1^*)} \right\}^{-1} \left\{ \frac{\sum_{i=1}^{N} I(T_i = 1) \hat{Y}_i^* G_i}{\hat{\pi}(X_i; \hat{\alpha}_1^*, \hat{\beta}_1^*)} \right\}, \]

and

\[ \hat{\mu}_0^* = \left\{ \frac{\sum_{i=1}^{N} I(T_i = 0) G_i}{1 - \hat{\pi}(X_i; \hat{\alpha}_1^*, \hat{\beta}_1^*)} \right\}^{-1} \left\{ \frac{\sum_{i=1}^{N} I(T_i = 0) \hat{Y}_i^* G_i}{1 - \hat{\pi}(X_i; \hat{\alpha}_1^*, \hat{\beta}_1^*)} \right\}, \]

with \( \hat{Y}_i^* = g_\xi(\hat{\gamma}^* Z_{\hat{\pi}^*,i}) \). It can be shown based on arguments similar to those in Tian et al. (2007) that the asymptotic distribution of \( n^{1/2}(\hat{\Delta} - \bar{\Delta}) \) coincides with that of \( n^{1/2}(\hat{\Delta}^* - \bar{\Delta}) \mid \mathcal{F} \). In this perturbation scheme, estimation of \( \alpha \), the double-index PS through kernel smoothing, and the final IPW estimators \( \hat{\mu}_1^* \) and \( \hat{\mu}_0^* \) does not technically need to be perturbed as they are estimated based on data from \( \mathcal{R} \), and their contributions to the asymptotic variance is of higher order when \( N \gg n \). However, we found that not perturbing these steps can have some impact on the standard error estimation in finite samples if \( N \) is not yet very large relative to \( n \) and chose to perturb these steps by default. We approximate the standard error of \( \hat{\Delta} \) based on the empirical standard deviation, or, as a robust alternative, the mean absolute deviation (MAD), of a large number of samples of \( \hat{\Delta}^* \) and construct confidence intervals (CI) based on the empirical percentiles.
2.7 Numerical Studies

2.7.1 Simulations

We performed simulations to assess the finite samples bias and relative efficiency (RE) of our proposed estimator (SS\textsubscript{DR}) compared to alternative estimators. In separate simulations we also examined the performance of the perturbation procedure for inference based on SS\textsubscript{DR}. For SS\textsubscript{DR}, we specified \( h(\cdot) \) in the imputation model in (2.6) as natural cubic splines with 6 knots specified at uniform quantiles. Natural cubic splines were also applied to \( \hat{U}_{\pi} \) for additional flexibility in the imputation model. Ridge regression with the tuning parameter chosen by cross-validation on the deviance was used for regularization in (2.7), and adaptive LASSO with initial weights estimated by ridge regression and tuning parameter chosen by minimizing a modified BIC criteria (Minnier et al., 2011) in (2.11) and (2.12). A plug-in estimate was used for the bandwidth in the smoothing for the double-index PS as in Chapter 1. Prior to smoothing the components of \( \hat{S} \) were standardized and transformed by a probability integral transform based on the normal cumulative distribution function to induce approximately uniformly distributed inputs, which can improve finite-sample performance (Wand et al., 1991). As we focused on binary outcomes, we specified \( g_\xi(u) = g_\pi(u) = g_\mu(u) = 1/(1 + e^{-u}) \) for the working models in (2.6) and (2.10).

For comparison, we considered common CC ATE estimators (Lunceford and Davidian, 2004; Kang and Schafer, 2007), including the standard IPW estimator (CC\textsubscript{IPW}), outcome regression estimator (CC\textsubscript{REG}), and the standard doubly-robust estimator (CC\textsubscript{DR}). We also considered two estimators that leverage \( \mathcal{Z} \). The first is a naive imputation approach (SS\textsubscript{Naive}), in which \( Y \) is imputed using a logistic regression of \( Y \) on \( V \) and \( T \), and the imputations are plugged into a standard IPW estimator. The second is adapted from a principled estimator for pretest-posttest randomized studies with \( Y \) missing at random (SS\textsubscript{PrePost}) (Davidian et al., 2005). We modified the estimator by replacing instances of the randomization probability by PS estimates \( \pi(X_i; \alpha) \). SS\textsubscript{PrePost} is also doubly-robust so that it is consistent for \( \Delta \) when either the PS model \( \pi(x; \alpha) \) or baseline outcome model \( \mu_k(x; \beta) \) is correctly specified, providing another approach to leverage \( \mathcal{Z} \) without requiring correct specification of an imputation model. However, SS\textsubscript{PrePost} is constructed to achieve the effi-
ciency bound in a model where the distribution of \((V^T, T)^T\) is unknown and \(Y\) is missing at random with the missingness proportion bounded away from 0. This bound differs from that of the SS model we consider, and the RE simulations correspondingly show that SS_{DR} is more efficient under a SS setup. For all reference methods, the same logistic regression models with main effects only were used for the underlying requisite PS, baseline outcome, and imputation models.

To mimic the EMR data, we considered the case with \(Y\) as binary and \(W\) as count variables. In all scenarios, data were generated according to \(X \sim N\{0, \sigma_x^2(1 - \rho_x)I + \sigma_x^2\rho_x\}\), \(T \mid X \sim Bernoulli\{\pi(X)\}\), \(Y \mid X, T \sim Bernoulli\{\mu_T(X)\}\), and \(W = [\Gamma(1, X^T, T, Y)^T + \epsilon]\), where \(\epsilon \sim N\{0, \sigma_w^2(1 - \rho_w)I + \sigma_w^2\rho_w\}\) and \(\lfloor \cdot \rfloor\) is the floor function. We considered \(p_x = 10\) baseline covariates and \(p_w = 5\) surrogates, with variances and correlations \(\sigma_x^2 = 1, \rho_x = .2, \sigma_w^2 = 5, \rho_w = .2\), and \(\Gamma_{5 \times 13} = (0_{5 \times 1}, 11_{5 \times 5}, -11_{5 \times 5}, 11_{5 \times 1}, (5, 5, 2.5, 0, 0)^T)\). We varied the simulations over different model specifications and sample sizes. The imputation model was misspecified across all settings, whereas three scenarios were considered for the baseline outcome model \(\mu_k(x; \beta)\) and the PS model \(\pi(x; \alpha)\): (1) both models are correctly specified, (2) the PS model is correctly specified but outcome model is misspecified, and (3) the outcome model is correctly specified but the PS model is misspecified. The true \(\pi(x)\) and \(\mu_k(x)\) for these three settings were specified as:

\[
(1) \text{Both correct}: \mu_k(x) = g_\mu(\beta_0 + \beta_1^T x + \beta_2 k), \quad \pi(x) = g_\pi(\alpha_0 + \alpha_1^T x)
\]

\[
(2) \text{Misspecified } \mu: \mu_k(x) = g_\mu\{\beta_0 + \beta_{1[1]}^T x(\beta_{1[2]}^T x + 1) + \beta_2 k\}, \quad \pi(x) = g_\pi(\alpha_0 + \alpha_1^T x)
\]

\[
(3) \text{Misspecified } \pi: \mu_k(x) = g_\mu(\beta_0 + \beta_1^T x + \beta_2 k), \quad \pi(x) = g_\pi\{\alpha_0 + \alpha_{1[1]}^T x(\alpha_{1[2]}^T x + 1)\},
\]

where \(g_\mu(u) = g_\pi(u) = 1/(1 + e^{-u})\) and parameter values were:

\[
\alpha_0 = -3, \quad \alpha_1 = 0.351_{1 \times 10}, \quad \beta_0 = -0.65, \quad \beta_1 = (11_{1 \times 3}, .51_{1 \times 3}, -1.15, -11_{1 \times 3})^T
\]

\[
\alpha_{1[1]} = .5(0, .35, 0, .35, 0, .35, 0, .35, 0, .35)^T, \quad \alpha_{1[2]} = (.35, 0, .35, 0, .35, 0, .35, 0, .35, 0)^T
\]

\[
\beta_{1[1]} = .5(1, 0, 1, 0, 5, 0, -5, 0, -1, 0)^T, \quad \beta_{1[2]} = (0, .5, 0, .5, 0, .5, 0, .5, 0, .5)^T.
\]

We considered sample sizes of (A) \(n = 100\) with \(N = 1, 112\), (B) \(n = 250\) with \(N = 5, 000\), and (C) \(n = 500\) with \(N = 125, 000\). The results in each scenario are summarized from 1,000 simulated datasets.
Table 2.1 presents the bias and root-mean square error (RMSE) across misspecification scenarios. SS_{DR}, SS_{PrePost}, and CC_{DR} exhibits low bias that diminishes to zero as sample size increased under all three scenarios, demonstrating their double-robustness. In contrast, CC_{REG} and CC_{IPW} exhibit substantial bias when working models \( \mu_k(x; \beta) \) and \( \pi(x; \alpha) \) are misspecified, respectively. SS_{Naive} requires correct specification of both the imputation model \( g_\xi(\gamma^Tz_\pi) \) and PS model \( \pi(x; \alpha) \) for consistency but shows negligible bias when both \( \mu_k(x; \beta) \) and \( \pi(x; \alpha) \) are correctly specified, as the logistic regression imputation model likely provides an adequate approximation under the given data generating process. Still, it incurs substantial bias if either \( \pi(x; \alpha) \) or \( \mu_k(x; \beta) \) are misspecified, where \( g_\xi(\gamma^Tz_\pi) \) becomes further misspecified in the later case.

Table 2.1: Bias and RMSE of estimators under different model misspecification scenarios over 1,000 simulated datasets.

<table>
<thead>
<tr>
<th>Size</th>
<th>Estimator</th>
<th>Both Correct Bias</th>
<th>RMSE</th>
<th>Misspecified ( \mu ) Bias</th>
<th>RMSE</th>
<th>Misspecified ( \pi ) Bias</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 100, N = 1112 )</td>
<td>CC_{IPW}</td>
<td>0.008</td>
<td>0.235</td>
<td>0.012</td>
<td>0.296</td>
<td>0.014</td>
<td>0.126</td>
</tr>
<tr>
<td></td>
<td>CC_{REG}</td>
<td>-0.001</td>
<td>0.096</td>
<td>0.023</td>
<td>0.127</td>
<td>0.003</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>CC_{DR}</td>
<td>-0.001</td>
<td>0.118</td>
<td>0.014</td>
<td>0.190</td>
<td>0.002</td>
<td>0.075</td>
</tr>
<tr>
<td></td>
<td>SS_{Naive}</td>
<td>-0.004</td>
<td>0.104</td>
<td>0.012</td>
<td>0.115</td>
<td>0.017</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>SS_{PrePost}</td>
<td>-0.004</td>
<td>0.094</td>
<td>0.013</td>
<td>0.126</td>
<td>-0.002</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>SS_{DR}</td>
<td>-0.006</td>
<td>0.077</td>
<td>-0.005</td>
<td>0.083</td>
<td>0.000</td>
<td>0.048</td>
</tr>
<tr>
<td>( n = 500, N = 12500 )</td>
<td>CC_{IPW}</td>
<td>-0.001</td>
<td>0.115</td>
<td>0.004</td>
<td>0.117</td>
<td>0.020</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>CC_{REG}</td>
<td>0.000</td>
<td>0.041</td>
<td>0.022</td>
<td>0.057</td>
<td>-0.002</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>CC_{DR}</td>
<td>-0.002</td>
<td>0.063</td>
<td>0.002</td>
<td>0.081</td>
<td>-0.002</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>SS_{Naive}</td>
<td>0.001</td>
<td>0.029</td>
<td>0.009</td>
<td>0.038</td>
<td>0.018</td>
<td>0.027</td>
</tr>
<tr>
<td></td>
<td>SS_{PrePost}</td>
<td>0.001</td>
<td>0.033</td>
<td>0.000</td>
<td>0.048</td>
<td>-0.002</td>
<td>0.021</td>
</tr>
<tr>
<td></td>
<td>SS_{DR}</td>
<td>-0.002</td>
<td>0.028</td>
<td>-0.001</td>
<td>0.037</td>
<td>-0.001</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Figure 2.1 presents the RE of various estimators relative to CC_{DR} across different scenarios. In the small sample case where \( n = 100 \) and \( N = 1,112 \), SS_{DR} is much more efficient than both the CC and other SS estimators, regardless of the specification scenario. It gains over the other SS estimators since its asymptotic variance approximates the SS efficiency bound and through the use of regularization to estimate nuisance parameters. The efficiency gain is most prominent under misspecification of \( \mu_k(x; \beta) \), which may be driven
by the lack of impact on the influence function by \( \mu_k(x; \beta) \) when \( \pi(x; \alpha) \) is correctly specified. In the large sample case where \( n = 500 \) and \( N = 125,000 \), \( \text{SS}_{\text{DR}} \) is still uniformly most efficient, but the gains are somewhat less pronounced. This is expected at least in part from the reduced role of regularization in large samples. Though \( \text{SS}_{\text{Naive}} \) may appear to be competitive in large samples if \( \pi(x; \alpha) \) is correctly specified, it may suffer drastic efficiency loss from misspecification of its imputation model under other data generating processes.

Figure 2.1: RE of estimators, defined as the ratio of mean square errors (MSE) relative to \( \text{CC}_{\text{DR}} \), by model misspecification scenarios over 1,000 simulated datasets. Higher values of RE denotes greater efficiency (lower MSE) relative to \( \text{CC}_{\text{DR}} \). Higher values of RE denotes greater efficiency (lower MSE) relative to \( \text{DR-CC} \).

To implement the perturbation procedure, we used the weights \( G_i \sim 4 \times \text{Beta}(0.5, 1.5) \) and 1,000 sets of \( G \) for SE and CI estimation. We considered evaluating the perturbations only in the scenario where both \( \mu_k(x; \beta) \) and \( \pi(x; \alpha) \) were correctly specified models. The results are presented in Table 2.2. In small samples, SEs estimated by the standard deviation tended to over-estimate due to the presence of outlying perturbed estimates, while SEs estimated by MAD tended to under-estimate. In larger samples, the SE estimation improves. The coverage probabilities are close to nominal levels but slightly under-cover in the sample sizes considered. In other simulations not reported we found that perturbation with weights sampled from a multinomial distribution of size \( N \) and \( N \) categories
with equal probabilities, which effectively implements the bootstrap, to exhibit improved coverage probabilities. However, justifying the bootstrap may be more involved due to the correlated weights.

Table 2.2: Performance of perturbation resampling for $S_{DR}$ in 1,000 simulated datasets when both $\mu_k(x; \beta)$ and $\pi(x; \alpha)$ are correctly specified. Emp SE: empirical SE of $S_{DR}$ over simulated datasets; ASE: average of estimated SE based on the standard deviation of perturbed estimates; $A_{SE\text{MAD}}$: average of SE based on MAD of perturbed estimates; RMSE: root-mean square error; Coverage: coverage of 95% percentile CIs.

<table>
<thead>
<tr>
<th>Size</th>
<th>Bias</th>
<th>Emp SE</th>
<th>ASE</th>
<th>$A_{SE\text{MAD}}$</th>
<th>RMSE</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n = 100, N = 1112$</td>
<td>-0.005</td>
<td>0.076</td>
<td>0.306</td>
<td>0.057</td>
<td>0.076</td>
<td>0.914</td>
</tr>
<tr>
<td>$n = 250, N = 5000$</td>
<td>-0.003</td>
<td>0.043</td>
<td>0.039</td>
<td>0.036</td>
<td>0.043</td>
<td>0.921</td>
</tr>
</tbody>
</table>

2.7.2 EMR Data Application

We applied $S_{DR}$ and the alternative estimators to compare the rates of treatment response to two biologic agents for treating patients with inflammatory bowel disease (IBD) using the EMRs of two large metropolitan academic medical centers. Though the efficacy and effectiveness of adalimumab (ADA) and infliximab (IFX) for the management of IBD have been established individually, few studies have offered a direct comparison. Consequently the choice of treatment in practice is often influenced by factors other than comparative performance (Ananthakrishnan et al., 2016). Randomized trials may be unfeasible due to the large number of patients that would be needed to detect the presumed small treatment difference, and other observational data lack clinical information needed to ascertain meaningful outcomes and covariates for adjustment. EMRs are thus uniquely positioned to provide evidence on the comparative effectiveness of these two therapies.

The data we considered consisted of $N = 1,243$ total IBD patients, including 200 who initiated treatment with ADA and 1043 with IFX. Through chart review by a gastroenterologist, a random subset of $n = 117$ records were labeled with the true treatment response status (responder vs. non-responder) within one year of treatment initiation. We included 12 baseline covariates to adjust for confounding in $X$, including demographics,
comorbidities, prior utilization, and inflammation biomarker levels. We also selected 35 post-treatment surrogates for \( W \), comprising of counts of NLP mentions of clinically relevant terms (e.g. “bleeding”, “fistula”, “tenesmus”) within one year of initiation. The transformation \( u \mapsto \log(1 + u) \) was applied to all count variables in \( V \) to mitigate instability in the estimation due to skewness in their distributions. Nonparametric bootstrap was used to estimate SEs and CIs for the alternative estimators and perturbation for \( SS_{DR} \), using the MAD of resampled estimates as a robust estimator of the SEs. In addition we calculated two-sided p-values based on inverting percentile CIs from the resampled estimates, using the well-known equivalence between significance tests and confidence sets (Liu and Singh, 1997; Davison et al., 2003).

As shown in Table 2.3, the point estimates of most estimators agreed that patients receiving ADA experienced lower rates of treatment response, after adjustment for confounding. \( SS_{DR} \) is estimated to achieve more than 600% efficiency gain over CC estimators and 450% efficiency gain over the other SS estimators based on the estimated variances. It is the only estimator that exhibits a difference that is significant at the .05 level, suggesting that patients receiving IFX experience a modest benefit in the rate of response.

Table 2.3: Point and SE estimates based on MAD for the ATE of ADA vs. IFX, with respect to one-year treatment response rate, among IBD patients in EMR data based on various methods, including the naive CC estimator (CC\( _{Naive} \)) that completely ignores confounding bias. 95% CIs are percentile-based CIs from resampling and p-values are for testing \( H_0 : \Delta = 0 \) based on inverting percentile CIs.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Estimate</th>
<th>SE</th>
<th>95% CI (Pct)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC( _{Naive} )</td>
<td>0.014</td>
<td>0.099</td>
<td>(-0.201, 0.177)</td>
<td>0.822</td>
</tr>
<tr>
<td>CC( _{IPW} )</td>
<td>-0.227</td>
<td>0.325</td>
<td>(-0.558, 0.164)</td>
<td>0.714</td>
</tr>
<tr>
<td>CC( _{REG} )</td>
<td>-0.067</td>
<td>0.123</td>
<td>(-0.273, 0.162)</td>
<td>0.732</td>
</tr>
<tr>
<td>CC( _{DR} )</td>
<td>-0.125</td>
<td>0.153</td>
<td>(-0.416, 0.164)</td>
<td>0.592</td>
</tr>
<tr>
<td>SS( _{Naive} )</td>
<td>-0.051</td>
<td>0.088</td>
<td>(-0.318, 0.065)</td>
<td>0.198</td>
</tr>
<tr>
<td>SS( _{PrePost} )</td>
<td>0.033</td>
<td>0.109</td>
<td>(-0.265, 0.180)</td>
<td>0.778</td>
</tr>
<tr>
<td>SS( _{DR} )</td>
<td>-0.067</td>
<td>0.036</td>
<td>(-0.164, -0.002)</td>
<td>0.044</td>
</tr>
</tbody>
</table>
2.8 Discussion

The lack of direct observation on gold-standard outcomes of interest makes it challenging to perform CER using EMR data. Under a SS setting where the true outcome is labeled for only a small subset of patients, we developed an efficient and robust estimator for the ATE that addresses the missingness in the outcome and confounding bias. The estimator adopts an imputation approach to leverage surrogate data from $U$ to improve efficiency. It is not only robust to misspecification of the imputation model but also possesses the traditional doubly-robust property, requiring only correct specification of either a PS or baseline outcome model to be consistent. We showed that it is locally semiparametric efficient under an ideal SS semiparametric model and demonstrated through simulations that it is more efficient than available CC and alternative SS estimators, even under misspecification of working models.

The efficiency gain over CC estimators under a nonparametric model is not obvious if the imputation model is badly misspecified. In a favorable scenario for CC estimators, at distributions of the data for which $\pi(x; \alpha)$ and $\mu_k(x; \beta)$ are correctly specified so that efficient CC estimators under the nonparametric model are available and yet $g_\xi(\gamma^Tz_{\pi})$ is misspecified, it can be shown by comparing the asymptotic variances that $\hat{\Delta}$ will still be more efficient if $E[U_2^2\{Y - g_\xi(\gamma^Tz_{\pi})\} \{g_\xi(\gamma^Tz_{\pi}) - \mu_T(x; \beta)\}] = 0$. This condition can be guaranteed, for example, if linear link functions are used for $g_\xi(\cdot)$ and $g_\mu(\cdot)$, and each model includes interactions between $U_2^2\hat{\pi}$ and its linear predictor. Another potential approach is to include $U_2^2\mu_T(x_i; \hat{\beta})$ among the covariates of the imputation model and to introduce a more general parameterization of the imputation model given by $g_{\xi, \rho}(\gamma^Tz_{\pi})$ such that $\rho$ is a parameter enforcing $E[U_2^2\{Y - g_{\xi, \rho}(\gamma^Tz_{\pi})\} g_{\xi, \rho}(\gamma^Tz_{\pi})] = 0$ at some $\rho = \rho_0$.

