Adapting Kernel Estimation to Uncertain Smoothness

Yulia Kotlyarova* Marcia M.A. Schafgans†

Victoria Zinde-Walsh‡

Discussion paper
No. EM/2011/557
April 2011

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 6674

---

* Department of Economics, Dalhousie University, Halifax, Nova Scotia.

† Department of Economics, London School of Economics. Corresponding author: Address: Houghton Street, London WC2A 2AE; Phone: +44.208955.7487; E-mail: m.schafgans@lse.ac.uk.

‡ 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

For local and average kernel based estimators, smoothness conditions ensure that the kernel order determines the rate at which the bias of the estimator goes to zero and thus allows the econometrician to control the rate of convergence. In practice, even with smoothness the estimation errors may be substantial and sensitive to the choice of the bandwidth and kernel. For distributions that do not have sufficient smoothness asymptotic theory may importantly differ from standard; for example, there may be no bandwidth for which average estimators attain root-n consistency. We demonstrate that non-convex combinations of estimators computed for different kernel/bandwidth pairs can reduce the trace of asymptotic mean square error relative even to the optimal kernel/bandwidth pair. Our combined estimator builds on these results. To construct it we provide new general estimators for degree of smoothness, optimal rate and for the biases and covariances of estimators. We show that a bootstrap estimator is consistent for the variance of local estimators but exhibits a large bias for the average estimators; a suitable adjustment is provided.

***Keywords**:* Nonparametric estimation, kernel based estimator, combined stimator, variance bootstrap.

JEL Classification: C14.

1. Introduction

Kernel estimation is a widely used method of nonparametric estimation that is becoming more prevalent in empirical research, in part because of software applications in statistical packages such as Stata, R, and XploRE. It is used to estimate density functions, conditional means, variances and covariances, as well as higher order moments and their derivatives. Important functionals are averages of these functions, e.g., the average derivative of the conditional mean used in semiparametric estimation of single index models (Powell, Stock and Stoker, 1989). Subject to suitable smoothness conditions, this averaging permits a parametric convergence rate, despite nonparametric kernel estimators typically exhibiting a slower rate of convergence.

The applications of kernel estimation in the empirical literature are varied: (un)conditional variance and covariance kernel estimates are, e.g., used for estimation of volatilities and correlation in finance (Hafner and Linton, 2010 and Long, Su and Ullah, 2011) and testing for affiliation in auction models (Jun, Pinkse and Wan, 2010); kernel estimation of the conditional mean is used in the analysis of the effect of governance on growth (Huynh and Jacho-Chavez, 2009), trade costs (Henderson and Millimet, 2008), Engel curves (Blundell and Duncan, 1998), and estimation of distributional policy effects (Rothe, 2010 and DiNardo and Tobias, 2001); average derivative estimation is used to assess nonlinear pricing in labour markets (Coppejans and Sieg, 2005) and for consumer demand analysis (Härdle, Hildenbrand and Jerison, 1991 and Blundell, Duncan and Pendakur, 1998); a recent application in kernel density estimation is in the analysis of bank loan recovery rates in Italy (Calabrese and Zenga, 2010).

Implementation of kernel estimation methods requires the researcher to select a kernel function $K(\cdot)$ and bandwidth parameter h. These choices typically are based on asymptotic results for these estimators that rely on smoothness assumptions. E.g., the rule-of thumb plug-in method of bandwidth selection offered for univariate density estimation in Silverman (1986) assumes that the density has at least two continuous derivatives and specifies the use of a second order kernel. However, with enough smoothness, improvements in efficiency

can be obtained by using higher order kernels (see, e.g., Pagan and Ullah, PU, 1999 for discussion), and the optimal bandwidth that balances the squared bias and variance needs to be adapted to this degree of density smoothness. The use of higher order kernels and smoothness requirements are instrumental in allowing claims of a parametric rate over a range of bandwidth choices for average kernel based estimators; again, optimal bandwidth and kernel choices are dependent on the assumed smoothness. For the average density weighted derivative estimator (ADE) of Powell, Stock and Stoker (PSS, 1998), the density of the k covariates is assumed to possess at least $(k+6)/2$ continuous derivatives and a kernel of order $(k+4)/2$ is needed; the (direct) average derivatives estimator of Stoker (1991) necessitates smoothness assumptions both on the density and conditional moment $E(y|X)$ (specifically, the existence of at least $k+2$ continuous derivatives) in conjunction with the use of a kernel of order $k+2$.

The main theoretical purpose of the various smoothness conditions in the literature is to ensure that the kernel order determines the rate at which the bias of the estimator goes to zero (and thus to control this rate via the choice of kernel). Although the theoretical results that utilize the smoothness assumptions (including the selection of optimal kernel and bandwidth, e.g., Powell and Stoker, 1996) provide the appropriate asymptotics, finite sample behaviour of the estimators even when these assumptions are satisfied still exhibits significant variability depending on the actual underlying distributions and may be very sensitive to the bandwidth choice and the choice of kernel; these results are documented in many papers, including, e.g., Hansen (2005).

Our simulations confirm these results. The better performing bandwidths (oversmoothed or undersmoothed) and kernels (second or higher order kernel) differ depending on the underlying distribution. Moreover, this dependence (in finite samples) is not restricted to the theoretical smoothness properties and may be affected by much subtler properties of the underlying distribution (e.g., magnitude of derivatives). Frequently encountered functions and distributions, such as mixtures of normals, while satisfying the smoothness assumptions often exhibit very high values of derivatives that are more reminiscent of lack of smoothness (see, e.g., Marron and Wand, 1992). Specifically, the simulations reveal that

the root mean squared error (RMSE), for the ADE with distribution of the regressors that satisfy all smoothness assumptions can be as much as 4 to 10 times that obtained under the Gaussian density (see Table 1 in Schafgans and Zinde-Walsh, SZW, 2010).[1] The root mean integrated error for the univariate kernel density estimation of a mixture of normal distribution was 3 to 4 times that obtained under a Gaussian density; this observed error discrepancy for the mixture of normal density estimate was comparable to that observed for a non-smooth density (see Table 1 in Kotlyarova and Zinde-Walsh, KZW, 2007).

Clearly, estimators that adapt to unknown smoothness are warranted. The literature does provide some solutions to the bandwidth and kernel selection that explicitly takes account of uncertainty about the underlying smoothness. In an early paper Woodroofe, 1970 proposed to estimate the smoothness of a density function; his approach was not given much prominence in the research that followed where sufficient smoothness was instead assumed. SZW, 2010 recently successfully implemented his approach in the context of ADE. The advantage is that the selection of the tuning parameter reflects the estimated smoothness in an adaptive way, thereby enabling to approach the optimal rate in various cases. KZW, 2006 proposed a combined estimator that was adaptive to the unknown smoothness and that could achieve asymptotically the best available (a priori unknown) rate. SZW, 2010 make an argument (in the ADE case) that a combined estimator with appropriate selection of tuning parameters can outperform the estimator with optimal bandwidth not only in case of insufficient smoothness (as in KZW, 2006) but with sufficient smoothness as well.

In this paper we pursue further the agenda of robustifying nonparametric estimators against lack of smoothness by estimating consistently the optimal rate under unknown smoothness (extending the result of Woodroofe, 1970 and SZW, 2010 to the general case)

[1] For the ADE with underlying Gaussian density in the two regressors, denoted (s,s), the RMSE using the better performing fourth order kernel varied from from 0.08 to 0.15 for the range of bandwidths considered; for a similar selection of bandwidths the ADE with underlying mixture of normal densities provided RMSE ranges such as 0.43-0.53 in the (s,m) setting using the better performing fourth order kernel and 0.76-1.49 in the (s,d) setting using the better performing second order kernel (here d and m refer to different mixtures).

and combining estimators with different kernels and bandwidths to reduce sensitivity. The results are applied to two classes of estimators: to *local kernel based estimators* such as univariate and multivariate density, density derivatives, and (weighted) conditional moment and (weighted) derivatives of conditional moment estimators, and to *average kernel based estimators* such as average density, average (weighted) conditional moments, and average ((density)-weighted) derivative of conditional moment estimators. A non-exaustive list of estimators considered here is presented in Table A.1 in the Appendix. In Table A2, various relevant results are summarized; they clarify that optimal rates are determined by the kernel order only when there is sufficient smoothness of functions that drive the bias expansion; in the absence of sufficient smoothness the term with the parametric rate of convergence for average kernel based estimators may be dominated by terms depending on the kernel and bandwidth.[2] It should be noted that our framework is not limited to these estimators but incorporates other estimators such as the average outer product of the gradient and average hessian estimator considered in Samarov (1993) and Donkers and Schafgans (2008). A similar analysis applies to some extremum estimators such as the smoothed maximum score where a combined estimator was examined in KZW, 2010; other estimators such as conditional quantiles could also be studied within the same approach.

The performance of estimators with different tuning parameters and possibly based on different kernels is evaluated by means of the $trAMSE$; this refers to the trace of the leading term in the asymptotic expansion of MSE, or if the leading term is parametric we consider that term and the next expansion term that depends on the bandwidth. We show that even with knowledge of the optimal bandwidth there always exists a linear combination of estimators that has a smaller $trAMSE$ than that of the optimal estimator. This result exploits the fact that the distribution of an oversmoothed estimator is dominated by bias, and that (like for a jackknife) one can find weights that will give a zero leading bias term in the linear combination while reducing the variance. This result was presented in SZW,

[2] As Dalalyan et al. (2006) document, even when there is sufficient smoothness for parametric rates the choice of bandwidth and kernel affects second-order terms in MSE which are often not much smaller than first-order terms.

2010 for ADE; here we give a general (and corrected) version.

We illustrate the proposed approach by summarizing some simulation results that show the advantages of using the combined estimator, especially in situations where the estimation errors are large relative to the magnitude of the value being estimated.

Section 2 introduces notation and the assumptions underlying the classes of estimators. Section 3 examines the estimation of the smoothness via rate of the bias; this provides an estimated optimal bandwidth rate. Section 4 demonstrates the existence of linear combinations of kernel estimators with different bandwidths that can provide a smaller $trAMSE$ than with the optimal bandwidth by automatically removing the asymptotic bias and possibly reducing the asymptotic variance. Implementation of the "combined estimator" requires estimation of biases and covariances and is considered in section 5. Section 6 summarizes obtained simulation results for various estimators and demonstrates that the combined estimator offers significant advantages when insufficient smoothness results in very large relative errors.

