Econometric Theory, **26**, 2010, 94–118.
doi:10.1017/S0266466609090628

OPTIMAL BANDWIDTH CHOICE FOR ESTIMATION OF INVERSE CONDITIONAL–DENSITY–WEIGHTED EXPECTATIONS

DAVID TOMÁS JACHO-CHÁVEZ
Indiana University

This paper characterizes the bandwidth value (h) that is optimal for estimating parameters of the form $\eta = \mathrm{E}\left[\omega / f_{V|\mathbf{U}}(V|\mathbf{U})\right]$, where the conditional density of a scalar continuous random variable V, given a random vector $\mathbf{U}$, $f_{V|\mathbf{U}}$, is replaced by its kernel estimator. That is, the parameter η is the expectation of ω inversely weighted by $f_{V|\mathbf{U}}$, and it is the building block of various semiparametric estimators already proposed in the literature such as Lewbel (1998), Lewbel (2000b), Honoré and Lewbel (2002), Khan and Lewbel (2007), and Lewbel (2007). The optimal bandwidth is derived by minimizing the leading terms of a second-order mean squared error expansion of an in-probability approximation of the resulting estimator with respect to h. The expansion also demonstrates that the bandwidth can be chosen on the basis of bias alone, and that a simple "plug-in" estimator for the optimal bandwidth can be constructed. Finally, the small sample performance of our proposed estimator of the optimal bandwidth is assessed by a Monte Carlo experiment.

1. INTRODUCTION

An important class of semiparametric estimators, first proposed by Lewbel (1998), involves the use of kernel-based nonparametric estimates in place of the true conditional density in objects of the form

$$\eta = \mathrm{E}\left[\frac{\omega}{f_{V|\mathbf{U}}(V|\mathbf{U})}\right], \tag{1}$$

where $\left\{\omega^{\top}, V, \mathbf{U}^{\top}\right\}$ is a random vector, and $f_{V|\mathbf{U}}(\cdot)$ denotes the conditional density function of a scalar continuous random variable V given the random subvector $\mathbf{U}$. This conditional density function is assumed to be estimated here by the

The author thanks the co-editor, Oliver B. Linton, and three anonymous referees for their helpful comments. In particular, many constructive remarks by the referees on the first draft of the paper greatly enhanced the overall presentation and the exposition of the results. The paper also benefited from discussions with colleagues at the London School of Economics, University of York, Michigan State University, and seminar participants at the Singapore and Canadian Econometric Study Group Meetings and the Far Eastern Meetings of the Econometric Society in Taiwan. Address correspondence to David T. Jacho-Chávez, Department of Economics, Indiana University, Wylie Hall 251, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA; e-mail: djachoch@indiana.edu.

 © Cambridge University Press, 2009 0266-4666/10 $15.00

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

ratio of kernel estimators for $f_{V\mathbf{U}}(\cdot)$ and $f_{\mathbf{U}}(\cdot)$, the joint and marginal densities of $\left(V, \mathbf{U}^\top\right)$ and $(\mathbf{U})$, respectively.

For limited dependent variable models, examples of estimators belonging to this class are Lewbel (1998), Lewbel (2000b), Honoré and Lewbel (2002), Khan and Lewbel (2007), and Lewbel (2007). Results derived in this paper are directly applicable to these estimators. Specifically, if one has a random sample $\left\{\omega_i^\top, v_i, \mathbf{u}_i^\top\right\}$ from the joint distribution of $\left\{\omega^\top, V, \mathbf{U}^\top\right\}$ for $i = 1, \ldots, N$, implementation of any of these estimators requires choosing the numerical value of a bandwidth parameter, h, for the nonparametric kernel estimator of $f_{V|\mathbf{U}}(\cdot)$ in (1). This paper discusses formally how to perform this selection. Given that the asymptotic (first-order) distribution of this semiparametric estimator, $\widetilde{\eta}(h)$, of (1) does not depend on the bandwidth h, any optimal bandwidth formula must be based on a higher-order approximation to such distribution.[1] However, this simple "plug-in" semiparametric estimator may not possess higher-order moments because of its random denominator; see, e.g., Srinivasan (1970).[2] Therefore, we find a formula for the optimal bandwidth that minimizes instead (with respect to h) the leading terms of an asymptotic approximation to

$$\mathrm{E}\left[\left\|\widehat{\eta}(h) - \eta\right\|^2\right], \tag{2}$$

where $\|\cdot\|$ is the standard Euclidean norm, and $\widehat{\eta}(h)$ represents $\widetilde{\eta}(h)$'s in-probability approximation, see Section 2 for definitions and details.[3]

Similar calculations to the ones derived here can be found in the literature of bandwidth selection for average derivative estimation; see, e.g., Härdle, Hart, Marron, and Tsybakov (1992); Härdle and Tsybakov (1993); and Powell and Stoker (1996). Our results are different from theirs in that the optimal bandwidth for semiparametric kernel estimators of (1) can be chosen on the basis of bias alone. In particular, we show that the leading terms in (2) are two biases. One is attributed to the pointwise "smoothing" bias of the kernel density estimator used, and the other to its variance. Linton (1991) called the latter "degrees-of-freedom" bias. Similar results were found by Jones and Sheather (1991) for the kernel-based integrated squared density derivatives estimator of Hall and Marron (1987), and by Ichimura and Linton (2005) for a kernel–based implementation of Hirano, Imbens, and Ridder's (2003) estimators of treatment effects. Linton (1991) discussed a similar result for the variance estimator in the presence of unknown mean. Furthermore, semiparametric estimation of (1) could include discrete elements through its nonparametric component (specifically in $\mathbf{U}$).

As could be expected, another conclusion from this paper is that the derived asymptotically optimal bandwidth, $h_{\mathbf{opt}}$, must shrink more rapidly to zero than it would be for optimal pointwise kernel estimation of $f_{V|\mathbf{U}}(\cdot)$, i.e., estimating this function at a point. In this sense, "asymptotic undersmoothing" is necessary for $\sqrt{N}$–consistent estimation of (1). This feature is explained in the unifying theory of Goldstein and Messer (1992), whose main focus was to highlight differences in the conditions of limiting theory between nonparametric and semiparametric

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

estimation, but did not address the issue of bandwidth selection for particular applications such as the one discussed here.

The remainder of the paper is organized as follows: Section 2, presents the notation and assumptions used throughout the paper, analyzes the sensitivity of kernel-based semiparametric estimator of (1) to the choice of bandwidth, and its kernel's order via a second–order asymptotic expansion of (2), the mean squared error (MSE) of an in-probability approximation to the estimator $\widetilde{\eta}(h)$; and makes explicit the difference between "nonparametric" and "semiparametric" optimal bandwidths. Section 3 discusses how to exploit the in-probability approximation of $\widetilde{\eta}(h)$ in order to construct a simple estimator of the optimal bandwidth and proves its consistency. In Section 4, a Monte Carlo experiment is performed to assess the small-sample behavior of the proposed plug-in estimator of the optimal bandwidth and we compares its performance against the least squares cross-validated bandwidth for the estimation of the nonparametric component $f_{V|\mathbf{U}}(\cdot)$. Section 5 summarizes and discusses the mixed case when there are continuous as well as discrete elements in $\mathbf{U}$. All proofs are presented in the Appendix.

2. ASYMPTOTIC MEAN SQUARE ERROR

First, we introduce some notation and definitions that will aid the latter discussion.

2.1. Framework

We assume that each observation in a data set, $\left\{\omega_i^\top, v_i, \mathbf{u}_i^\top\right\}$, is an independently, identically distributed draw from the joint distribution of $\left\{\omega^\top, V, \mathbf{U}^\top\right\}$ for $i = 1, \ldots, N$, where $\mathbf{U}$ is a $d-1$ vector, V is a scalar, and ω another $\dim(\omega) \times 1$ vector of random variables or known functions of observed random variables. The distributions of $\mathbf{U}$ and $(V, \mathbf{U}^\top)$ are absolutely continuous with respect to some Lebesgue measures, with Radon–Nikodym densities $f_{\mathbf{U}}(\mathbf{u})$ and $f_{V\mathbf{U}}(v, \mathbf{u})$ with bounded supports $\Omega_{\mathbf{U}}$ and $\Omega_{V\mathbf{U}}$, respectively.

For a bandwidth sequence $h \equiv h(N) \to 0$ and $N \to \infty$, the nonparametric estimators of the unknown densities $f_{\mathbf{U}}(\mathbf{u})$ and $f_{V\mathbf{U}}(v, \mathbf{u})$ used here are the well known kernel smoothers:

$$\widehat{f}_{\mathbf{U}}(\mathbf{u}_i; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d-1}} \mathcal{K}\left(\frac{\mathbf{u}_j - \mathbf{u}_i}{h}\right), \quad \text{and} \tag{3}$$

$$\widehat{f}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d}} W\left(\frac{v_j - v_i}{h}\right) \mathcal{K}\left(\frac{\mathbf{u}_j - \mathbf{u}_i}{h}\right), \tag{4}$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

respectively. Here

$$\mathcal{K}(x_1,\ldots,x_{d-1}) = \prod_{j=1}^{d-1} K\left(x_j\right), \qquad x = (x_1,\ldots,x_{d-1}) \in \Re^{d-1},$$

where K and W are one-dimensional bounded symmetric kernel functions that integrate to one. We have also used the "leave-one-out" paradigm in the construction of our smoothers above. A natural estimator[4] for $f_{V\mathbf{U}}(v_i|\mathbf{u}_i)$ is then given by $\widehat{f}_{V|\mathbf{U}}(v_i|\mathbf{u}_i;h) = \widehat{f}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h)/\widehat{f}_{\mathbf{U}}(\mathbf{u}_i;h)$, and its inverse can be estimated[5] by

$$\widehat{l}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h) = \frac{\widehat{f}_{\mathbf{U}}(\mathbf{u}_i;h)}{\widehat{f}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h)},$$

and an estimator of $\eta \equiv \mathrm{E}[\omega/f(v|\mathbf{u})]$ is then given by

$$\widetilde{\eta}(h) = N^{-1}\sum_{i=1}^{N} \omega_i \widehat{l}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h). \tag{5}$$

As previously noted, this estimator may not possess higher-order moments given the presence of the stochastic denominator in $\widehat{l}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h)$. Therefore, by using identity (A.19) in Appendix A, we also define its in-probability approximation, which will be the basis of our analysis below,

$$\widehat{\eta}(h) = N^{-1}\sum_{i=1}^{N} \omega_i \widehat{L}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h), \tag{6}$$

where

$$\widehat{L}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h) = \frac{f_{\mathbf{U}i}}{f_{V\mathbf{U}i}} + 2\frac{\widehat{f}_{\mathbf{U}i}}{f_{V\mathbf{U}i}} - 2\frac{f_{\mathbf{U}i}\widehat{f}_{V\mathbf{U}i}}{f^2_{V\mathbf{U}i}} - \frac{\widehat{f}_{V\mathbf{U}i}\widehat{f}_{\mathbf{U}i}}{f^2_{V\mathbf{U}i}} + \frac{f_{\mathbf{U}i}\widehat{f}^2_{V\mathbf{U}i}}{f^3_{V\mathbf{U}i}}, \tag{7}$$

and $\widehat{f}_{\mathbf{U}i} \equiv \widehat{f}_{\mathbf{U}}(\mathbf{u}_i;h)$, $\widehat{f}_{V\mathbf{U}i} \equiv \widehat{f}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h)$.

