OPTIMAL BANDWIDTH SELECTION FOR NONPARAMETRIC CONDITIONAL DISTRIBUTION AND QUANTILE FUNCTIONS

JUAN LIN
SCHOOL OF ECONOMICS AND BUSINESS ADMINISTRATION, BEIJING NORMAL UNIVERSITY
BEIJING, 100875, PRC

QI LI
DEPARTMENT OF ECONOMICS, TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-4228, USA

JEFFREY S. RACINE
DEPARTMENT OF ECONOMICS, MCMASTER UNIVERSITY
HAMILTON, ONTARIO, CANADA, L8S 4M4

Abstract. Li & Racine (2008) consider the nonparametric estimation of conditional cumulative distribution functions (CDF) in the presence of discrete and continuous covariates along with the associated conditional quantile function. However, they did not propose an optimal data-driven method of bandwidth selection and left this important problem as an 'open question'. In this paper we propose an automatic data-driven method for selecting these bandwidths, establish the asymptotic optimality of our approach, and derive asymptotic normality results for the resulting nonparametric estimator. By solving this 'open question' we thereby provide practitioners with an optimal nonparametric approach for estimating conditional CDF and quantile functions.

1. Introduction

Though the nonparametric estimation of conditional probability density functions (PDF) has received substantial attention in the literature (Fan & Yim (2004), Hall, Racine & Li (2004), Chung & Dunson (2009)), certain problems such as the estimation of conditional quantiles require the estimation of conditional cumulative distribution functions (CDF). Nonparametric estimation of the latter has proven more formidable but has drawn the attention of a growing number of researchers (Bashtannyk & Hyndman (2001), Hyndman & Yao (2002), Li & Racine (2008) among others).

In a recent paper Li & Racine (2008) propose a nonparametric kernel-based CDF estimation method. They consider a very general setting allowing for both continuous and discrete covariates, while the dependent variable(s) can also be discrete or continuous. They also provide rates of convergence and asymptotic normality results for their proposed estimators. However, they come up short on the possibility of using optimal automatic data-driven methods for selecting the bandwidths. They state "Unfortunately, to the best of our knowledge, there does not exist an *automatic* data-driven method for optimally selecting bandwidths when estimating a conditional CDF in the sense that a weighted integrated MSE is minimized" (Li & Racine (2008, page 426)). As a compromise, they rely on data-driven methods that are optimal for selecting bandwidths for a conditional PDF as proposed by Hall et al. (2004).

Racine would like to thank the Shared Hierarchical Academic Research Computing Network (SHARCNET:www.sharcnet.ca) for their ongoing support and to gratefully acknowledge financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC:www.nserc.ca) and from the Social Sciences and Humanities Research Council of Canada (SSHRC:www.sshrc.ca). We would also like to extend our thanks to Laine Ruus of the University of Toronto Data Library Service for assistance in procuring the public use versions of the Survey of Household Spending/Family Expenditure Surveys.

Two problems immediately surface when deploying bandwidths that are optimal for the conditional PDF in the conditional CDF setting: (i) the rates of convergence of the optimal bandwidths differ in the two settings, and (ii) the optimal constants associated with the bandwidths differ in the two settings. More specifically, let $x = (x^c, x^d)$ denote the covariates, where $x^c = (x_1^c, \ldots, x_q^c)$ and $x^d = (x_1^d, \ldots, x_r^d)$ are the q continuous and r discrete covariates, and let $h = (h_1, \ldots, h_q)$ and $\lambda = (\lambda_1, \ldots, \lambda_r)$ be the corresponding bandwidths. The optimal bandwidths have the following forms: $h_s = c_s n^{-1/\alpha}$ $(s = 1, \ldots, q)$ and $\lambda_s = b_s n^{-1/\beta}$ $(s = 1, \ldots, r)$ for some constants $\alpha, \beta > 0$. α and β differ depending on whether one estimates a conditional CDF or a conditional PDF. Letting α_c and α_p denote the optimal rate constants for estimating a CDF and PDF, respectively, and letting $d = 1/\alpha_p - 1/\alpha_c$, then one can multiply n^{-1/α_p} by a factor n^d to obtain the desired rate of $n^{-1/\alpha_p} n^d = n^{-1/\alpha_c}$. How one estimates the optimal constant α (and β) is a more formidable task. The optimal constants appropriate for PDF estimation can lie far from those for estimating a CDF. In fact, if the optimal rate of a bandwidth h is $h \sim n^{-1/\alpha}$, then the selection of $h = cn^{-1/\alpha}$ will satisfy the optimal rate of convergence for any finite positive constant c. However, the value of c directly impacts the finite sample efficiency of the resulting estimator. Therefore, choosing c optimally is of paramount importance in applied settings.

In this paper we propose a data-driven method for selecting bandwidth parameters optimally when estimating a conditional CDF, and thereby close the open question raised in Li & Racine (2008). The rest of this paper proceeds as follows. In Section 2 we outline the proposed approach when all variables are presumed to be relevant. In Section 3 we consider the empirically relevant case where some of the covariates may in fact be irrelevant but this is not known a priori. Section 4 considers the estimation of conditional quantile functions which constitute an extremely popular estimation methodology (Koenker (2005)) and may be predicated directly on an estimated conditional CDF as proposed by Li & Racine (2008). Section 5 assesses the finite sample performance of the proposed method relative to that employed in Li & Racine (2008) and considers an empirical application. All proofs are relegated to the appendices.

2. Conditional CDF Bandwidth Selection: Relevant Variables

We consider the case for which x is a vector containing mixed discrete and continuous variables. Let $x = (x^c, x^d) \in (S^c, S^d)$, where x^c is a q-dimensional continuous random vector, and where x^d is an r-dimensional discrete random vector. We shall allow for both ordered and unordered discrete datatypes (Li & Racine (2008)). Let x_{is}^d (x_s^d) denote the sth component of x_i^d (x^d), $s = 1, \ldots, r$; $i = 1, \ldots, n$, where n is the sample size. Let λ denote the bandwidth for a discrete variable. For an ordered variable, we use the following kernel:

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

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

$$l(x_{is}^d, x_s^d, \lambda_s) = \begin{cases} 1, & \text{if } x_{is}^d = x_s^d, \\ \lambda_s, & \text{if } x_{is}^d \neq x_s^d. \end{cases} \tag{2}$$

We assume that x_s takes values in $\{0, 1, \ldots, c_s - 1\}$, where $c_s \geq 2$ is a positive integer. We write the product (discrete variable) kernel as $L_\lambda(x_i^d, x^d, \lambda) = \prod_{s=1}^r l(x_{is}^d, x_s^d, \lambda_s)$. The product kernel function used for the continuous variables is given by $W_h(x_i^c, x^c) = \prod_{s=1}^q h_s^{-1} w((x_{is}^c - x_s^c)/h_s)$, where $w(\cdot)$ is a univariate kernel function for a continuous variable. x_{is}^c (x_s^c) denotes the s^{th} component of x_i^c (x^c) and h_s is the bandwidth associated with x_s^c.

The kernel function for the vector of mixed variables $x = (x^c, x^d)$ is simply the product of $W_h(\cdot)$ and $L_\lambda(\cdot)$ which we denote a 'generalised product kernel' given by $K_\gamma(x_i, x) = W_h(x_i^c, x^c) L_\lambda(x_i^d, x^d, \lambda)$, where $\gamma = (h, \lambda)$.

2.1. **The scalar y case.** We use $F(y|x)$ to denote the conditional CDF of Y given $X = x$ and let $f(x)$ denote the marginal density of X. In this paper we consider three estimators that may be of general interest. We will use (a), (b) and (c) to distinguish the three estimators defined below. The first one (e.g., Li & Racine (2008)) smooths the covariates x (but not y) and is given by:

$$\hat{F}_a(y|x) = n^{-1} \sum_{j=1}^{n} \mathbf{I}(y_j \leq y) K_\gamma(x_j, x) / \hat{f}(x), \tag{3}$$

where $\mathbf{I}(A)$ denotes an indicator function that assumes the value 1 if A occurs and 0 otherwise, where $\hat{f}(x) = n^{-1} \sum_{j=1}^{n} K_\gamma(x_j, x)$ is the kernel estimator of the design density $f(x)$.

The advantage of using $\hat{F}_a(y|x)$ to estimate $F(y|x)$ is that it is applicable whether y_j is a continuous or a discrete variable.

The second estimator proposed by Li and Racine smooths the dependent variable y_j (assuming that y_j is a continuous variable) and is defined by

$$\hat{F}_b(y|x) = n^{-1} \sum_{j=1}^{n} G((y - y_j)/h_0) K_\gamma(x_j, x) / \hat{f}(x), \tag{4}$$

where $G(\cdot)$ is a CDF function defined by $G(v) = \int_{-\infty}^{v} w(u)du$ (because $w(\cdot)$ is a kernel density function), h_0 is the bandwidth associated with y.

When y is a discrete variable, we propose a third estimator that also smooths both x and y using a discrete support kernel for y,

$$\hat{F}_c(y|x) = n^{-1} \sum_{j=1}^{n} \mathcal{L}(y_j, y, \lambda_0) K_\gamma(x_j, x) / \hat{f}(x), \tag{5}$$

where $\mathcal{L}(y_j, y, \lambda_0) = \sum_{z \leq y} l(y_j, z, \lambda_0)$ is the cumulative discrete kernel function based on (1) or (2) depending on whether y is an ordered or an unordered discrete variable.

In all three cases we suggest choosing bandwidths by minimizing the following cross-validation function,

$$CV(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \int \left\{ \mathbf{I}(y_i \leq y) - \hat{F}_{-i}(y|x_i) \right\}^2 \mathcal{M}(x_i) M(y) dy, \tag{6}$$

where $\mathcal{M}(\cdot)$ and $M(\cdot)$ are trimming functions with bounded support. If y is a discrete variable, then one should replace $\int dy$ by $\sum_{y \in D_y}$ in (6), where D_y is the support of y_i (discrete), and

$$\hat{F}_{-i}(y|x_i) = \begin{cases} \hat{F}_{a,-i}(y|x_i) \stackrel{def}{=} n^{-1} \sum_{j \neq i}^{n} \mathbf{I}(y_j \leq y) K_\gamma(x_j, x_i) / \hat{f}_{-i}(x_i) & \text{for case } (a), \\ \hat{F}_{b,-i}(y|x_i) \stackrel{def}{=} n^{-1} \sum_{j \neq i}^{n} G((y - y_j)/h_0) K_\gamma(x_j, x_i) / \hat{f}_{-i}(x_i) & \text{for case } (b), \\ \hat{F}_{c,-i}(y|x_i) \stackrel{def}{=} n^{-1} \sum_{j \neq i}^{n} \mathcal{L}(y_j, y, \lambda_0) K_\gamma(x_j, x_i) / \hat{f}_{-i}(x_i) & \text{for case } (c), \end{cases} \tag{7}$$

is the leave-one-out estimator of $F(y|x_i)$, while $\hat{f}_{-i}(x_i) = (n-1)^{-1} \sum_{j \neq i} K_\gamma(x_j, x_i)$ is the leave-one-out estimator of the design density. Again, note that for case (a), y can be either a continuous or a discrete variable, while case (b) applies only to continuous y and case (c) only to discrete y.

For the proofs below we make the following assumptions.

Condition 1. *$\{X_j, Y_j\}_{i=1}^{n}$ are independent and identically distributed as (X, Y), $f(x)$ and $F(y|x)$ have uniformly continuous third-order partial derivative functions with respect to x^c and y (if y is a continuous variable).*

Condition 2. *$w(\cdot)$ is a non-negative, symmetric and bounded second order kernel function with $\int w(v)|v|^4 dv$ being a finite constant.*

Condition 3. *As $n \to \infty$, $h_s \to 0$ for $s = 0, 1, \ldots q$, $\lambda_s \to 0$ for $s = 0, 1, \ldots r$, $n(h_1 \ldots h_q) \to \infty$.*

We will first present results on the leading terms of $CV(\cdot)$, and for this we need to obtain leading bias and variance terms. To describe the leading bias term associated with the discrete variables, we need to introduce some notation. When x_s^d is an unordered categorical variable, define an indicator function $\mathbf{I}_s(\cdot,\cdot)$ by

$$\mathbf{I}_s(x^d, z^d) = \mathbf{I}(x_s^d \neq z_s^d) \prod_{t\neq s}^{r} \mathbf{I}(x_t^d = z_t^d).$$

$\mathbf{I}_s(x^d, z^d)$ equals 1 if and only if x^d and z^d differ only in their sth component, and is zero otherwise. For notational simplicity, when x_s^d is an ordered categorical variable, we shall assume that x_s^d assumes (finitely many) consecutive integer values, and $\mathbf{I}_s(\cdot,\cdot)$ is defined by

$$\mathbf{I}_s(x^d, z^d) = \mathbf{I}(|x_s^d - z_s^d| = 1) \prod_{t\neq s}^{r} \mathbf{I}(x_t^d = z_t^d).$$

Note that $\mathbf{I}_s(x^d, z^d)$ equals 1 if and only if x^d and z^d differ by one unit only in the sth component, and is zero otherwise.

For $s = 1, \ldots, q$, let $F_s(y|x) = \partial F(y|x)/\partial x_s$ and $F_{ss}(y|x) = \partial^2 F(y|x)/\partial x_s^2$. Let $F_0(y|x) = \partial F(y|x)/\partial y$, $F_{00}(y|x) = \partial^2 F(y|x)/\partial y^2$, $\kappa_2 = \int w(v)v^2 dv$, and $\nu_0 = \int W(v)^2 dv$.

The next theorem gives the leading terms for $CV(\cdot)$[1].

Theorem 2.1. *Letting $CV(\gamma)$ be defined in (6) and also assuming that conditions (C1) to (C3) hold, then the leading term of $CV(\cdot)$ is given by $CV_L(\cdot)$, which is defined as follows (where $\int dx = \sum_{x^d\in D_x} \int dx^c$, D_x is the support of x_i^d):*

For case (a) (no smoothing for y)

$$CV_{a,L}(\gamma) = \iint \left\{ \left[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \right]^2 + \frac{\Sigma_{y|x}}{n h_1 \cdots h_q} \right\} f(x)\mathcal{M}(x)M(y)dxdy,$$

while if y is discrete, $\int dy$ above needs to be replaced by $\sum_{y\in D_y}$.

For case (b) (smoothing for continuous y)

$$CV_{b,L}(\gamma) = \iint \left\{ \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \right]^2 + \frac{\Sigma_{y|x} - h_0\Omega_1}{n h_1 \cdots h_q} \right\} f(x)\mathcal{M}(x)M(y)dxdy.$$

For case (c) (discrete support smoothing for discrete y)

$$CV_{c,L}(\gamma) = \sum_{y\in D_y} \int \left\{ \left[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=0}^{r} \lambda_s B_{2s}(y|x) \right]^2 + \frac{\Sigma_{y|x} + \lambda_0\Omega_2}{n h_1 \cdots h_q} \right\} f(x)\mathcal{M}(x)dx,$$

where $B_{10} = \frac{\kappa_2}{2} F_{00}(y|x)$, $B_{1s}(y|x) = \frac{\kappa_2}{2}\big[f(x)F_{ss}(y|x) + 2f_s(x)F_s(y|x)\big]/f(x)$, for $s = 1, \ldots, q$, $B_{20}(y|x) = C_y - F(y|x)$, where $C_y = \sum_{z\le y} 1$. $B_{2s}(y|x) = \sum_{z^d\in S^d} \mathbf{I}_s(z^d, x^d)\big[F(y|x^c, z^d) - F(y|x)\big]f(x^c, z^d)/f(x)$, for $s = 1, \ldots, r$, $\Sigma_{y|x} = \nu_0[F(y|x) - F(y|x)^2]/f(x)$, $\Omega_1 = \nu_0 C_w F_0(y|x)/f(x)$, $\Omega_2 = 2\nu_0\big[F(y|x)^2 - F(y|x)\big]/f(x)$, $C_w = 2\int G(v)w(v)vdv$.

Theorem 2.1 is proved in Appendix A.

For $\hat{F}_a(y|x)$ defined in (3), Li & Racine (2008) have shown that the estimation MSE has the following leading term,

$$MSE_L[\hat{F}_a(y|x)] = \Big[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \Big]^2 + \frac{\Sigma_{y|x}}{n h_1 \cdots h_q}, \tag{8}$$

[1]When we say that CV_L is the leading term of CV, it means that $CV = CV_L + (s.o.)$, where $(s.o.)$ denotes terms having probability order smaller than CV_L and terms unrelated to the bandwidths.

Comparing $CV_{a,L}(\cdot)$ given in case (a) of Theorem 2.1 with (8), we observe that

$$CV_{a,L} = \iint MSE_L[\hat{F}_a(y|x)]f(x)\mathcal{M}(x)M(y)dxdy.$$

Hence, the CV selected bandwidth is asymptotically optimal because the leading term from the CV function equals the leading term of the weighted integrated estimation MSE. Therefore, the CV selected bandwidths lead to an estimator that minimizes a weighted integrated MSE. Similar results hold true for cases (b) and (c).

Using the results of Theorem 2.1 we obtain the main result of the paper which describes the asymptotic behavior of CV selected bandwidths.

Theorem 2.2. *Under conditions (C1) - (C3), we have*

(i) $n^{1/(4+q)}\hat{h}_s \xrightarrow{p} a_s^0$, $s = 1, \ldots, q$;
(ii) $n^{2/(4+q)}\hat{\lambda}_s \xrightarrow{p} b_s^0$, $s = 1, \ldots, r$;
(iii) $n^{1/(4+q)}\hat{h}_0 \xrightarrow{p} a_0^0$,
(iv) $n^{2/(4+q)}\hat{\lambda}_0 \xrightarrow{p} b_0^0$,

where a_s^0 $(s = 1, \ldots, q)$ *are positive constants,* a_0^0, *and* b_s^0 $(s = 0, 1, \ldots, r)$ *are non-negative constants.*

Note that Theorem 2.2 should be understood as follows: Results (i) and (ii) are relevant for case (a) because in case (a) we do not smooth y, hence there are no bandwidths involved for y (i.e. $\hat{h}_0$ and $\hat{\lambda}_0$) for case (a). Similarly, (i) to (iii) apply to case (b) (continuous y), while (i), (ii) and (iv) apply to case (c) (discrete y).

The results of Theorem 2.2 can be interpreted as follows. If one defines some optimal non-stochastic bandwidths, say $h_s^0 = a_s^0 n^{-1/(4+q)}$ and $\lambda_s^0 = b_s^0 n^{-2/(4+q)}$, that minimize the leading terms of the weighted integrated estimation MSE (with weight function given by $\mathcal{M}(x)M(y)$), and we write $\hat{h}_s = \hat{a}_s n^{-1/(4+q)}$ and $\hat{\lambda}_s = \hat{b}_s n^{-2/(4+q)}$, then we have $\hat{a}_s \xrightarrow{p} a_s^0$ and $\hat{b}_s \xrightarrow{p} b_s^0$. Thus, the CV selected bandwidths are asymptotically equivalent to the optimal non-stochastic bandwidths.

Using the results of Theorem 2.2, we obtain the following asymptotic normality result for $\hat{F}(y|x)$.

Theorem 2.3. *Under conditions (C1) - (C3), we have*

$$\sqrt{n\hat{h}_1 \cdots \hat{h}_q}\Big[\hat{F}(y|x) - F(y|x) - \sum_{s=0}^{q} \hat{h}_s^2 B_{1s}(y|x) - \sum_{s=0}^{r} \hat{\lambda}_s B_{2s}(y|x)\Big] \xrightarrow{d} N(0, \Sigma_{y|x}), \tag{9}$$

where for case (a), $\hat{h}_0^2 B_{10}(y|x)$ *and* $\hat{\lambda}_0 B_{20}(y|x)$ *should be removed from equation (9) as there is no* $\hat{h}_0$ *and* $\hat{\lambda}_0$ *for case* (a). *Similarly,* $\hat{\lambda}_0 B_{20}(y|x)$ *and* $\hat{h}_0^2 B_{10}(y|x)$ *should be removed for cases* (b) *and* (c), *respectively.*

One problem with the $CV(\cdot)$ function defined in (6) is that it involves (numerical) integration, which can be computationally prohibitive. Below we propose an alternative cross-validation function which replaces the integration over y by a sample average over the $y_j s$. Therefore, one can also choose the bandwidths by minimizing the following alternative cross-validation objective function:

$$CV_\Sigma(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{n-1} \sum_{j \neq i}^{n} \Big[\mathbf{I}\,(y_i \leq y_j) - \hat{F}_{-i}(y_j|x_i)\Big]^2 \mathcal{M}_i = \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j \neq i}^{n} \Big[\mathbf{I}\,(y_i \leq y_j) - \hat{F}_{-i}(y_j|x_i)\Big]^2 \mathcal{M}_i, \tag{10}$$

where $\mathcal{M}_i = \mathcal{M}(X_i)$ is the same weight function used in (6).

