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.

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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).

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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

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

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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 $n|\mathbf{H}_n|^2 \not\to \infty$. 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

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

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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:

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

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

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

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:

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

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 $d \geq 3$.

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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 $\mathbb{E}[\hat{\boldsymbol{\theta}}_n^{**}(\mathbf{H}_n)] - \boldsymbol{\theta} \approx$

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

$n^{-1}|\mathbf{H}_n|^{-1}\mathcal{B}_0 + \mathcal{S}(\mathbf{H}_n)$, 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) $S \geq 3 + 2/d$.

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

Part (i) of the theorem shows that even under the (seemingly) minimal moment requirement $S = 2$, consistency of $\hat{\mathbf{\Sigma}}_n$ 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 $\hat{\mathbf{\Sigma}}_n$ 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:

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

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

$$\text{Model 3} \;:\; \mathbf{x}_i \sim_{\text{iid}} \begin{bmatrix} (\chi^2_4 - 4)/\sqrt{8} \\ \mathcal{N}(\mathbf{0}_2, \mathbf{V}_3) \end{bmatrix}, \quad \mathbf{V}_3 = \begin{bmatrix} 2/3 & 1/4 \\ 1/4 & 1 \end{bmatrix},$$

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 $\hat{\mathbf{\Sigma}}_n(\mathbf{H}_n)$ 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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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).

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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).

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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

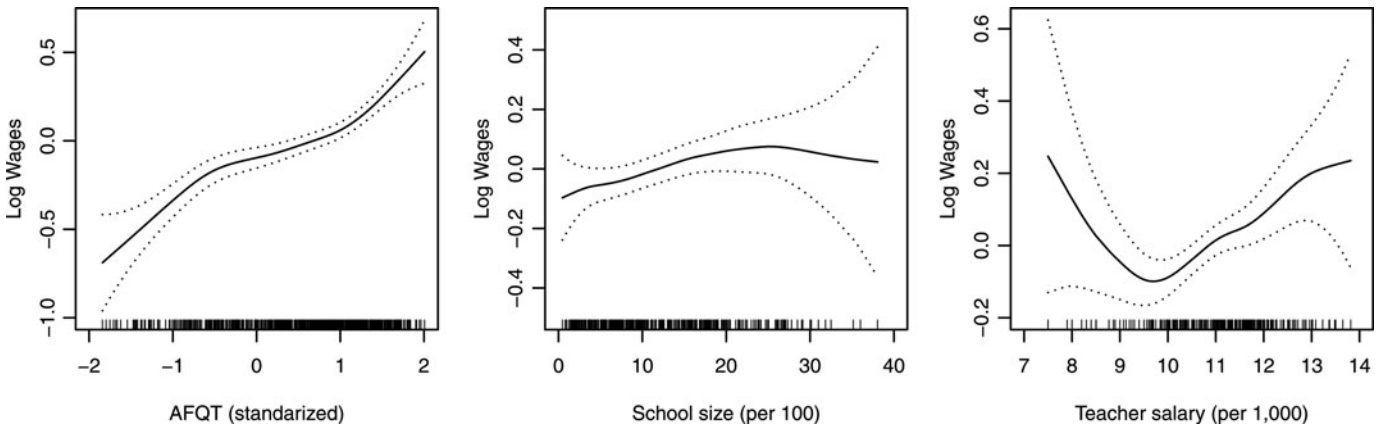

Figure 1. Smoothing splines estimates for univariate conditional expectations.

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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 $\hat{\mathbf{f}}(\mathbf{x}) = \partial f(\mathbf{x})/\partial\mathbf{x}$, 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

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

(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 $\mathbf{x} \in \mathcal{W}$,

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

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

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

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]

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

Comment

Donglin ZENG

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

Cattaneo, Grump, and Jansson (2013) present an interesting estimator, namely the generalized jackknife estimator, for estimating weighted average derivatives. Starting with a high-order (in this case, second-order) linearization of the estimating equation, they obtain the asymptotic approximation under a weak bandwidth selection which does not require the standard convergence rate of the nonparametric estimator faster than $n^{1/4}$. Specifically, an asymptotic approximation of $\widehat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ is given when $n|\mathbf{H}_n|^{3/2}\lambda_{\min}(\mathbf{H}_n)/\log(n)^{3/2} \to \infty$. The polynomial expression of the asymptotic bias in $\widehat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ in terms of $\mathbf{H}_n$ further motivates the construction of the generalized jackknife estimator $\widetilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, c)$, which eliminates the asymptotic bias. They present a number of simulation studies demonstrating that $\widetilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, c)$ leads to noticeable bias reduction with small bandwidths. Another contribution includes a proof of the uniform convergence of the kernel estimators.

1. ASYMPTOTIC BIAS REDUCTION

Under a weak assumption on the bandwidth, this work handles bias reduction via a second-order linearization of $\widehat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ in terms of the plug-in kernel estimator for $f(x)$. A similar technique was used by Robins et al. (2008) who addressed the convergence rate with high-dimensional covariates. As pointed out by Robins et al. (2008), the same technique can be carried out for a cubic or even higher-order linearization if the estimating function is sufficiently smooth in $f(x)$. Then, an even weaker bandwidth assumption is needed when a generalized jackknife estimator is constructed, although the simulation evidence suggests that the current second-order linearization is sufficient to render a negligible bias relative to its standard deviation for the sample sizes used.

In nonparametric or semiparametric literature, an alternative approach to perform bias reduction is to use a high-order kernel which has high-order zero moments. Consider the one-dimensional case. A high-order kernel function $K(x)$ satisfies $\int x^l K(x)dx = 0$ for $|l| \le P$. Then the asymptotic bias in $\widehat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ will be in the form of $\int \Omega(x + \mathbf{H}_n)K(x)dx$ so, by the Taylor expansion and assuming $\Omega(x)$ is sufficiently smooth, this bias is asymptotically equivalent to $O(\mathbf{H}_n^{P+1})$. Therefore, a weak assumption on $\mathbf{H}_n$ is required to eliminate this bias.

The comparison between these two approaches can be summarized in the following way. The first approach, which is implemented in the current article, is to directly study the influence of the bandwidth $\mathbf{H}_n$ on the estimating function, which in turn relies on the smoothness of the estimating function as a functional of $f(x)$. In contrast, the second approach uses the high-order kernel function to examine the influence of $\mathbf{H}_n$ on the plug-in estimator $\widehat{f}(x)$ so mostly relies on the smoothness of $f(x)$. From this point of view, it is evident that the former is useful in semiparametric estimation when some functional of $f(x)$ instead of $f(x)$ itself is of interest. However, when a class of functionals of $f(x)$, for example, $\boldsymbol{\theta} = E[w(x)\nabla g(x)]$ when $w(x)$ belongs to a class of weights, it may be difficult for the first method to identify a uniform $\mathbf{H}_n$ to eliminate all the bias in estimating the whole class of functionals; instead, the second method has its advantage as it only relies on the smoothness of $f(x)$ regardless of the number of $w(x)$'s in consideration.

2. DATA-ADAPTIVE JACKKNIFE ESTIMATOR

In the construction of the generalized jackknife estimator $\widetilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, c)$, one has to determine the order J (satisfying $J < 1 + d/2$ and $J \ge (d-2)/8$) so that

$$\sum_{j=0}^{J} w_j(c_j)E[\widehat{\boldsymbol{\theta}}_n^{**}(c_j\mathbf{H}_n)] - \boldsymbol{\theta} = o(n^{-1/2}),$$

where $w_j(c)$ is given in Section 3.2 of the article. The simulations use $J = 2$. A more data-adaptive construction of the jackknife estimator can be performed as follows. We again use the fact that the asymptotic bias is in a polynomial order of bandwidth. Thus, for c chosen from a reasonable range, consider fitting the following regression model:

$$\widehat{\boldsymbol{\theta}}(c\mathbf{H}_n) = \boldsymbol{\theta} + c^{-d}(b_0 + b_1c^2 + \cdots + b_Jc^{2J}) + \epsilon,$$

where ϵ is a stochastic term with mean zero and variance of order $n^{-1/2}$ and $J < 1 + d/2$. However, since $\widehat{\boldsymbol{\theta}}(c\mathbf{H}_n)$ is from the same data, this regression is no longer stochastic.

To this end, divide the whole data into N independent data of equal sizes and choose $c_1, \ldots, c_N$. For each c_k, we calculate $\widehat{\boldsymbol{\theta}}(c_k\mathbf{H}_n)$ using the kth data and denote it by $\widehat{\boldsymbol{\theta}}_k$. Then, the above regression model implies

$$\widehat{\boldsymbol{\theta}}_k = \boldsymbol{\theta} + c_k^{-d}\left(b_0 + b_1c_k^2 + \cdots + b_Jc_k^{2J}\right) + \epsilon_k,$$

where $\epsilon_k, k = 1, \ldots, N$ are iid and asymptotically follow the normal distribution with mean zero and covariance $\Sigma/(n/N)$. Therefore, we can regress $\{\widehat{\boldsymbol{\theta}}_k\}$ on $(1, c_k^{-d}, \ldots, c_k^{-d+2J})$ to

1. first, we implement the AIC or BIC to choose J;
2. we estimate $\boldsymbol{\theta}$ after J is chosen;
3. we estimate Σ using the residual variance–variance matrix.

