Statistica Sinica **14**(2004), 485-512

CROSS-VALIDATED LOCAL LINEAR NONPARAMETRIC REGRESSION

Qi Li and Jeff Racine

Texas A&M University and Syracuse University

Abstract: Local linear kernel methods have been shown to dominate local constant methods for the nonparametric estimation of regression functions. In this paper we study the theoretical properties of cross-validated smoothing parameter selection for the local linear kernel estimator. We derive the rate of convergence of the cross-validated smoothing parameters to their optimal benchmark values, and we establish the asymptotic normality of the resulting nonparametric estimator. We then generalize our result to the mixed categorical and continuous regressor case which is frequently encountered in applied settings. Monte Carlo simulation results are reported to examine the finite sample performance of the local-linear based cross-validation smoothing parameter selector. We relate the theoretical and simulation results to a corrected AIC method (termed AIC_c) proposed by Hurvich, Simonoff and Tsai (1998) and find that AIC_c has impressive finite-sample properties.

Key words and phrases: Asymptotic normality, data-driven bandwidth selection, discrete and continuous data, local polynomial regression.

1. Introduction

There exists a rich body of literature on the estimation of unknown regression functions using kernel weighted local linear methods; see Fan (1992, 1993), Ruppert and Wand (1994), Fan and Gijbels (1995), among others. The local linear estimator has many attractive properties including the fact that it is minimax efficient and is one of the best known approaches for boundary correction. While practitioners often encounter a mix of discrete and continuous data types in applied settings, existing local linear methods do not handle the presence of discrete data in a satisfactory manner. In this paper we propose a new local linear estimator which smooths both the discrete and continuous regressors using the method of kernels. Since it is widely appreciated that data-driven smoothing parameter selection is a necessity in applied nonparametric settings, we propose using least squares cross-validation (CV) for selecting smoothing parameters for both types of regressors. In particular, we derive the rate of convergence of the cross-validated smoothing parameters to their optimal benchmark values, and we establish the asymptotic normality of the resulting nonparametric estimator.

The results contained herein are new even when considering the case for which there exist only continuous regressors.

The CV method is one of the most widely used bandwidth selectors for kernel smoothing, despite the fact that the relative error of the cross-validated bandwidths may be higher than that for some alternative selection methods, for example, the plug-in method. In the presence of discrete regressors, however, the CV method is particularly attractive because it has the ability to automatically remove irrelevant discrete regressors by smoothing them out; see Hall, Racine and Li (2004) for a more detailed discussion on this and related issues. In this paper we explicitly address the case for which each regressor has a unique bandwidth (i.e., the vector-valued smoothing parameter case). This leads to a set of conditions that ensure that cross-validation will lead to optimal smoothing for the local linear kernel estimator, and illustrates how plug-in methods may face some practical problems because it can be difficult to select good initial smoothing parameter values that are required by the plug-in method. We show via simulations that the cross-validated local linear estimator is capable of out-performing the local constant estimator in the presence of mixed data types. We also find that the corrected AIC method proposed by Hurvich, Simonoff and Tsai (1998) has impressive finite-sample properties. After the submission of this paper, a work by Xia and Li (2002) was brought to our attention in which they study the asymptotic behavior of cross-validated bandwidth selection for local polynomial fitting in a time series regression model with a univariate continuous regressor; our paper differs from Xia and Li's in that (i) we consider multivariate regression models and (ii) we allow for the presence of mixed discrete and continuous regressors.

2. Cross-Validation and the Local Linear Estimator: The Continuous Regressor Case

Consider a nonparametric regression model

$$y_j = g(x_j) + u_j, \qquad j = 1, \dots, n, \tag{2.1}$$

where x_j is a continuous random vector of dimension q. Define the derivative of $g(x)$: $\beta(x) \stackrel{def}{=} \nabla g(x) \equiv \partial g(x)/\partial x$ ($\nabla g(\cdot)$ is a $q \times 1$ vector).

Define $\delta(x) = (g(x), \beta(x)')'$, so $\delta(x)$ is a $(q+1) \times 1$ vector-valued function whose first component is $g(x)$ and whose remaining q components are the first derivatives of $g(x)$. Taking a Taylor series expansion of $g(x_j)$ at x_i, we get $g(x_j) = g(x_i) + (x_j - x_i)'\beta(x_i) + R_{ij}$, where $R_{ij} = g(x_j) - g(x_i) - (x_j - x_i)'\beta(x_i)$. We write (2.1) as

$$\begin{aligned} y_j &= g(x_i) + (x_j - x_i)'\nabla g(x_i) + R_{ij} + u_j \\ &= (1, (x_j - x_i)')\delta(x_i) + R_{ij} + u_j. \end{aligned} \tag{2.2}$$

A leave-one-out local linear kernel estimator of $\delta(x_i)$ is obtained by a kernel weighted regression of y_j on $(1,(x_j-x_i)')$ given by

$$\hat{\delta}_{-i}(x_i)=\begin{pmatrix}\hat{g}_{-i}(x_i)\\ \hat{\beta}_{-i}(x_i)\end{pmatrix}$$
$$=\left[\sum_{j\neq i}W_{h,ij}\begin{pmatrix}1, & (x_j-x_i)'\\ x_j-x_i, & (x_j-x_i)(x_j-x_i)'\end{pmatrix}\right]^{-1}\sum_{j\neq i}W_{h,ij}\begin{pmatrix}1\\ x_j-x_i\end{pmatrix}y_j, \quad (2.3)$$

where $W_{h,ij}=\prod_{s=1}^{q}h_s^{-1}w((x_{js}-x_{is})/h_s)$ is the product kernel function and $h_s=h_s(n)$ is the smoothing parameter associated with the sth component of x.

Define a $(q+1)\times 1$ vector e_1 whose first element is one with all remaining elements being zero. The leave-one-out kernel estimator of $g(x_i)$ is given by $\hat{g}_{-i}(x_i)=e_1'\hat{\delta}_{-i}(x_i)$, and we choose $h_1,\ldots,h_q$ to minimize the least-squares cross-validation function given by

$$CV(h_1,\ldots,h_q)=\sum_{i=1}^{n}[y_i-\hat{g}_{-i}(x_i)]^2. \quad (2.4)$$

We use $\hat{h}=(\hat{h}_1,\ldots,\hat{h}_q)$ to denote the cross-validation choices of $h_1,\ldots,h_q$ that minimize (2.4). Having computed $\hat{h}$ we then estimate $\delta(x)$ by

$$\hat{\delta}(x)=\begin{pmatrix}\hat{g}(x)\\ \hat{\beta}(x)\end{pmatrix}$$
$$=\left[\sum_{i=1}^{n}W_{\hat{h},ix}\begin{pmatrix}1, & (x_i-x)'\\ x_i-x, & (x_i-x)(x_i-x)'\end{pmatrix}\right]^{-1}\sum_{i=1}^{n}W_{\hat{h},ix}\begin{pmatrix}1\\ x_i-x\end{pmatrix}y_i,$$

where $W_{\hat{h},ix}=\prod_{s=1}^{q}\hat{h}_s^{-1}w((x_{is}-x_s)/\hat{h}_s)$, and we estimate $g(x)$ by $\hat{g}(x)=e_1'\hat{\delta}(x)$.

The following assumptions are used to establish the convergence of $\hat{h}_1,\ldots,\hat{h}_q$ to their optimal benchmark values and to establish the asymptotic normality of $\hat{g}(x)$.

(A1) (i) (x_i,y_i) are i.i.d. as (X,Y); $\mathcal{S}$, the support of X, is a compact set; $E(y_i|x_i)=g(x_i)$ almost surely; $u_i=y_i-g(x_i)$ has finite 4th moments. (ii) $\inf_{x\in\mathcal{S}}f(x)\geq\epsilon>0$ for some (small) $\epsilon>0$. (iii) $g(x)$, $f(x)$ and $\sigma^2(x)=E(u_i^2|x_i=x)$ are all fourth order differentiable in $\mathcal{S}$. (iv) Letting $g_{ss}(x)$ denote the second order derivative of g with respect to x_s, then $\int g_{ss}(x)^2f(x)dx>0$ for all $s=1,\ldots,q$.

(A2) $w(\cdot):\mathcal{R}\rightarrow\mathcal{R}$ is a bounded symmetric density function with $\int w(v)v^4dv<\infty$, and is m times differentiable. Letting $w^{(s)}(\cdot)$ denote the sth order derivative of $w(\cdot)$, $\int|w^{(s)}(v)v^s|dv<\infty$ for all $s=1,\ldots,m$, where $m>\max\{2+4/q,1+q/2\}$ is a positive integer.

(A3) $(\hat{h}_1, \dots, \hat{h}_q) \in H_n = \{(h_1, \dots, h_q) | (h_1, \dots, h_q) \in [0, \eta]^q$, and $n\hat{h}_1 \cdots \hat{h}_q \geq t_n\}$, where $\eta = \eta(n)$ is a positive sequence that goes to zero slower than the inverse of any polynomial in n, and t_n is a sequence that diverges to $+\infty$.

(A1) (iv) requires that g is not linear in any of its components. The assumption that $h_1, \dots, h_q$ lie in a shrinking set given in (A3) is not as restrictive as it appears, since otherwise the kernel estimator will have a non-vanishing bias term resulting in an inconsistent estimator when the model is nonlinear. We rule out the case for which $g(x)$ is linear in any of its components x_s. The two conditions on H_n in (A3) are similar to those used in Härdle and Marron (1985), and they basically require that $h_s \to 0$ for all s, and $nh_1 \cdots h_q \to \infty$ as $n \to \infty$.

In Appendix A we show that the leading term of the cross-validation function is given by

$$CV_L(h_1, \dots, h_q) = \int \left[\frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) h_s^2 \right]^2 f(x) dx + \frac{B_0}{nh_1 \cdots h_q}, \tag{2.5}$$

where $g_{ss}(x)$ is the second order derivative of g with respect to x_s, $B_0 = \kappa^q \int \sigma^2(x) dx$, $\kappa = \int w(v)^2 dv$ and $\kappa_2 = \int w(v) v^2 dv$.

Define a_s via $h_s = a_s n^{-1/(q+4)}$ for $s = 1, \dots, q$. Then we have $CV_L(h_1, \dots, h_q) = n^{-4/(q+4)} \chi(a_1, \dots, a_q)$, where

$$\chi(a_1, \dots, a_q) = \int \left[\frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) a_s^2 \right]^2 f(x) dx + \frac{B_0}{a_1 \cdots a_q}. \tag{2.6}$$

Let $a_1^0, \dots, a_q^0$ denote values of $a_1, \dots, a_q$ that minimize χ subject to being non-negative. Note that if $a_s^0 = 0$ for some s, then we must have $a_t^0 = \infty$ for some $t \neq s$. Since we have assumed that g is not linear in any of its components, we rule out the case for which $a_t^0 = \infty$ and assume that, for $s = 1, \dots, q$,

$$\text{each } a_s^0 \text{ is uniquely defined and is finite.} \tag{2.7}$$

It is easy to see that (2.7) requires that, for all $s = 1, \dots, q$, $g_{ss}(x)$ does not vanish almost everywhere (our assumption (A3)), for otherwise $a_s^0 = \infty$.

Below we provide a necessary and sufficient condition for (2.7). Let $z_s = a_s^2$ $(s = 1, \dots, q)$, and let A denote a $q \times q$ positive semidefinite matrix having its (t, s)th element given by $A_{t,s} = (\kappa_2/2) \int g_{tt}(x) g_{ss}(x) f(x) dx$. Then (2.6) can be re-written as

$$\chi_z(z_1, \dots, z_q) = z' A z + \frac{B_0}{\sqrt{z_1 \cdots z_q}}, \tag{2.8}$$

where $z = (z_1, \dots, z_q)'$ is a $q \times 1$ vector. Let $z_1^0, \dots, z_q^0$ denote the values of $z_1, \dots, z_q$ that minimize $\chi_z(z_1, \dots, z_q)$ subject to the requirement that each of them be non-negative. Then it is easy to see that each z_s^0 is uniquely defined and is finite if and only if A is a positive definite matrix. A being positive definite ensures that each z_s^0 is finite, for otherwise $z'Az = \infty$ (hence $\chi_z = \infty$). Given that each z_0^s is finite, we must have $z_s^0 > 0$ because otherwise $B_0/(z_1^0 \cdots z_q^0)^{1/2} = \infty$. Thus, each z_s^0 must be positive and finite, which in turn implies that each $a_s^0 = (z_s^0)^{1/2}$ is positive and finite. Thus, (2.7) holds true if and only if A is positive definite.

