Optimal Bandwidth Choice for Estimation of Inverse Conditional–Density–Weighted Expectations

David Tomás Jacho-Chávez*
Indiana University

Abstract

This paper characterizes the bandwidth value (h) that is optimal for estimating parameters of the form $\eta = E\left[\omega / f_{V|\mathbf{U}}\left(V|\,\mathbf{U}\right)\right]$, where $f_{V|\mathbf{U}}$ (the conditional density of a scalar continuous random variable V, given a random vector $\mathbf{U}$) is replaced by its kernel estimator. The results in this paper are directly applicable to semiparametric estimators proposed in Lewbel (1998), Lewbel (2000b), Honoré and Lewbel (2002), Khan and Lewbel (2007), and Lewbel (2006). The optimal bandwidth is derived by minimizing the leading terms of a second–order mean squared error expansion 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. We then use our methodology in an empirical application of the female labour force participation in Ecuador.

Keywords: Optimal bandwidth; Semiparametric estimation; Density 'Plug-in' Estimators; Trimming

JEL classification: C13; C14; C25

*Department of Economics, Indiana University, Wylie Hall 251, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. Phone: +1 (812) 855 7928. Fax: +1 (812) 855 3736. E-mail: djachoch@indiana.edu. Web Page: http://mypage.iu.edu/∼djachoch/

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

$$\eta = E\left[\frac{\omega}{f_{V|\mathbf{U}}\left(V|\,\mathbf{U}\right)}\right], \tag{1.1}$$

where $\left\{\omega^{\top}, V, \mathbf{U}^{\top}\right\}$ is a random vector, and $f_{V|\mathbf{U}}\left(\cdot\right)$ 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 ratio of kernel estimators for $f_{V\mathbf{U}}\left(\cdot\right)$ and $f_{\mathbf{U}}\left(\cdot\right)$, the joint and marginal densities of $\left(V, \mathbf{U}^{\top}\right)$ and $\left(\mathbf{U}\right)$ 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 (2006). 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 $f_{V|\mathbf{U}}\left(\cdot\right)$ in (1.1). This paper discusses formally how to perform this selection. Given that the asymptotic (first–order) distribution of this semiparametric estimator, $\widetilde{\eta}\left(h\right)$, of (1.1) does not depend on the bandwidth[1] h, any optimal bandwidth formula must be based on a higher–order approximation to such distribution. Technically, such approximations become more complex in the presence of stochastic denominators in a simple 'plug–in' semiparametric estimator of (1.1) as explained above. Therefore, we take an alternative approach. We first show that $\widetilde{\eta}\left(h\right)$ is asymptotically equivalent to a linear combination of functions of U–statistics, which we call its 'asymptotic representation', $\widehat{\eta}\left(h\right)$, and does not have a stochastic denominator. This asymptotic representation includes functions of a U–statistic of order one (a simple sample average), and two data dependent (via the bandwidth parameter h) second–order U–statistics. Finally, we find a formula for the optimal bandwidth that minimizes (with respect to h) the leading terms of an asymptotic approximation to

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

where $\left\|\cdot\right\|$ is the standard Euclidean[2] norm.

Related 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

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

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

theirs in that the optimal bandwidth for semiparametric kernel estimators of (1.1) can be chosen on the basis of bias alone. In particular, we show that the leading terms in the Mean Squared Error (MSE) 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 (2003)'s estimators of treatment effects. Linton (1991) discussed a similar result for the variance estimator in the presence of unknown mean. Furthermore, unlike the standard case in average derivative estimation[3], semiparametric estimation of (1.1) could include discrete elements (specifically in **U)** through its nonparametric component without the need of additional conditions. We explain this extension in greater detail in our discussion below.

As it 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 (1.1). 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 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. In this section, we analyze the sensitivity of kernel–based semiparametric estimator of (1.1) to the choice of bandwidth, and its kernel's order via a second–order asymptotic expansion of its MSE. We also make explicit the difference between 'nonparametric' and 'semiparametric' optimal bandwidths. Section 3 discusses how to exploit the asymptotic representation of $\widetilde{\eta}(h)$ in order to construct a simple estimator of the optimal bandwidth. We also prove its consistency. In Section 4.1, a Monte Carlo experiment is performed to assess the small sample behavior of the proposed 'plug–in' estimator of the optimal bandwidth. We also compare its performance against other reference rules proposed in the literature for estimation of the nonparametric component $f_{V|\mathbf{U}}(\cdot)$. This section also contains an empirical implementation of the proposed methodology to the semiparametric estimation of the binary response model of labour market participation of women in Ecuador. Section 5 examines how results in Section 2 can be extended to cases when some components of $\mathbf{U}$ are discrete. Section 6 summarizes and gives concluding remarks. All proofs are presented in the Appendix.

[3]Horowitz and Härdle (1996) adapted the average derivative estimator to allow for some discrete components. This requires additional conditions than in the standard case.

2 Asymptotic Mean Square Error

Firstly, 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 Nadaraya–Watson kernel smoothers:

$$\widehat{f}_{\mathbf{U}}(\mathbf{u}_i; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d-1}} \mathcal{K}\left(\frac{\mathbf{u}_j - \mathbf{u}_i}{h}\right), \text{ and} \tag{2.1}$$

$$\widehat{f}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d}} W\left(\frac{v_j - v_i}{h}\right) \mathcal{K}\left(\frac{\mathbf{u}_j - \mathbf{u}_i}{h}\right) \tag{2.2}$$

respectively. Here

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

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

$$\widetilde{\eta}(h) = N^{-1} \sum_{i=1}^{N} \omega_i \widehat{l}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h). \tag{2.3}$$

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

As previously noted, this estimator is technically inconvenient to handle given the presence of the stochastic denominator in $\widehat{l}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h)$. Therefore, we also define an asymptotic representation which will be the basis of our analysis below,

$$\widehat{\eta}(h) = N^{-1} \sum_{i=1}^{N} \omega_i \widehat{L}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h), \tag{2.4}$$

where

$$\begin{aligned} \widehat{L}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h) = & \frac{f_{\mathbf{U}i}}{f_{V\mathbf{U}i}} + 2\frac{\widehat{f}_{\mathbf{U}i}}{f_{V\mathbf{U}i}} - 2\frac{f_{\mathbf{U}i}\widehat{f}_{V\mathbf{U}i}}{f^2_{V\mathbf{U}i}} \\ & - \frac{\widehat{f}_{V\mathbf{U}i}\widehat{f}_{\mathbf{U}i}}{f^2_{V\mathbf{U}i}} + \frac{f_{\mathbf{U}i}\widehat{f}^2_{V\mathbf{U}i}}{f^3_{V\mathbf{U}i}}, \end{aligned} \tag{2.5}$$

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: $\widehat{\delta}_1 \equiv N^{-1}\sum_{i=1}^{N} f_{\mathbf{U}i}/f_{V\mathbf{U}i}$, $\widehat{\delta}_2(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i (\widehat{f}_{\mathbf{U}i}/f_{V\mathbf{U}i})$, $\widehat{\delta}_3(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i (f_{\mathbf{U}i}\widehat{f}_{V\mathbf{U}i}/f^2_{V\mathbf{U}i})$, $\widehat{\delta}_4(h) \equiv N^{-1}\sum_{i=1}^{N}\omega_i \widehat{f}_{V\mathbf{U}i}\widehat{f}_{\mathbf{U}i}/f^2_{V\mathbf{U}i}$, 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 that

$$\widehat{\eta}(h) = \widehat{\delta}_1 + 2\widehat{\delta}_2(h) - 2\widehat{\delta}_3(h) - \widehat{\delta}_4(h) + \widehat{\delta}_5(h), \tag{2.6}$$

That is, $\widehat{\eta}(h)$ can be written as a linear combination of functions of certain U–statistics. In particular, $\widehat{\delta}_2(h)$ and $\widehat{\delta}_3(h)$ are generic second–order U–statistics:

$$\begin{aligned} \widehat{\delta}_2(h) &= \binom{N}{2}^{-1} \sum_{i<j} \frac{\varpi_{2i} + \varpi_{2j}}{2h^{d-1}} \mathcal{K}\left(\frac{\mathbf{u}_i - \mathbf{u}_j}{h}\right) \\ &\equiv \binom{N}{2}^{-1} \sum_{i<j} p_2(\mathbf{t}_{2i}, \mathbf{t}_{2j}; h), \text{ and} \\ \widehat{\delta}_3(h) &= \binom{N}{2}^{-1} \sum_{i<j} \frac{\varpi_{3i} + \varpi_{3j}}{2h^{d}} W\left(\frac{v_i - v_j}{h}\right) \mathcal{K}\left(\frac{\mathbf{u}_i - \mathbf{u}_j}{h}\right) \\ &\equiv \binom{N}{2}^{-1} \sum_{i<j} p_3(\mathbf{t}_{3i}, \mathbf{t}_{3j}; h), \end{aligned}$$

where $\mathbf{t}_{2i}^{\top} = (\varpi_{2i}^{\top}, \mathbf{u}_i^{\top})$, and $\mathbf{t}_{3i}^{\top} = (\varpi_{3i}^{\top}, v_i, \mathbf{u}_i^{\top})$, with $\varpi_{2i} \equiv \omega_i/f_{V\mathbf{U}i}$, and $\varpi_{3i} \equiv \omega_i f_{\mathbf{U}i}/f^2_{V\mathbf{U}i}$ respectively. By simple inspection, we notice that these U–statistics 'kernel' functions $p_2(\cdot)$ and $p_3(\cdot)$ are symmetric – that is, $p_2(\mathbf{t}_{2i}, \mathbf{t}_{2j}; h) = p_2(\mathbf{t}_{2j}, \mathbf{t}_{2i}; h)$ and $p_3(\mathbf{t}_{3i}, \mathbf{t}_{3j}; h) = p_3(\mathbf{t}_{3j}, \mathbf{t}_{3i}; h)$. 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 $\varpi_{4i} \equiv \omega_i / f^2_{V\mathbf{U}i}$, and $\varpi_{5i} \equiv \omega_i f_{\mathbf{U}i} / f^3_{V\mathbf{U}i}$. It then follows, under conditions explained below, that

$$\begin{aligned}\eta &= \lim_{h\to 0} E\left[\widehat{\eta}\left(h\right)\right] \\ &= E\left[\widehat{\delta}_1\right] + 2 \times \lim_{h\to 0} E\left[\widehat{\delta}_2\left(h\right)\right] - 2 \times \lim_{h\to 0} E\left[\widehat{\delta}_3\left(h\right)\right] - \lim_{h\to 0} E\left[\widehat{\delta}_4\left(h\right)\right] + \lim_{h\to 0} E\left[\widehat{\delta}_5\left(h\right)\right] \\ &= \eta + 2\eta - 2\eta - \eta + \eta.\end{aligned}$$

Also, notice that by construction

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

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

$$\begin{aligned}\widehat{\vartheta}_{1i}\left(h\right) &\equiv (\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i})^2 (\widehat{f}_{\mathbf{U}i} - f_{\mathbf{U}i}) / (f^2_{V\mathbf{U}i} \widehat{f}_{V\mathbf{U}i}), \text{ and} \\ \widehat{\vartheta}_{2i}\left(h\right) &\equiv (\widehat{f}_{V\mathbf{U}i} - f_{V\mathbf{U}i})^3 f_{\mathbf{U}i} / (f^3_{V\mathbf{U}i} \widehat{f}_{V\mathbf{U}i}).\end{aligned}$$

2.2 Sensitivity Analysis

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

ASSUMPTION A:

(A1) The kernels $W : [-1,1] \to \Re$, and $K : [-1,1] \to \Re$ are bounded, continuously differentiable, symmetric such that $\int W\left(c\right) dc = \int K\left(c\right) dc = 1$. There exists constants k_v, k_v^*, k_u and $k_u^* > 0$, such that $W^2\left(v\right) \geq k_v^* 1\left(\|v\| \leq k_v^{-1}\right)$, $K^2\left(u\right) \geq k_u^* 1\left(\|u\| \leq k_u^{-1}\right)$.

