Smoothness Adaptive Average Derivative Estimation[±]

Marcia M.A. Schafgans[*] Victoria Zinde-Walshyz[†]

Discussion paper
No. EM/2008/529
August 2008

The Suntory Centre
Suntory and Toyota International Centres for Economics and Related Disciplines
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
Tel: 020 7955 6679

[±] The authors would like to thank an anonymous referee and Richard Smith for their comments and suggestions.

[*] 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 recherché sur la société et la culture (FRQSC) .

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. 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 trace of the part of the estimated asymptotic mean squared error that depends on the bandwidth. Estimation of the components of the AMSE, of the optimal bandwidths, 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.

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 a 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 (weighted) average derivatives of the conditional mean 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, hereafter referred to as ADEs, 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 estimating the coefficients in the single index model to the presence of discrete covariates; Donkers and Schafgans (2008) extended the ADE approach to multiple index models; Chaudhuri et al. (1997) 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).

To formulate the ADE under consideration in our paper, let $g(x) = E(y|x)$ with $y \in R$ and $x \in R^k$, and define

$$\delta_0 = E(f(x)g'(x)), \tag{1}$$

with $g'(x)$ the derivative of the unknown conditional mean function and $f(x)$ the density of x. Recognizing that $\delta_0 = -2E(f'(x)y)$ under certain regularity conditions, Powell, Stock

and Stoker (1989), hereafter referred to as PSS, introduced the estimator

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

(which is the sample analogue of $-2E(f'(x)y)$ with

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

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

In all of the literature on ADE, asymptotic theory was provided for parametric rates of convergence. Even though the estimators are based on nonparametric kernel estimators which depend on the kernel and bandwidth and converge 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, $f(x)$. This assumption is not based on any a priori theoretical considerations, also, there is no consistent statistical procedure that can verify assumptions of distribution smoothness (Lalley and Nobel, 2003). Izenman and Sommer (1988), for example, show various multimodal distributions that are encountered in biomedical and statistical studies; 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. Even when there is sufficient smoothness for parametric rates the choice of bandwidth and kernel affects second-order terms which are often not much smaller than first-order terms (see, e.g. Dalalyan et al., 2006).

Our concern with the possible violation of assumed high degree of density smoothness led us to extend the existing asymptotic results for ADE by relaxing the smoothness assumptions on the density. We show that insufficient smoothness will result in possible

asymptotic bias and may easily lead to non-parametric rates. Since selection of optimal kernel (order) and bandwidth (Powell and Stoker, 1996) presumes the knowledge of the degree of density smoothness, uncertainty about the degree of density smoothness poses an additional concern.

In principle, smoothness properties of density $f(x)$ could differ for different components of the vector $x = (x_1, ...x_k)^T$ which could lead to possibly different rates for the component bandwidths, h_ℓ, $\ell = 1, .., k$, (e.g., Li and Racine, 2007). Even when all the rates are the same it may be advantageous to use different bandwidths in finite sample. To make explicit the use of different bandwidths we specify a bandwidth vector $h = (h_1, ...h_k)^T$.The kernel smoothing function is constructed from the product of univariate kernel functions, where $K(\frac{x_i - x_j}{h}) = K_1(\frac{x_{i1} - x_{j1}}{h_1}) \times ... \times K_k\left(\frac{x_{ik} - x_{jk}}{h_k}\right)$. We assume for simplicity that all the univariate kernels in the product kernel are the same while allowing for possibly different bandwidths. Denote by $\mathbf{h^{-1}}$ the diagonal matrix

$$\mathbf{h^{-1}} = diag(h_1^{-1}, ...h_k^{-1}); \tag{3}$$

for equal bandwidths $\mathbf{h^{-1}}$ can be read as the scalar h^{-1}. The vector $\frac{\partial}{\partial x}K(x)$ is denoted $K'(x)$; then $\frac{\partial}{\partial x}K(\frac{x - x_j}{h}) = \mathbf{h^{-1}}K'(\frac{x - x_j}{h})$. The ADE $\hat{\delta}_N(K, h)$ is given by

$$\hat{\delta}_N(K, h) = -\frac{2}{N(N-1)\Pi_{\ell=1}^k h_\ell} \sum_{j \neq i}^{N} \mathbf{h^{-1}} K'(\frac{x_i - x_j}{h}).$$

Insufficient smoothness affects the asymptotic variance and bias of the ADE. We show that if the degree of smoothness is known an optimal asymptotic rate of the bandwidth can still be derived. With an unknown degree of smoothness we derive consistent estimators of the rate of the bias, the bias and optimal bandwidth rate for a given kernel by exploiting properties of oversmoothed and undersmoothed estimators. Without knowledge of the form of the density optimal choice between different kernels (e.g. lower order versus higher order) may not be possible. The importance of kernel choice was investigated by Hansen (2005) for density estimation; he showed the order of the kernel to have a large impact on its finite-sample MISE. Kotlyarova, Zinde-Walsh (2007) also obtained highly varying results

for density estimators based on kernels of different orders, or even the same order, but different shapes (including asymmetric).

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. We consider linear combinations of the ADE estimators for different kernel/bandwidth pairs (K_s, h_s), $s = 1, ...S$. The estimator $\hat{\delta}^*_{N,comb}$ that results when we minimize the trace of estimated asymptotic MSE is the combined estimator (following Kotlyarova and Zinde-Walsh, 2006). In their paper, hereafter referred to as KZW, the combined estimator is shown to provide the best rate available among all the rates of estimators in the linear combination even without knowledge of which is the best estimator. Combining estimators to achieve adaptive property was recently investigated in statistical literature (e.g., Yang, 2000, Juditsky and Nemirovski, 2000).

To ensure rate efficiency the bandwidth that provides the best rate estimator should be approached by some sequence of h_s as $N \to \infty$, e.g. Horowitz and Spokoiny (2001) achieve that by expanding the set of dyadic bandwidths with sample size. Here we provide a different approach by consistently estimating the optimal bandwidth rate for a given kernel and including the corresponding estimator in the linear combination. We show that with this choice of the set of bandwidths the combined estimator achieves second-order rate efficiency in the parametric case and first order rate efficiency when parametric rate cannot be achieved. The weights are not restricted to be nonnegative giving the combined estimator the "jackknife" property that asymptotically eliminates leading terms in the bias and thus automatically bias-corrects.[1]

Our finite sample investigation uses 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). There, we demonstrate that there is no clear guidance on the choice of suitable kernel bandwidth pair. Even though in these cases the smoothness assumptions hold, the

[1] For references on the jackknife see Miller (1974) and Schucany, Gray and Owen (1971).

high modal nature of these mixture distributions leads to large partial derivatives that undermine the performance of ADE. At the same time, the combined estimator provides reliable results in all cases.

The paper is organized as follows. In section 2 we provide the assumptions, where we relax the usual high smoothness assumptions common in the literature. In section 3 we derive the asymptotic properties of the density-weighted ADE under various assumptions about density smoothness and joint asymptotics for ADE estimators based on different bandwidth kernels pairs, then we develop the combined estimator and prove its adaptive property. Section 4 provides the Monte Carlo study results and Section 5 concludes.

2. ASSUMPTIONS

The assumptions here keep some conditions common in the literature on ADE but relax the usual higher smoothness assumptions.

The first two assumptions are similar to PSS; they restrict x to be random variables that are continuously distributed with no component of x functionally determined by other components of x (y could be discrete, e.g. a binary variable) and impose the minimal smoothness assumption of continuous differentiability on f and g.

Assumption 1. *Let $z_i = (y_i, x_i^T)^T$, $i = 1, .., N$ be a random sample drawn from a distribution that is absolutely continuous in x. The support Ω of the density of x, $f(x)$, is a convex (possibly unbounded) subset of R^k with nonempty interior Ω_0.*

Assumption 2. *The density function $f(x)$ is continuous over 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 the conditional mean function $g(x)$ 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 detailed differentiability conditions subsequently 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 $0 < \alpha_0 \leq 1$:

$$\left|(fg)'(x + \Delta x) - (fg)'(x)\right| \leq \omega_{(fg)'}(x) \left\|\Delta x\right\|^{\alpha_0}$$

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

The kernel K satisfies a standard assumption.

