Centre interuniversitaire de recherche
en économie quantitative

CIREQ

Cahier 12-2007

ROBUST AVERAGE DERIVATIVE ESTIMATION

Marcia M.A. SCHAFGANS and
Victoria ZINDE-WALSH

CIREQ

Le **Centre interuniversitaire de recherche en économie quantitative (CIREQ)** regroupe des chercheurs dans les domaines de l'économétrie, la théorie de la décision, la macroéconomie et les marchés financiers, la microéconomie appliquée et l'économie expérimentale ainsi que l'économie de l'environnement et des ressources naturelles. Ils proviennent principalement des universités de Montréal, McGill et Concordia. Le CIREQ offre un milieu dynamique de recherche en économie quantitative grâce au grand nombre d'activités qu'il organise (séminaires, ateliers, colloques) et de collaborateurs qu'il reçoit chaque année.

*The **Center for Interuniversity Research in Quantitative Economics (CIREQ)** regroups researchers in the fields of econometrics, decision theory, macroeconomics and financial markets, applied microeconomics and experimental economics, and environmental and natural resources economics. They come mainly from the Université de Montréal, McGill University and Concordia University. CIREQ offers a dynamic environment of research in quantitative economics thanks to the large number of activities that it organizes (seminars, workshops, conferences) and to the visitors it receives every year.*

Cahier 12-2007

ROBUST AVERAGE DERIVATIVE ESTIMATION

Marcia M.A. SCHAFGANS and Victoria ZINDE-WALSH

CIREQ, Université de Montréal
C.P. 6128, succursale Centre-ville
Montréal (Québec) H3C 3J7
Canada

téléphone : (514) 343-6557
télécopieur : (514) 343-5831
cireq@umontreal.ca
http://www.cireq.umontreal.ca

Université de Montréal McGill Concordia UNIVERSITY

Dépôt légal - Bibliothèque nationale du Canada, 2007, ISSN 0821-4441
Dépôt légal - Bibliothèque et Archives nationales du Québec, 2007
ISBN-13 : 978-2-89382-553-3

Robust Average Derivative Estimation

Marcia M.A. Schafgans* Victoria Zinde-Walsh†

July 2007

Abstract. Many important models, such as index models widely used in limited dependent variables, partial linear models and nonparametric demand studies utilize estimation of average derivatives (sometimes weighted) of the conditional mean function. Asymptotic results in the literature focus on situations where the ADE converges at parametric rates (as a result of averaging); this requires making stringent assumptions on smoothness of the underlying density; in practice such assumptions may be violated. We extend the existing theory by relaxing smoothness assumptions and obtain a full range of asymptotic results with both parametric and non-parametric rates. We consider both the possibility of lack of smoothness and lack of precise knowledge of degree of smoothness and propose an estimation strategy that produces the best possible rate without a priori knowledge of degree of density smoothness. The new combined estimator is a linear combination of estimators corresponding to different bandwidth/kernel choices that minimizes the estimated asymptotic mean squared error (AMSE). Estimation of the AMSE, selection of the set of bandwidths and kernels are discussed. Monte Carlo results for density weighted ADE confirm good performance of the combined estimator.

Keywords: Nonparametric estimation, density weighted average derivative estimator, combined estimator.

JEL Classification: C14.

*Department of Economics, London School of Economics. Mailing address: Houghton Street, London WC2A 2AE, United Kingdom.

†Department of Economics, McGill University and CIREQ. This work was supported by the Social Sciences and Humanities Research Council of Canada (SSHRC) and by the *Fonds québecois de la recherche sur la société et la culture* (FRQSC) .

1. Introduction

Many important models, such as index models widely used in limited dependent variables, partial linear models and nonparametric demand studies utilize estimation of average derivatives (sometimes weighted) of the conditional mean function. Härdle, Hildenbrand, Jerison (1991) and Blundell, Duncan, and Pendakur (1998), amongst others, advocated the derivative based approach in the analysis of consumer demand, where nonparametric estimation of Engel curves has become common place (e.g., Yatchew, 2003). Powell, Stock and Stoker (1989) popularized the use of average derivatives of the conditional mean (or conditional mean weighted by some function) in the semiparametric estimation of index models by pointing out that the average derivatives in single index models identify the parameters "up to scale".

A large literature is devoted to the asymptotic properties of nonparametric estimators of average derivatives and to their use in estimation of index models and testing of coefficients. Asymptotic properties of average density weighted derivatives are discussed in Powell, Stock and Stoker (1989) and Robinson (1989); Härdle and Stoker (1989) investigated the properties of the average derivatives themselves; Newey and Stoker (1993) addressed the choice of weighting function. Horowitz and Härdle (1996) extended the ADE approach in the estimation of coefficients in the single index model to the presence of discrete covariates; Donkers and Schafgans (2005) extended the ADE approach to multiple index models; Chaudhury et al. (1998) investigated the average derivatives in quantile regression; Li et al. (2003) investigated the local polynomial fitting to average derivatives and Banerjee (2007) provided a recent discussion on estimating the average derivatives using a fast algorithm. Higher order expansions and the properties of bootstrap tests of ADE for hypotheses are investigated in Nichiyama and Robinson (2000, 2005).

In all of the literature on ADE estimation asymptotic theory was provided for parametric rates of convergence. Even though the estimators are based on a nonparametric kernel estimator of the conditional mean which depends on the kernel and bandwidth and converges at a nonparametric rate, averaging can produce a parametric convergence rate

thus reducing dependence on selection of the kernel and bandwidth which do not appear in the leading term of the AMSE expansion. This parametric rate of convergence (and thus the results in this literature), however, relies on the assumption of sufficiently high degree of smoothness of the underlying density of the regressors. This assumption is not based on any a priori theoretical considerations nor is it supported by empirical verification. For example many multimodal distributions, even if they are sufficiently smooth, possess derivatives that are large enough to cause problems (see discussion in Marron and Wand, 1992, for examples of normal mixtures that exhibit features usually thought of as characteristic of non-smooth densities.). Various examples of multimodal distributions are encountered in biomedical and statistical studies, e.g., Izenman and Sommer (1988).

Our concern with the assumed high degree of density smoothness led us to extend the existing asymptotic results by relaxing assumptions on the density. We show that insufficient smoothness will result in possible asymptotic bias and may easily lead to non-parametric rates. The selection of optimal kernel order and optimal bandwidth (Powell and Stoker, 1996) in the absence of sufficient smoothness moreover presumes the knowledge of the degree of density smoothness. Thus an additional concern for us is the possible uncertainty about the degree of density smoothness. To address problems associated with an incorrect choice of a bandwidth/kernel pair we construct an estimator that optimally combines estimators for different bandwidths and kernels to protect against the negative consequences of errors in assumptions about the order of density smoothness.

We construct a linear combination of density weighted average derivative estimators for different bandwidth/kernel choices, with weights chosen to minimize the estimated asymptotic MSE; the resulting estimator we call the combined estimator. Kotlyarova and Zinde-Walsh (2006) have shown that the weights in this combination provide the best rate available among all the rates without a priori knowledge of degree of smoothness, thus protecting against making a bandwidth/kernel choice that relies on incorrect smoothness assumptions and would yield high asymptotic bias. Without prior knowledge of smoothness there is no guidance for the choice of the bandwidth or kernel. In this circumstance, the combined estimator provides a robust alternative to specifying a particular bandwidth/kernel pair.

Our approach is supported by Hansen (2005). In discussing the choice of kernel for density estimation, Hansen showed the order of the kernel to have a large impact on its finite-sample MISE. With the ideal kernel order depending on the unknown smoothness of the density Hansen proposed a criterion of minimax regret. Whereas Hansen (2005) observed little difference in performance among symmetric kernels of the same order, within our combined estimator, we argue that gains are achievable.

Using a Monte Carlo experiment for the Tobit model, for a variety of distributions for the explanatory variables (gaussian, tri-modal gaussian mixture and the "double claw" and "discrete comb" mixtures from Marron and Wand, 1992), we demonstrate that there is no clear guidance on the choice of suitable kernel bandwidth pair. Even in these cases, where the smoothness assumptions hold, the high modal nature of these mixture distributions exhibit large partial derivatives that undermine the performance of ADE. At the same time, the combined estimator delivers robust reliable results in all cases.

The paper is organized as follows. In section 2 we discuss the general set-up and assumptions. In section 3 we derive the asymptotic properties of the density-weighted ADE under various assumptions about density smoothness, derive the joint asymptotic distribution for ADE estimators based on different bandwidth kernels pairs, and develop the combined estimator. Section 4 provides the Monte Carlo study results and Section 5 concludes.

2. General set-up and assumptions

The unknown conditional mean function can be represented as

$$g(x) = E(y|x) = \int y \frac{f^*(x,y)}{f(x)} dy = \frac{G(x)}{f(x)},$$

with dependent variable $y \in R$ and explanatory variables $x \in R^k$. The joint density of (y, x) is denoted by $f^*(y, x)$, the marginal density of x is denoted by $f(x)$ and $G(x)$ denotes the function $\int y f^*(y, x) dy$.

Since the regression derivative, $g'(x)$, can be expressed as

$$g'(x) = \frac{G'(x)}{f(x)} - g(x)\frac{f'(x)}{f(x)},$$

the need to avoid imprecise contributions to the average derivative for observations with low densities emanates from the presence of the density in the denominator. One way of doing this is to employ some weighting function, $w(x)$; for example, the density weighted average derivative estimator of Powell,Stock and Stoker (1989), hereafter referred to as PSS, utilizes $w(x) = f(x)$. In Härdle and Stoker (1989) trimming on the basis of the density takes the place of the weighting function, that is they consider $w_N(x) = 1(f(x) > b_N)$ where $b_N \to 0$. On the other hand, Fan (1992, 1993), Fan and Gijbels (1992) avoid weighting by use of regularization whereby n^{-2} is added to the denominator of the estimator. In this paper we focus on the PSS estimator.

We now turn to the fundamental assumptions. The first two assumptions are common in this literature, restricting x to be a continuously distributed random variable, where no component of x is functionally determined by other components of x, imposing a boundary condition allowing for unbounded x's and requiring differentiability of f and g.

Assumption 1. *The underlying measure of (y, x) can be written as $v_y \times v_x$, where v_x is Lebesque measure. The support Ω of f is a convex (possibly unbounded) subset of R^k with nonempty interior Ω_0.*

Assumption 2. *The density function $f(x)$ is continuous in the components of x for all $x \in R^k$, so that $f(x) = 0$ for all $x \in \partial\Omega$, where $\partial\Omega$ denotes the boundary of Ω. f is continuously differentiable in the components of x for all $x \in \Omega_0$ and g is continuously differentiable in the components of all $x \in \bar{\Omega}$, where $\bar{\Omega}$ differs from Ω_0 by a set of measure 0.*

Additional requirements involving the conditional distribution of y given x as well as more smoothness conditions need to be added. The conditions are slightly amended from how they appear in the literature, in particular we use the weaker Hölder conditions instead of Lipschitz conditions in the spirit of weakening smoothness assumptions as much as possible.

Assumption 3. *(a)* $E(y^2|x)$ *is continuous in* x

(b) The components of the random vector $g'(x)$ *and matrix* $f'(x)[y, x']$ *have finite second moments;* $(fg)'$ *satisfies a Hölder condition with* $0 < \alpha \leq 1$:

$$\left|(fg)'(x + \Delta x) - (fg)'(x)\right| \leq \omega_{(fg)'}(x) \left\|\Delta x\right\|^{\alpha}$$

and $E(\omega^2_{(fg)'}(x)[1 + |y| + \|x\|]) < \infty$.

For kernel estimators considered in the literature both the choice of the kernel (its order) and the selection of bandwidth have played a crucial role in ensuring that the asymptotic bias for the nonparametric estimates of the derivative based functionals (averages) vanishes sufficiently fast subject to a high degree of density smoothness. The kernel smoothing function is assumed to satisfy a fairly standard assumption here, except for the fact that we allow for the kernel to be non-symmetric.

Assumption 4. *(a) The kernel smoothing function* $K(u)$ *is a continuously differentiable function with bounded support* $[-1, 1]^k$.