We have assumed that the true outcomes $Y$ are labeled completely at random, which is usually reasonable since researchers can control the labeling. This assumption could be restrictive if labeling was stratified by some known factors or if some records that are available were not labeled for research purposes, in which case the labeling decision may not have been random. One possible approach to address the case where $Y$ are missing at random is to apply weighting or semiparametric efficient methods (Robins et al., 1994,
difficulty. Interaction effects between outcome model under a fully nonparametric model. The results in Appendix B show that
In the following, the supporting lemmas of Appendix A identify rates of convergence for 2.9 Appendix
1995; Robins and Rotnitzky, 1995) to the estimating equation when estimating
required for the double-index PS in Appendix A of Chapter 1 hold. In the case where W is high dimensional, group LASSO (Yuan and Lin, 2006) where the basis expansion functions for each surrogate variable are grouped together can also potentially be used to improve efficiency in finite-samples. It would also be of interest to extend the theoretical results to the case where \( p_x \) and \( p_w \) are allowed to diverge with \( n \).

2.9 Appendix
In the following, the supporting lemmas of Appendix A identify rates of convergence for frequently encountered quantities and also identify the efficient influence function for \( \hat{\Delta}^* \) under a fully nonparametric model. The results in Appendix B show that \( \hat{\Delta} \) is consistent and asymptotically linear, deriving its influence function. The results in Appendix C establish the semiparametric efficiency bound under the SS model and shows that \( \hat{\Delta} \) achieves this bound at particular distributions for the data so that it is locally semiparametric efficient. Throughout this Appendix we assume that mild regularity conditions required for the double-index PS in Appendix A of Chapter 1 hold.

The following notations facilitate the subsequent derivations. Let \( \pi_k(x) = \mathbb{P}(T = k \mid X = x) \) for \( k = 0, 1 \). Let \( \pi_k(x; \alpha_1, \beta_1) = \mathbb{P}(T = k \mid \alpha_1^T X = \alpha_1^T x, \beta_1^T X = \beta_1^T x) \), \( \pi_k(x; \alpha) = \pi(x; \alpha)^k \{1 - \pi(x; \alpha)\}^{1-k} \), and \( \hat{\pi}_k(x; \alpha_1, \beta_1) = \hat{\pi}(x; \alpha_1, \beta_1)^k \{1 - \hat{\pi}_k(x; \alpha_1, \beta_1)\}^{1-k} \) for given \( \alpha_1, \beta_1 \in \mathbb{R}^p \) and \( \alpha \in \mathbb{R}^{p+1} \) and \( k = 0, 1 \). Moreover, let \( \hat{\vartheta} = (\hat{\alpha}_1^T, \hat{\beta}_1^T)^T \), \( \hat{\vartheta} = (\alpha_1^T, \beta_1^T)^T \), \( \pi_k(x; \hat{\vartheta}) = \pi_k(x; \alpha_1, \beta_1), \hat{\pi}_k(x; \hat{\vartheta}) = \hat{\pi}_k(x; \alpha_1, \beta_1), \) and \( \hat{\pi}_k(x; \hat{\vartheta}) = \hat{\pi}_k(x; \alpha_1, \beta_1) \). Let the working imputation model be denoted by \( \xi_T(V; \gamma, \pi) = g_\xi(\gamma^T (1, h(V)^T, T, U)^T) \), where \( U_\pi = I(T = 1)/\pi(X) - I(T = 0)/(1 - \pi(X)) \), given some PS \( \pi \). Let \( \omega_{k,i} = I(T_i = k)/\pi_k(X_i), \hat{\omega}_{k,i} = I(T_i = k)/\pi_k(X_i, \hat{\vartheta}), \) \( \hat{\omega}_{k,i} = I(T_i = k)/\hat{\pi}_k(X_i, \hat{\vartheta}), \) and \( S = (\alpha_1, \beta_1)^T X \) with \( \hat{S}_i = (\hat{\alpha}_1, \hat{\beta}_1)^T X_i \) for \( k = 0, 1 \) and \( i = 1, \ldots, N \).
2.9.1 Appendix A: Supporting Lemmas

Lemma 1. The rates of uniform convergence for kernel estimators we use are as follows:

\[
\sup_x \| \hat{\pi}_k(x; \tilde{\theta}) - \pi_k(x; \bar{\theta}) \| = O_p(\bar{a}_N) \quad (2.18)
\]

\[
\sup_x \left\| \frac{\partial}{\partial \alpha_1^i} \hat{\pi}_k(x; \tilde{\theta}) - \frac{\partial}{\partial \alpha_1^i} \pi_k(x; \bar{\theta}) \right\| = O_p(\bar{b}_N) \quad (2.19)
\]

\[
\sup_x \left\| \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \tilde{\theta}) - \frac{\partial}{\partial \beta_1} \pi_k(x; \bar{\theta}) \right\| = O_p(\bar{b}_N) \quad (2.20)
\]

\[
\sup_x \| \hat{\pi}_k(x; \tilde{\theta}) - \pi_k(x; \bar{\theta}) \| = O_p(a_n), \quad (2.21)
\]

where:

\[ \bar{a}_N = h^q + \left( \frac{\log N}{Nh^2} \right)^{1/2}, \quad \bar{b}_N = h^q + \left( \frac{\log N}{Nh^2} \right)^{1/2}, \text{ and } a_n = n^{-1/2} + n^{-1/2} \bar{b}_N + \bar{a}_N. \]

Proof. The uniform rates for fixed \( \alpha_1 \) and \( \beta_1 \) in the first three equations follow directly from the uniform convergence rates of kernel smoothers and their first derivatives (Hansen, 2008b). To establish the uniform convergence rate for DiPS, we first note that:

\[
\sup_x \| \hat{\pi}_k(x; \tilde{\theta}) - \pi_k(x; \tilde{\theta}) \| \leq \sup_x \| \hat{\pi}_k(x; \tilde{\theta}) - \hat{\pi}_k(x; \bar{\theta}) \| + \sup_x \| \hat{\pi}_k(x; \bar{\theta}) - \pi_k(x; \bar{\theta}) \|. \]

The first term on the right-hand side can be written:

\[
\sup_x \| \hat{\pi}_k(x; \tilde{\theta}) - \hat{\pi}_k(x; \bar{\theta}) \| \leq \sup_x \left\| \frac{\partial}{\partial \alpha_1} \hat{\pi}_k(x; \alpha_1, \beta_1)(\alpha_1 - \bar{\alpha}_1) + \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \alpha_1, \beta_1)(\beta_1 - \bar{\beta}_1) \right\|
\]

\[
+ \sup_x \left\| \left\{ \frac{\partial}{\partial \alpha_1^i} \hat{\pi}_k(x; \alpha_1, \beta_1) - \frac{\partial}{\partial \alpha_1^i} \pi_k(x; \alpha_1, \beta_1) \right\} (\alpha_1 - \bar{\alpha}_1) \right\|
\]

\[
+ \sup_x \left\| \left\{ \frac{\partial}{\partial \beta_1} \hat{\pi}_k(x; \alpha_1, \beta_1) - \frac{\partial}{\partial \beta_1} \pi_k(x; \alpha_1, \beta_1) \right\} (\beta_1 - \bar{\beta}_1) \right\|
\]

\[
+ O_p(\| \alpha_1 - \bar{\alpha}_1 \|^2 + \| \beta_1 - \bar{\beta}_1 \|^2 + \| \alpha_1 - \bar{\alpha}_1 \| \| \beta_1 - \bar{\beta}_1 \|). \]

We obtain the desired rate by collecting terms and applying the other rates from above, using that \( \frac{\partial}{\partial \alpha_1^i} \pi_k(x; \alpha_1, \beta_1) \) and \( \frac{\partial}{\partial \beta_1} \pi_k(x; \alpha_1, \beta_1) \) are continuous in \( x \), and \( x \) lies in a compact covariate space. \( \square \)
Lemma 2. Let $\zeta_i = g(Z_i)$ be some integrable function of $Z_i = (V_i^T, T_i)^T$, for $i = 1, \ldots, N$. Then:

$$\sqrt{N} \sum_{i=1}^{N} \tilde{\omega}_{k,i} \zeta_i = \mathbb{E}(\tilde{\omega}_{k,i} \zeta_i) + O_p(c_n),$$

(2.22)

where $c_n = \tilde{a}_N + n^{-1/2}N^{-1}h^{-3}$.

Proof. Consider the decomposition:

$$N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{k,i} \zeta_i = S_{1,k} + S_{2,k} + S_{3,k},$$

where:

$$S_{1,k} = N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{i,k} \zeta_i, \quad S_{2,k} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{\tilde{\pi}_k(X_i; \bar{\theta})} - \frac{1}{\pi_k(X_i; \bar{\theta})} \right\} I(T_i = k) \zeta_i,$$

and

$$S_{3,k} = N^{-1} \sum_{i=1}^{N} \left\{ \frac{1}{\pi_k(X_i; \bar{\theta})} - \frac{1}{\tilde{\pi}_k(X_i; \bar{\theta})} \right\} I(T_i = k) \zeta_i.$$

The second term can be bounded:

$$|S_{2,k}| \leq \sup_x \left\| \tilde{\pi}_k(x; \bar{\theta}) - \pi_k(x; \bar{\theta}) \right\| N^{-1} \sum_{i=1}^{N} \frac{I(T_i = k) \zeta_i}{\tilde{\pi}_k(X_i; \bar{\theta}) \pi_k(X_i; \bar{\theta})} = O_p(\tilde{a}_N).$$

The third term can be written:

$$S_{3,k} = N^{-1} \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_1} I(T_i = k) \zeta_i (\bar{\alpha}_1 - \alpha_1) + \frac{\partial}{\partial \beta_1} I(T_i = k) \zeta_i (\bar{\beta}_1 - \beta_1)
+ O_p \left( \|\bar{\alpha}_1 - \alpha_1\|^2 + \|\bar{\beta}_1 - \beta_1\|^2 + \|\bar{\alpha}_1 - \alpha_1\| \|\bar{\beta}_1 - \beta_1\| \right)
= O_p \left\{ (1 + N^{-1/2}h^{-1} + N^{-1}h^{-3})n^{-1/2} \right\}$$

where we use that $\frac{\partial}{\partial \alpha_1} 1/\tilde{\pi}_k(X_i; \alpha_1, \beta_1)$ and $\frac{\partial}{\partial \beta_1} 1/\tilde{\pi}_k(X_i; \alpha_1, \beta_1)$ are Lipshitz continuous in $\alpha_1$ and $\beta_1$ for the first equality and the rate deduced from an analogous term in Chapter 1 for the second equality. The desired result follows from collecting the dominant rates.

Lemma 3. Let $Y_i^\dagger = \xi_{T_i}(V_i; \tilde{\gamma}, \pi)$ for $i = 1, \ldots, N$. Then:

$$\sqrt{\frac{N}{n}} \sum_{i=1}^{N} \tilde{\omega}_{k,i} (Y_i^\dagger - \bar{\mu}_k) = O_p(1 + d_n),$$

(2.23)

where $d_n = \nu_n^{1/2}N^{1/2}h^q + \nu_n^{1/2}N^{-1/2}h^{-2} + N^{-1/2}h^{-1} + N^{-1}h^{-3}$. 

57
Proof. Consider the decomposition:

\[
\frac{\sqrt{n}}{N} \sum_{i=1}^{N} \tilde{\omega}_{k,i} (Y_i^\dagger - \bar{\mu}_k) = \tilde{W}_{1,1,k} + \tilde{W}_{2,1,k} + \tilde{W}_{3,1,k},
\]

where:

\[
\tilde{W}_{1,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \tilde{\omega}_{k,i} (Y_i^\dagger - \bar{\mu}_k)
\]

\[
\tilde{W}_{2,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left\{ \frac{1}{\hat{\pi}_k(X_i; \vartheta)} - \frac{1}{\hat{\pi}_k(X_i; \theta)} \right\} I(T_i = k) (Y_i^\dagger - \bar{\mu}_k)
\]

\[
\tilde{W}_{3,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \left\{ \frac{1}{\hat{\pi}_k(X_i; \vartheta)} - \frac{1}{\hat{\pi}_k(X_i; \theta)} \right\} I(T_i = k) (Y_i^\dagger - \bar{\mu}_k).
\]

The first term is a scaled sum of iid centered terms so that:

\[
\tilde{W}_{1,1,k} = \nu_n^{1/2} N^{-1/2} \sum_{i=1}^{N} \tilde{\omega}_{k,i} (Y_i^\dagger - \bar{\mu}_k) = \nu_n^{1/2} O_p(1).
\]

The V-statistic arguments similar to those from Chapter 1, the second term can be written:

\[
\tilde{W}_{2,1,k} = - \nu_n^{1/2} N^{-1/2} \sum_{i=1}^{N} \mathbb{E}(Y_i^\dagger \mid \bar{S}_i, T_i = k) \left\{ \frac{I(T_i = k)}{\hat{\pi}_k(X_i; \vartheta)} - 1 \right\} + O_p \left\{ \nu_n^{1/2} (h^q + N^{-1/2} h^{-2}) \right\} + O_p \left\{ \nu_n^{1/2} (h^q + N^{1/2} h^q + N^{-1/2} h^{-2}) \right\} + O_p (\nu_n^{1/2} N^{1/2} \bar{a}_N^2)
\]

\[
= O_p (\nu_n^{1/2} N^{1/2} h^q + \nu_n^{1/2} N^{-1/2} h^{-2}).
\]

The final term can be written:

\[
\tilde{W}_{3,1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^{N} \frac{\partial}{\partial \hat{\alpha}_1^T} I(T_i = k) (Y_i^\dagger - \bar{\mu}_k) (\hat{\alpha}_1 - \alpha_1) + \frac{\partial}{\partial \hat{\beta}_1} I(T_i = k) (Y_i^\dagger - \bar{\mu}_k) (\hat{\beta}_1 - \beta_1)
\]

\[
+ O_p \left\{ n^{1/2} \left( \| \hat{\alpha}_1 - \alpha_1 \|^2 + \| \hat{\beta}_1 - \beta_1 \|^2 + \| \hat{\alpha}_1 - \alpha_1 \| \| \hat{\beta}_1 - \beta_1 \| \right) \right\} + O_p (1 + N^{-1/2} h^{-1} + N^{-1/2} h^{-3}) O_p(1).
\]

where we use that \( \frac{\partial}{\partial \hat{\alpha}_1} 1/\hat{\pi}_k(X_i; \alpha, \beta) \) and \( \frac{\partial}{\partial \hat{\beta}_1} 1/\hat{\pi}_k(X_i; \alpha, \beta) \) are Lipshitz continuous in \( \alpha \) and \( \beta \) for the first equality and used the rate deduced from an analogous term from Chapter 1 for the second equality. The desired result follows from collecting the rates. \( \square \)
Lemma 4. Let $\mathcal{M}_{NP} = \{f_{Y,Z}(y,z) : \text{there exists a } \epsilon_\pi > 0 \text{ such that } \pi_1(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for all } x \text{ where } f_X(x) > 0 \}$ be a nonparametric model for the distribution of $(Y,Z)$, where $z = (v^T, t)^T$. Let $M_{NP,sub} = \{f_{Y,Z}(y,z; \theta) : \theta \in \Theta \}$ denote a regular parametric submodel of $\mathcal{M}_{NP}$, where $\theta$ is a finite-dimensional parameter and the true density is at $\theta = \theta^*$. Let $\mathcal{P}_{NP}$ be the collection of all such regular parametric submodels that satisfy:

1. $E_\theta[E_\theta^*(Y \mid X, T = k)]^2$ is continuous in $\theta$ at $\theta = \theta^*$ for $k = 0, 1$, where $E_\theta(\cdot)$ and $E_\theta(\cdot \mid \cdot)$ denote expectation and conditional expectation with respect to $f(\cdot; \theta)$ and $f(\cdot \mid \cdot; \theta)$

2. The score at $\theta^*$, satisfies $S_{Y,W,T,X}(\theta^*) = S_{Y|W,T,X}(\theta^*) + S_{W|T,X}(\theta^*) + S_{T|X}(\theta^*) + S_{X}(\theta^*)$, where $S_{Y|W,T,X}(\theta^*)$, $S_{W|T,X}(\theta^*)$, $S_{T|X}(\theta^*)$ and $S_{X}(\theta^*)$ denote the scores in implied parametric submodels for the respective conditional and marginal distributions at $\theta^*$.

3. $\frac{\partial}{\partial \theta} E_\theta^* \{E_\theta(Y \mid X, T = k)\} \bigg|_{\theta^*} = E_\theta^* \{\frac{\partial}{\partial \theta} E_\theta(Y \mid X, T = k)\} \bigg|_{\theta^*}$ and $E_\theta^* \{\frac{\partial}{\partial \theta} E_\theta(Y \mid W, X, T = k)\} \bigg|_{\theta^*} = E_\theta^* \{\frac{\partial}{\partial \theta} E_\theta(Y \mid W, X, T = k)\} \bigg|_{\theta^*}$ for $k = 0, 1$.

4. $E_\theta^* \{E_\theta(Y \mid W, X, T = k)\}^2 \mid X, T = k$ is continuous in $\theta$ at $\theta^*$ for $k = 0, 1$.

5. $E_\theta(Y^2 \mid W, X, T = k)$ is continuous in $\theta$ at $\theta^*$ for $k = 0, 1$.

The efficient influence function for $\bar{\Delta}^* = E\{E(Y \mid X, T = 1) - E(Y \mid X, T = 0)\}$ in $\mathcal{M}_{NP}$ with respect to $\mathcal{P}_{NP}$ is:

$$\Psi_{eff} = E(Y \mid X, T = 1) - E(Y \mid X, T = 0) + \left\{ \frac{I(T = 1)}{\pi_1(X)} - \frac{I(T = 0)}{\pi_0(X)} \right\} \{Y - E(Y \mid X, T)\} - \bar{\Delta}^*.$$

The semiparametric efficiency bound for $\bar{\Delta}^*$ under $\mathcal{M}_{NP}$ with respect to $\mathcal{P}_{NP}$ is $E(\Psi_{eff}^2)$.

Proof. The derivation of the semiparametric efficiency bound for $\bar{\Delta}^*$ under $\mathcal{M}_{NP}$ directly follows arguments from the well-known works of Robins et al. (1994) and Hahn (1998). It can be shown that the availability of $W$ in our framework does not alter the bound as $\mathcal{M}_{NP}$ is a model for the distribution of data where $Y$ is fully observed. We omit repeating the arguments here for brevity.
2.9.2 Appendix B: Consistency and Asymptotic Linearity of $\hat{\Delta}$

**Theorem 1.** Under the identification assumptions from (2)-(4) of the main text, given a bandwidth of $h = O(N^{-\alpha})$ for $\frac{1-\beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{4})$ and $n = O(N^{1-\beta})$ with $\frac{1}{q+1} < \beta < 1$, $\hat{\Delta} - \Delta = O_p(n^{-1/2})$ when either $\pi_1(x; \alpha)$ or $\mu_k(x; \beta)$ is correctly specified.

