ESTIMATING DENSITY RATIO OF MARGINALS TO JOINT: APPLICATIONS TO CAUSAL INFERENCE

YUKITOSHI MATSUSHITA, TAISUKE OTSU, AND KEISUKE TAKAHATA

Abstract. In various fields of data science, researchers often face problems of estimating the ratios of two probability densities. Particularly in the context of causal inference, the product of marginals for a treatment variable and covariates to their joint density ratio typically emerges in the process of constructing causal effect estimators. This paper applies the general least square density ratio estimation methodology by Kanamori, Hido and Sugiyama (2009) to the product of marginals to joint density ratio, and demonstrates its usefulness particularly for causal inference on continuous treatment effects and dose-response curves. The proposed method is illustrated by a simulation study and an empirical example to investigate the treatment effect of political advertisements in the U.S. presidential campaign data.

1. Introduction

In various fields of data science, researchers often face problems of estimating the ratios of two probability densities (or referred to as importance). Examples include important sampling, distribution comparison, outlier detection, mutual information estimation, covariate shift adaptation, to name a few. One of the major findings in the machine learning literature on this topic is that it is possible and advisable to estimate directly the density ratios rather than estimating separately the denominator and numerator densities; see, Sugiyama, Suzuki and Kanamori (2012a) for an overview. In particular, Sugiyama *et al.* (2008) and Kanamori, Hido and Sugiyama (2009) proposed direct estimation strategies for

Our research is supported by the JSPS KAKENHI (26780133, 18K01541) (Matsushita).

density ratios by minimizing the Kullback-Leiber divergence and least square criterions, respectively.

One of the most popular examples of the density ratio is the joint to (product of) marginals ratio, which emerges in mutual information estimation and conditional probability estimation (see, e.g., Chapters 11-12 of Sugiyama, Suzuki and Kanamori, 2012a). On the other hand, its reciprocal (i.e., the *marginals to joint* ratio) is less studied in the literature. In this paper, however, we argue that such reciprocals are also useful and empirically relevant objects especially in the context of causal inference. It should be noted that the reciprocals of the existing joint to marginals ratio estimators may perform poorly because it does not directly estimate the original object of interest.

In this paper, we apply the general least square density ratio estimation methodology by Kanamori, Hido and Sugiyama (2009) to the product of marginals to joint density ratio and demonstrate its usefulness in the context of causal inference. In particular, we study estimation of causal effects of continuous treatment effects and dose-response curves based on the density ratio estimation method. Our simulation and empirical examples illustrate usefulness of the proposed method. Compared to the conventional nonparametric plug-in estimation methods which involve three separate nonparametric density estimation problems (i.e., two marginals and joint densities), one major attractiveness of the direct density ratio estimation method is that it involves only one nonparametric estimation problem and tends to provide stable and practical estimates for the final object of interest; cf. Vapnik's (1998) principle: *one should avoid solving more difficult intermediate problems when solving a target problem.*

A major contribution of this paper is to propose new applications of the density estimation methodology in the context of causal inference. Most existing papers consider modeling and estimating binary treatment effects under unconfounded treatment assignments, such as inverse propensity score weighting (e.g., Rosenbaum and Rubin, 1983, and Hirano, Imbens and Ridder, 2003), matching (e.g., Heckman, Ichimura and Todd, 1998, Imbens, 2004, and Abadie and Imbens, 2006), and regression adjustment (e.g., Angrist and Pischke, 2008). Also efficient estimation of treatment effects is investigated by Robins, Rotnitzky and Zhao (1994), Hahn (1998), Hirano, Imbens and Ridder (2003), and Graham, Pinto and Egel (2012), for example. Our density ratio-based estimator presented in Section 3.1 may be applied to binary treatments (see Remark 5 below). In this case, our methodology may be interpreted as an inverse propensity score weighting estimator, such as Hirano, Imbens and Ridder (2003), where the reciprocal of the propensity score function is directly estimated by the density ratio estimator instead of taking reciprocal after estimating the propensity score function.

The usefulness of our density ratio-based approach becomes clearer when we consider continuous treatment effect analyses, where the density ratio involves three continuous functions (the joint and two marginal densities for the treatment and covariates). Section 3 demonstrates this point by applying the density ratio estimators to continuous treatment effects and dose-response curves. Several estimation methods are proposed in the literature of continuous treatment effect analysis, such as the parametric generalized propensity score approach (Imbens, 2004, and Imai and van Dyk, 2004), control function approach (Florens *et al.*, 2008), stabilized weighting (Galvao and Wang, 2015), kernel smoothing (Kennedy *et al.*, 2017), propensity score balancing (Fong, Hazlett and Imai, 2018), and a unified weighting approach (Yiu and Su, 2018). However these exiting estimation approaches for continuous treatment effects are either semiparametrically inefficient or possibly biased if parametrized components are misspecified. A general semiparametric efficient estimation framework for continuous (and also discrete) treatment effects is developed in a recent work by Ai *et al.* (2021). By investigating the semiparametric efficiency bound of the continuous treatment effect, Ai *et al.* (2021) clarify moment conditions to be satisfied by the stabilized weights (Robins, HernÃ¡n and Brumback, 2000) to achieve efficiency, estimate those weights by an entropy maximization method, and propose a semiparametrically efficient estimator for continuous treatment effects by utilizing the estimated stabilized weights. Our density ratio-based estimator for continuous treatment effects presented in Section 3.1 is asymptotically equivalent to the one by Ai *et al.* (2021). Indeed our density ratio estimator may be interpreted as alternative stabilized weights to achieve semiparametric efficiency for estimating causal effects.

The rest of the paper is organized as follows. In Section 2, we introduce the general idea of density ratio estimation (Section 2.1), develop the marginals to joint density ratio estimator (Section 2.2), and study its asymptotic properties (Section 2.3). Section 3 presents applications of the proposed estimator for causal effects of continuous treatment effects (Section 3.1) and dose-response curves (Section 3.2). In Section 4, we conduct a simulation study, and Section 5 presents a real data example to study a causal effect of political advertisements on the amount of donations.

2. Estimation of density ratio of marginals to joint

2.1. **Least square density ratio estimation.** As a basis of our estimation strategy, we first introduce the general idea of the least square density ratio estimation proposed by Kanamori, Hido and Sugiyama (2009). Let $\{Z_i^{\mathrm{nu}}\}_{i=1}^{n_{\mathrm{nu}}}$ and $\{Z_j^{\mathrm{de}}\}_{j=1}^{n_{\mathrm{de}}}$ be random samples on $\mathbb{R}^d$ generated from the distributions having the Lebesgue densities f_{nu} and f_{de}, respectively. We are interested in the density ratio

$$R(z) = \frac{f_{\mathrm{nu}}(z)}{f_{\mathrm{de}}(z)}.$$

Instead of estimating the densities f_{nu} and f_{de} separately, Kanamori, Hido and Sugiyama (2009) proposed to estimate directly the ratio $R(z)$ by approximating $R(z)$ by linear models and minimizing a least square type criterion function.

More precisely, we approximate $R(z)$ by the linear model

$$R(z;\theta) = \sum_{\ell=1}^{K} \theta_\ell \psi_\ell(z) = \psi(z)'\theta,$$

where $\psi = (\psi_1, \ldots, \psi_K)'$ is a vector of basis functions and $\theta = (\theta_1, \ldots, \theta_K)'$ is a vector of parameters. To estimate θ, we consider the minimization of the squared error weighted by f_{de}:

$$\frac{1}{2}\int \{R(z;\theta) - R(z)\}^2 f_{\mathrm{de}}(z)dz.$$

By ignoring the term that does not depend on θ and estimating the expectations by the empirical averages, the parameter vector θ can be estimated as a solution of $\min_\theta \hat{Q}(\theta)$, where

$$\hat{Q}(\theta) \;=\; \frac{1}{2n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} R(Z_j^{\mathrm{de}};\theta)^2 - \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} R(Z_i^{\mathrm{nu}};\theta) = \frac{1}{2}\theta'\hat{H}\theta - \hat{h}'\theta,$$

with $\hat{H} = \frac{1}{n_{\mathrm{de}}}\sum_{j=1}^{n_{\mathrm{de}}} \psi(Z_j^{\mathrm{de}})\psi(Z_j^{\mathrm{de}})'$ and $\hat{h} = \frac{1}{n_{\mathrm{nu}}}\sum_{i=1}^{n_{\mathrm{nu}}} \psi(Z_i^{\mathrm{nu}})$.

If $\hat{H}$ is invertible, this optimization problem can be explicitly solved as $\hat{\theta} = \hat{H}^{-1}\hat{h}$ and the density ratio $R(z)$ can be estimated by $R(z;\hat{\theta}) = \psi(z)'\hat{\theta}$. Sugiyama, Suzuki and Kanamori (2012a) surveyed theoretical properties of this estimator and discussed its applications in various contexts of statistical data analysis. In the next subsection, we apply this general approach for estimation of the (product of) marginals to joint density ratio. To the best of our knowledge, such an application of the density estimation technique seems to be new in the literature.

2.2. **Estimation of marginals to joint density ratio.** We now adapt the idea of the least square density ratio estimation to our object of interest, marginals to joint density ratio. Let $\{T_i, X_i\}_{i=1}^n$ be a random sample of $(T, X) \in \mathcal{T} \times \mathcal{X}$. In our examples on causal inference below, T is a treatment variable and X is a vector of covariates. This paper is concerned with estimation of the product of marginals to joint density ratio

$$r_0(t, x) = \frac{f_T(t) f_X(x)}{f_{TX}(t, x)},$$

where f_T and f_X are the marginal densities of T and X, respectively, and f_{TX} is the joint density of (T, X). As in the last subsection, we approximate the ratio $r_0(t, x)$ by the linear model

$$r(t, x; \alpha) = \sum_{\ell=1}^{K} \alpha_\ell \psi_\ell(t, x) = \psi(t, x)'\alpha, \tag{1}$$

where $\psi = (\psi_1, \ldots, \psi_K)'$ is a vector of basis functions, $\alpha = (\alpha_1, \ldots, \alpha_K)'$ is a vector of parameters, and K is the length of series approximation. For the asymptotic analysis below, we let $K \to \infty$ as the sample size n increases.

