DEPARTMENT OF ECONOMICS

WORKING PAPER SERIES

2012-10

McMASTER UNIVERSITY

Department of Economics
Kenneth Taylor Hall 426
1280 Main Street West
Hamilton, Ontario, Canada
L8S 4M4

http://www.mcmaster.ca/economics/

OPTIMAL BANDWIDTH SELECTION FOR NONPARAMETRIC CONDITIONAL DISTRIBUTION AND QUANTILE FUNCTIONS

QI LI
DEPARTMENT OF ECONOMICS, TEXAS A&M UNIVERSITY
COLLEGE STATION, TX 77843-4228, USA; AND
SCHOOL OF ECONOMICS AND MANAGEMENT, CAPITAL UNIVERSITY OF ECONOMICS AND BUSINESS,
BEIJING 100070, CHINA

JUAN LIN
RISK MANAGEMENT INSTITUTE, NATIONAL UNIVERSITY OF SINGAPORE
21 HENG MUI KENG TERRACE, SINGAPORE 119613

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

Abstract. We propose a data-driven least squares cross-validation method to optimally select smoothing parameters for the nonparametric estimation of conditional cumulative distribution functions and conditional quantile functions. We allow for general multivariate covariates that can be continuous, categorical or a mix of either. We provide asymptotic analysis, examine finite-sample properties via Monte Carlo simulation, and consider an application involving testing for first order stochastic dominance of children's health conditional on parental education and income.

1. Introduction

The nonparametric estimation of conditional probability density functions (PDFs) has received a substantial amount of attention in the literature (see e.g. Fan & Yim (2004), Hall, Racine & Li (2004), Chung & Dunson (2009), and Efromovich (2010)). In contrast, certain problems such as the estimation of conditional quantiles can require the estimation of conditional cumulative distribution functions (CDFs). Nonparametric estimation of the latter has proven more formidable but has drawn the attention of a growing number of researchers (see e.g. Bashtannyk & Hyndman (2001), Hyndman & Yao (2002), among others). Since the seminal work of Koenker & Bassett (1978) appeared, quantile regression has exploded in popularity and a range of innovative approaches towards the estimation of conditional quantiles have been proposed including Yu & Jones (1998), Koenker & Xiao (2004), Li, Liu & Zhu (2007), and Peng & Huang (2008) to name but a few. The distinguishing feature of the approach considered herein, however, is the direct estimation of the conditional distribution function

Date: August 2, 2012.

using kernel methods that smooth both continuous and categorical variables in a particular manner – optimal bandwidth selection for conditional CDFs is therefore central to this approach.

It is widely appreciated that smoothing parameter selection is a key determinant of sound nonparametric estimation. A range of effective methods have been developed for selecting smoothing parameters optimally for the estimation of unconditional PDFs, conditional mean functions and conditional PDFs (see Marron, Jones & Sheather (1996), Hall, Li & Racine (2007), and Hall et al. (2004), among others). However, to the best of our knowledge, there does not exist in the literature an optimal data-driven method for choosing bandwidths when estimating conditional CDF and conditional quantile functions in general multivariate nonparametric settings. Although in principle one can always compute optimal smoothing parameters using some 'plug-in' methods, in a general multivariate setting, particularly when some of the covariates are independent of the response variables, 'plug-in' methods become infeasible (see e.g. Hall et al. (2004)). Recent work on nonparametric estimation of conditional CDF and quantile functions has pointed out that bandwidth selection for nonparametric quantile estimators remains an open topic of research (Li & Racine (2008, page 429)). In this paper we aim to fill this gap in the literature by providing an automatic data-driven method for selecting the smoothing parameters optimally in the sense that a weighted estimation mean squared error (MSE) of the conditional CDF and conditional quantile is minimized.

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 constitutes a popular estimation methodology (see Koenker (2005)) and may be predicated directly on an estimated conditional CDF. Section 5 assesses the finite-sample performance of the proposed method 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 categorical/discrete and continuous variables. Let $x = (x^c, x^d)$, 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. Let x^d_{is} (x^d_s) denote the s^{th} component of x^d_i (x^d), $s = 1, \dots, r$; $i = 1, \dots, n$, where n is the sample size. We assume that x_s takes values in $\{0, 1, \dots, c_s - 1\}$, where $c_s \geq 2$ is a positive integer.

Let λ denote the bandwidth for a discrete variable. For ordered discrete variables we use the kernel $l(X_{is}^d, x_s^d, \lambda_s) = \lambda_s^{|X_{is}^d - x_s^d|}$ (with $\lambda_s^0 = 1$ and $0^0 = 1$), and for unordered discrete variables we use the kernel $l(X_{is}^d, x_s^d, \lambda_s) = \mathbf{1}(X_{is}^d = x_s^d) + \lambda_s \mathbf{1}(X_{is}^d \neq x_s^d)$, where $\mathbf{1}(A) = 1$ if A holds, and 0 otherwise. 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 $L_\lambda(\cdot)$ given by $K_\gamma(x_i, x) = W_h(x_i^c, x^c) \times L_\lambda(x_i^d, x^d, \lambda)$, 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. We propose two conditional CDF estimators. The first one (case (a)) is given by

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

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 $f(x)$. Note that in $\hat{F}(y|x)$, y_j can be either a continuous or a discrete variable.

The second estimator (case (b)) smooths the dependent variable y_j (assuming that y_j is a continuous variable) and is defined by

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

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) and where h_0 is the bandwidth associated with y.

We suggest choosing bandwidths by minimizing the following cross-validation function,

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

where $\mathcal{M}(\cdot)$ and $M(\cdot)$ are trimming functions with bounded support. We need to introduce the trimming functions M and $\mathcal{M}$ that have bounded support to ensure the CV objective function is finite even for X and Y that have unbounded support. When X and Y have bounded support, trimming functions are not needed and one can simply replace M and $\mathcal{M}$ by 1. In the simulations reported in Section 5, we generated (X, Y) having a joint normal distribution, which has unbounded support. But simulated random draws are always bounded, and replacing M and $\mathcal{M}$ by 1 does not create any problems in our simulations.

If y is a discrete variable, then one should replace $\int dy$ by $\sum_{y \in D_y}$ in (2), where D_y is the support of y_i (discrete), and

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

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 $f(x_i)$.

We make the following assumptions.

Condition 1. *$\{X_i, Y_i\}_{i=1}^n$ 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 $s = 1, \ldots r$, $n(h_1 \ldots h_q) \to \infty$.*

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 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 $s = 1, \dots, q$, let $F_s(y|x) = \partial F(y|x)/\partial x_s$, $F_{ss}(y|x) = \partial^2 F(y|x)/\partial x_s^2$, $\kappa_2 = \int w(v)v^2 dv$, and $\nu_0 = \int W(v)^2 dv$. The next theorem gives the leading terms for $CV(\cdot)$ (note that 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).

Theorem 2.1. *Letting $CV(\gamma)$ be defined in (2) and also assuming that conditions 1 to 3 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):*

$$CV_L(\gamma) = \iint \left\{ \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \right]^2 + \frac{\Sigma_{y|x}}{nh_1 \cdots h_q} \right\} f(x)\mathcal{M}(x)M(y)dxdy,$$

while if y is discrete, $\int dy$ above needs to be replaced by $\sum_{y \in D_y}$, where $B_{10}(y|x) = 0$ for case (a), and $B_{10}(y|x) = \frac{\kappa_2}{2} F_{00}(y|x)$ for case (b), $B_{1s}(y|x) = \frac{\kappa_2}{2}\big[f(x)F_{ss}(y|x) + 2f_s(x)F_s(y|x)\big]/f(x)$, for $s = 1, \dots, q$, $B_{2s}(y|x) = \sum_{z^d \in S^d} \mathbf{I}_s(z^d, x^d)\big[F(y|x^c, z^d) - F(y|x)\big] f(x^c, z^d)/f(x)$, for $s = 1, \dots, r$, $\Sigma_{y|x} = \nu_0[F(y|x) - F(y|x)^2]/f(x)$, $\Omega_1 = \nu_0 C_w F_0(y|x)/f(x)$, $\Omega_2 = 2\nu_0\big[F(y|x)^2 - F(y|x)\big]/f(x)$, $C_w = 2\int G(v)w(v)vdv$.

Theorem 2.1 is proved in Appendix A.

It can be shown that the estimation MSE of $\hat{F}(y|x)$ has the following leading term,

$$MSE_L[\hat{F}(y|x)] = \Big[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \Big]^2 + \frac{\Sigma_{y|x}}{nh_1 \cdots h_q}, \tag{3}$$

Comparing $CV_L(\cdot)$ of Theorem 2.1 with (3), we observe that

$$CV_L = \iint MSE_L[\hat{F}(y|x)] f(x)\mathcal{M}(x)M(y)dxdy.$$

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.

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 1 - 3, we have*

(i) $n^{1/(4+q)}\hat{h}_s \xrightarrow{p} a_s^0$, $s = 1, \dots, q$ for case (a), and $s = 0, 1, \dots, q$ for case (b);

(ii) $n^{2/(4+q)}\hat{\lambda}_s \xrightarrow{p} b_s^0$, $s = 1, \dots, r$,

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

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.

We can obtain the rate of convergence of the CV selected bandwidths if we strengthen Condition 1 to the following condition:

Condition 4. *$\{X_i, Y_i\}_{i=1}^n$ are independent and identically distributed as (X, Y), $f(x)$ and $F(y|x)$ have uniformly continuous fourth-order partial derivative functions with respect to x^c and y (if y is a continuous variable).*

Condition 4 requires that $f(x)$ and $F(y|x)$ have uniformly continuous fourth-order partial derivative functions. The next theorem gives the rate of convergence for the CV selected smoothing parameters.