**Proof.** We first show that $\hat{\Delta} - \Delta = O_p(n^{-1/2})$ where $\Delta = \mu_1 - \mu_0$. If this can be shown, the limiting estimate is:

$$\Delta = \mathbb{E} \left\{ \frac{I(T = 1)Y}{\pi_1(X; \theta)} - \frac{I(T = 0)Y}{\pi_0(X; \theta)} \right\} = \Delta,$$

where the first equality follows from the argument in the main text and the second equality holds when either $\pi_1(x; \alpha)$ or $\mu_k(x; \beta)$ are correctly specified, as in Chapter 1. It suffices to show that $\hat{\mu}_k - \mu_k = O_p(n^{-1/2})$, for $k = 0, 1$. First note that the normalizing constant can effectively be ignored. By application of Lemma 2 with $\zeta_i = 1$, the normalizing constant is:

$$N^{-1} \sum_{i=1}^N \omega_{k,i} = 1 + O_p(c_n). \tag{2.24}$$

We can now write the standardized mean for the $k$-th group as:

$$n^{1/2}(\hat{\mu}_k - \mu_k) = \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(\hat{Y}^\dagger_i - \mu_k) + \left\{ N^{-1} \sum_{i=1}^N \omega_{k,i} \right\}^{-1} - 1 \left[ \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(\hat{Y}^\dagger_i - \mu_k) \right]$$

$$= \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(\hat{Y}^\dagger_i - \mu_k) + O_p(c_n), \tag{2.25}$$

where the last equality follows provided that the main term is $O_p(1)$. Denote:

$$\hat{W}_k = \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(\hat{Y}^\dagger_i - \mu_k).$$

This can be decomposed as $\hat{W}_k = \hat{W}_{1,k} + \hat{W}_{2,k}$, where:

$$\hat{W}_{1,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(Y^\dagger_i - \mu_k) \text{ and } \hat{W}_{2,k} = \frac{\sqrt{n}}{N} \sum_{i=1}^N \omega_{k,i}(\hat{Y}^\dagger_i - Y^\dagger_i). \tag{2.26}$$

First, focusing on $\hat{W}_{2,k}$, we expand $\hat{Y}^\dagger_i = \xi T_i(V_i; \hat{\gamma}, \hat{\pi})$ around $Y^\dagger_i = \xi T_i(V_i; \gamma, \pi)$:

$$\hat{W}_{2,k} = \hat{W}_{2,k}^\pi + \hat{W}_{2,k}^\gamma + O_p \left\{ \sup_x \left\| \hat{\pi}(x; \hat{\theta}) - \pi(x; \hat{\theta}) \right\|^2 \right\} + O_p(\|\hat{\gamma} - \gamma\|^2), \tag{2.27}$$

60
Lemma 1 and Lemma 2 taking

where

bounded:

again taking

accounts for estimating the DiPS in the imputation with

\[ \dot{g}_x(\tilde{\gamma}^T Z_{x,i}) = -\frac{I(T_i=1)}{\pi_k(x_i; \theta)^2} \] and \( \dot{g}_x(u) = dg_x(u)/du \) and

\[
\hat{\mathcal{W}}_{2,k} = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \partial \xi_{T_i}(V_i; \hat{\gamma}, \pi) \{ \hat{\pi}_k(X_i; \hat{\theta}) - \hat{\pi}_k(X_i; \hat{\theta}) \}.
\]

accounts for estimating \( \gamma \) in the imputation. We can further decompose \( \hat{\mathcal{W}}_{2,k} = \hat{\mathcal{W}}_{2,k}^{\text{pa}} + \hat{\mathcal{W}}_{2,k}^{\text{np}} \), where:

\[
\hat{\mathcal{W}}_{2,k}^{\text{pa}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \partial \xi_{T_i}(V_i; \hat{\gamma}, \pi) \{ \hat{\pi}_k(X_i; \hat{\theta}) - \hat{\pi}_k(X_i; \hat{\theta}) \}
\]

\[
\hat{\mathcal{W}}_{2,k}^{\text{np}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \partial \xi_{T_i}(V_i; \hat{\gamma}, \pi) \{ \hat{\pi}_k(X_i; \alpha_1, \beta_1) - \hat{\pi}_k(X_i; \hat{\theta}) \}.
\]

The \( \hat{\mathcal{W}}_{2,k}^{\text{np}} \) term accounts for the parametric estimation of \( \alpha_1 \) and \( \beta_1 \) in DiPS and is:

\[
\hat{\mathcal{W}}_{2,k}^{\text{np}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_{k,i} \partial \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \frac{\partial}{\partial \alpha_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \beta_1) (\hat{\alpha} - \alpha) 
\right.
\]

\[
+ \frac{\partial}{\partial \beta_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \beta_1) (\hat{\beta} - \beta) + O_p(||\hat{\alpha} - \alpha||^2 + ||\hat{\beta} - \beta||^2 + ||\hat{\alpha} - \alpha|| ||\hat{\beta} - \beta||) \right\}
\]

\[
= N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \partial \xi_{T_i}(V_i; \hat{\gamma}, \pi) \left\{ \frac{\partial}{\partial \alpha_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \beta_1) n^{1/2} (\hat{\alpha} - \alpha) + \frac{\partial}{\partial \beta_1} \hat{\pi}_k(X_i; \hat{\alpha}_1, \beta_1) n^{1/2} (\hat{\beta} - \beta) \right\}
\]

\[
+ O_p(1 + c_n)O_p(\hat{b}_N)O_p(\nu_n^{1/2}) + O_p(1 + c_n)O_p(\hat{b}_N)O_p(1)
\]

\[
= O_p(1 + c_n) + O_p(\hat{b}_N),
\]

where the first equality uses the Lipshitz continuity of \( K(\cdot) \), the second equality applies Lemma 1 and Lemma 2 taking \( \zeta_i = \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \), and the last equality applies lemma 2 again taking \( \zeta_i = \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \frac{\partial}{\partial \alpha_1} \pi(X_i; \hat{\theta}) \) as well as \( \zeta_i = \frac{\partial}{\partial \beta} \xi_{T_i}(V_i; \hat{\gamma}, \pi) \frac{\partial}{\partial \beta_1} \pi(X_i; \hat{\theta}) \).

The \( \hat{\mathcal{W}}_{2,k}^{\text{np}} \) term accounts for the nonparametric smoothing in DiPS and can be bounded:

\[
\hat{\mathcal{W}}_{2,k}^{\text{np}} \leq n^{1/2} \sup_{x} \| \hat{\pi}_k(x; \hat{\theta}) - \hat{\pi}_k(x; \hat{\theta}) \| N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i}(V_i; \hat{\gamma}, \pi)
\]

(2.32)
\[
\text{where }
\mu = O_p(1 + c_n) = O_p(n^{1/2} \tilde{a}_N).
\]

Returning to (2.28), the following term accounts for the parametric estimation of \( \gamma \):

\[
\mathcal{W}_{2,k}^\gamma = N^{-1} \sum_{i=1}^{N} \hat{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi (V_i; \hat{\gamma}, \pi)n^{1/2}(\hat{\gamma} - \gamma) + O_p(1 + c_n)O_p(a_n)O_p(1)
\]

\[
= O_p(1 + c_n)O_p(1) + O_p(a_n) = O_p(1 + c_n + a_n),
\]

where the first equality applies Lemma 2 taking \( \zeta_i = \frac{\partial}{\partial \gamma} \xi (V_i; \hat{\gamma}, \pi) \) as well as Lemma 1, and the second equality follows from application of Lemma 2 again taking \( \zeta_i = \frac{\partial}{\partial \gamma} \xi (V_i; \hat{\gamma}, \pi) \). Finally, collecting all the terms, we find:

\[
n^{1/2}(\hat{\mu}_k - \bar{\mu}_k) = \mathcal{W}_{1,k} + \mathcal{W}_{2,k}^{\text{pa, } \pi} + \mathcal{W}_{2,k}^{\text{mp, } \pi} + \mathcal{W}_{2,k}^{\gamma} + O_p(a_n^2 + n^{-1}) + O_p(c_n)
\]

\[
= O_p(1 + d_n) + O_p(1 + c_n + \tilde{b}_N) + O_p(n^{1/2} \tilde{a}_N) + O_p(1 + c_n + a_n)
\]

\[
+ O_p(a_n^2 + n^{-1}) + O_p(c_n)
\]

\[
= O_p(1),
\]

where the second to last equality applies Lemma 3 and the last equality follows when \( h = O(N^{-\alpha}) \) for \( \frac{1 - \beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{4}) \) and \( n = O(N^{1-\beta}) \) for \( \frac{1}{q+1} < \beta < 1 \). This shows that \( \hat{\mu}_k - \bar{\mu}_k = O_p(n^{-1/2}) \).

**Theorem 2.** Let \( \mathcal{W}_k = n^{1/2}(\hat{\mu}_k - \bar{\mu}_k) \) for \( k = 0, 1 \) so that \( n^{1/2}(\hat{\Delta} - \Delta) = \mathcal{W}_1 - \mathcal{W}_0 \). Given a bandwidth of \( h = O(N^{-\alpha}) \) for \( \frac{1 - \beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{4}) \) and \( n = O(N^{1-\beta}) \) with \( \frac{1}{q+1} < \beta < 1 \), then \( \mathcal{W}_k \) has the influence function expansion of the form:

\[
\mathcal{W}_k = n^{-1/2} \sum_{i=1}^{n} (v_{\beta_{1,k}}^T + u_{\text{pa, } \pi, k}^T) \varphi_{\beta_{1,i}} + u_{\gamma, k}^T \varphi_{\gamma,i} + o_p(1),
\]

where \( v_{\beta_{1,k}} = 0 \) when \( \pi_1(x; \alpha) \) is correctly specified and \( u_{\text{pa, } \pi, k} = 0 \) when either \( \pi_1(x; \alpha) \) or \( \xi_k(v; \gamma, \pi) \) without the utility covariate is correctly specified.

**Proof.** As in (2.25) and (2.26) the standardized mean for the \( k \)-th group can be written:

\[
\mathcal{W}_k = \mathcal{W}_k + O_p(c_n) = \mathcal{W}_{1,k} + \mathcal{W}_{2,k} + O_p(c_n),
\]

where the first equality follows provided \( \mathcal{W}_k = O_p(1) \). The first term can be written:

\[
\mathcal{W}_{1,k} = \nu_n^{1/2} \sum_{i=1}^{N} \hat{\omega}_{k,i} (Y_i^T - \bar{\mu}_k)
\]

\[62\]
\[
\begin{align*}
&= \nu_n^{1/2} \tilde{W}_{1,1,k} + \nu_n^{1/2} \tilde{W}_{2,1,k}^d + O_p \left\{ \nu_n^{1/2} \left( h^q + N^{1/2} h^q + N^{-1/2} h^{-2} + N^{1/2} \tilde{a}_N \right) \right\} \\
&+ \left\{ \nu_{\beta_1,k} + O_p \left( N^{-1/2} h^{-1} + N^{-1} h^{-3} \right) \right\} n^{1/2} \left( \tilde{\beta}_1 - \beta_1 \right) + O_p \left\{ \nu_n^{1/2} \left( 1 + N^{-1/2} h^{-1} + N^{-1} h^{-3} \right) \right\},
\end{align*}
\]

where:

\[
\begin{align*}
\tilde{W}_{1,1,k} &= N^{-1/2} \sum_{i=1}^{N} \tilde{\omega}_{k,i} (Y_i^\dagger - \bar{\mu}_k) \\
\tilde{W}_{2,1,k}^d &= -N^{-1/2} \sum_{i=1}^{N} E(Y^\dagger | S, T = k) (\tilde{\omega}_{k,i} - 1) + O_p (h^q + N^{-1/2} h^{-2}) \\
v_{\beta_1,k} &= E \left\{ \hat{K}_h \left( \frac{S_j - \bar{S}_i}{h} \right)^T (1 - \tilde{\omega}_{k,i}) \frac{l_k(X_i; \tilde{\theta})}{\bar{\mu}_k} (X_j^\dagger - X_i^\dagger)^T \right\},
\end{align*}
\]

with \( l_k(x; \tilde{\theta}) = \pi_k(x; \tilde{\theta}) f(x; \tilde{\theta}) \) and \( f(x; \tilde{\theta}) \) being the joint density of \( S \), and \( x^\dagger = (x, 0) \) for any vector \( x \). Here \( v_{\beta_1,k} = O_p(1) \) in general and is 0 when \( \pi(X; \alpha) \) is correctly specified, as in Chapter 1. As in (2.27) and (2.30) of Theorem 1, the second term from (2.37) can be written:

\[
\tilde{W}_{2,k} = \tilde{W}_{2,k}^{pa,\pi} + \tilde{W}_{2,k}^{np,\pi} + \tilde{W}_{2,k}^{\gamma} + O_p (a_n^2 + n^{-1}).
\]

Continuing the expansion of \( \tilde{W}_{2,k}^{pa,\pi} \) from (2.31):

\[
\begin{align*}
\tilde{W}_{2,k}^{pa,\pi} &= N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i} (V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1} \pi_k (X_i; \tilde{\theta}) n^{1/2} (\tilde{\beta}_1 - \beta_1) + O_p \left\{ (1 + c_n) \left( \nu_n^{1/2} + \tilde{b}_N \right) \right\} \\
&= E \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i} (V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1} \pi_k (X_i; \tilde{\theta}) \right\} n^{1/2} (\tilde{\beta}_1 - \beta_1) + O_p \left\{ c_n + (1 + c_n) \left( \nu_n^{1/2} + \tilde{b}_N \right) \right\} \\
&= n^{-1/2} \sum_{i=1}^{n} u_{pa,\pi,k}^T \varphi_{\beta_1,i} + o_p (1) + O_p (c_n + \tilde{b}_N),
\end{align*}
\]

by repeated application of Lemma 2, where:

\[
u_{pa,\pi,k}^T = E \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \pi} \xi_{T_i} (V_i; \gamma, \pi) \frac{\partial}{\partial \beta_1} \pi_k (X_i; \tilde{\theta}) \right\}
\]

is a constant that is \( 0^T \) when either \( \pi_k(x; \alpha) \) or \( \xi_k(v; \gamma, \pi) \) without the utility covariate is correctly specified and \( \varphi_{\beta_1,i} \) is the influence function for \( \tilde{\beta} \). For \( \tilde{W}_{2,k}^{np,\pi} \) from (2.32) we have that \( \tilde{W}_{2,k}^{np,\pi} = O_p (n^{1/2} \tilde{a}_N) \). For \( \tilde{W}_{2,k}^{\gamma} \), continuing from (2.34):

\[
\tilde{W}_{2,k}^{\gamma} = N^{-1} \sum_{i=1}^{N} \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_{T_i} (V_i; \tilde{\gamma}, \pi) n^{1/2} (\tilde{\gamma} - \gamma) + O_p (a_n)
\]

63
\[
\mathbb{E} \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_{T_i}(V_i; \gamma, \pi) \right\} n^{1/2} (\check{\gamma} - \gamma) + O_p(c_n)O_p(1) + O_p(a_n)
\]
\[
= n^{1/2} \sum_{i=1}^{n} u_{\gamma,i}^T \varphi_{\gamma,i} + o_p(1) + O_p(c_n + a_n),
\]
where:
\[
u_n = \mathbb{E} \left\{ \tilde{\omega}_{k,i} \frac{\partial}{\partial \gamma} \xi_{T_i}(V_i; \gamma, \pi) \right\}
\]
is some constant and \( \varphi_\gamma \) is the influence function for \( \check{\gamma} \).

Collecting the results from above, we find:
\[
\hat{W}_k = n^{1/2} \sum_{i=1}^{n} (v_{\beta,i}^T + u_{\psi,x,i}^T) \varphi_{\beta,i} + u_{\gamma,i}^T \varphi_{\gamma,i} + O_p \left\{ \nu_n^{1/2} \left( N^{-1/2} h^3 + N^{-1/2} h^{-2} \right) \right\}
\]
\[
+ O_p(N^{-1/2} h^{-1} + N^{-1} h^{-3} + \tilde{b}_N + n^{1/2} a_N + c_n + a_n)
\]
\[
= n^{1/2} \sum_{i=1}^{n} (v_{\beta,i}^T + u_{\psi,x,i}^T) \varphi_{\beta,i} + u_{\gamma,i}^T \varphi_{\gamma,i} + O_p \left\{ \nu_n^{1/2} N^{1/2} h^3 + \frac{(\log N)^{1/2}}{N^{1/2} h^2} + \nu_n^{1/2} (\log N)^{1/2} \right\}.
\]
The error terms are \( o_p(1) \) when \( h = O(N^{-\alpha}) \) for \( \frac{1-\beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{4}) \) and \( n = O(N^{1-\beta}) \) for \( \frac{1}{q+1} < \beta < 1 \).

### 2.9.3 Appendix C: Semiparametric Efficiency

**Theorem 3.** Let \( \mathcal{M}_{SS} = \{ f_{Y|Z}(y, z) = f_{Y|z}(y \mid z) f_{Z}^2(z) : f_{Z}^2(z) \text{ is a known density such that there exists a } \epsilon_\pi > 0 \text{ where } \pi_1^*(x) \in [\epsilon_\pi, 1 - \epsilon_\pi] \text{ for all } x \text{ with } f_{X}(x) > 0 \} \) be an ideal semiparametric semi-supervised model where the distribution of \( Z = (V^T, T)^T \) is known, with \( z = (v^T, t)^T \) and \( \pi_1^*(x) \) and \( f_{Z}^2(x) \) being the implied PS and density of \( X \) under \( f_{Z}^2(z) \). Let \( \mathcal{M}_{SS, sub} = \{ f_{Y|Z}(y, z; \theta) f_{Z}^2(z) : \theta \in \Theta \} \) denote a regular parametric submodel of \( \mathcal{M}_{SS} \), where \( \theta \) is a finite-dimensional parameter, and the true density is at \( \theta = \theta^* \). Let \( \mathcal{P}_SS \subseteq \mathcal{P}_{NP} \) be the subcollection of all such regular parametric submodels among \( \mathcal{P}_{NP} \). The efficient influence function for \( \Delta^* \) in \( \mathcal{M}_{SS} \) with respect to \( \mathcal{P}_{SS} \) is:

\[
\varphi_{eff} = \left\{ \frac{I(T = 1)}{\pi_1(X)} - \frac{I(T = 0)}{\pi_0(X)} \right\} \{Y - \xi_T(V)\},
\]
\[
(2.39)
\]
and the semiparametric efficiency bound for \( \Delta^* \) under \( \mathcal{M}_{SS} \) with respect to \( \mathcal{P}_{SS} \) is \( E(\varphi^2_{eff}) \). Furthermore, the efficiency bound under \( \mathcal{M}_{SS} \) is lower than or equal to the efficiency bound under the fully nonparametric model \( \mathcal{M}_{NP} \) where the distribution of \( Z \) is unknown. That is,

\[
\mathbb{E}(\varphi^2_{eff}) \leq \mathbb{E}(\Psi^2_{eff}).
\]

64
Proof. Let \( \mathcal{L}_2^0 \) denote the Hilbert space of mean 0 square-integrable functions of \((Y, Z^T)^T\) at the true distribution, with inner product of \( v_1, v_2 \in \mathcal{L}_2^0 \) defined by \( \langle v_1(Y, Z), v_2(Y, Z) \rangle = \mathbb{E}\{v_1(Y, Z)v_2(Y, Z)\} \). We first show that the tangent space of \( \mathcal{M}_{SS} \) with respect to \( \mathcal{P}_{SS} \) at the true distribution is the closure of \( \{s(Y, Z) \in \mathcal{L}_2^0 : \mathbb{E}\{s(Y, Z) | Z\} = 0\} \), denoted by \( \Lambda_{SS} \).

Let \( s(Y, Z) \) be any bounded element belonging to \( \Lambda_{SS} \). Consider the parametric submodel given by \( \mathcal{M}_{SS,ilt} = \{f_{Y|Z}(y | z; \theta) = f_{Y|Z}(y | z) f_z^*(z) : \theta \in (-\varepsilon, \varepsilon)\} \) for some sufficiently small \( \varepsilon > 0 \), where:

\[
f_{Y|Z}(y | z; \theta) = f_{Y|Z}(y | z)\{1 + \theta s(y, z)\}
\]

with \( f_{Y|Z}(y | z) \) being the true density. The true density is thus at \( \theta = \theta^* = 0 \). It can be shown that \( \mathcal{M}_{SS,ilt} \) and the implied conditional and marginal submodels have proper densities, are regular, and the respective score can be written as the derivative of the log density with respect to \( \theta \). It can also be shown through calculations similar to those in analogous arguments for the derivation of Lemma 4 that \( \mathcal{M}_{SS,ilt} \) belongs in \( \mathcal{P}_{SS} \subseteq \mathcal{P}_{NP} \).

The score for \( \mathcal{M}_{SS,ilt} \) at \( \theta = \theta^* = 0 \) is \( S_{Y,Z}(\theta^*) = s(Y, Z) \), so any bounded element in \( \Lambda_{SS} \) belongs in the tangent space of \( \mathcal{M}_{SS} \) with respect to \( \mathcal{P}_{SS} \) at the true distribution. Since the bounded elements are dense in \( \Lambda_{SS} \) and the tangent space is closed, any element \( r(Y, W, T, X) \in \Lambda_{SS} \) also belongs in the tangent space. Any element of the tangent space at the true distribution also belongs in \( \Lambda_{SS} \) by the regularity of the parametric submodels and properties of scores. This verifies that the tangent space of \( \mathcal{M}_{SS} \) with respect to \( \mathcal{P}_{SS} \) at the true distribution is \( \Lambda_{SS} \).

