Efficient Estimation of Average Treatment Effects With Mixed Categorical and Continuous Data [*]

Qi Li
Department of Economics
Texas A&M University
Bush Academic Building West
College Station, TX 77843-4228

Jeff Racine
Department of Economics &
Center for Policy Research
Syracuse University
Syracuse, NY 13244-1020

Jeff Wooldridge
Department of Economics
Michigan State University
East Lansing, MI 48824-1038

June 18, 2004

Abstract

In this paper 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 to select the smoothing parameters, has the amazing ability of automatically removing 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 existing kernel approaches 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 and phrases: Average Treatment, Discrete Covariates, Kernel Smoothing, Bootstrap, Asymptotic normality.

[*] Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine's research is supported by NSF grant BCS 0320284.

1 Introduction

The measurement of average treatment effects, 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 towards the measurement of treatment effects involves estimation of a 'propensity score'. Estimation of the propensity score (conditional probability of receiving treatment) was originally undertaken with parametric index models such as the Logit or Probit, though there is an expanding literature on the semiparametric and nonparametric estimation thereof (see Hahn (1998) and Hirano et al. (2002)). The advantage of nonparametric approaches in this setting is rather obvious, as 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. Though the appeal of robust nonparametric methods is obvious in this setting, existing nonparametric approaches split the sample into 'cells' in the presence of categorical covariates, resulting in a loss of efficiency. Given that datasets used to assess treatment effects frequently contain a preponderance of categorical data,[1] it is not uncommon that the number of discrete cells are larger than the sample sizes. In such cases the sample splitting frequency method become infeasible. On the other hand, the kernel-based smoothing cross-validation method can (asymptotically) automatically detect and remove irrelevant covariates,[2] leading to a feasible and accurate nonparametric estimation of the average treatment effects. Another obvious symptom of the sample-splitting frequency method would be a loss in power of *tests* of whether a treatment effect differs from that of no effect.

In this paper 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 datatypes. 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 the least-squares conditional cross-validation method to select smoothing parameters for both the categorical and continuous variables proposed by Hall et al. (2004a), who demonstrate that cross-validation produces asymptotically optimal smoothing for relevant components, while it eliminates irrelevant components by oversmoothing. Indeed, in the problem of nonparametric estimation of a conditional density with mixed categorical and continuous data, cross-validation comes into its own as a method with no obvious peers.

The rest of the paper 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, which demonstrate that the proposed method is capable of outperforming existing kernel approaches that require splitting the sample into subsamples ('discrete cells'). 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.

[1] In the typical medical study, it is common to encounter exclusively categorical data types.

[2] It is not clear to us how to extend this property of kernel smoothing, which automatically removes irrelevant covariates, to nonparametric series methods.

2 The Model

For what follows, we use a dummy variable, $t_i \in \{0,1\}$, to indicate whether or not an individual has received treatment. We let $t_i = 1$ for the treated, 0 for the untreated. Letting $y_i(t_i)$ denote the outcome, then, for $i = 1, \dots, n$, 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 pre-treatment 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)** (Unfoundedness):

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 1 one can easily show that

$$\tau(x) = E[y_i|t_i = 1, x_i = x] - E[y_i|t_i = 0, x_i = x]. \tag{2.3}$$

The two terms on the right-hand side of (2.3) can be estimated consistently by any nonparametric estimation technique. Therefore, under **A1**, the average treatment effects can be obtained via simple averaging over $\tau(x)$.

$$\tau = E[\tau(x_i)]. \tag{2.4}$$

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

$$y_i = g(x_i, t_i) + u_i, \tag{2.5}$$

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 re-write (2.5) as

$$\begin{aligned} y_i &= g_0(x_i) + [g_1(x_i) - g_0(x_i)]t_i + u_i \\ &= g_0(x_i) + \tau(x_i)t_i + u_i, \end{aligned} \tag{2.6}$$

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

From (2.6) it is easy to show that $\tau(x_i) = cov(y_i, t_i|x_i)/var(t_i|x_i)$. Letting $\mu(x_i) = Pr(t_1 = 1|x_i) \equiv E(t_i|x_i)$ (because $t_i = \{0,1\}$), we may write

$$\tau = E[\tau(x_i)] = E\left\{\frac{(t_i - \mu_i)y_i}{var(t_i|x_i)}\right\}. \tag{2.7}$$

We now turn to the discussion of the nonparametric estimation of τ based on (2.7).

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 $x_i^c \in R^q$ 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 by $W_h(x_i^c, x_j^c) = \prod_{s=1}^q h_s^{-1} w\left(\frac{x_{is}^c - x_{js}^c}{h_s}\right)$, where x_{is}^c is the sth component of x_i^c.

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 $\tilde{x}_i^d$ denote a r_1-vector (say, the first r_1 components of x_i^d, $0 \leq r_1 \leq r$) of discrete covariates that have a natural ordering ($0 \leq r_1 \leq r$), and let $\bar{x}_i^d$ denote the remaining $r_2 = r - r_1$ discrete covariates that do not have a natural ordering. We use x_{it}^d to denote the tth component of x_i^d ($t = 1, \dots, r$).

For an ordered variable, we use the Habbema kernel:

$$\tilde{l}(\tilde{x}_{it}^d, \tilde{x}_{jt}^d, \lambda_t) = \begin{cases} 1, & \text{if } \tilde{x}_{it}^d = \tilde{x}_{jt}^d, \\ \lambda_t^{|\tilde{x}_{it}^d - \tilde{x}_{jt}^d|}, & \text{if } \tilde{x}_{it}^d \neq \tilde{x}_{jt}^d. \end{cases} \tag{2.8}$$

When $\lambda_t = 0$ ($\lambda_t \in [0,1]$), $l(\tilde{x}_{it}^d, \tilde{x}_{jt}^d, \lambda_t = 0)$ becomes an indicator function, and when $\lambda_t = 1$, $l(\tilde{x}_{it}^d, \tilde{x}_{jt}^d, \lambda_t = 1) = 1$ becomes a uniform weight function.

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

$$\bar{l}(\bar{x}_{it}^d, \bar{x}_{jt}^d) = \begin{cases} 1, & \text{if } \bar{x}_{it}^d = \bar{x}_{jt}^d, \\ \lambda_t, & \text{otherwise.} \end{cases} \tag{2.9}$$

Again $\lambda_t = 0$ leads to an indicator function, $\lambda_t = 1$ to a uniform weight function.

Let $\mathbf{1}(A)$ denote an indicator function that assumes the value 1 if A occurs and 0 otherwise. Combining (2.8) and (2.9), we obtain the product kernel function given by

$$L(x_i^d, x_j^d, \lambda) = \left[\prod_{t=1}^{r_1} \lambda_t^{|\tilde{x}_{it}^d - \tilde{x}_{jt}^d|}\right] \left[\prod_{t=r_1+1}^{r} \lambda_t^{\mathbf{1}(\bar{x}_{it}^d \neq \bar{x}_{jt}^d)}\right]. \tag{2.10}$$

We note that there does not exist a plug-in or even an ad-hoc formula for selecting λ_t in this setting. Hence we recommend using least squares cross-validation for selecting λ_t ($t = 1, \dots, r$). 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 (see Hall et al. (2004a,b)). 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 paper. We let $\hat{t}(x_i)$ be the nonparametric estimator of $\mu_i \equiv \mu(x_i)$ defined by

$$\hat{t}(x_i) = \frac{\sum_{j=1}^n t_j K_{n,ij}}{\sum_{j=1}^n K_{n,ij}}, \tag{2.11}$$

where $K_{n,ij} = W_h(x_i^c, x_j^c) L(x_i^d, x_j^d, \lambda)$. By noting that $var(t_i|x_i) = \mu_i(1 - \mu_i)$, one can estimate the average treatment effects by

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \frac{(t_i - \hat{t}(x_i))y_i M_{ni}}{\hat{t}(x_i)(1-\hat{t}(x_i))} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i y_i}{\hat{t}(x_i)} - \frac{(1-t_i)y_i}{1-\hat{t}(x_i)}\right] M_{ni}, \tag{2.12}$$

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

With the exception of the presence of the trimming function M_{ni}, (2.12) is exactly the same as the propensity score based estimator considered by Hirano et al. (2002), who used series methods for estimating $\mu(x_i)$. As we mentioned earlier, it is not unclear how to use nonparametric series methods to automatically remove irrelevant discrete covariates. Therefore, in this paper we will consider only the kernel-based estimation method, which has the advantage of automatically removing irrelevant covariates (be they continuous or discrete).

To derive the asymptotic distribution of $\hat{\tau}$ we make the following regularity assumptions. 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) 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, $\mu(x^c, x^d) \in \mathcal{G}_\nu^4$, $f(x^c, x^d) \in \mathcal{G}_{\nu-1}^4$, where $\nu \geq 2$ and $\nu > q-2$ is a positive integer. (iii) $\inf_{x\in\mathcal{S}} f(x) \geq \eta$ for some $\eta > 0$, where $\mathcal{S}$ is the support of x_i. (iv) $\sigma^2(x,t) = var(u_i|x_i = x, t_i = t)$ 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 $n \to \infty$) $\mathbf{1}(x \in \mathcal{S})$, where $\mathbf{1}(\cdot)$ is the usual indicator function, and $\mathcal{S}$ is the support of $f(x)$.

Assumption **(A3):** (i) $w(\cdot)$ is νth order kernel; it is bounded, symmetric and differentiable up to order ν . (ii) As $n \to \infty$, $n\sum_{s=1}^q h_s^{2\nu+4} \to 0$, and $nh_1^2 \dots h_q^2 \to \infty$.