In this setup, we consider the least square-type population criterion function for α:

$$\begin{aligned} \frac{1}{2}E[\{\psi(T, X)'\alpha - r_0(T, X)\}^2] &= \frac{1}{2}\int r(t, x; \alpha)^2 f_{TX}(t, x) dt dx - \int r(t, x; \alpha) f_T(t) f_X(x) dt dx \\ &\quad + \frac{1}{2}\int r_0(t, x)^2 f_{TX}(t, x) dt dx. \end{aligned} \tag{2}$$

The least square parameter vector α^* is defined as the minimizer of (2), that is,

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2}E[\{\psi(T, X)'\alpha - r_0(T, X)\}^2]. \tag{3}$$

For estimation, we ignore the third term in (2), which does not depend on α. The first and second terms in (2) can be estimated by the empirical moments. Thus, the parameter vector α^* can be estimated by $\hat{\alpha} = \arg\min_\alpha \hat{Q}(\alpha)$, where

$$\hat{Q}(\alpha) = \frac{1}{2n}\sum_{i=1}^{n} r(T_i, X_i; \alpha)^2 - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} r(T_i, X_j; \alpha) = \frac{1}{2}\alpha'\hat{H}\alpha - \hat{h}'\alpha, \tag{4}$$

with $\hat{H} = \frac{1}{n}\sum_{i=1}^{n}\psi(T_i, X_i)\psi(T_i, X_i)'$ and $\hat{h} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(T_i, X_j)$. If $\hat{H}$ is invertible, our estimator for the product of marginals to joint density ratio $r_0(t, x)$ is explicitly written as

$$\hat{r}(t, x) = r(t, x; \hat{\alpha}), \quad \text{where} \quad \hat{\alpha} = \hat{H}^{-1}\hat{h}. \tag{5}$$

Alternatively, to obtain more stable estimates especially when $\hat{H}$ is nearly singular, we can introduce a quadratic (or ridge-type) regularizer and minimize

$$\hat{Q}(\alpha) + \frac{\lambda}{2}\alpha'\alpha$$

with respect to α, where λ is a tuning parameter. In this case, the estimator for $r_0(t, x)$ is obtained as

$$\tilde{r}(t, x) = r(t, x; \tilde{\alpha}), \quad \text{where} \quad \tilde{\alpha} = (\hat{H} + \lambda I)^{-1}\hat{h}. \tag{6}$$

Remark 1. (Comparison with joint to marginals density ratio) In contrast to the estimator in (5), the joint to marginals density ratio $\frac{f_{TX}(t,x)}{f_T(t)f_X(x)}$ (i.e., the reciprocal of $r_0(t, x)$) can be estimated by $r(t, x; \hat{\alpha}_{\text{joint to marginals}})$, where $\hat{\alpha}_{\text{joint to marginals}} = \hat{H}_J^{-1}\hat{h}_J$ with $\hat{H}_J = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(T_i, X_j)\psi(T_i, X_j)'$ and $\hat{h}_J = \frac{1}{n}\sum_{i=1}^{n}\psi(T_i, X_i)$. Obviously, the reciprocal of $r(t, x; \hat{\alpha}_{\text{joint to marginals}})$ is different from our direct estimator $r(t, x; \hat{\alpha})$, and such a reciprocal would be numerically unstable to estimate $r_0(t, x)$ particularly when the estimate $r(t, x; \hat{\alpha}_{\text{joint to marginals}})$ is close to zero. See Section 4 for some numerical illustration.

Remark 2. (Selections of series length and tuning parameter by cross validation) To implement our ratio estimators, we need to choose the series length K for (5) and additionally the tuning parameter λ for (6). One practical way to choose K is based on the (M-fold) cross validation to minimize the (dominant part of) L^2-risk in (4). More precisely, divide the sample $S = \{T_i, X_i\}_{i=1}^n$ into M groups (say, $M = 5$ or 10), and let $S_m = \{T_i^{(m)}, X_i^{(m)}\}_{i=1}^{n/M}$ be the m-th group for $m = 1, \ldots, M$. The cross validation criterion for K is written as

$$CV(K) = \sum_{m=1}^{M} \left(\frac{1}{2} \hat{\alpha}'_{-m} \hat{H}_m \hat{\alpha}_{-m} - \hat{h}'_m \hat{\alpha}_{-m} \right), \tag{7}$$

where $\hat{\alpha}_{-m}$ is $\hat{\alpha}$ based on the subsample $S \backslash S_m$, $\hat{H}_m = \frac{1}{(n/M)} \sum_{i=1}^{n/M} \psi(T_i^{(m)}, X_i^{(m)}) \psi(T_i^{(m)}, X_i^{(m)})'$, and $\hat{h}_m = \frac{1}{(n/M)^2} \sum_{i=1}^{n/M} \sum_{j=1}^{n/M} \psi(T_i^{(m)}, X_j^{(m)})$. Then we can choose K as the minimizer of $CV(K)$. This method is employed in our numerical illustrations.

Similarly, in order to choose K and λ to implement the regularized estimator in (6), the cross validation criterion may be constructed as

$$CV(K, \lambda) = \sum_{m=1}^{M} \left(\frac{1}{2} \tilde{\alpha}'_{-m} \hat{H}_m \tilde{\alpha}_{-m} - \hat{h}'_m \tilde{\alpha}_{-m} \right),$$

where $\tilde{\alpha}_{-m}$ is $\tilde{\alpha}$ based on the subsample $S \setminus S_m$.

Remark 3. (Imposing non-negativity on density ratio estimators) We note that the density ratio estimators $\hat{r}(t, x)$ and $\tilde{r}(t, x)$ may yield negative estimates in finite samples even though such a feature does not affect their asymptotic properties below. There are some ways to impose non-negativity on the density ratio estimator. For example, Kanamori *et al.* (2009) proposed to estimate α in (1) by the constrained optimization:

$$\bar{\alpha} = \arg\min_{\alpha} \left[\frac{1}{2} \alpha' \hat{H} \alpha - \hat{h}' \alpha + \lambda \alpha' 1_K \right], \quad \text{s.t. } \alpha \geq 0, \tag{8}$$

where $\hat{H}$ and $\hat{h}$ are constructed by using non-negative basis functions $\psi = (\psi_1, \ldots, \psi_K)'$, 1_K is the K-dimensional vector of ones, and λ is a positive regularization parameter. Kanamori *et al.* (2009) pointed out that although this optimization problem is efficiently implemented by utilizing regularization path tracking, it tends to be numerically unstable especially when there are many change points in the regularization path.

An alternative way to impose non-negativity is to simply modify $\hat{r}(t, x)$ (or $\tilde{r}(t, x)$) as

$$\hat{r}^+(t, x) = \max\{0, \hat{r}(t, x)\},$$

see Sugiyama, Suziki and Kanamori (2012b). We note that the convergence rates of $\hat{r}^+(t, x)$ derived in Section 2.3 are never worse than those of $\hat{r}(t, x)$ since the density ratio function is non-negative by definition. A practical advantage of $\hat{r}^+(t, x)$ compared to the estimator based on (8) is that it does not involve numerical optimization.

Remark 4. (Estimation based on other divergences) Although this paper focuses on the least square criterion to construct the marginals to joint density ratio estimators, other criteria may be employed. For example, as in Sugiyama, Suzuki and Kanamori (2012b), we can also consider to choose the parameters α to minimize the Bregman divergence from $r_0(t, x)$ to $r(t, x; \alpha)$ associated with a strictly convex differentiable function φ, that is

$$\begin{aligned} & BR_{\varphi}(r_0(\cdot, \cdot) \| r(\cdot, \cdot; \alpha)) \\ = & \int f_{TX}(t, x)[\varphi(r_0(t, x)) - \varphi(r(t, x; \alpha)) - \varphi'(r(t, x; \alpha))\{r_0(t, x) - r(t, x; \alpha)\}] dt dx, \end{aligned}$$

where φ' is the derivative of φ. This formulation covers various density estimation methods, such as the least squares ($\varphi(a) = (a-1)^2/2$), Kullback-Leibler divergence

($\varphi(a) = t\log t - t$), and Basu *et al.*'s (1998) power divergence ($\varphi(a) = (t^{1+\alpha} - t)/\alpha$ for $\alpha > 0$). Although it is beyond the scope of this paper, we expect that analogous theoretical properties can be established for this estimation approach.

Remark 5. (Case of discrete T) Although we present our methodology for the case of continuous T, our estimation approach can be also adapted to the case where T is discrete. In this case, we understand $f_T(t)$ and $f_{TX}(t,x)$ as the marginal probability mass function for T and its joint density, respectively, and construct the ratio estimator in the same manner.

2.3. **Theoretical properties.** This subsection presents the asymptotic properties of our estimators $\hat{r}(t,x) = r(t,x;\hat{\alpha})$ in (5) and $\tilde{r}(t,x) = r(t,x;\tilde{\alpha})$ in (6) for the marginals to joint density ratio $r_0(t,x) = \frac{f_T(t)f_X(x)}{f_{TX}(t,x)}$. Let $\|\cdot\|$ be the Euclidean norm, $\|a\|_{P,2} = \sqrt{E[\|a(T,X)\|^2]}$, and $\xi_K = \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}\|\psi(t,x)\|$. We impose the following assumptions.

Assumption.

- **(i):** *$\{T_i, X_i\}_{i=1}^n$ is an i.i.d. sample of $(T,X) \sim P$ with support $\mathcal{T}\times\mathcal{X}$ and* $0 < \inf_{(t,x)\in\mathcal{T}\times\mathcal{X}} r_0(t,x) \leq \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} r_0(t,x) < \infty$.
- **(ii):** *$K = K_n \in \mathbb{N}$ is a sequence depending on the sample size n such that $K_n \to \infty$. Uniformly over all n, the eigenvalues of $H = E[\psi(T,X)\psi(T,X)']$ are bounded above and away from zero. Furthermore, $\xi_K^2 \log K/n \to 0$ as $n \to \infty$.*
- **(iii):** *For each $K \in \mathbb{N}$, there exist finite constants c_K and ℓ_K such that*

$$\|r_0 - \psi'\alpha^*\|_{P,2} \leq c_K, \qquad \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |r_0(t,x) - \psi(t,x)'\alpha^*| \leq \ell_K c_K.$$

Assumption (i) is on the data. The i.i.d. assumption may be relaxed to weakly dependent data by replacing the stochastic bounds in the proof with dependent counterparts.

Assumption (ii) is on the basis functions $\{\psi_\ell\}_{\ell=1}^K$, and Assumption (iii) says the marginals to joint density ratio r_0 is well approximated by the basis functions. These assumptions are common in the literature of nonparametric series or sieve estimation and satisfied by popular basis functions, such as polynomials, splines, and wavelets (see, e.g., Newey, 1997).

Let $a \lesssim_P b$ mean $a = O_p(b)$. Under the above assumptions, the asymptotic properties of our estimator $\hat{r} = \psi'\hat{\alpha}$ for r_0 are obtained as follows.

Theorem 1. *(Convergence rates of estimator) Suppose that Assumptions (i)-(iii) hold true. Then*

$$\|\hat{r} - r_0\|_{P,2} \lesssim_P \sqrt{\frac{K}{n}} + c_K, \tag{9}$$

$$\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\hat{r}(t,x) - r_0(t,x)| \lesssim_P \xi_K \sqrt{\frac{K}{n}} + \ell_K c_K. \tag{10}$$

These convergence rates are standard in the literature. The first terms in the rates are due to the estimation risk for the best linear approximation $\psi'\alpha^*$ by the estimator $\hat{r}$, and the second terms in the rates come from the approximation bias for r_0 by $\psi'\alpha^*$. Similar to Theorem 1, we can obtain the convergence rates of the regularized estimator.

Theorem 2. *(Convergence rates of regularized estimator) Suppose that Assumptions (i)-(iii) hold true. Then for $\lambda = o(n^{-1/2})$,*

$$\|\tilde{r} - r_0\|_{P,2} \lesssim_P \sqrt{\frac{K}{n}} + c_K, \tag{11}$$

$$\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\tilde{r}(t,x) - r_0(t,x)| \lesssim_P \xi_K \sqrt{\frac{K}{n}} + \ell_K c_K. \tag{12}$$

Similar comments to Theorem 1 apply. The tuning constants K and λ may be chosen by the cross validation method as in Remark 2.

In the next section, we present applications of the proposed density estimation methods in the context of causal inference.

3. Applications to causal inference

3.1. **Causal effect of continuous treatment.** In this subsection, we employ the framework of Ai *et al.* (2021) for continuous treatment effect analysis, and propose an alternative estimation procedure based on our density ratio estimator for the causal effect.

