Rates of expansions for functional estimators

Yulia Kotlyarova* Marcia M.A. Schafgans† Victoria Zinde-Walsh‡

Abstract. In this paper, we summarize results on convergence rates of various non- and semiparametric estimators focusing on the impact of insufficient smoothness, possibly unknown smoothness and even non-existence of density. In the presence of a possible lack of smoothness and the uncertainty about smoothness, methods of safeguarding against this uncertainty are surveyed with emphasis on nonconvex model averaging. This approach can be implemented via a combined estimator that selects weights based on minimizing the asymptotic mean squared error. We provide evidence about the importance of accounting for possible lack of smoothness when evaluating the finite sample performance of the estimators.

Keywords: Nonparametric estimation, kernel based estimation, model averaging, combined estimator, convergence rates, degree of smoothness.

JEL Classification: C14.

*Department of Economics, Dalhousie University.

†Department of Economics, London School of Economics.

‡Department of Economics, McGill University and CIREQ. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

1. INTRODUCTION

This paper focuses on the convergence rate of non- and semiparametric estimators, a rate which is instrumental for stochastic expansions, limit processes and limit moments. The original Nagar expansions (Nagar, 1959, Nagar and Ullah, 1970) were specific to parametric econometric estimators, such as the k-class and 3SLS estimator, that can be written as $\sqrt{N}\left(\hat{\delta}_n - \delta_0\right) = A^{-1}a$ with $A = \bar{A} + \Delta A$, where $\bar{A}$ is a finite matrix and $\Delta A = O(n^{-1/2})$. A general form of Nagar's expansion was provided in Sargan (1974). To further generalize stochastic expansions to non- and semiparametric estimators we can define an expansion for the estimator $\hat{\delta}_N$ of δ_0 in the following way

$$N^{\tau_o}\left(\hat{\delta}_n - \delta_0\right) = e_1 + \Sigma_{s=2}^{m} e_s/N^{\tau_s} + r_m/N^{\tau_{m+1}}, \tag{1}$$

where $\tau_0 > 0$ (originally considered to be 1/2), $\{\tau_s\}_{s=2}^{m+1}$ a strictly increasing positive deterministic sequence, with e_s, $s = 1, ..., m$, and r_m being $O_p(1)$. Nagar's approach was to approximate the moments of the estimator $\hat{\delta}_n$ by moments of terms in the expansion. In particular, assuming that the expectations of e_s and r_m are $O(1)$, the suitable bias expansion would be given by $N^{\tau_o}E\left(\hat{\delta}_n - \delta_0\right) = E(e_1) + \Sigma_{s=2}^{m} Ee_s/N^{\tau_s} + O\left(N^{-\tau_{m+1}}\right)$. The variance of the distribution can similarly be approximated; ignoring r_m requires furthermore $E(r_m e_s) = O(1)$ for all $s \leq m$ and $E(r_m r_m') = O(1)$, with Srinivasan (1970) providing cautionary arguments about the conditions.

A crucial aspect of Nagar's and similar expansions is that the rate of convergence, N^{τ_o}, is known. In fact, to our knowledge, there are no contributions where it is not parametric. Nagar-type expansions for semiparametric kernel estimators in Linton (1995) (partial linear regression model) and Ichimura and Linton (2005) (semiparametric programme evaluation estimator) enjoyed the parametric rate under additional regularity conditions such as smoothness and behavior of tails of r_m. The parametric rate, though, could only be obtained with appropriate smoothness, and for kernel estimators requires the bandwidth to be chosen over a very restrictive range. Insufficient smoothness or incorrect choice of the bandwidth would jeopardize τ_0 equalling 1/2 and could result in lower than parametric rates of convergence even when a stochastic expansion of the form (1) may exist. A discussion of potentially nonparametric rates for semiparametric kernel estimators, which depend on smoothness and bandwidth choice, is provided e.g. in Schafgans and Zinde-Walsh (2010) (henceforth SZW, 2010) for the density weighted average derivative estimator and in Kotlyarova et al. (2016) (henceforth KSZW, 2016) for the general class of semiparametric estimators. The fragility of the leading rate for semiparametric estimators has implications not only for the existence of Nagar-type stochastic expansions, but also for the limiting moments and asymptotic distribution. In this paper, our focus is on the **convergence rates** of kernel based estimators and the impact associated with insufficient smoothness, possibly unknown smoothness, and possibly even non-existence of density. We provide an overview of available theoretical results for kernel-based estimators and discuss the importance of suitably evaluating their finite sample validity accounting for possibly deficient smoothness and uncertainty. While the focus of this paper is on kernel estimators, limit

properties for many functional estimators, such as series estimators are affected by the smoothness (Holder) class of functions (see, e.g, Ichimura and Newey, 2017). Similarly, rates for adaptive estimators (which may rely on kernels) such as Lepski and Spokoiny (1997) and Mukherjee et al., (2016) reflect the order of the smoothness class. Some of these estimators may improve on the properties of kernel estimators, e.g., sieve estimators (Ai and Chen, 2003) can achieve root-n consistency for the parameter estimators in a semiparametric context without requiring higher differentiability for the unknown function; they assume that the unknown function is in a Holder class and require a uniform approximation rate over the sequence of sieve spaces.

In general, some of the terms in the expansion of $\hat{\delta}_N - \delta_0$ and its moments for non- and semiparametric estimators can only be obtained by relying on smoothness of the density (and other functions). Such smoothness assumptions are nonparametrically non-testable (see e.g. Lalley and Nobel, 2003), they are arbitrary and often not easy to justify. Indeed, there is a lot of uncertainty about the true degree of smoothness of density of such variables as income or labor supply, where "bunching" is well documented; see Kleven (2016) for an interesting review on bunching.

We introduce a **degree of smoothness** parameter, $\bar{v}$, that captures insufficient smoothness settings, where, say, the parametric rate of convergence for semiparametric kernel estimators is not possible. The main issue in relation to the expansion (1) is that only when the degree of smoothness is known, is it possible to specify the rate N^{τ_o} with the value of τ_0 reflecting the degree of smoothness. As it may not be realistic to assume the level of smoothness is known, we analyze the **uncertainty** associated with the true degree of smoothness. Following Woodroofe (1970), SZW (2010) and KSZW (2016) explored methods of **estimating the degree of smoothness**. To obtain the limit process for the kernel estimator of the density in the absence of smoothness of the distribution function, Zinde-Walsh (2008, 2017) (henceforth ZW, 2008 and ZW, 2017) employ the space of generalized functions, where the differentiation operator can be defined. Tuvaandorj and Zinde-Walsh (2014) (henceforth TZW, 2014), extend these results to the kernel estimator of conditional distribution and conditional density function.

Acknowledging lack of smoothness and uncertainty about smoothness requires developing methods of safeguarding against this uncertainty. One approach to deal with lack of smoothness is bias reduction via various methods that has been favored in the literature. The other is a **model-averaging procedure** that could be better suited for dealing with uncertainty about smoothness.

