

Twicing Kernels and a Small Bias Property of Semiparametric Estimators*

Whitney K. Newey
Department of Economics
M.I.T.

Fushing Hsieh
Department of Statistics
UC Berkeley

James Robins
School of Public Health
Harvard University

First Draft, 1991.
This draft, March 2002.

JEL Classification:

Keywords: Semiparametric estimation, twicing kernels, bias, idempotent transformations.

*This is a revised version of a paper that was formerly titled Undersmoothing and Bias Corrected Functional Estimation. The NSF provided partial financial support for this work. Helpful comments were provided by P. Bickel, A. Monfort, Y. Ritov, and three referees. Ziad Nejmeldien provided capable research assistance.

1 Introduction

There are now many important econometric estimators that depend on nonparametric regression or density estimates, including average derivatives, (Powell, Stock, and Stoker, 1989) and nonparametric consumer surplus (Hausman and Newey, 1995). The purpose of this note is to show that particular nonparametric estimates give semiparametric estimators with a small bias property (SBP), that the bias converges to zero faster than the bias of the nonparametric estimate. Estimators with the SBP have the theoretical advantage that their bias shrinks to zero at a faster rate. We find in some examples that this theoretical advantage can lead to significant small sample improvements in mean-square error.

Kernel estimators have been popular and are the standard method for important estimators that depend on densities, such as weighted average derivatives, Powell, Stock, and Stoker (1989) and inverse density weighted estimators, e.g. Ruud (1986) and Lewbel (1998). We find that a twicing kernel (defined below) gives the SBP. In particular, we show that the bias of a twicing kernel average derivative has smaller order than the nonparametric bias. We also show that twicing kernels can give \sqrt{n} -consistency of a semiparametric m-estimator without requiring that \sqrt{n} times that nonparametric bias goes to zero. In addition, we find that the SBP is present when nonparametric estimation does not affect the asymptotic variance, by showing \sqrt{n} -consistency with a standard kernel without requiring that \sqrt{n} times that nonparametric bias goes to zero.

It is known that other kinds of nonparametric estimators also confer the SBP. The Bickel and Ritov (1988) split sample estimator of the average density has the SBP. Newey (1994) showed that series nonparametric regression confers the SBP and Shen (1997) that sieve maximum likelihood estimators do also. To help explain why the SBP property holds for series and sieve estimators, as well as for twicing kernels, we interpret these estimators as idempotent transformations of the empirical distribution and show that idempotent linear transformations of the empirical distribution confer the SBP.

Throughout this note we focus on simple settings to aid in understanding the SBP

and its affect on the performance of semiparametric estimators. We leave to other work the formulation and proof of more general results. Such results might include the SBP being present for any estimator where the nonparametric estimate does not affect the asymptotic variance, the SBP being conferred by any nonparametric estimator that is an idempotent transformation of the empirical distribution, and a full stochastic expansion and higher-order mean-square error derivation for kernel based nonlinear semiparametric estimators.

One distinguishing feature of the SBP is that undersmoothing may not be needed for \sqrt{n} -consistency. That is, the nonparametric estimator may both converge at its optimal nonparametric rate and yield a \sqrt{n} -consistent semiparametric estimator. Bickel and Ritov (2000) explore this property (calling it the plug in property, PIP for short), showing that certain smoothness restrictions are necessary for existence of such an estimator.¹ We find that the extent of the SBP, i.e. the amount by which the semiparametric bias is smaller than the nonparametric bias, depends on how smooth certain functions are.

In some settings there may be better bias reduction methods. Stoker (1993) provides nice results for estimation of average derivatives, showing that a particular instrumental variables estimator greatly reduces bias. The virtue of twicing kernels is that they confer the SBP on essentially any semiparametric estimator.

In Section 2 we consider twicing kernels and derive the mean-square error of kernel average derivatives. These results show that the twicing kernel confers the SBP, and help quantify the effect of using a twicing kernel on higher-order variance. We also show \sqrt{n} -consistency of a general semiparametric m-estimator, with the SBP resulting from use of a twicing kernel or from no effect on the asymptotic variance from nonparametric estimation. Section 3 explains the SBP as the result of an adjoint transformation. Section 4 reports some Monte Carlo results.