Now let us define the following quantities: $\widehat{\delta}_1 \equiv N^{-1}\sum_{i=1}^{N}\omega_i f_{\mathbf{U}i}/f_{V\mathbf{U}i}$, $\widehat{\delta}_2(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i(\widehat{f}_{\mathbf{U}i}/f_{V\mathbf{U}i})$, $\widehat{\delta}_3(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i(f_{\mathbf{U}i}\widehat{f}_{V\mathbf{U}i}/f^2_{V\mathbf{U}i})$, $\widehat{\delta}_4(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i\widehat{f}_{V\mathbf{U}i}\widehat{f}_{\mathbf{U}i}/f^2_{V\mathbf{U}i}$, and $\widehat{\delta}_5(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i f_{\mathbf{U}i}\widehat{f}^2_{V\mathbf{U}i}/f^3_{V\mathbf{U}i}$.

It then follows from identity (A.19) in Appendix A, and some algebra, that

$$\widehat{\eta}(h) = \widehat{\delta}_1 + 2\widehat{\delta}_2(h) - 2\widehat{\delta}_3(h) - \widehat{\delta}_4(h) + \widehat{\delta}_5(h), \tag{8}$$

That is, the in-probability approximation of $\widetilde{\eta}(h)$ can be written as a linear combination of functions of certain U-statistics. In particular, $\widehat{\delta}_2(h)$ and $\widehat{\delta}_3(h)$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

are generic second-order U-statistics:

$$\widehat{\delta}_2(h) = \binom{N}{2}^{-1} \sum_{i<j} \frac{\varpi_{2i} + \varpi_{2j}}{2h^{d-1}} \mathcal{K}\left(\frac{\mathbf{u}_i - \mathbf{u}_j}{h}\right)$$

$$\equiv \binom{N}{2}^{-1} \sum_{i<j} p_2\left(\mathbf{t}_{2i}, \mathbf{t}_{2j}; h\right), \quad \text{and}$$

$$\widehat{\delta}_3(h) = \binom{N}{2}^{-1} \sum_{i<j} \frac{\varpi_{3i} + \varpi_{3j}}{2h^{d}} W\left(\frac{v_i - v_j}{h}\right) \mathcal{K}\left(\frac{\mathbf{u}_i - \mathbf{u}_j}{h}\right)$$

$$\equiv \binom{N}{2}^{-1} \sum_{i<j} p_3\left(\mathbf{t}_{3i}, \mathbf{t}_{3j}; h\right),$$

where $\mathbf{t}_{2i}^{\top} = \left(\varpi_{2i}^{\top}, \mathbf{u}_i^{\top}\right)$ and $\mathbf{t}_{3i}^{\top} = \left(\varpi_{3i}^{\top}, v_i, \mathbf{u}_i^{\top}\right)$, with $\varpi_{2i} \equiv \omega_i / f_{V\mathbf{U}i}$ and $\varpi_{3i} \equiv \omega_i f_{\mathbf{U}i} / f_{V\mathbf{U}i}^2$, respectively. By simple inspection, we notice that these U–statistics 'kernel' functions $p_2(\cdot)$ and $p_3(\cdot)$ are symmetric—that is, $p_2\left(\mathbf{t}_{2i}, \mathbf{t}_{2j}; h\right) = p_2\left(\mathbf{t}_{2j}, \mathbf{t}_{2i}; h\right)$ and $p_3\left(\mathbf{t}_{3i}, \mathbf{t}_{3j}; h\right) = p_3\left(\mathbf{t}_{3j}, \mathbf{t}_{3i}; h\right)$. Powell, Stock, and Stoker (1989) derived first–order limiting theory for this type of linear functions that involves data–dependent (via the bandwidth parameter h) U–statistics. Similarly, we also define $\varpi_{4i} \equiv \omega_i / f_{V\mathbf{U}i}^2$ and $\varpi_{5i} \equiv \omega_i f_{\mathbf{U}i} / f_{V\mathbf{U}i}^3$. It then follows, under conditions explained below, that

$$\eta = \lim_{N \to \infty} \mathrm{E}\left[\widehat{\eta}(h)\right].$$

Also, notice that by construction

$$\widetilde{\eta}(h) - \widehat{\eta}(h) = \widehat{\vartheta}(h),$$

where $\widehat{\vartheta}(h) = N^{-1} \sum_{i=1}^{N} (\widehat{\vartheta}_{1i}(h) - \widehat{\vartheta}_{2i}(h)) \omega_i$, with

$$\widehat{\vartheta}_{1i}(h) \equiv (\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i})^2 (\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}) / (f_{V\mathbf{U}i}^2 \widehat{f}_{V\mathbf{U}i}), \quad \text{and}$$

$$\widehat{\vartheta}_{2i}(h) \equiv (\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i})^3 f_{\mathbf{U}i} / (f_{V\mathbf{U}i}^3 \widehat{f}_{V\mathbf{U}i}).$$

2.2. Assumptions and Sensitivity Analysis

The objective of this paper is to characterize the optimal bandwidth $h_{\mathbf{opt}}$ for computing η. Toward that end, we make the following assumptions.

Assumption A1. The kernels $W : [-1, 1] \to \Re$, and $K : [-1, 1] \to \Re$ are bounded, continuously differentiable, and symmetric such that $\int W(c)\, dc =$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

$\int K(c)\,dc = 1$. There exist constants k_v, k_v^*, k_u, and $k_u^* > 0$, such that $W^2(v) \geq k_v^* \mathbb{I}\left(\|v\| \leq k_v^{-1}\right)$, $K^2(u) \geq k_u^* \mathbb{I}\left(\|u\| \leq k_u^{-1}\right)$.

Assumption A2. Kernels $W(c)$, and $K(c)$ have order P, that is, there exists a positive integer $P \geq 2$ such that $\int c^j W(c)\,dc = \int c^j K(c)\,dc = 0$, $j = 1, \ldots, P-1$, $\int c^P W(c)\,dc = d_W \neq 0$, and $\int c^P K(c)\,dc = d_K \neq 0$.

Assumption A3. The continuous-density functions $f_{\mathbf{u}}(\mathbf{u})$, $f_{V\mathbf{U}}(v, \mathbf{u})$ exist and are bounded away from zero. The functions $\pi_1(\mathbf{u}) = \mathrm{E}[\varpi_1 | \mathbf{U} = \mathbf{u}]$, $\pi_2(\mathbf{u}) = \mathrm{E}[\varpi_2 | \mathbf{U} = \mathbf{u}]$, $\widetilde{\pi}_1(v, \mathbf{u}) = \mathrm{E}[\varpi_1 | V = v, \mathbf{U} = \mathbf{u}]$, $\widetilde{\pi}_2(v, \mathbf{u}) = \mathrm{E}[\varpi_2 | V = v, \mathbf{U} = \mathbf{u}]$, $\pi_l(v, \mathbf{u}) = \mathrm{E}[\varpi_l | V = v, \mathbf{U} = \mathbf{u}]$, for $l = 3, 4, 5$ exist and have bounded continuous partial derivatives up to the order P on their compact supports $\mathbf{\Omega}_{\mathbf{U}} \equiv \prod_{j=1}^{d-1} \left[\underline{U}_j, \overline{U}_j\right]$ and $\mathbf{\Omega}_{v\mathbf{u}} \equiv \left[\underline{V}, \overline{V}\right] \times \mathbf{\Omega}_{\mathbf{U}}$, respectively, for $-\infty < \underline{V} < \overline{V} < \infty$, and $-\infty < \underline{U}_j < \overline{U}_j < \infty$, for $j = 1, \ldots, d-1$.

Assumption A4. $\sup_{\mathbf{\Omega}_{V\mathbf{U}}} \|\omega\| < \infty$, and $\mathrm{E}\left[\|\omega\|^\epsilon | V = v, \mathbf{U} = \mathbf{u}\right]$ has bounded continuous partial derivatives up to order P on their compact support, for $\epsilon = 1, 2, 3, 4$.

Assumption A5. $h_{\mathbf{opt}} \propto N^{-1/(P+d)}$.

Assumptions A1–A3 are standard conditions when using kernel smoothers ensuring the regularity of W, $\mathcal{K}$, $f_{V\mathbf{U}}$, and $f_{\mathbf{U}}$. Assumption A4 will facilitate the proofs and can be relaxed at the expense of more complicated mathematics.[6] Assumption A5 predefines the optimal rate of $h_{\mathbf{opt}}$, which is derived below. The following lemma establishes the validity of $\widehat{\eta}(h)$ as $\widetilde{\eta}(h)$'s in-probability approximation.

LEMMA 1 (In-probability approximation). *Under Assumptions* A1–A5,

$$\sqrt{N}\widehat{\vartheta}(h) = o_p\left(N^{-(P-d)/(P+d)}\right), \qquad as\ N \to \infty.$$

Proof. See Appendix A. ■

We make an additional technical assumption before we state the main result of this paper:

Assumption A6. The vectors of errors $\varepsilon_1 = \varpi_1 - \pi_1(\mathbf{u})$, $\widetilde{\varepsilon}_1 = \varpi_1 - \widetilde{\pi}_1(v, \mathbf{u})$, $\varepsilon_2 = \varpi_2 - \pi_2(\mathbf{u})$, $\widetilde{\varepsilon}_2 = \varpi_2 - \widetilde{\pi}_2(v, \mathbf{u})$, and $\varepsilon_l = \varpi_l - \pi_l(v, \mathbf{u})$ for $l = 3, 4, 5$ are such that $\sigma_1^2(\mathbf{u}) = \mathrm{E}\left[\varepsilon_1^\top \varepsilon_1 \middle| \mathbf{U} = \mathbf{u}\right]$, $\widetilde{\sigma}_1^2(v, \mathbf{u}) = \mathrm{E}\left[\widetilde{\varepsilon}_1^\top \widetilde{\varepsilon}_1 \middle| V = v, \mathbf{U} = \mathbf{u}\right]$, and $\sigma_m^2(\mathbf{u}) = \mathrm{E}\left[\varepsilon_m^\top \varepsilon_m \middle| \mathbf{U} = \mathbf{u}\right]$ for $m = 2, 3, 4, 5$; $\widetilde{\sigma}_{st}(v, \mathbf{u}) = \mathrm{E}\left[\widetilde{\varepsilon}_s^\top \varepsilon_t | V = v, \mathbf{U} = \mathbf{u}\right]$ for $s = 1, 2$ and $t = 3, 4, 5$; and $\sigma_{12}(v, \mathbf{u})$, $\sigma_{34}(v, \mathbf{u})$, $\sigma_{35}(v, \mathbf{u})$, and $\sigma_{45}(v, \mathbf{u})$, where $\sigma_{st}(\mathbf{u}) = \mathrm{E}\left[\varepsilon_s^\top \varepsilon_t \middle| \mathbf{U} = \mathbf{u}\right]$ are bounded on their respective compact supports $\mathbf{\Omega}_{\mathbf{U}}$ and $\mathbf{\Omega}_{V\mathbf{U}}$.