3. VARIANCE ESTIMATION

Unfortunately, the variance estimates reported in the simulations perform rather poorly. My experience is that one may need larger bandwidths than the ones used in point estimation to

Donglin Zeng, University of North Carolina, Biostatistics, Chapel Hill, USA (E-mail: *dzeng@email.unc.edu*).

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.854172

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

estimate the nonparametric quantity in the variance estimation. Alternatively, the bootstrap approach may be worth pursuing, especially smoothed bootstrapping, where bootstrap samples are simulated from a kernel density estimator of (Y, X). The asymptotic properties of the bootstrap estimator can be established along the same lines as in the current article.

4. USE OF EMPIRICAL PROCESS THEORY

Empirical process theory has been a powerful tool to establish the uniform convergence of many estimators. In this case, it can be used to derive a similar result (but with stronger bandwidth condition) to Lemma B-1 regarding the kernel estimator. For example, consider $d = 1$. First, $\widehat{\psi}_n(x) - \psi_n(x) = n^{-1/2}\mathbf{G}_n[k_{\mathbf{H}_n}(x - X)Y]$, where $\mathbf{G}_n$ denotes the empirical process. Consider the class of functions $\mathcal{F} = \{k_{\mathbf{H}_n}(x - X)Y : x \in \chi_n\}$. From Assumption B2, we note

$$\begin{aligned}&\left|k_{\mathbf{H}_n}(x - X)Y - k_{\mathbf{H}_n}(x^* - X)Y\right| \\ &\quad\le \|x - x^*\||Y| \sup_x k^*\left(\mathbf{H}_n^{-1}x\right)|\mathbf{H}_n|^{-2}.\end{aligned}$$

Therefore, this class function has an envelop function given by $F = \mathbf{H}_n^{-1}|Y|$ and has a finite bracket entropy integral, that is,

$$\int_0^1 \sqrt{1 + \log N_{[]}(\epsilon\|F\|, \mathcal{F}, \|\cdot\|_{L_2(P)})}d\epsilon < \infty.$$

Following Theorem 2.14.2 in van der Vaart and Wellner (1996), it yields

$$\|\sup_{\mathcal{F}}|\mathbf{G}_n|\| = O_p\left(\|F\|_{L_2(P)}\right) = O_p\left(\mathbf{H}_n^{-1}\right).$$

This gives

$$\sup_{x\in\chi_n}|\widehat{\psi}_n(x) - \psi_n(x)| = O_p\left(\frac{1}{\sqrt{n\mathbf{H}_n^2}}\right).$$

5. EXTENSION TO MORE GENERAL SEMIPARAMETRIC ESTIMATION

The same technique can be applied to a more general semiparametric estimation where the parameter of interest, $\boldsymbol{\theta}$, implicitly solves an estimating function $E[g(\boldsymbol{\theta}, f, f', f,'' \ldots)] = 0$, where $f(x)$ is the density function of f and f' is its first derivative and so on . These kinds of estimating equations often arise from modeling certain stochastic dynamic systems, for instance, HIV dynamics. It will be interesting to see how the method can be carried out in this general context.

REFERENCES

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013), "Generalized Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1256. [1257]

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. [1257]

van der Vaart, A. W., and Wellner, J. (1996), *Weak Convergence and Empirical Process*, New York: Springer. [1258]

Comment

Holger DETTE

The article of Cattaneo, Crump, and Jansson (2013) makes three important contributions to weighted average derivative estimation. It provides a new first-order asymptotic approximation based on a quadratic expansion of the estimating equation. With this approach nonparametric estimators with a slower rate of convergence can be used for weighted derivative estimation. Moreover, from a technical point of view, an asymptotic analysis under substantially weaker conditions on the moments of the dependent variable and on the bandwidths is possible. Additionally, an interesting method for the elimination of an asymptotic bias is proposed which is based on jackknife methodology.

For the sake of brevity, the focus of this discussion is on the jackknife methodology. A careful investigation of this approach in the case of weighted average derivative estimation would be too technical and beyond the scope of a discussion. Therefore, we will raise some general questions regarding the elimination of the bias by jackknife methodology in the context of "classical" density estimation. All observations carry obviously over to the more complicated case of weighted derivative estimation. In particular, I will comment on the choice of c_i for two reasons:

1. I do not think that there exists an optimal choice of the weights c_i in the jackknife approach, at least if one applies the "usual" mathematical machinery.
2. Some care is necessary in the application of the jackknifing methodology, because in finite samples one pays a serious price for the bias reduction in terms of variance.

Notation. We consider the classical setup of one-dimensional density estimation, where $X_1, \ldots, X_n$ are independent identically distributed random variables with density f. The classical density estimate is defined by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right). \tag{1}$$

Holger Dette is Professor, Fakultät für Mathematik, Ruhr-Universität Bochum, 44780 Bochum, Germany (E-mail: *holger.dette@rub.de*).

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.859516

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

If f is twice differentiable and the kernel K is symmetric, then the bias of this estimate is given by

$$\mathbb{E}[\hat{f}_h(x)] = \frac{h^2 f''(x)}{2} + o(h^2). \tag{2}$$

Similarly, the variance is obtained as

$$\operatorname{var}(\hat{f}_h(x)) = \frac{f(x)}{nh} \int K^2(u)du \cdot (1 + o(1)). \tag{3}$$

The impact of bias correction on the variance. The jackknife approach (see, e.g., Schucany and Sommers 1977) is based on formula (2) and considers (in the simplest case) an estimator of the form

$$\hat{g}_{c_1,c_2}(x) = w_1 \hat{f}_{c_1 h}(x) + w_2 \hat{f}_{c_2 h}(x),$$

where the weights w_1, w_2 are determined such that $w_1 + w_2 = 1$ and the dominating term in

$$\mathbb{E}[\hat{g}_{c_1,c_2}(x)] = \left(w_1 c_1^2 + w_2 c_2^2\right) \frac{h^2 f''(x)}{2} + o(h^2)$$

vanishes, that is,

$$w_1 = \frac{c_2^2}{c_2^2 - {c_1}^2}; \quad w_2 = \frac{-c_1^2}{c_2^2 - c_1^2}$$

(note that we basically construct a Lagrange interpolation function $w_1 + w_2 x^2$ with values 1 and 0 at the points 1 and c_2/c_1). For this choice, we obtain a density estimate with bias $\mathbb{E}[\hat{g}_{c_1,c_2}(x)] = o(h^2)$. Now we investigate the variance of the estimator $\hat{g}_{c_1,c_2}(x)$, that is,

$$\begin{aligned}\operatorname{var}(\hat{g}_{c_1,c_2}(x)) = {} & w_1^2 \operatorname{var}(\hat{f}_{c_1 h}(x)) + w_2^2 \operatorname{var}(\hat{f}_{c_2 h}(x)) \\ & + 2 w_1 w_2 \operatorname{cov}(\hat{f}_{c_1 h}(x), \hat{f}_{c_2 h}(x)).\end{aligned}$$

A standard calculation yields

$$\operatorname{cov}(\hat{f}_{c_1 h}(x), \hat{f}_{c_2 h}(x)) = \frac{f(x)}{nhc_2} \int K(u) K\left(\frac{c_1}{c_2} u\right) du (1 + o(1)),$$

and we obtain

$$\begin{aligned}&\operatorname{var}(\hat{g}_{c_1,c_2}(x)) \\ &\quad \geq \left\{ \left(\frac{w_1^2}{c_1} + \frac{w_2^2}{c_2} \right) \frac{f(x)}{nh} \int K^2(x)du \right. \\ &\qquad \left. + \; 2 w_1 w_2 \frac{f(x)}{nhc_2} \left(\int K^2(u)du \int K^2\left(\frac{c_1}{c_2} u\right) du \right)^{1/2} \right\} \\ &\qquad \times (1 + o(1)),\end{aligned}$$

where we used the Cauchy Schwarz inequality and the fact that $w_1 w_2 \leq 0$. Finally, a substitution in the integral $\int K^2(\frac{c_1}{c_2}u)du$ and a simple calculation gives

$$\begin{aligned}\operatorname{var}(\hat{g}_{c_1,c_2}(x)) &\geq \alpha^2(c_1, c_2) \frac{f(x)}{nh} \int K^2(u)du(1 + o(1)) \\ &= \alpha^2(c_1, c_2) \operatorname{var}(\hat{f}_h(x))(1 + o(1))\end{aligned} \tag{4}$$

as a lower bound for the variance of the jackknife estimate, where the factor $\alpha^2 = \alpha^2(c_1, c_2)$ is defined by

$$\alpha^2(c_1, c_2) := \left(\frac{w_1}{\sqrt{c_1}} + \frac{w_2}{\sqrt{c_2}} \right)^2. \tag{5}$$