Assumptions (A2) (i)-(iv) are standard smoothness and moment conditions. (A2) (v) implies that, asymptotically, we only trim a negligible amount of data (near the boundary) so that $\hat{\tau}$ is asymptotically efficient (see Theorem 2.1). A trimming set is used in (2.12) for theoretical reasons. Given that the support of x is a compact and 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 $A_{\delta_n} = \prod_{s=1}^q [-\delta_s, \delta_s]$, where $\delta_s = \delta_{sn} < 1$ converges to 0 as $n \to \infty$. To avoid boundary bias, one can choose $\delta_s = O\left(h_s^\alpha\right)$ for some $0 < \alpha < 1$, and define $M_n(x_i) = \mathbf{1}(x_i \in A_{\delta_n})$. 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.

In order to appreciate the restrictions imposed by (A3), let us assume that $h_s = h$ for all s's. In this case, (A3) (ii) requires that $\nu + 2 > q$. Using a second order kernel ($\nu = 2$) implies that $q < 4$ or $q \leq 3$ since q is a positive integer. Thus, a second order kernel can satisfy (A3) if $q \leq 3$. When $q \geq 4$ (A3) requires the use of a higher order kernel function.

Remark 2.1 *If $1 \leq q \leq 3$ and one uses a second order kernel ($\nu = 2$), then (A3) allows optimal smoothing. To see this, note that when $\nu = 2$, the optimal smoothing is $h_s = O\left(n^{-1/(4+q)}\right)$. (A3) (ii) becomes (assuming $h_s = h$) $nh^8 \to 0$ and $nh^{2q} \to \infty$; optimal smoothing $h \sim n^{-1/(4+q)}$ satisfies these conditions for $q = 1, 2, 3$. For $q > 3$, (A3) (ii) rules out optimal smoothing.*

We will choose the smoothing parameters *based on*, but not the same as (if $q \geq 4$), the least squares cross-validation method. The leave-one-out kernel estimator of $E(y_i|x_i) = \mu(x_i)$ is given by

$$\hat{t}_{-i}(x_i) = \frac{\sum_{j\neq i}^{n} t_j K_{n,ij}}{\sum_{j\neq i}^{n} K_{n,ij}}, \tag{2.13}$$

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

$$CV(h, \lambda) = \frac{1}{n}\sum_{i=1}^{n} \left[t_i - \hat{t}_{-i}(x_i)\right]^2 S_n(x_i), \tag{2.14}$$

where $S_n(\cdot)$ is a weight function that trims out observations near the boundary of the support of x_i (avoiding the excessive boundary bias).

Hall et al (2004a,b) 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. Given that the cross-validation method can automatically remove the irrelevant covariates, we will derive the asymptotic distribution of $\hat{\tau}$ under the condition that all the q continuous variables and the r discrete variables are relevant ones, i.e., assuming that the irrelevant covariates have already been removed when computing $\hat{\tau}$. When all the covariates are the relevant ones, we request that h_ss and λ_ss are chosen from some shrinking set, $h_s \in (0, \eta_n]$ $(s = 1, ..., q)$ and $\lambda_s \in (0, \eta_n]$ $(s = 1, ..., r)$, where $\eta_n \to 0$ as $n \to \infty$ at a rate slower than any inverse polynomial of n. This assumption can be relaxed as in Hall et al (2004a,b).

We let $(\bar{h}_1, \ldots, \bar{h}_q, \bar{\lambda}_1, \ldots, \bar{\lambda}_q)$ denote the cross-validation choices of $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_q)$ that minimize (2.14). It is well known that the cross-validation method leads to optimal smoothing, but our assumption (A2) (ii) rules out optimal smoothing when $q > 3$. Therefore, we suggest using $\hat{h}_s = \bar{h}_s n^{1/(2\nu+q)} n^{-1/(q+\nu+2)}$ and $\hat{\lambda}_s = \bar{\lambda}_s n^{\nu/(2\nu+q)} n^{-\nu/(q+\nu+2)}$. Thus we have $\hat{h}_s \sim n^{-1/(q+\nu+2)}$ and $\hat{\lambda}_s \sim n^{-\nu/(q+\nu+2)}$, satisfying (A2) (ii). When $1 \leq q \leq 3$, we know that we can choose $\nu = 2$ (second order kernel), which leads to $\hat{h}_s \equiv \bar{h}_s$ and $\hat{\lambda}_s \equiv \bar{\lambda}_s$.

Following the proofs of Hall et al. (2004a), one can show the following:

Lemma 2.2 *Under the assumptions (A1) to (A3), and the condition of (A.6) given in the Appendix A, we have*

$$\begin{aligned} \hat{h}_s &= a_s^0 n^{-1/(\nu+q+2)} + o_p\left(n^{-1/(\nu+q+2)}\right), & \text{for } s = 1, ..., q; \\ \hat{\lambda}_s &= b_s^0 n^{-2/(\nu+q+2)} + o_p\left(n^{-2/(\nu+q+2)}\right), & \text{for } s = 1, ..., r. \end{aligned}$$

where a_s^0s are finite positive constants, and b_s^0s are non-negative finite constants.

The proof of Lemma 2.2, as well as consistent estimators of V_1, V_2 and $B_{h,\lambda}$, is given in Appendix A.

The empirical applications and simulation results presented in Hall et al. (2004a,b) reveal that nonparametric estimation based on cross-validated bandwidth selection performs much better than a conventional frequency estimator (which corresponds to $\lambda_s = 0$ for all $s = 1, \ldots, r$) because the former does not split the sample in finite-sample applications, which creates efficiency losses.

Having obtained the $\hat{h}_s$s and $\hat{\lambda}_s$s based on the cross-validation method, we estimate τ using expression (2.12) with $\hat{t}(x_i)$ computed using $\hat{h}_s$s and $\hat{\lambda}_s$s. To avoid introducing too many notations, we will still use $\hat{\tau}$ to denote the resulting estimator of τ.

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

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 $B_{h,\lambda} = \sum_{s=1}^{q} B_{1s}(x)\hat{h}_s^{\nu} - \sum_{s=1}^{r} B_{2s}(x)\hat{\lambda}_s$*,* $B_{1s}(x)$ *and* $B_{2s}(x)$ *are defined in lemma B.2 of Appendix B,* $V_1 = var(\tau(x_i))$*,* $V_2 = E\left\{\sigma^2(x_i,t_i)(t_i-\mu_i)^2/[\mu_i^2(1-\mu_i)^2]\right\}$*, 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.

Theorem 2.1 shows that our kernel-based estimator of $\hat{\tau}$ is semiparametrically efficient. 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 $\int dx = \sum_{x^d}\int dx^c$, $\mu_x = \mu(x)$, and 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

$$\begin{aligned}
V_2 &= E\left\{\sigma^2(x_i)(t_i-\mu_i)^2/\left[\mu_i^2(1-\mu_i)^2\right]\right\} \\
&= \sum_{t=1,0}\int f_x(x)p(t|x)\left\{\sigma^2(x,t)(t_i-\mu_x)^2/\left[\mu_x^2(1-\mu_x)^2\right]\right\}dx \\
&= \int \frac{f_x(x)\mu_x\sigma^2(x,1)(1-\mu_x)^2}{\mu_x^2(1-\mu_x)^2}dx + \int \frac{f_x(x)(1-\mu_x)^2\sigma^2(x,0)}{\mu_x^2(1-\mu_x)^2}dx \\
&= E\left\{\frac{\sigma^2(x_i,1)}{\mu_i} + \frac{\sigma^2(x_i,0)}{1-\mu_i}\right\}.
\end{aligned} \tag{2.16}$$

Equation (2.16) matches the expression given in Hahn (1998). Thus, $V_1 + V_2$ coincides with the semiparametric efficient bound for this model.

Hirano et al. (2002) consider the problem of estimating average treatment effects using series estimation methods, and they observe that if one uses the true $cov(t_i|x_i) = \mu_i(1-\mu_i)$ to replace the estimated covariance $\hat{t}_i(1-\hat{t}_i)$ 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 2.3 *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.,*

$$\tilde{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{(t_i-\hat{t}_i)y_iM_{ni}}{\mu_i(1-\mu_i)}, \tag{2.17}$$

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 that 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 2.3 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 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 2.3. In Appendix A we also show that $\sqrt{n}(\hat{\tau} - \tilde{\tau}) = -Z_{n3} + o_p(1)$. Hence,

$$\begin{aligned} \sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda}) &= \sqrt{n}(\hat{\tau} - \tilde{\tau}) + \sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}) \\ &= \{-Z_{n3} + o_p(1)\} + \{Z_{n1} + Z_{n2} + Z_{n3} + o_p(1)\} \\ &= Z_{n1} + Z_{n2} + o_p(1) \to N(0, V_1 + V_2) \text{ in distribution,} \end{aligned} \tag{2.21}$$

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 that using an estimated variance $\hat{v}ar(t_i|x_i)$ 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 $\hat{\tau}$ since $\sqrt{n}(\tilde{\tau} - \tau)$ is asymptotically normal with zero mean 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.

