A Unified Framework for Efficient Estimation of General Treatment Models *

Chunrong Ai [†⋆], Oliver Linton [‡⋆], Kaiji Motegi [§†], and Zheng Zhang [¶‡]

[⋆]*School of Management and Economics, Chinese University of Hong Kong, Shenzhen*
[⋆]*Faculty of Economics, University of Cambridge*
[†]*Graduate School of Economics, Kobe University*
[‡]*Center for Applied Statistics, Institute of Statistics & Big Data, Renmin University of China*

February 15, 2021

Abstract

This paper presents a weighted optimization framework that unifies the binary, multi-valued, and continuous treatment—as well as mixture of discrete and continuous treatment—under unconfounded treatment assignment. With a general loss function, the framework includes the average, quantile and asymmetric least squares causal effect of treatment as special cases. For this general framework, we first derive the semiparametric efficiency bound for the causal effect of treatment, extending the existing bound results to a wider class of models. We then propose a generalized optimization estimator for the causal effect with weights estimated by solving an expanding set of equations. Under some sufficient conditions, we establish the consistency and asymptotic normality of the proposed estimator of the causal effect and show that the estimator attains the semiparametric efficiency bound, thereby extending the existing literature on efficient estimation of causal effect to a wider class of applications. Finally, we discuss estimation of some causal effect functionals such as the treatment effect curve and the average outcome. To evaluate the finite sample performance of the proposed procedure, we conduct a small-scale simulation study and find that the proposed estimation has practical value. In an empirical application, we detect a significant causal effect of political advertisements on campaign contributions in the binary treatment model, but not in the continuous treatment model.

*All authors contribute to the paper equally.

[†]E-mail: chunrongai@cuhk.edu.cn
[‡]E-mail: obl20@cam.ac.uk
[§]E-mail: motegi@econ.kobe-u.ac.jp
[¶]E-mail: zhengzhang@ruc.edu.cn

Keywords: Causal effect; Entropy maximization; Treatment effect; Semiparametric efficiency; Sieve method; Stabilized Weights.

1 Introduction

Modeling and estimating the causal effect of certain treatments or policies is of major interest in economics and social science more generally (see, e.g., Hirano, Imbens, and Ridder, 2003, Imbens, 2004, Abadie, 2005, Heckman and Vytlacil, 2005, Chernozhukov, Fernández-Val, and Melly, 2013, Athey, Imbens, and Wager, 2018, Wager and Athey, 2018). Most existing studies focus on the binary treatment where an individual either receives the treatment or does not, ignoring the treatment intensity. In many applications, however, the treatment intensity is a part of the treatment, and its causal effect is also of great interest to decision makers. For example, in evaluating how financial incentives affect health care providers, the causal effect may depend on not only the introduction of incentive but also the level of incentive. Similarly, in studying how taxes affect addictive substance usages, the causal effect may depend not only on the imposition of tax but also on the tax rate. In finance, there are many plausible examples of interest. For example, in evaluating the effect of corporate bond purchase schemes on market quality, the causal effect may depend not just on whether the bond is selected into the scheme but on how much of it is purchased (see Boneva, Elliott, Kaminska, Linton, McLaren, and Morley, 2018).

In recognition of the importance of the treatment intensity, the binary treatment literature has been extended to the multi-valued treatment (e.g., Imbens, 2000, Cattaneo, 2010) and continuous treatment (e.g., Hirano and Imbens, 2004, Imai and van Dyk, 2004, Florens, Heckman, Meghir, and Vytlacil, 2008, Fong, Hazlett, and Imai, 2018, Yiu and Su, 2018). The parameter of primary interest in this literature is the average causal effect of treatment, defined as the difference in response to two levels of treatment by the same individual, averaged over a set of individuals. The identification and estimation difficulty is that each individual only receives one level of treatment. To overcome this difficulty, researchers impose the *unconfounded treatment assignment* condition, which allows them to find statistical matches for each observed individual from all other treatment levels.

The main objective of this paper is to present a weighted optimization estimation framework that unifies the binary, multi-valued, and continuous treatment—as well as the mixture of discrete and continuous treatment—and to identify and estimate causal effect parameters through a population minimization problem under the unconfounded treatment assignment condition. The weights are called the *stabilized weights* by Robins, Hernán, and Brumback (2000) and are defined as the ratio of the marginal probability distribution

of the treatment status over the conditional probability distribution of the treatment status given covariates. We first compute the semiparametric efficiency bound (Bickel, Klaassen, Ritov, and Wellner, 1993) of the causal effect of treatment, extending the results of Hahn (1998), Firpo (2007), and Cattaneo (2010) from the binary treatment to a variety of treatments and to parameters defined through a population minimization problem. Our bound reveals that the weighted optimization with known stabilized weights does not produce efficient estimation since it fails to account for the information restricting the stabilized weights. This observation was made by Hirano, Imbens, and Ridder (2003) in the binary treatment case; here we show that their observation holds true for a much wider class of treatment models. We exploit the information that the stabilized weights satisfy certain moment conditions (an expanding number thereof) by estimating the stabilized weights from those equations by a novel entropy maximization method; we then estimate the causal effect by the generalized optimization method with the true stabilized weights replaced by the estimated weights. Under some sufficient conditions, we show that our proposed estimator is consistent and asymptotically normally distributed and, more importantly, it attains the semiparametric efficiency bound. We propose consistent standard errors based on the same sieve methodology. We propose a tuning parameter selection methodology to guide the practical implementation. We also discuss estimation of the full nonparametric effect curve and establish its pointwise asymptotic normality and uniform consistency.

We next present some simulation evidence that the proposed methodology operates well in finite samples and is robust to misspecification, whereas the existing methodology of Fong, Hazlett, and Imai (2018) is somewhat fragile. We apply our methodology to the study of the effect of political advertisements on campaign contributions using data considered by Urban and Niebler (2014) and Fong, Hazlett, and Imai (2018). We detect a significant causal effect of advertisements on contributions in the binary treatment model, but not in the continuous treatment model. The former result is consistent with Urban and Niebler (2014), while the latter is consistent with Fong, Hazlett, and Imai (2018).

Literature review. In the binary treatment case with *unconfounded treatment assignment*, the average causal effect is estimated by the difference of the weighted average responses with the propensity scores as weights (see, e.g., Rosenbaum and Rubin, 1983, Hirano, Imbens, and Ridder, 2003). Other popular methods include regression adjustment (Angrist and Pischke, 2008), matching (Imbens, 2004, Abadie and Imbens, 2006, 2016) , imputation (Heckman, Ichimura, and Todd, 1998, Cattaneo and Farrell, 2011), and hybrid method (Farrell, 2015, Słoczyński and Wooldridge, 2018). The efficiency bound of the average causal effect in this model is derived by Robins, Rotnitzky, and Zhao (1994) and Hahn (1998), and efficient estimation is proposed by Robins, Rotnitzky, and Zhao (1994), Hahn

(1998), Hirano, Imbens, and Ridder (2003), Graham, Pinto, and Egel (2012), and Chan, Yam, and Zhang (2016). Of particular interest in this literature is the study by Hirano, Imbens, and Ridder (2003) which shows that the weighted average difference estimator attains the semiparametric efficiency bound if the weights are estimated by the empirical likelihood estimation.

In the multi-valued treatment case, Imbens (2000) generalizes the propensity score, and Cattaneo (2010) derives the efficiency bound and proposes an estimator that attains the efficiency bound. In the continuous treatment case, Hirano and Imbens (2004) and Imai and van Dyk (2004) parameterize the generalized propensity score function and propose a consistent estimator of the average causal effect. Their estimators are not efficient and could be biased if the generalized propensity score function is misspecified. Florens, Heckman, Meghir, and Vytlacil (2008) use a control function approach to identify the average causal effect in the continuous treatment and propose a consistent estimation. It is unclear if their estimation is efficient. Galvao and Wang (2015) estimate the continuous treatment effect through stabilized weighting. They do not study how to construct the stabilized weights such that their estimator is efficient. Kennedy, Ma, McHugh, and Small (2017) propose a nonparametric kernel estimator for the treatment effects curve, again the efficient estimation is still unclear. Fong, Hazlett, and Imai (2018) propose an estimator of the average causal effect of continuous treatment but do not establish consistency of their estimation. In fact, their simulation results indicate their estimation could be seriously biased. Yiu and Su (2018) study the average causal effect of both discrete and continuous treatment by parameterizing the propensity score. Their estimator is generally biased if their parameterization is incorrect.

In addition to the average causal effect of treatment (ATE), it is also important to investigate the distributional impact of treatment. For instance, a decision maker may be interested in the causal effect of a treatment on the outcome dispersion or on the lower tail of the outcome distribution. Firpo (2007) computes the efficiency bound and proposes an efficient estimation of quantile causal effect of treatment (QTE) for the binary treatment. For additional studies on QTE, we refer to Chernozhukov and Hansen (2005), Angrist and Pischke (2008), and Donald and Hsu (2014).

To the best of our knowledge, we are unaware of any previous work that computes the efficiency bound and proposes efficient estimation of the causal effect in the continuous treatment or mixture of discrete and continuous treatment under a general minimization problem that permits ATE and QTE. The present paper fills this gap rigorously.

The paper is organized as follows. Section 2 sets up the basic framework, Section 3 computes the semiparametric efficiency bound of the causal effect of treatment, Section 4

presents the generalized optimization estimator, Section 5 establishes the large sample properties of the proposed estimator. Section 6 constructs confidence intervals based on plug-in and simulation-based approaches. In Section 7 we propose two data-driven approaches for selecting tuning parameters. In Section 8 we discuss some extensions. Section 9 reports on a simulation study, while Section 10 presents an empirical application, followed by some concluding remarks in Section 11. All technical proofs and extra simulation results are relegated to the supplemental material Ai, Linton, Motegi, and Zhang (2020).

2 Basic framework and notation

Let T denote the observed treatment status variable with support $\mathcal{T} \subset \mathbb{R}$, where $\mathcal{T}$ is either a discrete set, a continuum, or a mixture of discrete and continuum subsets, and T has a marginal probability distribution function $F_T(t)$. Let $Y^*(t)$ denote the potential response when treatment $T = t$ is assigned. Let $L(\cdot)$ denote a known convex loss function whose derivative, denoted by $L'(\cdot)$, exists almost everywhere. For the leading part of the paper, we shall maintain that there exists a parametric causal effect function $g(t;\beta)$ with the unknown value $\boldsymbol{\beta}^* \in \mathbb{R}^p$ (with $p \in \mathbb{N}$) uniquely solving the minimization problem below, i.e.,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \int_{\mathcal{T}} \mathbb{E}\left[L\left(Y^*(t) - g(t;\boldsymbol{\beta})\right)\right] dF_T(t). \tag{2.1}$$

The parameterization of the causal effect is restrictive, but quite common in applications. Some extensions to the unspecified causal effect function shall be discussed later in the paper (see Section 8).

Model (2.1) includes many prominent models in the literature as special cases. For example, it includes: the average causal effect of binary treatment studied in Hahn (1998) and Hirano, Imbens, and Ridder (2003) (i.e., $\mathcal{T} = \{0,1\}$, $L(v) = v^2$ and $g(t;\ \boldsymbol{\beta}) = \beta_0 + \beta_1 t$), the quantile causal effect of binary treatment studied in Firpo (2007) (i.e., $\mathcal{T} = \{0,1\}$, $L(v) = v(\tau - I(v \leq 0))$ is an almost everywhere differentiable function with $\tau \in (0,1)$ and $g(t;\boldsymbol{\beta}) = t\beta_1 + (1-t)\beta_0$), the average causal effect of multi-valued treatment studied in Cattaneo (2010) (i.e., $\mathcal{T} = \{0,1,\ldots,J\}$ for some $J \in \mathbb{N}$, $L(v) = v^2$ and $g(t;\boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j I(t = j)$), and the average causal effect of continuous treatment studied in Hirano and Imbens (2004) (i.e., $L(v) = v^2$ and $\mathbb{E}[Y^*(t)] = g(t;\boldsymbol{\beta})$ is a parametric model indexed by $\boldsymbol{\beta}$ for the potential outcome means, which is also called a *marginal structural model* in Robins, Hernán, and Brumback (2000). Examples include the linear marginal structure model $\mathbb{E}[Y^*(t)] = \beta_0 + \beta_1 \cdot t$, and the nonlinear marginal structure model $\mathbb{E}[Y^*(t)] = \beta_0 \cdot t + 1/(t+\beta_1)^2$ studied in Hirano and Imbens (2004)). It also includes the

quantile causal effect of multi-valued (i.e., $L(v) = v(\tau - I(v \leq 0))$ with $\tau \in (0,1)$ and $g(t;\boldsymbol{\beta}) = \sum_{j=0}^{J} \beta_j I(t=j)$) and continuous treatment (i.e., $L(v) = v(\tau - I(v \leq 0))$ and $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = g(t;\boldsymbol{\beta})$ is a parametric model indexed by $\boldsymbol{\beta}$ for the potential outcome quantiles. Examples include the linear model $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = \beta_0 + \beta_1 \cdot t$ and the Box-Cox transformation model $\inf\{q : \mathbb{P}(Y^*(t) \geq q) \leq \tau\} = h_\lambda(\beta_0 + \beta_1 \cdot t)$ studied in Buchinsky (1995), where $h_\lambda(z) = (\lambda z + 1)^{-1/\lambda}$). The latter has so far not been covered by the existing literature. Moreover, with $L(v) = v^2 |\tau - I(v \leq 0)|$, our framework covers the asymmetric least squares estimation of the causal effect of (binary, multi-valued, continuous, mixture of discrete and continuous) treatment. The asymmetric least squares regression received attention from some noted econometricians (see Newey and Powell, 1987) but zero attention in the causal effect literature. Our framework can also accommodate vector-valued treatment T and the inclusion of multiple variables in $g(\cdot)$, although they would add to the dimensionality problem. (We are grateful for an anonymous referee for pointing out these possible extensions.)

The problem with (2.1) is that the potential outcome $Y^*(t)$ is not observed for all t. Let $Y := Y^*(T)$ denote the observed response. One may attempt to solve the following optimization problem:

$$\min_{\boldsymbol{\beta}} \mathbb{E}[L(Y - g(T;\boldsymbol{\beta}))].$$

However, if there exists a selection into treatment, the true value $\boldsymbol{\beta}_0$ does not solve the above minimization problem. Indeed, in this case, the observed response and treatment assignment data alone cannot identify $\boldsymbol{\beta}^*$. To address this identification issue, studies in the literature impose a selection on observable condition (e.g., Hirano, Imbens, and Ridder, 2003, Imai and van Dyk, 2004, Fong, Hazlett, and Imai, 2018). Specifically, let $\boldsymbol{X}$ denote a vector of covariates. The following condition shall be maintained throughout the paper.