This condition imposes some restrictions on the second order derivative functions g_{ss} ($s = 1, \ldots, q$), and is more intuitive than (2.7). For example, if $q = 1$, it requires that $g_{11}(x_1)$ is not a 'zero function' (i.e., cannot be equal to zero a.e.). When $q = 2$, it assumes that $g_{ss}(x)$ is not identically zero for $s = 1, 2$, and that $[\int g_{11}(x)^2 dF(x)][\int g_{22}(x)^2 dF(x)] > [\int g_{11}(x) g_{22}(x) dF(x)]^2$ (F is the distribution function of X). This last condition is equivalent to the requirement that $g_{11}(x) - c\, g_{22}(x)$ is not identically zero for any constant c.

While it is easy to obtain a closed form solution for a_0^s from (2.6) for $q = 1, 2$, in the general multivariate q case there do not exist closed form solutions for the a_s^0's ($s = 1, \ldots, q$), even though they are well defined for any values of q. Therefore, a plug-in method based on (2.6) does not possess closed form solutions, and it seems difficult to obtain good initial values for the h_s's ($s = 1, \ldots, q$) that are required by the plug-in method.

We note here that it is important to explicitly allow for different values of h_s for the different components of x_s ($s = 1, \ldots, q$). If one were to use a scalar $h_1 = \cdots = h_q = h$, as is often done to simplify the theoretical derivations (e.g., Racine and Li (2003)), then one would not get the positive definiteness of A. To see this, note that if one were to use $h_1 = \cdots = h_q = h$, and $a_1 = \cdots = a_q = a$, then (2.6) becomes

$$\chi(a) = a^4 \int \left[\frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) \right]^2 f(x) dx + \frac{B_0}{a^q}. \tag{2.9}$$

Therefore, there exists a unique positive and finite a^0 that minimizes $\chi(a)$ if $\sum_{s=1}^q g_{ss}(x)$ is not a zero function. But this condition clearly does not give applied researchers correct guidance as it would assert that h_s converges to zero *even if* $g(x)$ is linear in x_s as long as $g(x)$ is non-linear in some other component such that $\sum_{s=1}^q g_{ss}(x)$ is not a zero function. Since in practice one never forces all h_s's to be the same, (2.9) fails to reveal the correct conditions that ensure (2.7).

Let $h_1^0, \ldots, h_q^0$ denote the values of $h_1, \ldots, h_q$ that minimize (2.5). We have $h_s^0 = a_s^0 n^{-1/(q+4)}$. Also, given the fact that CV_L is the leading term of CV, one can show that $\hat{h}_s = h_s^0 + o_p(h_s^0)$.

Theorem 2.1. *Under* (A1) *through* (A3) *and* (2.7), *we have, for all* $s = 1, \ldots, q$,

$$(\hat{h}_s - h_s^0)/h_s^0 = O_p(n^{-\epsilon/(4+q)}) \quad \textit{with} \quad \epsilon = \min\{q/2, 2\}, \quad \textit{where} \quad h_s^0 = a_s^0 n^{-1/(q+4)}.$$

For results on cross-validated local constant kernel regression, see Härdle, Hall and Marron (1988, 1992), and see Chen (1996) on using extra information for nonparametric smoothing in order to improve efficiency. With our result one can establish the asymptotic normality of $\hat{g}(x)$.

Theorem 2.2. *Under assumptions* (A1) *through* (A3), *and assuming that* $f(x) > 0$, *then*

$$\sqrt{n\hat{h}_1 \cdots \hat{h}_q}\left[\hat{g}(x) - g(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} g_{ss}(x)\hat{h}_s^2\right] \to N(0, \Omega_x) \quad \textit{in distribution,}$$

where $\Omega_x = \kappa^q \sigma^2(x)/f(x)$.

Under the assumption that $g(\cdot)$ is a smooth function with non-vanishing second-order derivatives, Theorems 2.1 and 2.2 show that $\hat{h}_s$ converges to zero at the rate $O_p(n^{-1/(4+q)})$ and that $\hat{g}(x)$ converges to $g(x)$ at the rate $O_p(n^{-2/(4+q)})$. In practice, some regression functions $g(\cdot)$ may have a linear regression functional form or be linear in some of their components (a partially linear specification). Our Theorem 2.1 does not cover such cases. However, it can be shown that in the case for which g is linear in some of its components, say x_s, then the cross-validation smoothing parameter $\hat{h}_s$ will tend to take large numerical values, indicating that the model is partially linear. Note that our Theorem 2.1 does not cover a partially linear model since for a partially linear model, $g_{ss}(x) = 0$ for some $s \in \{1, \ldots, q\}$, and Assumption (A1) (iv) is violated. In Section 3, we use simulations to investigate the distribution of $\hat{h}_s$ in this case. Our results explain how the use of cross-validated local linear kernel methods in empirical settings may result in large smoothing parameters for some regressors and small smoothing parameters for others, a feature often exhibited in applied settings.

Up to now, we have restricted attention to the use of least squares cross-validation (CV) when selecting smoothing parameters. Härdle, Hall and Marron (1988) have shown that, for the local-constant estimator, the CV smoothing parameter selectors are asymptotically equivalent to generalized CV (GCV) selectors, which include Akaike's (1974) information criterion, Shibata's (1981) model selector, and Rice's (1984) T selector, among others. It can easily be shown that the same conclusions hold true for the local linear method, that is, that the local-linear based CV smoothing parameter selector is asymptotically equivalent to the local-linear based GCV selector. This follows the *exact* same proof as in Härdle, Hall and Marron (1988, p.95) as local-constant and local-linear estimators have the same rate of convergence (when both use a second order kernel).

Recently, Hurvich, Simonoff and Tsai (1998) suggested a corrected (improved) AIC criterion (termed AIC_c) as a smoothing parameter selector, and their simulations show that the AIC_c selector performs quite well compared with the plug-in method (when it is available) and with a number of generalized CV methods. While there is no theoretical result available for the AIC_c selector, we conjecture that the AIC_c selector is asymptotically equivalent to the (generalized) CV method, and simulation results are consistent with this conjecture. We find that, for small samples, AIC_c tends to perform better than the CV method, while for large samples there is no appreciable difference between the two methods.

Härdle, Hall and Marron (1988) also consider the intermediate benchmark case of selecting h_s's by minimizing the average square error given by $ASE = n^{-1}\sum_i[\hat{g}(x_i) - g(x_i)]^2$ for the univariate x case. They use $\hat{h}^0$ to denote the values of h that minimize ASE, and they further show that $\hat{h} - \hat{h}^0 = O_p(n^{-1/10}\hat{h}_0) = O_p(n^{-3/10})$. This is the same rate as for $\hat{h} - h_0$ stated in our Theorem 2.1 for $q = 1$. We conjecture that Theorem 2.1 holds true when one replaces h_s^0 by $\hat{h}_s^0$. This is because one can show that $CV = ASE + O_p(\eta_2^3 + \eta_1(h_1 \cdots h_q)^{1/2}) = ASE + O_p(ASE)O_p(\eta_2 + (h_1 \cdots h_q)^{1/2})$, where $\eta_2 = \sum_{s=1}^q h_s^2$ and $\eta_1 = (nh_1 \cdots h_q)^{-1}$. From this we expect that $\hat{h}_s = \hat{h}_s^0 + O_p(\hat{h}_s^0)O_p((h_s^0)^{\min\{2,q/2\}})$, or equivalently that $\hat{h}_s - \hat{h}_s^0 = O_p(n^{-1/(q+4)}n^{-\min[2,q/2]/(q+4)})$. However, a rigorous proof of this result lies beyond the scope of this paper.

3. Local Linear Cross Validation with Mixed Continuous and Discrete Regressors

In this section we consider the case where a subset of regressors are categorical and the remaining are continuous. Although it is well known that one can use a nonparametric frequency method to handle the discrete regressors (theoretically), such an approach cannot be used in practice if the number of discrete cells is large relative to the sample size, as is often the case with economic data sets containing mixed data types. Borrowing from Aitchison and Aitken's (1976) approach, we elect to smooth the discrete regressors to circumvent this problem; see Hall (1981), Grund and Hall (1993), and the monographs by Scott (1992) and Simonoff (1996) for further discussion on the kernel smoothing of discrete variables.

Let x_i^d denote a $r \times 1$ vector of regressors that assume discrete values and let $x_i^c \in R^q$ denote the remaining continuous regressors. It should be mentioned that Ahmad and Cerrito (1994) and Bierens (1983, 1987) also consider the case of estimating a regression function with mixed categorical and continuous regressors, but they did not study the theoretical properties associated with using data-driven methods (cross-validation) when selecting smoothing parameters. Furthermore, both works only consider the local constant kernel estimator.

For a discrete regressor, we use a variation on Aitchison and Aitken's (1976) kernel function defined by

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

The range of λ_s is [0,1]. Note that when $\lambda_s = 0$ the above kernel function becomes an indicator function, and when $\lambda_s = 1$, it is a constant function. That is, the x_s^d regressor is removed (smoothed out) if $\lambda_s = 1$. Let $\mathbf{1}(A)$ denote an indicator function which assumes the value 1 if A holds true and 0 otherwise. Then the product kernel function for a vector of discrete regressors is given by

$$L(x_i^d, x_j^d, \lambda) = \left[\prod_{s=1}^{r} \lambda_s^{1-\mathbf{1}(x_{is}^d = x_{js}^d)} \right].$$

Now define the partial derivative of $g(x) = g(x^c, x^d)$ with respect to x^c: $\beta(x) \stackrel{def}{=} \nabla g(x) \equiv \partial g(x^c, x^d)/\partial x^c$, and define $\delta(x) = (g(x), \beta(x)')'$. Also, we use the short-hand notation $K_{h,ij} = W_{h,ij} L_{\lambda,ij}$, where $W_{h,ij} = \prod_{s=1}^{q} h_s^{-1} w((x_{is}^c - x_{js}^c)/h_s)$ and $L_{\lambda,ij} = \prod_{s=1}^{r} l(x_{is}^d, x_{js}^d, \lambda_s)$. Then the leave-one-out kernel estimator of $\delta(x_i) \equiv \delta(x_i^c, x_i^d)$ is given by

$$\begin{aligned} \hat{\delta}_{-i}(x_i) &= \begin{pmatrix} \hat{g}_{-i}(x_i) \\ \hat{\beta}_{-i}(x_i) \end{pmatrix} \\ &= \left[\sum_{j \neq i} K_{h,ij} \begin{pmatrix} 1, & (x_j^c - x_i^c)' \\ x_j^c - x_i^c, & (x_j^c - x_i^c)(x_j^c - x_i^c)' \end{pmatrix} \right]^{-1} \sum_{j \neq i} K_{h,ij} \begin{pmatrix} 1 \\ x_j^c - x_i^c \end{pmatrix} y_j. \quad (3.1) \end{aligned}$$

Note that (3.1) treats the continuous regressor x^c in a local linear fashion and the discrete regressor x^d in a local constant one. Again, $\hat{g}_{-i}(x_i) = e_1' \hat{\delta}_{-i}(x_i)$ ($e_1 = (1, 0, \ldots, 0)'$), and we choose (h, λ) to minimize

$$CV(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \hat{g}_{-i}(x_i) \right]^2. \quad (3.2)$$

We use $(\hat{h}_1, \ldots, \hat{h}_q, \hat{\lambda}_1, \ldots, \hat{\lambda}_r)$ to denote values of $(h_1, \ldots, h_q, \lambda_1, \ldots, \lambda_r)$ that minimize (3.2). We then estimate $g(x)$ by $\hat{g}(x) = e_1' \hat{\delta}(x)$, where

$$\begin{aligned} \hat{\delta}(x) &= \begin{pmatrix} \hat{g}(x) \\ \hat{\beta}(x) \end{pmatrix} \\ &= \left[\sum_{i} K_{\hat{h},ix} \begin{pmatrix} 1, & (x_i^c - x^c)' \\ x_i^c - x^c, & (x_i^c - x^c)(x_i^c - x^c)' \end{pmatrix} \right]^{-1} \sum_{i} K_{\hat{h},ix} \begin{pmatrix} 1 \\ x_i^c - x^c \end{pmatrix} y_i, \end{aligned}$$

with $K_{\hat{h},ix} = \prod_{s=1}^{q} \hat{h}_s^{-1} w\left((x_{is}^c - x_s^c)/\hat{h}_s\right) \prod_{s=1}^{r} l(x_{is}^d, x_s^d, \hat{\lambda}_s)$.