An alternative estimator for τ

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

Note that (2.6) can also be viewed as a functional coefficient model (smooth coefficient model) as considered by Chen and Tsay (1993), Cai, Fan and Yao (2000), Cai, Fan and Li (2000), and Li et al. (2002), among others. 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

$$\begin{pmatrix} \hat{g}_0(x_i) \\ \hat{\tau}_n(x_i) \end{pmatrix} = \left[n^{-1} \sum_{j \neq i}^{n} \begin{pmatrix} 1 \\ t_j \end{pmatrix} (1, t_j) W_{h,ij} L_{\hat{\lambda},ij} \right]^{-1} \left[n^{-1} \sum_{j \neq i}^{n} \begin{pmatrix} 1 \\ t_j \end{pmatrix} y_j W_{h,ij} L_{\hat{\lambda},ij} \right], \tag{2.22}$$

where $W_{h,ij} = W_h(x_j^c, x_i^c)$ and $L_{\hat{\lambda},ij} = L(x_i^d, x_j^d, \hat{\lambda})$. (2.22) gives consistent estimator of $g_0(x_i)$ and $\tan(x_i)$. For example, the resulting estimate of $\tau(x_i)$ is given by

$$\hat{\tau}_n(x_i) = \frac{\hat{E}(y_i t_i|x_i) - \hat{E}(y_i|x_i)\hat{E}(t_i|x_i)}{\hat{t}(x_i)(1 - \hat{t}(x_i))}, \tag{2.23}$$

where $\hat{E}(y_i t_i|x_i) = n^{-1} \sum_{j=1}^{n} t_j y_j K_{n,ij} / \hat{f}(x_i)$, $\hat{E}(y_i|x_i) = n^{-1} \sum_{j=1}^{n} y_j K_{h,ij} / \hat{f}(x_i)$, $\hat{t}(x_i) = n^{-1} \sum_{j=1}^{n} t_j K_{n,ij} / \hat{f}(x_i)$, and $\hat{f}(x_i) = n^{-1} \sum_{j=1}^{n} K_{n,ij}$.

From (2.23) we readily obtain a consistent estimator of $E(y_i|x_i)$ by $\hat{g}_0(x_i) + \hat{\tau}_n(x_i)\, t_i$. One can also estimate τ by $\hat{\tau}_n = n^{-1} \sum_{i=1}^n \hat{\tau}_n(x_i)$.

2.3 Testing no Effect Based on Bootstrapping

The result of Theorem 2.1 can be used to test the null hypothesis of 'no effect' ($\tau = 0$) of a treatment. We know that nonparametric estimation results can be sensitive to the choice of smoothing parameters. Therefore, test results based on asymptotic distributions may also be sensitive to the selection of smoothing parameters. In order to obtain a more robust test, we suggest using resampling methods to approximate the null distribution of the test statistic.

Below we present two bootstrap procedures. The first procedure does not involve any null hypothesis and is useful for the construction of error bounds, while the second procedure imposes the null hypothesis of no treatment effect.

Let $z_i \equiv \{y_i, t_i, x_i^c, x_i^d\}_{i=1}^n$, the vector of realizations on the outcome, treatment, and conditioning information. We wish to construct the sampling distribution of $\hat{\tau}$, and do so with the following resampling procedure.

1. 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 now wish to construct the sampling distribution of $\hat{\tau}$ under the null of no treatment effect, and do so with the following bootstrap procedure.

1. Randomly select from $\{y_j\}_{j=1}^n$ from $\{z_j\}_{j=1}^n$ with replacement and call this $\{y_j^*\}_{j=1}^n$. Next, call $\{z_i^*\}_{i=1}^n \equiv \{y_j^*, t_j, x_j^c, x_j^d\}_{j=1}^n$ the bootstrap sample. Note that we have broken any systematic relationships between the outcome and covariates, thereby imposing the null of no treatment effect on the sample $\{z_i^*\}_{i=1}^n$.

2. Use the bootstrap sample to compute the bootstrap statistic $\hat{\tau}^*$ using the same cross-validated smoothing parameters as were used for $\hat{\tau}$ ($\hat{\tau}_n$).

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 $\hat{\tau}$ under the null.

4. A test of the null of no treatment follows directly. Let $\{\hat{\tau}*_j\}_{j=1}^B$ be the ordered (in an ascending order) statistic of the B bootstrap statistics, and let $\hat{\hat{\tau}}_\alpha^*$ denote the αth percentile of $\{\hat{\tau}_j^*\}_{j=1}^B$. We reject H_0 if $\hat{\tau} > \hat{\tau}_\alpha^*$ at the level α.

Let $\hat{V}_1$, $\hat{V}_2$ and $\hat{B}_{h,\lambda}$ denote some consistent estimators of V_1, V_2 and $B_{h,\lambda}$, respectively (say, as given in Appendix A), and let $\hat{V}_1^*$, $\hat{V}_2^*$ and $\hat{B}_{h,\lambda}^*$, denote their bootstrap counterparts, obtained by replacing (y_i, x_i, t_i) with (y_i^*, x_i^*, t_i^*) in $\hat{V}_1$, $\hat{V}_2$ and $\hat{B}_{h,\lambda}$, respectively. Note that the bootstrap counterpart quantities use the same smoothing parameters (they do not require re-cross-validation). The following theorem shows that the bootstrap method works.

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

$$T_n^*|\{x_i, t_i, y_i\}_{i=1}^n \to N(0,1)$$

in distribution in probability.[3]

The proof of Theorem 2.2 is similar to the proof of Theorem 2.1 and is thus omitted here.

Let $\hat{T}_n = \sqrt{n}(\hat{\tau} - \hat{B}_{h,\lambda})/\sqrt{\hat{V}_1 + \hat{V}_2}$. Then under H_0 both $\hat{T}_n$ and $\hat{T}_n^*$ have asymptotic standard normal distributions. Compare the differences between the numerators of the two test statistics: $\sqrt{n}(\hat{\tau} - \hat{\tau}^*) + \sqrt{n}(B_{h,\lambda}^* - B_{h,\lambda}) = \sqrt{n}(\hat{\tau} - \hat{\tau}^*) + \sqrt{n}O_p\left(\sum_{s=1}^q h_s^{\nu+2}\right) = \sqrt{n}\left(\hat{\tau} - \hat{\tau}^*\right) + o_p(1)$ because both $\hat{B}_{h,\lambda}$ and $\hat{B}_{h,\lambda}^*$ are of order $O\left(\sum_{s=1}^q h_s^{\nu}\right)$ and their differences are of smaller order $\hat{B}_{h,\lambda} - \hat{B}_{h,\lambda}^* = O_p\left(\sum_{s=1}^q h_s^{\nu+2}\right)$. Therefore, when one uses bootstrap procedures to conduct the test, one does not need to compute $B_{h,\lambda}$ (and $B_{h,\lambda}^*$). This is yet another advantage of using bootstrap procedures in this setting.

3 Simulations

In this section we report simulations designed to examine the finite sample performance of the proposed methods. We highlight performance in mixed data settings, a feature that existing methods do not handle well, and consider testing for the null of no treatment effect using two kernel methods, a nonparametric propensity score model, and a nonparametric frequency propensity score model (traditional cell-based estimator).

We consider the following data generating process (DGP):

$$\begin{aligned} y_i &= g(x_i^c, x_i^d, t_i) + \epsilon_i \\ &= g_0(x_i^c, x_i^d) + [g_1(x_i^c, x_i^d) - g_0(x_i^c, x_i^d)]t_i + \epsilon_i \\ &= g_0(x_i^c, x_i^d) + \tau(x_i^c, x_i^d)t_i + \epsilon_i \\ &= \beta_0 + \beta_1 x_{i1}^c + \beta_2 (x_{i1}^c)^2 + \beta_3 x_{i1}^d + \beta_4 x_{i2}^d + \tau(x_i^c, x_i^d)t_i + \epsilon_i, \end{aligned} \tag{3.25}$$

where x_1^c is $U[-1,1]$ and $x_1^d \in \{0,1\}$ with $P[x_1^d = 1] = 0.2$ and $x_2^d \in \{0,1\}$ with $P[x_1^d = 1] = 0.6$, and let $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \tau)' = (1, 2, 2, 2, 0, \tau)$, while $\sigma_\epsilon = 1$ (x_2^d is irrelevant). These parameter choices yield an adjusted R^2 of around 66%.

Our model for the propensity score (Probit) is

$$\begin{aligned} t_i^* &= \gamma_0 + \gamma_1 x_{i1}^c + \gamma_2 (x_{i1}^c)^2 + \gamma_3 x_{i1}^d + \gamma_4 x_{i2}^d + \eta_i, \\ t_i &= \begin{cases} 1 & \text{if } \Phi(t_i^*) > 0.5 \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.26}$$

where $\sigma_\eta = 1$ and Φ is the standard normal CDF. We set $(\gamma_0, \gamma_1, \gamma_2, \gamma_4, \gamma_4)' = (-1/2, 1/4, 1/4, 1/4, 0)$ (x_2^d is irrelevant). We fix the sample size at for $n = 250$. These parameter choices yield a correct classification ratio of roughly 60%.

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 $\{t, y\}$.

[3] For the definition of convergence in distribution in probability, see Li et al (2003).

2. Using $\{t_i, y_i, x^c_{i1}, x^d_{i1}, x^d_{i2}\}_{i=1}^n$, compute $\hat{\tau}$ using each of the four kernel methods.

3. Compute tests for the null of no effect for each of the four kernel methods based on 199 bootstrap replications under H_0 at nominal levels $\alpha = 0.10, 0.05, 0.01$.

4. Repeat steps 1 through 3 $B = 1000$ times for values of τ in $\{0.0, 0.25, 0.50, 0.75\}$.

All bandwidths were selected via cross-validation based upon two restarts of a multidimensional numerical search routine allowing for different bandwidths for all variables.

The results are summarized in tables 1 and 2. Table 1 presents results for the proposed method which employs nonparametric kernel approach appropriate for mixed data ('smooth'), while Table 2 presents the traditional frequency approach, which involves splitting the data into subsamples ('non-smooth').

Table 1: Empirical Rejection Frequencies for Smooth Propensity Score Model.