Let $T \in \mathcal{T}$ be a continuous treatment variable, $Y(t)$ be the potential outcome that would have been observed under treatment level t, and X be a vector of covariates. For each unit i, we observe T_i, X_i, and the outcome $Y_i = Y(T_i)$. Suppose we specify the random outcome function $Y(t)$ by a parametric model $g(t;\beta_0)$, where β_0 solves

$$\beta_0 = \arg\min_{\beta\in\mathcal{B}} \int E[L(Y(t) - g(t;\beta))] f_T(t) dt, \tag{13}$$

for parameter space $\mathcal{B}$, and L is a user-specified loss function. The expectation above is taken with respect to $Y(t)$ for fixed t. Under the ignorability condition (i.e., $Y(t)$ and T are independent conditionally on X for each $t \in \mathcal{T}$), the above criterion for β_0 can be written as (see, Hirano and Imbens, 2004, and Imai and van Dyk, 2004)

$$\int E[L(Y(t) - g(t;\beta))] f_T(t) dt = E[L(Y - g(T;\beta)) r_0(T, X)].$$

By estimating $r_0(t, x)$ by $r(t, x; \hat{\alpha})$ in (5) and evaluating the expectation with the empirical average, β_0 can be estimated by

$$\hat{\beta} = \arg\min_{\beta\in\mathcal{B}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i - g(T_i;\beta)) r(T_i, X_i; \hat{\alpha}). \tag{14}$$

The estimator $\tilde{\beta}$ using the regularized version $\tilde{\alpha}$ in (6) is defined analogously. We impose the following assumptions to study asymptotic properties of $\hat{\beta}$. Let $L'(a) = \frac{dL(a)}{da}$ and $W_i(\beta) = L'(Y_i - g(T_i;\beta))\frac{\partial g(T_i;\beta)}{\partial \beta}$.

Assumption.

- **(iv):** *$Y(t)$ and T are independent conditionally on X for each $t \in \mathcal{T}$. The parameter space $\mathcal{B} \subset \mathbb{R}^p$ is compact. β_0 is a unique solution of (13) and is in the interior of $\mathcal{B}$. $L(Y - g(T;\beta))$ is continuous in β, and $E[\sup_{\beta\in\mathcal{B}} |L(Y - g(T;\beta))|] < \infty$.*
- **(v):** *L is twice continuously differentiable almost everywhere. $g(t;\beta)$ is twice continuously differentiable in β. $G = E\left[r_0(T,X)\frac{\partial W(\beta_0)}{\partial \beta'}\right]$ is non-singular, $E\left[\sup_{\beta\in\mathcal{N}}\left\|\frac{\partial W(\beta)}{\partial \beta'}\right\|\right] < \infty$ for a neighborhood $\mathcal{N}$ around β_0, and $E\left[\|W(\beta_0)\|^{2+\delta}\right] < \infty$ for some $\delta > 0$.*

Assumptions (iv)-(v) are simplifications of those in Ai *et al.* (2021) by assuming twice differentiability of L. Assumption (iv) is used to show consistency of $\hat{\beta}$, which contains unconfoundedness assumption and some regularity conditions on the objective function. Assumption (v) is imposed to obtain the asymptotic distribution of $\hat{\beta}$. Twice continuous differentiability of L may be relaxed by applying an analogous argument as in Ai *et al.* (2021) based on Chen, Linton and Van Keilegom (2003). Based on these assumptions, the asymptotic properties of $\hat{\beta}$ are obtained as follows.

Proposition 1. *In addition to Assumptions (iv)-(v), suppose that Assumptions (i)-(iii) hold true with $\ell_K c_K = O(K^{-\eta_1})$ for some $\eta_1 > 0$, and there exists $\gamma^* \in \mathbb{R}^K$ such that*

$$\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |E[\nabla_\beta g(T;\beta_0)L'(Y - g(T;\beta_0))|T = t, X = x] - \psi(t,x)'\gamma^*| = O(K^{-\eta_2}), \quad (15)$$

for some $\eta_2 > 0$*. Furthermore,* $\sqrt{n}K^{-(\eta_1+\eta_2)} \to 0$*,* $K^2/n \to 0$*, and* $\xi_K^2 K/n \to 0$ *as* $n \to \infty$*. Then* $\hat{\beta} \xrightarrow{p} \beta_0$ *and*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V), \tag{16}$$

where $V = G^{-1}E[\xi(Y,T,X;\beta_0)\xi(Y,T,X;\beta_0)'](G^{-1})'$ *and*

$$\begin{aligned} \xi(Y,T,X;\beta_0) &= r_0(T,X)\nabla_\beta g(T;\beta_0)L'(Y - g(T;\beta_0)) - r_0(T,X)\nabla_\beta g(T;\beta_0)\epsilon(T,X;\beta_0) \\ &\quad + E[r_0(T,X)\nabla_\beta g(T;\beta_0)\epsilon(T,X;\beta_0)|T] + E[r_0(T,X)\nabla_\beta g(T;\beta_0)\epsilon(T,X;\beta_0)|X]. \end{aligned}$$

This theorem says that our estimator $\hat{\beta}$ for β_0 is consistent and asymptotically normal, and the asymptotic variance V is equivalent to the efficiency bound derived in Ai *et al.* (2021). This asymptotic variance can be estimated by nonparametric methods. In our numerical studies below, we estimate V by series estimation and bootstrap methods.

Remark 6. (Comparison with Ai *et al.* (2021)) It is interesting to note that in the setup of this subsection for continuous treatment effect analysis, the estimation methods for β_0 by Ai *et al.* (2021) and this paper can be presented in a unified way. To be precise, consider the following weight estimation problem

$$\min_{\pi_1,\ldots,\pi_n} \sum_{i=1}^n \rho(\pi_i) \quad \text{s.t.} \quad \frac{1}{n}\sum_{i=1}^n \pi_i \psi(T_i, X_i) = \frac{1}{n^2}\sum_{i=1}^n \sum_{j=1}^n \psi(T_i, X_j), \tag{17}$$

where ψ contains a constant function so that $\sum_{i=1}^n \pi_i = n$. Then the weight function $r(T_i, X_i; \hat{\alpha})$ for the objective function (14) can be obtained as the solution for π_i with $\rho(\pi) = (\pi - 1)^2$. On the other hand, the weight function for Ai *et al.* (2021) is given by the solution for π_i with $\rho(\pi) = \pi \log \pi$ and multiplicative basis functions $\psi(T,X) = \psi_a(T)\psi_b(X)$ for some ψ_a and ψ_b. As pointed out by Ai *et al.* (2021), the above minimization problem is different from the one for generalized empirical likelihood

(Newey and Smith, 2004) due to the right hand side of the constraint for π_i, which is random in the current context. However, Proposition 1 indicates that such difference of weight constructions for the objective function in (14) is asymptotically negligible as far as we are concerned with the first-order asymptotic properties of the estimator for β_0.

One practical advantage of our approach is that our estimated weights $r(T_i, X_i; \hat{\alpha})$ in (14) do not require numerical optimization while the weights for Ai *et al.*'s (2021) method need to numerically solve the dual problem of (17). Indeed Ai *et al.* (2021, p. 801) reported some numerical issues associated with this optimization when the dimension of covariates is large.

3.2. **Dose-response curve.** Based on the setup in Section 3.1, we can also consider to estimate the dose-response curve:

$$m_0(t) = E[Y(t)] = E[r_0(T, X)Y|T = t], \quad \text{for } t \in \mathcal{T}.$$

Several papers studied different estimation methods for $m_0(t)$, such as Kennedy *et al.* (2017), Fong, Hazlett and Imai (2018), Ai *et al.* (2021), and Semenova and Chernozhukov (2020).

For example, the estimation approach by Ai *et al.* (2021) can be adapted as follows. By estimating $r_0(t, x)$ by $r(t, x; \hat{\alpha})$ and evaluating the conditional expectation with the series method, $m_0(t)$ can be estimated by $\hat{m}(t) = u(t)'\hat{\delta}$, where

$$\hat{\delta} = \arg\min_{\delta} \frac{1}{n} \sum_{i=1}^{n} \{r(T_i, X_i; \hat{\alpha})Y_i - u(T_i)'\delta\}^2,$$

and $u(t) = (u_1(t), \ldots, u_K(t))'$ is a vector of basis functions. The consistency and pointwise asymptotic normality of $\hat{m}(t)$ can be established by applying the arguments in Ai *et al.* (2021).

Alternatively, we can construct a double robust version of the above estimator based on Semenova and Chernozhukov (2020). Note that the dose-response curve can be written as

$$m_0(t) = E\left[r_0(T,X)(Y - E[Y|T,X]) + \int E[Y|T,x] f_X(x)dx \middle| T = t \right], \tag{18}$$

which is conditionally orthogonal to the nuisance functions $r_0(t,x)$ and $E[Y|t,x]$. Define

$$\tilde{Y}_i = r(T_i, X_i; \hat{\alpha})(Y_i - \hat{\mu}(T_i, X_i)) + \frac{1}{n}\sum_{j=1}^{n} \hat{\mu}(T_i, X_j),$$

where $\hat{\mu}(t,x)$ is a nonparametric estimator of $E[Y|t,x]$. Based on (18), $m_0(t)$ can be estimated by $\tilde{m}(t) = u(t)'\tilde{\delta}$, where

$$\tilde{\delta} = \arg\min_{\delta} \frac{1}{n}\sum_{i=1}^{n} \{\tilde{Y}_i - u(T_i)'\delta\}^2.$$

The asymptotic properties of $\tilde{m}(t)$ can be obtained by adapting the results in Semenova and Chernozhukov (2020) to our context.

4. Simulation

4.1. **Setting.** In this section we illustrate the finite sample performance of the proposed estimation method by numerical experiments. We focus on the continuous treatment effect analysis in Section 3.1, and adopt the simulation designs in Ai *et al.* (2021). In particular, we consider the data generating processes (DGPs):

$$\text{(DGP-L2) } T = 1 + 0.2(X_1 + X_2) + \xi, \quad Y = 1 + \frac{1}{2}(X_1 + X_2) + T + \epsilon,$$

$$\text{(DGP-N2) } T = 0.1(X_1 + X_2)^2 + \xi, \quad Y = \frac{1}{2} + \frac{1}{4}(X_1 + X_2)^2 + T + \epsilon,$$

where X_1, X_2, ξ, and ϵ are mutually independent and follow $N(0,1)$. The notation of the DGPs corresponds to the one in Ai *et al.* (2021), where "L2" means that T and Y depend linearly on the two covariates (X_1, X_2), while "N2" means that T and Y depend non-linearly on (X_1, X_2). We employ the quadratic loss function $L(u) = u^2/2$ and the linear link function $g(T;\beta) = \beta_1 + \beta_2 T$ with the true value $(\beta_{01}, \beta_{02}) = (1,1)$.

For the basis functions $\psi(T,X)$ to approximate $r_0(T,X)$, we consider separable functions in T and X. Then the series approximation can be written as

$$r(T,X) = u_{K_1}(T)' A v_{K_2}(X),$$

where $u_{K_1}(T)$ and $v_{K_2}(X)$ are vectors of basis functions and A is a $K_1 \times K_2$ matrix with the (i,j)-th element corresponding to the coefficient parameter for the product of the i-th element of $u_{K_1}(T)$ and the j-th element of $v_{K_2}(X)$. For $u_{K_1}(T)$, we consider

$$u_2(T) = (1,T)', \qquad u_3(T) = (1,T,T^2)', \qquad u_4(T) = (1,T,T^2,T^3)'.$$

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

$$\begin{aligned} v_3(X) &= (1,X_1,X_2)', \qquad v_6(X) = (1,X_1,X_2,X_1^2,X_2^2,X_1X_2)', \\ v_{10}(X) &= (1,X_1,X_2,X_1^2,X_2^2,X_1X_2,X_1^3,X_2^3,X_1^2X_2,X_1X_2^2)'. \end{aligned}$$

Our estimator $\hat{\beta}$ in (14) is computed for the sample sizes $n = 100, 500, 1000$, and the number of Monte Carlo replications is 1000. For the regularization parameter to implement (6), we consider $\lambda \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$.

4.2. **Cross validation, point estimation, and interval estimation.** Table 1 illustrates the result of the cross validation method to choose the series lengths (K_1, K_2)

described in Remark 2. We apply the 5-fold cross validation for each replicated dataset and report the most frequently chosen pair (K_1, K_2) over 1000 datasets. For both DGPs, a basis of higher degree is chosen as the sample size grows and the appropriate model is chosen with $n = 1000$ in the sense of the degree of basis. This result suggests that the cross validation works effectively for model selection in our method.