In the following, we will argue that for reasonable choices of the parameters c_1 and c_2 we have $\alpha^2(c_1, c_2) \geq 1$, which implies that the reduction of the bias comes usually with an increase in variance. For this purpose, we display in Table 1 the value of α^2 for various choices of c_1 and c_2 and make the following observations:

Table 1. The value α^2 in (5) for various choices of c_1 and c_2

c_1	c_2	α^2	c_1	c_2	α^2
0.5	1	2.41	0.5	0.7	2.71
0.7	1	1.91	0.3	0.6	4.23
0.9	1	1.65	0.2	0.6	5.54
0.8	1.2	1.64	1.2	1.6	1.14
0.8	1.4	1.56	1.2	1.8	1.10
0.8	1.6	1.51	1.4	1.8	0.99

1. For reasonable choices of c_1 and c_2, the factor α^2 is always larger than 1. This means the bias reduction is obtained at a cost of a larger variance (note that the right-hand side of Equation (4) provides a lower bound for the variance of $\hat{g}_{c_1,c_2}(x)$).
2. For increasing values of $c_1, c_2 \to \infty$, the first-order approximation for the variance of $\hat{g}_{c_1,c_2}$ becomes arbitrarily small. Thus, in principle there does not exist any optimal choice of the constants c_1 and c_2. Moreover, this reduction is obtained by an increase of the bias in the terms of order h^3, h^4, etc. Thus, these first-order considerations might be misleading.

A similar problem occurs in the application of higher-order kernels. Consider, for example, the Epanechnikov kernel

$$K_1(x) = \frac{3}{4}(1 - x^2) I_{[-1,1]}(x),$$

which is of order 2 (see Gasser, Müller, and Mammitzsch 1985 for a precise definition) and yields a bias of order $O(h^2)$. Now the kernel

$$K_2(x) = \frac{15}{32}(1 - x^2)(3 - 7x^2) I_{[-1,1]}(x)$$

is of order 4 and yields a bias of order $O(h^4)$. However, we obtain for the corresponding terms in the variance

$$\int K_1^2(x)dx = \frac{3}{5}, \qquad \int K_2^2(x)dx = \frac{5}{4},$$

which means that the kernel density estimate (1) based on the kernel K_2 has a 108% larger variance than the corresponding estimate based on the kernel K_1. Similarly, if the kernel of order 6

$$K_3(x) = \frac{15}{256}(1 - x^2)(35 - 250x^2 - 231x^4 + 231x^6)$$

is used, the asymptotic variance increases by a factor 3.15. Gasser, Müller, and Mammitzsch (1985) realized these problems and proposed to choose the kernel K, such that it minimizes the first-order approximation of the mean squared error if an asymptotic optimal bandwidth has been used. While this method yields an improvement in kernel density and regression estimation, it seems to be difficult to develop an analog concept for the jackknife methodology.

REFERENCES

Cattaneo, M. D., Crump, R. K., and Jansson, M. (2013), "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1256. [1258]

Gasser, T., Müller, H. G., and Mammitzsch, V. (1985), "Kernels for Nonparametric Curve Estimation," *Journal of the Royal Statistical Society*, Series B, 47, 238–252. [1259]

Schucany, W. R., and Sommers, J. P. (1977), "Improvement of Kernel Type Density Estimators," *Journal of the American Statistical Association*, 72, 420–423. [1259]

Comment

Enno MAMMEN

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

Professors M. D. Cattaneo, R. K. Crump, and M. Jansson are to be congratulated for an interesting article with a new point of view on semiparametrics. Their nonstandard way to look at semiparametric estimation problems is very innovative and it is motivating for further research.

The article studies what happens if one goes beyond the border of standard asymptotics. For a specific example, the article discusses a semiparametric estimation problem, where the nonparametric estimator has a poorer asymptotic performance than required from classical semiparametric theory. This is an important problem, in the concrete setting of the article and also in general theory. Often, in semiparametrics, assumptions are made on the nonparametric estimator that are not realistic. An example would be higher dimensional nonparametric regression functions where higher order smoothness assumptions are made that allow $o_P(n^{-1/4})$ convergence of the nonparametric estimator. There are some concerns in nonparametrics about the sense of such higher order smoothness conditions for moderate sample sizes, see, for example, Marron and Wand (1992). It is natural to argue that also in semiparametric contexts it is questionable if these higher order assumptions make sense. This motivates an asymptotic framework in semiparametrics, where such assumptions are avoided and where this problem is not neglected in the asymptotic limit. That is exactly what the authors of this article have done. I think that the article addresses a central question of mathematical statistics.

As mentioned in the article, the discussions of the article are related to recent work of L. Li, J. Robins, E. Tchetgen, and A. van der Vaart, but a different point of view is taken here. It is assumed that the bias of the nonparametric estimator is negligible and does not influence the first-order asymptotics of the parametric estimator. Then the asymptotics of the parametric part is only affected by the stochastic part of the nonparametric estimator. As was shortly mentioned in the article, this relates the article to discussions on high-dimensional parametric models. Nonparametric regression can be interpreted as parametrics with increasing dimension. Then the nuisance nonparametric component is related to a nuisance parameter with increasing dimension in a purely parametric model. In the following I will give a more detailed discussion of this relation in the context of this article.

1. DIMENSION ASYMPTOTICS

High-dimensional models are a central example where asymptotic frameworks are used that do not neglect an important finite-sample feature in the asymptotic limit. Here, the important feature is the high dimensionality of the model. For high-dimensional models, this can be easily done by letting the dimension of the model grow with increasing sample size. Recently, there has been a huge amount of research on high-dimensional models under sparsity constraints. This has also motivated investigators to revisit older strands of research and to study high-dimensional models without sparsity, see, for example, Belloni, Chernozhukov, and Fernandez-Val (2011) who considered high-dimensional linear quantile regression. Early papers on dimension asymptotics in linear models were Huber (1973) and Portnoy (1984, 1985, 1986). High-dimensional log-linear models were considered in Haberman (1977a,b) and Ehm (1991). The latter papers discuss applications to large contingency tables where the minimal cell expectations do not converge to infinity. Exponential families with increasing dimension were studied in Portnoy (1988) and Belloni and Chernozhukov (2012). For linear and log–linear models, Mammen (1989) and Sauermann (1989) showed consistency of bootstrap for linear contrasts under conditions where the normal approximation fails because of bias effects. These two papers are closely related in spirit to the findings in the article of M. D. Cattaneo, R. K. Crump, and M. Jansson. I will outline this below for robust linear regression. I would like to mention other papers, where dimension asymptotics lead to insights that were hidden by asymptotics with fixed dimension. Bickel and Freedman (1983) proved consistency of bootstrap for least-squares estimation in high-dimensional linear models that includes cases where the asymptotic distribution is nonnormal. This was the first article where it was shown that bootstrap works in a setting where classical approaches fail. Bootstrap and Wild Bootstrap were compared in Mammen (1993), again including settings where the

Enno Mammen is Professor in Statistics, Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany (E-mail: *emammen@rumms.uni-mannheim.de*). The author acknowledges support by the DFG project FOR916.

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.829000

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

normal approximation fails. Mammen (1996) showed that for ML estimation in high-dimensional linear models the empirical distribution of residuals is biased toward the assumed error distribution.

2. NUISANCE PARAMETERS WITH INCREASING DIMENSION

I now outline the relation between a parametric model with a high-dimensional nuisance parameter and the semiparametric estimation problem of the article by M. D. Cattaneo, R. K. Crump, and M. Jansson. I will do this by using the example of robust regression in a high-dimensional linear model. Suppose one observes $Y_i = X_i^\top \beta + \varepsilon_i$ with deterministic covariables $X_i \in \mathbb{R}^p$ and iid errors with $\mathbb{E}[\psi(\varepsilon_i)] = 0$ for a function $\psi : \mathbb{R} \to \mathbb{R}$. Consider an M-estimator $\hat{\beta}_n$ with M-function ψ:

$$\sum_{i=1}^n X_i \psi\left(Y_i - X_i^\top \hat{\beta}_n\right) = 0. \tag{1}$$

W.l.o.g. we assume that $\sum_{i=1}^n X_i X_i^\top = I_p$, where I_p is the $p \times p$ identity matrix. Then $p = \text{trace}\,[\sum_{i=1}^n X_i X_i^\top] = \text{trace}\,[\sum_{i=1}^n X_i^\top X_i] = \sum_{i=1}^n \|X_i\|^2$. For simplicity, we make the assumption that the design vectors are of the same order of size, in the sense that $\max_{1\le i\le n} \|X_i\|^2 = O(p/n)$. For dimension p fixed one has under regularity assumptions that $\hat{\beta}_n - \beta$ converges in distribution to $N(0, \rho_0\rho_1^{-2} I_p)$, where $\rho_0 = \mathbb{E}[\psi^2(\varepsilon_i)]$ and $\rho_1 = \mathbb{E}[\psi'(\varepsilon_i)]$. In particular, for $c_n \in \mathbb{R}^p$ with norm $\|c_n\| = 1$ one gets that the linear contrast $c_n^\top(\hat{\beta}_n - \beta)$ has a normal limit $N(0, \rho_0\rho_1^{-2})$.

