NONPARAMETRIC ESTIMATION OF CONDITIONAL CDF AND QUANTILE FUNCTIONS WITH MIXED CATEGORICAL AND CONTINUOUS DATA

QI LI AND JEFF RACINE

ABSTRACT. We propose a new nonparametric conditional CDF kernel estimator that admits a mix of discrete and categorical data along with an associated nonparametric conditional quantile estimator. Bandwidth selection for kernel quantile regression remains an open topic of research. We employ a conditional PDF based bandwidth selector proposed by Hall, Racine & Li (2004) that can automatically remove irrelevant variables and has impressive performance in this setting. We provide theoretical underpinnings including rates of convergence and limiting distributions. Simulations demonstrate that this approach performs quite well relative to its peers, while two illustrative examples serve to underscore its value in applied settings.

1. INTRODUCTION

Perhaps the most common econometric application of nonparametric techniques has been the estimation of a regression function (i.e., a conditional mean). However, often it is of interest to model conditional quantiles (e.g., a median or quartile), particularly when it is felt that the conditional mean may not be representative of the impact of the covariates on the dependent variable. For example, modeling conditional quantiles rather than conditional means is often desirable when the dependent variable is left (or right) censored, which distorts the relationship given by the conditional mean estimator, whereas quantiles above (or below) the censoring point are robust to the presence of censoring. Furthermore, the quantile regression function in general provides a much more comprehensive picture of the conditional distribution of a dependent variable than the conditional mean function.

The literature on quantile regression has blossomed in recent years, originating with the seminal work of Koenker & Bassett (1978), who introduced this approach in a parametric framework. This literature continues to evolve in a number of exciting directions as exemplified by the work of Powell (1986) and Buchinsky & Hahn (1998), who considered regression quantiles for censored data, Chaudhuri, Doksum & Samarov (1997), who considered nonparametric average derivative estimation, Horowitz (1998), who considered resampling methods for such models, Yu & Jones (1998), who considered local polynomial estimation of regression quantiles, Koenker & Xiao (2002), who proposed new tests of location and

Date: December 17, 2005.

Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine's research is partially supported by the NSF, SSHRC, and NSERC.

location-scale shifts, Koenker & Xiao (2004), who studied inference on unit root quantile regression models, Cai (2002), and Cai & Xu (2004), who considered quantile regressions with time series data, to name but a few. These approaches share one defining feature - least absolute deviation estimation via the use of the so-called 'check function' defined in Section 3.1 below.

A natural way to model a conditional quantile function is to invert a conditional cumulative distribution function (CDF) at the desired quantile. Of course, the conditional CDF is unknown and must be estimated, and methods that are robust to functional misspecification have obvious appeal. The nonparametric estimation of conditional CDFs has received much attention as of late; see by way of example Hall, Wolff & Yao (1999), Cai (2002), and Hansen (2004). Each of these authors considers the case of continuous conditioning variables only. In applied settings, however, one frequently encounters a mix of discrete and continuous data. For example, nonparametric estimates of conditional CDFs are used to recover the distribution of unobserved private values in auctions, and auction data frequently includes mixed discrete and continuous conditional covariates (see Guerre, Perrigne & Vuong (2000)[1], Li, Perrigne & Vuong (2000), among others); nonparametric estimates of conditional CDFs are also used to obtain consistent estimates of nonadditive (non-separable) random functions that may involved discrete data (Matzkin (2003)). Furthermore, quantile estimation methods are widely used in labor settings where most of the covariates are discrete in nature (see, e.g., Buchinsky (1994)).

Rather than using the conventional frequency approach in dealing with the discrete variables, in this paper we smooth both the discrete and the continuous variables. It is well known that the selection of smoothing parameters is of crucial importance for sound nonparametric estimation. Unfortunately, there do not exist automatic data-driven methods for optimal selection of smoothing parameters when estimating a conditional CDF or a conditional quantile function. Although there exist various plug-in methods in the pure continuous covariate case, no general plug-in formula is available for the mixed variable case. The plug-in method, even after adaptation to mixed data, requires choices of 'pilot' smoothing parameters, and it is not clear how to best make such a selection for discrete variables. While the conventional frequency-based method may seem a natural choice, in applications involving economic data one often encounters a large number of discrete cells, and the frequency method may simply not be feasible in such cases. Moreover, the plug-in method does not seem to be able to handle the irrelevant covariate case. Therefore, in this paper we suggest adopting a conditional probability

[1]Although Guerre et al. (2000) allow for smoothing the discrete variables (e.g., Bierens (1987)), they do not address the issue of how to use data-driven methods for selecting the smoothing parameters associated with the discrete variables.

density function (PDF) method of bandwidth selection proposed by Hall et al. (2004) in the context of estimating a conditional CDF or a conditional quantile function. Smoothing categorical covariates may introduce some estimation bias, it can reduce the variance significantly, resulting in a smaller estimation mean squared error. This approach has the advantage that it is an automatic data-driven procedure, no 'pilot' estimators are required, and it has the ability to remove irrelevant covariates. Thus, the new nonparametric estimator of a conditional CDF (or a conditional quantile estimator) proposed in this paper has a number of features including (i) it admits both continuous and categorical data, (ii) the bandwidth selection method is capable of removing irrelevant variables, (iii) we smooth the dependent variable, unlike a number of 'non-smoothing' approaches towards conditional CDF estimation that have been proposed, and (iv) we adopt a conditional PDF method of bandwidth selection proposed by Hall et al. (2004) that overcomes one weakness common to many of the existing methods - the lack of an *automatic* appropriate data-driven bandwidth selector.

The remaining part of the paper is organized as follows. In Section 2 we propose two smoothing based conditional CDF estimators and establish their asymptotic distributions, and we also discuss how to select the smoothing parameters in applied settings. In Section 3 we suggest obtaining conditional quantile estimators by inverting a mixed data conditional CDF, and we study the asymptotic behavior of the proposed conditional quantile estimators. Section 4 reports simulation results along with two empirical applications.

2. Estimating A Conditional Cumulative Distribution Function with Mixed Discrete and Continuous Covariates

2.1. **Smoothing the Covariates Only.** We first consider an estimator of a conditional CDF that smooths only the explanatory variables (i.e., it does not smooth the dependent variable). We consider the case for which x is a vector containing a mix of discrete and continuous variables. Let $x = (x^c, x^d)$, where $x^c \in R^q$ 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 (i.e., both 'ordinal' and 'nominal' discrete datatypes). Let X^d_{is} (x^d_s) denote the sth component of X^d_i (x^d), $s = 1, \dots, r$; $i = 1, \dots, n$, where n is the sample size. For an ordered variable, we use the following kernel:

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

Note that λ_s is a bandwidth having the following properties; when $\lambda_s = 0$ ($\lambda_s \in [0,1]$), $l(X_{is}^d, X_{js}^d, 0)$ becomes an indicator function, and when $\lambda_s = 1$, $l(X_{is}^d, X_{js}^d, 1) = 1$ becomes a uniform weight function.

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

$$l(X_{is}^d, X_{js}^d, \lambda_s) = \begin{cases} 1, & \text{if } X_{is}^d = X_{js}^d, \\ \lambda_s, & \text{otherwise.} \end{cases} \tag{2}$$

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

Without loss of generality we assume that the first r_1 components of x^d are ordered variables and the remaining components are unordered. Let $I(A)$ denote an indicator function that assumes the value 1 if A occurs and 0 otherwise. Combining (1) and (2), we obtain the product kernel function given by

$$L_\lambda(X_i^d, X_j^d) = \left[\prod_{s=1}^{r_1} \lambda_s^{|X_{is}^d - X_{js}^d|}\right]\left[\prod_{s=r_1+1}^{r} \lambda_s^{I(X_{is}^d \neq X_{js}^d)}\right]. \tag{3}$$

We shall use $w(\cdot)$ to denote a univariate kernel function for a continuous variable. The product kernel function used for the continuous variables is given by $W_h(X_i^c, X_j^c) = \prod_{s=1}^q h_s^{-1} w\left(\frac{X_{is}^c - X_{js}^c}{h_s}\right)$, where X_{is}^c (x_s^c) is the sth component of X_i^c (x^c), and h_s is the bandwidth associated with x_s^c.

We use $F(y|x)$ to denote the conditional CDF of Y given $X = x$, and let $\mu(x)$ denote the marginal density of X. We shall estimate $F(y|x)$ by

$$\tilde{F}(y|x) = \frac{n^{-1}\sum_{i=1}^n I(Y_i \leq y) K_\gamma(X_i, x)}{\hat{\mu}(x)}, \tag{4}$$

where $\hat{\mu}(x) = n^{-1}\sum_{i=1}^n K_\gamma(X_i, x)$ is the kernel estimator of $\mu(x)$, $K_\gamma(X_i, x) = W_h(X_i^c, x^c) L_\lambda(X_i^d, x^d)$ is the product kernel. Note that (4) is similar to the conditional mean function estimator considered in Racine & Li (2004) where one replaces $I(Y_i \leq y)$ in (4) by Y_i.

In order to describe the leading bias term associated with the discrete variables, we need to introduce some notation. When x_s^d is an unordered categorical variable, define an indicator function $I_s(\cdot,\cdot)$ by

$$I_s(x^d, z^d) = I\left(x_s^d \neq z_s^d\right) \prod_{t \neq s}^{r} I\left(x_t^d = z_t^d\right). \tag{5}$$

$I_s(x^d, z^d)$ equals one if and only if x^d and z^d differ only in the 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 $I_s(\cdot,\cdot)$ is defined by

$$I_s(x^d, z^d) = I\left(|x_s^d - z_s^d| = 1\right)\prod_{t\neq s}^{r} I\left(x_t^d = z_t^d\right). \tag{6}$$

Note that $I_s(x^d, z^d)$ equals one if and only if x^d and z^d differ by one unit only in the sth component, and is zero otherwise. We make the following assumptions.