TABLE 1. Most frequently chosen pair (K_1, K_2) by the 5-fold cross validation

		$n = 100$	$n = 500$	$n = 1000$
DGP-L2	(K_1, K_2)	$(2, 3)$	$(2, 3)$	$(3, 3)$
	λ	0.05	0.01	0.01
DGP-N2	(K_1, K_2)	$(2, 3)$	$(2, 6)$	$(3, 6)$
	λ	0.1	0.05	0.05

Tables 2 and 3 summarize the simulation results on point estimation of the intercept β_1 and slope β_2 under DGP-L2 and DGP-N2, respectively. For comparison, the simulation results based on the joint to marginals density ratio described in Remark 1 and the results from Ai *et al.* (2021) are also included in the tables. For our estimator ("Ratio") and the joint to marginals estimator ("Ratio JM"), the results of the regularized estimator chosen by the 5-fold cross validation are shown for each sample size. For Ai *et al.*'s estimator ("ALMZ"), the results for the 5-fold cross validation are borrowed from Tables 5-8 in their supplemental material. In Appendix A.4, we present additional simulation results for different choices of the series lengths (K_1, K_2) and estimator without regularization. The results are similar to the ones in Tables 2 and 3.

Overall, our estimator performs better than the other estimators. In comparison with the joint to marginals estimator, our estimator has much smaller standard deviations (SDs) and root mean squared errors (RMSEs). Table 4 describes the quantiles of the SDs of the estimated density ratio $\hat{r}_0$ by our estimator and the joint to the marginals estimator, $r(t, x; \hat{\alpha})$ and $1/r(t, x; \hat{\alpha}_{\text{joint to marginals}})$, over 1000 replicated datasets, where the sample size is $n = 1000$ and the basis degree is $(K_1, K_2) = (2, 6)$ without regularization.

As mentioned in Remark 1, the reciprocal of the joint to marginals estimator tends to have a larger SD than our estimator for both DGPs, which leads to the unstable results of the joint to marginals estimator. These results show the advantage of the proposed method. In comparison with Ai *et al.*'s estimator, the results are similar under DGP-L2 (linear system), whereas our estimator outperforms Ai *et al.*'s under DGP-N2 (non-linear system) in most of the cases, as the sample size grows.

Note that the results of Ai *et al.* are obtained based on "trimmed samples" in the sense that, when calculating the bias, SD, and RMSE, they discard samples in which the mean of the estimated ratios is not included in $[0.5, 2]$ to eliminate the impact of numerical instability. In contrast, the mean of the estimated ratios in our (non-regularized) estimator is always one because of its construction. Even though that is not true for our regularized estimator, we observe that the mean of the estimated ratios is included in $[0.9, 1.1]$ for all the Monte Carlo replications and it does not need such a trimming. In this sense, our estimators are more tractable than the one by Ai *et al.* (2021).

TABLE 2. Simulation results on point estimation for DGP-L2

		$n = 100$			$n = 500$			$n = 1000$		
	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
β_1	Ratio	−0.028	0.172	0.174	−0.009	0.074	0.075	−0.012	0.054	0.055
	Ratio JM	−0.025	0.870	0.870	−0.083	2.260	2.262	−0.121	2.255	2.259
	ALMZ	0.001	0.174	0.174	−0.003	0.074	0.074	−0.004	0.054	0.054
β_2	Ratio	0.023	0.112	0.115	0.010	0.048	0.049	0.007	0.034	0.035
	Ratio JM	0.038	0.918	0.919	0.057	1.643	1.644	0.187	4.095	4.099
	ALMZ	0.003	0.113	0.113	0.005	0.052	0.052	0.006	0.037	0.038

TABLE 3. Simulation results on point estimation for DGP-N2

		$n = 100$			$n = 500$			$n = 1000$		
	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
β_1	Ratio	−0.041	0.119	0.126	0.004	0.056	0.056	0.008	0.041	0.041
	Ratio JM	0.062	3.429	3.430	−0.097	2.033	2.035	0.045	1.570	1.571
	ALMZ	−0.041	0.130	0.136	−0.027	0.060	0.065	−0.019	0.045	0.049
β_2	Ratio	0.165	0.134	0.212	0.020	0.047	0.051	0.025	0.034	0.042
	Ratio JM	−0.103	4.121	4.122	0.098	2.603	2.605	0.046	1.917	1.917
	ALMZ	0.126	0.138	0.187	0.100	0.089	0.134	0.083	0.079	0.115

TABLE 4. Quantiles of the standard deviations of the estimated density ratio $\hat{r}_0$ over the replicated datasets

		Quantile				
		0	25	50	75	100
DGP-L2	Ratio	0.209	0.313	0.342	0.376	0.499
	Ratio JM	0.207	0.329	0.413	0.679	289.148
DGP-N2	Ratio	0.187	0.287	0.318	0.351	0.483
	Ratio JM	0.161	0.340	0.512	0.995	299.376

Tables 5 and 6 report the simulation results on interval estimation, where the coverage ("CV") is the proportion of the times that the true value is included in the estimated 95% confidence interval and the average length ("AL") is the mean of the length of the estimated 95% confidence interval. For simplicity, the degree of the basis function is fixed as $(K_1, K_2) = (2, 6)$ for our estimator and Ai *et al.*'s and the regularization parameter λ is set to 0 for our estimator. To estimate several conditional expectations for the asymptotic variance in (16), we use a series approximation as in the density ratio estimation. We employ the same choice of the degree of basis and regularization parameter as the density ratio estimation, i.e., $(K_1, K_2) = (2, 6)$ and $\lambda = 0$.

For both DGPs, the coverage of our estimator is reasonably close to 95% even with the small sample of $n = 100$, which indicates that the asymptotic approximation works well to perform hypothesis testing with a finite sample. The average length becomes smaller as the sample size grows for all the cases. In the estimation of the slope β_2, which is of interest in causal analysis, the coverage of our estimator increases with a smaller average length, while with Ai *et al.*'s estimator the smaller average length leads to the smaller coverage, as the sample size grows. This characteristic of our estimator may also be advantageous for hypothesis testing.

4.3. **Sensitivity analysis.** Figures 1 and 2 describe the sensitivity of the RMSE of the estimated slope $\hat{\beta}_2$ against the choice of the regularization parameter λ for each (K_1, K_2). Note that the horizontal axis of the graphs is not in a linear scale. For both DGPs, the

Table 5. Simulation results on interval estimation for DGP-L2

		$n = 100$		$n = 500$		$n = 1000$	
	Estimator	CV	AL	CV	AL	CV	AL
β_1	Ratio	0.935	0.662	0.954	0.293	0.960	0.207
	ALMZ	0.967	0.722	0.960	0.304	0.959	0.221
β_2	Ratio	0.926	0.445	0.938	0.192	0.952	0.136
	ALMZ	0.956	0.494	0.945	0.205	0.929	0.149

Table 6. Simulation results on interval estimation for DGP-N2

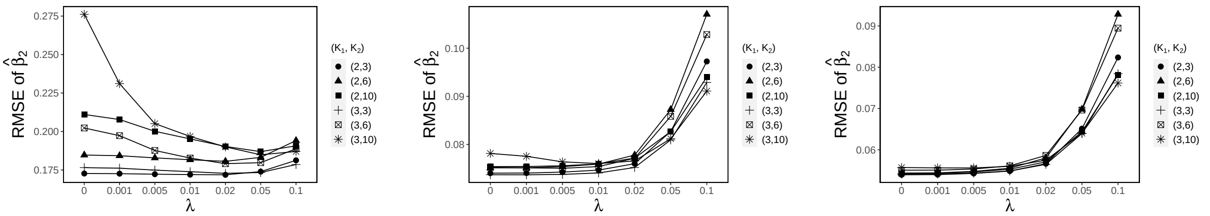
		$n = 100$		$n = 500$		$n = 1000$	
	Estimator	CV	AL	CV	AL	CV	AL
β_1	Ratio	0.937	0.469	0.929	0.208	0.920	0.147
	ALMZ	0.969	0.537	0.953	0.236	0.967	0.171
β_2	Ratio	0.931	0.474	0.960	0.221	0.970	0.166
	ALMZ	0.965	0.498	0.937	0.208	0.934	0.153

case with the small sample ($n = 100$) and the cases with the medium to large samples ($n = 500, 1000$) exhibit similar behaviors, respectively. For $n = 100$, a too large model (e.g. $(K_1, K_2) = (3, 10)$) with respect to the sample size has a large RMSE. The RMSEs decrease as λ increases and they are not so sensitive to a large λ. For $n = 500$ and 1000, under DGP-L2 (Figure 1), most of the (K_1, K_2) pairs exhibit a similar behavior with respect to the RMSE as λ increases. On the other hand, for $n = 500$ and 1000 under DGP-N2 (Figure 2), it becomes clear that too simplified models (e.g. $(K_1, K_2) = (2, 3)$ and $(3, 3)$) have large RMSEs, while sufficiently large models (e.g. $(K_1, K_2) = (3, 6)$ and $(3, 10)$) show a reasonable behavior as in the cases of DGP-L2. Importantly, for both DGPs, the RMSE is not so sensitive around the λ chosen by the cross validation (see Table 1) with sufficiently large models.[1] This result may add to the reliability of the results shown above.

5. Empirical example

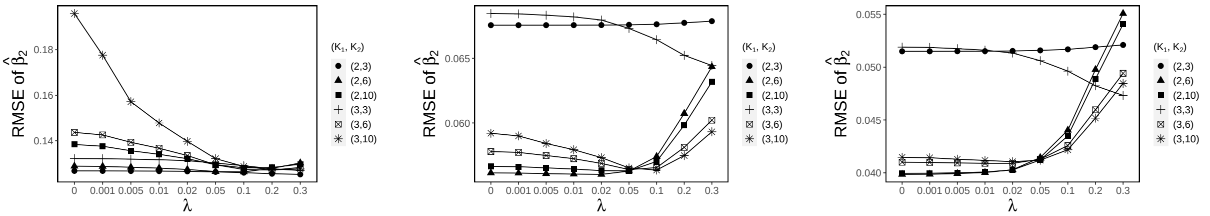
We apply the proposed estimation method to the U.S. presidential campaign data, which is originally analyzed by Urban and Niebler (2014) and followed by Fong, Hazlett

[1] Note that λ is chosen to minimize the cross validation criterion (7) so it does not necessarily minimize the RMSE of $\hat{\beta}_2$.



(A) $n = 100$ (B) $n = 500$ (C) $n = 1000$

FIGURE 1. Sensitivity analysis for DGP-L2



(A) $n = 100$ (B) $n = 500$ (C) $n = 1000$

FIGURE 2. Sensitivity analysis for DGP-N2

and Imai (2018) and Ai *et al.* (2021). In presidential campaigns, competitive states usually gather candidates' attention and most advertising efforts. However, if political advertisements contribute to more donations, candidates may also want to distribute their advertising efforts for non-competitive states. Therefore, the purpose of the study is to find if there is a causal effect of political advertisements on the amount of donations. A more detailed background of the study can be found in Fong, Hazlett and Imai (2018, Section 2).

The treatment of interest is the number of political advertisements aired in each zip code (across $n = 16265$ zip codes), which ranges from 0 to 22379 and can be considered as a continuous variable. Urban and Niebler (2014) conduct their analysis by dichotomizing the variable based on whether it is greater than 1000 or not and conclude that there is a significant effect of the political advertisements on donations. However, Fong, Hazlett and Imai (2018) point out that, due to the dichotomization, Urban and Niebler's (2014) result is difficult to interpret and there is a large information loss. To mitigate this issue, Fong, Hazlett and Imai (2018) apply the covariate balancing generalized propensity score methodology, where they deal with the treatment variable as a continuous variable. In contrast to Urban and Niebler (2014), their result suggests that there is no significant effect of the political advertisements on donations. In addition, Ai *et al.* (2021) report similar results to Fong, Hazlett and Imai (2018). Here, we investigate how those results will change when using the proposed estimation method for the continuous treatment.

