Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data

Qi Li
Department of Economics, Texas A&M University, College Station, TX 77843-4228 (*qi@econmail.tamu.edu*)

Jeffrey S. Racine
Department of Economics, McMaster University, Hamilton, ON Canada L8S 4M4 (*racinej@mcmaster.ca*)

Jeffrey M. Wooldridge
Department of Economics, Michigan State University, East Lansing, MI 48824-1038 (*wooldri1@msu.edu*)

In this article, we consider the nonparametric estimation of average treatment effects when there exist mixed categorical and continuous covariates. One distinguishing feature of the approach presented herein is the use of kernel smoothing for both the continuous and the discrete covariates. This approach, together with the cross-validation method. which we use for selecting the smoothing parameters, has the ability to automatically remove irrelevant covariates. We establish the asymptotic distribution of the proposed average treatment effects estimator with data-driven smoothing parameters. Simulation results show that the proposed method is capable of performing much better than the conventional kernel approach whereby one splits the sample into subsamples corresponding to ''discrete cells.'' An empirical application to a controversial study that examines the efficacy of right heart catheterization on medical outcomes reveals that our proposed nonparametric estimator overturns the controversial findings of Connors et al. (1996), suggesting that their findings may be an artifact of an incorrectly specified parametric model.

KEY WORDS: Asymptotic normality; Average treatment effect; Bootstrap; Discrete covariates; Kernel smoothing.

1. INTRODUCTION

The measurement of average treatment effects (ATEs), initially confined to the assessment of dose-response relationships in medical settings, is today widely used across a range of disciplines. Assessing human-capital losses arising from war (Ichino and Winter-Ebmer 1998) and the effectiveness of job training programs (Lechner 1999) are but two examples of the wide range of potential applications.

Perhaps the most widespread approach toward the measurement of treatment effects involves estimation of a ''propensity score'' (i.e., the conditional probability of receiving treatment). Estimation of the propensity score was originally undertaken with parametric index models such as the Logit or Probit. Recently, there has been a surge in the literature on semiparametric and nonparametric estimation of treatment effects (Hahn 1998; Hirano et al. 2003). The advantage of pursuing a nonparametric approach in this setting is rather obvious, because misspecification of the propensity score may impact significantly upon the magnitude and even the sign of the estimated treatment effect. In many settings, mismeasurement induced by misspecification can be extremely costly—envision for a moment the societal cost of incorrectly concluding that a novel and beneficial cancer treatment, in fact, causes harm.

Datasets used to assess treatment effects frequently contain a preponderance of categorical data (in the typical medical study, it is common to encounter categorical data types exclusively). Though the appeal of robust nonparametric methods is obvious in this setting, conventional nonparametric approaches split the sample into ''cells'' in the presence of categorical covariates, resulting in a loss of efficiency (we shall refer to this conventional nonparametric approach as a ''frequency-based'' method for what follows). It is not uncommon to encounter situations in which the number of cells is comparable to or even exceeds the sample size. In such cases, the frequency-based kernel approach becomes infeasible. In addition to these issues, another undesirable side-effect of frequency-based methods is a loss of power for *tests* of whether a treatment effect differs from that of no effect. However, strong theoretical and practical reasons support the application of kernel smoothing methods to a mix of continuous and discrete (i.e., nominal and ordinal categorical) data types, where both the continuous *and* the discrete data are smoothed in a particular fashion. These recently developed cross-validated kernel smoothing methods not only admit both categorical and continuous covariates, but can also automatically detect and remove irrelevant covariates (asymptotically). Such approaches lead to feasible and efficient nonparametric estimation when confronted with a mix of categorical and continuous data, and can be used to construct tests that do not suffer from power loss exhibited by the conventional kernel approach that arises from sample splitting. Note that Hahn's (1998) and Hirano, Imbens and Ridder's (2003) ATE estimators are based on nonparametric series estimation

Journal of Business & Economic Statistics
April 2009, Vol. 27, No. 2
DOI 10.1198/jbes.2009.0015

methods. All series-based nonparametric estimators use the indicator function/frequency method to handle the presence of discrete covariates. In finite-sample applications, this may become infeasible if the number of ''discrete cells'' is not substantially smaller than the sample size.

In this article, we propose a kernel-based nonparametric method for measuring and testing for the presence of treatment effects that is ideally suited to datasets containing a mix of categorical (nominal and ordinal) and continuous data types. One distinguishing feature of the proposed approach is the use of kernel smoothing for both the continuous and the discrete covariates. We elect to use a least-squares conditional cross-validation method to select smoothing parameters for both the categorical and continuous variables; this method was proposed by Hall et al. (2006), who demonstrate that cross-validation produces asymptotically optimal smoothing for relevant components while eliminating irrelevant components by oversmoothing. Indeed, for the problem of nonparametric estimation with mixed categorical and continuous data, cross-validation comes into its own as a method with no obvious peers.

In addition to deriving the asymptotic distribution of our proposed kernel-based ATE estimator, we also propose using a bootstrap method to better approximate the finite-sample distribution of the ATE estimator, and we prove that the bootstrap method works.

The rest of the article proceeds as follows. In Section 2, we outline the nonparametric model and derive the distribution of the resultant average treatment effect. In Section 3, we undertake some simulation experiments designed to demonstrate that the proposed method is capable of outperforming existing kernel approaches that require splitting the sample into cells. We also report simulation results that compare our proposed kernel estimator to a nonparametric series-based estimator. An empirical application presented in Section 4 involving a study that examines the efficacy of right heart catheterization on medical outcomes reveals that our approach negates the controversial findings of Connors et al. (1996), suggesting that their result may be an artifact of an incorrectly specified parametric model. Main proofs appear in the appendices.

2. THE MODEL

For what follows, we use a dummy variable, $t_i \in \{0, 1\}$, to indicate whether an individual has received treatment. We let $t_i = 1$ for the treated, and 0 for the untreated. Letting $y_i(t_i)$ denote the outcome, then, for $i = 1, \ldots, n$, we write

$$y_i = t_i y_i(1) + (1 - t_i) y_i(0).$$

Interest lies in the average treatment effect defined as follows:

$$\tau = E[y_i(1) - y_i(0)].$$

Let x_i denote a vector of pretreatment variables. One issue that instantly surfaces in this setting is that, for each individual i, we either observe $y_i(0)$ or $y_i(1)$, but not both. Therefore, in the absence of additional assumptions, the treatment effect is not consistently estimable. One popular assumption is the ''unconfoundedness condition'' (Rosenbaum and Rubin 1983).

Assumption (A1) (Unconfoundedness):

Conditional on x_i, the treatment indicator t_i is independent of the potential outcome.

Define the conditional treatment effect by $\tau(x) = E[y_i(1) - y_i(0)|X = x]$. Under Assumption (A1) one can easily show that (e.g., Theorem 4 of Rosenbaum and Rubin (1983))

$$\tau(x) = E[y_i|t_i = 1, x_i = x] - E[y_i|t_i = 0, x_i = x]. \quad (1)$$

The two terms on the right-hand side of (1) can be estimated consistently by any nonparametric estimation technique. It is not clear to us how to extend the property of kernel smoothing noted in Section 1 to other nonparametric estimation methods such as series methods (i.e., the ability to automatically remove irrelevant covariates). Therefore, we restrict our attention to nonparametric kernel methods in this section. Under Assumption (A1), the average treatment effect can be obtained *via* simple averaging over $\tau(x)$ and is given by

$$\tau = E[\tau(x_i)]. \quad (2)$$

Letting $E(y_i|x_i, t_i)$ be denoted by $g(x_i, t_i)$, we then have

$$y_i = g(x_i, t_i) + u_i, \quad (3)$$

with $E(u_i|x_i, t_i) = 0$.

Defining $g_0(x_i) = g(x_i, t_i = 0)$ and $g_1(x_i) = g(x_i, t_i = 1)$, we can rewrite (3) as

$$\begin{aligned} y_i &= g_0(x_i) + [g_1(x_i) - g_0(x_i)]t_i + u_i \\ &= g_0(x_i) + \tau(x_i)t_i + u_i, \end{aligned} \quad (4)$$

where $\tau(x_i) = g_1(x_i) - g_0(x_i)$.

From (4), it is easy to show that $\tau(x_i) = \text{cov}(y_i, t_i|x_i)/\text{var}(t_i|x_i)$. Letting $\mu(x_i) = \Pr(t_i = 1|x_i) \equiv E(t_i|x_i)$ (because t_i equals 0 or 1), we may write

$$\tau = E[\tau(x_i)] = E\left\{\frac{(t_i - \mu_i(x_i))y_i}{\text{var}(t_i|x_i)}\right\}. \quad (5)$$

We now turn to the discussion of the nonparametric estimation of τ based on (5) in the presence of a mix of continuous and categorical covariates, some of which, in fact, may be irrelevant.

2.1. Nonparametric Estimation of the Propensity Score

We use x_i^c and x_i^d to denote the continuous and discrete components of x_i, with $x_i^c \in \mathbb{R}^q$ and x_i^d being of dimension r. Let $w(\cdot)$ denote a univariate kernel function for the continuous variables, and define the product kernel function for the continuous variables by

$$W_h(x_i^c, x_j^c) = \prod_{s=1}^{q} h_s^{-1} w\left(\frac{x_{is}^c - x_{js}^c}{h_s}\right), \quad (6)$$

where x_{is}^c is the sth component of x_i^c and h_s is the corresponding smoothing parameter ($s = 1, \ldots, q$).

We assume that some of the discrete variables have a natural ordering, examples of which would include preference orderings (like, indifference, dislike), health conditions (excellent, good, poor), and so forth. Let $x_i^{d,o}$ denote an r_o vector ($0 \le r_o \le r$) of discrete covariates that have a natural ordering, and let $x_i^{d,u}$ denote the remaining $r_u = r - r_o$ discrete covariates that do not

have a natural ordering (e.g., race or industry of occupation). We use x_{it}^d to denote the tth component of $x_i^d (t = 1, \ldots, r)$.

As in Hall et al. (2006), for an ordered variable, we use the following kernel:

$$l_o(x_{is}^d, x_{js}^d, \lambda_s) = \begin{cases} 1, & \text{if } x_{is}^d = x_{js}^d, \\ \lambda_s^{|x_{is}^d - x_{js}^d|}, & \text{if } x_{is}^d \neq x_{js}^d. \end{cases} \tag{7}$$

Note that when $\lambda_s = 0$ ($\lambda_s \in [0, 1]$), $l_o(x_{is}^d, x_{js}^d, \lambda_s = 0)$ becomes an indicator function, and when $\lambda_s = 1, l_o(x_{is}^d, x_{js}^d, \lambda_s = 1) = 1$ becomes a uniform weight function.

For an unordered variable, we use a variation on Aitchison and Aitken's (1976) kernel function defined by

$$l_u(x_{is}^d, x_{js}^d) = \begin{cases} 1, & \text{if } x_{is}^d = x_{js}^d, \\ \lambda_s, & \text{otherwise.} \end{cases} \tag{8}$$

Again, note that $\lambda_s = 0$ leads to an indicator function and $\lambda_s = 1$ leads to a uniform weight function.

Let $1(A)$ denote an indicator function that assumes the value 1 if A occurs and 0 otherwise. Combining (7) and (8), we obtain the product kernel function for the categorical variables (ordered and unordered), which we denote by

$$L(x_i^d, x_j^d, \lambda) = \left[\prod_{s \in S_o} \lambda_s^{|x_{is}^d - x_{js}^d|}\right] \left[\prod_{s \in S_u}^{r} \lambda_s^{1\left(x_{is}^d \neq x_{js}^d\right)}\right], \tag{9}$$

where S_o and S_u denote the index sets for ordered and unordered components of x^d.

When there exists a mix of categorical and continuous variables, we obtain an appropriate kernel function by simply taking the product of (9) and (6). The use of this ''generalized product kernel'' along with a particular data-driven method of smoothing parameter selection is supported by strong theoretical and practical reasons, which we now briefly discuss.

We note that there does not exist a plug-in or even an ad-hoc formula for selecting the smoothing parameters associated with the categorical variables in this setting (i.e., λ_s, $S = 1, \ldots, r$). Hence, we recommend using least-squares cross-validation for selecting λ_s ($S = 1, \ldots, r$). Our recommendation is based not only on the mean-square-error optimality of least-squares cross-validation, but also due to its automatic ability to (asymptotically) remove irrelevant discrete covariates (Hall et al. 2004, 2006). This property bears highlighting as we have observed that irrelevant variables tend to occur surprisingly often in practice. Thus, cross-validation provides an efficient way of guarding against overspecification of nonparametric models, and thereby mitigates the ''curse of dimensionality'' often associated with kernel methods.

Since $\mu(x_i) = Pr(t_i = 1|x_i) = E(t_i|x_i)$, we can use either a conditional probability estimator or a conditional mean estimator to estimate $\mu(x_i)$. We will use the latter in this article. We let $\hat{t}(x_i)$ be the nonparametric estimator of $\mu_i \equiv \mu(x_i)$ defined by

$$\hat{t}(x_i) = \frac{\sum_{j=1}^n t_j K_{\gamma,ij}}{\sum_{j=1}^n K_{\gamma,ij}}, \tag{10}$$

where $K_{\gamma,ij} = W_h(x_i^c, x_j^c) L(x_i^d, x_j^d, \lambda) (\gamma = (h, \lambda))$ with $W_h(\cdot)$ and $L(\cdot)$ being the kernel functions defined in (6) and (9), respectively. By noting that var $(t_i|x_i) = \mu_i(1 - \mu_i)$, one can estimate the average treatment effect by

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \frac{(t_i - \hat{t}(x_i)) y_i M_{ni}}{\hat{t}(x_i)(1 - \hat{t}(x_i))} \equiv \frac{1}{n} \sum_{i=1}^n \left[\frac{t_i y_i}{\hat{t}(x_i)} - \frac{(1 - t_i) y_i}{1 - \hat{t}(x_i)}\right] M_{ni}, \tag{11}$$

where $M_{ni} = M_n(x_i)$ is a trimming set that trims out observations near the boundary.

Equation (11) is a nonparametric version of the Horvitz and Thomson (1952) estimator. Hirano et al. (2003) studied the Horvitz-Thomson estimator with the propensity score estimated by series methods. As we mentioned earlier, it is not clear how to use nonparametric series methods to automatically remove irrelevant discrete covariates. Therefore, in this paper, we will focus on kernel-based estimation methods, which has the advantage of automatically removing irrelevant covariates (be they discrete or continuous).

We will choose the smoothing parameters based on a least-squares cross-validation method. The leave-one-out kernel estimator of $E(t_i|x_i) = \mu(x_i)$ is given by

$$\hat{t}_{-i}(x_i) = \frac{\sum_{j \neq i}^n t_j K_{\gamma,ij}}{\sum_{j \neq i}^n K_{\gamma,ij}}, \tag{12}$$

where, as defined previously, $K_{\gamma,ij} = W_h(x_i^c, x_j^c) L(x_i^d, x_j^d, \lambda)$. We choose $(h, \lambda) = (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ by minimizing the following least-squares cross-validation function:

$$CV(h, \lambda) = \frac{1}{n} \sum_{i=1}^n [t_i - \hat{t}_{-i}(x_i)]^2 S(x_i), \tag{13}$$

where $S(\cdot)$ is a weight function that trims out observations near the boundary of the support of x_i (thereby avoiding excessive boundary bias). For example, if a continuous covariate, say, x_{is}^c, takes values in [0,1], then $S(\cdot)$ will trim out data near the boundary and only use the data for $x_{is}^c \in [\delta, 1 - \delta]$ for some small $\delta \in (0, 1/2)$.

Hall et al. (2004, 2006) have shown that, when $x_s^d (x_s^c)$ is an irrelevant covariate, the cross-validation selected smoothing parameter λ_s (h_s) will converge to 1 (∞) in probability; hence, irrelevant covariates (discrete or continuous) will be automatically smoothed out.

We let $(\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ denote the cross-validation choices of $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ that minimize (13). Without loss of generality, we assume that the first q_1 components of x_i^c, $1 \leq q_1 \leq q$, and the first r_1 components of x_i^d, $0 \leq r_1 \leq r$, are the relevant covariates, while the remaining covariates are irrelevant; Condition (C1) provides a rigorous definition. Then, similar to the proofs of Hall et al. (2006), one can show the following.

Lemma 2.1. Under conditions (C1)–(C5) given in Appendix A, we have

$$\hat{h}_s = a_s^0 n^{-1/(\nu + q_1 + 2)} + o_p(n^{-1/(\nu + q_1 + 2)}), \text{ for } s = 1, \ldots, q_1,$$

$$\hat{\lambda}_s = b_s^0 n^{-2/(\nu + q_1 + 2)} + o_p(n^{-2/(\nu + q_1 + 2)}), \text{ for } s = 1, \ldots, r_1,$$

$$P[\hat{h}_s > C] \to 1 \text{ for any } C > 0, \text{for } s = q_1 + 1, \ldots, q$$

$$\hat{\lambda}_s \to 1 \text{ in probability for } s = r_1 + 1, \ldots, r,$$

where the a_s^o's are finite positive constants, and the b_s^o's are nonnegative finite constants.