Theorem 2.3. *Let $h_{0,s} = a_s^0 n^{-1/(4+q)}$ for $s = 1, \dots, q$ for case (a), and $s = 0, 1, \dots, q$ for case (b), and $\lambda_{0,s} = b_s^0 n^{-2/(4+q)}$ for $s = 1, \dots, r$. Then under conditions 2 - 4, we have*

(i) If $q \le 3$, $(\hat{h}_s - h_{0,s})/h_{0,s} = O_p(n^{-q/[2(4+q)]})$ for $s = 0, 1, \dots, q$, and $\hat{\lambda}_s - \lambda_{s,0} = O_p(n^{-1/2})$ for $s = 1, \dots, r$;

(ii) If $q \le 3$, $(\hat{h}_s - h_{0,s})/h_{0,s} = O_p(h_{s,0}^2) = O_p(n^{-2/(4+q)})$ for $s = 0, 1, \dots, q$, and $\hat{\lambda}_s - \lambda_{s,0} = O_p(h_{s,0}^4) = O_p(n^{-4/(4+q)})$ for $s = 1, \dots, r$.

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

Theorem 2.4. *Under conditions 1 - 3, we have*

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

One problem with the $CV(\cdot)$ function defined in (2) is that it involves numerical integration, which can be computationally costly. 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:

(4)

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

where $\mathcal{M}_i = \mathcal{M}(X_i)$ is the same weight function used in (2). The advantage of using (4) 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 (4) is similar to those described by Theorem 2.2, while the resulting estimator has the same asymptotic distribution as described in Theorem 2.4. We direct the interested reader to an implementation of (4) that is available in the R (R Core Team (2012)) package 'np' (see the function 'npcdistbw', version 0.40-14 and higher, Hayfield & Racine (2008)).

Theorem 2.5. *Let $M(y) = g(y)$, where $g(y)$ is the marginal density (probability function) of y (y can be either continuous or discrete), then under conditions 1 to 3, $CV(\gamma)$ defined in (2) and $CV_\Sigma(\gamma)$ defined in (4) are asymptotically equivalent in the sense that they have the same leading term, i.e.,*

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

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 2.5 is given in Appendix A.

From Theorem 2.5 we immediately obtain the following useful results.

Theorem 2.6. *If one chooses the bandwidths by minimizing $CV_\Sigma(\cdot)$, then Theorem 2.2 and Theorem 2.4 remain valid with the only modification being that one replaces $M(\cdot)$ by $g(\cdot)$.*

Theorem 2.6 follows directly from theorems 2.2, 2.4 and 2.5. Therefore, its proof is omitted.

2.2. **The Multivariate y Case.** When y is multivariate we write $y = (y_1, \dots, y_p) = (y_1^c, \dots, y_{q_y}^c, y_1^d, \dots, y_{r_y}^d)$ 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. For expositional simplicity, we will only discuss the non-smooth Y case (a) in the multivariate y setting (the subscript m below is taken to mean 'multivariate' y),

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

where $\mathbf{I}(y_j \le y) = \prod_{s=1}^{p} \mathbf{I}(y_{js} \le y_s)$ is the product of indicator functions.

We again propose selecting bandwidths via leave-one-out cross-validation by minimizing (where $\int dy = \sum_{y^d \in D_y} \int dy^c$)

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

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

where $\hat{F}_{-i}(y|x_i)$ is the leave-one-out estimator of $F(y|x_i)$ given by $\hat{F}_{m,-i} = (n-1)^{-1} \sum_{j \neq i} \mathbf{I}(y_j \leq y) K_\gamma(x_j, x_i)/\hat{f}_{-i}(x_i)$, and $\hat{F}_{-i}(y_j|x_i)$ is obtained from $\hat{F}_{-i}(y|x_i)$ with y replaced by y_j.

It is easy to show that theorems 2.2 to 2.6 remain valid except that now $F(y|x)$ is understood to be $F(y_1, \ldots, y_p|x)$.

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

Next, we consider the case for which one or more of the covariates 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' covariates 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 et al. (2004) we shall assume that

$$\bar{x}, y \text{ is independent of } \tilde{x}. \tag{5}$$

Assumption (5) 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

$$\text{Conditional on } \bar{x}, \text{ the variables } \tilde{x} \text{ and } y \text{ are independent.} \tag{6}$$

However, using (6) will cause some technical difficulties in the proof of our main result. Therefore, in this paper we will only consider unconditional independence given in (5) though we point out that extensive simulations carried out (not reported here for space considerations) indicate that all results indeed follow under (6).

Note that the conditional CDF of $F(y|x)$ is $F(y|\bar{x})$. This is because under Assumption (5), 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 covariates $x = (\bar{x}, \tilde{x})$, still using Equation (1). 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. (2004). Define

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

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

$n^{\epsilon - 1} \leq H \leq n^{-\epsilon}$, $0 < h_0 \leq n^{-c}$, $n^{-c} < h_s < n^{c}$ for all $s = 1, \ldots, q$; the kernel $w(\cdot)$ is a symmetric, compactly supported, Hölder-continuous probability density; and $w(0) > w(\delta)$ for all $\delta > 0$.

The above conditions basically ask that each h_s, $s = 1, \ldots, q$, does not converge to zero, or to infinity, too fast, and that $nh_1 \ldots h_{q_1} \to \infty$ as $n \to \infty$.

We use $\mathcal{H}$ to denote the permissible set for $(h_0, h_1, \ldots, h_q)$ that satisfies (3). 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 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 (9) 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

$$\eta(y, \bar{x}) = \bar{f}(\bar{x})^{-1} E\big[(F(y|\bar{x}_j) - F(y|\bar{x}_i)) \bar{K}_{\bar{\gamma},ji} | \bar{x}_i = \bar{x}\big]. \tag{7}$$

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

$$\bar{\mathcal{M}}(\bar{x}) = \int \tilde{f}(\tilde{x}) \mathcal{M}(x) d\tilde{x}. \tag{8}$$

We assume that

$$\iint[\eta(y,\bar{x})]^2\bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})M(y)d\bar{x}dy, \text{ as a function of } h_1,\dots,h_{q_1} \text{ and } \lambda_1,\dots,\lambda_{r_1},$$

$$\text{vanishes if and only if all of the bandwidths vanish.} \tag{9}$$

In Lemma B.4 in Appendix B we show that (3) and (9) imply that as $n\to\infty$, $h_s\to 0$ for $s=1,\dots,q_1$ and $\lambda_s\to 0$ for $s=1,\dots,r_1$. Therefore, the bandwidths associated with the relevant covariates all vanish asymptotically. In Appendix B, we also show that $h_s\to\infty$ for all $s=q_1+1,\dots,q$ and $\lambda_s\to 1$ for all $s=r_1+1,\dots,r$. 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 addition to conditions 1 to 4, assume that conditions (3), (9) and (B.10) also hold, and let* $\hat{h}_1,\dots,\hat{h}_q,\hat{\lambda}_1,\dots,\hat{\lambda}_r$ *denote the bandwidths that minimize* $CV(\gamma)$. *Then*

$n^{1/(q_1+4)}\hat{h}_s\to a_s^0$ *in probability,* $1\le s\le q_1$ *for case* (a), *and* $0\le s\le q_1$ *for case* (b),

$P(\hat{h}_s>C)\to 1$ *for* $q_1+1\le s\le q$ *and for all* $C>0$,

$n^{2/(q_1+4)}\hat{\lambda}_s\to b_s^0$ *in probability for* $1\le s\le r_1$,

$\hat{\lambda}_s\to 1$ *in probability for* $r_1+1\le s\le r$.

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 MSE (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 integrated MSE (with only relevant covariates used in the estimation). Therefore, the CV method leads to optimal smoothing in the sense of minimizing a weighted integrated MSE asymptotically.

From Theorem 3.1 one can easily obtain the following result.

Theorem 3.2. *Under the same conditions given in Theorem 3.1, when x is an interior point of $S = S^c \times S^d$ (the support of X), then*

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

where $\bar{B}_{10}(y|\bar{x}) = 0$ for case (a), and $\bar{B}_{10}(y|\bar{x}) = \frac{\kappa_2}{2} F_{00}(y|\bar{x})$ for case (b), $\bar{B}_{1s}(y|\bar{x})$ 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.

Note that our justification for smoothing the discrete covariates relies mainly on the fact that there may be 'irrelevant' discrete covariates in which case smoothing out these irrelevant discrete covariates will reduce estimation variance without increasing estimation bias. However, for discrete covariates that are not irrelevant, we have not provided a formal justification for smoothing the discrete covariates. An anonymous referee suggested that one might be able to justify smoothing to discrete covariates from a Bayesian perspective if one's prior has the conditional distributions of Y being dependent across different (adjacent) values of the regressors. We are certainly sympathetic to this point of view. The relationship between smoothing discrete covariates and Bayesian priors in a regression setting has recently been established in Kiefer & Racine (2009), who provide a deeper understanding of kernel smoothing methods for discrete data by leveraging the unexplored links between hierarchical Bayes models and kernel methods for discrete processes. As a detailed discussion on this issue in the setting we consider is beyond the scope of the current paper, we direct the interested reader to in Kiefer & Racine (2009), and leave this issue as a subject for further study.

4. Estimating Conditional Quantile Functions

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

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x). \tag{10}$$

Since $F(y|x)$ is (weakly) monotone in y, inverting (10) leads to a unique solution for $q_\alpha(x)$. In this section we will focus on using $\hat{F}(y|x)$ to obtain a quantile estimator for $q_\alpha(x)$. Therefore, we propose the following estimator for estimating $q_\alpha(x)$:

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\}, \tag{11}$$

where $\hat{F}(y|x)$ is defined in Section 2 with CV selected bandwidths. The CV objective function can be either $CV(\cdot)$ defined in (2) or $CV_\Sigma(\cdot)$ defined in (4).

Because $\hat{F}(y|x)$ is monotone in y, (11) 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 $q_\alpha(x)$.

Because $\hat{F}(y|x)$ lies between zero and one and is monotone in y, $\hat{q}_\alpha(x)$ 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,

$$\hat{q}_\alpha(x) = \arg\min_{q_\alpha} |\, \alpha - \hat{F}(q_\alpha|x)\,|. \tag{12}$$

That is, the value of q_α that minimizes (12) gives us $\hat{q}_\alpha(x)$. We make the following assumption.

Condition 5. *The conditional PDF $f(y|x)$ 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 $B_n(y|x) = [\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x)]$ is the leading bias term of $\hat{F}(y|x)$ (with $y = q_\alpha(x)$). Then, under conditions 1 to 5, we have*

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

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 similar arguments as the proof of Theorem 3 of Cai (2002) given the results of Theorem 3.2 above. Thus, the proof of Theorem 4.1 is omitted.

