Robust Average Derivative Estimation

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

February 2007

(Preliminary and Incomplete – Do not quote without permission)

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 dicussed. Monte Carlo results for density weighted ADE confirm good performance of the combined estimator.

*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 rely on estimation of average derivatives (ADE) of the conditional mean function (averaged response coefficients); the most widely used such model is the single index model where the conditional mean function can be represented as a univariate function of a linear combination of conditioning variables. Index representations are ubiquitous in econometric studies of limited dependent variable models, partial linear models and in nonparametric demand analysis. Estimation of coefficients in single index models relies on the fact that averaged derivatives of the conditional mean (or conditional mean weighted by some function) are proportional to the coefficients, thus a non-parametric estimator of the derivative of the conditional mean function provides estimates of the coefficients (up to a multiplicative factor). This method does not require assumptions about the functional form of either the density of the data or of the true regression function.

Powell, Stock and Stoker (1989) and Robinson (1989) examined density weighted average derivatives, while Härdle and Stoker (1989) investigated the properties of the average derivatives themselves; one important difference is the need to introduce some form of trimming when there is no weighting by density since the estimator of density appears in the denominator of the ADE and may be close to zero. Newey and Stoker (1993) addressed the issue of efficiency related to 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) addressed the lack of identification associated with the estimation of coefficients in single index models for cases where the derivative of the unknown function on average equals zero; they propose an estimator based on the average outer product of derivatives which resolves this lack of identification while at the same time enabling the estimation of parameters in multiple index models.

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. However, other terms are sensitive to bandwidth/kernel choice. Powell and Stoker (1996) address the optimal bandwidth choice for (weighted) average derivative estimation. Further results including finite sample performance of average derivatives and corrections to improve finite-sample properties are discussed in Robinson (1995) and Nichiyama and Robinson (2000, 2005). Parametric rates of convergence and thus all the results in this literature rely on the assumption of sufficiently high degree of smoothness of the underlying density.

In this paper we are motivated by a concern about the assumed high degree of density smoothness. There is some empirical evidence that for many variables the density may not be sufficiently smooth and may have shapes that are not consistent with a high level of smoothness: peaks and cusps and even discontinuity of density functions are not uncommon (references). We 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 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. Incorrect assumptions about smoothness may lead to using an estimator that suffers from problems associated with under- or oversmoothing. 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 examine a variety of estimators corresponding to different kernels and bandwidth rates and derive the joint limit process for those estimators. When each estimator is normalized appropriately (with different rates) we obtain a joint Gaussian limit process which possibly exhibits an asymptotic bias and possibly some degeneracy. Any linear combination of such estimators is asymptotically Gaussian and we are able select a combination

that minimizes the estimated asymptotic MSE. The resulting estimator is what we call the combined estimator. Kotlyarova and Zinde-Walsh (2006) have shown that the weights in this combination will be such that they 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.

Performance of the combined estimator relies on good estimators for the asymptotic variances and biases that enter into the combination; a method of estimation that does not depend on our knowledge about the degree of density smoothness is required. Variances can be estimated without much difficulty, e.g. by bootstrap. In Kotlyarova and Zinde-Walsh (2006) a method of estimation of the asymptotic bias of a (possibly) oversmoothed estimator that utilizes asymptotically unbiased undersmoothed estimators is proposed; here we add bootstrapping to improve the properties of this estimator of asymptotic bias. Without prior knowledge of smoothness the bandwidth choices must be such that bandwidths optimal for smooth densities (obtained by rule-of-thumb or by cross-validation) should be included to cover the possibility of high smoothness; since such choices will correspond to oversmoothing if density is not sufficiently smooth. Some lower bandwidths determined e.g. as percentiles of the "optimal" bandwidth should also be considered. Our method requires utilization of undersmoothed estimators to determine asymptotic bias, thus it is important to consider fairly small bandwidths. We select kernels of different orders for the combination; most of the Monte Carlo results both here and for other combined estimators (for SMS in binary choice model and for density estimation, Kotlyarova, 2005) are not very sensitive to kernel choices.

Monte Carlo results here are for the density weighted ADE for a single index model. We demonstrate here that even in the case where the smoothness assumptions hold the combined estimator performs similarly 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 results in cases where the density is not sufficiently smooth or while smooth has a shape that gives high values for low-order derivatives (e.g. a trimodal mix-

ture of normals) indicate gains from the combined estimator relative to the optimal ADE estimator.

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 several estimators and the combined estimator. Section 4 provides the result of a Monte Carlo study analysing for the Tobit model the performance of the combined estimator vis-a-vis single bandwidth/kernel based estimators for the density-weighted ADE in cases with different smoothness conditions.

2. General set-up and assumptions

We should have a brief intro to this section maybe?

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, $w(x)$; 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 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 $b_N \to 0$. An alternative is the density weighted average derivative estimator of Powell,Stock and Stoker (1989), PSS, with $w(x) = f(x)$. Here we focus on the PSS estimator.

The nonparametric estimates for the various unknown derivative based functionals make use of kernel smoothing functions. E.g., the nonparametric estimate for the derivative of the density is given by

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

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

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. *Let $z_i = (y_i, x_i^T)^T$, $i = 1, .., N$ be a random sample drawn from $f^*(y, x)$, with $f^*(y, x)$ the density of (y, x). 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 $E(\omega^2_{(fg)'}(x)[1 + |y| + \|x\|]) < \infty$.

Both the choice of the kernel (its order) and the selection of bandwidth have played a crucial role in the literature 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, except for the fact that we allow for the kernel to be asymmetric.

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 concerning the smoothness of the density in the literature (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):

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

for every $x, x+\Delta x \in E$.