The proof of Lemma 2.1 follows in a manner similar to the proof of Theorem 3.1 of Hall et al. (2006). A sketch of the proof of Lemma 2.1 is given in Appendix A.

Lemma 2.1 shows that the cross-validation method can (asymptotically) automatically remove irrelevant covariates. If $q_1 \leq 3$, we will use the cross-validation-selected smoothing parameters to estimate $\hat{\tau}$. However, if $q_1 \geq 4$, we suggest using $\bar{h}_s = \hat{h}_s n^{1/(2\nu+q_1)} n^{-1/(q_1+\nu+2)}$ and $\bar{\lambda}_s = \hat{\lambda}_s n^{2/(2\nu+q_1)} n^{-\nu/(q_1+\nu+2)}$ for computing $\hat{\tau}$. The reason for doing so is that, when $q_1 \geq 4$, regularity conditions that ensure the $\sqrt{n}$-normality result for $\hat{\tau}$ require us to undersmooth the data; the discussion following Assumption (A3) provides more detail.

The empirical applications and simulation results presented in Hall et al. (2004, 2006) reveal that nonparametric estimation based on cross-validated smoothing parameter selection performs much better than a frequency-based estimator (which, as noted, corresponds to $\lambda_s = 0$ for all $s = 1, \ldots, r$).

Having obtained the $\bar{h}_s$'s and $\bar{\lambda}_s$'s based on cross-validation, we estimate τ using expression (11) with $\hat{t}(x_i)$ computed using the $\bar{h}_s$'s and $\bar{\lambda}_s$'s. To avoid introducing superfluous notation, we continue to use $\hat{\tau}$ to denote the resulting estimator of τ.

2.2. The Asymptotic Distribution of $\hat{\tau}$

We shall require some regularity assumptions to derive the asymptotic distribution of $\hat{\tau}$. Following Robinson (1988), we use $\mathcal{G}_\nu^\alpha$ (ν is a positive integer) to denote the smooth class of functions such that, if $g \in \mathcal{G}_\nu^\alpha$, then g is ν-times differentiable, and g and its partial derivatives (up to order ν) all satisfy some Lipschitz conditions and are all bounded by functions with finite αth moments.

Assumption (A2):

(i) (y_i, x_i, t_i) are independently and identically distributed as (y_i, x_i, t_i). (ii) x_i^d takes finitely many different values; for each x^d, the support of $f(x^c, x^d)$ is a compact convex set in x^c (the support of $f(.,.)$ is defined as $\{(x^c, x^d) \in R^{q+r} \mid f(x^c, x^d) > 0\}$), $\mu(x^c, x^d) \in \mathcal{G}_\nu^4$, $f(x^c, x^d) \in \mathcal{G}_{\nu-1}^4$, where $\nu \geq 2$ and $\nu > q - 2$ is a positive integer. (iii) $\inf_{x \in S} f(x) \geq \eta$ for some $\eta > 0$, where S is the support of x_i. (iv) $\sigma^2(x, t) = \mathrm{var}(u_i | x_i = x, t_i = t)$ is bounded below by a positive constant on the support of (x_i, t_i). (v) The trimming function, $M_n(x)$, converges to an indicator function (as $n \to \infty$) $1(x \in S)$, where $1(\cdot)$ is the usual indicator function and S is the support of $f(x)$.

Assumption (A3):

(i) $w(\cdot)$ is a compactly supported νth order kernel; it is bounded, symmetric, and differentiable up to order ν. (ii) As $n \to \infty$, $n \sum_{s=1}^{q_1} h_s^{2\nu+4} \to 0$, and $n(h_1 \ldots h_{q_1})^2 \to \infty$.

Assumptions (A2) (i)-(iv) are standard smoothness and moment conditions. Assumption (A2) (v) implies that, asymptotically, we only trim a negligible amount of data (near the boundary) so that $\hat{\tau}$ is asymptotically efficient (Theorem 2.1). A trimming set is used in (11) for theoretical reasons. Given that the support of x is a compact convex set (in x^c), without loss of generality, one can assume that $x^c \in [-1, 1]^q$. Then, one can define a set $A_{\delta_n} = \prod_{s=1}^{q} [-1 + \delta_s, 1 - \delta_s]$, where $\delta_s = \delta_{sn} < 1$ converges to 0 as $n \to \infty$. To avoid boundary bias, one can choose $\delta_s = O(h_s^\alpha)$ for some $0 < \alpha < 1$, and define $M_n(x_i) = 1(x_i \in A_{\delta_n})$. In this way, the boundary effects disappear asymptotically. In practice, boundary trimming does not appear to be necessary. In both the simulations and the empirical application reported in Sections 3 and 4, we do not resort to trimming. In the presence of outliers, however, one might wish to consider trimming. (A3) requires that $w(\cdot)$ has compact support; this assumption is used in Hall et al. (2006) and can be relaxed, however, at the cost of a much lengthier proof. In particular, the Gaussian kernel can be used in practice.

Note that with a νth order kernel, $\hat{h}_s \sim n^{-1/(2\nu+q_1)}$ and $\hat{\lambda}_s \sim n^{-2/(2\nu+q_1)}$. However, our Assumption (A3) (ii) rules out optimal smoothing when $q_1 \geq 4$. Note that, by Assumption (A2) (ii), we know that we can choose $h_s \sim n^{-1/\alpha}$ $(s = 1, \ldots, q_1)$ for any α such that $2q_1 < \alpha < 2\nu + 4$. Here, we choose $\alpha = q_1 + \nu + 2$, the mean value of $2q_1$, and $2\nu + 4$, in $\bar{h}_s \sim n^{-1/\alpha}$ and $\bar{\lambda} \sim n^{-2/\alpha}$. This is why we use $\bar{h}_s$ and $\bar{\lambda}_s$ to replace $\hat{h}_s$ and $\hat{\lambda}_s$ when $q_1 \geq 4$. To see this clearly, let us assume that $h_s = h$ for all $s = 1, \ldots, q_1$. In this case, Assumption (A3) (ii) requires that $\nu + 2 > q_1$. If one uses a second-order kernel ($\nu = 2$), this will imply that $q_1 < 4$ or $q_1 \leq 3$, because q_1 is a positive integer. Thus, a second-order kernel satisfies Assumption (A3) if $q_1 \leq 3$. When $q_1 \geq 4$, Assumption (A3) requires the use of a higher order kernel function.

Remark 2.1. If $1 \leq q_1 \leq 3$ and one uses a second-order kernel ($\nu = 2$), then Assumption (A3) allows for optimal smoothing. To see this, note that, when $\nu = 2$, optimal smoothing requires that $h_s = O(n^{-1/(4+q_1)})$. Assumption (A3) (ii) becomes (assuming $h_s = h$) $nh^8 \to 0$ *and* $nh^{2q_1} \to \infty$; optimal smoothing, *i.e.*, $h \sim n^{-1/(4+q_1)}$, satisfies these conditions for $q_1 = 1, 2, 3$, where $A \sim B$ means that A and B have the same order of magnitude.

The next theorem provides the asymptotic distribution of $\hat{\tau}$.

Theorem 2.1. Under Assumptions (A1)–(A3), we have

$$\sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda}) \to N(0, V_1 + V_2) \text{ in distribution,}$$

where $B_{h,\lambda} = \sum_{s=1}^{q_1} C_{1s}(x) \hat{h}_s^\nu - \sum_{s=1}^{r_1} C_{2s}(x) \hat{\lambda}_s$, $C_{1s}(x)$ and $C_{2s}(x)$ are defined in Lemma B.2 of Appendix B, $V_1 = \mathrm{var}(\tau(x_i))$, $V_2 = E\{\sigma^2(x_i, t_i)(t_i - \mu_i)^2 / [\mu_i^2 (1 - \mu_i)^2]\}$, and $\sigma^2(x_i, t_i) = E(u_i^2 | x_i, t_i)$.

The proof of Theorem 2.1 is given in Appendix A.

Let $f(x, t)$ and $f_x(x)$ denote the joint and marginal densities of (x_i, t_i) and x_i, respectively, and let $p(t_i|x_i)$ be the conditional probability of t_i given x_i. Letting $\int dx = \sum_{x^d} \int dx^c$, $\mu_x = \mu(x)$, using $f(x_i, t_i) = p(t_i|x_i) f_x(x_i)$, and noting that $p(t_i = 1|x_i) = \mu_i$ and $p(t_i = 0|x_i) = 1 - \mu_i$, we have

$$\begin{aligned} V_2 &= E\{\sigma^2(x_i)(t_i - \mu_i)^2 / [\mu_i^2 (1 - \mu_i)^2]\} \\ &= \sum_{t=1,0} \int f_x(x) p(t|x) \{\sigma^2(x, t)(t_i - \mu_x)^2 / [\mu_x^2 (1 - \mu_x)^2]\} dx \\ &= \int \frac{f_x(x) \mu_x \sigma^2(x, 1)(1 - \mu_x)^2}{\mu_x^2 (1 - \mu_x)^2} dx \\ &\quad + \int \frac{f_x(x)(1 - \mu_x) \sigma^2(x, 0) \mu_x^2}{\mu_x^2 (1 - \mu_x)^2} dx \\ &= E\left\{ \frac{\sigma^2(x_i, 1)}{\mu_i} + \frac{\sigma^2(x_i, 0)}{1 - \mu_i} \right\}. \end{aligned} \tag{14}$$

Equation (14) matches the expression given in Hahn (1998). Thus, $V_1 + V_2$ coincides with the semiparametric efficiency bound for this model. Therefore, Theorem 2.1 shows that our kernel-based estimator of $\hat{\tau}$ is semiparametrically efficient.

Hirano et al. (2003) consider the problem of estimating average treatment effects using series estimation methods. They observe that, if one uses the true var $(t_i|x_i) = \mu_i(1-\mu_i)$ to replace the estimated variance $\hat{t}_i(1-\hat{t}_i)$ in (the denominator of) $\hat{\tau}$, then it results in a *less* efficient estimator of τ. The same result holds true for our kernel-based estimator, as the next lemma shows.

Lemma 2.2. If one replaces the denominator $\hat{t}_i(1-\hat{t}_i)$ in $\hat{\tau}$ by $\mu_i(1-\mu_i)$, and lets $\tilde{\tau}$ denote the resulting estimator of τ, i.e.,

$$\tilde{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{(t_i - \hat{t}_i) y_i M_{ni}}{\mu_i(1-\mu_i)}, \tag{15}$$

then

$$\sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}) \to N(0, V_1 + V_2 + V_3) \text{ in distribution,}$$

where V_1 and V_2 are the same as those given in Theorem 2.1, while V_3 is given by

$$V_3 = E\left\{ \left[\frac{(t_i - \mu_i)^2}{\mu_i(1-\mu_i)} - 1 \right]^2 \tau_i^2 \right\}.$$

The proof of Lemma 2.2 is given in Appendix A.

We observe how using the true var $(t_i|x_i)$ yields a less efficient estimator than $\hat{\tau}$, which uses the estimated var $(t_i|x_i)$. The reason for this result is that one can express $\sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda})$ as

$$\sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda}) = \sqrt{n}(\hat{\tau} - \tilde{\tau}) + \sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}).$$

In Appendix A, we show that $\sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}) = Z_{n1} + Z_{n2} + Z_{n3} + o_p(1) \to N(0, V_1 + V_2 + V_3)$ in distribution, where the Z_{nl}'s ($l = 1, 2, 3$) are *three* asymptotically uncorrelated terms, having asymptotic $N(0, V_l)$ distributions, respectively ($l = 1, 2, 3$, with definitions appearing in Appendix A). This yields Lemma 2.2. In Appendix A, we also show that $\sqrt{n}(\hat{\tau} - \tilde{\tau}) = -Z_{n3} + o_p(1)$. Hence,

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda}) &= \sqrt{n}(\hat{\tau} - \tilde{\tau}) + \sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}) \\
&= \{-Z_{n3} + o_p(1)\} \\
&\quad + \{Z_{n1} + Z_{n2} + Z_{n3} + o_p(1)\} \\
&= Z_{n1} + Z_{n2} + o_p(1) \\
&\quad \to N(0, V_1 + V_2) \text{ in distribution,}
\end{aligned}$$

resulting in Theorem 2.1. That is, since the leading term in $\sqrt{n}(\hat{\tau} - \tilde{\tau})$ cancels one of the leading terms in $\sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda})$, this gives rise to the result whereby using an estimated variance var$(\hat{t}_i|x_i)$ is more efficient than using the true variance var $(t_i|x_i)$ when estimating τ. If one uses the true propensity score μ_i in both the numerator and denominator of $\hat{\tau}$, then one gets $\tilde{\tau}$, which is more efficient than $\hat{\tau}$, because $\sqrt{n}(\tilde{\tau} - \tau)$ is asymptotically normal with mean zero and asymptotic variance V_1. Of course, $\tilde{\tau}$ is not a feasible estimator. Thus, among the class of feasible (''regular'') estimators, $\hat{\tau}$ is asymptotically efficient.

To construct a consistent estimator for the asymptotic variance $V_1 + V_2$, we need to obtain, among other things, a consistent estimator of the error u_i. The estimator $\hat{\tau}$ proposed previously is based on an estimated propensity score, so it does not estimate the regression mean function directly. In this subsection, we consider an alternative estimator for τ, which is based on the direct estimation of $E(y_i|x_i, t_i)$, which of course also leads to a direct estimator of u_i.

Note that (4) can also be viewed as a functional coefficient model (i.e., a smooth coefficient model), which has been considered by a number of authors including Chen and Tsay (1993), Cai, Fan, and Yao (2000), Cai, Fan, and Li (2000), and Li et al. (2002). Thus, an alternative estimator of $\tau(x_i)$ can be obtained by a local regression of y_i on $(1, t_i)$ using kernel weights. In this way, we obtain a nonparametric estimator of $(g_0(x_i), \tau(x_i))'$ given by

$$\begin{aligned}
\begin{pmatrix} \hat{g}_0(x_i) \\ \hat{\tau}_n(x_i) \end{pmatrix} &= \left[n^{-1} \sum_{j \neq i}^{n} \begin{pmatrix} 1 \\ t_j \end{pmatrix} (1, t_j) W_{h,ij} L_{\hat{\lambda},ij} \right]^{-1} \\
&\quad \times \left[n^{-1} \sum_{j \neq i}^{n} \begin{pmatrix} 1 \\ t_j \end{pmatrix} y_j W_{h,ij} L_{\hat{\lambda},ij} \right],
\end{aligned} \tag{16}$$

where $W_{h,ij} = W_h\left(x_j^c, x_i^c\right)$ and $L_{\hat{\lambda},ij} = L(x_i^d, x_j^d, \hat{\lambda})$. Equation (16) provides consistent estimators of $g_0(x_i)$ and $\tau(x_i)$. For example, the resulting estimators of $g_0(x_i)$ and $\tau(x_i)$ are given by

$$\hat{g}_0(x_i) = \frac{\hat{E}(t_i|x_i)[\hat{E}(y_i|x_i) - \hat{E}(y_i t_i|x_i)]}{\hat{t}(x_i)(1-\hat{t}(x_i))} \tag{17}$$

and

$$\hat{\tau}_n(x_i) = \frac{\hat{E}(y_i t_i|x_i) - \hat{E}(y_i|x_i)\hat{E}(t_i|x_i)}{\hat{t}(x_i)(1-\hat{t}(x_i))}, \tag{18}$$

where $\hat{E}(y_i t_i|x_i) = n^{-1}\sum_{j=1}^{n} t_j y_j K_{\gamma,ij}/\hat{f}(x_i)$, $\hat{E}(y_i|x_i) = n^{-1}\sum_{j=1}^{n} y_j K_{h,ij}/\hat{f}(x_i)$, $\hat{t}(x_i) = n^{-1}\sum_{j=1}^{n} t_j K_{\gamma,ij}/\hat{f}(x_i)$, and $\hat{f}(x_i) = n^{-1}\sum_{j=1}^{n} K_{\gamma,ij}$. Similarly, one can obtain consistent estimators for V_1, V_2 and $B_{h,\lambda}$, which we denote by $\hat{V}_1$, $\hat{V}_2$, and $\hat{B}_{h,\lambda}$; the explicit definitions can be found in Appendix A. The following result follows directly from Theorem 2.1:

$$\hat{T}_n \overset{\text{def}}{=} \sqrt{n}(\hat{\tau} - \tau - \hat{B}_{h,\lambda})/\sqrt{\hat{V}_1 + \hat{V}_2} \to N(0,1) \text{ in distribution.} \tag{19}$$

2.3 A Bootstrap Test for the Presence of Treatment Effects

The estimation of treatment effects goes hand in hand with the central issue of testing whether the estimated effect differs from that of no effect, and the result of Theorem 2.1 can be used to test the null hypothesis of no ''effect'' of a treatment (i.e., the null hypothesis is H_0: $\tau = 0$). It is well known that valid bootstrap procedures often provide more accurate finite-sample estimates for confidence intervals. Subsequently, we present a bootstrap procedure and we will use nonparametric bootstrap confidence intervals to test the null hypothesis of no treatment effect.

