NOTE

NONPARAMETRIC ESTIMATION OF REGRESSION FUNCTIONS IN THE PRESENCE OF IRRELEVANT REGRESSORS

Peter Hall, Qi Li, and Jeffrey S. Racine*

Abstract—In this paper we consider a nonparametric regression model that admits a mix of continuous and discrete regressors, some of which may in fact be redundant (that is, irrelevant). We show that, asymptotically, a data-driven least squares cross-validation method can remove irrelevant regressors. Simulations reveal that this "automatic dimensionality reduction" feature is very effective in finite-sample settings.

I. Introduction and Background

Nonparametric kernel methods appeal because they are robust to functional form specification, though they are often criticized because they suffer from the curse of dimensionality. If, however, some regressors are in fact irrelevant, then they should be removed, producing a lower dimensional model and thereby alleviating the curse of dimensionality. In order for kernel methods to remove a regressor, however, oversmoothing must occur and the bandwidth must not converge to 0 as the sample size increases. If the irrelevant regressors are independent of both the dependent variable and the relevant regressors, we can show theoretically that least squares kernel smoothing of continuous and categorical regressors can (asymptotically) remove irrelevant regressors, though simulations clearly reveal that strict independence among the regressors is not necessary. Hall, Racine, and Li (2004) establish similar results for conditional kernel density estimation. In this paper we extend Hall et al.'s results to the regression setting. For related literature on testing for irrelevant regressors in a nonparametric regression framework, see Fan and Li (1996) and Lavergne and Vuong (1996), among others. The results contained in the current paper clearly demonstrate that pretesting is totally unnecessary when one employs cross-validated bandwidth selection. We view this as a powerful result that has the potential to extend the reach of nonparametric methods. The rest of this paper is organized as follows. Section II presents the main results of the paper by showing that the cross-validation method has the ability to remove irrelevant regressors. Simulations are presented in section III. The appendix provides proofs of results presented in section II.

II. Regression Models Having Irrelevant Regressors

Consider a nonparametric model with both categorical and continuous regressors. Let X_i^{d} be a $q \times 1$ vector of discrete regressors and let $X_i^{\mathrm{c}} \in \mathbb{R}^p$ be the continuous ones. We use X_{is}^{d} to denote the sth component of X_i^{d}, assume that X_{is}^{d} takes $c_s \geq 2$ different values, and use $\mathcal{S}^{\mathrm{d}}$ to denote the support of X^{d}. We wish to estimate $E(Y_i|X_i)$ nonparametrically, where $X_i = (X_i^{\mathrm{d}}, X_i^{\mathrm{c}})$. However, often in applied settings not all $q + p$ regressors in X_i are relevant. Without loss of generality, assume that the first p_1 ($1 \leq p_1 \leq p$) components of X^{c} and the first q_1 ($0 \leq q_1 \leq q$) components of X^{d} are "relevant" in the sense defined below.

Let $\bar{X}$ consist of the first p_1 relevant components of X^{c} and the first q_1 relevant components of X^{d}, and let $\tilde{X} = X \backslash \{\bar{X}\}$ denote the remaining irrelevant components of X. One way of defining $\bar{X}$ to be relevant and $\tilde{X}$ to be irrelevant is to ask that

$$(\bar{X}, Y) \text{ is independent of } \tilde{X}. \tag{1}$$

Clearly, condition (1) implies that $E(Y|X) = E(Y|\bar{X})$. A weaker condition would be to ask that

$$\begin{aligned} &\text{conditional on } \bar{X}, \text{ the variables } \tilde{X} \text{ and } Y \text{ are} \\ &\text{independent.} \end{aligned} \tag{2}$$

Though more appealing, condition (2) creates quite difficult technical hurdles, so we proceed under (1) for our proofs; extensive simulations reveal that our results clearly hold under (2) (see section III).

We shall assume that the true regression model is given by $Y_i = \bar{g}(\bar{X}_i) + u_i$ where $\bar{g}(\cdot)$ is of unknown form, and where $E(u_i|\bar{X}_i) = 0$. We assume that the relevant regressors are unknown ex ante, hence one estimates $\bar{g}(\cdot)$ potentially using the superset of regressors $X = (\bar{X}, \tilde{X})$. We use $f(x)$ to denote the joint density function of X_i, and we use $\bar{f}(\bar{x})$ and $\tilde{f}(\tilde{x})$ to denote the marginal densities of $\bar{X}_i$ and $\tilde{X}_i$, respectively.

For the discrete regressors X_i^{d}, we will first consider the case for which there is no natural ordering in X_i^{d}. The extension to the case whereby some of the discrete regressors have natural orderings will be discussed at the end of this section. For $1 \leq s \leq q$, we define the kernel function for discrete regressors as

$$l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = \begin{cases} 1 & \text{if } X_{is}^{\mathrm{d}} = x_s^{\mathrm{d}}, \\ \lambda_s & \text{if } X_{is}^{\mathrm{d}} \neq x_s^{\mathrm{d}}, \end{cases} \tag{3}$$

where $0 \leq \lambda_s \leq 1$ is the smoothing parameter for x_s^{d}. Therefore, the product kernel for $x^{\mathrm{d}} = (x_1^{\mathrm{d}}, \ldots, x_q^{\mathrm{d}})$ is given by $K^{\mathrm{d}}(x^{\mathrm{d}}, X_i^{\mathrm{d}}) = \prod_{s=1}^{q} l(X_{is}^{\mathrm{d}}, x_s^{\mathrm{d}}, \lambda_s) = \prod_{s=1}^{q} \lambda_s^{I(X_{is}^{d} \neq x_s^{d})}$, where $I(X_{is}^{\mathrm{d}} \neq x_s^{\mathrm{d}})$ is an indicator function that equals 1 when $X_{is}^{\mathrm{d}} \neq x_s^{\mathrm{d}}$, and 0 otherwise. Note that when $\lambda_s = 1$, $K^{\mathrm{d}}(x^{\mathrm{d}}, X_i^{\mathrm{d}})$ is unrelated to $(x_s^{\mathrm{d}}, X_{is}^{\mathrm{d}})$ (that is, the sth component of x^{d} is smoothed out).

For the continuous regressors $x^{\mathrm{c}} = (x_1^{\mathrm{c}}, \ldots, x_p^{\mathrm{c}})$ we use the product kernel given by $K^{\mathrm{c}}(x^{\mathrm{c}}, X_i^{\mathrm{c}}) = \prod_{s=1}^{p} h_s^{-1} K\left(\frac{x_s^c - X_{is}^c}{h_s}\right)$, where K is a symmetric, univariate density function, and h_s is the smoothing parameter for x_s^{c}. The kernel function for the mixed regressor case $x = (x^{\mathrm{c}}, x^{\mathrm{d}})$ is simply the product of K^{c} and K^{d}, that is, $\mathcal{H}(x, X_i) = K^{\mathrm{c}}(x^{\mathrm{c}}, X_i^{\mathrm{c}}) K^{\mathrm{d}}(x^{\mathrm{d}}, X_i^{\mathrm{d}})$. Thus we estimate $E(Y|X = x)$ by

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \mathcal{H}(x, X_i)}{\sum_{i=1}^{n} \mathcal{H}(x, X_i)}.$$

Received for publication September 7, 2005. Revision accepted for publication August 23, 2006.