2. Notation and Classes of Estimators

We assume that the data represent an i.i.d. sample of observations that could be given by $x_i \in R^k$ or $(y_i, x_i^T)^T$ where $y_i \in R$ is the dependent variable (y could be discrete, e.g. a binary variable) and $x_i \in R^k$ continuous explanatory variables.[3]

The two types of estimators we consider, *local kernel based estimators* and *average kernel based estimators,* involve the choice of a kernel K and bandwidth h such that $h \to 0$ and $N \to \infty$ and are generically denoted as $\hat{\delta}_N(K, h)$. The function, value of the function at a particular point (e.g. density), or a parameter vector that is being estimated is denoted δ_0, a notation also used in SZW, 2010. In Table A.1 in the Appendix relevant expressions of $\hat{\delta}_N(K, h)$ for each estimator are given.

[3]In some cases one can consider discrete regressors for which special kernels have been developed, e.g., see Racine and Li (2007). Härdle and Horowitz (1996) consider the ADE estimator in the presence of discrete regressors; they provide a separate noniterative estimator for the parameters of the discrete regressors.

The kernel function $K : R^k \to R$ is defined to have the order $v(K)$ and satisfies standard assumptions, e.g., PU, 1999. The kernel does not need to be symmetric; as argued in KZW, 2007, asymmetric functions may pick up some irregularities that will be discarded by symmetric smoothing functions (see also Abadir and Lawford, 2004).

The papers KZW, 2007 and SZW, 2010 have examined the behaviour of some of these estimators under relaxed smoothness conditions on the functions that drive the bias expansion; they demonstrate that the optimal rates are determined by the kernel order only when there is sufficient smoothness of these functions. Denote by $f(x)$ the density of x, and by $g(x)$ a conditional moment of interest (e.g. $g(x) = E(y|X = x)$ or $g(x) = E(y^r|X = x)$ for given $r > 0$). Assume that the support of the density of x is Ω (a convex (possibly unbounded) subset of R^k) with nonempty interior Ω_0 and $f(x) = 0$ for all $x \in \partial\Omega$, where $\partial\Omega$ denotes the boundary of Ω as in, e.g., Härdle and Stoker (1989) and PSS, 1989.

For any "smooth" function φ and $x \in R^k$ let $\varphi'(x)$ stand for the vector $(\partial\varphi(x)/\partial x_1, .., \partial\varphi(x)/\partial x_k)^T$, and $\varphi^{(m)}(x)$ denote an m^{th} partial derivative of $\varphi(x)$ given by $\partial^m\varphi(x)/(\partial^{m_1}x_1...\partial^{m_k}x_k)$, where $m_1 + ... + m_k = m$. We follow SZW, 2010 in formalizing the degree of smoothness of functions $\varphi(x)$, defined on some support Ω, in terms of the Hölder space of functions, $C_{m+\alpha}(\Omega)$, with integer m and $0 < \alpha \leq 1$, where any $\varphi(x) \in C_{m+\alpha}(\Omega)$ is m times continuously differentiable on Ω with $\varphi^{(m)}(\cdot)$, satisfying Hölder's condition of order α :

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

It can be said that $v = m + \alpha$ is the degree of smoothness of φ. Alternatively, the modulus of continuity could be used to indicate the degree of smoothness.

In the multivariate case it may be desirable to specify different smoothness conditions for the different components of a function, such as the density. To streamline exposition here we abstract from that possibility and assume the same smoothness conditions for all the components of a function. SZW, 2010 details the possible different treatment for smoothness with respect to the different components of x; the approach there can be extended to the other estimators considered in this paper. We consider estimators that use

the same bandwidth for all components.[4]

The two main high level assumptions of our estimator $\hat{\delta}(K,h)$, describing the bias and variance, are presented next.

Assume that the degree of smoothness of the functions that are relevant for the bias expansion of the estimator (such as the density $f(x)$ or its derivative $f'(x)$ and the conditional moment $g(x)$) is v; see the relevant assumptions in, e.g. PU, 1999. Denote by $B(K,h)$ the bias of the estimator $\hat{\delta}(K,h)$, $E(\hat{\delta}_N(K,h)-\delta_0)$, and define

$$\bar{v} = \min(v, v(K));$$

then in this notation for all the estimators that we consider (see Table A.1) we get $|B(K,h)| \leq \omega h^{\bar{v}}$. With insufficient smoothness, the rate at which the bias of the estimator goes to zero is not determined by the choice of the order of the kernel but by the degree of smoothness of appropriate functions. We make a stronger assumption on the bias, namely, that it is stabilized at this rate; this is the assumption made by Woodroofe (1970) for density estimation and is also made in SZW, 2010 for ADE.

Assumption 1. *As $N \to \infty$, $h \to 0$ and $h = O(N^{-\epsilon})$ with $\epsilon > \epsilon_L > 0$*

$$h^{-\bar{v}} bias(\hat{\delta}_N(K,h)) \to \mathcal{B}(K), \tag{1}$$

for some $\bar{v} > 0$, where the vector $\mathcal{B}(K) = (\mathcal{B}_1(K), ...\mathcal{B}_k(K))'$ is such that $0 < |\mathcal{B}_\ell(K)| < \infty$ for $\ell = 1, ..., k$.

The bound ϵ_L may be needed to ensure that the rest of the bias expansion converges to zero sufficiently fast.

The assumption on the variance below differs for local and averaged estimators:

Assumption 2. *As $N^\omega h^{d(k)} \to \infty$, $h \to 0$, for some $\omega \geq 0$, $d(k) \geq 1$*
(a) for local kernel based estimators: there is a finite positive definite matrix $\Sigma(K)$ such that

$$N^\omega h^{d(k)} var(\hat{\delta}_N(K,h)) \to \Sigma(K)$$

[4] We do not assume, however, that the optimal bandwidth is the same for all components.

(b) for average kernel based estimators: there exist finite positive definite matrices $\Sigma_1(K)$ *and* Σ_2 *such that an expansion for the variance is*

$$var(\hat{\delta}_N(K,h)) = N^{-\omega}h^{-d(k)}\left[\Sigma_1(K) + o(h^{\alpha})\right] + N^{-1}\left[\Sigma_2 + o(h^{\alpha})\right]$$

Conditions that guarantee this high level assumption include the existence of various second moments and continuity of $E(y^{2r}|x)$. The assumption on the variance holds for kernel density and conditional mean estimators where the asymptotic variance is of the form $(Nh^k)^{-1}\Sigma(K)$ and for m^{th} partial derivatives of kernel density with $(Nh^{k+m})^{-1}\Sigma(K)$. For the average kernel based estimators, this assumption highlights that there are two possible leading terms. Given sufficient smoothness, averaging can yield a parametric rate of convergence for a range of bandwidths; the non-parametric term could determine the overall rate in the case of insufficient smoothness or poor bandwidth rate choice; even when the parametric term dominates, the nonparametric term which depends on the kernel and bandwidth could be important in finite sample. For example, for the ADE PSS estimator the variance is expressed as $N^{-2}h^{-(k+2)}\left[\Sigma_1(K) + o(h^{\alpha})\right] + N^{-1}\left[\Sigma_2 + o(h^{\alpha})\right]$.

We summarize representative results about the estimators in Table A.2 in the Appendix. The table lists the rates of the leading terms in the AMSE expansion, the functions whose degree of smoothness is specified as v, the optimal rate that depends on $\bar{v}$ and may differ from standard under insufficient smoothness. As in SZW, 2010 the optimal rate is defined to balance the bandwidth dependent part in the expression in (b) with the bias, and provides the optimal rate when the parametric rate is not achievable because of lack of smoothness. The optimal bandwidth is defined to have the rate $N^{-\eta(\bar{v})}$ where $\eta(\bar{v}) \equiv \frac{\omega}{2\bar{v}+d(k)}$; for the estimators in Table A.2 ω can be 1 or 2 and $d(k)$ is k or $k+2$. When smoothness holds, the order of the kernel determines the asymptotic results that we list from the literature (see, e.g. PU, 1999 or Li and Racine, 2007). When smoothness assumption is violated we list (in the notation of this paper) the non-standard results from the KZW, 2007 and SZW, 2010 papers for density and ADE; the results for average density and for the conditional mean are obtained similarly. Similar results were also obtained for the SMS estimator (KZW, 2010). The general conclusion is that with insufficient smoothness the difference in

asymptotic performance may be substantial.

3. ESTIMATION OF ASYMPTOTIC RATE OF THE BIAS

From the table we can see that knowledge of $\bar{v}$ would allow one to find the optimal rate of the bandwidth that would give the smallest trace of asymptotic MSE. Under the Assumptions 1 and 2 $\bar{v}$ can be consistently estimated; this idea was applied by Woodroofe, 1970 to density estimation.

Denote by h_o some oversmoothed bandwidth. We assume that such a bandwidth can be obtained. For example, it would be provided by an "optimal" plug-in bandwidth computed on the basis of $v(K)$ rather than $\bar{v}$; such a bandwidth would provide oversmoothing if $\bar{v} < v(k)$; to cover the smooth case as well it could be magnified by some N^{ε} for a small $\varepsilon > 0$. In SZW, 2010 the generalized cross-validation bandwidth was used, since it is known to oversmooth in the ADE PSS case.

Define a sequence of bandwidths $\{h_t\}_{t=1}^{H}$ such that $h_t = c_t h_o N^{\gamma_t}$ for some $c_t > 0$; $0 \le \gamma_1 < ... < \gamma_H$ where γ_H is such that $h_H = c_H h_o N^{\gamma_H} \to 0$. E.g. for ADE if h_o is given by cross-validation that has the rate $N^{-\frac{1}{2\bar{v}+k}}$ select $\gamma_H < \frac{1}{2\bar{v}+k}$. Let $\mathcal{T}$ define a subset of all pairs $\{(h_t, h_{t'}),\ \ t, t' = 1, ...H$ with $t' < t\}$ with cardinality Q: $2 \le Q \le \frac{H(H+1)}{2}$.

Theorem 1. *Under Assumptions 1 and 2 the estimator for* $\bar{v}, \widehat{\bar{v}}$, *given by*

$$\widehat{\bar{v}} = \frac{\sum_{(t,t')\in\mathcal{T}} \ln\left[\left(\hat{\delta}_N(K,h_t) - \hat{\delta}_N(K,h_{t'})\right)^2\right] \cdot \left(\ln {h_t}^2 - \frac{1}{Q}\sum_{(t,t')\in\mathcal{T}} \ln {h_t}^2\right)}{\sum_{(t,t')\in\mathcal{T}} \left(\ln {h_t}^2 - \frac{1}{Q}\sum_{(t,t')\in\mathcal{T}} \ln {h_t}^2\right)^2}, \qquad (2)$$

satisfies $\widehat{\bar{v}} - \bar{v} = o_p((\ln N)^{-1})$. *A bandwidth vector with optimal rate is consistently estimated by* $\widehat{h^{opt}} = cN^{-\eta(\widehat{\bar{v}})}$.

