Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Generalized Jackknife Estimators of Weighted Average Derivatives

Matias D. CATTANEO, Richard K. CRUMP, and Michael JANSSON

With the aim of improving the quality of asymptotic distributional approximations for nonlinear functionals of nonparametric estimators, this article revisits the large-sample properties of an important member of that class, namely a kernel-based weighted average derivative estimator. Asymptotic linearity of the estimator is established under weak conditions. Indeed, we show that the bandwidth conditions employed are necessary in some cases. A bias-corrected version of the estimator is proposed and shown to be asymptotically linear under yet weaker bandwidth conditions. Implementational details of the estimators are discussed, including bandwidth selection procedures. Consistency of an analog estimator of the asymptotic variance is also established. Numerical results from a simulation study and an empirical illustration are reported. To establish the results, a novel result on uniform convergence rates for kernel estimators is obtained. The online supplemental material to this article includes details on the theoretical proofs and other analytic derivations, and further results from the simulation study.

KEY WORDS: Bias correction; Semiparametric estimation; Uniform consistency.

1. INTRODUCTION

Two-step semiparametric *m*-estimators are an important and versatile class of estimators whose conventional large-sample properties are by now well understood. These procedures are constructed by first choosing a preliminary nonparametric estimator, which is then “plugged in” in a second step to form the semiparametric estimator of the finite-dimensional parameter of interest. Although the precise nature of the high-level assumptions used in conventional approximations varies slightly, it is possible to formulate sufficient conditions so that the semiparametric estimator is $\sqrt{n}$-consistent (where n denotes the sample size) and asymptotically linear (i.e., asymptotically equivalent to a sample average based on the influence function). These results lead to a Gaussian distributional approximation for the semiparametric estimator that, together with valid standard-error estimators, theoretically justify classical inference procedures, at least in large samples. Newey and McFadden (1994, Sec. 8), Chen (2007, Sec. 4), and Ichimura and Todd (2007, Sec. 7), among others, gave detailed surveys on semiparametric inference in econometric theory, and further references in statistics and econometrics.

A widespread concern with these conventional asymptotic results is that the (finite sample) distributional properties of semiparametric estimators are widely believed to be much more sensitive to the implementational details of its nonparametric ingredient (e.g., bandwidth choice when the nonparametric estimator is kernel-based) than predicted by conventional asymptotic theory, according to which semiparametric estimators are asymptotically linear with influence functions that are invariant with respect to the choice of nonparametric estimator (e.g., Newey 1994a, Proposition 1). Conventional approximations rely on sufficient conditions carefully tailored to achieve asymptotic linearity, thereby assuming away additional approximation errors that may be important in samples of moderate size. In particular, whenever the preliminary nonparametric estimator enters nonlinearly in the construction of the semiparametric procedure, a common approach is to linearly approximate the underlying estimating equation to characterize the contribution of the nonparametric ingredient to the distributional approximation. This approach leads to the familiar sufficient condition that requires the nonparametric ingredient to converge at a rate faster than $n^{1/4}$, effectively allowing one to proceed “as if” the semiparametric estimator depends linearly on its nonparametric ingredient, which in turn guarantees an asymptotic linear representation of the semiparametric estimator under appropriate sufficient conditions.

In this article we study the large-sample properties of a kernel-based estimator of weighted average derivatives (Stoker 1986; Newey and Stoker 1993), and propose a new first-order asymptotic approximation for the semiparametric estimator based on a quadratic expansion of the underlying estimating equation. The key idea is to relax the requirement that the convergence rate of the nonparametric estimator be faster than $n^{1/4}$, and to rely instead on a quadratic expansion to tease out further information about the dependence of the semiparametric estimator on its nonparametric ingredient, thereby improving upon the conventional (first-order) distributional approximation available in the literature. Although our idea leads to an improved understanding of the differences between linear and nonlinear functionals of nonparametric estimators in some generality, we focus attention on weighted average derivatives to keep the results as

Matias D. Cattaneo is Associate Professor of Economics, Department of Economics, University of Michigan, Ann Arbor, MI 48109-1220 (E-mail: *cattaneo@umich.edu*). Richard K. Crump is Senior Economist, Federal Reserve Bank of New York, 33 Liberty Street, New York, NY 10045 (E-mail: *richard.crump@ny.frb.org*). Michael Jansson is Professor of Economics, Department of Economics, UC Berkeley, 530 Evans Hall #3880, Berkeley, CA 94720-3880 (E-mail: *mjansson@econ.berkeley.edu*). For comments and suggestions, we thank Enno Mammen, Whitney Newey, Jim Powell, Rocio Titiunik, seminar participants at Brown, CEMFI/Universidad Carlos III de Madrid, Harvard/MIT, Mannheim, Michigan, NYU, Toulouse School of Economics, UC-Davis, UCSD, UIUC, UPenn, and Yale, and conference participants at the 2010 World Congress of the Econometric Society, 2011 WEAI Conference, 2012 Winter Meetings of the Econometric Society, 2012 Tsinghua International Conference in Econometrics, and 2012 SETA conference. We also thank the Associate Editor and a referee for helpful recommendations. The first author gratefully acknowledges financial support from the National Science Foundation (SES 0921505 and SES 1122994). The third author gratefully acknowledges financial support from the National Science Foundation (SES 0920953 and SES 1124174) and the research support of *CREATES* (funded by the Danish National Research Foundation).

Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2012.745810

interpretable as possible, and because this estimand is popular in theoretical and empirical works. Indeed, it should be conceptually straightforward to apply the methodology employed herein to other kernel-based semiparametric *m*-estimators at the expense of more considerable notation and technicalities.

We obtain several new results for the kernel-based weighted average derivatives estimator. First, under standard kernel and bandwidth conditions we establish asymptotic linearity of the estimator and consistency of its associated "plug-in" variance estimator under a weaker-than-usual moment condition on the dependent variable. Indeed, the moment condition imposed would appear to be (close to) minimal, suggesting that these results may be of independent theoretical interest in the specific context of weighted average derivatives. More broadly, the results (and their derivation) may be of interest as they are achieved by judicial choice of estimator, and by employing a new uniform law of large numbers specifically designed with consistency proofs in mind.

Second, we also establish asymptotic linearity of the weighted average derivative estimator under weaker-than-usual bandwidth conditions. This relaxation of bandwidth conditions is of practical usefulness because it permits the employment of kernels of lower-than-usual order (and, relatedly, enables us to accommodate unknown functions of lower-than-usual degree of smoothness). More generally, the derivation of these results may be of interest because of its "generic" nature and because of its ability to deliver an improved understanding of the distributional properties of other semiparametric estimators that depend nonlinearly on a nonparametric component.

These results are based on a stochastic expansion retaining a "quadratic" term that is treated as a "remainder" term in conventional derivations. Retaining this term not only permits the relaxation of sufficient (bandwidth) conditions for asymptotic linearity, but also enables us to establish necessity of these sufficient conditions in some cases and, most importantly, to characterize the consequences of further relaxing the bandwidth conditions. Indeed, the third (and possibly most important) type of result we obtain shows that in general the nonlinear dependence on a nonparametric estimator gives rise to a nontrivial "bias" term in the stochastic expansion of the semiparametric estimator. Being a manifestation of the well-known curse of dimensionality of nonparametric estimators, this "nonlinearity bias" is a generic feature of nonlinear functionals of nonparametric estimators whose presence can have an important impact on distributional properties of such functionals.

Because the "nonlinearity bias" is due to the (large) variance of nonparametric estimators, attempting to remove it by means of conventional bias reduction methods aimed at reducing "smoothing" bias, such as increasing the order of the kernel, does not work. Nevertheless, it turns out that this "nonlinearity bias" admits a polynomial expansion (in the bandwidth), suggesting that it should be amenable to elimination by means of the method of generalized jackknifing (Schucany and Sommers 1977). Making this intuition precise is the purpose of the final type of result presented herein. Although some details of this result are specific to our weighted average derivative estimator, the main message is of much more general validity. Indeed, an inspection of the derivation of the result suggests that the fact that removal of "nonlinearity bias" can be accomplished by means of generalized jackknifing is a property shared by most (if not all) kernel-based semiparametric two-step estimators.

The article proceeds as follows. After briefly discussing the related literature in the remaining of this section, Section 2 introduces the model and estimator(s) under study. Our main theoretical results are presented in Section 3, including implementational recommendations for the estimators. Numerical results from a Monte Carlo and an empirical illustration are given in Section 4. Section 5 offers concluding remarks. Appendix A contains proofs of the theoretical results, while Appendix B contains some auxiliary results (of possibly independent interest) about uniform convergence of kernel estimators. The online supplemental material includes details on the theoretical proofs and other analytic derivations, and further results from the simulation study.

1.1 Related Literature

Our results are closely related and contribute to the important literature on semiparametric averaged derivatives (Stoker 1986; see also, e.g., Härdle and Stoker 1989; Härdle et al. 1992; Horowitz and Härdle 1996), in particular shedding new light on the problem of semiparametric weighted average derivative estimation (Newey and Stoker 1993). This problem has wide applicability in statistics and econometrics, as we further discuss in the following section. This problem is conceptually and analytically different from the problem of semiparametric density-weighted average derivatives because a kernel-based density-weighted average derivative estimator depends on the nonparametric ingredient in a linear way (Powell, Stock, and Stoker 1989), while the kernel-based weighted average derivative estimator has a nonlinear dependence on a nonparametric estimator. As a consequence, the alternative first-order distributional approximation obtained by Cattaneo, Crump, and Jansson (2010, in press) for a kernel-based density-weighted average derivatives estimator is not applicable to the estimator studied herein and our main findings are qualitatively different from those obtained in our earlier work. Indeed, a crucial finding in this article is that considering "small bandwidth asymptotics" for the kernel-based weighted average derivative estimator leads to a first-order bias contribution to the distributional approximation (rather than a first-order variance contribution, as in the case of the kernel-based density-weighted average derivative estimator), which in turn requires bias-correction of the estimator (rather than adjustment of the standard-error estimates, as in the case of the kernel-based density-weighted average derivative estimator).

From a more general perspective, our findings are also connected to other results in the semiparametric literature. Mammen (1989) studied the large sample properties of a nonlinear least-squares estimator when the (effective) dimension of the parameter space is allowed to increase rapidly, and found a first-order bias effect qualitatively similar to the one characterized herein. The "nonlinearity bias" we encounter is also analogous in source to the so-called "degrees of freedom bias" discussed by Ichimura and Linton (2005) for the case of a univariate semiparametric estimation problem, but due to the different nature of our asymptotic experiment its presence has first-order consequences herein. Nonnegligible biases in models with covariates of large dimension (i.e., "curse of dimensionality" effects of first order) were also found by Abadie and Imbens (2006), but in the case of their matching estimator the bias in question does not seem to be attributable to nonlinearities. Finally, the recent work by Robins et al. (2008) on higher-order influence functions is also related to our results insofar as it relaxes the underlying convergence rate requirement for the nonparametric estimator. Whereas Robins et al. (2008) were motivated by a concern about the plausibility of the smoothness conditions needed to guarantee existence of $n^{1/4}$-consistent nonparametric estimators in models with large-dimensional covariates, our work seeks to relax this underlying convergence rate requirement for the nonparametric estimator to improve the accuracy of the distributional approximation even in cases where lots of smoothness is assumed. Indeed, our results highlight the presence of a leading, first-order bias term that is unrelated to the amount of smoothness assumed (but clearly related to the dimensionality of the covariates).

2. PRELIMINARIES

2.1 Model and Estimand

We assume that $\mathbf{z}_i = (y_i, \mathbf{x}_i')'$, $i = 1, \ldots, n$, are iid observed copies of a vector $\mathbf{z} = (y, \mathbf{x}')'$, where $y \in \mathbb{R}$ is a dependent variable and $\mathbf{x} = (x_1, x_2, \ldots, x_d)' \in \mathbb{R}^d$ is a continuous explanatory variable with density $f(\cdot)$. A weighted average derivative of the regression function $g(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ is defined as

$$\boldsymbol{\theta} = \mathbb{E}\left[w(\mathbf{x})\frac{\partial}{\partial \mathbf{x}}g(\mathbf{x})\right], \tag{1}$$

where $w(\cdot)$ is a known scalar weight function. (Further restrictions on $w(\cdot)$ will be imposed below.) As illustrated by the following examples, $\boldsymbol{\theta}$ is an estimand that has been widely considered in both theoretical and empirical works.

Example 1. Semilinear Single-Index Models. Let $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ and $g(\mathbf{x}) = G(\mathbf{x}_1'\boldsymbol{\beta}, \mathbf{x}_2)$ with $G(\cdot)$ unknown and $\boldsymbol{\beta}$ the parameter of interest. Partition $\boldsymbol{\theta}$ conformably with $\mathbf{x}$ as $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$. Under appropriate assumptions, $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\theta}_1$ because

$$\boldsymbol{\theta}_1 = \mathbb{E}\left[w(\mathbf{x})\dot{\mathbf{G}}_1(\mathbf{x}_1'\boldsymbol{\beta}, \mathbf{x}_2)\right]\boldsymbol{\beta}, \qquad \dot{\mathbf{G}}_1(\mathbf{u}, \mathbf{x}_2) = \frac{\partial}{\partial \mathbf{u}}G(\mathbf{u}, \mathbf{x}_2).$$

This setup covers several problems of interest. For example, single-index limited dependent variable models (e.g., discrete choice, censored and truncated models) are included with $\mathbf{x}_1 = \mathbf{x}$ and $G(\cdot)$ the so-called link function. Another class of problems fitting in this example are partially linear models of the form $G(\mathbf{x}_1'\boldsymbol{\beta}, \mathbf{x}_2) = \phi_1(\mathbf{x}_1'\boldsymbol{\beta} + \phi_2(\mathbf{x}_2))$ with $\phi_1(\cdot)$ a link function and $\phi_2(\cdot)$ another unknown function. For further discussion on these and related examples, see Stoker (1986), Härdle and Stoker (1989), Newey and Stoker (1993), and Powell (1994).