Assumption 1 (*Unconfounded Treatment Assignment*)**.** *T is independent of $Y^*(t)$ for all $t \in \mathcal{T}$ given $\boldsymbol{X}$, i.e., $Y^*(t) \perp T|\boldsymbol{X}$.*

Let $F_{T|X}$ denote the conditional probability distribution of T given the observed covariates $\boldsymbol{X}$ and let $dF_{T|X}$ denote the corresponding probability measure. In the literature, $dF_{T|X}$ is called the *generalized propensity score* (Hirano and Imbens, 2004, Imai and van Dyk, 2004). Suppose that $dF_{T|X}(T|\boldsymbol{X})$ is positive everywhere and let

$$\pi_0(T, \boldsymbol{X}) := \frac{dF_T(T)}{dF_{T|X}(T|\ \boldsymbol{X})}.$$

The function $\pi_0(T, \boldsymbol{X})$ is called the *stabilized weight* in Robins, Hernán, and Brumback

(2000). Under Assumption 1, we obtain

$$\mathbb{E}[\pi_0(T, \boldsymbol{X})L(Y - g(T; \boldsymbol{\beta}))] = \int \mathbb{E}\left[L(Y^*(t) - g(t; \boldsymbol{\beta}))\right] dF_T(t) \tag{2.2}$$

(see Appendix A for derivation), and hence the true value $\boldsymbol{\beta}^*$ solves the weighted optimization problem

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathbb{E}[\pi_0(T, \boldsymbol{X})L(Y - g(T; \boldsymbol{\beta}))]. \tag{2.3}$$

This result is very insightful. It tells us that the selection bias in the *unconfounded treatment assignment* can be corrected through covariate-balancing. More importantly, it says that the true value $\boldsymbol{\beta}^*$ can be identified from the observed data. The weighted optimization (2.3) provides a unified framework for estimating the causal effect of a variety of treatments, including binary, multi-level, continuous, and mixture of discrete and continuous treatment, and under a general loss function. The goal of this paper is to compute the semiparametric efficiency bound and to present an efficient estimator for $\boldsymbol{\beta}^*$ under this general framework.

Although the parametric specification of $g(t; \boldsymbol{\beta})$ is somewhat restrictive, it is useful from a practical point of view. First, if T is a discrete variable, model misspecification is not an issue since the coefficient $\boldsymbol{\beta}^*$ has a clear causal interpretation. Second, if T is a continuous variable, usually a parametric specification may suffer from the model misspecification problem. Since T is univariate, the true response model can be well approximated through several polynomials of t. Third, a parametric specification of $g(t; \boldsymbol{\beta})$ allows us to infer the parameters at $\sqrt{N}$-consistent rate and to construct the most efficient estimator. Fourth, the proposed framework (2.1) is more general than the existing literature of continuous treatment (Hirano and Imbens, 2004, Fong, Hazlett, and Imai, 2018), where either a regression model $\mathbb{E}[Y|T, \boldsymbol{X}]$ or a response model $\mathbb{E}[T|\boldsymbol{X}]$ is often required. In Section 8, we also consider fully nonparametric estimation of $g(t)$ under several important cases. The fully nonparametric estimation of $g(t)$ within the general framework (2.1) is beyond the scope of this article, and it will be pursued in a future work.

3 Efficiency bound

We begin by applying the approach of Bickel, Klaassen, Ritov, and Wellner (1993) to compute the semiparametric efficiency bound of the parameter $\boldsymbol{\beta}^*$ defined by (2.1) under Assumption 1. This gives the least possible variance achievable by a regular estimator in the semiparametric model. The result is presented in the following theorem.

Theorem 1. *Suppose that $g(T;\boldsymbol{\beta})$ is twice differentiable with respect to $\boldsymbol{\beta}$ in the parameter space $\Theta \subset \mathbb{R}^p$, with $m(T;\boldsymbol{\beta}^*) := \nabla_\beta g(T;\boldsymbol{\beta}^*)$, and $\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|Y,\boldsymbol{X}\right]$ is differentiable with respect to $\boldsymbol{\beta} \in \Theta$. Denote $\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*) := \mathbb{E}[L'(Y-g(T;\boldsymbol{\beta}^*))|T,\boldsymbol{X}]$, $H_0 := -\nabla_\beta \mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\ \boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]\big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$, and*

$$\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) := \pi_0(T,\ \boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'(Y-g(T;\boldsymbol{\beta}^*)) - \pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)$$
$$+\mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|T\right] + \mathbb{E}\left[\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta}^*)\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)|\boldsymbol{X}\right].$$

Suppose that H_0 is nonsingular and $\mathbb{E}\left[\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)\psi(Y,T,\boldsymbol{X};\ \boldsymbol{\beta}^*)^\top\right]$ exists and is finite. Under Assumption 1, namely $Y^*(t) \perp T|\boldsymbol{X}$ for all $t \in \mathcal{T}$, and model (2.1), the efficient influence function of $\boldsymbol{\beta}^*$ is given by

$$S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) = H_0^{-1}\psi(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*).$$

Consequently, the efficient variance bound of $\boldsymbol{\beta}^*$ is

$$V_{eff} = \mathbb{E}\left[S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)S_{eff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)^\top\right]. \tag{3.1}$$

The proof of Theorem 1 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2020, Section 2.1). It is worth noting that our bound V_{eff} is equal to: the bound of Hahn (1998) for the case of binary average treatment, the bound of Cattaneo (2010) for the case of multi-valued average treatment, and the bound of Firpo (2007) for the case of binary quantile treatment (see Ai, Linton, Motegi, and Zhang, 2020, Sections 2.2-2.4). Moreover, our bound applies to a much wider class of models, including quantile causal effect of multi-valued, continuous, and mixture of discrete and continuous treatment as well as the asymmetric least squares estimation of the causal effect of all kinds of treatments.

Based on the expression of the efficient influence function, many papers construct an efficient estimator by solving the estimated efficient score equation (Athey, Imbens, Pham, and Wager, 2017, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins, 2018). Such estimators typically have the double or multiple robustness property. However, in our case the efficient influence function $S_{eff}(T,\ \boldsymbol{X},Y;\boldsymbol{\beta})$ involves five unknown functionals $f_T(T)$, $f_{T|X}(T|\boldsymbol{X})$, $\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta})$, $\mathbb{E}[\pi_0(T,\boldsymbol{X})\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta})m(T,\boldsymbol{\beta})|T]$, and $\mathbb{E}[\pi_0(T,\boldsymbol{X})\varepsilon(T,\boldsymbol{X};\boldsymbol{\beta})m(T,\boldsymbol{\beta})|\boldsymbol{X}]$. Estimation of these functionals is difficult in practice, and we expect that the finite sample performance of the estimated $\boldsymbol{\beta}^*$ would be poor. Instead of explicitly estimating the efficient influence function S_{eff}, we propose a simple weighted optimization estimator based on (2.3) by estimating the stabilized weights $\pi_0(T,\boldsymbol{X})$. This procedure is remarkably stable numerically and performs well statistically in small samples as we demonstrate in the Monte Carlo section.

It is also worth noting that, if the stabilized weights are known and $g(t;\boldsymbol{\beta}^*)$ is correctly specified, one can estimate $\boldsymbol{\beta}^*$ by solving the sample analogue of the weighted optimization (2.3). The asymptotic variance of this estimator is

$$V_{ineff} = \mathbb{E}\left[S_{ineff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)S_{ineff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*)^\top\right],$$

with

$$S_{ineff}(Y,T,\boldsymbol{X};\boldsymbol{\beta}^*) = H_0^{-1}\cdot\pi_0(T,\boldsymbol{X})m(T;\boldsymbol{\beta}^*)L'\left\{Y-g(T;\boldsymbol{\beta}^*)\right\}.$$

It is easy to show that $V_{ineff} > V_{eff}$ (see Proposition C.1 of Appendix C), implying that the weighted optimization estimator is not efficient. This follows because the weighted optimization does not account for the restriction on the stabilized weight $\pi_0(t,\boldsymbol{x})$ that

$$\mathbb{E}\left[\pi_0(T,\boldsymbol{X})u(T)v(\boldsymbol{X})\right] = \mathbb{E}[u(T)]\cdot\mathbb{E}[v(\boldsymbol{X})] \tag{3.2}$$

holds for any suitable functions $u(t)$ and $v(\boldsymbol{x})$. Incorporating restriction (3.2) into the estimation of the causal effect can improve efficiency. A similar observation was made by Hirano, Imbens, and Ridder (2003) in the binary treatment. Exactly how to incorporate restriction (3.2) into the estimation is the subject of the next section.

4 Efficient estimation

One way to incorporate (3.2) into the estimation is to estimate the stabilized weights from (3.2) and then implement (2.3) with the estimated weights. But before doing so, we must verify that (3.2) uniquely identifies $\pi_0(T,\boldsymbol{X})$.

Theorem 2. *For any integrable functions $u(T)$ and $v(\boldsymbol{X})$, $\mathbb{E}\left[\pi(T,\boldsymbol{X})u(T)v(\boldsymbol{X})\right] = \mathbb{E}[u(T)]\cdot\mathbb{E}[v(\boldsymbol{X})]$ holds if and only if $\pi(T,\boldsymbol{X}) = \pi_0(T,\boldsymbol{X})$ a.s.*

The proof is presented in Appendix B. Therefore, condition (3.2) identifies the stabilized weights. The challenge now is that (3.2) implies an infinite number of moment conditions. With a finite sample of observations, it is impossible to solve an infinite number of equations. To overcome this difficulty, we approximate the (infinite dimensional) function space with the (finite dimensional) sieve space. Specifically, let $u_{K_1}(T) = (u_{K_1,1}(T),\ldots,u_{K_1,K_1}(T))^\top$ and $v_{K_2}(\boldsymbol{X}) = (v_{K_2,1}(\boldsymbol{X}),\ldots,v_{K_2,K_2}(\boldsymbol{X}))^\top$ denote the known basis functions with dimensions $K_1\in\mathbb{N}$ and $K_2\in\mathbb{N}$ respectively, and let $K := K_1\cdot K_2$. The functions $u_{K_1}(t)$ and $v_{K_2}(\boldsymbol{x})$ are called the *approximation sieves* that can approximate any suitable functions $u(t)$ and $v(\boldsymbol{x})$ arbitrarily well (see Newey, 1997, Chen, 2007, for more

discussion on sieve approximation). Since the sieve approximating space is also a subspace of the function space, $\pi_0(T, \boldsymbol{X})$ satisfies

$$\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) u_{K_1}(T) v_{K_2}(\boldsymbol{X})^\top\right] = \mathbb{E}[u_{K_1}(T)] \cdot \mathbb{E}[v_{K_2}(\boldsymbol{X})]^\top. \tag{4.1}$$

Let $\{T_i, \boldsymbol{X}_i, Y_i\}_{i=1}^N$ denote an independently and identically distributed sample of observations drawn from the joint distribution of $(T, \mathbf{X}, Y)$. We propose to estimate the stabilized weights $\pi_i = \pi_0(T_i, \mathbf{X}_i)$ by solving the entropy maximization problem

$$\begin{cases} \max\left\{-\sum_{i=1}^N \pi_i \log \pi_i\right\} \\ \text{subject to } \frac{1}{N}\sum_{i=1}^N \pi_i u_{K_1}(T_i) v_{K_2}(\boldsymbol{X}_i)^\top = \left(\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\right)\left(\frac{1}{N}\sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j)^\top\right). \end{cases} \tag{4.2}$$

Noting $\sum_{i=1}^N N^{-1}\pi_i = 1$ (since both $u_{K_1}(T)$ and $v_{K_2}(\boldsymbol{X})$ contain the constant 1) and

$$\max\left\{-\sum_{i=1}^N \pi_i \log \pi_i\right\} = -\min\left\{\sum_{i=1}^N \{N^{-1}\pi_i\} \cdot \log \frac{N^{-1}\pi_i}{N^{-1}}\right\},$$

the formulation (4.2) can be interpreted as the minimization of the Kullback-Leibler divergence between the estimated weights $\{N^{-1}\pi_i\}_{i=1}^N$ and the empirical frequencies $\{N^{-1}\}$ subject to the empirical moment constraints (4.1). This idea is similar to the exponential tilting (ET) idea developed in Kitamura and Stutzer (1997) and Imbens, Spady, and Johnson (1998). The difference is that they consider a parametric problem and we consider a nonparametric problem.

The primal problem (4.2) is hard to solve numerically. We instead consider its dual problem, which can be solved by numerically efficient and stable algorithms. Specifically, let $\rho(v) := -e^{-v-1}$ for any $v \in \mathbb{R}$, by Tseng and Bertsekas (1991), we can show that the dual solution is given by

$$\hat{\pi}_K(T_i, \boldsymbol{X}_i) := \rho'\left(u_{K_1}(T_i)^\top \hat{\Lambda}_{K_1\times K_2} v_{K_2}(\boldsymbol{X}_i)\right), \tag{4.3}$$

where $\hat{\Lambda}_{K_1\times K_2}$ is the maximizer of the strictly concave function $\hat{G}_{K_1\times K_2}$ defined by

$$\widehat{\Lambda}_{K_1\times K_2} = \arg\max_{\Lambda} \hat{G}_{K_1\times K_2}(\Lambda) := \frac{1}{N}\sum_{i=1}^N \rho\left(u_{K_1}(T_i)^\top \Lambda v_{K_2}(\boldsymbol{X}_i)\right) - \left(\frac{1}{N}\sum_{i=1}^N u_{K_1}(T_i)\right)^\top \Lambda \left(\frac{1}{N}\sum_{j=1}^N v_{K_2}(\boldsymbol{X}_j)\right). \tag{4.4}$$

By the first order condition, the constraints of (4.2) are automatically satisfied by $\{\hat{\pi}_K(T_i, \boldsymbol{X}_i)\}_{i=1}^N$. The duality between (4.2) and (4.4) is shown in Appendix D. Having estimated the weights, we now estimate $\boldsymbol{\beta}^*$ by solving the generalized optimization problem, that is,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^N \hat{\pi}_K(T_i, \boldsymbol{X}_i) L\left(Y_i - g(T_i; \boldsymbol{\beta})\right). \tag{4.5}$$

Remarks:

1. Alternatively, one can estimate the stabilized weights by estimating the generalized propensity score function as well as the marginal distribution of the treatment variable nonparametrically (e.g., kernel estimation). But these alternative estimated weights do not satisfy the empirical moment condition in (4.2). Kang and Schafer (2007) argued that the inverse probability weighting method is likely to produce extreme weights and unstable estimates. If the number of moment restrictions (i.e., K) is large enough, our method is unlikely to produce extreme weights, thereby improving the finite sample performance of $\widehat{\boldsymbol{\beta}}$. See Imai and Ratkovic (2014) for simulation evidence on how the covariate balancing method dramatically improves the poor performance of the propensity score matching and weighting estimator, reported by Smith and Todd (2005) and Kang and Schafer (2007).

2. The primal problem (4.2) is different from the empirical likelihood approach (Smith, 1997, Imbens, 2002). Notice that $\rho(v) = -e^{-v-1}$ satisfies the invariance property (i.e., $-\rho''(v) = \rho'(v)$). It turns out that this invariance property is critical for establishing consistency of the generalized optimization estimator. Any other choice of $\rho(\cdot)$ that does not have the invariance property may result in biased causal effect estimation.

3. The proposed estimation (4.5) is a semiparametric estimation problem that contains both finite dimensional and infinite unknown parameters. The general semiparametric estimation problem has been studied by Ai and Chen (2003) and Chen, Linton, and Van Keilegom (2003). Ai and Chen (2003) study the large sample properties in the smooth objective function case, while Chen, Linton, and Van Keilegom (2003) extend the analysis to criterion functions that are not necessarily smooth. Equation (4.5) is a special case of the general setting of Chen, Linton, and Van Keilegom (2003), and we will indeed apply their Theorem 2 (page 1594) to derive the asymptotic properties of $\hat{\boldsymbol{\beta}}$. There is a major difference between the present paper and Chen, Linton, and Van Keilegom (2003), however. Our focus is on the efficiency bound derivation and efficient estimation, whereas their focus is on deriving the asymptotic properties of the sequential estimator under high level conditions (e.g., Condition 2.6, page 1594). These high level conditions are nontrivial to verify. Most of our derivations are indeed verifying those high level conditions; see Section 4.2 of the supplemental material Ai, Linton, Motegi, and Zhang (2020).