Let $z_i \equiv \{y_i, x_i, t_i\}_{i=1}^{n}$, i.e., the vector of realizations on the outcome, treatment, and conditioning information, respectively.

We wish to construct the sampling distribution of $\hat{\tau}$, and do so with the following resampling procedure.

(1) Letting $z_j = (y_j, x_j, t_j)$, randomly select from $\{z_j\}_{j=1}^n$ with replacement, and call $\{z_i^*\}_{i=1}^n$ the bootstrap sample.
(2) Use the bootstrap sample to compute the bootstrap statistic $\hat{\tau}^*$ using the same cross-validated smoothing parameters as were used for $\hat{\tau}$.
(3) Repeat steps 1 and 2 a large number of times, say, B times. The empirical distribution function of $\{\hat{\tau}_j^*\}_{j=1}^B$ will be used to approximate the finite-sample distribution of $\hat{\tau}$.

We point out that the bootstrap counterpart quantities (i.e., $\hat{T}_{n,(j)}^*$) use the same smoothing parameters as the original statistic (i.e., $\hat{T}_n$) (they do not require recross-validation). The following theorem shows that the bootstrap method works.

Theorem 2.2. Under the same conditions as in Theorem 2.1, define $T_n^* = \sqrt{n}(\hat{\tau}^* - \hat{\tau} - \hat{B}_{h,\lambda}^*)/\sqrt{\hat{V}_1^* + \hat{V}_2^*}$. Then,

$$\sup_{z\in\mathbb{R}} |P\big(T_n^* \le z | \{y_i, x_i, t_i\}_{i=1}^n\big) - \Phi(z)| = o_p(1) \tag{20}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

Since $\Phi(z)$ is a continuous function, by Theorem 1 of Tucker (1967), we know that (20) follows from

$$\begin{aligned}&\big|P\big(T_n^* \le z | \{x_i, t_i, y_i\}_{i=1}^n\big) - \Phi(z)\big| \\ &\quad = o_p(1) \text{ for any fixed value of } z \in \mathbb{R}.\end{aligned} \tag{21}$$

The proof of (21) (i.e., the proof of Theorem 2.2) is given in Appendix C.

3. SIMULATIONS

In this section, we report on simulations designed to examine the finite-sample performance of the proposed methods. We highlight performance in mixed data settings, a feature that existing frequency-based methods do not handle well, and consider three estimators of the average treatment effect (''$\hat{\tau}$'') differing in the respective estimator of the propensity score (''$\hat{t}_i(x_i)$''); (1) the proposed nonparametric propensity score estimator that smooths both continuous and discrete variables in a particular manner (''NP (CVLS)''), (2) a nonparametric frequency-based propensity score estimator (''NP (FREQ)''), and (3) an alternative nonparametric frequency-based propensity score estimator, namely, a B-spline approach (''Spline''). That is, we are comparing the proposed estimator that smooths both the continuous and discrete data in a particular way with two *nonsmooth* conventional approaches that require sample splitting to be used in the presence of categorical covariates.

To provide a sound basis for comparison, leave-one-out cross-validation is used to select the smoothing parameters for the continuous and discrete variables for NP (CVLS), for the smoothing parameter for the continuous variable for NP (FREQ), and for the knot and parameter selection for Spline. Minimization of the cross-validation function is achieved using multidimensional numerical search routines that allow for different smoothing parameters for all variables. Restarting is used to avoid the presence of local minima. The second-order Gaussian kernel is used for the continuous kernel, and the Aitchison and Aitken kernel is used for the discrete kernel. Further details and code are available upon request.

For what follows, we consider the following nonlinear data generating process (DGP):

$$\begin{aligned}y_i &= g_0\big(x_i^c, x_i^d\big) + \big[g_1\big(x_i^c, x_i^d\big) - g_0\big(x_i^c, x_i^d\big)\big]t_i + \epsilon_i \\ &= g_0\big(x_i^c, x_i^d\big) + \tau\big(x_i^c, x_i^d\big)t_i + \epsilon_i \\ &= 1 + 2x_{i1}^c/(4\pi) + 2x_{i1}^d + \big(x_{i1}^c\big)^2/(4\pi) + \tau t_i + \epsilon_i,\end{aligned}$$

where x_1^c is $U[-\pi, \pi]$ and $x_1^d \in \{0, 1\}$ with $P\big[x_1^d = 1\big] = 0.5$ and $x_2^d \in \{0, 1\}$ with $P\big[x_1^d = 1\big] = 0.5$, while $\sigma_\varepsilon = 1/2$ and x_2^d are irrelevant. Our model for the propensity score (Probit) is

$$T_i = -\pi/2 + \pi \sin\left(x_{i1}^c\right)/2 + \pi x_{i1}^d + \pi x_{i1}^d \cos\left(x_{i1}^c\right)/2 + \eta_i,$$

$$t_i = \begin{cases} 1 & \text{if } \Phi\left(T_i\right) > 0.5 \\ 0 & \text{otherwise,}\end{cases}$$

where $\sigma_\eta = 3$ and $\Phi(\cdot)$ is the standard normal CDF, while x_d^2 is irrelevant.

We first examine the sampling properties of each method, and then examine size and power of tests of H_0: $\tau = 0$ based on the estimators (1)–(3) mentioned earlier. We vary both τ and n, and consider $\tau \in \{0, 1/4, 1/2, 3/4\}$ and $n \in \{200, 300, 400, 500\}$.

3.1 Sampling Performance

To assess the sampling performance of the proposed method relative to its frequency-based peers, we draw $M = 1{,}000$ Monte Carlo replications from the preceding DGP for a given value of τ, and summarize the sampling performance of each estimator by reporting its median square error relative to the proposed smooth approach (i.e., we report ratios of $\text{med}_i\left(\hat{\tau}_i - \tau\right)^2,\ i = 1, \ldots, M$).Values of relative median square error greater than 1 indicate a loss in efficiency relative to the proposed method. Results are summarized in Table 1.

Table 1 reveals that the proposed smooth approach yields more efficient estimators (in finite-sample applications) of τ in finite samples than either the kernel or spline-based frequency approach. One important reason for the efficiency loss for the frequency-based estimators is that these estimators split the sample also for discrete cells related to x_2^d, while our kernel smoothed estimator does not as we can smooth out the irrelevant covariate x_2^d.

An anonymous referee has suggested to us that even when all variables are relevant, the nonparametric approach that smooths both the discrete and continuous variables in a particular manner may still have better finite-sample properties than the frequency estimator. Additionally, as the number of categories increases for the discrete variables, the benefits from smoothing discrete variables should be even more pronounced. Therefore, we also ran a set of simulations focusing on the issue of the efficiency gain arising by smoothing over discrete variables when, in fact, all variables are relevant. We modify the propensity score used previously so that $x_1^d, x_2^d \in \{0, 1, 2\}, p\big(x_1^d = x_2^d = 0\big) = 1/4, p\big(x_1^d = x_2^d = 1\big) = 1/2$, and $p\big(x_1^d = x_2^d = 2\big) = 1/4$, while $\sigma_\varepsilon = 1/2$. Our model for the propensity score (Probit) is now given by

Table 1. Relative sampling performance given by the ratio of median square errors where NP (CVLS) is the numeraire (four data cells)

n	NP (CVLS)	NP (FREQ)	Spline
	$\tau = 0.00$		
200	1.00	1.44	1.84
300	1.00	1.24	1.80
400	1.00	1.26	1.50
500	1.00	1.17	1.52
	$\tau = 0.25$		
200	1.00	1.22	1.70
300	1.00	1.39	1.53
400	1.00	1.35	1.35
500	1.00	1.40	1.54
	$\tau = 0.50$		
200	1.00	1.29	1.96
300	1.00	1.24	1.47
400	1.00	1.44	1.32
500	1.00	1.43	1.42
	$\tau = 0.75$		
200	1.00	1.51	1.96
300	1.00	1.30	1.45
400	1.00	1.34	1.20
500	1.00	1.39	1.16

Table 2. Relative sampling performance given by the ratio of median square errors where NP (CVLS) is the numeraire (nine data cells)

n	NP (CVLS)	NP (FREQ)
	$\tau = 0.00$	
200	1.00	1.19
300	1.00	1.19
400	1.00	1.18
500	1.00	1.18
	$\tau = 0.25$	
200	1.00	1.19
300	1.00	1.18
400	1.00	1.17
500	1.00	1.18
	$\tau = 0.50$	
200	1.00	1.24
300	1.00	1.17
400	1.00	1.20
500	1.00	1.17
	$\tau = 0.75$	
200	1.00	1.23
300	1.00	1.17
400	1.00	1.16
500	1.00	1.14

$$T_i = -\pi/2 + \pi x_{i2}^d \sin(x_{i1}^c)/2 + \pi x_{i1}^d + \pi x_{i1}^d \cos(x_{i1}^c)/2 + \eta_i,$$

$$t_i = \begin{cases} 1 & \text{if } \Phi(T_i) > 0.5 \\ 0 & \text{otherwise,} \end{cases}$$

where $\sigma_\eta = 3$ and $\Phi(\cdot)$ is the standard normal CDF, and both x_{i1}^d and x_{i2}^d are relevant (i.e., all variables are now relevant).

Table 2 presents relative efficiency for the case described previously where all variables are relevant (we do not include spline estimation results, because as the number of discrete cells increases, the frequency-based spline method is more prone to failing to converge, which leads to substantial increases in its relative square error performance).

The results in Table 2 clearly show that it is not just the presence of irrelevant variables, but also the number of cells in the data that drives the relative efficiency gains for our proposed smooth estimator.

3.2 Testing for the Null of No Effect

Next, we consider the performance of the proposed test for the null of no effect (i.e., H_0: $\tau = 0$). For a given value of τ, we generate each replication in the following manner.

(1) Draw a sample of size n for $\{x_1^c, x_1^d, x_2^d, \eta, \epsilon\}$, which then determines the values of $\{t, y\}$.
(2) Using $\{t_i, y_i, x_{i1}^c, x_{i1}^d, x_{i2}^d\}_{i=1}^n$, compute $\hat{\tau}$.
(3) Test for the null of no effect based on $B = 1{,}000$ bootstrap replications.
(4) Repeat steps 1 through 3 $M = 1{,}000$ times for a given value of τ.

We then construct empirical rejection frequencies, and the results are summarized in Table 3 for tests conducted at a nominal size of $\alpha = 0.05$ (results for 0.01 and 0.10 were also computed and are available upon request but are omitted for space considerations).

Examining Table 3, we observe that the smooth approach is correctly sized (i.e., when $\tau = 0$) and has power that increases with n and the magnitude of τ. The conventional frequency-based propensity score estimators suffer from substantial size distortions, suggesting that they are more susceptible to efficiency losses arising from sample splitting than the proposed smooth approach. It is clear that, at least for high-frequency nonlinear DGPs such as that considered previously, these frequency-based approaches are to be avoided altogether.

Note that we have only considered one binary irrelevant covariate in this experiment. When there exist more irrelevant covariates, or one irrelevant covariate that assumes more

Table 3. Size and power comparisons, $\alpha = 0.05$

n	NP (CVLS)	NP (FREQ)	Spline
	Size ($\tau = 0.00$)		
200	0.062	0.218	0.007
300	0.061	0.203	0.003
400	0.056	0.187	0.004
500	0.059	0.182	0.004
	Power ($\tau = 0.25$)		
200	0.116	0.312	0.013
300	0.133	0.328	0.051
400	0.125	0.276	0.095
500	0.166	0.286	0.171
	Power ($\tau = 0.50$)		
200	0.220	0.386	0.151
300	0.301	0.375	0.306
400	0.374	0.458	0.523
500	0.476	0.513	0.612
	Power ($\tau = 0.75$)		
200	0.411	0.475	0.373
300	0.581	0.592	0.631
400	0.692	0.661	0.745
500	0.816	0.783	0.814

than two unique values, the differences between the smooth and frequency-based approaches become even more pronounced. Summarizing, the proposed smooth test is more powerful than a conventional frequency-based test when confronted with categorical data, which is often the case in applied settings.

4. AN EMPIRICAL APPLICATION

In this section, we consider the performance of parametric and nonparametric propensity score models based upon data taken from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The data were obtained from the Department of Health Evaluation Sciences at the University of Virginia, and we are most grateful to Drs. B. Knaus and F. Harrell, Jr. for making these data available to us. These data were used in a study by Connors et al. (1996) who considered 30-day, 60-day, and 180-day survival, and they also considered categories of admission diagnosis and categories of comorbidities illness as covariates. We restrict attention to 180-day survival by way of example, while we ignore admission diagnosis and comorbidities illness due to the prevalence of missing observations among these covariates. As it is our intention to demonstrate the utility of the proposed methods on actual data and not to become immersed in ad hoc adjustments that must be made to handle the prevalence of missing data for these additional covariates, we beg the reader's forgiveness in this matter. Nevertheless, even though we omit admission diagnosis and comorbidities illness as covariates, we indeed observe results that are qualitatively and quantitatively similar to those reported in Connors et al. (1996) and Lin et al. (1998).

- Y: Outcome-1 if death occurred within 180 days, zero otherwise
- T: Treatment-1 if a Swan-Ganz catheter was received by the patient when they were hospitalized, zero otherwise.
- X_1: Sex-0 for female, 1 for male
- X_2: Race-0 if black, 1 if white, 2 if other
- X_3: Income-0 if under 11 K, 1 if 11-25 K, 2 if 25-50 K, 3 if over 50 K
- X_4: Primary disease category-1 if acute respiratory failure, 2 if congestive heart failure, 3 if chronic obstructive pulmonary disease, 4 if cirrhosis, 5 if colon cancer, 6 if coma, 7 if lung cancer, 8 if multiple organ system failure with malignancy, 9 if multiple organ system failure with sepsis
- X_5: Secondary disease category-1 if cirrhosis, 2 if colon cancer, 3 if coma, 4 if lung cancer, 5 if multiple organ system failure with malignancy, 6 if multiple organ system failure with sepsis, 7 if NA
- X_6: Medical insurance-1 if medicaid, 2 if medicare, 3 if medicare and medicaid, 4 if no insurance, 5 if private, 6 if private and medicare
- X_7: Age-age (converted to years from Y/M/D data stored with 2 decimal accuracy)

Table 4 presents some summary statistics on the variables described previously. The number of cells in this dataset is 18,144, which exceeds the number of records, 5,735.

Note that, as was found by Connors et al. (1996), those receiving right-heart catheterization are more likely to die within 180 days than those who did not. Interestingly, Lin et al. (1998) also find that, when further adjustments were made, the risk of death is lower than that reported by Connors et al. (1996) and they conclude that ''results of our sensitivity analysis provide additional insights into this important study and imply perhaps greater uncertainty about the role of RHC than those stated in the original report.''

Table 4. Summary statistics

Variable	Mean	StdDev	Minimum	Maximum
Outcome	0.65	0.48	0	1
Treatment	0.38	0.49	0	1
Sex	0.56	0.50	0	1
Race	0.90	0.46	0	2
Income	0.75	0.99	0	3
Primary disease category	3.98	3.34	1	9
Secondary disease category	6.66	0.84	1	7
Medical insurance	3.81	1.79	1	6
Age	61.38	16.68	18	102

Lin et al. (1998) note that cardiologists' and intensive care physicians' belief in the efficacy of RHC for guiding therapy for certain patients is so strong that ''it has prevented the conduct of a randomized clinical trial'' (RCT), while Connors et al. (1996) note that ''the most recent attempt at an RCT was stopped because most physicians refused to allow their patients to be randomized.''

The confusion matrices for the parametric and nonparametric propensity score models are given in Tables 5 and 6. A ''confusion matrix'' is simply a tabulation of the actual outcomes (A) versus those predicted (P) by a model. The diagonal elements contain correctly predicted outcomes, while the off-diagonal ones contain incorrectly predicted (confused) outcomes. The classification ratio (CR(0–1)) is the number of correctly predicted outcomes, while CR(0) and CR(1) denote the number of correctly predicted zeros and ones, respectively.

An examination of these confusion matrices demonstrates how, for this dataset, the nonparametric approach is better able to predict who receives treatment and who does not than the parametric Logit model. The parametric approach correctly predicts 3,828 of the 5,735 patients, while the nonparametric approach correctly predicts 3,976 patients, thereby predicting an additional 148 patients correctly. The differences between the parametric and nonparametric versions of the weighting estimator reflect this additional number of correctly classified patients along with differences in the estimated probabilities of treatment themselves. The increased risk suggested by the parametric model drops from a 7% increase for those receiving RHC to 0% when the proposed smooth nonparametric method is used.