Assumption 4. *(a) The kernel smoothing function $K(u)$ is a symmetric[2] 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 $v(K)$.

Density smoothness plays a role in controlling the rate for the bias of the PSS estimator; the bias is

$$Bias\hat{\delta}_N(K, h) = E(\hat{\delta}_N(K, h) - \delta_0) = -2Ey \int \left[f'(x - hu) - f'(x)\right] K(u)du. \qquad (4)$$

We formalize the degree of density smoothness in terms of the Hölder space of functions, $C_{m+\alpha}(\Omega)$ for integer m, $0 < \alpha \leq 1$. Any $f \in C_{m+\alpha}(\Omega)$ is m times continuous differentiable with all the m^{th} order derivatives satisfying Hölder's condition of order α :

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

[2]In Schafgans Zinde-Walsh (2007) we discusses the possibility of non-symmetric kernels and derive results for that case.

Assumption 5. *$f \in C_{m+\alpha}(\Omega)$ with $m \geq 1, 0 < \alpha \leq 1$ and $E(\omega^2_{f^{(m)}}[1 + |y|^2 + \|x\|]) < \infty$.*

Note that in the case $m - 1 = 0$ there may be no more than Hölder continuity of some partial derivative without further differentiability significantly relaxing the usual assumptions in the literature. Assumption (5) implies that each component of the derivative of density $f'(x)$ belongs to $C_{m-1+\alpha}(\Omega)$; denote $m - 1 + \alpha$ by v; we refer to v as the degree of smoothness.

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 $v(K)-$th order Taylor expansion of $f'(x_i - uh)$ around $f'(x_i)$). If differentiability conditions typically assumed do not hold, then the bias does not vanish sufficiently fast even for bandwidths such that $Nh^{2v(K)} = o(1)$. All we can state is the rate $O(\|h\|^{\bar{v}})$ for the bias in (4).

It is possible that some components of the derivative vector $f'(x)$ are smoother than others, then $f'(x)_\ell \in C_{m_\ell - 1 + \alpha_\ell}(\Omega)$ with $v_\ell = m_\ell - 1 + \alpha_\ell > v$. Define the vector $\bar{v} = (\bar{v}_1, .., \bar{v}_k)$, where $\bar{v}_\ell = \min(v_\ell, v(K)), \ell = 1, ...k$. We introduce a diagonal matrix

$$\mathbf{h}^{\bar{\mathbf{v}}} = diag(h_1^{\bar{v}_1}, h_2^{\bar{v}_2}, ...h_k^{\bar{v}_k}) \tag{5}$$

and denote by $\mathbf{h}^{-\bar{\mathbf{v}}}$ the inverse of the matrix in (5). We strengthen Assumption 5 by requiring exact rates for the individual components of the bias vector, $Bias(\hat{\delta}_N(K, h))$.

Assumption 6. *(a) As $N \to \infty, \|h\| \to 0$*

$$\boldsymbol{h}^{-\bar{\mathbf{v}}} Bias(\hat{\delta}_N(K, h)) \to \mathcal{B}(K), \tag{6}$$

where the vector $\mathcal{B}(K) = (\mathcal{B}_1(K), ...\mathcal{B}_k(K))'$ is such that $0 < |\mathcal{B}_l(K)| < \infty$, $l = 1, ..., k$;

(b) additionally,

$$Bias(\hat{\delta}_N(K, h)) = \mathbf{h}^{\bar{\mathbf{v}}} \left(\mathcal{B}(K) + o(\mathbf{h}^{\boldsymbol{\gamma}})\right) \tag{7}$$

for some $\gamma > 0$.

(c) $\bar{v}_\ell = \bar{v} = const.$

When the rates of the bandwidths differ, the leading bias term for each component of the derivative is determined by the corresponding component of the bandwidth vector. Part (b) provides a more specific description for the convergence for the bias; it will be used in deriving the adaptive estimator but is not needed for expressing the MSE. Part (c) assumes the same smoothness for the different derivatives. When all the bandwidths are the same and $\bar{v}_\ell$ is constant for all components, the matrix $\mathbf{h}^{\bar{v}}$ (and $\mathbf{h}^{-\bar{v}}$) in Assumption 6 can be read as a scalar.

3. Main Results

We extend the existing asymptotic results for ADE by relaxing the smoothness assumptions on the density and obtain optimal bandwidth rates. When the degree of smoothness is not known we propose a new combined estimator that is smoothness adaptive.

3.1. Asymptotic results for estimators based on a specific kernel and bandwidth vector. Consider the ADE (2) under the Assumptions 1-6(a) of the previous section for all possible degrees of smoothness and kernel orders (for $\bar{v} = \min(v, v(K))$. The variance is derived in the appendix, and for notation purposes we recall (3): that $\mathbf{h^{-1}}$ is a diagonal matrix. Both $\mathbf{h^{\bar{v}}}$ and $\mathbf{h^{-1}}$ can be read as scalars in the equal bandwidth setting (and Assumption 6(c)). The product of bandwidths $\Pi_{\ell=1}^{k} h_\ell$ is denoted by the scalar $\left(h^k\right)$. The expression for the variance, based on result (A.5) derived in the Appendix, is given by

$$Var(\hat{\delta}_N(K,h)) = N^{-2}\left(h^k\right)^{-1}\mathbf{h}^{-1}\left(\Sigma_1(K)+o(1)\right)\mathbf{h}^{-1} + \left(\Sigma_2+o(1)\right)N^{-1}, \tag{8}$$

where $\Sigma_1(K)$ depends on the kernel, but Σ_2 does not (defined in (A.7)). The bias is given by Assumption 6(a), thus

$$\begin{aligned} &MSE(\hat{\delta}_N(K,h)) \\ &= \left(N^{-2}h^{-k}\right)\mathbf{h^{-1}}\left(\Sigma_1(K)+o(1)\right)\mathbf{h}^{-1} + \left(\Sigma_2+o(1)\right)N^{-1} + \mathbf{h^{\bar{v}}}\left(\mathcal{B}(K)\mathcal{B}^T(K)+o(1)\right)\mathbf{h^{\bar{v}}}. \end{aligned} \tag{9}$$

The following Theorem summarizes all the possible convergence rates and limit features of the ADE, $\hat{\delta}_N(K,h)$, for different choices of bandwidth and kernel and gives the optimal bandwidth rate.

Theorem 1. *Under Assumptions 1–6(a)*

(a) *If the density is sufficiently smooth and the order of kernel is sufficiently high: all* $\bar{v}_\ell > \frac{k+2}{2}, \ell = 1, ..., k$, *the rate* $O(N^{-1})$ *for the MSE and the parametric rate* $\sqrt{N}$ *for the ADE can be achieved for a range of bandwidths* $H(\sqrt{N}) = \left\{ h : \left[N^2(h^k)\mathbf{h^2}\right]^{-1} = O(1); N\mathbf{h^{2\bar{v}}} = O(1)\right\}$. *Outside this range when* $N^2\left(h^k\right)\mathbf{h^2} \to 0$ *the asymptotic variance depends on the kernel, if* $N\mathbf{h^{2\bar{v}}} \to \infty$ *asymptotic bias dominates.*

(b) *If the density is not smooth enough or the order of the kernel is too low: some* $\bar{v}_\ell < \frac{k+2}{2}$ *the parametric rate cannot be obtained. The asymptotic variance depends on the kernel. Depending on* $\bar{v}$ *and bandwidth/kernel pair* (K, h), *a diagonal matrix of rates* $\mathbf{r}_N$, $\mathbf{r}_N \to \infty$, *such that* $\mathbf{r}_N\left(\hat{\delta}_N(K,h) - \delta_0\right)$ *has finite first and second moments obtains. If* $\mathbf{r}_N\mathbf{h}^{\bar{v}} \to 0$, *ADE has no asymptotic bias at the rate of convergence* $\mathbf{r}_N$.