* Department of Mathematics and Statistics, University of Melbourne; Department of Economics, Texas A&M University; and Department of Economics, McMaster University, respectively.

We would like to thank three anonymous referees and the editor for their helpful comments that led to a much-improved version of this paper. Li's research is partially supported by the Private Enterprise Research Center, Texas A&M University and Tsinghua University. Racine would like to gratefully acknowledge support from National Sciences and Engineering Research Council of Canada (NSERC: www.nserc.ca), the Social Sciences and Humanities Research Council of Canada (SSHRC: www.sshrc.ca), and the Shared Hierarchical Academic Research Computing Network (SHARCNET: www.sharcnet.ca).

We choose $(h,\lambda) = (h_1, \ldots, h_p, \lambda_1, \ldots, \lambda_q)$ by minimizing the cross-validation function given by

$$CV(h,\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i}(X_i))^2 w(X_i), \tag{4}$$

where $\hat{g}_{-i}(X_i) = \sum_{j\neq i}^{n} Y_j \mathcal{H}(X_i, X_j)/\sum_{j\neq i}^{n} \mathcal{H}(X_i, X_j)$ is the leave-one-out kernel estimator of $E(Y_i|X_i)$, and $0 \leq w(\cdot) \leq 1$ is a weight function that serves to avoid difficulties caused by dividing by 0, or by the slower convergence rate arising when X_i lies near the boundary of the support of X.

Define $\sigma^2(\bar{x}) = E(u_i^2|\bar{X}_i = \bar{x})$ and let $\mathcal{S}^c$ denote the support of w. We also assume that

$$\begin{aligned}&\text{the data are i.i.d. and } u_i \text{ has finite moments}\\ &\text{of any order; } g, f, \text{ and } \sigma^2 \text{ have two continuous}\\ &\text{derivatives; } w \text{ is continuous, nonnegative, and has}\\ &\text{compact support; } f \text{ is bounded away from 0 for}\\ &x = (x^c, x^d) \in \mathcal{S} = \mathcal{S}^c \times \mathcal{S}^d.\end{aligned} \tag{5}$$

We impose the following conditions on the bandwidth and kernel functions. Define

$$H = \Big(\prod_{s=1}^{p_1} h_s\Big) \prod_{s=p_1+1}^{p} \min(h_s, 1). \tag{6}$$

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

$$\begin{aligned}&n^{\epsilon-1} \leq H \leq n^{-\epsilon};\ n^{-c} < h_s < n^c \text{ for all } s =\\ &1, \ldots, p; \text{ the kernel } K \text{ is a symmetric,}\\ &\text{compactly supported, Hölder-continuous}\\ &\text{probability density; and } K(0) > K(\delta) \text{ for all } \delta > 0.\end{aligned} \tag{7}$$

The above conditions basically ask that each h_s does not converge to 0, or to infinity, too fast, and that $nh_1 \ldots h_{p1} \to \infty$. It would be convenient to further assume that $h_s \to 0$ for $s = 1, \ldots, p_1$, and that $\lambda_s \to 0$ for $s = 1, \ldots, q_1$, however, for practical reasons we choose not to assume that the relevant components are known a priori. Therefore, we shall assume the following condition holds. Defining $\bar{\mu}_g(\bar{x}) = E|\hat{g}(x)\hat{f}(x)]/E[\hat{f}(x)]$,[1] we assume that

$$\begin{aligned}&\int_{\text{supp}w}[\bar{\mu}_g(x) - \bar{g}(\bar{x})]^2 \bar{w}(\bar{x})\bar{f}(\bar{x})d\bar{x},\\ &\text{a function of } h_1, \ldots, h_{p1}, \text{ and } \lambda_1, \ldots, \lambda_{q1},\\ &\text{vanishes if and only if all of the smoothing}\\ &\text{parameters vanish,}\end{aligned} \tag{8}$$

where $\bar{w}(\bar{x}) = \int \tilde{f}(\tilde{x})w(x)d\tilde{x}$. In the appendix we show that (7) and (8) imply that as $n \to \infty$, $h_s \to 0$ for $s = 1, \ldots, p_1$ and $\lambda_s \to 0$ for $s = 1, \ldots, q_1$. Therefore, the smoothing parameters associated with the relevant regressors all vanish asymptotically.

Define an indicator function $I_s(v^d, x^d) = I(v_s^d \neq x_s^d)\prod_{t\neq s, t=1}^{q} I(v_t^d = x_t^d)$. Note that $I_s(v^d, x^d) = 1$ if and only if v^d and x^d differ in their sth component only. Also define $\int d\bar{x} = \sum_{\bar{x}^d} \int d\bar{x}^c$. Letting m_s and m_{ss} denote the first and second derivatives of $m(x)$ with respect to x_s^c $(m = \bar{g}, \bar{f})$, in the appendix we show that the leading term of CV is

$$\begin{aligned}&\int \frac{\kappa^{p1}\sigma^2(\bar{x})}{nh_1 \ldots h_{p1}} w(x)\tilde{R}(x)\tilde{f}(\tilde{x})dx + \int\Bigg(\sum_{s=1}^{q1}\lambda_s \sum_{\bar{v}^d}[I_s(^d,\bar{x}^d)\\ &\times\{\bar{g}(\bar{x}^c,\bar{v}^d) - \bar{g}(\bar{x})\}(\bar{x}^c,\bar{v}d)] + \frac{1}{2}\kappa_2\sum_{s=1}^{p1}h_s^2\\ &\times\{\bar{g}_{ss}(\bar{x})\bar{f}(\bar{x}) + 2\bar{f}_s(\bar{x})\bar{g}_s(\bar{x})\}\Bigg)^2 \bar{f}(\bar{x})^{-1}\bar{w}(\bar{x})d\bar{x},\end{aligned} \tag{9}$$

with $\kappa = \int K(v)^2 dv, \kappa_2 = \int K(v)v^2 dv$, and where $\tilde{R}(x) = \tilde{R}(x, h_{p1+1}, \ldots, h_p, \lambda_{q1+1}, \ldots, \lambda_q)$ is given by $\tilde{R}(x) = v_2(x)/[v_1(x)]^2$, where for $j = 1, 2$, $v_j(x) = E\left(\left[\prod_{s=p1+1}^{p} h_s^{-1} K\left(\frac{X_{is}^c - x_s^c}{h_s}\right)\prod_{s=q1+1}^{q}\lambda_s^{I(X_{is}^d \neq x_s^d)}\right]^j\right)$.

In equation (9) the irrelevant regressor $\tilde{x}$ appears in $\tilde{R}$. By Hölder's inequality, $\tilde{R}(x) \geq 1$ for all choices of x, $h_{p1+1}, \ldots, h_p$, and $\lambda_{q1+1}, \ldots \lambda_q$. Also, $\tilde{R} \to 1$ as $h_s \to \infty$ $(p_1 + 1 \leq s \leq p)$ and $\lambda_s \to 1$ $(q_1 + 1 \leq s \leq q)$. It can be shown that the only smoothing parameter values for which $\tilde{R}(x, h_{p1+1}, \ldots, h_p, \lambda_{q1+1}, \ldots, \lambda_q) = 1$ are $h_s = \infty$ for $p_1 + 1 \leq s \leq p$, and $\lambda_s = 1$ for $q_1 + 1 \leq s \leq q$ (see Hall, Li, & Racine, 2005). Therefore, in order to minimize equation (9), the smoothing parameters corresponding to the irrelevant regressors must all converge to their upper extremities, so that $\tilde{R}(x) \to 1$ as $n \to \infty$ for all $x \in \mathcal{S}$. Thus, the irrelevant components are asymptotically smoothed out.