We now formulate the asymptotic expansion of (2), in terms of the dominant components.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

THEOREM 1. *If Assumptions A1–A3, and A6, hold, then*

$$\mathrm{E}\left[\left\|\widehat{\eta}(h)-\eta\right\|^2\right] = O\left(N^{-1}\right) + \left\|\mathfrak{B}_1 h^P + \mathfrak{B}_2 N^{-1} h^{-d}\right\|^2 + O\left(\frac{h^P}{N}+\frac{1}{N^2}\right) + O\left(\frac{h^P}{N}+\frac{1}{N^2 h^{2d}}+h^{2P}\right), \tag{9}$$

as $N \to \infty$, *where*

$$\mathfrak{B}_1 = \int \pi_2(\mathbf{u})\, S_K(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} - \int \pi_3(\upsilon,\mathbf{u})\, S_{WK}(\upsilon,\mathbf{u})\, f_{V\mathbf{U}}(\upsilon,\mathbf{u})\, d\upsilon\, d\mathbf{u}, \tag{10}$$

$$\mathfrak{B}_2 = C_{W\mathcal{K}} \int \pi_3(\upsilon,\mathbf{u})\, f_{V\mathbf{U}}(\upsilon,\mathbf{u})\, d\upsilon\, d\mathbf{u}, \quad \textit{and} \tag{11}$$

$$C_{WK} = \left[\int W^2(c)\, dc\right]\left[\int K^2(c)\, dc\right]^{d-1},$$

$$S_K(\mathbf{u}) = d_K \frac{1}{P!} \sum_{j=1}^{d-1} \frac{\partial^P f_{\mathbf{U}}(\mathbf{u})}{\partial u_j^P}, \quad \textit{and}$$

$$S_{WK}(\upsilon,\mathbf{u}) = \frac{1}{P!}\left[d_W \frac{\partial^P f_{V\mathbf{U}}(\upsilon,\mathbf{u})}{\partial \upsilon^P} + d_K \sum_{j=1}^{d-1} \frac{\partial^P f_{V\mathbf{U}}(\upsilon,\mathbf{u})}{\partial u_j^P}\right].$$

Proof. See Appendix A. ■

The first bias, $\mathfrak{B}_1$, is related to the smoothing bias of the kernel smoother used, while the second bias, $\mathfrak{B}_2$, comes from its pointwise variance. This degrees-of-freedom bias dominates the $O\left(N^{-2}h^{-d}\right)$ variance term that would otherwise appear in the expansion (see Powell and Stoker, 1996, for such calculation).

The result of Theorem 1 can be used to perform a sensitivity analysis with respect to bandwidth choice and order of kernel.

Choice of h. The asymptotically optimal bandwidth is obtained by minimizing (9) on the basis of h. This is achieved when

$$h_{\mathbf{opt}} = C_0 \times N^{-1/(P+d)}, \tag{12}$$

where C_0 is a proportionality constant. The choice of bandwidth equates the leading orders of both biases, $\mathfrak{B}_1 h^P$ and $\mathfrak{B}_2 N^{-1} h^{-d}$. By choosing this bandwidth,

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

we have

$$\mathrm{E}\left[\left\|\widehat{\eta}(h)-\eta\right\|^2\right] = O\left(N^{-1}\right) + \left\|\mathfrak{B}_1 C_0^P + \mathfrak{B}_2 C_0^{-d}\right\|^2 N^{-2P/(P+d)} + O\left(N^{-2}\right) + o(N^{-2P/(P+d)}), \quad \text{as } N \to \infty. \tag{13}$$

That is, the decreasing rate of the best bandwidth optimizing second–order terms is of order $N^{-1/(P+d)}$, which results in an optimal rate of convergence of N^{-1}. In comparison with the leading term, the second term in (13) is not small in general, since their ratio is $O(N^{-(P-d)/(P+d)})$. This means that very large values of N are needed before its influence eventually disappears.

Interestingly, unlike other semiparametric estimators (see Hall and Marron, 1987, and Linton, 1995) the use of "leave-one-out" estimators ((3) and (4)) has not fully eliminated the degrees–of–freedom bias of order $O(N^{-2}h^{-2d})$.[7]

Choice of *P*. If we believe that $f_{V\mathbf{U}}$, $f_{\mathbf{U}}$, $\pi_2(\mathbf{u})$, and $\pi_3(v,\mathbf{u})$ are infinitely many times continuously differentiable, it follows from (13) that the best $O\left(N^{-1}\right)$ rate is not attained unless $P > d$. For example, in the case $d = 2$, we shall use $P > 2$. As Assumption A2 permits, a higher value for P must be chosen for larger values of d. In this sense, the use of oscillating higher–order kernels guarantees the best rate of convergence.

2.3. Nonparametric versus Semiparametric Optimal Bandwidths

For the case $d = 2$, the asymptotic properties of kernel-based estimator $\widehat{f}_{V|\mathbf{U}}(v|\mathbf{u};h)$, used in (5), were first derived by Hyndman, Bashtannyk, and Grunwald (1996), and discussed further by Chen, Linton, and Robinson (2001). When $d > 2$, it follows from their results that the integrated MSE–minimizing optimal bandwidth is

$$h^+_{\mathbf{opt}} \propto N^{-1/(2P+d)}. \tag{14}$$

A direct comparison with (12) indicates that in the semiparametric case, the optimal bandwidth shrinks to 0 at a faster rate of its nonparametric component's optimal bandwidth $h^+_{\mathbf{opt}}$. This phenomenon is known as asymptotic undersmoothing. Other semiparametric estimators sharing this feature are Robinson (1988), Powell, Stock, and Stoker (1989), Härdle and Stoker (1989), and Härdle, Hart, Marron, and Tsybakov (1992), among others.

It should also be noticed that this comparison does not imply that $h_{\mathbf{opt}}$ is numerically smaller than $h^+_{\mathbf{opt}}$ in any particular case or sample size. Particularly, let

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

A_0 be the proportionality constant in (14).[8] It then follows that

$$h_{\mathbf{opt}} = D_0 \mathrm{E}_N h^+_{\mathbf{opt}}, \quad \text{where}$$

$$D_0 = C_0 / A_0,$$

$$\mathrm{E}_N = N^{-P/[(2P+d)(P+d)]}.$$

D_0 is an adjustment factor that depends on the underlying structure of the bias and variance of $\widehat{f}_{V|\mathbf{U}}$ and $\widetilde{\eta}$. In general, $D_0 \lesseqgtr 1$, and it does not vary with sample size. On the other hand, the adjustment term for sample size, E_N, is always less than 1 when $N \geq 2$. Therefore, whether $h_{\mathbf{opt}}$ is larger or smaller than $h^+_{\mathbf{opt}}$ will depend on C_0 being larger or smaller than A_0, and this, in turn, relies on the bias and variance structure in a particular application.

3. OPTIMAL PLUG-IN BANDWIDTH ESTIMATOR

If we knew $\mathfrak{B}_1$ and $\mathfrak{B}_2$ in (13), we could define C_0 (and therefore $h_{\mathbf{opt}}$) by the following minimization problem:

$$C_0 = \underset{C_0 \in \Re_{++}}{\arg\min} \left\| \mathfrak{B}_1 C_0^P + \mathfrak{B}_2 \frac{1}{C_0^d} \right\|^2.$$

As these quantities are unknown in general, a feasible procedure will be to replace them by consistent estimators based on empirical implementations of (10) and (11). These estimators for the smoothing and degrees-of-freedom bias terms are denoted here as $\widehat{\mathfrak{B}}_1$, and $\widehat{\mathfrak{B}}_2$, respectively. However, with a sample of practical size, any kernel-based estimator may be affected by boundary effects that are endemic in kernel density estimation; see Silverman (1986). In view of Assumption A3, this technicality is resolved here by using a known asymptotic trimming function, $a_\tau(\upsilon, \mathbf{u})$ in their construction;[9] that is

$$\widehat{\mathfrak{B}}_1(h_0) = \frac{\widetilde{\eta}(\Delta h_0) - \widetilde{\eta}(h_0)}{h_0^P\left(\Delta^P - 1\right)}, \tag{15}$$

$$\widehat{\mathfrak{B}}_2(h_*) = \frac{C_{W\mathcal{K}}}{N} \sum_{i=1}^{N} \widehat{\varpi}_{*3\tau i}, \tag{16}$$

where $\widehat{\varpi}_{*3\tau i} = w_i a_\tau(\upsilon_i, \mathbf{u}_i) \widehat{f}_{*\mathbf{U}i} / \widehat{f}^2_{*V\mathbf{U}i}$, $\varpi_{3\tau i} = w_i a_\tau(\upsilon_i, \mathbf{u}_i) f_{\mathbf{U}i} / f^2_{V\mathbf{U}i}$, and Δ is a known constant that is greater than 1. Here we have used $\widehat{f}_{*V\mathbf{U}i} \equiv \widehat{f}_{V\mathbf{U}}(\upsilon_i, \mathbf{u}_i; h_*)$, $\widehat{f}_{*\mathbf{U}i} \equiv \widehat{f}_{\mathbf{U}}(\mathbf{u}_i; h_*)$, $f_{V\mathbf{U}i} \equiv f_{V\mathbf{U}}(\upsilon_i, \mathbf{u}_i)$, and $f_{\mathbf{U}i} \equiv f_{\mathbf{U}}(\mathbf{u}_i)$ in order to ease notation, where $\widehat{f}_{\mathbf{U}}(\cdot)$, and $\widehat{f}_{V\mathbf{U}}(\cdot)$ are given by (1), and (2), respectively. The estimator $\widetilde{\eta}(\cdot)$ is like (5), after replacing ω_i by $\omega_i a_\tau(\upsilon_i, \mathbf{u}_i)$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

everywhere. The estimator (15) is similar to the average derivative as proposed by Powell and Stoker (1996). The quantities h_* and h_0 are pilot bandwidths that must be chosen beforehand.