(c) *The optimal bandwidth can be obtained by minimizing trace of* $MSE(\hat{\delta}_N(K,h)$; *for bandwidths with the same rates and* $\bar{v}_l = \bar{v}$ *(Assumption 6(c)) this provides*

$$h^{opt} = O(N^{-2/(2\bar{v}+k+2)}). \tag{10}$$

Proof. See Appendix.

The Theorem provides a full description of the asymptotic behavior of the moments of the estimator allowing for different bandwidth rates for different components. Different rates for bandwidth components may result in some components of the bias vanishing faster than others, thus having asymptotically no impact on $trMSE$. When the $trMSE$ criterion is used, typically the same rate for bandwidth components is used (though not necessarily identical bandwidths in finite samples). For equal (rate) bandwidths under 6(c) we get the usual scalar conditions $N^2h^{k+2} \to 0$; $Nh^{2\bar{v}} \to \infty$. Then the Powell, Stock and Stoker (1989) results with the parametric rate holds for sufficiently smooth $f(x)$ (permitting $\bar{v} > (k+2)/2$) with $h \to 0$ and $Nh^{k+2} \to \infty$. In the absence of the high degree of differentiability the asymptotic variance (as the asymptotic bias) does depend on the weighting used in the local averaging – involves $\Sigma_1(K)-$ yielding a non-parametric rate. Selection of the optimal

bandwidth and kernel (order) that minimize the mean squared error critically depends on our knowledge of the degree of smoothness of the density.

3.2. The combined estimator. We construct an estimator that achieves rate efficiency by optimally combining PSS estimators even when the degree of smoothness is not known. Consider ADE for various kernel/bandwidth pairs (K_s, h_s), $s = 1, ...S$, and consider the linear combination:

$$\hat{\delta}_N^* = \sum_{s=1}^{S} a_s \hat{\delta}_N(K_s, h_s) \text{ with } \sum_{s=1}^{S} a_s = 1. \tag{11}$$

The estimator that results when we choose the weights so as to minimize the trace of estimated asymptotic MSE of $\hat{\delta}_N^*$ is the combined estimator, $\hat{\delta}_{N,comb}^*$ (similarly to KZW). The weights a_s, $s = 1, ..., S$ sum to one and are not restricted to be non-negative to allow bias trade-offs.

To obtain the MSE of the linear combination in (11), we need to establish the first and second moments for the stacked vector $\mathbf{r_{Ns}} \left(\hat{\delta}_N(K_s, h_s) - \delta_0 \right)$, $s = 1, ...S$, with $\mathbf{r_{Ns}}$ the diagonal matrix of rates associated with kernel/bandwidth pair (K_s, h_s). First moments are given by Assumption 6. Next, we derive the limit covariances between the estimators.

Theorem 2. *Under Assumptions 1-6(a), the limit covariance matrix for the vector with components* $\mathbf{r_{Ns}} \left(\hat{\delta}_N(K_s, h_{Ns}) - \delta_0 \right)$, $s = 1, ..., S$, *has* $k \times k$ *blocks*

$$\Gamma_{s_1 s_2} = N^{-2} \left(h_{s_2}^k \right)^{-1} \mathbf{h}_{s_1}^{-1} \left(\Sigma_1(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) + o(1) \right) \mathbf{h}_{s_2}^{-1} + \left(\Sigma_2 + o(1) \right) N^{-1} \tag{12}$$

$s_1, s_2 = 1, ..., S$, *with* $\Sigma_1(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2})$ *defined in (A.6). Covariance (matrix) between estimators converging at different rates goes to zero. When* $s_1 = s_2$ *this provides the variance in (8).*

Proof. See Appendix.

Note that the fact that some estimators have zero covariances in the limit indicates that they provide complimentary information.

The trace of the asymptotic MSE of $\hat{\delta}_N^*$, used to define the combined estimator, can then be written as

$$trAMSE(\hat{\delta}_N^*) = \sum_{s_1,s_2=1}^{S} a_{s_1} a_{s_2} (\tilde{\mathcal{B}}_{s_1}^T \tilde{\mathcal{B}}_{s_2} + tr\tilde{\Gamma}_{s_1 s_2}) = a' D a, \tag{13}$$
$$\text{where } \{D\}_{s_1 s_2} = \mathcal{B}_{s_1}^T \mathcal{B}_{s_2} + tr\Gamma_{s_1 s_2},$$

$\tilde{\mathcal{B}}_s = \mathbf{r}_{\mathbf{Ns}}^{-\mathbf{1}} \mathcal{B}_s$, for $s = s_1, s_2$, and $\tilde{\Gamma}_{s_1 s_2} = \mathbf{r}_{\mathbf{Ns_1}}^{-\mathbf{1}} \mathbf{r}_{\mathbf{Ns_2}}^{-\mathbf{1}} \Gamma_{s_1 s_2}$. The weight vector a, needs to be chosen so as to minimize the $trAMSE(\hat{\delta}_N^*)$ subject to the restriction that $a'\iota = 1$ where ι is a $(S \times 1)$ vector of ones. The linear combination that minimizes (13) provides an estimator that converges at a rate no worse than that of the AMSE for the fastest converging individual estimator (see KZW). Moreover, this result holds whether we use the (true) AMSE or the AMSE based on consistent estimators of biases and variances to provide a feasible combined estimator (as shown in Theorem 3 in KZW).

3.3. The adaptive property of the combined estimator. Here we construct consistent estimators of variances, covariances, biases and the optimal bandwidth rate to arrive at a combined estimator that is second-order rate efficient in the parametric convergence case and is rate efficient when the parametric rate does not obtain.

A consistent estimate for the asymptotic covariances (and variances) that does not rely on the degree of smoothness is given by the bootstrap (see Samarov, 1993). It is obtained as

$$\widehat{\tilde{\Gamma}_{s_1,s_2}} = \frac{1}{B} \sum_{b=1}^{B} \left(\hat{\delta}_{b,N}(K_{s_1}, h_{s_1}) - \hat{\delta}_N(K_{s_1}, h_{s_1}) \right) \left(\left(\hat{\delta}_{b,N}(K_{s_2}, h_{s_2}) - \hat{\delta}_N(K_{s_2}, h_{s_2}) \right) \right)^T, \tag{14}$$

where for each of the B bootstrapped samples estimates $\hat{\delta}_{b,N}(K_s, h_s)$ are obtained for $s = 1, .., S$.

Estimation of the bias is more difficult; Theorem 3 below provides the technical details.

First, we construct a sequence of bandwidths $\{h_t\}_{t=1}^H$ for which the corresponding estimators are asymptotically biased (oversmoothed). One such bandwidth is given by the usual cross-validation for nonparametric regression, with $h^{gcv} = cN^{-1/(2\bar{v}+2)}$ (see Stone,

1982) for $\bar{v} = \min(v, v(K))$, c some positive constant. By differencing these estimators for different bandwidth rates we obtain a consistent estimator for $\bar{v}$, $\widehat{\bar{v}}$.

Next, using a pair of estimators $\hat{\delta}_N$, one of which is based on an oversmoothed bandwidth, and the other on a somewhat undersmoothed one we consistently estimate the bias for the oversmoothed estimator, and then use the estimated $\bar{v}$ to construct consistent estimators of all biases since for different bandwidths the leading terms of the bias differ by the ratio of bandwidths to power $\bar{v}$.

Finally, we construct an adaptive combined estimator.

Theorem 3. *Under Assumptions 1-6*

(a) Consider a sequence of bandwidth vectors $\{h_t\}_{t=1}^{H}$, *such that* $h_t = c_t h_{gcv} N^{\gamma_t}$ *for some positive constants* c_t *with* $0 \leq \gamma_1 < ... < \gamma_H < \frac{1}{2v(K)+k}$. *For any* Q *pairs of distinct bandwidths:* $2 \leq Q \leq \frac{H(H+1)}{2}$ *a consistent estimator for* $\bar{v}, \widehat{\bar{v}}$, *is given by*

$$\widehat{\bar{v}} = \frac{\sum_{t>t'=1}^{Q} \ln\left|\hat{\delta}_N(K,h_t)_\ell - \hat{\delta}_N(K,h_{t'})_\ell\right| \cdot \left(\ln h_{t\ell} - \frac{1}{Q}\Sigma_q \ln h_{q\ell}\right)}{\sum_{t=1}^{Q}\left(\ln h_{t\ell} - \frac{1}{Q}\Sigma_q \ln h_{q\ell}\right)^2}, \tag{15}$$

for any $\ell = 1, ..., k$ *Given* $\widehat{\bar{v}}$ *the optimal rate for the bandwidth is consistently estimated by* $\widehat{h^{opt}} = O(N^{-2/(2\widehat{\bar{v}}+k+2)})$.