To analyze the behavior of the smoothing parameters associated with the relevant regressors, we replace $\tilde{R}(x)$ by 1 in equation (9), thus the first term on the right side of equation (9) becomes $\int \frac{\kappa^{p_1}\sigma^2(\bar{x})}{nh_1 \ldots h_{p_1}} \bar{w}(x)d\bar{x}$. Next, defining $a_s = h_s n^{1/(q_1+4)}$ and $b_s = \lambda_s n^{2/(q_1+4)}$, then equation (9) (with $\tilde{R}$ replaced by 1 as its first term since $\tilde{R}(x) \to 1$) becomes $n^{-4/(q_1+4)}\bar{x}(a_1, \ldots, a_{p1}, b_1, \ldots, b_{q1})$, where

$$\begin{aligned}\bar{x}(a_1, \ldots, a_{p1}, b_1, \ldots, b_{q1}) = &\int\Bigg(\sum_{s=1}^{q1} b_s \sum_{\bar{v}^d} I_s(\bar{v}\mathrm{d},\bar{x}^d)(\bar{g}(\bar{x}^c,\bar{v}^d)\\ &- \bar{g}(\bar{x}))\bar{f}(\bar{x}^c,\bar{v}^d) + \frac{1}{2}\kappa_2\sum_{s=1}^{p_1}\\ &\times a_s^2(\bar{g}_{ss}(\bar{x})\bar{f}(\bar{x}) + 2\bar{f}_s(\bar{x})\bar{g}_s(\bar{x}))\Bigg)^2 \bar{f}(\bar{x})^{-1}\bar{w}(\bar{x})d\bar{x}\\ &+ \int\frac{\kappa^p\sigma^2(\bar{x})}{a_1 \ldots a_{p1}}\bar{w}(\bar{x})d\bar{x}.\end{aligned} \tag{10}$$

Let $a_1^0, \ldots, a_{p1}^0, b_1^0, \ldots, b_{q1}^0$ denote values of $a_1, \ldots, a_{p1}, b_1, \ldots, b_{q1}$ that minimize $\bar{x}$ subject to each of them being nonnegative. We require that

$$\begin{aligned}&\text{each } a_s^0 \text{ is positive and each } b_s^0 \text{ nonnegative,}\\ &\text{all are finite and uniquely defined.}\end{aligned} \tag{11}$$

The above analysis leads to the following result.

[1] Note that $\bar{\mu}_g(\bar{x})$ does not depend on $\tilde{x}$, nor does it depend on $(h_{p1+1}, \ldots, h_p, \lambda_{q+1}, \ldots, \lambda_q)$, because the components in the numerator related to the irrelevant regressors cancel with those in the denominator.

TABLE 1.—OUT-OF-SAMPLE PREDICTED PMSE PERFORMANCE MEDIAN [5TH PERCENTILE, AND 95TH PERCENTILE] OF PMSE

n_1	NP CV	NP CV-FR	P int(2)	P int(3)
		$\rho = 0.00$		
100	1.14 [1.02, 1.31]	1.23 [1.09, 1.44]	1.52 [1.19, 2.09]	15.00 [3.94, 61.26]
250	1.07 [0.98, 1.15]	1.11 [1.01, 1.21]	1.13 [1.02, 1.24]	1.69 [1.27, 2.65]
		$\rho = 0.75$		
100	1.13 [1.01, 1.29]	1.17 [1.04, 1.34]	1.47 [1.14, 2.52]	33.42 [3.64, 1055.68]
250	1.06 [0.98, 1.15]	1.09 [1.00, 1.18]	1.12 [1.02, 1.25]	1.71 [1.22, 3.79]

Theorem 2.1 *Assume conditions (1), (5), (7), (8), and (11) hold, and let* $\hat{h}_1, \ldots, \hat{h}_p, \hat{\lambda}_1, \ldots, \hat{\lambda}_q$ *denote the smoothing parameters that minimize CV. Then*

$n^{1/(p_1+4)}\hat{h}_s \to a_s^0$ *in probability for* $1 \le s \le p_1$, $P(\hat{h}_s > C) \to 1$ *for* $p_1 + 1 \le s \le p$ *and for all* $C > 0$, $n^{2/(p_1+4)}\hat{\lambda}_s \to b_s^0$ *in probability for* $1 \le s \le q_1$, $\hat{\lambda}_s \to 1$ *in probability for* $q_1 + 1 \le s \le q$.

The proof of theorem 2.1 is given in the appendix. Theorem 2.1 states that the cross-validated smoothing parameters will behave so that the smoothing parameters for the irrelevant components converge in probability to the upper extremities of their respective ranges. Therefore, all irrelevant regressors are (asymptotically) automatically smoothed out. Next we present the asymptotic normality result for $\hat{g}(x)$.

Theorem 2.2 *Under the same conditions as in theorem 2.1, and letting* x *be an interior point to* $\mathcal{S} = \mathcal{S}^c \times \mathcal{S}^d$ *with* $f(x) > 0$, *then*

$$(n\hat{h}_1 \ldots \hat{h}_{p1})^{1/2}\left[\hat{g}(x) - \bar{g}(\bar{x}) - \sum_{s=1}^{q_1} B_{1s}(\bar{x})\hat{\lambda}_s^2 - \sum_{s=1}^{p_1} B_{2s}(\bar{x})\hat{h}_s^2\right] \to N(0,\Omega(\bar{x})) \text{ in distribution}, \tag{12}$$

where $B_{1s}(\bar{x}) = \Sigma_{\bar{v}}^{\mathrm{d}} I_s(\bar{v}d,\bar{x}^{\mathrm{d}})\,(\bar{g}\,(\bar{x}^c, \bar{v}^{\mathrm{d}}) - \bar{g}(\bar{x}))\bar{f}(\bar{x}^c, \bar{v}^{\mathrm{d}})\bar{f}(\bar{x})^{-1}$, $B_{2s}(\bar{x}) = \frac{1}{2}\kappa_2\left(\bar{g}_{ss}(\bar{x}) + 2\frac{\bar{f}_s(\bar{x})\bar{g}_s(\bar{x})}{f(\bar{x})}\right)$, *and* $\Omega(\bar{x}) = \kappa^{p1}\sigma^2(\bar{x})/\bar{f}(\bar{x})$ *are terms related to the asymptotic bias and variance, respectively.*

Theorem 2.2 follows from theorem 2.1 and its proof is therefore omitted.