An associate editor has noted that one of the fundamental problems with quantile estimation in practice is the problem of quantile crossing, resulting in bizarre conclusions like a 70th percentile exceeding, say, a 75th percentile conditional on covariates. This arises due to the use of quantile models that are predicated on a regression-type model where the least squares objective function has been replaced with the so-called check function. The approach we propose, however, is immune to this drawback *provided* that the user uses a second order nonnegative kernel function, for example the second order Epanechnikov or Gaussian weight functions (i.e. Condition 2 holds). This is because we are estimating a conditional CDF directly, and the estimator $\hat{F}(y|x)$ is guaranteed to be monotone in y when the kernel functions are nonnegative, which in turn implies that our quantile estimator $\hat{q}_\alpha(x)$ is monotone in α.

In certain semiparametric settings, when one wishes to establish a parametric $\sqrt{n}$ rate of convergence for some finite dimensional parameters, typically one has to use higher order kernels to reduce the bias so that the estimation error is of smaller order than $O(n^{-1/2})$. Since we only consider fully nonparametric quantile estimators, hence do not have finite dimensional parameters in our setup, there is no need to use higher order kernel functions here. However, if one were tempted to use a higher order kernel for Y to reduce bias in our framework, quantile crossing could occur, i.e., the estimated quantile function $\hat{q}_\alpha(x)$ may not be monotone in α, hence, for different α, estimated quantile curves could well cross each other. In such cases, in order to avoid the quantile crossing problem one could use the 'rearrangement' method to generate a new quantile estimator, say, $\hat{q}^*_\alpha(x)$ (obtained based on rearrangement of $\hat{q}_\alpha(x)$), and that $\hat{q}^*_\alpha(x)$ is monotone in α (see Chernozhukov, Fernndez-Val & Galichon (2010, Page 1097) for a detailed description as how to generate $\hat{q}^*_\alpha(x)$). Our recommendation, however, is to restrict attention to second order kernel functions only thereby guaranteeing that quantiles cannot cross and sidestepping this issue completely, which we believe is one of the strengths of our approach.

5. Monte Carlo Simulations and Empirical Application

In this section we examine the finite-sample performance of the proposed method of cross-validated conditional CDF bandwidth selection. We numerically minimize the objective functions $CV(h_x) = n^{-1}\sum_{i=1}^n \int_{-\infty}^{\infty} \left\{\mathbf{I}(y_i \le y) - \hat{F}_{-i}(y|x_i)\right\}^2 dy$ and $CV(h_x) = n^{-2}\sum_{i=1}^n \sum_{j=1}^n \left\{\mathbf{I}(y_i \le y_j) - \hat{F}_{-i}(y_j|x_i)\right\}^2$. Having computed the bandwidths we then compute the sample MSE of the estimators of $F(y|x)$ for both the conditional CDF and PDF-based (Hall et al. (2004)) bandwidths via

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

In the tables that follow we report the ratio of the median MSE of the proposed method versus that of Hall et al. (2004) where the median of each approach over all Monte Carlo replications is compared (i.e. we compare the location of the distribution of MSEs for each approach). Alternatively one could compute the median of the set of ratios for each draw, though this does not appear to qualitatively alter our findings. A second order Gaussian kernel was used throughout.

5.1. **Comparison of Integral Versus Summation Approach.** We first assess how the integration-based approach compares with the summation-based one in finite-sample settings. 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 the proposed method and that appropriate for conditional PDF estimation (Hall et al. (2004)) that has been used in the literature given the lack of a method tailored to the estimation of the conditional CDF.

Table 1. Relative median efficiency of kernel estimators of conditional CDFs using the proposed bandwidth method versus that appropriate for conditional PDFs. Numbers less than 1 indicate superior MSE performance.

Nonsmooth Y summation variant, leftmost table, smooth Y summation variant, rightmost table

	$n=50$	$n=100$	$n=200$
$\rho=0.95$	0.97	0.97	0.98
$\rho=0.85$	0.97	0.98	1.00
$\rho=0.75$	0.99	0.97	1.01
$\rho=0.50$	0.99	0.99	0.97
$\rho=0.25$	1.04	0.95	0.93

	$n=50$	$n=100$	$n=200$
$\rho=0.95$	0.87	0.88	0.89
$\rho=0.85$	0.88	0.87	0.92
$\rho=0.75$	0.92	0.91	0.92
$\rho=0.50$	0.90	0.91	0.89
$\rho=0.25$	0.97	0.91	0.87

Nonsmooth Y integration variant, leftmost table, smooth Y integration variant, rightmost table

	$n=50$	$n=100$	$n=200$
$\rho=0.95$	0.97	0.99	0.97
$\rho=0.85$	0.97	0.96	1.00
$\rho=0.75$	0.99	0.97	1.05
$\rho=0.50$	0.99	0.92	0.97
$\rho=0.25$	1.02	0.98	0.95

	$n=50$	$n=100$	$n=200$
$\rho=0.95$	0.87	0.88	0.89
$\rho=0.85$	0.88	0.87	0.91
$\rho=0.75$	0.91	0.89	0.93
$\rho=0.50$	0.87	0.89	0.88
$\rho=0.25$	0.92	0.86	0.85

Table 1 reveals that the proposed method delivers bandwidths that dominate those based on conditional PDF bandwidth selection in finite-sample settings, while smoothing Y improves finite-sample MSE relative to the nonsmooth Y variant. This is an expected result as conditional PDF bandwidth selection will not be optimal for conditional CDFs. Furthermore, the computational burden associated with numerical integration does not appear to be necessary in order to achieve the MSE improvement (e.g. the median search time for $n=50$, smooth Y approach, $K=1$ increases by a factor of 30 when using numerical integration). Note also that numerical integration for the nonsmooth Y approach degrades computation time even further.

5.2. **Irrelevant Categorical Covariates.** Next, we take the data generating process used above but now add an additional categorical covariate $z \in \{0, 1\}$, $\Pr(z = 1) = 0.5$, that is independent of y, i.e., satisfying (5), but this is not presumed to be known a-priori hence is included in the covariate set. Results are presented in Table 2 below for the summation-based approach only in light of the performance of the summation versus integration approach reported in Table 1 above (similar results are obtained when the additional covariate is continuous and are not reported for space considerations).

TABLE 2. Irrelevant z summation-based relative median efficiency of kernel estimators of conditional CDFs using the proposed method versus that appropriate for conditional PDFs. Numbers less than 1 indicate superior MSE performance.

Nonsmooth Y variant, leftmost table, smooth Y variant, rightmost table

	$n = 50$	$n = 100$	$n = 200$
$\rho = 0.95$	0.87	0.94	0.96
$\rho = 0.85$	0.88	0.94	0.97
$\rho = 0.75$	0.92	0.95	0.96
$\rho = 0.50$	0.91	0.93	0.93
$\rho = 0.25$	0.98	0.92	0.92

	$n = 50$	$n = 100$	$n = 200$
$\rho = 0.95$	0.85	0.87	0.89
$\rho = 0.85$	0.86	0.90	0.92
$\rho = 0.75$	0.91	0.90	0.90
$\rho = 0.50$	0.92	0.94	0.89
$\rho = 0.25$	0.99	0.92	0.91

Table 2 reveals again that the smooth Y variant has improved finite-sample MSE relative to the nonsmooth Y variant, which becomes more pronounced as n increases. We also note that the bandwidth λ_z for the categorical 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 (bandwidth results are not reported here for space considerations).

5.3. **Testing for Stochastic Dominance – Children's Health and Parental Education.** The relationship between children's health and parental endowments such as income and education level touches a wide number of disciplines. A popular measure of children's health is based on hemoglobin, the protein molecule that is present in red blood cells (it carries oxygen from the lungs to the body's tissues and returns carbon dioxide from the tissues to the lungs). Anemia (a low hemoglobin level) is associated with malnourishment and is widely used as a measure of health by the World Health Organization, among others. Normal results for men range from 13-18 grams per decilitre (g/dl). For women the normal range is 12-16 g/dl.

We consider a dataset used by Maasoumi & Lugo (2008) that comes from the 2000 Indonesian Family Life Survey conducted by RAND, UCLA and the Demographic Institute of the University of Indonesia for a group in Indonesia, the 'Sunda' (the Sunda are the largest unreached people group in

Indonesia, and they reside in the province of West Java). Interest lies in the outcome children's 'health' (hemoglobin, g/dl) conditional on the covariates level of education of the household head ('education') and household income per capita ('income' per household member). We consider the following variables for $n = 4254$ observations, i) hb.ema: levels of hemoglobin adjusted by gender and age ('health'), ii) hdeduc00: level of education of the head of the household ('education'), and iii) pce00: per capita expenditures ('income') (in December 2000, the exchange rate for the Rupiah was Rp.9,480 / 1 US dollar).

In what follows we treat the covariate education as an ordered factor (i.e. discrete) and income as continuous, while health is treated as continuous. We consider testing for dominance relationships of

$$G = F(\text{health} \mid \text{education=low}) \text{ versus } F = F(\text{health} \mid \text{education=high}),$$

and also we consider testing for dominance relationships of

$$G = F(\text{health} \mid \text{education=low, income}) \text{ versus } F = F(\text{health} \mid \text{education=high, income}),$$

where we additionally control for the level of income given potential dependence among the covariates education and income (income is held constant at its median value). For low and high values of education we use 2 and 12 years corresponding to the 15th and 85th percentiles of the data, respectively. That is, we first assess whether differences in parental education are associated with different health outcomes, and then assess whether this difference reflects instead differences in parental resources as measured by per capita income.

5.3.1. *Testing First Order Stochastic Dominance: Definitions and Tests.* Let W and V denote a health variable measured, say, at either two different points in time, or for different regions or countries. Let $W_1, W_2, \ldots, W_n$ be n observations on W, and $V_1, V_2, \ldots, V_m$ be similar observations on V. Let $\mathcal{U}$ denote the class of all utility functions s such that $s' \geq 0$, (increasing). Let $W_{(i)}$ and $V_{(i)}$ denote the i-th order statistics, and assume $F(x)$ and $G(x)$ are continuous and monotonic CDF's of W and V, respectively.