Example 2. Nonseparable Models. Let $\mathbf{x} = (\mathbf{x}_1', \mathbf{x}_2')'$ and $y = m(\mathbf{x}_1, \varepsilon)$ with $m(\cdot)$ unknown and ε an unobserved random variable. Under appropriate assumptions, including $\mathbf{x}_1 \perp\!\!\!\perp \varepsilon \mid \mathbf{x}_2$, a population parameter of interest is given by

$$\boldsymbol{\theta}_1 = \mathbb{E}\left[w(\mathbf{x})\frac{\partial}{\partial \mathbf{x}_1}m(\mathbf{x}_1, \varepsilon)\right] = \mathbb{E}\left[w(\mathbf{x})\frac{\partial}{\partial \mathbf{x}_1}g(\mathbf{x}_1, \mathbf{x}_2)\right],$$

which captures the (weighted) average marginal effect of $\mathbf{x}_1$ on $m(\cdot)$ over the population $(\mathbf{x}_1', \varepsilon)'$. As in the previous example, $\boldsymbol{\theta}_1$ is the first component of the weighted average derivative $\boldsymbol{\theta}$ partitioned conformably with $\mathbf{x}$. The parameter $\boldsymbol{\theta}_1$ is of interest in policy analysis and treatment effect models. A canonical example is given by the linear random coefficients model $y = \beta_0(\varepsilon) + \mathbf{x}_1'\boldsymbol{\beta}_1(\varepsilon)$, where the parameter of interest reduces to $\boldsymbol{\theta}_1 = \mathbb{E}\left[w(\mathbf{x})\boldsymbol{\beta}_1(\varepsilon)\right]$ under appropriate assumptions. For further discussion on averaged derivatives in non-separable models see, for example, Matzkin (2007), Imbens and Newey (2009), and Altonji, Ichimura, and Otsu (2012).

Example 3. Applications in Economics. In addition to the examples discussed above, weighted average derivatives have also been employed in several specific economic applications that do not necessarily fit the previous setups. Some examples are: (i) Stoker (1989) proposed several tests statistics based on averaged derivatives obtained from economic-theory restrictions such as homogeneity or symmetry of cost functions; (ii) Härdle, Hildenbrand, and Jerison (1991) developed a test for the law of demand using weighted-average derivatives; (iii) Deaton and Ng (1998) employed averaged derivatives to estimate the effect of a tax and subsidy policy change on individuals' behavior; (iv) Coppejans and Sieg (2005) developed a test for nonlinear pricing in labor markets based on averaged derivatives obtained from utility maximization; and (v) Campbell (2011) used averaged derivatives to evaluate empirically the simplifying assumption of large market competition without strategic interactions.

2.2 Estimator and Known Results

Newey and Stoker (1993) studied estimands of the form (1) and gave conditions under which the semiparametric variance bound for $\boldsymbol{\theta}$ is $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\psi}(\mathbf{z})\boldsymbol{\psi}(\mathbf{z})']$, where $\boldsymbol{\psi}(\cdot)$, the pathwise derivative of $\boldsymbol{\theta}$, is given by

$$\boldsymbol{\psi}(\mathbf{z}) = w(\mathbf{x})\frac{\partial}{\partial \mathbf{x}}g(\mathbf{x}) - \boldsymbol{\theta} + [y - g(\mathbf{x})]\,\mathbf{s}(\mathbf{x}),$$

$$\mathbf{s}(\mathbf{x}) = -\frac{\partial}{\partial \mathbf{x}}w(\mathbf{x}) + w(\mathbf{x})\boldsymbol{\ell}(\mathbf{x}), \qquad \boldsymbol{\ell}(\mathbf{x}) = -\frac{\partial f(\mathbf{x})/\partial \mathbf{x}}{f(\mathbf{x})}.$$

The following assumption, which we make throughout the article, guarantees existence of the parameter $\boldsymbol{\theta}$ and semiparametrically efficient estimators thereof.

Assumption 1. (a) For some $S \geq 2$, $\mathbb{E}[|y|^S] < \infty$ and $\mathbb{E}[|y|^S|\mathbf{x}]f(\mathbf{x})$ is bounded. (b) $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{\psi}(\mathbf{z})\boldsymbol{\psi}(\mathbf{z})']$ is positive definite. (c) w is continuously differentiable, and w and its first derivative are bounded. (d) $\inf_{\mathbf{x}\in\mathcal{W}} f(\mathbf{x}) > 0$, where $\mathcal{W} = \{\mathbf{x} \in \mathbb{R}^d : w(\mathbf{x}) > 0\}$. (e) For some $P_f \geq 2$, f is $(P_f + 1)$ times differentiable, and f and its first $(P_f + 1)$ derivatives are bounded and continuous. (f) g is continuously differentiable, and e and its first derivative are bounded, where $e(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$. (g) $\lim_{\|\mathbf{x}\|\to\infty}[f(\mathbf{x}) + |e(\mathbf{x})|] = 0$, where $\|\cdot\|$ is the Euclidean norm.

The restrictions imposed by Assumption 1 are fairly standard and relatively mild, with the possible exception of the "fixed trimming" condition in part (d). This condition simplifies the exposition in our article, allowing us to avoid tedious technical arguments. It may be relaxed to allow for nonrandom asymptotic trimming, but we decided not to pursue this extension to avoid cumbersome notation and other associated technical distractions.

Under Assumption 1, it follows from integration by parts that $\boldsymbol{\theta} = \mathbb{E}[y\mathbf{s}(\mathbf{x})]$. A kernel-based analog estimator of $\boldsymbol{\theta}$ is therefore given by

$$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) = \frac{1}{n}\sum_{i=1}^{n} y_i \hat{\mathbf{s}}_n(\mathbf{x}_i; \mathbf{H}_n),$$

$$\hat{\mathbf{s}}_n(\mathbf{x}_i; \mathbf{H}_n) = -\frac{\partial}{\partial \mathbf{x}} w(\mathbf{x}) - w(\mathbf{x}) \frac{\partial \hat{f}_n(\mathbf{x}; \mathbf{H}_n)/\partial \mathbf{x}}{\hat{f}_n(\mathbf{x}; \mathbf{H}_n)},$$

where

$$\hat{f}_n(\mathbf{x}; \mathbf{H}_n) = \frac{1}{n}\sum_{j=1}^{n} K_{\mathbf{H}_n}(\mathbf{x} - \mathbf{x}_j), \qquad K_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{x}),$$

for some kernel $K : \mathbb{R}^d \to \mathbb{R}$ and some sequence $\mathbf{H}_n$ of diagonal, positive definite $d \times d$ (bandwidth) matrices. By not requiring $\mathbf{H}_n \propto \mathbf{I}_d$ our results allow for different bandwidth sequences for each coordinate of the covariates $\mathbf{x} \in \mathbb{R}^d$. (We thank the Associate Editor for encouraging us relax the restriction $\mathbf{H}_n \propto \mathbf{I}_d$ imposed in an earlier version of the article.)

As defined, $\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ depends on the user-chosen objects K and $\mathbf{H}_n$, but because our main interest is in the sensitivity of the properties of $\hat{\boldsymbol{\theta}}_n$ with respect to the bandwidth matrix $\mathbf{H}_n$, we suppress the dependence of $\hat{\boldsymbol{\theta}}_n$ on K in the notation (and make the dependence on $\mathbf{H}_n$ explicit).

The following assumption about the kernel K will be assumed to hold. [In Assumption 2(c), and elsewhere in the article, we use the notational convention that if $\mathbf{l} = (l_1, l_2, \ldots, l_d)' \in \mathbb{Z}_+^d$ and if $\mathbf{u} = (u_1, u_2, \ldots, u_d)' \in \mathbb{R}^d$, then $\mathbf{u}^{\mathbf{l}}$ denotes $u_1^{l_1} u_2^{l_2}, \ldots, u_d^{l_d}$.]

Assumption 2. (a) K is even, bounded, and twice differentiable, and its first two derivatives are bounded. (b) $\int_{\mathbb{R}^d} \|\dot{\mathbf{K}}(\mathbf{u})\| d\mathbf{u} < \infty$, where $\dot{\mathbf{K}}(\mathbf{u}) = \partial K(\mathbf{u})/\partial \mathbf{u}$. (c) For some $P_K \geq 2$, $\int_{\mathbb{R}^d} |K(\mathbf{u})|(1 + \|\mathbf{u}\|^{P_K}) d\mathbf{u} < \infty$ and for $\mathbf{l} = (l_1, \ldots, l_d)' \in \mathbb{Z}_+^d$,

$$\int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u}) d\mathbf{u} = \begin{cases} 1 & \text{if } l_1 = \cdots = l_d = 0 \\ 0 & \text{if } 0 < l_1 + \cdots + l_d < P_K \end{cases}.$$

(d) $\int_{\mathbb{R}} \bar{K}(u) du < \infty$, where $\bar{K}(u) = \sup_{\|\mathbf{r}\| \geq u} \|\partial(K(\mathbf{r}), \dot{\mathbf{K}}(\mathbf{r})')/\partial \mathbf{r}\|$.

With the possible exception of Assumption 2(d), the restrictions imposed on the kernel are fairly standard. Assumption 2(d) is inspired by Hansen (2008) and holds if K has bounded support or if K is a Gaussian density-based higher-order kernel.

If Assumptions 1 and 2 hold (with P_f and P_K large enough), it is easy to give conditions on the bandwidth vector $\mathbf{H}_n$ under which $\hat{\boldsymbol{\theta}}_n$ is asymptotically linear with influence function $\boldsymbol{\psi}(\cdot)$. For instance, proceeding as by Newey (1994a, 1994b) it can be shown that if Assumptions 1 and 2 hold and if

$$n\lambda_{\max}\left(\mathbf{H}_n^{2P}\right) \to 0, \qquad P = \min(P_f, P_K) \tag{2}$$

and

$$\frac{n\,|\mathbf{H}_n|^2\,\lambda_{\min}\left(\mathbf{H}_n^4\right)}{(\log n)^2} \to \infty, \tag{3}$$

then

$$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) - \boldsymbol{\theta} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{z}_i) + o_p\left(n^{-1/2}\right), \tag{4}$$

where in conditions (3) and (2), and elsewhere in the article, $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the smallest and largest eigenvalue, respectively, of the argument. Moreover, under the same conditions, the variance $\boldsymbol{\Sigma}$ is consistently estimable, as we discuss in more detail in Section 3.4.

The lower bound on (the diagonal elements of) $\mathbf{H}_n$ implied by condition (3) helps ensure that the estimation error of the nonparametric estimator $\hat{f}_n$ is $o_p(n^{-1/4})$ in an appropriate (Sobolev) norm, which in turn is a high-level assumption featuring prominently in Newey's (1994a) work on asymptotic normality of semiparametric m-estimators and in more recent refinements thereof (see, e.g., Chen 2007, for references).

This article explores the consequences of employing bandwidths that are "small" in the sense that Equation (3) is violated. Three main results will be derived. The first result, given in Theorem 1, gives sufficient conditions for Equation (4) that involve a weaker lower bound on $\mathbf{H}_n$ than Equation (3). For $d \geq 3$, the weaker lower bound takes the form $n|\mathbf{H}_n|^2 \to \infty$. The second result, given in Theorem 2, shows that $n|\mathbf{H}_n|^2 \to \infty$ is also necessary for Equation (4) to hold (if $d \geq 3$). More specifically, Theorem 2 finds that if $d \geq 3$, then $\hat{\boldsymbol{\theta}}_n$ has a nonnegligible bias when $n|\mathbf{H}_n|^2 \nrightarrow \infty$. The third result, given in Theorem 3, shows that while $n|\mathbf{H}_n|^2 \to \infty$ is necessary for asymptotic linearity of $\hat{\theta}_n$ (when $d \geq 3$), a bias-corrected version of $\hat{\boldsymbol{\theta}}_n$ enjoys the property of asymptotic linearity under the weaker condition

$$\frac{n\,|\mathbf{H}_n|^{\frac{3}{2}}\,\lambda_{\min}(\mathbf{H}_n)}{(\log n)^{3/2}} \to \infty. \tag{5}$$

In addition, we provide some implementational recommendations. First, in Section 3.3 we derive an "optimal" choice of $\mathbf{H}_n$ based on an asymptotic expansion of the (approximate) mean squared error of $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and used this bandwidth choice to construct a feasible implementation of the bias-corrected version of $\hat{\boldsymbol{\theta}}_n$ proposed in Theorem 3. Second, in Section 3.4, Theorem 4 shows that a modest strengthening of Assumption 1(a) is sufficient to obtain consistency of the conventional plug-in standard-error estimator even when the lower bound on the bandwidth is given by Equation (5).

Remark 1. (i) Most statements involving $\mathbf{H}_n$ can be simplified somewhat in the important special case when $\mathbf{H}_n \propto \mathbf{I}_d$, as $|\mathbf{H}_n| = h_n^d$ and $\lambda_{\min}(\mathbf{H}_n^p) = \lambda_{\max}(\mathbf{H}_n^p) = h_n^p$ (for any $p \in \mathbb{R}$) when $\mathbf{H}_n = h_n\mathbf{I}_d$. For instance, conditions (2) and (5) become $nh_n^P \to 0$ and $nh_n^{3d/2+1}/(\log n)^{3/2} \to \infty$, respectively, when $\mathbf{H}_n = h_n\mathbf{I}_d$. (ii) Imposing Equations (2) and (3), and making assumptions similar to Assumptions 1 and 2, Newey and McFadden (1994, pp. 2212–2214) established asymptotic linearity of the alternative kernel-based estimator

$$\check{\boldsymbol{\theta}}_n = \frac{1}{n}\sum_{i=1}^{n} w(\mathbf{x}_i) \frac{\partial}{\partial \mathbf{x}} \hat{g}_n(\mathbf{x}_i),$$

$$\hat{g}_n(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^{n} K_{\mathbf{H}_n}(\mathbf{x} - \mathbf{x}_j) y_j / \hat{f}_n(\mathbf{x}; \mathbf{H}_n).$$

Their analysis (assumes $\mathbf{H}_n = h_n\mathbf{I}_d$ and) requires $S \geq 4$ to handle the presence of $\hat{g}_n$. The fact that $\hat{\boldsymbol{\theta}}_n$ does not involve $\hat{g}_n$ enables us to develop distribution theory for it under the seemingly minimal condition $S = 2$.