The advantage of using (10) is that it is less computationally onerous as it does not involve (numerical) integration.

It can be shown that the asymptotic behavior of the bandwidths selected by minimizing (10) is similar to those described by Theorem 2.2, while the resulting estimator has the same asymptotic distribution as described in Theorem 2.3.

Theorem 2.4. *If one chooses $M(y) = g(y)$, where $g(y)$ is the marginal density (probability function) of y (y can be either continuous or discrete), then $CV(\gamma)$ defined in (6) and $CV_\Sigma(\gamma)$ defined in (10) are asymptotically equivalent in the sense that*

$$CV_{\Sigma,L}(\gamma) = CV_L(\gamma),$$

where $CV_{\Sigma,L}$ is the leading term of $CV_\Sigma(\gamma)$, CV_L is the leading term of $CV(\gamma)$.

A sketch of the proof of Theorem 2.4 is given in Appendix A.

From Theorem 2.4 we immediately obtain the following useful results.

Theorem 2.5. *If one chooses the bandwidths by minimizing $CV_\Sigma(\cdot)$, then Theorem 2.2 and Theorem 2.3 remain valid with the only modification being that one replaces $M(\cdot)$ by $g(\cdot)$.*

The conclusion of Theorem 2.5 follows directly from theorems 2.2, 2.3 and 2.4. Therefore, its proof is omitted.

2.2. **The Multivariate y Case.** When y is multivariate we write $y = (y_1, \ldots, y_p) = (y_1^c, \ldots, y_{q_y}^c, y_1^d, \ldots, y_{r_y}^d)$ which is of dimension $p = q_y + r_y$, where the first q_y are continuous variables and the last r_y are discrete ones. Our method outlined earlier can be generalized to cover the multivariate y case in a straightforward manner. We consider two estimators for multivariate y, one that does not smooth y which we again call case (a) (the subscript m below is taken to mean 'multivariate' y),

$$\hat{F}_{m,a}(y|x) = n^{-1}\sum_{j=1}^{n} \mathbf{I}(y_j \le y) K_\gamma(x_j, x)/\hat{f}(x), \tag{11}$$

where $\mathbf{I}(y_j \le y) = \prod_{s=1}^{p} \mathbf{I}(y_{js} \le y_s)$ is the product of indicator functions, while the second estimator smooths both x and y (call it case (b)),

$$\hat{F}_{m,b}(y|x) = n^{-1}\sum_{j=1}^{n} \mathcal{K}(y_j, y, \gamma_0) K_\gamma(x_j, x)/\hat{f}(x), \tag{12}$$

where $\mathcal{K}(y_j, y, \gamma_0) = G\left(\frac{y^c - y_j^c}{h_0}\right)\mathcal{L}(y_j^d, y^d, \lambda_0)$, $G\left(\frac{y^c - y_j^c}{h_0}\right) = \prod_{s=1}^{q_y} G\left(\frac{y_s^c - y_{js}^c}{h_{0,s}}\right)$ and $\mathcal{L}(y_j^d, y^d, \lambda_0) = \prod_{s=1}^{r_y} \mathcal{L}(y_{js}^d, y_s^d, \lambda_{0,s})$. We again propose selecting bandwidths via leave-one-out cross-validation by minimizing (where $\int dy = \sum_{y^d \in D_y} \int dy^c$)

$$CV_m = n^{-1}\sum_{i=1}^{n} \int \left\{\mathbf{I}(y_i \le y) - \hat{F}_{-i}(y|x_i)\right\}^2 \mathcal{M}(x_i) M(y) dy, \text{ or}$$

$$CV_{m,\Sigma} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \left\{\mathbf{I}(y_i \le y_j) - \hat{F}_{-i}(y_j|x_i)\right\}^2 \mathcal{M}_i,$$

where $\hat{F}_{-i}(y|x_i)$ is the leave-one-out estimator of $F(y|x_i)$ and it can be either $\hat{F}_{m,a,-i} = (n-1)^{-1}\sum_{j\ne i} \mathbf{I}(y_j \le y) K_\gamma(x_j, x_i)/\hat{f}_{-i}(x_i)$ or $\hat{F}_{m,b,-i}(y|x_i) = (n-1)^{-1}\sum_{j\ne i}^{n} \mathcal{K}(y_j, y, \gamma_0) K_\gamma(x_j, x_i)/\hat{f}_{-i}(x_i)$, and $\hat{F}_{-i}(y_j|x_i)$ is obtained from $\hat{F}_{-i}(y|x_i)$ with y replaced by y_j.

For case (a) (no smoothing of y), it is easy to show that theorems 2.2 and 2.3 (case (a)) remain valid except now that $F(y|x)$ is understood to be $F(y_1, \ldots, y_p|x)$.

For case (b), Theorem 2.2 is modified by replacing $n^{-1/(4+q)}\hat{h}_0 \xrightarrow{p} a_0^0$ by $n^{-1/(4+q)}\hat{h}_{0,s} \xrightarrow{p} a_{0,s}^0$ for $s = 1, \ldots, q_y$, and replacing $n^{-2/(4+q)}\hat{\lambda}_0 \xrightarrow{p} b_0^0$ by $n^{-2/(4+q)}\hat{\lambda}_{0,s} \xrightarrow{p} b_{0,s}^0$ for $s = 1, \ldots, r_y$, where $a_{0,s}^0$ and $b_{0,s}^0$ are non-negative constants.

Finally we present the asymptotic distribution of $\hat{F}_{m,a}(y|x)$ and $\hat{F}_{m,b}(y|x)$ defined in (11) and (12) with CV selected bandwidths.

Condition 4. *As $n \to \infty$, $h_{0,s} \to 0$ for $s = 1, \ldots q_y$, $\lambda_{0,s} \to 0$ for $s = 1, \ldots r_y$.*

Theorem 2.6. *Under (C1) to (C4), we have*

(a): $\sqrt{n\hat{h}_1 \ldots \hat{h}_q}[\hat{F}_{m,a}(y|x) - F(y|x) - \sum_{s=1}^{q} \hat{h}_s^2 B_{1s}(y|x) - \sum_{s=1}^{r} \hat{\lambda}_s B_{2s}(y|x)] \xrightarrow{d} N(0, \Sigma_{y|x})$;

(b): $\sqrt{n\hat{h}_1 \ldots \hat{h}_q}[\hat{F}_{m,b}(y|x) - F(y|x) - \sum_{s=1}^{q} \hat{h}_s^2 B_{1s}(y|x) - \sum_{s=1}^{q_y} \hat{h}_{0,s}^2 B_{0,1s}(y|x)$
$- \sum_{s=1}^{r} \hat{\lambda}_s B_{2s}(y|x) - \sum_{s=1}^{r_y} \hat{\lambda}_{0,s} B_{0,2s}(y|x)] \xrightarrow{d} N(0, \Sigma_{y|x})$,

where the definition of $B_{1s}(y|x)$, $B_{2s}(y|x)$ and $\Sigma_{y|x}$ are the same as defined in Theorem 2.1 except that now $y = (y_1, \ldots, y_p)$, where the definitions of $B_{0,1s}(y|x)$ and $B_{0,2s}(y|x)$ are defined in Appendix A (in the proof of Theorem 2.6).

A sketch of the proof of Theorem 2.6 is given in Appendix A.

3. Conditional CDF Bandwidth Selection in the Presence of Irrelevant Regressors

Next, we consider the case for which one or more of the regressors may be irrelevant, which can occur surprisingly often in practice. Without loss of generality, we assume that only the first q_1 ($1 \leq q_1 \leq q$) components of x^c and the first r_1 ($0 \leq r_1 \leq r$) components of x^d are "relevant" regressors in the sense defined below. Let $\bar{x}$ consist of the first q_1 relevant components of x^c and the first r_1 relevant components of x^d, and let $\tilde{x} = x \setminus \bar{x}$ denote the remaining irrelevant components of x. We assume there exists at least one relevant continuous variable (i.e. $q_1 \geq 1$).

Similar to the definition given in Hall, Li & Racine (2007), we shall assume that

$$\bar{x}, y \text{ is independent of } \tilde{x}. \tag{13}$$

Assumption (13) is quite strong as it requires independence not only between $\tilde{x}$ and y but also between $\tilde{x}$ and $\bar{x}$. A weaker assumption would be to require that

$$\text{Conditional on } \bar{x}, \text{ the variable } \tilde{x} \text{ and } y \text{ are independent.} \tag{14}$$

However, using (14) will cause some technical difficulties for the proof of our main result. Therefore, in the paper we will only consider unconditional independence given in (13) though we point out that extensive simulations carried out by Hall et al. (2007) indicate that all results indeed follow under (14).

For ease of presentation we will focus on the CDF estimator $\hat{F}_a(y|x)$ first. We generalize our conclusion to include cases $\hat{F}_b(y|x)$ and $\hat{F}_c(y|x)$ in the end of this section. Note that the conditional CDF of $F(y|x)$ is $F(y|\bar{x})$. This is because under the assumption of (13), we get $F(y|x) = E[\mathbf{I}(y_i \leq y)|x_i = x] = E[\mathbf{I}(y_i \leq y)|\bar{x}_i = \bar{x}] = F(y|\bar{x})$. We shall consider the case for which the exact number of relevant variables is unknown, and where one estimates the conditional CDF based upon (possibly) a larger set of regressors $x = (\bar{x}, \tilde{x})$, still using equation (3). We use $f(x)$ to denote the joint density function of $x = (x^c, x^d)$, and we use $\bar{f}(\bar{x})$ and $\tilde{f}(\tilde{x})$ to denote the marginal densities of $\bar{x}_i$ and $\tilde{x}_i$, respectively.

We impose similar conditions on the bandwidth and kernel functions as Hall et al. (2007). Define

$$H = \left(\prod_{s=1}^{q_1} h_s \right) \prod_{s=q_1+1}^{q} \min(h_s, 1). \tag{15}$$

Letting $0 < \epsilon < 1/(p+4)$ and for some constant $c > 0$, we further assume that

$$\begin{aligned} &n^{\epsilon-1} \leq H \leq n^{-\epsilon}; n^{-c} < h_s < n^c \text{ for all } s = 1, \ldots, q; \text{ the kernel } w(\cdot) \text{ is a symmetric,} \\ &\text{compactly supported, Hölder-continuous probability density;} \\ &\text{and } w(0) > w(\delta) \text{ for all } \delta > 0. \end{aligned} \tag{16}$$

The above conditions basically ask that each h_s does not converge to zero, or to infinity, too fast, and that $nh_1 \ldots h_{q_1} \to \infty$ as $n \to \infty$ ($h_0 \to 0$ and $\lambda_0 \to 0$ as $n \to \infty$ will be always assumed throughout this paper).

We use $\mathcal{H}$ to denote the permissible set for $(h_1, \ldots, h_q)$ that satisfies (16). The range for $(\lambda_1, \ldots, \lambda_r)$ is $[0,1]^r$, and we use $\Gamma = \mathcal{H} \times [0,1]^r$ to denote the range for the bandwidth vector $\gamma \equiv (h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$. We maintain the assumption that $h_0 \to 0$ and $\lambda_0 \to 0$ as $n \to \infty$.

We expect that, as $n \to \infty$, the bandwidths associated with the relevant covariates will converge to zero, while those associated with the irrelevant covariates will not. It would be convenient to further assume that $h_s \to 0$ for $s = 1, \ldots, q_1$, and that $\lambda_s \to 0$ for $s = 1, \ldots, r_1$. However, for practical reasons we choose not to assume that the relevant components are known a priori, but rather assume that assumption (19) given below holds. We write $K_{\gamma,ij} = \bar{K}_{\bar{\gamma},ij} \tilde{K}_{\tilde{\gamma},ij}$, where $\bar{\gamma} = (h_1, \ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{r_1})$, and $\tilde{\gamma} = (h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r)$ so that $\bar{K}$ and $\tilde{K}$ are the product kernels associated with the relevant and the irrelevant covariates, respectively. We define

$$\eta(y, \bar{x}) = \bar{f}(\bar{x})^{-1} E\big[(F(y|\bar{x}_j) - F(y|\bar{x}_i))\bar{K}_{\bar{\gamma},ji} | \bar{x}_i = \bar{x}\big]. \tag{17}$$

Note that $\eta(y, \bar{x})$ defined above only depends on the bandwidths associated with the relevant covariates, that is, it is unrelated to $(\tilde{h}, \tilde{\lambda})$, the bandwidths associated with the irrelevant covariates.

Define

$$\bar{\mathcal{M}}(\bar{x}) = \int \tilde{f}(\tilde{x}) \mathcal{M}(x) d\tilde{x}. \tag{18}$$

We assume that

$$\begin{aligned} &\iint [\eta(y, \bar{x})]^2 \bar{f}(\bar{x}) \bar{\mathcal{M}}(\bar{x}) M(y) d\bar{x} dy, \text{ as a function of } h_1, \ldots, h_{q_1} \text{ and } \lambda_1, \ldots, \lambda_{r_1}, \\ &\text{vanishes if and only if all of the bandwidths vanish.} \end{aligned} \tag{19}$$

In Lemma B.4 in Appendix B we show that (16) and (19) imply that as $n \to \infty$, $h_s \to 0$ for $s = 1, \ldots, q_1$ and $\lambda_s \to 0$ for $s = 1, \ldots, r_1$. Therefore, the bandwidths associated with the relevant covariates all vanish asymptotically. In Appendix B, we also show that $h_s \to \infty$ for all $s = q_1 + 1, \ldots, q$ and $\lambda_s = 1$ for all $s = r_1 + 1, \ldots, r$. This means that all irrelevant variables will be smoothed out asymptotically. Therefore, the leading term of CV is the same as the result in Theorem 2.1 except that one has q_1 and r_1 replacing q and r in Theorem 2.1. This leads to the following main result of this section.

Theorem 3.1. *In additional to conditions (C1) to (C4), assume that conditions (16), (19) and (B.10) also hold, and let $\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_r$ denote the bandwidths that minimize $CV_a(\gamma)$. Then*

$$\begin{aligned} &n^{1/(q_1+4)} \hat{h}_s \to a_s^0 \text{ in probability for } 1 \le s \le q_1, \\ &P(\hat{h}_s > C) \to 1 \text{ for } q_1 + 1 \le s \le q \text{ and for all } C > 0, \\ &n^{2/(q_1+4)} \hat{\lambda}_s \to b_s^0 \text{ in probability for } 1 \le s \le r_1, \\ &\hat{\lambda}_s \to 1 \text{ in probability for } r_1 + 1 \le s \le r, \\ &n^{2/(q_1+4)} \hat{h}_0 \to a_0^0 \text{ in probability for } r_1 + 1 \le s \le r, \\ &n^{2/(q_1+4)} \hat{\lambda}_0 \to b_0^0 \text{ in probability.} \end{aligned} \tag{20}$$

Theorem 3.1 states that the bandwidths associated with the irrelevant covariates all converge to their upper bounds, so that, asymptotically, all irrelevant covariates are smoothed out, while the bandwidths associated with the relevant covariates all converge to zero at a rate that is optimal for minimizing asymptotic mean square error (i.e., without the presence of the irrelevant covariates).

Similar to the result given in Section 2, one can show that the leading term of the CV function equals a weighted IMSE (with only relevant covariates used in the estimation). Therefore, the CV method leads to optimal smoothing in the sense of minimizing a weighted IMSE asymptotically.

From Theorem 3.1 one can easily obtain the following result.

Theorem 3.2. *Under the same conditions given in Theorem 3.1, for $x \in$ interior to $S = S^c \times S^d$, then*

$$\sqrt{n\hat{h}_1 \dots \hat{h}_{q1}} \left[\hat{F}_a(y|x) - F(y|\bar{x}) - \sum_{s=1}^{q_1} \hat{h}_s^2 \bar{B}_{1s}(y|\bar{x}) - \sum_{s=0}^{r_1} \hat{\lambda}_s \bar{B}_{2s}(y|\bar{x}) \right] \xrightarrow{d} N(0, \bar{\Sigma}_{y|\bar{x}}), \tag{21}$$

where $\bar{B}_{1s}(y|\bar{x})$ and $\bar{B}_{2s}(y|\bar{x})$ are defined in (B.3) and (B.4), while $\bar{\Sigma}_{y|\bar{x}}$ is defined in (B.5).

Theorem 3.2 shows that the asymptotic normality of the conditional CDF estimator in the presence of irrelevant covariates is the same as the estimator with only relevant covariates.

4. Estimating Conditional Quantile Functions

With the nonparametric conditional CDF estimator in hand, it is straightforward to obtain a conditional quantile estimator. A conditional τ^{th} quantile of y given x is defined by ($\tau \in (0,1)$)

$$q_\tau(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x). \tag{22}$$

Since $F(y|x)$ is (weakly) monotone in y, inverting (22) leads to a unique solution for $q_\tau(x)$. In this section we will focus on using $\hat{F}(y|x)$ to obtain a quantile estimator for $q_\tau(x)$. Therefore, we propose the following estimator for estimating $q_\tau(x)$:

$$\hat{q}_\tau(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\}, \tag{23}$$

where $\hat{F}(y|x)$ can be any of the three estimators discussed in Section 2 with CV selected bandwidths. The CV objective function can be either $CV(\cdot)$ defined in (6) or CV_Σ defined in (10). Of course, $\hat{F}_b(y|x)$ and $\hat{F}_c(y|x)$ are only applicable to continuous and discrete y, respectively.

Because $\hat{F}(y|x)$ is monotone in y, (23) leads to a computationally simple estimator relative to, say, the check function approach where one needs to minimize a nonlinear function in order to obtain an estimator for $q_\tau(x)$.

Because $\hat{F}(y|x)$ lies between zero and one and is monotone in y, $\hat{q}_\alpha(x)$ ($\tilde{g}_\alpha(x)$) always exists. Therefore, once one obtains $\hat{F}(y|x)$, it is trivial to compute $\hat{q}_\alpha(x)$, for example, by choosing q_α to minimize the following objective function,

$$\hat{q}_\alpha(x) = \arg\min_{q_\alpha} | \, \alpha - \hat{F}(q_\alpha|x) \, |. \tag{24}$$

That is, the value of q_α that minimizes (24) gives us $\hat{q}_\alpha(x)$. We make the following assumption.

Condition (C5): The conditional PDF $g_\alpha(y|x)$ is continuous in x^c, $f(q_\alpha(x)|x) > 0$.

We use $f(y|x) \equiv F_0(y|x) = \frac{\partial}{\partial y} F(y|x)$ to denote the conditional PDF of y given x. Below we present the asymptotic distribution of $\hat{q}_\alpha(x)$.

Theorem 4.1. *Define $B_{n,\alpha}(x) = B_n(q_\alpha(x)|x)/f(q_\alpha(x)|x)$, where $B_n(y|x) = [\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=0}^{r} \lambda_s B_{2s}(y|x)]$ is the leading bias term of $\hat{F}(y|x)$ (with $y = q_\alpha(x)$). Then, under (C1) to (C5), we have*

$$(nh_1 \dots h_q)^{1/2}[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)] \to N(0, V_\alpha(x)) \text{ in distribution,}$$

where $V_\alpha(x) = \alpha(1-\alpha)\nu_0/[f^2(q_\alpha(x)|x)f(x)] \equiv V(q_\alpha(x)|x)/f^2(q_\alpha(x)|x)$ (since $\alpha = F(q_\alpha(x)|x)$).

The proof of Theorem 4.1 follows the same arguments as the proof of Theorem 3.1 of Li & Racine (2008) given the results of Theorem 3.2 above. Thus, the proof of Theorem 4.1 is omitted.

5. Monte Carlo Simulations and Empirical Applications

In this section we examine the finite-sample performance of proposed method of cross-validated conditional CDF bandwidth selection.

We numerically minimize the following objective functions:

$$CV(h_y, h_x) = n^{-1} \sum_{i=1}^{n} \int_{-\infty}^{\infty} \left\{ \mathbf{I}(y_i \leq y) - \tilde{F}_{-i}(y|x_i) \right\}^2 dy, \tag{25}$$

$$CV(h_y, h_x) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left\{ \mathbf{I}(y_i \leq y_j) - \tilde{F}_{-i}(y_j|x_i) \right\}^2, \tag{26}$$

where $\mathbf{1}(\cdot)$ is the usual indicator function and where $\tilde{F}_{-i}(\cdot)$ is the leave-one-out kernel estimator.