Condition (C1): The data $\{Y_i, X_i\}_{i=1}^n$ is independent and identically distributed as (Y, X). Both $\mu(x)$ and $F(y|x)$ have continuous second order partial derivatives with respect to x^c. For fixed values of y and x, $\mu(x) > 0$, $0 < F(y|x) < 1$.

Condition (C2): $w(\cdot)$ is a symmetric, bounded, and compactly supported density function.

Condition (C3): As $n \to \infty$, $h_s \to 0$ for $s = 1, \ldots, q$, $\lambda_s \to 0$ for $s = 1, \ldots, r$, and $(nh_1 \ldots h_q) \to \infty$.

To avoid having to deal with a random denominator in $\tilde{F}(y|x)$, define $\tilde{M}(y|x) = [\tilde{F}(y|x) - F(y|x)]\hat{\mu}(x)$. Then $\tilde{F}(y|x) - F(y|x) = \tilde{M}(y|x)/\hat{\mu}(x)$. For $s = 1, \ldots, q$, let $F_s(y|x) = \frac{\partial F(y|x)}{\partial x_s}$, $\mu_s(x) = \frac{\partial \mu(x)}{\partial x_s}$, and $F_{ss}(y|x) = \frac{\partial^2 F(y|x)}{\partial x_s^2}$. Also define $|h| = \sum_{s=1}^q h_s$ and $|\lambda| = \sum_{s=1}^r \lambda_s$. The next theorem provides the asymptotic distribution of $\tilde{F}(y|x)$.

THEOREM 2.1. *Under conditions (C1) to (C3), we have*

(i) $MSE[\tilde{M}(y|x)] = \mu(x)^2 \left\{\sum_{s=1}^q h_s^2 B_{1s}(y|x) + \sum_{s=1}^r \lambda_s B_{2s}(y|x)\right\}^2 + \frac{\mu(x)^2 V(y|x)}{(nh_1 \ldots h_q)} + o(\eta_{1n})$, *where* $\eta_{1n} = |h|^4 + |\lambda|^2 + (nh_1 \ldots h_q)^{-1}$.

(ii) *If* $(nh_1 \ldots h_q)^{1/2}(|h|^2 + |\lambda|) = O(1)$, *then*

$$(nh_1 \ldots h_q)^{1/2}\left[\tilde{F}(y|x) - F(y|x) - \sum_{s=1}^q h_s^2 B_{1s}(y|x) - \sum_{s=1}^r \lambda_s B_{2s}(y|x)\right] \to N(0, V(y|x)) \text{ in distribution,}$$

where $V(y|x) = \kappa^q F(y|x)[1 - F(y|x)]/\mu(x)$, $B_{1s}(y|x) = (1/2)\kappa_2[2F_s(y|x)\mu_s(x) + \mu(x)F_{ss}(y|x)]/\mu(x)$, $B_{2s}(y|x) = \mu(x)^{-1}\sum_{z^d \in D} I_s(z^d, x^d)[F(y|x^c, z^d)\mu(x^c, z^d) - F(y|x)\mu(x)]/\mu(x)$, $\kappa = \int w(v)^2 dv$, $\kappa_2 = \int w(v)v^2 dv$, *and* D *is the support of* X^d.

The proof of Theorem 2.1 is given in Appendix A. Note that Theorem 2.1 holds true regardless of whether Y is a continuous or a discrete variable.

2.2. **Smoothing the Dependent Variable and Covariates.** When the dependent variable Y is a continuous random variable, one can use an alternative estimator that also smooths the dependent variable y. That is, one can also estimate $F(y|x)$ by

$$\hat{F}(y|x) = \frac{n^{-1}\sum_{i=1}^{n} G\left(\frac{y-Y_i}{h_0}\right) K_\gamma(X_i,x)}{\hat{\mu}(x)}, \tag{7}$$

where $G(v) = \int_{-\infty}^{v} w(u)du$ is the distribution function derived from the density function $w(\cdot)$, and h_0 is the bandwidth associated with Y_i.

We will show below that the optimal h_0 should have an order smaller than that of h_s for $s = 1, \dots, q$. Therefore, we will make the following assumption.

Condition (C4): $F(y|x)$ is twice continuously differentiable in (y, x^c), and $h_0 = o((nh_1 \dots h_q)^{-1/4})$.

Defining $\hat{M}(y|x) = [\hat{F}(y|x) - F(y|x)]\hat{\mu}(x)$, we then have $\hat{F}(y|x) - F(y|x) = \hat{M}(y|x)/\hat{\mu}(x)$. Letting $F_0(y|x) = \frac{\partial F(y|x)}{\partial y}$, $F_{00}(y|x) = \frac{\partial^2 F(y|x)}{\partial y^2}$, we obtain the following results.

THEOREM 2.2. *Under conditions (C1) to (C4) we have*

(i) $E[\hat{M}(y|x)] = \mu(x)\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \mu(x)\sum_{s=0}^{r}\lambda_s B_{2s}(y|x) + o(h_0^2) + o(|h|^2 + |\lambda|)$, *where* $B_{10}(y|x) = F_{00}(y|x)$, *while* κ *and the remaining* B_{1s}*'s and* B_{2s}*'s are the same as that defined in Theorem 2.1; the* $o(|h|^2 + |\lambda|)$ *term does not depend on* h_0.

(ii) $Var[\hat{M}(y|x)] = (nh_1 \dots h_q)^{-1}\mu(x)^2[V(y|x) - h_0 C_w \Omega(y|x)] + o(\eta_{2n}) + o(\eta_{3n})$, *where* $C_w = 2\int G(v)w(v)vdv$, $V(y|x)$ *is the same as that defined in Theorem 2.1;* $\eta_{2n} = (nh_1 \dots h_q)^{-1}$, *which does not depend on* h_0, *and* $\eta_{3n} = (nh_1 \dots h_q)^{-1}h_0$.

(iii) *If* $(nh_1 \dots h_q)^{1/2}(|h|^2 + |\lambda|) = O(1)$, *then*

$$(nh_1 \dots h_q)^{1/2}\left[\hat{F}(y|x) - F(y|x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) - \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\right] \to N(0, V(y|x)) \text{ in distribution.}$$

The proof of Theorem 2.2 is given in Appendix A.

2.3. **Selection of Smoothing Parameters.** In this subsection we will mainly focus on how to choose the smoothing parameters for $\hat{F}(y|x)$. Theorem 2.2 implies that the MSE of $\hat{F}(y|x)$ is

$$MSE[\hat{F}(y|x)] = \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)\right]^2 + \frac{V(y|x) - h_0\Omega(y|x)}{nh_1 \dots h_q} + o(\eta_{4n}) + o(\eta_{5n}), \tag{8}$$

where $\eta_{4n} = |h|^4 + |\lambda|^2 + (nh_1 \dots h_q)^{-1}$, which does not depend on h_0, and $\eta_{5n} = h_0^4 + (nh_1 \dots h_q)^{-1}h_0$.

One may choose the bandwidths to minimize the *leading* term of a weighted integrated $MSE[\hat{F}(y|x)]$ given by:

$$IMSE_{\hat{F}} = \int \left\{ \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) \right]^2 + \frac{[V(y|x) - h_0 \Omega(y|x)]}{nh_1 \dots h_q} \right\} s(y,x) dy dx, \quad (9)$$

where $\int dxdy = \sum_{x^d \in D} \int dx^c dy$ and where $s(\cdot,\cdot)$ is a non-negative weight function. We now analyze the order of optimal smoothing. First, consider the simplest case for which $q = 1$ and $r = 0$. From (9) we know that in this case $IMSE_{\hat{F}}$ is given by

$$\begin{aligned} IMSE_{\hat{F}} &= \int MSE[\hat{F}(y|x)] s(y,x) dy dx \\ &= A_0 h_0^4 + A_1 h_0^2 h_1^2 + A_2 h_1^4 + A_3 (nh_1)^{-1} - A_4 h_0 (nh_1)^{-1}, \end{aligned} \quad (10)$$

where the A_j's are all constants given by $A_0 = \int B_0(y|x)^2 s(y,x)dydx$, $A_1 = 2\int B_0(y|x)B_1(y|x)s(y,x)dydx$, $A_2 = \int B_1(y|x)^2 s(y,x)dydx$, $A_3 = V(y|x)s(y,x)dydx$, and $A_4 = \Omega(y|x)s(y,x)dydx$. All the A_j's, except A_1, are positive, while A_1 can be positive, negative, or zero.

The first order conditions required for minimization of (10) are:

$$\frac{\partial IMSE_{\hat{F}}}{\partial h_0} = 4A_0 h_0^3 + 2A_1 h_0 h_1^2 - A_4 (nh_1)^{-1} \overset{set}{=} 0 \quad (11)$$

$$\frac{\partial IMSE_{\hat{F}}}{\partial h_1} = 2A_1 h_0^2 h_1 + 4A_2 h_1^3 - A_3 (nh_1^2)^{-1} \overset{set}{=} 0. \quad (12)$$

From (11) and (12), one can easily see that h_0 and h_1 will not have the same order. Further it can be shown that h_0 must have an order smaller than that of h_1. Let us assume that $h_0 = c_0 n^{-\alpha}$ and $h_1 = c_1 n^{-\beta}$. Then we must have[2] $\alpha > \beta$. From (12) we get $\beta = 1/5$, and substituting this into (11) yields $\alpha = 2/5$. Thus, optimal smoothing requires that $h_0 \sim n^{-2/5}$ and $h_1 \sim n^{-1/5}$.