Definition 5.1. *W First Order Stochastic Dominates V, denoted W FSD V, if and only if any one of the following equivalent conditions holds:*

(1) $E[u(W)] \geq E[u(V)]$ for all $u \in \mathcal{U}$, with strict inequality for some u.

(2) $F(x) \leq G(x)$ for all $x \in \mathcal{R}$ with strict inequality for some x.

We consider testing for FSD based upon (2) in Definition 5.1 using the following Kolmogorov-Smirnov (KS) statistic:

$$D = \min\left\{\sup_x (G(x) - F(x)), \sup_x (F(x) - G(x))\right\},$$

where $G(x)$ and $F(x)$ are CDFs that differ in their covariate values (below G and F are the CDFs of health holding income and/or education fixed at specific values), and the CDFs are evaluated at a common grid of x values. $D < 0$ indicates FSD of F over G (or G over F), while $D \geq 0$ indicates no such dominance. We therefore consider testing the hypothesis $H_0 : D \geq 0$ versus $H_1 : D < 0$. Note that when we reject H_0 we conclude that F FSDs G (from the estimated curves we know F FSDs G).

We elect to use a nonparametric bootstrap method whereby we impose the null that $G = F$ (i.e. $D \geq 0$). We do this by first drawing a bootstrap sample pairwise, and then once more bootstrapping (i.e. shuffling in place leaving the remaining variables unchanged) the covariate variable(s) only thereby removing any systematic relationship between the covariates variable(s) and the outcome for the bootstrap sample. We can then compute the nonparametric P-value which is given by

$$\hat{P} = B^{-1} \sum_{b=1}^{B} I_{D < D_b^*},$$

where B is the number of bootstrap replications and where $I_{D<D_b^*}$ is an indicator function equal to one when the sample statistic D is less than the bootstrap statistic computed under the null (D_b^*) and zero otherwise. In other words, it is the proportion of bootstrap statistics more extreme than the sample statistic (i.e. more negative).

5.3.2. *Results and Discussion.* Figure 1 presents the conditional CDFs G and F for low/high education using the proposed bandwidth method (we restrict the plotting range of the X axis to the .005th and .995th quantiles of the data, while plots that condition on income are similar to those presented and are not included for space considerations). Bandwidths were $(\hat{h}_{\mathrm{h}}, \hat{h}_{\mathrm{e}}) = (0.328, 0.042)$ and $(\hat{h}_{\mathrm{h}}, \hat{h}_{\mathrm{e}}, \hat{h}_{\mathrm{i}}) = (0.392, 0.640, 108869)$ when covariates are education and both education and income, respectively. A dominance relationship is apparent in both figures indicating that education levels are associated with divergent health outcomes even after controlling for parental incomes per capita.

For each test conducted, we conduct $B = 999$ bootstrap replications and evaluate F and G on a common set of 1000 points. We then report the statistic D along with its P-value. For these types of tests, comparison of the distributions is performed over the range for which either $\hat{F}$ or $\hat{G}$ is above ϵ and below $1 - \epsilon$ (we set $\epsilon = 0.025$) as tail noise is known to impact power.

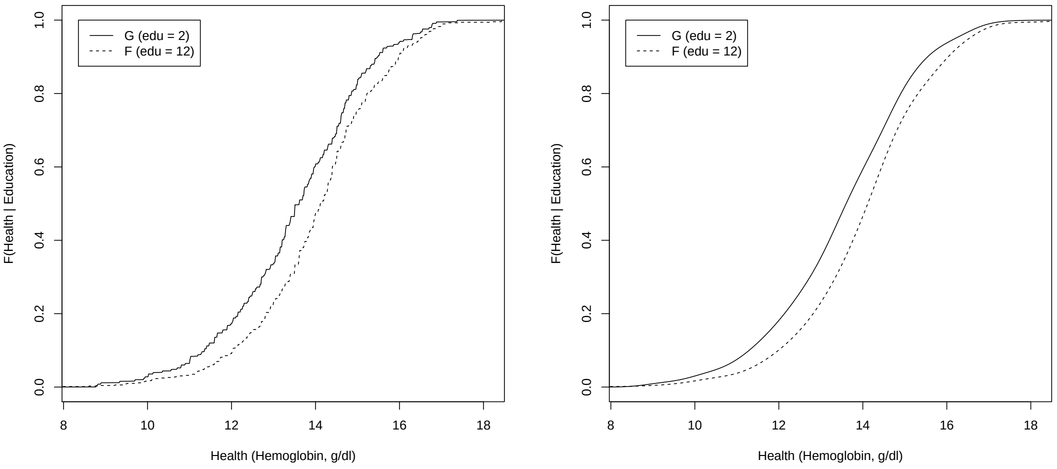

Figure 1. Nonsmooth Y (left figure) and smooth Y (right figure) conditional CDF plots.

The smooth Y D statistics from the nonparametric SD tests are -0.0113 and -0.00589 when covariates are education and both education and income, respectively. The smooth Y P-values from the nonparametric SD tests are 0.00501 and 0.018 when covariates are education and both education and income, respectively. On the basis of this test we would conclude that there is statistical evidence that the distribution of children's health in households having a high education household head FSDs those having a low education household head, even after controlling for differences in household per capita income. The nonsmooth Y D statistics from the nonparametric SD tests are -0.0102 and -0.00506 when covariates are education and both education and income, respectively. The nonsmooth P-values from the nonparametric SD tests are 0.001 and 0.00701 when covariates are education and both education and income, respectively.

These results indicate that high levels of parental education are associated with improved health outcomes for their children even after controlling for potential differences in household per capita income.

Stochastic dominance testing is also proving quite useful in the growing literature on inference in (conditional) moment inequalities, and the approach detailed in this illustrative example could be directly applied to this and a range of other applications.

6. Acknowledgements

We would like to thank the Associate Editor, Co-Editor, and two anonymous referees whose insightful comments and suggestions helped to greatly improve this paper. 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). Li's research is partially supported by research funds from the Social Sciences and Humanities Research Council of Canada, and from the National Natural Science Foundation of China (Projects # 70773005 and # 71131008).

Appendix A. Proofs of Theorems 2.1, 2.2, 2.3, 2.4 and 2.5

Throughout Appendices A and B we will only consider/prove the non-smooth Y case (a) for all the theorems as the proofs of these theorems for the smooth Y case (b) are similar to that of case (a).

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) Unless otherwise stated, we will use $\sum_i = \sum_{i=1}^n$, $\sum_{j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n$, $\sum_{j\neq i}\sum_{l\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n \sum_{l=1,l\neq i}^n$, $\sum_{l\neq j\neq i} = \sum_{i=1}^n \sum_{j=1,j\neq i}^n \sum_{l=1,l\neq i,l\neq j}^n$, $\sum_{j>i} = \sum_{i=1}^{n-1} \sum_{j>i}^n$, $\sum_{l>j>i} = \sum_{i=1}^{n-2} \sum_{j>i}^{n-1} \sum_{l>j}^n$.

(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 $|h|^2 = \sum_{s=1}^q h_s^2$, $|\lambda|^2 = \sum_{s=1}^r \lambda_s^2$, $\zeta_{1n} = |h|^2 + |\lambda|$, $\zeta_n = \zeta_{1n}^2 + \ln(n)/(nh_1 \dots h_q)$ and $\xi_n = \zeta_n^{3/2} + (\zeta_{1n} + n^{-1/2})(nh_1 \dots h_q)^{-1}$.

The only exceptions to the rule (2) above are $f_{i,0}$ and $f_{i,1s}$ defined below (A.4). From $f_{i,0} = (n-1)^{-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\mathbf{I}(x_j^d = x_i^d)$, it is obvious that here $\sum_{j\neq i} = \sum_{j=1,j\neq i}^n$ involves only a single summation because the left hand side of the equation depends on i.

Proof of Theorem 2.1. Denote by $\hat{F}_{-i} = \hat{F}_{-i}(y|x_i)$. We need to show that $CV(\cdot) = CV_L(\cdot) + (s.o.)$, where $(s.o.)$ contains terms unrelated to bandwidths or terms having smaller order than $CV_L(\cdot)$. Also, the smaller order terms are uniformly small for all $\gamma \in \Gamma$ (as defined in Section 3). We rewrite (2) as (by adding/subtracting terms: $(\hat{F}_{-i} - \mathbf{I}_i)^2 = (\hat{F}_{-i} - F_i + F_i - \mathbf{I}_i)^2$),

$$CV(\cdot) = \frac{1}{n}\sum_i \int \big[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i) + (F_i - \mathbf{I}_i)^2\big]\mathcal{M}_i M(y)dy.$$

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 $CV(\cdot)$ over $(h_1, \dots, h_q, \lambda_1, \dots, \lambda_r)$ is equivalent to minimizing $CV_1(\cdot)$, where $CV_1(\cdot)$ is defined as

$$\begin{aligned} CV_1(\cdot) &= \frac{1}{n}\sum_i \int \big[(\hat{F}_{-i} - F_i)^2 - 2(\hat{F}_{-i} - F_i)(\mathbf{I}_i - F_i)\big]\mathcal{M}_i M(y)dy \\ &= \int \Big[\frac{1}{n(n-1)^2}\sum_{j\neq i}\sum_{l\neq i}\int (\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/\hat{f}^2_{-i} \\ &\qquad - \frac{2}{n(n-1)}\sum_{j\neq i}\int (\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}/\hat{f}_{-i}\Big]\mathcal{M}_i M(y)dy \\ &= \int (S_{1n} - 2S_{2n})M(y)dy \qquad \text{(A.1)} \end{aligned}$$

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

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

$$\text{(A.2)} \quad S_{1n} = \int \Big\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\Big\}^2 f(x)\mathcal{M}(x)dx + \int \frac{\Sigma_{y|x}}{nh_1 \dots h_q} f(x)\mathcal{M}(x)dx + O_p(\zeta_{1n}\zeta_n)$$

$$\text{(A.3)} \quad S_{2n} = O_p\big((n^{-1/2}\zeta_n) + (n(h_1 \dots h_q)^{1/2})^{-1}\big).$$

Combining (A.1), (A.2) and (A.3), we have shown that

$$CV_1(\cdot) = CV_L(\cdot) + O_p\big(\zeta_{1n}\zeta_n + (n^{-1/2}\zeta_n) + (n(h_1 \dots h_q)^{1/2})^{-1}\big),$$

where CV_L is defined in Theorem 2.1.