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 $E(\omega_{f^{(m)}}(x))^2[1+|y|^2+\|x\|]) < \infty$.*

The assumption implies that each component of the derivative of density $f'(x) \in C_{m-1+\alpha}(\Omega)$ and thus for every component of the derivative of density 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:

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

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 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]

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

where the latter equality uses the Hölder inequality. If differentiability conditions typically assumed to ensure that $\bar{v} > \frac{k+2}{2}$ 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), is defined as

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

Given Assumptions 1-3, (3) can be represented as

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

(see Lemma 2.1 in PSS).

[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 where $f'(x)$ is replaced by a consistent nonparametric estimate, i.e.,

$$\hat{\delta}_N(K,h) = \frac{-2}{N}\sum_{i=1}^{N} \hat{f}'_{(K,h)}(x_i)y_i, \tag{4}$$

where

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

K is the kernel smoothing function (which PSS assume to be symmetric) and h is a smoothing parameter that depends on the sample size N, with $h \to 0$ as $N \to \infty$.

We derive the variance of $\hat{\delta}_N(K,h)$ without relying on results on $U-$statistics to accomodate possibly non-symmetric kernels. This is provided in Lemma 1 in the Appendix. We obtain the following expression for this variance:

$$Var(\hat{\delta}_N(K,h)) = \Sigma_{1\delta}(K)N^{-2}h^{-(k+2)} + \Sigma_{2\delta}N^{-1} + O(N^{-2}) \tag{5}$$

where

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

$\Sigma_{2\delta}$ 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 (5) 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 (or when oversmoothing) 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

$$Bias(\hat{\delta}_N(K,h)) = -2E(A(K,h,x_i)y_i). \tag{6}$$

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.*

The asymptotic bias of the estimator $\hat{\delta}_N(K,h)$ can then be written as

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

and vanishes as $h \to 0$. We note that assumption (6) could hold as a results of primitive moment assumptions on y_i, $f(x_i)$, and $g(x_i)$.

Let "$d(N) \approx O(1)$" denote the case when both $d(N))$ and $1/d(N)$ are $O(1)$ as $N \to \infty$. Assume that $C = \lim_{N\to\infty} Nh^{k+2}$ always exists and $C \in [0,\infty]$.

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. *choosing $h : Nh^{k+2} = o(1), N^2h^{k+2} \to \infty$ provides an unbiased but not efficient estimator*

$$Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(0, \Sigma_{1\delta}(K));$$

ii. *if $h : Nh^{k+2} \to C, 0 < C < \infty$;*

$$\sqrt{N}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(0, C\Sigma_{1\delta}(K) + \Sigma_{2\delta});$$

iii. *when $h : Nh^{k+2} \to \infty$, $Nh^{2\bar{v}} = o(1)$* the same result as 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. *if $h : Nh^{k+2} \to \infty$*, but $Nh^{2\bar{v}} \approx O(1)$, a biased asymptotically normal estimator results:

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

v. *if* $h : Nh^{2\bar{v}} \to \infty$; 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* $h : Nh^{k+2+2\bar{v}} = o(1)$,$N^2h^{k+2} \to \infty$ in the limit there is normality, no bias:

$$Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(0, \Sigma_{1\delta}(K));$$

ii. *for* $h : Nh^{k+2+2\bar{v}} \approx O(1), N^2h^{k+2} \to \infty$ in the limit there is normality with asymptotic bias:

$$Nh^{\frac{k+2}{2}}\left(\hat{\delta}_N(K,h) - \delta_0\right) \xrightarrow{d} N(\mathcal{B}(K), \Sigma_{1\delta}(K));$$

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

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

Proof. See Appendix (where the variances and covariances are derived for any kernel but parts of the normality proof are provided for the case of a symmetric kernel).

Selection of the optimal bandwidth as minimizing the mean squared error critically depends on our knowledge of the degree of smoothness of the density. Let v denote the true differentiability (smoothness) of f' and choose the order of our kernel $v(K) \leq [v]$. The $MSE(\hat{\delta}_N(K,h))$ can then be represented as

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

and the optimal bandwidth yields $h^{opt} = cN^{-2/(2v(K)+k+2)}$, where the problem of efficient estimation is to find an appropriate c (e.g., Powell and Stoker (1996)). If higher order derivatives exist, further improvements in efficiency can be obtained by using a higher

order kernel to reduce the bias. In any case, to ascertain a parametric rate of our limiting distribution for $\hat{\delta}_N(K,h)$ with the use of the higher order kernel (as long as the density is sufficiently differentiable to have $v(K) \leq [v]$), our bandwidth sequence needs to satisfy $Nh^{2v(K)} \to 0$, and the degree of smoothness of the derivative of the density v needs be in excess of $\frac{k+2}{2}$ (with $Nh^{k+2} \approx O(1)$ to guarantee boundedness of the variance of$\sqrt{N}$ $\hat{\delta}_N(K,h)$). 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 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 degree of density smoothness is not known there is no guidance for the choice of kernel and bandwidth: a higher order kernel and larger bandwidth that could be better if there were more smoothness could lead to substantial bias if the density is less smooth. Without making further assumptions about knowledge of the degree of smoothness of the density, all that is known is that for some rate of $h \to 0$ there is undersmoothing: no asymptotic bias and a limiting Gaussian distribution, and for some slower convergence rate of h there is oversmoothing. An optimal rate may exist, but to determine it, and to identify bandwidths which lead to under- and over- smoothing precise knowledge of v, density smoothness, is required.