Similarly, for the general case for which $q > 1$, by symmetry, all h_s should have the same order, and all λ_s should have the same order (with $\lambda_s \sim h_s^2$), say $h_s \sim n^{-\beta}$ for $s = 1, \dots, q$, $\lambda_s \sim n^{-2\beta}$ for $s = 1, \dots, r$, and $h_0 \sim n^{-\alpha}$. Then it is easy to show that $\beta = 1/(4+q)$ and $\alpha = 2/(4+q)$. Thus, the optimal $h_s \sim n^{-1/(4+q)}$ for $s = 1, \dots, q$, $\lambda_s \sim n^{-2/(4+q)}$ for $s = 1, \dots, r$, and $h_0 \sim n^{-2/(4+q)}$.

The above analysis provides only the optimal rate at which the bandwidths converge to zero. Although in principle one can compute plug-in bandwidths based on (9), the caveats noted earlier suggest

[2] One can also try to use $\alpha = \beta$, which will lead to $\beta = 1/5$ from (12) and $\beta = 1/4$ from (11), a contradiction. Similarly, assuming that $\alpha < \beta$ also leads to a contradiction. Therefore, we must have $\alpha > \beta$.

that this will not be feasible in applied settings. A plug-in method would first require estimation of the $B_{1s}(y|x)$'s, the $B_{2s}(y|x)$'s, along with $V(y|x)$, and $\Omega(y|x)$, which requires one to choose some initial 'pilot' bandwidths, while the accurate numerical computation of $(q+1)$-dimensional integration is challenging to say the least. Moreover, when some of the covariates are irrelevant variables, in the sense that they are in fact independent of the dependent variable, then some of the $B_{1s}(\cdot,\cdot)$'s or $B_{2s}(\cdot,\cdot)$'s are functions and (9) is no longer a valid expression for the leading MSE term of $\hat{F}(y|x)$. Therefore, it is highly desirable to be able to construct *automatic* data-driven bandwidth selection procedures that hopefully can also admit the existence of irrelevant covariates. Here, by *automatic* we mean that the method does not rely on pilot nonparametric estimators, i.e., it does not require one to choose some ad hoc pilot bandwidth values to estimate unknown (density) functions.

Unfortunately, to the best of our knowledge, there does not exist an *automatic* data-driven method for optimally selecting bandwidths when estimating a conditional CDF in the sense that a weighted integrated MSE is minimized. However, there do exist well developed *automatic* data-driven methods for selecting bandwidths when estimating the closely related conditional PDF. In particular, Hall et al. (2004) have considered the estimation of conditional probability density functions when the conditioning variables are a mix of discrete and continuous datatypes. Let $f(y|x)$ denote the conditional probability density function of Y given $X=x$, and let $g(y,x)$ denote the joint density of (Y,X). We estimate $f(y|x)$ by $\hat{f}(y|x)=\hat{g}(y,x)/\hat{\mu}(x)$, where $\hat{g}(y,x)=n^{-1}\sum_{i=1}^{n} w_{h_0}(Y_i,y)K_\lambda(X_i,x)$ is a kernel estimator of $g(y,x)$, and where $w_{h_0}(Y_i,y)=h_0^{-1}w\left(\frac{Y_i-y}{h_0}\right)$. Hall et al. (2004) have shown that a weighted square difference between $\hat{f}(y|x)$ and $f(y|x)$, i.e.,

$$WSQ_f \stackrel{def}{=} \int [\hat{f}(y|x)-f(y|x)]^2\mu(x)s(y,x)dxdy,$$

can be well approximated by a feasible cross-validation objective function $CV_f(h,\lambda)$, where

$$CV_f(h,\lambda)=\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{G}_{-i}(X_i)s(Y_i,X_i)}{\hat{\mu}_{-i}(X_i)^2}-\frac{2}{n}\sum_{i=1}^{n}\frac{\hat{g}_{-i}(Y_i,X_i)s(Y_i,X_i)}{\hat{\mu}_{-i}(X_i)} \tag{13}$$

where $\hat{\mu}_{-i}(X_i)=(n-1)^{-1}\sum_{j\neq i}K_\gamma(X_i,X_j)$ and $\hat{\mu}_{-i}(X_i)=(n-1)^{-1}\sum_{j\neq i}w_{h_0}(Y_i,Y_j)K_\gamma(X_i,X_j)$ are leave-one-out estimators of $\mu(X_i)$ and $g(Y_i,X_i)$ respectively, and

$$\hat{G}_{-i}(X_i)=\frac{1}{(n-1)^2}\sum_{j\neq i}^{n}\sum_{l\neq i}^{n}K_\gamma(X_i,X_j)K_\gamma(X_i,X_l)\int w_{h_0}(y,Y_j)w_{h_0}(y,Y_l)dy.$$

Note that $CV_f(h, \lambda)$ defined in (13) does not involve numerical integration, nor does it require any pilot estimators of nonparametric (density) functions. Therefore, one can use *automatic* data-driven methods for selecting the bandwidths (h, λ) by minimizing the cross-validation function $CV_f(h, \lambda)$. Hall et al. (2004) further show that the CV procedure can automatically remove irrelevant covariates via over-smoothing of the irrelevant variables. More specifically, Hall et al. (2004) show that, if x_s^d is an irrelevant variable, then the CV selected $\lambda_s \to 1$ in probability so that $l(x_{is}^d, x_s^d, \lambda_s) \to l(x_{is}^d, x_s^d, 1) \equiv 1$, and the resulting conditional PDF (or CDF) estimator will be unrelated to x_s^d, i.e., the irrelevant variable x_s^d is removed from the conditional PDF (CDF) estimator. Similarly, if x_s^c is an irrelevant variable, the CV selected $h_s \to \infty$ in probability so that $w((x_{is}^c - x_s^c)/h_s) \to w(0)$ (a constant), then the resulting conditional PDF (or CDF) estimator will be unrelated to x_s^c, so the irrelevant variable x_s^d is removed from the conditional PDF (CDF) estimator. We view this as a powerful result, as it extends the reach of nonparametric estimation techniques to high dimensional data when there exist irrelevant covariates, which appears to occur surprisingly often, particularly when modeling data in the social sciences.

Given the close relationship between a conditional PDF and a conditional CDF, intuitively it makes sense to adopt the automatic data-driven selected bandwidths from conditional PDF estimation and use them for conditional CDF estimation. Therefore, in this paper we recommend using the least squares cross-validation (LSCV) method based on conditional PDF estimation when selecting bandwidths and using them when estimating the conditional CDF function with either $\hat{F}(y|x)$ or $\tilde{F}(y|x)$.[3]

Letting $\bar{h}_s$ and $\bar{\lambda}_s$ denote the values of h_s and λ_s that minimize the cross-validation function $CV_f(h, \lambda)$ (from conditional PDF estimation), using second order kernels one can show that the optimal bandwidths are of order $\bar{h}_s \sim n^{-1/(5+q)}$ for $s = 0, 1, \ldots, q$, and $\bar{\lambda}_s \sim n^{-2/(5+q)}$ for $s = 1, \ldots, r$. Note that the exponential factor is $1/(5+q)$ rather than $1/(4+q)$ because the dimension of (y, x^c) in conditional PDF estimation is $q+1$ rather than q.

To obtain bandwidths that have the correct optimal rates for $\hat{F}(y|x)$, we recommend using $\hat{h}_0 = \bar{h}_0 n^{\frac{1}{5+q} - \frac{2}{4+q}}$, $\hat{h}_s = \bar{h}_s n^{\frac{1}{5+q} - \frac{1}{4+q}}$ $(s = 1, \ldots, q)$, and $\hat{\lambda}_s = \bar{\lambda}_s n^{\frac{2}{5+q} - \frac{2}{4+q}}$ $(s = 1, \ldots, r)$. By Theorem 4.1 of

[3] Of course, for $\tilde{F}(y|x)$ that does not smooth y, we would use only those bandwidths associated with x.

Hall et al. (2004), we know that

$$\begin{aligned}
\hat{h}_0 &= a_0^0 n^{-2/(4+q)} + o_p\left(n^{-2/(4+q)}\right), \\
\hat{h}_s &= a_s^0 n^{-1/(4+q)} + o_p\left(n^{-1/(4+q)}\right) \text{ for } s = 1, \dots, q, \\
\hat{\lambda}_s &= b_s^0 n^{-2/(4+q)} + o_p\left(n^{-2/(4+q)}\right) \text{ for } s = 1, \dots, r,
\end{aligned} \tag{14}$$

where the a_s^0's are uniquely defined finite positive constants, and the b_s^0's are uniquely defined finite non-negative constants. Equation (14) gives the correct optimal rates for the h_s's and λ_s's when estimating $F(y|x)$ via $\hat{F}(y|x)$.

The conclusions of Theorem 2.1 and Theorem 2.2 remain valid with the above data-driven (i.e., stochastic) bandwidths. This can be proved by using arguments similar to those found in the proof of Theorem 4.2 of Hall et al. (2004).

3. Estimating Conditional Quantile Functions with Mixed Data Covariates

A conditional αth quantile of a conditional distribution function $F(\cdot|x)$ is defined by ($\alpha \in (0,1)$)

$$q_\alpha(x) = \inf\{y : F(y|x) \geq \alpha\} = F^{-1}(\alpha|x). \tag{15}$$

Or equivalently, $F(q_\alpha(x)|x) = \alpha$. Therefore, we can estimate the conditional quantile function $q_\alpha(x)$ by inverting the estimated conditional CDF function defined above. By inverting $\hat{F}(\cdot|\cdot)$ we obtain

$$\hat{q}_\alpha(x) = \inf\{y : \hat{F}(y|x) \geq \alpha\} \equiv \hat{F}^{-1}(\alpha|x). \tag{16}$$

Or by inverting the indicator function based CDF estimate $\tilde{F}(y|x)$, we get

$$\tilde{q}_\alpha(x) = \inf\{y : \tilde{F}(y|x) \geq \alpha\} \equiv \tilde{F}^{-1}(\alpha|x). \tag{17}$$

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

$$\hat{q}_\alpha(x) = \arg\min_q |\, \alpha - \hat{F}(q|x)\, |. \tag{18}$$

That is, the value of q that minimizes (18) gives us $\hat{q}_\alpha(x)$. We make the following assumption.