Having computed the bandwidths we then compute the sample MSE of the estimators of $F(y|x)$ for both the CCDF and CPDF-based bandwidths via

$$MSE = n^{-1} \sum_{i=1}^{n} (F(y_i|x_i) - \tilde{F}(y_i|x_i))^2. \tag{27}$$

5.1. **Comparison of Integral Versus Summation Approach.** We first assess how the integration-based method compares with the summation-based version given in (25) above. We draw 1,000 Monte Carlo replications from a joint normal distribution with correlation ρ for a range of sample sizes That is, $(y, x)' \sim N(\mu, \Sigma)$ with $\mu = (0, 0)'$ and $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. For each replication we conduct cross-validation using both (25) ('CCDF') and that appropriate for PDF estimation ('CPDF').

TABLE 1. Summation-based relative median efficiency of kernel estimators of CCDFs using the proposed CCDF-based bandwidth method versus that appropriate for CPDFs. Numbers less that 1 indicate superior MSE performance.

	$n = 25$	$n = 50$	$n = 75$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0.95$	0.83	0.87	0.89	0.90	0.89	0.91
$\rho = 0.85$	0.88	0.89	0.92	0.90	0.92	0.90
$\rho = 0.75$	0.88	0.90	0.90	0.90	0.92	0.92
$\rho = 0.50$	0.86	0.92	0.91	0.90	0.92	0.89
$\rho = 0.25$	1.01	0.95	0.92	0.89	0.86	0.89
$\rho = 0.00$	1.17	1.09	1.11	1.08	1.07	1.18

TABLE 2. Integration-based relative median efficiency of kernel estimators of CCDFs using the proposed CCDF-based bandwidth method versus that appropriate for CPDFs. Numbers less that 1 indicate superior MSE performance.

	$n = 25$	$n = 50$	$n = 75$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0.95$	0.77	0.81	0.83	0.84	0.88	0.88
$\rho = 0.85$	0.80	0.84	0.85	0.85	0.89	0.91
$\rho = 0.75$	0.80	0.84	0.86	0.86	0.91	0.90
$\rho = 0.50$	0.83	0.85	0.86	0.88	0.87	0.89
$\rho = 0.25$	0.89	0.93	0.89	0.85	0.87	0.88
$\rho = 0.00$	1.02	0.99	1.07	1.03	1.09	1.16

Tables 1 and 2 reveal that a) when there is no relationship between y and x ($\rho = 0$) the bandwidth selector of Hall et al. (2004) performs better in that it has a higher probability of removing the irrelevant variable x (in this case the appropriate h_x is ∞). However, when $\rho \neq 0$ it is seen that the proposed method delivers bandwidths that dominate those based on PDF bandwidth selection in finite-sample settings.

5.2. **Empirical Rates of Convergence of the Cross-Validated Bandwidths.** We can use these simulation results to examine the rate at which the bandwidths h_x and h_y converge to zero empirically when x is relevant ($\rho \neq 0$) by simple regression of the logarithm of the bandwidth on the log of the sample size (the coefficient will be the parameter α in the expression cn^{α}). In particular, we take the median values of, say, h_y for each n in the tables above and conduct a log-log regression of this median on n. The coefficient on $\log(n)$ indicates the rate at which h_y (h_x) approaches zero as $n \to \infty$. For instance, if $h_y \propto n^{-1/3}$, the coefficient on $\log(n)$ would be around -0.33, while if $h_y \propto n^{-1/5}$, the coefficient on $\log(n)$ would be around -0.20. Similarly, if $h_x \propto n^{-1/6}$, then the coefficient on $\log(n)$ would be around -0.17.

For the proposed CDF method we obtain a coefficient on $\log(n)$ of -0.31 ($\approx -1/3$) for h_y and -0.16 ($\approx -1/6$) for h_x. For the PDF method however we obtain $-0.21(\approx -1/5)$ for h_y and -0.15 ($\approx -1/6$) for h_x. These results are in line with the theoretical results presented above and in Hall et al. (2004) which confirms that the proposed method delivers bandwidths that indeed mirror the theoretical rates of convergence.

5.3. **Irrelevant Categorical Covariates.** Next, we take the DGP used above but now add an additional covariate z that is uncorrelated with y but this is not presumed to be known a-priori. Results are presented in Table 3 below. We note that the bandwidth λ_z for the discrete variable takes its upper bound with high probability as it should given that z is 'irrelevant', while the method otherwise continues to perform as expected.

TABLE 3. Irrelevant z summation-based relative median efficiency of kernel estimators of CCDFs using the proposed CCDF-based bandwidth method versus that appropriate for CPDFs. Numbers less that 1 indicate superior MSE performance.

	$n = 25$	$n = 50$	$n = 75$	$n = 100$	$n = 150$	$n = 200$
$\rho = 0.95$	0.79	0.83	0.89	0.88	0.88	0.90
$\rho = 0.85$	0.89	0.90	0.87	0.91	0.89	0.91
$\rho = 0.75$	0.87	0.86	0.89	0.89	0.93	0.91
$\rho = 0.50$	0.91	0.89	0.89	0.92	0.90	0.87
$\rho = 0.25$	1.00	0.97	0.96	0.87	0.86	0.86
$\rho = 0.00$	1.06	0.98	0.93	0.95	0.98	0.95

5.4. **Application - Out-of-Pocket Drug Expenditures.** Prescription drug cost containment is an issue that has been hotly debated in Canada as of late. Canadian provincial government drug subsidy programs have recently begun to change the basis of subsidy from age (age 65+) to financial need (defined as high drug costs relative to income, regardless of age), in an attempt to improve the distributive equity of their programs. We consider using quantile methods in order to assess the distributive features of out-of-pocket prescription drug expenditures.

Our data is taken from the public use versions of the Statistics Canada Family Expenditure Surveys (FAMEX) and the Surveys of Household Spending (SHS), which replaced the FAMEX in 1997. These surveys collect information on annual household level income, spending on various goods and services, including prescription drugs, as well as information on household living arrangements. We consider data for 2008 for British Columbia and restrict attention to households having positive out-of-pocket expenditures for prescription drugs (i.e., who have positive levels of cost sharing) for which there were $n = 679$ households. We make use of the variables prescription drug share ('rxshare'), sex, marital status, age category, and (log) household expenditure ('lrex'). Our dependent variable is rxshare when the remaining being predictors of which only expenditure is continuous while the remaining are categorical. We apply the proposed method of bandwidth selection which is summarized in the following table.

TABLE 4. Bandwidth summary for the prescription drug illustration.

Variable	Bandwidth	$\lambda_{\max}$	c
male	0.753	1	NA
married	0.396	1	NA
hagecat	0.0173	1	NA
lrex	0.382	NA	2.03
rxshare	0.000525	NA	0.15

Table 4 reveals that the categorical variable 'male' receives the most smoothing while that for age 'hagecat' receives substantially less, while the continuous predictor household expenditure 'lrex' receives a fair bit of smoothing, the dependent variable drug share 'rxshare' receiving less. Note that an empirical CDF approach that did not smooth the dependent variable would have $\lambda_{\text{rxshare}} = 0$ hence positive smoothing is deemed appropriate by the proposed method.

Figure 1 presents partial quantile plots for $\tau = \{0.5, 0.6, 0.7, 0.8\}$ which range from median out-of-pocket expenditure (0.5) to the 80th percentile (0.8). For these figures variables not appearing on the axes are held constant at their mode (for categorical predictors) and median (for the continuous predictor).

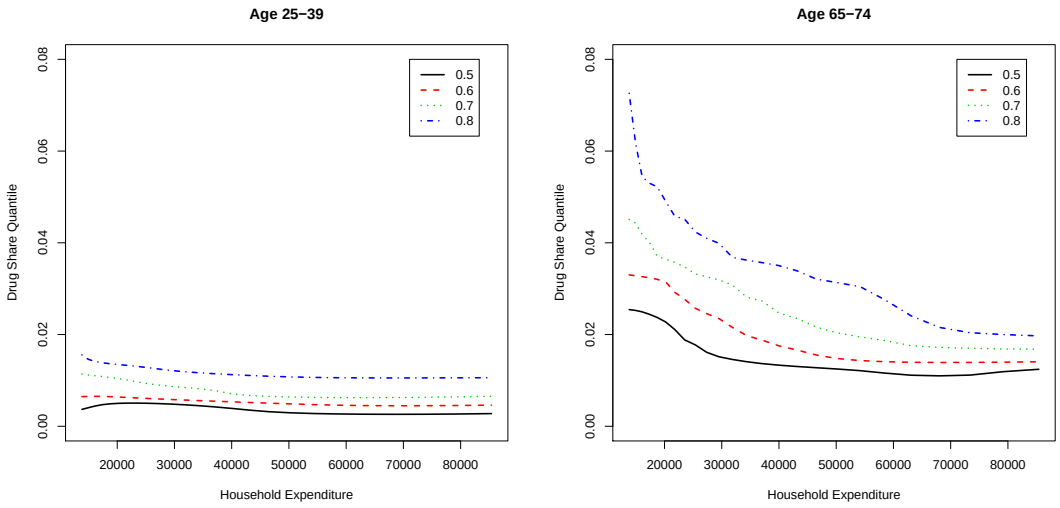

FIGURE 1. The figure on the left represents drug share quantiles for persons aged 25-39, while that on the right is that for persons aged 65-74.

Figure 1 reveals that those aged 25-39 have out-of-pocket drug shares in the neighborhood of around 1% conditional on positive expenditures in this category for all ranges of income. However, for those aged 65-74 the picture is quite different and reveals the regressive nature of out-of-pocket prescription drug expenditures among seniors. Those with higher incomes spend a markedly smaller fraction on out-of-pocket prescription drug costs than those with lower incomes.

6. CONCLUSION

In this paper we have solved an 'open problem', namely, the optimal selection of bandwidths for conditional CDF estimation. However, in this paper we only consider the independent data case. We note that Cai (2002) and Cai & Xu (2008) have considered the problem of estimating conditional quantile functions with weakly dependent data (β-mixing and α-mixing processes), though these authors only

consider the case where all covariates are both relevant and continuous. By combining the methods outlined in this paper with those in Cai (2002) and Cai & Xu (2008), one can readily generalize the results in this paper to cover the weakly dependent time series data case.

We hope that the method proposed in this paper proves useful to those interested in the estimation of conditional CDFs, and in particular to those who wish consistent nonparametric estimates of conditional quantile functions.

References

Aitchison, J. & Aitken, C. G. G. (1976), 'Multivariate binary discrimination by the kernel method', *Biometrika* **63**(3), 413–420.

Bashtannyk, D. M. & Hyndman, R. J. (2001), 'Bandwidth selection for kernel conditional density estimation', *Computational Statistics and Data Analysis* **36**, 279–298.

Cai, Z. (2002), 'Regression quantiles for time series', *Econometric Theory* **18**, 169–192.

Cai, Z. & Xu, X. (2008), 'Nonparametric quantile estimations for dynamic smooth coefficient models', *Journal of the American Statistical Association* **103**(484), 1595–1608.

Chung, Y. & Dunson, D. B. (2009), 'Nonparametric bayes conditional distribution modeling with variable selection', *Journal of the American Statistical Association* **104**(488), 1646–1660.

Fan, J. & Yim, T. H. (2004), 'A crossvalidation method for estimating conditional densities', *Biometrika* **91**(4), 819–834.

Hall, P., Li, Q. & Racine, J. S. (2007), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors', *The Review of Economics and Statistics* **89**, 784–789.

Hall, P., Racine, J. S. & Li, Q. (2004), 'Cross-validation and the estimation of conditional probability densities', *Journal of the American Statistical Association* **99**(468), 1015–1026.

Hyndman, R. J. & Yao, Q. (2002), 'Nonparametric estimation and symmetry tests for conditional density functions', *Journal of Nonparametric Statistics* **18**(3), 439–454.

Koenker, R. (2005), *Quantile Regression*, Cambridge University Press, New York.

Lee, J. (1990), *U-statistics: Theory and practice*, Marcel Dekker, New York.

Li, Q. & Racine, J. S. (2008), 'Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data', *Journal of Business and Economic Statistics* **26**(4), 423–434.

Li, Q. & Zhou, J. (2005), 'The uniqueness of cross-validation selected smoothing parameters in kernel estimation of nonparametric models', *Econometric Theory* **21**(5), 1017–1025.

Masry, E. (1996), 'Multivariate local polynomial regression for time series: uniform strong consistency and rates', *Journal of Time Series Analysis* **17**, 571–599.

Rosenthal, H. P. (1970), 'On the subspace of l^p $(p \geq 1)$ spanned by sequences of independent random variables', *Israel Journal of Mathematics* **8**, 273–303.

Appendix A: Proofs of Theorems 2.1, 2.2 , 2.3 and 2.6

To simplify the derivations that follow, it is necessary to introduce some shorthand notation and preliminary manipulations.

(1) Let $f_i = f(x_i)$, $\hat{f}_{-i} = \hat{f}_{-i}(x_i)$, $K_{\gamma,ji} = K_\gamma(x_j, x_i)$. $\mathbf{I}_i = \mathbf{I}(y_i \leq y)$, $F_i = F(y|x_i)$, $\mathcal{M}_i = \mathcal{M}(x_i)$.

(2) We define $\sum_i = \sum_{i=1}^n$, $\sum\sum_{j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n$, $\sum\sum_{j\neq i}\sum_{l\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n \sum_{l=1,l\neq i}^n$, $\sum\sum\sum_{l\neq j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n \sum_{l=1,l\neq i,l\neq j}^n$, $\sum\sum_{j>i} = \sum_{i=1}^{n-1}\sum_{j>i}^n$, $\sum\sum\sum_{l>j>i} = \sum_{i=1}^{n-2}\sum_{j>i}^{n-1}\sum_{l>j}^n$.

(3) We write $A_n = B_n + (s.o.)$ to denote the fact that B_n is the leading term of A_n, where $(s.o.)$ denotes terms that have orders smaller than B_n. Also, we write $A_n \sim B_n$ to mean that A_n and B_n have the same order of magnitude in probability.

(4) For notational simplicity we often ignore the difference between n^{-1} and $(n-1)^{-1}$ (or $(n-k)^{-1}$ for any fixed finite integer k) simply because this will have no effect on the asymptotic analysis.

(5) Define $|h|^2 = \sum_{s=1}^q h_s^2$, $|\lambda|^2 = \sum_{s=1}^r \lambda_s^2$, $\zeta_{1n} = |h|^2 + |\lambda|$ and $\zeta_n = \zeta_{1n}^2 + (nh_1 \dots h_q)^{-1}$.

In the proofs that follow we make use of U-statistic H-decomposition and Rosenthal's Inequality repeatedly. We present the results below for the reader's convenience.

The H-decomposition for U-Statistics. Let $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ denote the number of combinations obtained by choosing k items from n (distinct) items. Then a general kth order U-statistic $U_{(k)}$ is defined by

$$U_{(k)} = \binom{n}{k}^{-1} \sum_{1 \le i_1 < \cdots < i_k \le n} H_n(x_{i_1}, \ldots, x_{i_k}),$$

where $H_n(x_{i_1}, \ldots, x_{i_k})$ is symmetric in its arguments and $E[H_n^2(x_{i_1}, \ldots, x_{i_k})] < \infty$. In our proofs we will often use the following H-decomposition for a second order U-statistic,

$$U_{(2)} = \theta + \frac{2}{n}\sum_i (H_{ni} - \theta) + \frac{2}{n(n-1)} \sum\sum_{j>i} [H_{n,ij} - H_{ni} - H_{nj} + \theta], \tag{A.1}$$

where $H_{n,ij} = H_n(x_i, x_j)$, $H_{ni} = E[H_{n,ij}|x_i]$ and $\theta = E[H_{n,ij}]$. We will also make use of the H-decomposition for a third order U-statistic,

$$\begin{aligned} U_{(3)} = \theta &+ \frac{3}{n}\sum_i (H_{ni} - \theta) + \frac{6}{n(n-1)} \sum\sum_{j>i} [H_{n,ij} - H_{ni} - H_{nj} + \theta] \\ &+ \frac{6}{n(n-1)(n-2)} \sum\sum\sum_{l>j>i} (H_{n,ijl} - H_{n,ij} - H_{n,jl} - H_{n,li} + H_{ni} + H_{nj} + H_{nl} - \theta), \end{aligned} \tag{A.2}$$

where $H_{n,ijl} = H_n(x_i, x_j, x_l)$, $H_{n,ij} = E[H_{n,ijl}|x_i, x_j]$, $H_{n,i} = E[H_{n,ij}|x_i]$ and $\theta = E[H_{n,ijl}]$. For an H-decomposition for a general k^{th} order U-statistic, see Lee (1990, page 26).

Rosenthal's Inequality. Let $p \ge 2$ be a positive constant and let $x_1, \ldots, x_n$ denote i.i.d random variables for which $E(x_i) = 0$ and $E(|x_i|^p) < \infty$. Then there exists a positive constant (which may depend on p) $C(p)$ such that

$$E\left(\left|\sum_{i=1}^n x_i\right|^p\right) \le C(p)\left\{\sum_{i=1}^n E(|x_i|^p) + \left[\sum_{s=1}^n E(x_i^2)\right]^{p/2}\right\}. \tag{A.3}$$

Equation (A.3) is known as Rosenthal's Inequality (Rosenthal (1970)).

Proof of Theorem 2.1 Case (a). Define $\hat{F}_{-i} = \hat{F}_{a,-i}(y|x_i)$. We need to show that $CV_a(\cdot) = CV_{a,L}(\cdot) + (s.o.)$, where $(s.o.)$ contains terms unrelated to bandwidths or terms having smaller order than $CV_{a,L}(\cdot)$. Also, the smaller order terms are uniformly small for all $\gamma \in \Gamma$ (as defined in Section 3). We rewrite (6) as (by adding/subtracting terms),

$$\begin{aligned} CV_a(\cdot) &= \frac{1}{n}\sum_i \int \left(\hat{F}_{-i} - F_i + F_i - \mathbf{I}_i\right)^2 \mathcal{M}_i M(y) dy \\ &= \frac{1}{n}\sum_i \int \left[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i) + (F_i - \mathbf{I}_i)^2\right] \mathcal{M}_i M(y) dy. \end{aligned}$$

Since $n^{-1}\sum_i \int (F_i - \mathbf{I}_i)^2 \mathcal{M}_i M(y)dy$ is unrelated to the bandwidths, it follows that minimizing $CV_a(\cdot)$ over $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ is equivalent to minimizing $CV_{a,1}(\cdot)$, where $CV_{a,1}(\cdot)$ is defined as

$$\begin{aligned} CV_{a,1}(\cdot) &= \frac{1}{n}\sum_i \int \big[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\big]\mathcal{M}_i M(y)dy \\ &= \int \Big[\frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}\int (\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/\hat{f}^2_{-i} \\ &\quad - \frac{2}{n(n-1)}\sum\sum_{j\neq i}\int (\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}/\hat{f}_{-i}\Big]\mathcal{M}_i M(y)dy \\ &= \int (S_{1n} - 2S_{2n})M(y)dy \qquad \text{(A.4)} \end{aligned}$$

where $S_{1n} = \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\mathcal{M}_i/\hat{f}^2_{-i}$, $S_{2n} = \frac{1}{n(n-1)}\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}\mathcal{M}_i/\hat{f}_{-i}$.

Lemma A.1 and Lemma A.2 show that (recall that $\zeta_n = |h|^4 + |\lambda|^2 + (nh_1\ldots h_q)^{-1}$)

$$\begin{aligned} S_{1n} &= \int \Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big\}^2 f(x)\mathcal{M}(x)dx \\ &\quad + \int \frac{\Sigma_{y|x}}{nh_1\ldots h_q} f(x)\mathcal{M}(x)dx + o_p(\zeta_n) \qquad \text{(A.5)} \\ S_{2n} &= O_p\big((n^{-1/2}\zeta_n) + (n(h_1\ldots h_q)^{1/2})^{-1}\big). \qquad \text{(A.6)} \end{aligned}$$

Combining (A.4), (A.5) and (A.6), we have shown that

$$CV_a(\cdot) = CV_{a,L}(\cdot) + (s.o.),$$

where $CV_{a,L}$ is defined in Theorem 2.1, $(s.o.)$ denotes terms having probability order (uniformly) smaller than $CV_{a,L}$ and a term that is unrelated to bandwidths.