Related methods

In the binary treatment effect model with $T \in \{0, 1\}$, the propensity score is defined by $\pi(\boldsymbol{X}) := P(T = 1|\boldsymbol{X})$. Hirano, Imbens, and Ridder (2003) estimate the propensity score function by fitting a logit regression for T onto $u_K(\boldsymbol{X})$. As K increases to infinity, their estimator attains the efficiency bound of ATE developed by Hahn (1998).

The propensity score satisfies the following covariate balancing equation:

$$\mathbb{E}\left[T \cdot \pi(\boldsymbol{X})^{-1} v(\boldsymbol{X})\right] = \mathbb{E}[v(\boldsymbol{X})]. \tag{4.6}$$

Based on (4.6), various estimators of average treatment effects have been proposed in the existing literature. Graham, Pinto, and Egel (2012) parametrically model the propensity score $\pi(\boldsymbol{X}) = \pi(\gamma^\top v^*(\boldsymbol{X}))$ by a finite dimensional parameter γ and known $v^*(\boldsymbol{X})$. They estimate γ by solving the empirical moment of (4.6) with $v(\boldsymbol{X}) = v^*(\boldsymbol{X})$. Their estimator attains the efficiency bound if both the propensity score function is correctly specified and the conditional potential outcomes $\{\mathbb{E}[Y^*(t)|\boldsymbol{X}], t \in \{0, 1\}\}$ are linear function of $v^*(\boldsymbol{X})$. Imai and Ratkovic (2014) parametrically model the propensity score by $\pi(\boldsymbol{X}; \gamma)$ and consider the overidentified moment condition with $v(\boldsymbol{X}) = v_K(\boldsymbol{X})$ being a specified K-dimensional vector of covariates, where K is possibly larger than the dimension of γ. They propose to estimate γ through generalized method of moments (GMM) and empirical likelihood (EL). We note neither GMM nor EL leads to the empirical moment of (4.6) because both of them are defined to be the maximizer of certain criterion functions rather than directly solving the empirical moment of (4.6). In addition, the estimation of Imai and Ratkovic (2014) is not guaranteed to attain the efficiency bound of ATE developed by Hahn (1998).

Fong, Hazlett, and Imai (2018) extend the covariate balancing propensity score approach to a continuous treatment by noticing the moment condition

$$\mathbb{E}\left[\pi(T, \boldsymbol{X}) \cdot \{T - \mathbb{E}[T]\} \cdot \{\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]\}\right] = 0. \tag{4.7}$$

They consider estimating the stabilized weights by balancing covariates such that weighted correlation between T and $\boldsymbol{X}$ is minimized. However, the equation (4.7) is of finite dimension and cannot identify $\pi(T, \boldsymbol{X})$. Hence, Fong, Hazlett, and Imai (2018) impose a parametric model for the stabilized weights in order to achieve consistent estimation.

5 Large sample properties

To establish the large sample properties of the generalized optimization estimator, we first show that the estimated weight function $\hat{\pi}_K(t, \boldsymbol{x})$ is consistent and compute its convergence

rates under both the L_∞ norm and the L_2 norm. The following conditions shall be imposed.

Assumption 2. *(i) The support $\mathcal{X}$ of $\boldsymbol{X}$ is a compact subset of $\mathbb{R}^r$. The support $\mathcal{T}$ of the treatment variable T is a compact subset of $\mathbb{R}$. (ii) There exist two positive constants η_1 and η_2 such that*

$$0 < \eta_1 \leq \pi_0(t, \boldsymbol{x}) \leq \eta_2 < \infty \ , \ \forall (t, \boldsymbol{x}) \in \mathcal{T} \times \mathcal{X} \ .$$

Assumption 3. *There exist $\Lambda_{K_1 \times K_2} \in \mathbb{R}^{K_1 \times K_2}$ and a positive constant $\alpha > 0$ such that*

$$\sup_{(t,\boldsymbol{x}) \in \mathcal{T} \times \mathcal{X}} \left| (\rho'^{-1}(\pi_0(t, \boldsymbol{x})) - u_{K_1}(t)^\top \Lambda_{K_1 \times K_2} v_{K_2}(\boldsymbol{x}) \right| = O(K^{-\alpha}),$$

where $\rho(v) = -\exp(-v-1)$.

Assumption 4. *(i) For every K_1 and K_2, the smallest eigenvalues of $\mathbb{E}\left[u_{K_1}(T)u_{K_1}(T)^\top\right]$ and $\mathbb{E}\left[v_{K_2}(\boldsymbol{X})v_{K_2}(\boldsymbol{X})^\top\right]$ are bounded away from zero uniformly in K_1 and K_2. (ii) There are two sequences of constants $\zeta_1(K_1)$ and $\zeta_2(K_2)$ satisfying $\sup_{t \in \mathcal{T}} \|u_{K_1}(t)\| \leq \zeta_1(K_1)$ and $\sup_{\boldsymbol{x} \in \mathcal{X}} \|v_{K_2}(\boldsymbol{x})\| \leq \zeta_2(K_2)$, $K = K_1(N)K_2(N)$ and $\zeta(K) := \zeta_1(K_1)\zeta_2(K_2)$, such that $\zeta(K)K^{-\alpha} \to 0$ and $\zeta(K)\sqrt{K/N} \to 0$ as $N \to \infty$.*

Assumption 2 (i) restricts both the covariates and treatment level to be bounded. This condition is restrictive but convenient for computing the convergence rate under L_∞ norm. It is commonly imposed in the nonparametric regression literature. This condition can be relaxed, however, if we restrict the tail behavior of the joint distribution of $(\boldsymbol{X}, T)$. Assumption 2 (ii) restricts the weight function to be bounded and bounded away from zero. Given Assumption 2 (i), this condition is equivalent to $dF_{T|X}(T|\boldsymbol{X})$ being bounded away from zero, meaning that each type of individual (denoted by $\boldsymbol{X}$) always have a sufficient portion participating in each level of treatment. This restriction is important for our analysis since each individual participates only in one level of treatment and this condition allows us to construct her statistical counterparts from all other treatments. Although Assumption 2 (ii) is useful in causal analysis and establishing the convergence rates, it is not essential and could be relaxed by allowing η_1 (resp. η_2) to depend on N and to go to zero (resp. infinity) slowly, as $N \to \infty$. Notice that $u_{K_1}(t)^\top \Lambda v_{K_2}(\boldsymbol{x})$ is a linear sieve approximation to any suitable function of $(\boldsymbol{X}, T)$.

Assumption 3 requires the sieve approximation error of $\rho'^{-1}(\pi_0(t, \boldsymbol{x}))$ to shrink at a polynomial rate. This condition is satisfied for a variety of sieve basis functions. For example, if both $\boldsymbol{X}$ and T are discrete, then the approximation error is zero for sufficiently large K and in this case Assumption 3 is satisfied with $\alpha = +\infty$. If some components of $(\boldsymbol{X}, T)$ are continuous, the polynomial rate depends positively on the smoothness of

$\rho'^{-1}(\pi_0(t,\boldsymbol{x}))$ in continuous components and negatively on the number of the continuous components; indeed, for power series and B-splines, $\alpha = -s/r$, where s is the smoothness of approximand and r is the dimension of $\boldsymbol{X}$. Hence, the proposed method still suffers from the curse of dimensionality that typically occurs in nonparametric estimation. We will show that the convergence rate of the estimated weight function (and consequently the rate of the generalized optimization estimator) is bounded by this polynomial rate.

Assumption 4 (i) essentially ensures the sieve approximation estimator is non-degenerate. Similar conditions are common in the sieve regression literature (Andrews, 1991, Newey, 1997). If the approximation error is nonzero, Assumption 4 (ii) requires it to shrink to zero at an appropriate rate as the sample size increases. Newey (1997) show that if $u_{K_1}(t)$ (resp. $u_{K_2}(\boldsymbol{x})$) is a power series then $\zeta_1(K_1) = O(K_1)$ (resp. $\zeta_2(K_2) = O(K_2)$), and if $u_{K_1}(t)$ (resp. $u_{K_2}(\boldsymbol{x})$) is a B-spline then $\zeta_1(K_1) = O(\sqrt{K_1})$ (resp. $\zeta_2(K_2) = O(\sqrt{K_2})$).

Under these conditions, we are able to establish the following theorem:

Theorem 3. *Suppose that Assumptions 2-4 hold. Then, we obtain the following:*

$$\int_{\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t,\boldsymbol{x}) - \pi_0(t,\boldsymbol{x})|^2 dF_{T,X}(t,\boldsymbol{x}) = O_p\left(\max\left\{K^{-2\alpha}, \frac{K}{N}\right\}\right),$$

$$\frac{1}{N}\sum_{i=1}^{N} |\hat{\pi}_K(T_i,\boldsymbol{X}_i) - \pi_0(T_i,\boldsymbol{X}_i)|^2 = O_p\left(\max\left\{K^{-2\alpha}, \frac{K}{N}\right\}\right).$$

The proof of Theorem 3 immediately follows from the supplemental material Ai, Linton, Motegi, and Zhang (2020, Lemma 3.1 & Corollary 3.3).

The following additional condition is needed to establish the consistency of the proposed estimator $\hat{\boldsymbol{\beta}}$.

Assumption 5. *(i) The parameter space $\Theta \subset \mathbb{R}^p$ is a compact set and the true parameter $\boldsymbol{\beta}^*$ is in the interior of Θ , where $p \in \mathbb{N}$. (ii) $L(Y - g(T;\boldsymbol{\beta}))$ is continuous in $\boldsymbol{\beta}$, $\sup_{\boldsymbol{\beta}\in\Theta} \mathbb{E}\left[|L(Y - g(T;\boldsymbol{\beta}))|^2\right] < \infty$ and $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta} |L(Y - g(T;\boldsymbol{\beta}))|\right] < \infty$.*

Assumption 5 (i) is commonly imposed in the nonlinear regression literature, but can be relaxed if $g(t;\boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$. Assumption 5 (ii) is an envelope condition that is sufficient for the applicability of the uniform law of large numbers. A similar condition is also imposed in Newey and McFadden (1994, Lemma 2.4).

Under these and other conditions, we establish the consistency of the generalized optimization estimator. The proof of Theorem 4 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2020, Section 4.1)

Theorem 4. *Suppose that Assumptions 1-5 hold. Then, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \xrightarrow{p} 0$.*

To establish the asymptotic distribution of the proposed estimator, we need some smoothness condition on the regression function and some under-smoothing condition on the sieve approximation (i.e., larger K than needed for consistency). We also have to address the possibility of a nonsmooth loss function. These conditions are presented below.

Assumption 6.

(i) *The loss function $L(v)$ is differentiable almost everywhere, $g(t;\boldsymbol{\beta})$ is twice continuously differentiable in $\boldsymbol{\beta} \in \Theta$ and we denote its first derivative by $m(t;\boldsymbol{\beta}) := \nabla_\beta g(t;\boldsymbol{\beta})$;*

(ii) *$\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]$ is differentiable with respect to $\boldsymbol{\beta}$ and $H_0 := -\nabla_\beta \mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\,\boldsymbol{\beta})\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$ is nonsingular;*

(iii) *$\varepsilon(t,\boldsymbol{x};\boldsymbol{\beta}^*) := \mathbb{E}[L'(Y-g(T;\boldsymbol{\beta}^*))|T=t,\boldsymbol{X}=\boldsymbol{x}]$ is continuously differentiable in $(t,\boldsymbol{x})$;*

(iv) *Suppose that $N^{-1}\sum_{i=1}^N \hat{\pi}_K(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\hat{\boldsymbol{\beta}}))m(T_i;\hat{\boldsymbol{\beta}}) = o_p(N^{-1/2})$ holds with probability approaching one.*

Assumption 7. *(i) $\mathbb{E}\left[\sup_{\boldsymbol{\beta}\in\Theta}|L'(Y-g(T;\boldsymbol{\beta}))|^{2+\delta}\right] < \infty$ for some $\delta > 0$; (ii) The function class $\{L'(y-g(t;\,\boldsymbol{\beta})) : \boldsymbol{\beta}\in\Theta\}$ satisfies:*

$$\mathbb{E}\left[\sup_{\boldsymbol{\beta}_1:\|\boldsymbol{\beta}_1-\boldsymbol{\beta}\|<\delta}|L'(Y-g(T;\,\boldsymbol{\beta}_1))-L'(Y-g(T;\boldsymbol{\beta}))|^2\right]^{1/2} \le a\cdot\delta^b$$

for any $\boldsymbol{\beta}\in\Theta$ and any small $\delta>0$ and for some finite positive constants a and b.

Assumption 6 (i) imposes sufficient regularity conditions on both the regression function and the loss function. These conditions permit nonsmooth loss functions and are satisfied by the examples mentioned in previous sections. Assumption 6 (ii) ensures that the efficient variance to be finite. Assumption 6 (iv) is essentially saying that the almost sure first order condition is approximately satisfied, see Pakes and Pollard (1989). Assumption 7 is a stochastic equicontinuity condition, which is needed for establishing weak convergence, see Andrews (1994). Again, it is satisfied by widely used loss functions such as $L(v)=v^2$, $L(v)=v\{\tau-I(v\le 0)\}$, and $L(v)=v^2\cdot|\tau-I(v\le 0)|$ discussed in Section 2.

Under the above sufficient conditions, we have the following theorem.

Theorem 5. *Suppose that Assumptions 1-7 hold, and strengthen Assumption 4 (ii) to*

$$\textit{Assumption 4 (ii)}' \ \ \zeta(K)\sqrt{K^2/N} \to 0 \textit{ and } \sqrt{N}K^{-\alpha} \to 0.$$

Then, $\sqrt{N}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*) \xrightarrow{d} N(0, V_{eff})$*, where* $V_{eff} = \mathbb{E}\left[S_{eff}(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*)S_{eff}(T, \boldsymbol{X}, Y; \boldsymbol{\beta}^*)^\top\right]$*. Therefore,* $\widehat{\boldsymbol{\beta}}$ *attains the semi-parametric efficiency bound of Theorem 1.*

Assumption 4 (ii)′ imposes further restrictions on the smoothing parameter (K) so that the sieve approximation is under-smoothed. This condition is stronger than Assumption 4 (ii) but it is commonly imposed in the semiparametric regression literature. The proof of Theorem 5 is given in the supplemental material Ai, Linton, Motegi, and Zhang (2020, Section 4).

6 Confidence interval and variance estimation

The asymptotic normality of $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \widehat{\beta}_1, ..., \widehat{\beta}_{p-1})^\top$ established in Theorem 5 has a direct implication for constructing the confidence interval of $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*, ..., \beta_{p-1}^*)^\top$. The 95% symmetric confidence interval for β_j^* is given by

$$\left[\hat{\beta}_j - 1.96 \cdot \widehat{SE}_j, \quad \hat{\beta}_j + 1.96 \cdot \widehat{SE}_j\right], \tag{6.1}$$

where $\widehat{SE}_j = \widehat{V}_{jj}^{1/2}/\sqrt{N}$ is the standard error of $\hat{\beta}_j$, and $\widehat{V}_{jj}$ is a consistent estimator for V_{jj}. Here, V_{ij} denotes the (i,j)-element of V_{eff}, the asymptotic covariance matrix of the estimator (recall (3.1)). Broadly, there are two approaches for computing the standard error $\widehat{SE}_j$: plug-in and simulation-based approaches. The plug-in approach is described in Section 6.1, and the simulation-based approach is described in Section 6.2.