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

(A3) The continuous density functions $f_{\mathbf{u}}\left(\mathbf{u}\right)$, $f_{V\mathbf{U}}(v, \mathbf{u})$ exist and are bounded away from zero. The functions $\pi_1\left(\mathbf{u}\right) = E\left[\varpi_1 | \, \mathbf{U} = \mathbf{u}\right]$, $\pi_2\left(\mathbf{u}\right) = E\left[\varpi_2 | \, \mathbf{U} = \mathbf{u}\right]$, $\widetilde{\pi}_1\left(v, \mathbf{u}\right) = E\, \varpi_1 | \, V = v$, $\mathbf{U} = \mathbf{u}]$, $\widetilde{\pi}_2\left(v, \mathbf{u}\right) = E\left[\varpi_2 | \, V = v, \mathbf{U} = \mathbf{u}\right]$, $\pi_3\left(v, \mathbf{u}\right) = E[\varpi_3 | \, V = v, \mathbf{U} = \mathbf{u}]$, $\pi_4\left(v, \mathbf{u}\right) = E[\varpi_4 | \, V = v, \mathbf{U} = \mathbf{u}]$, $\pi_5\left(v, \mathbf{u}\right) = E[\varpi_5 | \, V = v, \mathbf{U} = \mathbf{u}]$ exist and have bounded continuous partial derivatives up to the order P on their compact supports $\Omega_{\mathbf{U}} \equiv \prod_{j=1}^{d-1} \left[\underline{U}_j, \overline{U}_j\right]$ and $\Omega_{v\mathbf{u}} \equiv \left[\underline{V}, \overline{V}\right] \times \Omega_{\mathbf{U}}$ respectively, for $-\infty < \underline{V} < \overline{V} < \infty$, and $-\infty < \underline{U}_j < \overline{U}_j < \infty$, for $j = 1, \ldots, d-1$.

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

(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. The last Assumption, (A5), predefines the optimal rate of $h_{\mathbf{opt}}$, which is derived below. The following Lemma guarantees that the MSE–expansion of $\widehat{\eta}(h)$ is equivalent to that of $\widetilde{\eta}(h)$ up to the third power.

Lemma 2.1 *(Asymptotic Representation) Under Assumptions* (A1)–(A5)*,*

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

Proof. See Appendix. ■

This Lemma guarantees the asymptotic equivalence between $\widehat{\eta}(\cdot)$ and $\widetilde{\eta}(\cdot)$, which means that (2.3) may be replaced by (2.4) for purpose of this analysis. We make an additional technical assumption before we state the main result of this paper:

(A6) The vectors of errors $\varepsilon_1 = \varpi_1 - \pi_1(\mathbf{u})$, $\widetilde{\varepsilon}_1 = \varpi_1 - \widetilde{\pi}_1(v,\mathbf{u})$, $\varepsilon_2 = \varpi_2 - \pi_2(\mathbf{u})$, $\widetilde{\varepsilon}_2 = \varpi_2 - \widetilde{\pi}_2(v,\mathbf{u})$, $\varepsilon_3 = \varpi_3 - \pi_3(v,\mathbf{u})$, $\varepsilon_4 = \varpi_4 - \pi_4(v,\mathbf{u})$, and $\varepsilon_5 = \varpi_5 - \pi_5(v,\mathbf{u})$ are such that $\sigma_1^2(\mathbf{u}) = E\left[\varepsilon_1^\top\varepsilon_1 \middle| \mathbf{U}=\mathbf{u}\right]$, $\widetilde{\sigma}_1^2(v,\mathbf{u}) = E\left[\widetilde{\varepsilon}_1^\top\widetilde{\varepsilon}_1 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\sigma_2^2(\mathbf{u}) = E\left[\varepsilon_2^\top\varepsilon_2 \middle| \mathbf{U}=\mathbf{u}\right]$, $\sigma_3^2(v,\mathbf{u}) = E\left[\varepsilon_3^\top\varepsilon_3 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\sigma_4^2(v,\mathbf{u}) = E[\varepsilon_4^\top\varepsilon_4 \middle| V=v,\mathbf{U}=\mathbf{u}]$, $\sigma_5^2(v,\mathbf{u}) = E\left[\varepsilon_5^\top\varepsilon_5 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\sigma_{12}(\mathbf{u}) = E\left[\varepsilon_1^\top\varepsilon_2 \middle| \mathbf{U}=\mathbf{u}\right]$, $\widetilde{\sigma}_{13}(v,\mathbf{u}) = E[\widetilde{\varepsilon}_1^\top\varepsilon_3 | V=v,\mathbf{U}=\mathbf{u}]$, $\widetilde{\sigma}_{14}(v,\mathbf{u}) = E\left[\widetilde{\varepsilon}_1^\top\varepsilon_4 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\widetilde{\sigma}_{15}(v,\mathbf{u}) = E[\widetilde{\varepsilon}_1^\top\varepsilon_5 \middle| V=v,\mathbf{U}=\mathbf{u}]$, $\widetilde{\sigma}_{23}(v,\mathbf{u}) = E\left[\widetilde{\varepsilon}_2^\top\varepsilon_3 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\widetilde{\sigma}_{24}(v,\mathbf{u}) = E[\widetilde{\varepsilon}_2^\top\varepsilon_4 \middle| V=v,\mathbf{U}=\mathbf{u}]$, $\widetilde{\sigma}_{25}(v,\mathbf{u}) = E\left[\widetilde{\varepsilon}_2^\top\varepsilon_5 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\sigma_{34}(v,\mathbf{u}) = E\left[\varepsilon_3^\top\varepsilon_4 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, $\sigma_{35}(v,\mathbf{u}) = E\left[\varepsilon_3^\top\varepsilon_5 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$, and $\sigma_{45}(v,\mathbf{u}) = E\left[\varepsilon_4^\top\varepsilon_5 \middle| V=v,\mathbf{U}=\mathbf{u}\right]$ are bounded on their respective compact supports $\Omega_{\mathbf{U}}$ and $\Omega_{V\mathbf{U}}$.

We now formulate the Mean Square Error of $\widehat{\eta}(h)$ for η, in terms of the dominant components in an asymptotic expansion.

Theorem 2.2 *If Assumptions* (A1)–(A3)*, and* (A6)*, hold, then*

$$E\left[\left\|\widehat{\eta}(h)-\eta\right\|^2\right] = O\left(N^{-1}\right) + \left\|\mathfrak{B}_1 h^P + \mathfrak{B}_2 N^{-1}h^{-d}\right\|^2 + O\left(\frac{h^P}{N}+\frac{1}{N^2}\right) + o\left(\frac{h^P}{N}+\frac{1}{N^2h^{2d}}+h^{2P}\right), \tag{2.7}$$

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

$$\mathfrak{B}_1 = \int \pi_2(\mathbf{u}) S_K(\mathbf{u}) f_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} - \int \pi_3(v, \mathbf{u}) S_{WK}(v, \mathbf{u}) f_{V\mathbf{U}}(v, \mathbf{u}) dv d\mathbf{u}, \tag{2.8}$$

$$\mathfrak{B}_2 = C_{WK} \int \pi_3(v, \mathbf{u}) f_{V\mathbf{U}}(v, \mathbf{u}) dv d\mathbf{u}, \tag{2.9}$$

and

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

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

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

Proof. See Appendix. ■

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).

2.2.1 Optimization

The result of Theorem 2.2 can be used to perform a sensitivity analysis with respect to bandwdith choice and order of kernel.

Choice of h

The asymptotically optimal bandwidth is obtained by minimizing (2.7) on the basis of h. This is achieved when

$$h_{\mathbf{opt}} = C_0 \times N^{-1/(P+d)}, \tag{2.10}$$

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, we have

$$\begin{aligned} E\left[\left\|\widehat{\eta} - \eta\right\|^2\right] &= O\left(N^{-1}\right) \\ &\quad + \left\|\mathfrak{B}_1 C_0^P + \mathfrak{B}_2 C_0^{-d}\right\|^2 N^{-2P/(P+d)} \\ &\quad + O\left(N^{-2}\right) + o(N^{-2P/(P+d)}), \text{ as } N \to \infty. \end{aligned} \tag{2.11}$$

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 MSE–rate of convergence of N^{-1}. In comparison with the leading term, the second term in (2.11) 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 ((2.1) and (2.2)) has not fully eliminated the 'degrees–of–freedom' bias[5] of order $O(N^{-2}h^{-2d})$.

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 (2.11) that the best $O\left(N^{-1}\right)$ rate of the MSE, 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' vs '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 (2.3), 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

$$h^{+}_{\mathbf{opt}} \propto N^{-1/(2P+d)}. \tag{2.12}$$

A direct comparison with (2.10) 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 A_0 be the proportionality

[5]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 (2003)'s estimator. This method can potentially be adapted to our framework.

constant[6] in (2.12). It then follows that

$$\begin{aligned} h_{\mathbf{opt}} &= D_0 E_N h^+_{\mathbf{opt}}, \text{ where} \\ D_0 &= C_0/A_0, \\ E_N &= N^{-P/[(2P+d)(P+d)]}. \end{aligned}$$

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 (2.11), we can 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 (2.8) and (2.9). 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 which 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, $a_\tau(v, \mathbf{u})$ in their construction[7], that is

$$\widehat{\mathfrak{B}}_1(h_0) = \frac{\widetilde{\eta}(\Delta h_0) - \widetilde{\eta}(h_0)}{h_0^P(\Delta^P - 1)}, \tag{3.1}$$

$$\widehat{\mathfrak{B}}_2(h_*) = \frac{C_{W\mathcal{K}}}{N} \sum_{i=1}^{N} \widehat{\varpi}_{*3\tau i}, \tag{3.2}$$

where $\widehat{\varpi}_{*3\tau i} = w_i a_\tau(v_i, \mathbf{u}_i) \widehat{f}_{*\mathbf{U}i}/\widehat{f}^2_{*V\mathbf{U}i}$, $\varpi_{3\tau i} = w_i a_\tau(v_i, \mathbf{u}_i) f_{\mathbf{U}i}/f^2_{V\mathbf{U}i}$, and Δ is a known constant which is greater than 1. Here we have used $\widehat{f}_{*V\mathbf{U}i} \equiv \widehat{f}_{V\mathbf{U}}(v_i, \mathbf{u}_i; h_*)$, $\widehat{f}_{*\mathbf{U}i} \equiv \widehat{f}_{\mathbf{U}}(\mathbf{u}_i; h_*)$, $f_{V\mathbf{U}i} \equiv f_{V\mathbf{U}}(v_i, \mathbf{u}_i)$, 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 (2.1), and (2.2) respectively. The estimator $\widetilde{\eta}(\cdot)$ is like (2.3), after replacing ω_i by $\omega_i a_\tau(v_i, \mathbf{u}_i)$ everywhere. The estimator (3.1) is similar to the average derivative as proposed

[6] See Bashtannyk and Hyndman (2001) and Chen, Linton, and Robinson (2001) for derivations.

[7] 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.

by Powell and Stoker (1996). The quantities h_* and h_0 are pilot bandwidths which must be chosen beforehand.

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

$$a_\tau(v_i, \mathbf{u}_i) = 1\left(v_i \in \left[\underline{V} + \tau, \overline{V} - \tau\right]\right) \prod_{j=1}^{d-1} 1\left(\mathbf{u}_i^{[j]} \in \left[\underline{U}_j + \tau, \overline{U}_j - \tau\right]\right), \tag{3.3}$$

where $1(\cdot)$ 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 j–th element of the vector $\mathbf{u}$. By using this type of trimming in the construction of our estimators (3.1) and (3.2), 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 are not affected.

(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 (3.3), 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 (3.3). 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

$$\begin{aligned} \widehat{h}_{\mathbf{opt}} &= \widehat{C}_0 \times N^{-1/(P+d)}, \text{ where} \\ \widehat{C}_0 &= \underset{C_0 \in \Re^{++}}{\arg\min} \left\| \widehat{\mathfrak{B}}_1 C_0^P + \widehat{\mathfrak{B}}_2 \frac{1}{C_0^d} \right\|^2. \end{aligned} \tag{3.4}$$

An interesting characteristic of estimators (3.1) and (3.2) 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 (3.4) is also computationally straightforward, because it only involves a univariate numerical search over strictly positive real numbers. The consistency[8] of this procedure is ensured by the following proposition:

[8] 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.

Proposition 3.1 *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. ■

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 tunning 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[9], see Silverman (1986); or using a reference rule for kernel–based conditional density estimators, see Section 4.1. 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 till certain pre–defined convergence criteria is 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 the proposed 'plug–in' estimator $\widetilde{\eta}(\widehat{h}_{\mathbf{opt}})$ is asymptotically equivalent to $\widehat{\eta}(h_{\mathbf{opt}})$. 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 it is used in the construction of $\widetilde{\eta}$. All the calculations used to derive the asymptotic MSE expansion in Theorem 2.2 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.

