1. ASYMPTOTIC BIAS REDUCTION

Under a weak assumption on the bandwidth, this work handles bias reduction via a second-order linearization of \( \hat{\theta}_n(\mathbf{H}_n) \) 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 \( |l| \leq P \). Then the asymptotic bias in \( \hat{\theta}_n(\mathbf{H}_n) \) 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 \( O(H^{p+1}) \). 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 \( \hat{f}(x) \) 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, \( \theta = E[w(x)\nabla g(x)] \) 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 \( \hat{\hat{\theta}}_n(\mathbf{H}_n, c) \), one has to determine the order \( J \) (satisfying \( J < 1 + d/2 \) and \( J \geq (d - 2)/8 \)) so that

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

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:

\[
\hat{\theta}(c\mathbf{H}_n) = \theta + c^{-d}(b_0 + b_1c^2 + \ldots + b_d c^{2J}) + \epsilon,
\]

where \( \epsilon \) is a stochastic term with mean zero and variance of order \( n^{-1/2} \) and \( J < 1 + d/2 \). However, since \( \hat{\theta}(c\mathbf{H}_n) \) 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 \( \hat{\theta}(c_k \mathbf{H}_n) \) using the \( k \)th data and denote it by \( \hat{\theta}_k \). Then, the above regression model implies

\[
\hat{\theta}_k = \theta + c_k^{-d}(b_0 + b_1c_k^2 + \ldots + b_d c_k^{2J}) + \epsilon_k,
\]

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 \( \hat{\theta}_k \) 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 \( \theta \) after \( J \) is chosen;
3. we estimate \( \Sigma \) 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...
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, \( \hat{\psi}_n(x) - \psi_n(x) = n^{-1/2}G_n[k_{H_n}(x - X)Y] \), where \( G_n \) denotes the empirical process. Consider the class of functions \( \mathcal{F} = \{ k_{H_n}(x - X)Y : x \in \chi_n \} \). From Assumption B2, we note

\[
|k_{H_n}(x - X)Y - k_{H_n}(x^* - X)Y| \\leq \|x - x^*\| \sup_x k^*(H_n^{-1}x)|H_n|^{-2}.
\]

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

\[
\int_0^1 \sqrt{1 + \log N_1(1, F)} |F| \cdot \|F\|_{L^1(P)} d\epsilon < \infty.
\]

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

\[
\|\sup_{\mathcal{F}} |G_n|\| = O_p \left( \|F\|_{L^1(P)} \right) = O_p \left( H_n^{-1} \right).
\]

This gives

\[
\sup_{x \in \chi_n} |\hat{\psi}_n(x) - \psi_n(x)| = O_p \left( \frac{1}{\sqrt{nH_n}} \right).
\]

5. EXTENSION TO MORE GENERAL SEMIPARAMETRIC ESTIMATION

The same technique can be applied to a more general semiparametric estimation where the parameter of interest, \( \theta \), implicitly solves an estimating function \( E[g(\theta, f, f', f'', \ldots)] = 0 \), 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


Comment

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 \( c \) for two reasons:

1. I do not think that there exists an optimal choice of the weights \( c \) 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

\[
\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^n K \left( \frac{X_i - x}{h} \right).
\]
If \( f \) is twice differentiable and the kernel \( K \) is symmetric, then the bias of this estimate is given by
\[
\mathbb{E}[\hat{f}_h(x)] = \frac{h^2 f''(x)}{2} + o(h^2).
\] (2)

Similarly, the variance is obtained as
\[
\text{var}(\hat{f}_h(x)) = \frac{f(x)}{nh} \int K^2(u)du \cdot (1 + o(1)).
\] (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

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

where the weights \( w_1, w_2 \) are determined such that \( w_1 + w_2 = 1 \) and the dominating term in
\[
\mathbb{E}[\hat{g}_{a,c_2}(x)] = (w_1 c_1^2 + w_2 c_2^2) \frac{h^2 f''(x)}{2} + o(h^2)
\]
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 \( \mathbb{E}[\hat{g}_{a,c_2}(x)] = o(h^2) \). Now we investigate the variance of the estimator \( \hat{g}_{a,c_2}(x) \), that is,
\[
\text{var}(\hat{g}_{a,c_2}(x)) = w_1^2 \text{var}(\hat{f}_{c_1,h}(x)) + w_2^2 \text{var}(\hat{f}_{c_2,h}(x)) + 2w_1w_2 \text{cov}(\hat{f}_{c_1,h}(x), \hat{f}_{c_2,h}(x)).
\]