(b) Given bandwidths $h_o = \widehat{h^{opt}} N^{\zeta}$, *with* $0 < \zeta < \frac{2}{2\widehat{\bar{v}}+k+2}$, *and* $h_u = \widehat{h^{opt}} N^{-\xi}$, *with* $\xi > \frac{\zeta(k+2)}{2\widehat{\bar{v}}}$; *a consistent estimate for bias* $Bias\hat{\delta}_N(K,h_o)$ *is provided by* $\widehat{Bias}\hat{\delta}_N(K,h_o) = \hat{\delta}_N(K,h_o) - \hat{\delta}_N(K,h_u)$. *Consistent estimates of* $Bias\hat{\delta}_N(K,h_t)$ *for* h_t *can be obtained as* $\mathbf{h_t^{\widehat{\bar{v}}} h_o^{-\widehat{\bar{v}}}} \left(\hat{\delta}_N(K,h_o) - \hat{\delta}_N(K,h_u)\right)$.

(c) Consider a set of kernel/bandwidth pairs $\{K_s, h_s\}_{s=1}^{S}$ *with* $h_s = \widehat{h^{opt}}$ *for some* s. *The combined estimator based on consistent estimators of covariances and biases, is asymptotically second-order efficient when* $\bar{v} \geq \frac{k+2}{2}$ *and* $h \in H(\sqrt{N})$ *and achieves first-order efficiency otherwise.*

Proof. See Appendix.

Note that all the components of each bandwidth vector in the set $\{h_t\}_{t=1}^{H}$ have the same rate by construction, but using different components in (15) leads to k consistent estimators for $\bar{v}$, which will differ in finite samples.

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. We consider the density weighted average derivative estimate (ADE) of this single-index model defined in (2) which identifies the parameters β "up to scale" without relying on the Gaussianity assumption on ε_i. Under the usual smoothness assumptions, the finite sample properties of the ADE for this Tobit model have been considered in the literature (Nichiyama and Robinson, 2005).

We use two explanatory variables and select $\beta = (1, 1)^T$. We make various assumptions about the distribution of our independent, explanatory variables. The base model uses two standard normal explanatory variables. In the other models various multimodal normal mixtures are considered, which while still being infinitely differentiable, do allow behavior resembling that of nonsmooth densities. In particular, we consider the trimodal normal

mixture used in KZW, $0.5\phi(x+0.767)+3\phi(\frac{x+0.767-0.8}{0.1})+2\phi(\frac{x+0.767-1.2}{0.1})$, and the "double claw" and "discrete comb" mixtures (Marron and Wand, 1992). The models are labelled using two indices (i_1, i_2) representing the distributions used for the two explanatory variables with each index $i. =$ s (standard normal), m (trimodal normal mixture), c (double claw) and d (discrete comb). The sample size is set at $N = 2000$ with 100 replications each.

The multivariate kernel function $K(\cdot)$ (on R^2) is chosen as the product of two univariate kernel functions. We use the quartic second order kernel (see e.g., Yatchew, 2003) and a fourth order kernel[3] 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.

First, we apply the usual cross-validation for nonparametric regression, yielding a bandwidth sequence $h^{gcv} = cN^{-1/(2\bar{v}+2)}$ with $\bar{v} = \min(v, v(K))$, c some positive constant. Even though the rates are the same, for computation of bandwidth vectors in our finite sample experiment we allow for differing bandwidths. We obtain them using a gridsearch.[4]

Next, we estimate $\bar{v}$ using bandwidths that satisfy the conditions of Theorem 3(a). We set $H = 6$, $\gamma_t = \frac{1}{2v(K)+2} \cdot \frac{t-1}{6}$, $t = 1, .., H$. The actual bandwidth sequences $\{c_t h^{gcv} N^{\gamma_t}\}$ are $\{h^{gcv}, 1.01h^{gcv}N^{1/36}, 0.98h^{gcv}N^{2/36}, 0.93h^{gcv}N^{3/36}, 0.86h^{gcv}N^{4/36}, 0.78h^{gcv}N^{5/36}\}$ for second order kernel; $\{h^{gcv}, 1.10h^{gcv}N^{1/60}, 1.16h^{gcv}N^{2/60}, 1.20h^{gcv}N^{3/60}, 1.21h^{gcv}N^{4/60}, 1.19h^{gcv}N^{5/60}\}$ for fourth order kernel. The reason for selecting the c_t's is to ensure a reasonable spread of bandwidths for $N = 2000$ (correspond to bandwidth sequences $\{h^{gcv}, 1.25h^{gcv}, 1.5h^{gcv}, 1.75h^{gcv}, 2.0h^{gcv}, 2.25h^{gcv}\}$). To estimate $\bar{v}$ we select a subset of Q bandwidths in the following way: select a range of consecutive bandwidths where differences (ℓhs of (A.9)) all have the same sign, if that is not possible, we use all $Q = \frac{H(H+1)}{2}$. The estimated $\bar{v}$, $(\widehat{\bar{v}}_1, \widehat{\bar{v}}_2)$,

[3] The fourth order kernel we use is given by $K(x) = \frac{105}{64}\left(-3x^6+7x^4-5x^2+1\right)1(|x| \leq 1)$.

[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. 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 $\binom{1.70}{0.97}$.

for our models on average provided in the (s,s) model $(1.99, 1.98)$ for the second order kernel, K_2, and $(3.68, 3.71)$ for the fourth order kernel, K_4; in the (s,m) model K_2 provided $(1.70, 1.51)$ and $K_4 : (3.16, 2.67)$; the (m,m) model $K_2 : (1.40, 1.37)$ and $K_4 : (1.98, 1.96)$ the (s,c) model $K_2 : (1.94, 1.87)$ and $K_4 : (3.56, 3.21)$; the (s,d) model $K_2 : (1.50, 1.04)$ and $K_4 : (3.36, 1.67)$; and the (c,d) model $K_2 : (1.49, 0.89)$ and $K_4 : (3.14, 1.68)$, which are reasonable. We use $\widehat{\bar{v}}_1$ and $\widehat{\bar{v}}_2$ as estimators of $\bar{v}$ relating them to their respective component in the ADE vector.

In accordance with Theorem 3(b) we choose $h_o = \widehat{h^{opt}} N^{0.05}$ and $h_u = \widehat{h^{opt}} N^{-0.07}$ and obtain $\widehat{Bias}\hat{\delta}(K, h_o)$. For the combined estimator we consider a range of bandwidths $\{h^{opt}N^{-0.04}, h^{opt}, h^{opt}N^{0.03}, h^{opt}N^{0.05}, h^{gcv}\}$ (or $\{0.75h^{opt}, h^{opt}, 1.25h^{opt}, 1.50h^{opt}, h^{gcv}\}$). Covariances are computed by bootstrap using (14); biases according to Theorem 3(b). The weights are then obtained by minimizing the trAMSE constructed according to (13) with estimated biases and covariances subject to the sum of the weights being equal to 1.[5] More weight, including of opposite signs, are given to the higher bandwidths for the second and fourth order kernel.

In Table 1, we report relative error: the ratio of the true finite sample Root Mean Squared Errors (RMSE) to δ_0 for ADE in the different models for the sample size $N = 2000$.