The type of asymptotic trimming used here is that proposed by Lewbel (2000a):

$$a_\tau(v_i, \mathbf{u}_i) = \mathbb{I}\left(v_i \in \left[\underline{V} + \tau, \overline{V} - \tau\right]\right) \prod_{j=1}^{d-1} \mathbb{I}\left(\mathbf{u}_i^{[j]} \in \left[\underline{U}_j + \tau, \overline{U}_j - \tau\right]\right), \tag{17}$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if its argument is true and zero otherwise, the values $\underline{U}_j$ and $\overline{U}_j$, were defined in Assumption A3, and $\mathbf{u}^{[j]}$ refers to the jth element of the vector $\mathbf{u}$. By using this type of trimming in the construction of our estimators (15) and (16), we set to zero all terms in these averages that have observations within a distance τ of the boundary of the support where bias of the kernel estimators are of a different order than for the interior points. The following assumption guarantees that the trimming induced bias goes to zero rapidly, so the consistency of the estimators is not affected.

Assumption A7. The value τ is such that $h_0/\tau \to 0$, and $N\tau^2 \to 0$ as $N \to \infty$.

This trimming has a disadvantage in that it requires knowledge of the support of $(v, \mathbf{u})$. Nevertheless, this support could be estimated in practice. For example, Khan and Lewbel (2007) proposed a data-dependent trimming function, by replacing $\underline{U}_j$, $\overline{U}_j$, $\underline{V}$, and $\overline{V}$ in (17), by the observed maximums and minimums from a sample of N observations of the corresponding variables. They showed that this data-dependent feasible trimming function is asymptotically equivalent to (17). Their result is applicable in situations where the boundary of the support is unknown, and τ equals the bandwidth used in the kernel estimators above.

Consequently, the optimal bandwidth is estimated as

$$\widehat{h}_{\mathbf{opt}} = \widehat{C}_0 \times N^{-1/(P+d)}, \quad \text{where}$$

$$\widehat{C}_0 = \underset{C_0 \in \Re_{++}}{\arg\min} \left\| \widehat{\mathfrak{B}}_1 C_0^P + \widehat{\mathfrak{B}}_2 \frac{1}{C_0^d} \right\|^2. \tag{18}$$

An interesting characteristic of estimators (15) and (16) is that they do not require estimation of higher-order derivatives of unknown functions. This feature makes their calculation computationally very simple. Likewise, the minimization problem in (18) is also computationally straightforward, because it only involves a univariate numerical search over strictly positive real numbers. The consistency of this procedure is ensured by the following proposition:[10]

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

PROPOSITION 1. *Let Assumptions* A1–A3, A4, A6, *and* A7 *hold. If* $h_* \to 0$, $h_0 \to 0$, *with* $Nh_*^d \to \infty$, *and* $Nh_0^{2P+d} \to \infty$ *as* $N \to \infty$, *then*

$$\widehat{\mathfrak{B}}_1(h_0) \xrightarrow{p} \mathfrak{B}_1,$$

$$\widehat{\mathfrak{B}}_2(h_*) \xrightarrow{p} \mathfrak{B}_2.$$

Proof. See Appendix A. ■

An important part of this estimator of the optimal bandwidth is the choice of pilot bandwidths h_*, h_0, constants Δ and δ, and trimming parameter τ. Having to select other tuning parameters is an unappealing feature, but practical guidance is provided here: Given the conditions on h_*, an obvious way of choosing this bandwidth would be by standard cross-validation methods, see Section 4.[11] The resulting bandwidth, $\widehat{h}_*$, would be of order $N^{-1/(2P+d)}$. We can then set $h_0 = \widehat{h}_* \times N^{\delta}$, where $0 < \delta < 1/(2P+d)$. As a result, only $\Delta > 1$ and $\tau \geq 0$ are left to be chosen. In practice, for a fixed number of observations, a feasible approach would be to fix the value τ, and choose a high value of Δ and then decrease it until $\widehat{\mathfrak{B}}_1$ does not vary significantly. We could iterate the above procedure until certain predefined convergence criteria are met.

A technical proviso explained by Powell and Stoker (1996) for the estimated optimal bandwidth of the average derivative estimator is also applicable in this framework. That is, we have not shown that Assumption A will guarantee that the proposed plug-in estimator $\widehat{\eta}(\widehat{h}_{\mathbf{opt}})$ is an in-probability approximation to $\widetilde{\eta}(h_{\mathbf{opt}})$. Firstly, the calculation of $\widetilde{\eta}$ itself would be subject to some trimming with a fixed-size sample. Doing this alone will increase the MSE by the square of the trimming-induced bias. Secondly, the (stochastic) bandwidth $\widehat{h}_{\mathbf{opt}}$ was calculated using the same data as in the construction of $\widetilde{\eta}$. All the calculations used to derive the asymptotic MSE expansion in Theorem 1 implicitly assume a fixed rather than a stochastic value of h. From this, it does not immediately follow that $\widehat{h}_{\mathbf{opt}}$ will be of the same order as $h_{\mathbf{opt}}$. Although addressing this question is beyond the scope of this paper, it is possible that solutions to this problem discussed in Powell and Stoker (1996), in the framework of density-weighted average derivative estimators, might be applicable.

4. MONTE CARLO EXPERIMENTS

Section 4 reports the results of a small-scale Monte Carlo investigation of the finite sample behavior of our proposed plug-in estimator for the optimal bandwidth and the behavior of the associated estimated η's. Samples were generated from the joint distribution given by

$$f_{VU}(v,u) = f_{V|U}(v|u)\, f_U(u),$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

where

$$f_{V|U}(v|u) = \frac{1}{\sqrt{2\pi\sigma^2(u)}} \exp\left\{-\frac{1}{2}\left[\frac{v - m(u)}{\sigma(u)}\right]^2\right\},$$

$$m(u) \equiv c/\left(1+u^2\right), \sigma^2(u) = 1/\left(1+u^2\right), \quad \text{and}$$

$$f_U(u) = \exp[-u/b]\exp\{-\exp[-u/b]\}/b.$$

For simplicity, we set $b \in \{3/4, 1\}$ and $c \in \{0, 1\}$, giving us a total of four different data-generating processes. Figure 1 displays their joint density functions. Notice that these designs are such that there is conditional mean and variance dependence between V and U, and therefore their joint densities are both multimodal and skewed. They also have unbounded supports.

The variable ω is set such that the object of interest becomes

$$\eta = \mathrm{E}\left[\mathbb{I}[-1 \le V \le 1] \times \mathbb{I}[-1 \le U \le 1]/f_{V|U}(V|U)\right].$$

As suggested in Section 2.2, we use a fourth-order Gaussian kernel ($P = 4$), $K(t) = \left(3/2 - t^2/2\right)\exp\left(-t^2/2\right)/\sqrt{2\pi}$ and set $W \equiv K$. Their associated constants are $d_K = -3$ and $C_K = 27/(32\sqrt{\pi})$.

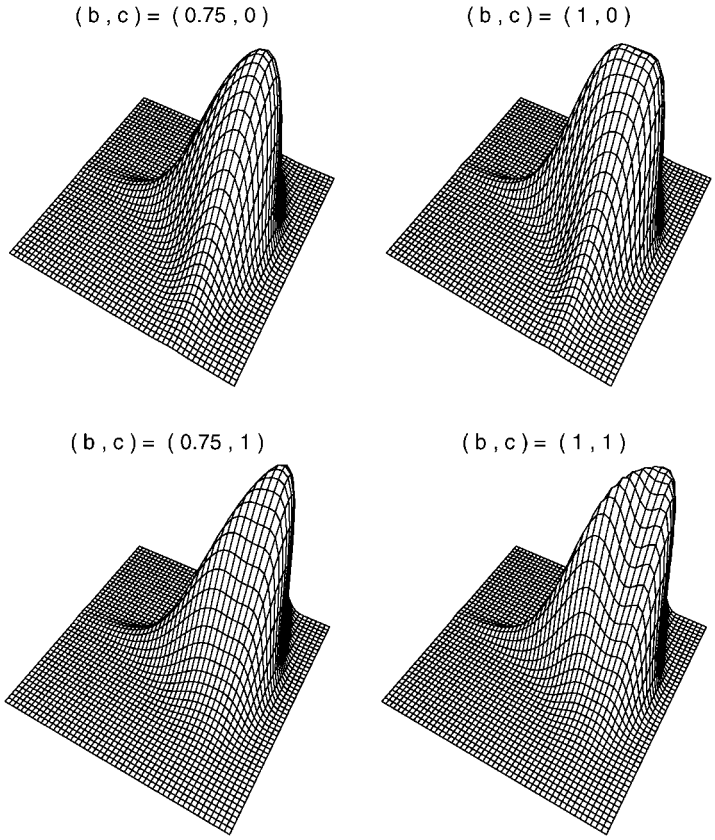
In each of 2,000 replications and for each sample size $N = 50$, 200, and 800, we proceed to calculate (14) by standard least squares cross-validation. Notice that the resulting bandwidth, $\widehat{h}^{+}_{\mathbf{opt}}$, is not optimal for η as discussed in Section 2.2, but it is used as an input in the proposed estimator because of its desirable asymptotic properties; see, e.g., Hall, Racine, and Li (2004). We then set $h_0 = \widehat{h}^{+}_{\mathbf{opt}} N^{\delta}$ and $h_* = \widehat{h}^{+}_{\mathbf{opt}}$. The other tuning parameters are kept constants throughout the experiments at $\delta = 1/12$, $\tau = 0$ (no trimming), and $\Delta = 2$.

The small sample performance of the estimators (15), and (16) is reported in Table 1. The first two rows report the theoretical biases for each design calculated using equations (10) and (11). Estimator (15) seems to be unbiased and displays less variation than its degrees-of-freedom counterparts.

In all designs, the degrees-of-freedom bias is numerically very large compared to the smoothing bias. Similar calculations for other designs (see Jacho-Chávez, 2006) also show these patterns. This lends support to the use of an explicit bias correction mechanisms as in Ichimura and Linton (2005), and it remains a topic for future research.

Table 2 shows the performance of the proposed estimator (18). The theoretical optimal bandwidths for η, $h_{\mathbf{opt}}$, are reported in the first row for each design. They are two to six times larger than their average pointwise optimal counterparts. A larger bandwidth is needed when V is not only conditionally heteroskedastic with respect to U, but its mean also depends on U. This table also shows that the proposed plug-in estimator, $\widehat{h}_{\mathbf{opt}}$, performs fairly well in most circumstances. Although the bandwidths' simulated standard deviations increase when the correlation between V and U increases, this dispersion is offset as the sample sizes increases.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press



Note: Each graph displays the joint density of the Monte Carlo designs implied by different combinations of b and c.

FIGURE 1. Joint density functions of the Monte Carlo designs.