¹Early versions of our work, including Newey, Hsieh, and Robins (1991), were done independently of and before Bickel and Ritov (2000).

2 Twicing Kernels and Semiparametric Bias

A twicing kernel is one with the form

$$K(u) = 2k(u) - \int k(u-v)k(v)dv, \quad (2.1)$$

where $k(u)$ is a kernel function, satisfying $\int k(u)du = 1$. We consider nonparametric estimates of $\gamma_0(x) = f(x)E[w|x]$, where $f(x)$ is the density of an $r \times 1$ vector x of continuously distributed variables and y is a vector of random variables². A kernel estimator of $\gamma_0(x)$ is given by

$$\hat{\gamma}(x) = \sum_{j=1}^n K_h(x - x_j)w_j/n, K_h(u) = h^{-r}K(u/h),$$

for a bandwidth h .

To facilitate comparison with previous results we first consider linear kernel averages like those Powell and Stoker (1996). Define the "leave one out" estimator

$$\hat{\gamma}_{-i}(x) = \sum_{j \neq i} K_h(x - x_j)w_j/n.$$

Also, let λ denote an $r \times 1$ vector of nonnegative integers, $|\lambda| = \sum_{j=1}^r \lambda_j$, $x^\lambda = \prod_{j=1}^r (x_j)^{\lambda_j}$, and $\partial^\lambda \gamma(x) = \partial^{|\lambda|} \gamma(x) / \partial^{\lambda_1} x_1 \dots \partial^{\lambda_r} x_r$. For a random variable y and some $\bar{\lambda}$ we consider a scalar estimator

$$\hat{\beta} = \sum_{i=1}^n \partial^{\bar{\lambda}} \hat{\gamma}_{-i}(x_i) y_i / n. \quad (2.2)$$

The leave one out feature of this estimator makes our results comparable to those of Powell, Stock, and Stoker (1989) and Powell and Stoker (1996), and also leads to the MSE converging to zero at a faster rate than would be obtained if the own observation were included. This estimator includes several well known examples, such as the average density, where $\bar{\lambda} = 0$ and $y = w = 1$, and the density weighted average derivative, where λ is a unit vector and $w = -2$.

²The density $f(x)$ is a component of $\gamma_0(x)$ when the corresponding component of w is 1.

Some notation is useful for the mean square error comparisons. Let $\beta_0 = E[\partial^{\bar{\lambda}}\gamma_0(x)y]$, $\zeta_\lambda = \int k(u)u^\lambda du$, and

$$\begin{aligned}\nu(x) &= (-1)^{|\bar{\lambda}|}\partial^{\bar{\lambda}}\{E[y|x]f(x)\}, \\ \psi(z) &= \partial^{\bar{\lambda}}\gamma_0(x)y - \beta_0 + \nu(x)w - E[\nu(x)w], \\ V &= \text{Var}(\psi(z)), Q = \int \{E[w^2|x]E[y^2|x] + (-1)^{|\bar{\lambda}|}E[wy|x]^2\}f(x)^2 dx, \\ \bar{K} &= \int [\partial^{\bar{\lambda}}K(u)]^2 du, \bar{k} = \int [\partial^{\bar{\lambda}}k(u)]^2 du, \\ P &= \sum_{|\lambda|=s, |\bar{\lambda}|=s} \zeta_\lambda \zeta_{\bar{\lambda}} \int \partial^\lambda \gamma_0(x) \partial^{\bar{\lambda}} v(x) dx / (s!)^2, p = \sum_{|\lambda|=s} \zeta_\lambda \int v(x) \partial^\lambda \gamma_0(x) dx / (s!).\end{aligned}$$

We assume that all of the integrals in these expression exist, but for notational simplicity postpone the statement of regularity conditions until the appendix.