Table 5. Parametric confusion matrix

A/P	0	1
0	2,841	710
1	1,197	987
Sample size		5,735
CR(0-1)		66.7%
CR(0)		80.0%
CR(1)		45.2%

Table 6. Nonparametric confusion matrix

A/P	0	1
0	2,916	635
1	1,092	1,092
Sample size		5,735
CR(0-1)		69.9%
CR(0)		82.1%
CR(1)		50.0%

Based upon the parametric propensity score estimate, the treatment effect is 0.072, while the nonparametric propensity score estimate yields a treatment effect of −0.001. We then bootstrapped the sampling distribution of these estimates and obtained 95% coverage error bounds of [0.044, 0.099] for the parametric approach and [−0.039, 0.010] for the nonparametric approach. Thus, we overturn the parametric testing result and conclude that patients receiving RHC treatment, in fact, do not appear to suffer an increased risk of death. These error bounds indicate that the parametric model suggests a statistically significant increased risk of death for those receiving RHC, while the nonparametric model yields no significant difference.

APPENDIX A

Definition A.1. (Definition of $\hat{V}_1, \hat{V}_2$ and $\hat{B}_{h,\lambda}$). $\hat{V}_1 = n^{-1}\sum_{i=1}^{n}[\hat{\tau}_{ni} - \hat{m}_\tau]^2$, where $\hat{\tau}_{ni}$ is defined in (18) and $\hat{m}_\tau = n^{-1}\sum_{j=1}^{n}\hat{\tau}_{nj}$ is the sample mean of $\hat{\tau}_n(x_j)$.

$\hat{V}_2 = n^{-1}\sum_{i=1}^{n}\hat{u}_i^2(t_i - \hat{t}_i)^2/[\hat{t}_i^2(1-\hat{t}_i^2)]$, where $\hat{t}_i$ is the kernel estimator of $E(t_i|x_i)$ defined in (10), $\hat{u}_i = \hat{g}_0(x_i) + t_i\hat{\tau}_n(x_i)$, and $\hat{g}_0(x_i)$ is defined in (16).

To estimate $B_{h,\lambda}$, we need to estimate B_{1s} $(s = 1, \ldots, q_1)$ and B_{2s} $(s = 1, \ldots, r_1)$. These quantities can be estimated by first estimating $f(x_i)$, $m(x_i) = g_0(x_i) + \mu(x_i)\tau(x_i)$ by $\hat{f}(x_i)$ and $\hat{m}(x_i) = \hat{g}_0(x_i) + \hat{\mu}(x_i)\hat{\tau}_n(x_i)$, respectively, and their derivative estimators can be obtained by taking derivatives (since the kernel function is differentiable up to order ν). Finally, replacing the population mean $E(\cdot)$ by the sample mean leads to consistent estimators for B_{1s} *and* B_{2s}, and hence for $B_{h,\lambda}$.

In Appendices A and B, because $M_n(x) \to 1$ on the support of $f(x)$, we will omit the trimming function M_{ni} to simplify notation. Also, we will use the notation (s.o.), which is defined as follows: when B_n is the leading term of A_n, we write $A_n = B_n +$ (s.o.), where (s.o.) denotes terms having probability order smaller than B_n. Also, when we write $A(x_i) = B(x_i) +$ (s.o.), it is always understood to mean that $n^{-1}\sum_{i=1}^{n}A(x_i) = n^{-1}\sum_{i=1}^{n}B(x_i) +$ (s.o.).

Proof of Lemma 2.1. We first list conditions that are needed to prove Lemma 2.1. We use $\bar{X}$ to denote the first q_1 relevant components of X^c and the first r_1 components of X^d. Let $\tilde{X} = X/\bar{X}$ denote the remaining components of X. We make the following assumptions:

(C1) $(Y, \bar{X})$ is independent of $\tilde{X}$.

(C2) The data are iid and u_i has finite moments of any order; g, f, and σ^2 have two continuous derivatives; $M(\cdot)$ is continuous, non-negative, and has compact support; and f is bounded away from zero for $x = (x^c, x^d) \in M \times S^d$.

(C3) Define $H = \left(\prod_{s=1}^{q_1}h_s\right)\prod_{s=q_1+1}^{q}\min(h_s, 1)$. Let $0 < \varepsilon < 1/(q+4)$ and for some constant $c > 0$, $n^{\varepsilon-1} \le H \le n^{-c}$; $n^{-c} < h_s < n^c$ for all $s = 1, \ldots, q$; the kernel function is a symmetric, compactly supported, Hölder-continuous probability density; $w(0) > w(\delta)$ for all $\delta > 0$.

(C4) Define $\bar{\mu}_t = E[\hat{t}(x)\hat{f}(x)]/E[\hat{f}(x)]$; then, $\int_{\mathrm{supp}(M)}[\bar{\mu}_t(x) - \mu(\bar{x})]\bar{M}(\bar{x})\bar{f}(\bar{x})d\bar{x}$, as a function of $h_1,\ldots,h_{q_1}$, and $\lambda_1,\ldots,\lambda_{r_1}$, vanishes if and only if all of the smoothing parameters vanish.

(C5) Let $\chi(a, b)$ be defined as in (A.5) subsequently. We assume that there exist unique finite positive constants α_s^0 $(s = 1, \ldots, q_1)$ and finite non-negative constants λ_s $(s = 1, \ldots, r_1)$ that minimize $\chi(a, b)$.

We use the notation $g_s^{(l)}(x)$ to denote $\partial^l g(x)/\partial\left(x_s^c\right)^l$, the lth order partial derivative of $g(\cdot)$ with respect to ${x_s}^c$. Also, when x_s^d is an unordered categorical variable, define an indicator function $I_s(.,.)$ by

$$I_s(z^d, x^d) = 1\left(z_s^d \ne x_s^d\right)\prod_{t\ne s}^{r_1}1\left(x_t^d = z_t^d\right). \tag{A.1}$$

When x_s^d is an ordered categorical variable, for notational simplicity, we assume that x_s^d takes (finitely many) consecutive integer values, and $I_s(.,.)$ is defined by

$$I_s(z^d, x^d) = 1\left(|z_s^d - x_s^d| = 1\right)\prod_{t\ne s}^{r_1}1\left(x_t^d = z_t^d\right). \tag{A.2}$$

When using a second-order kernel ($\nu = 2$), Hall et al. (2004, 2006) have shown that $\hat{h}_s \to \infty$ for $s = q_1 + 1, \ldots, q$, that $\hat{\lambda}_s \to 1$ for $s = r_1 + 1, \ldots, r$, and that

$$\begin{aligned} CV(h,\lambda)|_{\nu=2} &= \sum_{x^d}\int\left\{\frac{\kappa_2}{2}\sum_{s=1}^{q_1}\left[(f\mu)_s^{(2)}(x) - \mu(x)f_s^{(2)}(x)\right]h_s^2\right. \\ &\quad \left. + \sum_{v^d}\sum_{s=1}^{r_1}I_s(v^d, x^d)[\mu(x^c, v^d) - \mu(x)]f(x^c, v^d)\lambda_s\right\}^2 \\ &\quad \times S(x)f(x)^{-1}dx^c \\ &\quad + \frac{\kappa^{q_1}}{nh_1\ldots h_{q_1}}\sum_{x^d}\int\sigma^2(x)S(x)dx^c + (\text{s.o.}), \end{aligned} \tag{A.3}$$

where $\kappa_2 = \int w(v)v^2dv, \kappa = \int w(v)^2dv$, and $I_s(\cdot)$ is defined in (A.1) and (A.2).

By following exactly the same derivation as in Hall et al. (2004, 2006), one can show that, with a ν*th* order kernel, $\hat{h}_s \to \infty$ for $s = q_1 + 1, \ldots, q$, that $\hat{\lambda}_s \to 1$ for $s = r_1 + 1, \ldots, r$, and that

$$\begin{aligned} CV(h,\lambda) &= \sum_{x^d}\int\left\{\sum_{s=1}^{q_1}B_{1s}(x)h_s^\nu + \sum_{s=1}^{r_1}B_{2s}(x)\lambda_s\right\}^2 \\ &\quad \times S(x)f(x)^{-1}dx^c \\ &\quad + \frac{\kappa^{q_1}}{nh_1\ldots h_{q_1}}\sum_{x^d}\int\sigma^2(x)S(x)\,dx^c + (\text{s.o.}), \end{aligned} \tag{A.4}$$

where $B_{1s}(x) = (\kappa_{q_1}/\nu!)[(f\mu)_s^{(\nu)}(x) - \mu(x)f_s^{(\nu)}(x)](s = 1, \ldots, q_1), \kappa_{q_1} = \int w(v)v^{q_1}dv$, and $B_{2s}(x) = \sum_{v^d}I_s(v^d, x^d)[\mu(x^c,$

$v^d) - \mu(x)]f(x^c, v^d)(s = 1, \ldots, r_1)$, and (s.o.) denotes terms having smaller probability orders, uniformly in $(h, \lambda) \in (0, \eta_n]^{q_1+r_1}$.

The only difference between (A.3) and (A.4) is that h_s^2 is replaced by h_s^ν and that the definition of B_{1s} is slightly different. Of course, (A.4) reduces to (A.3) if $\nu = 2$.

Defining a_s via $h_s = a_s n^{-1/(2\nu+q_1)}(s = 1, \ldots, q_1)$ and b_s via $\lambda_s = b_s n^{-\nu/(2\nu+q_1)}(s = 1, \ldots, r_1)$, (A.4) can be written as $CV(h, \lambda) = n^{-2\nu/(2\nu+q_1)}\chi(a, b) + (\text{s.o.})$ uniformly in $(h, \lambda) \in (0, \eta_n]^{q_1+r_1}$, where

$$\chi(a,b) = \sum_{x^d} \int \left\{ \sum_{s=1}^{q_1} B_{1s}(x) a_s^\nu + \sum_{s=1}^{r_1} B_{2s}(x) b_s \right\}^2 S(x) f(x)^{-1} dx^c + \frac{\kappa^{q_1}}{a_1 \ldots a_{q_1}} \sum_{x^d} \int \sigma^2(x) S(x) dx^c. \tag{A.5}$$

Note that the $\chi(a, b)$ function defined previously is the leading term of $CV(h, \lambda)$ (up to a factor $n^{-4/(4+q_1)}$, which is also the leading term of the weighted integrated mean-square error when estimating $\mu(x) = E(t_i|x_i = x)$). Li and Zhou (2005) provide a necessary and sufficient condition for Condition (C.5), i.e., for the existence of unique $(a_1, \ldots, a_{q_1}, b_1, \ldots, b_{r_1})$ that minimize $\chi(a, b)$. Li and Zhou (2005) also provide some intuitive explanations of Condition (C.5).

From (A.4), (A.5), and Condition (C5), we obtain $\hat{h}_s = a_s^0 n^{-1/(2\nu+q_1)} + o_p\left(n^{-1/(2\nu+q_1)}\right)$ and $\hat{\lambda}_s = b_s^0 n^{-\nu/(2\nu+q_1)} + o_p\left(n^{-\nu/(2\nu+q_1)}\right)$. ■

To make the proof of Theorem 2.1 more manageable, we make a number of simplifying assumptions. (1) We replace $\hat{t}(x_i)$ in the definition of $\hat{\tau}$ by the leave-one-out estimator $\hat{t}_{-i}(x_i)$ (or one can redefine $\hat{\tau}$ by replacing $\hat{t}(x_i)$ by $\hat{t}_{-i}(x_i)$ in $\hat{\tau}$). (2) We replace $\hat{h}_s$ by the nonstochastic quantity $h_s^0 = a_s^0 n^{-1/(\nu+q_1+2)}$ $(s = 1, \ldots, q_1)$, and $\hat{\lambda}_s$ by $\lambda_s^0 = b_s^0 n^{-2/(\nu+q_1+2)}(s = 1, \ldots, r_1)$. (3) When we evaluate the probability *order* of a term, we sometimes assume that $h_s = h$ for all $s = 1, \ldots, q_1$ and that $\lambda_s = \lambda$ for all $s = 1, \ldots, r_1$ to simplify notation. For example, we will write $O(h^\nu)$ for $O\left(\sum_{s=1}^{q_1} h_s^\nu\right)$ and $O(\lambda)$ for $O\left(\sum_{s=1}^{r_1} \lambda_s\right)$ to save space. Note that the proof carries through without making these simplifying assumptions. For example, ignoring the leave-one-out estimator only introduces some extra smaller order terms. Lemma 2.1 shows that $\hat{h}_s/h_s^0 - 1 = o_p(1)$ and $\hat{\lambda}_s/\lambda_s^0 - 1 = o_p(1)$, and by the stochastic equicontinuity result of Ichimura (2000) and Hsiao et al. (2006), we know that the asymptotic distribution of $\hat{\tau}$ remains the same whether one uses the $\hat{h}_s$'s and $\hat{\lambda}_s$'s or their nonstochastic leading terms (i.e., the h_s^0's and λ_s^0's). Or, alternatively, one can use tightness and stochastic-equicontinuity arguments to prove this result (e.g., Hsiao et al. (2006)).

We will repeatedly use the *u* statistic *H* decomposition in the subsequent proof. When we evaluate the *order* of some terms, we sometimes write n^{-1} for $(n-1)^{-1}$, because this approximation does not affect the *order* of any quantities considered.

We will use the short-hand notation $\hat{t}_i = \hat{t}_{-i}(x_i)$ and $\hat{f}_i = \hat{f}_{-i}(x_i)$; i.e.,

$$\hat{t}_i = \frac{n^{-1}\sum_{j\neq i} t_j K_\gamma(x_j, x_i)}{\hat{f}_i}, \tag{A.6}$$

with $\hat{f}_i = n^{-1}\sum_{j\neq i} K_\gamma(x_j, x_i)$.

Defining $v_i = t_i - E(t_i|x_i) \equiv t_i - \mu_i$, so that $t_i = \mu_i + v_i$, and replacing t_j by $\mu_j + v_j$ in the right-hand-side of (A.6), we have

$$\hat{t}_i = \hat{\mu}_i + \hat{v}_i, \tag{A.7}$$

where $\hat{\mu}_i = n^{-1}\sum_{j\neq i} \mu_j K_\gamma(x_j, x_i)/\hat{f}_i$, and $\hat{v}_i = n^{-1}\sum_{j\neq i} v_j K_\gamma(x_j, x_i)/\hat{f}_i$.

We use the short-hand notation w_i and $\tilde{w}_i$ defined by

$$w_i = \mu_i(1-\mu_i) \text{ and } \tilde{w}_i = \hat{t}_i(1-\hat{t}_i). \tag{A.8}$$

We use the following identities to handle the random denominator of $\hat{\tau}$:

$$\frac{1}{\tilde{w}_i} = \frac{1}{w_i} + \frac{w_i - \tilde{w}_i}{w_i^2} + \frac{(w_i - \tilde{w}_i)^2}{w_i^2 \tilde{w}_i}. \tag{A.9}$$

Proof of Theorem 2.1. We have defined $\tilde{\tau}$ in (15). We now define another intermediate quantity $\bar{\tau}$ ($v_i = t_i - \mu_i$ and we omit M_{ni} for notational simplicity):

$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^n \frac{(t_i - \mu_i) y_i}{w_i} \equiv \frac{1}{n}\sum_{i=1}^n \frac{v_i y_i}{w_i}. \tag{A.10}$$

By adding and subtracting terms in $\sqrt{n}(\hat{\tau} - \tau)$, we obtain

$$\begin{aligned} \sqrt{n}(\hat{\tau} - \tau) &= \sqrt{n}[(\hat{\tau} - \tilde{\tau}) + (\tilde{\tau} - \bar{\tau}) + (\bar{\tau} - \tau)] \\ &= J_{1n} + J_{2n} + J_{3n}, \end{aligned} \tag{A.11}$$

where $J_{1n} = \sqrt{n}(\hat{\tau} - \tilde{\tau}), J_{2n} = \sqrt{n}(\tilde{\tau} - \bar{\tau})$, and $J_{3n} = \sqrt{n}(\bar{\tau} - \tau)$.

Lemma A.4 gives the leading terms of J_{2n}. Recall that $w_i = \mu_i(1-\mu_i)$, $\tilde{w}_i = \hat{t}_i(1-\hat{t}_i)$. Using (A9), from (C4), we obtain

$$\begin{aligned} \hat{\tau} &= \frac{1}{n}\sum_{i=1}^n [t_i - \hat{t}_i]\, y_i \left[\frac{1}{w_i} + \frac{w_i - \tilde{w}_i}{w_i^2} + \frac{(w_i - \tilde{w}_i)^2}{w_i^2 \tilde{w}_i}\right] \\ &\equiv L_{1n} + L_{2n} + L_{3n}, \end{aligned} \tag{A.12}$$

where $L_{1n} = n^{-1}\sum_{i=1}^n [(t_i - \hat{t}_i) y_i]/w_i \equiv \tilde{\tau}, L_{2n} = n^{-1}\sum_{i=1}^n [(t_i - \hat{t}_i)(w_i - \tilde{w}_i) y_i]/[w_i^2]$, and $L_{3n} = n^{-1}\sum_{i=1}^n [(t_i - \hat{t}_i)(w_i - \tilde{w}_i)^2 y_i]/[w_i^2 \tilde{w}_i]$.