Until now we have considered only the case for which the discrete regressors are unordered. If, however, some of the discrete regressors are ordered, one should use alternative kernel functions that reflect this fact. In this case we suggest the use of the following simple kernel function for ordered regressors defined by $K^{\mathrm{d}}(x_s^{\mathrm{d}}, v_s^{\mathrm{d}}) = \lambda_s^{|x_s^{\mathrm{d}} - v_s^{\mathrm{d}}|}$. The range of λ_s is [0,1]. Again, when $\lambda_s = 1$, $K^{\mathrm{d}}(x_s^{\mathrm{d}}, v_s^{\mathrm{d}}) \equiv 1$ for all values of $x_s^{\mathrm{d}}, v_s^{\mathrm{d}} \in \mathcal{S}^{\mathrm{d}}$, and x_s^{d} is completely smoothed out from the regression function.

III. Finite-Sample Behavior

We now assess the effectiveness of our cross-validatory approach. We consider three models: (i) a parametric model (P); (ii) a nonparametric frequency model having cross-validated bandwidths for the continuous regressors (NP CV-FR); (iii) the proposed nonparametric cross-validated method (NP CV).

For $i = 1, \ldots, n$ we generate the following random variables: $(z_{i1}, z_{i2}) \in \{0, 1\}$, $\Pr[z_{i1} = 1] = 0.69$, $\Pr[z_{i2} = 1] = 0.73$, $x_{i1} \sim N(0, 1)$, $x_{i2} \sim N(0, 1)$, and $u_i \sim N(0, 1)$. The regressors x_{1i}, x_{2i}, z_{1i}, z_{2i} vary in their degree of correlation, $\rho = \{0.0, 0.75\}$, while the regressors and u_i are independent of one another. However, not all the regressors are relevant, as we generate y_i according to $y_i = z_{i1} + x_{i1} + \epsilon_i$, $i = 1,2, \ldots, n$, so that z_2 and x_2 are irrelevant. Note that, when $\rho \neq 0$, the relevant *and* irrelevant regressors are correlated. Simulation results clearly demonstrate that cross-validation does indeed smooth out irrelevant regressors regardless of whether they are independent ($\rho = 0$) or correlated ($\rho \neq 0$) with the relevant regressors. We consider samples of size $n_1 = 100$ and 250, then evaluate a model's performance on *independent* data drawn from the same DGP of size $n_2 = 1{,}000$. Predictive performance is computed as $\text{PMSE} = n_2^{-1}\,\Sigma_i(\hat{y}_i - y_i)^2$.

We consider two parametric models, both of which include quadratic terms in x_1 and x_2 along with interaction terms, one having interaction terms of order two (P int(2)) and the other of order three (P int(3)). The nonparametric model is a local constant one with a Gaussian kernel for the continuous regressors and the kernel for the discrete regressors given in section II. All models therefore include the same conditioning information, $(z_{i1},z_{i2},x_{i1},x_{i2})$. PMSE results are presented in table 1.

Table 1 reveals that, in *finite-sample* settings, the proposed nonparametric approach (NP CV) not only has better out-of-sample performance than the cross-validated frequency approach (NP CV-FR), it is also capable of outperforming parametric models containing irrelevant regressors. Next we consider the behavior of the cross-validated bandwidths summarized in table 2.

Table 2 reveals that cross-validation indeed displays a tendency to "smooth out" or remove irrelevant regressors. The irrelevant discrete regressor is smoothed out when its bandwidth $\hat{\lambda}_{z2}$ takes on its upper

TABLE 2.—SUMMARY OF CROSS-VALIDATED BANDWIDTHS FOR THE NP CV ESTIMATOR MEDIAN, [10TH PERCENTILE, 90TH PERCENTILE] OF $\hat{\lambda}$, $\hat{h}$.

n_1	$\hat{\lambda}_{z_1}$	$\hat{\lambda}_{z_2}$	$\hat{h}_{x_1}$	$\hat{h}_{x_2}$
		$\rho = 0.00$		
100	0.05 [0.00, 0.14]	1.00 [0.13, 1.00]	0.39 [0.24, 0.50]	3,072,495.00 [0.76, 20,670,300.00]
250	0.03 [0.00, 0.06]	1.00 [0.21, 1.00]	0.33 [0.23, 0.40]	1,633,310.00 [0.87, 11,928,360.00]
		$\rho = 0.75$		
100	0.00 [0.00, 0.28]	1.00 [0.05, 1.00]	0.36 [0.21, 0.49]	1,938,935.00 [0.52, 12,731,400.00]
250	0.00 [0.00, 0.03]	1.00 [0.15, 1.00]	0.30 [0.20, 0.37]	700,582.50 [0.61, 7,030,172.00]

bound value of 1, while the irrelevant continuous regressor is effectively smoothed out when its bandwidth $\hat{h}_{x2}$ exceeds just a few standard deviations of the data. Note that the median value of $\hat{\lambda}_{z2}$ is indeed 1 for all sample sizes considered, while that for $\hat{\lambda}_{x2}$ is orders of magnitude larger than the standard deviation of x_2 ($\sigma_{x_2} = 1$). Simulation results with higher nonparametric dimensions are qualitatively similar and are available from the authors upon request.

APPENDIX

PROOF OF THEOREM 2.1

In this appendix we provide a sketch of the proof of theorem 2.1. A more detailed proof is given in Hall et al. (2005).

Step (i): Preparations. Letting $g_i = \bar{g}(\bar{x}_i)$, $\hat{g}_{-i} = \hat{g}_{-i}(x_i)$, $\hat{f}_{-i} = \hat{f}_{-i}(x_i)$, and $w_i = w(x_i)$, we have

$$CV(h,\lambda) = \frac{1}{n}\sum_{i=1}^{n}(g_i - \hat{g}_{-i})^2 w_i + \frac{2}{n}\sum_{i=1}^{n} u_i(g_i - \hat{g}_{-i})w_i + \frac{1}{n}\sum_{i=1}^{n} u_i^2 w_i. \tag{A1}$$

The third term on the right side of equation (A1) does not depend on (h, λ). It can be shown that the second term has an order smaller than the first term (see Hall et al., 2005). Therefore, the first term is the leading term of *CV*. Define $\hat{m}_{1,-i}(x_i) = (n-1)^{-1}\sum_{j\neq i}^{n}[\bar{g}(\bar{x}_j) - \bar{g}(\bar{x}_i)]\mathcal{H}(x_j, x_i)$ and let $\hat{m}_{2,-i}(x_i) = (n-1)^{-1}\sum_{j\neq i}^{n} u_j\mathcal{H}(x_j, x_i)$. Then the first term of equation (A1) can be written as

$$\frac{1}{n}\sum_{i=1}^{n}\hat{m}_{1,-i}^2 w_i/\hat{f}_{-i}^2 + \frac{1}{n}\sum_{i=1}^{n}\hat{m}_{2,-i}^2 w_i/\hat{f}_{-i}^2 + \frac{2}{n}\sum_{i=1}^{n}\hat{m}_{1,-i}\hat{m}_{2,-i}w_i/\hat{f}_{-i}^2 \equiv S_1 + S_2 + 2S_3, \tag{A2}$$

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

$$\zeta_n = \left(\sum_{s=1}^{p_1} h_s^2 + \sum_{s=1}^{q_1}\lambda_s\right)^2. \tag{A3}$$