We now start a heuristic discussion for the case that $p \to \infty$. By Taylor expansion of the left-hand side of Equation (1) one gets that $0 \approx \sum_{i=1}^n X_i\psi(\varepsilon_i) - \sum_{i=1}^n X_i X_i^\top \psi'(\varepsilon_i)(\hat{\beta}_n - \beta) + (1/2)\sum_{i=1}^n X_i[X_i^\top(\hat{\beta}_n - \beta)]^2\psi''(\varepsilon_i)$. This gives with $\rho_2 = \mathbb{E}[\psi''(\varepsilon_i)]$, $\rho_3 = \mathbb{E}[\psi(\varepsilon_i)[\psi'(\varepsilon_i) - \rho_1]]$ and $\psi_1(x) = \psi'(x) - \rho_1$

$$\begin{aligned}
\hat{\beta}_n - \beta &\approx \rho_1^{-1}\sum_{i=1}^n X_i\psi(\varepsilon_i) - \rho_1^{-2}\sum_{i,j=1}^n X_i\psi_1(\varepsilon_i)\left(X_i^\top X_j\right)\psi(\varepsilon_j) \\
&\quad + \frac{1}{2}\rho_1^{-3}\sum_{i,j,k=1}^n X_i\psi''(\varepsilon_i)\left(X_i^\top X_j\right)\psi(\varepsilon_j)(X_i^\top X_k)\psi(\varepsilon_k) \\
&\approx \rho_1^{-1}\sum_{i=1}^n X_i\psi(\varepsilon_i) - \rho_1^{-2}\rho_3\sum_i^n X_i(X_i^\top X_i) \\
&\quad + \frac{1}{2}\rho_1^{-3}\rho_2\rho_0\sum_{i,j=1}^n X_i\left(X_i^\top X_j\right)^2 \\
&= \rho_1^{-1}\sum_{i=1}^n X_i\psi(\varepsilon_i) + \rho_1^{-3}\left(\frac{1}{2}\rho_2\rho_0 - \rho_1\rho_3\right) \\
&\quad \times \sum_i^n X_i\|X_i\|^2.
\end{aligned} \tag{2}$$

Under appropriate conditions, this expansion is valid with rest terms of order $p^{3/2}\log(n)^{3/2}/n$. This can be shown with the methods developed in Mammen (1989). For a linear contrast $c_n^\top(\hat{\beta}_n - \beta)$ with $\|c_n\| = 1$ one gets that $c_n^\top(\hat{\beta}_n - \beta) - c_n^\top b_n$ has a normal limit $N(0, \rho_0\rho_1^{-2})$ where $b_n = \rho_1^{-3}(\frac{1}{2}\rho_2\rho_0 - \rho_1\rho_3)\sum_i^n X_i\|X_i\|^2$. The bias term is of order $O(pn^{-1/2})$. This follows from $\|b_n\| = O(pn^{-1/2})$. Note that for a vector e with $\|e\| = 1$ it holds that

$$\begin{aligned}
|e^\top b_n| &\le Cn^{1/2}\left[\sum_{i=1}^n (e^\top X_i\|X_i\|^2)^2\right]^{1/2} \\
&\le n^{1/2}\max_{1\le i\le n}\|X_i\|^2\left[\sum_{i=1}^n (e^\top X_i)^2\right]^{1/2} = O(pn^{-1/2})
\end{aligned}$$

because of $\sum_{i=1}^n X_i X_i^\top = I_p$ and $\max_{1\le i\le n}\|X_i\|^2 = O(pn^{-1})$.

One can write $X_i^\top\beta = X_{i,1}\beta_1 + X_{i,-1}^\top\beta_{-1}$, where β_1 is the first element of β and where β_{-1} contains the remaining elements of β. If $X_{i,-1}^\top\beta_{-1}$ is a series expansion of a nonparametric function and if β_1 is the parameter of interest and β_{-1} a nuisance parameter we are in a semiparametric model as is the case in the article by Cattaneo, Crump, and Jansson. Note also that in their article bias terms of the nonparametric estimators are neglected in the chosen asymptotic setting. With the choice $c_n = (1, 0, \ldots, 0)^\top$, we get from the above discussion the following conclusions. As long as $p^{3/2}\log(n)^{3/2}/n \to 0$, it holds

(1) that $\hat{\beta}_{n,1} - \beta_1$ has an asymptotic bias $b_{n,1}$ which is of order $O(pn^{-1/2})$,
(2) and that for $\hat{\beta}_{n,1} - \beta_1 - b_{n,1}$ the same stochastic expansion $\rho_1^{-1}\sum_{i=1}^n X_{i,1}\psi(\varepsilon_i)$ holds as for $\hat{\beta}_{n,1} - \beta_1$ if p is fixed.

Analogous statements hold for the estimator $\hat{\theta}_n(H_n)$ of the article. This follows from their Theorem 2. Note that one has to compare $\hat{\beta}_{n,1} - \beta_1$ with $\sqrt{n}(\hat{\theta}_n(H_n) - \theta)$. The dimension p of the linear model corresponds to $(h_1 \cdot \ldots \cdot h_d)^{-1} = |H_n|^{-1}$. With this relation, we get from part (a) of Theorem 2 that the bias terms of $\hat{\beta}_{n,1}$ and of $\hat{\theta}_n(H_n)$ are of the same order. The validity (2) of the linear stochastic expansion is stated in part (b) of Theorem 2. Even the rest terms in the asymptotic expansions are comparable, at least for d large. This all suggests that the discussion of Cattaneo, Crump, and Jansson apply to a much larger class of models than considered in their article. These are not only further semiparametric models but also high-dimensional models where the dimension of a nuisance parameter converges to infinity.

The above asymptotic expansions also give some insights for higher dimensional models where $p^{3/2}/n$ does not converge to 0. For the case that $p^{3/2}/n \to \infty$ one has to apply Taylor expansions around $\beta - b_n$ instead of expansions around β. The first term in the stochastic expansion (2) of $\hat{\beta}_n - \beta$ now becomes $[\sum_{i=1}^n X_i X_i^\top \mathbb{E}[\psi'(\varepsilon_i - X_i^\top b_n)]^{-1}\sum_{i=1}^n X_i\psi(\varepsilon_i - X_i^\top b_n)$. Because now in general $X_i^\top b_n$ does not converge to zero this term has another variance as the first term in Equation (2). Also the second term in Equation (2) becomes nonrandom, in general.

[Received April 2013. Revised July 2013.]

REFERENCES

Belloni, A., and Chernozhukov, V. (2012), "Posterior Inference in Curved Exponential Families Under Increasing Dimension," Working Paper, MIT, Economics of Department, arXiv:0904.3132. [1260]

Belloni, A., Chernozhukov, V., and Fernandez-Val, I. (2011), "Conditional Quantile Processes Based on Series or Many Regressors," Working Paper, MIT, Economics of Department, arXiv:1105.6154. [1260]

Bickel, P., and Freedman, D. (1983), "Bootstrapping Regression Models With Many Parameters," in *A Festschrift for Erich L. Lehmann in Honor of his Sixty-fifth Birthday*, eds. P. J. Bickel, K. A. Doksum, and J. L. Hodges, Jr., Belmont, CA: Wadsworth, pp. 28–48. [1260]
Ehm, W. (1991), *Statistical Problems With Many Parameters: Critical Quantities for Approximate Normality and Posterior Density Based Inference*, Habilitationsschrift: University of Heidelberg. [1260]
Haberman, J. (1977a), "Log-Linear and Frequency Tables With Small Expected Cell Counts," *The Annals of Statistics*, 5, 1148–1169. [1260]
——— (1997b), "Maximum Likelihood Estimates in Exponential Response Models," *The Annals of Statistics*, 5, 815–841. [1260]
Huber, P. J. (1973), "Robust Regression: Asymptotics, Conjectures and Monte Carlo," *The Annals of Statistics*, 1, 799–821. [1260]
Mammen, E. (1993), "Bootstrap and Wild Bootstrap for High-Dimensional Linear Models," *The Annals of Statistics*, 21, 255–285. [1260]
——— (1996), "Empirical Process of Residuals for High-Dimensional Linear Models," *The Annals of Statistics*, 24, 307–335. [1261]
Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736. [1260]
Portnoy, S. (1984), "Asymptotic Behavior of M-Estimators of p Regression Parameters When p^2/n is Large. I. Consistency," *The Annals of Statistics*, 12, 1298–1309. [1260]
——— (1985), "Asymptotic Behavior of M-Estimators of p Regression Parameters When p^2/n is Large. II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417. [1260]
——— (1986), "Asymptotic Behavior of the Empiric Distribution of M-Estimated Residuals From a Regression Model With Many Parameters," *The Annals of Statistics*, 14, 1152–1170. [1260]
——— (1988), "Asymptotic Behavior of Likelihood Methods for Exponential Families When the Number of Parameters Tends to Infinity," *The Annals of Statistics*, 16, 356–366. [1260]
Sauermann, W. (1989), "Bootstrapping the Maximum Likelihood Estimator in High-Dimensional Log-Linear Models," *The Annals of Statistics*, 17, 1198–1216. [1260]