This completes the proof of case (a) of Theorem 2.1. □

A technical difficulty in handling (A.4) arises from the presence of the random denominator $\hat{f}_{-i} = \hat{f}_{-i}(X_i)$. We will use the following identity to handle the random denominator:

$$\frac{1}{\hat{f}_{-i}} = \frac{1}{f_i} + \frac{f_i - \hat{f}_{-i}}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3} + \frac{(f_i - \hat{f}_{-i})^3}{f_i^3\hat{f}_{-i}}. \qquad \text{(A.7)}$$

Define $f_{i,0} = (n-1)^{-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\mathbf{I}(x_j^d = x_i^d)$, $f_{i,1s} = (n-1)^{-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\mathbf{I}_s(x_j^d, x_i^d)$. We have uniformly in $1 \le i \le n$,

$$\begin{aligned} f_i - \hat{f}_{-i} &= f_i - \frac{1}{n-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)L_\lambda(x_j^d, x_i^d, \lambda) \\ &= f_i - \frac{1}{n-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\big[\mathbf{I}(x_j^d = x_i^d) + \sum_{s=1}^r \lambda_s\ \mathbf{I}_s(x_j^d, x_i^d) + O(|\lambda|^2)\big] \\ &= (f_i - f_{i,0}) - \sum_{s=1}^r \lambda_s f_{i,1s} + O_p(|\lambda|^2). \qquad \text{(A.8)} \end{aligned}$$

Let $\mathcal{S}$ denote the intersection of the support of X_i and the support of the trimming set $\mathcal{M}(X_i)$. Then equation (A.8) implies that, uniformly in $1 \leq i \leq n$ and in $x \in \mathcal{S}$,

$$f_i - \hat{f}_{-i} = O_p\left(\frac{(ln(n))^{1/2}}{(nh_1 \dots h_q)^{1/2}} + |h|^2 + |\lambda|\right), \tag{A.9}$$

because $\sup_{1\leq i\leq n} |f_i - \hat{f}_{-i}| \leq \sup_{x\in S} |f(x) - n^{-1}\sum_i W_h(x_j^c, x_i^c)\mathbf{I}(x_j^d = x_i^d)| + O(n^{-1}) = O_p\Big(\frac{(ln(n))^{1/2}}{(nh_1\dots h_q)^{1/2}} + |h|^2\Big)$ (because S is bounded) and $\sup_{1\leq i\leq n} |f_{i,1s}| = O_p(1)$.

Therefore, we have

$$|f_i - \hat{f}_{-i}|^3 = O_p\left(\left(\frac{ln(n)}{nh_1 \dots h_q}\right)^{3/2} + |h|^6 + |\lambda|^3\right) = o\left((nh_1 \dots h_q)^{-1} + |h|^4 + |\lambda|^2\right). \tag{A.10}$$

Substituting (A.8) and (A.10) into (A.7), we obtain uniformly in $1 \leq i \leq n$ and $x \in \mathcal{S}$

$$\frac{1}{\hat{f}_{-i}} = \frac{1}{f_i} + \frac{(f_i - \hat{f}_{-i})}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3} + o\left((nh_1 \dots h_q)^{-1} + |h|^4 + |\lambda|^2\right). \tag{A.11}$$

From (A.11), we also obtain uniformly in $1 \leq i \leq n$ and $x \in \mathcal{S}$

$$\frac{1}{\hat{f}_{-i}^2} = \frac{1}{f_i^2} + \frac{2(f_i - \hat{f}_{-i})}{f_i^3} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^4} + o\left((nh_1 \dots h_q)^{-1} + |h|^4 + |\lambda|^2\right). \tag{A.12}$$

Both (A.11) and (A.12) will be used to handle the random denominator in the proofs that follow.

Lemma A.1. *Equation (A.5) holds true.*

Proof. We omit the weighted function $\mathcal{M}_i$ for notational simplicity. Define S_{1n}^0 by replacing $\hat{f}_{-i}^{-1}$ in S_{1n} with f_i^{-1}. We will show that (A.5) holds true with S_{1n} being replaced by S_{1n}^0 and that $S_{1n} - S_{1n}^0 = o_p(\zeta_n)$.

$$\begin{aligned}
S_{1n}^0 &= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2 \\
&= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)^2K_{\gamma,ji}^2/f_i^2 \\
&\quad + \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2 \\
&= S_{1n,1} + S_{1n,2}, \qquad \text{(A.13)}
\end{aligned}$$

where the definitions of $S_{1n,1}$ and $S_{1n,2}$ should be apparent.

First, we consider $S_{1n,2}$, which can be written as a third-order U-statistic. $S_{1n,2} = 1/(n(n-1)^2)\sum\sum\sum_{l\neq j\neq i} Q_{ijl}$, where Q_{ijl} is a symmetrized version of $(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2$. Define $Q_{ij} = E(Q_{ijl}|x_i, x_j)$ and $Q_i = E(Q_{ijl}|x_i)$. Then by U-statistic H-decomposition, we have

$$\begin{aligned}
S_{1n,2} &= EQ_i + \frac{3}{n}\sum_i(Q_i - EQ_i) + \frac{6}{n(n-1)}\sum\sum_{j>i}(Q_{ij} - Q_i - Q_j + EQ_i) \\
&\quad + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i}(Q_{ijl} - Q_{ij} - Q_{jl} - Q_{li} + Q_i + Q_j + Q_l - EQ_i) \\
&= J_0 + J_1 + J_2 + J_3
\end{aligned}$$

where the definition of J_0, J_1, J_2 and J_3 should be clear.

$$\begin{aligned} J_0 &= E(Q_i) = E(Q_{ijl}) = E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\ &= E\Big\{E\big[(\mathbf{I}_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2 \\ &= E\Big\{E\big[(F_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2. \end{aligned} \tag{A.14}$$

We first compute $E\big[(F_j - F_i)K_{\gamma,ji}|x_i\big]$.

$$\begin{aligned} &E\big[(F_j - F_i)K_{\gamma,ji}|x_i\big] \\ &= \sum_{z^d \in S^d} L(z^d, x_i^d, \lambda) \int [F(y|x_i^c + hv, z^d) - F(y|x_i)] f(x_i^c + hv, z^d) W(v) dv \\ &= \sum_{z^d \in S^d} \left[\mathbf{I}(z^d = x_i^d) + \sum_{s=1}^r \lambda_s \mathbf{I}_s(z^d, x_i^d) + O(|\lambda|^2)\right] \times \int \Big\{\big[F(y|x_i^c, z^d) - F(y|x_i) \\ &\quad + \sum_{s=1}^q F_s(y|x_i^c, z^d)h_s v_s + (1/2)\sum_{s=1}^q \sum_{t=1}^q F_{st}(y|x_i^c, z^d)h_s h_t v_s v_t + o(|h|^2)\big] \\ &\quad \big[f(x_i^c, z^d) + \sum_{s=1}^q f_s(x_i^c, z^d)h_s v_s + O(|h|^2)\big]\Big\} W(v) dv \\ &= \frac{\kappa_2}{2}\sum_{s=1}^q h_s^2\big[f(x_i)F_{ss}(y|x_i) + 2f_s(x_i)F_s(y|x_i)\big] \\ &\quad + \sum_{s=1}^r \lambda_s \sum_{z^d \in S^d} \mathbf{I}_s(z^d, x_i^d)\big[F(y|x_i^c, z^d) - F(y|x_i)\big] f(x_i^c, z^d) + o(\zeta_{1n}^2). \end{aligned} \tag{A.15}$$

Plugging (A.15) into (A.14), we have

$$\begin{aligned} J_0 &= E\Big\{\frac{\kappa_2}{2}\sum_{s=1}^q h_s^2\big[f(x_i)F_{ss}(y|x_i) + 2f_s(x_i)F_s(y|x_i)\big] f_i^{-1} \\ &\quad + \sum_{s=1}^r \lambda_s \sum_{z^d \in S^d} \mathbf{I}_s(z^d, x_i^d)\big[F(y|x_i^c, z^d) - F(y|x_i)\big] f(x_i^c, z^d) f_i^{-1}\Big\}^2 + o(\zeta_{1n}^2) \\ &= E\Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x_i) + \sum_{s=1}^r \lambda_s B_{2s}(y|x_i)\Big\}^2 + o(\zeta_{1n}^2) \\ &= \int \Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big\}^2 f(x)dx + o(\zeta_{1n}^2), \end{aligned}$$

where $B_{1s}(y|x)$ and $B_{2s}(y|x)$ are defined in Theorem 2.1.

It is obvious that $E(J_1) = 0$ and it is easy to show that $E(J_1^2) = O(n^{-1}\zeta_{1n}^2)$. Hence, $J_1 = O_p(n^{-1/2}\zeta_{1n})$. Similarly, $J_2 = O_p(n^{-1}\zeta_{1n})$, $J_3 = O_p(n^{-3/2}\zeta_{1n})$. Therefore, the leading term of $S_{1,2}$ is J_0. Thus, we have shown that

$$S_{1n,2} = \int \Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big\}^2 f(x)dx + o_p(\zeta_{1n}^2)$$

Next, we consider $S_{1n,1}$, which can be written as a second-order U-statistic. Define $\mathcal{Q}_{ij} = (1/2)[(\mathbf{I}_j - F_i)^2 f_i^{-2} + (\mathbf{I}_i - F_j)^2 f_j^{-2}]K_{\gamma,ji}^2$, $\mathcal{Q}_i = E[\mathcal{Q}_{ij}|x_i]$. Then

$$\begin{aligned} S_{1n,1} &= \frac{1}{n}\Big(E\mathcal{Q}_i + \frac{2}{n}\sum_i(\mathcal{Q}_i - E\mathcal{Q}_i) + \frac{2}{n(n-1)}\sum\sum_{j>i}[\mathcal{Q}_{ij} - \mathcal{Q}_i - \mathcal{Q}_j + E\mathcal{Q}_i]\Big) \\ &= \mathcal{J}_0 + \mathcal{J}_1 + \mathcal{J}_2, \end{aligned}$$

where the definitions of $\mathcal{J}_0$, $\mathcal{J}_1$ and $\mathcal{J}_2$ should be apparent.

$$\begin{aligned} \mathcal{J}_0 &= n^{-1}E(\mathcal{Q}_i) = n^{-1}E(\mathcal{Q}_{ij}) = n^{-1}E\big\{(\mathbf{I}_j - F_i)^2 K_{\gamma,ji}^2 f_i^{-2}\big\} \\ &= n^{-1}E\big\{(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)K_{\gamma,ji}^2 f_i^{-2}\big\} \\ &= n^{-1}E\Big\{E\big[(F_j - 2F_iF_j + F_i^2)K_{\gamma,ji}^2|x_i\big]f_i^{-2}\Big\} \\ &= E\big[\nu_0(nh_1\dots h_q)^{-1}(F_i - F_i^2)f_i^{-1}\big] + O((nh_1\dots h_q)^{-1}\zeta_n) \\ &= E\left(\frac{\Sigma_{y|x_i}}{nh_1\dots h_q}\right) + O((nh_1\dots h_q)^{-1}\zeta_n) \\ &= \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)dx + O((nh_1\dots h_q)^{-1}\zeta_n) \end{aligned} \tag{A.16}$$

where $\Sigma_{y|x}$ is defined in Theorem 2.1.

Similarly, one can easily show that $\mathcal{J}_1 = O_p(n^{-1/2}(nh_1\dots h_q)^{-1})$ and $\mathcal{J}_2 = O_p(n^{-1}(nh_1\dots h_q)^{-1})$. Hence, the leading term of $S_{1n,1}$ is $\mathcal{J}_0$. Thus, we have shown that

$$S_{1n,1} = \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)dx + o_p((nh_1\dots h_q)^{-1}).$$

Summarizing the above we have shown that

$$\begin{aligned} S_{1n}^0 = &\int \Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big\}^2 f(x)\mathcal{M}(x)dx \\ &+ \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)\mathcal{M}(x)dx + o_p(\zeta_n). \end{aligned}$$

Next we show that $S_{1n} - S_{1n}^0 = o_p(\zeta_n)$. By Equation (A.12),

$$\begin{aligned} S_{1n} - S_{1n}^0 &= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\Big(\frac{1}{\hat{f}_i^2} - \frac{1}{f_i^2}\Big) \\ &= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\Big[\frac{2(f_i - \hat{f}_{-i})}{f_i^3} \\ &\quad + \frac{(f_i - \hat{f}_{-i})^2}{f_i^4} + o\big((nh_1\dots h_q)^{-1} + |h|^4 + |\lambda|^2\big)\Big] \\ &= o_p\left(\zeta_n\right). \end{aligned} \tag{A.17}$$

This is because the two terms $2/(n(n-1)^2)\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}(f_i - \hat{f}_{-i})/f_i^3$ and $1/(n(n-1)^2)\sum\sum_{j\neq i}\sum_{l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}(f_i - \hat{f}_{-i})^2/f_i^4$ can be written as fourth-order and fifth-order U-statistics, respectively. Tedious but straightforward calculations can show that both these two terms are $o_p\left(\zeta_n\right)$. Intuitively these results are quite easy to understand, as these two terms have an extra factor $(f_i - \hat{f}_{-i})$ and $(f_i - \hat{f}_{-i})^2$ compared to the leading term. Therefore, both terms have probability orders smaller than ζ_n. □

Lemma A.2. *Equation (A.19) holds true.*

Proof. Define S_{2n}^0 by replacing $\hat{f}_{-i}^{-1}$ in S_{2n} with f_i^{-1}. We will show that (A.19) holds true with S_{2n} being replaced by S_{2n}^0. Because $E\big[F(y|x_i) - \mathbf{I}(y_i \le y)|x_i\big] = 0$ and $E\big[F(y|x_j) - \mathbf{I}(y_j \le y)|x_j\big] = 0$, S_{2n}^0 can be written as a second order degenerate U-statistic.

$$\begin{aligned}
E[(S_{2n}^0)^2] &= \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i} E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)^2(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&= \frac{1}{n^2(n-1)^2}\sum\sum\sum_{l\neq j\neq i} E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)^2(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&\quad + \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i} E\big[(\mathbf{I}_j - F_i)^2(\mathbf{I}_i - F_i)^2K_{\gamma,ji}^2/f_i^2\big] \\
&= O(n^{-1}(\zeta_n) + O((n^2h_1\dots h_q)^{-1}).
\end{aligned}$$

Hence,

$$S_{2n}^0 = O_p\big((n^{-1/2}\zeta_{1n} + (n(h_1\dots h_q)^{1/2})^{-1}\big).$$

Next, using (A.11) we have

$$\begin{aligned}
S_{2n} - S_{2n}^0 &= \frac{1}{n(n-1)}\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}\Big(\frac{1}{f_i} - \frac{1}{\hat{f}_{-i}}\Big) \\
&= \frac{1}{n(n-1)}\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}\Big[\frac{(f_i - \hat{f}_{-i})}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3} \\
&\quad + o\big((nh_1\dots h_q)^{-1} + |h|^4 + |\lambda|^2\big) \\
&= o_p\big(\zeta_n\big).
\end{aligned}$$

The last equality follows from U-statistic H-decomposition, because $1/(n(n-1))\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}(f_i - \hat{f}_{-i})/f_i^2$ and $1/(n(n-1))\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}(f_i - \hat{f}_{-i})^2/f_i^3$ can be written as third and fourth order U-statistics, the leading terms are the mean values of the U-statistics. Given that they have either an extra factor $(f_i - \hat{f}_{-i})$, or $(f_i - \hat{f}_{-i})^2$, it can be shown that they both have probability orders smaller than the leading order of ζ_n. □

Proof of Theorem 2.1 Case (b). Define $G_i = G((y - y_i)/h)$, $\hat{F}_{-i} = \hat{F}_{b,-i}(y|x_i)$, $f_{s,i} = \partial f(x_i)/\partial x_s^c$, $f_{ss,i} = \partial^2 f(x_i)/\partial (x_s^c)^2$, $F_{s,i} = F_s(y|x_i)$, $F_{ss,i} = F_{ss}(y|x_i)$, $s = 0, 1, \dots, q$.

Similar to the proof for case (a), minimizing $CV_b(\cdot)$ over $(h_0, h_1, \dots, h_q, \lambda_1, \dots, \lambda_r)$ is equivalent to minimizing $CV_{b,1}(\cdot)$, where $CV_{b,1}(\cdot)$ is defined as

$$\begin{aligned}
CV_{b,1}(\cdot) &= \frac{1}{n}\sum_i \int \big[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\big]\mathcal{M}_i M(y)dy \\
&= \int \big(S_{1n,b} - 2S_{2n,b}\big)M(y)dy
\end{aligned}$$

where $S_{1n,b} = n^{-1}\sum_i(\hat{F}_{-i} - F_i)^2\mathcal{M}_i$ and $S_{2n,b} = n^{-1}\sum_i(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\mathcal{M}_i$.

Lemma A.1 and Lemma A.2 below show that (recall that $\zeta_{1n} = |h|^2 + |\lambda|$)

$$S_{1n,b} = \int\Big\{\Big[\sum_{s=0}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big]^2 + \frac{V(y|x) - h_0\Omega}{nh_1\dots h_q}\Big\}f(x)\mathcal{M}(x)dx + o_p(h_0^4 + \zeta_n) \tag{A.18}$$

$$S_{2n,b} = O_p\big(n^{-1/2}(h_0^4 + \zeta_n) + (n(h_1\dots h_q)^{1/2})^{-1}\big). \tag{A.19}$$

This completes the proof of Theorem 2.1 Case (b). □

In the proof that follows, we will use two results from Li & Racine (2008), that is,

$$E\Big[G\Big(\frac{y-Y_i}{h_0}\Big)|X_i\Big] = F(y|X_i) + (1/2)\kappa_2 h_0^2 F_{00}(y|X_i) + o(h_0^2) \tag{A.20}$$

$$E\Big[G^2\Big(\frac{y-Y_i}{h_0}\Big)|X_i\Big] = F(y|X_i) - h_0 C_w F_0(y|X_i) + O(h_0^2) \tag{A.21}$$

where $F_0(y|x) = \partial F(y|x)/\partial y$, $F_{00}(y|x) = \partial^2 F(y|x)/\partial y^2$. The above results are proved in Lemma A.5 of Li & Racine (2008).

Lemma A.3. *Equation (A.18) holds true.*

Proof. We omit the weight function $\mathcal{M}_i$ for notational simplicity. Define $S^0_{1n,b}$ by replacing $1/\hat{f}_{-i}$ in $S_{1n,b}$ with $1/f_i$.

$$\begin{aligned} S^0_{1n,b} &= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i}(G_j - F_i)(G_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2 \\ &= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}(G_j - F_i)^2 K^2_{\gamma,ji}/f_i^2 \\ &\quad + \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i}(G_j - F_i)(G_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2 \\ &= S_{1,1,b} + S_{1,2,b}, \end{aligned}$$

where the definitions of $S_{1,1,b}$ and $S_{1,2,b}$ should be apparent.

First, we consider $S_{1,2,b}$, which can be written as a third-order U-statistic. The leading term of $S_{1,2,b}$ is $E[(G_j - F_i)(G_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2]$.

$$E\big[(G_j - F_i)(G_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] = E\Big\{E\big[(G_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2. \tag{A.22}$$

We first compute $E\big[(G_j - F_i)K_{\gamma,ji}|x_i\big]$.

$$\begin{aligned} E\big[(G_j - F_i)K_{\gamma,ji}|x_i\big] &= E\big[K_{\gamma,ji}E(G_j|x_j)|x_i\big] - E[K_{\gamma,ji}F_i|x_i] \\ &= E\big[K_{\gamma,ji}(F_j + \frac{\kappa_2}{2}h_0^2 F_{00,j})|x_i\big] - E[K_{\gamma,ji}F_i|x_i] + o(h_0^4), \end{aligned} \tag{A.23}$$

where (A.20) is used in the last equality. It's easy to see that

$$\begin{aligned} E[F_j K_{\gamma,ji}|x_i] &= f_i F_i + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[f_i F_{ss,i} + 2f_{s,i}F_{s,i} + F_i f_{ss,i}] \\ &\quad + \sum_{s=1}^{r}\lambda_s \sum_{z_i^d \in D}\mathbf{I}_s(z_i^d, x_i^d)f(x_i^c, z_i^d)F(y|x_i^c, z_i^d) + o(\zeta_{1n}) \end{aligned} \tag{A.24}$$

$$E[F_i K_{\gamma,ji}|x_i] = f_i F_i + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss,i}F_i + \sum_{s=1}^{r}\lambda_s \sum_{z_i^d \in D}\mathbf{I}_s(z_i^d, x_i^d)f(x_i^c, z_i^d)F_i + o(\zeta_{1n}) \tag{A.25}$$

$$E\Big[\frac{\kappa_2}{2}h_0^2 F_{00,j}K_{\gamma,ji}|x_i\Big] = \frac{\kappa_2}{2}h_0^2 F_{00,i}f_i + o(h_0^2) \tag{A.26}$$

Substituting (A.24), (A.25) and (A.26) into (A.23), we have,

$$E\big[(G_j - F_i)K_{\gamma,ji}|x_i\big] = \frac{\kappa_2}{2}h_0^2 f_i F_{00,i} + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[f_i F_{ss,i} + 2f_{s,i}F_{s,i}]$$

$$+ \sum_{s=1}^{r}\lambda_s \sum_{z_i^d \in D} I_s(z_i^d, x_i^d) f(x_i^c, z_i^d)[F(y|x_i^c, z_i^d) - F_i] + o(h_0^2 + \zeta_{1n}). \tag{A.27}$$

Substituting (A.27) into (A.22), we have

$$E\big[(G_j - F_i)(G_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big]$$

$$= \int \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\right]^2 f(x)dx + o(h_0^4 + \zeta_{1n}^2) \tag{A.28}$$

where $B_{1s}(y|x)$ for $s = 0, \ldots, q$, $B_{2s}(y|x)$ for $s = 1, \ldots, r$ are defined in Theorem 2.1.