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

4 Numerical Results

4.1 Monte Carlo

This section 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 a two–dimensional random variable (V,U) having a bivariate normal distribution doubly truncated with respect to both variables. The joint distribution is given by

$$f_{VU}(v,u) = g(v,u)/G, \quad \underline{v} \leq v \leq \overline{v}, \underline{u} \leq u \leq \overline{u},$$

where

$$g(v,u) = \frac{1}{2\pi\sigma_v\sigma_u\sqrt{1-\rho^2}} \times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{v-\mu_v}{\sigma_v}\right)^2 - 2\rho\left(\frac{v-\mu_v}{\sigma_v}\right)\left(\frac{u-\mu_u}{\sigma_u}\right) + \left(\frac{u-\mu_u}{\sigma_u}\right)^2\right]\right\},$$

and

$$G = \int_{\underline{u}}^{\overline{u}}\int_{\underline{v}}^{\overline{v}} g(v,u)\,dvdu.$$

The marginal density of U is then given by

$$f_U(u) = h(u)/G, \quad \underline{u} \leq u \leq \overline{u},$$

where

$$h(u) = \frac{1}{\sigma_u}\phi\left(\frac{u-\mu_u}{\sigma_u}\right)\left[\Phi\left(\left(\left(\frac{\overline{v}-\mu_v}{\sigma_v}\right) - \rho\left(\frac{u-\mu_u}{\sigma_u}\right)\right)/\sqrt{1-\rho^2}\right) - \Phi\left(\left(\left(\frac{\underline{v}-\mu_v}{\sigma_v}\right) - \rho\left(\frac{u-\mu_u}{\sigma_u}\right)\right)/\sqrt{1-\rho^2}\right)\right],$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ represent the density and cumulative distribution of a standard normal random variable. The object of interest in this simulation is

$$\eta = E\left[1/f_{V|U}(V|U)\right],$$

where

$$f_{V|U}(v|u) = g(v,u)/h(u).$$

For simplicity, we set $\underline{v} = \underline{u} = -3$, $\overline{v} = \overline{u} = 3$, $\mu_v = 0$, $\sigma_v^2 = \sigma_u^2 = 6$, and consider 3 designs

Design 1: $\mu_u = 0$;

Design 2: $\mu_u = 1$;

Design 3: $\mu_u = 2$.

For each design, we consider 2 cases based on possible values for ρ: (a) $\rho = 0$, and (b) $\rho = 1/4$. Their associated joint, conditional and marginal densities can be visualized in Figures 1, 2, and 3. These designs were chosen so that their associated marginal densities are bounded well above zero at the boundary of their support. A similar property is displayed by their conditional densities.

We set $P = 2$, and set $W \equiv K$ to be a gaussian second–order kernel. Their associated constants are $d_K = 1$, and $C_K = 1/2\sqrt{\pi}$.

Reference Rules

Preliminary bandwidths employed in this simulation study are based on the following three assumptions underlying the joint distribution of (V, U):

(R1) $f_{VU}(v,u) \equiv g(u,v)$, with $\underline{v} = \underline{u} = -\infty$, $\overline{v} = \overline{u} = +\infty$, $\mu_v = \mu_u = u$, and $\sigma_v^2 = \sigma_u^2 = \sigma^2$. Under this assumption, Chen, Linton, and Robinson (2001) calculated

$$A_0 = \sigma \left[\frac{16\pi\sqrt{2}\left(1-\rho^2\right)^{5/2} C_K^2}{\left(15\rho^4 - 50\rho^2 + 39\right) d_K^2} \right]^{1/6} \equiv A_{R1}.$$

(R2) $V \,|\, U = u \sim N(c + du, (p+qu)^2)$, U is uniform over $[\underline{u}, \overline{u}]$, with $-\infty \leq V \leq +\infty$. Under this assumption, similar calculations to those in Bashtannyk and Hyndman (2001) shows

$$A_0 = \left[\frac{256 q \sqrt{\pi} C_K^2}{3z\left(4 + w + 8d^2 - 12q^2\right) d_K^2} \right]^{1/6} \equiv A_{R2},$$

where $w = 19q^4 + 4d^4 + 28q^2d^2$, and $z = \left[(p+q\overline{u})^4 - (p+q\underline{u})^4\right] / \left(p+q\overline{u}\right)^4 (p+q\underline{u})^4$.

(R3) $f_{VU}(v,u) \equiv g(u,v)$, with $\underline{v} = \underline{u} = -\infty$, $\overline{v} = \overline{u} = +\infty$, $\mu_v = \mu_u = 0$, $\rho = 0$, and $\sigma_v^2 = \sigma^2$. Under this assumption, $f_{V|U}(v|u) = f_V(v)$, for which Silverman (1986), pages 45–47, calculated

$$A_0 = \left[\frac{8\sqrt{\pi} C_K}{3 d_K^2} \right]^{1/5} \sigma \equiv A_{R3}.$$

We make these reference rules operational by making A_{R1}, A_{R2}, and A_{R3} vary with each replication. We define these quantities as $\widehat{A}_{R1}$, $\widehat{A}_{R2}$, and $\widehat{A}_{R3}$ respectively. Specifically, let $\{v_i^s, u_i^s\}_{i=1}^{N^s}$ be a size–N^s generated data set at draw s, then $\widehat{A}_{R1}$ is obtained by replacing σ^2, and ρ by $\widehat{\sigma}^2 = N^{-1}\sum_{i=1}^N (v_i^s - \overline{v}^s)$, and $\widehat{\rho} = [\widehat{\sigma}^2 (N-1)]^{-1} \sum_{i=1}^N (v_i^s - \overline{v}^s)(u_i^s - \overline{u}^s)$ respectively. Likewise, $\widehat{A}_{R2}$ is calculated by setting $\underline{u} = \min_{i=1,\ldots,N} u_i^s$, $\overline{u} = \max_{i=1,\ldots,N} u_i^s$, (c,d) as the least squares coefficients from a regression of v on u, and (p,q) as the least squares coefficients from a regression of the squared residuals from the previous regressions on u including a constant term. Similarly, $\widehat{A}_{R3}$ is made operational by replacing σ^2 by $\widehat{\sigma}^2$ as calculated above.

Hence, we calculate $\widetilde{\eta}$ using $\widehat{h}_{R1} = \widehat{A}_{R1}N^{-1/6}$, $\widehat{h}_{R2} = \widehat{A}_{R2}N^{-1/6}$, and $\widehat{h}_{R3} = \widehat{A}_{R3}N^{-1/6}$. Of course, in our designs these bandwidths are neither optimal for η, nor do they have the optimal rate of convergence derived in Section 2.2.1. However, we have chosen them for comparison purposes because of their computational simplicity, as well as the fact that they were the most likely to be chosen by a practitioner prior to the results discussed in this paper.

We also look at the behavior of $\widetilde{\eta}$ using our 'plug–in' estimator for the optimal bandwidth explained in Section 3. We implement this estimator by setting $\widehat{h}_0 = \widehat{h}^*_{Rl}N^{\delta}$, where $\widehat{h}^*_{Rl} \equiv \widehat{h}_{R1}$ for $l = 1, 2, 3$. Other parameters are chosen accordingly and kept constant throughout the experiments, i.e. $\delta = 1/12$, $\tau = 0$ (no trimming) and $\Delta = 2$. The results of 2000 replications are presented in Tables 1 to 6.

Tables 1, 3, and 5 report the small sample performance of the proposed 'plug–in' estimator for the optimal bandwidth under different conditions. The true optimal bandwidths $h_{\mathbf{opt}}$, are also reported in the first row for each case. As we would expect, higher correlation between V and U entails a larger bandwidth in each design. These results show that the proposed 'plug–in' estimator performs fairly well in all circumstances. This good performance seems not to be affected by the choice of pilot bandwidths, h_* and h_0 in large samples. On the other hand, there is more variation among the bandwidths predicted by the reference rules than among the estimated ones. Numerically, differences among them become more evident when samples sizes are large. The bandwidths' simulated standard deviations increase as we increase the theoretical mean of U. The use of trimming could reduce these variances.

The respective MSE are presented in Tables 2, 4, and 6. We notice that the main component of these simulated MSE is bias instead of variance in each case, as is predicted by the expansion derived here. The use of either the theoretical or estimated optimal bandwidths dominates the use of those predicted by the reference rules in terms of MSE, for all sample sizes, designs and scenarios. The MSE associated to the estimated optimal bandwidths are numerically very close to the simulated theoretical ones.

In our designs, it is also the case that the 'degrees–of–freedom' bias is numerically large, up to 10 times greater than the 'smoothing' bias. Similar calculations for other designs (not presented here) have also shown such a pattern. This lends support to the use of an explicit bias correction mechanism for such term, see for example Ichimura and Linton (2005). This remains a topic for future research.

Finally, Figures 4, 5 and 6 show how close the theoretically optimal bandwidths are to the actual MSE–minimizing bandwidths. The MSE for $\widetilde{\eta}$ are obtained by simulation as functions of a grid of fixed bandwidth parameters. The vertical gray lines represent the optimal bandwidths predicted by Theorem 2.2 in each case. Note that even for small sample sizes, the approximation results are very good. However, the quality of the approximation may deteriorate

in situations where trimming is necessary.

4.2 Empirical Application

We now apply the results of this paper to the problem of selecting the bandwidth when estimating a binary response model of female labour market participation. The data set used for this analysis is a sample from the Ecuadorian Household Income and Expenditure Survey, which provides information on income, employment status, household composition and other socioeconomic characteristics. The data are from interview year 2004 and were collected by the Ecuadorian National Institute of Statistics, INEC. The analysis is limited to women aged 60 or below, whose husbands or unmarried partners earned labour income in 2004. The resulting sample data set contains 3447 women, 50% of whom were working for wages. Descriptive statistics are presented in Table 7. The model is

$$d_i = 1\left(\alpha v_i + \mathbf{x}_i^\top \beta + e_i > 0\right),$$

where d_i is the indicator of the i-th woman's response, v_i is the log of the partner's monthly income (LNHINC) normalized to have zero mean, and the other regressors $\mathbf{x}_i$ are a constant, the number of children not older than 3 (YCHILD), the number of older children (CHILD), years of formal education (EDUC), age divided by 10 (AGE), and age squared divided by 100 (AGE2). For more information about the model, and alternative semiparametric estimators, see Gerfin (1996), and Martins (2001).

Table 8 displays estimation results for parametric and semiparametric specifications, after using the normalization $|\alpha| = 1$. By assuming $e_i \sim N\left(0, \sigma_i^2\right)$, the first column from the left ($\sigma_i^2 = 1$), and the second ($\sigma_i^2 = \exp\left(\mathbf{z}^\top \gamma\right)$) are obtained by maximum likelihood estimation. The third and fourth columns report results using estimators proposed by Lewbel (2000b) and Lewbel and Schennach (2007) respectively. Standard errors are in parenthesis.

Lewbel's (2000b) estimator is equivalent to a standard linear least squares regression of $\left[d - 1\left(v > 0\right)\right] / \widehat{f}_{V|\mathbf{X}}\left(v|\,\mathbf{x}\right)$ on $\mathbf{x}$, where $\widehat{f}_{V|\mathbf{X}}$ is calculated using the gaussian kernel. The bandwidth was chosen using the procedure described in Section 3 where the object of interest was in this case $E[\mathbf{x}\left(d - 1\left(v > 0\right)\right) / f_{V|\mathbf{X}}\left(v|\,\mathbf{x}\right)]$. Figure 7 shows that the estimated optimal bandwidth is approximately 0.88. By using this bandwidth, 45 observations were excluded for which the estimator of f_{VX} was exactly zero. The same bandwidth was used to construct their asymptotic standard errors. The fourth estimator in Table 8 is the ordered data estimator in Lewbel and Schennach (2007, Corollary 7) which does not require any smoothing under rather strong conditions.

The two probit and kernel-based sets of estimates are quite similar in general. The kernel-based least squares parameter estimates are all within two standard errors of the simple and

heteroskedastic probit estimates. All set of estimates predicts that the effect of the number of children older than 3 years is small and statistically insignificant. On the other hand, this variable can significantly explain heteroskedasticity in a parametric fashion. However, for all other regressors, the ordered data estimates are generally different with bigger standard errors.

5 Discrete Covariates

In this section, we examine the situation in which the conditioning variables, $\mathbf{U}$, have continuous as well as discrete components. In this case, the order of magnitude of the optimal bandwidth only depends on the number of continuously distributed elements of the random vector $\left(V, \mathbf{U}^\top\right)$.