The situation where there is uncertainty about smoothness of density was considered in Kotlyarova and Zinde-Walsh (2006), hereafter referred to as KZW. Theorem 1 corresponds to Assumption 1 of that paper and demonstrates that when our Assumptions 1–6 are satisfied the estimator satisfies their Assumption 1. We next establish that Assumption 2 of that paper is satisfied as well. Consider several kernel/bandwidth sequence pairs $(h_{Nj}, K_j), j = 1, ...J$ and the corresponding estimators, $\hat{\delta}_N(K_j, h_{Nj})$. If all satisfy the assumptions of Theorem 1 then there exist corresponding rates r_{Nj} for which the joint limit process of $r_{Nj}\left(\hat{\delta}_N(K_j,h_{Nj}) - \delta_0\right)$ is non-zero Gaussian, possibly degenerate.

Theorem 2. *Under the Assumptions of Theorem 1 the joint limit process for the vector with components* $r_{Nj}\left(\hat{\delta}_N(K_j,h_{Nj}) - \delta_0\right), j = 1, ...J$ *is Gaussian with the covariance matrix*

such that for components that correspond to different rates covariances are zero.

Proof. See appendix.

Consider a linear combination of the estimators[2]

$$\hat{\delta}_N^* = \sum_j a_j \hat{\delta}_N(K_j, h_{Nj}) \text{ with } \sum_{j=1}^{J} a_j = 1.$$

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

$$\sum_{t_1,s_1}\sum_{t_2,s_2} a_{t_1,s_1} a_{t_2,s_2} Cov(\hat{\delta}_N(K_{t_1}, h_{s_1}), \hat{\delta}_N(K_{t_2}, h_{s_2})) \equiv \sum a_{j_1} a_{j_2} \Gamma_{j_1 j_2},$$

where (see appendix)

$$\Gamma_{j_1 j_2} = \Sigma_{1\delta}(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) N^{-2} h_{s_2}^{-(k+1)} h_{s_1}^{-1} + \Sigma_{2\delta} N^{-1} + O(N^{-2})$$

with

$$\begin{aligned} & \Sigma_{1\delta}(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) \\ = \; & 4E\left[y^2 f(x_i) \mu_2(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) + \mu_2^*(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2})(gf)(x_i) y_i\right] \end{aligned}$$

and $\Sigma_{2\delta}$ as before.

The $MSE(\hat{\delta}_N^*) = MSE(\sum_{t,s} a_{t,s} \hat{\delta}_N(K_t, h_s))$ can then be represented as

$$MSE(\hat{\delta}_N^*(K_t, h_s)) = \sum a_{j_1} a_{j_2} (\mathcal{B}_{j_1} \mathcal{B}_{j_2}^T + \Gamma_{j_1 j_2})$$

with

$$\mathcal{B}(K_j) = -2A(K_j)$$

To optimally choose the weights a_{j_1}, we will minimize the trace of the AMSE as in KZW.[3]

$$\begin{aligned} & tr(AMSE(\hat{\delta}_N^*(K_t, h_s)) = \sum a_{j_1} a_{j_2} (\tilde{\mathcal{B}}_{j_1}^T \tilde{\mathcal{B}}_{j_2} + tr\tilde{\Gamma}_{j_1 j_2}) = a' D a, \\ & \text{where } \{D\}_{j_1 j_2} = \mathcal{B}_{j_1}^T \mathcal{B}_{j_2} + tr\Gamma_{j_1 j_2}, \end{aligned}$$

[2] Alternatively, more complicated though, we could consider $\tilde{\delta}_N = \frac{-2}{N} \sum_{i=1}^{N} \sum_{s=1}^{S} w_s(x_i) \hat{f}'_{h_s, K_s}(x_i) y_i$, with $\sum_{s=1}^{S} w_s(x_i) = 1$

[3] 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.

$\tilde{\mathcal{B}}_j = \mathcal{B}_j / r_N(t_j, s_j)$, and $\tilde{\Gamma}_{j_1 j_2} = \Gamma_{j_1 j_2} / (r_N(t_{j_1}, s_{j_1}) * r_N(t_{j_2}, s_{j_2}))$.

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

KZW discusses the optimal weights that minimize the (consistently) estimated $tr(AMSE(\hat{\delta}_N^*)$ subject to $\sum_j a_j = 1$; here we summarize the results. After ranking the pairs (K_{t_j}, h_{s_j}) in declining order of rates $r_N(t_j, s_j)$, denote by D^I the largest invertible submatrix of D and by D_{11} its square submatrix associated with estimators having the fastest rate of convergence; note that it can have entries associated with at most one oversmoothed estimator to be of full rank. Then $a' D^I a$ (subject to $\sum_j a_j = 1$) is minimized by

$$a^I_{\lim} = \left(\left(\frac{1}{\iota' D_{11}^{-1} \iota} \right) D_{11}^{-1} \iota, 0, ..., 0 \right)',$$

that is by weights equalling $\left(\frac{1}{\iota' D_{11}^{-1} \iota} \right) D_{11}^{-1} \iota$ to the kernel/bandwidth combinations having the fastest rate of convergence and zero weight to all combinations with slower rate of convergence. Note that the weights in the limiting linear combination are non-negative for estimators corresponding to D^I (at most one asymptotically biased estimator). If $D^I \neq D$, then $D - D^I = D^{II}$ is of rank one and corresponds to oversmoothed estimators only. D^{II} has dimension more than one (otherwise the only oversmoothed estimator would have been included in D^I) ; note that then there always exist vectors $a^{II}_{\lim}$ such that

$$a^{II\prime}_{\lim} D^{II} a^{II}_{\lim} = 0; \ \Sigma a^{II}_{\lim i} = 1,$$

in other words, it is possible to automatically bias-correct by using the combined estimator with weights that are not restricted to be non-negative. Finally, the vector of weights in the combined estimator approaches an optimal linear combination of $a^I_{\lim}$ and $a^{II}_{\lim}$. The combined estimator thus has the trace of AMSE that converges at the rate no worse than that of the trace of AMSE for the fastest converging individual estimator. The combined estimator provides a useful mechanism for reducing the uncertainty about the degree of smoothness and thus about the best rate (bandwidth) and automatically selects the best

rate from those available even though it is not known a priori which of the estimators converges faster.