In Appendix B we show that the leading term of the cross-validation function is given by

$$CV_L(h,\lambda) = \sum_{x^d} \int \left\{ \frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) h_s^2 + \sum_{s=1}^{r} D_s(x)\lambda_s \right\}^2 f(x)dx^c + \frac{B_0}{nh_1\cdots h_q}, \tag{3.3}$$

where $g_{ss}(x)$ is the second order derivative of $g(x)$ with respect to x_s^c, $B_0 = \kappa^q \sum_{x^d} \int \sigma^2(x)\, dx^c$, $D_s(x) = \sum_{v^d} [\mathbf{1}_s(v^d, x^d)\, g(x^c, v^d) - g(x)]\, f(x^c, v^d)$ with $\mathbf{1}_s(x^d, v^d) = \mathbf{1}(x_s^d \neq v_s^d) \prod_{t\neq s} \mathbf{1}(x_t^d = v_t^d)$, and $\mathbf{1}_s(x^d, v^d) = 1$ if x^d and v^d differs only in the sth component, and is 0 otherwise.

Define $a_1, \ldots, a_q, b_1, \ldots, b_r$ via $h_s = a_s n^{-1/(q+4)}$ ($s = 1, \ldots, q$) and $\lambda_s = b_s n^{-2/(q+4)}$ ($s = 1, \ldots, r$). Then we have $CV_L(h,\lambda) = \chi(a,b)$, where

$$\chi(a,b) = \sum_{x^d} \int \left\{ \frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) a_s^2 + \sum_{s=1}^{r} D_s(x) b_s \right\}^2 f(x)dx^c + \frac{B_0}{h_1\cdots h_q}.$$

Letting $(a_1^0, \ldots, a_q, b_1^0, \ldots, b_r^0)$ denote the values of $(a_1, \ldots, a_q, b_1, \ldots, b_r)$ that minimize $\chi(a,b)$ subject to the restriction that they are non-negative, we further assume that

$$\text{each of the } a_s^0\text{'s and } b_s^0\text{'s is uniquely defined and is finite.} \tag{3.4}$$

Let $h_1^0, \ldots, \lambda_r^0$ denote the values of $h_1, \ldots, \lambda_r$ that minimize (3.3). Then obviously we have $n^{1/(q+4)} h_s^0 \sim a_s^0$ for $s = 1, \ldots, q$, and $n^{2/(q+4)} \lambda_s^0 \sim b_s^0$ for $s = 1, \ldots, r$. In Appendix B we show that $\hat{h}_s = h_s^0 + o_p(h_s^0)$ for $s = 1, \ldots, q$, and that $\hat{\lambda}_s = \lambda_s^0 + o_p(\lambda_s^0)$ for $s = 1, \ldots, r$.

Theorem 3.1. *Under* (B1) *and* (B2) *given in Appendix* B, *and* (3.4), *we have* $(\hat{h}_s - h_s^0)/h_s^0 = O_p(n^{-\epsilon_1/(4+q)})$ *for* $s = 1, \ldots, q$, *and* $\hat{\lambda}_s - \lambda_s^0 = O_p(n^{-\epsilon_2})$ *for* $s = 1, \ldots, r$, *where* $\epsilon_1 = \min\{q/2, 2\}$, $\epsilon_2 = \min\{1/2, 4/(4+q)\}$.

Combining Theorem 3.1's rate of convergence result with a Taylor expansion argument, it is easy to establish the asymptotic normal distribution of $\hat{g}(x)$ as the next theorem shows. The argument is sketched in Appendix B.

Theorem 3.2. *Under the conditions of Theorem* 3.1, *we have*

$$\sqrt{n\hat{h}_1\cdots\hat{h}_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{2}(\kappa_2/2) g_{ss}(x)\hat{h}_s^2 - \sum_{s=1}^{r}\hat{\lambda}_s D_s(x)\right) \to N(0, \Omega_x)$$

in distribution,

where $D_s(x) = \sum_{v^d}[\mathbf{1}_s(v^d, x^d)g(x^c, v^d) - g(x)]f(x^c, v^d)$, $\Omega_x = \kappa^q \sigma^2(x)/f(x)$.

From the discussion found in Section 2 we know that when $g(x)$ is linear in x_s^c, $\hat{h}_s$ will not converge to zero, rather, $\hat{h}_s$ will tend to take large values. Similarly, if $g(x)$ turns out to be unrelated to x_s^d (x_s^d is an irrelevant regressor), it can be shown that $\hat{\lambda}_s$ will not converge to zero, rather, it will tend to the upper bound value of 1. The theoretical results presented in this section do not cover these cases. We rely on some simulation exercises to examine the finite sample behavior of $\hat{h}_s$ and $\hat{\lambda}_s$ when $g(x)$ is linear in x_s^c and/or is unrelated to x_s^d.

The Ordered Categorical Regressor Case

Up to now we have only considered the case for which x^d is unordered. If x_s^d is an ordered regressor, we use the following kernel function:

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

The range of λ_s is $[0, 1]$. Again when $\lambda_s = 0$, $(lx_{is}^d, x_{js}^d, \lambda_s = 0)$ becomes an indicator function, and when $\lambda_s = 1$, $l(x_{is}^d, x_{js}^d, \lambda_s = 1) = 1$ is a uniform weight function.

It is easy to show that the results of Theorems 3.1 and 3.2 remain valid provided we redefine $\mathbf{1}_s(v^d, x^d)$ by $\mathbf{1}_s(v^d, x^d) = \mathbf{1}(|x_s^d - v_s^d| = 1) \prod_{t \neq s} \mathbf{1}(x_t^d = v_t^d)$ when x_s^d is an ordered regressor.

4. Monte Carlo Results

In this section we examine the finite-sample behavior of cross-validated local linear regression in the presence of mixed data types. In particular, we consider three data generating processes (DGPs), one that is nonlinear in the continuous regressors, one that is linear, and one that lies in-between (i.e., is partially linear). These are given by

$$\text{DGP}_1\text{: } y_i = 1 + z_{i1} + z_{i2} + x_{i1}x_{i2} + \sin(2\pi x_{i1}) + \sin(2\pi x_{i2}) + u_i,$$

$$\text{DGP}_2\text{: } y_i = 1 + z_{i1} + z_{i2} + x_{i1} + x_{i2} + x_{i1}x_{i2} + 2\sin(2x_{i2}) + u_i,$$

$$\text{DGP}_3\text{: } y_i = 1 + z_{i1} + z_{i2} + x_{i1} + x_{i2} + u_i,$$

where $X_1 \sim N(0,1)$, $X_2 \sim N(0,1)$, $Z_1 \in \{0,1\}$ with $Pr[Z_1 = 1] = 0.4$, $Z_2 \in \{0,1\}$ with $Pr[Z_2 = 1] = 0.6$, and $U \sim N(0, \sigma^2)$ with $\sigma = 1.0$. We compare the performance of six estimators: LIN: OLS assuming linearity and no interaction; DGP: OLS based upon the correct DGP; LL(CV): Cross-validated local linear regression; LL(AIC_c): AIC_c local linear regression; LC(CV): Cross-validated local constant regression; LC(AIC_c): AIC_c local constant regression.

For the nonparametric LL and LC estimators we employ both the proposed cross-validation method and the corrected AIC method proposed by Hurvich, Simonoff and Tsai (1998) to select the smoothing parameters. We conduct five restarts of the multidimensional numerical search algorithm, each time beginning from different random initial bandwidth values, retaining those bandwidths that resulted in a minimum over the five restarts in an attempt to avoid local minima. A second-order Gaussian kernel is used for the continuous regressors.

For each DGP, 1,000 Monte Carlo replications having estimation samples of size $n_1 = (100, 200, 400)$ and independent evaluation samples of size $n_2 = 1,000$ are drawn. For each Monte Carlo replication, the models are estimated on n_1 observations drawn from a given DGP, and then predictions $(\hat{g}(x_i))$ are generated based upon the regressors in the evaluation sample of size n_2. The mean square estimation error is computed as $MSEE = (1/n_2) \sum_{i=1}^{n_2} (g(x_i) - \hat{g}(x_i))^2$, where $g(x_i)$ is the systematic component of the true DGP.

We consider two cases, one for which all regressors are relevant, and one for which the discrete regressor Z_2 is in fact irrelevant (we remove Z_2 from the DGP). The median MSEEs for each estimated model taken over the 1,000 Monte Carlo replications are tabulated along with their interquartile ranges. The parametric model based on the true DGP is, of course, expected to perform the best (it serves as a benchmark), and we focus attention upon the relative performance of the remaining methods.

4.1. Out-of-sample MSEE results: all regressors relevant

Tables 1 through 3 present the median MSEEs for each estimated model along with their interquartile ranges. An examination of these tables reveals the consistent nature of the cross-validated nonparametric estimators via a reduction in their medians and interquartile ranges for MSEE as the sample size increases.

Note that DGP_1 is the 'most nonlinear' one, DGP_2 is partially linear, while DGP_3 is fully linear. An examination of Table 1 reveals that, for DGP_1 (the most nonlinear DGP), the local linear AIC_c estimator performs the best in small samples, but for larger samples ($n > 200$), the cross-validated local constant estimator outperforms all others. As expected, the misspecified linear model performs worst overall. Table 2 reveals that, for the partially linear DGP_2 (nonlinear in X_2), the local linear estimators outperform the local constant estimators, the local linear AIC_c estimator performs the best while, as the sample size increases, the performance of the cross-validated local linear estimator and the local linear AIC_c estimator become indistinguishable from one another. For both of these DGPs, the misspecified parametric model is inconsistent.

Table 1. DGP_1 median MSEE results, all regressors relevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC_c)	LC(CV)	LC(AIC_c)	Linear
100	1.82	1.63	1.79	1.75	2.17
	(1.57,2.43)	(1.49,1.87)	(1.61,1.98)	(1.61,1.93)	(2.06,2.31)
200	1.38	1.32	1.34	1.39	2.08
	(1.20,1.81)	(1.16,1.74)	(1.24,1.46)	(1.29,1.52)	(2.00,2.17)
400	1.06	0.98	0.94	1.09	2.05
	(0.92,1.35)	(0.88,1.24)	(0.87,1.03)	(1.03,1.17)	(1.97,2.13)

Table 2. DGP_2 median MSEE results, all regressors relevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC_c)	LC(CV)	LC(AIC_c)	Linear
100	0.84	0.66	1.01	0.99	2.95
	(0.63,1.27)	(0.52,0.87)	(0.86,1.20)	(0.86,1.17)	(2.79,3.11)
200	0.44	0.37	0.64	0.65	2.82
	(0.34,0.61)	(0.30,0.48)	(0.56,0.74)	(0.57,0.74)	(2.70,2.96)
400	0.23	0.20	0.41	0.42	2.76
	(0.19,0.33)	(0.17,0.27)	(0.37,0.47)	(0.37,0.48)	(2.65,2.88)

Table 3. DGP_3 median MSEE results, all regressors relevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC_c)	LC(CV)	LC(AIC_c)	Linear
100	0.15	0.13	0.44	0.42	0.04
	(0.11,0.23)	(0.09,0.17)	(0.35,0.54)	(0.34,0.52)	(0.03,0.07)
200	0.07	0.06	0.27	0.27	0.02
	(0.05,0.09)	(0.05,0.08)	(0.23,0.32)	(0.23,0.32)	(0.01,0.03)
400	0.03	0.03	0.17	0.17	0.01
	(0.02,0.05)	(0.02,0.04)	(0.15,0.20)	(0.15,0.20)	(0.01,0.02)

From Table 3 we see that, for the linear DGP, DGP_3, the local linear estimators outperform the local constant estimators to a greater extent than was the case for DGP_2. Also, the relative performance of the cross-validated and the AIC_c local linear methods become indistinguishable from one another as n gets large.