In particular, let us consider the case when the random vector, $\mathbf{U}$, can be partitioned as $\mathbf{U} = \left(\mathbf{U}^{(1)\top}, \mathbf{U}^{(2)\top}\right)$, with $\mathbf{U}^{(1)} \in \Omega_{\mathbf{U}^{(1)}}$, and $\mathbf{U}^{(2)} \in \Omega_{\mathbf{U}^{(2)}}$, where $\Omega_{\mathbf{U}^{(1)}} \subset \Re^{d^{(1)}-1}$, and $\Omega_{\mathbf{U}^{(2)}} \subset \Re^{d^{(2)}}$ is a set with finite number of real points, such that $d^{(1)} + d^{(2)} = d$ with $d^{(1)} \geq 1$ as before. Let $f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}\left(v, \mathbf{u}^{(1)} \middle| \mathbf{u}^{(2)}\right)$ be the probability density of $\left(V, \mathbf{U}^{(1)}\right)$ conditional on $\mathbf{U}^{(2)} = \mathbf{u}^{(2)}$, let $f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}\left(\mathbf{U}^{(1)} \middle| \mathbf{U}^{(2)}\right)$ be the probability density of $\mathbf{U}^{(1)}$ conditional on $\mathbf{U}^{(2)} = \mathbf{u}^{(2)}$, and let $p\left(\mathbf{u}^{(2)}\right)$ be the probability mass that $\mathbf{u}^{(2)} \in \Omega_{\mathbf{U}^{(2)}}$. Then

$$f_{V\mathbf{U}}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv f_{V\mathbf{U}}(v, \mathbf{u}) = f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v, \mathbf{u}^{(1)} \Big| \mathbf{u}^{(2)}) p(\mathbf{u}^{(2)}), \text{ and}$$
$$f_{\mathbf{U}}(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv f_{\mathbf{U}}(\mathbf{u}) = f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)} \Big| \mathbf{u}^{(2)}) p(\mathbf{u}^{(2)}).$$

We also replace (2.1) and (2.2) with

$$\widehat{f}_{\mathbf{U}}(\mathbf{u}_i^{(1)}, \mathbf{u}_i^{(2)}; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d^{(1)}-1}} \mathcal{K}\left(\frac{\mathbf{u}_j^{(1)} - \mathbf{u}_i^{(1)}}{h}\right) 1\left(\mathbf{u}_j^{(2)} = \mathbf{u}_i^{(2)}\right), \text{ and} \tag{5.1}$$

$$\widehat{f}_{V\mathbf{U}}(v_i, \mathbf{u}_i^{(1)}, \mathbf{u}_i^{(2)}; h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^{d^{(1)}}} W\left(\frac{v_j - v_i}{h}\right) \mathcal{K}\left(\frac{\mathbf{u}_j^{(1)} - \mathbf{u}_i^{(1)}}{h}\right) 1\left(\mathbf{u}_j^{(2)} = \mathbf{u}_i^{(2)}\right), \tag{5.2}$$

respectively, and recalculate (2.3) and (2.4). We also redefine $\pi_1(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \pi_1(\mathbf{u})$, $\pi_2(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \pi_2(\mathbf{u})$, $\widetilde{\pi}_1\left(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) \equiv \widetilde{\pi}_1(v, \mathbf{u})$, $\widetilde{\pi}_2\left(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) \equiv \widetilde{\pi}_2(v, \mathbf{u})$, $\pi_3(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \pi_3(v, \mathbf{u})$, $\pi_4(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \pi_4(v, \mathbf{u})$, $\pi_5(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \pi_5(v, \mathbf{u})$, $\sigma_1^2(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \sigma_1^2(\mathbf{u})$, $\sigma_2^2(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \sigma_2^2(\mathbf{u})$, $\sigma_{12}\left(\mathbf{u}^{(1)}, \mathbf{u}^{(2)}\right) \equiv \sigma_{12}(\mathbf{u})$, $\widetilde{\sigma}_1^2(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_1^2(v, \mathbf{u})$, $\widetilde{\sigma}_{13}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{13}(v, \mathbf{u})$, $\widetilde{\sigma}_{14}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{14}(v, \mathbf{u})$, $\widetilde{\sigma}_{15}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{15}(v, \mathbf{u})$, $\widetilde{\sigma}_{23}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{23}(v, \mathbf{u})$, $\widetilde{\sigma}_{24}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{24}(v, \mathbf{u})$, $\widetilde{\sigma}_{25}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{25}(v, \mathbf{u})$, $\sigma_3^2(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \sigma_3^2(v, \mathbf{u})$, $\widetilde{\sigma}_{34}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{34}(v, \mathbf{u})$, $\widetilde{\sigma}_{35}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{35}(v, \mathbf{u})$, $\sigma_4^2(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)})$, $\widetilde{\sigma}_{45}(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \widetilde{\sigma}_{45}(v, \mathbf{u})$, and $\sigma_5^2(v, \mathbf{u}^{(1)}, \mathbf{u}^{(2)}) \equiv \sigma_5^2(v, \mathbf{u})$.

In order to extend our results to this mixed case, we need to re–state Assumptions (A3) and (A6) as:

(A3*) $f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\pi_1(\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$ and $\pi_2(\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, understood as functions of $\mathbf{u}^{(1)}$, exist and have bounded continuous partial derivatives up to the order P on $\Omega_{\mathbf{U}^{(1)}} \equiv \prod_{j=1}^{d^{(1)}-1}\left[\underline{U}_j^{(1)},\overline{U}_j^{(1)}\right]$, where $-\infty < \underline{U}_j^{(1)} < \overline{U}_j^{(1)} < \infty$, for $j = 1,\ldots,d^{(1)}-1$. Furthermore, $f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\widetilde{\pi}_1(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) \times f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\widetilde{\pi}_2(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\pi_3(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\pi_4(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, and $\pi_5(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}|\mathbf{u}^{(2)})$ understood as functions of v and $\mathbf{u}^{(1)}$, exist and have bounded continuous partial derivatives up to the order P on $\Omega_{V\mathbf{U}^{(1)}} \equiv \left[\underline{V},\overline{V}\right] \times \Omega_{\mathbf{U}^{(1)}}$, for $-\infty < \underline{V} < \overline{V} < \infty$. The probability mass function $p(\mathbf{u}^{(2)}) > 0$.

(A6*) The functions, $\sigma_1^2(\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\sigma_2^2(\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$ and $\sigma_{12}\left(\mathbf{u}^{(1)},\mathbf{u}^{(2)}\right)f_{\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, understood as functions of $\mathbf{u}^{(1)}$, are bounded on their compact support $\Omega_{\mathbf{U}^{(1)}}$. Similarly, $\widetilde{\sigma}_1^2(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\sigma_3^2(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})$ $f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\sigma_4^2(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, $\sigma_5^2(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}$ $(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, and $\widetilde{\sigma}_{lk}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})f_{V\mathbf{U}^{(1)}|\mathbf{U}^{(2)}}(v,\mathbf{u}^{(1)}\big|\,\mathbf{u}^{(2)})$, for $\forall l,k = 1,\,2,\,3,\,4$ such that $l \neq k$, understood as functions of v and $\mathbf{u}^{(1)}$, are bounded on their compact support $\Omega_{V\mathbf{U}^{(1)}}$.

As expected, in this mixed case scenario, similar conditions have to be imposed on the continuous part of the problem, but no new techniques are required in order to prove the following corollary:

Corollary 5.0.1 *Let Assumptions* (A1), (A2) *hold, and Assumptions* (A3*) *and* (A6*) *hold for every* $\mathbf{u}^{(2)} \in \Omega_{\mathbf{u}^{(2)}}$*, then*

$$h_{\mathbf{opt}}^{(1)} = C_0^{(1)} \times \left(\frac{1}{N}\right)^{1/(P+d^{(1)})}, \text{ where}$$
$$C_0^{(1)} = \underset{C_0^{(1)}\in\Re^{++}}{\arg\min} \left\| \mathfrak{B}_1^{(1)} C_0^P + \mathfrak{B}_2^{(1)} \frac{1}{C_0^d} \right\|^2 .$$

and

$$\mathfrak{B}_1^{(1)} = \sum_{\mathbf{u}^{(2)}\in\Omega_{\mathbf{u}^{(2)}}} \int \pi_2(\mathbf{u}^{(1)},\mathbf{u}^{(2)}) S_K^{(1)}(\mathbf{u}^{(1)},\mathbf{u}^{(2)}) f_{\mathbf{U}}(\mathbf{u}^{(1)},\mathbf{u}^{(2)}) d\mathbf{u}^{(1)}$$
$$- \sum_{\mathbf{u}^{(2)}\in\Omega_{\mathbf{u}^{(2)}}} \int \pi_3(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) S_{W\mathcal{K}}^{(1)}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) f_{V\mathbf{U}}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) dv d\mathbf{u}^{(1)}, \tag{5.3}$$

$$\mathfrak{B}_2^{(1)} = C_{W\mathcal{K}} \sum_{\mathbf{u}^{(2)}\in\Omega_{\mathbf{u}^{(2)}}} \int \pi_3(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) f_{V\mathbf{U}}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) dv d\mathbf{u}^{(1)}, \tag{5.4}$$

with

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

$$S_K^{(1)}(\mathbf{u}^{(1)},\mathbf{u}^{(2)}) = d_K \frac{1}{P!}\sum_{j=1}^{d^{(1)}-1}\frac{\partial^P}{(\partial u_j^{(1)})^P} f_{\mathbf{u}}(\mathbf{u}^{(1)},\mathbf{u}^{(2)}),$$

$$S_{W\mathcal{K}}^{(1)}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) = \frac{1}{P!}\left[d_W \frac{\partial^P}{\partial v^P} f_{V\mathbf{U}}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)}) + d_K \sum_{j=1}^{d^{(1)}-1}\frac{\partial^P}{(\partial u_j^{(1)})^P} f_{V\mathbf{U}}(v,\mathbf{u}^{(1)},\mathbf{u}^{(2)})\right].$$

Proof. See proof of Theorem 2.2 in the Appendix. ■

From this result, and similarly to other nonparametric and semiparametric models (see Delgado and Mora (1995)), we note that the MSE–minimizing rate of bandwidth shrinkage in our case only depends on the number of continuously distributed random variables in our sample from $(V,\mathbf{U})$. The unknown constants (5.4) and (5.3) can be consistently estimated, by simple extensions of the estimators described in Section 3. Likewise, trimming (if needed) should be performed only with respect to the continuously distributed variables, in particular

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

6 Conclusion

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 asymptotically represented as 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. The derived formula for the optimal bandwidth equates the order of magnitude arising from the squared 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 towards zero at a faster rate than the optimal for its nonparametric component. In this 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 formulae. 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 which 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. An empirical implementation is also performed regarding the labour force participation decision for Ecuadorian women in 2004. We also discuss how the formula for the optimal bandwidth can be adapted when continuous as well as discrete elements are present in the weighted averages.

References

Bashtannyk, D. M., and R. J. Hyndman (2001): "Bandwidth Selection for Kernel Conditional Density Estimation," *Computational Statistics & Data Analysis*, 36, 279–298.

Chen, X., O. B. Linton, and P. M. Robinson (2001): "The Estimation of Conditional Densities," in *In: Asymptotics in Statistics and Probability: Papers in Honor of George Gregory Roussas*, ed. by M. L. Puri, pp. 71–84. VSP International Science Publishers, The Netherlands, 1 edn.

Collomb, G., and 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.

Delgado, M. A., and J. Mora (1995): "On Asymptotic Inferences in Non-parametric and Semiparametric Models with Discrete and Mixed Regressors," *Investigaciones Económicas*, 19(3), 435–467.

Fernandes, M., and P. K. Monteiro (2005): "Central Limit Theorem for Asymmetric Kernel Functionals," *Annals of the Institute of Statistical Mathematics*, 57(3), 425–442.

Gasser, T., H.-G. Müller, and V. Mammitzsch (1985): "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society. Series B (Methodological)*, 47, 238–252.

Gerfin, M. (1996): "Parametric and Semi-parametric Estimation of the Binary Response Model of Labour Market Participation," *Journal of Applied Econometrics*, 11(3), 321–339.

Goldstein, L., and K. Messer (1992): "Optimal Plug-in Estimators for Nonparametric Functional Estimation," *The Annals of Statistics*, 20(3), 1306–1328.

Hall, P., and J. S. Marron (1987): "Estimation of Integrated Squared Density Derivatives," *Statistica and Probability Letters*, 6(2), 109–115.

Härdle, W., J. D. Hart, J. S. Marron, and A. B. Tsybakov (1992): "Bandwidth Choice for Average Derivative Estimation," *Journal of The American Statistical Association*, 87(417), 218–226.

Härdle, W., and 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., and A. B. Tsybakov (1993): "How Sensitive are Average Derivatives?," *Journal of Econometrics*, 58, 31–48.

Hirano, K., G. W. Imbens, and G. Ridder (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

Honoré, B. E., and A. Lewbel (2002): "Semiparametric Binary Choice Panel Data Models Without Strictly Exogenous Regressors," *Econometrica*, 70(5), 2053–2063.

Horowitz, J. L., and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632–1640.

Hyndman, R. J., D. M. Bashtannyk, and G. K. Grunwald (1996): "Estimating and Visualizing Conditional Densities," *Journal of Computational and Graphical Statistics*, 5(4), 315–336.