Proof. The proof requires comparison of the asymptotic bias and variance contribution in the stochastic expansion of the estimator. It is essentially the same as that given in SZW, 2010 Theorem 3.3a; the only difference being that there specific γ_H and rate of h^{opt} are used. ■

In SZW, 2010 the constants were selected close to 1 but so as to ensure a spread of bandwidths for the given sample size.

4. ASYMPTOTIC OPTIMALITY OF LINEAR COMBINATIONS OF ESTIMATORS

It was argued in KZW, 2006 that linear combinations of estimators based on different bandwidths and kernels could provide the rate associated with the best of those estimators, where performance is evaluated in terms of minimizing the trace of the asymptotic MSE. Linear combinations of estimators typically used in the literature consider convex combinations; KZW, 2006 proposed using weights of different signs in the case of insufficient smoothness where bias is a prominent obstacle to reducing the estimation error.

To compute the trace of MSE for linear combinations of estimators in addition to the bias and variance of Assumptions 1 and 2, covariances of the estimators are needed. The covariances were derived for the density, SMS and ADE estimators in the respective papers KZW (2007,2010) and SZW (2010); in this paper they are summarized in the Appendix for the cases of conditional mean and average density as well.

SZW, 2010 gave a theoretical basis for combining estimators for the ADE: it was shown that there exists a linear combination of kernel estimators with different bandwidths such that it asymptotically outperforms the estimator that uses the optimal bandwidth. In the cases of ADE and average density when there is sufficient smoothness for the parametric term to determine the rate of AMSE, there is still an advantage in reducing the second term in the expansion of the variance and the result would still apply to the case of a parametric rate. For the cases of possibly parametric rates the theorem considers then the second order (bandwidth dependent) terms in the expansion. The theorem below provides this result in the general case; the proof in the Appendix details the general case and also corrects an inaccuracy in SZW, 2010.

Theorem 2. *Under the Assumptions 1 and 2 with* $\bar{v} \leq 2$*, for any kernel* K *and given an optimal bandwidth vector* h^{opt} *there exists a set of bandwidth vectors* $h_1, .., h_S$ *with* $h_s = c_s h^{opt}$ *for* $c_s > 1$*, and a corresponding set of weights,* $\{a_s\} : \sum_{s=1}^{S} a_s = 1$ *such that the*

linear combination, $\sum_{s=1}^{S} a_s \hat{\delta}_N(K, h_s)$ *provides*

$$trAMSE(\textstyle\sum_{s=1}^{S} a_s \hat{\delta}_N(K, h_s)) < trAMSE\left(\delta_N(K, h^{opt})\right). \tag{3}$$

Proof. See Appendix. ■

The proof gives a specific example of a set of bandwidths and weights that satisfy (3). The proof of this result relies on the fact that with weights of different signs the leading terms in the biases can be eliminated and the weights can be selected in a way that reduces the variance. One kernel is examined in the proof; more kernels would allow for more flexibility in the choice of bandwidths. This theorem could be modified (as in SZW, 2010) to account for unequal bandwidths for the different components of the vector $\hat{\delta}_N$.

The condition $\bar{v} \leq 2$ in the Theorem 2 holds if K is a second order kernel, and also for higher order kernels when bias goes to zero no faster than h^2. The proof can be modified to allow for higher $\bar{v}$, but we focus here on insufficient smoothness when the errors from a mistaken choice of bandwidth are substantial. It can be seen from the construction in the proof that a larger S allows more flexibility in the choice of the constants c_s that define the bandwidths.

5. Combined estimator: implementation

The theoretical results of the previous section give guidance for selection of estimators (corresponding to bandwidths indicated by Theorem 2) to include into the linear combination; in this section we discuss the issue of finding the coefficients that would minimize the trace of estimated MSE. This requires the estimation of the biases and covariances between the different estimators.

5.1. Bias estimation. The theorem below provides a consistent estimator for the asymptotic bias. The estimator uses the difference between an oversmoothed estimator, at a bandwidth h_o, that converges at the rate $h_o^{-\bar{v}}$ to the true parameter vector (δ_0) plus the asymptotic bias, and an undersmoothed estimator, at a bandwidth h_u, that converges to δ_0 plus a random variable that goes to zero at the rate $\left(N^{-\omega} h_u^{-d(k)}\right)^{\frac{1}{2}}$. The difference is

constructed in a way that the term $h_o^{\bar{v}}\mathcal{B}(K)$ dominates the difference and thereby provides a consistent bias estimator at h_o. Define h_o as $h_o = \widehat{h^{opt}}N^{\zeta}$, with $\max\{0, \zeta_L\} < \zeta < \zeta_H$ where $\zeta_L = \eta(\widehat{\bar{v}}) - \frac{1}{2\widehat{\bar{v}}}$; $\zeta_H = \eta(\widehat{\bar{v}})$ and $h_u = \widehat{h^{opt}}N^{-\xi}$, with $0 < \xi < \xi_H$ and $\xi_H = \frac{2\widehat{\bar{v}}\zeta}{d(k)}$. E.g. for ADE the appropriate choices were $\zeta_L = (1 - \frac{k+2}{2\widehat{\bar{v}}})\frac{1}{2\widehat{\bar{v}}+k+2}$, $\zeta_H = \frac{2}{2\widehat{\bar{v}}+k+2}$, and $\xi_H = \frac{2\widehat{\bar{v}}\zeta}{k+2}$.

Theorem 3. *A consistent estimator of the asymptotic bias for the oversmoothed estimator* $\hat{\delta}_N(K, h_o)$ *is provided by*

$$\widehat{bias}\hat{\delta}_N(K, h_o) = \hat{\delta}_N(K, h_o) - \hat{\delta}_N(K, h_u),$$

with h_o, h_u *defined above. A consistent estimator of the bias for* $\hat{\delta}_N(K, h)$ *with* $h \to 0$ *as* $N \to \infty$ *and* $h = O(N^{-\epsilon})$ *with* $\epsilon > \epsilon_L > 0$ *is given by*

$$h^{\widehat{\bar{v}}}h_o^{-\widehat{\bar{v}}}\widehat{bias}\hat{\delta}_N(K, h_o).$$

Proof. The proof requires comparison of the asymptotic bias and variance contribution in the stochastic expansion of the estimator and is the same as for Theorem 3.3b in SZW, 2010. ■

5.2. Covariance estimation. The covariances can be estimated by constructing appropriately consistent plug-in estimators for the leading terms in the asymptotic expansion of the covariances, or alternatively, by bootstrap. The Appendix provides the bootstrap derivations for the covariances. For validity of bootstrap, standard stronger moment assumptions such as boundedness of conditional fourth moments of $a(x, y_i)$ defined in the Appendix are required: of course, when $a(x, y_i)$ is bounded as for density, no additional conditions are needed.

Bootstrap for covariances of local estimators is straightforward; it is sketched in the Appendix.

For average estimators, estimating covariances by the resampling bootstrap leads to a significant bias in the nonparametric term in the bias expansion. Cattaneo et al. (2010) demonstrated this for the ADE. Here we derive a similar result for the non-derivative-based estimators, such as average density, average density weighted moments. For all the average

estimators considered, the leading non-parametric term in the bootstrap estimator is three times the leading non-parametric term of the variance of the estimator. Considering the fact that under our assumptions this term may well dominate the variance, this bias may be overwhelming. Of course, knowing this, we can correct by dividing the estimator by three. Even when the parametric part dominates, since for the purposes of our analysis only the nonparametric part matters for the trade-offs in the $trAMSE$ for the combined estimator, dividing the bootstrap covariance estimators by three is appropriate. Alternatively (as in Cattaneo et al., 2010), a bias correction would result if in bootstrap variance estimation the bandwidth h of $\hat{\delta}$ were replaced by $h_{var} = 3^{-\frac{1}{d(k)}}h$; this will automatically reduce the nonparametric part by a factor of 3.

Minimization of the estimated trace of MSE provides the weights for the different estimators in the linear combination.

6. PERFORMANCE OF THE COMBINED ESTIMATOR: SUMMARY OF THE EVIDENCE

In SZW, 2010 it was shown how linear combinations can automatically eliminate bias and perform better than bias corrected "optimal" bandwidth estimators. Their simulations, with sample size 1000, are summarized in Table 1.

Table 1: ADE - RMSE comparison.

Model	Best K/h	RMSE range, %	h^{opt}, K_4	Comb
(s,s)	K_4/h_3	$7.8 - 23.4$	8.5	9.6
(s,m)	K_4/h_0	$42.7 - 60.7$	49.5	56.1
(m,m)	K_2/h_0	$67.2 - 93.4$	81.1	86.9
(s,c)	K_4/h_2^{opt}	$44.4 - 49.9$	44.4	46.5
(s,d)	K_2/h_5^{gcv}	$76.6 - 153.8$	103.8	87.2
(c,d)	K_2/h_5^{gcv}	$47.9 - 105.4$	63.2	69.0

Here all the asymptotic conditions of PSS are satisfied and in theory all these estimators should be converging at a parametric rate. The models are represented by the underlying distributions of the two regressors: s (standard normal), m (trimodal normal mixture), c

(double claw) and d (discrete comb). For second-order terms a higher order kernel should be advantageous; the result should be stable over a range of bandwidths. The wide range of results here indicates that none of these conclusions are valid. By contrast, we see that the estimated optimal rate is not far from the best, which is an advantage; the combined estimator further improves where the errors are large.

For density estimation in KZW, 2007 the root mean integrated squared error, RMISE was evaluated over a range of bandwidths and kernels. The results from their simulations, with a sample size of 2000, are summarized in Table 2.

Table 2: Density estimation - RIMSE comparison.

Model	Best K (at h^{gcv})	RMISE range %	Comb24
normal	K_4	$2.4 - 2.8$	2.5
mixed normal	K_4	$6.5 - 7.1$	6.5
non-smooth	K_2	$6.8 - 6.7$	6.4

Here the error for normal mixture is much larger than for the Gaussian and is comparable to a non-smooth example. The combined estimator using two kernels with a range of bandwidths avoids the penalty associated with the incorrect choice and provides improvements over the best in problematic cases.

KZW, 2010 study a combined smoothed maximum score estimator. Whereas this estimator does not fit within the two classes of estimators considered, using the combined estimator provides similar benefits. With a sample size of 4000 their results are presented in Table 3. The estimator error depends very much on the selected kernel, with the 4th order kernel $f4$ not always the best, and sometimes the worst. Their results reveal, moreover, that two kernels of the same 4th order (labelled $f4$ and $g4$) may give strikingly different results even in the smooth case despite having the same asymptotic theory. The labels of the models which end with an H are heteroskedastic, the others homoskedastic; S stands for Gaussian model, M for mixture of normal, and NS for a non-smooth model. The combined estimator is often the best, or at least close to the best.