This completes the proof of Theorem 2.1. □

A technical difficulty in handling (A.1) 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:

$$\frac{1}{\hat{f}_{-i}} = \frac{1}{f_i} + \frac{f_i - \hat{f}_{-i}}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3} + \frac{(f_i - \hat{f}_{-i})^3}{f_i^3 \hat{f}_{-i}}. \tag{A.4}$$

Define $f_{i,0} = (n-1)^{-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\mathbf{I}(x_j^d = x_i^d)$, $f_{i,1s} = (n-1)^{-1}\sum_{j\neq i} W_h(x_j^c, x_i^c)\mathbf{I}_s(x_j^d, x_i^d)$ and using $\mathbf{I}(x_j^d = x_i^d) + \sum_{s=1}^r \lambda_s \ \mathbf{I}_s(x_j^d, x_i^d) + O(|\lambda|^2)$, we have uniformly in $1 \leq i \leq n$,

$$f_i - \hat{f}_{-i} = (f_i - f_{i,0}) - \sum_{s=1}^r \lambda_s f_{i,1s} + O_p(|\lambda|^2). \tag{A.5}$$

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

Therefore, we have

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

Substituting (A.5) and (A.6) into (A.4), we obtain uniformly in $1 \leq i \leq n$ and $x \in \mathcal{S}$

$$\frac{1}{\hat{f}_{-i}} = \frac{1}{f_i} + \frac{(f_i - \hat{f}_{-i})}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3} + O_p(\zeta_n^{3/2}). \tag{A.7}$$

From (A.7), we also obtain uniformly in $1 \leq i \leq n$ and $x \in \mathcal{S}$

$$\frac{1}{\hat{f}_{-i}^2} = \frac{1}{f_i^2} + \frac{2(f_i - \hat{f}_{-i})}{f_i^3} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^4} + O_p(\zeta_n^{3/2}). \tag{A.8}$$

Both (A.7) and (A.8) will be used to handle the random denominator in the proofs that follow.

Lemma A.1. *Equation (A.2) holds true.*

Proof. We omit the weight 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.2) holds true with S_{1n} being replaced by S_{1n}^0 and that $S_{1n} - S_{1n}^0 = o_p(\zeta_n)$.

$$\begin{aligned} S_{1n}^0 &= \frac{1}{n(n-1)^2}\sum_{j\neq i}(\mathbf{I}_j - F_i)^2 K_{\gamma,ji}^2/f_i^2 + \frac{1}{n(n-1)^2}\sum_{l\neq j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2 \\ &= S_{1n,1} + S_{1n,2}, \end{aligned} \tag{A.9}$$

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. $S_{1n,2} = 1/(n(n-1)^2)\sum_{l\neq j\neq i} Q_{ijl}$, 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

$$\begin{aligned} S_{1n,2} &= EQ_i + \frac{3}{n}\sum_i (Q_i - EQ_i) + \frac{6}{n(n-1)}\sum_{j>i}(Q_{ij} - Q_i - Q_j + EQ_i) \\ &\quad + \frac{6}{n(n-1)(n-2)}\sum_{l>j>i}(Q_{ijl} - Q_{ij} - Q_{jl} - Q_{li} + Q_i + Q_j + Q_l - EQ_i) \\ &= J_0 + J_1 + J_2 + J_3 \qquad \text{(A.10)} \end{aligned}$$

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

$$\begin{aligned} J_0 &= E(Q_i) = E(Q_{ijl}) = E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}/f_i^2\big] \\ &= E\Big\{E\big[(\mathbf{I}_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2 = E\Big\{E\big[(F_j - F_i)K_{\gamma,ji}|x_i\big]/f_i\Big\}^2. \qquad \text{(A.11)} \end{aligned}$$

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

$$\begin{aligned} E\big[(F_j - F_i)K_{\gamma,ji}|x_i\big] &= \sum_{z^d\in S^d} L(z^d, x_i^d, \lambda)\int [F(y|x_i^c + hv, z^d) - F(y|x_i)]f(x_i^c + hv, z^d)W(v)dv \\ &= \sum_{z^d\in S^d}\left[\mathbf{I}(z^d = x_i^d) + \sum_{s=1}^r \lambda_s \mathbf{I}_s(z^d, x_i^d) + O(|\lambda|^2)\right] \times \int \Big\{\big[F(y|x_i^c, z^d) - F(y|x_i) \\ &\quad + \sum_{s=1}^q F_s(y|x_i^c, z^d)h_s v_s + (1/2)\sum_{s=1}^q\sum_{t=1}^q F_{st}(y|x_i^c, z^d)h_s h_t v_s v_t + O(|h|^3)\big] \\ &\quad \big[f(x_i^c, z^d) + \sum_{s=1}^q f_s(x_i^c, z^d)h_s v_s + O(|h|^2)\big]\Big\}W(v)dv \\ &= \frac{\kappa_2}{2}\sum_{s=1}^q h_s^2\big[f(x_i)F_{ss}(y|x_i) + 2f_s(x_i)F_s(y|x_i)\big] \\ &\quad + \sum_{s=1}^r \lambda_s \sum_{z^d\in S^d}\mathbf{I}_s(z^d, x_i^d)\big[F(y|x_i^c, z^d) - F(y|x_i)\big]f(x_i^c, z^d) + O(\zeta_{1n}^2), \qquad \text{(A.12)} \end{aligned}$$

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

Plugging (A.12) into (A.11), we have

$$\begin{aligned}
J_0 &= E\Big\{ \frac{\kappa_2}{2} \sum_{s=1}^{q} h_s^2 \big[f(x_i)F_{ss}(y|x_i) + 2f_s(x_i)F_s(y|x_i)\big] f_i^{-1} \\
&\qquad + \sum_{s=1}^{r} \lambda_s \sum_{z^d \in S^d} \mathbf{I}_s(z^d, x_i^d)\big[F(y|x_i^c, z^d) - F(y|x_i)\big] f(x_i^c, z^d) f_i^{-1} \Big\}^2 + O(\zeta_{1n}^3) \\
&= E\Big\{ \sum_{s=1}^{q} h_s^2 B_{1s}(y|x_i) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x_i) \Big\}^2 + O(\zeta_{1n}^3) \\
&= \int \Big\{ \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \Big\}^2 f(x) dx + O(\zeta_{1n}^3), \qquad \text{(A.13)}
\end{aligned}$$

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

$$S_{1n,2} = \int \Big\{ \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \Big\}^2 f(x) dx + O_p(\zeta_{1n}^3 + n^{-1/2}\zeta_{1n}).$$

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

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

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

$$\begin{aligned}
\mathcal{J}_0 &= n^{-1}E(\mathcal{Q}_i) = n^{-1}E(\mathcal{Q}_{ij}) = n^{-1}E\big\{(\mathbf{I}_j - F_i)^2 K^2_{\gamma,ji} f_i^{-2}\big\} \\
&= n^{-1}E\big\{(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)K^2_{\gamma,ji} f_i^{-2}\big\} \\
&= n^{-1}E\Big\{E\big[(F_j - 2F_iF_j + F_i^2)K^2_{\gamma,ji}|x_i\big] f_i^{-2}\Big\} \\
&= E\big[\nu_0(nh_1\dots h_q)^{-1}(F_i - F_i^2)f_i^{-1}\big] + O((nh_1\dots h_q)^{-1}\zeta_n) \\
&= E\left(\frac{\Sigma_{y|x_i}}{nh_1\dots h_q}\right) + O((nh_1\dots h_q)^{-1}\zeta_n) \\
&= \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)dx + O((nh_1\dots h_q)^{-1}\zeta_n) \qquad \text{(A.14)}
\end{aligned}$$

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

$$S_{1n,1} = \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)dx + O_p((n^{-1/2} + \zeta_{1n})(nh_1\dots h_q)^{-1}).$$

Summarizing the above we have shown that

$$S^0_{1n} = \int \Big\{\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\Big\}^2 f(x)\mathcal{M}(x)dx + \int \frac{\Sigma_{y|x}}{nh_1\dots h_q} f(x)\mathcal{M}(x)dx + O_p(\xi_n),$$

where $\xi_n = \zeta_n^{3/2} + (\zeta_{1n} + n^{-1/2})(nh_1\dots h_q)^{-1}$.

Next we show that $S_{1n} - S^0_{1n} = O_p(\xi_n)$. By Equation (A.8),

$$\begin{aligned}
S_{1n} - S^0_{1n} &= \frac{1}{n(n-1)^2}\sum_{j\neq i}(I_j - F_i)^2_{\gamma,ji}\Big(\frac{1}{\hat{f}_i^2} - \frac{1}{f_i^2}\Big) \\
&\quad + \frac{1}{n(n-1)^2}\sum_{j\neq l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\Big(\frac{1}{\hat{f}_i^2} - \frac{1}{f_i^2}\Big) \\
&= \frac{1}{n(n-1)^2}\sum_{j\neq l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\Big[\frac{2(f_i - \hat{f}_{-i})}{f_i^3} \\
&\qquad + \frac{(f_i - \hat{f}_{-i})^2}{f_i^4} + O_p\big(n^{-1}\zeta_{1n} + \zeta_n^{3/2}\big)\Big] \\
&= O_p\left(\xi_n\right), \qquad \text{(A.15)}
\end{aligned}$$

where $\xi_n = \xi_n = \zeta_n^{3/2} + (\zeta_{1n} + n^{-1/2})(nh_1\dots h_q)^{-1}$. This is because the two terms $2/(n(n-1)^2)\sum_{j\neq l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}(f_i - \hat{f}_{-i})/f_i^3$ and $1/(n(n-1)^2)\sum_{j\neq l\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}(f_i - \hat{f}_{-i})^2/f_i^4$

can be written as fourth-order and fifth-order U-statistics, respectively. Tedious but straightforward calculations can show that both of these two terms are $O_p(\xi_n)$. 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 the leading order $O(|h|^4 + |\lambda|^2 + (nh_1 \dots h_q)^{-1})$. □

Lemma A.2. *Equation (A.3) 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.3) 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,

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

Hence,

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

Note that since S_{2n}^0 has zero mean, the above results imply that we can write S_{2n}^0 as

$$S_{2n}^0 = n^{-1/2}\zeta_{1n}\tilde{\mathcal{X}}_2 + (n^2 h_1 \dots h_q)^{-1/2}\tilde{\mathcal{X}}_3, \tag{A.16}$$

where $\tilde{\mathcal{X}}_2$ and $\tilde{\mathcal{X}}_3$ are zero mean $O_p(1)$ random variables. Equation (A.16) will be used in the proof of Theorem 2.3.