Similar to parametric regression analysis, model averaging approaches for nonparametric regression analysis have been developed under uncertain covariates to deal with the bias and inconsistency in estimation and testing associated with model selection. Typically convex weights are applied, which can be selected based on minimizing a least squares cross-validation criterion (Breiman, 1996), least squares weights (Li, Linton and Lu, 2015), information criteria or jackknife model averaging approaches (see, e.g., Hansen and Racine, 2012, and Ullah and Wang, 2013, for a review). Here we consider the model-averaging procedure as a means to deal with the uncertainty about how to select user-defined choices like bandwidth and kernel. Henderson and Parmeter (2016), also, consider such model-

averaging procedure for nonparametric regression analysis. They reveal the potential benefits of such an approach, where they consider weights based on minimizing a least squares cross-validation criterion (as in Liu, 2018) and using least squares weights (as in Li et al., 2015).

While it is common to use convex weights when averaging estimators, non-convex combinations may offer an advantage when bias plays a prominent role and using negative weights could permit trade-offs between large biases to reduce the overall impact of bias. This was accomplished by the use of jackknife-type estimator in Schucany and Sommers (1977), and similar trade-offs are considered in Bierens (1987), Powell et al. (1989) (henceforth PSS, 1989), and Cattaneo et al. (2013). Crucially, the degree of smoothness and thus the rates for different bandwidths would have to be known to implement each of these approaches. Kotlyarova and Zinde-Walsh (2006) (henceforth KZW, 2006) examined the problem of uncertainty about the smoothness and proposed model averaging by demonstrating the existence of (possibly non-convex) linear combinations of several estimators corresponding to different bandwidth-kernel combinations that would give the best available, but unknown a priori, convergence rate. This theoretical result shows the possibility of improved performance of averaged estimators, also referred to as combined estimators. Kotlyarova and Zinde-Walsh, 2007 (henceforth KZW, 2007) and SZW (2010) construct robust estimators of densities and density weighted average derivatives using estimated weights that minimize the asymptotic MSE (AMSE) and KSZW (2016) generalize the results to local and general averaged kernel based estimators.

The possibility of a lack of smoothness needs to be reflected when evaluating the finite sample performance of the estimators. In the econometrics literature, typically, distributions used in simulations are very smooth: the use of uniform and normal distributions are commonplace, occasionally chi-squared, in Cattaneo et al. (2014a, 2014b) or a mixture of two normals in Newey, Hsieh & Robins (2004, 1998) and Henderson and Parmeter (2016), and a mixture of three normals in Stoker (1993) are considered. None of these distributions is capable of capturing the effect of non-smoothness, or even of smooth functions in the presence of very large derivatives. Marron and Wand (1992) and Härdle et al. (1998) in the statistics literature investigated the estimation of density for cases with large derivatives, but such distributions are rarely considered elsewhere. This has prompted KZW (2007), SZW (2010), KSZW (2016) as well as Kankanala and Zinde-Walsh (2020) to explore performance of kernel estimators and kernel-based statistics in cases of normal **mixture with peaked normals** where derivatives could vastly exceed those for the standard cases. Simulations with such mixtures produce dramatically different results from those in the literature.

Section 2 gives an overview of non- and semiparametric estimators with known (but possibly insufficient) degree of smoothness. Section 3 examines the more realistic setting where the degree of smoothness is unknown and where the density may not even exist. In 3.1 some smoothness is satisfied but the degree itself is not known; we provide convergence results about estimated degree of smoothness and estimated optimal bandwidth rates. In 3.2, the possible absence of smoothness is addressed, e.g. density may not exist, and we provide theoretical results where a parametric rate in the space of stochastic generalized

functions is achieved. Results for several widely used non- and semiparametric estimators under different smoothness conditions are provided in Table 1. Section 4 reviews approaches to improved estimation: bias reduction and model averaging. Section 5 demonstrates the impact of distributions used in simulations on claims regarding finite sample performance.

2. RATES WITH KNOWN SMOOTHNESS

We consider non- and semiparametric kernel based estimators, generically denoted as $\hat{\delta}_N(K,h)$, with K specifying the kernel function and h the bandwidth parameter. Depending on the type of functional of interest, this may represent a local kernel based estimator, where the interest is the value of a function, e.g., density, at a particular point, or an average kernel based estimator, where the interest is an expectation. We use standard assumptions on the kernel function $K(x)$ for $x \in R^k$ (see, e.g. Li and Racine, 2007, Pagan and Ullah, 1999) and denote by $v(K)$ the order of the kernel. 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 bandwidth is assumed to be dependent on the sample size, N, such that $h \to 0$, and $h^k N \to \infty$ as $N \to \infty$. With $\mathbf{h} = (h_1, ..., h_k)$ denoting a vector of bandwidths, here, for simplicity, we assume that $h = \max\{h_i\}$ and $h_i/h_j = O(1)$ for all $i, j = 1, ..., k$.

Smoothness requirements for kernel-based estimation typically refer to smoothness of the density function in nonparametric estimation, but smoothness assumptions relate to other functions, such as conditional means, as well. To streamline the exposition we focus here on smoothness (and existence) of density, assuming that the other functions satisfy sufficient smoothness requirements. We use v as a measure that describes the smoothness of the density $f(x)$. Suppose $f(x)$ is not differentiable over some domain $\Omega \subseteq R^k$ but for some $0 < \alpha \leq 1$ satisfies a Holder condition at every $x, x + \Delta x \in \Omega$:

$$|f(x + \Delta x) - f(x)| \leq \omega_f(x) \|\Delta x\|^{\alpha}.$$

We then say that f has the smoothness order α (fractional) and set $v = \alpha$. If for some integer m it can be assumed that all partial derivatives of order m given by $\partial^m f(x)/(\partial^{m_1} x_1 ... \partial^{m_k} x_k)$, with $m_1 + ... + m_k = m$, are (Holder) continuous, then f has smoothness order no less than m and we set $v = m$.

The **degree of smoothness parameter**, $\bar{v}$, which determines the convergence rate for the estimator $\hat{\delta}_N(K,h)$, can be described by a function of density smoothness and kernel order:

$$\bar{v} = \bar{v}(v, v(K)) \tag{2}$$

and is often defined as the minimum of the density smoothness and the kernel order, $\bar{v} = \min(v, v(K))$. Indeed, in general, assume that v is the minimal degree of smoothness of the functions that are relevant for bias expansion of the estimator (such as the density $f(x)$ or its derivative $f'(x)$ and the conditional moment $g(x) = E(y|x)$). For some widely used estimators $\hat{\delta}(K,h)$ (listed in Table 1) $|B(K,h)| \leq \omega h^{\min(v, v(K))}$, where $B(K,h)$ denotes the bias of the estimator $E(\hat{\delta}_N(K,h) - \delta_0)$. With insufficient smoothness: $v < v(K)$, 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. For example, for a second-order kernel density estimator it is known that if the second derivative of the density is continuous at x, the bias expansion provides the leading term $h^2\mathcal{B}_2(K)$, with $\mathcal{B}_2(K) = 0.5f''(x)\int w^2K(w)\,dw$. Similarly, for a fourth-order kernel density estimator if the fourth derivative of the density is continuous at x the leading term in the expansion is $h^4\mathcal{B}_4(K)$, with $\mathcal{B}_4(K) = \frac{1}{4!}f^{(4)}(x)\int w^4K(w)\,dw$. However, if the density function is only twice continuously differentiable, the leading term in the bias expansion of the fourth-order kernel density estimator could be of order no better than $O(h^{2+\delta})$ for some small $\delta > 0$.