Table 3: SMS - RMSE comparison.

Model	Best K (at h^{opt} bias-corrected	RMSE range %	Comb
S	$f4$	$4.0 - 6.0$	4.7
SH	$f2$	$4.7 - 6.6$	4.9
M	$f4$	$2.8 - 4.1$	2.4
MH	$g4$	$1.3 - 2.6$	1.2
NS	$f4$	$9.6 - 14.6$	10.2
NSH	$f4$	$2.2 - 2.9$	2.2

7. CONCLUSIONS

We briefly summarize our findings here. Smoothness requirements lie at the heart of asymptotic properties of kernel based estimators. For distributions with insufficient smoothness, asymptotic theory may importantly differ from standard; for example; there may be no bandwidth for which average estimators attain root-N consistency. As we show, even for distributions such as mixtures of normals that deviate from Gaussian but still satisfy the assumptions for asymptotic efficiency of the estimator, the estimation errors may be substantial and very sensitive to the choice of the bandwidth and kernel.

To overcome these problems we propose an estimator that takes account of the (unknown) rate of the bias for any given kernel and combines estimators with different kernels and bandwidths. We estimate the bias rate and optimal bandwidth rate. We demonstrate that non-convex combinations of estimators computed for different kernel/bandwidth pairs can reduce the trace of asymptotic mean square error relative to the optimal kernel/bandwidth pair; we indicate that such combined estimators require some oversmoothed bandwidths relative to the estimated optimal rate to trade off the leading bias terms. To construct the combined estimator, weights that minimize the trace of estimated asymptotic mean square error need to be found; we provide estimators for the biases and covariances of our estimators using different kernels and bandwidths. We investigate the resampling bootstrap estimator for variances and show consistency for the class of local estimators.

For average estimators the resampling bootstrap exhibits a large bias that is thrice the nonparametric term in the variance expansion (under the insufficient smoothness conditions we are concerned that this term may easily be the leading one); the finding is similar to what Cattaneo et al. (2010) found for the ADE estimator and extends their result to other average estimators. With suitable adjustments the bootstrap variance estimator can be used in the procedure for the combined estimation.

8. APPENDIX

We provide the results for average density and the Nadaraya-Watson estimator that confirm that they can satisfy Assumption 1 and provide the covariances. The results for the covariances can be adapted easily to allow for unequal bandwidths for the different components; the derivations are similar to those in SZW and are omitted.

Average Density Estimator.

Consider the average density estimator:

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

We have

$$\begin{aligned} E(\hat{\delta}_N(K,h) &= h^{-k} E\left[E(K(\frac{x_i - x_j}{h})|x_j)\right] \\ &= h^{-k} E\left[\int K(\frac{x_i - x_j}{h}) f(x_i) dx_i\right] \\ &= E\left[\int K(u) f(x_j + uh) du\right] = E(f(x)) + h^{\bar{v}} \mathcal{B}(K) + o(h^{\bar{v}}) \end{aligned}$$

and the covariance is provided in Lemma 1.

Lemma 1. *Under Assumptions 1 and 2, if* $h_s \to 0$ *and* $N^2 h_s^k \to \infty$ *for* $s = 1,..,S$, *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* $s_1, s_2 = 1,..,S$ *is*

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

with

$$\Sigma_1(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) = 2E\left[f(x_i)\right]\kappa_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2});$$
$$\kappa_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) = \int K_{s_1}(u)K_{s_2}(u\frac{h_{s_1}}{h_{s_2}})du; \text{ and}$$
$$\Sigma_2 = 4E\left[(f(x_i) - Ef(x_i))^2\right].$$

Conditional Mean Estimator.

Consider the Nadarya Watson Kernel regression estimator:

$$\hat{\delta}_N(K, h, x) = \hat{g}(x) = \frac{\frac{1}{Nh^k}\sum_{i=1}^N K(\frac{X_i - x}{h})Y_i}{\frac{1}{Nh^k}\sum_{i=1}^N K(\frac{X_i - x}{h})}$$

Following the notation in Li and Racine, 2007 (page 61),

$$\hat{g}(x) - g(x) = \frac{(\hat{g}(x) - g(x))\,\hat{f}(x)}{\hat{f}(x)} \equiv \frac{\hat{m}(x)}{\hat{f}(x)} = \frac{\hat{m}_1(x) + \hat{m}_2(x)}{\hat{f}(x)} = \frac{\hat{m}_1(x) + \hat{m}_2(x)}{f(x) + o_p(1)}$$

where

$$Y_i = g(X_i) + u_i$$
$$\hat{m}_1(x) = \frac{1}{Nh^k}\sum_{i=1}^N (g(X_i) - g(x))K(\frac{X_i - x}{h}); \; \hat{m}_2(x) = \frac{1}{Nh^{-k}}\sum_{i=1}^N u_i K(\frac{X_i - x}{h})$$

Assuming that $f(x) > 0$, $\hat{g}(x) - g(x) = O_p\left(\frac{\hat{m}(x)}{f(x)+o_p(1)}\right)$.

$$\begin{aligned}
E(\hat{m}_1(x)) &= h^{-k}\left[E((g(z) - g(x))K(\frac{z - x}{h})\right] \\
&= h^{-k}\left[\int (g(z) - g(x))f(z)K(\frac{z - x}{h})dz\right] \\
&= \left[\int f(x + uh)(g(x + uh) - g(x))K(u)du\right] = h^{\bar{v}}\mathcal{B}_m(K, x) + o(h^{\bar{v}}) \\
E(\hat{m}_2(x)) &= 0
\end{aligned}$$

we have $E\left(\hat{g}(x) - g(x)\right) = h^{\bar{v}}\mathcal{B}_m(K, x)/f(x) + o(h^{\bar{v}})$ and $B(K, x) = B_m(K, x)/f(x)$. The covariance is provided in Lemma 2.

Lemma 2. *Under Assumptions 1 and 2, if* $h_s \to 0$ *and* $Nh_s^k \to \infty$ *for* $s = 1, .., S$, *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* $s_1, s_2 = 1, .., S$ *is*

$$\Gamma_{s_1,s_2} = N^{-1} h_{s_2}^{-k} \left(\Sigma_1(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) + o(1) \right),$$

with

$$\Sigma_1(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) = \frac{\sigma^2(x)}{f(x)} \kappa_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}); \text{ and}$$
$$\kappa_2(K_{s_1}, K_{s_2}, h_{s_1}, h_{s_2}) = \int K_{s_1}(u) K_{s_2}(u \frac{h_{s_1}}{h_{s_2}}) du.$$

Tables.

Table A.1: Local and Average Kernel based estimators.

Estimator		$\hat{\delta}_N(K,h)$
Local		
Density	$\hat{\delta}_f(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-k} K(\frac{x_i - x}{h})$
Derivative of Density	$\hat{\delta}_{f'}(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-(k+1)} K'(\frac{x_i - x}{h})$
Partial Derivative of Density	$\hat{\delta}_{f^{(m)}}(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-(k+m)} K^{(m)}(\frac{x_i - x}{h})$
Conditional moment (CM)[5]	$\hat{\delta}_g(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-k} K(\frac{x_i - x}{h}) w(x) \hat{\delta}_f(x)^{-1} y_i^r$
Density weighted CM	$\hat{\delta}_{fg}(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-k} K(\frac{x_i - x}{h}) y_i^r$
Derivative of CM	$\hat{\delta}_{g'}(x) =$	$\frac{1}{N}\sum_{i=1}^{N} h^{-(k+1)} K'(\frac{x_i - x}{h}) w(x) \hat{\delta}_f(x)^{-1} \left[y_i^r - \hat{\delta}_g(x) \right]$
Average		
Density	$\hat{\delta}_{Ef} =$	$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-k} K(\frac{x_i - x_j}{h})$
Derivative of Density	$\hat{\delta}_{Ef'} =$	$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-(k+1)} K'(\frac{x_i - x_j}{h})$
Partial Derivative of Density	$\hat{\delta}_{Ef^{(m)}} =$	$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-(k+m)} K^{(m)}(\frac{x_i - x_j}{h})$
Conditional moment	$\hat{\delta}_{Ewg} =$	$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-k} K(\frac{x_i - x_j}{h}) w(x_j) \hat{\delta}_f(x_j)^{-1} y_i^r$
Derivative of CM (direct)	$\hat{\delta}_{Eg'} =$	$\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-(k+1)} K'(\frac{x_i - x}{h}) w(x_j) \hat{\delta}_f(x_j)^{-1} \left[y_i^r - \hat{\delta}_g(x_j)) \right]$
Derivative of CM (indirect)	$\hat{\delta}_{E\left(-\frac{f'}{f}y\right)} =$	$-\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-(k+1)} K'(\frac{x_i - x_j}{h}) w(x_j) \hat{\delta}_f(x_j)^{-1} y_i^r$
Density Weighted Deriv. of CM (indirect)	$\hat{\delta}_{Efg'} =$	$-\frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N} h^{-(k+1)} K'(\frac{x_i - x_j}{h}) y_i^r$

[5] Here (and elsewhere) $w(\cdot)$ could be a weight to insure that the denominator is separated from zero.

Table A.2: AMSE rates

Estimated δ functions in smoothness assumption	**AMSE terms**	**optimal rate** $h=h^{opt}(\bar{v})$	**smooth** $h=h^{opt}(v(K))$	**non smooth** $h=h^{opt}(v(K))$
Density $f(x)$	$h^{2\bar{v}}B(K)B(K)^T$ $+(Nh^k)^{-1}\Sigma(K)$	$h^{opt} =$ $O(N^{-\frac{1}{2\bar{v}+k}})$	$\Rightarrow$ Gaussian MSE rate: $N^{-\frac{2v(K)}{2v(K)+k}}$ (PU, 1999)	$\hat{\delta}(K,h)-\delta_0 = h^{\bar{v}}(B(K)+o_p(1))$ MSE rate: $h^{2\bar{v}}$ (KZW,07)
Cond'l Mean $f(x), g(x)$	$h^{2\bar{v}}B(K)B(K)^T$ $+(Nh^k)^{-1}\Sigma(K)$	$h^{opt} =$ $O(N^{-\frac{1}{2\bar{v}+k}})$	$\Rightarrow$ Gaussian MSE rate: $N^{-\frac{2v(K)}{2v(K)+k}}$ (PU, 1999)	$\hat{\delta}(K,h)-\delta_0 = h^{\bar{v}}(B(K)+o_p(1))$ MSE rate: $h^{2\bar{v}}$
Avg Density $f(x)$	$h^{2\bar{v}}B(K)B(K)^T$ $+N^{-2}h^{-k}\Sigma_1(K)$ $+N^{-1}\Sigma_2$	$h^{opt} =$ $O(N^{-\frac{2}{2\bar{v}+k}})$	$\Rightarrow$ Gaussian MSE rate: N^{-1}	Either $\hat{\delta}(K,h)-\delta_0 = h^{\bar{v}}(B(K)+o_p(1))$ MSE rate: $h^{2\bar{v}}$, or $var(\hat{\delta}(K,h))$ $= N^{-2}h^{-k}(\Sigma_1(K)+o(1))$ MSE rate:$N^{-2}h^{-k}$
ADE (PSS) $f'(x)$	$h^{2\bar{v}}B(K)B(K)^T$ $+N^{-2}h^{-(k+2)}\Sigma_1(K)$ $+N^{-1}\Sigma_2$	$h^{opt} =$ $O(N^{-\frac{2}{2\bar{v}+k+2}})$	$\Rightarrow$ Gaussian MSE rate: N^{-1} (PU, 1999)	Either $\hat{\delta}(K,h)-\delta_0 = h^{\bar{v}}(B(K)+o_p(1))$ MSE rate: $h^{2\bar{v}}$, or $var(\hat{\delta}(K,h))$ $= N^{-2}h^{-(k+2)}(\Sigma_1(K)+o(1))$ MSE rate:$N^{-2}h^{-(k+2)}$ (SZW,10)

Proof of Theorem 2.