Following the previous studies mentioned above, we include the eight variables into our analysis as covariates: total population, population density, median income, percentage of Hispanic, percentage of black, percentage of over age 65, percentage of college graduates, and a binary indicator of whether it is possible to commute to the zip code from a competitive state. We apply the log-transformation on the outcome (amount of donations) Y, the treatment (number of advertisements) T, and the first three variables in the covariates. For the link function, we consider the quadratic function $g(T, \beta) = \beta_1 + \beta_2 T + \beta_3 T^2$. For the degree of the basis, we consider up to the third order for T and X respectively and choose an optimal combination by the cross validation method described in Remark 2 from the following list:

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

$$\begin{aligned} v_9(X) &= (1, X_1, \ldots, X_8), \\ v_{44}(X) &= (1, X_1, \ldots, X_8, X_1^2, X_1X_2, \ldots, X_7^2, X_7X_8), \text{ or} \\ v_{156}(X) &= (1, X_1, \ldots, X_8, X_1^2, X_1X_2, \ldots, X_7^2, X_7X_8, X_1^3, X_1^2X_2, \ldots, X_7^2X_8). \end{aligned}$$

Note that X_8 is a binary variable so the elements including X_8 to the second/third powers are dropped, which leads to $K_1 \in \{2, 3, 4\}$ and $K_2 \in \{9, 44, 156\}$. Similar to Section

4, we consider $\lambda \in \{0.001, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.5\}$ for the regularization parameter. As a result of the 5-fold cross validation, $(K_1, K_2) = (4, 44)$ and $\lambda = 0.005$ are chosen. As in Section 4, we use a series approximation to estimate the asymptotic variance in (16) with the same degree of basis and the same regularization parameter, $(K_1, K_2) = (4, 44)$ and $\lambda = 0.005$.

Table 7 reports the point estimates for $(\beta_1, \beta_2, \beta_3)$, 95% asymptotic confidence intervals ("Asymptotic CI"), and Bootstrap 95% confidence intervals ("Bootstrap CI") for our estimator. The asymptotic confidence interval of each coefficient is obtained as $\hat{\beta}_j \pm 1.96\sqrt{\hat{V}_{jj}/n}$, where $\hat{V}_{jj}$ is the estimates of (j, j)-element of the asymptotic variance in (16). The number of bootstrap replications is 1000. In our example, the asymptotic and bootstrap confidence intervals are reasonably close. The confidence intervals for β_2 and β_3 both contain zero, that is, the treatment (number of advertisements) has no significant impact on the outcome. This result supports the existing literature that deals with the treatment of interest as a continuous variable (Fong, Hazlett and Imai, 2018; Ai *et al.*, 2021), while being against the original analysis where the treatment variable is dichotomized (Urban and Niebler, 2014).

Table 7. Result of the empirical analysis: the estimated coefficients of the link function

	Estimate	C.I. (Asymptotics)	C.I. (Bootstrap)
β_1	1.248	$(1.207, 1.288)$	$(1.198, 1.299)$
β_2	0.010	$(-0.016, 0.036)$	$(-0.029, 0.041)$
β_3	-0.001	$(-0.004, 0.002)$	$(-0.004, 0.003)$

Appendix A. Mathematical appendix

A.1. **Proof of Theorem 1.** We normalize $H = I$ without loss of generality. We first prove (9). Note that

$$\|\hat{r} - r_0\|_{P,2} \leq \|\hat{r} - \psi'\alpha^*\|_{P,2} + \|\psi'\alpha^* - r_0\|_{P,2} \leq \|\psi'(\hat{\alpha} - \alpha^*)\|_{P,2} + c_K,$$

where the first inequality follows from the triangle inequality, and the second inequality follows from Assumption (iii). Under the normalization $H = I$, we have

$$\|\psi'(\hat{\alpha} - \alpha^*)\|_{P,2} = \left[\int (\hat{\alpha} - \alpha^*)'\psi(t,x)\psi(t,x)'(\hat{\alpha} - \alpha^*) dP(t,x)\right]^{1/2} = \|\hat{\alpha} - \alpha^*\|.$$

The matrix law of large numbers (see Belloni, Chernozhukov, Chetverikov and Kato 2015, Lemma 6.2) implies

$$\|\hat{H} - H\| = o_p(1) \qquad \text{if } \frac{\xi_K^2 \log K}{n} \to 0. \tag{19}$$

Hence, with probability approaching one, all eigenvalues of $\hat{H}$ are bounded away from zero, and then

$$\|\hat{\alpha} - \alpha^*\| = \|\hat{H}^{-1}(\hat{h} - \hat{H}\alpha^*)\| \lesssim_P \|\hat{h} - \hat{H}\alpha^*\|.$$

Therefore, it is sufficient for the conclusion to show that

$$\|\hat{h} - \hat{H}\alpha^*\| \lesssim_P \sqrt{\frac{K}{n}}. \tag{20}$$

Now, let $\varphi(t) = \int \psi(t,x) f_X(x)dx$. Observe that

$$\begin{aligned}
& E\left[\|\hat{h} - \hat{H}\alpha^*\|^2\right] \\
= \ & E\left[\left\|\left\{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(T_i,X_j) - \frac{1}{n}\sum_{i=1}^{n}\varphi(T_i)\right\} - \left\{\hat{H}\alpha^* - \frac{1}{n}\sum_{i=1}^{n}\varphi(T_i)\right\}\right\|^2\right] \\
\leq \ & 2E\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{n}\sum_{j=1}^{n}\{\psi(T_i,X_j) - \varphi(T_i)\}\right\}\right\|^2\right] + 2E\left[\left\|\hat{H}\alpha^* - \frac{1}{n}\sum_{i=1}^{n}\varphi(T_i)\right\|^2\right] \\
= \ & 2E\left[\left\|\frac{1}{n}\sum_{i=1}^{n}\{\psi(T_i,X_i)\psi(T_i,X_i)'\alpha^* - \varphi(T_i)\}\right\|^2\right] + O\left(\frac{K}{n}\right) \\
= \ & \frac{2}{n}E\left[\|\psi(T,X)\psi(T,X)'\alpha^* - \varphi(T)\|^2\right] + O\left(\frac{K}{n}\right) \\
\leq \ & \frac{4}{n}E\left[\|\psi(T,X)\|^2\right]E\left[\|\psi(T,X)'\alpha^*\|^2\right] + \frac{4}{n}E\left[\|\varphi(T)\|^2\right] + O\left(\frac{K}{n}\right) \\
\leq \ & \frac{4}{n}\left\{E\left[\|\psi(T,X)\|^2\right](\|r_0\|_{P,2}^2 + \|r_0 - \psi'\alpha^*\|_{P,2}^2)\right\} + \frac{4}{n}\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} r_0(t,x)E\left[\|\psi(T,X)\|^2\right] + O\left(\frac{K}{n}\right) \\
= \ & O\left(\frac{K}{n}\right),
\end{aligned}$$

where the first equality follows from the definition of $\hat{h}$, the first inequality follows from $\|A - B\|^2 \geq 0$, the second equality follows from the definition of $\hat{H}$ and Assumption (ii), the third equality follows from Assumption (i) (implying that $\{\psi(T_i,X_i)\psi(T_i,X_i)'\alpha^* - E_X[\psi(T_i,X)]\}_{i=1}^N$ is an independent zero mean sequence), the second inequality follows from the Cauchy-Schwarz inequality, the last inequality follows from the triangle inequality and

$E[\|\varphi(T)\|^2] \leq \int\int \|\psi(t,x)\|^2 f_T(t)f_X(x)dtdx$ (by the Cauchy-Schwarz inequality), and the last equality follows from Assumption (ii)-(iii). Thus, the Markov inequality yields (20) and the conclusion follows.

We now prove (10). This theorem follows by

$$\begin{aligned}\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}|\hat{r}(t,x)-r_0(t,x)| &\leq \xi_K \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}\left|\left(\frac{\psi(t,x)}{\|\psi(t,x)\|}\right)'(\hat{\alpha}-\alpha^*)\right| + \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}|\psi(t,x)'\alpha^* - r_0(t,x)| \\ &\lesssim_P \xi_K\sqrt{\frac{K}{n}} + \ell_K c_K,\end{aligned}$$

where the first inequality follows from the triangle inequality and definitions of $q(t,x)$ and ξ_K, and the second inequality follows from (20) and Assumption (iii).

A.2. **Proof of Theorem 2.** We first show (11). Note that

$$\|\tilde{r}-r_0\|_{P,2} \leq \|\tilde{r}-\psi'\alpha^*\|_{P,2} + \|\psi'\alpha^* - r_0\|_{P,2} \leq \|\psi'(\tilde{\alpha}-\alpha^*)\|_{P,2} + c_K,$$

where the first inequality follows from the triangle inequality, and the second inequality follows from Assumption (iii). Under the normalization $H = I$, we have

$$\|\psi'(\tilde{\alpha}-\alpha^*)\|_{P,2} = \left[\int(\tilde{\alpha}-\alpha^*)'\psi(t,x)\psi(t,x)'(\tilde{\alpha}-\alpha^*)dP(t,x)\right]^{1/2} = \|\tilde{\alpha}-\alpha^*\|.$$

Also note that

$$\begin{aligned}\|\tilde{\alpha}-\alpha^*\| &= \|(\hat{H}+\lambda I)^{-1}\{\hat{h}-(\hat{H}+\lambda I)\alpha^*\}\| \lesssim_P \|\hat{h}-(\hat{H}+\lambda I)\alpha^*\| \\ &\lesssim_P \|\hat{h}-\hat{H}\alpha^*\| + \lambda\|\alpha^*\| \lesssim_P \sqrt{\frac{K}{n}} + \lambda\sqrt{K}, \qquad (21)\end{aligned}$$

where the first inequality follows from the fact that all eigenvalues of $\hat{H}$ are bounded away from zero with probability approaching one (by 19), the second inequality follows from the triangle inequality, and the third inequality follows from (20) and definition of α^* in (3). Combining these results with $\lambda = o(n^{-1/2})$, the conclusion follows.

We now show (12). This theorem follows by

$$\begin{aligned}
\sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\tilde{r}(t,x) - r_0(t,x)| &\leq \xi_K \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} \left| \left(\frac{\psi(t,x)}{\|\psi(t,x)\|} \right)' (\tilde{\alpha} - \alpha^*) \right| + \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\psi(t,x)'\alpha^* - r_0(t,x)| \\
&\lesssim_P \xi_K \sqrt{\frac{K}{n}} + \ell_K c_K,
\end{aligned}$$

where the first inequality follows from the triangle inequality and definitions of $q(t,x)$ and ξ_K, and the second inequality follows from (21) and Assumption (iii).

A.3. **Proof of Proposition 1.** We first show consistency of $\hat{\beta}$. By the triangle inequality, we have

$$\begin{aligned}
& \sup_{\beta\in\mathcal{B}} \left| \frac{1}{n}\sum_{i=1}^{n} L(Y_i - g(T_i;\beta)) r(T_i, X_i; \hat{\alpha}) - E[L(Y - g(T;\beta)) r_0(T,X)] \right| \\
\leq & \sup_{\beta\in\mathcal{B}} \left| \frac{1}{n}\sum_{i=1}^{n} L(Y_i - g(T_i;\beta)) \{ r(T_i, X_i; \hat{\alpha}) - r_0(T_i, X_i) \} \right| \\
& + \sup_{\beta\in\mathcal{B}} \left| \frac{1}{n}\sum_{i=1}^{n} L(Y_i - g(T_i;\beta)) r_0(T_i, X_i) - E[L(Y - g(T;\beta)) r_0(T,X)] \right| \\
\equiv & T_1 + T_2.
\end{aligned}$$

For T_2, the uniform law of large numbers (Lemma 2.4 in Newey and McFadden, 1994) under Assumption (iv) implies $T_2 \xrightarrow{p} 0$. For T_1, we also have

$$T_1 \leq \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |r(t,x;\hat{\alpha}) - r_0(t,x)| \sup_{\beta\in\mathcal{B}} \left(\frac{1}{n}\sum_{i=1}^{n} |L(Y_i - g(T_i;\beta))| \right) \xrightarrow{p} 0,$$

where the convergence follows from Theorem 1 and the uniform law of large numbers. Therefore, by applying Newey and McFadden (1994, Theorem 2.1), we obtain the consistency $\hat{\beta} \xrightarrow{p} \beta_0$.