The respective MSEs are presented in Table 3. It should be noticed that the main component of these simulated MSEs is bias instead of variance for most cases. In particular, both simulated smoothing and degrees-of-freedom biases, see Table 1, are considerably larger than the simulated variance in all designs. The use of either the theoretical or estimated optimal bandwidths dominates the use of that predicted by the standard least squares cross-validated bandwidth in terms of MSE, for all sample sizes, and designs. The MSEs associated to the estimated optimal bandwidths are numerically very close to the simulated theoretical ones. Finally, it is worth mentioning that the simulated variance of all estimators is not sensible to the choice of bandwidth, and not their associated simulated bias as predicted in this paper.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

5. CONCLUSIONS AND EXTENSION

A crucial part of estimators with a nonparametric component is the choice of the smoothing parameter. Our main objective in this paper is to provide some guidance for choice of bandwidth for a class of semiparametric estimators that employ kernel estimators in the form of inverse conditional-density-weighted averages. By exploiting the fact that these estimators can be approximated in-probability by a linear combination of functions of U-statistics, we derive a formula for the optimal bandwidth based on a second-order mean squared error expansion of this approximation. The derived formula for the optimal bandwidth equates the order of magnitude arising from the square of the sum of two biases: a smoothing bias and a degrees-of-freedom bias. This formula shows that the optimal bandwidth for estimating the parameter of interest must decrease toward zero at a faster rate than the optimal for its nonparametric component. In this

TABLE 1. Monte Carlo bias decomposition

	$N = 50$		$N = 200$		$N = 800$	
	Mean	SD	Mean	SD	Mean	SD
$b = 3/4, c = 0$						
$\mathfrak{B}_1$	0.4594	–	0.4594	–	0.4594	–
$\mathfrak{B}_2$	2.5440	–	2.5440	–	2.5440	–
$\widehat{\mathfrak{B}}_1$	0.4593	0.3992	0.4593	0.1561	0.4593	0.1130
$\widehat{\mathfrak{B}}_2$	2.5119	0.4507	2.6009	0.1800	2.5926	0.0833
$b = 3/4, c = 1$						
$\mathfrak{B}_1$	−0.0064	–	−0.0064	–	−0.0064	–
$\mathfrak{B}_2$	4.5834	–	4.5834	–	4.5834	–
$\widehat{\mathfrak{B}}_1$	−0.0060	0.5196	−0.0068	0.2702	−0.0062	0.1581
$\widehat{\mathfrak{B}}_2$	3.3157	1.5739	4.1065	1.2633	4.4131	0.4784
$b = 1, c = 0$						
$\mathfrak{B}_1$	0.1842	–	0.1842	–	0.1842	–
$\mathfrak{B}_2$	2.5440	–	2.5440	–	2.5440	–
$\widehat{\mathfrak{B}}_1$	0.1846	0.5401	0.1839	0.1973	0.1846	0.1318
$\widehat{\mathfrak{B}}_2$	2.5761	0.4462	2.6250	0.1882	2.6126	0.0886
$b = 1, c = 1$						
$\mathfrak{B}_1$	−0.4506	–	−0.4506	–	−0.4506	–
$\mathfrak{B}_2$	4.5834	–	4.5834	–	4.5834	–
$\widehat{\mathfrak{B}}_1$	−0.4506	0.5102	−0.4507	0.3067	−0.4506	0.1784
$\widehat{\mathfrak{B}}_2$	3.2984	1.8613	4.0548	1.2418	4.3382	0.4913

Notes: Means and standard deviations (SD) are based on 2,000 replications. The theoretical smoothing bias (2.8) and degrees-of-freedom bias, (2.9) for each design are presented in the first two rows under each combination of b and c. The proposed estimators (3.1) and (3.2) are calculated as explained in the main text, by setting auxiliary bandwidths as $h_0 = \widehat{h}^{+}_{\mathbf{opt}} N^{1/12}$ and $h_* = \widehat{h}^{+}_{\mathbf{opt}}$, as described in Section 3, where $\widehat{h}^{+}_{\mathbf{opt}}$ is found by standard least squares cross-validation.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

sense, asymptotic undersmoothing (as explained in Powell and Stoker, 1996) is needed.

A plug-in estimator of the optimal bandwidth is also constructed exploiting the semiparametric estimator's biases formulas. The problem of random denominators is also addressed in the construction of the proposed estimator through the use of a trimming function. This trimming function, proposed by Lewbel (2000a), is set to give zero-weights in the averages to observations that are within a certain distance of the boundary of the observed support of the distribution. This estimator is shown to perform fairly well in small samples in a Monte Carlo experiment.

5.1. Discrete Conditioning Variables

It is also straightforward to adapt results of this paper to the situation in which the conditioning variables, $\mathbf{U}$, have continuous as well as discrete components. In this mixed case scenario, when using indicator functions to handle the discrete variables, similar conditions to Assumptions A1–A6 have to be imposed on the continuous part of the problem, but no new techniques are required to show that the

TABLE 2. Plug-in optimal bandwidth estimator

	$N = 50$		$N = 200$		$N = 800$	
	Mean	SD	Mean	SD	Mean	SD
$b = 3/4, c = 0$						
$h_{\mathbf{opt}}$	0.6174	–	0.4900	–	0.3889	–
$\widehat{h}_{\mathbf{opt}}$	0.6162	0.0172	0.4918	0.0057	0.3902	0.0022
$\widehat{h}^{+}_{\mathbf{opt}}$	0.2877	0.0295	0.2589	0.0138	0.2259	0.0062
$b = 3/4, c = 1$						
$h_{\mathbf{opt}}$	1.5588	–	1.2372	–	0.9820	–
$\widehat{h}_{\mathbf{opt}}$	1.5116	0.3666	1.2125	0.3023	0.9826	0.2413
$\widehat{h}^{+}_{\mathbf{opt}}$	0.3001	0.0294	0.2701	0.0132	0.2374	0.0062
$b = 1, c = 0$						
$h_{\mathbf{opt}}$	0.7190	–	0.5706	–	0.4529	–
$\widehat{h}_{\mathbf{opt}}$	0.7202	0.0209	0.5736	0.0073	0.4550	0.0035
$\widehat{h}^{+}_{\mathbf{opt}}$	0.2767	0.0327	0.2432	0.0144	0.2118	0.0062
$b = 1, c = 1$						
$h_{\mathbf{opt}}$	0.7669	–	0.6087	–	0.4831	–
$\widehat{h}_{\mathbf{opt}}$	0.7262	0.0501	0.5965	0.0240	0.4787	0.0088
$\widehat{h}^{+}_{\mathbf{opt}}$	0.2972	0.0330	0.2598	0.0142	0.2266	0.0062

Notes: Means and standard deviations (SD) are based on 2,000 replications. The theoretical optimal bandwidth for each design, $h_{\mathbf{opt}}$, are presented in the first row under each combination of b and c. The proposed plug-in bandwidth, $\widehat{h}_{\mathbf{opt}}$, is calculated by setting auxiliary bandwidths as $h_0 = \widehat{h}^{+}_{\mathbf{opt}} N^{1/12}$ and $h_* = \widehat{h}^{+}_{\mathbf{opt}}$, as described in Section 3, where $\widehat{h}^{+}_{\mathbf{opt}}$ is found by standard least squares cross-validation.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

TABLE 3. Simulated mean squared error decomposition

	$N = 50$			$N = 200$			$N = 800$		
	Bias	SD	Sqrt. MSE	Bias	SD	Sqrt. MSE	Bias	SD	Sqrt. MSE
$b = 3/4$, $c = 0$									
$\widehat{\eta}(h_{\mathbf{opt}})$	−0.0393	0.1681	0.1189	−0.0127	0.0804	0.0536	−0.0045	0.0381	0.0265
$\widehat{\eta}(\widehat{h}_{\mathbf{opt}})$	−0.0418	0.1712	0.1220	−0.0128	0.0810	0.0542	−0.0044	0.0382	0.0264
$\widehat{\eta}(\widehat{h}^{+}_{\mathbf{opt}})$	−0.3560	0.1623	0.3560	−0.1088	0.0777	0.1095	−0.0336	0.0360	0.0357
$b = 3/4$, $c = 1$									
$\widehat{\eta}(h_{\mathbf{opt}})$	0.0587	0.2333	0.1656	−0.0311	0.1056	0.0750	−0.0605	0.0495	0.0609
$\widehat{\eta}(\widehat{h}_{\mathbf{opt}})$	0.0684	0.3038	0.1863	−0.0137	0.1673	0.0860	−0.0480	0.0818	0.0565
$\widehat{\eta}(\widehat{h}^{+}_{\mathbf{opt}})$	−0.5130	0.1701	0.5130	−0.1952	0.0953	0.1952	−0.0584	0.0447	0.0587
$b = 1$, $c = 0$									
$\widehat{\eta}(h_{\mathbf{opt}})$	−0.0118	0.1679	0.1153	−0.0043	0.0830	0.0551	−0.0029	0.0403	0.0272
$\widehat{\eta}(\widehat{h}_{\mathbf{opt}})$	−0.0114	0.1714	0.1186	−0.0039	0.0837	0.0555	−0.0029	0.0405	0.0273
$\widehat{\eta}(\widehat{h}^{+}_{\mathbf{opt}})$	−0.3822	0.1741	0.3822	−0.1286	0.0848	0.1286	−0.0418	0.0387	0.0435
$b = 1$, $c = 1$									
$\widehat{\eta}(h_{\mathbf{opt}})$	−0.1238	0.2048	0.1667	−0.0606	0.1019	0.0800	−0.0315	0.0485	0.0389
$\widehat{\eta}(\widehat{h}_{\mathbf{opt}})$	−0.1312	0.2080	0.1744	−0.0626	0.1016	0.0811	−0.0314	0.0482	0.0390
$\widehat{\eta}(\widehat{h}^{+}_{\mathbf{opt}})$	−0.4912	0.1820	0.4912	−0.2067	0.1001	0.2067	−0.0684	0.0460	0.0685

Note: Biases, standard deviations (SD), and squared root mean squared error (Sqrt. MSE) are based on 2,000 replications.

order of magnitude of the optimal bandwidth only depends on the number of continuously distributed elements of the random vector $\left(V, \mathbf{U}^{\top}\right)$; see Jacho-Chávez (2006) for details. Other possibilities include smoothing the discrete components, as in Hall, Racine, and Li (2004), which remains a topic for future research.

NOTES

1. See Lewbel (1998), Lewbel (2000a), Lewbel (2000b), Honoré and Lewbel (2002), and Khan and Lewbel (2007) for precise derivations.

2. We thank the co-editor, Oliver B. Linton, for pointing this out.

3. Similarly, we could replace $\|\mathbf{a}\|$ everywhere in this paper by $\|\mathbf{a}\|_{\mathbf{W}} = \mathbf{a}^{\top}\mathbf{W}\mathbf{a}$, where $\mathbf{W}$ is any positive semidefinite weighting matrix. The results will not change.

4. This estimator was first proposed by Rosenblatt (1969), for the case $d = 2$, and later analyzed by Hyndman, Bashtannyk, and Grunwald (1996).

5. For notational simplicity, the same bandwidth is assumed to be used for both the numerator, otherwise b, and denominator, otherwise h. Results in Section 2.2 can be readily adapted to this situation by setting $b \equiv ch$, where c is a proportionality constant. We thank an anonymous referee for pointing this out.