To provide a proof it is sufficient to give a set of bandwidths and a corresponding set of weights, such that the leading bias terms will cancel out and the variance will not inflate.

Consider each i^{th} component of $\hat{\delta}_N(K, h_s)$ separately and suppress the subscript i.

We start by finding for any given set of bandwidths, $h_s, s = 1, ...S$ the weights, a_s, such that they sum to one, eliminate the leading bias term and give a vector with the smallest norm.

To do this solve

$$\min \Sigma_{s=1}^{S} a_s^2, \text{ subject to } \Sigma_{s=1}^{S} a_s = 1; \Sigma_{s=1}^{S} a_s h_s^{\bar{v}} = 0, \tag{A.1}$$

noting that $\Sigma_{s=1}^{S} a_s \mathcal{B}_i(K) h_s^{\bar{v}} = 0$ implies $\Sigma_{s=1}^{S} a_s h_s^{\bar{v}} = 0$. Denoting $h_s^{\bar{v}}$ by b_s, the Lagrangean is

$$\Sigma_{s=1}^{S} a_s^2 - \lambda(\ \Sigma_{s=1}^{S} a_s - 1) - \theta \Sigma_{s=1}^{S} a_s b_s.$$

From the FOC, we obtain

$$\lambda = 2\Sigma a_s^2; \theta = \frac{2 - 2S\Sigma a_s^2}{\Sigma b_s}; \text{ and } a_s = \frac{1}{2}(\lambda + \theta b_s).$$

Denoting Σa_s^2 by α, we obtain $a_s = \alpha + \frac{1 - S\alpha}{\Sigma b_s} b_s$. By squaring and summing a_s for $s = 1, ...S$**,** we get

$$\alpha = S\alpha^2 + 2\alpha(1 - S\alpha) + (1 - S\alpha)^2 \frac{\Sigma b_s^2}{(\Sigma b_s)^2}.$$

This quadratic equation for α has a root of $\frac{1}{S}$ as a solution to the FOC, the other root is $\alpha = \frac{\Sigma b_s^2}{S\Sigma b_s^2 - (\Sigma b_s)^2}$; this provides the general form for α.

With such weights the trace AMSE of a linear combination reduces to the trace of the asymptotic variance of the linear combination, that includes the covariance terms. Denote $\hat{\delta}_N(K, h_s)$ by $\hat{\delta}_s$, by Σ_{ii} the i^{th} diagonal element of the $N^{-\omega}\Sigma(K)$ matrix in Assumption

2(a) or of $N^{-\omega}\Sigma_1(K)$ in Assumption 2(b), whichever is appropriate.

$$\begin{aligned}
& tr(Avar(\textstyle\sum_{s=1}^{S} a_s\hat{\delta}_s)) = \sum_{s_1,s_2}^{S} a_{s_1}a_{s_2} \sum_{i=1}^{k} Acov(\hat{\delta}_{s_1,i}, \hat{\delta}_{s_2,i}) \\
\leq\ & \sum_{s_1,s_2}^{S} |a_{s_1}a_{s_2}| \sum_{i=1}^{k} \left| Acov(\hat{\delta}_{s_1,i}, \hat{\delta}_{s_2,i}) \right| \\
\leq\ & \sum_{s_1,s_2}^{S} |a_{s_1}a_{s_2}| \sum_{i=1}^{k} \max_{j=1,2} Avar(\hat{\delta}_{s_j})_i \\
=\ & \sum_{s_1,s_2}^{S} |a_{s_1}a_{s_2}| \max_{j=1,2} h_{s_j}^{-d(k)} \sum_{i=1}^{k} \Sigma_{ii} \\
\leq\ & \sum_{s_1,s_2}^{S} |a_{s_1}a_{s_2}| \left(\min_{j=1,2,i=1,..,k} c_{s_j,i} \right)^{-d(k)} \sum_{i=1}^{k} \left(h_i^{opt}\right)^{-d(k)} \Sigma_{ii} \\
\leq\ & \left(\min_{s,i} c_{s,i} \right)^{-d(k)} S\alpha \cdot tr\left(Avar\hat{\delta}^{opt} \right).
\end{aligned}$$

Here the second inequality is the usual bound for covariance via variances, then Assumption 2 for the variance is used. The values of the selected bandwidths $(c_{s_j,i}h_i^{opt}) = h_{s_j}, j = 1, 2$ are substituted: note that the optimal bandwidth has components with the same rates but may vary in the constant since $\mathcal{B}_i(K)$ may differ for different i, so $c_{s,i}h_i^{opt} = c_{s,1}h_1^{opt}$; to bound we use the smallest of all the $c_{s,i}$. Then we also use the bound

$$\begin{aligned}
\Sigma_{s_1,s_2=1}^{S} |a_{s_1i}a_{s_2i}| &\leq S \left(\Sigma_{s_1,s_2=1}^{S} |a_{s_1i}|^2\right)^{1/2} \left(\Sigma_{s_1,s_2=1}^{S} |a_{s_2i}|^2\right)^{1/2} \\
&= S\alpha^{1/2}\alpha^{1/2} = S\alpha.
\end{aligned}$$

Recall that $\alpha = \frac{\Sigma b_s^2}{S\Sigma b_s^2 - (\Sigma b_s)^2}$. Superiority of the combination will follow if we can show that there exist $c_{s,i}$ such that

$$(\min c_{s,i})^{-d(k)} S \frac{\Sigma c_{s,i}^{2\bar{v}}}{S\Sigma c_{s,i}^{2\bar{v}} - \left(\Sigma c_{s,i}^{\bar{v}}\right)^2} < 1.$$

Suppose that h_1^{opt} is the largest among the components of the optimal bandwidth. Then $c_{s,1}$ is the smallest among $c_{s,i}$ for a fixed s. Then it is sufficient to show

$$(\min c_{s,1})^{-d(k)} S \frac{\Sigma c_{s,1}^{2\bar{v}}}{S\Sigma c_{s,1}^{2\bar{v}} - \left(\Sigma c_{s,1}^{\bar{v}}\right)^2} < 1.$$

Equivalently, (dropping the subscript 1)

$$\left((\min c_s)^{-d(k)} - 1\right) S\Sigma c_s^{2\bar{v}} + \left(\Sigma c_s^{\bar{v}}\right)^2 < 0. \tag{A.2}$$

This is monotone in $d(k)$ and $d(k) \geq 1$. Thus (A.2) would hold for any $d(k)$ as long as it holds for $d(k) = 1$. Thus set $d(k) = 1$; define $c_s = (1+x)^{\frac{s}{v}}$, $x > 0$. Then

$$\begin{aligned}
&\left((\min c_s)^{-1} - 1\right) = (1+x)^{-\frac{1}{v}} - 1 = -e(x,v) < 0 \text{ for } \nu > 0; \\
&\Sigma_{s=1}^{S} c_s^{2\bar{v}} = \Sigma_{s=1}^{S} (1+x)^{2s} = (1+x)^2 \tfrac{\left[(1+x)^{2S}-1\right]}{(1+x)^2-1} \\
&\Sigma_{s=1}^{S} c_s^{\bar{v}} = \Sigma_{s=1}^{S} (1+x)^{s} = (1+x) \tfrac{\left[(1+x)^{S}-1\right]}{(1+x)-1} = (1+x) \tfrac{\left[(1+x)^{S}-1\right]}{x}.
\end{aligned}$$

Substituting these expressions into (A.2)) yield

$$\begin{aligned}
& (1+x)^2 \tfrac{-e(x,v)Sx\left((1+x)^S-1\right)\left((1+x)^S+1\right)+(x+2)\left((1+x)^S-1\right)^2}{x^2(x+2)} \\
= \; & (1+x)^2 \left((1+x)^S - 1\right) \tfrac{-e(x,v)Sx\left((1+x)^S+1\right)+(x+2)\left((1+x)^S-1\right)}{x^2(x+2)}.
\end{aligned}$$

With $x > 0$, and denoting $\kappa(x,v,S) = -e(x,v)S(1+x)^2 \frac{\left[(1+x)^{2S}-1\right]}{x(x+2)} + (1+x)^2 \frac{\left[(1+x)^S-1\right]^2}{x^2}$, we need to prove $\kappa(x,v,S) < 0$. Equivalently, we show

$$-e(x,v)Sx\left((1+x)^S+1\right) + (x+2)\left((1+x)^S-1\right) \equiv \tilde{\kappa}(x,v,S) < 0$$

with $\tilde{\kappa}(x,v,S)$ increasing in v :

$$\frac{\partial \tilde{\kappa}}{\partial v} = \frac{2}{\nu^2} \ln(x+1)(x+1)^{-\frac{2}{\nu}} Sx\left[(1+x)^S+1\right] > 0.$$

For $v = 2$ we get $e(x,v) = 1 - (1+x)^{-\frac{1}{2}}$ and for $\tilde{\kappa}(x,v,S)$ to be negative, we need $(x+2)\left[(1+x)^S - 1\right] < \left(1-(1+x)^{-\frac{1}{2}}\right) S\left[(1+x)^S+1\right]$. This inequality will be true if $x+2 \leq x\left(1-(1+x)^{-\frac{1}{2}}\right)S$, or $S \geq \frac{(x+2)}{x\left(1-(1+x)^{-\frac{1}{2}}\right)}$. For example, for $x = 2, S = 5$.

This demonstrates that for any kernel there exists a set of bandwidths that in a linear combination removes the leading term of the asymptotic bias while reducing the variance. ■

Bootstrap for the covariances.