We follow KSZW (2016) in assuming that the bias is stabilized at the rate $\bar{v}$; this is the assumption made by Woodroofe (1970) for density estimation and is also made in SZW (2010) for density weighted average derivative (henceforth PSS-ADE). This is given by Assumption 1, where the bound $\bar{\epsilon} > 0$ may be needed to ensure that the rest of the bias expansion converges to zero sufficiently fast.

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

$$h^{-\bar{v}}bias(\hat{\delta}_N(K,h)) \to \mathcal{B}(K), \tag{3}$$

for some $\bar{v} > 0$, where the vector $B(K) = (B_1(K), ...B_L(K))'$ is such that $0 < |\mathcal{B}_\ell(K)| < \infty$ for $\ell = 1, ..., L$ with $L = \dim(\delta_0)$.

The rate of convergence of the variance, $Var(\hat{\delta}_N(K,h))$, which differs for local and averaged kernel based estimators, is characterized in Assumption 2 from KSZW (2016):

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

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

(b) for average kernel based estimators: there exist finite positive definite matrices $\Sigma_1(K)$ and Σ_2 and $r > 0$ such that an expansion for the variance is

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

Specifics for different widely used estimators are provided in Table 1. The rates are usually associated with limiting computations for the moments. The results typically apply to a random sample setting, but some are valid in the presence of heteroskedasticity and weak dependence (e.g. mixing).

Assumptions 1 and 2 imply $\hat{\delta}_N - \delta_0 = O_p(h^{\bar{v}} + \left(N^{-\omega}h^{-d(k)}\right)^{1/2})$ for local kernel based estimators, while for averaged kernel based estimators $\hat{\delta}_N - \delta_0 = O_p(h^{\bar{v}} + \left(N^{-\omega}h^{-d(k)}\right)^{1/2} + N^{-1/2})$. For the estimators considered in Table 1, ω could be 1 or 2 and $d(k)$ could be either k or $k+2$. For kernel density and conditional mean estimators, $\omega = 1$ and $d(k) = k$ (for

Table 1: Kernel based estimators: Smooth and Nonsmooth results

Estimator $\hat{\delta}_N(K,h)$	Results
Density	
$\frac{1}{Nh^k}\sum_{i=1}^{N} K(\frac{X_i-x}{h}) := \hat{f}_{(K,h)}(x)$ Parzen (1962), Rosenblatt (1956) KZW (2007) (non-smooth) ZW (2008, 2017) (gen fun)	AMSE terms: $h^{2\bar{v}}\mathcal{B}(K)\mathcal{B}(K)^T + (Nh^k)^{-1}\Sigma(K)$ optimal bandwidth rate: $N^{-\frac{1}{2\bar{v}+k}}$ MSE rate: $N^{-\frac{2v(K)}{2v(K)+k}}$ if $\bar{v} = v(K)$ (smooth) $h^{2\bar{v}}$ if $\bar{v} < v(K)$ (non-smooth) as random generalized function: bias rate: $h^{\nu(K)}$ undersmoothed convergence rate: $N^{-1/2}$
Conditional mean	
$\frac{1}{Nh^k}\sum_{i=1}^{N} K(\frac{X_i-x}{h})Y_i \;/\; \hat{f}_{(K,h)}(x)$ Nadaraya (1964), Watson (1964) KSZW (2016) (non-smooth) ZW (2008, 2017) (gen fun)	AMSE terms: $h^{2\bar{v}}\mathcal{B}(K)\mathcal{B}(K)^T + (Nh^k)^{-1}\Sigma(K)$ optimal bandwidth rate: $N^{-\frac{1}{2\bar{v}+k}}$ MSE rate: $N^{-\frac{2v(K)}{2v(K)+k}}$ if $\bar{v} = v(K)$ (smooth) $h^{2\bar{v}}$ if $\bar{v} < v(K)$ (non-smooth) numerator as random generalized function: bias rate: $h^{\nu(K)}$ undersmoothed convergence rate: $N^{-1/2}$
Conditional density	
$\frac{1}{Nh^{k_x}h_0^{k_y}}\sum_{i=1}^{N} K(\frac{X_i-x}{h})K_0(\frac{Y_i-y}{h_0}) \;/\; \hat{f}_{(K,h)}(x)$ Chen, Linton, Robinson (2001)**, Racine, Li, Zhu (2004) (smooth) ZW (2013), TZW (2014)** (gen fun)	AMSE terms*: $h^{2\bar{v}}\mathcal{B}(K,K_0)\mathcal{B}(K,K_0)^T + (Nh^{k_x+k_y})^{-1}\Sigma(K,K_0)$ optimal bandwidth rate: $N^{-\frac{1}{k_x+k_y+2\bar{v}}}$ MSE rate: $N^{-\frac{v(K)}{k_x+k_y+v(K)}}$ (smooth) as random generalized function: undersmoothed convergence rate: $N^{-1/2}$

* with $\bar{v} = v(K) = 2$.