(b) The kernel function $K(u)$ *obeys*

$$\begin{aligned}
&\int K(u)du = 1, \\
&\int u_1^{i_1} ... u_k^{i_k} K(u)du = 0 \qquad i_1 + ... + i_k < v(K) \\
&\int u_1^{i_1} ... u_k^{i_k} K(u)du \neq 0 \qquad i_1 + ... + i_k = v(K)
\end{aligned}$$

where $(i_1, .., i_k)$ *is an index set,*

(c) The kernel smoothing function $K(u)$ *is differentiable up to the order* $v(K)$.

Various further assumptions have been made in the literature concerning the smoothness of the density (higher degree of differentiability, Lipschitz and boundedness conditions) to ensure parametric rates of convergence. We formalize the degree of density smoothness in terms of the Hölder space of functions. This space for integer $m \geq 0$ and $0 < \alpha \leq 1$ is defined as follows. For a set $E \subseteq R^k$ the space $C_{m+\alpha}(E)$ is a Banach space of bounded and

continuous functions which are m times continuous differentiable with all the m^{th} order derivatives satisfying Hölder's condition of order α (see Mathematicheskaya Encyclopedia. English., ed. M. Hazewinkel) for every $x, x + \Delta x \in E$:

$$\left|f^{(m)}(x + \Delta x) - f^{(m)}(x)\right| \leq \omega_{f^{(m)}}(x) \left\|\Delta x\right\|^{\alpha}.$$

Assumption 5. *$f \in C_{m+\alpha}(\Omega)$ where $C_{m+\alpha}(\Omega)$ is the Hölder space of functions on $\Omega \subset R^k$ with $m \geq 1, 0 < \alpha \leq 1$ and $E(\omega^2_{f^{(m)}}(x)[1 + |y|^2 + \|x\|]) < \infty$.*

The assumption implies that each component of the derivative of density $f'(x) \in C_{m-1+\alpha}(\Omega)$ thus for every component of $f'(x)$ continuous derivatives of order $m-1$ exist (if $m-1=0$ there is just Hölder continuity of derivative). This permits the following expansion for $c = 0, 1$ with $c = 0$ for the expansion of density and $c = 1$ for the expansion of the derivative of the density function:

$$\begin{aligned} & f^{(c)}(x + \Delta x) \qquad (1) \\ = & \sum_{p=c}^{m-1} \left\{ \sum_{i_1+...i_k=p-c} \tfrac{1}{i_1!...i_k!} f^{(p)}(x) \Delta x^{\iota} + \sum_{i_1+...i_k=m-c} \tfrac{1}{i_1!...i_k!} f^{(m)}(x + \zeta \Delta x) \Delta x^{\iota} \right\} \\ = & \sum_{p=c}^{m} \left\{ \sum_{i_1+...i_k=p-c} \tfrac{1}{i_1!...i_k!} f^{(p)}(x) \Delta x^{\iota} + \sum_{i_1+...i_k=m-c} \tfrac{1}{i_1!...i_k!} \left[f^{(m)}(x + \zeta \Delta x) - f^{(m)}(x)\right] \Delta x^{\iota} \right\}, \end{aligned}$$

where Δx denotes the vector $(\Delta x_1, .., \Delta x_k)$, Δx^{ι} the product $\Delta x_1^{i_1} \cdots \Delta x_k^{i_k}$ with ι the index set $(i_1, \ldots, i_k)$, and $f^{(m)}(x)$ the derivative $\partial^{m-c} f^{(c)} / (\partial x)^{\iota}$, also $\zeta : 0 \leq \zeta \leq 1$. The first equality is obtained by Taylor expansion (with the remainder term in Lagrange form) and the second equality is obtained by adding and subtracting the terms with $f^{(m)}(x)$. By Assumption (5) the $\left[f^{(m)}(x + \zeta \Delta x) - f^{(m)}(x)\right]$ in the last sum satisfies the Hölder inequality and thus the last sum is $O(\|\Delta x\|^{m-c+\alpha})$.

Lack of smoothness of the density can readily be shown to affect the asymptotic bias of derivative based estimators since the biases of those estimators can be expressed via the bias of the kernel estimator of the derivative of density. Let v be the degree of smoothness of the derivative of the density (equal to $m - 1 + \alpha$ by Assumption (5)), and $v(K)$, the order of the kernel. Define $\bar{v} = \min(v, v(K))$. Provided $\bar{v} = v(K) \leq v$, the bias of the

derivative of the density, $E(\hat{f}'_{(K,h)}(x_i) - f'(x_i)) = E\left(\int K(u)(f'(x_i - uh) - f'(x_i))du\right)$, is as usual $O(h^{v(K)})$ (by applying the usual $\bar{v}^{th}$ order Taylor expansion of $f'(x_i - uh)$ around $f'(x_i)$). We next show that with $\bar{v} = v < v(K)$, the bias of the derivative vanishes at the lower rate $O(h^v)$. In the latter case substituting (1), with $c = 1$, $\Delta x = -hu$, into the bias expression and using kernel order, yields[1]

$$\begin{aligned} & E\left(\int \left[f'(x - hu) - f'(x)\right] K(u)du\right) \qquad (2) \\ = \; & E\left(\int \sum_{\iota_1 + ...i_k = m-1} \tfrac{1}{i_1!...i_k!} h^{m-1} \cdot (-1)^{m-1} \left[f^{(m)}(x_i - \widetilde{hu}) - f^{(m)}(x_i)\right] K(u)\, u^{\iota} du\right) \\ = \; & O(h^{m-1+\alpha}) \equiv O(h^v), \end{aligned}$$

where the latter equality uses the Hölder inequality. If differentiability conditions typically assumed do not hold, then even for bandwidths such that $Nh^{2v(K)} = o(1)$ the bias does not vanish sufficiently fast. With $\bar{v} = \min(v, v(K))$ all we can state is the rate $O(h^{\bar{v}})$ for the bias:

$$E\left(\int \left[f'(x - hu) - f'(x)\right] K(u)du\right) = O(h^{\bar{v}}).$$

3. AVERAGE DENSITY WEIGHTED DERIVATIVE ESTIMATOR

The average density weighted derivative, introduced in Powell, Stock and Stoker (1989), PSS, is defined as

$$\delta_0 = E(f(x)g'(x)). \qquad (3)$$

Given Assumptions 1-3, via integration by parts, (3) can be represented (see Lemma 2.1 in PSS) as:

$$\delta_0 = -2E(f'(x)y). \qquad (4)$$

$z_i = (y_i, x_i^T)^T, i = 1, ...N$ is a random sample from the distribution of $z = (y, x^T)^T$.

[1] $\left|\int \left[f'(x - hu) - f'(x)\right] K(u)du\right| \leq h^{m-1+\alpha} \omega_{f^{(m)}}(x) \int \|K(u)\| \cdot \|u\|\, du \cdot O(1)$, where Assumption 4(a) implies that $\|K(u)\|$ is bounded (since it is continuous on a closed bounded set), and $\|u\|$ is bounded on the support of K, Assumption 5 ensures boundedness of $E\left\|w_{f^{(m)}}(x)\right\|$.

The estimator of δ_0 proposed by PSS uses the sample analogue of (4), where $f'(x)$ is replaced by a consistent nonparametric estimate, i.e.,

$$\hat{\delta}_N(K,h) = \frac{-2}{N}\sum_{i=1}^{N} \hat{f}'_{(K,h)}(x_i)y_i, \tag{5}$$

with

$$\hat{f}'_{(K,h)}(x_i) = \tfrac{1}{N-1}\sum_{j\neq i}^{N}\left(\frac{1}{h}\right)^{k+1} K'(\frac{x_i - x_j}{h}).$$

K is the kernel smoothing function and h is a smoothing parameter that depends on the sample size N, with $h \to 0$ as $N \to \infty$.

The variance of $\hat{\delta}_N(K,h)$ is given by

$$Var(\hat{\delta}_N(K,h)) = \Sigma_{1\delta}(K)N^{-2}h^{-(k+2)} + \Sigma_{2\delta}N^{-1} + O(N^{-2}) \tag{6}$$

where

$$\begin{aligned}
\Sigma_{1\delta}(K) &= 4E\left[y^2 f(x_i)\mu_2(K) + \mu_2^*(K)(gf)(x_i)y_i\right];\\
\text{with } \mu_2(K) &= \int K'(u)K'(u)^T du;\\
\mu_2^*(K) &= \int K'(u)K'(-u)^T du, \text{ (under symmetry } \mu_2^*(K) = -\mu_2(K));\\
\text{and } \Sigma_{2\delta} &= 4\left\{E([(g'f)(x_i) - (y_i - g(x_i))f'(x_i)]\,[(g'f)(x_i) - (y_i - g(x_i))f'(x_i)]^T)\right\} - 4\delta_0\delta_0^T.
\end{aligned}$$

Note that $\Sigma_{2\delta}$ for sufficiently smooth $f(x)$ coincides with the asymptotic variance of $\sqrt{N}\hat{\delta}_N(K,h)$ considered in PSS, when $Nh^{k+2} \to \infty$. For a symmetric kernel, $\Sigma_{1\delta}(K)$ simplifies to $4\mu_2(K)E\left[\sigma^2(x_i)f(x_i)\right]$, with the conditional variance $\sigma^2(x) = E(y^2|x) - E(y|x)^2$. For this case, Powell and Stoker (1996) discuss the rates of the asymptotic variance in (6) with a view to selecting the optimal for MSE bandwidth rate.

The asymptotic variance does not depend on the kernel function when the bandwidth satisfies $Nh^{k+2} \to \infty$, but only if we have a certain degree of smoothness of the density: $v > (k+2)/2$. In the absence of this degree of differentiability the asymptotic variance (as the asymptotic bias) does depend on the weighting used in the local averaging possibly yielding a non-parametric rate. To express the asymptotic bias of the estimator $\hat{\delta}_N(K,h)$ define

$$A(K,h,x_i) = E_{z_i}\left[\hat{f}'_{(K,h)}(x_i) - f'(x_i)\right] = \int K(u)(f'(x_i - uh) - f'(x_i))du.$$

Then

$$E(\hat{\delta}_N(K,h) - \delta_0) = -2E(A(K,h,x_i)y_i). \tag{7}$$

As shown in Section 2, $EA(K,h,x_i)$ is $O(h^{\bar{v}})$. We assume

Assumption 6. *As $N \to \infty$, $-2h^{-\bar{v}}E(A(K,h,x_i)y_i) \to \mathcal{B}(K)$, where $|\mathcal{B}(K)| < \infty$ holds.*

Then

$$E(\hat{\delta}_N(K,h) - \delta_0) = h^{\bar{v}}\mathcal{B}(K) + o(h^{\bar{v}})$$

The asymptotic bias of the estimator $r_N\left(\hat{\delta}_N(K,h) - \delta_0\right)$ for some rate r_N can then be written as

$$Bias(r_N\left(\hat{\delta}_N(K,h) - \delta_0\right)) = r_N h^{\bar{v}}\mathcal{B}(K). \tag{8}$$

and vanishes if $r_N h^{\bar{v}} \to 0$. We note that Assumption 6 could hold as a results of primitive moment assumptions on y_i, $f(x_i)$, and $g(x_i)$.

The following theorem describes the statistical properties of the PSS estimator in the cases: (a) of sufficient smoothness: $\bar{v} > \frac{k+1}{2}$, where the only new contribution relative to PSS is in considering the possibility of sub-optimal bandwidth choice (subcases (a)i.,ii.) and providing expressions for the moments for possibly non-symmetric kernels; (b) of marginally enough smoothness; (c) of smoothness (or kernel order) insufficient to allow for a parametric rate: $\bar{v} < \frac{k+1}{2}$. The latter cases were not examined in the literature where the stringent smoothess assumptions made the appropriate optimal choice of kernel order and bandwidth rate straightforward.