The optimality property of the combined estimator relies on consistent estimation of biases and covariances.[4] To provide a consistent estimate for the asymptotic variance that does not rely on the degree of smoothness, we apply the bootstap, which is obtained as

$$\begin{aligned}\widehat{\widehat{\Gamma}}_{12} &= \widehat{Cov}_B(\hat{\delta}_N(K_{t_1},h_{s_1}),\hat{\delta}_N(K_{t_2},h_{s_2})) \\ &= \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\delta}_{b,N}(K_{t_1},h_{s_1})-\hat{\delta}_N(K_{t_1},h_{s_1})\right)\left(\left(\hat{\delta}_{b,N}(K_{t_2},h_{s_2})-\hat{\delta}_N(K_{t_2},h_{s_2})\right)\right)',\end{aligned} \tag{8}$$

To provide us with a consistent estimator 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, a consistent estimator for the bias is obtained as:

$$\widehat{\widehat{\mathcal{B}}}_j \equiv \widehat{Bias}(\hat{\delta}_N(K_{t_j},h_{s_j})) = \hat{\delta}_N(K_{t_j},h_{s_j}) - \frac{1}{B}\sum_{b=1}^{B}\hat{\delta}_{b,N}(K_{t_j},h_{s_0}).$$

Alternatively the bootstrapped averaged estimates at the lowest bandwidth for all the kernels, $i=1,...,m$ could be used in bias estimation:

$$\widehat{\widehat{\mathcal{B}}}_j \equiv \widehat{Bias}(\hat{\delta}_N(K_{t_j},h_{s_j})) = \hat{\delta}_N(K_{t_j},h_{s_j}) - \frac{1}{B}\sum_{b=1}^{B}\frac{1}{m}\sum_{i=1}\hat{\delta}_{b,N}(K_{ti},h_{s_0}).$$

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

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

[4]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 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 the 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 explanatory variables. For the first model, we use two independent standard normal explanatory variables, i.e., $f_1(x_1, x_2) = \phi(x_1)\phi(x_2)$. This density is infinitely differentiable and very smooth; thus, the ADE estimator evaluated at the "optimal" bandwidth should be a good choice. This model, which we label (s,s), is considered to demonstrate that even in the case where the smoothness assumptions hold, the combined estimator performs similar to the ADE estimator evaluated at the optimal bandwidth. For the second model we use one standard normal explanatory variable and one mixture of normals, and in the third model both explanatory variables are mixtures of normals. We label these models respectively (s,m) and (m,m). As in the first model, we assume independence of the explanatory variables. Mixtures of normal, while still being infinitely differentiable, do allow behaviour resembling that of nonsmooth densities, e.g., the double claw density and the discrete comb density (see Marron and Wand (1992)). We consider here the trimodal normal mixture given by $f_m(x) = 0.5\phi(x + 0.767) + 3\phi(\frac{x+0.767-0.8}{0.1}) + 2\phi(\frac{x+0.767-1.2}{0.1})$.

So $f_2(x_1, x_2) = \phi(x_1)f_m(x_2)$ and $f_3(x_1, x_2) = f_m(x_1)f_m(x_2)$.[5] The sample size is set at 1000 and 100 replications are drawn 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 minimal 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. Other simulations will consider the use of asymmetric kernels, which may yield for the combined estimator further improvements.

For each kernel we consider three different bandwidths. The largest bandwidth is chosen on the basis of a generalized cross validation method where a gridsearch algorithm and 50 simulations are used. The cross-validation bandwidth is given by the optimal bandwidth sequence $h^{gcv} = cn^{-1/(2p+2)}$ (see Stone (1982)) with p equalling the order of the kernel (so that here $p = v(K)$). For densities of sufficient smoothness, this bandwidth does not represent the undersmoothing required to ensure asymptotic unbiasedness. When densities are not sufficiently smooth, $\overline{v} = v < v(K)$, h^{gcv} will even correspond to oversmoothing as we will have $Nh^{2\overline{v}} \to \infty$, providing cases (a)v, (b)iii, or (c)iii in Theorem 1. The smallest bandwidth for each kernel is chosen as $0.5h^{gcv}$, it needs to be be sufficiently small so as to ensure the required level of undersmoothing. In addition we take the intermediate bandwidth $0.75h^{gcv}$.

The generalized cross validation method applied is not that typically applied for non-parametric regression, but is specialized to the derivative of the regression function $g'(x)$. We use the usual generalized cross validation ($\min_h \sum_{i=1}^n (y_i - \hat{g}_h(x_i))^2$) to obtain numerical derivatives of $g(x)$, evaluated at a 50×50 uniform grid of the x's, which we denote as $\tilde{g}'_{h_{pgcv}}(x)$. The optimal bandwidth h for the derivative of the regression function is then obtained by minimizing $\left\| \tilde{g}'_{h_{pgcv}}(x) - \hat{g}'_h(x) \right\|^2$. The bandwidth obtained this way yielded smaller bandwidths than the usual cross validation method, which accorded well with se-

[5] We are planning an analysis using the claw density and discrete comb density (Marron and Wand (1992)) and are also exploring the selection of a density with a precise order or non-smoothness.

lecting the bandwidth by minimizing the mean squared error of the nonparametrically estimated moments (Donkers and Schafgans (2005)), a method which only can be applied in a simulation setting through the knowledge of the true data generating process.[6] Consistent estimators for biases and covariances of the density weighted ADE are obtained by bootstrap (with 250 bootstraps) as discussed in the previous section.