Here we sketch the results for the bootstrap estimators. In the case of local estimators, bootstrap estimators of the covariances are straightforward and suitable for obtaining the weights for the combined estimator. The averaged case presents some extra problems. It was examined in detail for ADE by Cattaneo et al. (2010), who have shown that in that

case the bootstrap estimator for the bandwidth dependent part of the variance is biased. Here we consider some of the other averaged estimators, e.g. averaged density, and indicate that similar results hold.

First, note that it is sufficient to examine variances.

Due to the linear structure of local and average kernel based estimators, a linear combination of such estimators, $a_1\hat{\delta}(K_{s_1},h_1)+a_2\hat{\delta}(K_{s_2},h_2)$, can be represented as yet another estimator, $\hat{\delta}_N(K,h)$. Consider $h_{s_2}/h_{s_1}=d$; $h_{s_1}=h$, then for $a_1\hat{\delta}(K_{s_1},h)+a_2\hat{\delta}(K_{s_2},dh)$ write

$$h_{s_2}^{-k}K_{s_2}(\frac{x_i-x}{h_{s_2}})=h^{-k}d^{-k}K_{s_2}(d^{-1}\frac{x_i-x}{h})=h^{-k}\bar{K}_{s_2}(\frac{x_i-x}{h});$$

similarly, for derivatives, e.g.,

$$\begin{aligned} h_{s_2}^{-(k+1)}K'_{s_2}(\frac{x_i-x}{h_{s_2}}) &= h^{-(k+1)}d^{-(k+1)}K'_{s_2}(d^{-1}\frac{x_i-x}{h}) \\ &= h^{-(k+1)}\bar{K}'_{s_2}(\frac{x_i-x}{h}), \end{aligned}$$

where $\bar{K}_{s_2}(w)=d^{-k}K_{s_2}(d^{-1}w)$. Then linear combinations $a_1\hat{\delta}(K_{s_1},h)+a_2\hat{\delta}(K_{s_2},dh)$ become $\hat{\delta}_N(K,h\ )$ for $K=a_1K_{s_1}+a_2\bar{K}_{s_2}$.

This means that in order to prove validity of bootstrap for covariance we only need to prove validity of bootstrap for the variance $var\hat{\delta}_N(K,h\ )$ for kernels and estimators that satisfy assumptions; the covariance $cov\left(\hat{\delta}_N(K_{s_1},h_{s_1}),\hat{\delta}_N(K_{s_2},h_{s_2})\right)$ can be expressed as $2var\hat{\delta}_N(K,h)-\frac{1}{2}\left[var\hat{\delta}_N(K_{s_1},h_{s_1})+var\hat{\delta}_N(K_{s_2},h_{s_2})\right]$ where $K=\frac{1}{2}\left(K_{s_1}+\bar{K}_{s_2}\right)$ is a kernel that satisfies assumptions on the kernel with order that is the lower of the two.

The subscript by N denotes the moments of the empirical distribution.

We consider the bootstrap variance in the following three settings.

I. Density, density weighted conditional moments at a point.

$$\hat{\delta}_I\equiv\hat{\delta}_I(x)=N^{-1}\Sigma_i a(x,y_i)h^{-k}K\left(\frac{x_i-x}{h}\right),$$

where for density $a(x,y_i)=1$, for density weighted conditional moment $a(x,y_i)=y_i^r$. For conditional moment $a(x,y_i)=\frac{1}{\hat{f}(x)}y_i^r$ and would require dealing with the

denominator; here we abstain from the estimation of a possible denominator, which vanishes when density weighting is used.

II. Order m derivatives of density, density weighted conditional moment at a point. Define for any vector of integers $(p) = (p_1, ...p_k)$, $p_1 + ... + p_k = p$, the operator $\partial^{(p)}$ applied to a differentiable function $q(x)$ as $\partial^{(p)} q(x) = \frac{\partial^p}{\partial x_1^{p_1} ... \partial x_k^{p_k}} q(x)$. Then

$$\begin{aligned} \hat{\delta}_{II} &\equiv \hat{\delta}_{II}(x) = \sum_{\text{all } (m)} \hat{\delta}_{II,(m)}; \\ \hat{\delta}_{II,(m)} &\equiv \hat{\delta}_{II,(m)}(x) = N^{-1} \Sigma_i a(x, y_i) h^{-(k+m)} \partial^{(m)} K\left(\frac{x_i - x}{h}\right), \end{aligned}$$

where for density $a(x, y_i) = 1$, for density weighted conditional moment $a(x, y_i) = y_i^r$.

III. Averages. Write each of the estimators in I and II as $\frac{1}{N}\Sigma_i \hat{\delta}_i(x)$, then the average estimator is

$$\hat{\delta}_{III} = \frac{1}{N(N-1)} \Sigma_{j=1}^N \Sigma_{i \neq j} \hat{\delta}_i(x_j).$$

In the following we restrict ourselves to looking at moments under the empirical distribution since the discrepancy with the bootstrap estimated moments can be controlled by suitably choosing the number of bootstraps.

Case I. Consider the bootstrapped estimator defining $\hat{\delta}_i \equiv h^{-k} a(x, y_i) K\left(\frac{x_i - x}{h}\right)$:

$$\hat{\delta}_I^* = \frac{1}{N} \Sigma_{i^*=1}^N \hat{\delta}_{i^*} = \frac{1}{N} \Sigma_{i^*=1}^N h^{-k} a(x, y_{i^*}) K\left(\frac{x_{i^*} - x}{h}\right)$$

and denoting $a(x, y_i)$ by a_i, $K\left(\frac{x_i - x}{h}\right)$ by K_i

$$\hat{\delta}_I^* = \frac{1}{N} \Sigma_{i^*=1}^N h^{-k} a_{i^*} K_{i^*}.$$

Then

$$\begin{aligned} E_N \hat{\delta}_I^* &= \left[\frac{1}{N} \Sigma_{i^*=1}^N E_N h^{-k} a_{i^*} K_{i^*}\right] = \frac{1}{N} \Sigma_{i=1}^N h^{-k} a_i K_i \qquad \text{(A.3)} \\ &= \frac{1}{N} \Sigma_{i=1}^N \hat{\delta}_i = \hat{\delta}_I \end{aligned}$$

Next consider the empirical variance of our bootstrap estimator:

$$var_N\hat{\delta}_I^* = E_N\hat{\delta}_I^{*2} - \left(E_N\hat{\delta}_I^*\right)^2.$$

Clearly

$$\hat{\delta}_I^{*2} = N^{-2}\Sigma_{i^*=1}^N\hat{\delta}_{i^*}^2 + N^{-2}\Sigma_{i_1^*=1}^N\Sigma_{i_2^*\neq i_1^*}^N\hat{\delta}_{i_1^*}\hat{\delta}_{i_2^*}.$$

Taking the empirical moment of the first term on the right hand side yields

$$\begin{aligned} E_N(N^{-2}\Sigma_{i^*=1}^N\hat{\delta}_{i^*}^2) &= \left[\frac{1}{N^2}\Sigma_{i^*=1}^N E_N(h^{-2k}a_{i^*}^2K_{i^*}^2)\right] \\ &= \frac{1}{N^2}\Sigma_{i=1}^N h^{-2k}a_i^2K_i^2. \end{aligned} \tag{A.4}$$

Similarly the second term yields

$$\begin{aligned} E_N\left[N^{-2}\Sigma_{i_1^*=1}^N\Sigma_{i_2^*\neq i_1^*}^N\hat{\delta}_{i_1^*}\hat{\delta}_{i_2^*}\right] &= N^{-2}\Sigma_{i_1^*=1}^N\Sigma_{i_2^*\neq i_1^*}^N E_N(\hat{\delta}_{i_1^*}\hat{\delta}_{i_2^*}) \\ &= N^{-2}\Sigma_{i_1^*=1}^N\Sigma_{i_2^*\neq i_1^*}^N\frac{1}{N^2}\Sigma_{i_1=1}^N\Sigma_{i_2=1}^N(\hat{\delta}_{i_1}\hat{\delta}_{i_2}) \\ &= \frac{N(N-1)}{N^2}\frac{1}{N^2}\Sigma_{i_1=1}^N\Sigma_{i_2=1}^N\hat{\delta}_{i_1}\hat{\delta}_{i_2} \\ &= \frac{N(N-1)}{N^2}\hat{\delta}_I^2 = \left(1-\frac{1}{N}\right)\hat{\delta}_I^2. \end{aligned} \tag{A.5}$$

So combining

$$var_N\hat{\delta}_I^* = \frac{1}{N^2}\Sigma_{i=1}^N h^{-2k}a_i^2K_i^2 - \frac{1}{N}\hat{\delta}_I^2. \tag{A.6}$$

Next, compute expectation $Evar_N\hat{\delta}_I^*$:

$$Evar_N\hat{\delta}_I^* = N^{-1}h^{-k}E(h^{-k}a_i^2K_i^2) - \frac{1}{N}E(\hat{\delta}_I^2), \tag{A.7}$$

where

$$\begin{aligned} E(\hat{\delta}_I^2) &= E(N^{-2}\Sigma_{i=1}^N\hat{\delta}_i^2 + N^{-2}\Sigma_{i_1=1}^N\Sigma_{i_2\neq i_1}^N\hat{\delta}_{i_1}\hat{\delta}_{i_2}) \\ &= N^{-1}E(\hat{\delta}_i^2) + \frac{N(N-1)}{N^2}E(\hat{\delta}_i)^2 \\ &= N^{-1}E(h^{-2k}a_i^2K_i^2) + \frac{N(N-1)}{N^2}\left[E(h^{-k}a_iK_i)\right]^2 \\ &= N^{-1}h^{-k}E(h^{-k}a_i^2K_i^2) + \frac{N(N-1)}{N^2}\left[E(h^{-k}a_iK_i)\right]^2 = O(1). \end{aligned}$$

Thus substituting into (A.7) we get

$$\begin{aligned} Evar_N\hat{\delta}_I^* &= N^{-1}h^{-k}E(h^{-k}a_i^2K_i^2) + O(N^{-1}) \\ &= N^{-1}h^{-k}[E(h^{-k}a_i^2K_i^2) + o(1)]. \end{aligned}$$

The expression for the variance $var\hat{\delta}_I$ is

$$\begin{aligned} var(\hat{\delta}_I) &= \left(E(\hat{\delta}_i^2) - E(\hat{\delta}_i)^2\right) N^{-1} \\ &= N^{-1}h^{-k}E(h^{-k}a_i^2K_i^2) - N^{-1}\left[E(h^{-k}a_iK_i)\right]^2 \\ &= N^{-1}h^{-k}[E(h^{-k}a_i^2K_i^2) + o(1)]. \end{aligned}$$