6.1 Plug-in approach

The plug-in approach is a conceptually straightforward approach which estimates V_{eff} by replacing unknown quantities in (3.1) with consistent estimators. This approach requires the consistent estimation of H_0 and $\psi(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*)$ (recall Theorem 1). Since the nonsmooth loss function may invalidate the exchangeability between the expectation and derivative operators, some care in the estimation of H_0 is warranted. Using the tower property of conditional expectation, H_0 can be rewritten as follows.

$$H_0 = -\nabla_\beta \mathbb{E}\left[\pi_0(T, \boldsymbol{X})\mathbb{E}\left[L'(Y - g(T; \boldsymbol{\beta}))|T, \boldsymbol{X}\right] m(T; \boldsymbol{\beta})\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$

$$= - \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) \nabla_\beta \mathbb{E}\left[L'(Y - g(T; \boldsymbol{\beta}))|T, \boldsymbol{X}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} m(T; \boldsymbol{\beta}^*)^\top\right]$$
$$- \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) \mathbb{E}\left[L'(Y - g(T; \boldsymbol{\beta}^*))|T, \boldsymbol{X}\right] \nabla_\beta m(T; \boldsymbol{\beta}^*)\right].$$

Applying integration by parts (see Appendix E), we obtain

$$\nabla_\beta \mathbb{E}\left[L'(Y - g(T; \boldsymbol{\beta}))|T = t, \boldsymbol{X} = \boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}$$
$$= \mathbb{E}\left[L'(Y - g(T; \boldsymbol{\beta}^*)) \frac{\partial}{\partial y} \log f_{Y,T,\boldsymbol{X}}(Y, T, \boldsymbol{X}) \middle| T = t, \boldsymbol{X} = \boldsymbol{x}\right] m(t; \boldsymbol{\beta}^*) \tag{6.2}$$

and consequently

$$H_0 = -\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) \left\{\frac{\partial}{\partial y} \log f_{Y,T,\boldsymbol{X}}(Y, T, \boldsymbol{X}) m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta})^\top + \nabla_\beta m(T; \boldsymbol{\beta}^*)^\top\right\}\right].$$

The log density $\log f_{Y,T,\boldsymbol{X}}(y, t, \boldsymbol{x})$ can be estimated via the widely used sieve extremum estimator (Chen, 2007, Example 2.6, page 5565):

$$\hat{f}_{Y,T,\boldsymbol{X}}(y, t, \boldsymbol{x}) := \frac{\exp\left(\hat{a}_{K_0}^\top r_{K_0}(y, t, \boldsymbol{x})\right)}{\int_{\mathcal{Y}\times\mathcal{T}\times\mathcal{X}} \exp\left(\hat{a}_{K_0}^\top r_{K_0}(y, t, \boldsymbol{x})\right) dydtd\boldsymbol{x}},$$

where $\hat{a}_{K_0} \in \mathbb{R}^{K_0}$ ($K_0 \in \mathbb{N}$) maximizes the following concave objective function

$$\hat{a}_{K_0} := \arg\max_{a \in \mathbb{R}^{K_0}} \frac{1}{N} \sum_{i=1}^N \left[a^\top r_{K_0}(Y_i, T_i, \boldsymbol{X}_i) - \log \int_{\mathcal{Y}\times\mathcal{T}\times\mathcal{X}} \exp\left(a^\top r_{K_0}(y, t, \boldsymbol{x})\right) dydtd\boldsymbol{x}\right],$$

and $r_{K_0}(t, y, \boldsymbol{x})$ is a K_0-dimensional sieve basis. Then, H_0 can be estimated by

$$\hat{H} := -\frac{1}{N} \sum_{i=1}^N \hat{\pi}_K(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i; \hat{\boldsymbol{\beta}})) \left\{\hat{a}_{K_0}^\top \frac{\partial}{\partial y} r_{K_0}(Y_i, T_i, \boldsymbol{X}_i) m(T_i; \hat{\boldsymbol{\beta}}) m(T_i; \hat{\boldsymbol{\beta}})^\top + \nabla_\beta m(T_i; \hat{\boldsymbol{\beta}})\right\}.$$

Also, $\psi(Y, T, \boldsymbol{X}; \boldsymbol{\beta}^*)$ can be directly estimated by the plug-in sieve estimator:

$$\hat{\psi}(Y, T, \boldsymbol{X}; \hat{\boldsymbol{\beta}}) = \hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}})) m(T; \hat{\boldsymbol{\beta}}) - \hat{\pi}_K(t, \boldsymbol{x}) \widehat{\mathbb{E}}\left[L'(Y - g(T; \hat{\boldsymbol{\beta}}))|T, \boldsymbol{X}\right] m(T; \hat{\boldsymbol{\beta}})$$
$$+ \widehat{\mathbb{E}}\left[\hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}}))|T\right] m(T; \hat{\boldsymbol{\beta}}) + \widehat{\mathbb{E}}\left[\hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}}))|\boldsymbol{X}\right] m(T; \hat{\boldsymbol{\beta}}),$$

where $\widehat{\mathbb{E}}[\hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}}))|T, \boldsymbol{X}]$ is the least square regression of $\hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}}))$ on a sieve basis $w_{K_0}(T, \boldsymbol{X})$; $\widehat{\mathbb{E}}[L'(Y - g(T; \hat{\boldsymbol{\beta}}))|T]$ and $\widehat{\mathbb{E}}[\hat{\pi}_K(T, \boldsymbol{X}) L'(Y - g(T; \hat{\boldsymbol{\beta}}))|\boldsymbol{X}]$ are defined similarly.

Finally, a consistent estimator of V_{eff} is given by

$$\widehat{V} := \hat{H}^{-1} \left\{\frac{1}{N} \sum_{i=1}^N \hat{\psi}(Y_i, T_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}}) \hat{\psi}(Y_i, T_i, \boldsymbol{X}_i; \hat{\boldsymbol{\beta}})^\top\right\} (\hat{H}^\top)^{-1}. \tag{6.3}$$

The sieve extreme estimator is uniformly strong consistent in the almost sure sense (see Chen, 2007, Theorem 3.1). Also from Theorems 3 and 4, we have $\sup_{(t,\boldsymbol{x})\in\mathcal{T}\times\mathcal{X}} |\hat{\pi}_K(t, \boldsymbol{x}) - \pi_0(t, \boldsymbol{x})| = o_p(1)$ and $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\| \to 0$. With these results, the consistency of $\hat{V}$ follows from standard arguments.

6.2 Simulation-based approach

The plug-in approach described in Section 6.1 is conceptually straightforward, but may be hard to implement from a practical point of view. In this section, we describe the Jackknife and bootstrap methods as alternative approaches. First, the Jackknife method proceeds as follows (Wasserman, 2013). The i^{th} Jackknife sample is constructed by deleting the i^{th} observation from the dataset:

$$\boldsymbol{J}^{[-i]} := \{T_j, \boldsymbol{X}_j, Y_j : j \in \{1, 2, ..., i-1, i+1, ..., N\}\}.$$

The i^{th} Jackknife replicate, denoted as $\widehat{\boldsymbol{\beta}}^{[-i]} = (\widehat{\beta}_0^{[-i]}, \widehat{\beta}_1^{[-i]}, ..., \widehat{\beta}_{p-1}^{[-i]})^\top$, is defined as the point estimator for $\boldsymbol{\beta}^*$ computed on the i^{th} Jackknife sample $\boldsymbol{J}_{[-i]}$. The Jackknife-based standard error of estimated β_j^* is given by

$$\widehat{SE}_j^{jack} = \left\{ \frac{N-1}{N} \sum_{i=1}^{N} \left(\widehat{\beta}_j^{[-i]} - \widehat{\beta}_j^{[\cdot]} \right)^2 \right\}^{\frac{1}{2}}, \tag{6.4}$$

where $\widehat{\beta}_j^{[\cdot]} = N^{-1} \sum_{i=1}^{N} \widehat{\beta}_j^{[-i]}$. Substitute (6.4) into (6.1) to compute the confidence interval.

Second, the bootstrap method proceeds as follows. The b^{th} bootstrap sample $\{T_i^{\{b\}}, \boldsymbol{X}_i^{\{b\}}, Y_i^{\{b\}}\}_{i=1}^N$ is resampled with replacement from the original sample $\{T_i, \boldsymbol{X}_i, Y_i\}_{i=1}^N$ with the uniform probability. The b^{th} bootstrap replicate, denoted by $\widehat{\boldsymbol{\beta}}^{\{b\}} = (\widehat{\beta}_0^{\{b\}}, \widehat{\beta}_1^{\{b\}}, ..., \widehat{\beta}_{p-1}^{\{b\}})^\top$, is defined as the point estimator for $\boldsymbol{\beta}^*$ computed on the b^{th} bootstrap sample. Repeat B times to get $\{\widehat{\boldsymbol{\beta}}^{\{b\}}\}_{b=1}^B$. The bootstrapped standard error of estimated β_j^* is given by

$$\widehat{SE}_j^{boot} = \left\{ \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{\beta}_j^{\{b\}} - \widehat{\beta}_j^{\{\cdot\}} \right)^2 \right\}^{\frac{1}{2}}, \tag{6.5}$$

where $\widehat{\beta}_j^{\{\cdot\}} = B^{-1} \sum_{b=1}^{B} \widehat{\beta}_j^{\{b\}}$. Substitute (6.5) into (6.1) to compute the confidence interval. (An alternative bootstrap approach can be found in Chen, Linton, and Van Keilegom, 2003, Section 3.3).

Bootstrapping provides another way to construct a confidence interval. Sort the B bootstrap replicates from the smallest to the largest, and relabel them as $\widehat{\beta}_j^{(1)} \leq \cdots \leq \widehat{\beta}_j^{(B)}$. The 95% bootstrapped equitailed confidence interval for β_j^* is given by

$$\left[\widehat{\beta}_j^{(0.025B)}, \quad \widehat{\beta}_j^{(0.975B)} \right]. \tag{6.6}$$

The entire confidence interval (6.1) can be replaced with (6.6). The former bootstrap approach (6.5) relies on the asymptotic normality result, while the latter approach (6.6) does not. We distinguish them hereafter, calling the former *bootstrap method I* and the latter *bootstrap method II*.

7 Selection of tuning parameters

The large sample properties of the proposed estimator permit a wide range of values of K_1 and K_2. This presents a dilemma for applied researchers who have only one finite sample and would like to have some guidance on the selection of smoothing parameters. Several data-driven methods for selecting tuning parameters in series estimation have been discussed in Li (1987) and Li and Racine (2007, Section 15.2). Based on that background, we present two data-driven approaches to select K_1 and K_2. The first one is simply to minimize a (penalized) loss function. Define $\bar{L}(K_1, K_2) := N^{-1} \sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i) L(Y_i - g(T_i; \hat{\boldsymbol{\beta}}))$. There are several ways to penalize using large K_1 or K_2:

No penalty. $\mathcal{L}(K_1, K_2) = \bar{L}(K_1, K_2)$.

Additive penalty. $\mathcal{L}(K_1, K_2) = (1 + 2(K_1 + K_2)/N) \times \bar{L}(K_1, K_2)$.

Multiplicative penalty. $\mathcal{L}(K_1, K_2) = (1 + 2K_1K_2/N) \times \bar{L}(K_1, K_2)$.

Choose (K_1^*, K_2^*) that minimizes $\mathcal{L}(K_1, K_2)$ in some choice sets $(K_1, K_2) \in \mathbb{K}_1 \times \mathbb{K}_2$. The second approach is the J-fold cross-validation (CV), which proceeds as follows.

1. Divide N samples into J groups, (say $J = 5$ or 10), and let $n = N/J$. The data in the j^{th} group is denoted by $S_j = \{\boldsymbol{X}_i^{(j)}, T_i^{(j)}, Y_i^{(j)} : i = 1, ..., n\}$ for $j \in \{1, .., J\}$.

2. For each $j \in \{1, ..., J\}$, compute the following quantities based on the dataset $S_{(-j)} = \{\boldsymbol{X}_i, T_i, Y_i\}_{i=1}^N / S_j$:

$$\begin{aligned}
\hat{\Lambda}_{K_1 \times K_2}^{(-j)} &= \arg\max_{\Lambda} \hat{G}_K^{(-j)}(\Lambda) \\
&= \frac{1}{N-n} \sum_{i \in S_{(-j)}} \rho\left(u_{K_1}^\top(T_i) \Lambda v_{K_2}(\boldsymbol{X}_i)\right) - \left[\frac{1}{N-n} \sum_{i \in S_{(-j)}} u_{K_1}^\top(T_i)\right] \Lambda \left[\frac{1}{N-n} \sum_{i \in S_{(-j)}} v_{K_2}(\boldsymbol{X}_i)\right], \\
\hat{\pi}_K^{(-j)}(T, \boldsymbol{X}) &= \rho'\left(u_{K_1}^\top(T) \hat{\Lambda}_{K_1 \times K_2}^{(-j)} v_{K_2}(\boldsymbol{X})\right), \\
\hat{\boldsymbol{\beta}}_K^{(-j)} &= \arg\min \sum_{i \in S_{(-j)}} \hat{\pi}_K^{(-j)}(T_i, \boldsymbol{X}_i) \left\{Y_i - g(T_i; \boldsymbol{\beta})\right\}^2.
\end{aligned}$$

3. Choose optimal K_1 and K_2 so that the following cross-validation criterion is minimized:

$$CV(K_1, K_2) = \sum_{j=1}^{J} \left[\sum_{k \in S_j} \hat{\pi}_K^{(-j)}(T_k, \boldsymbol{X}_k) \left\{Y_k - g\left(T_k; \hat{\boldsymbol{\beta}}_K^{(-j)}\right)\right\}^2\right].$$

When $J = N$, the second approach coincides with the leave-out cross-validation (Stone, 1974). Li (1987) shows that the above procedures to select K_1 and K_2 are asymptotically optimal in the sense of minimizing a weighted loss function for regression.

It should be noted that the K_1 and K_2 chosen by the above criteria are not guaranteed to satisfy the undersmoothing conditions Assumption 4 (ii'), which has been pointed out by Li and Racine (2007, Section 15.2). Linton (1995) and Donald and Newey (2001) develop second order theory to determine the optimal tuning parameters with respect to higher order MSE for a class of semiparametric estimation problems. In general, the optimal rates for K_1 and K_2 according to this criterion are larger reflecting the need for undersmoothing. This suggests that in practice one should take the K_1 and K_2 determined by CV or $\mathcal{L}$ as a lower bound.

8 Some extensions

The condition (2.1) that the causal effect is parameterized may be restrictive for some applications. To relax this condition, we can consider the nonparametric specification

$$\min_{g(\cdot)} \int_{\mathcal{T}} \mathbb{E}\left[L\left(Y^*(t) - g(t)\right)\right] dF_T(t).$$

Under Assumption 1, the above optimization is equivalent to

$$\min_{g(\cdot)} \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L(Y - g(T))\right].$$

We can estimate $g(\cdot)$ through the weighted nonparametric sieve regression:

$$\min_{g(\cdot)\in\mathcal{H}_{K_1}} \sum_{i=1}^{N} \hat{\pi}_K(T_i, \boldsymbol{X}_i) L(Y_i - g(T_i)),$$

where $\mathcal{H}_{K_1} := \left\{g(\cdot) : \mathcal{T} \to \mathbb{R}, g(t) = \lambda^\top u_{K_1}(t) : \lambda \in \mathbb{R}^{K_1}\right\}$ is a specified sieve space. The extension to the general loss function requires considerable derivation and shall be dealt with in a separate paper. In this section, we only consider three specific cases: first, the dose-response curve $\theta_t := \mathbb{E}[Y^*(t)]$, which corresponds to $L(v) = v^2$; second, the average treatment effects (ATE), which is defined by $\theta_{t_1,t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)]$ for $t_1 \neq t_0$; third, the average treatment effects on the treated (ATT), which is defined by $\theta_{t_1,t_0|t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)|T = t_0]$ for $t_1 \neq t_0$.