** under mixing conditions

Table 1 (Cont'd): Kernel based estimators: Smooth and Nonsmooth results

Estimator $\hat{\delta}_N(K,h)$	Results
Average density	
$\frac{1}{N}\sum_{i=1}^{N}\hat{f}_{(K,h)}(X_i)$ Newey et al. (1998) KSZW (2016) (non-smooth)	AMSE terms: $h^{2\bar{v}}\mathcal{B}(K)\mathcal{B}(K)^T + N^{-2}h^{-k}\Sigma_1(K) + N^{-1}\Sigma_2$ optimal bandwidth rate: $N^{-\frac{2}{2\bar{v}+k}}$ MSE rate: N^{-1} if $\bar{v} = v(K)$ (smooth) $h^{2\bar{v}}$ or $N^{-2}h^{-k}$ if $\bar{v} < v(K)$ (non-smooth)
Density weighted average derivative estimator (PSS-ADE)	
$-\frac{2}{N}\sum_{i=1}^{N}\frac{\partial\hat{f}_{(K,h)}(X_i)}{\partial x}Y_i$ PSS (1989), Powell and Stoker (1996) SZW (2010), KSZW (2016) (non-smooth)	AMSE terms: $h^{2\bar{v}}\mathcal{B}(K)\mathcal{B}(K)^T + N^{-2}h^{-k}\Sigma_1(K) + N^{-1}\Sigma_2$ optimal bandwidth rate: $N^{-\frac{2}{2\bar{v}+k+2}}$ MSE rate: N^{-1} if $\bar{v} = v(K)$ (smooth) $h^{2\bar{v}}$ or $N^{-2}h^{-(k+2)}$ if $\bar{v} < v(K)$ (non-smooth)
Smoothed maximum score	
$\widehat{b} = \arg\max\frac{1}{n}\sum y_i \cdot \Phi(\frac{x_i'b}{h})$ subject to $\lvert b_1\rvert = 1$ $\Phi()$ is an integral of $K()$ Horowitz (1992) Pollard (1993), KZW (2010) (non-smooth)	AMSE terms: $h^{2\bar{v}}\mathcal{B}(K)\mathcal{B}(K)^T + (Nh)^{-1}\Sigma(K)$ optimal bandwidth rate: $N^{-\frac{1}{2\bar{v}+1}}$ MSE rate: $N^{-\frac{2v(K)}{2v(K)+1}}$ if $\bar{v} = v(K)$ (smooth) $h^{2\bar{v}}$ if $\bar{v} < v(K)$ (non-smooth)

the m^{th} partial derivative of kernel density, not listed, we would have $d(k) = k + m$). The average density and the average derivative kernel based estimators have $\omega = 2$; parameter $d(k)$ equals k for the average density and $k+2$ for the PSS-ADE. For averaged kernel based estimators, with $h = cN^{-\zeta}$, Assumptions 1 and 2 result in the following expansion

$$\hat{\delta}_N - \delta_0 = O_p\left(N^{-\zeta\bar{v}}\right) + O_p\left(N^{-\omega+d(k)\zeta}\right)^{1/2} + O_p\left(N^{-1/2}\right). \tag{4}$$

Estimators not listed in Table 1 include Linton (1995) and Ichimura and Linton (2005), where Nagar expansions with terms corresponding to the rates in (4) were derived under smoothness assumptions, and other two-step semiparametric estimators such as the weighted average derivative estimator (Cattaneo et al., 2013). For Robinson's (1989) two-step semiparametric estimator for the partial linear regression model, Linton (1995) developed a Nagar expansion (under sufficient smoothness for the nonparametric regression function) to obtain a second-order optimal bandwidth. Ichimura and Linton (2005) develop a Nagar expansion for a popular two-step treatment effect estimator (of Hirano et al., 2003). Bias representations typically include terms in addition to those associated with the standard (asymptotic) bias rate based on kernel order (also called smoothing bias terms) and may include degrees of freedom bias, leave-in bias or nonlinearity bias (see also Cattaneo et al., 2018). These expansions permit to exploit the trade-offs between first-order and second-order bandwidth-dependent terms even in settings where parametric rate of convergence is attainable. As a result, improved finite sample performance can be achieved. For two-step semiparametric estimators, where the interest is in finite dimensional parameters in the presence of infinite dimensional nuisance parameters, it is well established (e.g., Andrews, 1994) that parametric rates of convergence are permitted in smooth settings when the rate of convergence of the nuisance parameters exceeds $N^{1/4}$. Cattaneo et al. (2018) exploit a leave-one-out kernel estimator to weaken this requirement somewhat by relying on a high level assumption of asymptotic separability in place of the usual stochastic equicontinuity assumption; the resulting rate requirement is $O\left(N^{1/6}\right)$. Nevertheless, even with insufficient smoothness the expansion (4), with possibly additional terms derived for the two-step estimators, could still hold, where the bias rate in the above expansion could be derived from the terms of order $O_p\left(N^{-\zeta\bar{v}}\right)$ and $O\left(\left(N^{-\omega+d(k)\zeta}\right)^{1/2}\right)$, with either one or the other possibly dominating and possibly overshadowing the parametric rate.

The following theorem provides the optimal bandwidth and rate under possibly insufficient but known smoothness. The rate balances the bandwidth dependent part in the variance $O\left(\left(N^{-\omega+d(k)\zeta}\right)^{1/2}\right)$ with the bias $O_p\left(N^{-\zeta\bar{v}}\right)$:

Theorem 1. *Under Assumptions 1 and 2 a bandwidth with the rate $N^{-\eta(\bar{v})}$ where*

$$\eta(\bar{v}) = \frac{\omega}{2\bar{v} + d(k)} \tag{5}$$

provides the best rate for the mean squared error of $\hat{\delta}_N$.

Theorem 1 provides the known optimal rate results for the bandwidth when the usual smoothness assumptions hold, but the rate determined by (5) would be the best even without sufficient smoothness. Details for various estimators are provided in Table 1.

3. RATES WITH UNKNOWN SMOOTHNESS

3.1. Unknown degree of smoothness with $\bar{v} > 0$. We focus here on the uncertainty about the true degree of smoothness and thus the rate of the bias. The density is assumed to exist and to satisfy some degree of smoothness, so that $\bar{v} > 0$.

KSZW (2016) explored the relation between smoothness and the bias and provided estimators for the rate of the bias, $\bar{v}$, which depends on the unknown degree of smoothness (of the underlying functions) and the asymptotic bias of kernel estimators. This, to some extent, followed the approach of Woodroofe (1970) for density estimation.

To estimate the rate of the bias, it is assumed that an oversmoothed bandwidth h_o 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 for average derivative estimation, as it is known to oversmooth in that case.

The consistent estimation of the rate of the bias then requires a sequence of bandwidths $\{h_t\}_{t=1}^{H}$ that satisfy $h_t = c_t h_o N^{\gamma_t}$ with $c_t > 0$, where $\gamma_t \geq 0$ is a strict increasing deterministic sequence such that $h_H = c_H h_o N^{\gamma_H} \to 0$. For PSS-ADE, for example, if h_o is given by cross-validation that has the rate $N^{-\frac{1}{2\bar{v}+k}}$, we should select $\gamma_H < \frac{1}{2\bar{v}+k}$. The estimator is based on a set $\mathcal{T}$ which comprises all pairs $\{(h_t, h_{t'}),\ \ t, t' = 1, ...H$ with $t' < t\}$ and has a cardinality Q: $2 \leq Q \leq \frac{H(H+1)}{2}$. The estimator for $\bar{v}$ can then be obtained easily as

$$\widehat{\bar{v}} = \frac{\sum_{(t,t')\in\mathcal{T}} \ln\left[\left(\hat{\delta}_N(K,h_t) - \hat{\delta}_N(K,h_{t'})\right)^2\right] \cdot \left(\ln {h_t}^2 - \frac{1}{Q}\sum_{(t,t')\in\mathcal{T}} \ln {h_t}^2\right)}{\sum_{(t,t')\in\mathcal{T}} \left(\ln {h_t}^2 - \frac{1}{Q}\sum_{(t,t')\in\mathcal{T}} \ln {h_t}^2\right)^2}. \tag{6}$$

The convergence and associated optimality result is established in the following theorem, see also Theorem 1 in KSZW (2016).