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

Comment

Xiaohong CHEN

1. INTRODUCTION

There is a great deal of literature on semiparametric two-step estimation of Euclidean parameters of interest in statistics and econometrics. Most of the existing results are about root-n asymptotically normal and efficient estimation of the Euclidean parameter in the second step when unknown nuisance functions are estimated in the first step. Surprisingly enough, there is little research on the finite sample behavior of the first-order asymptotically normal approximation when the Euclidean parameter is a nonlinear functional of the unknown nuisance functions. Cattaneo, Crump, and Jansson (CCJ) are to be congratulated for this excellent article addressing the important issue of nonlinearity bias within the class of root-n asymptotically normal (or regular and asymptotically linear) estimators. In the context of kernel plug-in estimation of a weighted average derivative (WAD) parameter, they (i) characterize the nonlinearity bias by a stochastic quadratic expansion; (ii) highlight that the nonlinearity bias is due to a large variance of nonparametric first-step kernel estimation, and hence could not be reduced by conventional nonparametric bias reduction methods such as increasing the order of the kernel; (iii) propose a clever generalized jackknife procedure to correct the nonlinearity bias; and (iv) establish the root-n asymptotic normality of the bias-corrected WAD estimator $\widetilde{\theta}$ and the consistency of their kernel estimator of the asymptotic variance of $\widetilde{\theta}$ under very weak bandwidth conditions. As a side but very useful technical result, they establish a new uniform convergence rate for kernel estimators.

In the following I make two general comments. First, in some applications, although the Euclidean parameter is nonlinear in one nuisance function, it can be also rewritten as a *linear* functional of another nuisance function that can be consistently estimated via the sieve method. This alternative way to eliminate nonlinearity bias might perform better in finite samples since it is based on estimation of a linear functional. Second, in other applications, there is no simple reparameterization that could convert a nonlinear functional of a nuisance function into a linear functional of another nuisance function. The insight of a stochastic quadratic expansion to characterize the nonlinearity bias suggested in this article should be widely applicable to other semiparametric estimators of nonlinear smooth functionals. The results of this article also call for additional research on how to provide easy-to-compute nonlinearity bias correction and more accurate variance estimation of bias-corrected semiparametric estimators.

2. SIEVE WEIGHTED AVERAGE DERIVATIVE ESTIMATORS

In many applications, although the Euclidean parameter of interest, θ, is a nonlinear functional of one nuisance function f, it could be expressed as a linear functional of another nuisance function g that could be estimated via the sieve method. For these applications, we suspect that a semiparametric two-step estimator of θ based on a nonparametric sieve estimation of g in the first step typically performs better in finite sample than another estimator of θ based on a nonparametric estimation of f in the first step. For example, consider the weighted average derivative parameter θ:

$$\theta = E\left[w(x)\frac{\partial}{\partial x}g_0(x)\right] \quad \text{with} \quad g_0(x) = E[y|X=x], \tag{1}$$

$$= -E\left[y\left(\frac{\partial}{\partial x}w(x) + w(x)\frac{\partial}{\partial x}\log f(x)\right)\right] \tag{2}$$

$$= -E\left[y\left(\frac{\partial}{\partial x}w(x) + w(x)\frac{\partial f(x)}{\partial x}/f(x)\right)\right], \tag{3}$$

Xiaohong Chen, Department of Economics, Yale University, 30 Hillhouse, Box 208281, New Haven, CT 06520 (E-mail: *xiaohong.chen@yale.edu*). The authors thank Matias Cattaneo and Jeffery Racine for very useful discussions.

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.855352

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

where $f()$ is the density of the regressor x and $g()$ is the conditional mean function of y given x. It is clear that θ is linear in nuisance function g_0 (see Equation (1)) and also linear in nuisance function $\log f$ (see Equation (2)), but is nonlinear in nuisance function f (see Equation (3)). CCJ considers estimation of θ based on Equation (3). Alternatively, one could estimate θ based on either Equation (1) or Equation (2).

Sieve WAD estimation based on Equation (1). Let $\widehat{g} = \arg\min_{g\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n [y_i - g(x_i)]^2$ be a sieve least squares (LS) estimator of $g_0(\cdot) = E[y|X = \cdot]$. Then the WAD parameter θ defined in Equation (1) can be estimated by the following sieve WAD estimator:

$$\widehat{\theta}_1 = \frac{1}{n}\sum_{i=1}^n w(x_i)\frac{\partial}{\partial x}\widehat{g}(x_i). \tag{4}$$

There is no universal "best" sieves $\mathcal{H}_n$ to use in terms of the convergence rate in mean squared error metric, since the rate depends on the function parameter space $\mathcal{H}$ to which g_0 belongs. For a typical function space such as a Sobolev space $W_2^m(\mathcal{X})$ or a Holder space $\Lambda^m(\mathcal{X})$, ($\mathcal{X}$ a subset in $\mathbb{R}^d$), we typically obtain $\|\widehat{g} - g_0\|_{L^2(X)} = O_P(n^{-m/(2m+d)})$ for tensor product linear sieves (or series), where the series LS estimator $\widehat{g}$ has a closed-form expression:

$$\widehat{g}(x) = p^{k_n}(x)'(P'P)^- \sum_{i=1}^n p^{k_n}(X_i)Y_i, \quad x \in \mathcal{X}, \tag{5}$$

where $\{p_j(), j = 1, 2, \ldots\}$ denotes a sequence of known basis functions that can approximate any square integrable functions of x well, $p^{k_n}(X) = (p_1(X), \ldots, p_{k_n}(X))'$, $P = (p^{k_n}(X_1), \ldots, p^{k_n}(X_n))'$ and $(P'P)^-$ the Moore–Penrose generalized inverse. This includes as special cases of tensor product polynomial splines, Fourier series, wavelets, Hermite polynomials, etc. (see Newey 1997; Huang 1998; Chen 2007 and the references therein). Therefore, linear sieves (or series) could achieve a convergence rate of $\|\widehat{g} - g_0\|_{L^2(X)} = o_P(n^{-1/4})$ if and only if $2m > d$. When $2m \le d$ it is better to either use some dimension reduction modeling techniques (such as additive models) or to use nonlinear sieves in purely nonparametric estimation of g_0 to achieve a convergence rate of $\|\widehat{g} - g_0\|_{L^2(X)} = o_P(n^{-1/4})$. For instance, a nonlinear sigmoid neural network sieve has a convergence rate of $\|\widehat{g} - g_0\|_{L^2(X)} = O_p([n/\log n]^{-(1+1/d)/[4(1+1/(2d))]}) = o_P(n^{-1/4})$ (see Chen and Shen 1998, Proposition 1), which is faster than the best rate achievable by any linear sieves whenever $2m \le d$.

Sieve WAD estimation based on Equation (2). Let $q_0(x) \equiv \log f(x)$ denote the log density of x. Then we could estimate $q_0(x)$ via the sieve maximum likelihood:

$$\widehat{q} = \arg\max_{q\in\mathcal{H}_n} \frac{1}{n}\sum_{i=1}^n \left[q(x_i) - \log\int_{\mathcal{X}} \exp q(z)dz\right].$$

Again, if $q_0(\cdot)$ belongs to a Sobolev space $W_2^m(\mathcal{X})$ or a Holder space $\Lambda^m(\mathcal{X})$, we could let $\mathcal{H}_n$ be a nonlinear sieve such as the artificial neural networks when $d \ge 2m$ (see, e.g., Chen and White 1999). When $d < 2m$ we could let $\mathcal{H}_n$ be a tensor product linear sieves, $\mathcal{H}_n = \{q : \mathcal{X} \to \mathbb{R}, q(x) = \Sigma_{j=1}^{k_n} a_j p_j(x) : \int_{\mathcal{X}} q(z)dz = 0, a_1, \ldots, a_{k_n} \in \mathbb{R}\}$, such as tensor product polynomial splines (see, e.g, Stone 1990). Let $\widehat{\log f}(x) = \widehat{q}(x) - \log\int_{\mathcal{X}} \exp\widehat{q}(z)dz$. Then the WAD parameter θ defined in Equation (2) can be estimated by the following sieve WAD estimator:

$$\widehat{\theta}_2 = -\frac{1}{n}\sum_{i=1}^n y_i\left(\frac{\partial}{\partial x}w(x_i) + w(x_i)\frac{\partial}{\partial x}\widehat{q}(x_i)\right). \tag{6}$$

We note that these two alternative sieve WAD estimators are linear in their respective nonparametric estimators of nuisance functions, and hence there is no bias due to nonlinearity. Moreover, unlike the kernel WAD estimator considered by CCJ, there is no trimming involved either so these sieve WAD estimators allow for wider class of weight functions $w()$ and the estimator (4) is extremely easy to compute.