8.1 Estimation of effect curve and average treatment effects

We begin with estimation of θ_t. Note that, for all $t \in \mathcal{T}$ and under Assumption 1, we can rewrite θ_t as

$$\theta_t := \mathbb{E}[Y^*(t)] = \mathbb{E}\left[\pi_0(T, \boldsymbol{X})Y | T = t\right].$$

With $\pi_0(T, \boldsymbol{X})$ replaced by $\widehat{\pi}_K(T, \boldsymbol{X})$, we estimate θ_t by regressing $\widehat{\pi}_K(T, \boldsymbol{X})Y$ on $u_{K_1}(t)$, thus

$$\widehat{\theta}_t := \left[\sum_{i=1}^{N} \widehat{\pi}_K(T_i, \boldsymbol{X}_i) Y_i u_{K_1}(T_i)^\top\right] \left[\sum_{i=1}^{N} u_{K_1}(T_i) u_{K_1}(T_i)^\top\right]^{-1} u_{K_1}(t).$$

To aid presentation of the asymptotic properties of $\widehat{\theta}_t$, define the following quantities:

$$\begin{aligned}
&\Phi_{K_1 \times K_1} := \mathbb{E}[u_{K_1}(T) u_{K_1}^\top(T)], \\
&b_{K_1}(T_i, \boldsymbol{X}_i, Y_i) := \pi_0(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) | T_i, \boldsymbol{X}_i\right] \\
&\qquad\qquad + \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) | \boldsymbol{X}_i\right] - \mathbb{E}\left[\pi_0(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i)\right], \\
&V_t := \mathbb{E}\left[\left\{u_{K_1}^\top(t) \Phi_{K_1 \times K_1}^{-1} b_{K_1}(T_i, \boldsymbol{X}_i, Y_i)\right\}^2\right] \\
&\quad = u_{K_1}^\top(t) \cdot \Phi_{K_1 \times K_1}^{-1} \cdot \mathbb{E}\left[b_{K_1}(T_i, \boldsymbol{X}_i, Y_i) b_{K_1}^\top(T_i, \boldsymbol{X}_i, Y_i)\right] \cdot \Phi_{K_1 \times K_1}^{-1} \cdot u_{K_1}(t).
\end{aligned}$$

Theorem 6. *Suppose* $\sup_{t \in \mathcal{T}} |\theta_t - (\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha} > 0$ *and* $\gamma^* \in \mathbb{R}^{K_1}$, $\lambda_{\min}\left\{\mathbb{E}\left[b_{K_1}(T, \boldsymbol{X}, Y) b_{K_1}^\top(T, \boldsymbol{X}, Y)\right]\right\} \geq \underline{c} > 0$, *and Assumptions 1-4 hold. Then:*

1. *(Consistency)*

$$\int_{\mathcal{T}} |\widehat{\theta}_t - \theta_t|^2 dF_T(t) = O_p\left(\zeta(K)^2 \left\{\frac{K}{N} + K^{-2\alpha}\right\} + K_1^{-2\tilde{\alpha}}\right).$$

$$\sup_{t \in \mathcal{T}} |\hat{\theta}_t - \theta_t| = O_p\left(\zeta_1(K_1)\left\{\zeta(K)\left(\sqrt{\frac{K}{N}} + K^{-\alpha}\right) + K_1^{-\tilde{\alpha}}\right\}\right).$$

2. *(Asymptotic Normality) suppose Assumption 4' and* $\sqrt{N} K_1^{-\tilde{\alpha}} \to 0$ *hold. Then for any fixed* $t \in \mathcal{T}$,

$$\sqrt{N} V_t^{-1/2}\left[\hat{\theta}_t - \theta_t\right] \xrightarrow{d} N(0, 1).$$

See Ai, Linton, Motegi, and Zhang (2020, Section 5.1) for a proof of Theorem 6.

The proposed estimation procedure can also be used to estimate the average treatment effects (ATE) which is defined by

$$\theta_{t_1, t_0} := \mathbb{E}[Y^*(t_1) - Y^*(t_0)] = \theta_{t_1} - \theta_{t_0} \text{ for } t_1 \neq t_0.$$

The estimator of θ_{t_1,t_0} is defined by $\widehat{\theta}_{t_1,t_0} := \widehat{\theta}_{t_1} - \widehat{\theta}_{t_0}$. Let

$$V_{t_1,t_0} := \mathbb{E}\left[\left\{u_{K_1}^\top(t_1)\Phi_{K_1\times K_1}^{-1} b_{K_1}(T_i, \boldsymbol{X}_i, Y_i) - u_{K_1}^\top(t_0)\Phi_{K_1\times K_1}^{-1} b_{K_1}(T_i, \boldsymbol{X}_i, Y_i)\right\}^2\right]$$
$$= \{u_{K_1}(t_1) - u_{K_1}(t_0)\}^\top \Phi_{K_1\times K_1}^{-1} \mathbb{E}\left[b_{K_1}(T_i, \boldsymbol{X}_i, Y_i) b_{K_1}^\top(T_i, \boldsymbol{X}_i, Y_i)\right] \Phi_{K_1\times K_1}^{-1} \{u_{K_1}(t_1) - u_{K_1}(t_0)\}.$$

Similar to prove Theorem 6, we have the following corollary:

Corollary 7. *Suppose* $\sup_{t\in\mathcal{T}} |\theta_t - (\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha} > 0$ *and* $\gamma^* \in \mathbb{R}^{K_1}$, $\lambda_{\min}\left\{\mathbb{E}\left[b_{K_1}(T, \boldsymbol{X}, Y) b_{K_1}^\top(T, \boldsymbol{X}, Y)\right]\right\} \geq \underline{c} > 0$, *Assumptions 1-4' hold, and* $\sqrt{N} K_1^{-\tilde{\alpha}} \to 0$. *Then*

$$\sqrt{N} V_{t_1,t_0}^{-1/2}\left[\widehat{\theta}_{t_1,t_0} - \theta_{t_1,t_0}\right] \xrightarrow{d} N(0,1).$$

Feasible versions of the above CLT's are implemented using plug-in sieve estimation of the unknown quantities. For example, V_t can be estimated by

$$\widehat{V}_t = \frac{1}{N}\sum_{i=1}^N \left\{u_{K_1}^\top(t)\widehat{\Phi}_{K_1\times K_1}^{-1} \widehat{b}_{K_1}(T_i, \boldsymbol{X}_i, Y_i)\right\}^2,$$

where $\widehat{\Phi}_{K_1\times K_1} := N^{-1}\sum_{i=1}^N u_{K_1}(T_i) u_{K_1}^\top(T_i)$,

$$\begin{aligned}\widehat{b}_{K_1}(T_i, \boldsymbol{X}_i, Y_i) :=& \widehat{\pi}_K(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) - \widehat{\mathbb{E}}\left[\widehat{\pi}_K(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) | T_i, \boldsymbol{X}_i\right] \\ &+ \widehat{\mathbb{E}}[\widehat{\pi}_K(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i) | \boldsymbol{X}_i] - \widehat{\mathbb{E}}[\widehat{\pi}_K(T_i, \boldsymbol{X}_i) Y_i \cdot u_{K_1}(T_i)]\end{aligned}$$

is the plug-in estimates of $b_{K_1}(T_i, \boldsymbol{X}_i, Y_i)$, and $\widehat{\mathbb{E}}[\widehat{\pi}_K(T, \boldsymbol{X}) Y u_{K_1}(T) | T, \boldsymbol{X}]$ is the least square regression of $\widehat{\pi}_K(T, \boldsymbol{X}) Y u_{K_1}(T)$ on a sieve basis $w_{K_0}(T, \boldsymbol{X})$, and $\widehat{\mathbb{E}}[\widehat{\pi}_K(T, \boldsymbol{X}) Y u_{K_1}(T) | \boldsymbol{X}]$ is the least square regression of $\widehat{\pi}_K(T, \boldsymbol{X}) Y u_{K_1}(T)$ on a sieve basis $v_{K_0}(\boldsymbol{X})$.

8.2 Average treatment effects on the treated

Another important parameter for program evaluation is the average treatment effects on the treated (ATT), which is defined by

$$\theta_{t_1,t_0|t_0} := \mathbb{E}\left[Y^*(t_1) - Y^*(t_0) | T = t_0\right] \equiv \theta_{t_1|t_0} - \theta_{t_0|t_0} \text{ for } t_1 \neq t_0.$$

Note that $\theta_{t_0|t_0} = \mathbb{E}[Y^*(t_0)|T = t_0] = \mathbb{E}[Y|T = t_0]$, so it can be estimated by regressing Y on $u_{K_1}(t_0)$:

$$\widehat{\theta}_{t_0|t_0} := \left[\sum_{i=1}^N Y_i \cdot u_{K_1}^\top(T_i)\right]\left[\sum_{i=1}^N u_{K_1}(T_i) u_{K_1}^\top(T_i)\right]^{-1} u_{K_1}(t_0).$$

The difficulty is to estimate $\theta_{t_1|t_0} = \mathbb{E}[Y^*(t_1)|T = t_0]$ owing to that $Y^*(t_1)$ cannot be observed under the treatment level $T = t_0$. Under Assumption 1, $\theta_{t_1|t_0}$ can be identified as follows:

$$\begin{aligned}
\theta_{t_1|t_0} =& \mathbb{E}\left[Y^*(t_1)|T = t_0\right] = \mathbb{E}\left[\mathbb{E}[Y^*(t_1)|\boldsymbol{X}, T = t_0]|T = t_0\right] \\
=& \mathbb{E}\left[\mathbb{E}[Y^*(t_1)|\boldsymbol{X}, T = t_1]|T = t_0\right] \quad \text{(by Assumption 1)} \\
=& \int \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = t_1] \cdot \frac{f_{X|T}(\boldsymbol{x}|t_0)}{f_{X|T}(\boldsymbol{x}|t_1)} \cdot f_{X|T}(\boldsymbol{x}|t_1) d\boldsymbol{x} \\
=& \int \mathbb{E}[Y|\boldsymbol{X} = \boldsymbol{x}, T = t_1] \cdot \frac{f_T(t_1)/f_{T|X}(t_1|\boldsymbol{x})}{f_T(t_0)/f_{T|X}(t_0|\boldsymbol{x})} \cdot f_{X|T}(\boldsymbol{x}|t_1) d\boldsymbol{x} \\
=& \mathbb{E}\left[\frac{\pi_0(T, \boldsymbol{X})}{\pi_0(t_0, \boldsymbol{X})} \cdot Y \middle| T = t_1\right] \\
=& \mathbb{E}\left[\frac{\pi_0(T, \boldsymbol{X})}{\pi_0(T - \delta, \boldsymbol{X})} \cdot Y \middle| T = t_1\right], \qquad (8.1)
\end{aligned}$$

where $\delta := t_1 - t_0$. Based on (8.1), we replace $\pi_0(\cdot)$ by the estimator $\widehat{\pi}_K(\cdot)$ then apply sieve regression on $u_{K_1}(t_1)$, so that $\theta_{t_1|t_0}$ can be estimated by

$$\widehat{\theta}_{t_1|t_0} := \left[\sum_{i=1}^{N} \frac{\widehat{\pi}_K(T_i, \boldsymbol{X}_i)}{\widehat{\pi}_K(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}^\top(T_i)\right] \left[\sum_{i=1}^{N} u_{K_1}(T_i) u_{K_1}^\top(T_i)\right]^{-1} u_{K_1}(t_1).$$

Therefore, $\theta_{t_1,t_0|t_0}$ can be estimated by

$$\widehat{\theta}_{t_1,t_0|t_0} := \widehat{\theta}_{t_1|t_0} - \widehat{\theta}_{t_0|t_0}.$$

To aid presentation of the asymptotic properties of $\widehat{\theta}_{t_1|t_0}$, define the following quantities:

$$\begin{aligned}
b_{1,K_1}(T_i, \boldsymbol{X}_i, Y_i) :=& \frac{f_T(T_i + \delta)}{f_T(T_i)} \cdot \mathbb{E}[Y|T = T_i + \delta, \boldsymbol{X} = \boldsymbol{X}_i] \cdot u_{K_1}(T_i + \delta) \\
&- \mathbb{E}\left[\frac{f_T(T_i + \delta)}{f_T(T_i)} \cdot \mathbb{E}[Y|T = T_i + \delta, \boldsymbol{X} = \boldsymbol{X}_i] \cdot u_{K_1}(T_i + \delta) \middle| \boldsymbol{X}_i\right] \\
&- \mathbb{E}\left[\frac{f_T(T_i + \delta)}{f_T(T_i)} \cdot \mathbb{E}[Y|T = T_i + \delta, \boldsymbol{X} = \boldsymbol{X}_i] \cdot u_{K_1}(T_i + \delta) \middle| T_i\right] \\
&+ \mathbb{E}\left[\frac{f_T(T_i + \delta)}{f_T(T_i)} \cdot \mathbb{E}[Y|T = T_i + \delta, \boldsymbol{X} = \boldsymbol{X}_i] \cdot u_{K_1}(T_i + \delta)\right], \\
b_{2,K_1}(T_i, \boldsymbol{X}_i, Y_i) :=& \frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i) - \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i) \middle| T_i, \boldsymbol{X}_i\right] \\
&+ \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i) \middle| \boldsymbol{X}_i\right] - \mathbb{E}\left[\frac{\pi_0(T_i, \boldsymbol{X}_i)}{\pi_0(T_i - \delta, \boldsymbol{X}_i)} \cdot Y_i \cdot u_{K_1}(T_i)\right],
\end{aligned}$$

$$b_{3,K_1}(T_i, Y_i) := u_{K_1}(T_i) \cdot \{Y_i - \mathbb{E}[Y_i|T_i]\}.$$

Note that the expectations of b_{1,K_1}, b_{2,K_1} and b_{3,K_1} are zeros. Let

$$V_{t_1,t_0|t_0} := \mathbb{E}\left[\left\{u_{K_1}^\top(t_1)\Phi_{K_1\times K_1}(b_{1,K_1} + b_{2,K_1}) - u_{K_1}^\top(t_0)\Phi_{K_1\times K_1}b_{3,K_1}\right\}^2\right] = \boldsymbol{w}^\top \Sigma_{2K_1\times 2K_1}\boldsymbol{w},$$

where $\boldsymbol{w} := \left(u_{K_1}^\top(t_1)\cdot\Phi_{K_1\times K_1}, u_{K_1}^\top(t_0)\cdot\Phi_{K_1\times K_1}\right)^\top \in \mathbb{R}^{2K_1}$ and

$$\Sigma_{2K_1\times 2K_1} := \mathbb{E}\begin{bmatrix} \{b_{1,K_1} + b_{2,K_1}\}\{b_{1,K_1} + b_{2,K_1}\}^\top, & -\{b_{1,K_1} + b_{2,K_1}\}b_{3,K_1}^\top \\ -b_{3,K_1}\{b_{1,K_1} + b_{2,K_1}\}^\top, & b_{3,K_1}b_{3,K_1}^\top \end{bmatrix}.$$

Theorem 8. *Suppose* $\sup_{t\in\mathcal{T}} |\mathbb{E}[\pi_0(T, \boldsymbol{X})Y/\pi_0(T-\delta, \boldsymbol{X})|T = t] - (\gamma^*)^\top u_{K_1}(t)| = O(K_1^{-\tilde{\alpha}})$ *holds for some* $\tilde{\alpha} > 0$ *and* $\gamma^* \in \mathbb{R}^{K_1}$, $\lambda_{\min}(\Sigma_{2K_1\times 2K_1}) \geq \underline{c} > 0$, *Assumptions 1-4' hold, and* $\sqrt{N}K_1^{-\tilde{\alpha}} \to 0$. *Then*

$$\sqrt{N}V_{t_1,t_0|t_0}^{-1/2}\left[\widehat{\theta}_{t_1,t_0|t_0} - \theta_{t_1,t_0|t_0}\right] \xrightarrow{d} N(0,1).$$

See Ai, Linton, Motegi, and Zhang (2020, Section 5.2) for a proof of Theorem 8. Feasible versions of the above CLT's are implemented using plug-in sieve estimation of the unknown quantities.