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

Note that $f(y|x) \equiv F_0(y|x) = \frac{\partial}{\partial y} F(y|x)$. Below we present the asymptotic distribution of $\hat{q}_\alpha(x)$.

THEOREM 3.1. *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 (C1) to (C5), we have*

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

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

The proof of Theorem 3.1 is given in Appendix A.

Remark 3.1. *We recommend that in practice one use the data-driven method discussed in Section 2 to select the smoothing parameters for estimating $\hat{F}(y|x)$. Theorem 3.1 holds true with stochastic data-driven bandwidths $\hat{h}_0$, $\hat{h}_s$, $\hat{\lambda}_s$ that satisfy* (14).

Similarly, one can obtain the asymptotic distribution of $\tilde{q}_\alpha(x)$.

THEOREM 3.2. *Define $\tilde{B}_{n,\alpha}(x) = [\sum_{s=1}^q h_s^2 B_{1s}(q_\alpha(x)|x) + \sum_{s=1}^r \lambda_s B_{2s}(q_\alpha(x)|x)]/f(q_\alpha(x)|x)$. Then under (C1), (C2), (C3) and (C5), we have*

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

where $V_\alpha(x)$ is defined in Theorem 3.1.

The proof of Theorem 3.2 is similar to the proof of Theorem 3.1 and is thus omitted here.

3.1. **The Check Function Approach.** One can also estimate a nonparametric quantile regression model using the so-called 'check function' approach. We discuss a local constant method (one can also use a local linear method), and we choose a to minimize the following objective function

$$\min_a \sum_{j=1}^n \rho_\alpha(Y_j - a) K_\lambda(X_i, x), \tag{19}$$

where $\rho_\alpha(v) = v[\alpha - I(v \leq 0)]$ is called the check function. Using a check function approach to estimate conditional quantiles when X_i does not include discrete components has been considered by Jones & Hall (1990), Chaudhuri (1991), Yu & Jones (1998), and Honda (2000). Also, He, Ng & Portony (1998) and He & Ng (1999) have considered nonparametric conditional quantile estimation using smoothing splines.

Letting $\tilde{a}_\alpha(x)$ denote the value of a that minimizes (19), then the asymptotic distribution of $\tilde{a}_\alpha(x)$ is the same to that of $\tilde{q}_\alpha(x)$ and is omitted here (e.g., Jones & Hall (1990)).

$\tilde{a}_\alpha(x)$ is a local constant quantile estimator because in (19) one approximates $q_\alpha(x)$ locally via the constant a. One can also use a local linear approximation where instead one minimizes

$$\sum_{j=1}^{n} \rho_\alpha(Y_j - a - b'(X_j^c - x^c))K_\lambda(X_i, x) \tag{20}$$

with respect to a and b (b is of dimension $q \times 1$, and the resulting estimator of a, say $\hat{a}_\alpha(x)$, is the local linear estimator of $q_\alpha(x)$). Note that in (20), one can only have a local linear expansion for the continuous covariate(s) x^c. If instead one were to also add a linear term in x^d, say $c'(X_j^d - x^d)$, the above minimization problem would not produce a consistent estimator.[4]

In this paper we only consider the independent data case. However, it is straightforward, though more tedious, to generalize the results of this paper to the weakly dependent data case such as β-mixing and α-mixing processes. For conditional quantile estimation with only continuous conditional covariates that allows for α-mixing data, see Cai (2002), Cai & Xu (2004), and the references therein.

4. Monte Carlo Simulation and Empirical Applications

4.1. **Monte Carlo Simulation: Mixed Data and Irrelevant Covariates.** We consider a variety of methods for estimating a conditional quantile. First we consider the proposed CDF-based quantile method and compare it with several existing approaches such as a local linear (LL) check function quantile estimator (Hall et al. (1999)) and a local constant (LC) variant (Yu & Jones (1998)). We entertain a variety of bandwidth selection methods including conditional PDF-based cross-validation, frequency, and ad hoc methods. Also, we consider a correctly specified parametric quantile model (SINQR) and a linear approximation (LINQR). The quantile regression approaches are based upon Koenker's (2004) R library which utilizes the interior point (Frisch-Newton) algorithm for quantile regression, while the CDF-based approaches are based on the N © C library.

[4]Because we allow for $\lambda_s = 0$ for $s = 1, \ldots, r$, in this case $L(X_j^d, x^d, \lambda) \neq 0$ only when $X_j^d = x^d$. However, this will cause $X_j^d - x^d = 0$, and consequently, $c'(X_j^d - x^d) = 0$, and the coefficient c is not well defined.

The DGP is given by

$$Y_i = \beta_1 + \beta_2 X^d_{i2} + \beta_3 X^d_{i3} + \beta_4 \sin(X^c_{i4}) + \epsilon_i, \quad i = 1, \ldots, n,$$

where $x^d_2 \in \{0, 1, 2\}$ with probability 0.49, 0.42, and 0.09, $x^d_3 \in \{0, 1, 2\}$ with probability 0.09, 0.42, and 0.49, and $x^c_4 \sim$ uniform$[-2\pi, 2\pi]$ on an equally spaced grid. We set $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)' = (1, 1, 0, 1)$ so that x^d_3 is an irrelevant discrete variable. For the simulations that follow, we set $n = 100$. By way of example, we consider a variety of error distributions, including

(1) Symmetric i.i.d.: $\epsilon_i \sim N(0, 1)$.

(2) Asymmetric i.i.d.: $\epsilon_i \sim \chi^2$ with 4 degrees of freedom.

(3) Heteroskedastic: $\epsilon_i \sim N(0, \sigma_i^2)$ with $\sigma_i = \frac{1}{\pi}\sqrt{(X^c_{i4})^2 + \pi}$.

We report the relative efficiency generated from 1,000 Monte Carlo replications drawn from the above DGP, i.e., a method's median MSE divided by the median MSE of the proposed method. We consider a variety of estimators and range of bandwidth selectors for h_y (where appropriate), $\lambda_{x^d_2}$, $\lambda_{x^d_3}$, and $h_{x^c_4}$. Rather than present page after page of tables, we instead summarize representative results for the 40th quantile ($\alpha = 0.4$) and median ($\alpha = 0.5$) in Table 1 below.[5] Extensive simulation results for this and larger sample sizes are available upon request from the authors.

Results summarized in Table 1 suggest that the proposed estimator dominates its peers by a wide margin when modeling quantiles defined over a mix of discrete and continuous covariates. The non-parametric competitors appear in lines 8[6] and 11[7], each having a relative efficiency of roughly 2 or higher, i.e., the mean square error of these methods is more than 100% larger than that for the proposed method. The $\chi^2_{df=4}$ error model is extremely noisy, and the misspecified linear parametric model appears to be marginally more efficient for this DGP, however, this vanishes rapidly and exceeds 1 for $n \geq 200$ (note that the variance of this process is eight times that for the $N(0, 1)$ DGP). Interestingly, the proposed approach using *non-stochastic* ad hoc bandwidths for y and x^c_4 combined with LSCV bandwidths for x^d_2 and x^d_3 appears to be marginally more efficient for the extremely noisy $\chi^2_{df=4}$ error

[5]Legend for Table 1: LSCV = PDF-based Least Squares Cross-Validation. LLCV = Local linear based Least Squares Cross-Validation. LCCV = Local constant based Least Squares Cross-Validation. FREQ = Sample Splitting. AH = Ad Hoc Normal Reference Rule-of-thumb, $h = 1.06\sigma n^{-1/(4+p+q)}$. EFF = Median MSE Relative to CDF with LSCV Bandwidths. NA = Not Applicable. LINQR = linear approximation parametric quantile model. SINQR = correctly specified parametric quantile function under conditional homoskedastic errors.

[6]The local linear check function estimator using an indicator function for Y (LLQR/LLCV/NA/FREQ/FREQ).

[7]The local constant check function estimator using an indicator function for Y (LCQR/LCCV/NA/FREQ/FREQ).

Table 1. Monte Carlo MSE summary results. The estimation method appears in column 1. Columns 2-5 indicate the bandwidth selector (h_y = bandwidth for y, $\lambda_{x_2^d}$ = bandwidth for discrete x_2^d, $\lambda_{x_3^d}$ = bandwidth for discrete x_3^d, $h_{x_4^c}$ = bandwidth for continuous x_4^c). The remaining columns present the relative efficiency of each approach with respect to the proposed approach. Entries greater than 1 indicate a loss in efficiency relative to the proposed approach.