We next show that \( \Psi_{eff} \) is one influence function for \( \Delta^* \) in \( \mathcal{M}_{SS} \) at the true distribution with respect to \( \mathcal{P}_{SS} \). Recall from Lemma 4 that \( \Psi_{eff} \) is the unique influence function for \( \Delta^* \) in \( \mathcal{M}_{NP} \) with respect to \( \mathcal{P}_{NP} \). This means that under any regular parametric submodel \( \mathcal{M}_{NP,sub} \) belonging to \( \mathcal{P}_{NP} \), \( \Psi_{eff} \) satisfies:

\[
\frac{\partial}{\partial \theta} \Delta^*(\theta) \bigg|_{\theta^*} = \mathbb{E}_{\theta^*}\{\Psi_{eff}S_{Y,W,T,X}(\theta^*)\}.
\]

Now since \( \mathcal{P}_{SS} \subseteq \mathcal{P}_{NP} \), pathwise differentiability of \( \Delta^*(\theta) \) at \( \theta = \theta^* \) also holds, in particular, under any regular parametric submodel in \( \mathcal{P}_{SS} \), with \( \Psi_{eff} \) being one influence function.
Finally, to obtain the efficient influence function for $\bar{\Delta}^*$ in $\mathcal{M}_{SS}$ with respect to $\mathcal{P}_{SS}$ at the true distribution, we identify the orthogonal projection of $\Psi_{eff}$ onto $\Lambda_{SS}$. It can be verified that this projection is $\Pi(\Psi_{eff} \mid \Lambda_{SS}) = \Psi_{eff} - \mathbb{E}(\Psi_{eff} \mid Z)$. The efficient influence function in $\mathcal{M}_{SS}$ is thus:

$$\varphi_{eff} = \Pi(\Psi_{eff} \mid \Lambda_{SS}) = \Psi_{eff} - \mathbb{E}(\Psi_{eff} \mid Z)$$

$$= (\mathbb{E}(Y \mid X, T = 1) + \left\{ \frac{I(T = 1)}{\pi_1(X)} \{ Y - \mathbb{E}(Y \mid X, T = 1) \} \right\})$$

$$- (\mathbb{E}(Y \mid X, T = 0) + \left\{ \frac{I(T = 0)}{\pi_0(X)} \{ Y - \mathbb{E}(Y \mid X, T = 0) \} \right\}) - \bar{\Delta}^*$$

$$- (\mathbb{E}(Y \mid X, T = 1) + \left\{ \frac{I(T = 1)}{\pi_1(X)} \{ \mathbb{E}(Y \mid Z) - \mathbb{E}(Y \mid X, T = 1) \} \right\})$$

$$+ (\mathbb{E}(Y \mid X, T = 0) + \left\{ \frac{I(T = 0)}{\pi_0(X)} \{ \mathbb{E}(Y \mid Z) - \mathbb{E}(Y \mid X, T = 0) \} \right\}) + \bar{\Delta}^*$$

$$= \left\{ \frac{I(T = 1)}{\pi_1(X)} - \frac{I(T = 0)}{\pi_0(X)} \right\} \{ Y - \mathbb{E}(Y \mid Z) \}.$$  

By the Pythagorean theorem, we can verify:

$$\mathbb{E}(\Psi^2_{eff}) = \| \Psi_{eff} \|_{L^2_0}^2 = \| \Pi(\Psi_{eff} \mid \Lambda_{SS}) \|_{L^2_0}^2 + \| \Psi_{eff} - \Pi(\Psi_{eff} \mid \Lambda_{SS}) \|_{L^2_0}^2$$

$$\geq \| \Pi(\Psi_{eff} \mid \Lambda_{SS}) \|_{L^2_0}^2$$

$$= \mathbb{E}(\varphi_{eff}^2).$$

\[ \square \]

**Corollary 1.** Given a bandwidth of $h = O(N^{-\alpha})$ for $\frac{1-\beta}{2q} < \alpha < \min(\frac{\beta}{2}, \frac{1}{2})$ and $n = O(N^{1-\beta})$ for $\frac{1}{q+1} < \beta < 1$, when $\pi_1(x; \alpha)$ and $\xi_k(v; \gamma, \pi)$ are correctly specified, then:

$$n^{1/2}(\hat{\Delta} - \bar{\Delta}^*) = n^{-1/2} \sum_{i=1}^{n} U_i \{ Y_i - \xi_{T_i}(V_i) \} + o_p(1).$$ (2.40)

That is, $\hat{\Delta}$ achieves the semiparametric efficiency bound in the ideal SS semiparametric model where the distribution of $Z$ is known.

**Proof.** From Theorem 2, given an appropriate bandwidth and order of labels, when $\pi_1(x; \alpha)$ is correctly specified:

$$n^{1/2}(\hat{\Delta} - \bar{\Delta}^*) = n^{-1/2} \sum_{i=1}^{n} (u^T_{\gamma,1} - u^T_{\gamma,0}) \varphi_{\gamma,i} + o_p(1),$$
where:

$$u_{\gamma,k}^T \varphi_{\gamma,i} = E \left\{ \omega_{k,i} \frac{\partial}{\partial \gamma^T} \xi_i (V_i; \gamma, \pi) \bigg| \gamma = \bar{\gamma} \right\} \left\{ \frac{\partial}{\partial \gamma^T} E \left[ Z_{\pi,i} \xi_i (V_i; \gamma, \pi) \bigg| \gamma = \bar{\gamma} \right] \right\}^{-1} Z_{\pi,i} \{ Y_i - \xi_{T,i} (V_i; \bar{\gamma}, \pi) \}$$

$$= E \left\{ \omega_{k,i} Z_{\pi,i}^T \hat{g} (\gamma^T Z_{\pi,i}) \right\} E \left\{ Z_{\pi,i} Z_{\pi,i}^T \hat{g} (\gamma^T Z_{\pi,i}) \right\}^{-1} Z_{\pi,i} \{ Y_i - \xi_{T,i} (V_i; \bar{\gamma}, \pi) \} .$$

The first and second equalities assume the usual regularity conditions to obtain the influence function of an estimator that is the solution of an estimating equation and exchange order of differentiation and integration. The influence function for $\hat{\Delta}$ can then be written as:

$$(u_{\gamma,1}^T - u_{\gamma,0}^T) \varphi_{\gamma,i} = E \left\{ U_{\pi,i} Z_{\pi,i}^T \hat{g} (\gamma^T Z_{\pi,i}) \right\} E \left\{ Z_{\pi,i} Z_{\pi,i}^T \hat{g} (\gamma^T Z_{\pi,i}) \right\}^{-1} Z_{\pi,i} \{ Y_i - \xi_{T,i} (V_i; \bar{\gamma}, \pi) \} .$$

The terms involving $Z_{\pi,i}$ is a population weighted least square projection of $U_i$ onto $Z_i$, weighted by $\hat{g} (\gamma^T Z_{\pi,i})$. But since $Z_{\pi,i}$ includes $U_{\pi,i}$, the influence function simplifies:

$$(u_{\gamma,1}^T - u_{\gamma,0}^T) \varphi_{\gamma,i} = U_{\pi,i} \{ Y_i - \xi_{T,i} (V_i; \bar{\gamma}, \pi) \} = U_{\pi,i} \{ Y_i - \xi_{T,i} (V_i) \} ,$$

where the second equality follows when $\xi_k (v; \gamma, \pi)$ is correctly specified. $\square$
Identifying Treatment Subgroups by Tree Approximations of Conditional Treatment Effects

David Cheng
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Lu Tian
Department of Biomedical Data Science
Stanford University

Tianxi Cai
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Hajime Uno
Division of Population Sciences
Dana-Farber Cancer Institute

Lee-Jen Wei
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Fabio Pellegrini
Biogen International
3.1 Introduction

There is wide interest in identifying patient subgroups with enhanced treatment effects in clinical trial data. It is now well-appreciated that naive post hoc subgroup analyses face substantial pitfalls that lead to inflated type I error and low power (Assmann et al., 2000; Rothwell, 2005; Wang et al., 2007). A variety statistical procedures have since been proposed to identify subgroups defined by combinations of multiple baseline covariates to improve power. Typically, methods involve modeling treatment effect heterogeneity, for example by regression (Cai et al., 2010; Qian and Murphy, 2011; Imai et al., 2013; Lu et al., 2013; Tian et al., 2014) or recursive-partitioning (Negassa et al., 2005; Su et al., 2011; Dusseldorp and Van Mechelen, 2014; Loh et al., 2015; Athey and Imbens, 2016), and thresholding the estimated treatment effects. Validating the value of identified subgroups is also crucial in practice (Zhao et al., 2013; Matsouaka et al., 2014; Li et al., 2016). A closely related literature on estimating optimal treatment regimes by value function maximization, which can be viewed as identifying subgroups for whom a treatment benefits, have also been rapidly developing (Zhang et al., 2012; Zhao et al., 2012, 2015). A more comprehensive review of the subject can be found in Lipkovich et al. (2017) and Sies and Van Mechelen (2017).

Recently, there is growing recognition that interpretability is a crucial feature when implementing these procedures in practice (Zhang et al., 2015; Lakkaraju and Rudin, 2017). Otherwise, procedures that identify a subgroup or a treatment regime without being transparent in its reasoning may make it difficult for clinicians and practitioners to appreciate their value. For example, let the baseline covariates of a patient be denoted by $X$ and suppose that their inclusion in the subgroup is defined by thresholding a linear combination score, as in $I(X^T\beta > s)$. A clinician may have a sense of what patients are more likely to be included by examining the coefficients $\beta$, but it is not immediately intuitive which patients are finally included in the subgroup. This issue in treatment selection for medical applications follows a broader trend in machine learning that concerns making predictions from black-box prediction models interpretable (Doshi-Velez and Kim, 2017). For this reason, methods based on recursive-partitioning have offered an attractive
approach for modeling treatment effect heterogeneity and identifying subgroups. However, methods based on modeling data directly with recursive partitioning can be highly variable. Small perturbations in the data can produce very different tree structures, which can potentially compromise the predictive accuracy of tree-based approaches (Petersen et al., 2016).

“Born again” or “representer” trees were proposed as a method for regression modeling, without regard to treatment effects, that seeks to balance interpretability and accuracy (Breiman and Shang, 1996). The general approach is to first fit a regression model \( \hat{\mu}(X) \) that predicts an outcome \( Y \) given \( X \). Additional covariate data \( X^* \) are then simulated from the distribution of \( X \) and used to generate predicted values \( \hat{\mu}(X^*) \). A tree can then be fit based on replicates of \((X^*, \hat{\mu}(X^*))\), simulating as much data as needed to accurately represent \( \hat{\mu} \). Provided that the original \( \hat{\mu} \) is accurate in predicting \( Y \), this “re-approximation” can also be accurate while maintaining the interpretability of a tree. In the medical domain, this has been applied for diagnosing depression (Gibbons et al., 2013). Recently, an active sampling strategy to simulate \( X^* \) was proposed to improve fidelity in approximating \( \hat{\mu} \) for a given size of the simulated data (Bastani et al., 2017). The broad strategy has also been adopted for subgroup identification by predicting outcomes given \( X \) for each treatment and then fitting a tree to model the predicted differences (Foster et al., 2011). However, they do not consider simulating \( X^* \) to improve the fidelity of the tree in approximating \( \hat{\mu} \). They also do not consider the issues of selecting the threshold for identifying the final subgroup and validating the final selected subgroup. In this paper, we consider more fully adopting the framework of Breiman and Shang (1996) for estimating conditional treatment effects. We consider a “cross-training-evaluation” procedure to select the threshold parameter for identifying the subgroup and validating it on out-of-sample data (Zhao et al., 2013). We apply the method to identify multiple sclerosis (MS) patients who experience reduced rates of 2-year relapse when treated with oral fumarate (BG-12).
3.2 Re-Approximation of Conditional Treatment Effects

3.2.1 Tree Approximation and Subgroup Definition

From a clinical trial, suppose that we observe data $D = \{(X_i^T, T_i, Y_i)^T : i = 1, \ldots, n\}$ from $n$ patients, where $X_i$ denotes baseline covariates of length $p$ taking values in covariate space $\mathcal{X}$, $T_i$ a randomized binary treatment, and $Y_i$ some outcome of interest. Without loss of generality, to mimic the data in our analysis, we assume that $Y$ is a detrimental outcome and that the active treatment, labeled as $T = 1$, aims to reduce $Y$. We are interested in approximating the treatment effect given $X$, $\Delta(X) = \mathbb{E}(Y \mid X, T = 1) - \mathbb{E}(Y \mid X, T = 0)$, using piecewise constant functions over partitions of $\mathcal{X}$ of the form:

$$\mathcal{T} = \left\{ h(X) = \sum_{\ell=1}^{L} I(X \in \mathcal{R}_\ell) \Delta_\ell : \mathcal{R}_\ell \cap \mathcal{R}_{\ell'} = \emptyset \text{ for } \ell \neq \ell', \bigcup_{\ell=1}^{L} \mathcal{R}_\ell = \mathcal{X} \right\},$$

where $L$ is the number of terminal leaves that could range from 1 up to some pre-specified bound, $\mathcal{R}_\ell \subseteq \mathcal{X}$ a rectangular region formed by recursive partitioning representing the $\ell$-th terminal leaf, and $\Delta_\ell = \mathbb{E}(Y \mid X \in \mathcal{R}_\ell, T = 1) - \mathbb{E}(Y \mid X \in \mathcal{R}_\ell, T = 0)$ the treatment effect among patients with $X \in \mathcal{R}_\ell$. By recursive partitioning we mean that $\mathcal{R}_\ell$ are identified by a recursion, where all of the data are initially split into two rectangular regions of the form $\{X : X_j \leq c\}$ and $\{X : X_j > c\}$, by evaluating some splitting criteria over all $(j, c)$. Each of these two rectangular regions are then further split into subset rectangular regions by evaluating the same criteria, and so on until some stopping rule is met. Informally, $\mathcal{T}$ can be viewed as a statistical model indexed by parameters $\theta = \{(\mathcal{R}_\ell, \Delta_\ell)\}_{\ell=1}^{L}$.

Rectangles constructed by recursive partitioning can be visualized as binary trees, making clear the reasoning for estimating the effect as $h(X)$ for a patient with given covariates $X$. Now, had the true $\Delta(X)$ been known, thresholding $\Delta(X)$ yields the Bayes rule for maximizing a cost-adjusted population average outcome (Matsouaka et al., 2014). Since $\Delta(X)$ is approximated by some $h(X) \in \mathcal{T}$, this suggests also thresholding $h(X)$ to define the final subgroup, which is equivalent to grouping the terminal regions or leaves $\mathcal{R}_\ell$ with treatment effects $\Delta_\ell$ less than some threshold $\delta$:

$$\mathcal{S} = \bigcup_{\ell: \Delta_\ell < \delta} \mathcal{R}_\ell.$$
Such an approach puts no constraints on which terminal leaves are grouped together and allows leaves that are far apart in $\mathcal{X}$ to be grouped together. This affords flexibility to the subgroup definition but also potentially increases the variability.

### 3.2.2 Estimation

To estimate $\theta$ and identify $S$, we split the data $\mathcal{D}$ into a training-test set $\mathcal{D}_{tts}$ and an validation set $\mathcal{D}_{val}$, leaving aside $\mathcal{D}_{val}$ to validate the identified subgroup in the end. Following the approach of Breiman and Shang (1996), we first estimate $\Delta(\mathbf{X})$ by a flexible method in $\mathcal{D}_{tts}$. This can be done, for example, using any regression method (e.g. generalized linear models, random forest, neural network, etc.) of $Y$ on $\mathbf{X}$, separately among $T = 1$ and $T = 0$. Denoting these respectively by $\hat{\mu}_1(\mathbf{X})$ and $\hat{\mu}_0(\mathbf{X})$, $\Delta(\mathbf{X})$ can be estimated as $\hat{\Delta}(\mathbf{X}) = \hat{\mu}_1(\mathbf{X}) - \hat{\mu}_0(\mathbf{X})$. To approximate the function $\hat{\Delta}$ with high fidelity, we simulate a large sample of new $\mathbf{X}^*$ from a mixed normal-binary distribution based on the marginal and correlation parameters estimated from the original $\mathbf{X}$ (Demirtas and Doganay, 2012). Let $\mathcal{D}_{sim} = \{(\mathbf{X}^*_i, \hat{\Delta}(\mathbf{X}^*_i))^T : i = 1, \ldots, N\}$ denote simulated data from this process consisting of an arbitrarily large number of $N$ samples. Among $\mathcal{D}_{sim}$, we estimate $h(\mathbf{X}^*)$ using the standard greedy CART algorithm (Breiman et al., 1984) to regress $\hat{\Delta}(\mathbf{X}^*)$ on $\mathbf{X}^*$ using, for example, Gini impurity as the splitting criteria for binary outcomes. This process yields the estimators $\hat{\theta} = \{(\hat{\mathcal{R}}_\ell, \hat{\Delta}_\ell)\}_{\ell=1}^L$, forming $\hat{h}(\mathbf{X}^*) = \sum_{\ell=1}^L I(\mathbf{X}^* \in \hat{\mathcal{R}}_\ell) \hat{\Delta}_\ell$ where $\hat{\Delta}_\ell = \left| \hat{\mathcal{R}}_\ell \right|^{-1} \sum_{\mathbf{X}_i^* \in \hat{\mathcal{R}}_\ell} \hat{\Delta}(\mathbf{X}^*_i)$. It is known that trees are consistent in its risk to the Bayes risk (Breiman et al., 1984), suggesting that a tree fit this way can approximate the original $\hat{\Delta}(\mathbf{X})$ to an arbitrary level of accuracy for large enough $N$. The tradeoff, however, is that an accurate tree may need to be extremely deep and complex. We consider the performance on a real dataset using mostly common parameter settings for estimating the tree in Section 3.4.

### 3.3 Tuning and Validation

It remains to estimate the threshold $\delta$ to form the final estimated subgroup $\hat{S} = \bigcup_{\ell : \hat{\Delta}_\ell < \delta} \hat{\mathcal{R}}_\ell$. In the cross-training-evaluation scheme, $\delta$ can be regarded as a tuning parameter defining
the final subgroup, which can be selected by cross-validation in $D_{tts}$. Cross-validation is needed in finite samples, as re-substituting the same training data used to estimate $\hat{\Delta}_t$ to also evaluate subgroups defined by $\hat{\Delta}_t$ results in over-optimistic biased evaluations (Zhao et al., 2013). For cross-validation, we partition $D_{tts}$ into $K$ disjoint sets $D_{tts}^{(1)}, \ldots, D_{tts}^{(K)}$. Let $\hat{h}^{(-k)}(X_i) = \sum_{t=1}^{L} I(X_i \in \mathcal{R}_t^{(-k)}) \hat{\Delta}_t^{(-k)}$ be the tree fitted in the large simulated data $D_{sim}^{(-k)} = \{(X_i^T, \hat{\Delta}_t^{(-k)}(X_i^*)) : i = 1, \ldots, N\}$, which has the same $X^*$ but estimated treatment differences based on initial estimates $\hat{\Delta}^{(-k)}(X)$ estimated among $D_{tts} \setminus D_{tts}^{(k)}$. Instead of cross-validating to directly select $\delta$, we consider a quantile scoring approach to form groups of sensible size regardless of the distribution of $\Delta_t$. That is, let $\hat{q}_g = \hat{F}^{-1}(k)(p_g)$ be the empirical quantiles of $\hat{h}^{(-k)}(X_i)$ evaluated over $X_i$ from $D_{tts}^{(k)}$, for $g = 1, \ldots, G$, for some grid of proportions $p_1 = .01, \ldots, p_G = .99$. For each proportion $p_g$, we calculate a scalar evaluation statistic $\hat{T}_g^{(k)}$ among those with $X_i \in \bigcup_t \Delta_t^{(-k)} < \hat{q}_g \hat{\Delta}_t$ in the test set $D_{tts}^{(k)}$. The statistic $\hat{T}_g^{(k)}$ evaluates the treatment effect for patients estimated to be among the “best” (e.g. lowest treatment differences) $p_g$ proportion of patients in $D_{tts}^{(k)}$. One example of $\hat{T}_g^{(k)}$ for binary data could be the chi-square statistic for a two-proportion test:

$$T_g^{(k)} = \frac{(\hat{p}_{g,1}^{(k)} - \hat{p}_{g,0}^{(k)})^2}{\hat{p}_g^{(k)} (1 - \hat{p}_g^{(k)}) (1/n_{g,1} + 1/n_{g,0})},$$

where $n_{g,t}^{(k)} = \sum_{(Y_i, T_i, X_i) \in D_{tts}^{(k)}} I(T_i = t, \hat{h}^{(-k)}(X_i) < \hat{q}_g)$ and $\hat{p}_{g,t}^{(k)} = n_{g,t}^{(k)} - \sum_{(Y_i, T_i, X_i) \in D_{tts}^{(k)}} I(T_i = k, \hat{h}^{(-k)}(X_i) < \hat{q}_g) Y_i$, and $\hat{p}_g^{(k)} = (n_{g,1}^{(k)} - \hat{p}_{g,1}^{(k)})/n_{g,1}^{(k)} + (n_{g,0}^{(k)} - \hat{p}_{g,0}^{(k)})/n_{g,0}^{(k)}$. Let $z_{1-\alpha}$ be the $1 - \alpha$ quantile of a standard normal distribution. Another possibly more interpretable statistic is the upper confidence limit for the difference in proportions:

$$T_g^{(k)} = \hat{p}_{g,1}^{(k)} - \hat{p}_{g,0}^{(k)} + z_{1-\alpha} \left\{ \frac{\hat{p}_{g,1}^{(k)} (1 - \hat{p}_{g,1}^{(k)})}{n_{g,1}^{(k)}} + \hat{p}_{g,0}^{(k)} (1 - \hat{p}_{g,0}^{(k)})/n_{g,0}^{(k)} \right\}^{1/2},$$

which is on the scale of the original treatment effect. Another benefit of using the upper CI is that it readily offers a one-sided level-$\alpha$ test of whether the treatment difference in the candidate grouping is less than some desired level, which can be considered in thresholding for selecting the final subgroup. These statistics encapsulate both the magnitude of treatment effect in the candidate subgroup as well as the degree of uncertainty to give a balanced evaluation. The process is repeated for $k = 1, \ldots, K$, and the results are averaged to yield $\hat{T}_g = K^{-1} \sum_{k=1}^K \hat{T}_g^{(k)}$. The optimal proportion $\hat{p}^{*}$ is the one that either yields
the optimal $\hat{T}_g$ or possibly the largest proportion such that $\hat{T}_g$ exceeds or falls below some predefined level (e.g., if $\hat{T}^{(k)}_g$ are confidence limits). Let $\hat{F}^{-1}(q)$ be the empirical quantiles of $\hat{h}(X_i)$ evaluated over $X_i$ in all of $D_{val}$, with $\hat{q}^* = \hat{F}^{-1}(\hat{p}^*)$. The final subgroup is defined by $\hat{S} = \bigcup_{\ell: \Delta_{\ell} < \hat{q}^*} \mathcal{R}_\ell$.

With the final subgroup identified by $\hat{S}$, we validate the results in $D_{val}$. That is, among data in $D_{val}$, we calculate the treatment differences, confidence interval, and p-value for patients with $X_i \in \hat{S}$. For contrast, we also check the same results among the remaining patients with $X_i \in \mathcal{X} \setminus \hat{S}$. Depending on the specific test statistic and tresholding method implemented, the final subgroup can be expected to have varying performance in $D_{val}$. For example, if maximizing a test statistic was used to select $\hat{p}^*$, it can be expected that the treatment effect among $\hat{S}$ would be the largest among candidate cumulative subgroups, after adjusting for uncertainty due to varying sample size. If the subgroup is selected by finding the largest subgroup with confidence limit under some threshold, the treatment effect can be expected to be below the threshold. Nevertheless, it is possible depending on the data that treatment effects in the complement group $\mathcal{X} \setminus \hat{S}$ are also significant and substantial, although they would be expected to be lower in magnitude than in $\hat{S}$.