Note that $L_{1n} = \tilde{\tau}$; therefore, by (A.12), we have

$$J_{1n} \equiv \sqrt{n}(\hat{\tau} - \tilde{\tau}) = \sqrt{n}(\hat{\tau} - L_{1n}) = \sqrt{n} L_{2n} + \sqrt{n} L_{3n}. \tag{A.13}$$

Lemma A.3 below gives the leading term of J_{1n}.

Using (4) and adding and subtracting terms, we write $J_{3n} = \sqrt{n}(\bar{\tau} - \tau)$ as

$$\begin{aligned} J_{3n} &= n^{-1/2}\sum_{i=1}^n [v_i y_i / w_i - \tau] \\ &= n^{-1/2}\sum_{i=1}^n (v_i y_i / w_i - \tau_i) + n^{-1/2}\sum_{i=1}^n (\tau_i - \tau) \\ &= n^{-1/2}\sum_{i=1}^n [v_i(g_{0i} + \tau_i t_i + u_i)/w_i - \tau_i] \\ &\quad + n^{-1/2}\sum_{i=1}^n (\tau_i - \tau), \end{aligned} \tag{A.14}$$

where $\tau_i = \tau(x_i)$.

By (A.14), Lemma A.3, and Lemma A.4, from (A.11), we obtain

$$\begin{aligned}\sqrt{n}(\hat{\tau}-\tau-B_{h,\lambda}) &= J_{1n}+J_{2n}-n^{1/2}B_{h,\lambda}+J_{3n}\\ &= n^{-1/2}\sum_{i=1}^{n} v_i[2\mu_i-1]\tau_i/w_i - n^{-1/2}\\ &\quad\times\sum_{i=1}^{n} v_i(g_{0i}+\tau_i\mu_i)/w_i]\\ &\quad+n^{-1/2}\sum_{i=1}^{n}\{v_i(g_{0i}+\tau_i t_i+u_i)/w_i-\tau_i\}\\ &\quad+n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau)+o_p(1)\\ &= n^{-1/2}\sum_{i=1}^{n}\frac{v_i u_i}{w_i}+n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau)+o_p(1)\\ &\equiv Z_{n2}+Z_{n3}+o_p(1), \end{aligned}\tag{A.15}$$

where the definitions of $Z_{n2}=n^{-1/2}\sum_{i=1}^{n} v_i u_i/w_i$ and $Z_{n3}=n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau)$. Also, in the preceding, we used the following cancellation result ($w_i=\mu_i(1-\mu_i)$):

$$\begin{aligned}&n^{-1/2}\sum_{i=1}^{n}\left[\frac{v_i(t_i-\mu_i)}{w_i}-1\right]\tau_i+n^{-1/2}\sum_i v_i\left[\frac{2\mu_i-1}{w_i}\right]\tau_i\\ &=n^{-1/2}\sum_{i=1}^{n}\left[\frac{v_i^2-\mu_i(1-\mu_i)+2v_i\mu_i-v_i}{w_i}\right]\tau_i\\ &=n^{-1/2}\sum_{i=1}^{n}\left[\frac{(\mu_i+v_i)^2-(\mu_i+v_i)}{w_i}\right]\tau_i\\ &=0, \end{aligned}\tag{A.16}$$

since $(\mu_i+v_i)^2-(\mu_i+v_i)\equiv t_i^2-t_i=0$ (because $t_i^2=t_i$).

Theorem 2.1 follows from (A.15) and the Lindeberg central limit theorem. ■

Proof of Lemma 2.2. From $\sqrt{n}(\tilde{\tau}-\tau-B_{h,\lambda})=\sqrt{n}(\tilde{\tau}-\bar{\tau}-B_{h,\lambda})+\sqrt{n}(\bar{\tau}-\tau)=J_{2n}-n^{1/2}B_{h,\lambda}+J_{3n}$, and using (A.14) and Lemma A.4, we have ($v_i=t_i-\mu_i$):

$$\begin{aligned}\sqrt{n}(\tilde{\tau}-\tau-B_{h,\lambda}) &= J_{2n}-n^{1/2}B_{h,\lambda}+J_{3n}\\ &= -n^{-1/2}\sum_{i=1}^{n} v_i(g_{0i}+\tau_i\mu_i)/w_i\\ &\quad+\ n^{-1/2}\sum_{i=1}^{n}[v_i(g_{0i}+\tau_i t_i+u_i)/w_i-\tau_i]\\ &\quad+\ n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau)+o_p(1)\\ &= n^{-1/2}\sum_{i=1}^{n}\left[\frac{v_i^2}{w_i}-1\right]\tau_i\\ &\quad+\ n^{-1/2}\sum_{i=1}^{n}\frac{v_i u_i}{w_i}+n^{-1/2}\\ &\quad\times\sum_{i=1}^{n}(\tau_i-\tau)+o_p(1)\\ &\equiv Z_{n1}+Z_{n2}+Z_{n3}+o_p(1)\\ &\to^{d} N(0,V_1+V_2+V_3) \end{aligned}$$

by the Lindeberg central limit theorem, (A.17)

where Z_{n1} and Z_{n2} are defined in (A.15), and $Z_{n3}=n^{-1/2}\sum_{i=1}^{n}\left[(v_i^2/w_i)-1\right]$. Note that by Lemma A.3 and (A.16), we know that $J_{1n}=-Z_{n3}+o_p(1)$. ■

Subsequently, we present some lemmas that are used in proving Theorem 2.1. We will use the following identity to handle the random denominator in the kernel estimator. For any positive integer p, we have

$$\frac{1}{\hat{f}_i}=\frac{1}{f_i}+\frac{f_i-\hat{f}_i}{f_i\hat{f}_i}=\frac{1}{f_i}+\sum_{l=1}^{p}\frac{(f_i-\hat{f}_i)^l}{f_i^{l+1}}+\frac{(f_i-\hat{f}_i)^{p+1}}{f_i^p\hat{f}_i}.\tag{A.18}$$

For example, in Lemma A.1 we need to evaluate a term such as $n^{-1}\sum_{i=1}^{n} v_i y_i\hat{v}_i/w_i^2$. $\hat{v}_i=n^{-1}\sum_{j\neq i} v_j K_{\gamma,ij}/\hat{f}_i$, which has a random denominator, namely $\hat{f}_i$. By computing the second moment of the term associated with $(f_i-\hat{f}_i)^l/f_i^{l+1}$, one can easily show that this term has an order smaller than the main term that is associated with $1/f_i$. Also, using the uniform convergence rate of $\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)|=O_p\left(\sum_{s=1}^{q_1}h_s^v+\ln n\left(nh_1\ldots h_{q_1}\right)^{-1}\right)$, together with $\inf_{x\in\mathcal{S}}f(x)\geq\delta>0$, one can easily show the last remainder term associated with $(f_i-\hat{f}_i)^{p+1}/(f_i^p\hat{f}_i)$ is of smaller order than the first leading term (by choosing p to be sufficiently large). Therefore, using (A.18), we have

$$n^{-1}\sum_{i=1}^{n} v_i y_i\hat{v}_i/w_i^2=n^{-1}\sum_{i=1}^{n} v_i y_i\hat{v}_i\hat{f}_i/(f_i w_i^2)+(\text{s.o.}).$$

Now the leading term $n^{-1}\sum_{i=1}^{n} v_i y_i\hat{v}_i\hat{f}_i/(f_i w_i^2)$ does not contain the random denominator $\hat{f}_i$; hence, its probability order can be easily evaluated by using H decomposition of u statistics.

Lemma A.1.

$$L_{2n}=n^{-1}\sum_i v_i(2\mu_i-1)\tau_i/w_i+o_p\left(n^{-1/2}\right).$$

Proof. Recalling that $w_i=\mu_i(1-\mu_i)$, $\tilde{w}_i=\hat{t}_i(1-\hat{t}_i)$, $t_i=\mu_i+v_i$, and $\hat{t}_i=\hat{\mu}_i+\hat{v}_i$, we have

$$\begin{aligned}L_{2n} &= n^{-1}\sum_{i=1}^{n} y_i(t_i-\hat{t}_i)[\mu_i-\hat{t}_i-(\mu_i^2-\hat{t}_i^2)]/w_i^2\\ &= n^{-1}\sum_{i=1}^{n} y_i(t_i-\hat{t}_i)(\mu_i-\hat{t}_i)[1-(\mu_i+\hat{t}_i)]/w_i^2\\ &= n^{-1}\sum_{i=1}^{n} y_i(\mu_i-\hat{\mu}_i+v_i-\hat{v}_i)(\mu_i-\hat{\mu}_i-\hat{v}_i)\\ &\quad\times[1-(\mu_i+\hat{\mu}_i+\hat{v}_i)]/w_i^2\\ &= -n^{-1}\sum_{i=1}^{n} y_i v_i\hat{v}_i[1-2\mu_i]/w_i^2+o_p\left(n^{-1/2}\right)\\ &\quad(\text{using }\hat{\mu}_i=\mu_i+(\hat{\mu}_i-\mu_i),\ \text{Lemma B.3, and (A3) (ii)})\\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i v_i v_i(2\mu_i-1)K_{\gamma,ij}/w_i^2+o_p\left(n^{-1/2}\right)\\ &= \frac{2}{n(n-1)}\sum_i\sum_{j>i} H_{n,a}(z_i,z_j)+(\text{s.o.}), \end{aligned}$$

where $H_{n,a}(z_i,z_j)=(1/2)v_i v_j\{y_i(1-2\mu_i)/[f_i w_i^2]+y_j(2\mu_j-1)/[f_j w_j^2]\}K_\gamma$, and $_{ij}$ and $z_i=(x_i,t_i,u_i)$.

Define $H_{1n,a}(z_i) = E[H_{n,a}(z_i, z_j)|z_i] = (1/2)\, v_i\tau_i(2\mu_i - 1)/w_i^2$ + (s.o.) by Lemma B.4 (i).

Hence, by the u statistic H decomposition, we have

$$\begin{aligned} L_{2n} &= \frac{2}{n(n-1)} \sum_i \sum_{j>i} H_{n,a}(z_i, z_j) + (\text{s.o.}) \\ &= 0 + (2/n) \sum_i H_{1n,a}(z_i) + \frac{2}{n(n-1)} \sum_i \sum_{j>i} \{H_{n,a}(z_i, z_j) \\ &\quad - H_{1n,a}(z_i) - H_{1n,a}(z_j) + 0\} + (\text{s.o.}), \\ &= n^{-1} \sum_i v_i\tau_i(2\mu_i - 1)/w_i + O_p((nh^{q_1/2})^{-1}) + (\text{s.o.}) \\ &= n^{-1} \sum_i v_i\tau_i(2\mu_i - 1)/w_i + o_p(n^{-1/2}) \end{aligned}$$

by Lemma B.4 and Assumption (A3), where we also used the fact that the degenerate u statistic $U_{n,a} \stackrel{\text{def}}{=} [2/n(n-1)] \sum_i \sum_{j>i} \{H_{n,a}(z_i, z_j) - H_{1n,a}(z_i) - H_{1n,a}(z_j)\}$ has a second moment given by $E[U_{n,a}^2] = O((n^2h^{q_1})^{-1})$, so $U_{n,a} = O_p((nh^{q_1/2})^{-1})$. ■

Lemma A.2.

$$L_{3n} = O(h^{2\nu} + h^2(nh^{q_1})^{-1}) = o_p(n^{-1/2}).$$

Proof. Using the identity

$$\frac{1}{\tilde{w}_i} = \frac{1}{w_i} + \sum_{l=1}^{p} \frac{(w_i - \tilde{w}_i)^l}{w_i^{l+1}} + \frac{(w_i - \tilde{w}_i)^{p+1}}{w_i^p \tilde{w}_i}, \tag{A.19}$$

one can show that the leading term of L_{3n} is $L_{3n,1} = n^{-1} \sum_i y_i(t_i - \hat{t}_i)(w_i - \tilde{w}_i)^2/w_i^3$. This follows because (1) it is easy show that (by computing their second moment) the term associated with $(w_i - \tilde{w}_i)^l/w_i^{l+1}$ has an order smaller than the main term that is associated with $1/w_i$ and also because (2) using the uniform convergence rate of $\sup_{x\in\mathcal{S}} |\hat{\mu}(x) - \mu(x)| = O_p\big(\sum_{s=1}^{q_1} h_s^{\nu} + \ln n\big(nh_1 \ldots h_{q_1}\big)^{-1}\big)$, together with $\inf_{x\in\mathcal{S}} \mu(x) \geq c > 0$ and $\sup_{x\in\mathcal{S}} \mu(x) \leq c^{-1} < 1$ $(0 < c < 1)$, one can easily show the last remainder term associated with $(w_i - \tilde{w}_i)^{p+1}/(w_i^p \tilde{w}_i)$ is of smaller order than the first leading term (by choosing p to be sufficiently large if needed).

By noting that $t_i = \mu_i + v_i$ and $w_i - \tilde{w}_i = (\mu_i - \hat{t}_i)[1 - (\mu_i + \hat{t}_i)]$, we have

$$\begin{aligned} L_{3n,1} &= n^{-1} \sum_i y_i(t_i - \hat{t}_i)(w_i - \tilde{w}_i)^2/w_i^3 \\ &= n^{-1} \sum_i [g_{0i} + \tau_i t_i + u_i][(\mu_i - \hat{t}_i) + v_i](\mu_i - \hat{t}_i)^2 \\ &\quad \times [1 - (\mu_i + \hat{t}_i)]^2/w_i^3 \\ &\sim n^{-1} \sum_{i=1} n[v_i(\mu_i - \hat{t}_i)^2 + (\mu_i - \hat{t}_i)^3] \\ &= O(h^{2\nu} + h^2(nh^{q_1})^{-1}) \end{aligned}$$

by Lemma B.3, where in the above $A \sim B$ means that $A = B +$ (s.o.). ■

Lemma A.3.

$$J_{1n} = n^{-1/2} \sum\nolimits_i v_i(2\mu_i - 1)\tau_i/w_i + o_p(1).$$

Proof. This follows from lemmas A.1 and A.2. ■

Lemma A.4.

$$J_{2n} = \sqrt{n}B_{h,\lambda} - (1/\sqrt{n}) \sum\nolimits_{i=1}^{n} v_i(g_{0i} + \tau_i\mu_i)/w_i + o_p(1).$$

Proof. Using $\hat{t}_i = \hat{\mu}_i + \hat{v}_i$, we have $J_{2n} = n^{-1/2} \sum_{i=1}^n (\mu_i - \hat{t}_i)y_i/w_i = n^{-1/2} \sum_{i=1}^n (\mu_i - \hat{\mu}_i)y_i/w_i - n^{-1/2} \sum_{i=1}^n \hat{v}_i y_i/w_i \equiv J_{2n,1} - J_{2n,2}$.

We consider $J_{n2,1}$ first:

$$\begin{aligned} J_{2n,1} &\equiv n^{-1/2} \sum_{i=1}^{n} (\mu_i - \hat{\mu}_i)y_i/w_i \\ &= \sqrt{n}B_{h,\lambda} + O_p(\sqrt{n}h^{\nu+2} + h(nh^{q_1})^{-1/2}) \end{aligned}$$

by Lemma B.2, where $B_{h,\lambda}$ is defined in Lemma B.2.

Next,

$$\begin{aligned} J_{2n,2} &= n^{-1/2} \sum_{i=1}^{n} \hat{v}_i\hat{f}_i y_i/(f_i w_i) + o_p(1) (\text{by using (A.18)}) \\ &= n^{-1/2}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j\neq i}^{n} v_j y_i K_{\gamma,ij}/(f_i w_i) \\ &= \frac{2}{n^{1/2}(n-1)} \sum_{i=1}^{n} \sum_{j>i} (1/2)\{v_j y_i/(f_i w_i) \\ &\quad + v_i y_j/(f_j w_j)\} K_{\gamma,ij} \\ &= n^{1/2} \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i} H_{n,b}(z_i, z_j), \end{aligned}$$

where $H_{n,b}(z_i, z_j) = (1/2)\{v_j y_i/(f_i w_i) + v_i y_j/(f_j w_j)\}K_{\gamma,ij}$, and $z_i = (x_i, t_i, u_i)$.