3. THEORETICAL RESULTS

Validity of the stochastic expansion (Equation (4)) can be established by exhibiting an approximation $\hat{\boldsymbol{\theta}}_n^A$ (say) to $\hat{\boldsymbol{\theta}}_n$ satisfying the following trio of conditions:

$$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) - \hat{\boldsymbol{\theta}}_n^A = o_p(n^{-1/2}), \qquad (6)$$

$$\hat{\boldsymbol{\theta}}_n^A - \mathbb{E}\big[\hat{\boldsymbol{\theta}}_n^A\big] = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{z}_i) + o_p(n^{-1/2}), \qquad (7)$$

$$\mathbb{E}\big[\hat{\boldsymbol{\theta}}_n^A\big] - \boldsymbol{\theta} = o(n^{-1/2}). \qquad (8)$$

Variations of this approach have been used in numerous papers, the typical choice being to obtain $\hat{\boldsymbol{\theta}}_n^A$ by "linearizing" $\hat{\boldsymbol{\theta}}_n$ with respect to the nonparametric estimator $\hat{f}_n$ and then establishing Equation (6) by showing in particular that the estimation error of $\hat{f}_n$ is $o_p(n^{-1/4})$ in a suitable norm. This general approach is now well-established in semiparametrics (see, e.g., Newey and McFadden 1994, Sec. 8; Chen 2007, Sec. 4; Ichimura and Todd 2007, Sec. 7, and references therein).

3.1 Asymptotic Linearity: Linear versus Quadratic Approximations

In the context of averaged derivatives, "linearization" amounts to setting $\hat{\boldsymbol{\theta}}_n^A$ equal to

$$\hat{\boldsymbol{\theta}}_n^*(\mathbf{H}_n) = \frac{1}{n}\sum_{i=1}^{n} y_i\hat{\mathbf{s}}_n^*(\mathbf{x}_i;\mathbf{H}_n),$$

where

$$\hat{\mathbf{s}}_n^*(\mathbf{x};\mathbf{H}_n) = \mathbf{s}(\mathbf{x}) - \frac{w(\mathbf{x})}{f(\mathbf{x})}\left[\frac{\partial}{\partial\mathbf{x}}\hat{f}_n(\mathbf{x};\mathbf{H}_n) + \ell(\mathbf{x})\hat{f}_n(\mathbf{x};\mathbf{H}_n)\right]$$

is obtained by linearizing $\hat{\mathbf{s}}_n$ with respect to $\hat{f}_n$. With this choice of $\hat{\boldsymbol{\theta}}_n^A$, conditions (6)–(8) will hold if Assumptions 1 and 2 are satisfied and if Equations (2) and (3) hold. In particular, Equation (3) serves as part of what would appear to be the best-known sufficient condition for the estimation error of $\hat{f}_n$ (and its derivative) to be $o_p(n^{-1/4})$, a property that in turn is used to establish Equation (6) when $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^*(\mathbf{H}_n)$.

In an attempt to establish Equation (6) under a bandwidth condition weaker than Equation (3), we set $\hat{\boldsymbol{\theta}}_n^A$ equal to a "quadratic" approximation to $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ given by

$$\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n) = \frac{1}{n}\sum_{i=1}^{n} y_i\hat{\mathbf{s}}_n^{**}(\mathbf{x}_i;\mathbf{H}_n),$$

where

$$\hat{\mathbf{s}}_n^{**}(\mathbf{x};\mathbf{H}_n) = \hat{\mathbf{s}}_n^*(\mathbf{x};\mathbf{H}_n) + \frac{w(\mathbf{x})}{f(\mathbf{x})^2}[\hat{f}_n(\mathbf{x};\mathbf{H}_n) - f(\mathbf{x})]\left[\frac{\partial}{\partial\mathbf{x}}\hat{f}_n(\mathbf{x};\mathbf{H}_n) + \ell(\mathbf{x})\hat{f}_n(\mathbf{x};\mathbf{H}_n)\right].$$

The use of a quadratic approximation to $\hat{\boldsymbol{\theta}}_n$ gives rise to a "cubic" remainder in Equation (6), suggesting that it suffices to require that the estimation error of $\hat{f}_n$ (and its derivative) be $o_p(n^{-1/6})$. In fact, the proof of the following result shows that the somewhat special structure of the estimator (i.e., $\hat{\mathbf{s}}_n$ is linear in the derivative of $\hat{f}_n$) can be exploited to establish sufficiency of a slightly weaker condition.

Theorem 1. Suppose Assumptions 1 and 2 are satisfied and suppose Equation (2) holds. Then Equation (4) is true if either of the following conditions is satisfied:

(i) $d = 1$ and $n|\mathbf{H}_n|^3 \to \infty$,
(ii) $d = 2$ and $n|\mathbf{H}_n|^2/(\log n)^{3/2} \to \infty$, or
(iii) $d \geq 3$ and $n|\mathbf{H}_n|^2 \to \infty$.

The proof of Theorem 1 verifies Equations (6)–(8) for $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$. Because the lower bounds on $\mathbf{H}_n$ imposed in cases (i) through (iii) are weaker than Equation (3) in all cases, working with $\hat{\boldsymbol{\theta}}_n^{**}$ when analyzing $\hat{\boldsymbol{\theta}}_n$ has the advantage that it enables us to weaken the sufficient conditions for asymptotic linearity to hold on the part of $\hat{\boldsymbol{\theta}}_n$. Notably, the existence of a bandwidth sequence satisfying the assumptions of Theorem 1 holds whenever $P > d$, a weaker requirement than the restriction $P > d + 2$ implied by the conventional conditions (2) and (3). In other words, Theorem 1 justifies the use of kernels of lower order, and thus requires less smoothness on the part of the density f, than do analogous results obtained using $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^*(\mathbf{H}_n)$. Moreover, working with $\hat{\boldsymbol{\theta}}_n^{**}$ enables us to derive necessary conditions for Equation (4) in some cases.

Theorem 2. Suppose Assumptions 1 and 2 are satisfied and suppose Equations (2) and (5) hold.

(a) Small bandwidth bias:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta} = \frac{1}{n\,|\mathbf{H}_n|}[\mathcal{B}_0 + o(1)] + o(n^{-1/2}), \qquad (9)$$

where

$$\mathcal{B}_0 = \left(-K(\mathbf{0}_d)\mathbf{I}_d + \int_{\mathbb{R}^d}\left[K(\mathbf{u})^2\mathbf{I}_d + K(\mathbf{u})\dot{\mathbf{K}}(\mathbf{u})\mathbf{u}'\right]d\mathbf{u}\right) \times \int_{\mathbb{R}^d} g(\mathbf{r})w(\mathbf{r})\ell(\mathbf{r})d\mathbf{r}.$$

(b) Asymptotic Linearity: If either (i) $d = 1$ and $n|\mathbf{H}_n|^3 \to \infty$ or (ii) $d \geq 2$, then

$$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) - \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}(\mathbf{z}_i) + o_p(n^{-1/2}).$$

The first part of Theorem 2 is based on an asymptotic expansion of the approximate bias $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta}$ and shows that, in general, the condition $n|\mathbf{H}_n|^2 \to \infty$ is necessary for Equation (8) to hold when $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$. (We know of no "popular" kernels and/or "plausible" examples of $g(\cdot)$, $w(\cdot)$, and $\ell(\cdot)$ for which $\mathcal{B}_0 = 0$.) The second part of Theorem 2 verifies Equations (6) and (7) for $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$ and can be combined with the first part to yield the result that the sufficient condition $n|\mathbf{H}_n|^2 \to \infty$ obtained in Theorem 1(iii) is also necessary (in general) when $d \geq 3$.

To interpret the matrix $\mathcal{B}_0$ in the (approximate) bias expression (9), it is instructive to decompose it as $\mathcal{B}_0 = \mathcal{B}_0^* + \mathcal{B}_0^{**}$, where

$$\mathcal{B}_0^* = -K(\mathbf{0}_d) \int_{\mathbb{R}^d} g(\mathbf{r}) w(\mathbf{r}) \ell(\mathbf{r}) \mathrm{d}\mathbf{r},$$

and

$$\mathcal{B}_0^{**} = \left(\int_{\mathbb{R}^d} \left[K(\mathbf{u})^2 \mathbf{I}_d + K(\mathbf{u}) \dot{\mathbf{K}}(\mathbf{u}) \mathbf{u}' \right] \mathrm{d}\mathbf{u} \right) \int_{\mathbb{R}^d} g(\mathbf{r}) w(\mathbf{r}) \ell(\mathbf{r}) \mathrm{d}\mathbf{r}.$$

The term $\mathcal{B}_0^*$ is a "leave in" bias term arising because each $\hat{\mathbf{s}}_n(\mathbf{x}_i; \mathbf{H}_n)$ employs a nonparametric estimator $\hat{\mathbf{s}}_n$ that uses the own observation $\mathbf{x}_i$. The other bias term, $\mathcal{B}_0^{**}$, is a "nonlinearity" bias term reflecting the fact that $\hat{\mathbf{s}}_n^{**}$ involves a nonlinear function of $\hat{f}_n$. The magnitude of this nonlinearity bias is $n^{-1}|\mathbf{H}_n|^{-1}$. This magnitude is exactly the magnitude of the pointwise variance of $\hat{f}_n$, which is no coincidence because $\hat{\mathbf{s}}_n^{**}$ involves a term that is "quadratic" in $\hat{f}_n$. (The approximation $\hat{\mathbf{s}}_n^{**}$ also involves a cross-product term in $\hat{f}_n$ and its derivative that, as shown in the proof of Lemma A-3, gives rise to a bias term of magnitude $n^{-1}|\mathbf{H}_n|^{-1}$ when K is even.)

Remark 2. (i) The leave-in-bias can be avoided simply by employing a "leave-one-out" estimator of f when forming $\hat{\mathbf{s}}_n$. (ii) Merely removing leave-in-bias does not automatically render $\hat{\boldsymbol{\theta}}_n$ asymptotically linear unless $n|\mathbf{H}_n|^2 \to \infty$, however, as the nonlinearity bias of the leave-one-out version of $\hat{\boldsymbol{\theta}}_n$ is identical to that of $\hat{\boldsymbol{\theta}}_n$ itself. (iii) Manipulating the order of the kernel (P_K) does not eliminate the nonlinearity bias either, as the magnitude, $n^{-1}|\mathbf{H}_n|^{-1}$, of the bias is invariant with respect to the order of the kernel.

3.2 Asymptotic Linearity Under Nonstandard Conditions

The second part of Theorem 2 suggests that if $d \geq 3$, then a bias-corrected version of $\hat{\boldsymbol{\theta}}_n$ might be asymptotically linear even if the condition $n|\mathbf{H}_n|^2 \to \infty$ is violated. Indeed, the method of generalized jackknifing can be used to arrive at an estimator $\tilde{\boldsymbol{\theta}}_n$ (say) whose (approximate) bias is sufficiently small also when $n|\mathbf{H}_n|^2 \nrightarrow \infty$. This approach is based on the following refinement of Theorem 2(a).

Lemma 1. Suppose the assumptions of Theorem 2 hold. Then, for any $c > 0$,

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(c\mathbf{H}_n)] - \boldsymbol{\theta} = \frac{c^{-d}}{n|\mathbf{H}_n|} \left[\mathcal{B}_0 + \sum_{j=1}^{\lfloor (P-1)/2 \rfloor} c^{2j} \mathcal{B}_j(\mathbf{H}_n) \right] + o(n^{-1/2}), \tag{10}$$

where $\{\mathcal{B}_j(\cdot) : 1 \leq j \leq \lfloor (P-1)/2 \rfloor\}$ are functions depending only on the kernel function and the data-generating process. (The $\{\mathcal{B}_j(\cdot)\}$ are defined in Lemma A-3 in the Appendix.)

Accordingly, let J be a positive integer with $J < 1 + d/2$, let $\mathbf{c} = (c_0, \ldots, c_J)' \in \mathbb{R}_{++}^{J+1}$ be a vector of distinct constants with $c_0 = 1$, and define

$$\begin{pmatrix} \omega_0(\mathbf{c}) \\ \omega_1(\mathbf{c}) \\ \vdots \\ \omega_J(\mathbf{c}) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & c_1^{-d} & \cdots & c_J^{-d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & c_1^{2(J-1)-d} & \cdots & c_J^{2(J-1)-d} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

It follows from Equation (10) that if the assumptions of Theorem 2 hold and if $J \geq (d-2)/8$, then

$$\sum_{j=0}^{J} \omega_j(\mathbf{c}) \mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(c_j \mathbf{H}_n)] - \boldsymbol{\theta} = o(n^{-1/2}).$$

As a consequence, we have the following result about the (generalized jackknife) estimator

$$\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c}) = \sum_{j=0}^{J} \omega_j(\mathbf{c}) \hat{\boldsymbol{\theta}}_n(c_j \mathbf{H}_n).$$

Theorem 3. Suppose Assumptions 1 and 2 are satisfied and suppose Equations (2) and (5) hold. If $(d-2)/8 \leq J < 1 + d/2$, then

$$\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c}) - \boldsymbol{\theta} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{z}_i) + o_p(n^{-1/2})$$

if either (i) $d = 1$ and $n|\mathbf{H}_n|^3 \to \infty$ or (ii) $d \geq 2$.

Theorem 3 gives a simple recipe for constructing an estimator of $\boldsymbol{\theta}$ that is semiparametrically efficient under relatively mild restrictions on the rate at which the bandwidth $\mathbf{H}_n$ vanishes.

