Comment

Donglin ZENG

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*).

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

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*).

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

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*).

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

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: Dimension Asymptotics and Semiparametrics

Enno MAMMEN

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.

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

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

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.

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

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]

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.

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

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]

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.

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

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).

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

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).

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

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

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]