To evaluate the results note that the relative errors for models (s,s), (s,c) are close for all bandwidths and kernels: range is 7.8%-11.3% for (s,s) and 44.4-49,9% for (s,c), so for these cases there is not much sensitivity to the choice of bandwidth/kernel order. There is somewhat more of a dispersion for the (s,m) case: the range is 42.7%-60.7%, but even in this case the price of an incorrect choice (associated with a too large bandwidth) is not that dramatic; here the optimal bandwidth (any kernel) or the combined estimator would be

[5] Ordering the kernel bandwidth pairs, $s = 1, .., 10$ as: $(K_2, h_1), .., (K_2, h_5), (K_4, h_1), ..., (K_4, h_5)$, on average the weights are $(-0.00, -0.03, 0.65, -0.45, -0.07, -0.09, -0.04, -0.23, 2.30, -1.05)$ for the (s,s) model; for (s,m) the weights are $(0.03, -0.01, 0.89, 1.00, -0.77, -0.38, -0.10, -0.22, 1.24, -0.70)$; for (m,m) $(0.02, 0.07, -0.92, 4.01, -2.40, -0.32, 0.14, -0.74, 1.43, -0.28)$; for (s,c) $(0.02, -0.11, 0.74, -0.14, -0.09, -0.14, -0.11, -0.08, 1.96, -1.05)$; for (s,d) $(0.03, 0.11, 0.30, 2.35, -0.60, -0.50, -0.36, -0.52, 1.10, -0.90)$; and for (c,d) $(0.05, 0.09, -0.29, 2.77, -0.85, -0.52, 0.65, -0.87, 1.62, -1.06)$.

Table 1: RMSE of the Density weighted ADE estimators/δ_0, 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 (h_u)	0.234	0.141	0.457	0.427	0.672	0.686
h_1	0.156	0.113	0.471	0.445	0.694	0.743
h_2 (h^{opt})	0.106	0.085	0.500	0.495	0.755	0.811
h_3	0.093	0.078	0.533	0.516	0.811	0.839
h_4 (h_o)	0.096	0.078	0.564	0.519	0.856	0.864
h_5 (h^{gcv})	0.113	0.083	0.607	0.543	0.910	0.934
Combined	0.096		0.561		0.869	
	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 (h_u)	0.499	0.455	1.487	1.538	1.054	1.015
h_1	0.465	0.447	1.319	1.271	0.900	0.785
h_2 (h^{opt})	0.470	0.444	1.168	1.038	0.728	0.632
h_3	0.480	0.447	1.033	0.995	0.613	0.671
h_4 (h_o)	0.486	0.451	0.895	0.925	0.517	0.671
h_5 (h^{gcv})	0.497	0.461	0.766	0.847	0.479	0.575
Combined	0.465		0.872		0.690	

beneficial. More striking consequences of choice are seen for the (m,m) case: the range of relative errors is 68.6%-93.4% (similarly to (s,m) incorrect choice involves oversmoothing); again the optimal bandwidths and combined estimator provide advantages over incorrect choices. The most dramatic cases are (s,d) with range 76.6%-153.8% and (c,d) with 47.9%-105.4% where now incorrect choice is associated with undersmoothing. In these cases the combined estimator gives results much closer to the lower bound than upper bound of the errors, and also presents a better choice than the optimal bandwidth.

We conclude that 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 optimal bandwidth compares favourably with many bandwidths (including cross validation), but there is no indication which order of kernel to use. The combined estimator offers reliably good performance and is often better than the optimal, especially in cases of large relative errors.

5. Conclusions

In this paper we relax the assumptions of high degree of density smoothness typically made in the literature on 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. We propose the use of the combined estimator when the degree of density smoothness is unknown and prove that it is adaptive in the sense that it obtains second-order rate efficiency when parametric rates are possible and is rate-efficient when parametric rates cannot be attained. Monte Carlo simulations demonstrate that even in the case where formally the smoothness assumptions hold, due to large values for the derivatives there is no guidance for selecting a bandwidth that will not lead to large errors for some distributions; thus the behavior of ADE based on an individual bandwidth becomes problematic. By not relying on a single kernel/bandwidth choice, the combined estimator reduces this sensitivity.

6. APPENDIX

Proof of Theorem 1. The proof relies on the expression for the MSE (9) that combines squared bias from Assumption 6 and variance as given in (8). The variance is obtained as a special case from the general formula for covariances (A.5) derived in the proof of Theorem 2. The variance has two leading parts, one that converges to Σ_2 at a parametric rate, $O(N^{-1})$, the other converges with rates $O(N^{-2}(h^k)^{-1}\mathbf{h}^{-2})$ to $\Sigma_1(K)$; the squared bias converges with rates $O(\mathbf{h}^{2\bar{v}})$.

In case (a) for $\left\{h : \left[N(h^k)\mathbf{h^2}\right]^{-1} = o(1); N\mathbf{h^{2\bar{v}}} = o(1)\right\}$ the term $N^{-1}\Sigma_2$ dominates the MSE; correspondingly a parametric rate holds for the estimator; the asymptotic normality result in PSS can easily adapt to accommodate different bandwidths and holds for this case. For $\left[N(h^k)\mathbf{h^2}\right]^{-1} = O(1)$ the parametric rate still holds but the variance may have a part that depends on the kernel. For $N\mathbf{h^{2\bar{v}}} = O(1)$ the rate is parametric, but asymptotic bias is present. When $N(h^k)\mathbf{h^2} \to 0$ (undersmoothing) the MSE is dominated by $N^{-2}(h^k)^{-1}\mathbf{h}^{-1}\Sigma_1(K)\mathbf{h}^{-1}$. The estimator has no asymptotic bias, but its variance depends on the kernel, convergence rate is $\mathbf{r_N^{-1}} = O(N^{-1}(h^k)^{-1}\mathbf{h}^{-1})$. If $N\mathbf{h^{2\bar{v}}} \to \infty$ (oversmoothing) the squared asymptotic bias dominates in the MSE and by standard arguments (Chebyshev's inequality) this situation results in the estimator converging in probability to $\mathcal{B}(K)$ with rates $\mathbf{r_N^{-1}} = O(N^{-1/2}\mathbf{h}^{-\bar{\mathbf{v}}})$.

In case (b) the range of bandwidths corresponding to parametric rates cannot be obtained. When $N^2(h^k)\mathbf{h^2h^{2\bar{v}}} \to 0$ the MSE is dominated by the term $N^{-2}(h^k)^{-1}\mathbf{h}^{-1}\Sigma_1(K)\mathbf{h}^{-1}$. The estimator has no asymptotic bias, convergence rate is $\mathbf{r_N^{-1}} = O(N^{-1}(h^k)^{-1}\mathbf{h}^{-1})$. If $N^2(h^k)\mathbf{h^2h^{2\bar{v}}} \to \infty$ the squared asymptotic bias dominates in the MSE and the estimator converges in probability to $\mathcal{B}(K)$ with rates $\mathbf{r_N^{-1}} = O(N^{-1/2}\mathbf{h}^{-\bar{\mathbf{v}}})$.