Ichimura, H., and O. B. Linton (2005): "Asymptotic Expansions for some Semiparametric Program Evaluation Estimators," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg,*, ed. by D. W. K. Andrews, and J. H. Stock, chap. 8, pp. 149–170. Cambridge University Press, Cambridge, 1 edn.

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., and 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., and 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.

——— (2000a): "Asymptotic Trimming for Bounded Density Plug-in Estimators," Boston College. Unpublished Manuscript.

——— (2000b): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97(1), 145–177.

——— (2006): "Endogenous Selection or Treatment Model Estimation," Forthcoming in Journal of Econometrics.

Lewbel, A., and S. M. Schennach (2007): "A simple ordered data estimator for inverse density weighted expectations," *Journal of Econometrics*, 127(1), 189–211.

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

——— (1995): "Second Order Approximation in the Partially Linear Regression Model," *Econometrica*, 63(3), 1079–1112.

Martins, M. F. O. (2001): "Parametric and Semiparametric Estimation of Sample Selection Models: An Empirical Application to the Female Labour Force in Portugal," *Journal of Applied Econometrics*, 16(1), 23–39.

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, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Powell, J. L., and 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 *Multivariate Analysis II*, ed. by P. R. Krishnaiah, pp. 25–31. Academic Press, New York.

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.

——— (1986): *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1 edn.

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

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):

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

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

Proof of Lemma 2.1

Firstly, from Assumption (A3), it follows that $\widehat{\vartheta}_{2i}\left(h\right)$ is

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

where (A-3) follows after observing that

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

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

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

as required.

Sketch of Proof of Theorem 2.2

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 (2.6). Firstly, let us denote $\delta_1 = E[\varpi_1]$, $\delta_2 = E[\pi_2(\mathbf{U}) f_{\mathbf{U}}(\mathbf{U})]$, $\delta_3 = E[\pi_3(V,\mathbf{U}) f_{V\mathbf{U}}(V,\mathbf{U})]$, $\delta_4 = E[\pi_4(V,\mathbf{U}) f_{\mathbf{U}}(\mathbf{U}) f_{V\mathbf{U}}(V,\mathbf{U})]$, and $\delta_5 = E\left[\pi_5(V,\mathbf{U}) f^2_{V\mathbf{U}}(V,\mathbf{U})\right]$. Then, by using the definitions in Section 2 and the properties of conditional expectations, it follows that

$$\delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = \eta.$$

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

$$\begin{aligned}
E\left[\|\widehat{\eta}(h) - \eta\|^2\right] &= E\left[\left\|\widehat{\delta}_1 - \delta_1\right\|^2\right] && \text{(A-4)}\\
&+ 4E\left[\left\|\widehat{\delta}_2(h) - \delta_2\right\|^2\right] && \text{(A-5)}\\
&+ 4E\left[\left\|\widehat{\delta}_3(h) - \delta_3\right\|^2\right] && \text{(A-6)}\\
&+ E\left[\left\|\widehat{\delta}_4(h)\right\|^2\right] && \text{(A-7)}\\
&+ E\left[\left\|\widehat{\delta}_5(h)\right\|^2\right] && \text{(A-8)}\\
&+ 4E\left[\left\langle\widehat{\delta}_1 - \delta_1, \widehat{\delta}_2(h) - \delta_2\right\rangle\right] && \text{(A-9)}\\
&- 4E\left[\left\langle\widehat{\delta}_1 - \delta_1, \widehat{\delta}_3(h) - \delta_3\right\rangle\right] && \text{(A-10)}\\
&- 2E\left[\left\langle\widehat{\delta}_1 - \delta_1, \widehat{\delta}_4(h) - \delta_4\right\rangle\right] && \text{(A-11)}\\
&+ 2E\left[\left\langle\widehat{\delta}_1 - \delta_1, \widehat{\delta}_5(h) - \delta_5\right\rangle\right] && \text{(A-12)}\\
&- 8E\left[\left\langle\widehat{\delta}_2(h) - \delta_2, \widehat{\delta}_3(h) - \delta_3\right\rangle\right] && \text{(A-13)}\\
&- 4E\left[\left\langle\widehat{\delta}_2(h) - \delta_2, \widehat{\delta}_4(h) - \delta_4\right\rangle\right] && \text{(A-14)}\\
&+ 4E\left[\left\langle\widehat{\delta}_2(h) - \delta_2, \widehat{\delta}_5(h) - \delta_5\right\rangle\right] && \text{(A-15)}\\
&+ 4E\left[\left\langle\widehat{\delta}_3(h) - \delta_3, \widehat{\delta}_4(h) - \delta_4\right\rangle\right] && \text{(A-16)}\\
&- 4E\left[\left\langle\widehat{\delta}_3(h) - \delta_3, \widehat{\delta}_5(h) - \delta_5\right\rangle\right] && \text{(A-17)}\\
&- 2E\left[\left\langle\widehat{\delta}_4(h) - \delta_4, \delta_5(h) - \delta_5\right\rangle\right]. && \text{(A-18)}
\end{aligned}$$

Let us define the following quantities:

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

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

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

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:	$\left\|\mathfrak{B}_{1,1}-\mathfrak{B}_{1,2}\right\|^2$	$2\left\langle\mathfrak{B}_{1,1}-\mathfrak{B}_{1,2},\mathfrak{B}_2\right\rangle$	$\left\|\mathfrak{B}_2\right\|^2$

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

$$\begin{aligned}
&h^{2P}\left\|\int \pi_2(\mathbf{u})\, S_{\mathcal{K}}(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} - \int \pi_3(v,\mathbf{u})\, S_{W\mathcal{K}}(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dvd\mathbf{u}\right\|^2 \\
&+ \frac{2C_{W\mathcal{K}}}{Nh^d}h^P\left\langle \int \pi_3(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dvd\mathbf{u}\right. \\
&\left., \int \pi_2(\mathbf{u})\, S_{\mathcal{K}}(\mathbf{u})\, f_{\mathbf{U}}(\mathbf{u})\, d\mathbf{u} - \int \pi_3(v,\mathbf{u})\, S_{W\mathcal{K}}(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dvd\mathbf{u}\right\rangle \\
&+ \frac{C_{W\mathcal{K}}^2}{N^2h^{2d}}\left\|\int \pi_3(v,\mathbf{u})\, f_{V\mathbf{U}}(v,\mathbf{u})\, dvd\mathbf{u}\right\|^2 \\
&= h^{2P}\left\|\mathfrak{B}_{1,1}-\mathfrak{B}_{1,2}\right\|^2 + 2\frac{h^P}{Nh^d}\left\langle\mathfrak{B}_2,\mathfrak{B}_{1,1}-\mathfrak{B}_{1,2}\right\rangle + \frac{1}{N^2h^{2d}}\left\|\mathfrak{B}_2\right\|^2 \\
&= \left\|h^P\mathfrak{B}_1 + \mathfrak{B}_2N^{-1}h^{-d}\right\|^2,
\end{aligned}$$

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

Proof of Proposition 3.1

In what follows, we make use of the following identities:

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

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

Term: $\widehat{\mathfrak{B}}_1(h_0)$

Firstly, it follows from (A-19) that

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

where (A-21) and (A-22) are $O_p(N^{-1}h_0^{-d}\log N+h_0^{2P})$ because of Assumptions (A1), (A2), (A3), (A4), and results (A-20), (A-1), (A-2). That is,

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

which means that $\widehat{\mathfrak{B}}_1(h_0)$ is the sum of two U–statistics plus a reminder 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, it then follows from Lemma 3.1 (page 1410) in Powell, Stock, and Stoker (1989), and Theorem A (page 4) in Lewbel (2000a), that $\left(\widetilde{\eta}\left(\Delta h_0\right)-\widetilde{\eta}\left(h_0\right)\right)/(h_0^P(\Delta^P-1))$ is consistent for

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

This is true because,

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

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

Term: $\widehat{\mathfrak{B}}_2\left(h_*\right)$

Notice that,

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

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

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

which is $o_p\left(1\right)$ by Assumption (A4), representation (A-20), and the assumptions of the proposition ($h_*\to0$ 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 conclude that

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

Appendix B: Technical Lemmas

Lemma B-1 *Let Assumptions* (A1)–(A3), (A4), (A6) *and* (A7) *hold, then*

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

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

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

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

where $c_\Delta=\left(1-\Delta^2\right)/\left(\Delta\left(1-\Delta^P\right)\right)^2$, and $\varrho_{0\tau}\left(v,\mathbf{u}\right)\equiv E\left[\varpi_{3\tau 1}\middle|V_1=v,\mathbf{U}_1=\mathbf{u}\right]$. The second equality uses the change of variables from $(y,\mathbf{y}^{\top},z,\mathbf{z}^{\top})$ to $(c=h_0^{-1}\left(y-z\right),\mathbf{c}^{\top}=h_0^{-1}(\mathbf{y}-\mathbf{z})^{\top},z,\mathbf{z}^{\top})$ 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 $a_\tau\left(z,\mathbf{z}\right)=0$ for all $(z,\mathbf{z})$ within a distance τ of the boundary of $\Omega_{V\mathbf{U}}$, 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 exact same arguments and therefore it is omitted. ■

Table 1: Monte Carlo results for Design 1: Bandwidth Estimation

		$N = 200$		$N = 400$		$N = 600$	
		$Mean$	SD	$Mean$	SD	$Mean$	SD
$\rho = 0$	$h_{\mathbf{opt}}$	0.6530	–	0.5491	–	0.4962	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7006	0.0583	0.5654	0.0228	0.5024	0.0151
	$\widehat{h}_{\mathbf{opt};R2}$	0.6994	0.0628	0.5651	0.0227	0.5024	0.0144
	$\widehat{h}_{\mathbf{opt};R3}$	0.6877	0.0311	0.5626	0.0146	0.5019	0.0104
	$\widehat{h}_{R1}$	0.5955	0.0220	0.5317	0.0135	0.4968	0.0102
	$\widehat{h}_{R2}$	0.6055	0.0395	0.5450	0.0237	0.5100	0.0178
	$\widehat{h}_{R3}$	0.6846	0.0252	0.6110	0.0155	0.5706	0.0117
$\rho = 1/4$	$h_{\mathbf{opt}}$	0.6664	–	0.5603	–	0.5063	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7024	0.0802	0.5698	0.0354	0.5059	0.0192
	$\widehat{h}_{\mathbf{opt};R2}$	0.7029	0.1012	0.5693	0.0399	0.5058	0.0195
	$\widehat{h}_{\mathbf{opt};R3}$	0.6872	0.0399	0.5636	0.0190	0.5033	0.0124
	$\widehat{h}_{R1}$	0.5918	0.0210	0.5283	0.0133	0.4938	0.0102
	$\widehat{h}_{R2}$	0.5931	0.0385	0.5336	0.0237	0.5003	0.0181
	$\widehat{h}_{R3}$	0.6823	0.0243	0.6083	0.0153	0.5686	0.0118

[a] Means and standard deviations (SD) are based on 2000 replications.

[b] Estimated bandwidths: $\widehat{h}_{\mathbf{opt};R1}$, $\widehat{h}_{\mathbf{opt};R2}$, and $\widehat{h}_{\mathbf{opt};R3}$ were calculated by setting $h_0 = \widehat{h}_{R1}N^{1/12}$, $h_0 = \widehat{h}_{R2}N^{1/12}$, and $h_0 = \widehat{h}_{R3}N^{1/12}$ in Section 3, respectively. Similarly, auxiliary bandwidths were set $h_* = \widehat{h}_{R1}$, $h_* = \widehat{h}_{R2}$, and $h_* = \widehat{h}_{R3}$ respectively.

Table 2: Monte Carlo results for Design 1: Parameter η

		$N = 200$			$N = 400$			$N = 600$		
		$Bias$	SD	MSE	$Bias$	SD	MSE	$Bias$	SD	MSE
$\rho = 0$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.8092	0.1598	0.6548	0.6593	0.0953	0.4347	0.5882	0.0731	0.3460
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7932	0.1594	0.6291	0.6583	0.0961	0.4333	0.5878	0.0741	0.3455
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7924	0.1595	0.6279	0.6585	0.0962	0.4336	0.5880	0.0741	0.3458
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7980	0.1576	0.6368	0.6596	0.0959	0.4350	0.5884	0.0740	0.3463
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8378	0.1658	0.7019	0.6616	0.0955	0.4377	0.5896	0.0737	0.3477
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8387	0.1641	0.7034	0.6637	0.0960	0.4405	0.5924	0.0747	0.3510
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.8124	0.1575	0.6600	0.6700	0.0978	0.4489	0.6062	0.0764	0.3675
$\rho = 1/4$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.7816	0.1650	0.6108	0.6395	0.1016	0.4090	0.5751	0.0773	0.3307
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7688	0.1641	0.5911	0.6357	0.1014	0.4041	0.5738	0.0780	0.3292
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7688	0.1725	0.5911	0.6355	0.1018	0.4039	0.5737	0.0780	0.3291
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7720	0.1591	0.5959	0.6371	0.1010	0.4059	0.5740	0.0778	0.3295
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8166	0.1855	0.6669	0.6453	0.1037	0.4164	0.5759	0.0776	0.3317
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8230	0.1989	0.6774	0.6468	0.1047	0.4183	0.5783	0.0780	0.3344
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.7851	0.1628	0.6164	0.6436	0.1018	0.4142	0.5868	0.0794	0.3444

[a] Simulated biases, standard deviations, and average Mean Squared Error (MSE) are based on 2000 replications.