By applying Lemma 5.1 of Newey (1994a) or Theorem 4.1 of Chen (2007),[1] the root-n asymptotic normality of these two sieve WAD estimators can be easily established under weak regularity conditions. For instance, Ai and Chen (2007, Example 2.1 and sec. 4.1) considered the sieve WAD estimator (4) when the conditional mean function $g_0(\cdot) = E[y|X = \cdot]$ might be potentially misspecified as a nonparametric additive form. Newey (1994a, Example 3 and Theorem 7.2) considered a linear sieve (series) estimation of average derivative parameter $E\left[\frac{\partial}{\partial x}g(x)\right]$. Moreover, Newey (1994a), Ai and Chen (2007), and others have shown how to consistently estimate the variance of a sieve semiparametric two-step estimator easily, while Newey (1994a) and Ackerberg, Chen, and Hahn (2012) provided a numerically equivalent way to compute standard errors of a large class of semiparametric two-step estimator when the first step nuisance functions are estimated via linear sieves. One additional benefit of using sieve estimation in the first step is that a cross-validated choice of sieve number of terms to get optimal mean squared error rate in the first step would typically lead to root-n asymptotic normality of the second step plug-in estimate of θ. See, for example, Newey (1994a) and Chen (2007).

The idea of removing nonlinearity bias completely by reexpressing the Euclidean parameter of interest as a linear functional of some nuisance functions is more broadly applicable. See, for example, Chen, Hong, and Tamer (2005), Chen, Hong, and Tarozzi (2008a,b), and Imbens and Wooldridge (2009) for the Euclidean parameters that could be expressed as either a nonlinear functional similar to Equation (3) or a linear functional similar to Equation (1) in nonclassical measurement error, missing data, program evaluation, and other settings.

3. ROOT-*N* ESTIMATION OF GENERAL NONLINEAR FUNCTIONALS

In some applications, there is no simple reparameterization that could convert a nonlinear functional of a nuisance function into a linear functional of another nuisance function. The insight of a stochastic quadratic expansion to characterize the nonlinearity bias suggested in this article should be widely applicable to other semiparametric estimators of nonlinear smooth functionals.

[1]Theorem 4.1 in Chen (2007) is a slight improvement of Theorem 2 in Chen, Linton, and Keilgom (2003).

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

This article proposes generalized jackknife to reduce nonlinearity bias, which, based on the Monte Carlo results, works quite well for kernel estimation of WAD. In principle, their jackknife bias correction idea is directly applicable to all other semiparametric nonlinear smooth functionals estimated via the kernel method in the first step. However, the generalized jackknife bias reduction needs additional choice of parameters (the vector valued **c** in this article).

This article proposes to compute the standard error of the bias-corrected kernel WAD estimator based on the asymptotic variance expression (Equation (12) in the article), which, based on the Monte Carlo results in the online appendix, seems have room for improvement. There are alternative consistent variance estimators that might have better finite sample performance: (a) a jackknife variance estimator (e.g., Shao and Wu (1989) and the references therein); (b) instead of computing a standard error based on the asymptotic variance expression, one could use a finite sample (or "fixing smoothing parameter") version such as in Newey (1994a,b), Ai and Chen (2007), Ackerberg, Chen, and Hahn (2012).

Instead of jackknife, bootstrap is another popular method to provide better finite sample approximation to estimators of smooth functionals in terms of both reducing bias and more accurate confidence sets. See, for example, Efron (1979), Mammen (1990), Horowitz (2003) and the references therein.

There is also a tradeoff between how smooth the functional is with respect to the nuisance function $f \in \mathcal{F}$ and how complex the function parameter space $\mathcal{F}$ is. See, for example, Shen (1997). If the functional is highly nonlinear but not very smooth or if the space $\mathcal{F}$ is too large (in terms of covering numbers, say), then at some point we would no longer be able to estimate the Euclidean parameter functional θ at a root-n rate. In the case of kernel WAD estimation, the nonlinear functional is smooth and this article presents clean necessary conditions on kernel bandwidth choice to ensure a root-n rate. Recently Li et al. (2011) considered quadratic expansion of a particular nonlinear functional allowing for slower than root-n case. I think the theoretical results developed in this article could be extended further to allow for slower than root-n estimated nonlinear functionals.

In summary, this article highlights the difficult issue of nonlinearity bias in semiparametric estimation of nonlinear functionals of nuisance functions estimated nonparametrically in the first step. The article makes significant progress in providing clever solutions to the nonlinearity bias issue in a class of widely used kernel WAD estimators. The Monte Carlo results of the article also call for additional research on exploring other solutions.

[Received April 2013. Revised July 2013.]

REFERENCES

Ackerberg, D., Chen, X., and Hahn, J. (2012), "A Practical Asymptotic Variance Estimator for Two-Step Semiparametric Estimators," *Review of Economics and Statistics*, 94, 482–498. [1263,1264]

Ai, C., and Chen, X. (2007), "Estimation of Possibly Misspecified Semiparametric Conditional Moment Restriction Models With Different Conditioning Variables," *Journal of Econometrics*, 141, 5–43. [1263,1264]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics* (Vol. 6B), eds. James J. Heckman and Edward E. Leamer, New York: Springer, pp. 5549–5632. [1263]

Chen, X., Hong, H., and Tamer, E. (2005), "Measurement Error Models With Auxiliary Data," *Review of Economic Studies*, 72, 343–366. [1263]

Chen, X., Hong, H., and Tarozzi, A. (2008a), "Semiparametric Efficiency in GMM Models With Auxiliary Data," *The Annals of Statistics*, 36, 808–843. [1263]

——— (2008b), "Semiparametric Efficiency in GMM Models of Nonclassical Measurement Errors, Missing Data and Treatment Effects," Cowles Foundation Discussion Paper d1644. [1263]

Chen, X., Linton, O., and van Keilegom, I. (2003), "Estimation of Semiparametric Models When the Criterion Function is not Smooth," *Econometrica*, 71, 1591–1608. [2]

Chen, X., and Shen, X. (1998), "Sieve Extremum Estimates for Weakly Dependent Data," *Econometrica*, 66, 289–314. [1263]

Chen, X., and White, H. (1999), "Improved Rates and Asymptotic Normality for Nonparametric Neural Network Estimators," *IEEE Transactions Information Theory*, 45, 682–691. [1263]

Efron, B. (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26. [1264]

Horowitz, J. L. (2003), "The Bootstrap," in *Handbook of Econometrics* (Vol. 5), eds. J. J. Heckmanand E. Leamer, North Holland: Elsevier Science B.V. [1264]

Huang, J. Z. (1998), "Projection Estimation in Multiple Regression With Application to Functional ANOVA Models," *The Annals of Statistics*, 26, 242–272. [1263]

Li, L., Tchetgen, E., van der Vaart, A., and Robins, J. (2011), "Higher Order Inference on a Treatment Effect Under Low Regularity Conditions," *Statistics and Probability Letters*, 81, 821–828. [1264]

Imbens, G., and Wooldridge, J. (2009), "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86. [1263]

Mammen, E. (1990), "Higher-Order Accuracy of Bootstrap for Smooth Functionals," Preprint SFB 123, University of Heidelberg. [1264]

Newey, W. K. (1994a), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382. [1263,1264]

——— (1994b), "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253. [1264]

——— (1997), "Convergence Rates and Asymptotic Normality for Series Estimators," *Journal of Econometrics*, 79, 147–168. [1263]

Shao, J., and Wu, C. F. J. (1989), "A General Theory for Jackknife Variance Estimation," *The Annals of Statistics*, 17, 1176–1197. [1264]

Shen, X. (1997), "On Methods of Sieves and Penalization," *The Annals of Statistics*, 25, 2555–2591. [1264]

Stone, C. J. (1990), "Large-Sample Inference for Log-Spline Models," *The Annals of Statistics*, 18, 717–741. [1263]

Rejoinder

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

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

We wish to thank our discussants Xiaohong Chen, Holger Dette, Enno Mammen, and Donglin Zeng for a very stimulating discussion of our article (Cattaneo, Crump, and Jansson, 2013a; CCJ, hereafter). We also acknowledge the fantastic work of Jun Liu, Xuming He, and Jin Sun in shaping this intellectual exchange. Participants at the 2013 JSM Meeting (*JASA* invited session) also provided useful comments.

Our discussants offered an array of insightful comments ranging from implementation issues to theoretical considerations. Our rejoinder is organized by topic to clarify the importance, overlap, and implications for present and future research of these comments.

1. BIAS REDUCTION AND VARIANCE INFLATION

The comments by Dette and Zeng both touch upon the relationship between generalized jackknifing and the use of higher-order kernels for the purpose of reducing bias. This is an important issue because, in conventional nonparametric problems, it is well known not only that higher-order kernels can reduce smoothing bias (provided enough smoothness of the underlying nonparametric function), but also that the method of generalized jackknifing generates a class of higher-order kernels. See, for example, Härdle (1989). An important finding in CCJ, however, is that the "equivalence" between higher-order kernels and generalized jackknifing breaks down when the nonlinearity bias, as opposed to the smoothing bias, of a semiparametric procedure is considered. Nonlinearity biases are potentially first-order biases arising in some semiparametric problems under "severe" undersmoothing (e.g., $h_n \to 0$ faster than usual), a situation where smoothing bias is less of a concern. (The smoothing bias is large when the bandwidth is "large".) Nevertheless, connections between higher-order kernels and generalized jackknifing could still be useful to better understand the features of a bias-corrected semiparametric estimator constructed using the generalized jackknifing.