9 Monte Carlo simulations

The large sample properties established in previous sections do not indicate how the generalized optimization estimator behaves in finite samples. To evaluate its finite sample performance, we conduct a simulation study on a continuous treatment. A simulation design is described in Section 9.1, and results are discussed in Section 9.2. To save space, the simulation study in the present section is kept compact; see the supplemental material Ai, Linton, Motegi, and Zhang (2020, Section 6) for a complete simulation study.

9.1 Simulation design

Let $X_i \overset{i.i.d.}{\sim} N(0,1)$ be a covariate. Error terms are drawn mutually independently as $\xi_i \overset{i.i.d.}{\sim} N(0,1)$ and $\epsilon_i \overset{i.i.d.}{\sim} N(0,1)$. We consider two data generating processes (DGPs):

DGP-L $T = 1 + 0.2X + \xi$ and $Y = 1 + X + T + \epsilon$. (X affects T and Y linearly.)

DGP-NL $T = 0.1X^2 + \xi$ and $Y = X^2 + T + \epsilon$. (X affects T and Y non-linearly.)

For each DGP, the true link function is $\mathbb{E}[Y(t)] = 1 + t$, a simple linear function with $\beta_1^* = \beta_2^* = 1$. We use a linear link function $g(T;\beta) = \beta_1 + \beta_2 T$, compute the generalized optimization estimator $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)^\top$ with the exponential tilting function $\rho(v) = -e^{-v-1}$, and examine its performance.

To compute the generalized optimization estimator, two sieve basis functions $u_{K_1}(T)$ and $v_{K_2}(X)$ need to be specified. For $u_{K_1}(T)$, we consider

$$u_2(T) = (1, T)^\top,\ u_3(T) = (1, T, T^2)^\top,\ u_4(T) = (1, T, T^2, T^3)^\top.$$

For $v_{K_2}(X)$, we consider

$$v_2(X) = (1, X)^\top,\ v_3(X) = (1, X, X^2)^\top,\ v_4(X) = (1, X, X^2, X^3)^\top.$$

Since $K_1, K_2 \in \{2, 3, 4\}$, there are 9 pairs of (K_1, K_2) in total. The 10-fold cross validation is employed to select an optimal pair (K_1^*, K_2^*) among the 9 pairs (recall Section 7). For comparison, simulation results with fixed $(K_1, K_2) = (2, 3)$ are also reported.

We also compute Fong, Hazlett, and Imai's (2018) covariate balancing generalized propensity score estimator with a linear model specification and the quadratic loss function. The linear specification is correct under DGP-L, while it is incorrect under DGP-NL. Comparing our estimator and the parametric estimator of Fong, Hazlett, and Imai (2018) allows us to highlight the robustness of the former to non-linear DGPs. Fong, Hazlett, and Imai (2018) also propose a nonparametric estimator in their Section 3.3. In their simulation study, the parametric and nonparametric estimators exhibit similar performance for each DGP considered (Fong, Hazlett, and Imai, 2018, Figure 2). Hence, the present paper focuses on the parametric version of their estimator to save space.

Our proposed estimator and the parametric version of Fong, Hazlett, and Imai's (2018) estimator are computed in a simulated sample with size $N \in \{100, 500\}$, after which another sample is generated and both estimators are computed again. This exercise is repeated $M = 1000$ times.

To evaluate the performance of point estimation, the bias, standard deviation, and root mean squared error (RMSE) of $\hat{\beta}_1$ and $\hat{\beta}_2$ are calculated from (a subset of) $M = 1000$ simulations. In a small portion of the $M = 1000$ samples, $\bar{\pi}_N \equiv (1/N)\sum_{i=1}^N \hat{\pi}_K(T_i, X_i)$, which should be equal to 1 in theory, takes a value far from 1 due to numerical instability in the computation of $\Lambda^*_{K_1 \times K_2}$. The numerical maximization with respect to Λ should lead to a global maximizer $\Lambda^*_{K_1 \times K_2}$ in theory, but optimizing the $K_1 \times K_2$ elements of Λ all at once is sometimes hard in practice. Hence, we calculate the bias, standard deviation, and RMSE from Monte Carlo samples such that $\bar{\pi}_N \in [0.5, 2]$. There can be a few samples in which

$\bar{\pi}_N \notin [0.5, 2]$, and these samples are simply discarded. (We admit that this computational problem becomes worse as the dimension of covariates $\boldsymbol{X}$ becomes larger.)

To evaluate the finite sample performance of the interval estimation associated with the proposed method, we implement the bootstrap method II with $B = 500$ iterations based on (6.6). In this method, we construct bootstrapped confidence intervals without using the asymptotic normality. For each of β_1 and β_2, we compute the 95% coverage probability and the average width of the 95% confidence intervals across $M = 1000$ Monte Carlo samples. For simplicity, the dimensions of the sieve basis functions are fixed at $(K_1, K_2) = (2, 3)$ when the performance of the interval estimation is evaluated.

9.2 Simulation results

Simulation results on point and interval estimation are reported in Tables 1 and 2, respectively. We discuss point estimation first, and then discuss interval estimation. Under DGP-L, the generalized optimization estimator (labeled as GOE) has reasonably small RMSE whether (K_1, K_2) are fixed at $(2, 3)$ or selected via the 10-fold cross validation. For the intercept parameter β_1, the RMSE of the parametric version of the covariate balancing generalized propensity score estimator (labeled as CBGPS) is even smaller than the RMSE of GOE. For the slope parameter β_2, the RMSE of CBGPS is as small as the RMSE of GOE. The sharp performance of CBGPS is not surprising, since it has a correct parametric specification under DGP-L1.

Under DGP-NL, GOE dominates CBGPS in terms of the estimation of β_2. When $N = 100$, the RMSEs with respect to β_2 are $\{0.104, 0.201, 0.267\}$ for GOE with fixed (K_1, K_2), GOE with the cross validation, and CBGPS, respectively. Similarly, when $N = 500$, the RMSEs are $\{0.048, 0.133, 0.211\}$. The bias and RMSE of GOE shrink to 0 as the sample size grows, indicating that GOE operates well under the non-linear DGP. CBGPS, by contrast, fails with considerable bias under DGP-NL. The bias of CBGPS is 0.189 for $N = 100$ and 0.194 for $N = 500$. These results suggest that GOE performs well for both linear and non-linear scenarios, while CBGPS performs well for linear scenarios only.

We now discuss the results on interval estimation associated with GOE. For each DGP, parameter, and sample size, the 95% coverage probability is nearly identical to 0.95. Further, the average width of the bootstrapped confidence intervals shrinks as the sample size grows, as expected. See β_2 under DGP-NL, for example. The coverage probabilities are 0.956 and 0.950 when $N \in \{100, 500\}$, respectively. Similarly, the average widths are 0.422 and 0.180. These results indicate that the bootstrap method II given in (6.6) operates sufficiently well under both linear and non-linear scenarios.

Table 1: Simulation results on point estimation

DGP-L: $T = 1 + X + \xi$ and $Y = 1 + X + T + \epsilon$

		Intercept β_1 (truth: $\beta_1^* = 1$)		Slope β_2 (truth: $\beta_2^* = 1$)	
		$N = 100$	$N = 500$	$N = 100$	$N = 500$
	(K_1, K_2)	Bias, Stdev, RMSE	Bias, Stdev, RMSE	Bias, Stdev, RMSE	Bias, Stdev, RMSE
GOE	$(2, 3)$	0.005, 0.187, 0.187	0.001, 0.083, 0.083	0.001, 0.107, 0.107	0.002, 0.050, 0.050
GOE	CV_{10}	-0.005, 0.190, 0.190	0.006, 0.080, 0.080	0.006, 0.112, 0.112	0.000, 0.047, 0.047
CBGPS	-	-0.005, 0.149, 0.149	0.001, 0.067, 0.067	0.003, 0.106, 0.106	-0.001, 0.049, 0.049

DGP-NL: $T = 0.1X^2 + \xi$ and $Y = X^2 + T + \epsilon$

		Intercept β_1 (truth: $\beta_1^* = 1$)		Slope β_2 (truth: $\beta_2^* = 1$)	
		$N = 100$	$N = 500$	$N = 100$	$N = 500$
	(K_1, K_2)	Bias, Stdev, RMSE	Bias, Stdev, RMSE	Bias, Stdev, RMSE	Bias, Stdev, RMSE
GOE	$(2, 3)$	0.002, 0.176, 0.176	0.004, 0.079, 0.079	0.004, 0.104, 0.104	-0.001, 0.048, 0.048
GOE	CV_{10}	-0.037, 0.176, 0.180	-0.012, 0.077, 0.078	0.102, 0.173, 0.201	0.080, 0.107, 0.133
CBGPS	-	-0.035, 0.179, 0.182	-0.021, 0.075, 0.078	0.189, 0.188, 0.267	0.194, 0.083, 0.211

"GOE" is the proposed generalized optimization estimator. K_1 and K_2 are the dimensions of the polynomials of T and X, respectively. CV_{10} signifies the 10-fold cross validation, where the choice set is $K_1, K_2 \in \{2, 3, 4\}$. "CBGPS" is Fong, Hazlett, and Imai's (2018) parametric covariate balancing generalized propensity score estimator. The sample size is $N \in \{100, 500\}$, and the number of Monte Carlo iterations is $M = 1000$.

10 Empirical study

We revisit the U.S. presidential campaign data analyzed by Urban and Niebler (2014) and Fong, Hazlett, and Imai (2018). The motivation of the original study, Urban and Niebler (2014), is well summarized in Fong, Hazlett, and Imai (2018, Section 2):

> Urban and Niebler (2014) explored the potential causal link between advertising and campaign contributions. Presidential campaigns ordinarily focus their advertising efforts on competitive states, but if political advertising drives more donations, then it may be worthwhile for candidates to also advertise in non-competitive states. The authors exploit the fact that media markets sometimes

Table 2: Simulation results on interval estimation (generalized optimization estimator)

	DGP-L				DGP-NL			
	Intercept β_1		Slope β_2		Intercept β_1		Slope β_2	
	CP95	AveW	CP95	AveW	CP95	AveW	CP95	AveW
$N = 100$	0.957	0.709	0.940	0.428	0.944	0.677	0.956	0.422
$N = 500$	0.966	0.311	0.947	0.184	0.942	0.305	0.950	0.180

DGP-L: $T = 1 + X + \xi$ and $Y = 1 + X + T + \epsilon$. DGP-NL: $T = 0.1X^2 + \xi$ and $Y = X^2 + T + \epsilon$. 95% confidence intervals on the target parameters (β_1, β_2) are constructed via the bootstrap with $B = 500$ iterations. The sieve basis functions are specified as $u_2(T) = (1, T)^\top$ (i.e., $K_1 = 2$) and $v_3(X) = (1, X, X^2)^\top$ (i.e., $K_2 = 3$). "CP95" signifies the 95% coverage probability, while "AveW" signifies the average width of the confidence intervals across $M = 1000$ Monte Carlo samples. The sample size is $N \in \{100, 500\}$.

> cross state boundaries. This means that candidates may inadvertently advertise in noncompetitive states when they purchase advertisements for media markets that mainly serve competitive states. By restricting their analysis to noncompetitive states, the authors attempt to isolate the effect of advertising from that of other campaigning, which do not incur these media market spillovers.

The treatment of interest, the number of political advertisements aired in each zip code, can be regarded as a continuous variable since it takes a range of values from 0 to 22379 across $N = 16265$ zip codes. Restricting themselves to a binary treatment framework, Urban and Niebler (2014) compared 5230 zip codes that received more than 1000 advertisements and 11035 zip codes that received less than 1000 advertisements. Their empirical results suggest that advertising in non-competitive states had a significant causal effect on the level of campaign contributions.

Fong, Hazlett, and Imai (2018) used the continuous treatment model, taking advantage of their proposed CBGPS method. Their empirical results suggest, contrary to Urban and Niebler (2014), that advertising in non-competitive states did *not* have a significant causal effect on the level of campaign contributions (cf. Fong, Hazlett, and Imai, 2018, Table 2).

Using the generalized optimization estimator, we analyze the impact of advertisements on contributions based on both binary and continuous treatment models. Let Y_i and T_i be the log of the campaign contribution and political advertisement in zip code $i \in \{1, \ldots, N\}$,

respectively. Stack eight covariates as

$$\boldsymbol{X} = \begin{bmatrix} \log(\text{Population}) \\ \%\text{Over 65} \\ \log(\text{Income} + 1) \\ \%\text{Hispanic} \\ \%\text{Black} \\ \log(\text{Population Density} + 1) \\ \%\text{College Graduates} \\ \text{Can Commute} \end{bmatrix}. \tag{10.1}$$

Subscript i is omitted for brevity, but (10.1) is defined for each zip code. The definition of each covariate is almost self-explanatory (see Fong, Hazlett, and Imai, 2018, Sec. 5 for more details). The log-transformation is implemented for Y, T, and some of the covariates in order to stabilize computation. Urban and Niebler (2014) made the data publicly available at the American Journal of Political Science (AJPS) Dataverse archive. See Section 10.1 for the binary treatment model and Section 10.2 for the continuous treatment model.