We next show the asymptotic distribution of $\hat{\beta}$. Let $\hat{r}_i = \hat{r}(T_i, X_i)$, $r_{0i} = r(T_i, X_i)$, $W_i = L'(Y_i - g(T_i; \beta_0))\frac{\partial g(T_i;\beta_0)}{\partial \beta}$, and $\omega(t, x) = E[W|T = t, X = x]$. An expansion of the first-order condition of $\hat{\beta}$ yields

$$0 = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i W_i(\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i W_i - \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{\partial W_i(\bar{\beta})}{\partial \beta'}(\hat{\beta} - \beta_0),$$

where $\bar{\beta}$ is a point on the line joining $\hat{\beta}$ and β_0. Thus, it is sufficient for the conclusion to show that

$$\hat{G} \equiv \frac{1}{n}\sum_{i=1}^{n} \hat{r}_i \frac{\partial W_i(\bar{\beta})}{\partial \beta'} \xrightarrow{p} G, \tag{22}$$

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \hat{r}_i W_i \xrightarrow{d} N(0, V). \tag{23}$$

For (22), the triangle inequality yields

$$\begin{aligned} |\hat{G} - G| \quad &\leq \quad \sup_{\beta \in \mathcal{N}} \left| \frac{1}{n}\sum_{i=1}^{n} r_{0i} \frac{\partial W_i(\beta)}{\partial \beta'} - E\left[r_{0i} \frac{\partial W_i(\beta)}{\partial \beta'} \right] \right| + \left| E\left[r_{0i} \frac{\partial W_i(\bar{\beta})}{\partial \beta'} \right] - G \right| \\ &\quad + \sup |\hat{r}_i - r_{0i}| \sup_{\beta \in \mathcal{N}} \frac{1}{n}\sum_{i=1}^{n} \left| \frac{\partial W_i(\beta)}{\partial \beta'} \right| \\ &\xrightarrow{p} \quad 0, \end{aligned}$$

where the convergence follows from the uniform law of large numbers under Assumption (v) and the continuous mapping theorem.

We now show (23). Note that $E[r_0(T,X)W] = \int_{\mathcal{T}}\int_{\mathcal{X}} \omega(t,x) f_T(t) f_X(x) dt dx = 0$. We decompose

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i W_i &= \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\{W_i - \omega(T_i,X_i)\} + \frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\omega(T_i,X_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} r_{0i}\{W_i - \omega(T_i,X_i)\} + \frac{1}{n}\sum_{i=1}^{n}\left\{\int \omega(T_i,x) f_X(x)dx + \int \omega(t,X_i) f_T(t) dt\right\} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_{0i})\{W_i - \omega(T_i,X_i)\} \\
&\quad + \left(\begin{array}{l}
\frac{1}{n}\sum_{i=1}^{n}\hat{r}_i\{\omega(T_i,X_i) - \psi(T_i,X_i)'\gamma^*\} \\
-\frac{1}{n}\sum_{i=1}^{n}\left\{\int \omega(T_i,x) f_X(x)dx + \int \omega(t,X_i) f_T(t)dt\right\} \\
+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\psi(T_i,X_j)'\gamma^*
\end{array}\right) \\
&\equiv M_1 + M_2 + R_1 + R_2.
\end{aligned}$$

By the facts that $E[r_{0i}W_i|T_i] = \int \omega(T_i,x) f_X(x)dx$ and $E[r_{0i}W_i|X_i] = \int \omega(t,X_i) f_T(t)dt$, M_2 is written as $M_2 = \frac{1}{n}\sum_{i=1}^{n}\{E[r_{0i}W_i|T_i] + E[r_{0i}W_i|X_i]\}$. Thus, the central limit theorem yields

$$\sqrt{n}(M_1 + M_2) \xrightarrow{d} N(0,V).$$

It remains to show that $R_1, R_2 = o_p(n^{-1/2})$. For R_1,

$$\begin{aligned}
R_1 &= \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - \psi(T_i,X_i)'\alpha^*)\{W_i - \omega(T_i,X_i)\} + \frac{1}{n}\sum_{i=1}^{n}(\psi(T_i,X_i)'\alpha^* - r_{0i})\{W_i - \omega(T_i,X_i)\} \\
&\lesssim_P (\hat{\alpha} - \alpha^*)'\frac{1}{n}\sum_{i=1}^{n}\psi(T_i,X_i)\{W_i - \omega(T_i,X_i)\} + \frac{\ell_K c_K}{\sqrt{n}} \\
&\lesssim_P \frac{K}{n} + \frac{\ell_K c_K}{\sqrt{n}},
\end{aligned}$$

where the first inequality follows from

$$
\begin{aligned}
& Var\left(\frac{1}{n}\sum_{i=1}^{n}(\psi(T_i,X_i)'\alpha^* - r_{0i})\{W_i - \omega(T_i,X_i)\}\right) \\
\leq \quad & \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}} |\psi(T_i,X_i)'\alpha^* - r_{0i}|^2 \, E\left[\frac{1}{n^2}\sum_{i=1}^{n}\|W_i - \omega(T_i,X_i)\|^2\right],
\end{aligned}
$$

combined with Assumption (iii), (v) and the Chebyshev inequality, and the second inequality follows from $\left\|\frac{1}{n}\sum_{i=1}^{n}\psi(T_i,X_i)\{W_i - \omega(T_i,X_i)\}\right\| \lesssim_P \sqrt{K/n}$. Since $K/\sqrt{n} \to 0$ and $\ell_K c_K \to 0$ under our assumptions, we obtain $R_1 = o_p(n^{-1/2})$.

For R_2, writing $\Delta(t,x) = \omega(t,x) - \psi(t,x)'\gamma^*$, we have

$$
\begin{aligned}
R_2 &= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_i\Delta(T_i,X_i) - \int \Delta(T_i,x)f_X(x)dx\right\} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\Delta(T_i,X_j) - \int \Delta(T_i,x)f_X(x)dx\right\} \\
&\quad + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\omega(T_i,X_j) - \int \omega(t,X_j)f_T(t)dt - \int \omega(T_i,x)f_X(x)dx\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_i\Delta(T_i,X_i) - \int \Delta(T_i,x)f_X(x)dx\right\} + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\hat{r}_i - r_{0i})\Delta(T_i,X_i) + \frac{1}{n}\sum_{i=1}^{n}\left\{r_{0i}\Delta(T_i,X_i) - \int \Delta(T_i,x)f_X(x)dx\right\} + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{r}_i - \psi(T_i,X_i)'\alpha^* + \psi(T_i,X_i)'\alpha^* - r_{0i}\right\}\Delta(T_i,X_i) + o_p(n^{-1/2}) \\
&= (\hat{\alpha} - \alpha^*)'\frac{1}{n}\sum_{i=1}^{n}\psi(T_i,X_i)\Delta(T_i,X_i) + O(K^{-(\eta_1+\eta_2)}) + o_p(n^{-1/2}) \\
&= \|\hat{\alpha} - \alpha^*\| \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}\|\psi(t,x)\| \sup_{(t,x)\in\mathcal{T}\times\mathcal{X}}\|\Delta(t,x)\| + O(K^{-(\eta_1+\eta_2)}) + o_p(n^{-1/2}) \\
&= O_p(\xi_K K^{1/2-\eta_2}/\sqrt{n}) + O(K^{-(\eta_1+\eta_2)}) + o_p(n^{-1/2}),
\end{aligned}
$$

where the second equality follows from the law of large numbers for the U-statistics, the fourth equality follows from $E[r_0(T,X)\Delta(T,X)] = \int \Delta(t,x)f_T(t)f_X(x)dtdx$, the Cauchy-Schwarz inequality and (15), the fifth equality follows from Assumption (iii) with $\ell_K c_K =$

$O(K^{-\eta_1})$ and (15), and the last equality follows from $\|\hat{\alpha} - \alpha^*\| \lesssim_P \sqrt{K/n}$ (by (20)) and (15).

Combining these results, we obtain (23) and the conclusion follows.

A.4. **Detailed simulation results.** Tables 8 and 9 (and 10 and 11) summarize the simulation results for estimation of the intercept β_1 and slope β_2 under DGP-L2 (and DGP-N2, respectively). We only show the one case per pair of a basis degree (K_1, K_2) and sample size which minimized the RMSE of $\hat{\beta}_2$, and omit the regularized parameters used for the reported results. The simulation results of Ai *et al.* (2021) are also included in the tables. The meanings of the abbreviations used in "Estimator" column are: "Ratio" for the proposed density ratio-based estimator, "Ratio w reg." for the proposed estimator with regularization, "Ratio JM" for the joint to marginals density ratio-based estimator, "Ratio JM w reg." for the joint to marginals density ratio-based estimator with regularization, and "ALMZ" for the estimator proposed by Ai *et al.* (2021).

Table 8. Simulation results on intercept β_1 under DGP-L2 ($\beta_1^* = 1$)