Remark 3. (i) An alternative, and perhaps more conventional, method of bias correction would employ (nonparametric) estimators of $\mathcal{B}_0$ and $\{\mathcal{B}_j(\cdot)\}$ and subtract an estimator of $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta}$ from $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$. In our view, generalized jackknifing is attractive from a practical point of view precisely because there is no need to explicitly (characterize and) estimate complicated functionals such as $\mathcal{B}_0$ and $\{\mathcal{B}_j(\cdot)\}$. (ii) Our results demonstrate by example that a more nuanced understanding of the bias properties of $\hat{\boldsymbol{\theta}}_n$ can be achieved by working with a "quadratic" (as opposed to "linear") approximation to it. It is conceptually straightforward to go further and work with a "cubic" approximation (say) to $\hat{\boldsymbol{\theta}}_n$. Doing so would enable a further relaxation of the bandwidth condition at the expense of a more complicated "bias" expression, but would not alter the fact that generalized jackknifing could be used to eliminate also the bias terms that become nonnegligible under the relaxed bandwidth conditions. The simulation evidence presented in Section 4 suggests that eliminating the biases characterized in Equation (10) suffices for the purposes of rendering the bias of the estimator negligible relative to its standard deviation in many cases, so for brevity we omit results based on a "cubic" approximation to $\hat{\boldsymbol{\theta}}_n$.

3.3 Tuning Parameters Choices

We briefly discuss an implementation approach for the point estimators $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n; \mathbf{c})$, focusing in particular on choosing $\mathbf{H}_n$ and $\mathbf{c}$.

First, we discuss the choice of bandwidth $\mathbf{H}_n$. With minor additional effort, the derivations upon which our results are based may be used to obtain an asymptotic expansion of the mean squared error (MSE) of $\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$, the "quadratic" approximation to $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$. [In turn, this approximation can be used to justify a second-order stochastic expansion of the estimator $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$.] It follows from Lemmas A-2 and A-3 in the appendix that the variance and bias of $\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$ satisfy, respectively, $\mathbb{V}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] \approx n^{-1}\boldsymbol{\Sigma}$ and $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta} \approx$

$n^{-1}|\mathbf{H}_n|^{-1}\mathcal{B}_0 + \mathcal{S}(\mathbf{H}_n)$, where $\mathcal{S}(\mathbf{H}_n) = O(\lambda_{\max}(\mathbf{H}_n^P))$ is the "smoothing" bias of $\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$ (see Lemma A-3(a) for the exact formula of $\mathcal{S}(\cdot)$). In these approximations only leading terms have been retained on the right hand side, and the corresponding remainder terms are of smaller order than the square of the leading term(s) in the bias expansion. As a consequence, choosing the bandwidth $\mathbf{H}_n$ in an attempt to make (approximate) MSE small amounts to selecting a value of $\mathbf{H}_n$ for which the outer product of the leading terms of $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta}$ is small: $\min_{\mathbf{H}_n} \text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)]$, where

$$\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] = \left(\frac{\mathcal{B}_0}{n|\mathbf{H}_n|} + \mathcal{S}(\mathbf{H}_n)\right)\left(\frac{\mathcal{B}_0}{n|\mathbf{H}_n|} + \mathcal{S}(\mathbf{H}_n)\right)'. \tag{11}$$

Unfortunately, this problem does not have a (closed-form) solution in general, but can usually be solved numerically.

If the same bandwidth h_n is used for each coordinate, then $\mathbf{H}_n = h_n\mathbf{I}_d$ and the approximate bias expression becomes $n^{-1}h_n^{-d}\mathcal{B}_0 + h_n^P\mathcal{S}(\mathbf{I}_d)$. Minimizing the asymptotic order of this expression requires $h_n \propto n^{-1/(P+d)}$, a rate of decay that is permitted by our main results. (Bandwidth sequences of this type violate the conventional condition (3) unless P is large enough.) For example, when the object of main interest is a linear combination of the form $\mathbf{a}'\boldsymbol{\theta}$ (for some $\mathbf{a} \in \mathbb{R}^d$), and $\mathbf{a}'\mathcal{B}_0 \neq 0$ and $\mathbf{a}'\mathcal{S}(\mathbf{I}_d) \neq 0$, then $\text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n\mathbf{I}_d)]$ is minimized by setting

$$h_n^* = \begin{cases} \left(\dfrac{|\mathbf{a}'\mathcal{B}_0|}{|\mathbf{a}'\mathcal{S}(\mathbf{I}_d)|}\dfrac{1}{n}\right)^{\frac{1}{P+d}} & \text{if } \operatorname{sgn}(\mathbf{a}'\mathcal{B}_0) \neq \operatorname{sgn}(\mathbf{a}'\mathcal{S}(\mathbf{I}_d)) \\ \left(\dfrac{d}{P}\dfrac{|\mathbf{a}'\mathcal{B}_0|}{|\mathbf{a}'\mathcal{S}(\mathbf{I}_d)|}\dfrac{1}{n}\right)^{\frac{1}{P+d}} & \text{if } \operatorname{sgn}(\mathbf{a}'\mathcal{B}_0) = \operatorname{sgn}(\mathbf{a}'\mathcal{S}(\mathbf{I}_d)) \end{cases}.$$

Implementation of the "optimal" bandwidth choice(s) based on minimizing $\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)]$ (or some variant thereof) requires knowledge or estimation of the constants underlying $\mathcal{B}_0$ and $\mathcal{S}(\mathbf{I}_n)$. A natural approach is to estimate these constants nonparametrically, using some preliminary choices of tuning parameters to construct the corresponding nonparametric estimators. This approach is standard and readily applicable, but requires constructing several (preliminary) nonparametric estimators.

A simpler alternative is to construct a Silverman-style rule-of-thumb (ROT) bandwidth estimator of $\mathbf{H}_n$. We derive three ROT bandwidth choices under the following assumptions: (i) $K(\mathbf{u}) = \prod_{j=1}^d k(u_j)$ and P even, (ii) $f(\mathbf{x}) = \prod_{j=1}^d \phi(x_j/\sigma_j)/\sigma_j$ with $\phi(x)$ the standard Gaussian density, (iii) $g(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$, and (iv) $w(\mathbf{x}) = f(\mathbf{x})$. The supplemental appendix includes all the derivations, and a few additional technical assumptions not listed here. Using these assumptions, we find simple expressions for $\mathcal{B}_0$ and $\mathcal{S}(\mathbf{I}_d)$, which depend only on the unknown but easy-to-estimate constants $(\sigma_1, \sigma_2, \ldots, \sigma_d)'$ and $\boldsymbol{\beta}$. We then employ these expressions to describe ROT bandwidth choices based on the following three problems: (i) $\min_{h_n}$ AMSE $[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n\mathbf{I}_d)]$, (ii) $\min_{h_n} \operatorname{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(h_n\mathbf{I}_d)])$, and (iii) $\min_{\mathbf{H}_n} \operatorname{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)])$. [We did not characterize the case $\min_{\mathbf{H}_n} \text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)]$ because some of the associated constants are zero.] For example, the ROT bandwidth choice based on $\text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n\mathbf{I}_d)]$ with $\mathbf{a} = (1, 0, 0, \ldots, 0)' \in \mathbb{R}^d$ is

$$h_{\text{ROT-1d},n}^* = \begin{cases} \left(\sigma_1^P \prod_{l=1}^d \sigma_l \dfrac{|C_{\mathcal{B}}|}{|C_{\mathcal{SH}}|}\dfrac{1}{n}\right)^{\frac{1}{P+d}} & \text{if } \operatorname{sgn}(C_{\mathcal{B}}) \neq \operatorname{sgn}(C_{\mathcal{SH}}) \\ \left(\sigma_1^P \prod_{l=1}^d \sigma_l \dfrac{d}{P}\dfrac{|C_{\mathcal{B}}|}{|C_{\mathcal{SH}}|}\dfrac{1}{n}\right)^{\frac{1}{P+d}} & \text{if } \operatorname{sgn}(C_{\mathcal{B}}) = \operatorname{sgn}(C_{\mathcal{SH}}) \end{cases},$$

where $C_{\mathcal{SH}} = (-1)^{3P/2}2^{1-d-P}\pi^{-d/2}\int_{\mathbb{R}} u^P k(u)\mathrm{d}u/\Gamma(P/2)$ and $C_{\mathcal{B}} = -k(0)^d + \frac{1}{2}(\int_{\mathbb{R}} k(u)^2\mathrm{d}u)^d$. If, in addition, $\sigma = \sigma_1 = \cdots = \sigma_d$, then we obtain $h_{\text{ROT-1d},n}^* \propto \sigma n^{-1/(P+d)}$. The supplemental appendix provides details on the ROT bandwidth choices mentioned before. We explore the performance of all three ROT choices in our simulations in Section 4.

Next, we discuss the choice of $\mathbf{c}$, which requires selecting J and the constants $c_1, c_2, \ldots, c_J$. Constructing "optimal" choices for the tuning parameters of a generalized jackknifing procedure is a hard problem, which has only been solved in special simple cases (e.g., Schucany 1988). Although it is beyond the scope of this article to derive "optimal" choices, we may still offer some heuristic recommendations based on our derivations and our simulation evidence. First, we recommend to choose $J = \lceil (d-2)/8 \rceil$, which amounts to remove only the first few leading bias terms characterized in Lemma 1. This recommendation is based on the observation that increasing J is likely to increase the variability of the resulting jackknife estimator $\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$, a fact confirmed in our simulation study. Second, having chosen J, a simple implementation approach to choose the constants $c_1, c_2, \ldots, c_J$ is to construct an evenly spaced grid starting from the value selected for $\mathbf{H}_n$. Because our results offer robustness properties for "small" bandwidths, we recommend to select $c_J < c_{J-1} < \cdots < c_2 < c_1 < c_0 = 1$. In our simulations, for instance, 5% reductions in bandwidth (i.e., $c_0 = 1$, $c_1 = 0.95$, $c_2 = 0.90$, etc.) led to generalized jackknife estimators that performed well in all the designs considered.

3.4 Standard Errors

The emphasis so far has been on demonstrating approximate normality of $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ even when the classical conditions imposed in the literature are not satisfied. For inference purposes, it is important to also have a consistent standard-error estimator. The purpose of the following result is to give conditions under which

$$\hat{\boldsymbol{\Sigma}}_n = \hat{\boldsymbol{\Sigma}}_n(\mathbf{H}_n) = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\psi}}_n(\mathbf{z}; \mathbf{H}_n)\hat{\boldsymbol{\psi}}_n(\mathbf{z}; \mathbf{H}_n)' \to_p \boldsymbol{\Sigma}, \tag{12}$$

where

$$\begin{aligned} \hat{\boldsymbol{\psi}}_n(\mathbf{z}; \mathbf{H}_n) &= w(\mathbf{x})\frac{\partial}{\partial \mathbf{x}}\hat{g}_n(\mathbf{x}; \mathbf{H}_n) - \hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) \\ &\quad + [y - \hat{g}_n(\mathbf{x}; \mathbf{H}_n)]\,\hat{\mathbf{s}}_n(\mathbf{x}; \mathbf{H}_n), \\ \hat{g}_n(\mathbf{x}; \mathbf{H}_n) &= \frac{\hat{e}_n(\mathbf{x}; \mathbf{H}_n)}{\hat{f}_n(\mathbf{x}; \mathbf{H}_n)}, \qquad \hat{e}_n(\mathbf{x}; \mathbf{H}_n) = \frac{1}{n}\sum_{j=1}^n K_{\mathbf{H}_n}(\mathbf{x} - \mathbf{x}_j)y_j. \end{aligned}$$

Theorem 4. Suppose Assumptions 1 and 2 are satisfied and suppose Equations (2) and (5) hold. Then Equation (12) is true if either (i) $S \geq 2$ and $n|\mathbf{H}_n|^2\lambda_{\min}(\mathbf{H}_n^2)/(\log n)^2 \to \infty$, (ii) $d = 1$, $n|\mathbf{H}_n|^3 \to \infty$ and $S > 3$, or (iii) $S \geq 3 + 2/d$.

Part (i) of the theorem shows that even under the (seemingly) minimal moment requirement $S = 2$, consistency of $\hat{\boldsymbol{\Sigma}}_n$ holds under conditions on $\mathbf{H}_n$ that are slightly weaker than the conventional conditions (2) and (3). Perhaps more importantly, parts (ii) and (iii) give conditions (on S) for consistency of $\hat{\boldsymbol{\Sigma}}_n$ to hold under the assumptions of Theorem 3.

The proof of Theorem 4 uses a (seemingly) novel uniform consistency result for kernel estimators (and their derivatives), given in Appendix B. It does not seem possible to establish part (i) using existing uniform consistency results for kernel estimators, as we are unaware of any such results (for objects like $\hat{g}_n$) that require only $S = 2$. For instance, assuming $\mathbf{H}_n = h_n\mathbf{I}_d$, a proof of Equation (12) based on Newey (1994b, Lemma B.1) requires $S > 4 - 4/(d+2)$ when the lower bound on the bandwidth is of the form $nh_n^{2d+2}/(\log n)^2 \to \infty$. (When the lower bound on the bandwidth is of the form (5), Newey (1994b, Lemma B.1) can be applied if $d \geq 2$ and $S > 6 - 8/(d+2)$.)

4. NUMERICAL RESULTS

We report the main findings from a simulation study and an empirical illustration employing the conventional estimator $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and the generalized jackknife estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$. The supplemental appendix includes a complete set of results from our simulation study.