Table 3: Monte Carlo results for Design 2: Bandwidth Estimation

		$N = 200$		$N = 400$		$N = 600$	
		$Mean$	SD	$Mean$	SD	$Mean$	SD
$\rho = 0$	$h_{\mathbf{opt}}$	0.6595	–	0.5545	–	0.5011	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7027	0.0826	0.5671	0.0331	0.5032	0.0175
	$\widehat{h}_{\mathbf{opt};R2}$	0.7011	0.0951	0.5665	0.0343	0.5031	0.0166
	$\widehat{h}_{\mathbf{opt};R3}$	0.6883	0.0408	0.5630	0.0186	0.5021	0.0112
	$\widehat{h}_{R1}$	0.5961	0.0214	0.5317	0.0134	0.4971	0.0103
	$\widehat{h}_{R2}$	0.6053	0.0398	0.5445	0.0244	0.5107	0.0185
	$\widehat{h}_{R3}$	0.6853	0.0246	0.6109	0.0154	0.5710	0.0118
$\rho = 1/4$	$h_{\mathbf{opt}}$	0.6755	–	0.5680	–	0.5133	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7090	0.1729	0.5722	0.0652	0.5084	0.0299
	$\widehat{h}_{\mathbf{opt};R2}$	0.7100	0.2604	0.5719	0.0724	0.5082	0.0322
	$\widehat{h}_{\mathbf{opt};R3}$	0.6906	0.0712	0.5655	0.0324	0.5050	0.0164
	$\widehat{h}_{R1}$	0.5917	0.0216	0.5276	0.0132	0.4934	0.0101
	$\widehat{h}_{R2}$	0.5924	0.0407	0.5332	0.0246	0.4996	0.0184
	$\widehat{h}_{R3}$	0.6820	0.0249	0.6071	0.0153	0.5681	0.0116

[a] Means and standard deviations (SD) are based on 2000 replications.

[b] Estimated bandwidths: $\widehat{h}_{\mathbf{opt};R1}$, $\widehat{h}_{\mathbf{opt};R2}$, and $\widehat{h}_{\mathbf{opt};R3}$ were calculated by setting $h_0 = \widehat{h}_{R1}N^{1/12}$, $h_0 = \widehat{h}_{R2}N^{1/12}$, and $h_0 = \widehat{h}_{R3}N^{1/12}$ in Section 3, respectively. Similarly, auxiliary bandwidths were set $h_* = \widehat{h}_{R1}$, $h_* = \widehat{h}_{R2}$, and $h_* = \widehat{h}_{R3}$ respectively.

Table 4: Monte Carlo results for Design 2: Parameter η

		$N = 200$			$N = 400$			$N = 600$		
		$Bias$	SD	MSE	$Bias$	SD	MSE	$Bias$	SD	MSE
$\rho = 0$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.8048	0.1655	0.6477	0.6640	0.0977	0.4409	0.5925	0.0729	0.3511
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7952	0.1638	0.6323	0.6621	0.0984	0.4383	0.5920	0.0737	0.3504
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7955	0.1643	0.6329	0.6622	0.0984	0.4385	0.5918	0.0737	0.3503
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7978	0.1601	0.6365	0.6634	0.0976	0.4401	0.5919	0.0736	0.3503
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8336	0.1814	0.6950	0.6667	0.0982	0.4445	0.5932	0.0734	0.3519
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8362	0.1852	0.6993	0.6682	0.0988	0.4465	0.5965	0.0744	0.3559
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.8098	0.1625	0.6558	0.6732	0.0991	0.4532	0.6100	0.0761	0.3721
$\rho = 1/4$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.7750	0.1804	0.6006	0.6410	0.1093	0.4109	0.5749	0.0779	0.3305
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7630	0.2567	0.5822	0.6363	0.1113	0.4049	0.5719	0.0778	0.3271
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7646	0.4242	0.5846	0.6365	0.1129	0.4052	0.5721	0.0779	0.3273
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7644	0.1704	0.5843	0.6375	0.1062	0.4064	0.5724	0.0775	0.3276
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8187	0.2796	0.6702	0.6470	0.1168	0.4186	0.5748	0.0787	0.3304
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8274	0.4854	0.6847	0.6486	0.1192	0.4207	0.5762	0.0795	0.3320
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.7811	0.1797	0.6101	0.6461	0.1082	0.4174	0.5844	0.0799	0.3416

[a] Simulated biases, standard deviations, and average Mean Squared Error (MSE) are based on 2000 replications.

Table 5: Monte Carlo results for Design 3: Bandwidth Estimation

		$N = 200$		$N = 400$		$N = 600$	
		$Mean$	SD	$Mean$	SD	$Mean$	SD
$\rho = 0$	$h_{\mathbf{opt}}$	0.6834	–	0.5747	–	0.5193	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7090	0.1552	0.5706	0.0672	0.5065	0.0395
	$\widehat{h}_{\mathbf{opt};R2}$	0.7070	0.1991	0.5694	0.0692	0.5061	0.0376
	$\widehat{h}_{\mathbf{opt};R3}$	0.6904	0.0694	0.5648	0.0311	0.5037	0.0200
	$\widehat{h}_{R1}$	0.5947	0.0213	0.5309	0.0133	0.4961	0.0100
	$\widehat{h}_{R2}$	0.6040	0.0444	0.5439	0.0264	0.5096	0.0197
	$\widehat{h}_{R3}$	0.6837	0.0245	0.6102	0.0153	0.5699	0.0115
$\rho = 1/4$	$h_{\mathbf{opt}}$	0.7101	–	0.5971	–	0.5396	–
	$\widehat{h}_{\mathbf{opt};R1}$	0.7169	0.4016	0.5797	0.1363	0.5138	0.0955
	$\widehat{h}_{\mathbf{opt};R2}$	0.7183	0.8889	0.5797	0.1708	0.5134	0.1141
	$\widehat{h}_{\mathbf{opt};R3}$	0.6937	0.1440	0.5699	0.0624	0.5085	0.0433
	$\widehat{h}_{R1}$	0.5919	0.0208	0.5273	0.0134	0.4931	0.0102
	$\widehat{h}_{R2}$	0.5910	0.0451	0.5305	0.0283	0.4972	0.0212
	$\widehat{h}_{R3}$	0.6816	0.0239	0.6068	0.0154	0.5675	0.0118

[a] Means and standard deviations (SD) are based on 2000 replications.

[b] Estimated bandwidths: $\widehat{h}_{\mathbf{opt};R1}$, $\widehat{h}_{\mathbf{opt};R2}$, and $\widehat{h}_{\mathbf{opt};R3}$ were calculated by setting $h_0 = \widehat{h}_{R1}N^{1/12}$, $h_0 = \widehat{h}_{R2}N^{1/12}$, and $h_0 = \widehat{h}_{R3}N^{1/12}$ in Section 3, respectively. Similarly, auxiliary bandwidths were set $h_* = \widehat{h}_{R1}$, $h_* = \widehat{h}_{R2}$, and $h_* = \widehat{h}_{R3}$ respectively.

Table 6: Monte Carlo results for Design 3: Parameter η

		$N = 200$			$N = 400$			$N = 600$		
		$Bias$	SD	MSE	$Bias$	SD	MSE	$Bias$	SD	MSE
$\rho = 0$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.7943	0.1692	0.6310	0.6539	0.1008	0.4276	0.5891	0.0756	0.3470
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7863	0.2013	0.6183	0.6515	0.1048	0.4244	0.5855	0.0779	0.3428
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7879	0.2506	0.6208	0.6515	0.1038	0.4244	0.5855	0.0771	0.3428
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7858	0.1639	0.6175	0.6510	0.1001	0.4238	0.5857	0.0759	0.3430
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8282	0.2195	0.6859	0.6600	0.1051	0.4357	0.5871	0.0764	0.3447
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8334	0.2643	0.6946	0.6602	0.1062	0.4359	0.5902	0.0771	0.3483
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.8020	0.1688	0.6433	0.6612	0.1014	0.4371	0.6027	0.0778	0.3632
$\rho = 1/4$	$\widetilde{\eta}\,(h_{\mathbf{opt}})$	0.7791	0.1957	0.6069	0.6487	0.1101	0.4208	0.5801	0.0818	0.3365
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R1})$	0.7721	0.7231	0.5961	0.6427	0.1535	0.4130	0.5741	0.1041	0.3296
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R2})$	0.7758	1.9588	0.6019	0.6431	0.2037	0.4136	0.5745	0.1269	0.3301
	$\widetilde{\eta}(\widehat{h}_{\mathbf{opt};R3})$	0.7718	0.2156	0.5957	0.6429	0.1079	0.4133	0.5738	0.0801	0.3293
	$\widetilde{\eta}(\widehat{h}_{R1})$	0.8268	0.8472	0.6835	0.6556	0.1488	0.4298	0.5787	0.0963	0.3349
	$\widetilde{\eta}(\widehat{h}_{R2})$	0.8383	0.8563	0.7027	0.6575	0.1869	0.4323	0.5810	0.1095	0.3375
	$\widetilde{\eta}(\widehat{h}_{R3})$	0.7903	0.2415	0.6246	0.6526	0.1115	0.4259	0.5878	0.0815	0.3455

[a] Simulated biases, standard deviations, and average Mean Squared Error (MSE) are based on 2000 replications.

Table 7: Descriptive Statistics

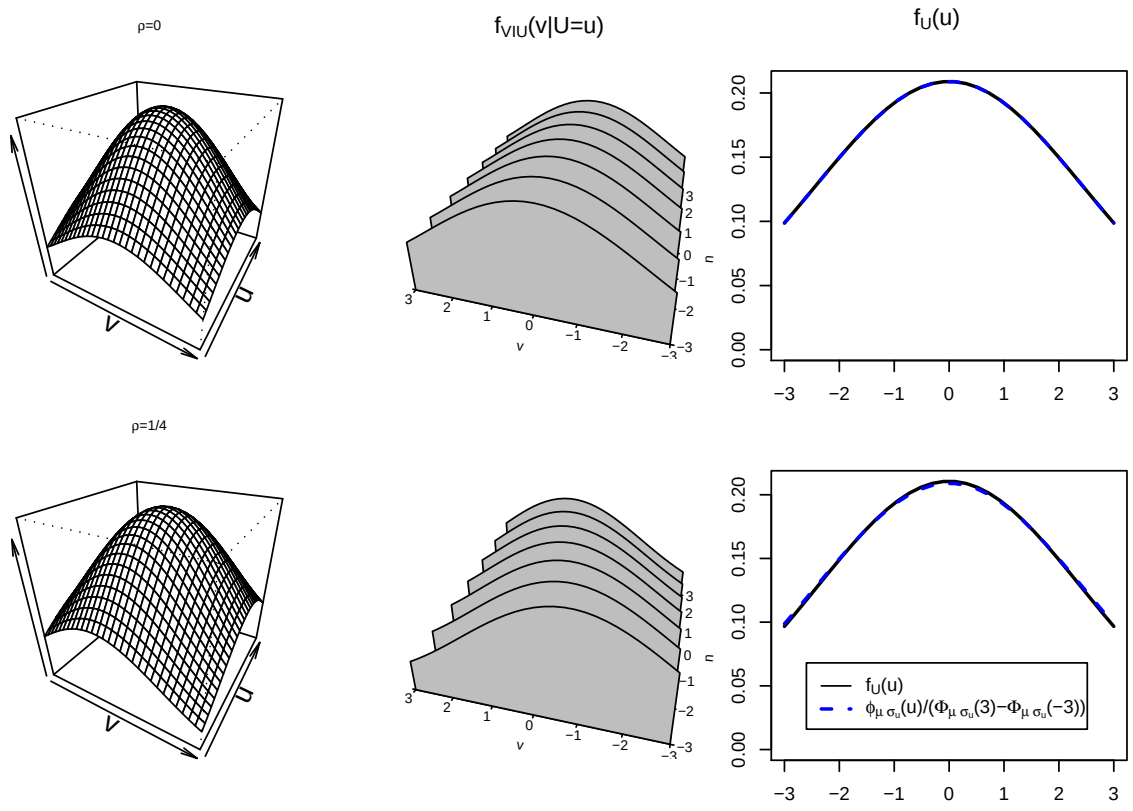
	Non-Participants (N = 1722)		Participants (N = 1725)		Total (N = 3447)	
Variable	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
CHILD	1.875	1.296	1.752	1.266	1.813	1.282
YCHILD	0.523	0.671	0.334	0.553	0.428	0.622
EDUC	9.107	3.844	10.627	4.662	9.867	4.340
LNHINC	0	0.587	0	0.614	0	0.601
AGE	3.416	1.003	3.639	0.871	3.528	0.946

[a] The original data set consisted of 9000 urban households. From these households we created a sample of 3447 women: 1) Aged 60 or below, married or living with partner; 2) They are not retired or in school; 3) Whose partners are present and report positive earnings in 2004.