Thus

$$Nh^k\left|var\hat{\delta}_I - Evar_N\hat{\delta}_I^*\right| = o(1).$$

Next, we show that the empirical (and bootstrap) variance estimator is consistent for the variance, in other words, we show that $Nh^k\left|var_N\hat{\delta}_I^* - var\hat{\delta}_I\right|$ converges to zero in probability. Indeed $Nh^k\left|var_N\hat{\delta}_I^* - var\hat{\delta}_I\right| \leq Nh^k|var_N\hat{\delta}_I^* - Evar_N\hat{\delta}_I^*| + Nh^k\left|var\hat{\delta}_I - Evar_N\hat{\delta}_I^*\right|$. By Chebyshev's inequality for any $\varepsilon > 0$

$$\begin{aligned} & \Pr(Nh^k\left|var_N\hat{\delta}_I^* - var\hat{\delta}_I\right| > \varepsilon) \\ \leq\ & \Pr(Nh^k\left|var_N\hat{\delta}_I^* - Evar_N\hat{\delta}_I^*\right| > \varepsilon - Nh^k\left|Evar_N\hat{\delta}_I^* - var\hat{\delta}_I\right|) \\ \leq\ & \frac{\left(Nh^k\right)^2}{\left(\varepsilon - Nh^k\left|Evar_N\hat{\delta}_I^* - var\hat{\delta}_I\right|\right)^2}\left[var\left(var_N\hat{\delta}_I^*\right)\right] \\ \leq\ & \frac{4\left(Nh^k\right)^2}{\varepsilon^2}\left[var\left(var_N\hat{\delta}_I^*\right)\right], \end{aligned}$$

where we consider N large enough that $Nh^k\left|Evar_N\hat{\delta}_I^* - var\hat{\delta}_I\right| < \frac{\varepsilon}{2}$ for the last inequality.

Now all that is needed is to evaluate the order of the terms in $\left(Nh^k\right)^2 var\left(var_N\hat{\delta}_I^*\right)$ and show that they go to zero.

Using the expression in (A.6) with $\hat{\delta}_i$ for brevity we can derive $var\left(var_N\hat{\delta}_I^*\right)$

$$\begin{aligned} var\left(var_N\hat{\delta}_I^*\right) &= N^{-5}(N-1)^2var(\widehat{\delta_i^2}) + 2N^{-5}(N-1)var(\hat{\delta}_i\hat{\delta}_{i'}) \\ &\quad + 4N^{-5}(N-1)(N-2)cov(\hat{\delta}_i\hat{\delta}_{i'}, \hat{\delta}_i\hat{\delta}_{i''}) - 4N^{-5}(N-1)^2cov\left(\hat{\delta}_i^2, \hat{\delta}_i\hat{\delta}_{i'}\right). \end{aligned} \tag{A.8}$$

which yield

$$\begin{aligned} & N^{-5}(N-1)^2 var(h^{-2k}K_i^2 a_i^2) + 2N^{-5}(N-1) var_{i_1 \neq i_2}(h^{-2k}K_{i_1}K_{i_2}a_{i_1}a_{i_2}) \\ & -4N^{-5}(N-1)^2 cov_{i_1 \neq i_2}\left(h^{-2k}K_{i_1}^2 a_{i_1}^2, h^{-2k}K_{i_1}K_{i_2}a_{i_1}a_{i_2}\right) \\ & +4N^{-5}(N-1)(N-2) cov_{i_1 \neq i_2 \neq i_3}(h^{-2k}K_{i_1}K_{i_2}a_{i_1}a_{i_2}, h^{-2k}K_{i_1}K_{i_3}a_{i_1}a_{i_3}) \\ = \; & O\left(N^{-3}h^{-3k} + N^{-4}h^{-2k} + N^{-3}h^{-2k} + N^{-3}h^{-k}\right) = O(N^{-3}h^{-3k}). \end{aligned}$$

The orders follow after noting, e.g., that $var(h^{-2k}K_i^2 a_i^2) = O(h^{-3k})$.

Substituting now into the Chebyshev inequality we obtain that the empirical variance converges in probability to the leading term in the variance.

Case II. The only difficulty comes from extra weights h^{-m} that will enter the appropriate rate; the rest of the derivation is similar to I.

Case III. Consider the average estimator that is based on I: $\hat{\delta}_{III} = \frac{1}{N(N-1)}\Sigma_{i=1}^N \Sigma_{j \neq i}\hat{\delta}_{i,j}$, where $\hat{\delta}_{i,j} = a(x_j, y_i)h^{-k}K\left(\frac{x_i - x_j}{h}\right)$.

Recall that the second moments for the estimator are as follows:

$$\begin{aligned} E\hat{\delta}_{III}^2 &= \frac{1}{N^2(N-1)^2} E\Sigma_{i=1}^N \Sigma_{j \neq i}\Sigma_{i'=1}^N \Sigma_{j' \neq i'}\hat{\delta}_{i,j}\hat{\delta}_{i',j'} \\ &= \frac{1}{N^2(N-1)^2}\left(\begin{array}{l} \Sigma_{i=1}^N \Sigma_{j \neq i} E\left(\hat{\delta}_{i,j}^2 + \hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\ +\Sigma_{i=1}^N \Sigma_{j \neq i}\Sigma_{j' \neq j \neq i} E\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + \hat{\delta}_{i,j}\hat{\delta}_{j',i} + \hat{\delta}_{j,i}\hat{\delta}_{i,j'} + \hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ +\Sigma_{i=1}^N \Sigma_{j \neq i}\Sigma_{i' \neq j \neq i}\Sigma_{j' \neq i' \neq j \neq i} E\hat{\delta}_{i,j}\hat{\delta}_{i',j'} \end{array}\right) \\ &= \frac{1}{\substack{N^2(N-1)^2 \\ i \neq j \neq j'}}\left(\begin{array}{l} N(N-1)E\hat{\delta}_{i,j}^2 + N(N-1)E\hat{\delta}_{i,j}\hat{\delta}_{j,i} \\ +N(N-1)(N-2)\left(E\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + E\hat{\delta}_{i,j}\hat{\delta}_{j',i} + E\hat{\delta}_{j,i}\hat{\delta}_{i,j'} + E\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ +N(N-1)(N-2)(N-3)\left(E\hat{\delta}_{i,j}\right)^2 \end{array}\right) \end{aligned}$$

and

$$var\hat{\delta}_{III} = \frac{1}{\substack{N(N-1) \\ i \neq j \neq j'}}\left(\begin{array}{l} E\hat{\delta}_{i,j}^2 + E\hat{\delta}_{i,j}\hat{\delta}_{j,i} - 2(2N-3)\left(E\hat{\delta}_{i,j}\right)^2 \\ +(N-2)\left(E\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + E\hat{\delta}_{i,j}\hat{\delta}_{j',i} + E\hat{\delta}_{j,i}\hat{\delta}_{i,j'} + E\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \end{array}\right), \quad \text{(A.9)}$$

using $var\hat{\delta}_{III} = E\hat{\delta}_{III}^2 - \left(E\hat{\delta}_{III}\right)^2 = E\hat{\delta}_{III}^2 - \left(\underset{i \neq j}{E\hat{\delta}_{i,j}}\right)^2$. Since $\underset{i \neq j}{E\hat{\delta}_{i,j}} = O(1), \underset{i \neq j}{E\hat{\delta}_{i,j}^2} = O(h^{-k}), \underset{i \neq j}{E\hat{\delta}_{i,j}\hat{\delta}_{j,i}} = O(h^{-k}), \underset{i \neq j \neq j'}{E\hat{\delta}_{i,j}\hat{\delta}_{i,j'}} = O(1)$, etc., the leading non-parametric term of the

variance is $N^{-2}\left(E\hat{\delta}_{i,j}^2 + E\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) = O\left(N^{-2}h^{-k}\right)$,whereas the leading parametric term is $N^{-1}\left(E\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + E\hat{\delta}_{i,j}\hat{\delta}_{j',i} + E\hat{\delta}_{j,i}\hat{\delta}_{i,j'} + E\hat{\delta}_{j,i}\hat{\delta}_{j',i} - 4\left(E\hat{\delta}_{i,j}\right)^2\right) = O(N^{-1})$.

Consider now the bootstrapped estimator for Case III

$$\hat{\delta}_{III}^* = \frac{1}{N(N-1)}\Sigma_{i^*=1}^N\Sigma_{j^*\neq i^*}\hat{\delta}_{i^*,j^*}I_{i^*j^*},$$

where $I_{i^*j^*} = I(x_{i^*} \neq x_{j^*})$. Note that it excludes combinations of observations for which $x_{i^*} = x_{j^*}$.[6]

$E_N\hat{\delta}_{III}^* = E_N\left(\underset{i^*\neq j^*}{\hat{\delta}_{i^*,j^*}I_{i^*j^*}}\right) = N^{-2}\Sigma_{i=1}^N\Sigma_{j=1}^N\hat{\delta}_{i,j}I_{ij} = N^{-2}\Sigma_{i=1}^N\Sigma_{j\neq i}\hat{\delta}_{i,j} = \frac{N-1}{N}\hat{\delta}_{III}$; for the original sample, $I_{ij} = 0$ iff $i = j$.

Thus, $E_N\hat{\delta}_{III}^*$ is the original full-sample estimator, $\hat{\delta}_{III}$, up to the multiple $(1 - \frac{1}{N})$.

Next consider the empirical variance of the bootstrap estimator: $var_N\hat{\delta}_{III}^* = E_N\hat{\delta}_{III}^{*2} - \left(E_N\hat{\delta}_{III}^*\right)^2$.