			$n = 100$			$n = 500$			$n = 1000$	
(K_1, K_2)	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
(2,3)	Ratio	0.010	0.172	0.173	−0.002	0.074	0.074	−0.004	0.054	0.054
	Ratio w reg.	−0.007	0.172	0.172	−0.002	0.074	0.074	−0.005	0.054	0.054
	Ratio JM	−0.025	0.870	0.870	−0.083	2.260	2.262	−0.121	2.255	2.259
	Ratio JM w reg	−0.032	0.417	0.418	−0.053	0.501	0.504	−0.096	0.267	0.283
	ALMZ	−0.005	0.171	0.171	0.000	0.073	0.073	−0.001	0.053	0.053
(2,6)	Ratio	0.011	0.184	0.185	−0.002	0.075	0.075	−0.004	0.054	0.054
	Ratio w reg.	−0.014	0.180	0.181	−0.002	0.075	0.075	−0.005	0.054	0.054
	Ratio JM	−0.093	2.139	2.141	−0.013	0.424	0.424	0.006	0.641	0.641
	Ratio JM w reg	−0.082	0.189	0.206	−0.061	0.096	0.113	−0.097	0.169	0.195
	ALMZ	−0.013	0.171	0.171	−0.010	0.081	0.082	−0.009	0.055	0.056
(2,10)	Ratio	0.009	0.211	0.211	−0.002	0.075	0.075	−0.004	0.054	0.054
	Ratio w reg.	−0.028	0.185	0.187	−0.002	0.075	0.075	−0.005	0.054	0.054
	Ratio JM	−0.092	2.153	2.155	0.017	1.063	1.063	−0.005	2.905	2.905
	Ratio JM w reg	−0.071	0.343	0.350	−0.065	0.229	0.238	−0.032	0.289	0.291
	ALMZ	−0.039	0.181	0.186	−0.034	0.078	0.085	−0.028	0.056	0.062
(3,3)	Ratio	0.008	0.176	0.177	−0.001	0.074	0.074	−0.004	0.054	0.054
	Ratio w reg.	−0.027	0.171	0.173	−0.001	0.074	0.074	−0.004	0.054	0.054
	Ratio JM	−0.013	0.541	0.541	0.150	3.496	3.499	0.035	4.520	4.520
	Ratio JM w reg	−0.079	0.320	0.330	−0.087	0.085	0.122	−0.084	0.132	0.157
	ALMZ	−0.016	0.169	0.170	−0.002	0.075	0.075	0.000	0.057	0.057
(3,6)	Ratio	0.009	0.202	0.202	−0.001	0.075	0.075	−0.004	0.055	0.055
	Ratio w reg.	−0.044	0.174	0.180	−0.002	0.075	0.075	−0.004	0.055	0.055
	Ratio JM	−0.063	1.275	1.277	−0.037	0.708	0.709	−0.004	0.626	0.626
	Ratio JM w reg	−0.049	0.171	0.178	−0.077	0.073	0.106	−0.079	0.051	0.094
	ALMZ	−0.027	0.195	0.197	−0.024	0.083	0.087	−0.026	0.061	0.066
(3,10)	Ratio	0.000	0.276	0.276	−0.003	0.078	0.078	−0.004	0.056	0.056
	Ratio w reg.	−0.053	0.179	0.187	−0.010	0.075	0.076	−0.008	0.055	0.056
	Ratio JM	−0.094	1.732	1.734	0.012	0.600	0.600	−0.004	1.185	1.185
	Ratio JM w reg	−0.060	0.170	0.180	−0.075	0.072	0.104	−0.067	0.054	0.086
	ALMZ	−0.032	0.202	0.205	−0.034	0.080	0.087	−0.030	0.058	0.065
(4,3)	Ratio	0.006	0.186	0.186	−0.001	0.074	0.074	−0.004	0.054	0.054
	Ratio w reg.	−0.026	0.172	0.173	−0.005	0.073	0.074	−0.005	0.054	0.054
	Ratio JM	0.729	22.376	22.387	0.087	2.313	2.314	0.276	8.618	8.622
	Ratio JM w reg	−0.092	0.317	0.330	−0.077	0.107	0.132	−0.085	0.082	0.118
	ALMZ	−0.014	0.179	0.180	−0.006	0.079	0.079	−0.008	0.056	0.056
(4,6)	Ratio	0.008	0.269	0.269	−0.001	0.078	0.078	−0.004	0.056	0.056
	Ratio w reg.	−0.039	0.175	0.180	−0.011	0.074	0.075	−0.009	0.055	0.055
	Ratio JM	−0.050	0.595	0.597	−0.036	0.991	0.991	−0.059	1.498	1.499
	Ratio JM w reg	−0.063	0.168	0.179	−0.073	0.074	0.104	−0.086	0.056	0.103
	ALMZ	−0.036	0.207	0.210	−0.029	0.082	0.087	−0.030	0.059	0.066
(4,10)	Ratio	0.008	0.500	0.500	−0.002	0.085	0.085	−0.004	0.058	0.058
	Ratio w reg.	−0.045	0.180	0.185	−0.014	0.074	0.076	−0.011	0.055	0.056
	Ratio JM	−0.024	0.635	0.636	−0.008	0.452	0.452	0.045	1.847	1.848
	Ratio JM w reg	−0.058	0.170	0.180	−0.046	0.092	0.103	−0.065	0.059	0.088
	ALMZ	−0.032	0.211	0.213	−0.030	0.082	0.088	−0.025	0.059	0.064

Table 9. Simulation results on intercept β_2 under DGP-L2 ($\beta_2^* = 1$)

		$n = 100$			$n = 500$			$n = 1000$		
(K_1, K_2)	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
(2,3)	Ratio	−0.007	0.116	0.116	0.004	0.048	0.048	0.002	0.035	0.035
	Ratio w reg.	0.007	0.114	0.114	0.004	0.048	0.048	0.002	0.035	0.035
	Ratio JM	0.038	0.918	0.919	0.057	1.643	1.644	0.187	4.095	4.099
	Ratio JM w reg	0.036	0.332	0.334	0.050	0.388	0.391	0.079	0.211	0.225
	ALMZ	0.001	0.108	0.108	0.002	0.047	0.047	0.000	0.035	0.035
(2,6)	Ratio	−0.007	0.130	0.130	0.004	0.049	0.049	0.002	0.035	0.035
	Ratio w reg.	0.013	0.123	0.123	0.004	0.049	0.049	0.002	0.035	0.035
	Ratio JM	0.075	1.383	1.385	0.024	0.341	0.342	−0.010	0.783	0.784
	Ratio JM w reg	0.074	0.144	0.162	0.056	0.076	0.095	0.082	0.139	0.162
	ALMZ	0.033	0.118	0.122	0.023	0.052	0.057	0.025	0.037	0.045
(2,10)	Ratio	−0.004	0.159	0.159	0.004	0.050	0.050	0.002	0.035	0.035
	Ratio w reg.	0.024	0.125	0.127	0.004	0.050	0.050	0.002	0.035	0.035
	Ratio JM	0.071	1.326	1.328	−0.026	1.195	1.195	−0.007	2.252	2.252
	Ratio JM w reg	0.071	0.279	0.288	0.062	0.167	0.178	0.021	0.275	0.276
	ALMZ	0.051	0.128	0.138	0.042	0.052	0.067	0.040	0.038	0.055
(3,3)	Ratio	−0.005	0.117	0.117	0.004	0.048	0.048	0.002	0.034	0.034
	Ratio w reg.	0.020	0.113	0.114	0.004	0.048	0.048	0.002	0.034	0.034
	Ratio JM	−0.012	1.416	1.416	−0.127	2.543	2.546	−0.130	6.554	6.556
	Ratio JM w reg	0.070	0.240	0.250	0.073	0.076	0.106	0.068	0.112	0.131
	ALMZ	0.019	0.116	0.117	0.019	0.052	0.056	0.015	0.038	0.041
(3,6)	Ratio	−0.005	0.137	0.137	0.003	0.049	0.049	0.002	0.035	0.035
	Ratio w reg.	0.033	0.117	0.121	0.004	0.049	0.049	0.002	0.035	0.035
	Ratio JM	0.154	3.682	3.685	0.037	0.573	0.574	0.005	0.516	0.516
	Ratio JM w reg	0.050	0.122	0.132	0.070	0.054	0.088	0.067	0.037	0.077
	ALMZ	0.023	0.133	0.135	0.029	0.054	0.062	0.031	0.040	0.050
(3,10)	Ratio	0.004	0.193	0.193	0.004	0.051	0.051	0.002	0.035	0.035
	Ratio w reg.	0.038	0.119	0.124	0.009	0.049	0.050	0.005	0.035	0.035
	Ratio JM	0.078	1.091	1.094	0.001	0.406	0.406	0.010	0.842	0.842
	Ratio JM w reg	0.062	0.119	0.135	0.072	0.049	0.087	0.060	0.049	0.077
	ALMZ	0.037	0.135	0.140	0.040	0.052	0.065	0.037	0.038	0.053
(4,3)	Ratio	−0.004	0.131	0.131	0.004	0.048	0.048	0.002	0.034	0.034
	Ratio w reg.	0.019	0.115	0.117	0.007	0.048	0.048	0.002	0.034	0.034
	Ratio JM	−0.551	15.097	15.107	−0.081	1.866	1.868	−0.191	5.424	5.427
	Ratio JM w reg	0.080	0.247	0.259	0.064	0.108	0.126	0.074	0.136	0.155
	ALMZ	0.024	0.122	0.125	0.021	0.052	0.056	0.023	0.037	0.044
(4,6)	Ratio	−0.002	0.192	0.192	0.003	0.051	0.052	0.002	0.036	0.036
	Ratio w reg.	0.027	0.121	0.124	0.011	0.049	0.050	0.006	0.035	0.036
	Ratio JM	0.041	0.432	0.434	0.023	0.518	0.519	0.043	1.146	1.147
	Ratio JM w reg	0.063	0.118	0.134	0.067	0.064	0.093	0.070	0.066	0.096
	ALMZ	0.028	0.141	0.144	0.035	0.057	0.066	0.039	0.038	0.054
(4,10)	Ratio	0.000	0.337	0.337	0.004	0.058	0.058	0.001	0.039	0.039
	Ratio w reg.	0.030	0.122	0.126	0.013	0.049	0.051	0.007	0.036	0.036
	Ratio JM	0.004	0.754	0.754	0.029	1.317	1.317	0.041	2.085	2.085
	Ratio JM w reg	0.062	0.120	0.135	0.046	0.093	0.104	0.060	0.072	0.093
	ALMZ	0.030	0.141	0.144	0.029	0.055	0.062	0.029	0.040	0.049

TABLE 10. Simulation results on intercept β_1 under DGP-N2 ($\beta_1^* = 1$)

		$n = 100$			$n = 500$			$n = 1000$		
(K_1, K_2)	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
(2,3)	Ratio	−0.043	0.119	0.127	−0.042	0.053	0.068	−0.035	0.038	0.051
	Ratio w reg.	−0.043	0.119	0.127	−0.042	0.053	0.068	−0.035	0.038	0.051
	Ratio JM	−0.075	0.961	0.964	−0.038	0.061	0.072	−0.034	0.038	0.051
	Ratio JM w reg	−0.032	0.136	0.139	−0.038	0.053	0.065	−0.033	0.038	0.050
	ALMZ	−0.043	0.118	0.126	−0.037	0.053	0.065	−0.038	0.036	0.053
(2,6)	Ratio	0.002	0.129	0.129	−0.003	0.056	0.056	0.002	0.040	0.040
	Ratio w reg.	0.004	0.126	0.126	−0.002	0.056	0.056	0.002	0.040	0.040
	Ratio JM	0.062	3.429	3.430	−0.097	2.033	2.035	−0.009	1.056	1.056
	Ratio JM w reg	−0.104	0.126	0.164	−0.116	0.060	0.130	−0.097	0.071	0.120
	ALMZ	−0.010	0.135	0.135	0.004	0.060	0.060	0.004	0.043	0.043
(2,10)	Ratio	0.005	0.138	0.138	−0.003	0.057	0.057	0.003	0.040	0.040
	Ratio w reg.	−0.001	0.129	0.129	−0.003	0.056	0.057	0.003	0.040	0.040
	Ratio JM	−0.073	0.687	0.691	−0.042	0.302	0.305	−0.009	0.958	0.958
	Ratio JM w reg	−0.117	0.116	0.165	−0.102	0.083	0.131	−0.114	0.088	0.144
	ALMZ	−0.030	0.132	0.136	−0.008	0.057	0.058	−0.007	0.043	0.044
(3,3)	Ratio	−0.050	0.122	0.132	−0.043	0.053	0.068	−0.036	0.038	0.052
	Ratio w reg.	−0.050	0.122	0.132	−0.043	0.053	0.068	−0.036	0.038	0.052
	Ratio JM	−0.028	0.224	0.226	−0.040	0.196	0.200	−0.036	0.100	0.106
	Ratio JM w reg	−0.044	0.122	0.130	−0.048	0.053	0.071	−0.040	0.038	0.055
	ALMZ	−0.052	0.125	0.136	−0.041	0.053	0.067	−0.037	0.039	0.053
(3,6)	Ratio	0.002	0.144	0.144	−0.003	0.058	0.058	0.002	0.041	0.041
	Ratio w reg.	−0.001	0.129	0.129	−0.003	0.057	0.057	0.002	0.041	0.041
	Ratio JM	−0.028	0.441	0.442	−0.022	0.725	0.725	0.045	1.570	1.571
	Ratio JM w reg	−0.096	0.117	0.151	−0.092	0.052	0.106	−0.098	0.037	0.104
	ALMZ	−0.030	0.137	0.140	−0.007	0.060	0.060	−0.005	0.045	0.045
(3,10)	Ratio	0.004	0.196	0.196	−0.004	0.059	0.059	0.002	0.041	0.041
	Ratio w reg.	−0.007	0.132	0.132	−0.004	0.058	0.058	0.002	0.041	0.041
	Ratio JM	−0.096	1.506	1.509	−0.031	0.334	0.335	−0.008	0.524	0.524
	Ratio JM w reg	−0.101	0.117	0.154	−0.094	0.053	0.107	−0.086	0.038	0.094
	ALMZ	−0.039	0.141	0.147	−0.017	0.060	0.062	−0.011	0.043	0.045
(4,3)	Ratio	−0.057	0.128	0.140	−0.045	0.053	0.070	−0.036	0.038	0.052
	Ratio w reg.	−0.051	0.123	0.133	−0.045	0.053	0.070	−0.036	0.038	0.052
	Ratio JM	−0.059	1.020	1.022	−0.035	0.202	0.205	−0.015	0.602	0.602
	Ratio JM w reg	−0.042	0.128	0.135	−0.048	0.054	0.072	−0.042	0.039	0.057
	ALMZ	−0.052	0.128	0.139	−0.038	0.055	0.067	−0.039	0.037	0.054
(4,6)	Ratio	0.003	0.173	0.173	−0.004	0.058	0.058	0.002	0.041	0.041
	Ratio w reg.	−0.002	0.131	0.131	−0.003	0.057	0.057	0.003	0.041	0.041
	Ratio JM	−0.015	0.603	0.603	0.038	1.875	1.875	−0.034	0.188	0.191
	Ratio JM w reg	−0.097	0.117	0.152	−0.104	0.052	0.116	−0.085	0.040	0.094
	ALMZ	−0.035	0.142	0.146	−0.012	0.061	0.062	−0.011	0.043	0.045
(4,10)	Ratio	0.005	0.320	0.320	−0.005	0.062	0.062	0.002	0.042	0.042
	Ratio w reg.	−0.008	0.133	0.133	−0.004	0.058	0.058	0.003	0.042	0.042
	Ratio JM	−0.051	0.483	0.486	0.176	6.108	6.110	−0.036	0.345	0.347
	Ratio JM w reg	−0.089	0.119	0.149	−0.105	0.052	0.117	−0.112	0.036	0.118
	ALMZ	−0.048	0.162	0.169	−0.018	0.061	0.063	−0.015	0.043	0.045