τ	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.00	0.10	0.04	0.01
0.25	0.65	0.46	0.15
0.50	0.97	0.92	0.58
0.75	1.00	0.99	0.94

Table 2: Empirical Rejection Frequencies for Non-Smooth Propensity Score Model.

τ	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
0.00	0.07	0.03	0.00
0.25	0.57	0.36	0.10
0.50	0.96	0.89	0.55
0.75	1.00	1.00	0.92

We observe first that the smooth approach is correctly sized and has power in the direction of the alternative DGP. The traditional nonsmooth propensity score approach suffers from minor size distortions, suggesting that it is more susceptible to efficiency losses arising from sample splitting than the nonsmooth functional coefficient approach. The important comparison lies with the performance of the proposed smooth and the traditional nonsmooth approaches. It can be seen that, as expected, sample splitting leads to efficiency loss, which manifests itself as a loss in power.

Note that we have only considered one binary irrelevant covariate case, when there exists more irrelevant covariates, or one irrelevant covariate that takes more than two different values, the power loss becomes more substantial. Summarizing, the proposed smooth test is more powerful that a traditional frequency-based (nonsmooth) test when confronted with mixed data settings, which are often encountered in applied settings.

4 An Empirical Application

We will be interested in models that make use of the following variables[4] which were taken from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The data was obtained from the Department of Health Evaluation Sciences at the University of Virginia[5]:

- 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 11K, 1 if 11-25K, 2 if 25-50K, 3 if over 50K
- 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 & Medicaid, 4 if No insurance, 5 if Private, 6 if Private & Medicare
- X_7: Age - age (converted to years from Y/M/D data stored with 2 decimal accuracy)

Table 3 presents some summary statistics on the variables described above. 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 that 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".

[4] The Connors et al (1996) study 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 detect results that are qualitatively and quantitatively similar to those reported in Connors et al (1996) and Lin et al (1998).

[5] We are most grateful to Dr. B. Knaus and Dr. F. Harrell Jr. for making this data available to us, and also to Luz Saavedra for helping with data conversion.

Table 3: 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	101.85

Lin et al (1998) note that cardiologists and intensive care physician's 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 matrix for the parametric propensity score model is

A/P	0	1
0	2841	710
1	1197	987
Sample Size		5735
CR(0-1)		66.7%
CR(0)		80.0%
CR(1)		45.2%

The confusion matrix for the nonparametric propensity score model is

A/P	0	1
0	2916	635
1	1092	1092
Sample Size		5735
CR(0-1)		69.9%
CR(0)		82.1%
CR(1)		50.0%

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 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 probability of treatment themselves. The increased risk suggested by the parametric model drops from a 7% increase for those receiving RHC to roughly 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 $[-0.038, 0.011]$ for the nonparametric approach. Thus, we overturn the parametric testing result and conclude that patients receiving RHC treatment does not suffer an increased risk.

We also conducted some sensitivity analysis. Using likelihood cross-validation rather than least-squares cross-validation yielded 95% coverage error bounds of $[-0.034, 0.013]$. Out of concern that the nonparametric results might reflect 'overfitting', we computed the leave-one-out kernel predictions, and again bootstrapped their error bounds. For the least-squares cross-validated estimates we obtained 95% coverage error bounds of $[-0.015, 0.037]$, while for the likelihood cross-validated estimates we obtained 95% coverage error bounds of $[-0.007, 0.039]$.

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. This does not appear to be a result of a loss of efficiency due to using the nonparametric propensity score rather than the parametric one as can be seen by a comparison of the out of sample prediction results of the confusion matrices, because if the parametric model is correctly specified, one should expect that the parametric model predicts better than a nonparametric model.

A Appendix A

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 (2.23), $\hat{m}_\tau = n^{-1}\sum_{j=1}^{n}\hat{\tau}_{nj}$ 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 $E(t_i|x_i)$, $\hat{u}_i = \hat{g}_0(x_i) + t_i\hat{\tau}_n(x_i)$, $\hat{g}_0(x_i)$ is defined in (2.22).

To estimate $B_{h,\lambda}$ we need to estimate B_{1s} $(s = 1, ..., q)$ and B_{2s} $(s = 1, ..., r)$, these quantities can be estimated by 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 derivatives estimators can be obtained by taking derivatives (since the kernel function is differentiable up to order ν). Finally, replacing the population mean $E(.)$ by sample mean lead 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}. 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 means that $n^{-1}\sum_{i=1}^{n}A(x_i) = n^{-1}\sum_{i=1}^{n}B(x_i) + (s.o.)$.

Proof of Lemma 2.2

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

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

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

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

When using a second order kernel ($\nu = 2$), Hall et al. (2004a,b) have shown that

$$\begin{aligned} CV(h,\lambda)|_{\nu=2} = \sum_{x^d} \int \Bigg\{ & \frac{\kappa_2}{2} \sum_{s=1}^{q} \left[(fg)_s^{(2)}(x) - g(x) f_s^{(2)}(x) \right] h_s^2 \\ & + \sum_{v^d} \sum_{s=1}^{r} I_s(v^d, x^d) \left[g(x^c, v^d) - g(x) \right] f(x^c, v^d) \lambda_s \Bigg\}^2 S(x) f(x)^{-1} dx^c \qquad \text{(A.3)} \\ & + \frac{\kappa^q}{nh_1 \dots h_q} \sum_{x^d} \int \sigma^2(x) S(x) dx^c + (s.o.), \end{aligned}$$

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

By following exactly the same derivation as in Hall et al. (2004a,b), one can show that, with a νth order kernel,

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

where $B_{1s}(x) = (\kappa_q/\nu!)[(fg)_s^{(\nu)}(x) - g(x) f_s^{(\nu)}(x)]$ $(s = 1, \dots, q)$, $\kappa_q = \int w(v) v^q dv$, and $B_{2s}(x) = \sum_{v^d} I_s(v^d, x^d)[g(x^c, v^d) - g(x)] f(x^c, v^d)$, and (s.o.) denote terms having smaller probability orders, uniformly in $(h, \lambda) \in (0, \eta_n]^{p+r}$.

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

Define a_s via $h_s = a_s n^{-1/(2\nu+q)}$ $(s = 1, ..., q)$, and b_s via $\lambda_s = b_s n^{-\nu/(2\nu+q)}$ $(s = 1, ..., r)$, then (A.4) can be written as $C(h, \lambda) = n^{-2\nu/(2\nu+q)} \chi(a, b) + (s.o.)$ uniformly in $(h, \lambda) \in (0, \eta_n]^{p+r}$, where

$$\begin{aligned} \chi(a,b) &= \sum_{x^d} \int \left\{ \sum_{s=1}^{q} B_{1s}(x) a_s^\nu + \sum_{s=1}^{r} B_{2s}(x) b_s \right\}^2 S(x) f(x)^{-1} dx^c \\ & \quad + \frac{\kappa^q}{a_1 \dots a_q} \sum_{x^d} \int \sigma^2(x) S(x) dx^c. \end{aligned} \tag{A.5}$$

We assume that

$$\begin{aligned} &\text{There exist unique finite positive constants } a_s^0 \ (s = 1, ..., q) \text{ and} \\ &\text{finite non-negative constants } \lambda_s \ (s = 1, ..., r) \text{ that minimzes } \chi(a, b). \end{aligned} \tag{A.6}$$

Define a $(q + r) \times (q + r)$ positive semidefinite matrix A via its tth row and sth column as $A_{t,s} = \sum_{x^d} \int \bar{B}_t(x) \bar{B}(x_s) S(x) f(x)^{-1} dx^c$, where $\bar{B}_t(x) = B_{1t}(x)$ for $t = 1, ..., q$, and $\bar{B}_{q+t}(x) = B_{2t}(x)$ for

$t = 1, ..., r$, then it can be easily shown that a sufficient condition for (A.6) to hold is that A is positive definite.

From (A.4), (A.5) and (A.6), we obtain that $\bar{h}_s = a_s^0 n^{-\nu/(2\nu+q)} + o_p(n^{-\nu/(2\nu+q)})$ and $\bar{\lambda}_s = \lambda_s^0 n^{-1/(2\nu+q)} + o_p(n^{-1/(2\nu+q)})$, Lemma 2.1 then follows from this since $\hat{h}_s = \bar{h}_s n^{1/(2\nu+q)} n^{-1/(\nu+q+2)}$ and $\hat{\lambda}_s = \bar{\lambda}_s n^{\nu/(2\nu+q)} n^{-\nu/(\nu+q+2)}$.

Proof of Theorem 2.1

In order to make the proof of Theorem 2.1 more manageable we make a number of simplifying assumptions. (i) 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 $\hat{\tau}$. (ii) we replace $\hat{h}_s$ by the non-stochastic quantity $h_s^0 = a_s^0 n^{-1/(\nu+q+2)}$ $(s = 1, ..., q)$, and $\hat{\lambda}_s$ by $\lambda_s^0 = b_s^0 n^{-2/(\nu+q+2)}$ $(s = 1, ..., r)$. (iii) when we evaluate the probability *order* of a term, we sometimes assume that $h_s = h$ for all $s = 1, \ldots, q$, to simplify the notation. For example, we will write $O(h^\nu)$ for $O(\sum_{s=1}^q h_s^\nu)$ 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 2.2 shows that $\hat{h}_s/h_s - 1 = o_p(1)$, and $\hat{\lambda}_s/\lambda_s - 1 = o_p(1)$, and by the stochastic equicontinuity result of Ichimura (2000), we know that the asymptotic distribution of $\hat{\tau}$ remains the same whether one uses $\hat{h}_s$'s and $\hat{\lambda}_s$'s, or the non-stochastic leading term of them (h_s^0's and λ_s^0's).