In steps (ii) to (iv) below we show that

$$S_1 = \int\left(\sum_{s=1}^{q_1}\lambda_s\sum_{\bar{v}^d}[I_s(\bar{v}^d,\bar{x}^d)\{\bar{g}(\bar{x}^c,\bar{v}^d) - \bar{g}(\bar{x})\}\bar{f}(\bar{x}^c,\bar{v}^d)] + \frac{1}{2}\kappa_2\sum_{s=1}^{p_1}h_s^2\{g_{ss}(\bar{x})(\bar{x}) + 2\bar{f}_s(\bar{x})g_s(\bar{x})\}\right)^2 \times \bar{f}(\bar{x})^{-1}\bar{w}(\bar{x})d\bar{x} + o_p(\zeta_n), \tag{A4}$$

$$S_2 = \int\frac{\kappa^{p_1}\bar{\sigma}^2(\bar{x})}{nh_1\ldots h_{p1}}w(x)\bar{R}(x)\bar{f}(\bar{x})dx + o_p((nH)^{-1}), \tag{A5}$$

$$S_3 = o_p(\zeta_n + (nH)^{-1}), \tag{A6}$$

where the $o_p(\cdot)$ terms are all uniform in (h, λ) such that

$$n^{\epsilon-1} \leq H \leq n^{-\epsilon}, n^{-c} < h_s < n^c \text{ for all } s = 1,\ldots,q, \text{ and } \lambda_s \in [0,1] \text{ for } 1 \leq s \leq q, \tag{A7}$$

for some $\epsilon \in \left(0, \frac{1}{4+p}\right)$. Therefore, the leading term of *CV* is given by equation (9), obtained from the leading terms of S_1 and S_2.

Letting $\mu_f(x_i) = E(\hat{f}_{-i}(x_i)|x_i)$, then for integers $k \geq 1$, define $H_k = E(\mathcal{H}(x_i, x_j)^k)/(E[\mu_f(x_i)])^k$. It can easily be shown that H_k is bounded below and above by some constants multiplying $H^{-(k-1)}$:

$$c^{-1}H^{-(k-1)} \leq H_k \leq cH^{-(k-1)}, \tag{A8}$$

for some $c > 1$, where $H = h_1\ldots h_{p1}\prod_{s=p1+1}^{p}\min(h_s, 1)$ as defined in condition (6).

Step (ii): Proof of equation (A4). Recall that $\mu_{f,i} = \mu_f(x_i) = E(\hat{f}_{-i}(x_i)|x_i)$, and define S_1^0 by replacing $\hat{f}_{-i}$ by $\mu_{f,i}$ in S_1, that is, $S_1^0 = n^{-1}\sum_{i=1}^{n}\hat{m}_{1,-i}^2 w_i/\mu_{f,i}^2$. We will show that equation (A4) holds true with S_1 replaced by S_1^0, and that $S_1 - S_1^0 = o_p(\zeta_n + (nH)^{-1})$ uniformly in (h, λ). Letting $\mathcal{H}_{ij} = \mathcal{H}(x_i, x_j)$, we write $S_1^0 = G_1 + G_2$, where $G_1 = \frac{1}{n(n-1)^2}\sum_i\sum_{j\neq i}(g_j - g_i)^2\mathcal{H}_{ij}^2 w_i/\mu_{f,i}^2$, and $G_2 = \frac{1}{n(n-1)^2}\sum\sum\sum_{l\neq j\neq i}(g_j - g_i)\mathcal{H}_{ij}(g_l - g_i)\mathcal{H}_{il}w_i/\mu_{f,i}^2$.

We first consider G_2, which can be written as a third-order U-statistic. Define Q_{ijl} as the symmetrized version of $(g_i - g_j)(g_i - g_l)\mathcal{H}_{ij}\mathcal{H}_{il}w_i/\mu_{f,i}^2$ (symmetric in *i,j,l*), let $Q_{ij} = E(Q_{ijl}|x_i, x_j)$, and let $Q_i = E(Q_{ij}|x_i)$. Then by the U-statistic Hoeffding-decomposition

$$G_2 = EQ_1 + \frac{3}{n}\sum_{i=1}^{n}[Q_i - EQ_1] + \frac{6}{n(n-1)}\sum_i\sum_{j>i}[Q_{ij} - Q_i - Q_j + EQ_1] + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i}[Q_{ijl} - Q_{ij} - Q_{il} - Q_{jl} + Q_+Q_j + Q_l - EQ_1] \equiv J_0 + J_1 + J_2 + J_3,$$

where the definition of J_j ($j = 0, \ldots, 3$) should be apparent.

In step (v) below we shall show that h_s for $s = 1, \ldots, p_1$ and λ_s for $s = 1, \ldots, q_1$ all converge to 0 as $n \to \infty$. Therefore, by a Taylor expansion argument, it can be shown that

$$E[(g_j - g_i)\mathcal{H}K_{ij}|x_i = x] = \left(\frac{\kappa_2}{2}\sum_{s=1}^{p_1}[\bar{g}_{ss}(\bar{x})\bar{f}(\bar{x}) + 2\bar{g}_s(\bar{x})\bar{f}_s(\bar{x})]h_s^2 + \sum_{\bar{v}^d}\lambda_s I_s(\bar{v}^d,\bar{x}^d)[\bar{g}(\bar{x}^c,\bar{v}^d) - \bar{g}(\bar{x})]\bar{f}(\bar{x}^c,\bar{v}^d)\right)v_1(x) + o(\zeta_n^{1/2})$$

uniformly in $x \in S$, with (h, λ) as prescribed by condition (A7), and where ζ_n is defined in equation (A3). Therefore

$$J_0 = \int\left(\frac{\kappa_2}{2}\sum_{s=1}^{p_1}[\bar{g}_{ss}(\bar{x})\bar{f}(\bar{x}) + 2\bar{g}_s(\bar{x})\bar{f}_s(\bar{x})]h_s^2 + \sum_{\bar{v}^d}\lambda_s I_s(\bar{v}^d,\bar{x}^d) \times [\bar{g}(\bar{x}^c,\bar{v}^d) - \bar{g}(\bar{x})]\bar{f}(\bar{x}^c,\bar{v}^d)\right)^2\bar{w}(\bar{x})(\bar{x})^{-1}d\bar{x} + o(\zeta_n), \tag{A9}$$

where in the above we have also used $\mu_f(x) = \bar{f}(\bar{x})v_1(x) + O_p(\zeta_n^{1/2})$ uniformly in $x \in \mathcal{S}$, (h, λ).

Next, we consider J_1. Noting that $E(Q_i^{\kappa}) = O(\zeta_n^{\kappa})$, then by Rosenthal's inequality, we have

$$E|J_1|^{2\kappa} \leq n^{-2\kappa}C_k(n^{\kappa}\zeta_n^{2\kappa} + n\zeta_n^{2\kappa}) = O(n^{-\kappa}\zeta_n^{2\kappa}), \tag{A10}$$

where C_κ is a constant. Using Markov's inequality one can show that equation (A10) implies that $J_1 = o_p(\zeta_n)$ uniformly in (h, λ).