6. Results in Sherman (1994) can potentially be used to relax this support restriction. We would like to thank an anonymous referee for pointing this out.

7. Ichimura and Linton (2005) proposed an explicit bias correction mechanism that indeed "knocked" this term out, allowing for a smaller MSE for Hirano, Imbens, and Ridder's (2003) estimator. This method can potentially be adapted to our framework.

8. See Bashtannyk and Hyndman (2001) and Chen, Linton, and Robinson (2001) for derivations.

9. Another possibility would be to use boundary kernels; see Gasser, Müller, and Mammitzsch (1985). Fernandes and Monteiro (2005) derived the asymptotic behavior of asymmetric kernel functionals.

10. An alternative estimator for $\mathfrak{B}_2$ is given by

$$\binom{N}{2}\sum_{i<j}\left(\frac{\widehat{\varpi}_{*3\tau i}\,\widehat{f}^{-1}_{*V\mathbf{U}i} + \widehat{\varpi}_{*3\tau j}\,\widehat{f}^{-1}_{*V\mathbf{U}j}}{4h_0^d}\right) W^2\left(\frac{v_i - v_j}{h_0}\right)\mathcal{K}^2\left(\frac{\mathbf{u}_i - \mathbf{u}_j}{h_0}\right),$$

and its consistency can be proven by the exact same arguments used in this section.

11. Wand and Jones (1995, Ch. 3 and 4), described in great detail many other (computationally simpler) bandwidth selection procedures that could be used instead.

REFERENCES

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

Chen, X., O.B. Linton, & P.M. Robinson (2001) The estimation of conditional densities. In M.L. Puri (ed.) *Asymptotics in Statistics and Probability: Papers in Honor of George Gregory Roussas*, 1st ed. pp. 71–84. VSP International Science.

Collomb, G. & W. Härdle (1986) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: Kernel regression estimation from dependent observations. *Stochastic Processes and their Application* 23(1), 77–89.

Fernandes, M. & P.K. Monteiro (2005) Central limit theorem for asymmetric kernel functionals. *Annals of the Institute of Statistical Mathematics* 57(3), 425–442.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

Gasser, T., H.-G. Müller, & V. Mammitzsch (1985) Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society. Series B (Methodological)* 47, 238–252.

Goldstein, L. & K. Messer (1992) Optimal plug-in estimators for nonparametric functional estimation. *Annals of Statistics* 20(3), 1306–1328.

Hall, P. & J.S. Marron (1987) Estimation of integrated squared density derivatives. *Statistics and Probability Letters* 6(2), 109–115.

Hall, P., J.S. Racine, & Q. Li (2004) Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association* 99(486), 1015–1026.

Härdle, W., J.D. Hart, J.S. Marron, & A.B. Tsybakov (1992) Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association* 87(417), 218–226.

Härdle, W. & T.M. Stoker (1989) Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association* 84, 986–995.

Härdle, W. & A.B. Tsybakov (1993) How sensitive are average derivatives? *Journal of Econometrics* 58, 31–48.

Hirano, K., G.W. Imbens, & G. Ridder (2003) Efficient estimation of average treatment effects using the estimated propensity Score. *Econometrica* 71(4), 1161–1189.

Honoré, B.E. & A. Lewbel (2002) Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica* 70(5), 2053–2063.

Hyndman, R.J., D.M. Bashtannyk, & G.K. Grunwald (1996) Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics* 5(4), 315–336.

Ichimura, H. & O.B. Linton (2005) Asymptotic expansions for some semiparametric program evaluation estimators. In D.W.K. Andrews and J.H. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,* 1st ed., pp. 149–170. Cambridge University Press.

Jacho-Chávez, D.T. (2006) Identification, Estimation and Efficiency of Nonparametric and Semiparametric Models in Microeconometrics. Ph.D. thesis, London School of Economics and Political Science.

Jones, M.C. & S.J. Sheather (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives. *Statistica and Probability Letters* 11(6), 511–514.

Khan, S. & A. Lewbel (2007) Weighted and two-stage least squares estimation of semiparametric truncated regression models. *Econometric Theory* 23(2), 309–347.

Lewbel, A. (1998) Semiparametric latent variable model estimation with endogenous or mismeasured regressors. *Econometrica* 66(1), 105–121.

Lewbel, A. (2000a) Asymptotic Trimming for Bounded Density Plug-in Estimators. Unpublished manuscript, Boston College.

Lewbel, A. (2000b) Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics* 97(1), 145–177.

Lewbel, A. (2007) Endogenous selection or treatment model estimation. *Journal of Econometrics* 141(2), 777–806.

Linton, O.B. (1991) Edgeworth Approximation in Semiparametric Regression Models. Ph.D. thesis, University of California, Berkeley.

Linton, O.B. (1995) Second order approximation in the partially linear regression model. *Econometrica*, 63(3), 1079–1112.

Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17(6), 571–599.

Powell, J.L., J.H. Stock, & T.M. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57(6), 1403–1430.

Powell, J.L. & T.M. Stoker (1996) Optimal bandwidth choice for density-weighted averages. *Journal of Econometrics* 75(2), 291–316.

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

Rosenblatt, M. (1969) Conditional probability density and regression estimators. In P.R. Krishnaiah (ed.). *Multivariate Analysis II*, pp. 25–31. Academic Press.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

Sherman, R.P. (1994) U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric Theory* 10(2), 372–395.

Silverman, B.W. (1978) Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Annals of Statistics* 6(1), 177–184.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis,* 1st ed. Chapman and Hall.

Srinivasan, T.N. (1970) Approximations to finite sample moments of estimators whose exact sampling distributions are unknown. *Econometrica* 38(3), 533–541.

Wand, M.P. & C. Jones (1995) *Kernel Smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability* 1st ed. Chapman and Hall.

APPENDIX A: Main Proofs

Let $\|\cdot\|$ denote euclidean norm, and let $\langle\cdot,\cdot\rangle$ represent the inner product when applied to vectors. We also use the following results from Masry (1996) (see Silverman, 1978, and Collomb and Härdle, 1986, for earlier results):

$$\max_{i=1,\ldots,N}\left|\widehat{f}_{V\mathbf{U}}(v_i,\mathbf{u}_i;h)-f_{V\mathbf{U}}(v_i,\mathbf{u}_i)\right|=O_p\left(\sqrt{\frac{\log N}{Nh^d}}+h^P\right), \tag{A.1}$$

$$\max_{i=1,\ldots,N}\left|\widehat{f}_{\mathbf{U}}(\mathbf{u}_i;h)-f_{\mathbf{U}}(\mathbf{u}_i)\right|=O_p\left(\sqrt{\frac{\log N}{Nh^{d-1}}}+h^P\right). \tag{A.2}$$

Proof of Lemma 1. Firstly, from Assumption A3, it follows that $\widehat{\vartheta}_{2i}(h)$ is

$$\begin{aligned}
\left|N^{-1/2}\textstyle\sum_{i=1}^N\widehat{\vartheta}_{2i}(h)\right| &\leq \left(N^{-1/2}\textstyle\sum_{i=1}^N\|\omega_i\|\,|f_{\mathbf{U}i}|\right)\left[\min_{i=1,\ldots,N}\left|\widehat{f}_{V\mathbf{U}i}\right|\right]^{-1}\\
&\quad\times\left[\max_{i=1,\ldots,N}\left|\widehat{f}_{V\mathbf{U}i}-f_{V\mathbf{U}i}\right|\right]^3\left[\min_{i=1,\ldots,N}\left|f^3_{V\mathbf{U}i}\right|\right]^{-1}\\
&\leq \left(N^{-1/2}\textstyle\sum_{i=1}^N\|\omega_i\|\,|f_{\mathbf{U}i}|\right)\left[\min_{i=1,\ldots,N}|f_{V\mathbf{U}i}|\right]^{-1} \qquad \text{(A.3)}\\
&\quad\times\left[\max_{i=1,\ldots,N}\left|\widehat{f}_{V\mathbf{U}i}-f_{V\mathbf{U}i}\right|\right]^3\left[\min_{i=1,\ldots,N}\left|f^3_{V\mathbf{U}i}\right|\right]^{-1}\\
&= O_p\left(\sqrt{N}\right)O_p\left(\left(\sqrt{\frac{\log N}{Nh^d}}+h^P\right)^3\right)\\
&= o_p\left(N^{-(P-d)/(P+d)}\right),
\end{aligned}$$

where (A.3) follows after observing that

$$\begin{aligned}
\inf_{\Omega_{V\mathbf{U}}}\widehat{f}_{V\mathbf{U}}(v,\mathbf{u};h) &\geq \inf_{\Omega_{V\mathbf{U}}}f_{V\mathbf{U}}(v,\mathbf{u})-\sup_{\Omega_{V\mathbf{U}}}\left|\widehat{f}_{V\mathbf{U}}(v,\mathbf{u};h)-f_{V\mathbf{U}}(v,\mathbf{u})\right|\\
&\geq \inf_{\Omega_{V\mathbf{U}}}f_{V\mathbf{U}}(v,\mathbf{u})+o_p(1),
\end{aligned}$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

and the last inequality follows from (A.1). Finally, by the same argument, it also follows that

$$\begin{aligned}\left|N^{-1/2}\Sigma_{i=1}^{N}\widehat{\vartheta}_{1i}(h)\right| &\leq \left(N^{-1/2}\Sigma_{i=1}^{N}\|\omega_i\|\right)\left[\min_{i=1,\ldots,N}|f_{V\mathbf{U}i}|\right]^{-1}\\ &\quad\times\left[\max_{i=1,\ldots,N}\left|\widehat{f}_{V\mathbf{U}i}-f_{V\mathbf{U}i}\right|\right]^{2}\left[\max_{i=1,\ldots,N}\left|\widehat{f}_{\mathbf{U}i}-f_{\mathbf{U}i}\right|\right]\\ &\quad\times\left[\min_{i=1,\ldots,N}\|f_{V\mathbf{U}i}\|^{2}\right]\\ &= o_p\left(N^{-(P-d)/(P+d)}\right),\end{aligned}$$

as required. ■

Sketch of Proof of Theorem 1. The proof of this theorem will require us to look at the contribution to the MSE from each of the elements on the right–hand side of (8). (The complete proof can be obtained from the author upon request.) First, let us denote $\delta_1 = \mathrm{E}[\varpi_1]$, $\delta_2 = \mathrm{E}[\pi_2(\mathbf{U})f_{\mathbf{U}}(\mathbf{U})]$, $\delta_3 = \mathrm{E}[\pi_3(V,\mathbf{U})f_{V\mathbf{U}}(V,\mathbf{U})]$, $\delta_4 = \mathrm{E}[\pi_4(V,\mathbf{U})f_{\mathbf{U}}(\mathbf{U})f_{V\mathbf{U}}(V,\mathbf{U})]$, and $\delta_5 = \mathrm{E}\left[\pi_5(V,\mathbf{U})f_{V\mathbf{U}}^2(V,\mathbf{U})\right]$. Then, by using the definitions in Section 2 and the properties of conditional expectations, it follows that $\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \eta$, and therefore

$$\eta = \mathrm{E}\left[\widehat{\delta}_1\right] + 2\times\lim_{N\to\infty}\mathrm{E}\left[\widehat{\delta}_2(h)\right] - 2\times\lim_{N\to\infty}\mathrm{E}\left[\widehat{\delta}_3(h)\right] - \lim_{N\to\infty}\mathrm{E}\left[\widehat{\delta}_4(h)\right] + \lim_{N\to\infty}\mathrm{E}\left[\widehat{\delta}_5(h)\right].$$