The cross-validated and the AIC_c local linear estimators perform quite well for partially linear and linear specifications even in these small-sample settings.

Next we turn to the case when there exist irrelevant regressors.

4.2. Out-of-sample MSEE results: Z_2 irrelevant

We now consider the case where one of the discrete regressors, Z_2, is in fact irrelevant. In this case, both the cross-validation and the AIC_c methods can automatically remove such regressors by assigning them a large value of $\hat{\lambda}_2$, the associated bandwidth. We base this simulation upon the same DGPs given above. However, now Z_2 is, in fact, irrelevant and is removed from the DGP when we generate Y. Furthermore, we do not assume that this information is known *a priori*. Therefore, Z_2 is still used for estimating the conditional mean of Y. MSEE results are presented in Tables 4 through 6.

Tables 4 through 6 illustrate that, in the presence of an irrelevant discrete regressor, the cross-validated local linear (constant) estimator and the AIC_c local linear (constant) estimator display behavior similar to the case for which all regressors are relevant. The cross-validated local constant estimator outperforms the cross-validated and AIC_c local linear estimators for the nonlinear DGP_1 for $n > 200$, while Tables 5 and 6 reveal that, for the partially linear DGP_2 (nonlinear in X_2) and the linear DGP_3, the local linear estimators outperform the local constant estimators.

Table 4. DGP_1 median MSEE results, Z_2 irrelevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC_c)	LC(CV)	LC(AIC_c)	Linear
100	1.66	1.49	1.57	1.57	2.17
	(1.40,2.27)	(1.33,1.84)	(1.43,1.78)	(1.42,1.73)	(2.06,2.29)
200	1.24	1.16	1.17	1.21	2.09
	(1.06,1.71)	(0.99,1.60)	(1.07,1.28)	(1.12,1.33)	(2.00,2.19)
400	0.96	0.88	0.77	0.90	2.04
	(0.81,1.31)	(0.78,1.14)	(0.70,0.84)	(0.80,1.00)	(1.96,2.13)

Table 5. DGP_2 median MSEE results, Z_2 irrelevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC_c)	LC(CV)	LC(AIC_c)	Linear
100	0.68	0.50	0.82	0.79	2.94
	(0.47,1.03)	(0.37,0.72)	(0.69,1.00)	(0.67,0.95)	(2.77,3.10)
200	0.34	0.26	0.51	0.49	2.81
	(0.24,0.51)	(0.20,0.39)	(0.44,0.61)	(0.42,0.58)	(2.68,2.95)
400	0.17	0.14	0.32	0.32	2.77
	(0.13,0.26)	(0.11,0.20)	(0.28,0.38)	(0.27,0.37)	(2.64,2.88)

Table 6. DGP_3 median MSEE results, Z_2 irrelevant (interquartile range in parentheses).

	Nonparametric				Parametric
n_1	LL(CV)	LL(AIC$_c$)	LC(CV)	LC(AIC$_c$)	Linear
100	0.10	0.07	0.33	0.30	0.05
	(0.06,0.17)	(0.05,0.11)	(0.26,0.42)	(0.24,0.37)	(0.03,0.07)
200	0.04	0.04	0.20	0.19	0.02
	(0.03,0.07)	(0.02,0.05)	(0.17,0.25)	(0.16,0.23)	(0.01,0.03)
400	0.02	0.02	0.13	0.12	0.01
	(0.01,0.03)	(0.01,0.03)	(0.11,0.15)	(0.10,0.14)	(0.01,0.02)

Next we consider the behavior of the cross-validated bandwidths. We expect that the local linear cross-validation method will tend to select a large bandwidth when the underlying DGP is in fact linear in a given continuous regressor, while the local constant estimator will display no such tendencies. For the cross-validated bandwidths for the irrelevant regressor Z_2, we have postulated that both the local linear and local constant cross-validation methods will tend to select a large bandwidth for an irrelevant discrete regressor, i.e., choose $\hat{\lambda}_2$ that tends toward its upper bound value of 1.

In an attempt to verify the above conjectures, we plot histograms of the cross-validated bandwidths for X_1 (h_1), X_2 (h_2), Z_1 (λ_1) and Z_2 (λ_2) for $n_1 = 200$ for DGP_2. The results are presented in Figures 1 and 2, while Figures 3 and 4 present comparable numbers for the AIC$_c$ approach.

The histogram on the upper left of each figure summarizes the bandwidths for X_1, the one on the upper right summarizes those for X_2, the one on the lower left summarizes those for Z_1, while that on the lower right summarizes those for Z_2. The uppermost histograms in Figure 1 (the partially linear DGP2) reveal how the local linear cross-validation method chooses much larger smoothing parameters for a continuous regressor that enters linearly (X_1) than for one that enters nonlinearly (X_2). In contrast, Figure 2 shows that the local constant cross-validation choices of $\hat{h}_1$ and $\hat{h}_2$ both assume (relatively) small values. Similar results hold when bandwidth choice is conducted via the AIC$_c$ approach.

While both the local linear and local constant cross-validation methods select small values of $\hat{\lambda}_1$, Figures 1 and 2 show that their choices of $\hat{\lambda}_2$ tend to assume large values close to their upper bound value of 1, thereby effectively removing the irrelevant regressor Z_2 from the nonparametric estimate. This 'automatic removal of irrelevant discrete variables' property is an appealing feature of the cross-validation method in applied settings. Similar results hold when bandwidth choice is conducted via the AIC$_c$ approach.

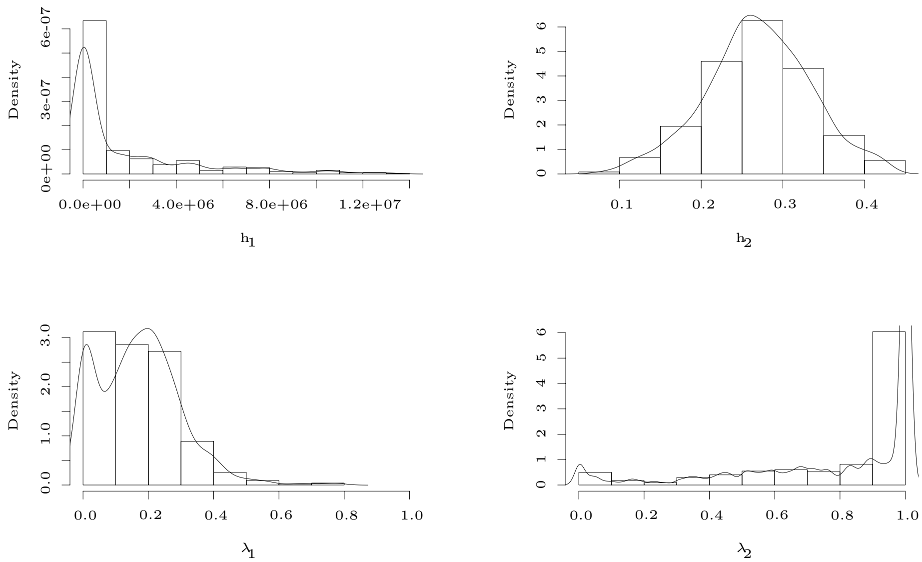

Figure 1. Histograms of LL(CV) smoothing parameters for DGP2, $n = 200$, Z_2 irrelevant.

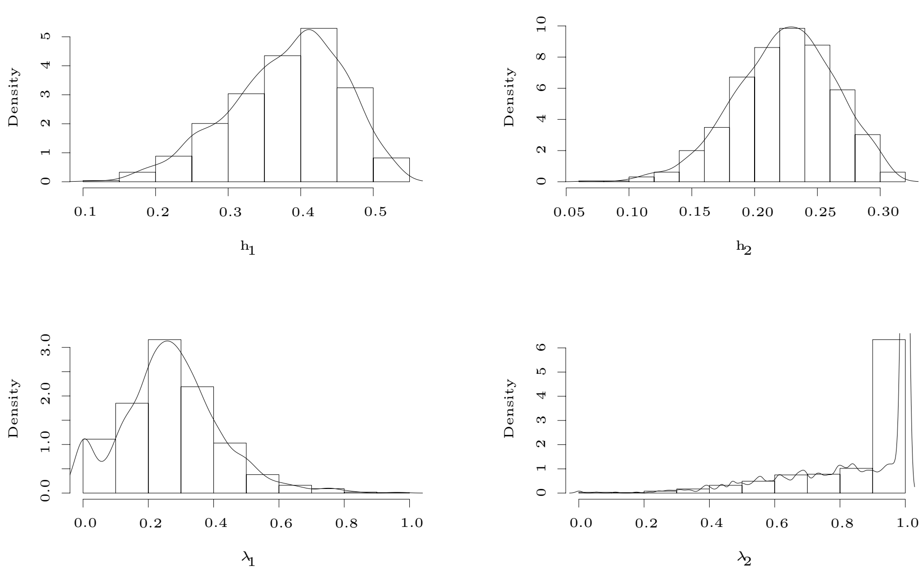

Figure 2. Histograms of LC(CV) smoothing parameters for DGP2, $n = 200$, Z_2 irrelevant.

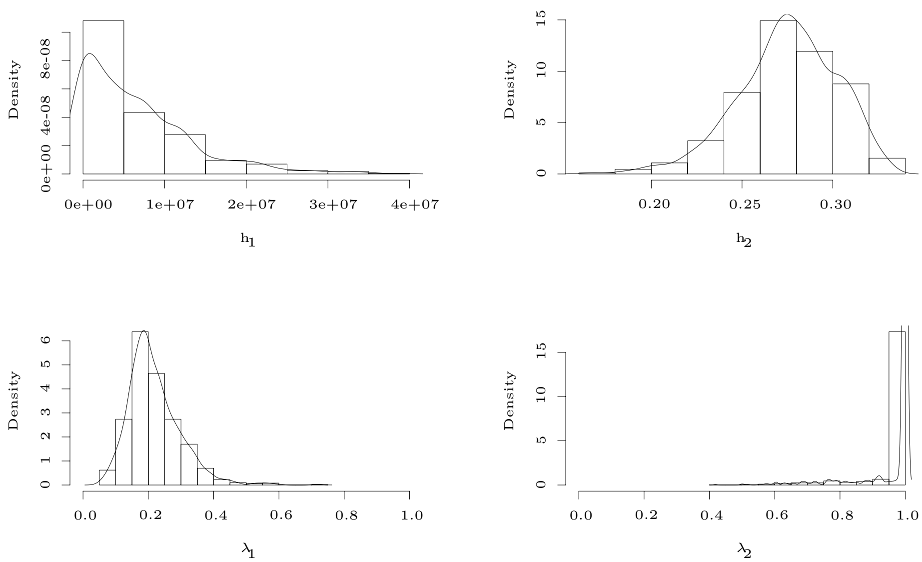

Figure 3. Histograms of LL(AIC$_c$) smoothing parameters for DGP2, $n = 200$, Z_2 irrelevant.

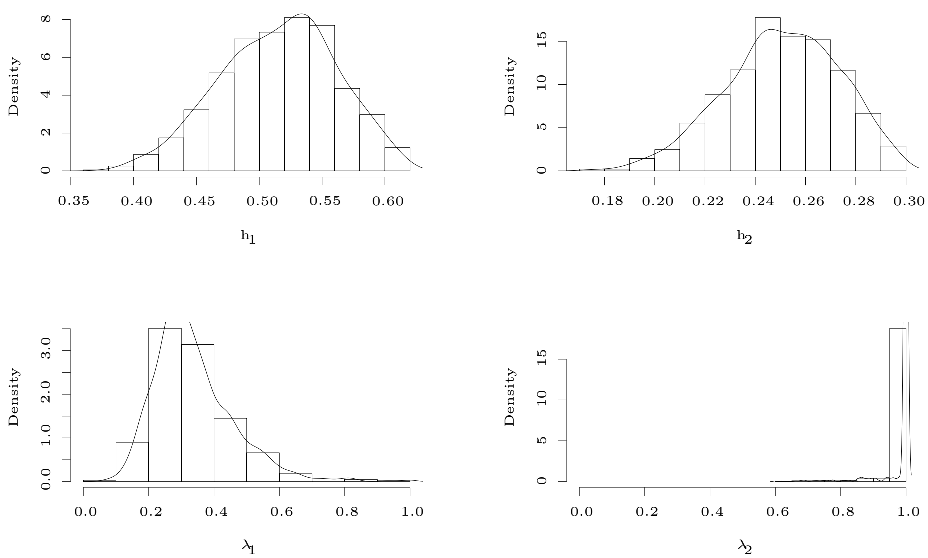

Figure 4. Histograms of LC(AIC$_c$) smoothing parameters for DGP2, $n = 200$, Z_2 irrelevant.