Next, using (A.7) we have

$$\begin{aligned}
S_{2n} - S_{2n}^0 &= \frac{1}{n(n-1)} \sum_{j\neq i} (\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i) K_{\gamma,ji} \Big(\frac{1}{f_i} - \frac{1}{\hat{f}_{-i}}\Big) \\
&= \frac{1}{n(n-1)} \sum_{j\neq i} (\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i) K_{\gamma,ji} \Big[\frac{(f_i - \hat{f}_{-i})}{f_i^2} + \frac{(f_i - \hat{f}_{-i})^2}{f_i^3}\Big] + O_p\big(\zeta_n^{3/2}\big) \\
&= O_p\big(\xi_n\big),
\end{aligned}$$

where $\xi_n = \zeta_n^{3/2}+(\zeta_{1n}+n^{-1/2})(nh_1\dots h_q)^{-1}$. The last equality follows from U-statistic H-decomposition. Because $1/(n(n-1))\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}(f_i - \hat{f}_{-i})/f_i^2$ and $1/(n(n-1))\sum_{j\neq i}(\mathbf{I}_j - F_i)(\mathbf{I}_i - F_i)K_{\gamma,ji}(f_i - \hat{f}_{-i})^2/f_i^3$ 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 by a factor of $\zeta_n^{1/2}$. □

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. The proof of Theorem 2.3 follows by combining the proofs for Theorem 2.1 and the proof method of Theorem 2.2 of Racine & Li (2004).

We provide a sketch of the proof here. To save space, we will make a notational simplification. We assume that $h_1 = \dots = h_q = h$ and $\lambda_1 = \dots = \lambda_r = \lambda$. Then the result of Theorem 2.1 simplifies to

$$CV_L = B_1 h^4 + B_2 h^2\lambda + B_3\lambda^2 + B_4(nh^q)^{-1},$$

where B_js are some constants, $j = 1, 2, 3, 4$. In fact by explicitly considering the other non-leading terms, one can show that

$$\begin{aligned} S_{1n} =& B_1 h^4 + B_2 h^2\lambda + B_3\lambda^2 + B_4(nh^q)^{-1} \\ & + B_5 h^6 + B_6 h^4\lambda + B_7 h^2\lambda^2 + B_8\lambda^3 + B_9(h^2+\lambda)(nh^q)^{-1} \\ & n^{-1/2}\zeta_{1n}\mathcal{X}_{1n} + (s.o.), \qquad \text{(A.17)} \\ S_{2n} =& n^{-1/2}\zeta_{1n}\tilde{\mathcal{X}}_2 + (nh^{q/2})^{-1}\tilde{\mathcal{X}}_3 + (s.o.), \qquad \text{(A.18)} \end{aligned}$$

where B_j's are some constants for $j = 1, \dots, 9$, and $\mathcal{X}_{jn}$, $j = 1, 2, 3$, are zero mean $O_p(1)$ random variables.

The derivations of (A.17) and (A.18) follow the same method as in the proof of Theorem 2.1. For example, by using Taylor expansion to higher order terms in the derivation of (A.13), one can easily show that

$$\begin{aligned} J_0 =& E\left\{h^2\tilde{B}_1 + \lambda\tilde{B}_2 + h^4\tilde{B}_3 + h^2\lambda\tilde{B}_4 + \lambda^2\ \tilde{B}_5\right\}^2 + (s.o.) \\ =& h^4\bar{B}_1 + h^2\lambda\bar{B}_2 + \lambda^2\bar{B}_3 + h^6\bar{B}_4 + h^4\lambda\bar{B}_5 + h^2\lambda^2\bar{B}_6 + \lambda^3\bar{B}_7 + (s.o), \qquad \text{(A.19)} \end{aligned}$$

where $\tilde{B}_j = \tilde{B}_j(y, x_i)$, the expectation is with respect to x_i, and $\bar{B}_j = \bar{B}_j(y)$.

Similarly, J_1 defined in (A.10) has zero mean and its second moment is of the order $O(n^{-1}\zeta_{1n}^2)$. Hence, if we write $J_1 = n^{-1/2}\zeta_{1n}^{1/2}\tilde{\mathcal{X}}_1$, or equivalently, $\tilde{\mathcal{X}}_1 \overset{def}{=} n^{1/2}\zeta_{1n}^{-1/2}J_1$, then obviously $\tilde{\mathcal{X}}_1$ is a zero mean $O_p(1)$ random variable. Also, (A.18) follows from (A.16). Now it should be clear that by explicitly considering smaller order terms in $CV_1(\cdot)$, one can prove (A.17) and (A.18).

The remaining steps are to show that (A.17) and (A.18) together imply the conclusion of Theorem 2.3. The detailed analysis follows *exactly* the same arguments as in the proof of Theorem 2.2 of Racine & Li (2004). Therefore, we omit the remaining proofs here. □

Proof of Theorem 2.4. Theorem 2.4 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 2.5. Recall that $CV_\Sigma = n^{-2}\sum_{j\neq i}^{n}\left[\hat{F}_{-i}(y_j|x_i) - \mathbf{I}(y_i \leq y_j)\right]^2\mathcal{M}_i$. Denote $\mathbf{I}_{ji} = \mathbf{I}(y_i \leq y_j)$, $\hat{F}_{-i,ji} = \hat{F}_{-i}(y_j|x_i)$, $F_{ji} = F(y_j|x_i)$. 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

$$CV_{\Sigma,1} = \frac{1}{n^2}\sum_{j\neq i}(\hat{F}_{-i,ji} - F_{ji})^2\mathcal{M}_i + \frac{2}{n^2}\sum_{j\neq i}(\hat{F}_{-i,ji} - F_{ji})(F_{ji} - \mathbf{I}_{ji})\mathcal{M}_i$$

$$= S_{\Sigma,1n} - S_{\Sigma,2n}, \tag{A.20}$$

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

$$S^0_{\Sigma,1n} = \frac{1}{n^4}\sum_{j\neq i}\sum_{l\neq i}\sum_{l'\neq i}(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(\mathbf{I}_{jl'} - F_{ji})K_{\gamma,il'}\mathcal{M}_i/f_i^2. \tag{A.21}$$

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

$$S^0_{\Sigma,1n,(i)} = \frac{1}{n^4}\sum_{i\neq j\neq l\neq l'}(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(\mathbf{I}_{jl'} - F_{ji})K_{\gamma,il'}\mathcal{M}_i/f_i^2. \tag{A.22}$$

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

$$E[S^0_{\Sigma,1n,(i)}] = \int g(y)E[(\mathbf{I}_{ly} - F_{iy})K_{\gamma,il}(\mathbf{I}_{l'y} - F_{iy})K_{\gamma,il'}\mathcal{M}_i/f_i^2]dy = \int g(y)S^0_{1n,1}(y)dy, \tag{A.23}$$

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.9) 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.9) and is one of the leading terms of S^0_{1n} (and of $CV(\cdot)$); see the proof of Theorem 2.1.

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

$$E[S^0_{\Sigma,1n,(ii)}] = n^{-1}\int g(y)E[(\mathbf{I}_{ly} - F_{iy})^2 K^2_{\gamma,il}\mathcal{M}_i/f_i^2]dy = \int g(y)S^0_{1n,2}(y)dy, \tag{A.24}$$

where $S^0_{1n,2}(y) = n^{-1}E[(\mathbf{I}_{ly} - F_{iy})^2 K^2_{\gamma,il}\mathcal{M}_i/f_i^2]$. By (A.9) 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.9) and is the second leading term of S^0_{1n} (and of $CV(\cdot)$).

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

$$\begin{aligned} E[S^0_{\Sigma,1n,(iii)}] &= n^{-1}E[(\mathbf{I}_{jl} - F_{ji})K_{\gamma,il}(1 - F_{ji})K_{\gamma,ij}\mathcal{M}_i/f_i^2] + (s.o.) \\ &= n^{-1}E[(F_{lj} - F_{ji})K_{\gamma,il}(1 - F_{ji})K_{\gamma,ij}\mathcal{M}_i/f_i^2] + (s.o.) \\ &= n^{-1}O(|h|^2 + |\lambda|) = O(n^{-1}\zeta_{1n}). \end{aligned} \tag{A.25}$$

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 is given by

$$CV_{\Sigma,L} = \int g(y)[S^0_{1n,1}(y) + S^0_{1n,2}(y)]dy, \tag{A.26}$$

which equals CV_L provided that one replaces $M(y)$ by $g(y)$ in $CV_L(\cdot)$. Hence, Theorem 2.5 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.26) will be written as $CV_{\Sigma,L} = \sum_j [S^0_{1n,1}(y_j) + S^0_{1n,2}(y_j)]g(y_j)$. This completes the proof of Theorem 2.5. □

Appendix B. Proof of Theorem 3.1 and Theorem 3.2

In Appendix B, we use F_i to denote the true conditional CDF $F(y|\bar{x}_i)$. 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: Following the same derivations that lead to (A.1), one can show that $CV(\cdot) = CV_1(\cdot)+$ a term unrelated to (h, λ), where

$$\begin{aligned} CV_1(\gamma) = \int \Big[& \frac{1}{n(n-1)^2} \sum_{j \neq i} \sum_{l \neq i} \big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big) K_{\gamma,ji} K_{\gamma,li} / \hat{f}_{-i}^2 \\ & - \frac{2}{n(n-1)} \sum_{j \neq i} \big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_i - F_i\big) K_{\gamma,ji} / \hat{f}_{-i} \Big] \mathcal{M}_i M(y) dy \\ = & \int (A_{1n} - 2A_{2n}) M(y) dy, \end{aligned}$$

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

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

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

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

where $\bar{\mathcal{M}}(\bar{x})$ is defined in (8).