For J_2, by noting that $E(Q_{ij}^{\kappa}) = O(\zeta_n^{\kappa/2}H_k)$ and applying Rosenthal's inequality, we obtain

$$E|J_2|^{2\kappa} \le C_\kappa n^{-4\kappa}\{n^{2\kappa}\zeta_n^\kappa H_2^\kappa\} + (s.o.) = O(n^{-2\kappa}\zeta_n^\kappa H_2^\kappa) = O(\zeta_n^\kappa n^{-\kappa}(nH)^{-\kappa}), \tag{A11}$$

where the last equality follows from equation (A8) and (*s.o.*) denotes smaller-order terms. By Markov's inequality, it follows that $J_2 = o_p((nH)^{-1})$ uniformly in (h, λ).

J_3 is a third-order U-statistic. By noting that $E(Q_{ijl}^\kappa) = O(\zeta_n^{\kappa/2} H_\kappa^2)$, it is easy to show that J_3 has an order smaller than that of J_2.

We now consider G_1. Define $\mathcal{D}_{ij} = (1/2)[w_i/\mu_{f,i}^2 + w_j/\mu_{f,j}^2](g_j - g_i)^2\mathcal{K}_{ij}^2$, $\mathcal{D}_i = E[\mathcal{D}_{ij}|x_i]$. Then

$$G_1 = \frac{1}{n-1}\left(E\mathcal{D}_1 + \frac{2}{n}\sum_{i=1}^n [\mathcal{D}_i - E\mathcal{D}_1] + \frac{2}{n(n-1)}\sum\sum_{j>i}[\mathcal{D}_{ij} - \mathcal{D}_i - \mathcal{D}_j + E\mathcal{D}_1]\right) = G_{1,0} + G_{1,1} + G_{1,2}.$$

Then by exactly the same arguments as in the analysis for G_2 above, one can easily show that

$$G_{1,0} = \frac{1}{n-1}E\mathcal{D}_{ij} = O(n^{-1}\zeta_n^{1/2}H_2^1) = O(\zeta_n^{1/2}(nH)^{-1}) = o((nH)^{-1})$$

uniformly in (h, λ).

Noting that $E[\mathcal{D}_i^\kappa] = O(\zeta_n^{\kappa/2}H_{2\kappa})$, by Rosenthal's inequality we obtain

$$E|G_{1,1}|^{2\kappa} \le C_\kappa n^{-4k}(n^\kappa\zeta_n^\kappa H_4^\kappa + n\zeta_n^\kappa H_{4\kappa}) = O(\zeta_n^\kappa n_{-3\kappa}H_4^\kappa) = O(\zeta_n^\kappa(nH)^{-3\kappa})$$

by equation (A8). From this, and using Markov's inequality, it can be shown that $G_{1,1} = o_p((nH)^{-1})$ uniformly in (h, λ). Similarly, one can show, that $G_{1,2} = o_p((nH)^{-1})$ uniformly in (h, λ).

Summarizing the above we have shown that $S_1^0 = G_1 + G_2 = J_0 + o_p(\zeta_n + (nH)^{-1})$, where J_0 is given in equation (A9). Thus, we have shown that equation (A4) holds true with S_1 replaced by S_1^0.

It remains to be shown that $S_1 - S_1^0 = o_p(\zeta_n + (nH)^{-1})$. Defining $\Delta_{f,-i} = E[\hat{f}_{-i}(x_i)|x_i]$, then by Rosenthal's inequality, we have

$$E\left(\left|\frac{\Delta_{f,-i}(x)}{\mu_{f,i}(x)}\right|^{2\kappa}\right) \le C_\kappa n^{-2\kappa}\{n^\kappa H_2^\kappa + nH_{2\kappa}\} = O((nH)^{-\kappa}), \tag{A12}$$

where we have used $H_2 = O(H^{-1})$ by equation (A8). By equation (A12) and Markov's inequality, one can show that $|\frac{\Delta_{f,-i}(x)}{\mu_f(x)}| = o((nH)^{-1/3})$. By exactly the same arguments one can show that $\sup|\hat{m}_{1-i}(x) - E\hat{m}_{1,-i}(x)| = O((nH)^{-1/3})$. Also, a simple Taylor expansion yields $\sup_{1\le i\le n, x, h, \lambda}|E\hat{m}_{1,-i}(x)| = O(\zeta_n^{1/2})$. Therefore,

$$\sup|\hat{m}_{1,-i}(x)| \le \sup|E\hat{m}_{1,-i}(x)| + \sup|\hat{m}_{1,-i} - E\hat{m}_{1,-i}(x)| = O(\zeta_n^{1/2} + (nH)^{-1/3}). \tag{A13}$$

Combining the above results we have shown that

$$S_1 - S_1^0 = \frac{1}{n}\sum_i \hat{m}_{1,-i}^2 w_i\left(\frac{1}{\hat{f}_{-i}^2} - \frac{1}{\mu_{f,i}^2}\right) \le C\sup|\hat{m}_{1,-i}(x)|^2\sup\left|\frac{\Delta_{f,-i}(x)}{\mu_{f,i}(x)}\right| = o_p(\zeta_n + (nH)^{-1}).$$

Step (iii): Proof of equation (A5). Define S_2^0 by replacing $\hat{f}_{-i}$ by $\mu_{f,i}$ in S_2. We will show that equation (A5) holds true with S_2 being replaced by S_2^0, and that $S_2 - S_2^0 = o_p(\zeta_n + (nH)^{-1})$ uniformly in (h, λ). Now,

$$S_2^0 = n^{-1}\sum_i \hat{m}_{2,-i}^2 w_i/\mu_{f,i}^2 = [n(n-1)^2]^{-1}\sum_i\sum_{j\ne i} u_j^2\mathcal{K}_{x_i,x_j}^2 w_i/\mu_{f,i}^2 + [n(n-1)^{-2}]^{-1}\sum_i\sum_{j\ne i}\sum_{l\ne i, l\ne j} u_j u_l\mathcal{K}_{x_i,x_j}\mathcal{K}_{x_i,x_l}w_i/\mu_{f,i}^2 \equiv D_1 + D_2.$$

Note that D_2 can be written as a third-order U-statistic. Let $\mathcal{V}_{ijl}$ denote the symmetrized version of $u_j u_l\mathcal{K}_{xi,xj}\mathcal{K}_{xi,xl}w_i/\mu_{f,i}^2$, and define $\mathcal{V}_{ij} = E[\mathcal{V}_{ijl}|z_i, z_j]$, $z_i = (x_i, u_i)$. Then by the U-statistic Hoeffding decomposition, we have

$$D_2^0 = \frac{6}{n(n-1)}\sum_i\sum_{j>i}\mathcal{V}_{ij} + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i} \times [\mathcal{V}_{ijl} - \mathcal{V}_{ij} - \mathcal{V}_{il} - \mathcal{V}_{jl}] \equiv D_{2,1} + D_{2,2}.$$

Here, $D_{2,1}$ is a second-order degenerate U-statistic, hence it is easy to show that $E[\mathcal{V}_{ij}^\kappa] = O(H_\kappa)$. By the same arguments used in the proof of step (v) (a), it can be shown that

$$E|D_{2,1}|^{2\kappa} \le n^{-4\kappa}C_\kappa\{n^{2\kappa}H_2^\kappa\} + (s.o.) = O(n^{-2\kappa}H_2^\kappa) = O(n^{-\kappa}(nH)^{-\kappa}). \tag{A14}$$

By equation (A14) and the Markov's inequality, one obtains that $D_{2,1} = o_p((nH)^{-1})$ uniformly in (h, λ).