Theorem 2. *Under Assumptions 1 and 2 the estimator for $\bar{v}, \widehat{\bar{v}}$, given by* (6) *satisfies $\widehat{\bar{v}} - \bar{v} = o_p((\ln N)^{-1})$. A bandwidth vector with the optimal rate is consistently estimated by $\widehat{h^{opt}} = cN^{-\eta(\widehat{\bar{v}})}$, with $\eta(\bar{v})$ specified in (5).*

While the theorem provides a consistency result for the degree of smoothness, the convergence rate is slow.[1]

Next we turn to the setting where the density may not even exist, in which case the standard non- and semiparametric kernel estimators may diverge point-wise.

[1] As a referee pointed out the estimation of the degree of smoothness bears similarity to the test for smoothness in Mukherjee et al. (2016) in the context of adaptive Lepski estimation of nonlinear functionals of density.

3.2. Unknown degree of smoothness; density may not exist. As is shown in ZW (2008), if the distribution is not absolutely continuous and the density at a point does not exist, then the kernel estimator at that point may diverge to infinity. Other non- and semiparametric estimators similarly lose their convergence properties when there is no smoothness. A way to examine the limit processes for such estimators is to consider the estimator and its functional object of interest as a generalized function: a functional on the space D_m of well behaved sufficiently (m times) continuously differentiable functions with either bounded support or strict tail conditions. So, e.g. while density does not exist pointwise because the distribution function is not absolutely continuous, in the dual space, D_m^*, of continuous linear functionals on D_m density is always defined as a generalized derivative of the (locally integrable) distribution function, $F \in D_m^*$. In the univariate case, the value of the functional of the density is given by $(f, \psi) = -(F, \psi') = -\int F(x)\,\psi'(x)\,dx$ for $\psi \in D_m(R)$ with $m \geq 1$.

Denote by $\tilde{F}$ the kernel estimator of the distribution function so that the kernel density estimator, $\hat{f}$, is such that $\int^x \hat{f}(w)\,dw = \tilde{F}(x)$, $x \in R^k$. The theorem below (a version of Theorem 1, ZW, 2017) shows that $\sqrt{N}\left(\tilde{f} - f, \psi\right)$ is distributed as a Gaussian generalized process with a (generalized) bias functional $h^{v(K)}B(h,K)$, where $v(K)$ denotes the order of the kernel K, a result also presented in Table 1.

Theorem 3. *If $h \to 0$ and $h^{2v(K)}N = O(1)$ as $N \to \infty$ the sequence of generalized random processes $\sqrt{N}\left(\widehat{f} - f - h^{v(K)}B(h,K)\right)$ converges to a generalized Gaussian process with mean functional zero and covariance functional C which for any $\psi_1, \psi_2 \in D_{k+v(K)}(R^k)$ provides*

$$(C, (\psi_1, \psi_2)) = E\left([\psi_1(x) - E\psi_1(x)]\,[\psi_2(x) - E\psi_2 x)]\right) = cov(\psi_1, \psi_2). \tag{7}$$

If $Nh^{2v(K)} \to 0$, then $\hat{f} - f$ converges at the parametric rate $\sqrt{N}$ to a generalized zero mean Gaussian process with covariance functional C in (7).

The zero mean generalized Gaussian process can be described as a generalized process representing the derivative, ∂U, of the Brownian bridge, U (see Gelfand and Vilenkin, 1964, p.250 for the description of the random process representing the generalized derivative of a Wiener process). Achieving the parametric rate this way does come at a cost as it does not permit the evaluation of pointwise behavior.

Similar results are discussed in TZW (2014) for the estimator of a conditional distribution (for a continuous but not necessarily absolutely continuous marginal distribution) in terms of its generalized function, see Table 1. It is defined as a generalized derivative of the joint distribution function (or of the copula) with respect to the marginal distribution function. Recall the usual kernel estimator of conditional distribution:

$$\hat{F}_{y|x}(x,y) = \frac{\sum_{i=1}^n \bar{G}(y - y_i)\,K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^n K\left(\frac{x_i - x}{h}\right)}, \tag{8}$$

where $\bar{G}$ is the indicator function $I(w > 0)$ and K is an integral of a 2nd order bounded product kernel on R^k. TZW (2014) provides a general result for samples under mixing conditions. The conditioning variable x is represented via the transform into marginals $F_x = (F_{x_1}, ..., F_{x_k})$, and the function $\psi(F_x)$ has $W = [0,1]^k$ as domain of definition. The operator ∂^k represents the differentiation $\frac{\partial^k}{\partial F_{x_1} ... \partial F_{x_k}}$ for functions of F_x.

Theorem 4. *Let $h = cN^{-\alpha}$, where $\alpha > 1/4$. The estimator $\hat{F}_{y|x}(x,y)$ as a generalized random function on $D_m(W)$, with $m \geq 2k + v(K)$, converges at the rate $N^{-1/2}$ to the conditional distribution generalized function $F_{y|x}$; the limit process for $\sqrt{N}(\hat{F}_{y|x} - F_{y|x})$ on $D_m(W)$ is given by a $\psi \in D_m(W)$ indexed random functional, $Q_{y|x}$ with $(Q_{y|x}, \psi) =$*

$$(-1)^k [\int F_{xy} \left(\sum_{i=1}^{k} \frac{\partial \partial^k(\psi(F_x))}{\partial F_{x_i}} U_{x_i} \right) dF_x + \int F_{xy} (\partial^k \psi)(F_x) dU_x + \int (\partial^k \psi)(F_x) U_{xy} dF_x],$$

Here $U_x := (U_{x_1}, .., U_{x_k})'$, and U_{xy} are Brownian bridge processes with dimension k and $k+1$, correspondingly; as a generalized random process the limit process $Q_{y|x}$ of $\sqrt{N}(\hat{F}_{y|x} - F_{y|x})$ is Gaussian with mean functional zero and covariance bilinear functional C, given for any ψ_1, ψ_2 by

$$(C, (\psi_1, \psi_2)) = cov[(Q_{y|x}, \psi_1), (Q_{y|x}, \psi_2).$$

Because of the continuity of the differentiation operator in generalized functions, similar theorems about conditional density are also derived in TZW (2014). While the results, within the space of generalized functions, appear somewhat complicated, the values of the functionals can be easily computed and are convenient to use for inference as in TZW (2014). They are reminiscent of statistics based on an infinite number of moment conditions.

4. IMPROVED ESTIMATORS

Whether smoothness levels are known or not, bias terms can dominate the stochastic expansions asymptotically and even if their asymptotic impact vanishes (i.e., they are not first order important) these terms can have a serious impact in finite samples. Removal of asymptotic bias is important for validity of inference on the nonparametric estimators and estimators that rely on them, and suitable approaches for this critically depend on whether smoothness levels are known. The finite sample performance, though, may also depend on the magnitude of derivatives suggesting that more conservative approaches may be called for.