We will repeatedly use the U-statistic H-decomposition in the proof below. We sometimes write n^{-1} for $(n-1)^{-1}$ since this approximation does not affect the order of any quantity 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.,

$$\hat{t}_i = \frac{n^{-1} \sum_{j \neq i}^n t_j K_n(x_j, x_i)}{\hat{f}_i}, \tag{A.7}$$

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

Define $v_i = t_i - E(t_i|x_i) \equiv t_i - \mu_i$, so that $t_i = \mu_i + v_i$, replacing t_j by $\mu_j + v_j$ in the right-hand-side of (A.7) we have

$$\hat{t}_i = \hat{\mu}_i + \hat{v}_i, \tag{A.8}$$

where

$$\hat{\mu}_i = \frac{n^{-1} \sum_{j \neq i}^n \mu_j K_h(\frac{x_j - x_i}{h})}{\hat{f}_i}, \tag{A.9}$$

and

$$\hat{v}_i = \frac{n^{-1} \sum_{j \neq i}^n v_j K_h(\frac{x_j - x_i}{h})}{\hat{f}_i}. \tag{A.10}$$

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

$$w_i = \mu_i(1 - \mu_i) \text{ and } \tilde{w}_i = \hat{t}_i(1 - \hat{t}_i). \tag{A.11}$$

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

$$\frac{1}{\tilde{w}_i} = \frac{1}{w_i} + \frac{w_i - \tilde{w}_i}{w_i^2} + \frac{(w_i - \tilde{w})^2}{w_i^2 \tilde{w}_i}. \tag{A.12}$$

Proof of Theorem 2.1

We have defined $\tilde{\tau}$ in (2.17). We now define another intermediate quantity $\bar{\tau}$ ($v_i = t_i - \mu_i$ and we omit M_{ni} for notational simplicity):

$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{(t_i-\mu_i)y_i}{w_i} \equiv \frac{1}{n}\sum_{i=1}^{n}\frac{v_i y_i}{w_i}. \tag{A.13}$$

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

$$\begin{aligned}\sqrt{n}(\hat{\tau}-\tau) &= \sqrt{n}[(\hat{\tau}-\tilde{\tau})+(\tilde{\tau}-\bar{\tau})+(\bar{\tau}-\tau)] \\ &= J_{1n}+J_{2n}+J_{3n},\end{aligned} \tag{A.14}$$

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

Recall that $w_i=\mu_i(1-\mu_i)$, $\tilde{w}_i=\hat{t}_i(1-\hat{t}_i)$. Using (A.12), we get from (2.12)

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}\left[t_i-\hat{t}_i\right]y_i\left[\frac{1}{w_i}+\frac{w_i-\tilde{w}_i}{w_i^2}+\frac{(w_i-\tilde{w}_i)^2}{w_i^2\tilde{w}_i}\right] \\ &\equiv L_{1n}+L_{2n}+L_{3n},\end{aligned} \tag{A.15}$$

where

$$\begin{aligned}L_{1n} &= n^{-1}\sum_{i=1}^{n}[(t_i-\hat{t}_i)y_i]/w_i \equiv \tilde{\tau}, \\ L_{2n} &= n^{-1}\sum_{i=1}^{n}[(t_i-\hat{t}_i)(w_i-\tilde{w}_i)y_i]/[w_i^2], \\ L_{3n} &= n^{-1}\sum_{i=1}^{n}[(t_i-\hat{t}_i)(w_i-\tilde{w}_i)^2y_i]/[w_i^2\tilde{w}_i].\end{aligned} \tag{A.16}$$

Note that $L_{1n}=\tilde{\tau}$, therefore, by (A.15) we have

$$J_{1n}=\sqrt{n}(\hat{\tau}-\tilde{\tau})=\sqrt{n}L_{2n}+\sqrt{n}L_{3n}. \tag{A.17}$$

Lemmas A.3 and A.4 (see below) give the leading terms of J_{1n} and J_{2n}.

Using (2.6) and adding and subtracting terms, we write $J_{3n}=\sqrt{n}(\bar{\tau}-\tau)$ as

$$\begin{aligned}J_{3n} &= n^{-1/2}\sum_{i=1}^{n}(v_iy_i/w_i-\tau) \\ &= n^{-1/2}\sum_{i=1}^{n}(v_iy_i/w_i-\tau_i)+n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau) \\ &= n^{-1/2}\sum_{i=1}^{n}[v_i(g_{0i}+\tau_it_i+u_i)/w_i-\tau_i]+n^{-1/2}\sum_{i=1}^{n}(\tau_i-\tau).\end{aligned} \tag{A.18}$$

By Eq. (A.18), Lemma A.3 and Lemma A.4, we obtain from Eq. (A.14) that

$$\begin{aligned}
\sqrt{n}(\hat{\tau} - \tau - B_{h,\lambda}) &= J_{1n} + J_{2n} - n^{1/2} B_{h,\lambda} + J_{3n} \\
&= n^{-1/2} \sum_{i=1}^{n} v_i [2\mu_i - 1] \tau_i / w_i - n^{-1/2} \sum_{i=1}^{n} v_i (g_{0i} + \tau_i \mu_i)/w_i] \\
&\quad + n^{-1/2} \sum_{i=1}^{n} \{ v_i (g_{0i} + \tau_i t_i + u_i)/w_i - \tau_i \} + n^{-1/2} \sum_{i=1}^{n} (\tau_i - \tau) + o_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} \frac{v_i u_i}{w_i} + n^{-1/2} \sum_{i=1}^{n} (\tau_i - \tau) + o_p(1) \\
&\equiv Z_{n2} + Z_{n3} + o_p(1), \qquad \text{(A.19)}
\end{aligned}$$

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 above we have used the following cancellation result ($w_i = \mu_i(1 - \mu_i)$):

$$\begin{aligned}
& n^{-1/2} \sum_{i=1}^{n} \left[\frac{v_i (t_i - \mu_i)}{w_i} - 1 \right] \tau_i + n^{-1/2} \sum_{i} v_i \left[\frac{2\mu_i - 1}{w_i} \right] \tau_i \\
= \; & n^{-1/2} \sum_{i=1}^{n} \left[\frac{v_i^2 - \mu_i(1 - \mu_i) + 2 v_i \mu_i - v_i}{w_i} \right] \tau_i \\
= \; & n^{-1/2} \sum_{i=1}^{n} \left[\frac{v_i^2 - \mu_i + \mu_i^2 + 2 v_i \mu_i - v_i}{w_i} \right] \tau_i \\
= \; & n^{-1/2} \sum_{i=1}^{n} \left[\frac{(\mu_i + v_i)^2 - (\mu_i + v_i)}{w_i} \right] \tau_i \\
= \; & 0, \qquad \text{(A.20)}
\end{aligned}$$

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.19) and the Lindeberg central limit theorem.

Proof of Lemma 2.3

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.18) and Lemma A.4, we have ($v_i = t_i - \mu_i$)

$$\begin{aligned}
\sqrt{n}(\tilde{\tau} - \tau - B_{h,\lambda}) &= J_{2n} - n^{1/2} B_{h,\lambda} + J_{3n} \\
&= -n^{-1/2} \sum_{i=1}^{n} v_i (g_{0i} + \tau_i \mu_i)/w_i \\
&\quad + \; n^{-1/2} \sum_{i=1}^{n} [v_i (g_{0i} + \tau_i t_i + u_i)/w_i - \tau_i] + n^{-1/2} \sum_{i=1}^{n} (\tau_i - \tau) + o_p(1) \\
&= n^{-1/2} \sum_{i=1}^{n} [\frac{v_i^2}{w_i} - 1] \tau_i + n^{-1/2} \sum_{i=1}^{n} \frac{v_i u_i}{w_i} + n^{-1/2} \sum_{i=1}^{n} (\tau_i - \tau) + o_p(1) \\
&\equiv Z_{n1} + Z_{n2} + Z_{n3} + o_p(1) \\
&\xrightarrow{d} N(0, V_1 + V_2 + V_3) \text{ by the Lindeberg central limit theorem,} \qquad \text{(A.21)}
\end{aligned}$$

where Z_{n1} and Z_{n2} are defined in (A.19), $Z_{n3} = n^{-1/2}\sum_{i=1}^{n}\left[\frac{v_i^2}{w_i} - 1\right]$. Note that by Lemma A.3 and (A.20) we know that $J_{1n} = -Z_{n3} + o_p(1)$.