Theorem 1. *Under Assumptions 1–6*

(a) *If the density is sufficiently smooth and order of kernel sufficiently high: $\bar{v} > \frac{k+2}{2}$*

i. *for h : $Nh^{k+2} = o(1)$, $N^2h^{k+2} \to \infty$ in the limit there is an asymptotically unbiased but not efficient estimator*

$$\begin{aligned} E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)) &\to 0; \\ E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right))(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)^T &\to \Sigma_{1\delta}(K); \end{aligned}$$

ii. *for* $h : Nh^{k+2} \to C, 0 < C < \infty$ (*so* $Nh^{2\bar{v}} = o(1)$) *similarly to i.*,

$$\begin{aligned} E(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)) &\to 0; \\ E(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right))(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)^T &\to C\Sigma_{1\delta}(K) + \Sigma_{2\delta}; \end{aligned}$$

iii. *for* $h : Nh^{k+2} \to \infty$, $Nh^{2\bar{v}} = o(1)$

$$\begin{aligned} E(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right) &\to 0; \\ E(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right))(N^{\frac{1}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)^T &\to \Sigma_{2\delta}; \end{aligned}$$

and the asymptotic normality result in PSS, Theorem 3.3 holds:

$$\sqrt{N}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(0, \Sigma_{2\delta});$$

iv. *for* $h : Nh^{k+2} \to \infty$, *but* $Nh^{2\bar{v}} \to C_1^2, 0 < C_1 < \infty$, *in the limit there is a biased asymptotically normal estimator:*

$$\sqrt{N}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(C_1\mathcal{B}(K), \Sigma_{2\delta});$$

v. *for* $h : Nh^{2\bar{v}} \to \infty$; *the bias dominates*

$$h^{-\bar{v}}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{p} \mathcal{B}(K).$$

(b) *For the case* $\bar{v} = \frac{k+2}{2}$ *i., ii. and v. of part (a) apply.*

(c) *If either the density is not smooth enough or the order of the kernel is low:* $\bar{v} < \frac{k+2}{2}$ *the parametric rate cannot be obtained:*

i. *for* $h : N^2h^{k+2+2\bar{v}} = o(1)$, $N^2h^{k+2} \to \infty$ *in the limit there is an asymptotically unbiased but not efficient estimator:*

$$\begin{aligned} E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)) &\to 0; \\ E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right))(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)^T &\to \Sigma_{1\delta}(K); \end{aligned}$$

ii. *for* $h : N^2h^{k+2+2\bar{v}} \to C_2^2, 0 < C_2 < \infty, N^2h^{k+2} \to \infty$ *in the limit there is asymptotic bias:*

$$\begin{aligned} E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)) &\to C_2\mathcal{B}(K); \\ E(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right))(Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right)^T &\to \Sigma_{1\delta}(K); \end{aligned}$$

iii. *for* $h : N^2h^{k+2+2\bar{v}} \to \infty$ *the bias dominates:*

$$h^{-\bar{v}}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{p} \mathcal{B}(K).$$

Proof. See Appendix.

The Theorem shows that selection of the optimal bandwidth and kernel order to minimize the mean squared error critically depends on our knowledge of the degree of smoothness of the density. The $MSE(\hat{\delta}_N(K,h))$ can be represented as

$$MSE(\hat{\delta}_N(K,h)) = (\Sigma_{1\delta}(K) + o(1))\, N^{-2}h^{-(k+2)} + (\Sigma_{2\delta} + o(1))\, N^{-1} + \left(\mathcal{B}(K)\mathcal{B}^T(K) + o(1)\right) h^{2\bar{v}},$$

and the optimal bandwidth yields

$$h^{opt} = cN^{-2/(2\bar{v}+k+2)}. \tag{9}$$

If v, the true differentiability (smoothness) of f' is known we can choose the order of kernel $v(K) \leq [v]$ and then since $\bar{v} = v(k)$ the problem of efficient estimation reduces to finding an appropriate c (e.g., Powell and Stoker, 1996). If derivatives of high order exist, further improvements in efficiency can be obtained by using a higher order kernel to reduce the bias. The advantage of being able to assume this high differentiability order is the insensitivity of the limit process to the bandwidth and kernel over a range of choices that satisfy the assumptions (among which $Nh^{k+2} \to \infty$). If, however, the density is not sufficiently smooth the parametric rate may not be achievable and bandwidth and kernel

choices become crucial in ensuring good performance. Moreover, if the degree of density smoothness is not known there is no guidance for the choice of kernel and bandwidth.

In Kotlyarova and Zinde-Walsh (2006), hereafter referred to as KZW, the situation where there is uncertainty about smoothness of density was considered. We note here that the results concerning MSE in KZW do not require that the limit processes be Gaussian and rely on moment conditions only. Theorem 1 here corresponds to the moment requirements of Assumption 1 of that paper and demonstrates that when our Assumptions 1–6 are satisfied the estimator satisfies the part of Assumption 1 that requires first and second limit moments. We next establish that the part of Assumption 2 of that paper as it relates to first and second moments is satisfied as well. Consider kernel/bandwidth pairs (K_t, h_p) for a set of kernels, $K_t, t = 1, ...T$ and bandwidths, $h_p, p = 0, 1, ...P$, and order those pairs: $(K_t, h_p) \equiv (K_s, h_s), s = 1, ...S$, (where some s corresponds to each pair) and the corresponding estimators, $\hat{\delta}_N(K_s, h_s)$. If all satisfy the assumptions of Theorem 1 then there exist corresponding rates r_{Ns} for each pair such that $r_{Ns}\left(\hat{\delta}_N(K_s, h_s) - \delta_0\right)$ has finite first and second moments. If we consider the vector with stacked vector components $r_{Ns}\left(\hat{\delta}_N(K_s, h_s) - \delta_0\right)$, $s = 1, ...S$ we need to establish the finite limit covariances between these components. Define

$$\begin{aligned}
\Sigma_{1\delta}(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) &= 4E\left[y^2 f(x_i)\mu_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) + \mu_2^*(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2})(gf)(x_i)y_i\right]; \\
\text{with } \mu_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) &= \int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})^T du; \qquad (10) \\
\mu_2^*(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) &= \int K'_{s_1}(u)K'_{s_2}(-u\frac{h_{s_1}}{h_{s_2}})^T du, \text{ (under symmetry } \mu_2^* = -\mu_2).
\end{aligned}$$

Theorem 2. *Under the Assumptions of Theorem 1 for the vector with components* $r_{Ns}\left(\hat{\delta}_N(K_s, h_{Ns}) - \delta_0\right), s = 1, ...S$ *the limit covariance matrix has* $k \times k$ *blocks corresponding to* s_1, s_2

$$\Gamma_{s_1 s_2} = \Sigma_{1\delta}(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2})N^{-2}h_{s_2}^{-(k+1)}h_{s_1}^{-1} + \Sigma_{2\delta}N^{-1} + O(N^{-2}) \qquad (11)$$

and for components that correspond to estimators that converge at different rates limit covariances are zero.

Proof. See appendix.

Consider a linear combination of the estimators

$$\hat{\delta}_N^* = \sum_s a_s \hat{\delta}_N(K_s, h_{Ns}) \text{ with } \sum_{s=1}^{S} a_s = 1.$$

We can represent the limit variance of $\hat{\delta}_N^*$ as

$$\sum_{s_1} \sum_{s_2} a_{s_1} a_{s_2} Cov(\hat{\delta}_N(K_{s_1}, h_{s_1}), \hat{\delta}_N(K_{s_2}, h_{s_2})) \equiv \sum a_{s_1} a_{s_2} \Gamma_{s_1 s_2},$$

where $\Gamma_{s_1 s_2}$ is given by (11) .

The $MSE(\hat{\delta}_N^*) = MSE(\sum_s a_s \hat{\delta}_N(K_s, h_s))$ can then be represented as

$$MSE(\hat{\delta}_N^*) = \sum a_{s_1} a_{s_2} (\mathcal{B}_{s_1} \mathcal{B}_{s_2}^T + \Gamma_{s_1 s_2}).$$

To optimally choose the weights a_s, we will minimize the trace of the AMSE as in KZW.[2]

$$tr(AMSE(\hat{\delta}_N^*) = \sum a_{s_1} a_{s_2} (\tilde{\mathcal{B}}_{s_1}^T \tilde{\mathcal{B}}_{s_2} + tr\tilde{\Gamma}_{s_1 s_2}) = a' D a,$$
$$\text{where } \{D\}_{s_1 s_2} = \mathcal{B}_{s_1}^T \mathcal{B}_{s_2} + tr\Gamma_{s_1 s_2},$$

$\tilde{\mathcal{B}}_s = \mathcal{B}_s / r_{Ns}$, and $\tilde{\Gamma}_{s_1 s_2} = \Gamma_{s_1 s_2} / (r_{Ns_1} * r_{Ns_2})$.

The combined estimator is defined as the linear combination with weights that minimize the estimated $tr(AMSE(\hat{\delta}_N^*))$.

KZW discuss the optimal weights that minimize the (consistently estimated) $tr(AMSE(\hat{\delta}_N^*))$ subject to $\sum_s a_s = 1$. They show that the trace of AMSE of the combined estimator converges at a rate no worse than that of the trace of AMSE for the fastest converging individual estimator. Moreover, it is possible for the combined estimator to converge at a faster rate than any of the individual estimators that enter into the combination and for ADE to achieve a parametric rate even in the case when due to lack of differentiability none of the individual estimators can. To illustrate this point we provide a hypothetical example.

[2]Note MSE only provides a complete ordering when $\hat{\delta}_N^*$ is a scalar, using a trace is one way to obtain a complete ordering. Depending on which scalar function of the AMSE is used the order might differ.

Example: Suppose that the density f satisfies Assumption (5) with $m = 2, \alpha > 0$, so that $v = \bar{v} = 2$. Assume also that $k = 3$. Then the optimal bandwidth rate for the ADE estimator by (9) is $N^{-\frac{2}{9}}$ and provides an oversmoothed estimator that converges at the nonparametric rate $N^{-\frac{4}{9}}$ regardless of the order of the kernel. Suppose now that for bandwidth $h = N^{-\frac{2}{9}}$ and for some kernels K_j the ADE estimator allows the following expansion $\hat{\delta}(h_{oi}, K_j) = \delta_0 + h_{oi}^2 \mathcal{B}(K_j) + h_{oi}^{2+\alpha} \mathcal{B}_\gamma(K_j) + (Nh_{oi}^2)^{-1} \tilde{\delta}(h_{oi}, K_j) + N^{-1/2} \tilde{\delta}_{1/2}(h_{oi}, K_j) + o_p(h_{oi}^{2+\alpha} + (Nh_{oi}^2)^{-1} + N^{-1/2})$ with $\mathcal{B}'s$ denoting some constant 3×1 vectors and $\tilde{\delta}'s$ denoting random vectors with zero means and bounded variances. Consider four estimators corresponding to different kernels with bandwidth $h = N^{-\frac{2}{9}}$. Consider a vector of weights $a = (a_1, a_2, a_3, a_4)'$ with $a_4 = 1 - a_1 - a_2 - a_3$; it can be chosen to satisfy $a_1 \mathcal{B}(K_1) + a_2 \mathcal{B}(K_2) + a_3 \mathcal{B}(K_3) + a_4 \mathcal{B}(K_4) = 0$. For such weights

$$\begin{aligned} & a_1 \delta(h_o, K_1) + a_2 \delta(h_o, K_2) + a_3 \delta(h_o, K_3) + a_4 \delta(h_o, K_4) \\ = \; & O_p(h_o^{2+\alpha} + (Nh_o^2)^{-1} + N^{-1/2}), \end{aligned}$$

note that $(Nh_o^2)^{-1} = N^{-\frac{5}{9}} = o(N^{-1/2})$, $(Nh_o^2)^{-1} = N^{-\frac{4-2\alpha}{9}} = o(N^{-1/2})$ for any $\alpha > .25$ and in such a case parametric rate is achieved for the combined estimator (by weights that minimize the MSE of the linear combination). Note that the specific bandwidth $h = N^{-\frac{2}{9}}$ was used to simplify the example, but it is easy to see that for a range of bandwidths $N^{-\frac{2}{9}+\varepsilon}$ for $0 \leq \varepsilon < \frac{4\alpha-1}{18(2+\alpha)}$ when $\alpha > .25$ weights that produce a parametrically convergent combined estimator exist.