### 3.4 Application to MS Data

MS is a neurodegenerative disease of the central nervous system, commonly treated by intramuscular (interferon beta) and subcutaneous injections (glatiramer acetate). BG-12 is an oral formulation of dimethyl fumarate that has been demonstrated to significantly reduce relapse and disability outcomes among patients with relapsing-remitting MS in two large, randomized, placebo-controlled, phase 3 studies (Fox et al., 2012; Gold et al., 2012). Specifically, BG-12 was found to reduce proportion of patients with relapse within 2 years by 12-20% in the overall populations of the two studies. In the following analysis, we pooled data from patients randomized to BG-12 and placebo arms of these two trials and sought to identify subgroups exhibiting enhanced treatment effects of BG-12 for reducing relapse.

Among the full data, we combined patients randomized to twice-daily and thrice-
daily BG-12 groups from the original trials into a single BG-12 group to assess effects of treatment at any dose of BG-12. The covariates considered are listed under Table 3.1, which includes demographics, disease history, and various physical functioning and neuropsychological tests. We considered a binary outcome of any relapse occurring within 2 years of follow-up accompanied by objective neurologic findings. Patients with missing values among any of the covariates or lacking outcome data were excluded in the analysis. Roughly 20% of patients were recorded in the data as being on-study for less than 600 days. We assumed loss-to-follow-up was non-differential between treatment groups and had minimal impact on treatment effect estimates in this analysis. However, this assumption warrants further investigation in future studies. The dataset used for the analysis included $n_1 = 1402$ patients in the BG-12 group and $n_0 = 697$ in the placebo group. Table 3.1 indicates that the majority of patients were female and white, having an average age of 38 and diagnosis duration of 5 years. These baseline characteristics appear to be well-balanced between treatment groups in the pooled data.

To implement our proposed procedure, we randomly split the entire data $D$ into $D_{tts}$ and $D_{val}$ at a 2:1 ratio, with sizes $|D_{tts}| = 1399$ and $|D_{val}| = 700$. For the simulated data, we generated $N = 10000$ samples of $X^*$, using the empirical means and variance-covariance estimates from the observed covariates $X_i$. Throughout, a simple logistic regression that includes only main effects for the covariates $X$ was used to obtain $\hat{\mu}_1(X)$ and $\hat{\mu}_0(X)$ and to construct the initial estimator $\hat{\Delta}(X) = \hat{\mu}_1(X) - \hat{\mu}_0(X)$. We chose logistic regression for simplicity and ease of approximation by trees, although more complex models could be considered. We fit trees throughout using the R package rpart (Therneau et al., 2015) with default settings, except we set the complexity parameter to be $cp=.001$ to increase the size of tree models for better accuracy. For cross-validation, we set $K = 4$ and repeated the cross-validation in $D_{tts}$ for 50 times and averaged over the results to mitigate variability in the results from the random partitioning. We used the chi-square test statistic for $\hat{T}_g^{(k)}$ and selected the proportion $\hat{p}^* = p_g$ that yielded the largest $\hat{T}_g$, over a grid of $G = 20$ approximately equidistant proportions ranging from $p_1 = .01$ to $p_G = .99$. As a reference, we also repeated the procedure without re-approximation by tree and cross-validated cumulative subgroups formed by quantile scoring of $\hat{\Delta}(X)$ directly.
Table 3.1: Baseline covariates of pooled data. Patients randomized to any dose of BG-12 from the original trials were combined into one group. \( p \)-values test for difference between the treatment groups using chi-square test for binary covariates t-test for continuous covariates (Yoshida and Bohn, 2015).

<table>
<thead>
<tr>
<th>Baseline Covariate</th>
<th>BG-12</th>
<th>PBO</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>1402</td>
<td>697</td>
<td></td>
</tr>
<tr>
<td>Female (%)</td>
<td>1001 (71.4)</td>
<td>502 (72.0)</td>
<td>0.804</td>
</tr>
<tr>
<td>Prior MS treatment (%)</td>
<td>667 (47.6)</td>
<td>329 (47.2)</td>
<td>0.909</td>
</tr>
<tr>
<td>McDonald criteria &gt; 1(%)</td>
<td>279 (19.9)</td>
<td>117 (16.8)</td>
<td>0.097</td>
</tr>
<tr>
<td>Non-white ethnicity (%)</td>
<td>215 (15.3)</td>
<td>112 (16.1)</td>
<td>0.709</td>
</tr>
<tr>
<td>Age (mean (sd))</td>
<td>38.18 (9.23)</td>
<td>37.67 (9.28)</td>
<td>0.232</td>
</tr>
<tr>
<td>Years since symptom onset (mean (sd))</td>
<td>8.09 (6.74)</td>
<td>8.01 (6.36)</td>
<td>0.807</td>
</tr>
<tr>
<td>Years since diagnosis (mean (sd))</td>
<td>5.09 (5.32)</td>
<td>5.16 (5.22)</td>
<td>0.775</td>
</tr>
<tr>
<td>Expanded disability scale status (mean (sd))</td>
<td>2.47 (1.22)</td>
<td>2.51 (1.20)</td>
<td>0.539</td>
</tr>
<tr>
<td>9 Hole peg test average score (mean (sd))</td>
<td>23.18 (13.97)</td>
<td>23.81 (15.97)</td>
<td>0.356</td>
</tr>
<tr>
<td>PASAT 3 score (mean (sd))</td>
<td>48.49 (11.14)</td>
<td>47.81 (11.50)</td>
<td>0.196</td>
</tr>
<tr>
<td>Timed 25 foot walk (mean (sd))</td>
<td>7.42 (9.33)</td>
<td>7.54 (10.32)</td>
<td>0.79</td>
</tr>
<tr>
<td>Visual functioning test 2.5% (mean (sd))</td>
<td>32.34 (12.4)</td>
<td>32.24 (12.18)</td>
<td>0.851</td>
</tr>
<tr>
<td>SF-36 MCS (mean (sd))</td>
<td>45.35 (10.9)</td>
<td>45.50 (10.91)</td>
<td>0.758</td>
</tr>
<tr>
<td>SF-36 PCS (mean (sd))</td>
<td>43.10 (10.21)</td>
<td>43.22 (9.98)</td>
<td>0.803</td>
</tr>
<tr>
<td>Number relapses 1 year prior (mean (sd))</td>
<td>1.32 (0.64)</td>
<td>1.31 (0.71)</td>
<td>0.844</td>
</tr>
<tr>
<td>Months since pre-study relapse (mean (sd))</td>
<td>6.44 (6.43)</td>
<td>6.39 (4.90)</td>
<td>0.857</td>
</tr>
</tbody>
</table>

3.4.1 Results

The results of the cross-validation are summarized in Figure 3.1. The average value of \( \hat{T}^{(k)} \) are relatively low for the initial cumulative subgroups. While the raw treatment effects in these initial subgroups may be large, they are relatively small when standardized by their standard error. The standardized effects increase in larger subgroups but tapers for subgroups that nearly include the entire population. This suggests that there exists a small subgroup of patients for whom the treatment is not effective. We find the optimal proportion to be patients with the lowest \( \hat{p}^* = .78 \) proportion of estimated treatment differences approximated by tree. Subgroups formed by treatment differences approximated by the tree closely approximated or even improved in their test set performance over subgroups formed by treatment effects estimated by the initial logistic regression estimator \( \hat{\Delta}(X) \). This could possibly result from a regularization effect from re-fitting \( \hat{\Delta}(X) \) using simple piecewise constant functions of the tree.
Figure 3.1: $\hat{T}_g$ averaged over cross-validation repetitions for cumulative subgroups defined by treatment effects approximated by the tree (red) and by the initial logistic regression estimator (black). Larger values denote larger subgroup treatment effects standardized by its standard error. The optimal proportion of patients to be included in the final subgroup was found to be $\hat{p}^* = .78$.

After identifying $\hat{p}^*$, we ran the tree approximation procedure to estimate $\hat{h}(X)$ in all of $D_{Ita}$ and identified $\hat{S}$. Table 3.2 presents the results when we validated $\hat{S}$ in $D_{val}$. Among the 700 patients in $D_{val}$, 509 (73%) belonged to $\hat{S}$. This does not match $\hat{p}^* = .78$ exactly due to the discreteness in the scores from $\hat{h}(X)$. The tree approach yielded a subgroup $\hat{S}$ in which the treatment effect was estimated to be 20% in favor of BG-12 (95% CI for treatment difference [-.292 , -.102] and chi-square test $p < .001$). In particular, among those in $\hat{S}$, placebo patients experienced high rates of relapse, up to 46%. In the complement subgroup $\mathcal{X} \setminus \hat{S}$, the treatment effect was around 1% in favor of placebo (95% CI [-.134,.148], $p = 1.000$). Among these patients the relapse rate was comparatively low for patients in both BG-12 and placebo groups, around 26-27%. For comparison, we also validated subgroups identified by cumulative subgroups formed directly by scoring with $\hat{\Delta}(X)$. The results parallel those of the tree approach, except that the discrepancies...
between the selected subgroup and the complement subgroup were less dramatic. The patients in the complement subgroup exhibited treatment effects of 11% in favor of BG-12 (95% CI [-.278,.069], \( p = .257 \)). The subgroup formed directly via \( \Delta(X) \) are thus relatively diluted and do not optimally differentiate patients with small treatment effects. This further reinforces that re-approximation with trees, possibly due to a regularization effect, can lead to consequential improvements in subgroup selection.

Table 3.2: Validation of identified subgroup \( \hat{S} \) in \( D_{val} \). Complement subgroup refers to patients belonging to \( \mathcal{X} \setminus \hat{S} \) in \( D_{val} \). Tree and regression refer to subgroups identified by approximating estimated treatment effects by tree and the initial direct estimates by logistic regression. 95% CIs constructed based on normal approximation. \( p \)-value refer to two-sided chi-square test against null that there is no treatment effect.

<table>
<thead>
<tr>
<th>Subgroup</th>
<th>Method</th>
<th>( n )</th>
<th>( \hat{p}_1 )</th>
<th>( \hat{p}_0 )</th>
<th>( \hat{\Delta} )</th>
<th>( X^2 )</th>
<th>95% CI</th>
<th>( p )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified</td>
<td>Tree</td>
<td>509</td>
<td>0.265</td>
<td>0.462</td>
<td>-0.197</td>
<td>18.367</td>
<td>(-0.292, -0.102)</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>Regression</td>
<td>548</td>
<td>0.262</td>
<td>0.413</td>
<td>-0.151</td>
<td>11.668</td>
<td>(-0.242, -0.06)</td>
<td>0.001</td>
</tr>
<tr>
<td>Complement</td>
<td>Tree</td>
<td>191</td>
<td>0.269</td>
<td>0.262</td>
<td>0.007</td>
<td>0.000</td>
<td>(-0.134, 0.148)</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>Regression</td>
<td>152</td>
<td>0.280</td>
<td>0.385</td>
<td>-0.105</td>
<td>1.283</td>
<td>(-0.278, 0.069)</td>
<td>0.257</td>
</tr>
</tbody>
</table>

Figure 3.2 plots the tree approximation estimated in \( D_{tts} \). The green terminal nodes represent the leaf regions that comprise \( \hat{S} \), whereas the red terminal nodes are those that comprise the complement subgroup \( \mathcal{X} \setminus \hat{S} \). The default \texttt{rpart} settings with complexity parameter \( cp=0.001 \) yields a reasonably sized tree that is not very difficult to interpret. In terms of composition, the complement subgroup \( \mathcal{X} \setminus \hat{S} \) include patients who are older (age \( \geq 37 \)) with either a short duration of symptom onset (onset years \( > 6.3 \)) or with long onset and older age (age \( \geq 45 \)) and short timed 25 foot walk (\( < 4.9 \)). That is, older patients having a short duration of symptoms and those with longer onset but still maintain steady walking ability exhibit low rates of relapse, regardless of treatment. Although BG-12 is efficacious for much of the the study population, there appears to be minimal benefit among this low-risk group who develops MS later on in life. BG-12 may not need to be prioritized for such patients, and other therapies may be sought to further mitigate future relapse in this group.
Figure 3.2: Tree estimated to approximate $\hat{\Delta}(X)$ in $D_{tts}$. Green nodes refer to those with treatment differences less than $\hat{q}^* = \hat{F}^{-1}(\hat{p}^*)$, where $\hat{F}^{-1}(q)$ refers to the empirical quantiles of $\hat{h}(X_i)$ evaluated over $X_i$ in $D_{val}$. Red nodes refer to those with treatment differences greater than $\hat{q}^*$. The final subgroup $\hat{S}$ is defined by the union of the terminal green nodes, while $\mathcal{X} \setminus \hat{S}$ the union of the terminal red nodes. The percentages refer to the proportion of patients $D_{tts}$ belonging to each terminal node. Plot is drawn using R package rpart.plot (Milborrow, 2016).

3.5 Discussion

Both performance in differentiating patients and interpretability are important for identifying subgroups with enhanced treatment effects. In this paper, we proposed to adapt the approach of “born again” trees (Breiman and Shang, 1996) for subgroup identification, by estimating treatment effects given $X$ with a flexible initial estimator and then obtaining an accurate re-approximation through fitting a single tree on synthetic data. We also proposed a data-driven approach to identify the final subgroup after re-approximation.
and considered validating the identified subgroup on independent validation data. Applying the approach to pooled trial data to evaluate the effects of BG-12 on preventing relapse among MS patients, we found that it was able to differentiate a sizable subgroup of low-risk patients for which BG-12 exhibited no significant treatment effects.

In the analysis we conducted, only a simple logistic regression approach was considered for the initial treatment effect estimator $\hat{\Delta}(X)$. More sophisticated approaches that account for non-linearity in effects such as random forests, neural networks, and other machine learning methods may improve the quality of $\hat{\Delta}(X)$ in estimating the true treatment effects $\Delta(X)$. However, more complex tree structures may also be needed in the re-approximation to benefit from refining the initial estimator. In the re-approximation, we generated $X^\ast$ from normal and Bernoulli distributions that match the empirical means and variance-covariances of the observed $X_i$. This is a crude approximation to the original distribution of $X_i$, and more faithful representation of the original distribution could enable $\hat{h}(X)$ to approximate $\hat{\Delta}(X)$ faster, by concentrating the synthetic data in the relevant regions of $\mathcal{X}$. Alternatively, an active sampling strategy might be used to sample data relevant for nodes the tree attempts to fit, which was found to improve the fidelity of the approximation (Bastani et al., 2017).

The improvement in performance of subgroups defined by the tree approximation over those defined by the initial regression estimator is somewhat surprising, since the tree was fit only to approximate the regression estimator rather than directly on observed data. If this is due to a regularization effect, it would be of interest to further examine the performance for trees and initial estimators of varying complexity to understand under what contexts can such an improvement be expected. A similar phenomenon has been observed for neural networks trained in the “born again” framework (Furlanello et al., 2017). In identifying the final subgroup $\hat{S}$, we chose a simple thresholding rule after fitting the tree based on approximating the Bayes rule. However, this could lead to unstable subgroup definitions when the leaves are small as there are no constraints on the distance between the leaf regions in forming $\hat{S}$. It would be of interest also to consider some form of a clustering or constrained grouping procedure that seeks to minimize distances between the leaves to yield more stable and plausible subgroups.
Adaptive Combination of Conditional Treatment Effect Estimators Based on Randomized and Observational Data

David Cheng
Department of Biostatistics
Harvard T.H. Chan School of Public Health

Ross Prentice
Department of Biostatistics
University of Washington School of Public Health and Community Medicine

Tianxi Cai
Department of Biostatistics
Harvard T.H. Chan School of Public Health
4.1 Introduction

Randomized clinical trials (RT) are regarded as the gold-standard for evaluating the causal effect of interventions (Jones and Podolsky, 2015). Fundamentally, randomization of treatment assignment mitigates confounding bias arising from both measured and unmeasured covariates. However, RTs are typically designed to evaluate some form of an average treatment effect (ATE) of a treatment $T$ on an outcome $Y$ in some specified population. To assess treatment effect heterogeneity and inform treatment decisions in practice, it is often also of interest to estimate conditional average treatment effects (CATE) given a set of observed covariates $X$. RT data may be undersized for estimation and inference about CATEs, as they are powered for testing ATEs (Rothwell, 2005). Moreover, RTs often enroll patient populations that are more narrowly defined than those of real-world clinical practice, sometimes raising questions about generalizability of findings (Cole and Stuart, 2010). This may also make it difficult to estimate CATEs due to the limited heterogeneity in observed patient profiles.

Observational studies (OS), in contrast, can potentially collect data for large samples from populations that are more representative than RTs. This makes it appealing to use OS data for estimating CATEs. But without random treatment assignments, adjustment for certain covariates is needed to avoid bias from confounding, and there is usually never a guarantee whether confounding has been adequately addressed. These complementary features of RT and OS raise the question of whether data from both sources could simultaneously be leveraged to yield a more effective CATE estimator, in situations where it is possible to observe parallel data on the same treatments, outcomes, and covariates from a RT and OS. Such data could arise, for example, when a pragmatic trial is conducted to confirm initial findings discovered in an OS. Availability of such parallel data can be expected to become more prevalent as electronic medical records become more widely used for research and interest grows in conducting pragmatic trials in real-world clinical settings (Ford and Norrie, 2016). This problem differs that of the developing literature on generalizing RT estimates (Cole and Stuart, 2010; Stuart et al., 2011; Hartman et al., 2015; Zhang et al., 2016b) in that $Y$ is also observed in the OS and it is of interest to combine
CATE estimates from both sources for efficiency gain, not only to extrapolate RT estimates to other populations.

Synthesizing evidence from RT and OS has long been a topic of interest in meta-analysis. A broad approach is to modify meta-analysis models for aggregate data to accommodate potential heterogeneity in treatment effects estimated from RT and OS data. For example, Prevost et al. (2000) proposed a three-level hierarchical model that allows study-specific treatment effects to be randomly drawn from different distributions for RT and OS studies. Another approach under a Bayesian model is to use OS estimates to formulate informative priors in models for the RT data (Schmitz et al., 2013). A comprehensive review of methods from the meta-analysis literature can be found in Verde and Ohmann (2015). When individual patient-level data is available for the RT and OS, the “Cross-Design Synthesis” method proposed to use ATE estimates from the OS to extrapolate estimates from the RT to populations not eligible for the RT, assuming the confounding bias for the OS estimator is constant between patients eligible and not eligible for the RT (Silberman et al., 1992; Kaizar, 2011). This assumption requires treatment effect modifiers to be unrelated to covariates that induce confounding, which can be difficult to justify. A related approach was taken in a series of analyses performed in the Women’s Health Initiative (WHI) combined RT and OS data to evaluate the effects of hormone therapy on cardiovascular, cancer, and other health outcomes (Prentice et al., 2006, 2008a,b). Finding that the residual bias in the OS hazard ratios to be similar for two hormone therapies, the analyses fit a Cox model among the combined RT and OS data to estimate the OS hazard ratio, adjusting for the residual bias by including terms for the interaction between treatment and study type and the type of hormone therapy. The aforementioned approaches, moreover, focus on estimating some form of an ATE and do not consider CATEs. More recently, in the context of A/B testing for experimenting with websites, Peysakhovich and Lada (2016) proposed to simultaneously leverage randomized and observational data by first obtaining initial CATE estimates through a linear model in the OS and then using the RT data to calibrate the initial estimate, noting that the initial estimate is a rank preserving estimate of the true effect under some assumptions for the unobserved confounding variables. This rank preservation assumption, however, may
be difficult to justify in general contexts.

In the following we consider an adaptive approach for combining CATE estimates from RT and OS data. Under some assumptions to address confounding and treatment effect heterogeneity, we propose “base” estimators constructed from the RT and OS data are consistent for the same target estimands. These are combined through a linear combination with weights that yield the optimal asymptotic variance when the estimators agree on a common target or aggressively weights toward the RT estimate if the estimators exhibit sufficient disagreement. The yields a combined estimator that that is an efficient combination when assumptions for combining are satisfied and smoothly adapts to discard the OS estimate in case the assumptions are not satisfied.