By U-statistic H-decomposition,

$$S_{1,2,b} = \int \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\right]^2 f(x)dx + o(h_0^4 + \zeta_{1n}^2). \tag{A.29}$$

Next, we consider $S_{1,1,b}$, which can be written as a second-order U-statistic. The leading term of $S_{1,1,b}$ is $E[(G_j - F_i)^2 K_{\gamma,ji}^2/f_i^2]$.

(A.30)

$$E[(G_j - F_i)^2 K_{\gamma,ji}^2/f_i^2] = E[(G_j^2 + F_i^2 - 2G_jF_i)K_{\gamma,ji}^2/f_i^2] = E\big\{E[(G_j^2 + F_i^2 - 2G_jF_i)K_{\gamma,ji}^2|x_i]/f_i^2\big\}.$$

We first compute $E[(G_j^2 + F_i^2 - 2G_jF_i)K_{\gamma,ji}^2|x_i]$.

$$E[(G_j^2 + F_i^2 - 2G_jF_i)K_{\gamma,ji}^2|x_i] = E\big\{[E(G_j^2 + F_i^2 - 2G_jF_i)K_{\gamma,ji}^2|x_i, x_j]|x_i\big\}$$

$$= E\Big\{\big[E(G_j^2|x_j) + F_i^2 - 2F_iE(G_j|x_j)\big]K_{\gamma,ji}^2|x_i\Big\} = \big[F_i - F_i^2 - h_0C_wF_{0,i} + O(h_0^2)\big]E[K_{\gamma,ji}^2|x_i]$$

$$= (h_1 \ldots h_q)^{-1}\nu_0(F_i - F_i^2 - h_0C_wF_{0,i})f_i + O((h_1 \ldots h_q)^{-1}(h_0^2 + |h|^2 + |\lambda|^2)). \tag{A.31}$$

Substituting (A.31) into (A.30), we have

$$E[(G_j - F_i)^2 K_{\gamma,ji}^2/f_i^2] = \int \frac{V(y|x) - h_0\Omega_1}{h_1 \ldots h_q} f(x)dx + O((h_1 \ldots h_q)^{-1}(h_0^2 + |h|^2 + |\lambda|^2))$$

where $V(y|x)$ and Ω_1 are defined in Theorem 2.1.

By U-statistic H-decomposition,

$$S_{1,1,b} = \int \frac{V(y|x) - h_0\Omega_1}{nh_1 \ldots h_q} f(x)dx + O((nh_1 \ldots h_q)^{-1}(h_0^2 + |h|^2 + |\lambda|^2)) + O((nh_1 \ldots h_q)^{-1}n^{-1/2}). \tag{A.32}$$

Summarizing (A.29) and (A.32), we have shown that

$$S_{1n,b}^0 = \int \left\{\Big[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\Big]^2 + \frac{V(y|x) - h_0\Omega_1}{nh_1 \ldots h_q}\right\} f(x)\mathcal{M}(x)dx + o_p(h_0^4 + \zeta_{1n}^2).$$

□

Lemma A.4. *Equation (A.19) holds true.*

Proof. Define $S_{2n,b}^0$ by replacing $\hat{f}_{-i}^{-1}$ in $S_{2n,b}$ with f_i^{-1}.

$$S_{2n,b}^0 = n^{-1}\sum_i (\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i) = \frac{1}{n(n-1)}\sum\sum_{j \neq i}(G_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}\mathcal{M}_i/f_i.$$

Because $E\big[\mathbf{I}(y_i \le y) - F(y|x_i)|x_i\big] = 0$, $S^0_{2n,b}$ can be written as a second order degenerate U-statistic.

$$\begin{aligned}
E[(S^0_{2n,b})^2] &= \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i} E\big[(G_j - F_i)(G_l - F_i)(\mathbf{I}_i - F_i)^2 K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&= \frac{1}{n^2(n-1)^2}\sum\sum\sum_{l\neq j\neq i} E\big[(G_j - F_i)(G_l - F_i)(\mathbf{I}_i - F_i)^2 K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&\quad + \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i} E\big[(G_j - F_i)^2(\mathbf{I}_i - F_i)^2 K^2_{\gamma,ji}/f_i^2\big] \\
&= O(n^{-1}(h_0^4 + \zeta_{1n}^2)) + O((n^2 h_1 \dots h_q)^{-1}).
\end{aligned}$$

Hence,

$$S^0_{2n,b} = O_p\big((n^{-1/2}(h_0^2 + \zeta_{1n}) + (n(h_1 \dots h_q)^{1/2})^{-1}\big).$$

□

Proof of Theorem 2.1 Case (c). Define $\mathcal{L}_j = \mathcal{L}(y, y_j, \lambda_0)$, $\hat{F}_{-i} = \hat{F}_{c,-i}(y|x_i)$.

By the same arguments used earlier we know that minimizing $CV_c(\cdot)$ is equivalent to minimizing $CV_{c,1}(\cdot)$, where $CV_{c,1}(\cdot)$ is defined as

$$\begin{aligned}
CV_{c,1}(\cdot) &= \frac{1}{n}\sum_i \sum_{y\in D_y} \big[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\big]\mathcal{M}_i \\
&= \sum_{y\in D_y} \big(S_{1n,c} - 2S_{2n,c}\big)
\end{aligned}$$

where $S_{1n,c} = n^{-1}\sum_i (\hat{F}_{-i} - F_i)^2\mathcal{M}_i$ and $S_{2n,c} = n^{-1}\sum_i (\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\mathcal{M}_i$.

Lemma A.5 and Lemma A.4 below show that

$$\begin{aligned}
S_{1n,c} &= \int \Big\{ \Big[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) + \lambda_0 B_{2,0}(y|x)\Big]^2 \\
&\qquad + \frac{V(y|x) + 2\lambda_0 F(y|x)^2}{n h_1 \dots h_q}\Big\} f(x)\mathcal{M}(x)dx + o_p(h_0^2 + \zeta_{1n}^2), \qquad \text{(A.33)} \\
S_{2n,c} &= O_p\big((n^{-1/2}(\lambda_0 + \zeta_{1n}) + (n(h_1 \dots h_q)^{1/2})^{-1}\big) = o_p(S_{1n}), \qquad \text{(A.34)}
\end{aligned}$$

where $\zeta_{1n} = |h|^2 + |\lambda|$.

Before we prove Lemma A.5 and Lemma A.4, we first calculate $E[\mathcal{L}(y, y_j, \lambda_0)|x_j]$ and $E[\mathcal{L}(y, y_j, \lambda_0)^2|x_j]$. Define $C_y = \sum_{z\le y} 1$.

$$\begin{aligned}
E[\mathcal{L}(y, y_j, \lambda_0)|x_j] &= \sum_{z\le y} E[l(z, y_j, \lambda_0)|x_j] = \sum_{z\le y}\big[f(z|x_j) + \lambda_0 \sum_{y_j\neq z} f(y_j|x_j)\big] \\
&= F(y|x_j) + \lambda_0 \sum_{z\le y}[1 - f(z|x_j)] \\
&= F(y|x_j) + \lambda_0[C_y - F(y|x_j)] + O(\lambda_0^2), \qquad \text{(A.35)}
\end{aligned}$$

$$
\begin{aligned}
E[\mathcal{L}(y,y_j,\lambda_0)^2|x_j] &= E\Big\{\big[\sum_{z\le y} l(z,y_j,\lambda_0)\big]^2|x_j\Big\} \\
&= E\big[\sum_{z\le y} l(z,y_j,\lambda_0)^2|x_j\big] + E\big[\sum_{z\le y}\sum_{z'\le y, z'\neq z} l(z,y_j,\lambda_0)l(z',y_j,\lambda_0)|x_j\big] \\
&= F(y|x_j) + 2\lambda_0 \sum_{z\le y}\sum_{z'\le y} f(z|x_j)\mathbf{I}(z\neq z') + O(\lambda_0^2) \\
&= F(y|x_j) + 2\lambda_0\Big[\sum_{z\le y}\sum_{z'\le y} f(z|x_j)\big(1-\mathbf{I}(z=z')\big)\Big] + O(\lambda_0^2) \\
&= F(y|x_j) + 2\lambda_0\big(C_y - 1\big)F(y|x_j) + O(\lambda_0^2). \qquad \text{(A.36)}
\end{aligned}
$$

Lemma A.5. *Equation (A.33) holds true.*

Proof. We omit the weighted function $\mathcal{M}_i$ for notational simplicity. Define $S^0_{1n,c}$ by replacing $\hat{f}^{-1}_{-i}$ in $S_{1n,c}$ with f_i^{-1}.

$$
\begin{aligned}
S^0_{1n,c} &= \frac{1}{n}\sum_i (\hat{F}_{-i} - F_i)^2 \\
&= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\sum_{k\neq i}(\mathcal{L}_j - F_i)(\mathcal{L}_k - F_i)K_{\gamma,ji}K_{\gamma,ki}/f_i^2 \\
&= \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}(\mathcal{L}_j - F_i)^2K^2_{\gamma,ji}/f_i^2 \\
&\quad + \frac{1}{n(n-1)^2}\sum\sum\sum_{k\neq j\neq i}(\mathcal{L}_j - F_i)(\mathcal{L}_k - F_i)K_{\gamma,ji}K_{\gamma,ki}/f_i^2 \\
&= S_{1,1,c} + S_{1,2,c},
\end{aligned}
$$

where the definitions of $S_{1,1,c}$ and $S_{1,2,c}$ should be apparent.

First, we consider $S_{1,2,c}$, which can be written as a third-order U-statistic. The leading term of $S_{1,2,c}$ is $E[(\mathcal{L}_j - F_i)(\mathcal{L}_k - F_i)K_{\gamma,ji}K_{\gamma,ki}/f_i^2]$.

$$
E\big[(\mathcal{L}_j - F_i)(\mathcal{L}_k - F_i)K_{\gamma,ji}K_{\gamma,ki}/f_i^2\big] = E\Big\{E\big[(\mathcal{L}_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2. \qquad \text{(A.37)}
$$

We first compute $E\big[(\mathcal{L}_j - F_i)K_{\gamma,ji}|x_i\big]$.

$$
\begin{aligned}
&E\big[(\mathcal{L}_j - F_i)K_{\gamma,ji}|x_i\big] = E\big[E(\mathcal{L}_j|x_j)K_{\gamma,ji}|x_i\big] - F_iE[K_{\gamma,ji}|x_i] \\
&= E\big[F(y|x_j)K_{\gamma,ji}|x_i\big] + \lambda_0 E\big[(C_y - F(y|x_j))K_{\gamma,ji}|x_i\big] - F_iE[K_{\gamma,ji}|x_i] \\
&= \frac{\kappa_2}{2}\sum_{s=1}^q h_s^2[f_iF_{ss,i} + 2f_{s,i}F_{s,i}] + \sum_{s=1}^r \lambda_s \sum_{z_i^d\in D} I_s(z_i^d,x_i^d)f(x_i^c,z_i^d)[F(y|x_i^c,z_i^d) - F_i] \\
&\quad + \lambda_0(C_y - F(y|x_j))f_i + o(\zeta_{1n}^2), \qquad \text{(A.38)}
\end{aligned}
$$

where (A.35) is used in the second equality.

Substituting (A.38) into (A.37), we have,

$$
\begin{aligned}
&E\big[(\mathcal{L}_j - F_i)(\mathcal{L}_k - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&= \int\Big[\sum_{s=1}^q h_s^2B_{1s}(y|x) + \sum_{s=0}^r \lambda_s B_{2s}(y|x)\Big]^2 f(x)dx + o(\zeta_{1n}^2), \qquad \text{(A.39)}
\end{aligned}
$$

where $B_{1s}(y|x)$ for $s = 1,\ldots,q$ and $B_{2s}(y|x)$ for $s = 0,1,\ldots,r$ are defined in Theorem 2.1.

By U-statistic H-decomposition,

$$S_{1,2,c} = \int \Big[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=0}^{r} \lambda_s B_{2s}(y|x)\Big]^2 f(x)dx + o(\zeta_{1n}^2). \tag{A.40}$$

Next, we consider $S_{1,1,c}$, which can be written as a second-order U-statistic. The leading term of $S_{1,1,c}$ is $E[(\mathcal{L}_j - F_i)^2 K_{\gamma,ji}^2/f_i^2]$.

$$\begin{aligned} E[(\mathcal{L}_j - F_i)^2 K_{\gamma,ji}^2/f_i^2] &= E[(\mathcal{L}_j^2 + F_i^2 - 2\mathcal{L}_j F_i)K_{\gamma,ji}^2/f_i^2] \\ &= E\big\{E[(\mathcal{L}_j^2 + F_i^2 - 2\mathcal{L}_j F_i)K_{\gamma,ji}^2|x_i]/f_i^2\big\}. \end{aligned} \tag{A.41}$$

We first compute $E[(\mathcal{L}_j^2 + F_i^2 - 2\mathcal{L}_j F_i)K_{\gamma,ji}^2 \mid x_i]$.

$$\begin{aligned} &E[(\mathcal{L}_j^2 + F_i^2 - 2\mathcal{L}_j F_i)K_{\gamma,ji}^2|x_i] \\ &= E\big\{[E(\mathcal{L}_j^2 + F_i^2 - 2\mathcal{L}_j F_i|x_i, x_j)]K_{\gamma,ji}^2|x_i\big\} \\ &= E[E(\mathcal{L}_j^2|x_j)K_{\gamma,ji}^2|x_i] + F_i^2 E[K_{\gamma,ji}^2|x_i] - 2F_i E[E(\mathcal{L}_j|x_j)K_{\gamma,ji}^2|x_i] \\ &= (h_1 \dots h_q)^{-1}\nu_0\big[F_i - F_i^2 + 2\lambda_0(F_i^2 - F_i)\big]f_i + O((h_1 \dots h_q)^{-1}(\lambda_0^2 + |h|^2 + |\lambda|^2), \end{aligned} \tag{A.42}$$

where (A.35) and (A.36) are used in the last equality.

Substituting (A.42) into (A.41), we have

$$E[(\mathcal{L}_j - F_i)^2 K_{\gamma,ji}^2/f_i^2] = \int \frac{V(y|x) + \lambda_0\Omega_2(y|x)}{h_1 \dots h_q} f(x)dx + O((h_1 \dots h_q)^{-1}(\lambda_0^2 + |h|^2 + |\lambda|^2)),$$

where $V(y|x)$ and $\Omega_2(y|x)$ are defined in Theorem 2.1.

By U-statistic H-decomposition,

$$S_{1,1,c} = \int \frac{V(y|x) + \lambda_0\Omega_2(y|x)}{nh_1 \dots h_q} f(x)dx + O((nh_1 \dots h_q)^{-1}(\lambda_0^2 + |h|^2 + |\lambda|^2)) + O((nh_1 \dots h_q)^{-1}n^{-1/2}). \tag{A.43}$$

Summarizing (A.40) and (A.43), we have shown that

$$\begin{aligned} S_{1n,c}^0 = \int \Big\{ &\Big[\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=0}^{r} \lambda_s B_{2s}(y|x)\Big]^2 \\ &+ \frac{V(y|x) + \lambda_0\Omega_2(y|x)}{nh_1 \dots h_q}\Big\} f(x)\mathcal{M}(x)dx + o_p(\lambda_0^2 + \zeta_{1n}^2). \end{aligned}$$

□

Lemma A.6. *Equation (A.34) holds true.*

Proof. Define $S_{2n,c}^0$ by replacing $\hat{f}_{-i}^{-1}$ in $S_{2n,c}$ with f_i^{-1}.

$$\begin{aligned} S_{2n,c}^0 &= n^{-1}\sum_i (\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i) \\ &= \frac{1}{n(n-1)}\sum\sum_{j\neq i} (\mathcal{L}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}\mathcal{M}_i/f_i. \end{aligned}$$

Because $E\big[\mathbf{I}(y_i \le y) - F(y|x_i)|x_i\big] = 0$, $S^0_{2n,c}$ can be written as a second order degenerate U-statistic.

$$\begin{aligned}
E[(S^0_{2n,c})^2] &= \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i}\sum_{l\neq i} E\big[(\mathcal{L}_j - F_i)(\mathcal{L}_l - F_i)(\mathbf{I}_i - F_i)^2 K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&= \frac{1}{n^2(n-1)^2}\sum\sum\sum_{l\neq j\neq i} E\big[(\mathcal{L}_j - F_i)(\mathcal{L}_l - F_i)(\mathbf{I}_i - F_i)^2 K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\
&\quad + \frac{1}{n^2(n-1)^2}\sum\sum_{j\neq i} E\big[(\mathcal{L}_j - F_i)^2(\mathbf{I}_i - F_i)^2 K^2_{\gamma,ji}/f_i^2\big] \\
&= O(n^{-1}(\lambda_0^2 + \zeta_{1n}^2)) + O((n^2 h_1 \dots h_q)^{-1}).
\end{aligned}$$

Hence,

$$S^0_{2n,c} = O_p\big((n^{-1/2}(\lambda_0 + \zeta_{1n}) + (n(h_1 \dots h_q)^{1/2})^{-1}\big).$$

□

□

Proof of Theorem 2.2. Theorem 2.2 is a special case of Theorem 3.1 with $q_1 = q$ and $r_1 = r$ (when there are no irrelevant covariates). □

Proof of Theorem 2.3. Theorem 2.3 is a special case of Theorem 3.2 with $q_1 = q$ and $r_1 = r$ (when there are no irrelevant covariates). □

Proof of Theorem 2.4. We will only prove case (a) as other cases can be proved similarly. Hence, we consider $CV_\Sigma = n^{-2}\sum_{i=1}^n\sum_{j\neq i}^n \left[\hat{F}_{-i}(y_j|x_i) - \mathbf{I}(y_i \le y_j)\right]^2 \mathcal{M}_i$, where $\hat{F}_{-i}(y_j|x_i) = \hat{F}_{a,-i}(y_j|x_i)$. Denote $\mathbf{I}_{ji} = \mathbf{I}(y_i \le y_j)$, $\hat{F}_{-i,ji} = \hat{F}_{a,-i}(y_j|x_i)$, $F_{ji} = F(y_j|x_i)$. Then similar to the proof of Theorem 2.1, by adding/subtracting F_{ji} between $\mathbf{I}_{ji}$ and $\hat{F}_{-i,ji}$ in CV_Σ, we obtain $CV_\Sigma = CV_{\Sigma,1} + (s.o.)$, where

$$\begin{aligned}
CV_{\Sigma,1} &= \frac{1}{n^2}\sum_i\sum_{j\neq i}(\hat{F}_{-i,ji} - F_{ji})^2\mathcal{M}_i + \frac{2}{n^2}\sum_i\sum_{j\neq i}(\hat{F}_{-i,ji} - F_{ji})(F_{ji} - \mathbf{I}_{ji})\mathcal{M}_i \\
&= S_{\Sigma,1n} - S_{\Sigma,2n}, \qquad \text{(A.44)}
\end{aligned}$$

where the definitions of $S_{\Sigma,1n}$ and $S_{\Sigma,2n}$ should be apparent. Using $\hat{F}_{-i,ji} = n^{-1}\sum_{l\neq i}\mathbf{I}_{jl}K_{\gamma,il}/\hat{f}_{-i}$ and $1/\hat{f}_{-i} = 1/f_i + (s.o.)$, we obtain $S_{\Sigma,1n} = S^0_{\Sigma,1n} + (s.o.)$, where

$$S^0_{\Sigma,1n} = \frac{1}{n^4}\sum_i\sum_{j\neq i}\sum_{l\neq i}\sum_{l'\neq i}(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(\mathbf{I}_{jl'} - F_{ji})K_{\gamma,il'}\mathcal{M}_i/f_i^2. \qquad \text{(A.45)}$$

We discuss several cases for $S^0_{\Sigma,1n}$: (i) all four indices i, j, l, l' differ from each other; (ii) $l = l'$ and $i \neq j \neq l$; (iii) $l = j$ and $i \neq j \neq l'$; (iv) $l' = j$ and $i \neq j \neq l$; (v) $l = l' = j$ and $j \neq i$.