The optimality property of the combined estimator relies on consistent estimation of biases and covariances.[3] To provide a consistent estimate for the asymptotic variance that does not rely on the degree of smoothness, we apply the bootstrap, which gives

$$\begin{aligned} \widehat{\tilde{\Gamma}}_{s_1,s_2} &= \widehat{Cov}_B(\hat{\delta}_N(K_{s_1}, h_{s_1}), \hat{\delta}_N(K_{s_2}, h_{s_2})) \qquad (12) \\ &= \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\delta}_{b,N}(K_{s_1}, h_{s_1}) - \hat{\delta}_N(K_{s_1}, h_{s_1}) \right) \left(\left(\hat{\delta}_{b,N}(K_{s_2}, h_{s_2}) - \hat{\delta}_N(K_{s_2}, h_{s_2}) \right) \right)', \end{aligned}$$

[3]Examples in KZW demonstrate that a combined estimator can reduce the AMSE relative to an estimator based on incorrectly assumed high smoothness level even when the weights are not optimally determined.

where subscript b indicates the estimator obtained from a bootstrapped sample. To provide estimates of the biases, we need to assume that for all kernels we consider an undersmoothed bandwidth, yielding an asymptotic bias equalling zero. Let $h_{s,0}$ denote the smallest bandwidth we consider for kernel K_s. The estimator for the bias is obtained as

$$\widehat{\widehat{\mathcal{B}}}_s \equiv \widehat{Bias}(\hat{\delta}_N(K_s, h_s)) = \hat{\delta}_N(K_s, h_s) - \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{B}\sum_{b=1}^{B}\hat{\delta}_{b,N}(K_t, h_{t,0})\right].$$

That is we subtract from $\hat{\delta}_N(K_s, h_s)$ the bootstrapped averaged estimates at the lowest bandwidths for all the kernels, $i = 1, ..., T$.

4. SIMULATION

In order to illustrate the effectiveness of the combined estimator, we provide a Monte Carlo study where we consider the Tobit model. The Tobit model under consideration is given by

$$\begin{aligned} y_i &= y_i^* \text{ if } y_i^* > 0, \quad y_i^* = x_i^T\beta + \varepsilon_i, \quad i = 1, ..., n \\ &= 0 \text{ otherwise,} \end{aligned}$$

where our dependent variable y_i is censored to zero for all observations for which the latent variable y_i^* lies below a threshold, which without loss of generality is set equal to zero.

We randomly draw $\{(x_i, \varepsilon_i)\}_{i=1}^n$, where we assume that the errors, drawn independently of the regressors, are standard Gaussian. Consequently, the conditional mean representation of y given x can be written as

$$g(x) = x^T\beta \cdot \Phi(x^T\beta) + \phi(x^T\beta),$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the standard normal cdf and pdf respectively. Irrespective of the distributional assumption on ε_i, this is a single index model as the conditional mean of y given x depends on the data only through the index $x^T\beta$. While MLE obviously offers the asymptotically efficient estimator of β, (density weighted) ADE offers a semiparametric estimator for β which does not rely on the Gaussianity assumption on ε_i. Under the usual

smoothness assumptions, the finite sample properties of ADE for this Tobit model have been considered in the literature (Nichiyama and Robinson, 2005).

We select two explanatory variables, and set $\beta = (1,1)^T$. We make various assumptions about the distribution of our independent, explanatory variables. Our base model, labeled (s,s), uses two independent standard normal explanatory variables, or $f_{ss}(x_1, x_2) = \phi(x_1)\phi(x_2)$. The other models introduce various mixtures of normal explanatory variables, which while still being infinitely differentiable, do allow behavior resembling that of non-smooth densities. In particular, we consider the trimodal normal mixture $0.5\phi(x+0.767) + 3\phi(\frac{x+0.767-0.8}{0.1}) + 2\phi(\frac{x+0.767-1.2}{0.1})$ (KZW) and the double claw and discrete comb mixture densities (Marron and Wand, 1992) which we denote respectively by $f_m(x)$, $f_c(x)$, and $f_d(x)$. The (s,m) model we consider uses $f_{sm}(x_1, x_2) = \phi(x_1)f_m(x_2)$; the (m,m) model $f_{mm}(x_1, x_2) = f_m(x_1)f_m(x_2)$; the (s,c) model $f_{sc}(x_1, x_2) = \phi(x_1)f_c(x_2)$; the (s,d) model $f_{sd}(x_1, x_2) = \phi(x_1)f_d(x_2)$; and the (c,d) model $f_{cd}(x_1, x_2) = f_c(x_1)f_d(x_2)$. We vary the sample size from 1000 to 2000 observations and draw 100 replications in each case.

The multivariate kernel function $K(\cdot)$ (on R^2) is chosen as the product of two univariate kernel functions. We use a second and fourth order kernel in our Monte Carlo experiment, where, given that we use two explanatory variables, the highest order satisfies the theoretical requirement for ascertaining a parametric rate subject to the necessary smoothness assumptions. Both are bounded, symmetric kernels, which satisfy the assumption that the kernel and its derivative vanish at the boundary.

For each kernel we consider five different bandwidths. First, we apply the usual cross-validation for nonparametric regression, yielding a bandwidth sequence $h^{gcv} = cN^{-1/(2\bar{v}+2)}$ (see Stone (1982)) with $\bar{v} = \min(v, v(K))$. We allow the cross validated bandwidths to be distinct for each explanatory variable and obtain them using a gridsearch[4]. For densities of sufficient smoothness, $\bar{v} = v(K)$, this cross validated bandwidth does not rep-

[4]The cross validated bandwidths for the second and fourth order kernel in the (s,s) model with $N = 2000$ were $\binom{0.66}{0.66}$ and $\binom{1.50}{1.50}$ respectively (compared to $\binom{0.74}{0.74}$ and $\binom{1.58}{1.56}$ with $N = 1000$). With $N = 2000$, the bandwidths for the (s,m) model were $\binom{0.63}{0.52}$ and $\binom{1.54}{0.92}$ respectively; the (m,m) model $\binom{0.52}{0.52}$ and $\binom{1.19}{1.18}$; the (s,c) model $\binom{0.61}{0.70}$ and $\binom{1.45}{1.57}$; the (s,d) model $\binom{0.69}{0.43}$ and $\binom{1.57}{0.94}$; and the (c,d) model $\binom{0.75}{0.39}$ and $\binom{1.70}{0.97}$.

resent the undersmoothing required to ensure asymptotic unbiasedness as $Nh^{2\overline{v}} \to \infty$, providing cases (a,b)v. in Theorem 1 – the optimal bandwidth minimizing the mean squared error is given by $h^{opt} = cN^{-2/(2\overline{v}+4)}$. When densities are not sufficiently smooth, $\overline{v} = v < v(K)$, h^{gcv} will also correspond to oversmoothing providing case (c)iii. in Theorem 1. In ascending order, the bandwidths considered are $h_0 = h^{gcv} \cdot N^{-2[2/(2\overline{v}+4)-1/(2\overline{v}+2)]}$, $h_1 = h^{gcv} \cdot N^{-[2/(2\overline{v}+4)-1/(2\overline{v}+2)]}$, $h_2 = h^{gcv} \cdot N^{-\frac{1}{2}[2/(2\overline{v}+4)-1/(2\overline{v}+2)]}$, $h_3 = h^{gcv}$, and $h_4 = h^{gcv} \cdot N^{\frac{1}{2}[2/(2\overline{v}+4)-1/(2\overline{v}+2)]}$. Clearly h_4 would yield an oversmoothed bandwidth, and h_1 would yield the rate appropriate for the optimal bandwidth h^{opt}. With $\bar{v}$ left unspecified, we suggest to evaluate $N^{-[2/(2\overline{v}+4)-1/(2\overline{v}+2)]}$, the optimal weighting of h^{gcv} (h_1), at the value of $\bar{v}$ giving the smallest weight (i.e., $\bar{v} = \sqrt{2}$ – note that the derivative of $-[2/(2\overline{v}+4) - 1/(2\bar{v}+2)]$ wrt $\bar{v}$ is zero at $\bar{v} = \sqrt{2}$) in an attempt to guard against insufficient undersmoothing. With $N = 1000$, this yields $h_0 = 0.3h^{gcv}$, $h_1 = 0.55h^{gcv}$, $h_2 = 0.74h^{gcv}$, and $h_4 = 1.34h^{gcv}$. Estimators for biases and covariances of the density weighted ADEs are obtained by bootstrap (with 250 bootstraps) as discussed in the previous section.

In table A1, in the Appendix, we report the true finite sample Root Mean Squared Errors (RMSE) of various density weighted average derivatives for the sample sizes $N = 1000$ and $N = 2000$. We consider two combined estimators depending on whether we use h_0 or h_1 as our smallest bandwidth. In all models the biases and standard deviations of the individual estimators on average (not reported) behave as expected: as the bandwidth increases, bias becomes more pronounced and the standard deviation declines. The theoretical standard deviation (using the leading two components of $\text{Var}(\hat{\delta}_N)$ given in (6) compares very well with the standard deviation based on the bootstrap.

We note large discrepancies in RMSE performance between the models (i.e., between the different distributions considered for the explanatory variables) and within each model for the range of bandwidths considered. All mixtures of normals considered exhibit large partial derivatives due to their high modal nature that, despite satisfying the usual smoothness assumptions, can be seen to clearly undermine the performance of the density weighted ADE. The RMSE of the ADE is largest for the (m,m) model, the model where both

explanatory variables are drawn from a trimodal normal mixture, ranging between 0.1296 and 0.1877 when $N = 1000$. This contrasts sharply with the RMSE in the (s,s) case where it ranges between 0.0045 and 0.0156. Comparing the mixtures models (s,m), (s,c), and (s,d), the discrete comb mixture, in particular, generates a large variation in RMSE performance for the range of bandwidths considered, whereas the performance in RMSE for the claw mixture is fairly stable. With $N = 1000$ the RMSE ranges between 0.0449 and 0.0618 in the (s,m) model, versus 0.0237–0.0300 and 0.0176–0.0690 in the (s,c) and (s,d) model respectively.

The following tables summarize the performance of various individual estimators as well as the combined estimators for the models considered. Our comparison is based on the RMSE; under "best" we list the estimator(s) with the lowest RMSE, under "worst" that with the largest RMSE (we note the ratio to RMSE of "best", $r > 1$, in brackets); "good" estimators with RMSE giving the ratio to "best" below 1/3 that of the worst: $< 1+\frac{1}{3}(r-1)$, "fair" with the ratio below 2/3: $1 + \frac{2}{3}(r-1)$, "bad" with the ratio higher than that.

The implication from these tables is clear. There is no rule regarding either kernel order or bandwidth that works uniformly (similar results found by Hansen, 2005): some individual estimators that are best for one model are worst for another; the behavior can change substantially with change in sample size. For example the advantage of using higher order kernel with smaller than cross-validated bandwidths for the (m,m) model contrasts with its poor performance in the (c,d) and (s,d) models, where second order kernel at higher than cross-validated bandwidths performs best. The individual estimators that are never bad are those with K_2 at h_1 and h_2 and with K_4 at h_1, h_2; however for all these estimators (except (K_2, h_2) that has 3 "good", 3 "fair" scores at both sample sizes) relative performance deteriorated as sample size increased, e.g. for (K_4, h_2) at 1000 there are 2 "best", 3 "good" and one "fair", going to 2 "best", 2"good" and 2 "fair" at 2000.