					$\epsilon_i \sim N(0,1)$		$\epsilon_i \sim \chi^2_{df=4}$		$\epsilon_i \sim N(0,\sigma_i^2)$	
Method	$h_{x_4^c}$	h_y	$\lambda_{x_2^d}$	$\lambda_{x_3^d}$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.4$	$\alpha=0.5$	$\alpha=0.4$	$\alpha=0.5$
CDF	LSCV	LSCV	LSCV	LSCV	1.00	1.00	1.00	1.00	1.00	1.00
CDF	LSCV	IND	LSCV	LSCV	1.28	1.30	1.11	1.10	1.25	1.27
CDF	AH	AH	LSCV	LSCV	1.32	1.31	0.92	0.92	2.54	1.35
CDF	AH	IND	FREQ	FREQ	1.88	1.88	2.40	2.48	1.53	1.53
CDF	LCCV	IND	LCCV	LCCV	1.11	1.12	1.20	1.17	1.25	1.28
LLQR	AH	NA	FREQ	FREQ	1.94	1.88	3.46	3.33	2.03	2.05
LLQR	LLCV	NA	FREQ	FREQ	2.21	2.15	3.40	3.29	2.47	2.48
LLQR	LLCV	NA	LLCV	LLCV	1.20	1.17	1.34	1.32	1.55	1.55
LCQR	AH	NA	FREQ	FREQ	1.80	1.79	2.59	2.73	1.53	1.53
LCQR	LCCV	NA	FREQ	FREQ	2.02	1.99	2.41	2.35	2.09	2.13
LCQR	LCCV	NA	LCCV	LCCV	1.13	1.13	1.17	1.16	1.25	1.28
LINQR	NA	NA	NA	NA	1.83	1.85	0.93	0.88	1.28	1.29
SINQR	NA	NA	NA	NA	0.20	0.20	0.34	0.41	0.20	0.17

process.[8] However, this likely reflects the additional variability inherent to data-driven methods for this DGP. Not surprisingly, the correctly specified parametric model (in line 14, SINQR) performs the best for all cases.[9] In contrast, the misspecified parametric linear quantile model performs poorly. The competitors to our nonparametric method remain the frequency-based nonparametric counterparts.

The superior performance of the CDF-based quantile estimator stems from the fact that it can automatically remove irrelevant variables and does not rely upon sample splitting. Note that rows 9 and 12 are for new estimators, ones extended to handle the presence of categorical data without resorting to sample splitting using the approach found in Li & Racine (2003), Racine & Li (2004), and so forth. Note also that the *extended* local constant and local linear quantile regression estimators have the ability to automatically remove *discrete* variables, though the extended local linear estimator cannot remove continuous ones. Even so, these estimators do not appear to dominate the proposed

[8]Results in line 4 (CDF/AH/AH/LSCV/LSCV).

[9]For the conditional heteroskedastic error case and with $\alpha \neq 0.5$, SINQR corresponds a (slightly) misspecified model. Simulation not reported here show that when n is sufficiently large and for $\alpha \neq 0.5$, the robust nonparametric estimators perform better than the SINQR estimator.

method in this setting, having relative efficiencies exceeding 1. This arises due to the fact that we smooth the dependent variable.

Summarizing, there are a number of novel features associated with the proposed approach embodied in the simulations. First, we consider *direct* estimation of a conditional quantile based upon robust estimates of a conditional CDF, $F(y|x)$. Second, our approach is able to handle a mix of discrete and continuous data without resorting to sample splitting. Third, we consider a fully data-driven method of bandwidth selection, based upon bandwidths that are optimal for estimating a conditional PDF, $f(y|x)$ (Hall et al. (2004)). Fourth, our approach is able to automatically remove irrelevant covariates. Note that bandwidth selection for nonparametric quantile estimators remains an open topic of research. In particular, existing approaches employ methods for bandwidth selection for regression mean estimation (see, e.g., Yu & Jones (1998)).

4.2. **Empirical Application: Conditional Value at Risk.** Financial instruments are subject to many types of risk, including interest risk, default risk, liquidity risk, and market risk. Value at risk (VaR) seeks to measure the latter, and is a single estimate of the amount by which one's position in an instrument could decline due to general market movements over a specific holding period. One can think of VaR as the maximal loss of a financial position during a given time period for a given probability; hence VaR is simply a measure of loss arising from an unlikely event under normal market conditions (Tsay (2002, pg 257)).

Letting $\Delta V(l)$ denote the change in value of a financial instrument from time t to $t + l$, we define VaR over a time horizon l with probability p as

$$p = P[\Delta V(l) \leq \text{VaR}] = F_l(\text{VaR}),$$

where $F_l(\text{VaR})$ is the unknown CDF of $\Delta V(l)$. VaR assumes negative values when p is small since the holder of the instrument suffers a loss.[10] Therefore, the probability that the holder will suffer a loss greater than or equal to VaR is p, and the probability that the maximal loss is VaR is $1 - p$. The quantity

$$\text{VaR}_p = \inf\{\text{VaR}|F_l(\text{VaR}) \geq p\}$$

[10]For a long position we are concerned with the lower tail of returns, i.e., a loss due to a fall in value. For a short position we are concerned with the upper tail, i.e., a loss due to a rise in value. In the latter case we simply model $1 - F_l(\text{VaR})$ rather than $F_l(\text{VaR})$. Of course, in either case we are modeling the tails of a distribution.

is the pth quantile of $F_l(\text{VaR})$; hence VaR_p is simply the pth quantile of the CDF $F_l(\text{VaR})$. The convention is to drop the subscript p and refer to, e.g, 'the 5% VaR' ($\text{VaR}_{0.05}$). Obviously the CDF $F_l(\text{VaR})$ is unknown in practice and must be estimated.

Conditional value at risk (CVaR) seeks to measure VaR when covariates are involved, which simply conditions one's estimate on a vector of covariates, X. Of course, the conditional distribution function $F_l(\text{VaR}|X)$ is again unknown and must be estimated. A variety of parametric methods can be found in the literature. In what follows, we compare results from five popular parametric approaches with the nonparametric quantile estimator developed in this paper. The econometric importance of this application is that such models are widely used in financial settings and potentially large sums of money may be involved, hence methods that are robust to functional specification ought to be of interest to practitioners and theoreticians alike.

We consider data used in Tsay (2002) for IBM stock (r_t, daily log returns (%) from July 3, 1962, through December 31, 1998). Log returns correspond approximately to percentage changes in the value of a financial position, and are used throughout. The CVaR computed from the quantile of the distribution of r_{t+1} conditional upon information available at time t is therefore in percentage terms, and so the dollar amount of CVaR is the cash value of one's position multiplied by the CVaR of the log return series. For what follows, we compute CVaR for a one day horizon ($l = 1$).

Following Tsay (2002), we model CVaR for daily log returns of IBM stock using the following explanatory variables:

(1) X_{1t}: an indicator variable for October, November, and December, equaling 1 if t is in the fourth quarter. This variable takes care of potential fourth-quarter effects (or year-end effects), if any, on the daily IBM stock returns.

(2) X_{2t}: an indicator variable for the behavior of the previous trading day, equaling 1 if and only if the log return on the previous trading day was $\leq 2.5\%$. This variable captures the possibility of panic selling when the price of IBM stock dropped 2.5% or more on the previous trading day.

(3) X_{3t}: a qualitative measurement of volatility, measured as the number of days between $t-1$ and $t-5$ having a log return with magnitude exceeding a threshold ($|r_{t-i}| \geq 2.5\%$, $i = 1, 2, \ldots, 5$).

(4) X_{4t}: an annual trend defined as (year of time t-1961)/38, used to detect any trend behavior of extreme returns of IBM stock.

(5) X_{5t}: a volatility series based on a Gaussian GARCH(1,1) model for the mean corrected series, equaling σ_t where σ_t^2 is the conditional variance of the GARCH(1,1) model.

Table 2 presents results reported in Tsay (2002, pg 282,295) along with our proposed approach for a variety of existing approaches towards measuring CVaR for December 31, 1998. The explanatory variables assumed values of $X_{1,9190} = 1$, $X_{2,9190} = 0$, $X_{3,9190} = 0$, $X_{4,9190} = 0.9737$, and $X_{5,9190} = 1.7966$. Values are based on an assumed long position of \$10 million, hence if value at risk is minus 2%, then we obtain VaR $= \$10,000,000 \times 0.02 = \$200,000$.

Table 2. Conditional Value at Risk for a long position in IBM stock

Model	5% CVaR	1% CVaR
Inhomogeneous Poisson, GARCH(1,1)	\$303,756	\$497,425
Conditional Normal, IGARCH(1,1)	\$302,500	\$426,500
AR(2)-GARCH(1,1)	\$287,700	\$409,738
Student-t_5 AR(2)-GARCH(1,1)	\$283,520	\$475,943
Extreme Value	\$166,641	\$304,969
LSCV CDF	\$258,727	\$417,192

Observe that, depending on one's choice of parametric model, one can obtain estimates that differ by as much 82% for 5% CVaR and 63% for 1% for this example. Of course, this variation arises due to model uncertainty. One might take a Bayesian approach to deal with this uncertainty, averaging over models, or alternatively consider nonparametric methods that are robust to functional specification. The proposed method provides guidance yielding sensible estimates that might be of value to practitioners.

Interestingly, Tsay (2002) finds that X_{1t} and X_{2t} are irrelevant explanatory variables. The LSCV bandwidths are consistent with this. However, X_{3t} is also found to be irrelevant; hence only one volatility measure is relevant in the sense of Hall et al. (2004).

4.3. **Empirical Application: Boston Housing Data.** We consider the 1970s era Boston housing data that has been extensively analyzed by a number of authors. This dataset contains $n = 506$ observations, and the response variable Y is the median price of a house in a given area. Following Chaudhuri et al. (1997, pg 724), we focus on three important covariates: RM = average number of rooms per house in the area, LSTAT = % of the population having lower economic status in the area, and DIS = weighted distance to five Boston employment centers. The variable RM was rounded to the nearest integer and assumes six unique values, $\{4, 5, \ldots, 9\}$. We treat RM as a discrete ordered variable and the cross-validated smoothing parameter for the entire sample was $\hat{\lambda} = 0.14$ indicating that some smoothing of this variable is appropriate. An interesting feature is that the data is right censored at \$50,000 (1970's housing prices) which makes this particularly well suited to quantile methods.