4.1 Simulation Setup

The Monte Carlo study is based on a Tobit model $y_i = \tilde{y}_i\mathbf{1}\{\tilde{y}_i \geq 0\}$ with $\tilde{y}_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$, so that $\boldsymbol{\theta} = \boldsymbol{\beta}\mathbb{E}[w(\mathbf{x})\Phi(\mathbf{x}'\boldsymbol{\beta})]$ with $\Phi(\cdot)$ the standard normal cdf. We set $d = 3$ and $\boldsymbol{\beta} = (1, 1, 1)'$, and assume that $\varepsilon_i \sim_{\text{iid}} \mathcal{N}(0, 1)$, $i = 1, 2, \ldots, n$, are independent of the covariates. We report results for three models, which depend on the distribution assumed on the vector of covariates. Specifically, for $i = 1, 2, \ldots, n$, we consider:

$$\text{Model 1} \;:\; \mathbf{x}_i \sim_{\text{iid}} \mathcal{N}(\mathbf{0}_3, \mathbf{V}_1), \quad \mathbf{V}_1 = \mathbf{I}_3,$$

$$\text{Model 2} \;:\; \mathbf{x}_i \sim_{\text{iid}} \mathcal{N}(\mathbf{0}_3, \mathbf{V}_2), \quad \mathbf{V}_2 = \begin{bmatrix} 1 & 1/4 & 1/4 \\ 1/4 & 2/3 & 1/4 \\ 1/4 & 1/4 & 1 \end{bmatrix},$$

$$\text{Model 3} \;:\; \mathbf{x}_i \sim_{\text{iid}} \begin{bmatrix} \left(\chi_4^2 - 4\right)/\sqrt{8} \\ \mathcal{N}(\mathbf{0}_2, \mathbf{V}_3) \end{bmatrix}, \quad \mathbf{V}_3 = \begin{bmatrix} 2/3 & 1/4 \\ 1/4 & 1 \end{bmatrix},$$

with $x_{1,i}$ independent of $(x_{2,i}, x_{3,i})'$. Consequently, Model 1 corresponds to independent, equal variance regressors; Model 2 corresponds to correlated, nonequal variance regressors; and Model 3 corresponds to asymmetric, partially correlated, nonequal variance regressors. We investigated many other configurations of data-generating processes, and in all cases we found qualitative similar results to those reported here (and in the supplemental appendix).

As for the choice of weight function, we use

$$w(\mathbf{x}; \gamma, \kappa) = \prod_{j=1}^{d} \exp\left[-\frac{x_j^{2\kappa}}{\tau_j^{2\kappa}\left(\tau_j^{2\kappa} - x_j^{2\kappa}\right)}\right]\mathbf{1}\{|x_j| < \tau_j\}.$$

The parameter κ governs the degree of approximation between $w(\cdot)$ and the rectangular function, the approximation becoming more precise as κ grows. (Being discontinuous, $w(\cdot)$ violates Assumption 1(c), so strictly speaking our theory does not cover the chosen weight function.) For specificity, we set $\kappa = 2$. When the covariates are jointly standard normal (Model 1), the trimming parameter $\tau_j = \tau(\gamma)$ is given by $\tau(\gamma) = \Phi^{-1}(1 - (1 - \sqrt[d]{1-\gamma})/2)$, where γ is the (symmetric) nominal amount of trimming (i.e., $\gamma = 0.15$ implies a nominal trimming of 15% of the observations). Thus, for Model 1, we set $\tau_j = \tau(\gamma)$ with $\gamma = 0.15$, while for the other models, we chose $(\tau_1, \tau_2, \tau_3)'$ so that approximately 15% of the observations were trimmed.

We construct the estimators using a Gaussian density-based multiplicative kernel with $P = 4$. (Note that since $d = 3$, choice of $P = 4$ would not be available under the conventional conditions (2) and (3).) The sample size is set to $n = 700$ for each replication, and the number of simulations is set to 5000.

4.2 Simulation Results

We investigate the performance of the estimators $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ for a variety of bandwidth choices, assuming both a common bandwidth ($\mathbf{H}_n = h_n\mathbf{I}_3$) and different bandwidths ($\mathbf{H}_n = \text{diag}(h_{1,n}, h_{3,n}, h_{3,n})$). For each case, we consider a grid of fixed (infeasible) bandwidths and the three ROT (data-driven, feasible) bandwidth choices introduced in Section 3.3.

The grid of bandwidth choices was constructed as follows. First, we computed the MSE "optimal" bandwidth choice for each model in each case, $\mathbf{H}_n = h_n\mathbf{I}_3$ and $\mathbf{H}_n = \text{diag}(h_{1,n}, h_{3,n}, h_{3,n})$, which we denote (abusing notation) $\mathbf{H}_n^* = h_n^*\mathbf{I}_3$ or $\mathbf{H}_n^* = \text{diag}(h_{1,n}^*, h_{2,n}^*, h_{3,n}^*)$, respectively. Second, we constructed a grid of bandwidths by setting $\mathbf{H}_n = \vartheta \cdot \mathbf{H}_n^*$ with $\vartheta \in \{0.50, 0.55, 0.60, \ldots, 1.45, 1.50\}$. Thus, $\vartheta = 1$ corresponds to using the infeasible, MSE optimal bandwidth choice for each of the six cases considered (three models for either common bandwidth or different bandwidths).

The ROT bandwidth choices were constructed as follows. First, we compute the scale of each covariate by $\hat{s}_j = \min\left\{S_j, \text{IQR}_j/1.349\right\}$ with S_j^2 and IQR_j denoting, respectively, the sample variance and interquartile range of the jth covariate ($j = 1, 2, 3$). We also estimated $\boldsymbol{\beta}$ by least-squares when needed. We report results for three feasible bandwidth choices: ROT bandwidth choice for (i) the first element of the AMSE (ROT-1d) with common bandwidth, (ii) the trace of the AMSE (ROT-tr) with common bandwidth, and (iii) the trace of the AMSE (ROT-tr) with different bandwidths. Abusing notation, we let $\hat{\mathbf{H}}_n$ denote any of these ROT bandwidth estimates.

The estimators $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\hat{\boldsymbol{\Sigma}}_n(\mathbf{H}_n)$ are computed for each point in the bandwidths grid and for the estimated ROT bandwidths. The generalized jackknife estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ was constructed as follows. First, for the bandwidths on the grid, $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ was computed by employing the adjacent bandwidth(s) to $\mathbf{H}_n$ on the grid, depending on the specific implementation (discussed next). [This approach implies that the actual constants $\mathbf{c} = (c_0, c_1, \ldots, c_J)'$ are slightly different along the grid.] Second, for the ROT estimated bandwidths, we constructed a five-point grid $\vartheta \cdot \hat{\mathbf{H}}_n$ with $\vartheta \in \{0.90, 0.95, 1, 1.05, 1.10\}$, and then implemented the estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ at $\vartheta = 1$ according to the specific implementation (discussed next).

As for the actual implementation of $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$, for a given $\mathbf{H}_n$, we consider five distinct approaches depending on the choice of $c_L \in \{0, 1, 2\}$ and $c_U \in \{0, 1, 2\}$. Specifically, c_L and c_U determine, respectively, how many grid points below and above the specific value $\mathbf{H}_n$ are used to construct $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$. (Hence, $J = c_L + c_U$.) In this section we only report results for $c_L = 1$ and $c_U = 0$, but in the supplemental appendix we include four other cases: $(c_L, c_U) = (2, 0)$, $(c_L, c_U) = (0, 1)$, $(c_L, c_U) = (1, 1)$, and $(c_L, c_U) = (0, 2)$.

Once the estimators $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ are constructed for each bandwidth value $\mathbf{H}_n$ (either on the grid or estimated using the ROT procedures), we computed MSE, squared-bias, variance, absolute-bias/square-root-variance, and coverage rates of 95% confidence intervals for each simulation design (Models 1–3, with either common or different bandwidths). In this section, for brevity and to facilitate the comparison between the two estimators, we only report two standardized measures: (i) MSE relative to MSE when employing the optimal common bandwidth, and (ii) absolute-bias divided by square-root of variance. Thus, we only include three short tables in the article, but the supplemental appendix includes all the results (30 long tables).

The results are presented in Tables 1–3 for Models 1–3, respectively. In all cases, we found that the generalized jackknife estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ leads to noticeable reductions in standardized bias, especially for "small" bandwidths (i.e., for smaller bandwidths than the MSE-optimal ones). This finding is consistent with our theory. In addition, we found that the MSE of $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ was also reduced in most cases relative to the MSE of $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$, suggesting that in our simulations employing generalized jackknifing does not increase the variability of the resulting estimator much (relative to the gains in bias-reduction). These findings highlight the potential sensitivity of the conventional estimator to perturbations of the bandwidth choice, which, in the case of the weighted average derivatives, leads to a nontrivial bias for "small" bandwidths, and therefore a need for bias correction.

Our simulations also suggest that the rule-of-thumb bandwidth selectors perform relatively well, providing a simple and easy-to-implement bandwidth choice. Although it is important to also consider consistent nonparametric bandwidth choices, our rule-of-thumbs seem to provide a natural and simple first bandwidth choice to employ.

We also explored the quality of the normal approximation to the distribution of the t-statistic (we do not report result here to conserve space). We found that the distribution of both $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ were close to Gaussian, although the classical estimator exhibited a nontrivial bias. In contrast, the generalized estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ was found to be approximately centered correctly, especially for "small" bandwidths.

Finally, we also explored the empirical coverage rates of the conventional and bias-corrected t-statistics. We found that neither the conventional nor the jackknife estimator succeeded in achieving empirical coverage rates near the nominal rate. This finding, together with the results reported above, suggests that the lack of good empirical coverage of the associated confidence intervals for the generalized jackknife procedure is due to the poor performance of the classical variance estimator commonly employed in the literature. Indeed, in the case of the conventional procedure, we found that both the bias properties and the performance of this variance estimator seem to be at fault for the disappointing empirical coverage rates found in the simulations.

Table 1. Classical and generalized jackknife estimators, Model 1

	$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$		$\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$	
	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$
(a) Common bandwidth, $J = 1$, $c_L = 1$, $c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot 0.591 \cdot \mathbf{I}_3$				
ϑ				
0.50	3.744	2.018	1.720	1.092
0.55	2.919	1.750	1.287	0.835
0.60	2.316	1.513	1.050	0.643
0.65	1.887	1.310	0.921	0.510
0.70	1.582	1.141	0.854	0.426
0.75	1.371	1.004	0.820	0.380
0.80	1.225	0.894	0.809	0.365
0.85	1.125	0.810	0.816	0.375
0.90	1.062	0.748	0.837	0.405
0.95	1.021	0.705	0.869	0.452
1.00	1.000	0.679	0.916	0.513
1.05	0.993	0.667	0.979	0.585
1.10	0.998	0.668	1.057	0.665
1.15	1.014	0.681	1.153	0.754
1.20	1.040	0.702	1.266	0.848
1.25	1.072	0.732	1.400	0.947
1.30	1.115	0.770	1.552	1.051
1.35	1.167	0.813	1.723	1.157
1.40	1.227	0.862	1.912	1.265
1.45	1.296	0.915	2.119	1.375
1.50	1.375	0.972	2.340	1.485
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-1d = 0.565	1.019	0.703	0.876	0.459
ROT-tr = 0.564	1.019	0.704	0.876	0.458
(b) Different bandwidths, $J = 1$, $c_L = 1$, $c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot \text{diag}(0.591, 0.591, 0.591)$				
ϑ				
0.50	3.744	2.018	1.720	1.092
0.55	2.919	1.750	1.287	0.835
0.60	2.316	1.513	1.050	0.643
0.65	1.887	1.310	0.921	0.510
0.70	1.582	1.141	0.854	0.426
0.75	1.371	1.004	0.820	0.380
0.80	1.225	0.894	0.809	0.365
0.85	1.125	0.810	0.816	0.375
0.90	1.062	0.748	0.837	0.405
0.95	1.021	0.705	0.869	0.452
1.00	1.000	0.679	0.916	0.513
1.05	0.993	0.667	0.979	0.585
1.10	0.998	0.668	1.057	0.665
1.15	1.014	0.681	1.153	0.754
1.20	1.040	0.702	1.266	0.848
1.25	1.072	0.732	1.400	0.947
1.30	1.115	0.770	1.552	1.051
1.35	1.167	0.813	1.723	1.157
1.40	1.227	0.862	1.912	1.265
1.45	1.296	0.915	2.119	1.375
1.50	1.375	0.972	2.340	1.485
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-tr = (0.565, 0.565, 0.565)	1.019	0.703	0.876	0.459

NOTE: (i) columns $\frac{\text{MSE}}{\text{MSE}^*}$ report MSE for each estimator divided by MSE of conventional estimator employing optimal common bandwidth; (ii) columns $\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$ report absolute bias divided by square root of variance for each estimator; (iii) upper part of panel (a) reports infeasible optimal bandwidth solving $\min_{h_n} \text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n \mathbf{I}_d)]$ with $\mathbf{a} = (1, 0, 0)'$, while upper part of panel (b) reports infeasible optimal bandwidths solving $\min_{\mathbf{H}_n} \text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)])$; (iv) lower parts of panels (a) and (b) report estimators employing ROT bandwidth choices, with average estimated bandwidths for each case (ROT-1d and ROT-tr corresponds to ROT estimates based on $\text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(\cdot)]$ and $\text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\cdot)])$, respectively).

Table 2. Classical and generalized jackknife estimators, Model 2

	$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$		$\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$	
	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$
(a) Common bandwidth, $J = 1, c_L = 1, c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot 0.58 \cdot \mathbf{I}_3$				
ϑ				
0.50	2.504	1.323	1.252	0.566
0.55	1.974	1.107	1.069	0.392
0.60	1.619	0.925	0.985	0.278
0.65	1.386	0.778	0.946	0.211
0.70	1.233	0.660	0.931	0.177
0.75	1.136	0.569	0.931	0.168
0.80	1.075	0.500	0.935	0.175
0.85	1.037	0.450	0.946	0.195
0.90	1.015	0.415	0.963	0.226
0.95	1.004	0.393	0.985	0.266
1.00	1.000	0.382	1.013	0.313
1.05	1.004	0.380	1.050	0.366
1.10	1.013	0.387	1.095	0.424
1.15	1.026	0.400	1.149	0.486
1.20	1.043	0.419	1.213	0.552
1.25	1.065	0.443	1.293	0.619
1.30	1.091	0.472	1.377	0.692
1.35	1.123	0.505	1.476	0.766
1.40	1.157	0.540	1.584	0.840
1.45	1.198	0.579	1.705	0.915
1.50	1.244	0.620	1.834	0.991
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-1d = 0.549	1.006	0.396	0.985	0.263
ROT-tr = 0.516	1.019	0.422	0.963	0.222
(b) Different bandwidths, $J = 1, c_L = 1, c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot \text{diag}(0.529, 0.551, 0.529)$				
ϑ				
0.50	3.060	1.501	1.496	0.729
0.55	2.397	1.280	1.200	0.521
0.60	1.929	1.085	1.054	0.370
0.65	1.608	0.919	0.981	0.269
0.70	1.392	0.781	0.946	0.206
0.75	1.246	0.670	0.931	0.172
0.80	1.149	0.581	0.927	0.159
0.85	1.084	0.512	0.931	0.161
0.90	1.043	0.459	0.940	0.175
0.95	1.017	0.420	0.950	0.198
1.00	1.002	0.393	0.968	0.230
1.05	0.994	0.377	0.987	0.268
1.10	0.994	0.369	1.015	0.311
1.15	0.996	0.368	1.047	0.360
1.20	1.004	0.374	1.086	0.412
1.25	1.015	0.386	1.134	0.469
1.30	1.030	0.402	1.192	0.528
1.35	1.052	0.422	1.386	0.547
1.40	1.071	0.448	1.463	0.619
1.45	1.097	0.476	1.420	0.716
1.50	1.127	0.508	1.509	0.788
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-tr = (0.563, 0.484, 0.564)	1.043	0.442	1.017	0.310

NOTE: (i) columns $\frac{\text{MSE}}{\text{MSE*}}$ report MSE for each estimator divided by MSE of conventional estimator employing optimal common bandwidth; (ii) columns $\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$ report absolute bias divided by square root of variance for each estimator; (iii) upper part of panel (a) reports infeasible optimal bandwidth solving $\min_{h_n} \text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n \mathbf{I}_d)]$ with $\mathbf{a} = (1, 0, 0)'$, while upper part of panel (b) reports infeasible optimal bandwidths solving $\min_{\mathbf{H}_n} \text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)])$; (iv) lower parts of panels (a) and (b) report estimators employing ROT bandwidth choices, with average estimated bandwidths for each case (ROT-1d and ROT-tr corresponds to ROT estimates based on $\text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(\cdot)]$ and $\text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\cdot)])$, respectively).