The number of contributions on bias reduction in the nonparametric literature is large. An important contribution in this literature is related to the use of higher order kernels, first introduced by Bartlett (1963), which improve the rate of the asymptotic bias without affecting the asymptotic variance. Subject to sufficient smoothness, higher order kernels play an important role in establishing the parametric rate of convergence of the density weighted average derivative (PSS-ADE), amongst others, by permitting the removal of the asymptotic bias of terms of order lower. Undersmoothing and plug-in bias correction methods, used to ensure that the bias becomes first order negligible, also require knowledge

about smoothness. Moreover, as argued in Calonico et al. (2018), plug-in bias correction methods are likely to have a first order impact for inference purposes (and construction of confidence intervals) for which they propose a suitable robust bias-corrected inference approach. Compared to the simpler undersmoothing approach, plug-in bias correction methods do typically require estimation of higher order derivatives and selection of associated bandwidths; of course, this holds only when smoothness conditions are satisfied. Cross fitting (sampling splitting) methods are known to eliminate "own observation" bias and have been used by Bickel and Ritov (1988) and PSS (1989) amongst others; double robust (Chernozhukov et al., 2018) and cross fit doubly robust estimators (Newey et al., 2018) that use separate subsamples to estimate nuisance functions in semiparametric estimation allow remainder terms to converge at faster rates.

Schucany and Sommers (1977) considered the use of a generalized jackknife approach for constructing a higher order kernel, an approach which also can be described as a model-averaging approach. They construct a suitable nonconvex weighting of two univariate 2nd order kernel density estimators that removes bias terms of order $O(h^2)$. In a similar vein, Bierens (1987) proposed a kernel regression estimator with asymptotic negligible bias by balancing the bias of two $v\,(K)$ order kernel based regression estimators with bandwidths $cN^{-1/(2v(K)+k)}$ and $cN^{-\delta/(2(v(K)+k)}$ for some $\delta \in (0,1)$; the resulting estimator attained the optimal rate $N^{v(K)/(2v(K)+k)}$. This rate can be made arbitrarily close to the parametric rate by increasing the order of the kernel $v(K)$ subject to smoothness requirements. PSS (1989) constructed a sequence of bandwidths that in a weighted average combination of PSS-ADE estimators leads to a bias reduction of order $v(K)$. More recently, Cattaneo et al. (2013) considered the generalized jackknife for kernel weighted average derivatives, where its use was based on the realization that the nonlinearity bias (not only the smoothing bias) admits a polynomial expansion in the bandwidth that makes it amenable to elimination by means of generalized jackknifing as well.

As the bias expansions critically depend on assumed smoothness requirements, the above bias reduction procedures may no longer be valid when there is insufficient smoothness. As shown in KSZW (2016) and discussed in the next section, the finite sample behavior of estimators shows a significant variability even when formally the smoothness assumption may be satisfied. This happens when the bounds of integrals of derivatives to appropriate order are large, and consequently there is a lot of uncertainly about how far to take the expansion of the bias function to obtain the best approximation in finite samples.

In an attempt to obtain an estimator that is adaptive to the unknown smoothness, while achieving asymptotically the best available (a priori unknown) rate, KZW (2006) propose a **model average estimator** that weights estimators calculated for different bandwidths and kernels. The choice of the weights is based on minimizing the asymptotic mean squared error (AMSE). Weights that minimize the estimated AMSE provide consistent estimates of the weights which would minimize trace of the true AMSE. In Henderson and Parmeter (2016) alternative weighting methods were proposed that, while suitable for nonparametric kernel regression, do not easily generalize to all estimators we consider. Specifically, they use least-squares weighting and cross validation, advocated by Li et al. (2015) and Liu (2018) respectively, to account for the uncertainty about bandwidth and kernel selection.

The benefit of choosing the weights based on the trace of the AMSE, is that it provides large bias-tradeoffs without affecting the variance. KZW (2007) and SZW (2010) studied the use of the model averaging **combined estimator** that minimizes the trace of the AMSE for density and average derivative estimators and KSZW (2016) generalized the results for local kernel based estimators and average kernel based estimators, see also Kotlyarova et al. (2011). They reveal that for both local and average kernel based estimators, appropriate selection of the 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. The following theorem from KSZW (2016) provides this important result

Theorem 5. *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 S 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*

$$trAMSE(\textstyle\sum_{s=1}^{S} a_s \hat{\delta}_N(K, h_s)) < trAMSE\left(\delta_N(K, h^{opt})\right). \tag{9}$$

More flexibility in the choice of bandwidth is permitted when we allow for multiple kernels, permit unequal bandwidths for the different components of the vector $\hat{\delta}_N$, or increase the number of bandwidth vectors S. The proof can be modified to allow for a higher smoothness parameter $\bar{v}$. The condition $\bar{v} \leq 2$ in Theorem 5 holds if K is a second order kernel, and also for higher order kernels when bias goes to zero no faster than $O(h^2)$.

The implementation of this robust averaging estimator does require the use of consistent estimators for the asymptotic bias and (co)variances that do not rely on smoothness assumptions. In particular, the estimation of the asymptotic bias is challenging. A comparison of the performance of the robust averaging estimator with its infeasible counterpart that utilizes the true bias in KSZW (2016) highlights this. While the infeasible averaged estimator provides clear improvements relative to using estimators based on a single kernel-bandwidth combination (even with an optimal bandwidth), the results for the feasible averaged estimators are somewhat more mixed. Nevertheless, the results are overall encouraging in the case of averaged estimators, such as PSS-ADE and smoothed maximum score, where careful implementation of the bias estimation strategy indeed reduces sensitivity to the kernel/bandwidth choice and provides more stable results. KSZW (2016) implemented two approaches for estimating the asymptotic bias. One based on the strategy of Woodroofe (1970) uses the estimate of the smoothness parameter $\bar{v}$ given in (6); SZW (2010) used this approach for the PSS-ADE. The other is based on a heuristic approach suggested in KZW (2007); it relies on averaging estimators over different kernels using small (undersmoothed) bandwidths.