Here we only report simulation results for the two-continuous regressor case. Simulations not repeated here also show that, for a model with more than one regressor entering the model linearly or being close to linear, the local linear cross-validation method provides even larger relative efficiency gains over the local constant method.

As noted in Section 2, Härdle, Hall and Marron (1988) demonstrated that, for the local constant estimator, CV smoothing parameter selectors are asymptotically equivalent to GCV selectors. We have included results based on the Hurvich, Simonoff and Tsai's (1998) AIC_c bandwidth selection criterion in Tables 1 through 6 which reveal that this approach indeed appears to be asymptotically equivalent to the CV method, and has excellent finite sample performance. We leave theoretical investigations of the AIC_c method (such as verifying our conjecture that AIC_c is asymptotically equivalent to the CV method) as a topic for future research.

5. Concluding Remarks

In this paper we present theoretical and simulation-based evidence in support of using data-driven methods such as cross-validation and AIC_c when choosing smoothing parameters for the local linear kernel estimator in the presence of mixed discrete and continuous data types. We find that the AIC_c approach has impressive finite-sample properties. We demonstrate that efficiency gains relative to the local constant estimator are not only theoretically possible but can be readily attained in finite-sample settings. The results presented in this paper also explain the observations of Li and Racine (2001) who found that nonparametric estimators with smoothing parameters chosen via cross-validation can yield superior predictions relative to commonly used parametric methods for U.S. patent application data, Spanish consumption data and U.S. and Swedish labor force participation data.

Acknowledgement

We would like to thank an anonymous referee, an associate editor, and a co-editor for their insightful comments which led to a substantially revised and improved version of the paper. An earlier version of this paper addressed only scalar smoothing parameter settings, while the current version deals with the practically important vector-valued smoothing parameter case which arose out of a suggestion from the associate editor. The nice small sample performance of the AIC_c method was suggested to us by a referee. Li's Research is partially supported by the Private Enterprise Research Center, Texas A&M University. Racine would like to thank the Center for Policy Research at Syracuse University for their ongoing support.

A. Proofs of Theorem 2.1 and 2.2

We will use the notation $A_n \sim B_n$ to denote that A_n has the same probability order as B_n. To simplify the proof, we first re-write (2.3) in an equivalent form. Define D_h^{-2}, a $q \times q$ diagonal matrix with its sth diagonal element given by h_s^{-2}, i.e., $D_h^{-2} = diag(h_s^{-2})$. Inserting the identity matrix $I_{q+1} = G_n^{-1} G_n$ into the middle of (2.3), where $G_n = \begin{pmatrix} 1, & 0 \\ 0, & D_h^{-2} \end{pmatrix}$, we get

$$\begin{aligned}
\hat{\delta}_{-i}(x_i) &= \Big[\sum_{j \neq i} W_{h,ij} G_n \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} (1, (x_j - x_i)') \Big]^{-1} \sum_{j \neq i} W_{h,ij} G_n \begin{pmatrix} 1 \\ x_j - x_i \end{pmatrix} y_j \\
&= \Big[\sum_{j \neq i} W_{h,ij} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} (1, (x_j - x_i)') \Big]^{-1} \\
&\quad \times \sum_{j \neq i} W_{h,ij} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} y_j. \qquad \text{(A.1)}
\end{aligned}$$

The advantage of using (A.1) in the proof is that $(1/n) \sum_{j \neq i} W_{h,ij} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} (1, (x_j - x_i)')$ converges in probability to a non-singular matrix. Hence, we can analyze the denominator and numerator of (A.1) separately and thus simplify the derivations.

Substituting the Taylor expansion (2.2) into (A.1), we have

$$\begin{aligned}
\hat{\delta}_{-i}(x_i) &= \delta(x_i) + \left[\frac{1}{n} \sum_{j \neq i} W_{h,ij} \begin{pmatrix} 1, & (x_j - x_i)' \\ D_h^{-2}(x_j - x_i), & D_h^{-2}(x_j - x_i)(x_j - x_i)' \end{pmatrix} \right]^{-1} \\
&\quad \times \Big\{ \frac{1}{n} \sum_{i} W_{h,ij} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} [R_{ij} + u_j] \Big\} \\
&\equiv \delta(x_i) + A_{2i}^{-1} A_{1i}, \\
A_{1i} &= \frac{1}{n} \sum_{j \neq i} W_{h,ij} \begin{pmatrix} 1 \\ D_h^{-2}(x_j - x_i) \end{pmatrix} [R_{ij} + u_j], \\
A_{2i} &= \begin{pmatrix} \hat{f}_i, & B_{1i}' \\ D_h^{-2} B_{1i}, & D_h^{-2} B_{2i} \end{pmatrix},
\end{aligned}$$

where $\hat{f}_i = n^{-1} \sum_{j \neq i} W_{h,ij}$, $B_{1i} = n^{-1} \sum_{j \neq i} W_{h,ij}(x_j - x_i)$, and $B_{2i} = n^{-1} \sum_{j \neq i} W_{h,ij}(x_j - x_i)(x_j - x_i)'$. It is easy to show $B_{1i} = O_p(\eta_2)$ and $B_{2i} = O_p(\eta_2)$ ($\eta_2 = \sum_{s=1}^{q} h_s^2$). Thus $D_h^{-2} B_{1i}$ and $D_h^{-2} B_{2i}$ are both $O_p(1)$ random variables.

Recall that e_1 is a $(q+1)$ column vector whose first element is one with all other elements being zero. Using the partitioned inverse, we have $e_1'\{A_{2i}\}^{-1} = (\hat{f}_i^{-1} + C_{1i}, -C_{2i})$, where $C_{1i} = \hat{f}_i^{-2} B_{1i}'[D_h^{-2}(B_{2i} - B_{1i}B_{1i}'\hat{f}_i^{-1})]^{-1} B_{1i}$, and $C_{2i} = \hat{f}_i^{-1} B_{1i}'[D_h^{-2}(B_{2i} - B_{1i}B_{1i}'\hat{f}_i^{-1})]^{-1}$. Note that both C_{1i} and C_{2i} are $O_p(\eta_2)$ random variables. Then

$$\begin{aligned}
\hat{g}_{-i}(x_i) &= e_1'\hat{\delta}_{-i}(x_i) = g(x_i) + e_1'[\,A_{2i}\,]^{-1}\{A_{1i}\} = g(x_i) + (\hat{f}_i^{-1} + C_{1i}, -C_{2i})A_{1i} \\
&= g(x_i) + \frac{1}{n}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j]/\hat{f}_i \\
&\quad + \frac{1}{n}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j][C_{1i} - C_{2i}D_h^{-2}(x_j - x_i)] \\
&\equiv g(x_i) + \frac{1}{n}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j]/\hat{f}_i + M_n,
\end{aligned}$$

where $M_n = n^{-1}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j][C_{1i} - C_{2i}D_h^{-2}(x_j - x_i)]$, which has an order smaller than $n^{-1}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j]/\hat{f}_i$ (smaller by a factor of $\eta_2 = \sum_{s=1}^q h_s^2$ since both C_{1i} and C_{2i} are $O_p(\eta_2)$).

Define $\mathcal{D}_i = n^{-1}\sum_{j\neq i} W_{h,ij}[R_{ij} + u_j]/\hat{f}_i$. Then we have $\hat{g}(x_i) = g(x_i) + \mathcal{D}_i + M_n \equiv \tilde{g}(x_i) + M_n$, where $\tilde{g}(x_i) = g(x_i) + \mathcal{D}_i$.

We use the short-hand notation $g_i = g(x_i)$, $\hat{g}_i = \hat{g}(x_i)$, $\tilde{g}_i = \tilde{g}(x_i)$. Define $CV_0(h)$ in the same manner as $CV(h)$ but with $\hat{g}_i$ being replaced by $\tilde{g}_i$. Then

$$\begin{aligned}
CV_0(h) &\stackrel{def}{=} \sum_i (y_i - \tilde{g}_i)^2 = \sum_i (g_i + u_i - \tilde{g}_i)^2 = \sum_i [u_i - \mathcal{D}_i]^2 \\
&= \sum_i \mathcal{D}_i^2 - 2\sum_i u_i\mathcal{D}_i + \sum_i u_i^2 \equiv CV_1(h) + n^{-1}\sum_i u_i^2, \qquad \text{(A.2)} \\
CV_1(h) &= \sum_i \mathcal{D}_i^2 - 2\sum_i u_i\mathcal{D}_i \\
&= n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}[R_{ij}R_{il} + u_ju_l + 2u_jR_{il}]W_{h,ij}W_{h,il}/\hat{f}_i^2 \\
&\quad -2n^{-2}\sum_i\sum_{j\neq i} u_i[R_{ij} + u_j]W_{h,ij}/\hat{f}_i. \qquad \text{(A.3)}
\end{aligned}$$

Note that minimizing $CV_0(h)$ over $h_1, \ldots, h_q$ is equivalent to minimizing $CV_1(h)$ because $n^{-1}\sum_i u_i^2$ is not related to $h_1, \ldots, h_q$.

A technical difficulty in handling (A.3) arises from the presence of the random denominator $\hat{f}_i$, but

$$\frac{1}{\hat{f}_i} = \frac{1}{f_i} + \frac{(f_i - \hat{f}_i)}{f_i^2} + \frac{(f_i - \hat{f}_i)^2}{f_i^2\hat{f}_i}. \qquad \text{(A.4)}$$

Define $CV_2(h)$ by replacing the random denominator $\hat{f}_i$ in $CV_1(h)$ by f_i.

$$\begin{aligned} CV_2(h) &\stackrel{def}{=} \Big\{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} R_{ij}R_{il}W_{h,ij}W_{h,il}/f_i^2\Big\} \\ &\quad +\Big\{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l W_{h,ij}W_{h,il}/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i} u_i u_j W_{h,ij}/f_i\Big\} \\ &\quad +2\Big\{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j R_{il} W_{h,ij}W_{h,il}/f_i^2 - n^{-2}\sum_i\sum_{j\neq i} u_i R_{ij} W_{h,ij}/f_i\Big\} \\ &\equiv \{S_1\}+\{S_2\}+2\{S_3\}, \end{aligned}$$

where the definition of S_j $(j=1,2,3)$ should be apparent.

Define $\eta_1=(nh_1\cdots h_q)^{-1}$ and $\eta_2=\sum_{s=1}^q h_s^2$. Lemmas A.1 to A.3 below show that $S_1=\int[(\kappa_2/2)\sum_{s=1}^q g_{ss}(x)h_s^2]^2 f(x)dx+O(\eta_2^3+\eta_1(h_1\cdots h_q)^{1/2}+n^{-1/2}\eta_2^2)$, $S_2=B_0(nh_1\cdots h_q)^{-1}+O(\eta_1(\eta_2+(h_1\cdots h_q)^{1/2}+n^{-1/2})$ and $S_3=O(n^{-1/2}\eta_2^2)$, where $B_0=\kappa^q\int\sigma^2(x)dx$. Therefore,

$$\begin{aligned} CV_2(h)=S_1+S_2+2S_3 &= \int\Big[\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x)h_s^2\Big]^2 f(x)dx+\frac{B_0}{nh_1\cdots h_q} \\ &\quad +O\left(\eta_2^3+\eta_1(\eta_2+(h_1\cdots h_q)^{-1}+n^{-1/2})\right). \end{aligned}$$

Let $CV_L(h)=\int[(\kappa_2/2)\sum_{s=1}^q g_{ss}(x)h_s^2]^2 f(x)dx+B_0(nh_1\cdots h_q)^{-1}$ denote the leading term of $CV_2(h)$.