In table 1, we report the Root Mean Squared Errors (RMSE) of various density weighted average derivatives together with the average bias and standard deviation (theoretical (T), sample (S), and bootstrapped (B)) of the average derivatives. The first three columns present the results using the 2nd order kernel (K_2) for various bandwidths, the next three columns present the results using the 4th order kernel (K_4) for various bandwidth, and the final column that of the combined estimator. The RMSEs using the different pairs of kernels and bandwidths should be compared with the RMSE of the combined estimator, which optimally chooses the weights.

In all three models we see that the biases and standard deviations of the individual estimators on average behave as expected: as the bandwidth increases, bias becomes more pronounced and the standard deviation declines. No kernel/bandwidth pair is the best in terms of RMSE among individual ones for all the models, although (K_4, h^{gcv}) is best for (s,s) and (s,m) and close to the best for (m,m).

The theoretical standard deviation (using the leading two components of $\text{Var}(\hat{\delta}_N)$ given in (5) compares very well with the standard deviation based on the bootstrap, where we note the importance of taking the kernel/bandwidth dependent component into account to ensure this close correspondence. The sample standard deviation still reveals a disparity (smaller for (s,s) and (s,m) versus larger for (m,m)) which might be the consequence of having set the number of simulations too low.

Table 1 shows that in terms of the RMSE of the ADE the combined estimator performs

[6] For the (s,s) model the bandwidths for the second and fourth order kernel were respectively $\left(\binom{1.1}{1.1}, \binom{1.8}{1.8}\right)$ for the nonparametric regression and $\left(\binom{1.0}{1.1}, \binom{1.1}{1.2}\right)$ for its derivative. For the (s,m) model they were $\left(\binom{1.1}{1.1}, \binom{1.7}{2.0}\right)$ and $\left(\binom{1.2}{1.0}, \binom{1.2}{1.1}\right)$ respectively, whereas for the (m,m) model they were $\left(\binom{1.1}{1.1}, \binom{1.9}{1.9}\right)$ and $\left(\binom{1.1}{1.0}, \binom{1.1}{1.1}\right)$ respectively.

Table 1: Density weighted ADE estimators

	$\binom{K_2,}{0.5h^{gcv}}$	$\binom{K_2,}{0.75h^{gcv}}$	$\binom{K_2,}{h^{gcv}}$	$\binom{K_4,}{0.5h^{gcv}}$	$\binom{K_4,}{0.75h^{gcv}}$	$\binom{K_4,}{h^{gcv}}$	**Combined**
Model 1 (s,s)							
RMSE	.0067	.0066	.0087	.0118	.0069	.0061	.0051
Bias	−.0008 −.0009	−.0013 −.0014	−.0032 −.0031	−.0000 −.0000	−.0005 −.0003	−.0015 −.0012	−.0006 −.0007
StdDev(T)	0.0041 0.0039	0.0029 0.0029	0.0025 0.0024	0.0083 0.0080	0.0045 0.0045	0.0035 0.0035	
StdDev(S)	0.0031 0.0030	0.0024 0.0024	0.0021 0.0021	0.0049 0.0050	0.0034 0.0032	0.0029 0.0029	0.0033 0.0034
StdDev(B)	0.0041 0.0039	0.0029 0.0028	0.0025 0.0024	0.0082 0.0081	0.0045 0.0045	0.0035 0.0035	0.0034 0.0033
Model 2 (s,m)							
RMSE	0.0142	0.0220	0.0343	0.0222	0.0172	0.0139	0.0119
Bias	−.0008 −.0009	−.0008 −.0009	−.0008 −.0009	−.0008 −.0009	−.0008 −.0009	−.0008 −.0009	−.0013 −.0021
StdDev(T)	0.0061 0.0104	0.0046 0.0070	0.0036 0.0043	0.0122 0.0171	0.0075 0.0130	0.0060 0.0095	
StdDev(S)	0.0047 0.0102	0.0037 0.0071	0.0030 0.0044	0.0081 0.0138	0.0055 0.0128	0.0048 0.0096	0.0060 0.0114
StdDev(B)	0.0061 0.0104	0.0046 0.0070	0.0036 0.0043	0.0122 0.0170	0.0075 0.0130	0.0060 0.0095	0.0063 0.0097
Model 3 (m,m)							
RMSE	0.0324	0.0557	0.0837	0.0469	0.0376	0.0340	0.0266
Bias	−.0118 −.0112	−.0379 −.0330	−.0584 −.0542	0.0118 0.0113	0.0078 0.0077	−.0147 −.0146	−.0047 −.0036
StdDev(T)	0.0183 0.0188	0.0108 0.0121	0.0060 0.0069	0.0301 0.0307	0.0239 0.0239	0.0166 0.0166	
StdDev(S)	0.0197 0.0177	0.0114 0.0118	0.0062 0.0063	0.0291 0.0247	0.0258 0.0238	0.0178 0.0161	0.0234 0.0199
StdDev(B)	0.0182 0.0187	0.0107 0.0120	0.0060 0.0068	0.0300 0.0306	0.0238 0.0237	0.0165 0.0165	0.0178 0.0185

better than the individual estimators in all cases. Where there is a clearly superior individual estimator it gets a higher weight on average and, 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. Specifically, the average weights of $((K_2, 0.75h^{gcv}), (K_2, h^{gcv}))$ and $((K_4, 0.75h^{gcv}), (K_4, h^{gcv}))$ have opposite signs in all models, and for (s,s) $(K_2, 0.75h^{gcv})$ gets a relatively large weight, for (s,m) so does (K_4, h^{gcv}), while for (m,m) $(K_2, 0.5h^{gcv})$ gets a large weight.[7]

In Table 2 the parameter estimates of the Tobit model are presented. Since the ADE allows for the estimation of $\beta = (\beta_1, \beta_2)^T$ up to scale, we report results of the parameter estimates of β_2 where β_1 is standardized to 1. For comparison, the Tobit MLE parameter estimates are reported as well. To ensure comparability with the semiparametric estimates, where β_1 is standardized to 1, we report $\hat{\beta}_2^{(t)}/\hat{\beta}_1^{(t)}$ for the Tobit regressions, where $\hat{\beta}^{(t)}$ are the Tobit parameter estimates (allowing for the estimation of an intercept β_0). Again the results are provided for each kernel/bandwidth pair selection as well as for the combined estimator.