For the estimation of the covariances, consistent plug-in estimators for the leading terms of the asymptotic expansion can be used, or alternatively, the bootstrap. Details for the asymptotic expansion of the covariance between pairs of estimators based on different kernel and/or bandwidths for the PSS-ADE, average density and the kernel regression estimator, are provided in SZW (2010) and KSZW (2016). While bootstrap estimates of

Table 2: Derivative comparison: Smooth and Nonsmooth results

density f	$\int (f''(x))^2\,dx$	$\int (f^{(3)}(x))^2 dx$	
$N(0,1)$	0.212	0.529	
$0.5N(-1,1)+0.5N(1.75,0.25)$	1.76	17.4	Henderson and Parmeter (2016)
$0.333N(-2.75,1)+0.334N(0,1)$ $+0.333N(2.75,1)$	0.05	0.19	Stoker (1993)
$0.5N(0,1)+0.3N(0.8,0.01)$ $+0.2N(1.2,0.01)$	3045	767×10^3	Härdle et al. (1998)
$\sum_{l=0}^{2}\frac{2}{7}N(\frac{12l-15}{7},\frac{2^2}{7^2})$ $+\sum_{l=8}^{10}\frac{1}{21}N(\frac{2l}{7},\frac{1}{21^2})$	9897	107×10^4	Marron and Wand (1992): Discrete Comb
$0.49N(-1,\frac{4}{9})+0.49N(1,\frac{4}{9})$ $+\sum_{l=0}^{6}\frac{1}{350}N(\frac{l-3}{2},0.01^2)$	1209×10^2	302×10^7	Marron and Wand (1992): Double Claw

covariances for local kernel based estimators are straightforward to obtain, bias corrections for estimators that are based on U-statistics, such as average kernel based estimators, may be required (see Cattaneo et al., 2014b, and Kotlyarova et al., 2011). Sample splitting could be employed to remove the need to estimate covariances; obviously sample splitting entails a loss of efficiency due to the reduced sample size used for each estimator.

Establishing a theoretical framework that accounts for the estimated weights of model averaging combined estimator is challenging. Limited results in this area are available, see, e.g. Hjort and Claeskens (2003), Hansen (2014), and Zhang and Liu (2019).

5. EVALUATION OF FINITE SAMPLE RESULTS

Even when asymptotically valid (with sufficient differentiability), bias expansions of kernel estimators often exhibit errors which are proportionate to higher-order derivatives, as in the case of density estimation. The finite-sample mean squared errors (MSE) for kernel-based estimators will reflect the values of those derivatives; mean integrated squared errors (MISE) are impacted by the integrals of squares of those derivatives. Consider a univariate kernel density estimator. For a second-order kernel, the asymptotic MISE at the MISE-minimizing bandwidth is proportional to $\left(\int (f''(x))^2\,dx\right)^{1/5}$ (Rosenblatt, 1956) and the accuracy of the asymptotic bias approximation depends on the bound L such that $\int (f^{(3)}(x))^2 dx \leq L$ (see the discussion in KSZW, 2016). In Table 2 below we report the values of the integrals of such derivatives for the standard normal density and a selection of mixtures of normal variables examined in the literature.

Normal mixtures can be associated with very high values of the bounds for the bias, while being infinitely differentiable. The typical error bounds for distributions used for evaluating finite sample performance of econometric estimators in the literature, including normal mixtures (e.g. in Bierens, 1987, Stoker , 1993, and Newey et al, 2004), do not

capture the possibility of the more extreme values. Henderson and Parmeter (2016), who also consider the benefits of model averaging, used a mixture of normals with integrals that are somewhat larger than that of the standard normal. The examples where the values of the bounds are particularly high are the trimodal mixture in Härdle et al. (1998) and the discrete comb and double claw in Marron and Wand (1992); these distributions are characterized by multimodality, wiggles and high peakedness. Making use of such distributions in simulations is warranted when evaluating the finite sample performance of non- and semiparametric estimators, as they may better capture the effect of non-smoothness and "bunching" present in the data. KZW (2007), SZW (2010), and KSZW (2016) considered these distributions when evaluating the performance of kernel estimators and kernel-based statistics.

For many econometric functionals, performance of kernel-based estimators where such bounds vastly exceed those for the Gaussian could be expected to be much worse than indicated by results obtained in typical smooth simulations settings. For the trimodal distribution from Härdle et al. (1998), the root mean integrated squared error of the kernel density estimator, for instance, is 2-3 times larger than RMISE of the normal density (KZW, 2007); the RMISE of the kernel density estimator for the trimodal distribution is comparable to the RMISE of the kernel density estimator for a non-smooth density. Similarly, the discrete comb and double claw mixtures of Marron and Wand (1992) can lead to large errors in finite sample performance of kernel estimators. Indeed, SZW (2010) report that RMSE for the PSS-ADE with regressors drawn from such mixtures are 4 to 10 times larger than for normally distributed regressors. The results in such experiments do not necessarily produce uniformly larger errors; simulations in SWZ (2010) for the PSS-ADE and in KSZW (2016) for the density estimator demonstrate a large variability of the RMSE over different choices of kernel and bandwidth which do not necessarily favor the theoretically optimal choice and thus undermine the usual methods of bandwidth selection.

These findings indicate the importance of extending finite sample evaluation of kernel-based estimators to more extreme (in terms of magnitude of derivatives) mixtures of normals. Such investigations will provide a more realistic evaluation of the estimators. If performance noticeably deteriorates, this can be taken as an indication for a need to implement procedures aimed at bias reduction such as model averaging via possibly non-convex combinations of estimators.

References

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

[2] Ai, C. and X. Chen (2003): "Efficient estimation of models with conditional moment restrictions containing unknown functions," *Econometrica*, **71**, 1795-1843.

[3] Andrews, D.W.K. (1994): "Asymptotics for semiparametric econometric models via stochastic equicontinuity," *Econometrica*, **62**, 43-72.

[4] Bartlett, M.S. (1963): "Statistical estimation of density functions," *Sankhya A,* **25**, 245-254.

[5] Bickel, P.J. and Y. Ritov (1988): "Estimating integrated squared density derivatives: Sharp best order of convergence estimates," *Sankhya: The Indian Journal of Statistics, Series A*, **50**, 381-393.

[6] Bierens, H.J. (1987) "Kernel Estimators of Regression Functions" in *Advances in Econometrics: Fifth World Congress*, Cambridge University Press, Cambridge, 99-144.

[7] Breiman, L. (1996) "Stacked regressions" in *Machine Learning*, **24**, 49-64.

[8] Calonico, S., M.D. Cattaneo and M.H. Farrell (2018): "On the effect of bias estimation on coverage accuracy in nonparametric inference," *Journal of the American Statistical Association*, **113**, 767-791.

[9] Cattaneo, M.D., R.K. Crump and M. Jansson (2013): "Generalized jackknife estimators of weighted average derivatives," *The Journal of the American Statistical Associated*, **108**, 1243-1256.

[10] Cattaneo, M.D., R.K. Crump and M. Jansson (2014a): "Small bandwidth asymptotics for density-weighted average derivatives," *Econometric Theory*, **30**, 176-200.

[11] Cattaneo, M.D., R.K. Crump and M. Jansson (2014b): "Bootstrapping density-weighted average derivatives," *Econometric Theory*, **30**, 1135-1164.

[12] Cattaneo, M.D. and M. Jansson (2018): "Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency," *Econometrica*, **86**, 955-995.

[13] Chen, X., Linton, O., P.M. Robinson (2001): "The estimation of conditional densities, " in *Asymptotics in Statistics and Probability, Festschrift for George Roussas*, ed. M.L. Puri. VSP International Science Publishers, Netherlands.