Letting $h_1^0,\ldots,h_q^0$ denote the values of $h_1,\ldots,h_q$ that minimize $CV_L(h)$, then obviously we have $h_s^0=n^{-1/(q+4)}a_s^0=O(n^{-1/(q+4)})$ for $s=1,\ldots,q$, where $a_0's$ are defined below (2.6). Recall that $\hat{h}_1,\ldots,\hat{h}_q$ are the values of $h_1,\ldots,h_q$ that minimize $CV(h)$. Based on the fact that $CV(h)=CV_L(h)+O(\eta_2^3+\eta_1\eta_2+\eta_1(h_1\cdots h_q)^{1/2})+$ terms not related to $h_1,\ldots,h_q$, we know that $\hat{h}_s=h_s^0+o_p(h_s^0)=O_p(n^{-1/(q+4)})$ for $s=1,\ldots,q$.

From $CV_1(h)=CV_L(h)+O(\eta_2^3+\eta_1\eta_2+\eta_1(h_1\cdots h_q)^{1/2}+n^{-1/2}\eta_2^2)$ and $h_s\sim n^{-1/(q+4)}$, we obtain $CV_1(h)=CV_L(h)+O(\eta_1(h_1\cdots h_q)^{1/2})$ if $q\le 3$, and $CV_1(h)=CV_L(h)+O(\eta_2^3)$ if $q\ge 4$. Using these results one can show that $\hat{h}_s=h_s^0+O_p(h_s^0 n^{-q/[2(q+4)]})$ if $q\le 3$, and $\hat{h}_s=h_s^0+O_p(h_s^0 n^{-2/(q+4)})$ if $q\ge 4$.

This completes the proof of Theorem 2.1.

Proof of Theorem 2.2. Define $\bar{g}(x)$ in the same manner as $\hat{g}(x)$, but with the $\hat{h}_s$'s in $\hat{g}(x)$ being replaced by h_s^0's. Then it is well established that $(nh_1^0\cdots h_q^0)^{1/2}$ $(\bar{g}(x)-\sum_{s=1}^q(h_s^0)^2\mu_s(x))\to N(0,\Omega_x)$ in distribution. Using the result of Theorem 2.1 and a standard Taylor expansion argument (e.g., Racine and Li (2004)), it is easy to check that $\hat{g}(x)-\bar{g}(x)=o_p(\sum_{s=1}^q(h_s^0)^2+(nh_1^0\cdots h_q^0)^{-1/2})$. Then using $\hat{h}_s=h_s^0+O_p(h_s^0 n^{-\epsilon/(4+q)})$, one has Theorem 2.2.

Below we present some lemmas that are used in the proof of Theorem 2.1. We write $\mathcal{A}_n = \mathcal{B}_n + (s.o.)$ to denote the fact that $\mathcal{B}_n$ is the leading order term of $\mathcal{A}_n$, while $(s.o.)$ denotes terms of smaller order than $\mathcal{B}_n$.

Lemma A.1. $S_1 = \int[(\kappa_2/2)\sum_{s=1}^q g_{ss}(x)h_s^2]^2 f(x)dx + O(\eta_2^3 + \eta_1(h_1\cdots h_q)^{1/2} + n^{-1/2}\eta_2^2)$.

Proof. $S_1 = n^{-3}\sum\sum\sum_{i\neq j\neq l} R_{ij}R_{il}W_{h,ij}W_{h,il}/f_i^2 + n^{-3}\sum\sum_{j\neq i} R_{ij}^2 W_{h,ij}^2/f_i^2 \equiv S_{1a} + S_{1b}$. Here $S_{1a} = [n^{-3}\sum\sum\sum_{i\neq j\neq l} H_{1a}(x_i,x_j,x_l)]$, where $H_{1a}(x_i,x_j,x_l)$ is a symmetrized version of $R_{ij}R_{il}W_{h,ij}W_{h,il}/f_i^2$ given by $H_{1a}(x_i,x_j,x_l) = (1/3)$ $\{R_{ij}R_{il}W_{h,ij}W_{h,il}/f_i^2 + R_{ji}R_{jl}W_{h,ij}W_{h,jl}/f_j^2 + R_{lj}R_{li}W_{h,lj}W_{h,il}/f_l^2\}$.

We first compute $E[R_{ij}W_{h,ij}f_i^{-1}|x_i]$. By the assumption that $g(.)$ is a four-time continuously differentiable function we have, uniformly in i,

$$\begin{aligned} E[R_{ij}W_{h,ij}f_i^{-1}|x_i] &= E\{\,[\,g_j - g_i - (x_j - x_i)'\nabla g_i\,]\,W_{h,ij}f_i^{-1}|x_i\,\} \\ &= (\kappa_2/2)\sum_{s=1}^q g_{ss}(x_i)h_s^2 + O(\eta_2^3), \end{aligned} \tag{A.5}$$

where $\kappa_2 = \int w(v)v^2dv$, $g_{ss}(x_i) = [\partial^2 g(x)/\partial x_s^2]|_{x=X_i}$. Using (A.5) we have

$$\begin{aligned} E[H_{1a}(x_i,x_j,x_l)] &= E\{E[R_{ij}W_{h,ij}f_i^{-1}|x_i]\}^2 \\ &= E\Big\{\Big[\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x_i)h_s^2\Big]^2\Big\} + O\left(\eta_2^3\right) \\ &= \int \Big[\frac{\kappa_2}{2}\sum_{s=1}^2\sum_{s=1}^q g_{ss}(x)h_s^2\Big]^2 f(x)dx + O\left(\eta_2^3\right), \end{aligned} \tag{A.6}$$

$$\begin{aligned} &E[H_{1a}(x_i,x_j,x_l)|x_i] \sim E[R_{ij}R_{il}W_{h,ij}W_{h,il}/f_i^2|x_i] \\ &= E[R_{ij}W_{h,ij}|x_i]E[R_{il}W_{h,il}|x_i]/f_i^2 = \{E[R_{ij}W_{h,ij}|x_i]/f_i\}^2 \\ &= \Big[\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x_i)h_s^2\Big]^2 + O\left(\eta_2^3\right). \end{aligned} \tag{A.7}$$

By (A.5), (A.6), (A.7), and the U-statistic H-decomposition, we have

$$\begin{aligned} S_{1a} &= E[H_{1a}(x_i,x_j,x_l)] + \frac{3}{n}\sum_i\{E[H_{1a}(x_i,x_j,x_l)|X_i] - E[H_{1a}(x_i,x_j,x_l)]\} + (s.o.) \\ &= E[H_{1a}(x_i,x_j,x_l)] + n^{-1/2}O\left(\eta_2^2\right) + O\left(\eta_1(h_1\cdots h_q)^{1/2}\right) \\ &= \Big[\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x)h_s^2\Big]^2 f(x)dx + O\left(\eta_2^3 + \eta_1(h_1\cdots h_q)^{1/2} + n^{-1/2}\eta_2^2\right). \end{aligned} \tag{A.8}$$

Note that in applying the U-statistic H-decomposition, we write the last term as $(s.o.)$ because the last term in the decomposition is a degenerate U-statistic, (the U-statistic $(2/n(n-1))\sum_i\sum_{j>i}H_n(z_i,z_j)$ is said to be a degenerate U-statistic if $E[H_n(z_i,z_j)|z_i]=0$), and it can be easily shown that it has an order of $O(\eta_2^{1/2}\eta_1(h_1\cdots h_q)^{1/2})=o(\eta_1(h_1\cdots h_q)^{1/2})$, so we write it as $(s.o.)$. The $\eta_2^{1/2}$ factor comes from R_{ij}, and $O(\eta_1(h_1\cdots h_q)^{1/2})$ comes from the standard degenerate U-statistic result.

Next, we consider S_{1b}. Defining $H_{1b}(x_i,x_j)=R_{ij}^2W_{h,ij}^2(1/f_i^2+1/f_j^2)/2$, then $S_{1b}=n^{-1}[n^{-2}\sum_i\sum_{j\neq i}H_{1b}(x_i,x_j)]$, and it is easy to see that $E[H_{1b}(x_i,x_j)]=E[R_{ij}^2W_{h,ij}^2/f_i^2]=O\left(\eta_2(h_1\cdots h_q)^{-1}\right)$.

Similarly, one can easily show that $E[H_{1b}(x_i,x_j)|x_i]=O\left(\eta_2(h_1\cdots h_q)^{-1}\right)$. Thus, by the H-decomposition,

$$S_{1b}=\frac{1}{n}\Big\{E[H_{1b}(x_i,x_j)]+2n^{-1}\sum_i\left(E[H_{1b}(x_i,x_j)|x_i]-E[H_{1b}(x_i,x_j)]\right)+(s.o.)\Big\}$$
$$=O\left(\eta_2\eta_1\right). \tag{A.9}$$

The lemma follows from (A.8) and (A.9).

Lemma A.2. $S_2=B_0(nh_1\cdots h_q)^{-1}+O\left(\eta_1(\eta_2+n^{-1/2}+(h_1\cdots h_q)^{1/2})\right)$, where $B_0=\kappa^q\int\sigma^2(x)dx$, with $\eta_1=(nh_1\cdots h_q)^{-1}$ and $\eta_2=\sum_{s=1}^q h_s^2$.

Proof. $S_2=n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i}u_ju_lW_{h,ij}W_{h,il}/f_i^2-2n^{-2}\sum_i\sum_{j\neq i}u_iu_jW_{h,ij}/f_i$ $=n^{-3}\sum_i\sum_{j\neq i}u_j^2W_{h,ij}^2/f_i^2+n^{-3}\sum\sum\sum_{i\neq j\neq l}u_j\,u_l\,W_{h,ij}\,W_{h,il}-2n^{-2}\sum_i\sum_{j\neq i}u_iu_jW_{h,ij}/f_i\equiv S_{2a}+S_{2b}-2S_{2c}$.

Define $H_{2a}(z_i,z_j)=(1/2)\,(u_i^2/f_i^2+u_j^2/f_j^2)\,W_{h,ij}^2$, then $S_{2a}=n^{-1}[n^{-2}\sum\sum_{i\neq j}H_{2a}(z_i,z_j)]$. Then $E[H_{2a}(z_i,z_j)]=E[u_i^2W_{h,ij}^2/f_i^2]=E[\sigma^2(x_i)W_{h,ij}^2/f_i^2]$ $=(h_1\cdots h_q)^{-1}[B_0+O(\eta_2)]$, where $B_0=\kappa^q\int\sigma^2(x)dx$ ($\kappa=\int w(v)^2dv$).

Next, we see that

$$\begin{aligned}E[H_{2a}(z_i,z_j)|z_i]&=(1/2)\{(u_i^2/f_i^2)E[W_{h,ij}^2|z_i]+E[(\sigma^2(x_j)/f_j^2)W_{h,ij}^2|z_i]\}\\&=(1/2)u_i^2f_i^{-2}\{E[W_{h,ij}^2|x_i]+(1/2)E[\sigma^2(x_j)W_{h,ij}^2/f_j^2|x_i]\\&=(1/2)(h_1\cdots h_q)^{-1}f_i^{-1}\{\kappa^q[u_i^2+\sigma^2(x_i)]+O(\eta_2)\}\\&=\mathcal{B}_{0i}(h_1\cdots h_q)^{-1}+O_p(\eta_2(h_1\cdots h_q)^{-1}),\end{aligned}$$

where $\mathcal{B}_{0i}=(\kappa^q/2)f_i^{-1}[u_i^2+\sigma^2(x_i)]$. It is easy to check that $B_0=E[\mathcal{B}_{0i}]$. Hence, by the H-decomposition we have

$$S_{2a}=n^{-1}\Big\{E[H_{2a}(z_i,z_j)]+2n^{-1}\sum_i\left(E[H_{2a}(z_i,z_j)|z_i]-E[H_{2a}(z_i,z_j)]\right)+(s.o.)\Big\}$$
$$=(nh_1\cdots h_q)^{-1}[B_0+O(\eta_2)]+O_p(n^{-1/2}\eta_1),$$

where the $O_p(n^{-1/2}\eta_1)$ term comes from the second term of the H-decomposition.

