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]