10.1 Binary treatment model

We dichotomize the treatment variable (i.e., the log-advertisement) as $D = \mathbf{1}(T > 4)$. This is equivalent to dichotomizing the advertisement at 100, and 7137 zip codes out of $N = 16265$ are above the cut-off level. The potential outcome model is written as

$$\mathbb{E}[Y^*(d)] = \beta_1 + \beta_2 \times d.$$

Then, the stabilized weight reduces to

$$\pi_0(D, \boldsymbol{X}) = D \times \frac{P(D = 1)}{P(D = 1 \mid \boldsymbol{X})} + (1 - D) \times \frac{P(D = 0)}{P(D = 0 \mid \boldsymbol{X})}.$$

The parameters of interest, $\boldsymbol{\beta} = (\beta_1, \beta_2)^\top$, are identified as

$$\begin{aligned} \beta_1 &= \mathbb{E}[Y^*(0)] = \frac{\mathbb{E}\left[(1 - D)\pi_0(D, \boldsymbol{X})Y\right]}{\mathbb{E}[(1 - D)\pi_0(D, \boldsymbol{X})]}, \\ \beta_2 &= \mathbb{E}[Y^*(1) - Y^*(0)] = \frac{\mathbb{E}\left[D\pi_0(D, \boldsymbol{X})Y\right]}{\mathbb{E}[D\pi_0(D, \boldsymbol{X})]} - \frac{\mathbb{E}\left[(1 - D)\pi_0(D, \boldsymbol{X})Y\right]}{\mathbb{E}[(1 - D)\pi_0(D, \boldsymbol{X})]}. \end{aligned}$$

The covariate balancing equation of propensity score becomes

$$\frac{\mathbb{E}\left[D\pi(D, \boldsymbol{X})v(\boldsymbol{X})\right]}{\mathbb{E}[D]} = \mathbb{E}[v(\boldsymbol{X})] = \frac{\mathbb{E}\left[(1 - D)\pi(D, \boldsymbol{X})v(\boldsymbol{X})\right]}{\mathbb{E}[1 - D]}.$$

Our proposed estimator of stabilized weights becomes

$$\hat{\pi}_K(D_i, \boldsymbol{X_i}) = D_i \rho' \left(\hat{\lambda}_{1K}^\top v_K(\boldsymbol{X}_i) \right) + (1 - D_i) \rho' \left(\hat{\lambda}_{2K}^\top v_K(\boldsymbol{X}_i) \right),$$

where

$$\hat{\lambda}_{1K} = \arg\max_{\lambda_1} \left\{ \frac{\sum_{i=1}^N D_i \rho \left(\lambda_1^\top v_K(\boldsymbol{X}_i) \right)}{\sum_{i=1}^N D_i} - \frac{1}{N} \sum_{i=1}^N \lambda_1^\top v_K(\boldsymbol{X}_i) \right\},$$
$$\hat{\lambda}_{2K} = \arg\max_{\lambda_2} \left\{ \frac{\sum_{i=1}^N (1 - D_i) \rho \left(\lambda_2^\top v_K(\boldsymbol{X}_i) \right)}{\sum_{i=1}^N (1 - D_i)} - \frac{1}{N} \sum_{i=1}^N \lambda_2^\top v_K(\boldsymbol{X}_i) \right\}.$$

Finally, the generalized optimization estimator for $\boldsymbol{\beta}$ is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^N (1 - D_i) \hat{\pi}_K(D_i, \boldsymbol{X}_i) Y_i}{\sum_{i=1}^N (1 - D_i) \hat{\pi}_K(D_i, \boldsymbol{X}_i)}, \qquad \hat{\beta}_2 = \frac{\sum_{i=1}^N D_i \hat{\pi}_K(D_i, \boldsymbol{X}_i) Y_i}{\sum_{i=1}^N D_i \hat{\pi}_K(D_i, \boldsymbol{X}_i)} - \hat{\beta}_1.$$

The sieve basis function is specified as $v_K(\boldsymbol{X}) = (1, \boldsymbol{X}^\top)^\top$ with $K = 9$, where the covariates are given in (10.1). The exponential tilting function $\rho(w) = -e^{-w-1}$ is used. As in the simulation study in Section 9, 95% confidence intervals for β_1 and β_2 are computed via the bootstrap method II with $B = 1000$ iterations; recall (6.6).

Our empirical results are as follows. First, $\hat{\beta}_1 = 1.227$ and the bootstrapped confidence interval is $[1.198, 1.257]$. Second, $\hat{\beta}_2 = 0.061$ and the bootstrapped confidence interval is $[0.003, 0.076]$. The latter result indicates that advertising in non-competitive states has a significantly positive causal effect on the level of campaign contributions at the 5% level, which is a consistent result with Urban and Niebler (2014).

10.2 Continuous treatment model

The procedure for the continuous treatment model is described in detail in Section 9, hence we refrain from repeating it here. The link function is specified as $g(T, \boldsymbol{\beta}) = \beta_1 + \beta_2 T + \beta_3 T^2$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^\top$. The sieve basis functions are specified as $u_{K_1}(T) = (1, T, T^2)^\top$ with $K_1 = 3$ and $v_{K_2}(\boldsymbol{X}) = (1, \boldsymbol{X}^\top)^\top$ with $K_2 = 9$, where the covariates are given in (10.1). The exponential tilting function $\rho(w) = -e^{-w-1}$ is used. 95% confidence intervals for $\boldsymbol{\beta}$ are computed via the bootstrap method II with $B = 1000$ iterations.

Our empirical results are as follows. First, $\hat{\beta}_1 = 1.100$ and the bootstrapped confidence interval is $[0.909, 1.320]$. Second, $\hat{\beta}_2 = 0.140$ and the confidence interval is $[-0.025, 0.232]$. Third, $\hat{\beta}_3 = -0.015$ and the confidence interval is $[-0.025, 0.001]$. The latter two results suggest that advertising in non-competitive states does not have a significant causal effect

on the level of campaign contributions, which is a consistent result with Fong, Hazlett, and Imai (2018).

The binary and continuous approaches lead to the opposite conclusions; the former finds the marginally significant impact of advertisements on campaign contributions at the 5% level, while the latter finds the marginally insignificant impact. These results suggest that the causal effect should be small if it exists at all. The binary model involves only one sieve basis function $v_9(\boldsymbol{X})$, while the continuous model involves two sieve basis functions $u_3(T)$ and $v_9(\boldsymbol{X})$. The latter requires the joint estimation of a relatively large-dimensional parameter matrix $\Lambda_{3\times 9}$; recall (4.4). This numerical complexity might be a reason why a significant causal effect is not detected under the continuous model.

11 Concluding remarks

The weighted optimization framework provides a unified approach towards estimation of treatment effects, under the condition of unconfounded treatment assignment. We established the semiparametric efficiency of our methodology, but perhaps the main advantage is its relatively simple form and good finite sample properties.

There are several extensions worth pursuing in future projects. First, estimation of the nonparametric causal effect function under general loss function has not been completely dealt with in this paper. But this is an important extension since it removes the burden of parameterizing the causal effect. Second, the extension of the current setting to allow for high dimensional covariates is also an important project. Third, panel data are common in the empirical literature. Our approach is readily applicable to those data, although the efficiency issue is more difficult. All these extensions shall be taken up in future studies.

Acknowledgement

We are grateful for the co-editor, Andres Santos, and three anonymous referees for their insightful comments and suggestions. We also thank Shigeyuki Hamori and Toru Kitagawa, seminar participants at Kobe University, and conference participants at the 15th International Conference of WEAI for their helpful comments. The first author, Chunrong Ai, acknowledges financial support from National Natural Science Foundation of China through project 71873138. The second author, Oliver Linton, acknowledges Cambridge INET for financial support. The third author, Kaiji Motegi, acknowledges the financial support of JSPS KAKENHI Grant Number 19K13670. The last author, Zheng Zhang, acknowledges the financial support from the National Natural Science Foundation of China

through project 12001535, and the fund for building world-class universities (disciplines) of the Renmin University of China.

References

ABADIE, A. (2005): "Semiparametric Difference-in-Differences Estimators," *The Review of Economic Studies*, 72(1), 1–19.

ABADIE, A., AND G. W. IMBENS (2006): "Large sample properties of matching estimators for average treatment effects," *Econometrica*, 74(1), 235–267.

——— (2016): "Matching on the estimated propensity score," *Econometrica*, 84(2), 781–807.

AI, C., AND X. CHEN (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, 71(6), 1795–1843.

AI, C., O. LINTON, K. MOTEGI, AND Z. ZHANG (2020): "Supplemental material for 'A Unified Framework for Efficient Estimation of General Treatment Models'," Discussion paper, Chinese University of Hong Kong, Shenzhen.

ANDREWS, D. W. K. (1991): "Asymptotic normality of series estimators for nonparametric and semiparametric regression models," *Econometrica*, 59(2), 307–345.

——— (1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4, chap. 37, pp. 2247–2294. Citeseer.

ANGRIST, J. D., AND J.-S. PISCHKE (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

ATHEY, S., G. IMBENS, T. PHAM, AND S. WAGER (2017): "Estimating average treatment effects: Supplementary analyses and remaining challenges," *American Economic Review*, 107(5), 278–81.

ATHEY, S., G. W. IMBENS, AND S. WAGER (2018): "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4), 597–623.

BICKEL, P. J., C. A. J. KLAASSEN, Y. RITOV, AND J. A. WELLNER (1993): *Efficient and Adaptive Estimation for Semiparametric Models*. The Johns Hopkins University Press.

BONEVA, L., D. ELLIOTT, I. KAMINSKA, O. LINTON, N. MCLAREN, AND B. MORLEY (2018): "The Impact of QE on liquidity: Evidence from the UK Corporate Bond Purchase Scheme," *Bank of England Working Paper*, Staff Working Paper No.782.

BUCHINSKY, M. (1995): "Quantile regression, Box-Cox transformation model, and the U.S. wage structure, 1963–1987," *Journal of Econometrics*, 65(1), 109–154.

CATTANEO, M. D. (2010): "Efficient semiparametric estimation of multi-valued treatment effects under ignorability," *Journal of Econometrics*, 155(2), 138–154.

CATTANEO, M. D., AND M. H. FARRELL (2011): "Efficient estimation of the dose-response function under ignorability using subclassification on the covariates," in *Missing Data Methods: Cross-sectional Methods and Applications*, vol. 27A, pp. 93–127. Emerald Group Publishing Limited.

CHAN, K. C. G., S. C. P. YAM, AND Z. ZHANG (2016): "Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(3), 673–700.

CHEN, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of Econometrics*, 6(B), 5549–5632.

CHEN, X., O. LINTON, AND I. VAN KEILEGOM (2003): "Estimation of Semiparametric Models When the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

CHERNOZHUKOV, V., D. CHETVERIKOV, M. DEMIRER, E. DUFLO, C. HANSEN, W. NEWEY, AND J. ROBINS (2018): "Double/debiased machine learning for treatment and structural parameters," *The Econometrics Journal*, 21(1), C1–C68.

CHERNOZHUKOV, V., I. FERNÁNDEZ-VAL, AND B. MELLY (2013): "Inference on counterfactual distributions," *Econometrica*, 81(6), 2205–2268.

CHERNOZHUKOV, V., AND C. HANSEN (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

DONALD, S. G., AND Y.-C. HSU (2014): "Estimation and inference for distribution functions and quantile functions in treatment effect models," *Journal of Econometrics*, 178(3), 383–397.

DONALD, S. G., AND W. K. NEWEY (2001): "Choosing the number of instruments," *Econometrica*, 69(5), 1161–1191.

FARRELL, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189(1), 1–23.

FIRPO, S. (2007): "Efficient Semiparametric Estimation of Quantile Treatment Effects," *Econometrica*, 75(1), 259–276.

FLORENS, J. P., J. J. HECKMAN, C. MEGHIR, AND E. VYTLACIL (2008): "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76(5), 1191–1206.

FONG, C., C. HAZLETT, AND K. IMAI (2018): "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements," *Annals of Applied Statistics*, 12(1), 156–177.

GALVAO, A. F., AND L. WANG (2015): "Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment," *Journal of the American Statistical Association*, 110(512), 1528–1542.

GRAHAM, B. S., C. C. D. X. PINTO, AND D. EGEL (2012): "Inverse probability tilting for moment condition models with missing data," *The Review of Economic Studies*, 79(3), 1053–1079.

HAHN, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," *Econometrica*, 66(2), 315–331.

HECKMAN, J. J., H. ICHIMURA, AND P. TODD (1998): "Matching as an econometric evaluation estimator," *The Review of Economic Studies*, 65(2), 261–294.

HECKMAN, J. J., AND E. VYTLACIL (2005): "Structural equations, treatment effects, and econometric policy evaluation," *Econometrica*, 73(3), 669–738.

HIRANO, K., AND G. W. IMBENS (2004): "The propensity score with continuous treatments," in *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, ed. by A. Gelman, and X.-L. Meng, chap. 7, pp. 73–84. John Wiley & Sons Ltd.

HIRANO, K., G. W. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

IMAI, K., AND M. RATKOVIC (2014): "Covariate balancing propensity score," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1), 243–263.

IMAI, K., AND D. A. VAN DYK (2004): "Causal inference with general treatment regimes: Generalizing the propensity score," *Journal of the American Statistical Association*, 99(467), 854–866.

IMBENS, G., R. SPADY, AND P. JOHNSON (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66(2), 333–357.

IMBENS, G. W. (2000): "The role of the propensity score in estimating dose-response functions," *Biometrika*, 87(3), 706–710.

——— (2002): "Generalized method of moments and empirical likelihood," *Journal of Business & Economic Statistics*, 20(4), 493–506.

——— (2004): "Nonparametric estimation of average treatment effects under exogeneity: A review," *The Review of Economics and Statistics*, 86(1), 4–29.

KANG, J. D. Y., AND J. L. SCHAFER (2007): "Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data," *Statistical Science*, 22(4), 523–539.

KENNEDY, E. H., Z. MA, M. D. MCHUGH, AND D. S. SMALL (2017): "Non-parametric methods for doubly robust estimation of continuous treatment effects," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(4), 1229–1245.

KITAMURA, Y., AND M. STUTZER (1997): "An information-theoretic alternative to generalized method of moments estimation," *Econometrica*, 65(4), 861–874.

LI, K.-C. (1987): "Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set," *The Annals of Statistics*, 15(3), 958–975.

LI, Q., AND J. S. RACINE (2007): *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

LINTON, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63(5), 1079–1112.

NEWEY, W., AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

NEWEY, W. K. (1997): "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79(1), 147–168.

NEWEY, W. K., AND J. L. POWELL (1987): "Asymmetric least squares estimation and testing," *Econometrica*, 55(4), 819–847.

PAKES, A., AND D. POLLARD (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57(5), 1027–1057.

ROBINS, J. M., M. A. HERNÁN, AND B. BRUMBACK (2000): "Marginal structural models and causal inference in epidemiology," *Epidemiology*, 11(5), 550–560.

ROBINS, J. M., A. ROTNITZKY, AND L. P. ZHAO (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.

ROSENBAUM, P. R., AND D. B. RUBIN (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

SŁOCZYŃSKI, T., AND J. M. WOOLDRIDGE (2018): "A general double robustness result for estimating average treatment effects," *Econometric Theory*, 34(1), 112–133.

SMITH, J. A., AND P. E. TODD (2005): "Does matching overcome LaLonde's critique of nonexperimental estimators?," *Journal of Econometrics*, 125(1-2), 305–353.

SMITH, R. J. (1997): "Alternative semi-parametric likelihood approaches to generalised method of moments estimation," *The Economic Journal*, 107(441), 503–519.

STONE, M. (1974): "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2), 111–133.

TSENG, P., AND D. P. BERTSEKAS (1991): "Relaxation methods for problems with strictly convex costs and linear constraints," *Mathematics of Operations Research*, 16(3), 462–481.

URBAN, C., AND S. NIEBLER (2014): "Dollars on the Sidewalk: Should U.S. Presidential Candidates Advertise in Uncontested States?," *American Journal of Political Science*, 58(2), 322–336.

WAGER, S., AND S. ATHEY (2018): "Estimation and inference of heterogeneous treatment effects using random forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

WASSERMAN, L. (2013): *All of Statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media.

YIU, S., AND L. SU (2018): "Covariate association eliminating weights: A unified weighting framework for causal effect estimation," *Biometrika*, 105(3), 709–722.