For case (i) we have

$$S^0_{\Sigma,1n,(i)} = \frac{1}{n^4}\sum\sum\sum\sum_{i\neq j\neq l\neq l'}(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(\mathbf{I}_{jl'} - F_{ji})K_{\gamma,il'}\mathcal{M}_i/f_i^2. \qquad \text{(A.46)}$$

$S^0_{\Sigma,1n,(i)}$ can be written as a fourth order U-statistic. By the H-decomposition we know that $S^0_{\Sigma,1n,(i)} = E[S^0_{\Sigma,1n,(i)}] + (s.o.)$. Denoting $\mathbf{I}_{ly} = \mathbf{I}(y_l \le y)$, $F_{iy} = F(y|x_i)$ and noting that y_j is independent of

$(y_l, y_{l'}, x_i, x_l, x_{l'})$, we have (recall that $g(\cdot)$ is the marginal density of y_j)

$$E[S^0_{\Sigma,1n,(i)}] = \int g(y)E[(\mathbf{I}_{ly} - F_{iy})K_{\gamma,il}(\mathbf{I}_{l'y} - F_{iy})K_{\gamma,il'}\mathcal{M}_i/f_i^2]dy$$
$$= \int g(y)S^0_{1n,1}(y)dy, \tag{A.47}$$

where $S^0_{1n,1}(y) = E[(\mathbf{I}_{ly} - F_{iy})K_{\gamma,il}(\mathbf{I}_{l'y} - F_{iy})K_{\gamma,il'}\mathcal{M}_i/f_i^2]$. From (A.13) we know that $S^0_{1n,1}(y) = E[S_{1n,2}]$ if one replaces $M(y)$ by $g(y)$ in the definition of $S_{1n,2}$, where $S_{1n,2}$ is defined in (A.13) and is one of the leading terms of S^0_{1n} (and of $CV_a(\cdot)$); see the proof of Theorem 2.1 case (a).

For case (ii), by H-decomposition we know $S^0_{\Sigma,1n,(ii)} = E[S^0_{\Sigma,1n,(ii)}] + (s.o.)$ and

$$E[S^0_{\Sigma,1n,(ii)}] = n^{-1}\int g(y)E[(\mathbf{I}_{ly} - F_{iy})^2K^2_{\gamma,il}\mathcal{M}_i/f_i^2]dy$$
$$= \int g(y)S^0_{1n,2}(y)dy, \tag{A.48}$$

where $S^0_{1n,2}(y) = n^{-1}E[(\mathbf{I}_{ly} - F_{iy})^2K^2_{\gamma,il}\mathcal{M}_i/f_i^2]$. By (A.13) we know that $S^0_{1n,2}(y) = E[S_{1n,1}]$ if one replaces $M(y)$ by $g(y)$ in the definition of $S_{1n,1}$, where $S_{1n,1}$ is defined in (A.13) and is the second leading term of S^0_{1n} (and of $CV_a(\cdot)$).

For case (iii) $l' = j$, by H-decomposition we know that $S^0_{\Sigma,1n,(iii)} = E[S^0_{\Sigma,1n,(iii)}] + (s.o.)$ and

$$E[S^0_{\Sigma,1n,(iii)}] = n^{-1}E[(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(1 - F_{ji})K_{\gamma,ij}\mathcal{M}_i/f_i^2] + (s.o.)$$
$$= n^{-1}E[(F_{lj} - F_{ji})K_{\gamma,il}(1 - F_{ji})K_{\gamma,ij}\mathcal{M}_i/f_i^2] + (s.o.)$$
$$= n^{-1}O(|h|^2 + |\lambda|) = O(n^{-1}\zeta_{1n}). \tag{A.49}$$

By symmetry, we know that case (iv) is the same as case (iii) so that we have $S^0_{\Sigma,1n,(iv)} = O(n^{-1}\zeta_{1n})$.

Finally, it is easy to see that $S_{\Sigma,1n,(v)} = O_p(n^{-2}(h_1 \dots h_q)^{-1})$.

Summarizing the above we have shown that the leading term of CV_Σ (for case (a)) is given by

$$CV_{\Sigma,L} = \int g(y)[S^0_{1n,1}(y) + S^0_{1n,2}(y)]dy, \tag{A.50}$$

which equals $CV_{a,L}$ provided that one replaces $M(y)$ by $g(y)$ in $CV_{a,L}(\cdot)$. Hence, Theorem 2.4 follows from Theorem 2.1.

So far we have assumed that y is a continuous random variable. For the discrete y case, we just need to replace the integral with the summation operator, that is, (A.50) will be written as $CV_{\Sigma,L} = \sum_j [S^0_{1n,1}(y_j) + S^0_{1n,2}(y_j)]g(y_j)$.

Cases (b) and (c) can be similarly proved. Thus, we have proved Theorem 2.4. □

Proof of Theorem 2.6. We will only provide a sketch of the proof for case (b) as the proof for case (a) follows exactly the same derivations as in the scalar y case of Theorem 2.3 (and using derivations similar to those used in the proof of Theorem 2.1).

Let $f(y^c, y^d)$ and $F(y^c, y^d)$ be the joint PDF and CDF of $y = (y^c, y^d)$. Define $m(y^c, z^d) = \int_{-\infty}^{y^c} f(y^c, z^d)dy^c$, then $\sum_{z^d \in y^d} m(y^c, z^d) = F(y^c, y^d)$. Define $m_{0,s}(y^c, y^d|x_j) = \partial m(y^c, y^d|x_j)/\partial y^c_s$, $m_{00,s}(y^c, y^d|x_j) = \partial^2 m(y^c, y^d|x_j)/\partial (y^c_s)^2$, $\mathbf{I}_{y^d_{js} = z_s} = \mathbf{I}(y^d_{js} = z_s)$, $F_{0,s}(y^c, y^d|x_j) = \partial F(y^c, y^d|x_j)/\partial y^c_s$, $F_{00,s}(y^c, y^d|x_j) = \partial^2 F(y^c, y^d|x_j)/\partial (y^c_s)^2$

We use $\hat{F}(y|x)$ to denote $\hat{F}_{m,b}(y|x)$ defined in (12). We write $\hat{F}(y|x) - F(y|x) = [\hat{F}(y|x) - F(y|x)]\hat{f}(x)/\hat{f}(x) \equiv \hat{p}(y|x)/\hat{f}(x)$, where $\hat{p}(y|x) = [\hat{F}(y|x) - F(y|x)]\hat{f}(x)$.

Define $\mathcal{K}_{y_j,y,\gamma_0} = \mathcal{K}(y_j, y, \gamma_0)$, $\mathcal{L}_{y_j,y,\gamma_0} = \mathcal{L}(y_j, y, \gamma_0)$, $F_{y|x} = F(y|x)$, $K_{\gamma,x_j,x} = K_\gamma(x_j, x)$, $f_{s,x} = \partial f(x)/\partial x^c_s$, $F_{s,x} = \partial F(y|x)/\partial x^c_s$, $F_{ss,x} = \partial^2 F(y|x)/(\partial x^c_s)^2$, $F_{0,s} = \partial F(y|x)/\partial y^c_s$, $F_{00,s} = \partial^2 F(y|x)/(\partial y^c_s)^2$. By Lemma A.7 (i), we know that

(*i*) $E[\hat{p}(y|x)] = E\big\{E[(\mathcal{K}_{y_j,y,\gamma_0}|x_j)K_{\gamma,x_j,x}]\big\} - E[F_{y|x}K_{\gamma,x_j,x}] = (\kappa_2/2)\sum_{s=1}^{q} h_s^2[f(x)F_{ss,x} + 2f_{s,x}F_{s,x}] + (\kappa_2^{q_y}/2)f(x)\sum_{s=1}^{q_y} h_{0,s}^2 F_{00,s} + \sum_{s=1}^{r}\lambda_s \sum_{z_i^d \in S^D} \mathbf{I}_s(z^d,x^d)[F(y|x^c,z^d) - F(y|x)]f(x^c,z^d) + \sum_{s=1}^{r_y}\lambda_{0,s}C_{1s,y,x} f(x) + o(\zeta_{1n} + |h_0|^2 + |\lambda_0|)$,
where $C_{1s,y,x}$ is defined in Lemma A.7 below, $|h_0|^2 = \sum_{s=1}^{q_y} h_{s,0}^2$ and $|\lambda_0| = \sum_{s=1}^{r_y}\lambda_{s,0}$.

Using Lemma A.7 (i) and (ii), we have

(*ii*) $Var[\hat{p}(y|x)] = n^{-1}\big\{E[(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})^2 K_{\gamma,x_j x}^2] - [E(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})K_{\gamma,x_j x}]^2\big\} = \frac{\nu_0}{nh_1\ldots h_q}\Big[F(y|x) - F(y|x)^2 - \sum_{s=1}^{q_y} h_{0,s}C_w F_{0,s} + \sum_{s=1}^{r_y}\lambda_{0,s}\big(C_{2s,y,x} - 2F(y|x)C_{1s,y,x}\big)\Big]f(x) + o((nh_1\ldots h_q)^{-1}(|\lambda_0| + |h_0|^2))$
where $C_{2s,y,x}$ is defined in (A.54) below and $C_w = 2\int G(v)w(v)vdv$.

From (i) and (ii) above and noting that $\hat{F}(y|x) - F(y|x) = \hat{p}(y|x)/\hat{f}(x) = \hat{p}(y|x)/f(x) + (s.o.)$, applying the Liapunov central limit theorem (CLT), we have

$$\sqrt{nh_1\ldots h_q}\left[\hat{F}_{m,a}(y|x) - F(y|x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) - \sum_{s=1}^{q_y} h_{0,s}^2 B_{0,1s}(y|x) - \sum_{s=1}^{r}\lambda_s B_{2s}(y|x) - \sum_{s=1}^{r_y}\lambda_{0,s}B_{0,2s}(y|x)\right] \xrightarrow{d} N(0,\Sigma_{y|x}),$$

where the definitions of $B_{1s}(y|x)$, $B_{2s}(y|x)$ and $\Sigma_{y|x}$ are the same as those given in Theorem 2.1 except that now $y = (y_1,\ldots,y_p)$, $B_{0,1s}(y|x) = (\kappa_2^{q_y}/2)F_{00,s}$ and $B_{0,2s}(y|x) = C_{1s,y,x}$. This completes the proof of Theorem 2.6. □

Lemma A.7. *Under the same conditions as given in Theorem 2.6, we have*

(i) $E[(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})K_{\gamma,x_j,x}] = \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[f(x)F_{ss,x} + 2f_{s,x}F_{s,x}] + (1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y} h_{0,s}^2 F_{00,s}f(x) + \sum_{s=1}^{r}\lambda_s\sum_{z_i^d\in S^D}\mathbf{I}_s(z^d,x^d)[F(y|x^c,z^d) - F(y|x)]f(x^c,z^d) + \sum_{s=1}^{r_y}\lambda_{0,s}C_{1s,y,x}f(x) + o(\zeta_{1n} + |\lambda_0| + |h_0|^2)$, *where* $C_{1s,y,x} = \sum_{y_j^d\in D_y^d}\Big[\sum_{z_s\le y_s^d}\boldsymbol{I}_{y_{js}^d\neq z_s}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l^d}\boldsymbol{I}_{y_{jl}^d = z_l}\Big]m(y^c,y_j^d|x)$.

(ii) $E[(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})^2 K_{\gamma,x_j x}^2] = (h_1\ldots h_q)^{-1}\nu_0\Big[F(y|x) - F(y|x)^2 - \sum_{s=1}^{q_y} h_{0,s}C_w F_{0,s} + \sum_{s=1}^{r_y}\lambda_{0,s}\big(C_{2s,y,x} - 2F(y|x)C_{1s,y,x}\big)\Big]f(x)$,

where $C_{2s,y,x} = 2\sum_{y_j^d\in D_y^d}\Big\{\sum_{z_s\le y_s^d}\sum_{z_s'\le y_s^d}\boldsymbol{I}_{y_{js}=z_s}\boldsymbol{I}_{y_{js}\neq z_s'}\Big\}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l}\boldsymbol{I}_{y_{jl}=z_l}m(y^c,y_j^d|x)$.

Proof. We first obtain some leading term expansions for the discrete kernel functions.

$$\begin{aligned}\mathcal{L}(y_j^d,y^d,\lambda_0) &= \prod_{s=1}^{r_y}\mathcal{L}(y_{js}^d,y_s^d,\lambda_{0,s}) = \prod_{s=1}^{r_y}\Big\{\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s} + \lambda_{0,s}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\Big\}\\ &= \prod_{s=1}^{r_y}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s} + \sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l^d}\mathbf{I}_{y_{jl}^d=z_l} + O(|\lambda_0|^2),\end{aligned} \tag{A.51}$$

$$\begin{aligned}\mathcal{L}^2_{y_j^d,y^d,\lambda_0} &= \prod_{s=1}^{r_y}\Big\{\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s} + \lambda_{0,s}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\Big\}^2\\ &= \prod_{s=1}^{r_y}\Big[\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s} + 2\lambda_{0,s}\sum_{z_s\le y_s^d}\sum_{z_s'\le y_s^d}\mathbf{I}_{y_{js}^d=z_s}\mathbf{I}_{y_{js}^d\neq z_s'}\Big] + O(|\lambda_0|^2)\\ &= \prod_{s=1}^{r_y}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s} + 2\sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\sum_{z_s'\le y_s^d}\mathbf{I}_{y_{js}^d=z_s}\mathbf{I}_{y_{js}^d\neq z_s'}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l^d}\mathbf{I}_{y_{jl}^d=z_l} + O(|\lambda_0|^2).\end{aligned} \tag{A.52}$$

Now we compute two intermediate results, $E[\mathcal{K}(y_j,y,\gamma_0)|x_j]$ and $E[\mathcal{K}(y_j,y,\gamma_0)^2|x_j]$. We will first use change-of-variable and Taylor expansion (which delivers $h_{0,s}^2$ terms) to handle the continuous variable

y^c, and then use $(A.51)$ to get an expansion of $\lambda_{0,s}$ for the discrete variable y^d so as to obtain the leading estimation bias term.

$$E[\mathcal{K}_{y_j,y,\gamma_0}|x_j] = E\left[G\left(\frac{y^c-y_j^c}{h_0}\right)\mathcal{L}(y_j^d,y^d,\lambda_0)|x_j\right] = \sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}\int f(y_j^c,y_j^d|x_j)G\left(\frac{y^c-y_j^c}{h_0}\right)dy_j^c$$

$$= -\sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}\int G(v)dm(y^c-h_0v,y_j^d|x_j) = \sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}\int w(v)m(y^c-h_0v,y_j^d|x_j)dv$$

$$= \sum_{y_j^d\in D_y^d}\mathcal{L}(y_j^d,y^d,\lambda_0)\Big[m(y^c,y_j^d|x_j)+(1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y}h_{0,s}^2m_{00,s}(y^c,y_j^d|x_j)+o(|h_0|^2)\Big]$$

$$= \sum_{y_j^d\in D_y^d}\Big[\prod_{s=1}^{r_y}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d=z_s}+\sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l^d\le y_l^d}\mathbf{I}_{y_{jl}^d=z_l}\Big]$$

$$\Big[m(y^c,y_j^d|x_j)+(1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y}h_{0,s}^2m_{00,s}(y^c,y_j^d|x_j)\Big]+o(\zeta_{1n}+|h_0|^2+|\lambda_0|)$$

$$= F(y^c,y^d|x_j)+(1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y}h_{0,s}^2F_{00,s}(y^c,y^d|x_j)$$

$$+\sum_{y_j^d\in D_y^d}\Big[\sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l^d}\mathbf{I}_{y_{jl}^d=z_l}\Big]m(y^c,y_j^d|x_j)+o(\zeta_{1n}+|h_0|^2+|\lambda_0|)$$

$$= F(y^c,y^d|x_j)+(1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y}h_{0,s}^2F_{00,s}(y^c,y^d|x_j)+\sum_{s=1}^{r_y}\lambda_{0,s}C_{1s,y,x}+o(\zeta_{1n}+|h_0|^2+|\lambda_0|). \tag{A.53}$$

where $C_{1s,y,x}=\sum_{y_j^d\in D_y^d}\Big[\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}^d\neq z_s}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l^d}\mathbf{I}_{y_{jl}^d=z_l}\Big]m(y^c,y_j^d|x)$.

Similarly, using Taylor expansion arguments and using $(A.52)$ we obtain

$$E[\mathcal{K}_{y_j,y,\gamma_0}^2|x_j] = E\left[G\left(\frac{y^c-y_j^c}{h_0}\right)^2\mathcal{L}_{y_j^d,y^d,\lambda_0}^2|x_j\right] = \sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}^2\int f(y_j^c,y_j^d|x_j)G\left(\frac{y^c-y_j^c}{h_0}\right)^2dy_j^c$$

$$= -\sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}^2\int G(v)^2dm(y^c-h_0v,y_j^d|x_j) = 2\sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}^2\int G(v)w(v)m(y^c-h_0v,y_j^d|x_j)dv$$

$$= \sum_{y_j^d\in D_y^d}\mathcal{L}_{y_j^d,y^d,\lambda_0}^2\Big[m(y^c,y_j^d|x_j)-\sum_{s=1}^{q_y}h_{0,s}C_wm_{0,s}(y^c,y_j^d|x_j)+O(h_0^2)\Big]$$

$$= \sum_{y_j^d\in D_y^d}\Big[\prod_{s=1}^{r_y}\sum_{z_s\le y_s^d}\mathbf{I}_{y_{js}=z_s}+2\sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\sum_{z_s'\le y_s^d}\mathbf{I}_{y_{js}=z_s}\mathbf{I}_{y_{js}\neq z_s'}\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l}\mathbf{I}_{y_{jl}=z_l}+O(|\lambda_0|^2)\Big]$$

$$\times\Big[m(y^c,y_j^d|x_j)-\sum_{s=1}^{q_y}h_{0,s}C_wm_{0,s}(y^c,y_j^d|x_j)+O(|h_0|^2)\Big]$$

$$= F(y^c,y^d|x_j)-\sum_{s=1}^{q_y}h_{0,s}C_wF_{0,s}(y^c,y^d|x_j)+2\sum_{y_j^d\in D_y^d}\Big[\sum_{s=1}^{r_y}\lambda_{0,s}\sum_{z_s\le y_s^d}\sum_{z_s'\le y_s^d}\mathbf{I}_{y_{js}=z_s}\mathbf{I}_{y_{js}\neq z_s'}$$

$$\prod_{l=1,l\neq s}^{r_y}\sum_{z_l\le y_l}\mathbf{I}_{y_{jl}=z_l}m(y^c,y_j^d|x_j)\Big]+O(|h_0|^2+|\lambda_0|^2)$$

$$= F(y^c,y^d|x_j)-\sum_{s=1}^{q_y}h_{0,s}C_wF_{0,s}(y^c,y^d|x_j)+\sum_{s=1}^{r_y}\lambda_{0,s}C_{2s,y,x}, \tag{A.54}$$

where $C_{2s,y,x} = 2\sum_{y_j^d \in D_y^d} \Big\{ \sum_{z_s \le y_s^d} \sum_{z_s' \le y_s^d} \mathbf{I}_{y_{js}=z_s} \mathbf{I}_{y_{js} \ne z_s'} \Big\} \prod_{l=1, l\ne s}^{r_y} \sum_{z_l \le y_l} \mathbf{I}_{y_{jl}=z_l} m(y^c, y_j^d | x_j)$.