At the same time the combined estimator, typically not providing the best RMSE, does provide more robust results – never delivering a RMSE worse than the poorest performing kernel bandwidth pair. Generally, the RMSE of the combined estimator performs better than individual estimators for a large range of bandwidths. In particular, the combined

Table 1: Tobit Model: Summary Performance RMSE ADE estimates $N = 1000$

Model	best	good	fair	bad	worst
(s,s)	(K_4, h_2)	$(K_2, h_1/h_2/h_3)$ $(K_4, h_1/h_3/h_4)$ $(comb, h_1)$	(K_2, h_4) $(comb, h_0)$	(K_4, h_0)	(K_2, h_0) [3.467]
(s,m)	$(K_4, h_1/h_2)$	$(K_2, h_0/h_1/h_2/h_3)$ $(K_4, h_0/h_3)$ $(comb, h_0/h_1)$		(K_4, h_4)	(K_2, h_4) [1.376]
(m,m)	(K_4, h_0)	$(K_2, h_0/h_1)$ (K_4, h_1) $(comb, h_1)$	(K_2, h_2) $(K_4, h_2/h_3)$ $(comb, h_0)$	$(K_2, h_3/h_4)$	(K_4, h_4) [1.448]
(s,c)	(K_4, h_1)	(K_2, h_1) $(K_4, h_0/h_2/h_3)$ $(comb, h_0/h_1)$	$(K_2, h_0/h_2/h_3)$	(K_4, h_4)	(K_2, h_4) [1.266]
(s,d)	(K_2, h_4)	(K_2, h_3) $(K_4, h_2/h_3/h_4)$	$(K_2, h_1/h_2)$ (K_4, h_1) $(comb, h_1)$	(K_4, h_0) $(comb, h_0)$	(K_2, h_0) [3.920]
(c,d)	(K_2, h_4)	$(K_2, h_1/h_2/h_3)$ $(K_4, h_1/h_2/h_3/h_4)$ $(comb, h_1)$		(K_2, h_0) $(comb, h_0)$	(K_4, h_0) [3.155]

Table 2: Tobit Model: Summary Performance RMSE ADE estimates $N = 2000$

Model	best	good	fair	bad	worst
(s,s)	(K_4, h_2)	$(K_2, h_1/h_2/h_3)$ $(K_4, h_1/h_3/h_4)$ $(comb, h_1)$	(K_2, h_4) (K_4, h_0) $(comb, h_0)$		(K_2, h_0) [4.000]
(s,m)	$(comb, h_0)$	$(K_2, h_0/h_1)$ $(K_4, h_0/h_1/h_2)$ $(comb, h_1)$	$(K_2, h_2/h_3)$ (K_4, h_3)	(K_4, h_4)	(K_2, h_4) [1.525]
(m,m)	(K_4, h_0)	$(K_2, h_0/h_1)$ $(K_4, h_0/h_1)$	$(K_2, h_2/h_3)$ $(K_4, h_1/h_2)$	(K_2, h_4) (K_4, h_3)	(K_4, h_4) [1.638]
(s,c)	(K_4, h_2)	$(K_2, h_1/h_2)$ $(K_4, h_0/h_1/h_3)$ $(comb, h_0/h_1)$	(K_2, h_3) (K_4, h_4)	(K_2, h_4)	(K_2, h_0) [1.263]
(s,d)	(K_2, h_4)	(K_2, h_3) $(K_4, h_3/h_4)$	$(K_2, h_1/h_2)$ $(K_4, h_1/h_2)$ $(comb, h_1)$	(K_2, h_0) $(comb, h_0)$	(K_4, h_0) [3.841]
(c,d)	(K_2, h_4)	$(K_2, h_2/h_3)$ $(K_4, h_1/h_2/h_3/h_4)$ $(comb, h_1)$	(K_2, h_1)	(K_4, h_0) $(comb, h_0)$	(K_2, h_0) [3.046]

estimator $(comb, h_1)$ performs stably with 5 "good" and 1 "fair" at both sample sizes even though some very badly behaved estimators enter into the combination. The estimator $(comb, h_1)$ for the base (s,s) model, reveals a performance similar to the optimal ADE estimator and does not exhibit much of an efficiency loss confirming the results about its being equivalent to the optimal rate estimator. The same can be said for the (s,c) model. The estimator $(comb, h_0)$ can be "bad"; as expected the RMSE performance of the combined estimator is related to the RMSE performance of the smallest bandwidth used for each kernel; nevertheless it is never as bad as the worst individual estimator. In the base model, the use of h_0 as the smallest bandwidth also worsens the RMSE performance of our combined estimator relative to h_1 due to the increased variability it imposes. On the other hand for the (s,m) model with $N = 2000$, the estimator $(comb, h_0)$ even outperforms the optimal ADE kernel. This reveals that in case where the density is not sufficiently smooth or while smooth as a shape that gives high values for low-order derivatives, gains from the combined estimator relative to the optimal ADE estimator can be obtained. Clear gains from using the combined ADE estimator for the models with the trimodal mixture of normals and/or discrete comb can be observed. Both combined estimators in the (s,m) and (m,m) models lie closer to the optimal ADE estimator than the worst ADE estimator. The large variation in RMSE performance for the range of bandwidths and choice of kernel considered in the (s,d) and (c,d) models illustrate to the potential gains of using the combined ADE estimator in the discrete comb setting. When comparing both combined estimators to the, fairly stable performing, individual (K_4, h_2) estimator we note that generally the relative performance of the combined estimator improves with the sample size. In table A.2 in the appendix, the ratio of the RMSE of our combined estimator to the individual (K_4, h_2) estimator for our models is provided.

In agreement with the results for the combined estimator, oversmoothed individual estimators get weights of different signs reflecting the tendency of the combined estimator to balance off the biases. With $N = 2000$, on average the weights $(a_{2,h_1}, ..., a_{2,h_4}; a_{4,h_1}, ..., a_{4,h_4})$ for (s,s) the model are (0.03, −0.07, 0.23, −0.80; 0.03, 0.10, 0.40, 1.07); for (s,m) the weights are (0.05, −0.20, 0.65, −1.88; 0.03, 0.36, 0.44, 1.55); for (m,m) (0.16, 0.04, 0.86,

-1.89; -0.15, 0.14, 1.30, 0.53); for (s,c) (0.03, -0.01, 0.24, -0.80, 0.01, 0.05, 0.65, 0.82); for (s,d) (-0.03, 0.06, 0.92, -1.55, 0.37, 0.14, 0.77, 0.32); and for (c,d) (0.05, -0.21, 0.83, -1.04, 0.44, 0.02, 0.76, 0.16) – comparable weights are obtained when h_0 is used as smallest bandwidth when giving zero weight to this bandwidth in the combination. More weight, including of opposite signs, are given to the higher bandwidths for the second and fourth order kernel.

In Table A3 we present selected RMSE estimates of the parameters in Tobit model. Since the ADE allows for the estimation of $\beta = (\beta_1, \beta_2)^T$ up to scale, we considered results of the parameter estimates subject to three possible normalisations: (i) where β_1 is standardized to 1 (unit normalization), (ii) where the estimated slope coefficient for β_2 is rescaled to have the sum of their absolute values equal to 2 (normalization considered by PSS), and (iii) where we consider the polar coordinate $\arctan(\beta_2/\beta_1)$. The results based on the density weighted ADE estimator are provided for each kernel/bandwidth pair selection as well as for the combined estimator. For comparison, RMSE of conformably normalized Tobit MLE parameter estimates are reported as well.

We make the following observations: Superiority in estimating ADE does not necessarily translate into better parameter estimators. Only in the (s,s) model does the optimal kernel bandwidth combination for parameter estimates (K_4, h^{gcv}) compare well with the optimal ADE kernel/bandwidth combination (K_4, h_2). The loss in efficiency arising from not knowing the distribution of the disturbances occurs as expected, but is within reason in this case: the standard deviation of the combined semiparametric estimator is less than double that of the Tobit MLE. In all other models, the optimal kernel bandwidth combination for parameter estimates typically differ substantially from that for the optimal ADE. Moreover in these models, the unit normalization typically gives rise to different optimal kernel bandwidth combinations than the PSS and polar normalizations, which by and large perform comparably. The reason for this is that, as argued also by PSS, the unit normalization, in particular when choosing small bandwidths, exhibit ill-behaved sample moments (arising from taking the ratio of two estimators which might be arbitrarily large). For the (m,m) model, where the optimal ADE kernel bandwidth combination (K_4, h_0) is,

this clearly points to a wedge between optimal ADE performance and good parameter estimates, but even for the PSS and polar normalization the results indicate that the bandwidth should be less undersmoothed than indicated for optimal ADE. For the (c,d) model, the performance of the PSS and polar normalized parameter estimates even suggest the use of a higher order kernel, not required for optimal ADE performance. While the results are suggestive that more robust parameter estimates can be obtained with the help of our combined ADE, it is noted that the weights are not chosen with optimality of RMSE for the parameter estimates in mind. This provides an interesting area of future research.

5. Conclusions

We have questioned in this paper the high degree of density smoothness assumed in the literature for obtaining the parametric rate for ADE. We show that insufficient smoothness will result in possible asymptotic bias and may easily lead to non-parametric rates. The selection of optimal kernel order and optimal bandwidth (Powell and Stoker, 1996) in the absence of sufficient smoothness moreover presumes the knowledge of the degree of density smoothness. Our Monte Carlo simulations demonstrate that even in the case where formally the smoothness assumptions hold, due to large values for the derivatives, the behavior of ADE becomes problematic. By not relying on a single kernel bandwidth choice, our combined estimator reduces this sensitivity.

6. Appendix

The proof of Theorems 1 and 2 relies on the following Lemmas 1 and 2, correspondingly, where moments are computed under the general assumptions of this paper.

We obtain the moments by direct computation for symmetric as well as non-symmetric kernels here.

Lemma 1. *Given Assumptions 1-4, the variance of* $\hat{\delta}_N(K,h)$ *can be expressed as*

$$\begin{aligned} & Var(\hat{\delta}_N(K,h)) \\ \equiv\ & (\Sigma_{1\delta} + o(1))\, N^{-2} h^{-(k+2)} + (\Sigma_{2\delta} + o(1))\, N^{-1} + O(N^{-2}) \end{aligned}$$

where

$$\begin{aligned}\Sigma_{1\delta} &= 4E\left[y_i^2 f(x_i)\mu_2(K) + \mu_2^*(K)g(x_i)f(x_i)y_i\right], \\ \Sigma_{2\delta} &= 4\left\{E(g'(x_i)f(x_i) - (y_i - g(x_i))f'(x_i))(g'(x_i)f(x_i) - (y_i - g(x_i)f'(x_i)))^T\right\} - 4\delta_0\delta_0^T,\end{aligned}$$

for

$$\begin{aligned}\mu_2(K) &= \int K'(u)K'(u)^T du \\ \mu_2^*(K) &= \int K'(u)K'(-u)^T du, \text{ (under symmetry } \mu_2^*(K) = -\mu_2(K)).\end{aligned}$$

Proof. First, recall that

$$\text{Bias}(\hat{\delta}_N(K,h)) = -2E(A(K,h,x_i)y_i = h^{\overline{v}}B(K) + o(h^{\overline{v}})$$

with

$$A(K,h,x_i) = \int K(u)(f'(x_i - uh) - f'(x_i))du. \tag{A.1}$$

To derive an expression for the Variance of $\hat{\delta}_N(K,h)$, we note

$$Var(\hat{\delta}_N(K,h)) = E(\hat{\delta}_N(K,h)\hat{\delta}_N(K,h)^T) - E\hat{\delta}_N(K,h)E\hat{\delta}_N(K,h)^T.$$

Let $I(a) = 1$, if the expression a is true, zero otherwise. We decompose the first term as follows

$$\begin{aligned}& E\left(\hat{\delta}_N(K,h)\hat{\delta}_N(K,h)^T\right) \qquad (A.2)\\ =\ & 4E\left\{\left[\tfrac{1}{N}\sum_{i=1}^N \hat{f}'_{(K,h)}(x_i)y_i\right]\left[\tfrac{1}{N}\sum_{i=1}^N \hat{f}'_{(K,h)}(x_i)y_i\right]^T\right\} \\ =\ & 4\left\{\tfrac{1}{N}E\left(\hat{f}'_{(K,h)}(x_i)\hat{f}'_{(K,h)}(x_i)^T y_i^2\right) + \tfrac{N-1}{N}E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1}y_{i_2}I(\ i_1 \neq i_2)\right)\right\}.\end{aligned}$$