Table 3. Classical and generalized jackknife estimators, Model 3

	$\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$		$\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$	
	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$	$\frac{\text{MSE}}{\text{MSE}^*}$	$\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$
(a) Common bandwidth, $J = 1, c_L = 1, c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot 0.466 \cdot \mathbf{I}_3$				
ϑ				
0.50	3.318	1.876	1.592	1.037
0.55	2.599	1.664	1.211	0.829
0.60	2.083	1.474	1.003	0.681
0.65	1.716	1.310	0.893	0.588
0.70	1.464	1.173	0.837	0.539
0.75	1.287	1.063	0.813	0.523
0.80	1.170	0.976	0.817	0.531
0.85	1.090	0.912	0.834	0.557
0.90	1.042	0.866	0.865	0.595
0.95	1.010	0.835	0.907	0.643
1.00	1.000	0.819	0.962	0.699
1.05	1.000	0.814	1.031	0.762
1.10	1.010	0.818	1.111	0.832
1.15	1.028	0.832	1.204	0.907
1.20	1.059	0.854	1.315	0.989
1.25	1.093	0.882	1.446	1.076
1.30	1.142	0.917	1.595	1.168
1.35	1.194	0.958	1.768	1.265
1.40	1.260	1.004	1.965	1.366
1.45	1.332	1.055	2.183	1.471
1.50	1.415	1.110	2.429	1.579
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-1d = 0.517	1.010	0.819	1.114	0.831
ROT-tr = 0.506	1.007	0.816	1.083	0.805
(b) Different bandwidths, $J = 1, c_L = 1, c_U = 0$				
$\mathbf{H}_n = \vartheta \cdot \text{diag}(0.491, 0.466, 0.456)$				
ϑ				
0.50	3.228	1.876	1.543	1.033
0.55	2.536	1.661	1.187	0.828
0.60	2.042	1.470	0.997	0.688
0.65	1.692	1.307	0.896	0.604
0.70	1.453	1.173	0.851	0.563
0.75	1.291	1.068	0.837	0.555
0.80	1.180	0.987	0.844	0.571
0.85	1.111	0.928	0.872	0.604
0.90	1.066	0.888	0.913	0.651
0.95	1.045	0.864	0.965	0.707
1.00	1.038	0.853	1.035	0.772
1.05	1.045	0.855	1.118	0.844
1.10	1.062	0.866	1.218	0.924
1.15	1.093	0.887	1.339	1.010
1.20	1.131	0.916	1.481	1.103
1.25	1.180	0.953	1.644	1.202
1.30	1.239	0.996	1.834	1.306
1.35	1.308	1.045	2.048	1.415
1.40	1.391	1.099	2.291	1.527
1.45	1.481	1.158	2.561	1.643
1.50	1.585	1.221	2.851	1.760
$\mathbf{H}_n = \hat{\mathbf{H}}_n$				
ROT-tr = (0.506, 0.488, 0.555)	0.997	0.796	1.083	0.790

NOTE: (i) columns $\frac{\text{MSE}}{\text{MSE*}}$ report MSE for each estimator divided by MSE of conventional estimator employing optimal common bandwidth; (ii) columns $\frac{\text{BIAS}}{\sqrt{\text{VAR}}}$ report absolute bias divided by square root of variance for each estimator; (iii) upper part of panel (a) reports infeasible optimal bandwidth solving $\min_{h_n} \text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(h_n \mathbf{I}_d)]$ with $\mathbf{a} = (1, 0, 0)'$, while upper part of panel (b) reports infeasible optimal bandwidths solving $\min_{\mathbf{H}_n} \text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)])$; (iv) lower parts of panels (a) and (b) report estimators employing ROT bandwidth choices, with average estimated bandwidths for each case (ROT-1d and ROT-tr corresponds to ROT estimates based on $\text{AMSE}[\mathbf{a}'\hat{\boldsymbol{\theta}}_n^{**}(\cdot)]$ and $\text{tr}(\text{AMSE}[\hat{\boldsymbol{\theta}}_n^{**}(\cdot)])$, respectively).

Further investigation into alternative variance estimation procedures, although beyond the scope of this article, is underway.

4.3 Empirical Illustration

To complement the simulation evidence reported above, we undertake a small empirical illustration that shows how our methods perform using real data. We focus on estimating the average marginal return to ability, employing a subset of the dataset constructed by Lang and Manove (2011). [The dataset is available at *http://www.aeaweb.org/articles.php?doi=10.1257/aer.101.4.1467*.]

The data comes from the National Longitudinal Survey of Youth (NLSY79), which follows individuals born in 1957–1964. This (panel) dataset provides not only demographic, economic, and educational information, but also includes a well-known proxy for ability (beyond schooling and work experience) for the individuals in the sample. Specifically, this data includes the results from the Armed Forces Qualification Test (AFQT) for those individuals who took the test in 1980, which provides a close-to-continuous measure that may be understood as a proxy for their intrinsic “ability.” This data has been used repeatedly to either control for or estimate the effects of “ability” in empirical studies in economics and related fields. For more details on this data and a discussion on the related literature, see Lang and Manove (2011) and references therein.

In our empirical illustration, we focus on estimating the (weighted) average marginal effect of an increase in AFQT on earnings while controlling for two other observed characteristics. In particular, we let $y_i = \log(\text{WAGE}_i)$ where WAGE_i denotes the mean adjusted hourly wages in 1996–2000 for individual i, and $\mathbf{x}_i = (\text{AFQT}_i, \text{SCHSZ}_i, \text{TEACHW}_i)'$ where AFQT_i denotes the (adjusted) standardized AFQT score for individual i, SCHSZ_i denotes the school size that individual i attended to, and TEACHW_i denotes the average teacher salary in the school that individual i attended to. Our parameter of interest is

$$\theta_1 = \mathbb{E}\left[w(\mathbf{x}_i) \frac{\partial}{\partial \text{AFQT}_i} g(\text{AFQT}_i, \text{SCHSZ}_i, \text{TEACHW}_i) \right],$$

where $g(\mathbf{x}_i) = \mathbb{E}[y_i|\mathbf{x}_i]$. To conduct the estimation, we restrict our sample to the subset of 15–19-year-old white males with 12–16 years of schooling in 1979. The final sample size is $n = 802$ individuals. Figure 1 plots nonparametric smoothing spline estimates of the univariate conditional expectations for each of the three covariates included in our sample, computed using the command `gam()` in `R` (*http://www.r-project.org*).

Table 4. Average marginal effect of ability on earnings ($\mathbf{c} = (1, 0.95)$)

	Coef.		Std. Err.
	$\hat{\boldsymbol{\theta}}_n(\hat{\mathbf{H}}_n)$	$\tilde{\boldsymbol{\theta}}_n(\hat{\mathbf{H}}_n, \mathbf{c})$	$\hat{\boldsymbol{\Sigma}}_n(\hat{\mathbf{H}}_n)$
Common bandwidth: ROT-1d			
$\hat{\mathbf{H}}_n = 0.48 \cdot \mathbf{I}_3$	0.536	0.484	0.023
$\hat{\mathbf{H}}_n = 0.9 \cdot 0.48 \cdot \mathbf{I}_3$	0.560	0.432	0.024
Common bandwidth: ROT-tr			
$\hat{\mathbf{H}}_n = 0.483 \cdot \mathbf{I}_3$	0.535	0.487	0.023
$\hat{\mathbf{H}}_n = 0.9 \cdot 0.483 \cdot \mathbf{I}_3$	0.559	0.433	0.024
Different bandwidths: ROT-tr			
$\hat{\mathbf{H}}_n = \text{diag}(0.48, 0.48, 0.48)$	0.536	0.484	0.023
$\hat{\mathbf{H}}_n = 0.9 \cdot \text{diag}(0.48, 0.48, 0.48)$	0.560	0.432	0.024

Figure 1 exhibits a nonlinear relationship between wages and ability, suggesting that different levels of ability will have differential effects on earnings for the individuals in this sample. The average derivative θ_1 provides an overall (weighted, averaged) marginal-effect measure for these individuals, after controlling for the other covariates.

Table 4 presents the empirical estimates of both the classical estimator $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and the generalized jackknife estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$. We employ the same weighting function introduced in the simulation section. To implement these estimators, we centered and scaled the covariates SCHSZ_i and TEACHW_i (without loss of generality), and then selected a trimming parameter for each dimension of $\mathbf{x}_i$ such that at least 1% of the sample was trimmed along each dimension. Based on our simulations, we selected $\mathbf{c} = (1, 0.95)$ to implement the generalized jackknife estimator. As for the bandwidth choice, we report results for all three ROT alternatives discussed previously.

Our empirical results suggest that in this illustration bias may be important. Indeed, while the point estimator $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ gives an average marginal return to ability of about 0.535, the generalized jackknife estimator $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ gives a point estimate of about 0.485. Interestingly, the 95% confidence interval based on $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$ does not include the point estimate $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$. (As

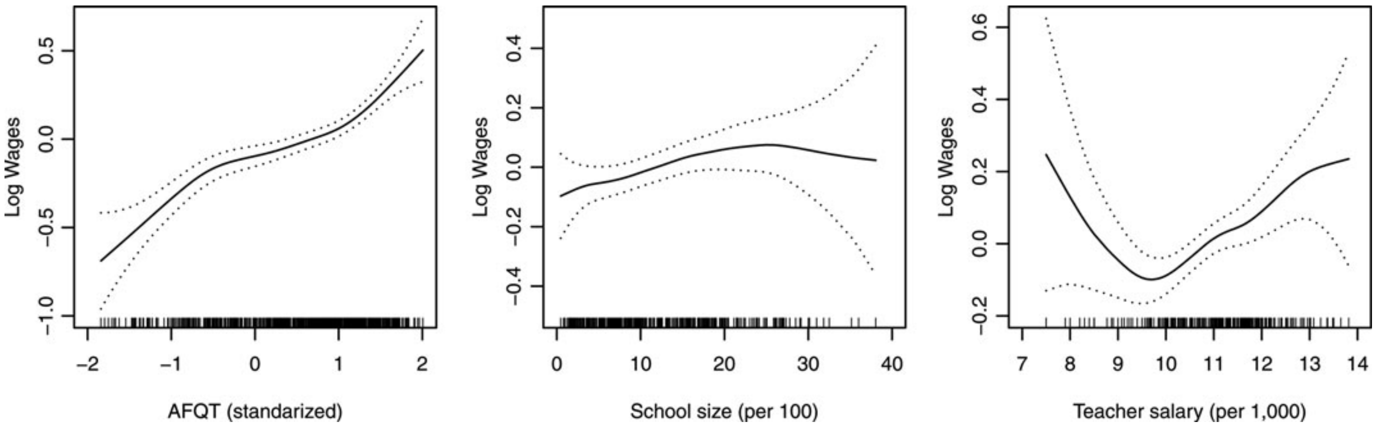

Figure 1. Smoothing splines estimates for univariate conditional expectations.

shown in the table, a 10% undersmoothing leads to even larger differences between the conventional and the generalized jackknife estimators.) As a consequence, this empirical illustration provides a simple empirical example where our procedure leads to a quantitatively different estimate than the conventional one.

5. CONCLUSION

This article has revisited the large-sample properties of a kernel-based weighted average derivative estimator. In important respects, this estimator can be viewed as a representative member of the much larger class of (kernel-based) semiparametric *m*-estimators. In particular, the "nonlinearity bias" highlighted by our development of asymptotics with smaller-than-usual bandwidths (i.e., larger-than-usual undersmoothing) is a generic feature of nonlinear functionals of nonparametric estimators and is likely to be quantitatively important in samples of moderate size also for estimators other than the one studied in this article.

To remove this "nonlinearity bias," we have employed the method of generalized jackknifing. Being "semiautomatic" in the sense that it requires knowledge only of the magnitudes of the terms in an asymptotic expansion of the "nonlinearity bias," that same method should be easily applicable whenever the nonparametric ingredient is a kernel estimator, as the variance properties of kernel estimators are very well understood. Partly because certain popular nonparametric estimators (notably series estimators) have variance properties that seem harder to analyze than those of kernel estimators, it would be useful to know if the validity of certain "fully automatic" bias correction methods and/or distributional approximations can be established under assumptions similar to those entertained in this article.

APPENDIX A: PROOFS

This appendix gives the proofs of Theorems 1–3. We first state four lemmas, the proofs of which are available in the supplemental appendix. We then employ these lemmas, together with the results for kernel-based estimators outlined in Appendix B, to prove the main theorems.