Now we are ready to prove the lemma. Using (A.53) we immediately have

$$
\begin{aligned}
&E[(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})K_{\gamma,x_j,x}] = E\{E[(\mathcal{K}_{y_j,y,\gamma_0}|x_j]K_{\gamma,x_j,x}\} - E[F_{y|x}K_{\gamma,x_j,x}] \\
&= \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2[f(x)F_{ss,x} + 2f_{s,x}F_{s,x}] + (1/2)\kappa_2^{q_y}\sum_{s=1}^{q_y} h_{0,s}^2 F_{00,s} f(x) \\
&+ \sum_{s=1}^{r}\lambda_s \sum_{z_i^d \in S^D} \mathbf{I}_s(z^d, x^d)[F(y|x^c, z^d) - F(y|x)]f(x^c, z^d) + \sum_{s=1}^{r_y}\lambda_{0,s}C_{1s,y,x}f(x) + o(\zeta_{1n} + |\lambda_0| + |h_0|^2).
\end{aligned} \tag{A.55}
$$

This proves Lemma A.7 (i). Next, using (A.53) and (A.54) we obtain

$$
\begin{aligned}
&E[(\mathcal{K}_{y_j,y,\gamma_0} - F_{y|x})^2 K_{\gamma,x_j x}^2] = E[(\mathcal{K}_{y_j,y,\gamma_0}^2 + F_{y|x}^2 - 2\mathcal{K}_{y_j,y,\gamma_0}F_{y|x})K_{\gamma,x_j x}^2] \\
&= E[E(\mathcal{K}_{y_j,y,\gamma_0}^2 + F_{y|x}^2 - 2\mathcal{K}_{y_j,y,\gamma_0}F_{y|x}|x_j)K_{\gamma,x_j x}^2] \\
&= E[E(\mathcal{K}_{y_j,y,\gamma_0}^2|x_j)K_{\gamma,x_j x}^2] + F_{y|x}^2 E[K_{\gamma,x_j x}^2] - 2F_{y|x}E[E(\mathcal{K}_{y_j,y,\gamma_0}|x_j)K_{\gamma,x_j x}^2] \\
&= (h_1 \dots h_q)^{-1}\nu_0\Big[F(y|x) - F(y|x)^2 - \sum_{s=1}^{q_y} h_{0,s}C_w F_{0,s} + \sum_{s=1}^{r_y}\lambda_{0,s}\big(C_{2s,y,x} - 2F(y|x)C_{1s,y,x}\big)\Big] f(x).
\end{aligned} \tag{A.56}
$$

Note that $(h_1 \dots h_q)^{-1}\nu_0[F(y|x) - F(y|x)^2]/f(x)$ is the leading variance term. We also obtain the terms related to $h_{0,s}$ and $\lambda_{s,0}$ in the above variance expansion so that one can see that the multivariate y case has a variance expression similar to that in the scalar y case. Indeed if $q_y = 1$ and $r_y = 0$; or $q_y = 0$ and $r_y = 1$, we obtain results for scalar y as special cases. □

Appendix B: Proof of Theorem 3.1 and Theorem 3.2

Proof of Theorem 3.1. In Appendix B, we use F_i to denote the true conditional CDF $F(y|\bar{x}_i)$. We will focus on proving case (a), and we will only provide sketches for the proofs of cases (b) and (c). We will use the notation that $\bar{\zeta}_{1n} = |\bar{h}|^2 + |\lambda|$, $|\bar{h}|^2 = \sum_{s=1}^{q_1} h_s^2$, $|\bar{\lambda}| = \sum_{s=1}^{r_1}\lambda_s$, and $\bar{\zeta}_n = \bar{\zeta}_{1n}^2 + (nh_1 \dots h_{q_1})^{-1}$.

Proof of Theorem 3.1: Case (a). Following the same derivations that lead to (A.4), one can show that $CV_a(\cdot) = CV_{a,1}(\cdot)+$ a term unrelated to (h, λ), where

$$
\begin{aligned}
CV_{a,1}(\gamma) &= \int \Big[\frac{1}{n(n-1)^2}\sum\sum_{j\ne i}\sum_{l\ne i}\big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big)K_{\gamma,ji}K_{\gamma,li}/\hat{f}_{-i}^2 \\
&\quad - \frac{2}{n(n-1)}\sum\sum_{j\ne i}\big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_i - F_i\big)K_{\gamma,ji}/\hat{f}_{-i}\Big]\mathcal{M}_i M(y)dy \\
&= \int (A_{a,1n} - 2A_{a,2n})M(y)dy,
\end{aligned}
$$

where the definitions of $A_{a,1n}$ and $A_{a,2n}$ should be obvious.

In Lemma B.1 and Lemma B.2 below we show, uniformly in $(h, \lambda) \in \Gamma$, that

$$
\begin{aligned}
A_{a,1n} &= \int \Big(\sum_{s=1}^{q_1} h_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1}\lambda_s \bar{B}_{2s}(y|\bar{x})\Big)^2 \hat{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})d\bar{x} \\
&\quad + \int \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \dots h_{q_1}}\tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)dx + (s.o.)
\end{aligned} \tag{B.1}
$$

$$
A_{a,2n} = O_p(n^{-1/2}\zeta_{1n} + (n^2 h_1 \dots h_{q_1})^{-1/2}) = o_p(A_{1n}), \tag{B.2}
$$

where $\bar{\mathcal{M}}(\bar{x})$ is defined in (18).

$$\bar{B}_{1s}(y|\bar{x}) = \frac{\kappa_2}{2}\big[\bar{f}(\bar{x})F_{ss}(y|\bar{x}) + 2\bar{f}_s(\bar{x})F_s(y|\bar{x})\big]/\bar{f}(\bar{x}) \tag{B.3}$$

$$\bar{B}_{2s}(y|\bar{x}) = \sum_{\bar{z}^d \in S^d} I_s(\bar{z}^d, \bar{x}^d)\big[F(y|\bar{x}^c, \bar{z}^d) - F(y|\bar{x}^c, \bar{x}^d)\big]\bar{f}(\bar{x}^c, \bar{z}^d)/\bar{f}(\bar{x}) \tag{B.4}$$

$$\bar{\Sigma}_{y|\bar{x}} = \kappa^{q_1}[F(y|\bar{x}) - F(y|\bar{x})^2]/\bar{f}(\bar{x}) \tag{B.5}$$

$F_s(y|\bar{x}) = \partial F(y|\bar{x})/\partial \bar{x}_s^c$, $F_{ss}(y|\bar{x}) = \partial^2 F(y|\bar{x})/\partial(\bar{x}_s^c)^2$, $\bar{f}_s(\bar{x}) = \partial \bar{f}(\bar{x})/\partial \bar{x}_s^c$. Let $\int d\bar{x} = \sum_{\bar{x}^d} \int d\bar{x}^c$, $\int dx = \sum_{x^d} \int dx^c$. $\tilde{R}(\tilde{x}) = \tilde{R}(\tilde{x}, h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r)$ is defined by

$$\tilde{R}(\tilde{x}) = \frac{\nu_2(\tilde{x})}{[\nu_1(\tilde{x})]^2} \tag{B.6}$$

where for $i = 1, 2$, $\nu_i(\tilde{x}) = E\Big(\big[\prod_{s=q_1+1}^{q} h_s^{-1} w(\frac{x_{is}^c - x_s^c}{h_s}) \prod_{s=r_1+1}^{r} l(x_{is}^d, x_s^d, \lambda_s)\big]^i\Big)$.

Hence, the leading term of $CV_{a,1}(\gamma)$ is

$$\begin{aligned} &\iint \Big(\sum_{s=1}^{q_1} h_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1} \lambda_s \bar{B}_{2s}(y|\bar{x})\Big)^2 \bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})M(y)d\bar{x}dy \\ &+ \iint \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \ldots h_{q_1}} \tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)M(y)dxdy. \end{aligned} \tag{B.7}$$

By Hölder's inequality, $\tilde{R}(\tilde{x}) \geq 1$ for all choices of $\tilde{x}, h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r$. Also, $\tilde{R}(\tilde{x}) \to 1$ as $h_s \to \infty (q_1 + 1 \leq s \leq q)$ and $\lambda_s \to 1 (r_1 + 1 \leq s \leq r)$. Therefore, in order to minimize (B.7), one needs to select $h_s (s = q_1 + 1, \ldots, q)$ and $\lambda_s (s = r_1 + 1, \ldots, r)$ to minimize $\tilde{R}(\tilde{x})$. In fact, we show that the only bandwidth values for which $\tilde{R}(\tilde{x}, h_{q_1+1}, \ldots, h_q, \lambda_{r_1+1}, \ldots, \lambda_r) = 1$ are $h_s \to \infty$ for $q_1 + 1 \leq s \leq q$, and $\lambda_s = 1$ for $r_1 + 1 \leq s \leq r$. To see this, let us define $\mathcal{V}_n = \prod_{s=q_1+1}^{q} h_s^{-1} w((x_{is}^c - x_s^c)/h_s) \prod_{s=r_1+1}^{r} l(x_{is}^d, x_s^d, \lambda_s)$. If at least one h_s is finite (for $q_1 + 1 \leq s \leq q$), or one $\lambda_s < 1$ (for $r_1 + 1 \leq s \leq r$), then by (16) ($w(0) > w(\delta)$ for all $\delta > 0$) we know that $Var(\mathcal{V}_n) = E[\mathcal{V}_n^2] - [E(\mathcal{V}_n)]^2 > 0$ so that $\tilde{R}(\tilde{x}) = E(\mathcal{V}_n^2)/[E(\mathcal{V}_n)]^2 > 1$. Only when, in the definition of $\mathcal{V}_n$, all $h_s = \infty$ and all $\lambda_s = 1$, do we have $\mathcal{V}_n \equiv w(0)^{q-q_1}$ (a constant) and $Var(\mathcal{V}_n) = 0$ so that $\tilde{R}(\tilde{x}) = 1$ only in this case.

Therefore, in order to minimize (B.7), the bandwidths corresponding to the irrelevant covariates must all converge to their upper bounds so that $\tilde{R}(\tilde{x}) \to 1$ as $n \to \infty$ for all $\tilde{x} \in \tilde{S}$ ($\tilde{S}$ is the support of $\tilde{x}$). Thus irrelevant components are asymptotically smoothed out.

To analyze the behavior of bandwidths associated with the relevant covariates, we replace $\tilde{R}(\tilde{x})$ by 1 in (B.7), thus the second term on the right-hand-side of (B.7) becomes

$$\iint \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \ldots h_{q_1}} \bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)M(y)dxdy. \tag{B.8}$$

Define $a_s = h_s n^{1/(q_1+4)}$ and $b_s = \lambda_s n^{2/(q_1+4)}$, then (B.7) (with (B.8) as its first term since $\tilde{R}(\tilde{x}) \to 1$) becomes $n^{-4/(q_1+4)}\bar{\mathcal{X}}(a_1, \ldots, a_{q_1}, b_1, \ldots, b_{r_1})$, where

$$\begin{aligned} \bar{\mathcal{X}}(a_1, \ldots, b_{r_1}) &= (a_1 \ldots a_{q_1})^{-1} \iint \bar{\Sigma}_{y|\bar{x}} \bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)M(y)dxdy \\ &+ \iint \Big(\sum_{s=1}^{q_1} a_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1} b_s \bar{B}_{2s}(y|\bar{x})\Big)^2 \bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})M(y)d\bar{x}dy. \end{aligned} \tag{B.9}$$

Let $(a_1^0, \ldots, a_{q_1}^0, b_1^0, \ldots, b_{r_1}^0)$ denote values of $(a_1, \ldots, a_{q_1}, b_1, \ldots, b_{r_1})$ that minimize $\bar{\mathcal{X}}$ subject to each of them being non-negative. We require that

(B.10) Each a_s^0 is positive and each b_s^0 non-negative, all are finite and uniquely defined.

The approach of Li & Zhou (2005) can be used to obtain primitive necessary and sufficient conditions that ensure that (B.10) holds true. The result of Theorem 3.1 case (a) immediately follows. □

Proof of Theorem 3.1: Case (b). Similar to the derivation of (B.7), one can show that for case (b) the leading term of $CV_{b,1}$ has the following expression (by similar derivations as those lead to case (b) of Theorem 2.1)

$$CV_{b,L} = \iint \Big(\sum_{s=0}^{q_1} h_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1} \lambda_s \bar{B}_{2s}(y|\bar{x}) \Big)^2 \bar{f}(\bar{x}) \bar{\mathcal{M}}(\bar{x}) M(y) d\bar{x} dy$$
$$+ \iint \frac{\bar{\Sigma}_{y|\bar{x}} - h_0 \bar{\Omega}_1(y|\bar{x})}{n h_1 \dots h_{q_1}} \tilde{R}(\tilde{x}) \bar{f}(\bar{x}) \tilde{f}(\tilde{x}) \mathcal{M}(x) M(y) dx dy, \tag{B.11}$$

where $\bar{B}_{10}(y|\bar{x}) = \frac{\kappa_2}{2} F_{00}(y|\bar{x})$ and $\bar{\Omega}_1(y|\bar{x}) = \nu_0 C_w F_0(y|\bar{x})/f(\bar{x})$. They are the same as the quantities defined in Theorem 2.1 except that one replaces x by $\bar{x}$. Other quantities are the same as defined in the proof of case (a).

From (B.11) and using exactly the same arguments as we did in the proof of case (a), one can easily show that $h_s \sim n^{-1/(4+q_1)}$ $(s = 1, \dots, q_1)$ and $\lambda_s \sim n^{-2/(4+q_1)}$ $(s = 1, \dots, r_1)$, $h_s \to \infty$ for $s = q_1 + 1, \dots, q$ and $\lambda_s \to 1$ for $s = r_1 + 1, \dots, r$. With these results it is easy to show that $h_0 \sim n^{-2/(4+q_1)}$ because we need to select h_0 to minimize the squared bias terms and the variance term that are associated with h_0, i.e., terms like $\{h_0^4, h_0^2 h_s^2, h_0^2 \lambda_j, h_0/(n h_1 \dots h_{q_1})\}$ $(s = 1, \dots, q_1,\ j = 1, \dots, r_1)$. For example, if the h_0^4 term needs to balance $h_0 h_s^2$ (assuming their coefficients have opposite signs), then we get $h_0 \sim h_s \sim n^{-1/(4+q)}$. It is easy to see that h_0 cannot have an order larger than $O(h^{-1/(4+q_1)})$ as this would lead the estimation MSE to have an order larger than $O(n^{-4/(4+q_1)})$. Hence, we obtain $h_0 \sim n^{-1/(4+q_1)}$. The remaining steps of proving case (b) follow the same arguments as in the proof of case (a) and thus are omitted. □

Proof of Theorem 3.1: Case (c). Similar to the derivation of (B.7), one can show that for case (c) the leading term of $CV_{c,1}$ has the following expression

$$CV_{c,L} = \sum_{y \in D_y} \int \Big(\sum_{s=1}^{q_1} h_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=0}^{r_1} \lambda_s \bar{B}_{2s}(y|\bar{x}) \Big)^2 \bar{f}(\bar{x}) \bar{\mathcal{M}}(\bar{x}) M(y) d\bar{x}$$
$$+ \sum_{y \in D_y} \int \frac{\bar{\Sigma}_{y|\bar{x}} + \lambda_0 \bar{\Omega}_2(y|\bar{x})}{n h_1 \dots h_{q_1}} \tilde{R}(\tilde{x}) \bar{f}(\bar{x}) \tilde{f}(\tilde{x}) \mathcal{M}(x) M(y) dx, \tag{B.12}$$

where $\bar{B}_{20} = C_y - F(y|\bar{x})$, $\bar{\Omega}_2 = 2\kappa^{q_1}[F(y|\bar{x})^2 - F(y|\bar{x})]/f(\bar{x})$, all other quantities are the same as defined in the proof of case (a).

From (B.12) and using exactly the same arguments as we did in the proof of case (a), one can easily show that $h_s \sim n^{-1/(4+q_1)}$ $(s = 1, \dots, q_1)$ and $\lambda_s \sim n^{-2/(4+q_1)}$ $(s = 1, \dots, r_1)$, $h_s \to \infty$ for $s = q_1 + 1, \dots, q$ and $\lambda_s \to 1$ for $s = r_1 + 1, \dots, r$. With these results it is easy to show that $h_0 \sim n^{-2/(4+q_1)}$ because we need to select h_0 to minimize the squared bias terms and the variance term that are associated with h_0, i.e., terms like $\{\lambda_0^2, \lambda_0^2 h_s^2, \lambda_0 \lambda_j, \lambda_0/(n h_1 \dots h_{q_1})\}$ $(s = 1, \dots, q_1,\ j = 1, \dots, r_1)$. For example, if the λ_0^2 term needs to balance $\lambda_0 h_s^2$, then we get $\lambda_0 \sim h_s^2 \sim n^{-2/(4+q)}$. From this we obtain $\lambda_0 \sim n^{-2/(4+q_1)}$. The remaining steps of proving case (c) follow the same arguments as in the proof of case (a) and thus are omitted. □

Lemma B.1. *Equation (B.1) holds true.*

Proof. By Lemma B.3 we know that $\hat{f}_{-i}(x)$ is the kernel estimator of $\mu(x) = \bar{f}(\bar{x})\nu_1(\tilde{x})$, where $\nu_1(\tilde{x}) = E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}]$. Therefore, we know that (see Lemma B.3) the leading term of $\hat{f}_{-i}(x_i)^{-1}$ is $\mu(x_i)^{-1}$. Define A_1^0 by replacing $\hat{f}_{-i}(x_i)^{-1}$ in A_1 by its leading term $\mu(x_i)^{-1}$. Then using the result of Lemma

B.3, it is easy to show that $A_{a,1n} = A^0_{a,1n} + (s.o.)$. Hence, we only need to consider $A^0_{a,1n}$ which is defined by

$$\begin{aligned} A^0_{a,1n} &= \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i} \big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big)K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i \\ &\quad + \frac{1}{n(n-1)^2}\sum\sum_{j\neq i}\big(\mathbf{I}_j - F_i\big)^2 K^2_{\gamma,ji}\mu(x_i)^{-2}\mathcal{M}_i \\ &= G_{1n} + G_{2n}, \end{aligned}$$

where the definitions for G_{1n} and G_{2n} should be apparent.

We first consider G_{1n}, which can be written as a third order U-statistic. By the U-statistic H-decomposition, one can show that $G_{1n} = E(G_{1n}) + (s.o.)$.

$$\begin{aligned} E(G_{1n}) &= E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i\big] \\ &= E\left\{\big(E[(\mathbf{I}_j - F_i)K_{\gamma,ji}/\mu(x_i)|x_i]\big)^2\mathcal{M}_i\right\}. \end{aligned} \tag{B.13}$$

We first compute $E[(\mathbf{I}_j - F_i)K_{\gamma,ji}\mu(x_i)^{-1}|x_i]$. Recalling that $\mu(x) = \bar{f}(\bar{x})\nu_1(\tilde{x})$, we have (noting that $E[\tilde{K}_{\tilde{\gamma},ij}/\nu_1(\tilde{x}_i)|\tilde{x}_i] = 1$)

$$\begin{aligned} &E[(\mathbf{I}_j - F_i)K_{\gamma,ji}\mu(x_i)^{-1}|x_i] \\ &= E[(F_j - F_i)K_{\gamma,ji}\mu(x_i)^{-1}|x_i] \\ &= E[(F_j - F_i)\bar{K}_{\bar{\gamma},ij}\bar{f}(\bar{x}_i)^{-1}|\bar{x}_i]E[\tilde{K}_{\tilde{\gamma},ij}/\nu_1(\tilde{x}_i)|\tilde{x}_i] \\ &= \bar{f}(\bar{x}_i)^{-1}\sum_{\bar{z}^d\in\bar{S}^d} L(\bar{z}^d,\bar{x}^d_i,\lambda)\int[F(y|\bar{x}^c_i + hv,\bar{z}^d) - F(y|\bar{x}^c_i,\bar{x}^d_i)]\bar{f}(\bar{x}^c_i + hv,\bar{z}^d)W(v)dv \\ &= \frac{\kappa_2}{2}\sum_{s=1}^{q_1} h_s^2\big[\bar{f}(\bar{x}_i)F_{ss}(y|\bar{x}_i) + 2\bar{f}_s(\bar{x}_i)F_s(y|\bar{x}_i)\big]/\bar{f}(\bar{x}_i) \\ &\quad + \sum_{s=1}^{r_1}\lambda_s\sum_{\bar{z}^d\in\bar{S}^d} I_s(\bar{z}^d,\bar{x}^d_i)\big[F(y|\bar{x}^c_i,\bar{z}^d) - F(y|\bar{x}^c_i,\bar{x}^d_i)\big]\bar{f}(\bar{x}^c_i,\bar{z}^d)/\bar{f}(\bar{x}_i) + o(\zeta_n) \\ &= \sum_{s=1}^{q_1} h_s^2\bar{B}_{1s}(y|\bar{x}_i) + \sum_{s=1}^{r_1}\lambda_s\bar{B}_{2s}(y|\bar{x}_i) + o(\bar{\zeta}_n), \end{aligned} \tag{B.14}$$

uniformly in $(h,\lambda)\in\Gamma$, where $\bar{B}_{1s}(y|\bar{x})$ and $\bar{B}_{2s}(y|\bar{x})$ are defined in (B.3) and (B.4).