4.2 Identification in RCT and OS

Let $Y$ denote an outcome, $T \in \{0, 1\}$ a binary treatment, $X$ a $p$-dimensional vector of baseline covariates, and $Z \in \{0, 1\}$ an indicator of participation in the RT ($Z = 0$) or OS ($Z = 1$) for a subject. We assume $Y$, $T$, and $X$ share common definitions across the two studies, though their distributions may differ. The RT and OS each include independent and identically distributed (iid) data of the form $D_z = \{(Z_i, X_i^T, T_i, Y_i)^T : Z_i = z\}$, for $z = 0$ and $z = 1$, respectively. The combined data is $D = D_0 \cup D_1$. Let $n_z = |D_z|$, with $n = n_0 + n_1$ being the total subjects among both studies. Given $D$, we want to make inferences about the CATE of $T$ given $X = x$, among the OS population:

$$\Delta_1(x) = \mathbb{E}\{Y(1) - Y(0) \mid X = x, Z = 1\}, \quad (4.1)$$

where $Y(t)$ denotes the counterfactual outcome had a subject received treatment $T = t$, for $t = 0, 1$. The CATE among the OS population is of particular interest because the OS study typically better reflects the real-world population facing the treatment decision in clinical practice. The method could be modified to obtain the effect in another population and is discussed in Section 4.5. We next formulate the assumptions to identify $\Delta_1(x)$ based on the RT and OS data distributions.
Basic assumptions for identifying causal effects from the RT data distribution include:

\[ T \perp \perp \{Y(1), Y(0), X\} \mid Z = 0 \quad (4.2) \]

\[ Y = TY(1) + (1 - T)Y(0) \text{ almost surely} \quad (4.3) \]

\[ \pi_1(x, z) \in [\epsilon_\pi, 1 - \epsilon_\pi], \text{ for } x, z \text{ such that } f(x, z) > 0, \quad (4.4) \]

where \( \pi_t(x, z) = \mathbb{P}(T = t \mid X = x, Z = z) \), for \( t = 0, 1 \), and \( \epsilon_\pi > 0 \) is some positive constant. The first condition specifies that \( T \) is independent of other variables among participants in RT due to randomization. The second is the standard consistency assumption linking counterfactual and observed outcomes. The last is a positivity assumption assuming treatment was not excluded from subgroups of patients defined by \( X = x \) in either the RT or OS. These basic conditions are assumed to hold across all potential scenarios that are considered.

For the RT and OS estimators to be consistent for a common estimand, \( X \) must be sufficiently rich to both alleviate confounding bias among the OS subjects and also capture treatment effect heterogeneity in the CATE between RT and OS patient populations. These assumptions can be formalized by the no unmeasured confounding (NUCA) and no unmeasured treatment effect modifiers (NUEM) conditions:

\[ T \perp \perp \{Y(1), Y(0)\} \mid X, Z = 1 \quad (4.5) \]

\[ Z \perp \perp \{Y(1) - Y(0)\} \mid X. \quad (4.6) \]

It is not uncommon for either one or both of these assumptions to not hold in practice, given the available \( X \). If they do hold, data from both the RT and OS could be used to obtain consistent estimators of the CATE, which could be further combined for optimal efficiency. Otherwise, if there remains residual confounding or treatment effect heterogeneity then consistent estimators are not simultaneously available for both data sources. These assumptions, unlike the basic assumptions (4.2)-(4.4), are viewed as ideal assumptions that may not hold in our setup. When they are violated, we consider a data-adaptive approach that asymptotically detects the violation and uses estimates from the RT to still estimate a conditional causal treatment effect.
Suppose for the moment that the PS function $\pi_t(x, z)$ known. In this case $\Delta_1(x)$ can be identified based on either RT or OS data through a simple modified outcomes approach (Signorovitch, 2007; Tian et al., 2014). Let the modified outcome be given by:

$$Y_t^* = \frac{I(T = t)}{\pi_t(X, Z)} Y.$$

Had one been interested in the CATE among the combined RT and OS population, it is identified under assumptions (4.2)-(4.5) through straightforward application of the g-formula (Robins, 1986) as:

$$\mathbb{E} \{Y(1) - Y(0) \mid X = x\} = \mathbb{E}(Y_1^* - Y_0^* \mid X = x).$$

There is no need to assume NUEM from (4.6) when the target estimand is among the combined populations. But when $\Delta_1(x)$ is specifically of interest, NUEM is needed to extrapolate effects estimated from the RT. In this case under assumptions (4.2)-(4.6):

$$\Delta_1(x) = \mathbb{E}(Y_1^* - Y_0^* \mid X = x, Z = 1) = \mathbb{E}(Y_1^* - Y_0^* \mid X = x, Z = 0).$$

### 4.2.1 Identification with Dimension Reduction of X

More than a few covariates may need to be included in $X$ to satisfy the assumptions (4.5)-(4.6). Direct nonparametric estimation of the above conditional means given $X = x$ would be infeasible due to the curse of dimensionality. An alternative strategy is to reduce $X$ by a low-dimensional score $S = r(X)$ of dimension $r \geq 1$, where $r(\cdot)$ is some transformation, potentially estimated from the data. This somewhat complicates identification as we have to acknowledge that any reduction $S$ may not fully capture the relevant variation in $X$. To make the analysis tractable, this next assumption formalizes a sufficient condition for identifying $\Delta_1(x)$ given the reduction $S$ instead of $X$:

$$X \perp \perp \{Y(1), Y(0)\} \mid S, Z \quad (4.7)$$

This simply states that the distribution of the potential outcomes depend on $X$ through $S$ among subjects in either the RT or OS population. As an example, suppose that $Y = \beta_0 + T\beta_1 + X^T\beta_2 + \varepsilon$, where $\varepsilon \sim N(0, 1)$ is independent of $(T, X^T)^T$. In this case (4.7) is satisfied with $S = X^T\beta_2$. 

86
For a sufficient reduction \( S \) satisfying (4.7), \( \Delta_1(x) \) is identified under the basic and ideal assumptions (4.2)-(4.6):

\[
\Delta_1(x) = \mathbb{E}(Y_1^* - Y_0^* \mid S = s, Z = 1) = \mathbb{E}(Y_1^* - Y_0^* \mid S = s, Z = 0).
\]

But if \( S \) is not a sufficient reduction, it is possible that \( \mathbb{E}(Y_1^* - Y_0^* \mid S = s, Z = 1) \neq \mathbb{E}(Y_1^* - Y_0^* \mid S = s, Z = 0) \) so that target estimands in RT and OS may not even agree. However, this problem can be alleviated by considering a further re-weighting of the RT outcomes. Let \( \varrho_z(x) = \mathbb{P}(Z = z \mid X = x) \) be the conditional probability of selection into the study \( z \), for \( z = 0, 1 \), given \( X = x \). For such a re-weighting to be successful, it must be that, among the combined population, there are no possible subjects with certain covariates \( X = x \) making them eligible for the OS and yet ineligible for the RT. In this case the distribution of the the data from the RT simply cannot provide information on the entirety of the OS population. This is formalized in the follow additional positivity assumption for \( \varrho_0(x) \) needed for re-weighting the RT outcomes:

\[
\varrho_0(x) \geq \epsilon_\varrho \text{ for } x \text{ such that } f(x) > 0,
\]

where \( \epsilon_\varrho > 0 \) is some constant. Under the basic and ideal assumptions (4.2)-(4.6) and this positivity condition, the following re-weighting of the RT outcomes leads to identification of a common CATE given \( S = s \), even when \( S \) is not sufficient:

\[
\mathbb{E} \left\{ I(Z = 0) \omega(X)(Y_1^* - Y_0^*) \mid S = s \right\} / \mathbb{P}(Z = 1 \mid S = s) = \mathbb{E}(Y_1^* - Y_0^* \mid S = s, Z = 1),
\]

\[
= \mathbb{E} \left\{ Y(1) - Y(0) \mid S = s, Z = 1 \right\},
\]

where \( \omega(x) = \varrho_1(x)/\varrho_0(x) \) is the odds ratio for trial selection.

### 4.3 Estimation and Adaptive Combination

#### 4.3.1 Estimation of Base Estimators

We now use the identification result from (4.9) to construct “base” estimators of \( \Delta_1(x) \) separately from \( D_1 \) and then \( D_0 \), respectively as \( \hat{\Delta}_1(s) \) and \( \hat{\Delta}_0(s) \). These two estimators are then combined in Section 4.3.2 in a data-adaptive fashion to accommodate cases when
working assumptions fail to hold. For the base estimators, we first posit parametric working models for the nuisance PS and trial selection functions:

\[
P(T = 1 \mid X = x, Z = z) = g_\pi(\alpha_0 + \alpha_1 z + \alpha_2 x z) = \pi_1(x, z; \alpha) \tag{4.10}
\]

\[
P(Z = 1 \mid X = x) = g_\theta(\gamma_0 + \gamma_1 x) = \varrho_1(x; \gamma), \tag{4.11}
\]

where \(g_\pi(\cdot)\) and \(g_\theta\) are specified link functions and \(\alpha = (\alpha_0, \alpha_1, \alpha_2^T)^T\) and \(\gamma = (\gamma_0, \gamma_1^T)^T\).

Let \(\pi_0(x, z; \alpha) = 1 - \pi_1(x, z; \alpha)\) and \(\varrho_0(x; \gamma) = 1 - \varrho_1(x; \gamma)\). The PS model (4.10) is guaranteed to be correctly specified when \(z = 0\) due to randomization assumption (4.2). Otherwise, these models are considered to be \textit{working} models that could possibly be misspecified. Let \(\hat{\alpha}\) and \(\hat{\gamma}\) be the maximum likelihood estimators (MLE) of \(\alpha\) and \(\gamma\) from maximizing the corresponding likelihoods based on these working models. Furthermore, let \(\hat{S}\) be a reduction of \(X\) that is estimated from the data, for example through estimating the regression model as in the example in Section 4.2.1. Let \(\hat{Y}_t^*, t = 0, 1\), be an estimated version of the modified outcome \(Y_t^*\) given by substituting \(\pi_t(X, Z; \hat{\alpha})\) for \(\pi_t(X, Z)\) in the modified outcome. Let the asymptotic limits of these estimators be denoted \(\tilde{\alpha}, \tilde{\gamma}, \tilde{S}, \tilde{Y}_t^*\), respectively.

The base RT estimator can be constructed based on an empirical version of (4.9), estimating conditional means by kernel smoothing:

\[
\hat{\Delta}_0(s) = \frac{\sum_{j=1}^n K_h(\bar{S}_j - s) I(Z_j = 0) \omega(X_j; \hat{\gamma})}{\sum_{j=1}^n K_h(\bar{S}_j - s)} \left\{ \frac{\hat{Y}_{j,0}^*}{\hat{\kappa}_{0,0}(s)} - \frac{\hat{Y}_{j,0}^*}{\hat{\kappa}_{j,0}(s)} \right\},
\]

where \(K(\cdot)\) is a \(q\)-th order \(r\)-variate kernel, \(h = O(n^{-\alpha})\) is a smoothing bandwidth for some appropriate \(\alpha > 0\), and \(K_h(\cdot) = K(\cdot/h)/h\). The constants \(\hat{\kappa}_{t,z}(s) = \sum_{j=1}^n K_h(\bar{S}_j - s) I(Z_j = z) \omega(X_j; \hat{\gamma})^{(1-z)} \{ I(T_j = t)/\pi_t(X_j, Z_j; \hat{\alpha}) \} / \sum_{j=1}^n K_h(\bar{S}_j - s)\) are normalization constants, for \(t = 0, 1\) and \(z = 0, 1\). When the trial selection model (4.11) is correctly specified, \(\hat{\kappa}_{t,0}(s)\) is consistent for \(P(Z = 1 \mid S = s)\), as required in (4.9). Normalizing the IPW weights is a way to stabilize the weights and tends to decrease the standard error of the final estimator. With the normalizing constant, it can be seen after simplification that this is simply a difference of kernel smoothers of \(Y\) over \(\hat{S}\) between the the two treatment groups of the RT data, weighting by \(\omega(X_j; \hat{\gamma})\). The second expression in (4.9) suggests a
similar smoothing approach for the base OS estimator to be:

\[
\tilde{\Delta}_1(s) = \frac{\sum_{j=1}^n K_h(\hat{S}_j - s) I(Z_j = 1)}{\sum_{j=1}^n K_h(\hat{S}_j - s)} \left\{ \frac{\hat{Y}^*_j}{\hat{\kappa}_{1,1}(s)} - \frac{\hat{Y}^*_j}{\hat{\kappa}_{0,1}(s)} \right\}.
\]

This can also been seen to be simply a difference of kernel smoothers of \(Y\) over \(\hat{S}\) between the treatment groups of the OS data, weighting by \(\pi_T(X_j; \hat{\alpha})^{-1}\). We next show that \(\hat{\Delta}_0(s)\) and \(\hat{\Delta}_1(s)\) admit an asymptotically linear-like representation, which shows the rate of convergence to the respective limiting value and later will help in the construction of the adaptive combination.

**Lemma 1.** Let \(\bar{\Delta}_z(s) = \bar{\mu}_{1,z}(s) - \bar{\mu}_{0,z}(s)\), be the asymptotic limits of \(\hat{\Delta}_z(s)\), where \(\bar{\mu}_{t,z}(s) = \mathbb{E}\{I(Z = z) \omega(X; \gamma)(1-z)Y_t^* | S = s\}/\mathbb{E}\{I(T = t, Z = z) \omega(X; \gamma)(1-z)\pi_T(X, Z; \alpha)^{-1} | S = s\}\), for \(t = 0, 1\) and \(z = 0, 1\). Let \(\hat{f}_{t,z}(s)\) be the asymptotic limit of the weighted kernel density estimator of \(\hat{S}\), \(\hat{f}_{t,z}(s) = n^{-1} \sum_{j=1}^n K_h(\hat{S}_j - s) I(T_j = t, Z_j = z) \omega(X_j; \gamma)(1-z)\pi_t(X_j, Z_j; \alpha)^{-z}\), for \(t = 0, 1\) and \(z = 0, 1\). Assume the reduction is of dimension \(r = 1\) and \(h = O(n^{-\alpha})\) with \(\alpha < 1/2\). The base estimators admit an asymptotically linear-like representation as:

\[
\hat{\Delta}_0(s) - \bar{\Delta}_0(s) = \frac{1}{nh} \sum_{i=1}^n \xi_{0,0,h}(s) + o_p\{(nh)^{-1/2}\},
\]

\[
\hat{\Delta}_1(s) - \bar{\Delta}_1(s) = \frac{1}{nh} \sum_{i=1}^n \xi_{1,1,h}(s) + o_p\{(nh)^{-1/2}\},
\]

where \(\xi_{i,z,h}(s) = \xi_{i,1,z,h}(s) - \xi_{i,0,z,h}(s)\), for \(z = 0, 1\), with:

\[
\xi_{i,t,z,h}(s) = K\{(\bar{S}_i - s)/h\} \omega(X_i; \gamma)(1-z)\pi_t(X_i, Z_i; \alpha)^{-z}I(T_i = t, Z_i = z)\{Y_j - \bar{\mu}_{t,z}(s)\} \hat{f}_{t,z}(s)^{-1}.
\]

The contributions \(\xi_{i,z,h}(s)\) are independent, identically distributed for a given \(h\), and have conditional mean 0 such that \(\mathbb{E}\{\xi_{i,z,h}(s) | \bar{S} = s\} = 0\).

We informally refer to \(\xi_{i,z,h}(s)\) as influence functions, although they are not identically distributed for different \(n\) when \(h\) varies with \(n\). This next result clarifies the interpretation of limiting estimands \(\bar{\Delta}_0(s)\) and \(\bar{\Delta}_1(s)\) under scenarios where working assumptions may or may not hold:

**Lemma 2.** Under the basic and ideal assumptions (4.2)-(4.6) and the positivity assumption in (4.8), when the PS model (4.10) and trial selection model (4.11) are correctly specified, the RT and
OS estimators share a common target estimand:
\[
\Delta_0(s) = \Delta_1(s) = \mathbb{E} \{ Y(1) - Y(0) \mid S = s, Z = 1 \}.
\]

In addition, if \( S \) is a sufficient reduction in the sense of (4.7), then also \( \Delta_0(s) = \Delta_1(s) = \Delta_1(x) \), for \( s = r(x) \). Under only the basic assumptions (4.2)-(4.4), \( \Delta_0(s) \) is a re-weighted CATE such that:
\[
\Delta_0(s) = \mathbb{E} \{ \mathbb{E}(Y(1) - Y(0) \mid X, Z = 0) \mathbb{P}(Z = 0 \mid X, Z = 0) \mid S = s \} / \tilde{\pi}(s),
\]
where \( \tilde{\pi}(s) = \mathbb{E} \{ \mathbb{P}(Z = 0 \mid X) \omega(X; \gamma) \mid S = s \} \) and \( \mathbb{E}_{X|S=s} (\cdot \mid S = s) \) denotes a conditional expectation under a tilted conditional distribution for \( X \), characterized by the tilted density \( \tilde{f}(x \mid S = s) = \mathbb{P}(Z = 0 \mid X = x) \omega(x; \gamma) / \tilde{\pi}(s) f(x \mid S = s) \).

This clarifies that there are two scenarios where \( \Delta_0(s) \) agrees with \( \Delta_1(s) \) and combination of the estimators is sensible. Outside of these two scenarios, the two estimators may not be consistent for the same estimand and are thus not suitable for combination. Nevertheless, due to the randomization in RT, \( \hat{\Delta}_0(s) \) always offers an estimator that is consistent for some re-weighted approximation of the desired CATE \( \Delta_1(x) \). Given these considerations, we next consider an adaptive combination of \( \hat{\Delta}_1(s) \) and \( \hat{\Delta}_0(s) \) that targets estimating \( \Delta_0(s) \). It is anchored on \( \Delta_0(s) \) because under full assumption \( \Delta_0(s) = \Delta_1(s) \) and under even minimal basic assumptions it still gives some re-weighted CATE. The proposed combination is adaptive in that if assumptions for combination are not satisfied then the estimator weights heavily toward \( \hat{\Delta}_0(s) \), which is always consistent for \( \Delta_0(s) \).

### 4.3.2 Adaptive Combination

Let \( \hat{\Delta}(s; \eta) = \Delta_0(s) + \eta \{ \hat{\Delta}_1(s) - \hat{\Delta}_0(s) \} \) be some linear combination of the base estimators with OS weight \( \eta \). Implicitly we allow \( \eta \) to be specific to each \( s \) so as to allow for different combination decisions at different \( s \) but suppress this dependence in the notation. To target \( \Delta_0(s) \), the combined estimator can use an estimate of \( \eta \) that minimizes the mean squared error (MSE) around \( \Delta_0(s) \):
\[
Q_s(\eta) = \mathbb{E} \left\{ (\hat{\Delta}(s; \eta) - \Delta_0(s))^2 \right\}. \tag{4.12}
\]
This next result shows that the MSE can be expressed in terms of a least squares criterion based on the influence functions.

**Lemma 3.** Under the positivity conditions (4.4) and (4.8), if \( Y \) is a bounded outcome, then the MSE criterion can be expressed as:

\[
Q_s(\eta) = \frac{1}{nh^2} \mathbb{E} \left[ \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} \right]^2 + \eta^2 \left\{ \bar{\Delta}_1(s) - \bar{\Delta}_0(s) \right\}^2 + o\left((nh)^{-1/2}\right).
\]

From this, an empirical estimate of a scaled MSE is given by:

\[
\tilde{Q}_s(\eta) = \sum_{i=1}^{n} \left[ \hat{\xi}_{i,0,h}(s) + \eta \left\{ \hat{\xi}_{i,1,h}(s) - \hat{\xi}_{i,0,h}(s) \right\} \right]^2 + n^{2(1-\beta)}h^2 \eta^2 \left\{ \hat{\Delta}_1(s) - \hat{\Delta}_0(s) \right\}^2 ,
\]  

(4.13)

where \( \hat{\xi}_{i,z,h}(s) \) is an empirical estimate of \( \xi_{i,z,h}(s) \) that plugs in estimators for the nuisance quantities, for \( t = 0, 1 \) and \( z = 0, 1 \) and \( \beta > 0 \) is some correction factor. This type of approach for estimating a weight for efficiency gain based on minimizing a least squares criteria based on estimated influence functions has previously been considered in the context of augmenting treatment contrasts in RT data (Tian et al., 2012). Generally, plugging in nuisance parameters in the influence functions asymptotically contributes negligible error. On the other hand, when the bias term \( \bar{\Delta}_1(s) - \bar{\Delta}_0(s) \) is estimated, a correction term \( n^{-2\beta} \) is needed to allow for \( \eta \) to asymptotically weight toward 0 under biased scenarios. The required order is given in the result below. It can now be seen that, with an appropriate bandwidth \( h \) and correction order \( \beta \), an estimate of \( \eta \) can be obtained by running a ridge regression of \( \hat{\xi}_{i,0,h}(s) \) on \( \hat{\xi}_{i,0,h}(s) - \hat{\xi}_{i,1,h}(s) \) with a weighted ridge penalty and no intercept. Alternatively, since the influence functions have mean 0, it can also be estimated directly by least squares:

\[
\tilde{Q}_s(\eta) = \sum_{i=1}^{n} \left( \hat{\xi}_{i,0,h}(s) - \eta \left[ \hat{\xi}_{i,0,h}(s) - \hat{\xi}_{i,1,h}(s) - n^{1/2-\beta}h \left\{ \hat{\Delta}_1(s) - \hat{\Delta}_0(s) \right\} \right] \right)^2
\]

In either case, with a suitable \( h \) and correction order \( \beta \), this least squares estimator is an adaptive weight \( \tilde{\eta} \) that asymptotically either optimally combines \( \hat{\Delta}_0(s) \) and \( \hat{\Delta}_1(s) \) for efficiency when there is no bias in the OS estimator or weights toward \( \hat{\Delta}_0(s) \) otherwise.