To be more specific, and following Dette, suppose $X_1, \ldots, X_n$ is a random sample from a univariate continuous distribution with density $f(\cdot)$ and consider the problem of estimating the value of f at some point x. The classical density estimate is

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x), \qquad K_h(u) = \frac{1}{h}K\left(\frac{u}{h}\right),$$

where K is a symmetric density and h is a bandwidth. Dette compared this estimator with the (generalized) jackknife estimator

$$\tilde{f}_{\mathbf{c},h}(x) = \frac{c_2^2}{c_2^2 - c_1^2}\hat{f}_{c_1 h}(x) - \frac{c_1^2}{c_2^2 - c_1^2}\hat{f}_{c_2 h}(x),$$

where $\mathbf{c} = (c_1, c_2)' \in \mathbb{R}^2_{++}$ is a vector of distinct positive constants, in an attempt to gain further intuition on the properties of $\hat{\boldsymbol{\theta}}_n(\mathbf{H}_n)$ and $\tilde{\boldsymbol{\theta}}_n(\mathbf{H}_n, \mathbf{c})$. It is argued that, although $\tilde{f}_{\mathbf{c},h}(x)$ has (smoothing) bias of smaller order than $\hat{f}_h(x)$, this reduction in bias typically comes at the expense of an increase in variance. In addition, the problem of choosing an "optimal" value of $\mathbf{c}$ is complicated by the fact that the (approximate) variance of $\tilde{f}_{\mathbf{c},h}(x)$ can be made arbitrarily small by increasing $\mathbf{c}$. For further discussion on these and related points see, for example, Jones and Foster (1993).

Indeed, defining $\tilde{h} = c_1 h$ and $\tilde{c} = c_2/c_1$, the estimator $\tilde{f}_{\mathbf{c},h}(x)$ can be written as

$$\tilde{f}_{\mathbf{c},h}(x) = \frac{1}{n}\sum_{i=1}^{n} \tilde{K}_{\tilde{c},\tilde{h}}(X_i - x),$$

$$\tilde{K}_{\tilde{c},\tilde{h}}(u) = K_{\tilde{h}}(u) + \frac{1}{\tilde{c}^2 - 1}[K_{\tilde{h}}(u) - K_{\tilde{c}\tilde{h}}(u)].$$

Thus, $\tilde{f}_{\mathbf{c},h}(x)$ can itself be interpreted as a kernel density estimator based on the kernel $\tilde{K}_{\tilde{c},\tilde{h}}$, which in turn can be thought of as a higher-order kernel obtained by means of a modification (indexed by $\tilde{c}$) of $K_{\tilde{h}}(\cdot)$. Because the modified kernel $\tilde{K}_{\tilde{c},\tilde{h}}(\cdot)$ is a higher-order kernel, estimators based upon it will "usually" have larger variance than estimators based on $K_{\tilde{h}}(\cdot)$. Interpreting $\tilde{f}_{\mathbf{c},h}(x)$ as a kernel estimator based on a higher-order kernel therefore provides an alternative explanation for Dette's observation that "usually" the variance of $\tilde{f}_{\mathbf{c},h}(x)$ exceeds that of $\hat{f}_h(x)$.

Furthermore, the reparameterization $(\mathbf{c}', h) \to (\tilde{c}, \tilde{h}) = (c_1/c_2, c_1 h)$ employed above also sheds light on Dette's observation about the difficulty of characterizing an "optimal" value of $\mathbf{c}$. In particular, the fact that $\tilde{h} = c_1 h$ can be thought of as the "effective" bandwidth of the kernel estimator based on $\tilde{K}_{\tilde{c},\tilde{h}}$ explains why an increase in $\mathbf{c}$ gives you "something for nothing" in the sense that it decreases the (approximate) variance of the generalized bandwidth estimator without affecting the order of magnitude of its bias.

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, University of California, Berkeley, 530 Evans Hall #3880, Berkeley, CA 94720-3880 (E-mail: *mjansson@econ.berkeley.edu*) and *CREATES*. 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).

© 2013 American Statistical Association
Journal of the American Statistical Association
December 2013, Vol. 108, No. 504, Theory and Methods
DOI: 10.1080/01621459.2013.856717

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

In addition to providing an alternative explanation for the findings of Dette, recognizing generalized jackknifing as a special case of employing a higher-order kernel when estimating the value of a density at a point is useful for the purpose of comparing that problem with the one addressed in our article. Zeng also offered some insightful comments about asymptotic (smoothing) bias reduction in general and about the relationship between generalized jackknifing and the use of higher-order kernels in particular.

All in all, three main points are highlighted in the discussions: (1) because generalized jackknifing is just like using a higher-order kernel one could think of using higher-order kernels more generically, (2) implementing generalized jackknife estimators requires choosing particular constants (e.g., $\mathbf{c}$) which is challenging in practice, and (3) generalized jackknifing will typically increase (higher-order) variance.

The main points above employ ideas from the nonparametric literature, and naturally apply to many problems where the concern is about smoothing bias (e.g., "large" bandwidths) as opposed to the nonlinearity bias (e.g., "small" bandwidths). In fact, many (but not all) linear functionals of a kernel estimator will not even have a nonlinearity bias (e.g., estimation of a density or regression function at a point). However, as shown in CCJ, not all of those ideas automatically apply when the object of interest is the nonlinearity bias, which naturally arises in the context of many nonlinear functionals of a kernel estimator. The weighted average derivative estimator studied in CCJ is just one example of a nonlinear functional of its nonparametric (kernel-based) ingredient. This distinction has two main implications. First, it implies that our generalized jackknife estimator cannot be interpreted as one based on a single higher-order kernel-based estimator. If anything, generalize jackknifing is altering the shape of the estimating equation and not of the kernel employed in the nonparametric estimator. Second, and perhaps more importantly, it implies that the bias problem addressed in the article cannot be solved simply by increasing the order of the kernel. Thus, point (1) above does not extend to the semiparametric problems considered in our article. On the other hand, points (2) and (3) above continue to be true insofar, first, it seems hard to propose a general selection rule for the constant $\mathbf{c}$ (see the discussion of Zeng for one such proposal) and, second, our generalized jackknife estimator is likely to have a larger finite-sample variance (our simulations provide supporting numerical evidence), although this variance inflation disappears asymptotically. The latter point implies that second-order efficiency considerations may be important, as mentioned by Dette.

2. THE ROLE OF NONLINEARITIES AND THE METHOD OF SIEVES

The main goal of our article was to highlight, in the context of semiparametrics, the presence of a potentially first-order bias arising from severe undersmoothing (i.e., for "small" bandwidths, $h_n \to 0$ faster than usual). Although the results in CCJ are obtained for a particular functional of a particular type of nonparametric estimator (namely, a kernel estimator), the consequences of nonlinearities in the estimating equation emphasized in our article will be shared also by other, but not all, semiparametric estimators based on the method of sieves. The comments of Chen and Mammen are both related to this point. As we further discuss in this section, we highlight that the presence and implications of the nonlinearity bias are crucially related to *both* the form of the estimating equation and the choice of nonparametric estimator (kernel-based, series-based, etc.). Furthermore, it appears difficult to separate the role of each of these two features of the semiparametric estimator. In other words, we can find "linear" and "nonlinear" population estimating equations that, when employed to construct semiparametric estimators using either kernels or sieves, will lead to estimators that may or may not exhibit a nonlinearity bias.

More specifically, Chen observed that while our chosen estimator can be motivated by the representation

$$\boldsymbol{\theta} = -\mathbb{E}\left[y \left(\frac{\partial}{\partial \mathbf{x}} w(\mathbf{x}) + w(\mathbf{x}) \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} / f(\mathbf{x}) \right) \right], \tag{C3}$$

sieve-based alternative estimators can be motivated by writing $\boldsymbol{\theta}$ as

$$\boldsymbol{\theta} = \mathbb{E}\left[w(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}} g(\mathbf{x}) \right], \qquad g(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}], \tag{C2}$$

or

$$\boldsymbol{\theta} = -\mathbb{E}\left[y \left(\frac{\partial}{\partial \mathbf{x}} w(\mathbf{x}) + w(\mathbf{x}) \frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}) \right) \right], \quad L(\mathbf{x}) = \log f(\mathbf{x}). \tag{C1}$$

As remarked by Chen, (1) the representations in (C1) and (C2) are linear in the nuisance functions $g(\cdot)$ and $L(\cdot)$, respectively, and (2) the nuisance functions $g(\cdot)$ and $L(\cdot)$ can be estimated using the method of sieves.