A.1 Useful Lemmas

The first lemma gives sufficient conditions for Equation (6) in terms of the magnitudes of $\Delta_{0,n}(\mathbf{H}_n) = \sup_{\mathbf{x}\in\mathcal{W}} |\hat{f}_n(\mathbf{x};\mathbf{H}_n) - f(\mathbf{x})|$ and $\Delta_{1,n}(\mathbf{H}_n) = \max\{\Delta_{0,n}(\mathbf{H}_n), \sup_{\mathbf{x}\in\mathcal{W}} \|\partial \hat{f}_n(\mathbf{x};\mathbf{H}_n)/\partial\mathbf{x} - \partial f(\mathbf{x})/\partial\mathbf{x}\|\}$.

Lemma A-1. Suppose Assumption 1 is satisfied and suppose $\Delta_{0,n} = o_p(1)$. Then Equation (6) is true if either (i) $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$ and $\Delta_{0,n}(\mathbf{H}_n)^2\Delta_{1,n}(\mathbf{H}_n) = o_p(n^{-1/2})$ or (ii) $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{*}(\mathbf{H}_n)$ and $\Delta_{0,n}(\mathbf{H}_n)\Delta_{1,n}(\mathbf{H}_n) = o_p(n^{-1/2})$.

The next result gives sufficient conditions for Equation (7).

Lemma A-2. Suppose Assumptions 1 and 2 are satisfied and suppose $\lambda_{\max}(\mathbf{H}_n) \to 0$ and $n|\mathbf{H}_n|\lambda_{\min}(\mathbf{H}_n^2) \to \infty$. Then Equation (7) is true for $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{*}(\mathbf{H}_n)$ and $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$.

Lemma 1 is a corollary of the following result, which can be used to evaluate $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^A] - \boldsymbol{\theta}$. To state the result succinctly, let $\dot{\mathbf{f}}(\mathbf{x}) = \partial f(\mathbf{x})/\partial\mathbf{x}$, let $\operatorname{diag}(\mathbf{h}_n) = \mathbf{H}_n$ (i.e., let $\mathbf{h}_n \in \mathbb{R}^d_{++}$ collect the diagonal elements of $\mathbf{H}_n$), and for any multi-index $\mathbf{l} = (l_1, l_2, \ldots, l_d)' \in \mathbb{Z}^d_+$ and any sufficiently smooth function $f(\cdot)$ (not necessarily equal to the density of $\mathbf{x}$), let

$$\mathbf{l}! = l_1!l_2!\ldots l_d!, \qquad \partial^{\mathbf{l}} f(\mathbf{x}) = \frac{\partial^{l_1+l_2+\cdots+l_d}}{\partial x_1^{l_1}\partial x_2^{l_2}\ldots\partial x_d^{l_d}} f(x_1, x_2, \ldots, x_d).$$

Also, for any $k \in \mathbb{Z}_+$, define $\mathbb{Z}^d_+(k) = \{(l_1, \ldots, l_d)' \in \mathbb{Z}^d_+ : l_1 + \cdots + l_d = k\}$.

Lemma A-3. Suppose Assumptions 1 and 2 are satisfied and suppose $\lambda_{\max}(\mathbf{H}_n) \to 0$. *(a)* Bias of $\hat{\boldsymbol{\theta}}_n^{*}(\mathbf{H}_n)$:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{*}(\mathbf{H}_n)] - \boldsymbol{\theta} = n^{-1}|\mathbf{H}_n|^{-1}\mathcal{B}_0^* + \mathcal{S}(\mathbf{H}_n) + o\left(\lambda_{\max}(\mathbf{H}_n^P)\right),$$

where

$$\mathcal{S}(\mathbf{H}_n) = (-1)^{P+1} \sum_{\mathbf{l}\in\mathbb{Z}^d_+(P)} \frac{\mathbf{h}_n^{\mathbf{l}}}{\mathbf{l}!} \left[\int_{\mathbb{R}^d} w(\mathbf{r})g(\mathbf{r})\left(\partial^{\mathbf{l}}\dot{\mathbf{f}}(\mathbf{r}) + \ell(\mathbf{r})\partial^{\mathbf{l}} f(\mathbf{r})\right)\mathrm{d}\mathbf{r}\right] \times \left[\int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u})\mathrm{d}\mathbf{u}\right].$$

(b) Nonlinearity bias:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n) - \hat{\theta}_n^{*}(\mathbf{H}_n)] = n^{-1}|\mathbf{H}_n|^{-1}\left[\mathcal{B}_0^{**} + \sum_{j=1}^{\lfloor (P-1)/2\rfloor} \mathcal{B}_j(\mathbf{H}_n)\right] + O\left(n^{-2}|\mathbf{H}_n|^{-2} + \lambda_{\max}(\mathbf{H}_n^{2P})\right),$$

where

$$\mathcal{B}_j(\mathbf{H}_n) = \sum_{\mathbf{l}\in\mathbb{Z}^d_+(2j)} \frac{\mathbf{h}_n^{\mathbf{l}}}{\mathbf{l}!}\mathbf{B}_z(\mathbf{l})B_K(\mathbf{l}) + \sum_{\mathbf{l}\in\mathbb{Z}^d_+(2j+1)} \frac{\mathbf{h}_n^{\mathbf{l}}}{\mathbf{l}!}\dot{B}_z(\mathbf{l})\mathbf{H}_n^{-1}\dot{\mathbf{B}}_K(\mathbf{l}),$$

with

$$B_K(\mathbf{l}) = \int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u})^2\mathrm{d}\mathbf{u}, \qquad \mathbf{B}_z(\mathbf{l}) = \int_{\mathbb{R}^d} g(\mathbf{r})\frac{w(\mathbf{r})}{f(\mathbf{r})}\ell(\mathbf{r})\partial^{\mathbf{l}} f(\mathbf{r})\mathrm{d}\mathbf{r},$$
$$\dot{\mathbf{B}}_K(\mathbf{l}) = \int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u})\dot{\mathbf{K}}(\mathbf{u})\mathrm{d}\mathbf{u}\,, \qquad \dot{B}_z(\mathbf{l}) = -\int_{\mathbb{R}^d} g(\mathbf{r})\frac{w(\mathbf{r})}{f(\mathbf{r})}\partial^{\mathbf{l}} f(\mathbf{r})\mathrm{d}\mathbf{r}.$$

The last lemma collects basic results about kernels-based integrals. Let $\dot{\mathbf{K}}_{\mathbf{H}}(\mathbf{x}) = \partial K_{\mathbf{H}}(\mathbf{x})/\partial\mathbf{x}$.

Lemma A-4. If Assumptions 1 and 2 are satisfied and if $\lambda_{\max}(\mathbf{H}_n) \to 0$, then *(a)* Uniformly in $\mathbf{x} \in \mathcal{W}$,

$$\begin{aligned} b(\mathbf{x};\mathbf{H}_n) &= \int_{\mathbb{R}^d} K_{\mathbf{H}_n}(\mathbf{x}-\mathbf{r})f(\mathbf{r})\mathrm{d}\mathbf{r} - f(\mathbf{x}) \\ &= (-1)^P \sum_{\mathbf{l}\in\mathbb{Z}^d_+(P)} \frac{\mathbf{h}_n^{\mathbf{l}}}{\mathbf{l}!}\partial^{\mathbf{l}} f(\mathbf{x})\left(\int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u})\,\mathrm{d}\mathbf{u}\right) + o\left(\lambda_{\max}(\mathbf{H}_n^P)\right) \\ &= O\left(\lambda_{\max}(\mathbf{H}_n^P)\right), \end{aligned}$$

$$\begin{aligned} \dot{\mathbf{b}}(\mathbf{x};\mathbf{H}_n) &= \int_{\mathbb{R}^d} \dot{\mathbf{K}}_{\mathbf{H}_n}(\mathbf{x}-\mathbf{r})f(\mathbf{r})\mathrm{d}\mathbf{r} - \partial f(\mathbf{x})/\partial\mathbf{x} \\ &= (-1)^{P+1} \sum_{\mathbf{l}\in\mathbb{Z}^d_+(P)} \frac{\mathbf{h}_n^{\mathbf{l}}}{\mathbf{l}!}\partial^{\mathbf{l}}\dot{\mathbf{f}}(\mathbf{x})\left(\int_{\mathbb{R}^d} \mathbf{u}^{\mathbf{l}} K(\mathbf{u})\mathrm{d}\mathbf{u}\right) + o\left(\lambda_{\max}(\mathbf{H}_n^P)\right) \\ &= O\left(\lambda_{\max}(\mathbf{H}_n^P)\right). \end{aligned}$$

(b) For any function F with $\mathbb{E}[F(\mathbf{z})^2] < \infty$,

(i) $\mathbb{E}[F(\mathbf{z}_1)^2 K_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_2)^2] = O(|\mathbf{H}_n|^{-1})$, (ii) $\mathbb{E}[F(\mathbf{z}_1)^2 \|\dot{\mathbf{K}}_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_2)\|^2] = O(|\mathbf{H}_n|^{-1}\lambda_{\max}(\mathbf{H}_n^{-2}))$, (iii) $\mathbb{E}[F(\mathbf{z}_1)^2 K_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_2)^2 K_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_3)^2] = O(|\mathbf{H}_n|^{-2})$, and (iv) $\mathbb{E}[F(\mathbf{z}_1)^2 K_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_2)^2 \|\dot{\mathbf{K}}_{\mathbf{H}_n}(\mathbf{x}_1 - \mathbf{x}_3)\|^2] = O(|\mathbf{H}_n|^{-2}\lambda_{\max}(\mathbf{H}_n^{-2}))$.

A.2 Proof of Theorems 1–3

Under the assumptions of the theorems, Equations (6) and (7) hold for $\hat{\boldsymbol{\theta}}_n^A = \hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)$. Validity of Equation (7) follows from Lemma A-2, while Equation (6) follows from Lemma A-1 because it can be shown that

$$\sup_{\mathbf{x}\in\mathcal{W}} \left|\hat{f}_n(\mathbf{x};\mathbf{H}_n) - f(\mathbf{x})\right| = O_p\left(\lambda_{\max}(\mathbf{H}_n^P) + \sqrt{\frac{\log n}{n|\mathbf{H}_n|}}\right) \quad \text{(A.1)}$$

and

$$\sup_{\mathbf{x}\in\mathcal{W}} \left\|\frac{\partial}{\partial\mathbf{x}}\hat{f}_n(\mathbf{x};\mathbf{H}_n) - \frac{\partial}{\partial\mathbf{x}} f(\mathbf{x})\right\| = O_p\left(\lambda_{\max}(\mathbf{H}_n^P) + \sqrt{\frac{\log n}{n|\mathbf{H}_n|\lambda_{\min}(\mathbf{H}_n^2)}}\right). \quad \text{(A.2)}$$

Specifically, Equation (A-1) holds because $\sup_{\mathbf{x}\in\mathcal{W}} |\mathbb{E}[\hat{f}_n(\mathbf{x};\mathbf{H}_n)] - f(\mathbf{x})| = O(\lambda_{\max}(\mathbf{H}_n^P))$ by Lemma A-4 (a) and because $\sup_{\mathbf{x}\in\mathcal{W}} |\hat{f}_n(\mathbf{x};\mathbf{H}_n) - \mathbb{E}[\hat{f}_n(\mathbf{x};\mathbf{H}_n)]| = O_p(\sqrt{\log n}/\sqrt{n\,|\mathbf{H}_n|})$ by Lemma B-1 with $(Y,\mathbf{X}) = (1,\mathbf{x})$, $\kappa = K$, and $\mathcal{X}_n = \mathcal{W}$. Similarly, Equation (A-2) can be shown by applying Lemma A-4 (a) and Lemma B-1 (with $\kappa(\mathbf{u})$ equal to an element of $\mathbf{H}_n \partial K(\mathbf{u})/\partial\mathbf{u}$).

Theorem 1 is a special case of Theorem 2. To complete the proof of Theorem 2, use Lemma A-3 to verify Equation (8). Similarly, the proof of Theorem 3 can be completed by using Lemma A-3 to verify Equation (10).

A.3 Proof of Theorem 4

It suffices to show that $\sum_{i=1}^n \|\hat{\boldsymbol{\psi}}_n(\mathbf{z}_i;\mathbf{H}_n) - \boldsymbol{\psi}(\mathbf{z}_i)\|^2 = o_p(n)$. To do so, it suffices to show that: (i) $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n) - \boldsymbol{\theta} = o_p(1)$, (ii) $\sup_{\mathbf{x}\in\mathcal{W}} \|\hat{\mathbf{s}}_n(\mathbf{x};\mathbf{H}_n) - \mathbf{s}(\mathbf{x})\| = o_p(1)$, (iii) $\sup_{\mathbf{x}\in\mathcal{W}} \|\hat{g}_n(\mathbf{x};\mathbf{H}_n) - g(\mathbf{x})\| = o_p(1)$, and (iv) $\sup_{\mathbf{x}\in\mathcal{W}} \|\partial\hat{g}_n(\mathbf{x};\mathbf{H}_n)/\partial\mathbf{x} - \partial g(\mathbf{x})/\partial\mathbf{x}\| = o_p(1)$.

It follows from Theorem 2 and its proof that (i) and (ii) hold. Also, Lemma B-1 (with $(Y,\mathbf{X}') = (y,\mathbf{x}')$, $s = S$, $\kappa = K$ and $\mathcal{X}_n = \mathcal{W}$) and routine arguments can be used to show that if Assumptions 1 and 2 are satisfied and if Equations (2) and (5) hold, then (iii) will be implied by $n^{1-1/S}|\mathbf{H}_n|/\log n \to \infty$. Similarly, (iv) can be established under the condition $n^{1-1/S}|\mathbf{H}_n|\lambda_{\min}(\mathbf{H}_n)/\log n \to \infty$. The latter holds if condition (i), (ii), or (iii) in the statement of the theorem is satisfied.

APPENDIX B: UNIFORM CONVERGENCE RATES FOR KERNEL ESTIMATORS

This Appendix derives uniform convergence rates for kernel estimators. Lemma B-1 is used in the proofs of the main results of this article. Because this result may be of independent interest, it is stated at a (slightly) greater level of generality than needed in the proofs of the other results in this article.