The first expectation yields

$$
\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_i)\hat{f}'_{(K,h)}(x_i)^T y_i^2\right) \qquad (A.3)\\
= \;& \left(\tfrac{1}{N-1}\right)^2 E\left\{E_{z_i}\left(y_i^2\left[\sum_{j\neq i}\left(\tfrac{1}{h}\right)^{k+1}K'(\tfrac{x_i-x_j}{h})\right]\left[\sum_{j\neq i}\left(\tfrac{1}{h}\right)^{k+1}K'(\tfrac{x_i-x_j}{h})\right]^T\right)\right\}\\
= \;& \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left[y_i^2E_{z_i}\left(K'(\tfrac{x_i-x_j}{h})K'(\tfrac{x_i-x_j}{h})^T I(\ i\neq j)\right)\right]+\\
& \tfrac{N-2}{N-1}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left[y_i^2E_{z_i}\left(K'(\tfrac{x_i-x_{j_1}}{h})K'(\tfrac{x_i-x_{j_2}}{h})^T I(\ i,j_1,j_2 \text{ pairwise distinct})\right)\right]\\
= \;& \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{k+2}E\left[y_i^2\int K'(u)K'(u)^T f(x_i-uh)du\right]+\\
& \tfrac{N-2}{N-1}\left(\tfrac{1}{h}\right)^{2k+2}E\left[E_{z_i}\left(y_iK'(\tfrac{x_i-x_{j_1}}{h})\right)E_{z_i}\left(y_iK'(\tfrac{x_i-x_{j_2}}{h})\right)^T I(\ i,j_1,j_2 \text{ pairwise distinct})\right]\\
= \;& \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{k+2}\left[Ey_i^2f(x_i)\mu_2(K)+O(h)\right]+\\
& \tfrac{N-2}{N-1}\cdot\left[E(f'(x_i)y_i)(f'(x_i)y_i)^T+O(h^{\overline{v}})\right],
\end{aligned}
$$

where for the third and the last equality we use change of variable in integration and independence of x_{j_1}, x_{j_2}; by Assumptions 4 and 5 the moments of the additional terms are correspondingly bounded. Further

$$
\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_i)\hat{f}'_{(K,h)}(x_i)^T y_i^2\right)\\
= \;& \{\tfrac{1}{N}\cdot\left(\tfrac{1}{h}\right)^{k+2}\left[Ey_i^2f(x_i)\mu_2(K)+O(h)\right]\\
& +\left[E(f'(x_i)y_i)(f'(x_i)y_i)^T+O(h^{\overline{v}})\right]\}\{1+O(N^{-1})\}.
\end{aligned}
$$

The second expectation yields,

$$
\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1}y_{i_2}I(i_1\neq i_2)\right)\\
= \;& \left(\tfrac{1}{N-1}\right)^2\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}\sum_{j_1\neq i_1}\sum_{j_2\neq i_2}K'(\tfrac{x_{i_1}-x_{j_1}}{h})K'(\tfrac{x_{i_2}-x_{j_2}}{h})^T\right)\\
= \;& \tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{j_1}}{h})K'(\tfrac{x_{i_2}-x_{j_1}}{h})^T I(\ i_1,i_2,j_1 \text{ pairwise distinct})\right)\\
& +\tfrac{1}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{i_2}}{h})K'(\tfrac{x_{i_2}-x_{i_1}}{h})^T I(\ i_1,i_2 \text{ pairwise distinct})\right)\\
& +\tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{i_2}}{h})K'(\tfrac{x_{i_2}-x_{j_2}}{h})^T I(\ i_1,i_2,j_2 \text{ pairwise distinct})\right)\\
& +\tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{j_1}}{h})K'(\tfrac{x_{i_2}-x_{i_1}}{h})^T I(\ i_1,i_2,j_1 \text{ pairwise distinct})\right)\\
& +\tfrac{(N-2)(N-3)}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{j_1}}{h})K'(\tfrac{x_{i_2}-x_{j_2}}{h})^T I(\ i_1,i_2,j_1,j_2 \text{ pairwise distinct})\right).
\end{aligned}
$$

Using the law of iterated expectations, we rewrite

$$
\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1} y_{i_2} I(i_1 \neq i_2)\right) \qquad (A.4)\\
= \; & \tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2} E\left(E_{z_{j_1}}\left[y_{i_1}K'(\tfrac{x_{i_1}-x_{j_1}}{h})\right]E_{z_{j_1}}\left[y_{i_2}K'(\tfrac{x_{i_2}-x_{j_1}}{h})\right]^T\right) + \\
& \tfrac{1}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2} E\left(y_{i_2}E_{z_{i_2}}\left[y_{i_1}K'(\tfrac{x_{i_1}-x_{i_2}}{h})K'(\tfrac{x_{i_2}-x_{i_1}}{h})^T\right]\right) + \\
& \tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2} E\left(E_{z_{i_2}}\left[y_{i_1}K'(\tfrac{x_{i_1}-x_{i_2}}{h})\right]E_{z_{i_2}}\left[y_{i_2}K'(\tfrac{x_{i_2}-x_{j_2}}{h})\right]^T\right) + \\
& \tfrac{N-2}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2} E\left(E_{z_{i_1}}\left[y_{i_1}K'(\tfrac{x_{i_1}-x_{j_1}}{h})\right]E_{z_{i_1}}\left[y_{i_2}K'(\tfrac{x_{i_2}-x_{i_1}}{h})\right]^T\right) + \\
& \tfrac{(N-2)(N-3)}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2} E\left(E_{z_{i_1}}\left[y_{i_1}K'(\tfrac{x_{i_1}-x_{j_1}}{h})\right]\right)E\left(E_{z_{i_2}}\left[y_{i_2}K'(\tfrac{x_{i_2}-x_{j_2}}{h})\right]\right)^T,
\end{aligned}
$$

where for brevity we omit terms such as $I(i_1 \neq i_2)$ in the terms of the expression.

Next follow details of derivation. Denote

$$
\begin{aligned}
A(K,h,x_i) &= E_{z_i}\left[\hat{f}'_{(K,h)}(x_i) - f'(x_i)\right] = \int K(u)(f'(x_i-uh)-f'(x_i))du\\
B(K,h,x_i) &= \int K'(u)K'(u)^T\left(f(x_i-uh)-f(x_i)\right)du.\\
C(K,h,x_i) &= -\int K(u)\left[(gf)'(x_i+uh)-(gf)'(x_i)\right]du\\
D(K,h,x_i) &= \int K'(u)K'(-u)^T\left[(gf)(x_i+uh)-(gf)(x_i)\right]du
\end{aligned}
$$

$$
\begin{aligned}
c(x_i) &= -(gf)'(x_i)\\
d(K,x_i) &= \mu_2^*(K)(gf)(x_i)\\
\mu_2(K) &= \int K'(u)K'(u)^T du\\
\mu_2^*(K) &= \int K'(u)K'(-u)^T du, \text{ (under symmetry } \mu_2^*(K) = -\mu_2(K)).
\end{aligned}
$$

Then write for terms in (A.4). First, $E_{z_i}\left[\left(\frac{1}{h}\right)^{k+1}K'(\frac{x_i-x_j}{h})y_i\right] = f'(x_i)y_i + A(K,h,x_i)y_i$. The remaining conditional moments are

$$
\begin{aligned}
E_{z_j}\left[\left(\tfrac{1}{h}\right)^{k+1}K'(\tfrac{x_i-x_j}{h})y_i\right] &= c(x_j) + C(K,h,x_j) \qquad (A.5)\\
E_{z_i}\left[\left(\tfrac{1}{h}\right)^{k}K'(\tfrac{x_j-x_i}{h})K'(\tfrac{x_i-x_j}{h})^T y_j\right] &= d(K,x_i) + D(K,h,x_i). \qquad (A.6)
\end{aligned}
$$

Indeed, for (A.5)

$$\begin{aligned} E_{z_j}\left[\left(\tfrac{1}{h}\right)^{k+1} K'(\tfrac{x_i-x_j}{h})y_i\right] &= \left(\tfrac{1}{h}\right)^{k+1}\int K'(\tfrac{x-x_j}{h})(gf)(x)dx \\ &= \left(\tfrac{1}{h}\right)\int K'(u)(gf)(x_i+uh)dx \text{ (integration by parts)} \\ &= -(gf)'(x_j) - \int K(u)\left[(gf)'(x_j+uh)-(gf)'(x_j)\right]du \end{aligned}$$

For (A.6)

$$\begin{aligned} & E_{z_i}\left[\left(\tfrac{1}{h}\right)^{k} K'(\tfrac{x_j-x_i}{h})K'(\tfrac{x_i-x_j}{h})^T y_j\right] \\ = & \left(\tfrac{1}{h}\right)^{k}\int g(x)K'(\tfrac{x-x_i}{h})K'(\tfrac{x_i-x}{h})^T f(x)dx \text{ c.o.v. } x-x_i = hu \\ = & \int K'(u)K'(-u)^T(gf)(x_i+uh)dudy \\ = & \int K'(u)K'(-u)^T(gf)(x_i)du + \int yK'(u)K'(-u)^T\left[(gf)(x_i+uh)-(gf)(x_i)\right]du \\ = & d(K,x_i)+D(K,h,x_i). \end{aligned}$$

It is useful to note here that

$$\begin{aligned} E\left[E_{z_i}\left[\left(\tfrac{1}{h}\right)^{k+1} K'(\tfrac{x_i-x_j}{h})y_i\right]\right] &= E\left[E_{z_j}\left[\left(\tfrac{1}{h}\right)^{k+1} K'(\tfrac{x_i-x_j}{h})y_i\right]\right] \\ E\left[f'(x_i)y_i + A(K,h,x_i)y_i\right] &= E\left[c(x_j)+C(K,h,x_j)\right]. \end{aligned}$$

Indeed it can easily be verified that $E(f'(x_i)y_i) = E(c(x_j))$.

Using (A.1), (A.5), and (A.6) we can express (A.4) as

$$\begin{aligned} & E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1}y_{i_2}\right) \qquad (A.7) \\ = & \tfrac{N-2}{(N-1)^2}E\left[\left(c(x_i)+C(K,h,x_i)\right)\left(c(x_i)+C(K,h,x_i)\right)^T\right] + \\ & \tfrac{1}{(N-1)^2}\left(\tfrac{1}{h}\right)^{k+2} E\left[d(K,x_i)y_i + D(K,h,x_i)y_i\right] \\ & \tfrac{N-2}{(N-1)^2}E\left[\left(c(x_i)+C(K,h,x_i)\right)\left(f'(x_{i_1})y_i + A(K,h,x_i)y_i\right)^T\right] + \\ & \tfrac{N-2}{(N-1)^2}E\left[\left(f'(x_i)y_i + A(K,h,x_i)y_i\right)\left(c(x_i)+C(K,h,x_i)\right)^T\right] \\ & \tfrac{(N-2)(N-3)}{(N-1)^2}E\left[f'(x_i)y_i + A(K,h,x_i)y_i\right]E\left[f'(x_i)y_i + A(K,h,x_i)y_i\right]^T \end{aligned}$$

Combining (A.2), (A.3), and (A.7) yields,

$$\begin{aligned}
& E\left(\hat{\delta}_N(K,h)\hat{\delta}_N(K,h)^T\right) \\
= \; & \tfrac{4}{N(N-1)}\left(\tfrac{1}{h}\right)^{k+2} E\left[y_i^2 f(x_i)\mu_2(K) + B(K,h,x_i)y_i^2 + d(K,x_i)y_i + D(K,h,x_i)y_i\right] \\
& +4\tfrac{N-2}{N(N-1)}E((f'(x_i)y_i + A(K,h,x_i)y_i)(f'(x_i)y_i + A(K,h,x_i)y_i)^T) \\
& +4\tfrac{N-2}{N(N-1)}E\left[(c(x_i) + C(K,h,x_i))\,(c(x_i) + C(K,h,x_i))^T\right] \\
& +4\tfrac{N-2}{N(N-1)}E\left[(c(x_i) + C(K,h,x_i))\,(f'(x_i)y_i + A(K,h,x_i)y_i)^T\right] \\
& +4\tfrac{N-2}{N(N-1)}E\left[(f'(x_i)y_i + A(K,h,x_i)y_i)\,(c(x_i) + C(K,h,x_i))^T\right] \\
& +\tfrac{(N-2)(N-3)}{N(N-1)}\left(E\hat{\delta}_N(K,h)\right)\left(E\hat{\delta}_N(K,h)\right)^T.
\end{aligned}$$