For estimators based on kernels, the relevant issue (from the perspective of our article) is not only whether the functional can be represented as a linear functional of some nuisance function that can be estimated using a kernel-based method. For instance, if $\hat{f}(\cdot)$ is a kernel estimator of $f(\cdot)$, then $\hat{L}(\cdot) = \log \hat{f}(\cdot)$ is a kernel-based estimator of $L(\cdot)$ in (C2), but of course the estimator based on evaluating the sample analog of (C2) at $L(\cdot) = \hat{L}(\cdot)$ is equivalent to our estimator based on (C3). Thus, at least in the case of kernels, the nuisance function has to be of the "right form" for it to be valuable to express the estimand as a linear functional thereof. As another example of the same point, consider the estimand $\theta = \mathbb{E}[f(\mathbf{x})] = \int_{\mathbb{R}^d} f(\mathbf{x})^2 \mathrm{d}\mathbf{x}$, and the associated plug-in kernel-based sample analogue estimators:

$$\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} \hat{f}(\mathbf{x}_i) \qquad \text{and} \qquad \hat{\theta}_2 = \int_{\mathbb{R}^d} \hat{f}(\mathbf{x})^2 \mathrm{d}\mathbf{x},$$

where $\hat{f}(\mathbf{x})$ is a classical kernel-based density estimator. Both of the estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ will exhibit leave-in bias and, furthermore, $\hat{\theta}_2$ will also exhibit nonlinearity bias. Therefore, it should be clear that studying the shape of the estimating equation alone is not enough to understand whether the semiparametric estimator will exhibit either leave-in-bias, nonlinearity bias, or both, at least when kernel-based estimators are employed. Indeed, in the case of kernels the relevant issue seems to be whether the estimand can be written as a linear functional of a nuisance function expressible as a density-weighted conditional expectation; that is, the nuisance function should be of the

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013

form $\gamma(\mathbf{x}) = \mathbb{E}[\mathbf{w}|\mathbf{x}]f(\mathbf{x})$, where $\mathbf{w}$ is some (possibly constant) observed variable.

We conjecture that similar remarks apply to estimators based on the method of sieves; that is, we suspect that also estimators based on the method of sieves can suffer from nonlinearity biases unless the estimand can be expressed as a linear functional of a nuisance function of the "right type." For sieve least-squares estimators, such as the estimator of $g(\cdot)$ in (C1) mentioned by Chen, it would appear that nuisance functions are of the "right type" when they are expressible as mean square projections (e.g., as a conditional expectation). Accordingly, we agree that it seems plausible that nonlinearity biases of the form highlighted by the article can be avoided by using the (least-squares) sieve-based estimator motivated by (C1). More generally, although we feel that more work is needed to understand the circumstances in which also nonlinear sieve estimators can be plugged into linear functionals without generating biases, we agree wholeheartedly with what we believe is the main message of Chen's comment: rather than basing the choice of nonparametric estimation method mainly on the ease of implementation one should pay careful attention to whether the nuisance function (estimator) can be chosen in such a way that the object of interest is a linear functional thereof.

As discussed in the article, the estimator we consider suffers from two distinct types of bias, namely nonlinearity bias and leave-in bias. Both biases are (of the same order of magnitude and) asymptotically nonnegligible only when the rate of convergence of the nonparametric ingredient is slower than $n^{1/4}$. Therefore, it is necessary to relax (among other assumptions) the assumption of $n^{1/4}$-consistency on the part of the nonparametric ingredient to uncover and characterize these biases. The extent to which this feature is shared by estimators based on the method of sieves would appear to be an open question. For instance, although we agree with Chen that analyzing sieve weighted average derivative estimators is easy once conventional assumptions such as $n^{1/4}$-consistency have been made, existing results such as Theorem 4.1 of Chen (2007) are silent about the consequences of employing severely undersmoothed nonparametric estimators (e.g., sieve estimators implemented using a larger-than-usual value of the tuning parameter k_n) when estimating finite-dimensional parameters. In particular, even if nonlinearity biases can be avoided by relying on the method of sieves, it would appear to be an open question whether any of the estimators proposed by Chen suffers from an analog of the leave-in bias discussed in the article.

Conversely to the discussion given so far, we also know of the existence of "nonlinear" estimands that lead to series-based estimators that do not exhibit either leave-in bias or nonlinearity bias. Specifically, the estimand of the parametric part of the partially linear model $y_i = \mathbf{x}_i'\boldsymbol{\beta} + g(\mathbf{z}_i) + \varepsilon_i$, with $\mathbb{E}[\varepsilon_i|\mathbf{z}_i, \mathbf{x}_i] = 0$ and other assumptions imposed, is given by

$$\boldsymbol{\beta} = (\mathbb{E}[(\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i|\mathbf{z}_i])\mathbf{x}_i'])^{-1}\mathbb{E}\left[(\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i|\mathbf{z}_i])y_i\right],$$

which could be regarded as a nonlinear estimating equation (i.e., the nuisance function $h(\mathbf{z}_i) = \mathbb{E}[\mathbf{x}_i|\mathbf{z}_i]$ enters nonlinearly). Nonetheless, Cattaneo, Jansson, and Newey (2012) showed that when $h(\cdot)$ is estimated by the method of linear sieves the resulting semiparametric estimator $\hat{\boldsymbol{\beta}}$ does not exhibit leave-in or nonlinearity biases. Furthermore, to make things more interesting, if undersmoothing is sufficiently severe (i.e., $K/n \to \alpha \in (0, 1)$), the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ exhibits a different, larger asymptotic variance instead of a bias, very much in line with the findings documented in Cattaneo, Crump, and Jansson (2010, 2013b) for a class of "linear" kernel-based semiparametric estimators.

For these reasons, we are currently developing distributional results for sieve-based semiparametric estimators under assumptions that permit (but do not necessarily require) the complexity of the sieve space to grow relatively rapidly with the sample size. Although doing so will require a possibly nontrivial relaxation of the methods used when establishing results such as Theorem 4.1 of Chen (2007), the comments of Mammen strongly suggest that, at least in some cases, significant progress toward a better theory-based understanding of the small-sample properties of sieve-based estimators is possible. We are very grateful to Mammen for not only clarifying the relationship between our work and his but, most importantly, for helping to place the work in a broader context and for providing a template for analyzing sieve-based estimators under weaker-than-usual assumptions about complexity of the sieve space.

3. THE ROLE OF DIMENSIONALITY AND BOOTSTRAPPING

The discussants raised a number of additional points. We found little to disagree with and would like to take this opportunity to thank the discussants for the numerous constructive suggestions. Among those, we would like to highlight two, one mainly conceptual and the other both theoretical and implementational. First, as pointed out by Mammen, our nonstandard asymptotics and the resulting biases in the distributional approximation also highlight an interesting role of the dimensionality of covariates, $\mathbf{x} \in \mathbb{R}^d$. In the context of kernel-based estimators, our article suggests that the larger d, the more important the nonlinearity and leave-in bias will be. As pointed out by Mammen, his work is closely related to this point insofar as nonlinear least-squares models with large-/high-dimensional covariates may also exhibit potentially first-order biases very similar in spirit, but different in form, from those we found in our work. It would certainly be of interest to deepen our understanding of these seemingly unrelated findings.

Second, as suggested by Mammen's comment, the idea of studying the properties of the bootstrap under the types of assumptions entertained in CCJ seems particularly interesting and promising. Despite the fact that severe undersmoothing of certain "linear" semiparametric estimators leads to invalidity of the bootstrap (Cattaneo, Crump, and Jansson 2014), in research currently under way we have addressed that very question and found that the bootstrap provides a method of (variance estimation and) bias correction that is valid under the assumptions made in CCJ. That is, we have shown that the bootstrap is indeed able to remove both nonlinearity and leave-in biases. Our current research is also extending the scope of this finding to a large class of possibly nonsmooth, nondifferentiable two-step semiparametric models.

REFERENCES

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. [1267]

——— (2013a), "Generalized Jackknife Estimators of Weighted Average Derivatives," *Journal of the American Statistical Association*, 108, 1243–1256. [1265]

——— (2013b), "Small Bandwidth Asymptotics for Density-Weighted Average Derivatives," *Econometric Theory, forthcoming*. [1267]

——— (2014), "Bootstrapping Density-Weighted Average Derivatives," *Econometric Theory, forthcoming*. [1267]

Cattaneo, M. D., Jansson, M., and Newey, W. K. (2012), "Alternative Asymptotics and the Partially Linear Model With Many Regressors," *Working paper, University of Michigan*. [1267]

Chen, X. (2007), "Large Sample Sieve Estimation of Semi-Nonparametric Models," in *Handbook of Econometrics, Volume VI*, eds. J. J. Heckmanand E. Leamer, New York: Elsevier Science B.V. [1267]

Härdle, W. (1989), *Applied Nonparametric Regression*, New York: Cambridge University Press. [1265]

Jones, M. C., and Foster, P. J. (1993), "Generalized Jackknifing and Higher Order Kernels," *Journal of Nonparametric Statistics*, 3, 81–94. [1265]

Downloaded by [University of California, Berkeley] at 13:56 19 December 2013