Next, S_{2b} can be written as a third-order U-statistic. $S_{2b} = [n^{-3}\sum\sum\sum_{i\neq j\neq l} H_{2b}(z_i, z_j, z_l)]$, where $H_{2b}(z_i, z_j, z_l)$ is a symmetrized version of $u_j u_l W_{h,ij} W_{h,il}/f_i^2$ given by

$$H_{2b}(z_i, z_j, z_l) = (1/3)[u_j u_l W_{h,ij} W_{h,il}/f_i^2 + u_i u_l W_{h,ij} W_{h,jl}/f_j^2 \\ + u_j u_i W_{h,lj} W_{h,il}/f_l^2].$$

Note that $E[H_{2b}(z_i, z_j, z_l)|z_j] = 0$ because $E(u_l|z_j) = 0$. Hence, the leading term of S_{2b} is a second-order degenerate U-statistic: $E[H_{2b}(z_i, z_j, z_l)|z_i, z_j] = (1/3)u_i u_j E[W_{h,lj} W_{h,il}/f_l^2|x_i, x_j]$.

Straightforward calculation shows that $E[W_{h,lj} W_{h,il}/f_l^2|x_i, x_j] = W_{h,ij}^{(2)}/f_i + O(\eta_2)$, where $W_{h,ij}^{(2)} = \prod_{s=1}^q h_s^{-1} w^{(2)}((x_{is} - x_{js})/h_s)$, and $w^{(2)}(v) \overset{def}{=} \int w(u)w(v + u)du$ is the two-fold convolution kernel derived from $w(\cdot)$. Hence,

$$\begin{aligned} S_{2b} &= 3\Big\{n^{-2}\sum\sum_{j\neq i} E[H_{2b}(z_i, z_j, z_l)|z_i, z_j] + (s.o)\Big\} \\ &= \Big\{n^{-2}\sum\sum_{j\neq i} u_i u_j E[W_{h,lj} W_{h,il}/f_l^2|z_i, z_j] + (s.o.)\Big\} \\ &= \Big[n^{-2}(h_1\cdots h_q)\sum\sum_{j\neq i} u_i u_j W_{h,ij}^{(2)}/f_i + (s.o.)\Big] \\ &= (n(h_1\cdots h_q)^{1/2})^{-1}\mathcal{Z}_{2b,n} + (s.o.), \end{aligned}$$

where $\mathcal{Z}_{2b,n} = (n(h_1\cdots h_q)^{1/2})\{n^{-2}\sum\sum_{j\neq i} u_i u_j W_{h,ij}^{(2)}/f_i\}$ is a zero mean $O_p(1)$ random variable.

Finally, $S_{2c} = n^{-2}\sum_i\sum_{j\neq i} u_i u_j W_{h,ij}/f_i = (n(h_1\cdots h_q)^{1/2})^{-1}\mathcal{Z}_{2c,n}$, where $\mathcal{Z}_{2c,n} = (n(h_1\cdots h_q)^{1/2})[n^{-2}\sum_i\sum_{j\neq i} u_i u_j W_{h,ij}/f_i]$ is a zero mean $O_p(1)$ random variable. The lemma follows.

Lemma A.3. $S_3 = O_p(\eta_2 n^{-1/2})$.

Proof. $S_3 = n^{-2}\sum_i\sum_{j\neq i} u_i R_{ij} W_{h,ij}/f_i - n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} R_{ij} u_l W_{h,ij} W_{h,il}/f_i^2 = n^{-2}\sum_i\sum_{j\neq i} u_i R_{ij} W_{h,ij}/f_i - n^{-3}\sum_i\sum_{j\neq i} R_{ij} u_j W_{h,ij}^2/f_i^2 - n^{-3}\sum\sum\sum_{i\neq j\neq l} R_{ij} u_l W_{h,ij} W_{h,il}/f_i^2 \equiv S_{3a} - S_{3b} - S_{3c}$.

$S_{3a} = n^{-2}\sum_i\sum_{j\neq i} H_{3a}(z_i, z_j)$, where $H_{3a}(z_i, z_j) = (1/2)[u_i R_{ij}/f_i + u_j R_{ji}/f_j] W_{h,ij}$.

We first compute $[H_{3a}(z_i, z_j)|z_i]$. $[H_{3a}(z_i, z_j)|z_i] = (1/2)(u_i/f_i)E[R_{ij} W_{h,ij}|x_i]$, and $E[R_{ij} W_{h,ij}|x_i] = (\kappa_2/2)\sum_{s=1}^q g_{ss}(x_i)h_s^2 + O_p(\eta_2^2)$. Thus, we have

$$[H_{3a}(z_i, z_j)|z_i] = (\kappa_2/4)(u_i/f_i)\Big\{\sum_{s=1}^q g_{ss}(x_i)h_s^2 + O_p(\eta_2^{3/2})\Big\} \equiv \sum_{s=1}^q \mathcal{B}_{3i}h_s^2 + (s.o.),$$

where $\mathcal{B}_{3i,s} = (\kappa_2/4)(u_i/f_i)g_{ss}(x_i)$.

Using H-decomposition and noting that $E[H_{3a}(z_i,z_j)] = 0$, we have $S_{3a} = 2n^{-1}\sum_i E[H_{3a}(z_i,z_j)|z_i] + (s.o.) = 2n^{-1}\sum_i \sum_{s=1}^q \mathcal{B}_{3i,s}h_s^2 + (s.o.) \equiv O_p(n^{-1/2}\eta_2)$, because $n^{-1/2}\sum_i \mathcal{B}_{3i,s}$ is a zero mean $O_p(1)$ random variable.

Next, for S_{3b} it is easy to see that $S_{3b} = (nh_1\cdots h_q)^{-1}O_p(S_{3a}) = O_p((nh_1\cdots h_q)^{-1}\eta_2 n^{-1/2})) = o_p(n^{-1/2}\eta_2)$.

Finally we consider S_{3c}. It can be written as a third-order U-statistic $S_{3c} = n^{-3}\sum\sum\sum_{i\neq j\neq l} H_{3c}(z_i,z_j,z_l)$, where $H_{3c}(z_i,z_j,z_l)$ is a symmetrized version of $u_l R_{ij} W_{h,ij} W_{h,il}/f_i^2$. Obviously $E[H_{3c,(i)}(z_i,z_j,z_l)] = 0$ and it can easily be verified that $E[H_{3c,(i)}(z_i,z_j,z_l)|z_i] = (1/3)u_i\sum_{s=1}^q \mathcal{D}_{3i,s}$, where $\mathcal{D}_{3i,s} = (\kappa_2/2)g_{ss}(x_i) + O(\eta_2)$. Therefore, by H-decomposition we have

$$S_{3c} = \frac{3}{n}\sum_i E[H_{3c}(z_i,z_j,z_l)|z_i] + (s.o.) = n^{-1/2}\Big[\sum_{s=1}^q n^{-1/2}\sum_i u_i\mathcal{D}_{3i,s}h_s^2\Big] + (s.o.)$$
$$\equiv O_p(\eta_2 n^{-1/2}),$$

because $n^{-1/2}\sum_i u_i\mathcal{D}_{3i,s}$ is a $O_p(1)$ random variable. The lemma follows.

B. Proof of Theorems 3.1 and 3.2

We first list the assumptions that will be used to prove Theorems 3.1 and 3.2.

Let $\mathcal{G}_\mu^\alpha$ denote the class of functions introduced in Robinson (1988) for $\alpha > 0$, and μ a positive integer: $m \in \mathcal{G}_\mu^\alpha$, if $m(x^c)$ is μ times differentiable, and $m(x^c)$ and its partial derivatives (up to order μ) are all bounded by functions that have finite αth moment.

(B1) (i) We restrict $(\hat{h}_1,\ldots,\hat{h}_q,\hat{\lambda}_1,\ldots,\hat{\lambda}_r) \in [0,\eta]^{q+r}$ to lie in a shrinking set, and $nh_1\cdots h_q \geq t_n$ ($t_n \to \infty$ as $n \to \infty$). (ii) The kernel function $w(\cdot)$ satisfies (A2). (iii) $f(x)$ is bounded below by a positive constant on $\mathcal{S}\times\mathcal{S}^d$, the support of $X = (X^c, X^d)$.

(B2) (i) $\{X_i, Y_i\}_{i=1}^n$ are independent and identically distributed as (X,Y), $u_i = Y_i - g(X_i)$ has finite fourth moment. (ii) Defining $\sigma^2(x) = E[u_i^2|X_i = x]$, $\sigma^2(\cdot,x^d)$, $g(\cdot,x^d)$ and $f(\cdot,x^d)$ all belong to $\mathcal{G}_2^4$ for all $x^d \in \mathcal{S}^d$. (iii) Define, with the $D_s(x)$'s defined in (3.3),

$$\int \Big\{\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x)a_s^2 + \sum_{s=1}^r D_s(x)b_s\Big\}^2 f(x)dx + \frac{B_0}{h_1\cdots h_q}$$

is uniquely minimized at $(a_1^0,\ldots,a_q^0,\lambda_1^0,\ldots,\lambda_r^0)$, and each a_s^0 and b_s^0 is finite.

Proof of Theorem 3.1. We first prove some intermediate results. Let $x_i = (x_i^c, x_i^d)$, and define $\beta(x_i) = [\partial g(x^c,x_i^d)/\partial x^c]|_{x^c=x_i^c}$. Now define $R_{ij} = g(x_j) -$

$g(x_i)-\beta(x_i)'(x_j^c-x_i^c)$, which is equivalent to $g(x_j)=g(x_i)+\beta(x_i)'(x_j^c-x_i^c)+R_{ij}$. Therefore, we have

$$y_j = g(x_j)+u_j = g(x_i)+(x_j^c-x_i^c)'\beta(x_i)+R_{ij}+u_j = (1,(x_j^c-x_i^c)')\delta(x_i)+R_{ij}+u_j, \qquad \text{(B.1)}$$

where $\delta(x_i)=(g(x_i),\beta(x_i)')'$.

We observe that (B.1) has a form similar to (2.2) for the continuous-regressor-only case. Therefore, by following the same arguments as in Appendix A, one can introduce $CV_0(h,\lambda)$, $CV_2(h,\lambda)$ and $CV_2(h,\lambda)$ in a manner analogous to the continuous-regressor-only case presented in Appendix A. By also using (A.4) with $\hat{f}_i = n^{-1}\sum_{j\neq i} K_{h,ij}$, and noting that $\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)| = o(1)$, one can show that

$$CV(h,\lambda) = CV_2(h,\lambda) + O_p(\eta_3+\eta_1^{-1/2})O_p(CV_2(h,\lambda)) + \frac{1}{n}\sum_i u_i^2, \qquad \text{(B.4)}$$

where $\eta_3 = \sum_{s=1}^q h_s^2 + \sum_{s=1}^r \lambda_s$, and

$$\begin{aligned} CV_2(h,\lambda) &= \{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} R_{ij}R_{il}K_{h,ij}K_{h,il}/f_i^2\} \\ &\quad +\{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l K_{h,ij}K_{h,il}/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i} u_i u_j K_{h,ij}/f_i\} \\ &\quad +2\{n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j R_{il}K_{h,ij}K_{h,il}/f_i^2 - \sum_i\sum_{j\neq i} u_i R_{ij}K_{h,ij}/f_i\} \\ &\equiv \{\mathcal{S}_1\} + \{\mathcal{S}_2\} + 2\{\mathcal{S}_3\}, \end{aligned} \qquad \text{(B.5)}$$

where the definition of $\mathcal{S}_j$ $(j=1,2,3)$ should be apparent.