Note that $D_{2,2}$ is a third-order U-statistic, and it is easy to show that $E[V_{ijl}^\kappa] = O(H_\kappa^2)$. By Rosenthal's inequality we know that

$$E|D_{2,2}|^{2\kappa} \le n^{-6\kappa}C_\kappa\{n^{3\kappa}H_2^{2\kappa}\} + (s.o.) = O(n^{-\kappa}(nH)^{-2\kappa}),$$

where (*s.o.*) denotes smaller-order terms (smaller than $O(n^{-\kappa}(nH)^{-2\kappa})$). We observe that $D_{2,2}$ has an order smaller than that of $D_{2,1}$. Therefore, we have shown that $D_2 = o_p((nH)^{-1})$ uniformly in (h, λ).

We now consider D_1. Define $V_{ij} = (1/2)[u_j^2 w_i/\mu_{f,i}^2 + u_i^2 w_j/\mu_{f,j}^2]\mathcal{K}_{x_i,x_j}^2$ and $V_i = E(V_{ij}|x_i, u_i)$. Then $D_1 = \frac{1}{n-1}\{EV_1 + \frac{2}{n}\Sigma_{i=1}^n[V_i - EV_1] + \frac{2}{n(n-1)}\Sigma\Sigma_{j>i}[V_{kij} - V_i - V_j + EV_1]\} \equiv B_0 + B_1 + B_2$.

We first consider $B_0 = (n-1)^{-1}E(V_{ij}) = (n-1)^{-1}E\{w(x_i)\mu_{f,i}^{-2}E[\sigma^2(\bar{x}_j)\mathcal{K}_{ij}^2 x_i]\}$. Applying a Taylor expansion we have

$$E[\sigma^2(x_j)\mathcal{K}_{ij}^2|x_i = x] = \kappa^{p_1}\sigma^2(\bar{x})\bar{f}(\bar{x})\nu_2(x)(h_1\ldots h_{p_1})^{-1} + O(\zeta_n^{1/2}(h_1\ldots h_{p_1})^{-1})$$

uniformly in (x, h, λ). Therefore, using $(n-1)^{-1} = n^{-1} + O(n^{-2})$, we have

$$B_0 = \kappa^{p_1}(nh_1\ldots h_{p1})^{-1}\int \sigma^2(\bar{x})w(x)(\nu_2(x)/\nu_1(x)^2)\bar{f}(\bar{x})dx + o((nH)^{-1}) \tag{A15}$$

uniformly in (x, h, λ).

For B_1, noting that $E(V_i^2) = O(H_4)$, by Rosenthal's inequality we have

$$E|B_1|^{2\kappa} \le C\kappa n^{-4\kappa}(n^\kappa H_4^\kappa + nH_{2\kappa}) = O(n^{-3\kappa}H_4^\kappa) = O((nH)^{-3\kappa}),$$

where the last equality follows from $H_\kappa = O(H^{-(\kappa-1)})$ by equation (A8). It follows by Markov's inequality that $B_1 = o_p((nH)^{-1})$ uniformly in (h, λ).

Noting that nB_2 is a second-order degenerate U-statistic, and that $E(V_{ij}^2) = O(H_4)$, we have

$$E|B_2|^{2\kappa} \le C_\kappa n^{-6\kappa}\{n^{2k}H_4^\kappa\} + (s.o.) = O(n^{-4k}H_4^\kappa) = O(n^{-\kappa}(nH)^{-3\kappa}).$$

We observe that B_2 has an order smaller than that of B_1. Summarizing the above we have shown that

$$S_2^0 = \frac{\kappa^{p1}}{nh_1\ldots h_{p1}}\int \sigma^2(\bar{x})w(x)[\nu_2(x)/\nu_1(x)^2]\bar{f}(\bar{x})dx + o_p((nH)^{-1}) \text{ uniformly in } (h, \lambda). \tag{A16}$$

Finally, following the same proof as for $\sup_{1\le i\le n,\xi}|\frac{\Delta_{f,-i}(x)}{\mu_f(x)}| = o_p((nH)^{-1/3})$, one can show that $\sup_{1\le i\le n,\xi}|\frac{\hat{m}_{2,-i}(x)}{\mu_f(x)}| = o_p((nH)^{-1/3})$, where

$\xi = (x,h, \lambda)$, the supremum is over $x \in S$, and (h, λ) is given in condition (A7). Therefore, we have uniformly in (h, λ), that

$$|S_2 - S_2^0| \le C\left[\sup_{1\le i\le n,\xi}\left|\frac{\hat{m}_{2,-i}(x)}{\mu_f(x)}\right|\right]^2 \sup_{1\le i\le n,\xi}|\frac{\Delta_{f,-i}(x)}{\mu_f(x)}| = o_p((nH)^{-1}).$$

Step (iv): Proof that $S_3 = o_p(\zeta_n + (nH)^{-1})$ uniformly in (h,λ) as prescribed in condition (A7). Define S_3^0 by replacing $\hat{f}_{-i}$ by $\mu_{f,i}$ in S_3, that is, $S_3^0 = n^{-1}\Sigma_i \hat{m}_{1,-i}\hat{m}_{2,-i}w_i/\mu_{f,i}^2 = n^{-3}\Sigma_i\Sigma_{j\ne i} u_j(g_i - g_j)\mathcal{K}^2_{x_i,x_j}w_i/\mu^2_{f,i} + n^{-3}\Sigma\Sigma\Sigma_{l\ne j\ne i} u_j(g_i - g_j)\mathcal{K}_{x_j,x_l}\mathcal{K}_{x_i,x_l}w_i/\mu_{f,i}^2 = M_1 + M_2$.

M_2 can be written as a third-order U-statistic. Letting $\mathcal{R}_{ijl}$ denote the symmetrized version of $u_j(g_i - g_l)\mathcal{K}_{x_i,x_j}\mathcal{K}_{x_i,x_l}w_i/\mu^2_{f,i}$, $\mathcal{R}_{ij} = E[\mathcal{R}_{ijl}|z_j, z_l]$, and $\mathcal{R}_i = E[\mathcal{R}_{ij}|z_i]$, then (note that $E\mathcal{R}_i = 0$)

$$\begin{aligned} M_2 &= \frac{3}{n}\sum_{i=1}^n \mathcal{R}_i + \frac{6}{n(n-1)}\sum\sum_{j>i}[\mathcal{R}_{ij} - \mathcal{R}_i - \mathcal{R}_j] \\ &\quad + \frac{6}{n(n-1)(n-2)}\sum\sum\sum_{l>j>i}[\mathcal{R}_{ijl} - \mathcal{R}_{ij} - \mathcal{R}_{il} - \mathcal{R}_{jl} + \mathcal{R}_i \\ &\quad + \mathcal{R}_j + \mathcal{R}_l] \equiv G_1 + G_2 + G_3. \end{aligned}$$

By noting that the kth moment of $\mathcal{R}_i$ has the same order as the kth moment of $\zeta_n^{1/2}u_i$, we have, for $k \ge 2$, $E[\mathcal{R}_i^k] = O(\zeta_n^{k/2})$. Then, by Rosenthal's inequality,

$$E|G_1|^{2k} \le C_k n^{-2k}\{n^k\zeta_n^k + n\zeta_n^k\} = O(\zeta_n^k n^{-k}). \tag{A17}$$

By Markov's inequality, for $\delta \in (0, \epsilon/2)$ and for all $C > 0$, we have

$$\begin{aligned} P(|G_1| > n^{-\delta}\zeta_n^{1/2}(nH)^{-1/2}) &\le C_k n^{2\delta k}(nH)^k n^{-k} \\ &= O(n^{-(\epsilon-2\delta)k}) = O(n^{-C}) \end{aligned} \tag{A18}$$

uniformly in (h, λ) because $(nH) < n^{1-\epsilon}$ by condition (A7).