The econometric importance of this application follows from the fact that hedonic pricing models are widely used in a variety of settings, hence methods that are robust to functional specification ought to be of interest to practitioners and theoreticians alike.

We first shuffle the data and create two independent samples of size $n_1 = 400$ and $n_2 = 106$. We then fit a linear parametric quantile model and nonparametric quantile model using the estimation sample of size n_1, and generate the predicted median of Y based upon the covariates in the independent hold-out data of size n_2. Finally, we compute the mean square prediction error defined as MSPE = $n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{q}_{0.5}(X_i))^2$ where $q_{0.5}(X_i)$ is the predicted median generated from either the parametric or nonparametric model, and Y_i is the actual value of the response in the hold-out dataset. To deflect potential criticism that our result reflects an unrepresentative split of the data, we repeat this process 100 times, each time computing the relative MSPE (i.e. the parametric MSPE divided by the nonparametric MSPE). For each split, we use the method of Hall et al. (2004) to compute data-dependent bandwidths. The median relative MSPE over the 100 splits of the data is 1.13 (lower quartile=1.03, upper quartile=1.20), indicating that the nonparametric approach is producing superior out-of-sample quantile estimates. Figure 1 presents a density estimate summarizing these results for the 100 random splits of the data.

Figure 1 reveals that 76% of all sample splits (i.e., the area to the right of the vertical bar) had a relative efficiency greater than 1, or equivalently, in 76 out of 100 splits, the nonparametric quantile model yields better predictions of the median housing price than the parametric quantile model. Given the small sample size and the fact that there exist three covariates, we feel this is a telling application for the proposed fully nonparametric method. Of course, we are not suggesting that we will outperform an approximately correct parametric model. Rather, we only wish to suggest that we can often outperform common parametric specifications that can be found in the literature.

5. Conclusions

We propose a conditional CDF estimator defined over a mix of discrete and continuous covariates and an associated quantile function estimator. We also propose a conditional PDF-based method of

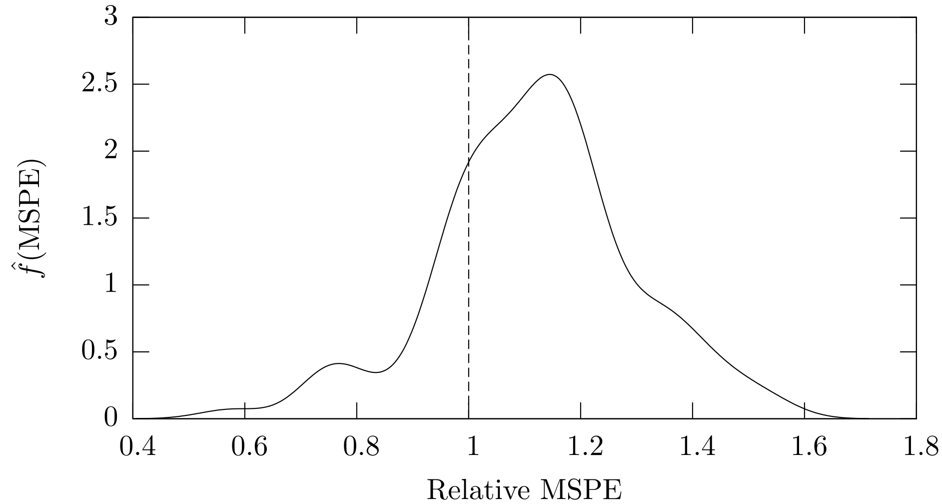

Figure 1. Density of the relative MSPE over the 100 random splits of the data. Values > 1 indicate superior out-of-sample performance of the proposed method for a given random split of the data [lower quartile=1.03, median=1.13, upper quartile=1.20].

data-driven bandwidth selection. In applied settings one frequently encounters a mix of datatypes; thus this estimator would be of value to applied researchers. Simulations demonstrate that the proposed approach performs quite well, and two applications highlight its value in applied settings. Future work includes a native method for data-driven bandwidth selection for the conditional CDF estimator.

References

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

Bierens, H. (1987), Kernel estimators of regression functions, *in* T. Bewley, ed., 'Advances in Econometrics: Fifth World Congress, Vol I', Cambridge University Press, pp. 99–144.

Buchinsky, M. (1994), 'Changes in the U.S. wage structure 1963-1987: Application of quantile regression', *Econometrica* **62**, 405–458.

Buchinsky, M. & Hahn, J. (1998), 'An alternative estimator for the censored quantile regression model', *Econometrica* **66**, 653–671.

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

Cai, Z. & Xu, X. (2004), Nonparametric quantile estimations for dynamic smooth coefficient models, Technical report, Department of Mathematics and Statistics, University of North Carolina.

Chaudhuri, P. (1991), 'Nonparametric estimates of regression quantiles and their local Bahadur representation', *Annals of Statistics* **19**, 760–777.

Chaudhuri, P., Doksum, K. & Samarov, A. (1997), 'On average derivative quantile regression', *Annals of Statistics* **25**(2), 715–744.

Guerre, E., Perrigne, I. & Vuong, Q. (2000), 'Optimal nonparametric estimation of first-price auctions', *Econometrica* **68**, 525–574.

Hall, P., Racine, J. & Li, Q. (2004), 'Cross-validation and the estimation of conditional probability densities', *Journal of the American Statistical Association* **99**(2), 1015–1026.

Hall, P., Wolff, R. C. L. & Yao, Q. (1999), 'Methods for estimating a conditional distribution function', *Journal of the American Statistical Association* **94**, 154–163.

Hansen, B. (2004), Nonparametric estimation of smooth conditional distributions, Technical report, University of Wisconsin.

He, X. & Ng, P. (1999), 'Quantile splines with several covariates', *Journal of Statistical Planning and Inference* **75**, 343–352.

He, X., Ng, P. & Portony, S. (1998), 'Bivariate quantile smoothing splines', *Journal of the Royal Statistical Society, Series B* **60**, 537–550.

Honda, T. (2000), 'Nonparametric estimation of a conditional quantile for α-mixing processes', *Annals of the Institute of Statistical Mathematics* **52**, 459–470.

Horowitz, J. (1998), 'Bootstrap methods for median regression models', *Econometrica* **66**, 1327–1352.

Hsiao, C., Li, Q. & Racine, J. (2004), A consistent model specification test with mixed categorical and continuous data, Unpublished manuscript, Department of Economics, Texas A & M University.

Jones, M. & Hall, P. (1990), 'Mean squared error properies of kernel estimates of regression quantiles', *Statistics and Probability Letters* **10**, 283–289.

Koenker, R. (2004), *quantreg: Quantile Regression*. R package version 3.72, Initial R port from Splus by Brian Ripley (ripley@stats.ox.ac.uk).

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

Koenker, R. & Xiao, Z. (2002), 'Inference on the quantile regression process', *Econometrica* **70**(4), 1583–1612.

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

Li, Q. & Racine, J. (2003), 'Nonparametric estimation of distributions with categorical and continuous data', *Journal of Multivariate Analysis* **86**, 266–292.

Li, T., Perrigne, I. & Vuong, Q. (2000), 'Conditional independent private information in OCS wildcat auctions', *Journal of Econometrics* **98**, 131–161.

Matzkin, R. L. (2003), 'Nonparametric estimation of nonadditive random functions', *Econometrica* **71**(5), 1339–1375.

Powell, J. L. (1986), 'Censored regression quantiles', *Journal of Econometrics* **32**, 143–155.

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

Tsay, R. S. (2002), *Analysis of Financial Time Series*, Wiley Series in Probability and Statistics.

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

Appendix A: Proofs of Theorem 2.1 and Theorem 2.2

Theorem 2.1 is proved in lemmas A.1 and A.2 below, while Theorem 2.2 is proved in lemmas A.3 and A.4.

Lemma A.1. *Let $B_{1s}(y|x)$ and $B_{2s}(y|x)$ be defined as in Theorem 2.1. Under conditions (C1) to (C3),*

$$E[\tilde{M}(y|x)] = \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) + o(|\lambda|^2 + |h|^4).$$

Proof. Letting $\mu_{ts} = \frac{\partial^2 \mu(x)}{\partial x_s \partial x_t}$ and $F_{ts} = \frac{\partial^2 F(y|x)}{\partial x_s \partial x_t}$, we have ($\int dx = \sum_{x \in D} \int dx^c$)

$$\begin{aligned}
&E[\tilde{F}(y|x)\hat{\mu}(x)] = E[K_\gamma(X_i, x)E(I(Y_i \le y)|X_i)] = E[K_\gamma(X_i, x)F(y|X_i)] \\
&= (h_1 \dots h_q)^{-1} \sum_{z^d \in D} \int \mu(z) W\left(\frac{z^c - x^c}{h}\right) L(z^d, x^d, \lambda) F(y|z) dz^c \\
&= \sum_{z^d \in D} L(z^d, x^d, \lambda) \int \mu(x^c + hv, z^d) F(y|x^c + hv, z^d) W(v) dv \\
&= \sum_{z^d \in D} \left[I_{z^d = x^d} + \sum_{s=1}^{r} \lambda_s \sum_{z^d \in D} I_s(z^d, x^d) + O(|\lambda|^2) \right] \\
&\times \left\{ \int \left[\mu(x^c, z^d) + \sum_{s=1}^{q} \mu_s(x^c, z^d) h_s v_s + \frac{1}{2} \sum_{s=1}^{q} \sum_{t=1}^{q} \mu_{st}(x^c, z^d) h_s h_t v_s v_t \right] \right. \\
&\times \left. \left[F(y|x^c, z^d) + \sum_{s=1}^{q} h_s F_s(y|x^c, z^d) v_s + \frac{1}{2} \sum_{s=1}^{q} \sum_{t=1}^{q} h_s h_t F_{st}(y|x^c, z^d) v_s v_t \right] W(v) dv + o(|h|^2) \right\} \\
&= \mu(x)F(y|x) + (1/2)\kappa_2 \sum_{s=1}^{q} h_s^2 \left[2\mu_s(x) F_s(y|x) + \mu(x) F_{ss}(y|x) + \mu_{ss}(x) F(y|x) \right] \\
&+ \sum_{z^d \in D} I_s(z^d, x^d) \sum_{s=1}^{r} \lambda_s \mu(x^c, z^d) F(y|x^c, z^d) + o(|h|^2 + |\lambda|).
\end{aligned} \tag{A.1}$$