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

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

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

$F_s(y|\bar{x}) = \partial F(y|\bar{x}) / \partial \bar{x}_s^c$, $F_{ss}(y|\bar{x}) = \partial^2 F(y|\bar{x}) / \partial (\bar{x}_s^c)^2$, $\bar{f}_s(\bar{x}) = \partial \bar{f}(\bar{x}) / \partial \bar{x}_s^c$. Let $\int d\bar{x} = \sum_{\bar{x}^d} \int d\bar{x}^c$, $\int dx = \sum_{x^d} \int dx^c$. $\tilde{R}(\tilde{x}) = \tilde{R}(\tilde{x}, h_{q_1+1}, \dots, h_q, \lambda_{r_1+1}, \dots, \lambda_r)$ 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 $CV_1(\gamma)$ is

$$\iint \Big(\sum_{s=1}^{q_1} h_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1} \lambda_s \bar{B}_{2s}(y|\bar{x}) \Big)^2 \bar{f}(\bar{x}) \bar{\mathcal{M}}(\bar{x}) M(y) d\bar{x} dy$$

$$+ \iint \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \ldots h_{q_1}} \tilde{R}(\tilde{x}) \bar{f}(\bar{x}) \tilde{f}(\tilde{x}) \mathcal{M}(x) M(y) dx dy. \tag{B.7}$$

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 λ_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 (3) ($w(0) > w(\delta)$ for all $\delta > 0$) we know that $\text{Var}(\mathcal{V}_n) = E[\mathcal{V}_n^2] - [E(\mathcal{V}_n)]^2 > 0$ 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 $\text{Var}(\mathcal{V}_n) = 0$ 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

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

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

$$\bar{\mathcal{X}}(a_1, \ldots, b_{r_1}) = (a_1 \ldots a_{q_1})^{-1} \iint \bar{\Sigma}_{y|\bar{x}} \bar{f}(\bar{x}) \tilde{f}(\tilde{x}) \mathcal{M}(x) M(y) dx dy$$

$$+ \iint \Big(\sum_{s=1}^{q_1} a_s^2 \bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1} b_s \bar{B}_{2s}(y|\bar{x}) \Big)^2 \bar{f}(\bar{x}) \bar{\mathcal{M}}(\bar{x}) M(y) d\bar{x} dy. \tag{B.9}$$

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 result of Theorem 3.1 immediately follows. □

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 $A_{1n} = A_{1n}^0 + (s.o.)$. Hence, we only need to consider A_{1n}^0 which is defined by

$$A_{1n}^0 = \frac{1}{n(n-1)^2}\sum_{l\neq j\neq i}\big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big)K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i + \frac{1}{n(n-1)^2}\sum_{j\neq i}\big(\mathbf{I}_j - F_i\big)^2 K_{\gamma,ji}^2\mu(x_i)^{-2}\mathcal{M}_i$$

$$= G_{1n} + G_{2n},$$

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 $G_{1n} = E(G_{1n}) + (s.o.)$.

$$E(G_{1n}) = E\big[(\mathbf{I}_j - F_i)(\mathbf{I}_l - F_i)K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i\big] = E\Big\{\big(E[(\mathbf{I}_j - F_i)K_{\gamma,ji}/\mu(x_i)|x_i]\big)^2\mathcal{M}_i\Big\}. \qquad \text{(B.11)}$$

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 $E[\tilde{K}_{\tilde{\gamma},ij}/\nu_1(\tilde{x}_i)|\tilde{x}_i] = 1$)

$$\begin{aligned}
E[(\mathbf{I}_j - F_i)K_{\gamma,ji}\mu(x_i)^{-1}|x_i] &= E[(F_j - F_i)K_{\gamma,ji}\mu(x_i)^{-1}|x_i] \\
&= E[(F_j - F_i)\bar{K}_{\bar{\gamma},ij}\bar{f}(\bar{x}_i)^{-1}|\bar{x}_i]E[\tilde{K}_{\tilde{\gamma},ij}/\nu_1(\tilde{x}_i)|\tilde{x}_i] \\
&= \bar{f}(\bar{x}_i)^{-1}\sum_{\bar{z}^d\in\bar{S}^d} L(\bar{z}^d, \bar{x}_i^d, \lambda)\int[F(y|\bar{x}_i^c + hv, \bar{z}^d) - F(y|\bar{x}_i^c, \bar{x}_i^d)]\bar{f}(\bar{x}_i^c + hv, \bar{z}^d)W(v)dv \\
&= \frac{\kappa_2}{2}\sum_{s=1}^{q_1} h_s^2\big[\bar{f}(\bar{x}_i)F_{ss}(y|\bar{x}_i) + 2\bar{f}_s(\bar{x}_i)F_s(y|\bar{x}_i)\big]/\bar{f}(\bar{x}_i) \\
&\quad + \sum_{s=1}^{r_1}\lambda_s\sum_{\bar{z}^d\in\bar{S}^d} I_s(\bar{z}^d, \bar{x}_i^d)\big[F(y|\bar{x}_i^c, \bar{z}^d) - F(y|\bar{x}_i^c, \bar{x}_i^d)\big]\bar{f}(\bar{x}_i^c, \bar{z}^d)/\bar{f}(\bar{x}_i) + o(\zeta_n) \\
&= \sum_{s=1}^{q_1} h_s^2\bar{B}_{1s}(y|\bar{x}_i) + \sum_{s=1}^{r_1}\lambda_s\bar{B}_{2s}(y|\bar{x}_i) + o(\bar{\zeta}_n), \qquad \text{(B.12)}
\end{aligned}$$

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.12) into (B.11), we immediately obtain (recall $\bar{\mathcal{M}}(\bar{x})$ is defined in (8))

$$E(G_{1n}) = \int\Big(\sum_{s=1}^{q_1} h_s^2\bar{B}_{1s}(y|\bar{x}) + \sum_{s=1}^{r_1}\lambda_s\bar{B}_{2s}(y|\bar{x})\Big)^2\bar{f}(\bar{x})\bar{\mathcal{M}}(\bar{x})d\bar{x} + o(\bar{\zeta}_n). \qquad \text{(B.13)}$$

Note that in the above we have only shown that for all fixed values of $(h, \lambda) \in \Gamma$, (B.13) holds true. By utilizing Rosenthal's and Markov's inequalities, it's straightforward to show (B.13) 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

$$\begin{aligned}
E(G_{2n}) &= n^{-1}E\Big[\big(\mathbf{I}_j - F_i\big)^2 K^2_{\gamma,ji}\mu(x_i)^{-2}\mathcal{M}_i\Big] \\
&= n^{-1}E\Big\{E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)K^2_{\gamma,ji}\mu(x_i)^{-2}|x_i\big]\mathcal{M}_i\Big\} \\
&= n^{-1}E\Big\{E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)\bar{K}^2_{\bar{\gamma},i}\bar{f}(\bar{x}_i)^{-2}|\bar{x}_i\big]\mathcal{M}_i\nu_2(\tilde{x}_i)\nu_1(\tilde{x}_i)^{-2}\Big\}. \qquad \text{(B.14)}
\end{aligned}$$

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 $s = 1, \dots, q_1$ and $\lambda_s \to 0$ for $s = 1, \dots, r_1$. Thus

$$
\begin{aligned}
&E\big[(\mathbf{I}_j - 2F_i\mathbf{I}_j + F_i^2)\bar{K}^2_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-2}|x_i\big] = E\big[(F_j - 2F_iF_j + F_i^2)\bar{K}^2_{\bar{\gamma},ji}\bar{f}(\bar{x}_i)^{-2}|x_i\big] \\
&= \frac{1}{h_1 \dots h_{q_1}} \sum_{\bar{z}^d \in \bar{S}^d} L(\bar{z}^d, \bar{x}_i^d, \lambda)^2 \int \big[F(y|\bar{x}_i^c + hv, \bar{z}^d) - 2F(y|\bar{x}_i^c, \bar{x}_i^d)F(y|\bar{x}_i^c + hv, \bar{z}^d) \\
&\quad + F(y|\bar{x}_i^c, \bar{x}_i^d)^2\big]\bar{f}(\bar{x}_i)^{-2}\bar{f}(\bar{x}_i^c + hv, \bar{z}^d)W(v)^2 dv \\
&= \frac{\bar{\Sigma}_{y|\bar{x}_i}}{h_1 \dots h_{q_1}} + O(\bar{\zeta}_{1n}^{-1/2}(h_1 \dots h_{q_1})^{-1}), \qquad \text{(B.15)}
\end{aligned}
$$

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

Substituting (B.15) into (B.14), we immediately obtain $E(G_{2n}) = \int \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \dots h_{q_1}}\tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)dx +$ $(s.o.)$, where $\tilde{R}(\tilde{x})$ is defined in (B.6). Hence,

$$
G_{2n} = E(G_{2n}) + (s.o.) = \int \frac{\bar{\Sigma}_{y|\bar{x}}}{nh_1 \dots h_{q_1}}\tilde{R}(\tilde{x})\bar{f}(\bar{x})\tilde{f}(\tilde{x})\mathcal{M}(x)dx + (s.o.).
$$

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_{2n}^0 denote 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 $A_{2n} = A_{2n}^0 + (s.o.)$. Hence, we only need to consider A_{2n}^0 which is defined by $A_{2n}^0 = (n(n-1))^{-1}\sum_{j \neq i}\big(\mathbf{I}_j - F_i\big)(\mathbf{I}_i - F_i)K_{\gamma,ji}\mu(x_i)^{-1}$. Notice that the part in A_{2n}^0 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_{2n}^0 we can ignore the irrelevant covariates part and need only consider

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

Note that $\bar{A}_{2n}^0$ only depends on $(h_1 \dots, h_{q_1}, \lambda_1, \dots, \lambda_{r_1})$. 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 $\bar{A}_{2n}^0$, so we omit $\mathcal{M}_i$ in the following proof of this Lemma.