**Theorem 1.** Let \( \beta \in (0, \frac{1-\alpha}{2}) \). The minimizer \( \tilde{\eta} = \arg \min_{\eta} \tilde{Q}_s(\eta) \) is an adaptive weight in that \( \tilde{\eta} = O_p\left\{ n^{\alpha+2\beta-1} \right\} \) when \( \bar{\Delta}_0(s) - \bar{\Delta}_1(s) \neq 0 \), and \( \tilde{\eta} - \eta = O_p\left\{ (nh)^{-1/2} \right\} \) when \( \bar{\Delta}_0(s) - \bar{\Delta}_1(s) = 0 \), where \( \eta \) is the weight such that the asymptotic variance of \( \hat{\Delta}(s; \eta) \) is minimized.
Applying this weight in \( \hat{\Delta}(s; \tilde{\eta}) \) yields an estimator that is always consistent for \( \Sigma_0(s) \), regardless of bias of \( \hat{\Delta}_1(s) \). Moreover, this result shows there is a tradeoff in how much to discount the bias term in the regressor when selecting \( \beta \). By selecting larger \( \beta \), the bias is discounted more and the procedure aims to directly estimate the weight that yields the optimal variance. But if there is a true bias, increasing \( \beta \) also slows the convergence of \( \tilde{\eta} \) to 0. In practice, we found that choosing \( \beta \) can be difficult for \( \tilde{\eta} \) and the convergence can be slow in either scenario. Instead of a weighted ridge penalty, we further consider a weighted \( \ell_1 \) penalty to speed up the convergence. Specifically, we replaced (4.13) with:

\[
\hat{Q}_s(\eta) = \sum_{i=1}^{n} \left[ \hat{\xi}_{i,0,h}(s) + \eta \left( \hat{\xi}_{i,1,h}(s) - \hat{\xi}_{i,0,h}(s) \right) \right]^2 + \lambda_n |\eta| \left( \hat{\Delta}_1(s) - \hat{\Delta}_0(s) \right)^2,
\]

where \( \lambda_n \) is a tuning parameter. This turns the problem into an adaptive LASSO (Zou, 2006) problem with a weight in the penalty such that the penalty is large when the estimated bias is large and vice versa when the estimated bias is small. By using the \( \ell_1 \) penalty, the procedure is allowed to set 0 exactly so as to completely discard \( \hat{\Delta}_1(s) \) when it is severely biased. Otherwise if \( \hat{\Delta}_1(s) \) is consistent, the first term of \( \hat{Q}_s(\eta) \) should dominate and lead to an estimator that optimizes the asymptotic variance in a way similar to \( \tilde{\eta} \). The theoretical properties of this modified procedure, which depends on the procedure used for tuning, are still under investigation. For a given \( \lambda_n \), the proposed adaptive estimator is \( \hat{\Delta}(s) = \hat{\Delta}(s; \tilde{\eta}) \), where \( \tilde{\eta} = \text{argmin}_{\eta} \hat{Q}_s(\eta) \).

To select the tuning parameter, we consider a cross-validation procedure that minimizes an out-of-sample integrated MSE. Specifically, we randomly partition \( \mathcal{D} \) into \( K \) subsets indexed by \( \mathcal{I}_1, \ldots, \mathcal{I}_K \). To estimate the integrated MSE without having to estimate the bias and provide a correction factor, we consider anchoring on a consistent estimator in the independent test set (Brookhart and Van Der Laan, 2006). That is, since \( \Sigma_0(s) \) in the MSE is unknown, we plug-in a consistent estimator \( \hat{\Delta}^{(k)}_0(s) \), estimated using observations in \( \mathcal{I}_k \) by the same smoothing approach in the base RT estimator. Let \( \hat{\Delta}^{(k)}_\lambda_n(s) \) be the adaptive estimator estimated from the training sample \{1, \ldots, n\} \( \setminus \mathcal{I}_k \) with a given \( \lambda_n \). We averaged the square differences between \( \hat{\Delta}^{(k)}_0(s) \) and \( \hat{\Delta}^{(k)}_\lambda_n(s) \) over the observed score reductions \( \hat{S}_i \) from the test set \( \mathcal{I}_k \), for a range of \( \lambda_n \) over a grid of plausible values over \( [0, 2\tilde{\eta}] \). For each \( \lambda_n \) the cross-validation evaluation averages the results over the \( K \) folds,
as given by:

\[ R(\lambda_n) = \sum_{k=1}^{K} \sum_{i \in I_k} \left( \hat{\Delta}^{(k)}_0(\hat{S}_i) - \hat{\Delta}^{(k)}_{\lambda_n}(\hat{S}_i) \right)^2. \]

To mitigate the variability in the results due to random partitioning in the cross-validation, we repeat this procedure, each time producing an evaluation \( R_l(\lambda) \) for each \( \lambda_n \), for \( l = 1, \ldots, L \). The results are averaged again before selecting the optimal parameter, as in \( \hat{\lambda}_n = \arg\min_{\lambda} L^{-1} \sum_{l=1}^{L} R_l(\lambda) \). This estimates an integrated MSE that averages the pointwise MSE at \( s \) across all \( s \) in the support of \( \hat{S} \), selecting one tuning parameter \( \hat{\lambda}_n \) for all \( s \). Data may be extremely limited for estimating a pointwise MSE for a given \( s \), and borrowing strength across \( s \) improves the stability. Conceptually, one may also want to make a single unified decision of whether to choose the optimal variance combination or \( \hat{\Delta}_0(s) \), as bias in \( \hat{\Delta}_1(s) \) at any \( s \) casts suspicion about its reliability at other \( s \). In expectation, this approach selects a \( \lambda_n \) that leads to a consistent estimator for \( \Delta_0(s) \) (Brookhart and Van Der Laan, 2006). Another potential approach is to estimate the MSE using weighted outcomes such as \( \hat{Y}_1^* - \hat{Y}_0^* \) to account for partially observed counterfactual outcomes, rather than \( \hat{\Delta}_0^{(k)}(s) \). However, anchoring the MSE directly on observed data may lead to highly variable estimators of the MSE, which may hinder the performance of the criteria in evaluating \( \lambda_n \). We found using the smoothing to estimate \( \Delta_0(s) \) tended to exhibit better tuning performance.

### 4.3.3 Inference

We consider estimating standard errors (SE) directly based on the influence functions. Since \( \hat{\Delta}(s) \) is a linear combination of estimators with a asymptotically linear-like representation, it also has a similar representation as:

\[
\hat{\Delta}(s) - \Delta_0(s) = (nh)^{-1} \sum_{i=1}^{n} \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} + \eta \left\{ \Delta_1(s) - \Delta_0(s) \right\} + (\hat{\eta} - \eta) \left\{ \Delta_1(s) - \Delta_0(s) \right\} + o_p \left\{ (nh)^{-1/2} \right\}.
\]

In the no bias case where \( \Delta_0(s) = \Delta_1(s) \), then the terms involving the bias are 0. On the other hand, provided that \( \hat{\eta} \) shrinks to 0 appropriately such that \( \mathbb{P}(\hat{\eta} = 0) \to 1 \) when
\( \Delta_0(s) \neq \Delta_1(s) \), the terms involving the bias are also \( o_p \{(nh)^{-1/2}\} \). If this holds, then there is no contribution to the variability from estimating \( \eta \) asymptotically. In this case, it can be shown using Liapunov’s central limit theorem that:

\[
(nh)^{1/2} \left\{ \hat{\Delta}(s) - \Delta_0(s) \right\} \xrightarrow{p} N \left\{ 0, \sigma^2(s) \right\},
\]

where \( \sigma(s)^2 = h^{-1}E \left\{ \eta \xi_{i,1,h}(s) + (1 - \eta) \xi_{i,0,h}(s) \right\}^2 \). The standard error of \( \hat{\Delta}(s) \) can thus be approximated through \( \hat{\sigma}^2(s) = (nh)^{-1} \sum_{i=1}^n \left\{ \hat{\eta} \xi_{i,1,h}(s) + (1 - \hat{\eta}) \xi_{i,0,h}(s) \right\}^2 \). Pointwise confidence intervals (CI) can be constructed through normal approximation given these estimated SEs.

### 4.4 Numerical Studies

#### 4.4.1 Simulations

We conducted simulations in finite samples to evaluate the bias and relative efficiency (RE) of \( \hat{\Delta}(s) \). We also evaluated the performance of the coverage of CIs for \( \hat{\Delta} \) constructed based on the influence functions. In addition to \( \hat{\Delta}(s) \) (CMB-HAT), we reported results for \( \hat{\Delta}_0(s) \) (RT), \( \hat{\Delta}_1(s) \) (OS), and the unregularized estimator \( \hat{\Delta}(s) = \Delta(s; \hat{\eta}) \) (CMB-TLD), which essentially optimizes the variance assuming no bias, for reference. For the reduction of \( X \), we modeled outcomes in the combined data through a main effects linear model \( Y = \beta_0 + T \beta_1 + X^T \beta_2 + \varepsilon \), taking \( \hat{\Sigma} = X \hat{\beta}_2 \). The estimators were evaluated for \( s \) in a central region belonging to the support of \( \hat{\Sigma} \) (\( s = -10 \) to \( s = 10 \)) that included over 70% of the population. In the smoothing we used a \( q = 2 \) order kernel with a plug-in bandwidth \( h = \hat{\sigma}_{\hat{\Sigma},1} n_1^{-1/5} \), where \( \hat{\sigma}_{\hat{\Sigma},1} \) denotes the sample standard deviation of \( \hat{\Sigma}_i \) among \( D_1 \) (Silverman, 1986). For cross-validation, we used \( K = 2 \) folds and repeated the procedure \( L = 5 \) times. When evaluating \( \hat{\Delta}^{(-k)}(s) \) in the test sets, we dropped observations in \( I_k \) with \( |\hat{\Sigma}| \geq 20 \) to mitigate undue influence on the results of cross-validation by outlying observations, for which smoothing estimates may be unreliable. Around 3% of the observations in the test sets were dropped this way.

Let \( \mathbf{1}_q \) and \( \mathbf{0}_q \) be vectors of 1’s and 0’s of length \( q \), respectively, and \( \mathbf{I}_q \) a \( q \times q \) identity matrix. We considered scenarios in which \( X \sim N \{ \mathbf{0}_p, .8 \mathbf{I}_p + .2 \} \), \( Z \mid X \sim \) ...
Bernoulli \(\{\varrho_1(X)\}, T \mid X, Z \sim \text{Bernoulli}\{\pi_1(X, Z)\}\), and an outcome model \(Y = .5 + T + X^T(1.751_p) + \varepsilon\), where \(\varepsilon \sim N(0, 1)\). We considered the case with \(p = 10\) covariates. The true trial selection function and PS were specified by:

\[
\varrho_1(X) = \expit\{2.5 + X^T(.051_p)\} \quad (4.14)
\]

\[
\pi_1(X, Z) = \expit\{-0.5Z + (X^TZ)^T(.315^T, 0_5^T)^T\} \quad (4.15)
\]

\[
\pi_1(X, Z) = .1\expit\{-4Z + (X^TZ)^T(.2110)\} + .9\expit\left[-5Z + \{(X^TZ)^T(.5110)\}^2\right]. \quad (4.16)
\]

The trial selection function is specified such that the OS represents a large proportion of the combined population, around 92%. There are also some weak selection effects such that the distribution of \(X\) is shifted towards larger values in the OS. The working model for \(\varrho_1(X)\) is assumed to be correctly specified. The PS is correctly specified by the working model in (4.15) and misspecified by a single-index model in (4.16), which exhibits non-linear effects in \(X\). We conducted the simulations with a combined size of \(n = 5,000\) and with 1,000 replications.

Figure 4.1 presents the bias results. In the top panels where \(\pi_1(X, Z)\) is correctly specified, \(\hat{\Delta}_0(s)\) and \(\hat{\Delta}_1(s)\) estimators both exhibit negligible bias across all \(s\) relative to the standard error. As a result, their combinations, including \(\bar{\Delta}(s)\) and \(\bar{\Delta}(s)\) also have negligible bias. In the bottom panels where \(\pi_1(X, Z)\) is misspecified, \(\hat{\Delta}_1(s)\) exhibits large biases that are several times larger than the standard error due to the misspecification. The RT estimator \(\hat{\Delta}_0(s)\) remains unbiased due to randomization. The unregularized combination \(\bar{\Delta}(s)\) weights toward \(\hat{\Delta}_0(s)\) somewhat but remains heavily biased. The regularized adaptive combination \(\bar{\Delta}(s)\) weights nearly entirely to \(\hat{\Delta}_0(s)\), leading it to also have negligible bias. In finite samples, this shows \(\bar{\Delta}(s)\) adapts so as to discard \(\hat{\Delta}_1(s)\) when strong bias is detected.

Figure 4.2 presents the RE results. Relative to \(\hat{\Delta}_0(s)\), \(\hat{\Delta}_1(s)\) and \(\bar{\Delta}(s)\) are around 5-18 times more efficient in terms of the MSE when \(\pi_1(X, Z)\) is correctly specified. These large RE reflects the large size of the OS relative to the RT. Since the OS selects subjects with large values of \(\hat{S}\), the efficiency are especially prominent is regions with large \(s\). However, when \(\pi_1(X, Z)\) is misspecified, \(\hat{\Delta}_1(s)\) and \(\bar{\Delta}(s)\) are much less efficient due to their high levels of bias. Our proposed estimator \(\bar{\Delta}(s)\) is as efficient as \(\bar{\Delta}(s)\) over a wide range of \(s\)
Figure 4.1: Empirical bias (left) standard error (right) of each estimator in simulations. The results were assessed for scenarios where working models for $\pi_1(X, Z)$ were correctly specified (top) or misspecified (bottom).

when $\pi_1(X, Z)$ is correctly specified, indicating there is a small price paid for regularization. On the other hand, under misspecification, the regularization prevents $\hat{\Delta}(s)$ from weighting towards a biased estimator. For $s$ belonging to the regions considered, there is virtually no loss of efficiency from $\hat{\Delta}_0(s)$ as $\hat{\eta}$ shrinks exactly to 0. There is only slight loss in efficiency in outlying regions of $s$, where data is more scarce.
Figure 4.2: Relative efficiency of $\hat{\Delta}_1(s)$, $\tilde{\Delta}(s)$, and $\hat{\Delta}(s)$ compared to $\hat{\Delta}_0(s)$, where relative efficiency was assessed as the ratio of empirical mean square error of $\hat{\Delta}_0(s)$ over that of each estimator, as estimated from the simulations. Higher values denotes more efficient estimators. The results were assessed for scenarios where working models for $\pi_1(X,Z)$ were correctly specified (left) or misspecified (right).

Figure 4.3 displays the ratio of the average of estimated SEs to empirical SEs over the simulations, as well as the empirical coverage of CIs based on normal approximations. Generally, the SEs estimated by influence functions for $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$ approximated the empirical standard errors within 10%, regardless of whether $\pi_1(X,Z)$ is correct or misspecified. The SE for $\hat{\Delta}(s)$ also similarly well-approximated the empirical SE, although there are some dips in outlying regions of $s$. Estimating $\eta$ does not contribute significantly to the variability. In terms of coverage, $\hat{\Delta}(s)$ exhibits nominal coverage under correct specification, as with $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$. Under misspecification, the coverage of $\hat{\Delta}_1(s)$ drops to 0 for most of $s$ considered due to its large bias. The adaptive estimator $\hat{\Delta}(s)$ retains coverage ranging between 92-96% for most of $s$ considered, although it drops to 90% in some outlying regions, likely due to the underestimation of the SE. As the SE for the base estimators are well-approximated even in outlying regions, this underestimation may result from some slight instability in the weight estimation in these regions with more scarce data.
Figure 4.3: Ratio of estimated SEs based on influence function approximation to empirical SEs (top) and the empirical coverage rate of associated CIs (bottom) in simulations. The results were assessed for scenarios where working models for $\pi_1(X, Z)$ were correctly specified (left) or misspecified (right).

4.4.2 WHI Data Analysis

The WHI was a collection of studies that investigated strategies for preventing common sources of mortality and morbidity among postmenopausal women, including cancer, cardiovascular disease, and fractures (Anderson et al., 1998). Among its various compo-
nents, the WHI included RTs and an OS examining the long-term health effects of hormone replacement therapy (HRT). In the following analysis, we focus on an analysis of the effects of estrogen plus progestin (E+P) combined hormone therapy on coronary heart disease (CHD) using data from the E+P trial and the cohort of women from the OS with a uterus who were either using combined hormone therapy or no hormone therapy.

The outcome was defined by occurrences of myocardial infarction (MI) and death due to CHD within 5-years of follow-up, which were locally and centrally adjudicated. The outcome was censored due to either loss to follow-up or administrative censoring at the end of the WHI core study in 2005. In this analysis we assume independent censoring and do not adjust for differential loss to follow-up, using an inverse probability weighting approach to estimate the 5-year risks on HRT and placebo. We included as covariates in $X$ age, ethnicity (white or non-white), education status (not high school graduate, high school graduate but not college graduate, college graduate or higher), ever smoking, a physical functioning score based on the SF-36 physical functioning subscore, lifetime E+P duration of use, previous occurrence of menopausal symptoms such as hot flashes or night sweats, and age at menopause. These covariates were selected to cover the basic adjustment variables used in many existing analyses of the WHI data. Duration of previous E+P use was also included, as it had been found to have a substantial impact on treatment effects upon adjustment (Prentice et al., 2005). Among subjects in the E+P trial and the OS cohort, we further excluded subjects without any outcome follow-up, without data on HRT use at baseline, and with missing data in any $X$. The final analytical dataset included $n_0 = 2,542$ subjects in the RT and $n_1 = 20,795$ in the OS. A substantial proportion of subjects are excluded due to missingness in the variable used for duration of previous E+P use.

We first conducted an initial Cox proportional hazards model analysis to estimate an overall adjusted hazard ratio in the RT and OS in the data sample before dropping subjects with missing E+P duration data, which comprised of 13,994 subjects in the RT and 47,976 in the OS. All covariates in $X$ except for duration of E+P use were included as main effects in the Cox model for adjustment. The hazard ratio for HRT was estimated to be $1.268 (p = .020)$ in the RT and $0.846 (p = .046)$ in the OS, reflecting the discrepancy in
overall treatment effects between the RT and OS initially reported in the WHI. We subsequently implemented the proposed method in the reduced analytical dataset, including duration of E+P use in X. Specifically, let $U$ be the underlying time to CHD and $C$ be the independent censoring time, with the observed time as $V = U \wedge C$. Let $G(v) = P(C > v)$ be the survival function for $C$ at time $v$ and $\hat{G}(v)$ be the Kaplan-Meier estimator of $G(v)$. We let the weighted outcome for each subject be $Y = I(U \leq C)I(V \leq t)/\hat{G}(V)$. Provided that $C$ is independent, similar arguments can be used to show the consistency of $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$ for the 5-year risk of CHD given $S$ under the assumptions discussed in this paper.

To estimate the score reduction, we fit a weighted logistic regression working model in the combined data:

$$P(V \leq t \mid X, T) = g_\pi(\beta_0 + T\beta_1 + X^T\beta_2),$$

weighting by $\{I(U \leq C)I(V \leq t) + I(V > t)\}/\hat{G}(V \wedge t)$ to account for censoring (Zheng et al., 2006). We then take $\hat{S} = X^T\hat{\beta}_2$. To identify values of $s$ for which to estimate $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$, we obtained 25 evenly spaced values of $s$ between the .05 and .95 quantiles of $\hat{S}$ estimated from the entire sample $D$. We applied cross-validation to identify $\hat{\lambda}_n$ with the same settings as in the simulations, except that we did not restrict the support of $\hat{S}$ in the test sets when calculating $R_t(\lambda_n)$. We calculated SEs and CIs for $\hat{\Delta}_0(s)$, $\hat{\Delta}_1(s)$, and $\hat{\Delta}(s)$ based on estimated influence functions as proposed in Section 4.3.3.

The results are summarized in Figure 4.4. Over the support of $\hat{S}$ that is considered, there is not strong disagreement between $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$ after adjustment with all of X, with some slight exceptions. Based on the estimated CIs, the estimated effects in the RT did not significantly differ from 0, whereas OS estimates indicated significant elevated risk for E+P for $s$ around 5.1 – 5.3, ranging from .7% to 1.1%. OS estimates also indicated significant elevated risks for $s$ around 3.5 – 3.7 and 4.25, ranging .1% to .3%. However, these effects may have also been considered to be non-significant had we considered simultaneous confidence bands rather than pointwise CIs. For $\hat{\Delta}(s)$, the cross-validation procedure selected $\hat{\lambda}_n = 0$ and allowed $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$ to combine for efficiency. This appears to be reasonable since $\hat{\Delta}_0(s)$ and $\hat{\Delta}_1(s)$ largely agree. There are some discrepancies in the point estimates for $s$ around 4.1 – 4.3, 5.5 – 5.7, and 6, among other locations.
But the variability of $\hat{\Delta}_0(s)$ is also relatively large in these regions. The cross-validation may have determined that shrinking the combination to $\hat{\Delta}_0(s)$ was not merited given the degree of uncertainty in $\hat{\Delta}_0(s)$. The combination $\hat{\Delta}(s)$ yields estimates that are much more precise than the RT estimates and gives some assurance that it does not incur bias from combining with the OS estimates. The combined estimates generally indicate there is no strong effect of HRT on 5-year CHD risk, although there may be slight elevated risk for subjects with $s$ around $5.1 - 5.3$.

Figure 4.4: Point estimates and CIs of the conditional treatment effects of combined HRT on 5-year CHD risk given $S$ in the WHI analytical dataset. The error bars indicate 95% pointwise CIs constructed based on influence function approximation.
4.5 Discussion

As interest grows in conducting RTs in real-world clinical settings, there may be increasing opportunities to simultaneously observe data from RTs and OSs evaluating common interventions. In this paper, we consider an approach for combining CATE estimates from parallel RT and OS data that seeks to address biases stemming from confounding and treatment effect heterogeneity. The combination is adaptive in that it smoothly weights toward the RT estimator when bias remains after adjustment for \( X \) or otherwise combines the estimators for optimal efficiency. This allows the data to directly assess the bias of the OS estimator and guide the decision to combine estimates across studies.