Similarly,

$$E[\hat{\mu}(x)] = \mu(x) + (1/2)\kappa_2 \sum_{s=1}^{q} h_s^2 \mu_{ss}(x) + \sum_{s=1}^{r} \lambda_s \sum_{z^d \in D} I_s(z^d, x^d) \mu(x^c, z^d) + o(|h|^2 + |\lambda|). \tag{A.2}$$

Therefore, (A.1) and (A.2) lead to

$$
\begin{aligned}
E[\tilde{M}(y|x)] &= E[\tilde{F}(y|x)\hat{\mu}(x)] - F(y|x)E[\hat{\mu}(x)] \\
&= (1/2)\kappa_2 \sum_{s=1}^{q} h_s^2[2F_s(y|x)\mu_s(x) + \mu(x)F_{ss}(y|x)] \\
&+ \sum_{z^d \in D} \lambda_s I_s(z^d, x^d)[F(y|x^c, z^d)\mu(x^c, z^d) - F(y|x)\mu(x)] + o(|h|^2 + |\lambda|) \\
&= \mu(x)\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \mu(x)\sum_{s=1}^{r} \lambda_s B_{2s}(y|x) + o(|h|^2 + |\lambda|). \qquad \text{(A.3)}
\end{aligned}
$$

$$
\begin{aligned}
Var\Big(\tilde{M}(y|x) - F(y|x)\hat{\mu}(x)\Big) &= n^{-1}Var[(I(Y_i \leq y) - F(y|x))K_\gamma(X_i, x)] \\
&= n^{-1}\left\{E[(I(y|X_i) - F(y|x))^2|X_i]K_\gamma(X_i, x)^2] + O(1)\right\} \\
&= n^{-1}E\left\{[F(y|X_i) - 2F(y|x)F(y|X_i) + F(y|x)^2]K_\gamma(X_i, x)^2]\right\} + O(n^{-1}) \\
&= n^{-1}\int \mu(z)\{[F(y|z) - 2F(y|x)F(y|z) + F(y|x)^2]K_\gamma(z, x)^2]\}dz + O(n^{-1}) \\
&= (nh_1 \dots h_q)^{-1}\sum_{z^d \in D}\int \mu(x^c + hv, z^d)\Big\{\Big[F(y|x^c + hv, z^d) - 2F(y|x)F(y|x^c + hv, z^d) \\
&\quad +F(y|x)^2\Big] \times W(v)^2 dv L(z^d, x^d, \lambda)^2]\Big\} + O(n^{-1}) \\
&= \kappa^q (nh_1 \dots h_q)^{-1}[F(y|x) - F(y|x)^2] + O((nh_1 \dots h_q)^{-1}(|h|^2 + |\lambda|) + n^{-1}). \qquad \text{(A.4)}
\end{aligned}
$$

□

Lemma A.2. *Under conditions (C1) to (C3),*

$$
(nh_1 \dots h_q)^{1/2}[\tilde{F}(y|x) - F(y|x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y|x) - \sum_{s=1}^{r}\lambda_s B_{2s}(y|x)] \to N(0, V),
$$

where $V = \kappa^q F(y|x)[1 - F(y|x)]/\mu(x)$.

Proof. Define $B_n = \mu(x)\sum_{s=1}^{q} h_s^2 B_{1s}(y|x) + \mu(x)\sum_{s=1}^{r}\lambda_s B_{2s}(y|x)$. Then

$$
\begin{aligned}
[\tilde{F}(y|x) - F(y|x) - B_n(y|x)] &= [\tilde{F}(y|x) - F(y|x) - B_n(y|x)]\hat{\mu}(x)/\hat{\mu}(x) \\
&= [\tilde{F}(y|x) - F(y|x) - B_n(y|x)]\hat{\mu}(x)/\mu(x) + o_p(1) \equiv \tilde{A}(y|x)/\mu(x) + o_p(1),
\end{aligned}
$$

where $\tilde{A}(y|x) = [\tilde{F}(y|x) - F(y|x) - B_n(y|x)]\hat{\mu}(x)$. By Theorem 2.1 (i) and (ii) we know that $E[\tilde{A}(y|x)] = o(|h|^2 + |\lambda|)$ and that $Var[\tilde{A}(y|x)] = (nh_1 \dots h_q)^{-1}\mu(x)^2 V(y|x) + o((nh_1 \dots h_q)^{-1})$. Hence, $(nh_1 \dots h_q)^{1/2}\tilde{A}(y|x) \to N(0, \mu(x)^2 V(y|x))$ in distribution by Liapunov's central limit theorem. Therefore, we have

$$
\begin{aligned}
(nh_1 \dots h_q)^{1/2}[\tilde{F}(y|x) - F(y|x) - B_n(y|x)] &= (nh_1 \dots h_q)^{1/2}\tilde{A}(y|x)/\mu(x) + o_p(1) \\
&\to \mu(x)^{-1}N(0, \mu(x)^2 V(y|x)) = N(0, V(y|x)) \text{ in distribution.}
\end{aligned}
$$

□

Lemma A.3. *Under conditions (C1) to (C4),*

(i) $E[\hat{M}(y|x)] = \sum_{s=0}^{q} h_s^2 B_{1s}(y|x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y|x) + o(|\lambda| + |h|^2)$,
(ii) $Var(\hat{M}(y|x)) = \kappa^q (nh_1 \dots h_q)^{-1}[F(y|x) - F(y|x)^2 - h_0 C_w F_0(y|x)]\mu(x)$.

Proof. First, by Lemma A.5 we know that

$$E\left[G\left(\frac{y-Y_i}{h_0}\right)\Big|X_i\right] = F(y|X_i) + (1/2)h_0^2\kappa_2 F_{00}(y|X_i) + o(h_0^2). \tag{A.5}$$

Using an approach similar to that used to prove (A.1) and also using (A.5), we have

$$\begin{aligned}
E[\hat{F}(y|x)\hat{\mu}(x)] &= E\left[G\left(\frac{y-Y_i}{h_0}\right)K_\gamma(X_i,x)\right] = E\left[K_\gamma(X_i,x)E\left(G\left(\frac{y-Y_i}{h_0}\right)\Big|X_i\right)\right]\\
&= E\left\{K_\gamma(X_i,x)[F(y|X_i) + (1/2)h_0^2\kappa_2 F_{00}(y|X_i) + o(h_0^2)]\right\}\\
&= \int \mu(z)K_\gamma(z,x)\left[F(y|z) + (1/2)h_0^2\kappa_2 F_{00}(y|z) + o(h_0^2)\right]dz\\
&= \sum_{z^d\in D} L(z^d,x^d,\lambda)\int \mu(x^c+hv,z^d)\left[F(y|x^c+hv,z^d) + (1/2)h_0^2\kappa_2 F_{00}(y|x^c+hv,z^d) + o(h_0^2)\right]W(v)dv\\
&= \left[\sum_{z^d\in D} I(z^d=x^d) + \sum_{s=1}^{r}\lambda_s \sum_{z^d\in D} I_s(z^d,x^d) + O(|\lambda|^2)\right]\\
&\times \int\left[\mu(x^c,z^d) + \sum_{s=1}^{q} h_s\mu_s(x^c,z^d)v_s + (1/2)\sum_{s=1}^{q}\sum_{t=1}^{q} h_s h_t \mu_{st}(x^c,z^d)v_s v_t + o(|h|^2)\right]\\
&\times \left[F(y|x^c,z^d) + \sum_{s=1}^{q} h_s F_s(y|x^c,z^d)v_s + (1/2)\kappa_2\sum_{s=1}^{q}\sum_{t=1}^{q} h_s h_t F_{st}(y|x^c,z^d)h_s h_t\right.\\
&\left. +(1/2)h_0^2\kappa_2 F_{00}(y|x^c,z^d) + o(h_0^2) + o(|h|^2)\right]W(v)dv\\
&= \mu(x)F(y|x) + \frac{1}{2}\kappa_2\mu(x)h_0^2 F_{00}(y|x) + \frac{1}{2}\kappa_2\sum_{s=1}^{q} h_s^2\,\left[\,\mu(x)F_{ss}(y|x) + 2\mu_s(x)F_s(y|x) + F(y|x)\mu_{ss}(x)\,\right]\\
&+ \sum_{s=1}^{r}\lambda_s\sum_{z^d\in D} I_s(z^d,x^d)\mu(x^c,z^d)F(y|x^c,z^d) + o(h_0^2) + o(|h|^2+|\lambda|).
\end{aligned} \tag{A.6}$$

Therefore, from (A.2) and (A.6) we obtain

$$\begin{aligned}
E[\hat{M}(y|x)] &= E[\hat{F}(y|x)\hat{\mu}(x)] - F(y|x)E[\hat{\mu}(x)]\\
&= (1/2)\kappa_2 h_0^2 F_{00}(y|x) + (1/2)\kappa_2\sum_{s=1}^{q} h_s^2[\mu(x)F_{ss}(y|x) + 2\mu_s(x)F_s(y|x)]\\
&+ \sum_{s=1}^{r}\lambda_s\sum_{z^d\in D} I_s(z^d,x^d)[F(y|x^c,z^d) - F(y|x)]\mu(x^c,z^d) + o(h_0^2) + o(|h|^2+|\lambda|)
\end{aligned} \tag{A.7}$$

and

$$Var(\hat{M}(y|x)) = n^{-1}Var\left[(G\left(\frac{y-Y_i}{h_0}\right) - F(y|x))K_\gamma(X_i,x)\right]$$
$$= n^{-1}\left\{E\left[E\left(\left(G\left(\frac{y-Y_i}{h_0}\right) - F(y|x)\right)^2 |X_i\right)K_\gamma(X_i,x)^2\right] + O(1)\right\}. \qquad (A.8)$$