Appendix

A Proof of (2.2)

Using the law of iterated expectation and Assumption 1, we can deduce that

$$\begin{aligned}
&\mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L\left(Y - g(T; \boldsymbol{\beta})\right)\right] \\
&= \int \pi_0(t, \boldsymbol{x}) \cdot \mathbb{E}[L(Y^*(T) - g(T; \boldsymbol{\beta}))|T = t, \boldsymbol{X} = \boldsymbol{x}]\, dF_{T|X}(t|\boldsymbol{x}) dF_X(\boldsymbol{x}) \\
&= \int \mathbb{E}\left[L(Y^*(t) - g(t; \boldsymbol{\beta}))|T = t, \boldsymbol{X} = \boldsymbol{x}\right] dF_T(t) dF_X(\boldsymbol{x}) \\
&= \int \mathbb{E}\left[L(Y^*(t) - g(t; \boldsymbol{\beta}))|\boldsymbol{X} = \boldsymbol{x}\right] dF_T(t) dF_X(\boldsymbol{x}) \ \text{(using Assumption 1)}
\end{aligned}$$

$$= \int \mathbb{E}\left[L(Y^*(t) - g(t;\boldsymbol{\beta}))\right] dF_T(t).$$

B Proof of Theorem 2

The sufficient part is obvious. We prove the necessary part. Let $u(T) = \exp(a \cdot T \cdot i)$ and $v(\boldsymbol{X}) = \exp(b^\top \boldsymbol{X} \cdot i)$ be the test functions, where $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. By assumption,

$$\begin{aligned}&\mathbb{E}\left[\{\pi(T,\boldsymbol{X}) - \pi_0(T,\boldsymbol{X})\}\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] + \mathbb{E}\left[\pi_0(T,\boldsymbol{X})\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] \\ =&\mathbb{E}\left[\exp(a \cdot T \cdot i)\right] \cdot \mathbb{E}\left[\exp(b^\top \boldsymbol{X} \cdot i)\right].\end{aligned}$$

By definition $\mathbb{E}\left[\pi_0(T,\boldsymbol{X})\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] = \mathbb{E}\left[\exp(a \cdot T \cdot i)\right] \cdot \mathbb{E}\left[\exp(b^\top \boldsymbol{X} \cdot i)\right]$. Then $\mathbb{E}\left[\{\pi(T,\boldsymbol{X}) - \pi_0(T,\boldsymbol{X})\}\exp\left\{a \cdot T \cdot i + b^\top \boldsymbol{X} \cdot i\right\}\right] = 0$ for all $a \in \mathbb{R}$ and $b \in \mathbb{R}^r$. By the uniqueness of Fourier transform, we can obtain $\pi(T,\boldsymbol{X}) = \pi_0(T,\boldsymbol{X})$ a.s.

C Asymptotic result when $\pi_0(T, \mathbf{X})$ is known

Suppose the stabilized weight function $\pi_0(T, \mathbf{X})$ is known, the weighted optimization estimator of $\boldsymbol{\beta}^*$, denoted by $\widehat{\boldsymbol{\beta}}_{known}$, is

$$\widehat{\boldsymbol{\beta}}_{known} = \min_{\beta} \sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L(Y_i - g(T_i;\beta)).$$

We also assume the asymptotic first order condition

$$\frac{1}{N}\sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known}) = o_p(N^{-1/2}) \tag{C.1}$$

holds with probability approaching to one.

Proposition B.1 Suppose Assumptions 5, 6 (i-ii), and 7 hold, and (C.1) holds, then we have

1. $\widehat{\boldsymbol{\beta}}_{known} \xrightarrow{p} \boldsymbol{\beta}^*$;
2. $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*) \xrightarrow{d} N(0, V_{ineff})$, where
$$V_{ineff} := H_0^{-1} \cdot \mathbb{E}\left[\pi_0(T,\boldsymbol{X})^2 L'(Y - g(T;\boldsymbol{\beta}^*))^2 m(T;\boldsymbol{\beta}^*)m(T;\boldsymbol{\beta}^*)^\top\right] \cdot H_0^{-1};$$
3. furthermore, if $\mathbb{E}\left[L'(Y(t) - g(t;\boldsymbol{\beta}^*))\right] = 0$ holds for all $t \in \mathcal{T}$, then $V_{ineff} \geq V_{eff}$ in the sense of that $c^\top \cdot V_{ineff} \cdot c \geq c^\top \cdot V_{eff} \cdot c$ for any vector $c \in \mathbb{R}^p$.

Proof. By Assumption 5 and the uniform law of large number, we obtain

$$\begin{aligned}&\frac{1}{N}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L\left\{Y_i-g(T_i;\boldsymbol{\beta})\right\}\\&\quad\to\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L\left\{Y-g(T;\boldsymbol{\beta})\right\}\right]\text{ in probability uniformly over }\boldsymbol{\beta},\end{aligned}$$

which implies the consistency result $\|\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*\|\xrightarrow{p}0$.

The first order condition (C.1) holds with probability approaching to one. Note that $L'(\cdot)$ may not be a differentiable function, e.g. $L'(v)=\tau-I(v<0)$ in quantile regression, we cannot simply apply Mean Value Theorem on (C.1) to obtain the expression for $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)$. To solve this problem, we resort to the empirical process theory in Andrews (1994). Define

$$f(\boldsymbol{\beta}):=\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right],$$

which is a differentiable function in $\boldsymbol{\beta}$ and by (2.3) $f(\boldsymbol{\beta}^*)=0$. Using Mean Value Theorem, we can obtain

$$0=\sqrt{N}f(\boldsymbol{\beta}^*)=\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})-\nabla_\beta f(\bar{\boldsymbol{\beta}})\cdot\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*),$$

where $\bar{\boldsymbol{\beta}}$ lies on the line joining $\widehat{\boldsymbol{\beta}}_{known}$ and $\boldsymbol{\beta}^*$. Because $\nabla_\beta f(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$ at $\boldsymbol{\beta}^*$, and $\|\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*\|\xrightarrow{p}0$, then we have

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)=\left[\nabla_\beta f(\boldsymbol{\beta}^*)\right]^{-1}\cdot\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})+o_p(1).$$

Define the empirical process

$$\nu_N(\boldsymbol{\beta})=\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left\{\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\boldsymbol{\beta}))m(T_i;\boldsymbol{\beta})-\mathbb{E}\left[\pi_0(T,\boldsymbol{X})L'(Y-g(T;\boldsymbol{\beta}))m(T;\boldsymbol{\beta})\right]\right\}.$$

By (C.1) and the definition of $\nu_N(\boldsymbol{\beta})$, we have

$$\begin{aligned}&\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known}-\boldsymbol{\beta}^*)\\=&\nabla_\beta f(\boldsymbol{\beta}^*)^{-1}\cdot\left\{\sqrt{N}f(\widehat{\boldsymbol{\beta}}_{known})-\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known})\right.\\&\qquad\left.+\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\pi_0(T_i,\boldsymbol{X}_i)L'(Y_i-g(T_i;\widehat{\boldsymbol{\beta}}_{known}))m(T_i;\widehat{\boldsymbol{\beta}}_{known})\right\}\\=&-\nabla_\beta f(\boldsymbol{\beta}^*)^{-1}\cdot\nu_N(\widehat{\boldsymbol{\beta}}_{known})+o_p(1)\\=&H_0^{-1}\cdot\left\{\left(\nu_N(\widehat{\boldsymbol{\beta}}_{known})-\nu_N(\boldsymbol{\beta}^*)\right)+\nu_N(\boldsymbol{\beta}^*)\right\}+o_p(1).\end{aligned}$$

By Assumptions 6, 7, Theorems 4 and 5 of Andrews (1994), we have that $\nu_N(\cdot)$ is stochastically equicontinuous, which implies $\nu_N(\widehat{\boldsymbol{\beta}}_{known}) - \nu_N(\boldsymbol{\beta}^*) \xrightarrow{p} 0$. Therefore,

$$\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*) = H_0^{-1} \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \pi_0(T_i, \boldsymbol{X}_i) L'(Y_i - g(T_i; \boldsymbol{\beta}^*)) m(T_i; \boldsymbol{\beta}^*) + o_p(1) \ ,$$

then we can conclude that the asymptotic variance of $\sqrt{N}(\widehat{\boldsymbol{\beta}}_{known} - \boldsymbol{\beta}^*)$ is V_{ineff}.

We next show $V_{ineff} \geq V_{eff}$. From Theorem 1, we have

$$\begin{aligned}
V_{eff} = H_0^{-1} \cdot \Big\{ & \mathbb{E}\left[\pi_0(T, \boldsymbol{X})^2 L'(Y - g(T; \boldsymbol{\beta}^*))^2 m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& - 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}_0)) m(T; \boldsymbol{\beta}^*) | T]^\top\right] \\
& + 2 \cdot \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1} \\
= H_0^{-1} \Big\{ & \mathbb{E}\left[\pi_0(T, \boldsymbol{X})^2 L'(Y - g(T; \boldsymbol{\beta}^*))^2 m(T; \boldsymbol{\beta}^*) m(T; \boldsymbol{\beta}^*)^\top\right] \\
& - \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& + \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1},
\end{aligned}$$

where the last equality holds by noting

$$\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T = t] = \mathbb{E}\left[L'(Y^*(t) - g(t; \boldsymbol{\beta}^*))\right] \cdot m(t; \boldsymbol{\beta}^*) = 0 \ ,$$

since the model is correctly specified, i.e. $\mathbb{E}\left[L'(Y^*(t) - g(t; \boldsymbol{\beta}^*))\right] = 0$ for $t \in \mathcal{T}$. Therefore,

$$\begin{aligned}
& V_{ineff} - V_{eff} \\
= & H_0^{-1} \Big\{ \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | T, \mathbf{X}]^\top\right] \\
& - \mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X}) L'(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right] \Big\} H_0^{-1} \\
\geq & 0,
\end{aligned}$$

where the last inequality holds by Jensen's inequality:

$$\mathbb{E}\left[\mathbb{E}[\pi_0(T, \boldsymbol{X})(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}] \cdot \mathbb{E}[\pi_0(T, \boldsymbol{X})(Y - g(T; \boldsymbol{\beta}^*)) m(T; \boldsymbol{\beta}^*) | \mathbf{X}]^\top\right]$$

$$<\mathbb{E}\left[\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]\cdot\mathbb{E}[\pi_0(T,\boldsymbol{X})(Y-g(T;\boldsymbol{\beta}^*))m(T;\boldsymbol{\beta}^*)|T,\mathbf{X}]^\top\right].$$

□

D Duality of primal problem (4.2)

We first introduce some notation:

- Let $m_K(T,\boldsymbol{X}) = \text{vec}\left(u_{K_1}(T)v_{K_2}^\top(\boldsymbol{X})\right)$ denote a K-dimensional column vector formed by the elements of the matrix $u_{K_1}(T)v_{K_2}^\top(\boldsymbol{X})$. Let $M_{K\times N} = (m_K(T_1,\boldsymbol{X}_1),\ldots,m_K(T_N,\boldsymbol{X}_N))$, which is a $K\times N$ matrix.
- Let $u_{K_1,k}(T)$ (resp. $v_{K_2,k'}(\boldsymbol{X})$) denote the k^{th} (resp. k'^{th}) component of $u_{K_1}(T)$ (resp. $v_{K_2}(\boldsymbol{X})$), and denote

$$\overline{u}_{K_1,k} = \frac{1}{N}\sum_{i=1}^N u_{K_1,k}(T_i) \text{ and } \overline{v}_{K_2,k'} = \frac{1}{N}\sum_{i=1}^N v_{K_2,k'}(\boldsymbol{X}_i).$$

  Let b_K be a K dimensional column vector whose elements are formed by $\{\bar{u}_{K_1,k}\bar{v}_{K_2,k'}; k = 1,...,K_1, k' = 1,...,K_2\}$.
- Denote $\boldsymbol{\pi} = (\pi_1,...,\pi_N)$ and $F(\boldsymbol{\pi}) = \sum_{i=1}^N \pi_i\log\pi_i$.

The primal optimization problem (4.2) can be written as

$$\begin{cases} \min_{\boldsymbol{\pi}} F(\boldsymbol{\pi}) \\ \text{subject to} \quad M_{K\times N}\cdot\boldsymbol{\pi} = N\cdot b_K \end{cases} \tag{D.1}$$

By Tseng and Bertsekas (1991), the conjugate convex function of $F(\cdot)$ is

$$F^*(\mathbf{z}) = \sup_{\boldsymbol{\pi}}\sum_{i=1}^N\{z_i\pi_i - \pi_i\log\pi_i\} = \sum_{i=1}^N\{z_i\pi_i^* - \pi_i^*\log\pi_i^*\},$$

where π_j^* satisfies the first order condition:

$$z_j = \log\pi_j^* + 1 \Rightarrow \pi_j^* = e^{z_j-1} = \rho'(z_i).$$

By substitution, we obtain

$$F^*(\mathbf{z}) = \sum_{i=1}^N\left\{z_ie^{z_i-1} - e^{z_i-1}(z_i-1)\right\} = \sum_{i=1}^N e^{z_i-1} = \sum_{i=1}^N -\rho(-z_i).$$

By Tseng and Bertsekas (1991), the dual problem of (D.1) is

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}^K}\left\{\lambda^\top(N\cdot b_K)-F^*\left(\lambda^\top M_{K\times N}\right)\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\sum_{i=1}^N\left\{\overline{u}_{K_1}^\top\Lambda\overline{v}_{K_2}+\rho\left(-u_{K_1}(T_i)^\top\Lambda v_{K_2}(\boldsymbol{X}_i)\right)\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\sum_{i=1}^N\left\{\rho\left(u_{K_1}(T_i)^\top\Lambda v_K(\boldsymbol{X}_i)\right)-\overline{u}_{K_1}^\top\Lambda\overline{v}_{K_2}\right\}\\
=&\max_{\Lambda\in\mathbb{R}^{K_1}\times\mathbb{R}^{K_2}}\hat{G}_{K_1\times K_2}(\Lambda).
\end{aligned}\tag{D.2}$$

Therefore, the dual solution of (4.2) is given by

$$\hat{\pi}_K(T_i,\boldsymbol{X}_i)=\rho'\left(u_{K_1}(T_i)^\top\hat{\Lambda}_{K_1\times K_2}v_{K_2}(\boldsymbol{X}_i)\right)\ ,$$

where $\hat{\Lambda}_{K_1\times K_2}$ is the maximizer of the strictly concave objective function $\hat{G}_{K_1\times K_2}$.

E Proof of (6.2)

$$\begin{aligned}
&\nabla_\beta\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}))|T=t,\boldsymbol{X}=\boldsymbol{x}\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\\
=&\nabla_\beta\left[\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}))f_{Y|T,X}(y|t,\boldsymbol{x})dy\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\\
=&\nabla_\beta\left[\int_{\mathbb{R}}L'(z)f_{Y|T,X}(z+g(t;\boldsymbol{\beta})|t,\boldsymbol{x})dz\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*}\ (\text{use } z=y-g(t;\boldsymbol{\beta}))\\
=&\int_{\mathbb{R}}L'(z)\cdot\frac{\partial}{\partial y}f_{Y|T,X}(z+g(t;\boldsymbol{\beta}^*)|t,\boldsymbol{x})dz\cdot m(t;\boldsymbol{\beta}^*)\\
=&\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\partial}{\partial y}f_{Y|T,X}(y|t,\boldsymbol{x})dy\cdot m(t;\boldsymbol{\beta}^*)\\
=&\int_{\mathbb{R}}L'(y-g(t;\boldsymbol{\beta}^*))\cdot\frac{\frac{\partial}{\partial y}f_{Y,T,X}(y,t,\boldsymbol{x})}{f_{Y,T,X}(y,t,\boldsymbol{x})}f_{Y|T,X}(y|t,\boldsymbol{x})dy\cdot m(t;\boldsymbol{\beta}^*)\\
=&\mathbb{E}\left[L'(Y-g(T;\boldsymbol{\beta}^*))\frac{\frac{\partial}{\partial y}f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})}{f_{Y,T,\boldsymbol{X}}(Y,T,\boldsymbol{X})}\middle|T=t,\boldsymbol{X}=\boldsymbol{x}\right]m(t;\boldsymbol{\beta}^*).
\end{aligned}$$