Below 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

$$\frac{1}{\hat{f}_i} = \frac{1}{f_i} + \frac{f_i - \hat{f}_i}{f_i\hat{f}_i} = \frac{1}{f_i} + \sum_{l=1}^{p}\frac{(f_i - \hat{f}_i)^l}{f_i^{l+1}} + \frac{(f_i - \hat{f}_i)^{p+1}}{f_i^p\hat{f}_i}. \tag{A.22}$$

For example, in Lemma A.1 below we need to evaluate a term like $n^{-1}\sum_{i=1}^{n} v_i y_i \hat{v}_i / w_i^2$. $\hat{v}_i = n^{-1}\sum_{j\neq i} v_j K_{n,ij}/\hat{f}_i$ has a random denominator $\hat{f}_i$. 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 smaller order than the main term that is associated with $1/f_i$. Also, using the uniform convergence rate of $sup_{x\in\mathcal{S}}|\hat{f}(x) - f(x)| = O_p(\sum_{s=1}^{q} h_s^{\nu} + ln\, n(nh_1...h_q)^{-1})$, 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.22) we have

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

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

$$\begin{aligned}
L_{2n} &= n^{-1}\sum_{i=1}^{n} y_i(t_i - \hat{t}_i)[\mu_i - \hat{t}_i - (\mu_i^2 - \hat{t}_i^2)]/w_i^2 \\
&= n^{-1}\sum_{i=1}^{n} y_i(t_i - \hat{t}_i)(\mu_i - \hat{t}_i)[1 - (\mu_i + \hat{t}_i)]/w_i^2 \\
&= n^{-1}\sum_{i=1}^{n} y_i(\mu_i - \hat{\mu}_i + v_i - \hat{v}_i)(\mu_i - \hat{\mu}_i - \hat{v}_i)[1 - (\mu_i + \hat{\mu}_i + \hat{v}_i)]/w_i^2 \\
&= -n^{-1}\sum_{i=1}^{n} y_i v_i \hat{v}_i[1 - 2\mu_i]/w_i^2 + o_p\left(n^{-1/2}\right) \quad \text{(using } \hat{\mu}_i = \mu_i + (\hat{\mu}_i - \mu_i) \text{ and Lemma B.3)} \\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} y_i v_i v_j (2\mu_i - 1)K_{n,ij}/w_i^2 + o_p\left(n^{-1/2}\right) \quad \text{(by Lemma B.2 and } E(v_i|x_i) = 0\text{)} \\
&= \frac{2}{n(n-1)}\sum_i\sum_{j>i} H_{n,a}(z_i, z_j) + (s.o.),
\end{aligned} \tag{A.24}$$

where $H_{n,a}(z_i, z_j) = (1/2)v_iv_j\{y_i(1-2\mu_i)/[f_iw_i^2] + y_j(2\mu_j - 1)/[f_jw_j^2]\}K_{n,ij}$ 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

$$
\begin{aligned}
L_{2n} &= \frac{2}{n(n-1)}\sum_i\sum_{j>i} H_{n,a}(z_i, z_j) + (s.o.) \\
&= 0 + (2/n)\sum_i H_{1n,a}(z_i) + \frac{2}{n(n-1)}\sum_i\sum_{j>i}\{H_{n,a}(z_i, z_j) - H_{1n,a}(z_i) - H_{1n,a}(z_j) + 0\} + (s.o.), \\
&= n^{-1}\sum_i v_i\tau_i(2\mu_i - 1)/w_i + O_p\left((nh^{q/2})^{-1}\right) + (s.o.) = n^{-1}\sum_i v_i\tau_i(2\mu_i - 1)/w_i + o_p\left(n^{-1/2}\right)
\end{aligned}
\tag{A.25}
$$

by Lemma B.4 and assumption (A3), where we also used the fact that the degenerate U-statistic $U_{n,a} \overset{def}{=} [2/n(n-1)]\sum_i\sum_{j>i}\{H_{n,a}(z_i, z_j) - H_{1n,a}(z_i) - H_{1n,a}(z_j)\}$ has a second moment of $E[U_{n,a}^2] = O\left((n^2h^q)^{-1}\right)$, so $U_{n,a} = O_p\left((nh^{q/2})^{-1}\right)$.

Lemma A.2 $L_{3n} = O\left(h^{2\nu} + h^2(nh^q)^{-1}\right) = o_p\left(n^{-1/2}\right)$.

Proof: Using the identity of

$$
\frac{1}{\tilde{w}_i} = \frac{1}{w_i} + \sum_{l=1}^{p}\frac{(w_i - \tilde{w}_i)^l}{w_i^{l+1}} + \frac{(w_i - \tilde{w}_i)^{p+1}}{w_i^p\tilde{w}_i}, \tag{A.26}
$$

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 is because, (i) it is easy show that (by computing the second moment of them) the term associated with $(w_i - \tilde{w}_i)^l/w_i^{l+1}$ has an smaller order than the main term that is associated with $1/w_i$. Also, using the uniform convergence rate of $sup_{x\in\mathcal{S}}|\hat{\mu}(x) - \mu(x)| = O_p(\sum_{s=1}^q h_s^\nu + ln\, n(nh_1...h_q)^{-1})$, 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

$$
\begin{aligned}
L_{3n,1} &= n^{-1}\sum_i y_i(t_i - \hat{t}_i)(w_i - \tilde{w}_i)^2/w_i^3 \\
&= n^{-1}\sum_i [g_{0i} + \tau_i t_i + u_i][(\mu_i - \hat{t}_i) + v_i](\mu_i - \hat{t}_i)^2[1 - (\mu_i + \hat{t}_i)]^2/w_i^3 \\
&\sim n^{-1}\sum_{i=1}^{n}[v_i(\mu_i - \hat{t}_i)^2 + (\mu_i - \hat{t}_i)^3] = O\left(h^{2\nu} + h^2(nh^q)^{-1}\right)
\end{aligned}
\tag{A.27}
$$

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

Lemma A.3 $J_{1n} = n^{-1/2}\sum_i v_i(2\mu_i - 1)\tau_i/w_i + o_p(1)$.

Proof: It follows from lemmas A.1 and A.2.

Lemma A.4 $J_{2n} = B_{h,\lambda} - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} v_{1i}(g_{0i} + \tau_i\mu_i)/w_i + o_p(1)$.

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.

$$J_{2n,1} \equiv n^{-1/2}\sum_{i=1}^n(\mu_i - \hat{\mu}_i)y_i/w_i = B_{h,\lambda} + O_p\left(n^{1/2}h^{\nu+2} + h(nh^q)^{-1/2}\right)$$

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

Next,

$$\begin{aligned} J_{2n,2} &= n^{-1/2}\sum_{i=1}^n \hat{v}_i \hat{f}_i y_i/(f_i w_i) + o_p(1) \quad \text{(by using Eq. (A.22))} \\ &= n^{-1/2}(n-1)^{-1}\sum_{i=1}^n\sum_{j\neq i}^n v_j y_i K_{n,ij}/(f_i w_i) \\ &= \frac{2}{n^{1/2}(n-1)}\sum_{i=1}^n\sum_{j>i}(1/2)\{v_j y_i/(f_i w_i) + v_i y_j/(f_j w_j)\}K_{n,ij} \\ &= n^{1/2}\frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j>i} H_{n,b}(z_i, z_j), \end{aligned} \tag{A.29}$$

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

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

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

by Lemma B.4 (iii).

Hence, by the U-statistic H-decomposition we have

$$\begin{aligned} J_{2n,2} &= -n^{1/2}\{0 + (2/n)\sum_{i=1}^n H_{1n,b}(z_i) + \frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j>i}[H_{n,b}(z_i,z_j) - H_{1n,b}(z_i) - H_{1n,b}(z_j) + 0] \\ &= n^{-1/2}\sum_{i=1}^n v_i(g_{0i} + \tau_i\mu_i)/w_i + O_p\left((nh^q)^{-1/2}\right) \end{aligned} \tag{A.31}$$

because the last term in the H-decomposition is a degenerate U-statistic which has an order $n^{1/2}O_p\left((nh^{q/2})^{-1}\right) = O_p\left((nh^q)^{-1/2}\right)$.

B 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 $f(x^d, x^c) \in \mathcal{G}_{\nu-1}$, $\nu \geq 2$ is an integer. Define $\eta_2 = \sum_{s=1}^q h_s^\nu + \sum_{s=1}^r \lambda_s$. Suppose the kernel function W satisfies (A2). Then, uniformly in x,*

(i) $E\{[g(X) - g(x)]K_n(X,x)\} = \sum_{s=1}^q C_{1s}(x)h_s^\nu + \sum_{s=1}^r C_{2s}(x)\lambda_s + O\left(\eta_2(\sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s)\right)$;

(ii) $E\left[K_n(X,x)\right] - f(x) = \sum_{s=1}^{q} D_{1s}(x)h_s^{\nu} + \sum_{s=1}^{r} D_{2s}(x)\lambda_s + O\left(\eta_2(\sum_{s=1}^{q} h_s^{\nu} + \sum_{s=1}^{r} \lambda_s)\right)$, *where* $C_{ls}(.)$ *and* $D_{ls}(.)$ *are defined in the proof below.*

Proof of (i):

$$\begin{aligned}
E\left\{[g(X)-g(x)]K_n(X,x)\right\} &= \sum_{z^d}\int f(z^c,z^d)\left[g(z^c,z^d)-g(x^c,x^d)\right]\times W_h(z^c,x^c)L(z^d,x^d,\lambda)dz^c \\
&= \int f(x^c+hv,x^d)\left[g(x^c+hv,x^d)-g(x^c,x^d)\right]W(v)dv \\
&\quad - \sum_{z^d\neq x^d}\int f(x^c,x^d)\left[g(x^c,x^d)-g(x^c,z^d)\right]W_h(z^c,x^c)L(z^d,x^d,\lambda_j)dz^c \\
&= \int\left\{(fg)(x^c+hv,x^d)-(fg)(x)-g(x)[f(x^c+hv,x^d)-f(x)]\right\}W(v)dv \\
&\quad + \sum_{s=1}^{r} I_s(z^c,x^d)f(x^c,x^d)\ \left[g(x^c,z^d)-g(x^c,x^d)\right]\ \lambda_s \\
&\quad + O(\eta_2(\sum_{s=1} h_s^2+\sum^{r}\lambda_s)) \\
&= \sum_{s=1}^{q} C_{1s}h_s^{\nu} + \sum_{s=1}^{r} C_{2s}(x)\lambda_s + O(\eta_2(\sum_{s=1}^{q} h_s^2+\sum_{s=1}^{r}\lambda_s))
\end{aligned} \tag{B.1}$$

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

$$C_{1s}(x) = (1/\nu!)\kappa_\nu[(gf)_s^{(\nu)}(x) - f(x)g_s^{(\nu)}(x)], \tag{B.2}$$