By Lemma A.5 below we know that

$$E\left\{\left[G\left(\frac{y-Y_i}{h_0}\right) - F(y|x)\right]^2 \middle| X_i\right\} = E\left[G^2\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right] - 2E\left[G\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right]F(y|x) + F(y|x)^2$$
$$= F(y|X_i) - h_0C_wF_0(y|x) - 2F(y|X_i)F(y|x) + F(y|x)^2 + O(h_0^2). \qquad (A.9)$$

Also, by the same calculation used in (A.4) we have

$$E[F(y|X_i)K_\gamma^2(X_i,x)] = \kappa^q(h_1\dots h_q)^{-1}[\mu(x)F(y|x) + O(h_0^2) + O(|h|^2+|\lambda|)]$$
$$E[F_0(y|X_i)K_\gamma^2(X_i,x)] = \kappa^q(h_1\dots h_q)^{-1}[\mu(x)F_0(y|x) + O(h_0^2) + O(|h|^2+|\lambda|)]. \qquad (A.10)$$

(A.8) to (A.10) lead to

$$Var(\hat{M}(y|x)) = \kappa^q(nh_1\dots h_q)^{-1}\mu(x)[F(y|x) - F(y|x)^2 - h_0C_wF_0(y|x)]$$
$$+ O((nh_1\dots h_q)^{-1}(h_0^2 + |h|^2 + |\lambda|) + n^{-1}). \qquad (A.11)$$

□

Lemma A.4. *Under conditions (C1) to (C4),*

$$(nh_1\dots h_q)^{1/2}[\hat{F}(y|x) - F(y|x) - \sum_{s=1}^{q} h_s^2B_{1s}(y|x) - \sum_{s=1}^{r}\lambda_sB_{2s}(y|x)] \to N(0,V) \text{ in distribution,}$$

where $V = \kappa^qF(y|x)[1-F(y|x)]/\mu(x)$.

Proof. This follows from Lemma A.3 and the Liapunov central limit theorem (using arguments similar to those used in the proof of Lemma A.2).

□

Lemma A.5. *Under conditions (C1) to (C4),*

(i) $E\left[G\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right] = F(y|X_i) + (1/2)\kappa_2h_0^2F_{00}(y|X_i) + o(h_0^2)$,

(ii) $E\left[G^2\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right] = F(y|X_i) - h_0C_wF_0(y|X_i) + O(h_0^2)$, *where* $C_w = 2\int G(v)w(v)vdv$.

Proof. ($\int = \int_{-\infty}^{\infty}$)

$$\begin{aligned}
E\left[G\left(\frac{y-Y_i}{h_0}\right)\Big|X_i\right] &= \int G\left(\frac{y-y_i}{h_0}\right) f(y_i|X_i)dy_i = -h_0 \int G(v)f(y-h_0v|X_i)dv \\
&= -\int G(v)dF(y-h_0v|X_i) = 0 + \int w(v)F(y-h_0v|X_i)dv \\
&= \int w(v)\left[F(y|X_i) - h_0F_0(y|X_i)v + (1/2)h_0^2F_{00}(y|X_i)v^2\right]dv + o(h_0^2) \\
&= F(y|X_i) + (1/2)\kappa_2 h_0^2 F_{00}(y|X_i) + o(h_0^2). \qquad \text{(A.12)}
\end{aligned}$$

Also,

$$\begin{aligned}
E\left[G^2\left(\frac{y-Y_i}{h_0}\right)\Big|X_i\right] &= \int G^2\left(\frac{y-y_i}{h_0}\right) f(y_i|X_i)dy_i = -h_0 \int G^2(v)f(y-h_0v|X_i)dv \\
&= -\int G^2(v)dF(y-h_0v|X_i) = 0 + 2\int G(v)w(v)F(y-h_0v|X_i)fv \\
&= 2\int G(v)w(v)[F(y|X_i) - h_0F_0(y|X_i)]dv + o(h_0^2) \\
&= F(y|X_i) - h_0C_wF_0(y|X_i) + O(h_0^2), \qquad \text{(A.13)}
\end{aligned}$$

where $C_w = 2\int G(v)w(v)vdv$ and the fact that $2\int G(v)w(v)dv = \int dG^2(v) = 1$.

□

We now turn to the proof of Theorem 3.1. We adopt an approach similar to that of Cai (2002). We first present a lemma.

Lemma A.6. *Under conditions (C1) to (C4),*

$$\hat{F}(y+\epsilon_n|x) - \hat{F}(y|x) = f(y|x)\epsilon_n + O_p(h_0^2) + o_p(\epsilon_n + (nh_1\dots h_q)^{-1/2}).$$

Proof. Let $A_n(\epsilon_n) = [\hat{F}(y+\epsilon_n|x) - \hat{F}(y|x)]\hat{\mu}(x)/\mu(x)$. Then $\hat{F}(y+\epsilon_n|x) - \hat{F}(y|x) = A_n(\epsilon_n)[1+o_p(1)]$. By Lemma A.13 we have

$$\begin{aligned}
E[A_n(\epsilon_n)] &= E\left\{E\left[G\left(\frac{y+\epsilon_n-Y_i}{h_0}\right) - G\left(\frac{y-Y_i}{h_0}\right)\Big|X_i\right]K_\gamma(X_i,x)\right\}/\mu(x) \\
&= E\{[F(y+\epsilon_n|X_i) - F(y|X_i) + O(h_0^2)]K_\gamma(X_i,x)\}/\mu(x) \\
&= E\{[f(y|X_i)\epsilon_n + O(\epsilon_n^2) + O(h_0^2)]K_\gamma(X_i,x)\} \\
&= \int \mu(x_i)[f(y|x_i)\epsilon_n + O(h_0^2) + O(\epsilon_n^2)]K_\gamma(x_i,x)\}/\mu(x) \\
&= f(y|x)\epsilon_n + O(\epsilon_n^2 + h_0^2).
\end{aligned}$$

Similarly,

$$Var[A_n(\epsilon_n)] \le n^{-1}\mu(x)^{-2}E\left\{\left[G\left(\frac{y+\epsilon_n-Y_i}{h_0}\right) - G\left(\frac{y-Y_i}{h_0}\right)\right]^2 K_\gamma^2(X_i,x)\right\} = O(\epsilon_n^2(nh_1\dots h_q)^{-1}).$$

Therefore,

$$\hat{F}(y+\epsilon_n|x) - \hat{F}(y|x) = A_n(\epsilon_n)[1+o_p(1)] = f(y|x)\epsilon_n + O_p(h_0^2) + o_p(\epsilon_n + (nh_1\dots h_q)^{-1/2}).$$

$\square$

Proof. Proof of Theorem 3.1

In the proof below we will only consider the non-stochastic bandwidth case with $h_0 = a_0^0 n^{-2/(4+q)}$, $h_s = a_s^0 n^{-1/(4+q)}$ $(s = 1, \ldots, q)$ and $\lambda_s = b_s^0 n^{-2/(4+q)}$. By using the tightness/stochastic equicontinuity argument found in Hsiao, Li & Racine (2004), one can show that this result holds true for stochastic bandwidths satisfying (14).

For any v, let $\epsilon_n = B_{n,\alpha}(x) + (nh_1 \ldots h_q)^{-1/2} V_\alpha(x)^{1/2} v$. Then

$$\begin{aligned} Q_\alpha(v) &\stackrel{def}{=} P\left[(nh_1 \ldots h_q)^{1/2} V_\alpha(x)^{-1/2}[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)] \leq v\right] \\ &= P\left[\hat{q}_\alpha(x) \leq q_\alpha(x) + \epsilon_n\right] = P\left[\hat{F}(q_\alpha(x) + \epsilon_n | x) \geq \alpha\right] \\ &= P\left[\hat{F}(q_\alpha(x)|x) \geq -f(q_\alpha(x)|x)\epsilon_n + \alpha\right] + o(1) \end{aligned} \tag{A.14}$$

by Lemma A.6 because $h_0^2 = o((nh_1 \ldots h_q)^{-1/2})$. Therefore, by noting that $f(q_\alpha(x)|x)B_{n,\alpha}(x) = B_n(q_\alpha(x)|x)$, $f(q_\alpha(x)|x)V_\alpha(x)^{1/2} = V(q_\alpha(x)|x)^{1/2}$, and $\alpha = F(q_\alpha(x)|x)$, we have from (A.14) that

$$\begin{aligned} Q_n(v) &= P\left[(nh_1 \ldots h_q)^{1/2} V_\alpha(x)^{-1/2}[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)] \leq v\right] \\ &= P\left[(nh_1 \ldots h_q)^{1/2} V(q_\alpha(x)|x)^{-1/2}\{\hat{F}(q_\alpha(x)|x) - \alpha - B_n(q_\alpha(x)|x)\} \geq -v\right] + o(1) \\ &\to \Phi(v), \end{aligned}$$

by Theorem 2.2, where $\Phi(\cdot)$ is the standard normal distribution. This completes the proof.

$\square$

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

Jeff Racine, Department of Economics, McMaster University, Hamilton, Ontario, Canada L8S 4M4 Email: racinej@mcmaster.ca