Result (A18) implies that $G_1 = o_p(\zeta_n^{1/2}(nH)^{-1/2}) = o_p(\zeta_n + (nH)^{-1})$ uniformly in (h, λ).

For G_2, noting that the kth moment of $\mathcal{R}_{ij}$ has the same order as the kth moment of $\zeta_n^{1/2}u_i(g_j - g_i)\mathcal{K}_{ij}/\mu_{f,i}$, we obtain $E[\mathcal{R}_{ij}^k] = O(\zeta_n^{k/2}H_k)$. Therefore, by exactly the same arguments used to prove step (v) (a), we have

$$\begin{aligned} E|G_2|^{2k} &\le C_k n^{-4k}(n^{2k}\zeta_n^k H_2^k) + (s.o.) = O(\zeta_n^k n^{-2k}H_2^k) \\ &= O(\zeta_n^k n^{-k}(nH)^{-k}). \end{aligned} \tag{A19}$$

By Markov's inequality one can show using equation (A19) that $G_2 = o_p((nH)^{-1})$ uniformly in (h, λ).

For G_3, arguments similar to those used above lead to

$$\begin{aligned} E|G_3|^{2k} &\le C_k n^{-6k}(n^{3k}\zeta_n^k H_2^{2k}) + (s.o.) = O(\zeta_n^k n^{-3k}H_2^{2k}) \\ &= O(\zeta_n^k n^{-k}(nH)^{-2k}). \end{aligned} \tag{A20}$$

Comparing (A20) with (A19) we know that G_3 has an order smaller than that of G_2. Therefore, we have shown that $G_2 + G_3 = o_p(\zeta_n + (nH)^{-1})$ uniformly in (h, λ).

Next, we consider M_1. Defining $R_{ij} = (1/2)[u_j(g_j - g_i)w_i/\mu_{f,i}^2 + \mu_i(g_i - g_j)w_j/\mu_{fj}^2]\mathcal{K}_{ij}$ and $R_i = E[R_{ij}|x_i, u_i]$, then $M_1 = \frac{1}{n-1}(\frac{2}{n}\Sigma_{i=1}^n R_i + \frac{2}{n(n-1)}\Sigma_{i=1}^n[R_{ij} - R_i - R_j])$. Using Rosenthal's and Markov's inequalities, it can be shown that $M_1 = o_p(\zeta_n)$ uniformly in (h, λ). Therefore, we have

$$S_3^0 = M_1 + M_2 = o_p(\zeta_n + (nH)^{-1}) \text{ uniformly in } (h,\lambda).$$

Finally, we have, uniformly in (h, λ), that

$$\begin{aligned} |S_3 - S_3^0| &\le C\left(\sup\left|\frac{\hat{m}_{1,-i}(x)}{\mu_f(x)}\right|\right)\left(\sup\left|\frac{\hat{m}_{2,-i}(x)}{\mu_f(x)}\right|\right)\left(\sup\left|\frac{\Delta_{f,-i}(x)}{\mu_f(x)}\right|\right) \\ &= o_p((nH)^{-1}), \end{aligned}$$

where the supremum is over $1 \le i \le n$, $x \in S$. This completes the proof of step (iv).

Step (v). Here we show that the cross-validated smoothing parameters associated with the relevant regressors all converge to 0 in probability. Since $E(u_i|\{x_j\}_{j=1}^n) = 0$, obviously, the only possible non $o_p(1)$ term in CV that is related to (h, λ) is S_1 defined in equation (A2). Moreover, it can be shown that $S_2 = G_2 + o_p(1)$, where $G_2 = \frac{1}{n(n-1)^2}\Sigma\Sigma\Sigma_{l\ne j\ne i}(g_i - g_j)\mathcal{K}_{ij}(g_i - g_l)\mathcal{K}_{il}w_i/\mu_{f,i}^2$. Furthermore, it is easy to show that $G_2 = E(G_2) + o_p(1)$ uniformly in (h, λ). By the independence of $(\bar{x}_i, u_i)$ with $\bar{x}_i$, we have

$$E(G_2) = \int[\bar{g}(\bar{x}) - \bar{\mu}_g(\bar{x})]^2\bar{w}(\bar{x})\bar{f}(\bar{x})d\bar{x},$$

where $\bar{w}(\bar{x})$ is defined below assumption (8). Hence, we have shown that

$$S_1 = \int[\bar{g}(\bar{x}) - \bar{\mu}_g(\bar{x})]^2\bar{w}(\bar{x})\bar{f}(\bar{x})d\bar{x} + O((nH)^{-1}). \tag{A21}$$

If the smoothing parameters $h_1, \ldots, h_{p1}, \lambda_1, \ldots, \lambda_{q1}$, along with remaining smoothing parameters, minimize CV, but do not all converge in probability to 0, then, by assumption (8), S_1 does not converge to 0, which implies that the probability that the minimum of S_1, over the smoothing parameters, exceeds δ (for some $\delta > 0$). However, choosing $h_1, \ldots, h_{p1}$ to be of size $n^{-1/(p1+4)}$, and $\lambda_1, \ldots, \lambda_{q1}$ to be of size $n^{-2/(p1+4)}$, letting $h_{p1+1}, \ldots, h_p$ diverge to infinity, and letting $\lambda_{q1+1}, \ldots, \lambda_{q1}$ converge to 1, one can easily show that S_1 converges in probability to 0. This contradicts the result obtained in the previous paragraph, and thus demonstrates that

> at the minimum of CV (the equivalent of S_1), the smoothing parameters $h_1, \ldots, h_{p1}, \lambda_1, \ldots, \lambda_{q1}$, for the relevant components of X, all converge in probability to 0.

REFERENCES

Aitchison, J., and C. G. G., Aitken, "Multivariate Binary Discrimination by the Kernel Method," *Biometrika* 63 (1976), 413–420.

Fan, Y., and Q. Li, "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica* 64 (1996), 865–890.

Hall, P., Q. Li, and J. Racine, "Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors," manuscript (2005).

Hall, P., J. Racine, and Q. Li, "Cross-Validation and the Estimation of Conditional Probability Densities," *Journal of the American Statistical Association 99* (2004), 1015–1026.

Lavergne, P., and Q. Vuong, (1996), "Nonparametric Selection of Regressors: The Nonnested Case," *Econometrica* 64 (1996), 207–219.