For (c) without loss of generality assume that h_1 has the slowest rate among the bandwidth components, then in terms of rates every other component is $h_\ell = O(h_1^{\sigma_\ell})$ with $\sigma_\ell \geq 1$. The part of $trMSE$ that depends on the bandwidth, $trMSE(h)$, then takes the form

$$N^{-2}h_1^{-\Sigma\sigma_\ell}\Sigma_{\ell=1}^k h_1^{-2\sigma_\ell}s_\ell + \Sigma_{\ell=1}^k h_1^{2\bar{v}_\ell\sigma_\ell}b_\ell$$

with positive coefficients, s_ℓ, b_ℓ. As h_1 increases the first term declines and the second term increases; in either of the cases (a) or (b) over the relevant range of h_1 the first term dominates the sum at low bandwidths, and the second at higher ones. As a continuous function of h_1 the $trMSE(h)$ attains a minimum over that range. If all the bandwidths are the same and equal to h_1, so that all $\sigma_\ell = 1, v_\ell = \bar{v} = const$ we get the optimal rate in (10) by equating the rates of the two components. ■

Proof of Theorem 2. To derive an expression for the covariance of $\hat{\delta}_N(K_{s_1}, h_{s_1})$ and $\hat{\delta}_N(K_{s_2}, h_{s_2})$, Γ_{s_1,s_2}, for $s_1, s_2 = 1, .., S$ we note

$$\Gamma_{s_1,s_2} = E\left[\hat{\delta}_N(K_{s_1}, h_{s_1})\hat{\delta}_N(K_{s_2}, h_{s_2})^T\right] - E\hat{\delta}_N(K_{s_1}, h_{s_1})E\hat{\delta}_N(K_{s_2}, h_{s_2})^T. \tag{A.1}$$

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

$$\begin{aligned} & E\left(\hat{\delta}_N(K_{s_1}, h_{s_1})\hat{\delta}_N(K_{s_2}, h_{s_2})^T\right) \qquad (A.2)\\ = \; & 4E\left\{\left[\tfrac{1}{N}\sum_{i=1}^N \hat{f}'_{(K_{s_1},h_{s_1})}(x_i)y_i\right]\left[\tfrac{1}{N}\sum_{i=1}^N \hat{f}'_{(K_{s_2},h_{s_2})}(x_i)y_i\right]^T\right\} \\ = \; & 4\left\{\tfrac{1}{N}E\left(\hat{f}'_{(K_{s_1},h_{s_1})}(x_i)\hat{f}'_{(K_{s_2},h_{s_2})}(x_i)^T y_i^2\right) + \right. \\ & \left. + \tfrac{N-1}{N}E\left(\hat{f}'_{(K_{s_1},h_{s_1})}(x_{i_1})\hat{f}'_{(K_{s_2},h_{s_2})}(x_{i_2})^T y_{i_1}y_{i_2}I(\, i_1 \neq i_2)\right)\right\}. \end{aligned}$$

For the first term of (A.2) we obtain

$$
\begin{aligned}
& E\left(\hat{f}'_{(K_{s_1},h_{s_1})}(x_i)\hat{f}'_{(K_2,h_2)}(x_i)^T y_i^2\right) \qquad (A.3)\\
= \; & \left(\tfrac{1}{N-1}\right)^2 E\left\{E_{z_i}\left(y_i^2\left[\sum_{j\neq i}\mathbf{h_{s_1}^{-1}}\left(h_{s_1}^k\right)^{-1}K'_{s_1}(\tfrac{x_i-x_j}{h_{s_1}})\right]\left[\sum_{j\neq i}\mathbf{h_{s_2}^{-1}}\left(h_{s_2}^k\right)^{-1}K'_{s_2}(\tfrac{x_i-x_j}{h_{s_2}})\right]^T\right)\right\}\\
= \; & \tfrac{1}{N-1}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left[y_i^2 E_{z_i}\left(K'_{s_1}(\tfrac{x_i-x_j}{h_{s_1}})K'_{s_2}(\tfrac{x_i-x_j}{h_{s_2}})^T I(\,i\neq j)\right)\right]\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{N-1}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left[y_i^2 E_{z_i}\left(K'_{s_1}(\tfrac{x_i-x_{j_1}}{h_{s_1}})K'_{s_2}(\tfrac{x_i-x_{j_2}}{h_{s_2}})^T I(\,i,j_1,j_2 \text{ pairwise distinct})\right)\right]\mathbf{h_{s_2}^{-1}}\\
= \; & \tfrac{1}{N-1}\cdot\left(h_{s_1}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left[y_i^2\int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})^T f(x_i-uh)du\right]\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{N-1}\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left[E_{z_i}\left(y_i K'_{s_1}(\tfrac{x_i-x_{j_1}}{h_{s_1}})\right)E_{z_i}\left(y_i K'_{s_2}(\tfrac{x_i-x_{j_2}}{h_{s_2}})\right)^T I(\,i,j_1,j_2 \text{ pairwise distinct})\right]\mathbf{h_{s_2}^{-1}}\\
= \; & \tfrac{1}{N-1}\cdot\left(h_{s_1}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}\left[Ey_i^2 f(x_i)\mu_2(K_{s_1},K_{s_2},h_{s_1}/h_{s_2})+o(1)\right]\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{N-1}\cdot\left[E\left(y_i^2 f'(x_i)f'(x_i)^T\right)+o(1)\right],
\end{aligned}
$$

where the third equality uses the independence of last equality x_{j_1} and x_{j_2}. The last equality uses integration by parts: $E_{z_i}\left(\left(h_t^k\right)^{-1}\mathbf{h_t^{-1}}y_i K'_t(\frac{x_i-x_j}{h_t})\right) = y_i\int K_t(u)f'(x_i-uh_t)du) = y_i f'(x_i)+y_i\int K_t(u)\left(f'(x_i-uh_t)-f'(x_i)\right)du) = y_i f'(x_i)+o(1)$, $t=s_1,s_2$.

For the second term of (A.2), where for brevity we omit terms such as $I(i_1\neq i_2)$ in the terms of the expression, we obtain

$$
\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1}y_{i_2}I(i_1\neq i_2)\right)\\
= \; & \tfrac{N-2}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(E_{z_{j_1}}\left[y_{i_1}K'_{s_1}(\tfrac{x_{i_1}-x_{j_1}}{h_{s_1}})\right]E_{z_{j_1}}\left[y_{i_2}K'_{s_2}(\tfrac{x_{i_2}-x_{j_1}}{h_{s_2}})\right]^T\right)\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{1}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(y_{i_2}E_{z_{i_2}}\left[y_{i_1}K'_{s_1}(\tfrac{x_{i_1}-x_{i_2}}{h_{s_1}})K'_{s_2}(\tfrac{x_{i_2}-x_{i_1}}{h_{s_2}})^T\right]\right)\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(E_{z_{i_2}}\left[y_{i_1}K'_{s_1}(\tfrac{x_{i_1}-x_{i_2}}{h_{s_1}})\right]E_{z_{i_2}}\left[y_{i_2}K'_{s_2}(\tfrac{x_{i_2}-x_{j_2}}{h_2})\right]^T\right)\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(E_{z_{i_1}}\left[y_{i_1}K'_{s_1}(\tfrac{x_{i_1}-x_{j_1}}{h_{s_1}})\right]E_{z_{i_1}}\left[y_{i_2}K'_{s_2}(\tfrac{x_{i_2}-x_{i_1}}{h_{s_2}})\right]^T\right)\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{(N-2)(N-3)}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(E_{z_{i_1}}\left[y_{i_1}K'_{s_1}(\tfrac{x_{i_1}-x_{j_1}}{h_{s_1}})\right]\right)E\left(E_{z_{i_2}}\left[y_{i_2}K'_{s_2}(\tfrac{x_{i_2}-x_{j_2}}{h_{s_2}})\right]\right)^T\mathbf{h_{s_2}^{-1}}.
\end{aligned}
$$

With $\left(h_t^k\right)^{-1}\mathbf{h_t^{-1}}E_{z_j}\left[K'_t(\frac{x_i-x_j}{h_t})y_i\right] = \mathbf{h_t^{-1}}\int K'_t(u)(gf)(x_j+uh_t)du = -\int K_t(u)(gf)'(x_j+uh_t)du = -(gf)'(x_j)+o(1)$ (using integration by parts); $\left(h_1^k\right)^{-1}E_{z_i}\left[K'_1(\frac{x_j-x_i}{h_1})K'_2(\frac{x_i-x_j}{h_2})^T y_j\right] =$