By noting that $E(v_i|x_i) = 0$, we have (using $y_j = g_{0j} + \tau_j(\mu_j + v_j) + u_j$)

$$H_{1n,b}(z_i) \stackrel{\text{def}}{=} E[H_{n,b}(z_i, z_j)\,|\,z_i] = (1/2)v_i(g_{0i} + \tau_i\mu_i)/w_i + (\text{s.o.})$$

by Lemma B.4 (iii).

Hence, by the u statistics H decomposition, we have

$$\begin{aligned} J_{2n,2} &= -n^{1/2} \Bigg\{ 0 + (2/n) \sum_{i=1}^{n} H_{1n,b}(z_i) \\ &\quad + \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i} \times [H_{n,b}(z_i, z_j) \\ &\quad - H_{1n,b}(z_i) - H_{1n,b}(z_j) + 0] \\ &= n^{-1/2} \sum_{i=1}^{n} v_i(g_{0i} + \tau_i\mu_i)/w_i + O_p((nh^q)^{-1/2}) \end{aligned}$$

because the last term in the H decomposition is a degenerate u statistic, which has an order $n^{1/2}O_p\big((nh^{q_1/2})^{-1}\big) = O_p\big((nh^{q_1})^{-1/2}\big)$. ■

APPENDIX B

Lemma B.1. Let $\mathcal{D}$ denote the support of x^d, for all $x^d \in \mathcal{D}$, let $g(x^d, x^c) \in \mathcal{G}_\nu$, and let $f(x^d, x^c) \in \mathcal{G}_{\nu-1}$, where $\nu \geq 2$ is an even integer. Define $\eta_2 = \sum_{s=1}^{q_1} h_s^{\nu} + \sum_{s=1}^{r_1} \lambda_s$. Suppose the kernel function $W(\cdot)$ satisfies Assumption (A2), all h_s have the same order, say, h, and all λ_s have the same order as h^{ν}. Then, uniformly in x,

(i) $E\big\{[g(X) - g(x)]K_\gamma(X, x)\big\} = \sum_{s=1}^{q_1} D_{1s}(x)h_s^{\nu} + \sum_{s=1}^{r_1} D_{2s}(x)\lambda_s + O\big(\eta_2 h^2\big)$;

(ii) $E\big[K_\gamma(X,x)\big]-f(x)\big] = \sum_{s=1}^{q_1}\bar{D}_{1s}(x)h_s^\nu + \sum_{s=1}^{r_1}\bar{D}_{2s}(x)\lambda_s + O\big(\eta_2 h^2\big)$, where $D_{ls}(\cdot)$ and $\bar{D}_{ls}(\cdot)$ are defined in the subsequent proof.

Proof of (i).

$$
\begin{aligned}
&E\big\{[g(X)-g(x)]K_\gamma(X,x)\big\}\\
&=\sum_{z^d}\int f(z^c,z^d)\big[g(z^c,z^d)-g(x^c,x^d)\big]\\
&\quad\times W_h(z^c,x^c)L(z^d,x^d,\lambda)\,dz^c\\
&=\int f(x^c+hv,x^d)\big[g(x^c+hv,x^d)-g(x^c,x^d)\big]W(v)\,dv\\
&\quad+\sum_{z^d\neq x^d}\int f(z^c,z^d)[g(z^c,z^d)\ -g(x^c,x^d)]W_h(z^c,x^c)\\
&\quad\times L(z^d,x^d,\lambda_j)\,dz^c\\
&=\int\{(fg)(x^c+hv,x^d)-(fg)(x)-g(x)[f(x^c+hv,x^d)\\
&\quad-f(x)]W(v)\,dv\sum_{s=1}^{r_1}I_s(z^d,x^d)f(x^c,x^d)[g(x^c,z^d)-g(x^c,x^d)]\lambda_s\\
&\quad+o\left(\sum_{s=1}^{q_1}h_s^\nu+\sum_{s=1}^{r_1}\lambda_s\right)=\sum_{s=1}^{q_1}D_{1s}h_s^\nu+\sum_{s=1}^{r_1}D_{2s}(x)\lambda_s+O\big(\eta_2h^2\big)
\end{aligned}
$$

by Taylor series expansion and the fact that $W(\cdot)$ is a νth-order kernel function, where

$$D_{1s}(x)=(1/\nu!)\kappa_\nu[(gf)_s^{(\nu)}(x)-g(x)f_s^{(\nu)}(x)], \tag{B.1}$$

$$\kappa_\nu=\int w(v)v^\nu dv, \text{ and}$$

$$D_{2s}(x)=I_s(z^d,x^d)f(x^c,x^d)\big[g(x^c,z^d)-g(x^c,x^d)\big]. \tag{B.2}$$

■

Proof of (ii).

$$
\begin{aligned}
&E\{[K_\gamma(X,x)-f(x)]\}\\
&=\sum_{z^d}\int f(z^c,z^d)W_h(z^c,x^c)L(z^d,x^d,\lambda)dz^d-f(x^c,x^d)\\
&=\int f(x^c+hv,x^d)W(v)dv-f(x^c,x^d)\\
&\quad+\sum_{s=1}^{r_1}I_s(z^d,x^d)f(x^c,z^d)\lambda_s+O\left(h^\nu\left(h^2+\sum_{s=1}^{r_1}\lambda_s\right)\right)\\
&=\sum_{s=1}^{q_1}\bar{D}_{1s}(x)h_s^\nu+\sum_{s=1}^{r_1}\bar{D}_{2s}(x)\lambda_s+O\big(\eta_2h^2\big),
\end{aligned}
$$

where $\bar{D}_{1s}(x)=(\kappa_\nu/\nu!)f_s^\nu(x)$ and $\bar{D}_{2s}=I_s(z^d,x^d)f(x^c,z^d)$. ■

Lemma B.2.

$$
\begin{aligned}
A_{1n}&\stackrel{\text{def}}{=}n^{-1}\sum_i(\mu_i-\hat{\mu}_i)y_i/w_i\\
&=B_{h,\lambda}+O_p(n^{-1/2}h^\nu+h\big(n^2h^{q_1}\big)^{-1/2}),
\end{aligned}
$$

where the definition of $B_{h,\lambda}$ is given in the following proof.

Proof. Using (A.18), we know that $A_{1n}=A_{1n,1}+$ (s.o.), where $A_{1n,1}=n^{-1}\sum_i(\mu_i-\hat{\mu}_i)\hat{f}_iy_i/(w_if_i)$. Note that $E(u_i|x_i)=0$ and $E(v_i|x_i)=0$. Letting $m(x)=E(y|x)=g_{01}(x)+\tau(x)\mu(x)$, we first compute $E(A_{1n})$. We observe that

$$
\begin{aligned}
E(A_{1n,1})&=E[(\mu_1-\mu_2)K_{\gamma,1,2}y_1/(f_1w_1)]\\
&=\sum_{x_1^d}\sum_{x_2^d}\int\int f(x_2)m(x_1)w(x_i)^{-1}(\mu_1-\mu_2)\\
&\quad\times W_{h,1,2}L_{\lambda,1,2,}dx_1^cdx_2^c\\
&=\sum_{x^d}\int\int m(x)w(x)^{-1}f(x^2+hv,x^d)\\
&\quad\times\big[\mu(x)-\mu(x^c+hv,x^d)\big]W(v)\,dv\,dx^c\\
&\quad+\sum_{x^d}\sum_{x_2^d\neq x^d}\int\int m(x)w(x)^{-1}f(x^c+hv,x^d)\\
&\quad\times\big[\mu(x)-\mu(x^c+hv,x^d)\big]W(v)L_{\lambda,1,2}dx^cdv\\
&=\sum_{s=1}^{q_1}C_{1s}h_s^\nu+\sum_{s=1}^{r_1}C_{2s}\lambda_s+O\big(h^{\nu+2}\big)\\
&\equiv B_{h,\lambda}+O\big(h^{\nu+2}\big)
\end{aligned}
$$

by the same proof used for Lemma B.1 (i), where $B_{h,\lambda}=\sum_{s=1}^{q_1}C_{1s}h_s^\nu+\sum_{s=1}^{r_1}C_{2s}\lambda_s$ with

$$
\begin{aligned}
C_{1s}&=(\kappa_\nu/\nu!)\sum_{x^d}\int m(x)w(x)^{-1}\\
&\quad\times[\nu(x)f_s^{(\nu)}(x)-(\mu f)_s^{(\nu)}(x)]dx^c,
\end{aligned}\tag{B.3}
$$

and

$$
\begin{aligned}
C_{2s}&=\int\sum_{z^d}\sum_{x^d}I_s(z^d,x^d)f(x)m(x)w(x)^{-1}f(x)\\
&\quad\times\big[\mu(x)-\mu(x^c,z^d)\big]dz^c.
\end{aligned}\tag{B.4}
$$

Next, we compute var $(A_{1n,1})=E[A_{1n,1}^2]-[E(A_{1n,1})]^2$. Note that

$$
\begin{aligned}
E(A_{1n,1}^2)&=n^{-4}\sum_{i_1}\sum_{j_1\neq i_1}\sum_{i_2}\sum_{j_2\neq i_2}E\big[(\mu_{i_1}-\mu_{j_1})K_{\gamma,i_1,j_1}y_{i_1}\\
&\quad\times(\mu_{i_2}-\mu_{j_2})K_{\gamma,i_2,j_2}y_{i_2}/(w_{i_1}w_{i_2})\big].
\end{aligned}
$$

We consider three cases: (1) the four indices i_1, j_1, i_2, and j_2 are all different; (2) the four indices assume three distinct values; and (3) the four indices assume two different values.

First, for case (1), it is easy to see that $E(A_{1n,1(i)}^2)$ will cancel the leading term of $[E(A_{1n,1})]^2$. Therefore, we have

$$
\begin{aligned}
E(A_{1n,1(i)}^2)-[E(A_{1n,1})]^2&=n^{-1}O([E(A_{1n,1})]^2)\\
&=O\big(n^{-1}h^{2\nu}\big).
\end{aligned}\tag{B.5}
$$

For case (2), using Lemma B.1 (i) with $h_s=h$ and $\lambda_s=O(h^\nu)$, we have

$$E(A_{1n,1(ii)}^2) \le Cn^{-4}n^3h^{2\nu}\{E[y_1^2] + E[|y_1y_3|]\} = O(n^{-1}h^{2\nu}). \tag{B.6}$$

Finally, for case (iii), we have

$$E(A_{1n,1(iii)}^2) \le Cn^{-4}n^2\{E[y_1^2(\mu_1-\mu_2)^2K_{\gamma,12}^2] + E[y_1y_3(\mu_1-\mu_3)^2K_{\gamma,13}^2]\} = n^{-2}O(h^{-q_1}h^2) = O((n^2h^{q_1-2})^{-1}). \tag{B.7}$$

Summarizing (B.5)–(B.7), we have shown that

$$\operatorname{var}(A_{1n,1}) = E[A_{1n,1}^2] - [E(A_{1n,1}]^2 = O(n^{-1}h^{2\nu} + (n^2h^{q_1-2})^{-1}). \tag{B.8}$$

Hence,

$$A_{1n,1} = E(A_{1n,1}) + O_p\Big(\sqrt{\operatorname{var}(A_{1n,1})}\Big) = \sum_{s=1}^{q_1} C_{1s}h_s^\nu + \sum_{s=1}^{r_1} C_{2s}\lambda_s + O_p(n^{-1/2}h^\nu + h\big(n^2h^{q1}\big)^{-1/2}).$$

■

Lemma B.3.

(i) $A_{2n} \overset{def}{=} n^{-1}\sum_i(\hat\mu_i-\mu_i)^2 = O_p(h^{2\nu} + h^2(nh^{q_1})^{-1})$.

(ii) $A_{3n} \overset{def}{=} n^{-1}\sum_i \hat v^2{}_i = O_p((nh^{q_1})^{-1})$.

(iii) $A_{4n} \overset{def}{=} n^{-1}\sum_i(\hat\mu_i-\mu_i)\hat v_i = O_p(h^{2\nu} + (nh^{q_1})^{-1})$.

(iv) $A_{5n} \overset{def}{=} n^{-1}\sum_i(\hat t_i-\mu_i)^2 = O_p(h^{2\nu} + (nh^{q_1})^{-1})$.

Proof of (i). Using (A.18), we have $A_{2n} \equiv n^{-1}\sum_i(\hat\mu_i-\mu_i)^2\hat f_i^2/\hat f_i^2 = n^{-1}\sum_i(\hat\mu_i-\mu_i)^2\hat f_i^2/f_i^2 + (\text{s.o.})$. Also, since $f(x)$ is bounded below by a positive constant, we only need to prove (i) for $A_{2n,1} \overset{def}{=} n^{-1}\sum_i(\hat\mu_i-\mu_i)^2\hat f_i^2$. Observe that

$$\begin{aligned}E[|A_{2n,1}|] &= E[(\hat\mu_1-\mu_1)^2\hat f_1^2]\\ &= \frac{1}{(n-1)^2}\sum_{i\ne1}^n\sum_{j\ne1}^n E\big[(\mu_i-\mu_1)K_{\gamma,i,1}(\mu_j-\mu_1)K_{\gamma,j,1}\big]\\ &= \frac{1}{(n-1)^2}\big\{(n-1)E[(\mu_i-\mu_1)^2K_{n,i,1}^2]\\ &\quad + (n-1)(n-2)E[(\mu_2-\mu_1)K_{\gamma,2,1}]E[(\mu_3-\mu_1)K_{\gamma,3,1}]\big\}\\ &= O(h^2(nh^{q_1})^{-1}) + O(h^{2\nu})\end{aligned}$$

by Lemma B.1, where we used $E[(\mu_i-\mu_1)^2K_{n,i,1}^2] = O((h^2+\sum_{s=1}^{r_1}\lambda_s)h^{-q_1}) = O\big((h^2+h^\nu)h^{-q_1}\big) = O\big(h^2h^{-q_1}\big)$, because $\lambda_j = O(h^\nu)$ and $\nu \ge 2$. Thus, $A_{2n,1} = O_p(h^2(nh^{q_1})^{-1}) + O\big(h^{2\nu}\big)$. ■

Proof of (ii). Similarly, by (A.18), we have $A_{3n} \equiv n^{-1}\sum_{i=1}^n \hat v_i^2\hat f_i^2/\hat f_i^2 = n^{-1}\sum_{i=1}^n \hat v_i^2\hat f_i^2/f_i^2 + (\text{s.o.})$. We only need to prove (ii) for $A_{3n,1} = n^{-1}\sum_i \hat\epsilon_i^2\hat f_i^2$ (since f_i^{-1} is bounded). Note that

$$\begin{aligned}E\big[|A_{3n,1}|\big] &= E\Big[\hat v_1^2\hat f_1^2\Big]\\ &= \frac{1}{(n-1)^2}\sum_{i\ne1}^E[v_i^2K_{\gamma,i,1}^2]\\ &= \frac{1}{n-1}E[v_2^2K_{\gamma,2,1}^2] = O((nh^{q_1})^{-1}).\end{aligned}$$

■

Proof of (iii). (iii) follows from (i) and (ii) and the Cauchy inequality. ■

Proof of (iv). Finally, (iv) follows from (i)–(iii), because $(\hat t_i-\mu_i)^2 = (\hat\mu_i-\mu_i)^2 + \hat v_i^2 + 2(\hat\mu_i-\mu_i)\hat v_i(\hat t_i = \hat\mu_i + \hat v_i)$. ■

Lemma B.4. Let $H_{n,a}(z_i, z_j)$ and $H_{n,b}(z_i, z_j)$ be defined as in Lemmas A.1 and A.4, respectively. Recalling that $A_i = B_i + (\text{s.o.})$ means that $n^{-1/2}\sum_{i=1}^n A_i = n^{-1/2}\sum_{i=1}^n B_i + (\text{s.o.})$, then we have

(i) $H_{1n,a}(z_i) \overset{\text{def}}{=} E[H_{n,a}(z_i,z_j)\,|\,z_i] = \tau_i\{2\mu_i-1\}/w_i + (\text{s.o.})$,

(ii) $H_{1n,b}(z_i) \overset{\text{def}}{=} E[H_{n,b}(z_i,z_j)\,|\,z_i] = (g_{0i}+\tau_i\mu_i)/w_i + (\text{s.o.})$.