A standard calculation yields
\[
\text{cov}(\hat{f}_{c_1,h}(x), \hat{f}_{c_2,h}(x)) = \frac{f(x)}{nc_2} \int K(u) K \left( \frac{c_1}{c_2} u \right) du (1 + o(1)),
\]
and we obtain
\[
\text{var}(\hat{g}_{a,c_2}(x)) \geq \left\{ \left( \frac{w_1^2}{c_1^2} + \frac{w_2^2}{c_2^2} \right) \frac{f(x)}{nh} \int K^2(u)du \right. \\
\left. + 2w_1w_2 \frac{f(x)}{nc_2} \left( \int K^2(u)du \int K^2 \left( \frac{c_1}{c_2} u \right) du \right)^{1/2} \right\} \\
\times (1 + o(1)),
\]
where we used the Cauchy Schwarz inequality and the fact that \( w_1, w_2 \leq 0 \). Finally, a substitution in the integral \( \int K^2(u)du \) and a simple calculation gives
\[
\text{var}(\hat{g}_{a,c_2}(x)) \geq \alpha^2(c_1, c_2) \frac{f(x)}{nh} \int K^2(u)du(1 + o(1))
\]
\[
= \alpha^2(c_1, c_2) \text{var}(\hat{f}_h(x))(1 + o(1))
\]
(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
\[
\alpha^2(c_1, c_2) := \left( \frac{w_1}{\sqrt{c_1}} + \frac{w_2}{\sqrt{c_2}} \right)^2.
\] (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 \( \alpha^2 \) for various choices of \( c_1 \) and \( c_2 \) and make the following observations:

1. For reasonable choices of \( c_1 \) and \( c_2 \), the factor \( \alpha^2 \) 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 \( \hat{g}_{a,c_2}(x) \)).

2. For increasing values of \( c_1, c_2 \rightarrow \infty \), the first-order approximation for the variance of \( \hat{g}_{a,c_2}(x) \) 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
\[
K_2(x) = \frac{15}{32}(1 - x^2)^3(3 - 7x^2)I_{[-1,1]}(x)
\]
is of order 4 and yields a bias of order \( O(h^4) \). However, we obtain for the corresponding terms in the variance
\[
\int K_1^2(x)dx = \frac{3}{5}, \quad \int K_2^2(x)dx = \frac{5}{4},
\]
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.
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 loglinear 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 Sauer (1979) 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...
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 \( Y_i = X_i^\top \beta + \epsilon_i \) with deterministic covariates \( X_i \in \mathbb{R}^p \) and iid errors with \( \mathbb{E}[\psi(\epsilon_i)] = 0 \) for a function \( \psi : \mathbb{R} \to \mathbb{R} \). Consider an M-estimator \( \hat{\beta}_n \) with M-function \( \psi : \)

\[
\sum_{i=1}^n X_i \psi(Y_i - X_i^\top \hat{\beta}_n) = 0.
\]

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 \( \beta = \text{trace} \left[ \sum_{i=1}^n X_i X_i^\top \right] = \text{trace} \left[ \sum_{i=1}^n X_i^\top X_i \right] = \sum_{i=1}^n \|X_i\|^2 \). For simplicity, we make the assumption that the design vectors are of the same order of size, in the sense that \( \max_{1 \leq i \leq n} \|X_i\|^2 = O(p/n) \). 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 \( \rho_0 = \mathbb{E}[\psi^2(\epsilon_i)] \) and \( \rho_1 = \mathbb{E}[\psi(\epsilon_i)] \). 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 \( 0 \approx \sum_{i=1}^n X_i \psi(\epsilon_i) - \sum_{i=1}^n X_i^\top \psi(\epsilon_i) (\hat{\beta}_n - \beta) + (1/2) \sum_{i=1}^n X_i (\hat{\beta}_n - \beta)^2 \psi''(\epsilon_i) \). This gives with \( \rho_2 = \mathbb{E}[\psi''(\epsilon_i)] \), \( \rho_3 = \mathbb{E}[\psi(\epsilon_i)[\psi'(\epsilon_i) - \rho_1]] \) and \( \psi(x) = \psi(x) - \rho_1 \nabla \psi(x) + \rho_2 \rho_3 \sum_{i=1}^n X_i \psi(\epsilon_i) + \rho_1 \rho_3 \sum_{i=1}^n X_i \|X_i\|^2 \).

Under appropriate conditions, this expansion is valid with rest terms of order \( p^{3/2} \log(n)/n \). 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 \( b_n = \rho_1^{-1/2} \rho_0 - \rho_1 \rho_3 \sum_{i=1}^n X_i \|X_i\|^2 \). 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

\[
|e^\top b_n| \leq Cn^{1/2} \left[ \sum_{i=1}^n (e^\top X_i \|X_i\|^2)^2 \right]^{1/2} \leq n^{1/2} \max_{1 \leq i \leq n} \|X_i\|^2 \left[ \sum_{i=1}^n (e^\top X_i)^2 \right]^{1/2} = O(pn^{-1/2})
\]

because of \( \sum_{i=1}^n X_i X_i^\top = I_p \) and \( \max_{1 \leq i \leq n} \|X_i\|^2 = O(p/n) \).