The proposed approach achieves adaptivity by anchoring the combined estimator on the RT estimator, under the assumption that the RT estimator always allows for consistent estimation of a relevant causal effect. In practice, poor design or execution of RTs may lead to issues such as extensive missing data, loss of follow-up, and measurement error, which can undermine the validity of RT estimates. Moreover, in some settings the OS study may be so well-conducted, with sources of potential sources of biases so thoroughly accounted, such that the results are potentially more trustworthy than a poorly-conducted RT. In such circumstances, and also more generally, it would be important to consider whether observed discrepancies between RT and OS estimates warrants confidence in the RT estimates. More broadly, in practice no OS estimator can be expected to be exactly free of bias, and it may be unclear when the estimators are allowed to combine for efficiency. But just as misspecified parametric models provide useful approximations, the no bias assumption serves as a useful guide for identifying cases in which combination should be allowed. In finite samples, we expect \( \hat{\lambda}_n \) to be 0 or small when the bias fall below detectable limits, although it would be interesting to consider the sensitivity of the procedure in the presence of low levels of bias.

It may be difficult to find a reduction \( S \) such that (4.7) is satisfied. As an alternative, one could estimate \( \hat{\Delta}_1(X) \) through a parametric or semiparametric model and use the initial estimator as \( S \). If the model was at least approximately correctly specified, the score could approximately satisfy \( X \perp \perp \{Y(1) - Y(0)\} \mid S, Z \), which would also be sufficient for
estimating the CATE. Dimension reduction in the context of causal inference has been an area of recent investigations (Luo et al., 2017; Persson et al., 2017; Huang and Chan, 2017), and methods based on sufficient dimension reduction, among others, are also possible. Another issue is that we proposed to re-weight \( \hat{\Delta}_0(s) \) to the OS population as the target population, but one may be interested in conditional effects in another population such as a subpopulation of the OS. In this case (4.11) may be modified to model the probability of selection into the target population, along with subsequently modifying the weights \( \omega(X; \gamma) \). The OS estimator \( \hat{\Delta}_1(s) \) may also need to be further re-weighted in a similar to target population of interest. In future work, to facilitate a unified decision in choosing either the RT or the efficient combination as the final estimator, a group sparsity penalty may be considered for \( \eta \) across \( s \) so as allow for simultaneous shrink of all \( \eta \) to 0. Another direction is to extend the proposed methods to accommodate multiple studies of each type.

4.6 Appendix

We assume throughout this appendix the following notations and regularity conditions. The covariates \( X \) belongs to a compact covariate space. Let the score reductions be \( S = s(X; \beta) \), for some finite dimensional \( \beta \), where \( s(x; \beta) \) is some function that is twice continuously differentiable in \( \beta \). Let \( \hat{\beta} \) be a \( n^{1/2} \)-consistent estimator of \( \beta \) with asymptotic limit \( \beta \), and let \( \hat{S} = s(X; \hat{\beta}) \) and \( S = s(X; \beta) \). We also assume that \( g_\pi(u), g_\rho(u), f_{t,z}(s), \mu_{t,z}(s) \), for \( t = 0, 1 \) and \( z = 0, 1 \), are all twice continuously differentiable. Let \( K(\cdot) \) be a bounded, integrable, and differentiable kernel function. Let \( \hat{K}(u) = \partial K(u)/\partial u \) be bounded and integrable.

**Lemma 1.** Let \( \Delta_z(s) = \pi_{1,z}(s) - \pi_{0,z}(s) \), be the asymptotic limits of \( \hat{\Delta}_z(s) \), where \( \pi_{t,z}(s) = \mathbb{E}\{I(T = z)\omega(X; \gamma)^{(1-z)} \overline{Y}_T \mid S = s\} / \mathbb{E}\{I(T = t, Z = z)\omega(X; \gamma)^{(1-z)} \pi_T(X, Z; \alpha)^{-1} \mid \overline{S} = s\} \), for \( t = 0, 1 \) and \( z = 0, 1 \). Let \( \overline{f}_{t,z}(s) \) be the asymptotic limit of the weighted kernel density estimator of \( \overline{S} \), \( \overline{f}_{t,z}(s) = n^{-1} \sum_{j=1}^n K_h(\hat{S}_j - s)I(T_j = t, Z_j = z)\omega(X_j; \gamma)^{(1-z)} \pi_t(X_j, Z_j; \alpha)^{-z} \), for \( t = 0, 1 \) and \( z = 0, 1 \). Assume the reduction is of dimension \( r = 1 \) and \( h = O(n^{-\alpha}) \) with
\( \alpha < 1/2 \). The base estimators admit an asymptotically linear-like representation as:

\[
\hat{\Delta}_0(s) - \overline{\Delta}_0(s) = \frac{1}{nh} \sum_{i=1}^{n} \xi_{i,0,h}(s) + o_p\left( (nh)^{-1/2} \right),
\]

\[
\hat{\Delta}_1(s) - \overline{\Delta}_1(s) = \frac{1}{nh} \sum_{i=1}^{n} \xi_{i,1,h}(s) + o_p\left( (nh)^{-1/2} \right),
\]

where \( \xi_{i,z,h}(s) = \xi_{i,1,z,h}(s) - \xi_{i,0,z,h}(s) \), for \( z = 0, 1 \), with:

\[
\xi_{i,t,z,h}(s) = K\{(\overline{S}_i - s)/h\} \omega(X_i; \overline{\gamma})^{(1-z)} \pi_t(X_i; Z_i; \overline{\alpha})^{-z} I(T_i = t, Z_i = z) \{ Y_j - \overline{\mu}_{t,z}(s) \} \overline{f}_{t,z}(s)^{-1}.
\]

The contributions \( \xi_{i,z,h}(s) \) are independent, identically distributed for a given \( h \), and have conditional mean 0 such that \( \mathbb{E}\{ \xi_{i,z,h}(s) \mid \overline{S} = s \} = 0 \).

**Proof.** We consider the case for \( \hat{\Delta}_0(s) \). The expansion for \( \hat{\Delta}_1(x) \) follows similarly. We first write \( \hat{\Delta}_0(s) - \overline{\Delta}_0(s) \) in terms of \( \hat{\Delta}_0(s) - \overline{\Delta}_0(s) = \{ \hat{\mu}_{1,0}(s) - \overline{\mu}_{1,0}(s) \} - \{ \hat{\mu}_{0,0}(s) - \overline{\mu}_{0,0}(s) \} \).

The centered mean in each group is:

\[
\hat{\mu}_{t,0} - \overline{\mu}_{t,0} = \frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \{ Y_j - \overline{\mu}_{t,0}(s) \} / \hat{f}_{t,0}(s)
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \{ Y_j - \overline{\mu}_{t,0}(s) \} \hat{f}_{t,0}(s) + O_p(n^{-1/2} \mathcal{E}_1),
\]

where \( \hat{f}(s) = n^{-1} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \) and \( \mathcal{E}_1 \) is an error term of the same order as the preceding term. This leaves a weighted kernel regression with estimated covariates \( \hat{S} \) that is centered in the main term. Continuing with the expansion:

\[
\frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \frac{Y_j - \overline{\mu}_{t,0}(s)}{\hat{f}_{t,0}(s)}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \frac{Y_j - \overline{\mu}_{t,0}(s)}{\hat{f}_{t,0}(s)} + O_p\left\{ (nh)^{-1/2} \mathcal{E}_2 \right\},
\]

where \( \mathcal{E}_2 \) is a term of the same order as the preceding term.

\[
\frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \frac{Y_j - \overline{\mu}_{t,0}(s)}{\hat{f}_{t,0}(s)}
\]

\[
= \frac{1}{n} \sum_{j=1}^{n} K_h(\hat{S}_j - s) I(Z_j = 0, T_j = t) \omega(X_j; \overline{\gamma}) \frac{Y_j - \overline{\mu}_{t,0}(s)}{\hat{f}_{t,0}(s)}
\]
where $\mathcal{E}_3$ is a term of the same order as the preceding terms and the last equality follows from application of rates of convergence for derivatives of kernel density estimators (Hansen, 2008b). The $O_p\{n^{-1/2}(\log(n)/(nh^3))^{1/2}\}$ is $o_p\{(nh)^{-1/2}\}$ when $n^{1/2}h \to \infty$, i.e. when $\alpha < 1/2$. Collecting higher order terms, we obtain:

$$
\hat{\mu}_{t,0}(s) - \mu_{t,0}(s) = \frac{1}{n} \sum_{j=1}^{n} K_h(\tilde{S}_j - s)I(Z_j = 0, T_j = t)\omega(X_j; \tilde{\gamma}) \frac{Y_j - \bar{\mu}_{t,0}(s)}{\bar{f}_{t,0}(s)} + O_p\{(nh)^{-1/2}\}
$$

$$
= \frac{1}{n} \sum_{i=1}^{n} \xi_{i,t,0,h}(s) + o_p\{(nh)^{-1/2}\}.
$$

\[\square\]

**Lemma 2.** Under the basic and ideal assumptions (4.2)-(4.6) and the positivity assumption in (4.8), when the PS model (4.10) and trial selection model (4.11) are correctly specified, the RT and OS estimators share a common target estimand:

$$
\overline{\Delta}_0(s) = \overline{\Delta}_1(s) = \mathbb{E}\{Y(1) - Y(0) \mid \tilde{S} = s, Z = 1\}.
$$

In addition, if $\tilde{S}$ is a sufficient reduction in the sense of (4.7), then also $\overline{\Delta}_0(s) = \overline{\Delta}_1(s) = \Delta_1(x)$, for $s = r(x)$. Under only the basic assumptions (4.2)-(4.4), $\overline{\Delta}_0(s)$ is a re-weighted CATE such that:

$$
\overline{\Delta}_0(s) = \mathbb{E}\{\mathbb{E}(Y(1) - Y(0) \mid X, Z = 0)\mathbb{P}(Z = 0 \mid X)\omega(X; \tilde{\gamma}) \mid \tilde{S} = s\}/\overline{\pi}(s)
$$

$$
= \mathbb{E}_X|\tilde{S} = s\{\mathbb{E}(Y(1) - Y(0) \mid X, Z = 0) \mid \tilde{S} = s\},
$$

where $\overline{\pi}(s) = \mathbb{E}\{\mathbb{P}(Z = 0 \mid X)\omega(X; \tilde{\gamma}) \mid \tilde{S} = s\}$ and $\mathbb{E}_{X|\tilde{S} = s}(\cdot \mid \tilde{S} = s)$ denotes a conditional expectation under a tilted conditional distribution for $X$, characterized by the tilted density $\tilde{f}(x \mid \tilde{S} = s) = \mathbb{P}(Z = 0 \mid X = x)\omega(x; \tilde{\gamma})/\overline{\pi}(s)f(x \mid \tilde{S} = s).$
Proof. We first consider the case where the full assumptions are met, except that $\overline{S}$ may not be a sufficient reduction. The limiting estimand for the RT estimator is $\overline{\Delta}_0(s) = \overline{\mu}_{t,0}(s) - \overline{\mu}_{0,0}$ with:

$$\overline{\mu}_{t,0}(s) = \mathbb{E} \left\{ I(Z = 0, T = t)\omega(X; \overline{\gamma}) Y \mid \overline{S} = s \right\} / \mathbb{E} \left\{ I(Z = 0, T = t)\omega(X; \overline{\gamma}) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y \mid X, T = t, Z = 0)\mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\} / \mathbb{E} \left\{ \mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y(t) \mid X, Z = 1)\mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\} / \mathbb{E} \left\{ \mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \{ Y(t) \mid \overline{S} = s, Z = 1 \},$$

where the second equality uses randomization (4.2) and that the trial selection model (4.11) is correctly specified, and the third equality uses assumptions (4.2)-(4.3) and NUCA (4.5). The positivity assumptions (4.4) and (4.8) are implicitly assumed throughout. For the OS estimator, the limiting estimand is $\overline{\Delta}_1(s) = \overline{\mu}_{t,1}(s) - \overline{\mu}_{0,1}(s)$ with:

$$\overline{\mu}_{t,1}(s) = \mathbb{E} \left\{ I(Z = 1, T = t)\pi_t(X, 1; \alpha)^{-1} Y \mid \overline{S} = s \right\} / \mathbb{E} \left\{ I(Z = 1, T = t)\pi_t(X, 1; \alpha)^{-1} \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y \mid X, T = t, Z = 1)\mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\} / \mathbb{E} \left\{ \mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y(t) \mid X, Z = 1)\mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\} / \mathbb{E} \left\{ \mathbb{P}(Z = 1 \mid X) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \{ Y(t) \mid \overline{S} = s, Z = 1 \},$$

where the second equality uses the correctly specified PS model (4.10), and the third equality uses consistency (4.3) and NUCA (4.5). When $\overline{S}$ is sufficient as in (4.7), then $\overline{\Delta}_0(s) = \overline{\Delta}_1(s) = \Delta_1(x)$ directly for $s = r(x)$. Under only the basic assumptions, $\overline{\Delta}_0(s) = \overline{\mu}_{t,0}(s) - \overline{\mu}_{0,0}(s)$ with:

$$\overline{\mu}_{t,0}(s) = \mathbb{E} \left\{ I(Z = 0, T = t)\omega(X; \overline{\gamma}) Y \mid \overline{S} = s \right\} / \mathbb{E} \left\{ I(Z = 0, T = t)\omega(X; \overline{\gamma}) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y \mid X, T = t, Z = 0)\mathbb{P}(Z = 0 \mid X)\omega(X; \overline{\gamma}) \mid \overline{S} = s \right\} / \mathbb{E} \left\{ \mathbb{P}(Z = 0 \mid X)\omega(X; \overline{\gamma}) \mid \overline{S} = s \right\}$$

$$= \mathbb{E} \left\{ \mathbb{E}(Y(t) \mid X, Z = 0)\mathbb{P}(Z = 0 \mid X)\omega(X; \overline{\gamma}) \mid \overline{S} = s \right\} / \pi(s)$$

$$= \mathbb{E}_{X \mid S = s} \{ \mathbb{E}(Y(t) \mid X, Z = 0) \mid \overline{S} = s \},$$

where we used randomization (4.2) in the second equality and (4.2)-(4.3) in the third equality. Again the positivity assumptions (4.4) and (4.8) are implicit.
Lemma 3. Under the positivity conditions (4.4) and (4.8), if $Y$ is a bounded outcome, then the MSE criterion can be expressed as:

$$Q_s(\eta) = \frac{1}{nh^2} \mathbb{E} \left[ \xi_{i,0,h}(s) + \eta \{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \} \right]^2 + \eta^2 \left\{ \overline{\Delta}_1(s) - \overline{\Delta}_0(s) \right\}^2 + o_p\{(nh)^{-1/2}\}. $$

Proof. Using Lemma 1, \( \hat{\Delta}(s; \eta) \) also admits an asymptotically linear-like representation as:

$$\hat{\Delta}(s; \eta) - \overline{\Delta}_0(s) = \overline{\Delta}_0(s) - \overline{\Delta}_0(s) + \eta \left\{ \hat{\Delta}_1(s) - \overline{\Delta}_1(s) - \overline{\Delta}_0(s) + \overline{\Delta}_0(s) \right\} + \eta \left\{ \overline{\Delta}_1(s) - \overline{\Delta}_0(s) \right\}$$

$$= \frac{1}{nh} \sum_{i=1}^{n} \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} + \eta \left\{ \overline{\Delta}_1(s) - \overline{\Delta}_0(s) \right\} + o_p\{(nh)^{-1/2}\}. $$

By the positivity and boundedness assumptions, the \( o_p\{(nh)^{-1/2}\} \) term can also be bounded. This allows terms converging in probability to 0 to also converge in expectation to 0 so that:

$$\mathbb{E} \left\{ \hat{\Delta}(s; \eta) - \overline{\Delta}_0(s) \right\}^2 = \mathbb{E} \left[ \frac{1}{nh} \sum_{i=1}^{n} \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} \right]^2 + \left[ \eta \left\{ \overline{\Delta}_1(s) - \overline{\Delta}_0(s) \right\} \right]^2 + o\{(nh)^{-1/2}\}$$

$$= \frac{1}{nh^2} \mathbb{E} \left[ \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} \right]^2 + \eta^2 \left\{ \overline{\Delta}_1(s) - \overline{\Delta}_0(s) \right\}^2 + o\{(nh)^{-1/2}\}. $$

\( \square \)

Theorem 1. Let \( \beta \in (0, \frac{1}{2}) \). The minimizer \( \bar{\eta} = \arg\min_{\eta} \hat{Q}_s(\eta) \) is an adaptive weight in that \( \bar{\eta} = O_p\{n^{\alpha+2\beta-1}\} \) when \( \overline{\Delta}_0(s) - \overline{\Delta}_1(s) \neq 0 \), and \( \bar{\eta} - \bar{\eta} = O_p\{(nh)^{-1/2}\} \) when \( \overline{\Delta}_0(s) - \overline{\Delta}_1(s) = 0 \), where \( \bar{\eta} \) is the weight such that the asymptotic variance of \( \Delta(s; \eta) \) is minimized.

Proof. We first consider the optimal asymptotic variance of \( \hat{\Delta}(s) \) in the case where \( \overline{\Delta}_0(s) = \overline{\Delta}_1(s) \). In this case, the normalized and centered estimator is given by:

$$\left\{ nh \right\}^{1/2} \left\{ \hat{\Delta}(s) - \overline{\Delta}_0(s) \right\} = \left\{ nh \right\}^{1/2} \left[ \overline{\Delta}_0(s) - \overline{\Delta}_0(s) + \eta \left\{ \hat{\Delta}_1(s) - \overline{\Delta}_1(s) - \overline{\Delta}_0(s) + \overline{\Delta}_0(s) \right\} \right]$$

$$= \left\{ nh \right\}^{-1/2} \sum_{i=1}^{n} \xi_{i,0,h}(s) + \eta \left\{ \xi_{i,1,h}(s) - \xi_{i,0,h}(s) \right\} + o(1),$$

which can be shown using Liapunov’s central limit theorem to converge in distribution to a mean 0 normal distribution with asymptotic variance.

107
$h^{-1}\mathbb{E} [\xi_{i,0,h}(s) + \eta \{\xi_{i,1,h}(s) - \xi_{i,0,h}(s)\}]^2$. The minimizer in $\eta$ of this asymptotic variance is then given by:

$$\eta^* = \frac{\mathbb{E} \xi_{i,0,h}(s)^2}{\mathbb{E} \{\xi_{i,1,h}(s) - \xi_{i,0,h}(s)\}^2}.$$  

Now, regardless of scenario, the initial estimator can be written as:

$$\tilde{\eta} = \sum_{i=1}^{n} \xi_{i,0,h}(s)^2 \left[ \sum_{i=1}^{n} \left\{ \tilde{\xi}_{i,0,h}(s) - \tilde{\xi}_{i,1,h}(s) \right\}^2 + n^{2(1-\beta)}h^2 \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 \right]$$

By scaling the numerator and denominator by $(nh)^{-1}$ using standard arguments for consistency of kernel smoothing estimators it can be shown that $\tilde{\eta} - \eta^* = O_{p}\{(nh)^{-1/2}\}$, provided that the correction term after normalizing $(nh)^{-1}n^{2(1-\beta)}h^2 \left\{ \Delta_1(s) - \Delta_0(s) \right\} = o_{p}(1)$. But when $\Delta_0(s) = \Delta_1(s)$:

$$(nh)^{-1}n^{2(1-\beta)}h^2 \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 = n^{1-2\beta}h \left\{ \Delta_1(s) - \Delta_1(s) - \Delta_0(s) + \Delta_0(s) \right\}^2$$

$$= n^{1-2\beta}h O_{p}\{(nh)^{-1}\}$$

$$= O_{p}(n^{-2\beta}),$$

which is $o_{p}(1)$ when $\beta > 0$. On the other hand if $\Delta_0(s) \neq \Delta_1(s)$ then:

$$\tilde{\eta} = (nh)^{-1}\sum_{i=1}^{n} \xi_{i,0,h}(s)^2 / \left[ (nh)^{-1}\sum_{i=1}^{n} \left\{ \tilde{\xi}_{i,0,h}(s) - \tilde{\xi}_{i,1,h}(s) \right\}^2 + n^{1-2\beta}h \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 \right]$$

$$= O_{p}(1) / \left[ O_{p}(1) + n^{1-2\beta}h \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 \right].$$

But the corrected bias term is of order:

$$n^{1-2\beta}h \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 = n^{1-2\beta}h \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2$$

$$+ n^{1-2\beta}h \left[ \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 - \left\{ \Delta_1(s) - \Delta_0(s) \right\}^2 \right]$$

$$= O_{p}(n^{1-2\beta-\alpha}),$$

which diverges for $\beta < (1 - \alpha)/2$. The result followings by scaling the numerator and denominator of $\tilde{\eta}$ both by $n^{\alpha+2\beta-1}$.

\[\square\]
References


URL http://CRAN.R-project.org/package=rpart.plot


the discrepancy between observational studies and the women’s health initiative clinical trial. *American journal of epidemiology* **162** 404–414.


Yoshida, K. and Bohn, J. (2015). Tableone: Create table 1 to describe baseline characteristics. *R package version 0.7 3*.