We are now able to write $\mathrm{E}\left[\left\|\widehat{\eta}(h)-\eta\right\|^2\right]$ as

$$\mathrm{E}\left[\left\|\widehat{\eta}(h)-\eta\right\|^2\right] = \mathrm{E}\left[\left\|\widehat{\delta}_1-\delta_1\right\|^2\right] \tag{A.4}$$

$$+4\mathrm{E}\left[\left\|\widehat{\delta}_2(h)-\delta_2\right\|^2\right] \tag{A.5}$$

$$+4\mathrm{E}\left[\left\|\widehat{\delta}_3(h)-\delta_3\right\|^2\right] \tag{A.6}$$

$$+\mathrm{E}\left[\left\|\widehat{\delta}_4(h)\right\|^2\right] \tag{A.7}$$

$$+\mathrm{E}\left[\left\|\widehat{\delta}_5(h)\right\|^2\right] \tag{A.8}$$

$$+4\mathrm{E}\left[\left\langle\widehat{\delta}_1-\delta_1,\widehat{\delta}_2(h)-\delta_2\right\rangle\right] \tag{A.9}$$

$$-4\mathrm{E}\left[\left\langle\widehat{\delta}_1-\delta_1,\widehat{\delta}_3(h)-\delta_3\right\rangle\right] \tag{A.10}$$

$$-2\mathrm{E}\left[\left\langle\widehat{\delta}_1-\delta_1,\widehat{\delta}_4(h)-\delta_4\right\rangle\right] \tag{A.11}$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

$$+2\mathrm{E}\left[\left\langle \widehat{\delta}_1-\delta_1, \widehat{\delta}_5(h)-\delta_5\right\rangle\right] \tag{A.12}$$

$$-8\mathrm{E}\left[\left\langle \widehat{\delta}_2(h)-\delta_2, \widehat{\delta}_3(h)-\delta_3\right\rangle\right] \tag{A.13}$$

$$-4\mathrm{E}\left[\left\langle \widehat{\delta}_2(h)-\delta_2, \widehat{\delta}_4(h)-\delta_4\right\rangle\right] \tag{A.14}$$

$$+4\mathrm{E}\left[\left\langle \widehat{\delta}_2(h)-\delta_2, \widehat{\delta}_5(h)-\delta_5\right\rangle\right] \tag{A.15}$$

$$+4\mathrm{E}\left[\left\langle \widehat{\delta}_3(h)-\delta_3, \widehat{\delta}_4(h)-\delta_4\right\rangle\right] \tag{A.16}$$

$$-4\mathrm{E}\left[\left\langle \widehat{\delta}_3(h)-\delta_3, \widehat{\delta}_5(h)-\delta_5\right\rangle\right] \tag{A.17}$$

$$-2\mathrm{E}\left[\left\langle \widehat{\delta}_4(h)-\delta_4, \delta_5(h)-\delta_5\right\rangle\right]. \tag{A.18}$$

Let us define the following quantities:

$$\mathfrak{B}_{1,1}=\int \pi_2(\mathbf{u})\, S_{\mathcal{K}}(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u},$$

$$\mathfrak{B}_{1,2}=\int \pi_3(v,\mathbf{u})\, S_{W\mathcal{K}}(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv d\mathbf{u}, \quad \text{and}$$

$$\mathfrak{B}_2=C_{W\mathcal{K}}\int \pi_3(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv d\mathbf{u},$$

where $S_{\mathcal{K}}(\mathbf{u})$ and $S_{W\mathcal{K}}(v,\mathbf{u})$ are defined in the main text. Then, under the assumptions of the theorem it is possible to show (see Jacho-Chávez, 2006) that the leading terms in the expansion of (A.4)– (A.18) are:

Term:	Contribution: h^{2P}	Contribution: $N^{-1}h^{P-d}$	Contribution: $N^{-2}h^{-2d}$
(A.4)	–	–	–
(A.5)	$+4\left\|\mathfrak{B}_{1,1}\right\|^2$	–	–
(A.6)	$+4\left\|\mathfrak{B}_{1,2}\right\|^2$	–	–
(A.7)	$+\left\|\mathfrak{B}_{1,1}+\mathfrak{B}_{1,2}\right\|^2$	–	–
(A.8)	$+4\left\|\mathfrak{B}_{1,2}\right\|^2$	$+4\left\langle \mathfrak{B}_2, \mathfrak{B}_{1,2}\right\rangle$	$\left\|\mathfrak{B}_2\right\|^2$
(A.9)	–	–	–
(A.10)	–	–	–
(A.11)	–	–	–
(A.12)	–	–	–
(A.13)	$-8\left\langle \mathfrak{B}_{1,1}, \mathfrak{B}_{1,2}\right\rangle$	–	–
(A.14)	$-4\left\langle \mathfrak{B}_{1,1}, \mathfrak{B}_{1,1}+\mathfrak{B}_{1,2}\right\rangle$	–	–
(A.15)	$+8\left\langle \mathfrak{B}_{1,1}, \mathfrak{B}_{1,2}\right\rangle$	$+4\left\langle \mathfrak{B}_2, \mathfrak{B}_{1,1}\right\rangle$	–
(A.16)	$+4\left\langle \mathfrak{B}_{1,2}, \mathfrak{B}_{1,1}+\mathfrak{B}_{1,2}\right\rangle$	–	–
(A.17)	$-8\left\|\mathfrak{B}_{1,2}\right\|^2$	$-4\left\langle \mathfrak{B}_2, \mathfrak{B}_{1,2}\right\rangle$	–
(A.18)	$-4\left\langle \mathfrak{B}_{1,1}+\mathfrak{B}_{1,2}, \mathfrak{B}_{1,2}\right\rangle$	$-2\left\langle \mathfrak{B}_2, \mathfrak{B}_{1,1}+\mathfrak{B}_{1,2}\right\rangle$	–
Net:	$\left\|\mathfrak{B}_{1,1}-\mathfrak{B}_{1,2}\right\|^2$	$2\left\langle \mathfrak{B}_{1,1}-\mathfrak{B}_{1,2}, \mathfrak{B}_2\right\rangle$	$\left\|\mathfrak{B}_2\right\|^2$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

There are also terms of smaller order, namely $O(N^{-1})$, $O(N^{-1}h^P + N^{-2})$, and $o(N^{-1}h^P + N^{-2}h^{-2d} + h^{2P})$. We conclude by grouping the leading terms as

$$
\begin{aligned}
&h^{2P}\left\|\int \pi_2(\mathbf{u})\, S_{\mathcal{K}}(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} - \int \pi_3(v,\mathbf{u})\, S_{W\mathcal{K}}(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv\, d\mathbf{u}\right\|^2 \\
&\quad + \frac{2C_{W\mathcal{K}}}{Nh^d} h^P \left\langle \int \pi_3(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv\, d\mathbf{u}\,, \right. \\
&\qquad \left. \int \pi_2(\mathbf{u})\, S_{\mathcal{K}}(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} - \int \pi_3(v,\mathbf{u})\, S_{W\mathcal{K}}(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv\, d\mathbf{u} \right\rangle \\
&\quad + \frac{C_{W\mathcal{K}}^2}{N^2h^{2d}} \left\| \int \pi_3(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dv\, d\mathbf{u} \right\|^2 \\
&= h^{2P} \left\| \mathfrak{B}_{1,1} - \mathfrak{B}_{1,2} \right\|^2 + 2\frac{h^P}{Nh^d} \left\langle \mathfrak{B}_2, \mathfrak{B}_{1,1} - \mathfrak{B}_{1,2} \right\rangle + \frac{1}{N^2h^{2d}} \left\| \mathfrak{B}_2 \right\|^2 \\
&= \left\| h^P \mathfrak{B}_1 + \mathfrak{B}_2 N^{-1} h^{-d} \right\|^2,
\end{aligned}
$$

where $\mathfrak{B}_1 = \mathfrak{B}_{1,1} - \mathfrak{B}_{1,2}$.

Proof of Proposition 1. In what follows, we make use of the following identities:

$$
\begin{aligned}
\frac{\widehat{f}_{\mathbf{U}i}}{\widehat{f}_{V\mathbf{U}i}} - \frac{f_{\mathbf{U}i}}{f_{V\mathbf{U}i}} &= \frac{\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}}{f_{V\mathbf{U}i}} - \frac{f_{\mathbf{U}i}\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)}{f_{V\mathbf{U}i}^2} \\
&\quad + \frac{f_{\mathbf{U}i}\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)^2}{f_{V\mathbf{U}i}^2 \widehat{f}_{V\mathbf{U}i}} - \frac{\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)\left(\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}\right)}{f_{V\mathbf{U}i}\widehat{f}_{V\mathbf{U}i}}, \quad \text{and}
\end{aligned}
\tag{A.19}
$$

$$
\begin{aligned}
\frac{\widehat{f}_{\mathbf{U}i}}{\widehat{f}_{V\mathbf{U}i}^2} - \frac{f_{\mathbf{U}i}}{f_{V\mathbf{U}i}^2} &= \frac{\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}}{f_{V\mathbf{U}i}^2} - \frac{f_{\mathbf{U}i}\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)\left(\widehat{f}_{V\mathbf{U}i} + f_{V\mathbf{U}i}\right)}{f_{V\mathbf{U}i}^4} \\
&\quad + \frac{f_{\mathbf{U}i}\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)^2\left(\widehat{f}_{V\mathbf{U}i} + f_{V\mathbf{U}i}\right)^2}{f_{V\mathbf{U}i}^4 \widehat{f}_{V\mathbf{U}i}^2} \\
&\quad - \frac{\left(\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i}\right)\left(\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}\right)\left(\widehat{f}_{V\mathbf{U}i} + f_{V\mathbf{U}i}\right)}{f_{V\mathbf{U}i}^2 \widehat{f}_{V\mathbf{U}i}^2}.
\end{aligned}
\tag{A.20}
$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