By lemmas B.1 through B.3 we know that

$$\begin{aligned} \mathcal{S}_1 &= \sum_{x^d}\int\Big\{\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x)h_s^2 + \sum_{s=1}^r D_s(x)\lambda_s\Big\}^2 \\ &\quad + O_p(\eta_3^3 + \eta_1(h_1\cdots h_q)^{1/2} + n^{-1/2}\eta_3^2), \\ \mathcal{S}_2 &= B_0(nh_1\cdots h_q)^{-1} + O_p(\eta_1(\eta_3+n^{-1/2}+(h_1\cdots h_q)^{1/2})) + (s.o.), \\ \mathcal{S}_3 &= O_p(n^{-1/2}\eta_3). \end{aligned} \qquad \text{(B.6)}$$

Note that the above results are almost the same as the continuous-regressor case except that $\eta_2=\sum_{s=1}^q h_s^2$ is replaced by $\eta_3=\sum_{s=1}^q h_s^2+\sum_{s=1}^r\lambda_s$, i.e., the bias term needs to be modified to include terms of order $O(\lambda_s)$ $(s=1,\ldots,r)$. The variance term remains unchanged.

Combining (B.4), (B.5) and (B.6), and also dropping $n^{-1}\sum_i u_i^2$, since it is independent of (h,λ), we get

$$CV_2 = \sum_{x^d}\int\Big\{\frac{\kappa_2}{2}\sum_{s=1}^q g_{ss}(x)h_s^2 + \sum_{s=1}^r D_s(x)\lambda_s\Big\}^2 f(x)dx^c + \frac{B_0}{nh_1\cdots h_q} + (s.o.). \qquad \text{(B.7)}$$

Define $a_1, \dots, a_q, b_1, \dots, b_r$ via $h_s = a_s n^{-1/(q+4)}$ ($s = 1, \dots, q$) and $h_s = b_s n^{-2/(q+4)}$ ($s = 1, \dots, r$). If $CV_L(h, \lambda)$ denotes the leading term of CV_2 at (B.7), then $CV_L(h, \lambda) = \chi(a, b)$, where

$$\chi(a,b) = \int \sum_{x^d} \Big\{ \frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) a_s^2 + \sum_{s=1}^{r} D_s(x) b_s \Big\}^2 dx^c + \frac{B_0}{h_1 \cdots h_q}.$$

Let $(a_1^0, \dots, b_r^0)$ denote the values of $(a_1, \dots, b_r)$ that minimize $\chi(a, b)$. By (3.4) we know that each of the a_s^0's and b_s^0's is uniquely defined and is finite. Letting $(h_1^0, \dots, \lambda_r^0)$ denote the values of $(h_1, \dots, \lambda_r)$ that minimize CV_L, then obviously $n^{1/(q+q)} h_s^0 \sim a_s^0$ for $s = 1, \dots, q$, and $n^{2/(q+4)} \lambda_s^0 \sim b_s^0$ for $s = 1, \dots, q$.

By arguments similar to those found in the proof of Theorem 2.1 of Racine and Li (2004), it can be shown that $CV = CV_L + O_p((h_1 \cdots h_q)^{1/2}) O_p(CV_L)$ if $q \le 3$, and $CV = CV_L + O_p(\sum_{s=1}^q h_s^2) O_p(CV_L)$ if $q \ge 4$. Using $h_s^0 = O\left(n^{-1/(q+4)}\right)$ we get $\hat{h}_s = h_s^0 + O_p(h_s^0 n^{-q/[2(q+4)]})$, $\hat{\lambda}_s = \lambda_s^0 + O_p(n^{-1/2})$, if $q \le 3$; $\hat{h}_s = h_s^0 + O_p(h_s^0 n^{-2/(q+4)})$, $\hat{\lambda}_s = \lambda_s^0 + O_p(n^{-4/(q+4)})$, if $q \ge 4$, where $s = 1, \dots, q$ for $\hat{h}_s$, and $s = 1, \dots, r$ for $\hat{\lambda}_s$. This completes the proof of Theorem 3.1.

Proof of Theorem 3.2. Define $\bar{g}(x)$ in the same manner as $\hat{g}(x)$ but with h_s^0's and λ_s^0's replacing $\hat{h}_s$' and $\hat{\lambda}_s$'s. Then it is easy to see that $(nh_1^0 \cdots h_q^0)^{1/2}(\bar{g}(x) - g(x) - (\kappa_2/2) \sum_{s=1}^q g_{ss}(x)(h_s^0)^2 - \sum_{s=1}^r D_s(x)\lambda_s^0) \to N(0, \Omega(x))$ in distribution.

Next, using the results of Theorem 3.1 and a Taylor expansion argument, it is easy to show that $(\hat{g}(x) - \bar{g}(x)) = o_p(n^{-2/(q+4)})$, also $\hat{h}_s^2 = (h_s^0)^2 + o_p(n^{-2/(q+4)})$ and $\hat{\lambda}_s = \lambda_s^0 + o_p(n^{-2/(q+4)})$. Theorem 3.2 follows from these results.

Lemma B.1. $\mathcal{S}_1 = \int \{ (\kappa_2/2) \sum_{s=1}^q g_{ss}(x) h_s^2 + \sum_{s=1}^r D_s(x) \lambda_s^2 \}^2 f(x) dx + O_p(\eta_3^3 + \eta_1 (h_1 \cdots h_q)^{1/2} + n^{-1/2} \eta_3^2)$, *where the* $D_{js}(x)$*'s are some functions defined in the proof of Theorem* 3.2.

Lemma B.2. $\mathcal{S}_2 = B_0 (nh_1 \cdots h_q)^{-1} + O_p(\eta_1(\eta_3 + n^{-1/2} + (h_1 \cdots h_q)^{1/2}))$, *where* $B_0 = \kappa^q \sum_{x^d} \int \sigma^2(x) dx^c$.

Lemma B.3. $\mathcal{S}_3 = O_p(n^{-1/2} \eta_3)$.

The proofs of Lemmas B.1 through B.3 proceed along the lines of the proofs of Lemmas A.1 through A.3. Below we provide outlines of proofs for Lemmas B.1 and B.2.

Proof of Lemma B.1. We have $S_1 = n^{-3} \sum \sum \sum_{i \ne j \ne l} R_{ij} R_{il} K_{h,ij} K_{h,il} / f_i^2 + n^{-3} \sum \sum_{j \ne i} R_{ij}^2 K_{h,ij}^2 / f_i^2 \equiv S_{1a} + S_{1b}$. The term S_{1a} can be written as a third order U-statistic whose leading term is $E[R_{ij} R_{il} K_{h,ij} K_{h,il} / f_i^2] = E\{E[R_{ij} K_{h,ij} / f_i | x_i]^2\}$. Now

$$E[R_{ij} K_{h,ij} f_i^{-1} | x_i] = E\{ [g_j - g_i - (x_j^c - x_i^c)' \nabla g_i] K_{h,ij} f_i^{-1} | x_i \}$$

$$= \frac{\kappa_2}{2}\sum_{s=1}^{q} g_{i,ss}h_s^2 + \sum_{s=1}^{q}\lambda_s \sum_{v^d}[\mathbf{1}_s(x_i^d, v^d)g(x_i^c, v^d) - g(x_i)] + O\left(\eta_3^2\right),$$

where $g_{i,ss} = [\partial^2/\partial(x_s^c)^2 g(x)]|_{x=X_i}$ is the second partial derivative of g with respect to x_s^c evaluated at x_i. Therefore,

$$\begin{aligned}
&E\{E[R_{ij}K_{h,ij}/f_i|x_i]^2\} \\
&= E\Big\{\frac{\kappa_2}{2}\sum_{s=1}^{q} g_{ss}(x_i)h_s^2 + \sum_{s=1}^{q}\lambda_s \sum_{v^d}[\mathbf{1}_s(x_i^d, v^d)g(x_i^c, v^d) - g(x_i)]\Big\}^2 + O\left(\eta_3^3\right) \\
&\equiv \sum_{x^d}\int \Big\{\frac{\kappa_2}{2}\sum_{s=1}^{q} g_{ss}(x)h_s^2 + \sum_{s=1}^{q}\lambda_s D_s(x)\Big\}^2 f(x)dx^c + O\left(\eta_3^3\right),
\end{aligned}$$

where $D_s(x) = \sum_{v^d}[\mathbf{1}_s(x^d, v^d)g(x^c, v^d) - g(x)]f(x^c, v^d)$.

Similar to the arguments used in the proof of Lemma A.1, one can show that $S_1 = E\{E[R_{ij}K_{h,ij}/f_i|x_i]^2\} + O_p(\eta_3^3 + \eta_1(h_1\cdots h_q)^{1/2} + n^{-1/2}\eta_3^2)$. This completes the proof.

Proof of Lemma B.2. $S_2 = n^{-3}\sum_i\sum_{j\neq i}\sum_{l\neq i} u_j u_l K_{h,ij}K_{h,il}/f_i^2 - 2n^{-2}\sum_i\sum_{j\neq i} u_i u_j K_{h,ij}/f_i = n^{-3}\sum_i\sum_{j\neq i} u_j^2 K_{h,ij}^2/f_i^2 + n^{-3}\sum\sum\sum_{i\neq j\neq l} u_j u_l K_{h,ij}K_{h,il} - 2n^{-2}\sum_i\sum_{j\neq i} u_i u_j K_{h,ij}/f_i \equiv S_{2a} + S_{2b} - 2S_{2c}$.

Along the lines of the proof of Lemma A.2, it can be shown that $S_2 = E(S_{2a}) + O(E(S_{2a})\left(\eta_3 + n^{-1/2} + (h_1\cdots h_q)^{1/2}\right)$. The leading term of S_2, $E[S_{2a}]$ is

$$\begin{aligned}
E[S_{2a}] &= n^{-1}E[u_i^2 K_{h,ij}^2/f_i^2] = n^{-1}E[\sigma^2(x_i)K_{h,ij}^2/f_i^2] \\
&= (nh_1\cdots h_q)^{-1}[B_0 + O(\eta_3)],
\end{aligned}$$

where $B_0 = \kappa^q E[\sigma^2(x_i)/f(x_i)]$, $\kappa = \int w(v)^2 dv$. Thus, we have

$$\begin{aligned}
S_2 &= E[S_{2a}] + O\left(E(S_{2a})(\eta_3 + n^{-1/2} + (h_1\cdots h_q)^{1/2})\right) \\
&= \frac{B_0}{nh_1\cdots h_q} + O\left(\eta_1(\eta_3 + n^{-1/2} + (h_1\cdots h_q)^{1/2})\right).
\end{aligned}$$

This completes the proof of Lemma B.2.

References

Ahmad, I. A. and Cerrito, P. B. (1994). Nonparametric estimation of joint discrete-continuous probability densities with applications, *J. Statist. Plann. Inference* **41**, 349-364.

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

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Trans. Automat. Control* **9**, 716-723.

Bierens, H. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 699-707.

Bierens, H. (1987). Kernel estimation of regression functions. In *Advances in Econometrics* (Edited by T. F. Bewley). Cambridge University Press, Cambridge.

Chen, R. (1996). Incorporating extra information in nonparametric smoothing. *J. Multivariate Anal.* **58**, 133-150.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.

Fan, J. (1993). Local linear smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196-216.

Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and its Applications.* Chapman and Hall.

Grund, B. and Hall, P. (1993). On the performance of kernel estimators for high-dimensional sparse binary data. *J. Multivariate Anal.* **44**, 321-344.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika* **68**, 287-294.

Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.* To accept.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86-99.

Härdle, W., Hall, P. and Marron, J. S. (1992). Regression smoothing parameters that are not far from their optimum. *J. Amer. Statist. Assoc.* **87**, 227-233.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13**, 1465-1481.

Hurvich, C. M. and Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. B* **60**, 271-293.

Li, Q. and Racine, J. (2001). Empirical applications of smoothing categorical variables. Unpublished manuscript.

Racine, J. and Li, Q. (2004). Nonparametric estimation of regression functions with both categorical and continuous Data. *J. Econometrics* **119**, 99-130.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12**, 1215-1230.

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

Ruppert, D. and Wand, M. P. (1994). Multivariate weighted least squares regression. *Ann. Statist.* **22**, 1346-1370.

Scott, D. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Xia, Y. C. and Li, W. K. (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. *J. Multivariate Anal.* **83**, 265-287.

Department of Economics, Texas A&M University, College Station, TX 77843-4228, U.S.A.
E-mail: qi@econmail.tamu.edu

Department of Economics and Center for Policy Research, Syracuse University, 426 Eggers Hall. Syracuse, NY 13244-1020, U.S.A.
E-mail: jracine@maxwell.syr.edu

(Received September 2002; accepted July 2003)