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

$$C_{2s}(x) = I_s(z^d,x^d)f(x^c,x^d)[g(x^c,z^d)-g(x^c,x^d)]. \tag{B.3}$$

Proof of (ii):

$$\begin{aligned}
E\{[K_n(X,x)-f(x)]\} &= \sum_{z^d}\int f(z^c,z^d)W_h(z^c,z^d)L(z^d,x^d,\lambda)dz^d - f(x^c,x^d) \\
&= \int f(x^c+hv,x^d)W(v)dv - f(x^c,x^d) + \sum_{s=1}^{r} I_s(z^d,x^d)f(x^c,z^d)\lambda_s \\
&\quad + O\left(h^\nu(h^2+\sum_{s=1}^{r}\lambda_s)\right) \\
&= \sum_{s=1}^{q} D_{1s}(x)h_s^{\nu} + \sum_{s=1}^{r} D_{2s}(x)\lambda_s + O\left(\eta_2(\sum_{s=1}^{q} h_s^2+\sum_{s=1}^{r}\lambda_s)\right),
\end{aligned} \tag{B.4}$$

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

Lemma B.2 *(i)* $A_{1n} \overset{def}{=} n^{-1}\sum_i(\mu_i - \hat{\mu}_i)y_i/w_i = B_{h,\lambda} + O_p(n^{-1/2}h^{\nu} + h(n^2h^q)^{-1/2})$,

(ii) $A_{2n} \overset{def}{=} n^{-1}\sum_i(\mu_i - \hat{\mu}_i)\tau_i(1-2\mu_i)/w_i = \bar{B}_{h,\lambda} + O_p(n^{-1/2}h^{\nu} + h(n^2h^q)^{-1/2})$, *where the definitions of* $B_{h,\lambda}$ *and* $\bar{B}_{h,\lambda}$ *are given in the proof below.*

Proof of (i): Using (A.22), we know that $A_{1n} = A_{1n,1} + (s.o.)$, where $A_{1n,1} = n^{-1}\sum_i(\mu_i - \hat{\mu}_i)\hat{f}_i y_i/(w_i f_i)$. By noting that $E(u_i|x_i) = 0$ and $E(v_i|x_i) = 0$, and denoting by $m(x) = E(y|x) = g_{01}(x) + \tau(x)\mu(x)$, we first compute $E(A_{1n})$.

$$\begin{aligned}
E(A_{1n,1}) &= E[(\mu_1 - \mu_2)K_{n,1,2}y_1/(f_1w_1)] \\
&= \sum_{x_1^d}\sum_{x_2^d}\int\int f(x_2)m(x_1)w(x_i)^{-1}(\mu_1-\mu_2)W_{h,1,2}L_{\lambda,1,2,}dx_1^c dx_2^c \\
&= \sum_{x^d}\int\int m(x)w(x)^{-1}\{f(x^2+hv,x^d)[\mu(x)-\mu(x^c+hv,x^d)]\}W(v)dvdx^c \\
&\quad + \sum_{x_1^d}\sum_{x_2^d \neq x^d}\int\int m(x)w(x)^{-1}\left\{\,[\,f(x^c+hv,x^d)-f(x)\,]-[(f\mu)(x^c+hv,x^d)-(f\mu)(x)\,]\right\}W(v)L_{\lambda,1,2}d \\
&= \sum_{s=1}^{q} B_{1s}h_s^{\nu} + \sum_{s=1}^{r} B_{2s}\lambda_s + O(h^{\nu+2}) \equiv B_{h,\lambda} + O(h^{\nu+2})
\end{aligned} \tag{B.5}$$

by the same proof of Lemma B.1 (i), where $B_{h,\lambda} = \sum_{s=1}^q B_{1s}h_s^{\nu} + \sum_{s=1}^r B_{2s}\lambda_s$ with

$$B_{1s} = -(\kappa_\nu/\nu!)E\left\{f(x_i)m(x_i)w(x_i)^{-1}\left[\mu_s^{(\nu)}(x_i) - (\mu f)_s^{(\nu)}(x_i)\right]\right\}, \tag{B.6}$$

and

$$B_{2s} = \sum_{x^d} E\left\{I_s(x^d,x_i^d)f(x_i)m(x_i)w(x_i)^{-1}f(x_i^c,x^d)(\mu(x_i)-\mu(x_i^c,x^d))\right\}. \tag{B.7}$$

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

$$E(A_{1n}^2) = n^{-4}\sum_{i_1}\sum_{j_1\neq i_1}\sum_{i_2}\sum_{j_2\neq i_2} E\left[(\mu_{i_1}-\mu_{j_1})K_{n,i_1,j_1}y_{i_1}(\mu_{i_2}-\mu_{j_2})K_{n,i_2,j_2}y_{i_2}/(w_{i_1}w_{i_2})\right].$$

We consider three cases: (i) the four indices i_1, j_1, i_2, and j_2 are all different, (ii) the four indices assume three distinct values, and (iii) the four indices assume two different values.

First for case (i), it is easy to see that in this case $E(A_{1n,(i)}^2) - [E(A_{1n})]^2 = n^{-1}O\left([E(A_{1n})]^2\right) = O\left(n^{-1}h^{2\nu}\right)$.

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

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

Finally, for case (iii) we have

$$\begin{aligned}
E(A_{1n}^2)_{(iii)} &\leq Cn^{-4}n^2\left\{E[y_1^2(\mu_1-\mu_2)^2K_{n,12}^2] + E[y_1y_3(\mu_1-\mu_3)^2K_{n,13}^2]\right\} \\
&= n^{-2}O(h^{-q}h^2) = O((n^2h^{q-2})^{-1}).
\end{aligned} \tag{B.9}$$

Summarizing the above results we have shown that

$$Var(A_{1n}) = O(n^{-1}h^{2\nu} + (n^2h^{q-2})^{-1}). \tag{B.10}$$

Hence, $A_{1n} = \sum_{s=1}^{q} B_{1s}h_s^{\nu} + \sum_{s=1}^{r} B_{2s}\lambda_s + O_p(n^{-1/2}h^{\nu} + h(n^2h^q)^{-1/2})$.

Proof of (ii): It follows exactly the same proof as in (i) above with $m(x)$ replaced by $\bar{m}(x) \overset{def}{=} \tau(x)(1-2\mu(x))$. Therefore, (ii) follows with $\bar{B}_{h,\lambda} = \sum_{s=1}^{q} \bar{B}_{1s}h_s^{\nu} + \sum_{s=1}^{r} \bar{B}_{2s}\lambda_s$,

$$\bar{B}_{1s} = (\kappa_\nu/\nu!)E\left\{f(x)\bar{m}(x)w(x)^{-1}\left[\mu(x)f_s^{(\nu)}(x) - (\mu f)_s^{(\nu)}(x)\right]\right\}, \tag{B.11}$$

and

$$\bar{B}_{2s} = \sum_{x_1^d}\sum_{x_2^d}\int I_s(x_2^d, x_1^d)f(x_1)\bar{m}(x_1)w(x_1)^{-1}f(x_1^c, x_2^d)(\mu(x_1) - \mu(x_1^c, x_2^d))dx_1^c. \tag{B.12}$$

Lemma B.3 *Let $\xi_i = \mu_i$, or μ_{2i} or μ_{3i}, $\epsilon_i = v_i$, or v_{2i} or v_{3i}, then we have*

(i) $A_{3n} \overset{def}{=} n^{-1}\sum_i(\hat{\xi}_i - \xi_i)^2 = O_p(h^{2\nu} + h^2(nh^q)^{-1})$.

(ii) $A_{4n} \overset{def}{=} n^{-1}\sum_i \hat{\epsilon}_i^2 = O_p((nh^q)^{-1})$.

(iii) $A_{5n} \overset{def}{=} n^{-1}\sum_i(\hat{\xi}_i - \xi_i)\hat{\epsilon}_i = O_p(h^{2\nu} + (nh^q)^{-1})$.

(iv) $A_{6n} \overset{def}{=} n^{-1}\sum_i(\hat{t}_i - \mu_i)^2 = O_p(h^{2\nu} + (nh^q)^{-1})$.

Since the proof for $\xi_i = \mu_i$, μ_{2i} or μ_{3i} are identical. We only prove the case of $\xi_i = \mu_i$, $\epsilon_i = v_i$.

Proof of (i). Using (A.22), we have $A_{3n} \equiv n^{-1}\sum_i(\hat{\xi}_i - \xi_i)^2\hat{f}_i^2/\hat{f}_i^2 = n^{-1}\sum_i(\hat{\xi}_i - \xi_i)^2\hat{f}_i^2/f_i^2 + (s.o.)$. Also, since $f(x)$ is bounded below by a positive constant, we only need to prove (i) for $A_{3n,1} \overset{def}{=} n^{-1}\sum_i(\hat{\xi}_i - \xi_i)^2\hat{f}_i^2$.

$$\begin{aligned}
&E[|A_{3n,1}|] = E[(\hat{\mu}_1 - \mu_1)^2\hat{f}_1^2] \\
&= \frac{1}{(n-1)^2}\sum_{i\neq 1}^{n}\sum_{j\neq 1}^{n} E\left[(\mu_i - \mu_1)K_{n,i,1}(\mu_j - \mu_1)K_{n,j,1}\right] \\
&= \frac{1}{(n-1)^2}\left\{(n-1)E\left[(\mu_i - \mu_1)^2K_{n,i,1}^2\right] + (n-1)(n-2)E\left[(\mu_2 - \mu_1)K_{n,2,1}\right]E\left[(\mu_3 - \mu_1)K_{n,3,1}\right]\right\} \\
&= O(h^2(nh^q)^{-1}) + O(h^{2\nu})
\end{aligned} \tag{B.13}$$

by Lemma B.1, where we used $E\left[(\mu_i - \mu_1)^2K_{n,i,1}^2\right] = O\left((h^2 + \sum_{s=1}^{r}\lambda_s)h^{-q}\right) = O\left((h^2 + h^{\nu})h^{-q}\right) = O\left(h^2h^{-q}\right)$ because $\lambda_j = O\left(h^{\nu}\right)$ and $\nu \geq 2$. Thus, $A_{3n,1} = O_p(h^2(nh^q)^{-1}) + O(h^{2\nu})$.