We have $E_N\hat{\delta}_{III}^{*2} = \frac{1}{N^2(N-1)^2}E_N\Sigma_{i^*=1}^N\Sigma_{j^*\neq i^*}\Sigma_{i'^*=1}^N\Sigma_{j'^*\neq i'^*}\hat{\delta}_{i^*,j^*}\hat{\delta}_{i'^*,j'^*}I_{i^*j^*}I_{i'^*j'^*}$

$$\begin{aligned}
&= \frac{1}{N^2(N-1)^2}\left(\begin{array}{l}
\Sigma_{i^*=1}^N\Sigma_{j^*\neq i^*}E_N\left(\hat{\delta}_{i^*,j^*}^2 + \hat{\delta}_{i^*,j^*}\hat{\delta}_{j^*,i^*}\right)I_{i^*j^*} \\
+\Sigma_{i^*=1}^N\Sigma_{j^*\neq i^*}\Sigma_{j'^*\neq j^*\neq i^*}E_N\left(\hat{\delta}_{i^*,j^*}\hat{\delta}_{i^*,j'^*} + \hat{\delta}_{i^*,j^*}\hat{\delta}_{j'^*,i^*} + \hat{\delta}_{j^*,i^*}\hat{\delta}_{i^*,j'^*} + \hat{\delta}_{j^*,i^*}\hat{\delta}_{j'^*,i^*}\right)I_{i^*j^*}I_{i^*j'^*} \\
+\Sigma_{i^*=1}^N\Sigma_{j^*\neq i^*}\Sigma_{i'^*\neq j^*\neq i^*}\Sigma_{j'^*\neq i'^*\neq j^*\neq i}E_N\hat{\delta}_{i^*,j^*}\hat{\delta}_{i'^*,j'^*}I_{i^*j^*}I_{i'^*j'^*}
\end{array}\right) \\
&= \frac{1}{N^2(N-1)^2}\left(\begin{array}{l}
N(N-1)N^{-2}\Sigma_{i=1}^N\Sigma_{j=1}^N\left(\hat{\delta}_{i,j}^2 + \hat{\delta}_{i,j}\hat{\delta}_{j,i}\right)I_{ij} \\
+N(N-1)(N-2)N^{-3}\Sigma_{i=1}^N\Sigma_{j=1}^N\Sigma_{j'=1}^N\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + \hat{\delta}_{i,j}\hat{\delta}_{j',i} + \hat{\delta}_{j,i}\hat{\delta}_{i,j'} + \hat{\delta}_{j,i}\hat{\delta}_{j',i}\right)I_{ij}I_{ij'} \\
+N(N-1)(N-2)(N-3)N^{-4}\Sigma_{i=1}^N\Sigma_{j=1}^N\Sigma_{i'=1}^N\Sigma_{j'=1}^N\hat{\delta}_{i,j}\hat{\delta}_{i',j'}I_{ij}I_{i'j'}
\end{array}\right) \\
&= \frac{1}{N^3(N-1)}\left(\begin{array}{l}
\Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}_{i,j}^2 + \hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\
+N^{-1}(N-2)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{j'\neq i}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'} + \hat{\delta}_{i,j}\hat{\delta}_{j',i} + \hat{\delta}_{j,i}\hat{\delta}_{i,j'} + \hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\
+N^{-2}(N-2)(N-3)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{i'=1}^N\Sigma_{j'\neq i'}\hat{\delta}_{i,j}\hat{\delta}_{i',j'}
\end{array}\right)
\end{aligned}$$

[6] For ADE this happens automatically for a symmetric kernel since then $K'(0) = 0$, but for average density this is not the case and $I_{i^*j^*}$ is needed.

and

$$\begin{aligned} var_N\hat{\delta}^*_{III} &= E_N\hat{\delta}^{*2}_{III} - \left(N^{-2}\Sigma_{i=1}^N\Sigma_{j\neq i}\hat{\delta}_{i,j}\right)^2 \\ &= \frac{1}{N^3(N-1)}\begin{pmatrix} \Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\ +N^{-1}(N-2)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{j'\neq i}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'}+\hat{\delta}_{i,j}\hat{\delta}_{j',i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j'}+\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ -2N^{-2}\left(2N-3\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{i'=1}^N\Sigma_{j'\neq i'}\hat{\delta}_{i,j}\hat{\delta}_{i',j'} \end{pmatrix}. \end{aligned}$$

Similarly to $var\hat{\delta}_{III}$ in (A.9) we express $var_N\hat{\delta}^*_{III}$ using non-overlapping indices in the multiple sums:

$$\begin{aligned} var_N\hat{\delta}^*_{III} &= \frac{1}{N^3(N-1)}\begin{pmatrix} \Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\ +N^{-1}(N-2)\Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j}+\hat{\delta}_{j,i}\hat{\delta}_{j,i}\right) \\ +N^{-1}(N-2)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{j'\neq i\neq j}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'}+\hat{\delta}_{i,j}\hat{\delta}_{j',i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j'}+\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ -2N^{-2}\left(2N-3\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\ -2N^{-2}\left(2N-3\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{j'\neq j\neq i}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'}+\hat{\delta}_{i,j}\hat{\delta}_{j',i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j'}+\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ -2N^{-2}\left(2N-3\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{i'\neq j\neq i}\Sigma_{j'\neq i'\neq j\neq i}\hat{\delta}_{i,j}\hat{\delta}_{i',j'} \end{pmatrix} \\ &= \frac{1}{N^5(N-1)}\begin{pmatrix} \left(3N^2-8N+6\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) \\ +\left(N^2-6N+6\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{j'\neq i\neq j}\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'}+\hat{\delta}_{i,j}\hat{\delta}_{j',i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j'}+\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \\ -2\left(2N-3\right)\Sigma_{i=1}^N\Sigma_{j\neq i}\Sigma_{i'\neq j\neq i}\Sigma_{j'\neq i'\neq j\neq i}\hat{\delta}_{i,j}\hat{\delta}_{i',j'} \end{pmatrix}. \end{aligned}$$

Consider next the convergence of each term to the corresponding one in $var\hat{\delta}_{III}$.

For example,

$$\frac{1}{N^3(N-1)}\Sigma_{i=1}^N\Sigma_{j\neq i}\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right)$$

has expectation

$$N^{-2}[E\hat{\delta}^2_{i,j}+E\hat{\delta}_{i,j}\hat{\delta}_{j,i}],$$

where $[E\hat{\delta}^2_{i,j}+E\hat{\delta}_{i,j}\hat{\delta}_{j,i}]=O(h^{-k})$ while the variance of this term is $O(N^{-6}h^{-3k})+O(N^{-5}h^{-2k})$; by Chebyshev's inequality this term converges to the corresponding term in the variance. Similarly, convergence can be shown for other terms.

Note that the expected value of $var_N\hat{\delta}^*_{III}$ is

$$E\left(var_N\hat{\delta}^*_{III}\right) = \underset{i\neq j\neq j'}{\frac{1}{N^4}} \left(\begin{array}{l} (3N^2-8N+6)\,E\left(\hat{\delta}^2_{i,j}+\hat{\delta}_{i,j}\hat{\delta}_{j,i}\right) - 2\,(2N-3)\,(N-2)(N-3)\left(E\hat{\delta}_{i,j}\right)^2 \\ +\,(N^2-6N+6)\,(N-2)E\left(\hat{\delta}_{i,j}\hat{\delta}_{i,j'}+\hat{\delta}_{i,j}\hat{\delta}_{j',i}+\hat{\delta}_{j,i}\hat{\delta}_{i,j'}+\hat{\delta}_{j,i}\hat{\delta}_{j',i}\right) \end{array} \right).$$

Although the leading parametric term here is the same as in $var\hat{\delta}_{III}$, the leading non-parametric term is three times the leading non-parametric term of $var\hat{\delta}_{III}$. Thus the bootstrap estimator is biased for the bandwidth dependent term. ■

REFERENCES

[1] Abadir, K.M. and S. Lawford (2004): "Optimal asymmetric kernels," *Economics Letters*, **83**, 61–68.

[2] Bertail, P., D.N. Politis, and J.P. Romano (1999): "On subsampling estimators with unknown rate of convergence," *Journal of the American Statistical Association* **94**, 569-579.

[3] Bickel, P. J. and D.A. Freedman (1981): "Some asymptotic theory for the bootstrap," *Annals of Statistics*, **9**, 1196-1217.

[4] Blundell, R. and A. Duncan (1998): "Kernel regression in empirical microeconomics," *Journal of Human Resources*, **33**, 62–87.

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

[6] Calabrese R. and M. Zenga (2010): "Bank loan recovery rates: Measuring and nonparametric density estimation," *Journal of Banking and Finance*, **34**, 903-911.

[7] Cattaneo, M.D, R.K. Crump and M. Jansson (2010): "Bootstrapping Density-Weighted Average Derivatives," manuscript.

[8] Coppejans, M. and H. Sieg (2005): "Kernels estimation and average derivatives and differences," *Journal of Business and Economic Statistics*, **23**, 211-225.

[9] 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.

[10] DiNardo, J. and J.L. Tobias (2001): "Nonparametric density and regression estimation," *The Journal of Economic Perspectives*, **15**, 11–28.

[11] Donkers, B. and M.M.A. Schafgans (2008): "Specification and estimation of semuparametric multiple-index models," *Econometrics Theory* **24**, 1584-1606.

[12] Hafner, C.M. and O. Linton (2010): "Efficient estimation of a multivariate multiplicative volatility model," *Journal of Econometrics*, **159**, 55–73.

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

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

[15] Härdle, W. & T.M. Stoker (1989): "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association,* **84**, 986–995.

[16] Henderson, D.J. and D.L. Millimet (2008): "Is gravity linear?" *Journal of Applied Econometrics*, **23**, 137–172.

[17] Horowitz, J. L. (1992): "A smoothed maximum score estimator for the binary response model," *Econometrica*, **60**, 505-531.

[18] 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.

[19] Huynh, K.P. and D.T. Jacho-Chavez (2009): "Growth and governance: A nonparametric analysis," *Journal of Comparative Economics*, **37**, 121-143.

[20] Jun, S.J., J. Pinkse and Y. Wan (2010), "A consistent nonparametric test of affiliation in auction models," *Journal of Econometrics*, **159**, 46–54.

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

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

[23] Kotlyarova, Y. and V. Zinde-Walsh (2010):"Robust estimation in binary choice models," *Communications in Statistics – Theory and Methods* **39**, 266-279.

[24] Li, Q. and J.Racine (2007): *Nonparametric Econometrics: Theory and Practice*, Princeton: Princeton University Press

[25] Long, X., L. Su and A. Ullah (2011): "Estimation and forecasting of dynamic conditional covariance: A semiparametric multivariate model," *Journal of Business and Economic Statistics* **29**, 109-125.

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

[27] Pagan, A. and A. Ullah (1999): *Nonparametric Econometrics*, Cambridge: Cambridge University Press

[28] Powell, J.L., J.H. Stock and T.M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica,* **57**, 1403–1430.

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

[30] Rothe, C. (2010): "Nonparametric estimation of distributional policy effects," *Journal of Econometrics*, **155**, 56–70.

[31] Samarov, A.M. (1993): "Exploring Regression Structure Using Nonparametric Functional Estimation," *Journal of the American Statistical Association* **88**, 836-847.

[32] Schafgans, M.M.A. and V. Zinde-Walsh (2010): "Smoothness Adaptive Average Derivative Estimation," *Econometrics Journal*, **13**, 40–62.

[33] Silverman, B.W. (1986): *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

[34] 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: Cambridge University Press.

[35] Stone, C.J. (1980): "Optimal rates of convergence for nonparametric estimators," *Annals of Statistics*, **8**, 1348–1360.

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

[37] Woodroofe, M. (1970): "On choosing a delta sequence," *The Annals of Mathematical Statistics,* **41**, 1665–1671.