When looking at table 2, we note that superiority in estimating ADE does not necessarily translate into better parameter estimators. If we judge performance on the RMSE, no individual estimator can be ranked to be the best in all models (and none is ranked above the combined estimator in all the models). The kernel bandwidth combination which is best for (s,s) and (m,m) is (K_2, h^{gcv}) compared to $(K_2, 0.5h^{gcv})$ for (s,m). Even though the combined estimator is not ranked best in RMSE sense in any of the models, its RMSE is relatively closer to the best individual estimator than the worst individual estimator. The same conclusions can be drawn if we judge the performance on the basis of absolute deviation of the mean of the individual estimator from the true value 1. In this case, $(K_2, 0.75h^{gcv})$ is best for (s,s) compared to $(K_4, 0.75h^{gcv})$ for (s,m) and (K_4, h^{gcv}) for (m,m), only the individual estimator $(K_4, 0.5h^{gcv})$ with this criterion is consistently worse than the combined estimator. A loss in efficiency arising from not knowing the distribution

[7] On average the weights are for (s,s) $(0, 34, 0.99, -0.42; 0.03, -0.47, 0.52)$, for (s,m) $(0.30, 0.39, -0.40; 0.05, -0.60, 1.27)$ and for (m,m) $(0.66, 1.58, -1.50; 0.07, -1.25, 1.44)$.

Table 2: Tobit Model: Single Index parameter estimates

	Parametric	Semiparametric, ADE based estimator						
	MLE	$\binom{K_2,}{0.5h^{gcv}}$	$\binom{K_2,}{0.75h^{gcv}}$	$\binom{K_2,}{h^{gcv}}$	$\binom{K_4,}{0.5h^{gcv}}$	$\binom{K_4,}{0.75h^{gcv}}$	$\binom{K_4,}{h^{gcv}}$	**Combined**
Model 1 (s,s)								
Mean	1.003	1.009	1.001	0.993	1.019	1.015	1.008	1.007
StdDev(T)	0.054	0.131	0.087	0.076	0.291	0.146	0.101	0.100
StdDev(S)	0.052	0.092	0.073	0.069	0.165	0.099	0.081	0.087
RMSE	0.054	0.131	0.087	0.076	0.292	0.147	0.101	0.100
Model 2 (s,m)								
Mean	1.018	0.844	0.777	0.680	0.818	0.910	0.803	0.812
StdDev(T)	0.070	0.168	0.127	0.092	0.261	0.196	0.148	0.147
StdDev(S)	0.064	0.155	0.126	0.092	0.196	0.181	0.144	0.167
RMSE	0.072	0.229	0.420	0.333	0.318	0.273	0.246	0.239
Model 3 (m,m)								
Mean	0.994	1.023	1.064	1.055	1.050	1.027	1.009	1.037
StdDev(T)	0.087	0.329	0.259	0.200	0.700	0.360	0.284	0.292
StdDev(S)	0.092	0.307	0.255	0.196	0.482	0.332	0.272	0.326
RMSE	0.087	0.330	0.335	0.265	0.702	0.361	0.284	0.294

of the disturbances occurs as expected, but is within reason; the standard deviation of the combined semiparametric estimator is less than double that of the Tobit MLE for (s,s). While the loss in efficiency arising from not knowing the distribution of the disturbances is more severe for (s,m) and (m,m), the potential gain from using the combined estimator over an incorrect kernel bandwidth combination is greater with non-smooth densities for the explanatory variables.

5. 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 do not use the theory of U statistics in the following lemma but obtain the moments by direct computation for symmetric as well as non-symmetric kernels.

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

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

where

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

for

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

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

$$\begin{aligned}
& E\left(\hat{f}'_{(K,h)}(x_i)\hat{f}'_{(K,h)}(x_i)^T y_i^2\right) \qquad (A.3)\\
= \; & \left(\tfrac{1}{N-1}\right)^2 E\left\{E_{z_i}\left(y_i^2\left[\sum_{j\neq i}\left(\tfrac{1}{h}\right)^{k+1}K'(\tfrac{x_i-x_j}{h})\right]\left[\sum_{j\neq i}\left(\tfrac{1}{h}\right)^{k+1}K'(\tfrac{x_i-x_j}{h})\right]^T\right)\right\}\\
= \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left[y_i^2E_{z_i}\left(K'(\tfrac{x_i-x_j}{h})K'(\tfrac{x_i-x_j}{h})^T I(\ i\neq j)\right)\right] +\\
& \tfrac{N-2}{N-1}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left[y_i^2E_{z_i}\left(K'(\tfrac{x_i-x_{j_1}}{h})K'(\tfrac{x_i-x_{j_2}}{h})^T I(\ i,i_1,i_2 \text{ pairwise distinct})\right)\right]\\
= \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{k+2}E\left[y_i^2\int K'(u)K'(u)^T f(x_i-uh)du\right] +\\
& \tfrac{N-2}{N-1}\left(\tfrac{1}{h}\right)^{2k+2}E\left[E_{z_i}\left(y_iK'(\tfrac{x_i-x_{j_1}}{h})\right)E_{z_i}\left(y_iK'(\tfrac{x_i-x_{j_2}}{h})\right)^T I(\ i,i_1,i_2 \text{ pairwise distinct})\right]\\
= \; & \tfrac{1}{N-1}\cdot\left(\tfrac{1}{h}\right)^{k+2}\left[Ey_i^2f(x_i)\mu_2(K)+O(h)\right] +\\
& \tfrac{N-2}{N-1}\cdot\left[E(f'(x_i)y_i)(f'(x_i)y_i)^T + O(h^{\overline{v}})\right],
\end{aligned}$$

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,

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

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 the term $I(i_1\neq i_2)$ in the terms of the expression.