Proof of (ii). Similarly, by (A.22), we have $A_{4n} \equiv n^{-1}\sum_{i=1}^{n}\hat{v}_i^2\hat{f}_i^2/\hat{f}_i^2 = n^{-1}\sum_{i=1}^{n}\hat{v}_i^2\hat{f}_i^2/f_i^2 + (s.o.)$. We only (since f_i^{-1} is bounded) need to prove (ii) for $A_{4n,1} = n^{-1}\sum_i\hat{\epsilon}_i^2\hat{f}_i^2$.

$$E\left[|A_{4n,1}|\right] = E\left[\hat{v}_1^2\hat{f}_1^2\right] = \frac{1}{(n-1)^2}\sum_{i\neq 1}E\left[v_i^2K_{n,i,1}^2\right] = \frac{1}{n-1}E\left[v_2^2K_{n,2,1}^2\right] = O((nh^q)^{-1}).$$

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

Finally, (vi) follow 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, recall that $A_i = B_i + (s.o.)$ means that $n^{-1/2}\sum_{i=1}^n A_i = n^{-1/2}\sum_{i=1}^n B_i + (s.o.)$, then we have*
(i) $H_{1n,a}(z_i) = E[H_{n,a}(z_i, z_j)|z_i] = \tau_i\{2\mu_i - 1\}/w_i + (s.o.)$,
(ii) $H_{1n,b}(z_i) = E[H_{n,b}(z_i, z_j)|z_i] = (g_{0i} + \tau_i\mu_i)/w_i + (s.o.)$.

Proof of (i)

$H_{n,a}(z_i, z_j) = (1/2)\{y_i v_i v_j(2\mu_i - 1)/(f_i w_i^2) + y_j v_i v_j(2\mu_j - 1)/(f_j w_j^2)\}K_{n,ij}$, where $z_i = (x_i, t_i, u_i)$. By noting that μ_i, w_i, f_i g_{0i} and τ_i are all functions of x_i and that $E(v_i|x_i) = 0$, we have

$E[y_i v_i v_j(2\mu_i - 1)K_{n,ij}/(f_i w_i^2)|z_i] = y_i v_i(2\mu_i - 1)(f_i w_i^2)^{-1}E\{E[v_j K_{n,ij}|x_j, z_i]|z_i\} = 0$. Also, using $y_j = g_{0j} + \tau_j t_j + u_j = g_{0j} + \tau_j(\mu_j + v_j) + u_j$, and $E(v_j|x_j) = 0$, we have

$$\begin{aligned}
H_{1n,a}(z_i) &= E[H_{n,a}(z_i, z_j)|z_i] \\
&= (1/2)\left\{0 + 2v_i E[v_j y_j \mu_j K_{n,ij}/(f_j w_j^2)|z_i] - v_i E[v_j y_j K_{n,ij}/(f_j w_j^2)|z_i]\right\} \\
&= (1/2)\left\{2v_i E[v_j^2 \tau_j \mu_j K_{n,ij}/(f_j w_j^2)|z_i] - v_i E[v_j^2 \tau_j K_{n,ij}/(f_j w_j^2)|z_i]\right\} \\
&= (1/2)v_i\left\{2E[\tau_j \mu_j K_{n,ij}/(f_j w_j)|z_i] - E[\tau_j K_{n,ij}/(f_j w_j)|z_i]\right\} \text{ (because } E(v_j^2|x_j) = var(t_j|x_j) = u \\
&= (1/2)v_i \tau_i\{2\mu_i - 1\}/w_i + (s.o.),
\end{aligned} \tag{B.14}$$

where we have used the change-of-variable argument: $E\left[\tau_j K_n((x_i - x_j)/h)/(f_j w_j)\right]|z_i] = \tau_i + O\left(h^\nu + \sum_{s=1}^r \lambda_s\right)$ and $E\left[\tau_j \mu_j K_n((x_i - x_j)/h)/(f_j w_j)\right]|z_i] = \tau_i\mu_i + O\left(h^\nu + \sum_{s=1}^r \lambda_s\right)$.

Proof of (ii)

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

$$\begin{aligned}
H_{1n,b}(z_i) &= E[H_{n,b}(z_i, z_j)|z_i] \\
&= (1/2)\left\{0 + v_i E[y_j K_{n,ij}/(f_j w_j)|z_i]\right\} \\
&= (1/2)v_i E\left\{[g_{0j} + \tau_j(\mu_j + v_j) + u_j]K_{n,ij}/(f_j w_j)|z_i]\right\} \\
&= (1/2)v_i E\left\{[(g_{0j} + \tau_j\mu_j)K_{n,ij}/(f_j w_j)|z_i]\right\} \\
&= (1/2)v_i(g_{0i} + \tau_i\mu_i)/w_i + (s.o.)
\end{aligned} \tag{B.15}$$

by the change-of-variable argument.

Lemma B.5 *Let ϵ_i be either v_i, v_{2i} or v_{3i}, then $n^{-1}\sum_{i=1}^n \hat{v}_i m(x_i) = n^{-1}\sum_{i=1}^n v_i m(x_i) + o_p(n^{-1/2})$, where $m(.)$ is a continuous function and $E[m(x_i)^4]$ is finite.*

Proof: We will only prove the case that $\epsilon_i = v_i$ for other two cases follow the identical proof.

$n^{-1}\sum_{i=1}^n \hat{v}_i m_i = [n(n-1)]^{-1}\sum_{i=1}^n v_j m_i K_{n,ij} = [n(n-1)]^{-1}\sum_{i=1}^n\sum_{j>i} H_{n,v}(z_i, z_j)$, where $H_{n,v}(z_i, z_j) = (v_j m_i + v_i m_j)K_{n,ij}$ and $z_i = (x_i, t_i)$.

$E[H_{n,v}(z_i, z_j)|z_i] = 0 + v_i E[m_j K_{n,ij}|x_i] = v_i m(x_i) + (s.o.)$. Hence, by the H-decomposition we have

$n^{-1}\sum_{i=1}^n \hat{v}_i m_i = 0 + \{\frac{2}{n}\sum_{i=1}^n v_i m(x_i) + o_p(n^{-1/2})\} + O_p((n^2 h_1 ... h_q)^{-1/2}) = n^{-1}\sum_{i=1}^n v_i m_i + o_p(n^{-1/2})$, where the $O_p((n^2 h_1 ... h_q)^{-1/2})$ comes from the last term of H-decomposition which is second order the degenerate U-statistic

REFERENCES

Aitchison, J. & Aitken, C. G. G. (1976), "Multivariate binary discrimination by the kernel method," *Biometrika*, 63, 413-420.

Angrist, J. (1991), "Instrumental Variables Estimation of Average Treatment Effects in Econometrics and Epidemiology," NBER Technical Working Paper No. 115.

Angrist, J., G.W. Imbens and D. Rubin (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association* 91, 444-455.

Barnow, B., G. Cain and A. Goldberger (1980), "Issues in the Analysis of Selectivity Bias," *Evaluation Studies* 5, 42-59.

Cai, Z., J. Fan, and R. Li (2000). Efficient estimation and inferences for varying-coefficient models, *Journal of American Statistical Association* 95, 888-902.

Cai, Z., J. Fan, and Q. Yao (2000). Functional-coefficient regression models for nonlinear time series, *Journal of American Statistical Association* 95, 941-956.

Chen, R. and R.S. Tsay (1993). Functional-coefficient autoregressive models, *Journal of the American Statistical Association* 88, 298-308.

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

Dehejia, R.H. and S. Wahba (1999), "Causal Effects in Non-Experimental Studies: Evaluating the Evaluation of Training Programs," *JASA* 94, 1053- 1062.

Garen, J. (1984), "The Returns to Schooling: A Selectivity Bias Approach With a Continuous Choice Variable," *Econometrica* 52, 1199-1218.

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

Hall, P., Q. Li, and J. Racine (2004a). "Nonparametric estimation of regression functions when there exist irrelevant regressors," unpublished manuscript.

Hall, P., J. Racine and Q. Li (2004b). "Cross-validation and the estimation of conditional probability densities," foothcoming in *Journal of American Statistical Association.*

Heckman, J. and B. Honoré (1990), "The Empirical Content of the Roy Model," *Econometrica* 58, 1121-1149.

Heckman, J., H. Ichimura and P. Todd (1997), "Matching as an Econometric Evaluation Estimator: Theory and Methods," *Review of Economic Studies* 64, 605-654.

Hirano, K., G.W. Imbens and G. Ridder (2002). "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica.*

Ichimura, H. (2000). Asymptotic distribution of non-parametric and semiparametric estimators with data dependent smoothing parameters. Unpublished munuscript.

Ichino, A. and R. Winter-Ebmer (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 and Economic Statistics 17, 74-90.

LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs," *American Economic Review* 76, 604-620.

Li, Q., C. Hsiao and J. Zinn, (2003). Consistent specification tests for semiparametric/nonparametric models based on series estimation methods, *Journal of Econometrics* 112, 295-325.

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

Robinson, P. (1988), "Root-N consistent semiparametric regression," *Econometrica* 56, 931-954.

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