[14] Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018): "Double/debiased machine learning for treatment and structural parameters," *Econometrics Journal*, **21**, C1-C68.

[15] Gelfand, I.M. and N. Ya. Vilenkin (1964): *Generalized Functions, Volume 4, Applications of harmonic analysis*, Academic Press.

[16] Hansen, B.E. (2014): "Nonparametric Sieve Regression: Least Squares, Averaging Least Squares, and Cross-Validation," in Racine, J.S., L. Su and A. Ullah (eds.) *The Oxford Handbook of Applied Nonparametric and Semiparametric Econometrics and Statistics.* Oxford University Press.

[17] Hansen, B.E. and J.S. Racine (2012): "Jackknife model averaging," *Journal of Econometrics*, **167**, 38-46.

[18] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A., 1998, *Wavelets, Approximation, and Statistical Applications* (New York: Springer-Verlag).

[19] Henderson, D.J. and C.F. Parmeter (2016): Model averaging over nonparametric estimators," in Gonzalez-Rivera, G., R.C. Hill and T.-H. Lee (eds.) *Advances in Econometrics, Vol. 36, Essays in Honor of Aman Ullah*, 561-589.

[20] Hirano, K., G. Imbens, and G. Ridder (2003): "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, **71**, 1161-1189.

[21] Hjort, N.L., and G. Claeskens (2003): "Frequentist model average estimators". *Journal of the American Statistical Association,* **98**: 879-899.

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

[23] Ichimura, H. and O. Linton (2005): "Asymptotic expansions for some semiparametric program evaluation estimators," in Andrews, D.W.K. and J. Stock (eds.), *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg.* Cambridge University Press, Cambridge, UK, pp. 149-170.

[24] Ichimura, H., and W.K. Newey (2017): "The influence function of semiparametric estimators," Cemmap Working paper, CWP06/17, **??** .

[25] Kotlyarova, Y., M. Schafgans and V. Zinde-Walsh (2011): "Adapting kernel estimation to uncertain smoothness," Sticerd Discussion Paper No. EM/2011/557, London School of Economics.

[26] Kotlyarova, Y., M. Schafgans and V. Zinde-Walsh (2016): "Exploration of smoothness: bias and efficiency of nonparametric kernel estimators," in Gonzalez-Rivera, G., R.C. Hill and T.-H. Lee (eds.) *Advances in Econometrics, Vol. 36, Essays in Honor of Aman Ullah*, 561-589.

[27] Kleven, H., (2016): "Bunching," *Annual Review of Economics*, **8**, 435-464,

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

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

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

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

[32] Lepski, O.V. and V.G. Spokoiny (1997): "Optimal pointwise adaptive methods in nonparametric estimation," *The Annals of Statistics*, **25**, 2512-2546.

[33] Li, Q., O. Linton, Z. Lu (2015): "A flexible semiparametric forecasting model for time series," *Journal of Econometrics*, **187**, 345-357.

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

[35] Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, **63**, 1079-1112.

[36] Liu, C.-A. (2018): "Averaging estimators for kernel regressions," *Economics Letters*, **171**, 102-105.

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

[38] Mukherjee, R., E. T. Tchetgen, and J. Robins (2016) : "Lepski's method and adaptive estimation of nonlinear integral functionals of density," **??** .

[39] Nadaraya, E.A. (1964): "On estimating regression," *Theory of Probability and its Applications*, **9**, 141-142.

[40] Nagar A.L. (1959) "The bias and moment matrix of the general k-class estimators of the parameters in simultaneous equations," *Econometrica*, **27**, 575-595.

[41] Nagar A.L. and A. Ullah (1970) "On Nagar's approximations to moments of the two-stage least-squares estimator," *Indian Economic Review*, **5**, 163-168.

[42] Newey, W.K., F. Hsieh and J.K. Robins (1998): "Undersmoothing and bias corrected functional estimation," working paper, MIT.

[43] Newey, W.K., F. Hsieh and J.K. Robins (2004): "Twicing kernels and a small bias property of semiparametric estimators," *Econometrica*, **72**, 947-962.

[44] Newey, W.K., and J.K. Robins (2018): "Cross-fitting and fast remainder rates for semiparametric estimation," **??**

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

[46] Parzen, E. (1962): "On estimation of a probability density function and mode," *The Annals of Mathematical Statitics*, **33**, 1065-1076.

[47] Pollard, D. (1993): "The asymptotics of a binary choice model," Technical Report, Department of Statistics, Yale University.

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

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

[50] Racine, J., Q. Li and X. Zhu (2004): "Kernel estimation of multivariate conditional distributions," *Annals of Economics and Finance*, **5**, 211-235.

[51] Robinson, P.M. (1988): "Root-N consistent semiparametric estimation," *Econometrica*, **56**, 931-954.

[52] Rosenblatt (1956): "Remarks on some nonparametric estimates of a density function," *The Annals of Mathematical Statitics*, **27**, 832-837.

[53] Sargan, J.D. (1974): "The validity of Nagar's expansion for the momnets of econometric estimators," *Econometrica*, **42**, 169-176.

[54] Schafgans, M.M.A. and V. Zinde-Walsh (2010): "Smoothness adaptive average derivative estimation," *Econometrics Journal*, **13**, 40–62.

[55] Schucany, W.R. and J.P. Sommers (1977): "Improvement of kernel type density estimators," *Journal of the American Statistical Association*, **72**, 420-423.

[56] Srinivasan, T.N. (1970): "Approximations to finite sample moments of estimators whose exact sampling distribution is unknown," *Econometrica*, **38**, 533-541

[57] Stoker, T.M. (1993): "Smoothing bias in density derivative estimation," *Journal of the American Statistical Association*, **88,** 855-863.

[58] Tukey, J.W. (1977): *Exploratory Data Analysis.* Reading, Massachusetts, Addison-Wesley.

[59] Tuvaandorj, P. and V. Zinde-Walsh (2014): "Limit theory and inference about conditional distributions," Working paper, McGill University.

[60] Ullah, A. and H. Wang (2013): "Parametric and nonparametric frequentist model selection and model averaging," *Econometrics*, MDPI, Open Access Journal, **1**, 1-23.

[61] Watson, G.S. (1964): "Smooth regression analysis," *Sankhya, Series A*, **26**, 359-372.

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

[63] Zhang, X. and C.A. Liu (2019): "Inference after model averaging in linear regression models," *Econometric Theory,* **35**: 816-841

[64] Zinde-Walsh, V. (2008): "Kernel estimation when density may not exist," *Econometric Theory*, **24**, 696-725.

[65] Zinde-Walsh, V. (2013): "Nonparametric functionals as generalized functions", arXiv:1303.1435, Statistics Theory (math.ST)

[66] Zinde-Walsh, V. (2017): "Kernel estimation when density may not exist – A corrigendum, *Econometric Theory*, **33**, 1259-1263.