Substituting (B.14) into (B.13), we immediately obtain (recall $\bar{\mathcal{M}}(\bar{x})$ is defined in (18))

$$E(G_1) = \int\Big(\sum_{s=1}^{q_1} h_s^2\bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1}\lambda_s\bar{B}_{2s}(y|\bar{x})\Big)^2\bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})d\bar{x} + o(\bar{\zeta}_n). \tag{B.15}$$

Note that in the above we have only shown that for all fixed values of $(h,\lambda)\in\Gamma$, (B.15) holds true. By utilizing Rosenthal's and Markov's inequalities, it's straightforward to show (B.15) holds true uniformly in $(h,\lambda)\in\Gamma$.

Next we consider G_{2n}. G_{2n} can be written as a second order U-statistic. By the U-statistic H-decomposition it is straightforward to show that $G_{2n} = E(G_{2n}) + (s.o.)$. Recalling $\mu(x) = \bar{f}(\bar{x})\nu_1(\tilde{x})$,

$\nu_2(\tilde{x}) = E[\tilde{K}^2_{\tilde{\gamma},ji}|\tilde{x}_i = \tilde{x}]$, we have

$$
\begin{aligned}
E(G_{2n}) &= n^{-1}E\Big[\big(\mathbf{I}_j - F_i\big)^2 K^2_{\gamma,ji}\mu(x_i)^{-2}\mathcal{M}_i\Big] \\
&= n^{-1}E\Big\{E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)K^2_{\gamma,ji}\mu(x_i)^{-2}|x_i\big]\mathcal{M}_i\Big\} \\
&= n^{-1}E\Big\{E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)\bar{K}^2_{\bar{\gamma},i}\bar{f}(\bar{x}_i)^{-2}|\bar{x}_i\big]\mathcal{M}_i\nu_2(\tilde{x}_i)\nu_1(\tilde{x}_i)^{-2}\Big\}. \qquad \text{(B.16)}
\end{aligned}
$$

We first compute $E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)\bar{K}^2_{\bar{\gamma},ji}/\bar{f}(\bar{x}_i)^2|x_i\big]$. By Lemma B.4 we know that $h_s \to 0$ for $s = 1, \ldots, q_1$ and $\lambda_s \to 0$ for $s = 1, \ldots, r_1$. Thus

$$
\begin{aligned}
&E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)\bar{K}^2_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-2}|x_i\big] \\
&= E\big[(F_j - 2F_iF_j + F_i^2)\bar{K}^2_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-2}|x_i\big] \\
&= \frac{1}{h_1\ldots h_{q_1}}\sum_{\bar{z}^d\in\bar{S}^d} L(\bar{z}^d, \bar{x}_i^d, \lambda)^2 \int \big[F(y|\bar{x}_i^c + hv, \bar{z}^d) - 2F(y|\bar{x}_i^c, \bar{x}_i^d)F(y|\bar{x}_i^c + hv, \bar{z}^d) \\
&\quad + F(y|\bar{x}_i^c, \bar{x}_i^d)^2\big]\bar{f}(\bar{x}_i)^{-2}\bar{f}(\bar{x}_i^c + hv, \bar{z}^d)W(v)^2dv \\
&= \frac{\bar{\Sigma}_{y|\bar{x}_i}}{h_1\ldots h_{q_1}} + O(\bar{\zeta}_{1n}^{-1/2}(h_1\ldots h_{q_1})^{-1}), \qquad \text{(B.17)}
\end{aligned}
$$

where $\bar{\Sigma}_{y|\bar{x}}$ is defined in (B.5).

Substituting (B.17) into (B.16), we immediately obtain

$$
\begin{aligned}
E(G_{2n}) &= \int \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1\ldots h_{q_1}}\tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)dx + O(\bar{\zeta}_{1n}^{-1/2}(nh_1\ldots h_{q_1})^{-1}) \\
G_{2n} = E(G_2) + (s.o.) &= \int \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1\ldots h_{q_1}}\tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)dx + (s.o.),
\end{aligned}
$$

where $\tilde{R}(\tilde{x})$ is defined in (B.6).

Moreover, by utilizing Rosenthal's and Markov's inequalities, one can show that the above result holds uniformly in $(h, \lambda) \in \Gamma$. □

Lemma B.2. *Equation (B.2) holds true.*

Proof. Let $A^0_{a,2n}$ denote $A_{a,2n}$ with $\hat{f}_{-i}(x_i)^{-1}$ being replaced by its leading term $\mu(x_i)^{-1}$. Then it can be shown that $A_{a,2n} = A^0_{a,2n} + (s.o.)$. Hence, we only need to consider $A^0_{a,2n}$ which is defined by $A^0_{a,2n} = (n(n-1))^{-1}\sum\sum_{j\neq i}\big(\mathbf{I}_j - F_i\big)(\mathbf{I}_i - F_i)K_{\gamma,ji}\mu(x_i)^{-1}$. Notice that the part in $A^0_{a,2n}$ that is related to the irrelevant covariates is $\tilde{K}_{\tilde{\gamma},ji}/\nu_1(\tilde{x})$, which is bounded. Therefore, when evaluating the order of $A^0_{a,2n}$ we can ignore the irrelevant covariates part and need only consider

$$
\bar{A}^0_{a,2n} = \frac{1}{n(n-1)}\sum\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)\bar{K}_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-1}\mathcal{M}_i.
$$

Note that $\bar{A}^0_{a,2n}$ only depends on $(h_1\ldots, h_{q_1}, \lambda_1, \ldots, \lambda_{r_1})$. By Lemma B.4 we know that these bandwidths all converge to zero as $n \to \infty$. Hence, we can use standard change-of-variable and Taylor expansion arguments to deal with the continuous covariates' kernel function, and use the polynomial expansion for the discrete kernel functions. Note that $\mathcal{M}_i$ does not influence the order of $\bar{A}^0_{a,2n}$, so we

omit $\mathcal{M}_i$ in the following proof of this Lemma.

$$\begin{aligned}
E[\bar{A}^0_{a,2n}]^2 &= \frac{1}{n^2(n-1)^2}\sum_i\sum_{j\neq i}\sum_{l\neq i} E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)^2(\mathbf{I}_l - F_i)\bar{K}_{\bar{\gamma},ji}\bar{K}_{\bar{\gamma},li}\bar{f}(\bar{x}_i)^{-2}\big] \\
&= \frac{1}{n^2(n-1)^2}\sum\sum\sum_{l\neq j\neq i} E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)^2(\mathbf{I}_l - F_i)\bar{K}_{\bar{\gamma},ji}\bar{K}_{\bar{\gamma},li}\bar{f}(\bar{x}_i)^{-2}\big] \\
&\quad + \frac{1}{n^2(n-1)^2}\sum_i\sum_{j\neq i} E\big[(\mathbf{I}_j - F_i)^2(\mathbf{I}_i - F_i)^2\bar{K}^2_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-2}\big] \\
&= O\big(n^{-1}\bar{\zeta}^2_{1n} + (n^2 h_1\ldots h_{q_1})^{-1}\big).
\end{aligned}$$

Hence

$$\bar{A}^0_{a,2n} = O_p(n^{-1/2}\bar{\zeta}_{1n} + (n(h_1\ldots h_{q_1})^{1/2})^{-1}). \tag{B.18}$$

Moreover, by utilizing Rosenthal's and Markov's inequalities, one can show that (B.18) holds uniformly in $(h,\lambda)\in\Gamma$. Therefore, (B.2) holds. □

Lemma B.3. *Defining* $\nu_1(\tilde{x}) = E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}]$ *and* $\mu(x) = \bar{f}(\bar{x})\nu_1(\tilde{x})$, *then* $\hat{f}_{-i}(x)^{-1} = \mu(x)^{-1} + O_p\left(\bar{\zeta}_{1n} + \big(\ln(n)\big)^{1/2}(nh_1\ldots h_{q_1})^{-1/2}\right)$ *uniformly in* $x\in S$ *and* $(h,\lambda)\in\Gamma$.

Proof. Defining $\hat{\mu}(x) = E[\hat{f}_{-i}(x_i)|x_i = x]$, then by the independence of $\tilde{x}_i$ and $\bar{x}_i, y_i$, we have

$$\begin{aligned}
\hat{\mu}(x) &= E[\bar{K}_{\bar{\gamma},ij}|\bar{x}_i = \bar{x}]E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}] \\
&= \{\bar{f}(\bar{x}) + O(\bar{\zeta}_{1n})\}E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}] \\
&= \mu(x) + O_p(\bar{\zeta}_{1n}).
\end{aligned} \tag{B.19}$$

Note $\hat{f}_{-i}(x) - \hat{\mu}(x)$ has zero mean. Following standard arguments used when deriving uniform convergence rates for nonparametric kernel estimators (e.g., Masry (1996)), we know that

$$\hat{f}_{-i}(x) - \hat{\mu}(x) = O_p\left(\big(\ln(n)\big)^{1/2}(nh_1\ldots h_{q_1})^{-1/2}\right), \tag{B.20}$$

uniformly in $x\in S$ and $(h,\lambda)\in\Gamma$.

Combining $(B.19)$ and $(B.20)$ we obtain

$$\hat{f}_{-i}(x) - \mu(x) = O_p\left(\bar{\zeta}_{1n} + \big(\ln(n)\big)^{1/2}(nh_1\ldots h_{q_1})^{-1/2}\right), \tag{B.21}$$

uniformly in $x\in S$ and $(h,\lambda)\in\Gamma$.

Using (B.21) and Taylor expansions, we obtain

$$\begin{aligned}
\hat{f}_{-i}(x)^{-1} &= \big[\mu(x) + \hat{f}_{-i}(x) - \mu(x)\big]^{-1} \\
&= \mu(x)^{-1} - \mu(x)^{-2}\big[\hat{f}_{-i}(x) - \mu(x)\big] + O_p\big(|\hat{f}_{-i}(x) - \mu(x)|^2\big) \\
&= \mu(x)^{-1} + O_p\left(\zeta_n^{1/2} + \big(\ln(n)\big)^{1/2}(nh_1\ldots h_{q_1})^{-1/2}\right).
\end{aligned}$$

□

Lemma B.4. $\hat{h}_s = o_p(1)$ *for* $s = 1,\ldots,q_1$ *and* $\hat{\lambda}_s = o_p(1)$ *for* $s = 1,\ldots,r_1$.

Proof. Without assuming that any of the bandwidths converge to zero, then the only possible non-$o_p(1)$ term in $CV(\gamma)$ is G_{1n}. It is fairly straightforward to see that $G_{1n} = \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i}\big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big)K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i + o_p(1) \equiv G_{1,0} + o_p(1)$, where $\mu(x_i) = \bar{f}(\bar{x})E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i]$ is defined in the proof of Lemma B.3.

Note that $G_{1,0}$ can be written as a third order U-statistic, hence by the H-decomposition of a U-statistic it is fairly straightforward to show that $G_{1,0} = E(G_{1,0}) + o_p(1)$. Furthermore, by the law of iterated expectations we have

$$\begin{aligned} E(G_{1,0}) &= E\Big\{ \big[\mu(x_i)^{-1} E\big((\mathbf{I}_j - F_i)K_{\gamma,ji}|x_i\big)\big]^2 \mathcal{M}(x_i)\Big\} \\ &= E\left\{ \Big[\bar{f}(\bar{x}_i)^{-1} E\big((F_j - F_i)\bar{K}_{\bar{\gamma},ji}|\bar{x}_i\big)\Big]^2 \mathcal{M}(x_i)\right\} \\ &= E\big\{[\eta(y,\bar{x}_i)]^2 \mathcal{M}(x_i)\big\} = \int [\eta(y,\bar{x})]^2 \bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})d\bar{x}, \end{aligned} \tag{B.22}$$

where $\eta(y,\bar{x})$ is defined in (17), $\bar{\mathcal{M}}(\bar{x})$ is defined in (18). Note that the right hand side of (B.22) does not depend on $(h_{q_1+1},\ldots,h_q,\lambda_{r_1+1},\ldots,\lambda_r)$ since $E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i]$ in the numerator cancels with the same quantity in the denominator (from $\mu(x_i)^{-1} = \bar{f}(\bar{x})^{-1}E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i]^{-1}$).

If the bandwidths $(h_1,\ldots,h_{q_1},\lambda_1,\ldots,\lambda_{r_1})$ that minimize $CV(\gamma)$ do not all converge in probability to zero, then by (19), $E(G_{1,0})$(or G_{1n}) does not converge to zero, which implies that the probability that the minimum of G_{1n} (over the bandwidths) exceeds δ, which does not converge to zero as $n \to \infty$ (for some $\delta > 0$).

However, choosing $h_1,\ldots,h_{q_1}$ to be of size $n^{-1/(q_1+4)}$, and $\lambda_1,\ldots,\lambda_{r_1}$ to be of size $n^{-2/(4+q)}$, letting $h_{q_1+1},\ldots,h_q$ diverge to infinity, and letting $\lambda_{r_1+1},\ldots,\lambda_r$ converge to 1, one can easily show G_{1n} converges in probability to zero. This contradicts the result obtained in the previous paragraph (the minimum of G_{1n} exceeds δ), and thus demonstrates that, at the minimum of $CV(\gamma)$, the bandwidths $(h_1,\ldots,h_{q_1},\lambda_1,\ldots,\lambda_{r_1})$, for the relevant components of x, all converge in probability to zero. □

□

Proof of Theorem 3.2. Proof. We will only prove case (a) as cases (b) and (c) can be proved similarly. By Theorem 3.1 we know that $\hat{h}_s \xrightarrow{p} +\infty$ for $s = q_1+1,\ldots,q$ and $\hat{\lambda}_s \xrightarrow{p} 1$ for $s = r_1+1,\ldots,r$. Therefore, we need only consider the case with all irrelevant covariates removed, i.e. we consider $\hat{F}_a(y|\bar{x}) = [\sum_{j\neq i}\bar{K}_{\hat{\gamma},ji}]^{-1}[\sum_{j\neq i}\mathbf{I}_j\bar{K}_{\hat{\gamma},ji}]$, where $\bar{K}_{\hat{\gamma},ji} = \big[\prod_{s=1}^{q_1}\hat{h}_s^{-1}w((x_{is}^c - x_s^c)/\hat{h}_s)\big]\big[\prod_{s=1}^{r_1} l(x_{is}^d, x_s^d, \hat{\lambda}_s)\big]$.

We first consider the benchmark case whereby we use non-stochastic bandwidths. Define $h_s^0 = a_s^0 n^{-1/(4+q_1)}$ for $s = 1,\ldots,q_1$, and $\lambda_s^0 = b_s^0 n^{-2/(4+q_1)}$ for $s = 1,\ldots,r_1$, where a_s^0 and b_s^0 are defined in (B.10). Also, define $\bar{F}(y|\bar{x}) = \big[\sum_{j\neq i}\bar{K}_{\gamma^0,ji}\big]^{-1}\big[\sum_{j\neq i}\mathbf{I}_j\bar{K}_{\gamma^0,ji}\big]$, where $\bar{K}_{\gamma^0,ji} = \big[\prod_{s=1}^{q_1}(h_s^0)^{-1}w((x_{is}^c - x_s^c)/h_s^0)\big]\big[\prod_{s=1}^{r_1} l(x_{is}^d, x_s^d, \lambda_s^0)\big]$. Then,

$$\bar{F}(y|\bar{x}) - F(y|\bar{x}) = \big[\sum_{j\neq i}\bar{K}_{\gamma^0,ji}\big]^{-1}\big[\sum_{j\neq i}\mathbf{I}_j\bar{K}_{\gamma^0,ji} - \sum_{j\neq i}\bar{K}_{\gamma^0,ji}F(y|\bar{x})\big], \tag{B.23}$$

where $F(y|\bar{x})$ is the true conditional CDF. By adding and subtracting terms, we obtain

$$\begin{aligned} \bar{F}(y|\bar{x}) - F(y|\bar{x}) &= \Big[\sum_{j\neq i}\bar{K}_{\gamma^0,ji}\Big]^{-1}\Big[\sum_{j\neq i}\bar{K}_{\gamma^0,ji}\big(\mathbf{I}_j - \bar{F}_j + \bar{F}_j - F(y|\bar{x})\big)\Big] \\ &= \big[A^0(\bar{x})\big]^{-1}\big[B^0(y|\bar{x}) + C^0(y|\bar{x})\big], \end{aligned}$$

where $A^0(\bar{x}) = n^{-1}\sum_{j\neq i}\bar{K}_{\gamma^0,ji}$, $B^0(y|\bar{x}) = n^{-1}\sum_{j\neq i}\bar{K}_{\gamma^0,ji}\big[\mathbf{I}_j - \bar{F}_j\big]$ and $C^0(y|\bar{x}) = n^{-1}\sum_{j\neq i}\bar{K}_{\gamma^0,ji}$ $\big[\bar{F}_j - F(y|\bar{x})\big]$.

By the same arguments as we used in the proof of Lemma B.3, one can show that $A^0(\bar{x}) = \bar{f}(\bar{x}) + o_p(1)$. Following the proof of Lemma B.1, one can show that $C^0(y|\bar{x}) = \bar{f}(\bar{x})\big[\sum_{s=1}^{q_1}(h_s^0)^2\bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1}\lambda_s^0\bar{B}_{2s}(y|\bar{x})\big] + o_p(\zeta_n^0)$, where $\zeta_n^0 = \sum_{s=1}^{q_1}(h_s^0)^2 + \sum_{s=1}^{r_1}\lambda_s^0$. Obviously, $B^0(y|\bar{x})$ has zero mean and its asymptotic variance is given by $(nh_1^0\ldots h_{q_1}^0)^{-1}\bar{\Sigma}_{y|\bar{x}}\bar{f}(\bar{x})^2$, where $\bar{\Sigma}_{y|\bar{x}}$ is defined in (B.5). By applying

a triangular-array CLT, we know that

$$\sqrt{nh_1^0 \dots h_{q_1}^0}\left[\bar{F}(y|\bar{x}) - F(y|\bar{x}) - \sum_{s=1}^{q_1}(h_s^0)^2 \bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\lambda_s^0 \bar{B}_{2s}(y|\bar{x})\right] \xrightarrow{d} N(0, \bar{\Sigma}_{y|\bar{x}}). \tag{B.24}$$

Next we consider $\hat{F}_a(y|x) = \big[\sum_{j\neq i}\bar{K}_{\hat{\gamma},ji}\big]^{-1}\big[\sum_{j\neq i}\mathbf{I}_j\bar{K}_{\hat{\gamma},ji}\big]$ with cross-validation selected bandwidths, where $\bar{K}_{\hat{\gamma},ji} = \big[\prod_{s=1}^{q_1}\hat{h}_s^{-1}w((x_{is}^c - x_s^c)/\hat{h}_s)\big]\big[\prod_{s=1}^{r_1} l(x_{is}^d, x_s^d, \hat{\lambda}_s)\big]$. Therefore, the only difference between $\hat{F}_a(y|x)$ and $\bar{F}(y|\bar{x})$ is that the former uses the cross-validated bandwidths, while the latter uses some benchmark non-stochastic bandwidths. By Theorem 3.1 we know that $\hat{h}_s/h_s^0 \xrightarrow{p} 1$ for $s = 1, \dots, q_1$, and $\hat{\lambda}_s/\lambda_s^0 \xrightarrow{p} 1$ for $s = 1, \dots, r_1$. By using stochastic equicontinuity arguments as in Hall et al. (2004), one can show that $\hat{D}(y|x) - \bar{D}(y|\bar{x}) = o_p((nh_1^0 \dots h_q^0)^{-1/2})$, where $\hat{D}(y|x) = \hat{F}_a(y|\bar{x}) - F(y|\bar{x}) - \sum_{s=1}^{q_1}(\hat{h}_s)^2\bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\hat{\lambda}_s\bar{B}_{2s}(y|\bar{x})$ and $\bar{D}(y|\bar{x}) = \bar{F}(y|\bar{x}) - F(y|\bar{x}) - \sum_{s=1}^{q_1}(h_s^0)^2\bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\lambda_s^0\bar{B}_{2s}(y|\bar{x})$. Hence, $\hat{F}_a(y|x)$ and $\bar{F}(y|\bar{x})$ have the same asymptotic distribution, i.e.,

$$\sqrt{n\hat{h}_1 \dots \hat{h}_{q_1}}\left[\hat{F}_a(y|x) - F(y|\bar{x}) - \sum_{s=1}^{q_1}\hat{h}_s^2 \bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\hat{\lambda}_s \bar{B}_{2s}(y|\bar{x})\right] \xrightarrow{d} N(0, \bar{\Sigma}_{y|\bar{x}}). \tag{B.25}$$

□

□