Table 8: Estimation Results

	Probit		Heter. Probit		Kernel		Ordered	
Variable	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.	Coeff.	Std. Err.
CHILD	-0.041	(0.023)	-0.043	(0.052)	-0.053	(0.030)	0.187	(0.134)
YCHILD	-0.228	(0.049)	-0.548	(0.123)	-0.277	(0.063)	-0.632	(0.254)
EDUC	0.107	(0.006)	0.140	(0.010)	0.170	(0.008)	0.364	(0.041)
LNHINC	-1	–	-1	–	-1	–	-1	–
AGE	1.767	(0.207)	2.409	(0.306)	2.002	(0.270)	2.548	(0.967)
AGE2	-0.215	(0.028)	-0.297	(0.040)	-0.234	(0.037)	-0.288	(0.130)
Constant	-4.250	(0.352)	-5.679	(0.535)	-5.449	(0.445)	-8.334	(1.828)
CHILD			0.385	(0.058)				
YCHILD			0.162	(0.127)				
N	3447		3447		3402		3447	
-Log L	2399.016		2268.178					

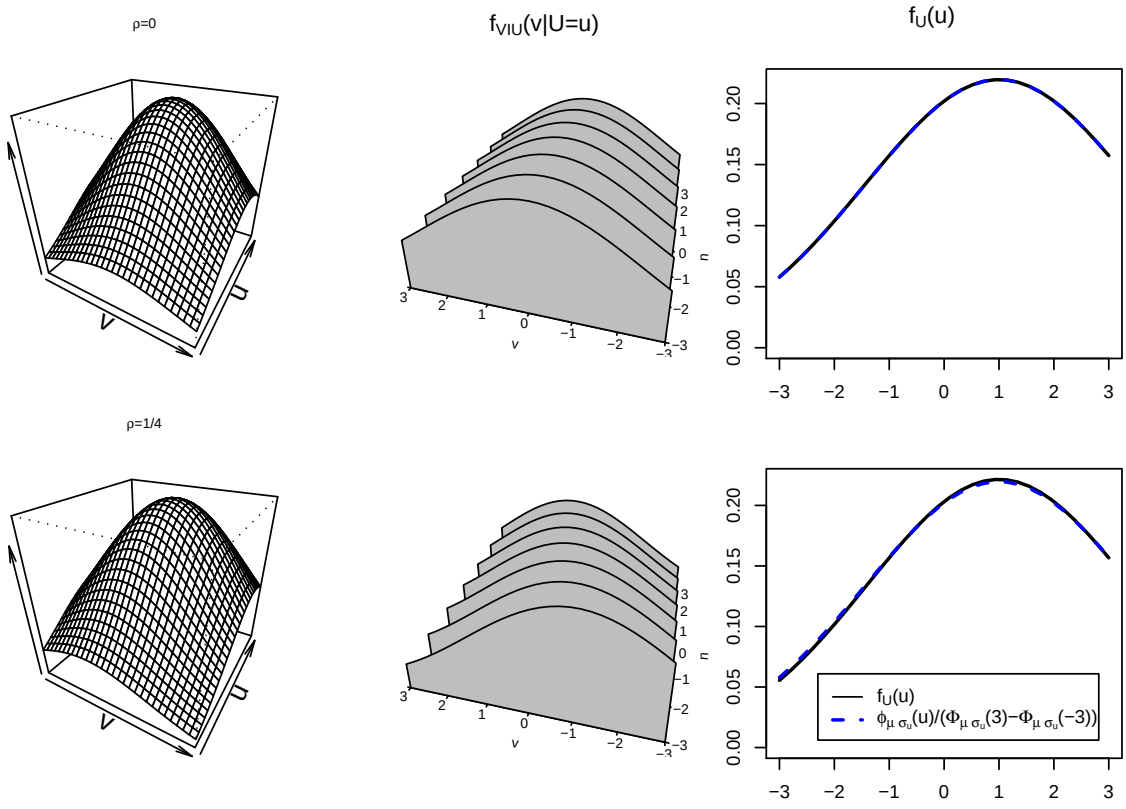
Figure 1: Visualization of Design 1



[a] Each row represents a variation of Design 1: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.

[b] First column from the left shows their joint densities, $f_{VU}(v, u)$. Middle column shows their associated conditional densities, $f_{V|U}(v|\, U = u)$, and last column shows their marginal distribution, $f_U(u)$, with respect to U, as well as that of a univariate truncated, $[-3, 3]$, normal with parameters: $\mu_u = 0$, and $\sigma_u^2 = 6$.

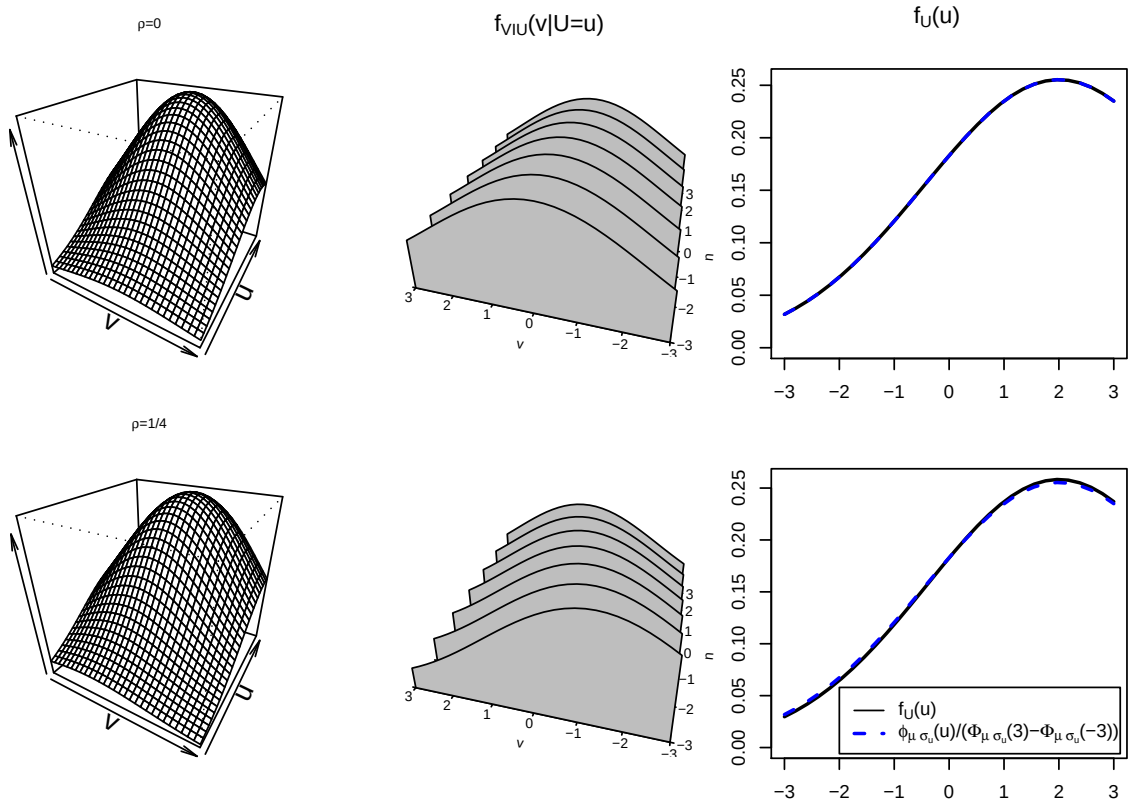
Figure 2: Visualization of Design 2



[a] Each row represents a variation of Design 2: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.

[b] First column from the left shows their joint densities, $f_{VU}(v, u)$. Middle column shows their associated conditional densities, $f_{V|U}(v|\, U = u)$, and last column shows their marginal distribution, $f_U(u)$, with respect to U, as well as that of a univariate truncated, $[-3, 3]$, normal with parameters: $\mu_u = 1$, and $\sigma_u^2 = 6$.

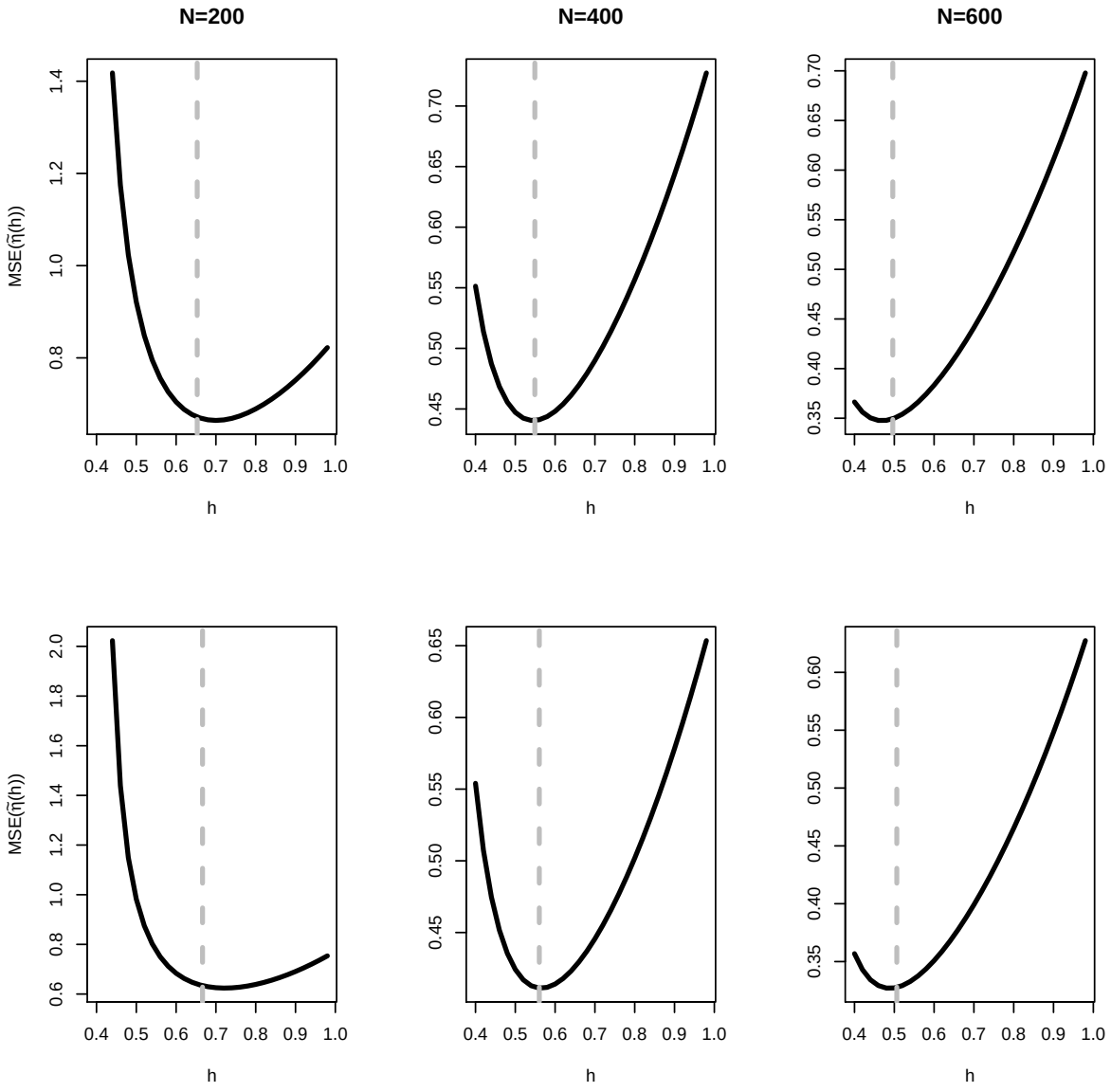
Figure 3: Visualization of Design 3



[a] Each row represents a variation of Design 3: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.

[b] First column from the left shows their joint densities, $f_{VU}(v, u)$. Middle column shows their associated conditional densities, $f_{V|U}(v|\, U = u)$, and last column shows their marginal distribution, $f_U(u)$, with respect to U, as well as that of a univariate truncated, $[-3, 3]$, normal with parameters: $\mu_u = 2$, and $\sigma_u^2 = 6$.