Next follow details of derivation. Denote

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

$$B(K,h,x_i) = \int K'(u)K'(u)^T\left(f(x_i - uh) - f(x_i)\right)du.$$

$$C(K,h,x_i) = -\int K(u)\left[(gf)'(x_i + uh) - (gf)'(x_i)\right]du$$

$$D(K,h,x_i) = \int K'(u)K'(-u)^T\left[(gf)(x_i + uh) - (gf)(x_i)\right]du$$

$$c(x_i) = -(gf)'(x_i)$$

$$d(K,x_i) = \mu_2^*(K)(gf)(x_i)$$

$$\mu_2(K) = \int K'(u)K'(u)^T du$$

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

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

$$E_{z_j}\left[\left(\tfrac{1}{h}\right)^{k+1} K'(\tfrac{x_i - x_j}{h})y_i\right] = c(x_j) + C(K,h,x_j) \tag{A.5}$$

$$E_{z_i}\left[\left(\tfrac{1}{h}\right)^{k} K'(\tfrac{x_j - x_i}{h})K'(\tfrac{x_i - x_j}{h})^T y_j\right] = d(K,x_i) + D(K,h,x_i). \tag{A.6}$$

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 $Nh^{k+2} \approx O(1)$ or $Nh^{k+2} \to \infty$. 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 of Theorem 1.

Three main situations have to be dealt with in the proof.

From Lemma 1 it follows that the variance has two leading parts, one that converges 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 $O(N^{-2}h^{-(k+2)})$. The bias converges at the rate $O(h^{\overline{v}})$.

The first situation arises when the rate of the bias dominates the rates for both leading terms in the variance: cases (a) v. (correspondingly in (b)) and (c) iii.. By standard arguments this situation clearly results in convergence in probability to $\mathcal{B}(K)$ as stated in the Theorem.

The second situation refers to parametric rate of the variance dominating (with or without bias). For this case Theorem 3.3 in PSS applies. Since the proof in PSS is based on the theory of $U-$statistics we make the additional assumption of symmetry of the kernel function (see the comment in Serfling (1980, p.172) to which PSS refer in footnote 7 re symmetrization - it is not actually clear to me how this will help in the proof for a non-symmetric kernel).

The third situation is when the condition $Nh^{k+2} \to \infty$ is violated; note that if the degree of smoothness, $v < \frac{k+2}{2}$ this condition regardless of kernel order could hold only in the case when the bias dominates. This possibility $Nh^{k+2} \to 0$ was not examined in the literature previously. We thus need to provide the proof of asymptotic normality for cases (a) i. (corresponding (b)) and (c) i..

Consider $Nh^{k+2} \to 0$. Sketch of proof.

We shall say that x_i, x_j are $h-$close if $|x_i - x_j| < h$; here $|w|$ indicates the maximum of the absolute value of the components of vector w. In the sample of $\{x_1, ...x_N\}$ denote by A_s the set $\{x_i \,|$ exactly $s-1$ other x_j with $j > i$ are $h-$close to $x_i\}$. Then A_1 is the set of "isolated" x_i, that do not have any other $h-$close sample points, A_2 is the set of points with exactly one $h-$close point, etc. Clearly, $\bigcup_{s=1}^{N-1} A_s$ represents a partition of the sample for a given h.

Step 1 of proof. We show that a small enough h results in the probablity measure of $\bigcup_{s=3}^{N-1} A_s$ going to zero fast enough; this implies that most of the non-zero contribution into $\hat{\delta}$ comes from A_2 (since A_1 does not add non-zero terms).

Step 2. Consider A_2. The contribution from the $x's$ in this set to $\hat{\delta}$ reduces to the sum (recall symmetry of the kernel)

$$\frac{1}{N}\sum_{x_i\in A_2}\frac{1}{N-1}\left(\frac{1}{h}\right)^{k+1}K'(\frac{x_i-x_j}{h})(y_i-y_j).$$

Since in view of the result in step 1 the x_j that is $h-$close to x_i with high probability is in A_2, we consider

$$\hat{\delta}_{A_2}=\frac{-2}{N}\sum_{\substack{x_i,x_j\in A_2\\ i=1,..N-1;j=i+1,..N}}\frac{1}{N-1}\left(\frac{1}{h}\right)^{k+1}K'(\frac{x_i-x_j}{h})(y_i-y_j). \tag{A.8}$$

The terms in (A.8) are i.i.d. (note that where a pair x_i, x_j is not in A_2 the contribution to the sum is zero.) The second moments of these terms were derived in Lemma 1. ?Note that for cross-products only the terms of the form

$$\tfrac{(N-2)(N-3)}{(N-1)^2}\cdot\left(\tfrac{1}{h}\right)^{2k+2}E\left(y_{i_1}y_{i_2}K'(\tfrac{x_{i_1}-x_{j_1}}{h})K'(\tfrac{x_{i_2}-x_{j_2}}{h})^T I(j_1\neq j_2\neq i_1\neq i_2)\right)$$

are relevant since in A_2 the terms are independent??? or something of that sort?? so that the variance will reflect the rate.... To be continued....