Proof of (i). $H_{n,a}(z_i, z_j) = (1/2)\{y_iv_iv_j(2\mu_i-1)/(f_iw_i^2) + y_jv_iv_j(2\mu_j-1)/(f_jw_j^2)\}K_{\gamma,ij}$, where $z_i = (x_i, t_i, u_i)$. By noting that μ_i, w_i, f_ig_{0i}, and τ_i are all functions of x_i and that $E(v_i|x_i) = 0$, we have

$$\begin{aligned}&E\big[y_iv_iv_j(2\mu_i-1)K_{\gamma,ij}/(f_iw_i^2)\,|\,z_i\big]\\ &\quad= y_iv_i(2\mu_i-1)(f_iw_i^2)^{-1}E\big\{E[v_jK_{\gamma,ij}\,|\,x_j,z_i]|\,z_i\big\}\\ &\quad= 0.\end{aligned}$$

Also, using $y_j = g_{0j} + \tau_jt_j + u_j = g_{0j} + \tau_j(\mu_j+v_j) + u_j$, and $E(v_j|x_j) = 0$, we have

$$\begin{aligned}H_{1n,a}(z_i) &= E[H_{n,a}(z_i,z_j)\,|\,z_i]\\ &= (1/2)\big\{0 + 2v_iE\big[v_jy_j\mu_jK_{\gamma,ij}/(f_jw_j^2)\,|\,z_i\big]\\ &\quad - v_iE[v_jy_jK_{\gamma,ij}/(f_jw_j^2)\,|\,z_i\big\}\\ &= (1/2)\big\{2v_iE[v_j^2\tau_j\mu_jK_{\gamma,ij}/(f_jw_j^2)\,|\,z_i]\\ &\quad - v_iE[v_j^2\tau_jK_{\gamma,ij}/(f_jw_j^2)|z_i]\big\}\\ &= (1/2)v_i\big\{2E[\tau_j\mu_jK_{\gamma,ij}/(f_jw_j)\,|\,z_i]\\ &\quad - E[\tau_jK_{\gamma,ij}/(f_jw_j)\,|\,z_i]\big\}\\ &\quad\times\ (\text{because } E(v_j^2\,|\,x_j) = \operatorname{var}(t_j\,|\,x_j) = w_j)\\ &= (1/2)v_i\tau_i\{2\mu_i-1\}/w_i + (\text{s.o.}),\end{aligned}$$

where we have used the change-of-variable argument: $E\big[\tau_jK_\gamma((x_i-x_j)/h)/(f_jw_j)\big]|z_i] = \tau_i + O(h^\nu+\lambda)$ and $E\big[\tau_j\mu_jK_\gamma((x_i-x_j)/h)/(f_jw_j)\big]|z_i] = \tau_i\mu_i + O(h^\nu+\lambda)$. ■

Proof of (ii). Note that $H_{n,b}(z_i, z_j) = (1/2)\{v_j y_i/(f_i w_i) + v_i y_j/(f_j w_j)\}K_{\gamma,ij}$, and $z_i = (x_i, t_i, u_i)$. By noting that $E(v_i|x_i) = 0$, we have

$$\begin{aligned} H_{1n,b}(z_i) &= E\big[H_{n,b}(z_i, z_j)\,|\,z_i\big] \\ &= (1/2)\big\{0 + v_i E\big[y_j K_{\gamma,ij}/(f_j w_j)\,|\,z_i\big]\big\} \\ &= (1/2)v_i E\big\{\big[\big(g_{0j}+\tau_j(\mu_j+v_j)+u_j\big)K_{\gamma,ij}/(f_j w_j)\,|\,z_i\big]\big\} \\ &= (1/2)v_i E\big\{[(g_{0j}+\tau_j\mu_j)K_{\gamma,ij}/(f_j w_j)\,|\,z_i]\big\} \\ &= (1/2)v_i(g_{0i}+\tau_i\mu_i)/w_i + (\text{s.o.}) \end{aligned}$$

by the change-of-variable argument. ■

APPENDIX C: PROOF OF THEOREM 2.2

First, we introduce some notation. We will write $A_n = o_p^*(1)$ if, for all $\varepsilon > 0$, $P^*(|A_n| > \varepsilon) \equiv P(|A_n| > \varepsilon|\{y_i, x_i, t_i\}_{i=1}^n) = o_p(1)$ (it is $o(1)$ in probability, not necessarily $o(1)$ almost surely). We will use the short-hand notation $\hat{t}_i^* = \hat{t}_{-i}^*(x_i^*)$ and $\hat{f}_i^* = \hat{f}_{-i}^*(x_i^*)$; i.e.,

$$\hat{t}_i^* = n^{-1}\sum_{j\neq i}^n t_j^* K_{n,ij}^*/\hat{f}_i^*, \tag{C.1}$$

with $\hat{f}_i^* \equiv \hat{f}_{-i}^*(x_i^*) = n^{-1}\sum_{j\neq i}^n K_{n,ij}^*$ and $K_{n,i_j}^* = K_n(x_i^*, x_j^*) = W_h(x_j^{c*}, x_j^{c*})L(x_i^{d*}, x_j^{d*}, \lambda)$.

Defining $v_i^* = t_i^* - E(t_i^*|x_i^*) \equiv t_i^* - \mu_i^*(\mu_i^* = \mu(x_i^*))$, so that $t_i^* = \mu_i^* + v_i^*$, and replacing t_j^* by $\mu_j^* + v_j^*$ in the right-hand-side of (C.1), we have

$$\hat{t}_i^* = \hat{\mu}_i^* + \hat{v}_i^*, \tag{C.2}$$

where $\hat{\mu}_i^* = n^{-1}\sum_{j\neq i}^n \mu_j^* K_{n,ij}^*/\hat{f}_i^*$, and $\hat{v}_i^* = n^{-1}\sum_{j\neq i}^n v_j^* K_{n,ij}^*/\hat{f}_i^*$.

We use the short-hand notation w_i* and $\tilde{w}_i^*$ defined by

$$w_i^* = \mu_i^*(1-\mu_i^*)\text{ and }\tilde{w}_i^* = \hat{t}_i^*(1-\hat{t}_i^*). \tag{C.3}$$

Then, we have

$$\hat{\tau}^* = \frac{1}{n}\sum_{i=1}^n \frac{(t_i^* - \hat{t}_i^*)y_i^*}{\tilde{w}_i^*}. \tag{C.4}$$

We use the following identities to handle the random denominator of $\hat{\tau}^*$:

$$\frac{1}{\tilde{w}_i^*} = \frac{1}{w_i^*} + \frac{w_i^* - \tilde{w}_i^*}{w_i^{*2}} + \frac{(w_i^* - \tilde{w}_i^*)^2}{w_i^{*2}\tilde{w}_i^*}. \tag{C.5}$$

Proof of Theorem 2.2. Similar to the proof of Theorem 2.1, define $\tilde{\tau}^*$ and $\bar{\tau}^*$ by

$$\tilde{\tau}^* = \frac{1}{n}\sum_{i=1}^n \frac{(t_i^* - \hat{t}_i^*)y_i^*}{\mu_i^*(1-\mu_i^*)}, \tag{C.6}$$

and ($v_i^* = t_i^* - \mu_i^*$)

$$\bar{\tau}^* \overset{def}{=} \frac{1}{n}\sum_{i=1}^n \frac{(t_i^* - \mu_i^*)y_i^*}{w_i^*} \equiv \frac{1}{n}\sum_{i=1}^n \frac{v_i^* y_i^*}{w_i^*}. \tag{C.7}$$

By adding and subtracting terms in $\sqrt{n}(\hat{\tau}^* - \tau)$, we obtain

$$\begin{aligned} \sqrt{n}(\hat{\tau}^* - \tau) &= \sqrt{n}[(\hat{\tau}^* - \tilde{\tau}^*) + (\tilde{\tau}^* - \bar{\tau}^*) + (\bar{\tau}^* - \tau)] \\ &= J_{1n}^* + J_{2n}^* + J_{3n}^*, \end{aligned} \tag{C.8}$$

where $J_{1n}^* = \sqrt{n}(\hat{\tau}^* - \tilde{\tau}^*), J_{2n}^* = \sqrt{n}(\tilde{\tau}^* - \bar{\tau}^*)$, and $J_{3n}^* = \sqrt{n}(\bar{\tau}^* - \tau)$.

Lemma C.4, as discussed subsequently, gives the leading terms of J_{2n}^*. Recall that $w_i^* = \mu_i^*(1-\mu_i^*)$, and $\tilde{w}_i^* = \hat{t}_i^*(1-\hat{t}_i^*)$. Using (C.5), from (C.4), we obtain

$$\begin{aligned} \hat{\tau}^* &= \frac{1}{n}\sum_{i=1}^n \big[t_i^* - \hat{t}_i^*\big]y_i^*\left[\frac{1}{w_i^*} + \frac{w_i^* - \tilde{w}_i^*}{w_i^{*2}} + \frac{(w_i^* - \tilde{w}_i)^{*2}}{w_i^{*2}\tilde{w}_i^*}\right] \\ &\equiv L_{1n}^* + L_{2n}^* + L_{3n}^*, \end{aligned} \tag{C.9}$$

where $L_{1n}^* = n^{-1}\sum_{i=1}^n[(t_i^* - \hat{t}_i^*)y_i^*]/w_i^*$, $L_{2n}^* = n^{-1}\sum_{i=1}^n [(t_i^* - \hat{t}_i^*)(w_i^* - \tilde{w}_i^*)y_i^*]/w_i^{*2}$, and $L_{3n}^* = n^{-1}\sum_{i=1}^n[(t_i^* - \hat{t}_i^*)(w_i^* - \tilde{w}_i)^{*2}y_i^*]/[w_i^{*2}\tilde{w}_i^*]$.

Note that $L_{1n}^* = \tilde{\tau}^*$, therefore, by (C.9) we have

$$\begin{aligned} J_{1n}^* \equiv \sqrt{n}(\hat{\tau}^* - \tilde{\tau}^*) &= \sqrt{n}(\hat{\tau}^* - L_{1n}^*) \\ &= \sqrt{n}L_{2n}^* + \sqrt{n}L_{3n}^*. \end{aligned} \tag{C.10}$$

Lemma C.3, as discussed subsequently, gives the leading term of J_{1n}^* .Using (C.8) and adding and subtracting terms, we write $J_{3n}^* = \sqrt{n}(\bar{\tau}^* - \tau^*)$ as

$$\begin{aligned} J_{3n}^* &= n^{-1/2}\sum_{i=1}^n \big[v_i^* y_i^*/w_i^* - \tau\big] = n^{-1/2}\sum_{i=1}^n (v_i^* y_i^*/w_i^* - \tau_i^*) \\ &\quad + n^{-1/2}\sum_{i=1}^n (\tau_i^* - \tau) \\ &= n^{-1/2}\sum_{i=1}^n [v_i^*(g_{0i}^* + \tau_i^* t_i^* + u_i^*)/w_i^* - \tau_i^*] \\ &\quad + n^{-1/2}\sum_{i=1}^n (\tau_i^* - \tau). \end{aligned} \tag{C.11}$$

By (C.11), Lemma C.3, and Lemma C.4, we obtain from (C.8) that

$$\begin{aligned} &\sqrt{n}(\hat{\tau}^* - \tau - \hat{B}_{h,\lambda}^*) \\ &\quad = J_{1n}^* + J_{2n}^* - n^{1/2}\hat{B}_{h,\lambda}^* + J_{3n}^* \\ &\quad = n^{-1/2}\sum_{i=1}^n v_i^*[2\mu_i^* - 1]\tau_i^*/w_i^* - n^{-1/2} \\ &\qquad \times \sum_{i=1}^n v_i^*(g_{0i}^* + \tau_i^*\mu_i)^*/w_i^* + n^{-1/2} \\ &\qquad \times \sum_{i=1}^n \{v_i^*(g_{0i}^* + \tau_i^* t_i^* + u_i^*)/w_i^* - \tau_i^*\} + n^{-1/2} \\ &\qquad \times \sum_{i=1}^n (\tau_i^* - \tau) + o_p^*(1) = n^{-1/2} \\ &\qquad \times \sum_{i=1}^n \frac{v_i^* u_i^*}{w_i^*} + n^{-1/2}\sum_{i=1}^n (\tau_i^* - \tau) + o_p^*(1) \\ &\quad \equiv Z_{n2}^* + Z_{n3}^* + o_p^*(1), \end{aligned} \tag{C.12}$$

where $Z^*_{n2} = n^{-1/2}\sum_{i=1}^n v_i^* u_i^*/w_i^*$ and $Z^*_{n3} = n^{-1/2}\sum_{i=1}^n (\tau_i^* - \tau)$.

Note that $E^*(Z^*_{n2}) = n^{-1/2}\sum_i v_i u_i/w_i = Z_{n2}$, and $E^*(Z^*_{n3}) = n^{-1/2}\sum_i(\tau_i - \tau) = Z_{n3}$. Hence, from (C.12), we obtain

$$\begin{aligned} E^*\{\sqrt{n}[\hat\tau^* - \tau - \hat B^*_{h,\lambda}]\} &= Z_{n2} + Z_{n3} + o_p(1) \\ &= \sqrt{n}\big(\hat\tau - \tau - \hat B_{h,\lambda}\big) + o_p(1), \end{aligned} \tag{C.13}$$

where the last equality follows from (A.15).

It can be shown that $\sqrt{n}[\hat B^*_{h,\lambda} - \hat B_{h,\lambda}] = o_p(1)$. This is because $\sqrt{n}[\hat B_{h,\lambda} - B_{h,\lambda}] = o_p(1)$, $\sqrt{n}[\hat B^*_{h,\lambda} - \tilde B_{h,\lambda}] = o_p(1)$, and $\sqrt{n}[\tilde B_{h,\lambda} - B_{h,\lambda}] = o_p(1)$, where $\tilde B_{h,\lambda}$ is defined in Lemma C.4. Then, from (C.13), we immediately have that

$$E^*[\sqrt{n}(\hat\tau^* - \hat\tau - \hat B^*_{h,\lambda})] = E^*(Z^*_{n2} + Z^*_{n3}) + o_p(1) = o_p(1). \tag{C.14}$$

Note that $\mathrm{var}^*(Z^*_{n2}) = \mathrm{var}^*(n^{-1/2}\sum_i u_i^* v_i^*/w_i^*) = n^{-1}\sum_i \mathrm{var}^*(u_i^* v_i^*/w_i^*) = n^{-1}\sum_i[u_i^2 v_i^2/w_i^2] - [n^{-1}\sum_i u_i v_i/w_i]^2 = \hat V_2 + o_p(1)$. Similarly, $\mathrm{var}^*(Z^*_{n3}) = \hat V_3 + o_p(1)$. Hence, we have

$$\begin{aligned} &\mathrm{var}^*(\sqrt{n}(\hat\tau^* - \hat\tau - \hat B^*_{h,\lambda})) \\ &\quad = \mathrm{var}^*(Z^*_{n2} + Z^*_{n3}) + o_p(1) = \hat V_1 + \hat V_2 + o_p(1). \end{aligned} \tag{C.15}$$

Equations (C.14) and (C.15) state that, conditional on the random sample $\mathcal{Z}_n$, $\sqrt{n}(\hat\tau^* - \hat\tau - \hat B^*_{h,\lambda})$ has mean $o_p(1)$ and variance $\hat V_1 + \hat V_2 + o_p(1)$. By taking the limit of $n \to \infty$, we know that $(\hat V_1 + \hat V_2)^{-1/2}\sqrt{n}(\hat\tau^* - \hat\tau - \hat B^*_{h,\lambda})$ has asymptotic mean zero and unit variance. It can also be shown that the conditions of the Liapunov (triangular array) central limit theorem hold for the leading terms of $Z^*_{n2}$and Z^*_{n3}. Hence, we know, for any $z \in \mathbb{R}$, that

$$\Big|\Pr[\sqrt{n}(\hat\tau^* - \hat\tau - \hat B^*_{h,\lambda}) < z|\mathcal{Z}_n] - \Phi(z)\Big| = o_p(1),$$

where $\Phi(\cdot)$ is the CDF for the standard normal distribution. ■

Subsequently, we present some lemmas that are used in proving Theorem 2.2. The same identity is used to handle the random denominator in the kernel estimator; i.e., for any positive integer p, we have

$$\begin{aligned} \frac{1}{\hat f_i^*} &= \frac{1}{f_i^*} + \frac{f_i^* - \hat f_i^*}{f_i^* \hat f_i^*} \\ &= \frac{1}{f_i^*} + \sum_{l=1}^{p}\frac{(f_i^* - \hat f_i^*)^l}{f_i^{*,l+1}} + \frac{(f_i^* - \hat f_i^*)^{p+1}}{f_i^{*,p}\hat f_i^*}. \end{aligned} \tag{C.16}$$

Lemma C.1.

$$L^*_{2n} = n^{-1}\sum\nolimits_i v_i^*(2\mu_i^* - 1)\tau_i^*/w_i^* + o_p^*\Big(n^{-1/2}\Big).$$