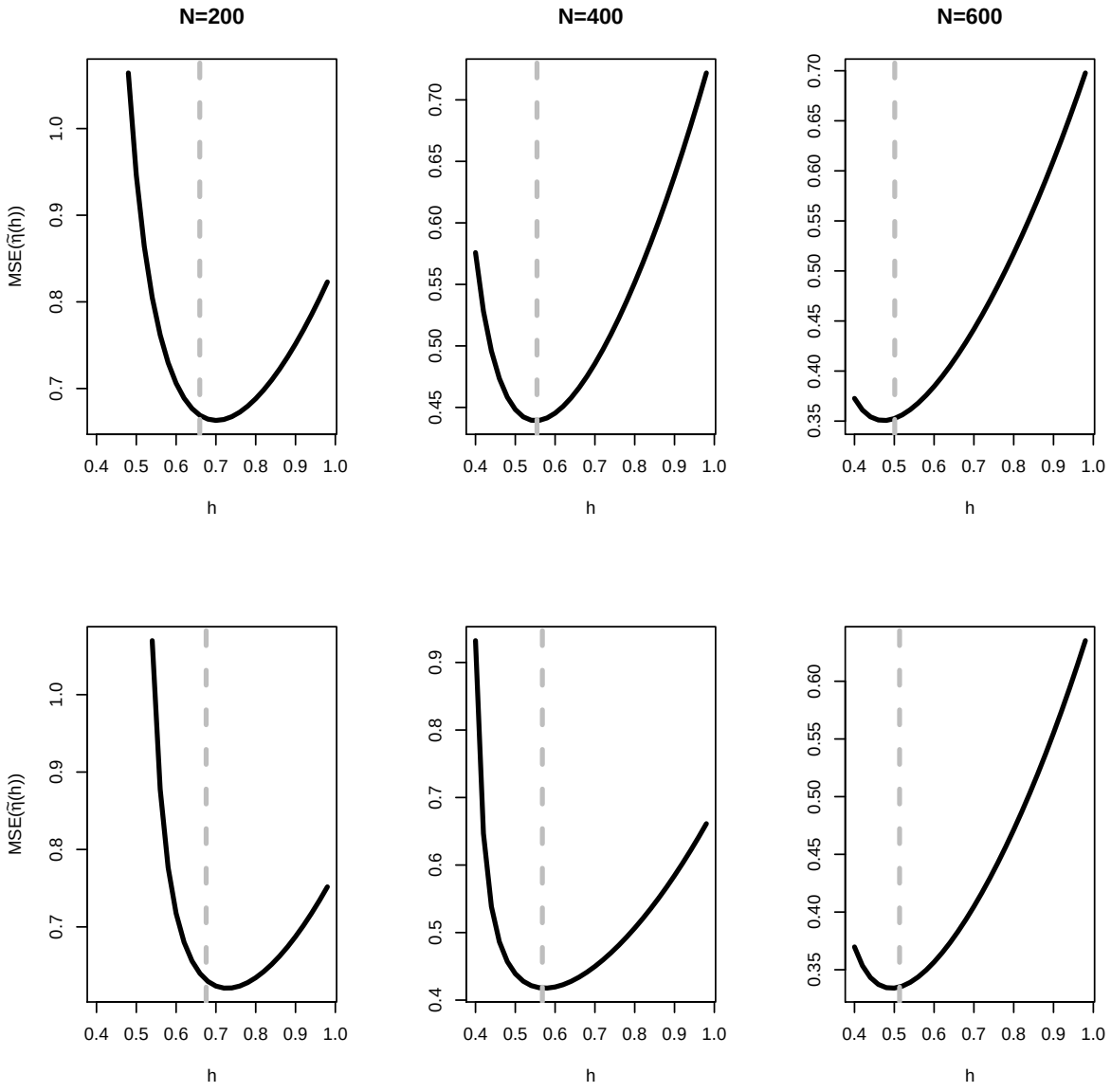
Figure 4: Simulated MSE of Design 1



[a] Each row represents a variation of Design 1: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.

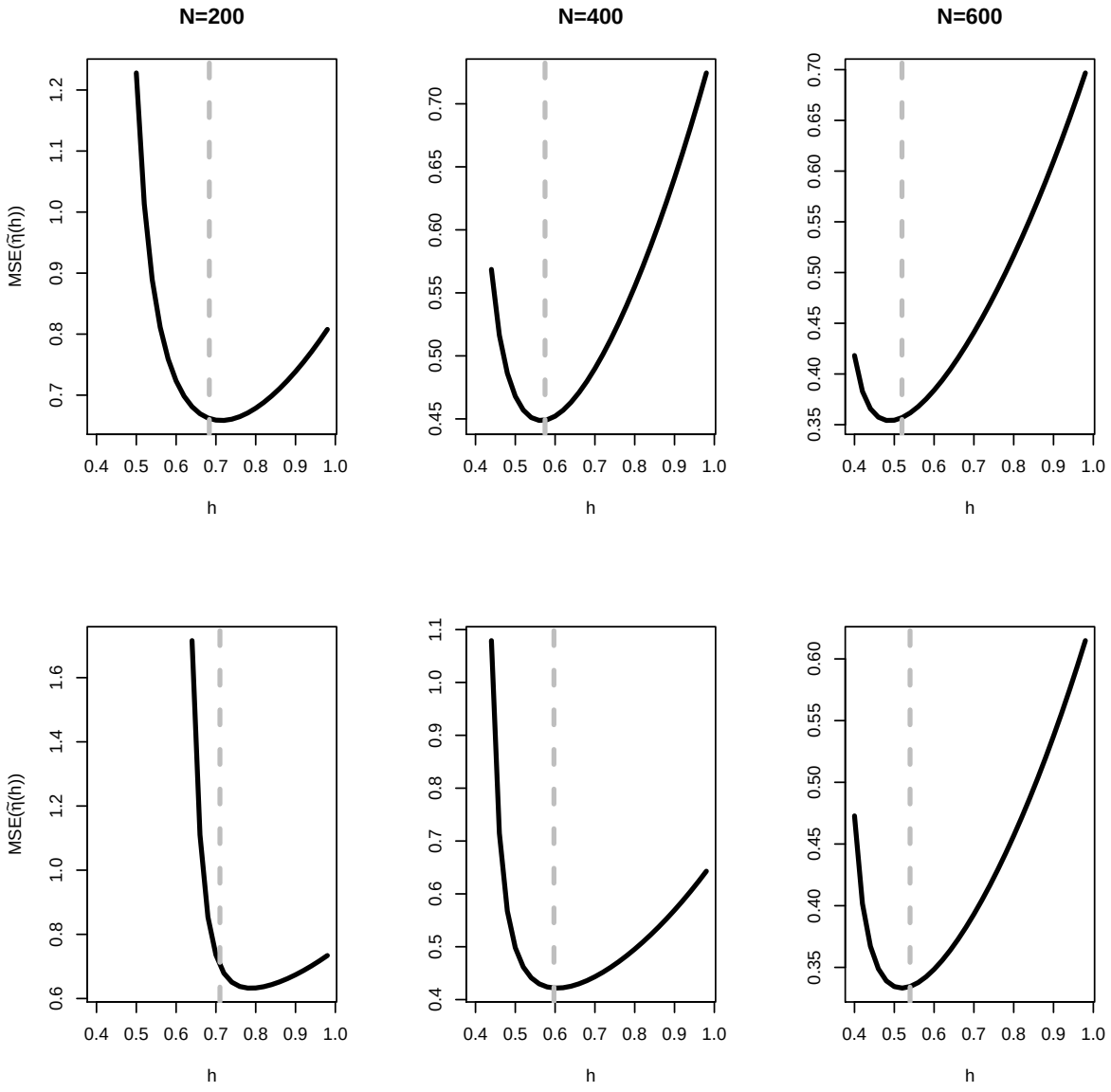
[b] Simulation based on 1000 replications. Dashed gray lines represent the optimal bandwidth predicted by our results in each case.

Figure 5: Simulated MSE of Design 2



[a] Each row represents a variation of Design 2: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.
[b] Simulation based on 1000 replications. Dashed gray lines represent the optimal bandwidth predicted by our results in each case.

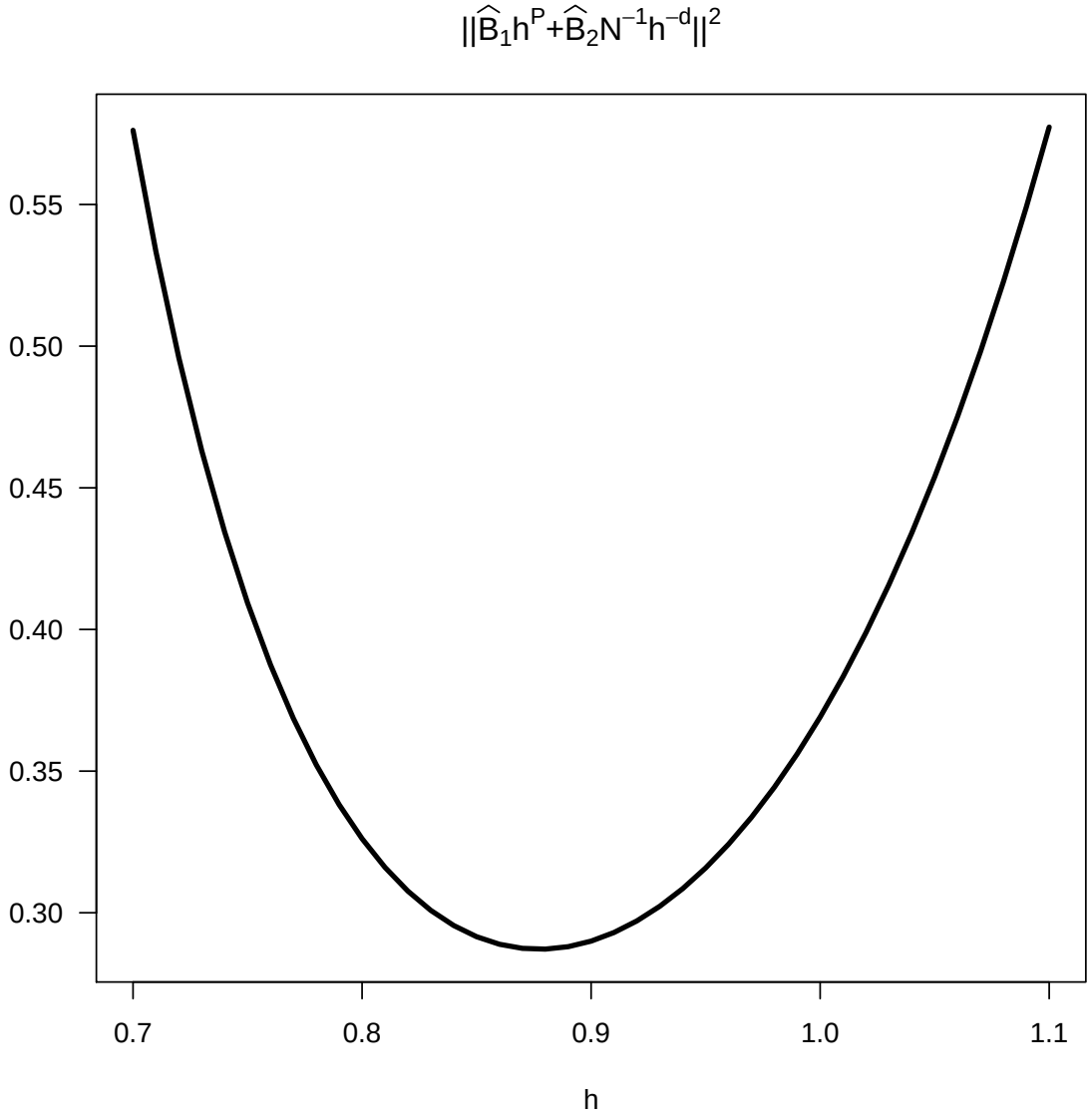
Figure 6: Simulated MSE of Design 3



[a] Each row represents a variation of Design 3: (a) $\rho = 0$, and (b) $\rho = 1/4$ in descending order.

[b] Simulation based on 1000 replications. Dashed gray lines represent the optimal bandwidth predicted by our results in each case.

Figure 7: Estimated Contribution to MSE



[a] Discrete regressors: CHILD, YCHILD and EDUC. Continuous regressors: LNHINC and AGE.

[b] Bandwidth $h_* \simeq 0.99$ was found by standard cross validation of the kernel estimator of f_{VX}. Similarly, $h_0 = 0.7$ and $\Delta = 2$.