Lemma 2. *Given Assumptions 1-4, the Var of $\hat{\delta}_N^*$ can be represented as*

$$\sum_{t_1,s_1}\sum_{t_2,s_2}a_{t_1,s_1}a_{t_2,s_2}Cov(\hat{\delta}_N(K_{t_1},h_{s_1}),\hat{\delta}_N(K_{t_2},h_{s_2}))\equiv\sum a_{j_1}a_{j_2}\Gamma_{j_1j_2}$$

where the covariance between $\hat{\delta}_N(K_{t_1}, h_{s_1})$ *and* $\hat{\delta}_N(K_{t_2}, h_{s_2})$, $\Gamma_{j_1 j_2}$, *is given by*

$$\begin{aligned}
\Gamma_{j_1 j_2} &= \tfrac{4}{N(N-1)} \left(\tfrac{1}{h_{s_2}}\right)^{k+1} \tfrac{1}{h_{s_1}} E\left[y_i^2 f(x_i) \mu_2(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2})\right] \\
&\quad + 4\tfrac{1}{N(N-1)} \left(\tfrac{1}{h_{s_2}}\right)^{k+1} \tfrac{1}{h_{s_1}} E\left[B(K_{t_1}, K_{t_2}, h_{s_1}, h_{s_2}, x_i) y_i^2\right] \\
&\quad + 4\tfrac{1}{N(N-1)} \left(\tfrac{1}{h_{s_2}}\right)^{k+1} \tfrac{1}{h_{s_1}} E\left[d(K_{t_1,K_{t_2}}, h_{s_1}/h_{s_2}, x_i) y_i + D(K_{t_1}, K_{t_2}, h_{s_1}, h_{s_2}, x_i) y_i\right] \\
&\quad + 4\tfrac{N-2}{N(N-1)} E((f'(x_i) y_i + A(K_{t_1}, h_{s_1}, x_i) y_i)(f'(x_i) y_i + A(K_{t_2}, h_{s_2}, x_i) y_i)^T) \\
&\quad + 4\tfrac{N-2}{N(N-1)} E\left[(c(x_i) + C(K_{t_1}, h_{s_1}, x_i))\,(c(x_i) + C(K_{t_2}, h_{s_2}, x_i))^T\right] \\
&\quad + 4\tfrac{N-2}{N(N-1)} E\left[(c(x_i) + C(K_{t_1}, h_{s_1}, x_i))\,(f'(x_i) y_i + A(K_{t_2}, h_{s_2}, x_i) y_i)^T\right] \\
&\quad + 4\tfrac{N-2}{N(N-1)} E\left[(f'(x_i) y_i + A(K_{t_1}, h_{s_1}, x_i) y_i)\,(c(x_i) + C(K_{t_2}, h_{s_2}, x_i))^T\right], \\
&\quad + \tfrac{-4N+6}{N(N-1)} E\left(\hat{\delta}_N(K_{t_1}, h_{s_1})\right) E\left(\hat{\delta}_N(K_{t_2}, h_{s_2})\right)^T \\
&\equiv \Sigma_{1\delta}(K_{t1}, K_{t2}, h_{s_1}/h_{s_2}) N^{-2} h_{s_2}^{-(k+1)} h_{s_1}^{-1} + \Sigma_{2\delta} N^{-1} + O(N^{-2})
\end{aligned}$$

where

$$\begin{aligned}
\mu_2(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) &= \int K'_{t_1}(u) K'_{t_2}(u \frac{h_{s_1}}{h_{s_2}})^T du \\
\mu_2^*(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) &= \int K'_{t_1}(u) K'_{t_2}(-u \frac{h_{s_1}}{h_{s_2}})^T du \\
B(K_{t_1}, K_{t_2}, h_{s_1}, h_{s_2}, x_i) &= \int K'_{t_1}(u) K'_{t_2}(u \frac{h_{s_1}}{h_{s_2}})^T \left(f(x_i - u h_{s_1}) - f(x_i)\right) du \\
d(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}, x_i) &= \mu_2^*(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2})(gf)(x_i) \\
D(K_{t_1}, K_{t_2}, h_{s_1}, h_{s_2}, x_i) &= \int y K'_{t_1}(u) K'_{t_2}(-u \frac{h_{s_1}}{h_{s_2}})^T \left((gf)(x_i + u h_{s_1}) - (gf)(x_i)\right) du,
\end{aligned}$$

and

$$\begin{aligned}
&\Sigma_{1\delta}(K_{t1}, K_{t2}, h_{s_1}/h_{s_2}) \\
&= 4E\left[y^2 f(x_i) \mu_2(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2}) + \mu_2^*(K_{t_1}, K_{t_2}, h_{s_1}/h_{s_2})(gf)(x_i) y_i\right]
\end{aligned}$$

Proof. TBC ■

Theorem 2:

Proof. Gaussianity is proved similarly to the proof of Theorem 1. If rates for two estimators differ, then the ratio $\frac{h_{s_1}}{h_{s_2}} \to \infty$ and due to finite support for the kernel the expressions for covariances converge to zero as $N \to \infty$. ■

References

[1] Donkers, B. and M. Schafgans (2005): "A method of moments estimator for semiparametric 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 (2004): "Robust kernel estimator for densities of unknown smoothness," mimeo, McGill University and CIREQ.

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

[7] 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.

[8] 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.

[9] Marron, J.S. and M.P. Wand (1992): "Exact Mean Integrated Squared Error," *Annals of Statistics*, 20, 712–736.

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

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

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

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

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

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

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

[17] 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.

[18] Stone, C.J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *Annals of Statistics*, 10, 1040–1053.