The final expression (using repeatedly Assumptions 3-5 to show convergence to zero of expectation of terms involving quantities denoted in capitals) is

$$\begin{aligned}
& E\left(\hat{\delta}_N(K,h)\hat{\delta}_N(K,h)^T\right) \\
= \; & \tfrac{4}{N^2}\left(\tfrac{1}{h}\right)^{k+2}\left(E\left[y_i^2 f(x_i)\mu_2(K) + y_i(gf)(x_i)\mu_2^*(K)\right] + o(1)\right) \\
& +4\tfrac{1}{N}\left(E(y_i^2(f'(x_i)(f'(x_i)^T + (gf)'(x_i)(gf)'(x_i)^T + y_i(gf)'(x_i)(f'(x_i)^T + y_i f'(x_i)(gf)'(x_i)^T\right) + o(1)\right) \\
& +\tfrac{(N-2)(N-3)}{N(N-1)}\left(E\hat{\delta}_N(K,h)\right)\left(E\hat{\delta}_N(K,h)\right)^T.
\end{aligned}$$

Alternatively, we can write the variance expression in the form given in the statement of the Lemma. ■

Remark. For $N \cdot Var(\hat{\delta}_N(K,h))$ to converge, we require $Nh^{k+2} \to C$ with $C \in (0,\infty)$. Notice that indeed, given $Nh^{k+2} \to \infty$ (regardless of whether we assume the kernel to be symmetric),

$$\begin{aligned}
& NVar(\hat{\delta}_N(K,h)) \\
\to \; & 4\left\{E\left[c(x_i)c(x_i)^T\right] + E\left[f'(x_i)c(x_i)^T + c(x_i)f'(x_i)^T)y_i + y_i^2 f'(x_i)f'(x_i)^T\right]\right\} \\
= \; & 4\left\{E((g'(x_i)f(x) - (y_i - g(x_i)f'(x_i))(g'(x_i)f(x) - (y_i - g(x_i)f'(x_i))^T\right\} - 4\delta_0\delta_0^T \\
= \Sigma_\delta \qquad & \text{as in PSS 1989.}
\end{aligned}$$

Proof. Theorem 1.

From Lemma 1 it follows that the variance has two leading parts, one that converges to $\Sigma_{2\delta}$ at a parametric rate, $O(N^{-1})$, requiring $Nh^{k+2} \to \infty$; when this condition on the rate of the bandwidth does not hold, the variance converges at the rate $O(N^{-2}h^{-(k+2)})$ to $\Sigma_{1\delta}$. The squared bias converges at the rate $O(h^{2\overline{v}})$.

In cases (a,b)v. and (c)iii. the rate of the squared bias dominates the rates for both leading terms in the variance. By standard arguments (Chebyshev's inequality) this situation clearly results in convergence in probability to $\mathcal{B}(K)$ as stated in the Theorem.

In cases (a)iii. and iv. the part of the variance with the parametric rate dominates (with or without bias), $Nh^{k+2} \to \infty$, and the results for the moments follow; similarly moment conditions in (a,b) ii. hold for the finite non-zero $C : Nh^{k+2} \to C$. For the case $Nh^{k+2} \to \infty$ Theorem 3.3 in PSS applies. We adapt the proof of normality in Theorem 3.1 in PSS with a minor change: we accommodate a possible non-symmetric kernel.

To accommodate a non-symmetric kernel we redefine $p_N(z_i, z_j)$ (defined in PSS, (3.11)) as

$$p_N(z_i, z_j) = -\left(\frac{1}{h}\right)^{k+1}\left[K'\left(\frac{x_i - x_j}{h}\right)y_i + K'\left(\frac{x_j - x_i}{h}\right)y_j\right]; \tag{A.8}$$

then the $t_N(z_i)$ defined by PSS (3.15) changes to

$$t_N(z_i) = \int K(u)\left[(gf)'(x_i + uh) - (gf)'(x_i)\right]du - y_i\int K(-u)\left[f'(x_i + uh) - f'(x_i)\right]du.$$

The rest of the asymptotic normality argument for U-statistics follows through noting that it is now $\hat{U}_N$ as defined in PSS (3.9) with the new $p_N(z_i, z_j)$ from (A.8) for which PSS, Lemma 3.1 holds.

When $Nh^{k+2} \to 0$ as in cases (a,b)i. and (c)i. and ii. the variance converges with the non-parametric rate, $Nh^{\frac{k+2}{2}}$, to $\Sigma_{1\delta}$. If the degree of smoothness is low, $v < \frac{k+2}{2}$, then regardless of the order of the kernel a parametric rate cannot be obtained. ■

Lemma 2. *Under the assumptions of Theorem 1 the covariance between* $\hat{\delta}_N(K_{s_1}, h_{s_1})$ *and* $\hat{\delta}_N(K_{s_2}, h_{s_2})$, $\Gamma_{s_1 s_2}$, *is given by*

$$\Gamma_{s_1 s_2} \equiv \left(\Sigma_{1\delta}(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2}) + o(1)\right)N^{-2}h_{s_2}^{-(k+1)}h_{s_1}^{-1} + \left(\Sigma_{2\delta} + o(1)\right)N^{-1} + O(N^{-2})$$

where

$$\begin{aligned} & \Sigma_{1\delta}(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2}) \\ = \; & 4E\left[y^2 f(x_i)\mu_2(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2}) + \mu_2^*(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2})(gf)(x_i)y_i\right] \end{aligned}$$

with

$$\begin{aligned} \mu_2(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2}) &= \int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})^T du; \\ \mu_2^*(K_{s_1}, K_{s_2}, h_{s_1}/h_{s_2}) &= \int K'_{s_1}(u)K'_{s_2}(-u\frac{h_{s_1}}{h_{s_2}})^T du. \end{aligned}$$

Proof. The proof is similar to the proof of Lemma 1, the only difference is that the different kernels and bandwidths in each term have to be considered.

Specifically, using $1, 2$ to replace s_1, s_2 in the derivation, we get for the analogue of (A.3),

$$\begin{aligned} & E\left(\hat{f}'_{(K_1,h_1)}(x_i)\hat{f}'_{(K_2,h_2)}(x_i)^T y_i^2\right) \qquad (A.9) \\ = \; & \left(\tfrac{1}{N-1}\right)^2 E\left\{E_{z_i}\left(y_i^2\left[\sum_{j\neq i}\left(\tfrac{1}{h_1}\right)^{k+1} K'_1(\tfrac{x_i-x_j}{h_1})\right]\left[\sum_{j\neq i}\left(\tfrac{1}{h_2}\right)^{k+1} K'_2(\tfrac{x_i-x_j}{h_2})\right]^T\right)\right\} \\ = \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h_1h_2}\right)^{k+1} E\left[y_i^2 E_{z_i}\left(K'_1(\tfrac{x_i-x_j}{h_1})K'_2(\tfrac{x_i-x_j}{h_2})^T I(\ i\neq j)\right)\right] + \\ & \tfrac{N-2}{N-1}\cdot\left(\tfrac{1}{h_1h_2}\right)^{k+1} E\left[y_i^2 E_{z_i}\left(K'_1(\tfrac{x_i-x_{j_1}}{h_1})K'_2(\tfrac{x_i-x_{j_2}}{h_2})^T I(\ i,j_1,j_2 \text{ pairwise distinct})\right)\right] \\ = \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h_1}\right)^{k+1}\frac{1}{h_2} E\left[y_i^2\int K'_1(u)K'_2(u\frac{h_1}{h_2})^T f(x_i-uh)du\right] + \\ & \tfrac{N-2}{N-1}\left(\tfrac{1}{h_1h_2}\right)^{k+1} E\left[E_{z_i}\left(y_iK'_1(\tfrac{x_i-x_{j_1}}{h_1})\right)E_{z_i}\left(y_iK'_2(\tfrac{x_i-x_{j_2}}{h_2})\right)^T I(\ i,j_1,j_2 \text{ pairwise distinct})\right] \\ = \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h_1}\right)^{k+1}\frac{1}{h_2}\left[Ey_i^2 f(x_i)\mu_2(K_1,K_2,h_1/h_2)+o(1)\right] + \\ & \tfrac{N-2}{N-1}\cdot\left[E(f'(x_i)y_i)(f'(x_i)y_i)^T+o(1)\right], \end{aligned}$$

and for the analogue of (A.7)

$$
\begin{aligned}
& E\left(\hat{f}'_{(K_1,h_1)}(x_{i_1})\hat{f}'_{(K_2,h_2)}(x_{i_2})^T y_{i_1}y_{i_2}I(i_1\neq i_2)\right) \qquad (A.10)\\
= \;& \tfrac{N-2}{(N-1)^2}E\left[(-(gf)'(x_i)+C_1)\left(-(gf')(x_i)+C_2\right)^T\right]+\\
& \tfrac{1}{(N-1)^2}\left(\tfrac{1}{h_1}\right)^{k+1}\frac{1}{h_2}E\left[\int K_1'(u)K_2'(-u\frac{h_1}{h_2})^T du(gf)(x_i)y_i+Dy_i\right]\\
& \tfrac{N-2}{(N-1)^2}E\left[(-(gf)'(x_i)+C_1)\left(f'(x_{i_1})y_i+A_2y_i\right)^T\right]+\\
& \tfrac{N-2}{(N-1)^2}E\left[(f'(x_i)y_i+A_1y_i)\left(-(gf)'(x_i)+C_2\right)^T\right]\\
& \tfrac{(N-2)(N-3)}{(N-1)^2}E\left[f'(x_i)y_i+A_1y_i\right]E\left[f'(x_i)y_i+A_2y_i\right]^T,
\end{aligned}
$$

where the terms denoted in capital letters are similar to the ones in Lemma 1 and are subscripted by the corresponding kernel/bandwidth number; they similarly contribute only to the relatively lower order terms of the expectation.

Combining we get

$$
\begin{aligned}
\Gamma_{s_1s_2} = \;& \tfrac{4}{N(N-1)}\left(\tfrac{1}{h_{s_2}}\right)^{k+1}\tfrac{1}{h_{s_1}}\left(E\left[y_i^2f(x_i)\mu_2(K_{s_1},K_{s_2},h_{s_1}/h_{s_2})\right]+o(1)\right)\\
& +4\tfrac{1}{N(N-1)}\left(\tfrac{1}{h_{s_2}}\right)^{k+1}\tfrac{1}{h_{s_1}}\left(E\left[\mu_2^*(K_{s_1},K_{s_2},h_{s_1}/h_{s_2},x_i)(gf)(x_i)y_i\right]+o(1)\right)\\
& +4\tfrac{N-2}{N(N-1)}\left(E((f'(x_i)y_i)(f'(x_i)y_i)^T)+o(1)\right)\\
& +4\tfrac{N-2}{N(N-1)}\left(E\left[(-(gf)'(x_i))\left(-(gf)'(x_i)\right)^T\right]+o(1)\right)\\
& +4\tfrac{N-2}{N(N-1)}\left(E\left[(-(gf)'(x_i))\left(f'(x_i)y_i\right)^T\right]+o(1)\right)\\
& +4\tfrac{N-2}{N(N-1)}\left(E\left[(f'(x_i)y_i)\left(-(gf)'(x_i)\right)^T\right]+o(1)\right)\\
& +\tfrac{-4N+6}{N(N-1)}E\left(\hat{\delta}_N(K_{s_1},h_{s_1})\right)E\left(\hat{\delta}_N(K_{s_2},h_{s_2})\right)^T.
\end{aligned}
$$

by comparing the orders of the terms the result follows. ■

Proof. **Theorem 2.** The limit covariances are provided by Lemma 2; the covariances can only converge at different rates if the bandwidths converge at different rates.

The expression for the covariance can also be written by interchanging $s_1$and s_2. Thus

without any loss of generality we can assume that $h_{s_1} = o(h_{s_2})$. Note that then

$$\begin{aligned}\mu_2 &= \int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})du \\ &= K'_{s_2}(0)\int K'_{s_1}(u)du + \int K''_{s_2}(\tilde{u})K'_{s_1}(u)udu \cdot \frac{h_{s_1}}{h_{s_2}} \\ &= \frac{h_{s_1}}{h_{s_2}}O(1)\end{aligned}$$

where $\tilde{u}$ lies between 0 and u. Similarly $\mu_2^* = \frac{h_{s_1}}{h_{s_2}}O(1)$. Only two cases of different rates are possible here: (a) a parametric rate for s_1 and a non-parametric for s_2, and (b) non-parametric (different) rates for both.