$\int K_1'(u)K_2'(-u\frac{h_1}{h_2})^T(gf)(x_i+uh)du = -\mu_2(K_1,K_2,h_1/h_2)(gf)(x_i)+o(1)$; and
$E\left[E_{z_i}\left(\left(h_t^k\right)^{-1}\mathbf{h_t^{-1}}y_iK_t'(\frac{x_i-x_j}{h_t})\right)\right] = (1/2)\left(E\hat{\delta}_N(K_t,h_t)\right)+o(1), t=s_1,s_2$, this gives

$$\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_{i_1})\hat{f}'_{(K,h)}(x_{i_2})^T y_{i_1}y_{i_2}I(i_1\neq i_2)\right) \qquad (A.4)\\
= \;& \tfrac{N-2}{(N-1)^2}\cdot\left[E\left((gf)'(x_i)\,(gf)'(x_i)^T\right)+o(1)\right]+\\
& \tfrac{1}{(N-1)^2}\cdot\left(h_{s_1}^k h_{s_2}^k\right)^{-1}\mathbf{h_{s_1}^{-1}}E\left(\int K'_{s_1}(u)K'_{s_2}(-u\frac{h_{s_1}}{h_{s_2}})^T du(gf)(x_i)y_i+o(1)\right)\mathbf{h_{s_2}^{-1}}+\\
& \tfrac{N-2}{(N-1)^2}\cdot E\left(-(gf)'(x_i)(f'(x_i)y_i)^T+o(1)\right)+\\
& \tfrac{N-2}{(N-1)^2}\cdot E\left((f'(x_i)y_i)\,(-(gf)'(x_i))^T+o(1)\right)+\\
& \tfrac{(N-2)(N-3)}{(N-1)^2}\cdot\frac{1}{4}\left[\left(E\hat{\delta}_N(K_{s_1},h_{s_1})\right)\left(\hat{\delta}_N(K_{s_2},h_{s_2})\right)^T+o(1)\right]
\end{aligned}$$

Substituting (A.3), (A.4) using (A.2) in (A.1) gives the expression

$$\Gamma_{s_1,s_2} = N^{-2}\left(h_{s_2}^k\right)^{-1}\mathbf{h}_{s_1}^{-1}\left(\Sigma_1(K_{s_1},K_{s_2},h_{s_1},h_{s_2})+o(1)\right)\mathbf{h}_{s_2}^{-1}+\left(\Sigma_2+o(1)\right)N^{-1} \qquad (A.5)$$

with

$$\begin{aligned}
& \Sigma_1(K_{s_1},K_{s_2},h_{s_1}/h_{s_2}) = 4E\left[y^2f(x_i)-(gf)(x_i)y_i\right]\mu_2(K_{s_1},K_{s_2},h_{s_1},h_{s_2}); \qquad (A.6)\\
& \mu_2(K_{s_1},K_{s_2},h_{s_1},h_{s_2}) = \int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})^T du;\ \text{and}\\
& \Sigma_2 = 4\left\{E([(g'f)(x_i)-(y_i-g(x_i))f'(x_i)]\,[(g'f)(x_i)-(y_i-g(x_i))f'(x_i)]^T)\right\}-4\delta_0\delta_0^T.
\end{aligned}$$

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

$$\begin{aligned}
\mu_2 &= \int K'_{s_1}(u)K'_{s_2}(u\frac{h_{s_1}}{h_{s_2}})^T du\\
&= \int K'_{s_1}(u)du\ K'_{s_2}(0)^T+\mathbf{h}_{s_1}\int K'_{s_1}(u)K''_{s_2}(\tilde{u})udu\ \mathbf{h}_{s_2}^{-1}\\
&= \mathbf{h}_{s_1}\mathbf{h_{s_2}^{-1}}O(1),
\end{aligned}$$

where $\tilde{u}$ lies between 0 and u.

Only two cases of different rates are possible here: (a) a parametric rate for s_2 and a non-parametric for s_1, and (b) non-parametric (different) rates for both.

Consider case (a): $N(h_{s_1}^k)\mathbf{h}_{\mathbf{s_1}}^{\mathbf{2}} \to 0$; $N(h_{s_2}^k)\mathbf{h}_{\mathbf{s_2}}^{\mathbf{2}} \to \infty$. Then

$$\begin{aligned} & Cov(N(h_{s_1}^{k/2})\mathbf{h}_{\mathbf{s_1}}\hat{\delta}_N(K_{s_1},h_{s_1}), \sqrt{N}\hat{\delta}_N(K_{s_2},h_{s_2})) \\ = \; & N^{3/2}(h_{s_1}^{k/2})\mathbf{h}_{\mathbf{s_1}}[N^{-2}(h_{s_2}^k)^{-1}\mathbf{h}_{\mathbf{s_2}}^{-\mathbf{2}}O(1) + N^{-1}O(1)] \\ = \; & O(\left[N^{1/2}(h_{s_1}^{k/2})\mathbf{h}_{\mathbf{s_1}}\right]\left[N^{-1}(h_{s_2}^k)^{-1}\mathbf{h}_{\mathbf{s_2}}^{-\mathbf{2}}\right]) + O(N^{1/2}(h_{s_1}^{k/2})\mathbf{h}_{\mathbf{s_1}}) = o(1). \end{aligned}$$

For case (b): $N(h_{s_1}^k)\mathbf{h}_{\mathbf{s_1}}^{\mathbf{2}}; N(h_{s_2}^k)\mathbf{h}_{\mathbf{s_2}}^{\mathbf{2}} \to 0$ we get

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

For $s_1 = s_2$ we get the variance expression in (8) with

$$\begin{aligned} & \Sigma_1(K) = 4E\left[y_i^2 f(x_i) - (gf)(x_i)y_i\right]\mu_2(K); \qquad (A.7) \\ & \mu_2(K) = \int K'(u)K'(u)^T du; \text{ and} \\ & \Sigma_2 = 4\left\{E([(g'f)(x_i) - (y_i - g(x_i))f'(x_i)]\,[(g'f)(x_i) - (y_i - g(x_i))f'(x_i)]^T)\right\} - 4\delta_0\delta_0^T. \end{aligned}$$

■

Proof of Theorem 3. (a) We utilize the expression in Assumption 6(b) component-wise

$$E(\hat{\delta}_N(K,h) - \delta_0) = \begin{pmatrix} h_1^{\bar{v}_1}\mathcal{B}_1(K) + o(h_1^{\bar{v}_1+\gamma}) \\ \vdots \\ h_k^{\bar{v}_k}\mathcal{B}_k(K) + o(h_k^{\bar{v}_k+\gamma}) \end{pmatrix}. \qquad (A.8)$$

Following 6(c) consider constant $\bar{v}$. Using (A.8) and (8) with Chebyshev's inequality we can write for bandwidth vector h_t

$$\hat{\delta}_N(K,h_t) = \delta_0 + \mathbf{h}_t^{\bar{v}}[\mathcal{B}(K) + o_p(h_t^{\gamma}) + \psi_t],$$

where $\psi_t = O_p(N^{-1}h_t^{-(\frac{k+2}{2})} + N^{-1/2})h_t^{-\bar{v}}$**.** For each kernel consider the sequence of bandwidths $h_t = c_t h^{gcv} N^{\gamma_t}$ with $\gamma_t > 0$ (to ensure oversmoothing) and $\gamma_t < \frac{1}{2\bar{v}+k}$ to ensure that the bandwidths converge to zero; the condition $\gamma < \frac{1}{2\bar{v}+k}$ relies on the unknown $\bar{v}$;

the more smoothness the tighter is the bound on γ_t; thus we can replace this condition by $\gamma_t < \frac{1}{2v(K)+k}$. Thus for some $H \geq 2$ we obtain a sequence of bandwidth vectors $\{h_t\}_{t=1}^{H}$ for which the bias term dominates in the MSE for this estimator so that $\psi_t = o_p(h_t^{\gamma})$. For this sequence of bandwidths then

$$\hat{\delta}_N(K, h_t) = \delta_0 + \mathbf{h}_t^{\bar{v}}[\mathcal{B}(K) + o_p(h_t^{\gamma})].$$

Difference these equations component-wise to get rid of δ_0; then for the $\ell-$th component based on two distinct bandwidth vectors $h_t, h_{t'}$

$$\left\{\hat{\delta}_N(K, h_t) - \hat{\delta}_N(K, h_{t'})\right\}_{\ell} = (h_{t\ell}^{\bar{v}} - h_{t'\ell}^{\bar{v}})\, \mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\bar{v}+\gamma} + h_{t'\ell}^{\bar{v}+\gamma}),\ \ell = 1, ..k.$$