Suppose $(Y_i, \mathbf{X}_i')'$, $i = 1,\ldots,n$, are iid copies of $(Y,\mathbf{X}')'$, where $\mathbf{X}\in\mathbb{R}^d$ is continuous with density $f_{\mathbf{X}}(\cdot)$. Consider the nonparametric estimator

$$\hat{\Psi}_n(\mathbf{x}) = \frac{1}{n}\sum_{j=1}^n \kappa_{\mathbf{H}_n}(\mathbf{x}-\mathbf{X}_j)Y_j, \qquad \kappa_{\mathbf{H}}(\mathbf{x}) = |\mathbf{H}|^{-1}\kappa(\mathbf{H}^{-1}\mathbf{x}),$$

where $\mathbf{H}_n$ is a sequence of diagonal, positive definite $d\times d$ bandwidth matrices and $\kappa:\mathbb{R}^d\to\mathbb{R}$ is a kernel-like function. To obtain uniform convergence rates for $\hat{\Psi}_n$, we make the following assumptions.

Assumption B-1. For some $s\geq 2$, $\mathbb{E}[|Y|^s] + \sup_{\mathbf{x}\in\mathbb{R}^d}\mathbb{E}[|Y|^s|\mathbf{X}=\mathbf{x}]f_{\mathbf{X}}(\mathbf{x}) < \infty$.

Assumption B-2. (a) $\sup_{\mathbf{u}\in\mathbb{R}^d}|\kappa(\mathbf{u})| + \int_{\mathbb{R}^d}|\kappa(\mathbf{u})|d\mathbf{u} < \infty$. (b) κ admits a $\delta_\kappa > 0$ and a function $\kappa^*:\mathbb{R}^d\to\mathbb{R}_+$ with $\sup_{\mathbf{u}\in\mathbb{R}^d}\kappa^*(\mathbf{u}) + \int_{\mathbb{R}^d}\kappa^*(\mathbf{u})d\mathbf{u} < \infty$, such that $|\kappa(\mathbf{u}) - \kappa(\mathbf{u}^*)| \leq \|\mathbf{u}-\mathbf{u}^*\|\kappa^*(\mathbf{u}^*)$ whenever $\|\mathbf{u}-\mathbf{u}^*\|\leq\delta_\kappa$.

Remark 4. Assumption B2(b) is adapted from the article by Hansen (2008). It holds if κ is differentiable with $\bar{\kappa}(0) + \int_{\mathbb{R}}\bar{\kappa}(u)du < \infty$, where $\bar{\kappa}(u) = \sup_{\|\mathbf{r}\|\geq u}\|\partial\kappa(\mathbf{r})/\partial\mathbf{r}\|$.

The first result gives an upper bound on the convergence rate of $\hat{\Psi}_n$ on (possibly) expanding sets of the form $\mathcal{X}_n = \{\mathbf{x}\in\mathbb{R}^d : \|\mathbf{x}\|\leq C_{\mathbf{X},n}\}$, where $C_{\mathbf{X},n}$ is a positive sequence satisfying

$$\overline{\lim}_{n\to\infty}\frac{\log(C_{\mathbf{X},n})}{\log n} < \infty. \quad \text{(B.1)}$$

Lemma B-1. Suppose Assumptions B1 and B2 are satisfied and suppose Equation (B.1) holds. If $\lambda_{\max}(\mathbf{H}_n)\to 0$ and $n^{1-1/s}|\mathbf{H}_n|/\log n\to\infty$, then

$$\sup_{\mathbf{x}\in\mathcal{X}_n}\left|\hat{\Psi}_n(\mathbf{x}) - \Psi_n(\mathbf{x})\right| = O_p(\rho_n), \quad \rho_n = \sqrt{\frac{\log n}{n|\mathbf{H}_n|}}\max\left\{1, \sqrt{\frac{\log n}{n^{1-2/s}|\mathbf{H}_n|}}\right\},$$

where $\Psi_n(\mathbf{x}) = \mathbb{E}[\hat{\Psi}_n(\mathbf{x})]$.

Remark 5. The natural "$s=\infty$" analog of Lemma B-1 holds if Y is bounded (e.g., if $Y\equiv 1$, as in the case of density estimation). In other words, the lower bound $n|\mathbf{H}_n|/\log n\to\infty$ suffices and ρ_n can be set equal to $\sqrt{\log n}/\sqrt{n|\mathbf{H}_n|}$ when Y is bounded.

Lemma B-1 generalizes Newey (1994b, Lemma B.1) in three respects. First, we obtain results allowing for matrix bandwidths as opposed to a scalar, common bandwidth for all the covariates. Second, by borrowing ideas from the article by Hansen (2008), we are able to accommodate kernels with unbounded support and to establish uniform convergence over certain types of expanding sets. Finally, and more importantly (for our purposes at least), Lemma B-1 relaxes the condition $n^{1-2/s}|\mathbf{H}_n|/\log n\to\infty$ imposed by Newey (1994b, Lemma B.1), when assuming $\mathbf{H}_n = h_n\mathbf{I}_d$. In typical applications of Newey (1994b, Lemma B.1), a condition like $s\geq 4$ is imposed to ensure that $n^{1-2/s}h_n^d/\log n\to\infty$ is implied by "natural" conditions on h_n, such as $nh_n^{2d}/(\log n)^2\to\infty$ (e.g., Newey, 1994b, Theorem 4.2; Newey and McFadden, 1994, Theorem 8.11). In contrast, only $s\geq 2$ is required for the condition imposed in Lemma B-1 to be implied by $nh_n^{2d}/(\log n)^2\to\infty$ (or its matrix analog $n|\mathbf{H}_n|^2/(\log n)^2\to\infty$).

If $n^{1-2/s}|\mathbf{H}_n|/\log n\to 0$, then the uniform rate obtained in Lemma B-1 falls short of the "usual" rate $\sqrt{n|\mathbf{H}_n|/\log n}$. This is potentially problematic if Lemma B-1 is used to establish uniform convergence with a certain rate (e.g., $n^{1/4}$ or $n^{1/6}$, as in proofs of results such as Equation (6)). On the other hand, the slower rate of convergence is of no concern when any rate of convergence will do (as in proofs of consistency results such as Equation (12)).

Because of their ability to control bias in some cases, leave-one-out estimators of the form

$$\hat{\Psi}_{n,i}(\mathbf{x}) = \frac{1}{n-1}\sum_{j=1,j\neq i}^n \kappa_{\mathbf{H}_n}(\mathbf{x}-\mathbf{X}_j)Y_j$$

are sometimes of interest. The next result extends Lemma B-1 to such estimators.

Lemma B-2. Suppose Assumptions B1 and B2 are satisfied and suppose Equation (B.1)holds. If $\lambda_{\max}(\mathbf{H}_n)\to 0$ and $n^{1-1/s}|\mathbf{H}_n|/\log n\to\infty$, then

$$\max_{1\leq i\leq n}\sup_{\mathbf{x}\in\mathcal{X}_n}\left|\hat{\Psi}_{n,i}(\mathbf{x}) - \Psi_{n,i}(\mathbf{x})\right| = O_p(\rho_n), \qquad \Psi_{n,i}(\mathbf{x}) = \mathbb{E}[\hat{\Psi}_{n,i}(\mathbf{x})].$$

Another corollary of Lemma B-1 is the following result, which can be useful when uniform convergence on the support of the empirical distribution of $\mathbf{X}$ suffices.

Lemma B-3. Suppose $\mathbb{E}[\|\mathbf{X}\|^{s_{\mathbf{X}}}] < \infty$ for some $s_{\mathbf{X}} > 0$ and suppose Assumptions B1 and B2 are satisfied. If $\lambda_{\max}(\mathbf{H}_n) \to 0$ and $n^{1-1/s}|\mathbf{H}_n|/\log n \to \infty$, then

$$\max_{1 \le i \le n} \left| \hat{\Psi}_n(\mathbf{X}_i) - \Psi_n(\mathbf{X}_i) \right| = O_p(\rho_n),$$

and

$$\max_{1 \le i \le n} \left| \hat{\Psi}_{n,i}(\mathbf{X}_i) - \Psi_{n,i}(\mathbf{X}_i) \right| = O_p(\rho_n).$$

Remark 6. Lemmas B-2 and B-3 are not used elsewhere in the article. We have included them because they may be of independent interest.

SUPPLEMENTARY MATERIALS

Supplementary appendix.

[Received November 2011. Revised August 2012.]

REFERENCES

Abadie, A., and Imbens, G. W. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267. [1245]

Altonji, J. G., Ichimura, H., and Otsu, T. (2012), “Estimating Derivatives in Nonseparable Models With Limited Dependent Variables,” *Econometrica*, 80, 1701–1719. [1245]

Campbell, J. R. (2011), “Competition in Large Markets,” *Journal of Applied Econometrics*, 26, 1113–1136. [1245]

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2010), “Robust Data-Driven Inference for Density-Weighted Average Derivatives,” *Journal of the American Statistical Association*, 105, 1070–1083. [1244]

——— (in press), “Small Bandwidth Asymptotics for Density-Weighted Average Derivatives,” *Econometric Theory*. [1244]

Chen, X. (2007), “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in *Handbook of Econometrics* (Vol. VI), eds. J. J. Heckman and E. Leamer, New York: Elsevier Science B.V., pp. 5549–5632. [1243,1246,1247]

Coppejans, M., and Sieg, H. (2005), “Kernel Estimation of Average Derivatives and Differences,” *Journal of Business and Economic Statistics*, 23, 211–225. [1245]

Deaton, A., and Ng, S. (1998), “Parametric and Nonparametric Approaches to Price and Tax Reform,” *Journal of the American Statistical Association*, 93, 900–909. [1245]

Hansen, B. E. (2008), “Uniform Convergence Rates for Kernel Estimation With Dependent Data,” *Econometric Theory*, 24, 726–748. [1246,1255]

Härdle, W., Hart, J., Marron, J., and Tsybakov, A. (1992), “Bandwidth Choice for Average Derivative Estimation,” *Journal of the American Statistical Association*, 87, 218–226. [1244]

Härdle, W., Hildenbrand, W., and Jerison, M. (1991), “Empirical Evidence on the Law of Demand,” *Econometrica*, 59, 1525–1549. [1245]

Härdle, W., and Stoker, T. (1989), “Investigating Smooth Multiple Regression by the Method of Average Derivatives,” *Journal of the American Statistical Association*, 84, 986–995. [1244,1245]

Horowitz, J., and Härdle, W. (1996), “Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates,” *Journal of the American Statistical Association*, 91, 1632–1640. [1244]

Ichimura, H., and Linton, O. (2005), “Asymptotic Expansions for Some Semiparametric Program Evaluation Estimators,” in *Identification and Inference in Econometric Models: Essays in Honor of Thomas J. Rothenberg*, eds. D. W. K. Andrews and J. H. Stock, New York: Cambridge University Press, pp. 149–170. [1244]

Ichimura, H., and Todd, P. E. (2007), “Implementing Nonparametric and Semiparametric Estimators,” in *Handbook of Econometrics* (Vol. VIB), eds. J. J. Heckman and E. Leamer, New York: Elsevier Science B.V., pp. 5370–5468. [1243,1247]

Imbens, G. W., and Newey, W. K. (2009), “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity,” *Econometrica*, 77, 1481–1512. [1245]

Lang, K., and Manove, M. (2011), “Education and Local Market Discrimination,” *American Economic Review*, 101, 1467–1496. [1253]

Mammen, E. (1989), “Asymptotics With Increasing Dimension for Robust Regression With Applications to the Bootstrap,” *The Annals of Statistics*, 17, 382–400. [1244]

Matzkin, R. L. (2007), “Nonparametric Identification,” in *Handbook of Econometrics* (Vol. VIB), eds. J. J. Heckman and E. Leamer, New York: Elsevier Science B.V., pp. 5307–5368. [1245]

Newey, W. K. (1994a), “The Asymptotic Variance of Semiparametric Estimators,” *Econometrica*, 62, 1349–1382. [1243,1246]

——— (1994b), “Kernel Estimation of Partial Means and a General Variance Estimator,” *Econometric Theory*, 10, 233–253. [1246,1250,1255]

Newey, W. K., and McFadden, D. L. (1994), “Large Sample Estimation and Hypothesis Testing,” in *Handbook of Econometrics* (Vol. IV), eds. R. F. Engle and D. L. McFadden, New York: Elsevier Science B.V., pp. 2111–2245. [1243,1246,1247,1255]

Newey, W. K., and Stoker, T. M. (1993), “Efficiency of Weighted Average Derivative Estimators and Index Models,” *Econometrica*, 61, 1199–1223. [1243,1244,1245]

Powell, J. L. (1994), “Estimation of Semiparametric Models,” in *Handbook of Econometrics* (Vol. IV), eds. R. F. Engle and D. McFadden, New York: Elsevier Science B.V., pp. 2443–2521. [1245]

Powell, J. L., Stocks, J. H., and Stoker, T. M. (1989), “Semiparametric Estimation of Index Coefficients,” *Econometrica*, 57, 1403–1430. [1244]

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. (2008), “Higher Order Influence Functions and Minimax Estimation of Nonlinear Functionals,” in *Probability and Statistics: Essays in Honor of David A. Freedman*, eds. D. Nolan and T. Speed, Beachwood, OH: Institute of Mathematical Statistics, pp. 335–421. [1245]

Schucany, W. R. (1988), “On Nonparametric Regression With Higher-Order Kernels,” *Journal of Statistical Planning and Inference*, 23, 145–151. [1249]

Schucany, W. R., and Sommers, J. P. (1977), “Improvement of Kernel Type Density Estimators,” *Journal of the American Statistical Association*, 72, 420–423. [1244]

Stoker, T. M. (1986), “Consistent Estimation of Scaled Coefficients,” *Econometrica*, 54, 1461–1481. [1243,1244,1245]

——— (1989), “Tests of Additive Derivative Constraints,” *Review of Economic Studies*, 56, 535–552. [1245]