Consider case (a): $Nh_{s_1}^{k+2} \to 0; Nh_{s_2}^{k+2} \to \infty$. Then

$$\begin{aligned}Cov(Nh_{s_1}^{\frac{k+2}{2}}\hat{\delta}_N(K_{s_1},h_{s_1}),\sqrt{N}\hat{\delta}_N(K_{s_2},h_{s_2})) &= N^{\frac{3}{2}}h_{s_1}^{\frac{k+2}{2}}[N^{-2}h_{s_2}^{-(k+2)}O(1)+N^{-1}O(1)] \\ &= O(N^{\frac{1}{2}}h_{s_1}^{\frac{k+2}{2}}N^{-1}h_{s_2}^{-(k+2)})+O(N^{\frac{1}{2}}h_{s_1}^{\frac{k+2}{2}}) = o(1).\end{aligned}$$

For case (b): $Nh_{s_1}^{k+2} \to 0; Nh_{s_2}^{k+2} \to 0$ we get

$$\begin{aligned}Cov(Nh_{s_1}^{\frac{k+2}{2}}\hat{\delta}_N(K_{s_1},h_{s_1}),Nh_{s_2}^{\frac{k+2}{2}}\hat{\delta}_N(K_{s_2},h_{s_2})) &= N^2h_{s_1}^{\frac{k+2}{2}}h_{s_2}^{\frac{k+2}{2}}[N^{-2}h_{s_2}^{-(k+2)}O(1)+N^{-1}O(1)] \\ &= O(h_{s_1}^{\frac{k+2}{2}}h_{s_2}^{-\frac{k+2}{2}})+O(Nh_{s_2}^{k+2}) = o(1).\end{aligned}$$

■

References

[1] Donkers, B. and M. Schafgans (2005): "A method of moments estimator for semi-parametric index models," Sticerd Discussion Paper No. EM/05/493, London School of Economics.

[2] Fan, J. (1992): "Design-adaptive nonparametric regression," *Journal of the American Statistical Association*, **87**, 998-1004.

[3] Fan, J. (1993): "Local linear regression smoothers and their minimax efficiencies," *The Annals of Statistics*, **21**, 196-216.

[4] Fan, J. and I. Gijbels (1992), "Variable bandwidth and local linear regression smoothers," *The Annals of Statistics*, **20**, 2008-2036.

[5] Kotlyarova, Y. and V. Zinde-Walsh (2007): "Robust kernel estimator for densities of unknown smoothness," *Journal of Nonparametric Statistics*, forthcoming.

[6] Kotlyarova, Y. and V. Zinde-Walsh (2006): "Non- and semi-parametric estimation in models with unknown smoothness," *Economics Letters*, **93**, 379-386.

[7] Hansen, B.E. (2005): "Exact mean integrated squared error of higher order kernel estimators", *Econometric Theory*, **21**, 1031–1057.

[8] 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.

[9] 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.

[10] Li, Q., X. Lu and A. Ullah (2003): "Multivariate local polynomial regression for estimating average derivatives", *Nonparametric Statistics*, **15**, 607–624.

[11] Marron, J.S. and M.P. Wand (1992): "Exact mean integrated squared error," *Annals of Statistics*, **20**, 712–736.

[12] *Mathematicheskaya Encyclopedia. English.*, ed. M. Hazewinkel (1988). Encyclopaedia of mathematics: an updated an annotated translation of the Soviet Mathematicheskaya Encyclopedia, Kluwer Academic Publishers.

[13] Newey, W.K. and T.M. Stoker (1993): "Efficiency of weighted average derivative estimators and index models," *Econometrica*, **61**, 1199–1223.

[14] Nichiyama, Y. and P.M. Robinson (2005): "The bootstrap and the edgeworth correction for semiparametric averaged derivatives," *Econometrica*, **73**, 903–948.

[15] Powell, J.L., J.H. Stock, and T.M. Stoker (1989): "Semiparametric estimation of weighted average derivatives," *Econometrica*, **57**, 1403–1430.

[16] Powell, J.L. and T.M. Stoker (1996): "Optimal bandwidth choice for density-weighted averages," *Journal of Econometrics,* **75**, 291–316.

[17] Robinson, P.M. (1995): "The normal approximation for semiparametric averaged derivatives," *Econometrica*, **63**, 667–680.

[18] Samarov, A.M., 1993, Exploring regression structure using nonparametric functional estimation, *Journal of the American Statistical Association,* **88**, 836-847.

[19] Stoker, T.M., 1991, Equivalence of direct, indirect, and slope estimators of average derivatives, in W.A. Barnett, J. Powell, and G.E. Tauchen, eds., *Nonparametric and Semiparametric Estimation Methods in Econometrics and Statistics*, Cambridge University Press, Cambridge.

[20] Stone, C.J. (1982): "Optimal global rates of convergence for nonparametric regression," *Annals of Statistics*, **10**, 1040–1053.

[21] Yatchew, A. (2003): *Semiparametric regression for the applied econometrician*, Cambridge University Press, Cambridge

Table A.1: RMSE of the Density weighted ADE estimators, N=1000

	Model (s,s)		Model (s,m)		Model (m,m)	
Bandw/Kernel	K_2	K_4	K_2	K_4	K_2	K_4
h_0	0.0156	0.0122	0.0495	0.0476	0.1332	0.1296
h_1	0.0059	0.0054	0.0467	0.0449	0.1464	0.1458
h_2	0.0053	0.0045	0.0500	0.0449	0.1585	0.1548
$h_3 = h^{gcv}$	0.0062	0.0046	0.0544	0.0488	0.1700	0.1680
h_4	0.0088	0.0070	0.0618	0.0578	0.1844	0.1877
Combined						
smallest: h_0	0.0105		0.0474		0.1329	
smallest: h_1	0.0051		0.0470		0.1502	
	Model (s,c)		Model (s,d)		Model (c,d)	
Bandw/Kernel	K_2	K_4	K_2	K_4	K_2	K_4
h_0	0.0270	0.0258	0.0690	0.0625	0.0644	0.0754
h_1	0.0255	0.0237	0.0443	0.0386	0.0354	0.0293
h_2	0.0260	0.0240	0.0352	0.0328	0.0302	0.0322
$h_3 = h^{gcv}$	0.0271	0.0248	0.0255	0.0282	0.0257	0.0278
h_4	0.0300	0.0281	0.0176	0.0268	0.0239	0.0271
Combined						
smallest: h_0	0.0245		0.0594		0.0613	
smallest: h_1	0.0246		0.0366		0.0281	

Table A.1: RMSE of the Density weighted ADE estimators, N=2000

	Model (s,s)		Model (s,m)		Model (m,m)	
Bandw/Kernel	K_2	K_4	K_2	K_4	K_2	K_4
h_0	0.0124	0.0087	0.0390	0.0377	0.1193	0.1147
h_1	0.0043	0.0037	0.0414	0.0399	0.1324	0.1418
h_2	0.0037	0.0031	0.0457	0.0428	0.1467	0.1493
$h_3 = h^{gcv}$	0.0045	0.0033	0.0505	0.0452	0.1600	0.1643
h_4	0.0071	0.0060	0.0575	0.0543	0.1745	0.1879
Combined						
smallest: h_0	0.0086		0.0365		0.1165	
smallest: h_1	0.0037		0.0409		0.1388	
	Model (s,c)		Model (s,d)		Model (c,d)	
Bandw/Kernel	K_2	K_4	K_2	K_4	K_2	K_4
h_0	0.0293	0.0239	0.0670	0.0749	0.0661	0.0651
h_1	0.0245	0.0233	0.0470	0.0431	0.0405	0.0329
h_2	0.0252	0.0232	0.0407	0.0399	0.0328	0.0353
$h_3 = h^{gcv}$	0.0260	0.0240	0.0305	0.0337	0.0254	0.0296
h_4	0.0284	0.0271	0.0195	0.0274	0.0217	0.0253
Combined						
smallest: h_0	0.0244		0.0676		0.0617	
smallest: h_1	0.0238		0.0422		0.0337	

Table A.2: Comparison RMSE of the Density weighted ADE estimators

	$RMSE(comb, h_0)/$ $RMSE(K_2, h_4)$		$RMSE(comb, h_1)/$ $RMSE(K_2, h_4)$	
Model	$N = 1000$	$N = 2000$	$N = 1000$	$N = 2000$
(s,s)	2.333	2.774	1.133	1.194
(s,m)	1.056	0.853	1.047	0.956
(m,m)	0.859	0.780	0.970	0.930
(s,c)	1.021	1.052	1.025	1.026
(s,d)	1.811	1.694	1.116	1.058
(c,d)	1.904	1.748	0.873	0.955

Table A.3: Tobit Model: RMSE Single Index parameter estimates

		Model (s,s)			Model (m,m)			Model (c,d)		
		Unit	PSS	Polar	Unit	PSS	Polar	Unit	PSS	Polar
					$N = 1000$					
Parametric: MLE		0.052	0.026	0.026	0.094	0.047	0.047	0.056	0.028	0.028
Nonparametric: ADE										
K_2	h_0	0.698	0.227	0.204	3.081	0.472	0.520	5.157	0.427	0.343
	h_1	0.140	0.066	0.066	1.042	0.242	0.309	8.928	0.304	0.363
	h_2	0.106	0.052	0.052	0.455	0.180	0.175	0.768	0.254	0.246
	$h_3 \equiv h^{gcv}$	0.094	0.047	0.047	0.340	0.156	0.153	0.656	0.231	0.224
	h_4	0.099	0.050	0.050	0.339	0.161	0.158	0.725	0.253	0.246
K_4	h_0	0.372	0.137	0.133	2.481	0.342	0.434	2.348	0.534	0.458
	h_1	0.119	0.057	0.057	0.737	0.201	0.192	0.660	0.229	0.218
	h_2	0.095	0.047	0.046	0.422	0.173	0.169	0.566	0.192	0.185
	$h_3 \equiv h^{gcv}$	0.090	0.045	0.044	0.418	0.188	0.183	0.738	0.252	0.244
	h_4	0.107	0.053	0.053	0.376	0.174	0.170	1.125	0.345	0.330
Combined										
	smallest: h_0	0.221	0.103	0.101	2.233	0.425	0.376	1.481	0.442	0.391
	smallest: h_1	0.103	0.051	0.050	0.529	0.196	0.189	0.771	0.260	0.252

Table A.3: Tobit Model: RMSE Single Index parameter estimates (Cont'd)

		Model (s,s)			Model (m,m)			Model (c,d)		
		Unit	PSS	Polar	Unit	PSS	Polar	Unit	PSS	Polar
					$N = 2000$					
Parametric: MLE		0.035	0.018	0.018	0.063	0.032	0.032	0.040	0.020	0.020
Nonparametric: ADE										
K_2	h_0	0.384	0.148	0.144	10.47	0.409	0.400	3.989	0.412	0.367
	h_1	0.089	0.043	0.042	0.452	0.190	0.185	1.115	0.324	0.310
	h_2	0.062	0.031	0.031	0.317	0.150	0.148	0.906	0.293	0.282
	$h_3 \equiv h^{gcv}$	0.059	0.030	0.030	0.270	0.131	0.128	0.708	0.248	0.241
	h_4	0.067	0.034	0.034	0.277	0.134	0.131	0.640	0.228	0.222
K_4	h_0	0.212	0.093	0.092	2.064	0.263	0.357	5.263	0.400	0.427
	h_1	0.072	0.035	0.035	0.427	0.151	0.146	0.727	0.238	0.229
	h_2	0.060	0.030	0.030	0.299	0.127	0.125	0.646	0.214	0.207
	$h_3 \equiv h^{gcv}$	0.056	0.028	0.028	0.281	0.134	0.132	0.773	0.266	0.258
	h_4	0.072	0.036	0.036	0.281	0.128	0.126	1.037	0.327	0.314
Combined										
	smallest: h_0	0.170	0.079	0.079	5.743	0.269	0.333	4.927	0.379	0.351
	smallest: h_1	0.064	0.032	0.033	0.382	0.158	0.154	0.773	0.287	0.277