When $h_{t'} = o(h_t)$ we get

$$\hat{\delta}_N(K, h_t)_\ell - \hat{\delta}_N(K, h_{t'})_\ell = h_{t\ell}^{\bar{v}} \mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\bar{v}+\xi}),\ \ell = 1, ..k \tag{A.9}$$

$\xi > 0$. For each ℓ we can consider Q such equations with $2 \leq Q \leq \frac{H(H+1)}{2}$. Take absolute values on both sides of (A.9):

$$\left|h_{t\ell}^{\bar{v}} \mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\bar{v}+\xi})\right| = h_{t\ell}^{\bar{v}} \left|\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right|.$$

Consider

$$\left|\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right| = \left(\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right) sgn\left(\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right),$$

where from Assumption 6(a)

$$sgn\left(\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right) = sgn\mathcal{B}_\ell(K) + o_p(1)$$

providing

$$\left|\mathcal{B}_\ell(K) + o_p(h_{t\ell}^{\xi})\right| = |\mathcal{B}_\ell(K)| + o_p(h_{t\ell}^{\xi}).$$

Then for each ℓ there are Q equations

$$\begin{aligned} \ln\left|\hat{\delta}_N(K, h_t)_\ell - \hat{\delta}_N(K, h_{t'})_\ell\right| &= \bar{v} \ln h_{t\ell} + \ln|\mathcal{B}_\ell(K)| + \ln(1 + o_p(h_{t\ell}^{\xi})) \\ &= \bar{v} \ln h_{t\ell} + \ln|\mathcal{B}_\ell(K)| + o_p(h_{t\ell}^{\xi}), \end{aligned}$$

where the last equality follows from the expansion of the ln function. For $r_{t\ell} = \ln h_{t\ell} - \frac{1}{Q}\Sigma_q \ln h_{q\ell}$ we can get the estimator

$$\begin{aligned}\widehat{\bar{v}}_\ell &= \frac{\Sigma_t \ln\left|\hat{\delta}_N(K,h_t)_\ell - \hat{\delta}_N(K,h_{t'})_\ell\right| r_{t\ell}}{\Sigma_t r_{t\ell}^2} \\ &= \bar{v} + \frac{\Sigma_t o_p(h_{t\ell}^{\xi}) r_{t\ell}}{\Sigma_t r_{t\ell}^2}.\end{aligned}$$

Noting from the bandwidth rates that $r_{t\ell} = O(\ln N)$; $\Sigma_t r_{t\ell}^2 = O\left((\ln N)^2\right)$, we get that for $\bar{h}_{t\ell}$ which goes to zero the slowest among all $h_{t\ell}, t = 1, ...Q$, the difference (for any ℓ) $\left|\widehat{\bar{v}}_\ell - \bar{v}\right| = o_p(\bar{h}_{t\ell}^{\xi}(\ln N)^{-1}) = o_p(1)$.

Substituting $\widehat{\bar{v}}$ for $\bar{v}$ in the expression (7) still gives the same leading term and with this substitution minimizing estimated $trMSE(h)$ provides similarly to (c) of Theorem 1 consistent estimators for rates in the optimal bandwidth vector, h^{opt}.

(b) Since the rates for the squared biases and variances for the h^{opt} are the same (to minimize $trMSE$) the relation between ζ and ξ makes sure that the variances of the components of $\hat{\delta}_N(K,h_u)$ go to zero faster than squared biases of $\hat{\delta}_N(K,h_o)$ thus

$$\hat{\delta}_N(K,h_o) - \hat{\delta}_N(K,h_u) = \mathbf{h}_o^{\bar{v}}[\mathcal{B}(K) + o_p(h_o^{\gamma})]; \gamma > 0.$$

Then $\left|\hat{\delta}_N(K,h_o)_\ell - \hat{\delta}_N(K,h_u)_\ell - Bias\hat{\delta}(K,h_o)_\ell\right| = o_p(h_{o\ell}^{\bar{v}})$. Since the leading terms in the biases for different bandwidths, e.g. h_t and h_o differ by the ratio $(\frac{h_t}{h_o})^{\bar{v}}$, a consistent estimator for $Bias\hat{\delta}(K,h_t)_\ell$ is provided by $(\frac{h_{t\ell}}{h_{o\ell}})^{\bar{v}}(\hat{\delta}_N(K,h_o)_\ell - \hat{\delta}_N(K,h_u)_\ell)$; substituting consistent estimates $\widehat{\bar{v}}$ for $\bar{v}$ will not affect consistency of the bias estimator.

(c) From Theorem 3 of KZW it follows that the combined estimator provides for the estimated $trMSE(h)$ the convergence rate that is no worse than the rate that results from the fastest converging individual estimator among $\hat{\delta}(K,h_t)$. Since the estimated h^{opt} provides the rate that converges to the best rate for the $trMSE(h)$ and this estimator is included in the combination, the combined estimator provides asymptotically the best rate for $trMSE(h)$. This rate characterizes first-order efficiency when the range $H(\sqrt{N})$ is empty and gives second-order efficiency for $h \in H(\sqrt{N})$. ■

References

[1] Banerjee, A.N. (2007): "A method of estimating the average derivative," *Journal of Econometrics*, **136**, 65-88.

[2] Blundell, R., A. Duncan, and K. Pendakur (1998): "Semiparametric estimation and consumer demand," *Journal of Applied Econometrics*, **13**, 435–461.

[3] Chaudhuri, P., Doksum, K. and Samarov, A. (1997): "On average derivative quantile regression," *Annals of Statistics*, **25**, 715-744.

[4] Dalalyan, A.S., G.K. Golubev and A.B. Tsybakov (2006): "Penalized maximum likelihood and semiparametric second order efficiency," *The Annals of Statistics*, **34**, 169-201.

[5] Donkers, B. and M. Schafgans (2008): "A method of moments estimator for semiparametric index models," *Econometric Theory*, **24**, forthcoming.

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

[7] Härdle,W., W. Hildenbrand and M. Jerison (1991): "Empirical evidence on the law of demand," *Econometrica,* **59**, 1525–1549.

[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] Horowitz, J.L. and V.G. Spokoiny (2001): "An adaptive, rate-optimal test of parametric mean-regression model against a nonparametric alternative", *Econometrica,* **69,** 599-631.

[11] Izenman, A. J., and C.J. Sommer (1988), "Philatelic mixtures and multimodal densities," *Journal of the American Statistical Association*, **83**, 941-953.

[12] Juditsky, A. and A. Nemirovski (2000): "Functional aggregation for nonparametric regression," *The Annals of Statistics*, **3**, 681–712.

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

[14] Kotlyarova, Y. and V. Zinde-Walsh (2007): "Robust kernel estimator for densities of unknown smoothness," *Journal of Nonparametric Statistics*, **19,** 89-101.

[15] Lalley, S.P. and A. Nobel (2003): "Indistinguishability of absolutely continuous and singular distributions," *Statistics and Probability Letter*, **62**, 145-154.

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

[17] Li, Q. and J. S. Racine (2006): Nonparametric Econometrics: Theory and Practice, Princeton University Press.

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

[19] Miller, R.G. (1974): "The jackknife - a review," *Biometrika*, **61**, 1–15.

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

[21] Nichiyama, Y. and P.M. Robinson (2000): "Edgeworth expansions for semiparametric averaged derivatives," *Econometrica,* **68**, 931–980.

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

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

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

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

[26] Robinson, P.M. (1989): "Hypothesis testing in semiparametric and nonparametric models for econometric time series. *Review of Economic Studies* **56**, 511–534.

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

[28] Schafgans, M.M.A. and V. Zinde-Walsh (2007): "Robust average derivative estimation," Department of Economics Working Paper 2007-12, McGill University, 2007.

[29] Schucany, W.R., H.L. Gray and D.B. Owen (1971): "On bias reduction in estimation," *Journal of the American Statistical Association,* **66**, 524–533

[30] 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.

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

[32] Yang, Y. (2000): "Combining different procedures for adaptive regression," *Journal of Multivariate Analysis,* **74**, 135–161.

[33] Yatchew, A. (2003): Semiparametric Regression for the Applied Econometrician, Cambridge University Press, Cambridge.