Term: $\widehat{\mathfrak{B}}_1(h_0)$
First, it follows from (A.19) that

$$\widetilde{\eta}(\Delta h_0)-\widetilde{\eta}(h_0)=\widehat{\delta}_2(\Delta h_0)-\widehat{\delta}_2(h_0)-\left[\widehat{\delta}_3(\Delta h_0)-\widehat{\delta}_3(h_0)\right]$$
$$+N^{-1}\sum_{i=1}^{N}(\widehat{\vartheta}_{1i}(\Delta h_0)-\widehat{\vartheta}_{2i}(\Delta h_0))\omega_i a_\tau(v_i,\mathbf{u}_i) \quad \textbf{(A.21)}$$
$$-N^{-1}\sum_{i=1}^{N}(\widehat{\vartheta}_{1i}(h_0)-\widehat{\vartheta}_{2i}(h_0))\omega_i a_\tau(v_i,\mathbf{u}_i), \quad \textbf{(A.22)}$$

where (A.21) and (A.22) are $O_p(N^{-1}h_0^{-d}\log N+h_0^{2P})$ because of Assumptions A1–A4 and results (A.20), (A.1), and (A.2). That is,

$$\frac{\widetilde{\eta}(\Delta h_0)-\widetilde{\eta}(h_0)}{h_0^P(\Delta^P-1)}=\binom{N}{2}^{-1}\sum_{i<j}\frac{p_2(\mathbf{t}_{2\tau i},\mathbf{t}_{2\tau j};\Delta h_0)-p_2(\mathbf{t}_{2\tau i},\mathbf{t}_{2\tau j};h_0)}{h_0^P(\Delta^P-1)}$$
$$-\binom{N}{2}^{-1}\sum_{i<j}\frac{p_3(\mathbf{t}_{3\tau i},\mathbf{t}_{3\tau j};\Delta h_0)-p_3(\mathbf{t}_{3\tau i},\mathbf{t}_{3\tau j};h_0)}{h_0^P(\Delta^P-1)}$$
$$+O_p\left(\left(Nh_0^{P+d}\right)^{-1}\log N+h_0^P\right),$$

which means that $\widehat{\mathfrak{B}}_1(h_0)$ is the sum of two U-statistics plus a remainder that is $o_p(1)$, because under the conditions of the proposition, $h_0\to 0$ and $Nh_0^{P+d}\to\infty$ as $N\to\infty$. Given Lemma B.1 in Appendix B, it then follows from Lemma 3.1 in Powell, Stock, and Stoker (1989, p. 1410) and Theorem A in Lewbel (2000a, p. 4), that $\left(\widetilde{\eta}(\Delta h_0)-\widetilde{\eta}(h_0)\right)/(h_0^P(\Delta^P-1))$ is consistent for

$$\mathrm{E}\left[\omega\left(\frac{S_{\mathcal{K}}(\mathbf{U})}{f_{V\mathbf{U}}(V,\mathbf{U})}-\frac{f_{\mathbf{U}}(\mathbf{U})S_{W\mathcal{K}}(V,\mathbf{U})}{f_{V\mathbf{U}}^2(V,\mathbf{U})}\right)a_\tau(V,\mathbf{U})\right].$$

This is true because

$$\left\|\mathrm{E}\left[\omega\left(\frac{S_{\mathcal{K}}(\mathbf{U})}{f_{V\mathbf{U}}(V,\mathbf{U})}-\frac{f_{\mathbf{U}}(\mathbf{U})S_{W\mathcal{K}}(V,\mathbf{U})}{f_{V\mathbf{U}}^2(V,\mathbf{U})}\right)(1-a_\tau(V,\mathbf{U}))\right]\right\|^2$$
$$\leq\left[\sup_{\Omega_{V\mathbf{U}}}\|\omega\|\sup_{\Omega_{V\mathbf{U}}}\left|\frac{S_{\mathcal{K}}(\mathbf{u})}{f_{V\mathbf{U}}(v,\mathbf{u})}-\frac{f_{\mathbf{U}}(\mathbf{u})S_{W\mathcal{K}}(v,\mathbf{u})}{f_{V\mathbf{U}}^2(v,\mathbf{u})}\right|\mathrm{E}[1-a_\tau(V,\mathbf{U})]\right]^2. \quad \textbf{(A.23)}$$

Now $\mathrm{E}[1-a_\tau(V,\mathbf{U})]$ equals the probability that $(v,\mathbf{u})$ is within a distance τ of the boundary of $\mathbf{\Omega}_{V\mathbf{U}}$, which is less than or equal to $\sup_{\Omega_{V\mathbf{U}}}f_{V\mathbf{U}}(v,\mathbf{u})$ times the volume of the space within a distance τ of the boundary or $\mathbf{\Omega}_{V\mathbf{U}}$. This volume is $O(\tau)$, so from Assumptions A3 and A6, we have that (A.23) is $O(\tau)=O(N^{-1/2}(N^{1/2}\tau))=o(N^{-1/2})$, where the last equality follows from Assumption A7. Therefore, under the conditions of the proposition, we conclude that $\widehat{\mathfrak{B}}_1(h_0)\xrightarrow{p}\mathfrak{B}_1$ as $N\to\infty$.

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

Term: $\widehat{\mathfrak{B}}_2(h_*)$
Notice that

$$\widehat{\mathfrak{B}}_2(h_*) = \frac{C_{W\mathcal{K}}}{N}\sum_{i=1}^{N}\varpi_{3\tau i} + \frac{C_{W\mathcal{K}}}{N}\sum_{i=1}^{N}\widehat{\varpi}_{*3\tau i} - \varpi_{3\tau i}, \tag{A.24}$$

where the second term on the right–hand side of (A.24) is bounded above by

$$C_{W\mathcal{K}}\left(\frac{1}{N}\sum_{i=1}^{N}\|\omega_i\|^2\right)\max_{i=1,\ldots,n}\left|\frac{\widehat{f}_{\mathbf{U}i}}{\widehat{f}^2_{V\mathbf{U}i}} - \frac{f_{\mathbf{U}i}}{f^2_{V\mathbf{U}i}}\right| = O_p\left(\sqrt{\frac{\log N}{Nh_*^d}} + h_*^P\right),$$

which is $o_p(1)$ by Assumption A4, representation (A.20), and the assumptions of the proposition ($h_* \to 0$ and $Nh_*^d \to \infty$ as $N \to \infty$). The result follows from Kolmogorov's Law of Large Numbers when applied to the first term in the right–hand side of (A.24), and we conclude that

$$\widehat{\mathfrak{B}}_2(h_*) \xrightarrow{p} \mathfrak{B}_2, \quad \text{as } N \to \infty.$$

APPENDIX B: Technical Lemmas

LEMMA B.1. *Let Assumptions* A1–A3*,* A4*,* A6*, and* A7 *hold; then*

$$\mathrm{E}\left[\|p_3(\mathbf{t}_{3\tau 1}, \mathbf{t}_{3\tau 2}; \Delta h_0) - p_3(\mathbf{t}_{3\tau 1}, \mathbf{t}_{3\tau 2}; h_0)\|^2 / \left[h_0^P\left(1 - \Delta^P\right)\right]^2\right] = o(N), \tag{B.1}$$

$$\mathrm{E}\left[\|p_2(\mathbf{t}_{2\tau 1}, \mathbf{t}_{2\tau 2}; \Delta h_0) - p_2(\mathbf{t}_{2\tau 1}, \mathbf{t}_{2\tau 2}; h_0)\|^2 / \left[h_0^P\left(1 - \Delta^P\right)\right]^2\right] = o(N). \tag{B.2}$$

Proof. Recall $\mathbf{t}_{3\tau 1}^\top = \left(\varpi_{3\tau 1}^\top, V_1, \mathbf{U}_1^\top\right)$, where $\varpi_{3\tau 1} = \varpi_{31} a_\tau(V_1, \mathbf{U}_1)$, and define $\varrho_{\epsilon\tau}(v, \mathbf{u}) \equiv \mathrm{E}[\|\varpi_{3\tau 1}\|^\epsilon \,|\, V_1 = v, \mathbf{U}_1 = \mathbf{u}]$ for $\epsilon = 1, 2, 3, 4$. Then

$$\begin{aligned}
&\mathrm{E}\left[\|p_3(\mathbf{t}_{3\tau 1}, \mathbf{t}_{3\tau 2}; \Delta h_0) - p_3(\mathbf{t}_{3\tau 1}, \mathbf{t}_{3\tau 2}; h_0)\|^2 / \left[h_0^P\left(1 - \Delta^P\right)\right]^2\right] \\
&\quad = \mathrm{E}\left[\left(\frac{c_\Delta}{h_0^{2P}}\right)\left\|\frac{\varpi_{3\tau 1} + \varpi_{3\tau 2}}{2h_0^d}\right\|^2 W^2\left(\frac{V_1 - V_2}{h_0}\right)\mathcal{K}^2\left(\frac{\mathbf{U}_1 - \mathbf{U}_2}{h_0}\right)\right] \\
&\quad = \int\left(\frac{c_\Delta}{4h_0^{2P+d}}\right) f_{V\mathbf{U}}(z + ch_0, \mathbf{z} + \mathbf{c}h_0)\, f_{V\mathbf{U}}(z, \mathbf{z}) \\
&\qquad \times \left[\varrho_{2\tau}(z + ch_0, \mathbf{z} + \mathbf{c}h_0) + \varrho_{2\tau}(z, \mathbf{z}) + 2\langle\varrho_{0\tau}(z + ch_0, \mathbf{z} + \mathbf{c}h_0), \varrho_{0\tau}(z, \mathbf{z})\rangle\right] \\
&\qquad \times W^2(c)\mathcal{K}^2(\mathbf{c})\, dz\, dc\, d\mathbf{z}\, d\mathbf{c} \\
&\quad = O(h_0^{-(2P+d)}) = O(N(Nh_0^{2P+d})^{-1}) = o(N),
\end{aligned}$$

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press

where $c_\Delta = \left(1-\Delta^2\right)/\left(\Delta\left(1-\Delta^P\right)\right)^2$, and $\varrho_{0\tau}(\upsilon,\mathbf{u}) \equiv \mathrm{E}\left[\varpi_{3\tau 1}|V_1=\upsilon,\mathbf{U}_1=\mathbf{u}\right]$. The second equality uses the change of variables from $(y,\mathbf{y}^\top,z,\mathbf{z}^\top)$ to $(c=h_0^{-1}(y-z),$ $\mathbf{c}^\top=h_0^{-1}(\mathbf{y}-\mathbf{z})^\top,z,\mathbf{z}^\top)$ with jacobian h_0^d. This change of variables is not affected by boundary effects because of Assumptions A1 and A2 and the fact that $a_\tau(z,\mathbf{z})=0$ for all $(z,\mathbf{z})$ within a distance τ of the boundary of $\mathbf{\Omega}_{V\mathbf{U}}$, with $h_0/\tau\to 0$. The last equality uses the continuity of the $\varrho_{\epsilon\tau}$'s and Assumption A7. (B.2) follows the same arguments, and therefore it is omitted. ■

https://doi.org/10.1017/S0266466609090628 Published online by Cambridge University Press