One can write \( X_i^\top \beta = X_i^\top \beta_1 + X_i^\top \beta_{1-1} \), where \( \beta_1 \) is the first element of \( \beta \) and \( \beta_{1-1} \) contains the remaining elements of \( \beta \). If \( \beta_1 \) is 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 \( p^{3/2} \log(n)/n \to 0 \), 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. that for \( \hat{\beta}_{n,1} - \beta_1 - b_{n,1} \) the same stochastic expansion

\[
\rho_1^{-1} \sum_{i=1}^n X_i \psi(\epsilon_i) \]

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 \( \hat{\theta}_{n,1} - \beta_1 \) with \( \sqrt{n} \hat{\beta}_1(H_n) - \beta_1 \). The dimension \( p \) of the linear model corresponds to \( (h_1, \ldots, h_2)^\top \)\( -1 \)\( = |H_n|^{-1} \). With this relation, we get from part (a) of Theorem 2 that the bias terms \( \beta_{n,1} \) and \( \hat{\theta}_n(H_n) \) 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 \( p^{3/2} \log(n)/n \) 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 \( \beta \). The first term in the stochastic expansion (2) of \( \hat{\beta}_n - \beta \) now becomes

\[
\sum_{i=1}^n X_i X_i^\top \mathbb{E}[\psi(\epsilon_i - X_i^\top b_n)]^2 \sum_{i=1}^n X_i \psi(\epsilon_i - X_i^\top b_n).
\]

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


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. In these applications, we suspect that a semiparametric two-step estimator of θ based on a nonparametric estimation of another estimator of θ in the first step typically performs better in finite sample than another estimator of θ based on a nonparametric sieve estimation 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:

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

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

\[
= -E \left[ y \left( \frac{\partial}{\partial x} w(x) + w(x) \frac{\partial f(x)}{\partial x}/f(x) \right) \right]. \tag{3}
\]
where \( f() \) is the density of the regressor \( x \) and \( g() \) is the conditional mean function of \( y \) given \( x \). It is clear that \( \theta \) 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)). CCIJ considers estimation of \( \theta \) based on Equation (3). Alternatively, one could estimate \( \theta \) based on either Equation (1) or Equation (2).

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

\[
\hat{\theta}_1 = \frac{1}{n} \sum_{i=1}^{n} w(x_i) \frac{\partial}{\partial x_i} \hat{g}(x_i), \tag{4}
\]

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 \( X \) to which \( g \) belongs. For a typical function space such as a Sobolev space \( W^m_r(X) \) or a Holder space \( \Lambda^m(X) \), \( X \) a subset in \( \mathbb{R}^d \), we typically obtain \( \| \hat{g} - g_0 \|_{L^2(X)} = O_P(n^{-1/4}) \) for tensor product linear sieves (or series), where the series LS estimator \( \hat{g} \) has a closed-form expression:

\[
\hat{g}(x) = \sum_{j=1}^{n} p^{b_j}(x)'(P'P)^{-1} p^{b_j}(X)_i Y_i, \quad x \in X, \tag{5}
\]

where \( \{ p_j(\cdot), j = 1, 2, \ldots \} \) denotes a sequence of known basis functions that can approximate any square integrable functions of \( x \) well, \( p^{b_j}(X) = (p_1(X), \ldots, p_b(X))^t, P = (p^{b_1}(X), \ldots, p^{b_b}(X))^t \) and \( (P'P)^{-1} \) 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 \( \| \hat{g} - g_0 \|_{L^2(X)} = o_p(n^{-1/4}) \) if and only if \( 2m > d \). When \( 2m \leq d \) 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 \( \| \hat{g} - g_0 \|_{L^2(X)} = o_P(n^{-1/4}) \). For instance, a nonlinear sigmoid neural network sieve has a convergence rate of \( \| \hat{g} - g_0 \|_{L^2(X)} = O_P(n^{-1/4}) \) (see Chen and Shen 1998, Proposition 1), which is faster than the best rate achievable by any linear sieves whenever \( 2m \leq d \).

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

\[
\hat{q} = \text{arg max}_{q \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \left[ q(x_i) - \log \int_X \exp q(z) dz \right].
\]

Again, if \( q_0(\cdot) \) belongs to a Sobolev space \( W^m_r(X) \) or a Holder space \( \Lambda^m(X) \), we could let \( \mathcal{H}_n \) be a nonlinear sieve such as the artificial neural networks when \( d \geq 2m \) (see, e.g., Chen and White 1999). When \( d < 2m \) we could let \( \mathcal{H}_n \) be a tensor product linear sieves, \( \mathcal{H}_n = \{ q : X \to \mathbb{R}, q(x) = \sum_{j=1}^{n} a_j q_j(x) \} \), such as tensor product polynomial splines (see, e.g., Stone 1990). Let \( \log f(x) = \hat{q}(x) - \log \int_X \exp \hat{q}(z) dz \). Then the WAD parameter \( \theta \) defined in Equation (2) can be estimated by the following sieve WAD estimator:

\[
\hat{\theta}_2 = -\frac{1}{n} \sum_{i=1}^{n} y_i \left( \frac{\partial}{\partial x_i} w(x_i) + w(x_i) \frac{\partial}{\partial x_i} \hat{g}(x_i) \right). \tag{6}
\]

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 CCIJ, 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 \( E \left[ \frac{\partial}{\partial g(x)} \right] \). 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 Ackermann, 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 \( \theta \). 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 Keilgorn (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 e 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 \( \theta \) 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 clearer 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