Table 11. Simulation results on intercept β_2 under DGP-N2 ($\beta_2^* = 1$)

		$n = 100$			$n = 500$			$n = 1000$		
(K_1, K_2)	Estimator	Bias	SD	RMSE	Bias	SD	RMSE	Bias	SD	RMSE
(2,3)	Ratio	0.161	0.134	0.210	0.180	0.058	0.189	0.183	0.041	0.187
	Ratio w reg.	0.161	0.134	0.210	0.180	0.058	0.189	0.183	0.041	0.187
	Ratio JM	0.071	2.139	2.140	0.162	0.759	0.776	0.187	0.051	0.194
	Ratio JM w reg	0.174	0.189	0.257	0.177	0.060	0.187	0.178	0.040	0.182
	ALMZ	0.172	0.129	0.215	0.185	0.060	0.194	0.184	0.040	0.188
(2,6)	Ratio	-0.003	0.124	0.124	0.001	0.048	0.048	0.001	0.034	0.034
	Ratio w reg.	0.025	0.118	0.121	0.003	0.048	0.048	0.001	0.034	0.034
	Ratio JM	-0.103	4.121	4.122	0.098	2.603	2.605	0.004	0.962	0.962
	Ratio JM w reg	0.064	0.133	0.147	0.070	0.072	0.100	0.063	0.113	0.130
	ALMZ	0.031	0.117	0.121	0.026	0.054	0.059	0.026	0.037	0.045
(2,10)	Ratio	-0.006	0.143	0.143	0.001	0.049	0.049	0.001	0.035	0.035
	Ratio w reg.	0.030	0.122	0.126	0.004	0.048	0.049	0.001	0.035	0.035
	Ratio JM	0.040	1.046	1.046	0.015	0.510	0.510	0.010	0.839	0.839
	Ratio JM w reg	0.065	0.118	0.135	0.061	0.101	0.118	0.075	0.146	0.164
	ALMZ	0.047	0.131	0.139	0.041	0.054	0.067	0.039	0.037	0.053
(3,3)	Ratio	0.153	0.135	0.204	0.176	0.059	0.186	0.181	0.041	0.185
	Ratio w reg.	0.153	0.135	0.204	0.176	0.059	0.186	0.181	0.041	0.185
	Ratio JM	0.189	0.563	0.594	0.174	0.795	0.814	0.168	0.517	0.544
	Ratio JM w reg	0.165	0.142	0.218	0.171	0.056	0.180	0.175	0.050	0.182
	ALMZ	0.163	0.131	0.209	0.180	0.060	0.189	0.181	0.041	0.186
(3,6)	Ratio	-0.001	0.135	0.135	0.001	0.048	0.048	0.001	0.034	0.034
	Ratio w reg.	0.036	0.119	0.125	0.005	0.048	0.048	0.001	0.034	0.034
	Ratio JM	0.051	0.667	0.669	0.103	2.862	2.864	0.046	1.917	1.917
	Ratio JM w reg	0.063	0.115	0.131	0.060	0.050	0.078	0.067	0.034	0.075
	ALMZ	0.027	0.132	0.134	0.031	0.054	0.062	0.031	0.039	0.050
(3,10)	Ratio	-0.006	0.180	0.180	0.001	0.050	0.050	0.001	0.034	0.034
	Ratio w reg.	0.041	0.122	0.128	0.010	0.048	0.049	0.002	0.034	0.034
	Ratio JM	0.043	0.497	0.498	0.003	0.348	0.348	0.054	1.404	1.405
	Ratio JM w reg	0.063	0.115	0.131	0.062	0.048	0.079	0.060	0.036	0.070
	ALMZ	0.044	0.139	0.146	0.036	0.053	0.063	0.036	0.040	0.054
(4,3)	Ratio	0.145	0.142	0.203	0.174	0.060	0.184	0.179	0.041	0.184
	Ratio w reg.	0.150	0.135	0.202	0.174	0.060	0.184	0.179	0.041	0.184
	Ratio JM	0.114	2.488	2.491	0.193	0.877	0.898	0.275	3.054	3.067
	Ratio JM w reg	0.171	0.226	0.283	0.171	0.118	0.207	0.171	0.070	0.185
	ALMZ	0.162	0.133	0.210	0.175	0.059	0.185	0.180	0.040	0.184
(4,6)	Ratio	-0.006	0.174	0.174	0.001	0.051	0.051	0.001	0.035	0.035
	Ratio w reg.	0.032	0.122	0.126	0.008	0.049	0.050	0.004	0.034	0.035
	Ratio JM	0.032	0.547	0.548	-0.094	2.417	2.419	0.019	0.383	0.383
	Ratio JM w reg	0.061	0.115	0.130	0.064	0.049	0.081	0.056	0.046	0.072
	ALMZ	0.038	0.133	0.138	0.034	0.055	0.064	0.033	0.040	0.052
(4,10)	Ratio	-0.022	0.296	0.297	0.002	0.056	0.056	0.001	0.036	0.037
	Ratio w reg.	0.036	0.124	0.130	0.010	0.050	0.051	0.005	0.035	0.036
	Ratio JM	0.059	0.854	0.856	-0.136	4.598	4.600	0.053	0.432	0.435
	Ratio JM w reg	0.055	0.118	0.130	0.066	0.050	0.083	0.070	0.035	0.078
	ALMZ	0.029	0.149	0.151	0.028	0.054	0.061	0.030	0.037	0.047

References

[1] Ai, C., Linton, O., Motegi, K. and Z. Zhang (2021) A unified framework for efficient estimation of general treatment models, *Quantitative Economics*, 12, 779-816.

[2] Abadie, A. and G. W. Imbens (2006) Large sample properties of matching estimators for average treatment effects, *Econometrica*, 74, 235-267.

[3] Angrist, J. D. and J. S. Pischke (2008) *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

[4] Basu, A., Harris, I. R., Hjort, N. L. and M. C. Jones (1998) Robust and efficient estimation by minimising a density power divergence, *Biometrika*, 85, 549-559.

[5] Belloni, A., Chernozhukov, V., Chetverikov, D. and K. Kato (2015) Some new asymptotic theory for least squares series: Pointwise and uniform results, *Journal of Econometrics*, 186, 345-366.

[6] Chen, X., Linton, O. and I. Van Keilegom (2003) Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, 71, 1591-1608.

[7] Florens, J. P., Heckman, J. J., Meghir, C. and E. Vytlacil (2008) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects, *Econometrica*, 76, 1191-1206.

[8] Fong, C., Hazlett, C. and K. Imai (2018) Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements, *Annals of Applied Statistics*, 12, 156-177.

[9] 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, 1528-1542.

[10] Graham, B. S., Pinto, C. C. D. X. and D. Egel (2012) Inverse probability tilting for moment condition models with missing data, *Review of Economic Studies*, 79, 1053-1079.

[11] Hahn, J. (1998) On the role of the propensity score in efficient semiparametric estimation of average treatment effects, *Econometrica*, 66, 315-331.

[12] Heckman, J. J., Ichimura, H. and P. Todd (1998) Matching as an econometric evaluation estimator, *Review of Economic Studies*, 65, 261-294.

[13] Hirano, K. and G. W. Imbens (2004) The propensity score with continuous treatments, in Gelman, A. and X.-L. Meng (eds.), *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, Chapter 7, pp. 73–84, Wiley.

[14] Hirano, K., Imbens, G. W. and G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity score, *Econometrica*, 71, 1161-1189.

[15] 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, 854-866.

[16] Imbens, G. W. (2004) Nonparametric estimation of average treatment effects under exogeneity: A review, *Review of Economics and Statistics*, 86, 4-29.

[17] Kanamori, T., Hido, S. and M. Sugiyama (2009) A least-squares approach to direct importance estimation, *Journal of Machine Learning Research*, 10, 1391-1445.

[18] Kennedy, E. H., Ma, Z., Mchugh, M. D. and D. S. Small (2017) Non-parametric methods for doubly robust estimation of continuous treatment effects, *Journal of the Royal Statistical Society*, B, 79, 1229-1245.

[19] Newey, W. K. (1997) Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, 79, 147-168.

[20] Newey, W. K. and R. J. Smith (2004) Higher order properties of GMM and generalized empirical likelihood estimators, *Econometrica*, 72, 219-255.

[21] Robins, J. M., HernÃ¡n, M. A. and B. Brumback (2000) Marginal structural models and causal inference in epidemiology, *Epidemiology*, 11, 550-560.

[22] Robins, J. M., Rotnitzky, A. and L. P. Zhao (1994) Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, 89, 846-866.

[23] Rosenbaum, P. R. and D. B. Rubin (1983) The central role of the propensity score in observational studies for causal effects, *Biometrika*, 70, 41-55.

[24] Semenova, V. and V. Chernozhukov (2020) Debiased machine learning of conditional average treatment effects and other causal functions, Working paper.

[25] Sugiyama, M., Suzuki, T. and T. Kanamori (2012a) *Density Ratio Estimation in Machine Learning*, Cambridge University Press.

[26] Sugiyama, M., Suzuki, T. and T. Kanamori (2012b) Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation, *Annals of Institute of Statistical Mathematics*, 64, 1009-1044.

[27] Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von BÃŒnau, P. and M. Kawanabe (2008) Direct importance estimation for covariate shift adaptation, *Annals of the Institute of Statistical Mathematics*, 60, 699-746.

[28] 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, 322-336.

[29] Vapnik, V. N. (1998) *Statistical Learning Theory*, Wiley.

[30] Yiu, S. and L. Su (2018) Covariate association eliminating weights: A unified weighting framework for causal effect estimation, *Biometrika*, 105, 709-722.

Graduate School of Economics, Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8601, Japan.

Email address: matsushita.y@r.hit-u.ac.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.

Email address: t.otsu@lse.ac.uk

Faculty of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan.

Email address: kei_t_12@keio.jp