Proof. Recall that $w_i^* = \mu_i^*(1 - \mu_i^*)$, $\tilde w_i^* = \hat t_i^*(1 - \hat t_i^*)$, $t_i^* = \mu_i^* + v_i^* v$, and $\hat t_i^* = \hat\mu_i^* + \hat v_i^*$. Then, by exactly the same arguments as we used in proving Lemma A.1 and using Lemma C.6, we have

$$\begin{aligned} L^*_{2n} &= n^{-1}\sum_{i=1}^n y_i^*(t_i^* - \hat t_i^*)\big[\mu_i^* - \hat t_i^* - (\mu_i^{*2} - \hat t_i^{*2})\big]/w_i^{*2} \\ &= n^{-1}\sum_{i=1}^n y_i^*(\mu_i^* - \hat\mu_i^* + v_i^* - \hat v_i^*)(\mu_i^* - \hat\mu_i^* - \hat v_i^*) \\ &\quad \times \big[1 - (\mu_i^* + \hat\mu_i^* + \hat v_i^*)\big]/w_i^{*2} \\ &= -n^{-1}\sum_{i=1}^n y_i^* v_i^* \hat v_i^*\big[1 - 2\mu_i^*\big]/w_i^{*2} + o_p^*(n^{-1/2}) \\ &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n y_i^* v_i^* v_j^*(2\mu_i^* - 1)K^*_{n,ij}/(f_i^* w_i^{*2}) \\ &\quad + o_p^*\big(n^{-1/2}\big) \\ &= \frac{1}{2n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n v_i^* v_j^* \\ &\quad \times\left[\frac{y_i^*(2\mu_i^* - 1)}{f_i^* w_i^{*2}} + \frac{y_j^*(2\mu_j^* - 1)}{f_j^* w_j^{*2}}\right]K^*_{n,ij} + o_p^*(n^{-1/2}) \\ &= \frac{1}{n}\sum_{i=1}^n \tau_i^* v_i^*(2\mu_i^* - 1)/w_i^* + o_p^*\big(n^{-1/2}\big), \end{aligned}$$

where the last equality follows from a u statistic H decomposition. Because for $j \neq i (K_{n,i^*j} = K_n(X_i^*, X_j))$,

$$\begin{aligned} &E^*[v_i^* v_j^* y_j^*(2\mu_j^* - 1)K^*_{n,ij}/(f_i^* w_i^{*2})|z_i^*] \\ &\quad = v_i^*\frac{1}{n}\sum_j v_j y_j(2\mu_j - 1)K_{n,i^*j}/(f_j w_j^2) \\ &\quad = v_i^*\frac{1}{n}\sum_j \tau_j v_j^2(2\mu_j - 1)K_{n,i^*j}/(f_j w_j^2) + (\text{s.o.}) \\ &\quad = v_i^* E\Big(\tau_j v_j^2(2\mu_j - 1)/(w_j^2)|x_j = x_i^*\Big) + (\text{s.o.}) \\ &\quad = v_i^*\tau_i^*(2\mu_i^* - 1)/w_i^* + (\text{s.o.}), \end{aligned}$$

while

$$\begin{aligned} &E^*[v_i^* v_j^* y_i^*(2\mu_i^* - 1)K^*_{n,ij}/(f_i^* w_i^{*2})|z_i^*] \\ &\quad = v_i^* y_i^*(2\mu_i^* - 1)K^*_{n,ij}/(f_i^* w_i^{*2})E^*(v_j^*|z_i^*) \\ &\quad = v_i^* y_i^*(2\mu_i^* - 1)K^*_{n,ij}/(f_i^* w_i^{*2})\left[n^{-1}\sum_j v_j K_{n,i^*j}\right], \end{aligned}$$

which has an order smaller than $v_i^* t_i^*(2m_i^* - 1)/w_i^*$, because $n^{-1}\sum_j v_j K_{n,i^*j} = O_p((nh^{q_1})^{-1})$. ■

Lemma C.2.

$$L^*_{3n} = O_p\big(h^{2\nu} + h^2(nh^{q_1})^{-1}\big) = o_p\big(n^{-1/2}\big).$$

The proof follows from similar arguments as in the proof of Lemma A.2 and is thus omitted here.

Lemma C.3.

$$J^*_{1n} = n^{-1/2}\sum\nolimits_i v_i^*(2\mu_i^* - 1)\tau_i^*/w_i^* + o_p(1).$$

Proof. Note that $J^*_{1n} = \sqrt{n}L^*_{2n} + \sqrt{n}L^*_{3n}$. Lemma C.2 follows from Lemmas C.1 and C.2.

Lemma C.4.

$$J^*_{2n} = \sqrt{n}\hat{B}_{h,\lambda} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_i^*(g^*_{0i} + \tau_i^*\mu_i^*)/w_i^* + o_p(1).$$

Proof. Using $\hat{t}_i^* = \hat{\mu}_i^* + \hat{v}_i^*$, we have $J^*_{2n} = n^{-1/2}\sum_{i=1}^n (\mu_i^* - \hat{t}_i^*)y_i^*/w_i^* = n^{-1/2}\sum_{i=1}^n(\mu_i^* - \hat{\mu}_i^*)y_i^*/w_i^* - n^{-1/2}\sum_{i=1}^n \hat{v}_i^*y_i^*/w_i^* \equiv J^*_{2n,1} - J^*_{2n,2}$.

We consider $J^*_{2n,1}$ first.

$$E^*(J^*_{2n,1}) = n^{-3/2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(\mu_i - \mu_j)y_iK_{n,ij}/w_i = \sqrt{n}\tilde{B}_{h,\lambda},$$

where $\tilde{B}_{h,\lambda} = n^{-1}\sum_{i=1}^n\sum_{j\neq i}^n(\mu_i - \mu_j)y_iK_{n,ij}/w_i$. It is easy to show that $E^*[(J^*_{2n,1})^2] = o_p(1)$. Hence, $J^*_{2n,1} = \sqrt{n}\tilde{B}_{h,\lambda} + (\text{s.o.})$.

Next,

$$\begin{aligned} J^*_{2n,2} &= n^{-3/2}\sum_{i=1}^{n}\sum_{j\neq i} v_j^*y_i^*K^*_{n,ij}/(f_i^*w_i^*) + (\text{s.o.}) \\ &= (1/2)n^{-3/2}\sum_i\sum_{j\neq i}\left[\frac{y_i^*v_j^*}{w_i^*f_i^*} + \frac{y_j^*v_i^*}{w_j^*f_j^*}\right]K^*_{n,ij} \qquad \text{(C.17)} \\ &= -\frac{1}{\sqrt{n}}\sum_i v_i^*\left(g^*_{0i} + \tau_i^*\mu_i^*\right)/w_i^* + (\text{s.o.}) \end{aligned}$$

by the u statistic H decomposition, because

$$\begin{aligned} E^*[y_j^*v_i^*K^*_{n,ij}/(w_j^*f_j^*)|z_i^*] &= v_i^*\left[n^{-1}\sum_{j\neq i}(y_j/f_jw_j)K_{n,i^*j}\right] \\ &= v_i^*E\left[y_j/w_j|x_j = x_i^*\right] + (\text{s.o.}) \\ &= v_i^*E\left[(g_{0j} + \tau_jg_{1j})/w_j|x_j = x_i^*\right] \\ &\quad + (\text{s.o.}) \\ &= v_i^*(g^*_{0i} + \tau_i^*g^*_{1i})/w_i^* + (\text{s.o.}), \end{aligned}$$

where we have used $E[y_j/w_j|x_j = x_i^*] = (g^*_{0i} + t_i^*g^*_{1i})/w_i^*$, because $E(v_j|x_j) = 0$ and $E(u_j|x_j) = 0$.

Also, $E^*[y_i^*v_j^*K^*_{n,ij}/(w_i^*f_i^*)|z_i^*] = (y_i^*/f_i^*w_i^*)[n^{-1}\sum_{j\neq i}v_jK_{n,i^*j}]$, which has an order smaller than $v_i^*(g^*_{0i} + \tau^*_{1i})/w_i^*$ because $n^{-1}\sum_{j\neq i}v_jK_{n,i^*j} = O_p\left((nh^{q_1})^{-1}\right)$. ■

Lemma C.5.

$$\begin{aligned} A^*_{1n} &\overset{def}{=} n^{-1}\sum_i(\mu_i^* - \hat{\mu}_i^*)y_i^*/w_i^* \\ &= \tilde{B}_{h,\lambda} + O_p(n^{-1/2}h^\nu + h\left(n^2h^{q_1}\right)^{-1/2}), \end{aligned}$$

where the definition of $\tilde{B}_{h,\lambda} = (1/n^2)\sum_i\sum_j(\mu_i - \mu_j)K_{n,ij}y_i/(f_iw_i)$is given in the subsequent proof.

Proof. Using (C.16), we know that $A^*_{1n} = A^*_{1n,1} +$ (s.o.), where $A^*_{1n,1} = n^{-1}\sum_i(\mu_i^* - \hat{\mu}_i^*)\hat{f}_i^*y_i^*/(w_i^*f_i^*)$.

$$\begin{aligned} E^*(A^*_{1n,1}) &= \frac{1}{n^2}\sum_i\sum_j(\mu_i - \mu_j)K_{n,ij}y_i/(f_iw_i) \\ &= \tilde{B}_{h,\lambda}. \qquad \text{(C.18)} \end{aligned}$$

By similar arguments, as in the derivation of (B.8), one can easily show that

$$\text{var}^*(A^*_{1n,1}) = O_p(n^{-1}h^{2\nu} + h^2\left(n^2h^{q_1}\right)^{-1}). \qquad \text{(C.19)}$$

Hence,

$$\begin{aligned} A^*_{1n,1} &= E(A^*_{1n,1}) + O_p\left(\sqrt{\text{var}(A^*_{1n,1})}\right) \\ &= \tilde{B}_{h,\lambda} + O_p(n^{-1/2}h^\nu + h\left(n^2h^{q_1}\right)^{-1/2}) \end{aligned}.$$

Lemma C.6.

(i) $A^*_{2n} \overset{def}{=} n^{-1}\sum_i(\hat{\mu}_i^* - \mu_i^*)^2 = O_p(h^{2\nu} + h^2(nh^{q_1})^{-1})$.

(ii) $A^*_{3n} \overset{def}{=} n^{-1}\sum_i \hat{v}_i^{*2} = O_p((nh^{q_1})^{-1})$.

(iii) $A^*_{4n} \overset{def}{=} n^{-1}\sum_i(\hat{\mu}_i^* - \mu_i^*)\hat{v}_i = O_p(h^{2\nu} + (nh^{q_1})^{-1})$.

(iv) $A_{5n} \overset{def}{=} n^{-1}\sum_i(\hat{t}_i^* - \mu_i^*)^2 = O_p(h^{2\nu} + (nh^{q_1})^{-1})$.

Proof of (i). Using (C.16), we have $A^*_{2n} \equiv n^{-1}\sum_i(\hat{\mu}_i^* - \mu_i^*)^2\hat{f}_i^{*2}/\hat{f}_i^{*2} = n^{-1}\sum_i(\hat{\mu}_i^* - \mu_i^*)^2.\hat{f}_i^{*2}/f_i^{*2} + (\text{s.o.})$ Also, since $f(\cdot)$ is bounded below by a positive constant, the leading term of A^*_{2n} is given by $A^*_{2n,1} \overset{\text{def}}{=} n^{-1}\sum_i(\hat{\mu}_i^* - \mu_i^*)^2\hat{f}_i^{*2}$. Hence

$$\begin{aligned} E^*\left[\left|A^*_{2n,1}\right|\right] &= E^*[(\hat{\mu}_i^* - \mu_i^*)^2\hat{f}_i^{*2}] \\ &= \frac{1}{n(n-1)^2}\sum_{i\neq l}^{n}\sum_{j\neq l}^{n}\sum_{l=1}^{n}[(\mu_i - \mu_l) \\ &\quad \times K_{n,il}(\mu_j - \mu_l)K_{n,jl}] \\ &= \frac{1}{n(n-1)^2}\sum_{i\neq l}^{n}\sum_{l=1}^{n}(\mu_i - \mu_1)^2K^2_{n,i1} \\ &\quad + \frac{1}{n(n-1)^2}\sum_{i\neq l}^{n}\sum_{j\neq i,l}^{n}\sum_{l=1}^{n}(\mu_i - \mu_1) \\ &\quad \times(\mu_j - \mu_l)K_{n,il}K_{n,jl} \\ &= O_p\left(h^2(nh^{q_1})^{-1} + h^{2\nu}\right) \end{aligned}$$

by Lemma B.3(i).

Proof of (ii). Similarly, by (C.16), we have $A^*_{3n} \equiv n^{-1}\sum_{i=1}^n \hat{v}_i^{*2}\hat{f}_i^{*2}/\hat{f}_i^{*2} = n^{-1}\sum_{i=1}^n \hat{v}_i^{*2}\hat{f}_i^{*2}/f_i^{*2} + (\text{s.o.}) \equiv A^*_{3n,1} + (\text{s.o.})$. Hence,

$$\begin{aligned} E^*[|A^*_{3n,1}|] &= E^*\left[\hat{v}_i^{*2}\hat{f}_i^{*2}\right] \\ &= \frac{1}{n(n-1)^2}\sum_{i\neq l}^{n}\sum_{j\neq l}^{n}\sum_{l=1}^{n}v_iv_jK_{n,il}K_{n,jl} = O_p((nh^{q_1})^{-1}). \end{aligned}$$

Proof of (iii). (iii) follows from (i) and (ii) and the Cauchy inequality.

Proof of (iv). Finally, (vi) follows from (i)–(iii), because $(\hat{t}_i^* - \mu_i^*)^2 = (\hat{\mu}_i^* - \mu_i^*)^2 + \hat{v}_i^{*2} + 2(\hat{\mu}_i^* - \mu_i^*)\hat{v}_i^*$.

ACKNOWLEDGMENTS

Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University. J.S. Racine gratefully acknowledges support from the Social Sciences and Humanities

Research Council of Canada (SSHRC:*www.sshrc.ca*) and the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca). We gratefully acknowledge comments from the editors and two anonymous referees that led to a much improved version of this paper. All errors remain, naturally, our own.

[Received June 2006. Revised April 2007]

REFERENCES

Aitchison, J., and Aitken, C. G. G. (1976), "Multivariate Binary Discrimination by the Kernel Method," *Biometrika,* 63, 413–420.

Cai, Z., Fan, J., and Li, R. (2000), "Efficient Estimation and Inferences for Varying-Coefficient Models," *Journal of the American Statistical Association,* 95, 888–902.

Cai, Z., Fan, J., and Yao, Q. (2000), "Functional-Coefficient Regression Models for Nonlinear Time Series," *Journal of the American Statistical Association,* 95, 941–956.

Chen, R., and Tsay, R. S. (1993), "Functional-Coefficient Autoregressive Models," *Journal of the American Statistical Association,* 88, 298–308.

Connors, A. F., Speroff, T., Dawson, N. V., Thomas, C., Harrell, F., Wagner, D., Desbiens, N., Goldman, L., Wu, A., Califf, R. M., Fulkerson, W. J., Vidaillet, H., Broste, S., Bellamy, P., Lynn, J., and Knaus, W. A. (1996) "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients," *Journal of the American Medical Association,* 11, 889–897.

Hahn, J. (1998), "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects." *Econometrica,* 66, 315–331.

Hall, P., Li, Q., and Racine, J. (2007), "Nonparametric Estimation of Regression Functions in the Presence of Irrelevent Regressors." *The Review of Economics and Statistics,* 89, 784–789.

Hall, P., Racine, J., and Li, Q. (2004), "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association,* 99, 1015–1026.

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using The Estimated Propensity Score," *Econometrica,* 71, 1161–1189.

Horvitz, D., and Thomson, D. (1952), "A Generalization of Sampling without Replacement from a Finite Population," *Journal of the American Statistical Association,* 47, 663–685.

Hsiao, C., Li, Q., and Racine, J. (2007), "A Consistent Model Specification Test with Mixed Categorical and Continuous Data," *Journal of Econometrics,* 410, 802–826.

Ichimura, H. (2000). "Asymptotic Distribution of Non-Parametric and Semiparametric Estimators with Data Dependent Smoothing Parameters," Unpublished Manuscript.

Ichino, A. and Winter-Ebmer, R. (1998). "The Long-Run Educational Cost of World War II: an Example of Local Average Treatment Effect," CEPR Publication DP1895.

Lechner, M. (1999), "Earnings and Employment Effects of Continuous Off-the-Job Training in East Germany after Unification," *Journal of Business & Economic Statistics,* 17, 74–90.

Li, Q., Huang, C. J., Li, D., and Fu, T. T. (2002), "Semiparametric Smooth Coefficient Models," *Journal of Business & Economic Statistics,* 20, 412–422.

Li, Q., and Zhou, J. X. (2005), "The Uniqueness of Cross-Validation Selected Smoothing Parameters in Kernel Estimation of Nonparametric Models," *Econometric Theory,* 21, 1017–1025.

Lin, D. Y., Pasty, B. M., and Kronmal, R. A. (1988), "Assessing the Sensitivity of Regression Results to Unmeasured Confounders in Observational Studies," *Biometrics,* 54, 948–963.

Robinson, P. (1988), "Root-N Consistent Semiparametric Regression," *Econometrica,* 56, 931–954.

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika,* 70, 41–55.

Tucker, H. (1967): *A Graduate Course in Probability*, New York: Academic Press.