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

Hence

$$\bar{A}_{2n}^0 = O_p(n^{-1/2}\bar{\zeta}_{1n} + (n(h_1\dots h_{q_1})^{1/2})^{-1}). \tag{B.16}$$

Moreover, by utilizing Rosenthal's and Markov's inequalities, one can show that (B.16) 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* $\hat{f}_{-i}(x)^{-1} = \mu(x)^{-1} + O_p\left(\bar{\zeta}_{1n} + \big(\ln(n)\big)^{1/2}(nh_1\dots h_{q_1})^{-1/2}\right)$ *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

$$\hat{\mu}(x) = E[\bar{K}_{\bar{\gamma},ij}|\bar{x}_i = \bar{x}]E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}] = \{\bar{f}(\bar{x}) + O(\bar{\zeta}_{1n})\}E[\tilde{K}_{\tilde{\gamma},ij}|\tilde{x}_i = \tilde{x}] = \mu(x) + O_p(\bar{\zeta}_{1n}). \tag{B.17}$$

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

$$\hat{f}_{-i}(x) - \hat{\mu}(x) = O_p\left(\big(\ln(n)\big)^{1/2}(nh_1\dots h_{q_1})^{-1/2}\right), \tag{B.18}$$

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

Combining $(B.17)$ and $(B.18)$ we obtain

$$\hat{f}_{-i}(x) - \mu(x) = O_p\left(\bar{\zeta}_{1n} + \big(\ln(n)\big)^{1/2}(nh_1\dots h_{q_1})^{-1/2}\right), \tag{B.19}$$

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

Using (B.19) and Taylor expansions, we obtain

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

Lemma B.4. $\hat h_s = o_p(1)$ *for* $s = 1, \dots, q_1$ *and* $\hat\lambda_s = o_p(1)$ *for* $s = 1, \dots, r_1$.

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 $G_{1n} = \frac{1}{n(n-1)^2}\sum_{l\neq j\neq i}\big(\mathbf{I}_j - F_i\big)\big(\mathbf{I}_l - F_i\big)$ $K_{\gamma,ji}K_{\gamma,li}\mu(x_i)^{-2}\mathcal{M}_i + o_p(1) \equiv G_{1,0} + o_p(1)$, 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

$$\begin{aligned}E(G_{1,0}) &= E\Big\{\big[\mu(x_i)^{-1}E\big((\mathbf{I}_j - F_i)K_{\gamma,ji}|x_i\big)\big]^2\mathcal{M}(x_i)\Big\}\\ &= E\left\{\Big[\bar f(\bar x_i)^{-1}E\big((F_j - F_i)\bar K_{\bar\gamma,ji}|\bar x_i\big)\Big]^2\mathcal{M}(x_i)\right\}\\ &= E\big\{[\eta(y,\bar x_i)]^2\mathcal{M}(x_i)\big\} = \int[\eta(y,\bar x)]^2\bar f(\bar x)\bar{\mathcal{M}}(\bar x)d\bar x, \qquad \text{(B.20)}\end{aligned}$$

where $\eta(y,\bar x)$ is defined in (7), $\bar{\mathcal{M}}(\bar x)$ is defined in (8). Note that the right hand side of (B.20) does not depend on $(h_{q_1+1}, \dots, h_q, \lambda_{r_1+1}, \dots, \lambda_r)$ 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 $(h_1, \dots, h_{q_1}, \lambda_1, \dots, \lambda_{r_1})$ that minimize $CV(\gamma)$ do not all converge in probability to zero, then by (9), $E(G_{1,0})$ (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 $h_1, \dots, h_{q_1}$ to be of size $n^{-1/(q_1+4)}$, and $\lambda_1, \dots, \lambda_{r_1}$ to be of size $n^{-2/(4+q)}$, letting $h_{q_1+1}, \dots, h_q$ diverge to infinity, and letting $\lambda_{r_1+1}, \dots, \lambda_r$ 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 $(h_1, \dots, h_{q_1}, \lambda_1, \dots, \lambda_{r_1})$, for the relevant components of x, all converge in probability to zero. □

Proof of Theorem 3.2. By Theorem 3.1 we know that $\hat{h}_s \xrightarrow{p} +\infty$ for $s = q_1 + 1, \dots, q$ and $\hat{\lambda}_s \xrightarrow{p} 1$ for $s = r_1 + 1, \dots, r$. Therefore, we need only consider the case with all irrelevant covariates removed, i.e. we consider $\hat{F}(y|\bar{x}) = [\sum_j \bar{K}_{\hat{\gamma},jx}]^{-1}[\sum_j \mathbf{I}_j \bar{K}_{\hat{\gamma},jx}]$, where $\bar{K}_{\hat{\gamma},jx} = \big[\prod_{s=1}^{q_1} \hat{h}_s^{-1} w((x_{js}^c - x_s^c)/\hat{h}_s)\big]\big[\prod_{s=1}^{r_1} l(x_{js}^d, x_s^d, \hat{\lambda}_s)\big]$.

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 $s = 1, \dots, q_1$, and $\lambda_s^0 = b_s^0 n^{-2/(4+q_1)}$ for $s = 1, \dots, r_1$, where a_s^0 and b_s^0 are defined in (B.10). Also, define $\bar{F}(y|\bar{x}) = \big[\sum_j \bar{K}_{\gamma^0,jx}\big]^{-1}\big[\sum_j \mathbf{I}_j \bar{K}_{\gamma^0,jx}\big]$, where $\bar{K}_{\gamma^0,jx} = \big[\prod_{s=1}^{q_1} (h_s^0)^{-1} w((x_{js}^c - x_s^c)/h_s^0)\big]\big[\prod_{s=1}^{r_1} l(x_{js}^d, x_s^d, \lambda_s^0)\big]$. Then,

$$\bar{F}(y|\bar{x}) - F(y|\bar{x}) = \big[\sum_j \bar{K}_{\gamma^0,j}\big]^{-1}\big[\sum_j \mathbf{I}_j \bar{K}_{\gamma^0,jx} - \sum_j \bar{K}_{\gamma^0,jx} F(y|\bar{x})\big], \tag{B.21}$$

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

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

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

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 $(nh_1^0 \dots h_{q_1}^0)^{-1}\bar{\Sigma}_{y|\bar{x}}\bar{f}(\bar{x})^2$, where $\bar{\Sigma}_{y|\bar{x}}$ is defined in (B.5). By applying a triangular-array CLT, we have that

$$\sqrt{nh_1^0 \dots h_{q_1}^0}\left[\bar{F}(y|\bar{x}) - F(y|\bar{x}) - \sum_{s=1}^{q_1}(h_s^0)^2 \bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\lambda_s^0 \bar{B}_{2s}(y|\bar{x})\right] \xrightarrow{d} N(0, \bar{\Sigma}_{y|\bar{x}}). \tag{B.22}$$

Next we consider $\hat{F}(y|x) = \big[\sum_{j\neq i}\bar{K}_{\hat{\gamma},ji}\big]^{-1}\big[\sum_{j\neq i}\mathbf{I}_j \bar{K}_{\hat{\gamma},ji}\big]$ 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 $\hat{F}(y|x)$ 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 $\hat{D}(y|x) = \hat{F}(y|\bar{x}) -$

$F(y|\bar{x}) - \sum_{s=1}^{q_1}(\hat{h}_s)^2\bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\hat{\lambda}_s\bar{B}_{2s}(y|\bar{x})$ 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, $\hat{F}(y|x)$ and $\bar{F}(y|\bar{x})$ have the same asymptotic distribution, i.e.,

$$\sqrt{n\hat{h}_1\dots\hat{h}_{q_1}}\left[\hat{F}(y|x) - F(y|\bar{x}) - \sum_{s=1}^{q_1}\hat{h}_s^2\bar{B}_{1s}(y|\bar{x}) - \sum_{s=1}^{r_1}\hat{\lambda}_s\bar{B}_{2s}(y|\bar{x})\right] \xrightarrow{d} N(0, \bar{\Sigma}_{y|\bar{x}}). \tag{B.23}$$

□

References

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.

Chernozhukov, V., Fernndez-Val, I. & Galichon, A. (2010), 'Quantile and probability curves without crossing', *Econometrica* **78**(3), 1093–1125.

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

Efromovich, S. (2010), 'Dimension reduction and adaptation in conditional density estimation', *Journal of the American Statistical Association* **105**, 761–774.

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.

Hayfield, T. & Racine, J. S. (2008), 'Nonparametric econometrics: The np package', *Journal of Statistical Software* **27**(5).
URL: *http://www.jstatsoft.org/v27/i05/*

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

Kiefer, N. M. & Racine, J. S. (2009), 'The smooth Colonel meets the Reverend', *Journal of Nonparametric Statistics* **21**, 521–533.

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

Koenker, R. & Bassett, G. (1978), 'Regression quantiles', *Econometrica* **46**, 33–50.

Koenker, R. & Xiao, Z. (2004), 'Unit root quantile autoregression inference', *Journal of the American Statistical Association* **99**, 775–787.

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, Y., Liu, Y. & Zhu, J. (2007), 'Quantile regression in reproducing kernel Hilbert spaces', **102**(477), 255–268.

Maasoumi, E. & Lugo, M. A. (2008), The information basis of multivariate poverty assessments, *in* 'Quantitative Approaches to Multidimensional Poverty Measurement', Palgrave-MacMillan.

Marron, J. S., Jones, M. C. & Sheather, S. J. (1996), 'A brief survey of bandwidth selection for density estimation', *Journal of the American Statistical Association* **91**, 401–407.

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

Peng, L. & Huang, Y. (2008), 'Survival analysis with quantile regression models', *Journal of the American Statistical Association* **103**(482), 637–649.

R Core Team (2012), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org/*

Racine, J. & Li, Q. (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data', *Journal of Econometrics* **119**(1), 99–130.

Yu, K. & Jones, M. C. (1998), 'Local linear quantile regression', *Journal of the American Statistical Association* **93**(441), 228–237.