Powell and Stoker (1996) show, for a version $\tilde{\beta}$ of this estimator where the original kernel $k(u)$ is used rather than the twicing kernel $K(u)$, $\gamma_0(x)$ is s times differentiable, and other regularity conditions hold, that

$$MSE(\tilde{\beta}) = V/n + n^{-2}h^{-r-2|\bar{\lambda}|}\bar{k}Q + h^{2s}p^2 + o(n^{-1} + n^{-2}h^{-r-2|\bar{\lambda}|} + h^{2s}).$$

For the twicing kernel we have the following result, with some assumptions given in the appendix:

Theorem 1: *If Assumptions A1-A3 are satisfied, $\gamma_0(x)$ and $v(x)$ are s times differentiable, $\zeta_\lambda = 0$ for $0 < |\lambda| < s$, and $h \rightarrow 0$ as $n \rightarrow \infty$, then*

$$MSE(\hat{\beta}) = V/n + n^{-2}h^{-r-2|\bar{\lambda}|}\bar{K}Q + h^{4s}P^2 + o(n^{-1} + n^{-2}h^{-r-2|\bar{\lambda}|} + h^{4s}).$$

For both $\tilde{\beta}$ and $\hat{\beta}$ the dominant part of the MSE consists of three terms. The first is an asymptotic variance term, the second a higher-order variance term, and the third a squared bias term. The asymptotic variance term is the same for both but the higher order variance and bias terms differ. The squared bias term shows the SBP conferred by the twicing kernel, because it converges to zero with h faster than the nonparametric rate h^{2s} . The higher-order variance term will generally be larger for the twicing kernel,

with $\bar{K} > \bar{k}$. However, this increase in variance only comes through the constant term, so that for a range of bandwidths the MSE of the twicing kernel version will be smaller in large samples.³ This variance increase is analogous to that found by Kauermann, Muller, and Carroll (1997).

An interesting example is a density weighted average derivative estimator like that of Powell, Stock, and Stoker (1989), where $|\bar{\lambda}| = 1$ and $w = -2$. Suppose that $k(u)$ is a symmetric density (with $s = 2$), $E[y|x]$ and $f(x)$ are three times continuously differentiable, and the other conditions of Theorem 1 are satisfied. Then

$$MSE(\hat{\beta}) = V/n + n^{-2}h^{-r-2}\bar{K}Q + h^8P^2 + o(n^{-1} + n^{-2}h^{-r-2} + h^8).$$

For an appropriate choice of bandwidth this estimator will be \sqrt{n} -consistent for $r \leq 6$.⁴ In comparison with Powell, Stock, and Stoker (1989), \sqrt{n} -consistency requires fewer derivatives of the density to exist but more of the regression $E[y|x]$.

The SBP conferred by a twicing kernel is different than the nonparametric bias reduction that results from a higher order kernel. As discussed more fully in Section 3, the SBP is a result of the bias for $\hat{\beta}$ being a product of smoothing biases for $v(x)$ and $\gamma_0(x)$ rather than just bias for $\gamma_0(x)$. This leads to the MSE formula, where bias of $\hat{\beta}$ goes to zero at the rate h^{2s} even though $\gamma_0(x)$ is only s times differentiable. The magnitude of this bias reduction depends on the smoothness (number of derivatives in existence) of $v(x)$. If $v(x)$ has fewer than s derivatives $\hat{\beta}$ will still have the SBP, but the bias of $\hat{\beta}$ would be of order h^{s+t} for some $t < s$. For simplicity we have excluded from consideration such weaker smoothness conditions.⁵

One consequence of the SBP is that undersmoothing may not be required for \sqrt{n} -consistency. To see how this works for twicing kernel averages, consider the case where

³For the twicing kernel the optimal choice of bandwidth, minimizing the dominant MSE terms is $h = [\bar{K}Q(r + 2|\bar{\lambda}|)/(P^24sn^2)]^{1/(4s+r+2|\bar{\lambda}|)}$. This could be estimated by replacing Q and P by estimates, similarly to Powell and Stoker (1996).

⁴When $r = 6$ the higher order terms will have the same magnitude bandwidth as the leading term, so that $\hat{\beta}$ will not have V as its asymptotic variance.

⁵For instance, the twicing kernel average density estimator from equation (2.2), where $w = y = 1$ and $\lambda = 0$ attains \sqrt{n} -consistency under the minimal Holder continuity conditions of Bickel and Ritov (1988), that do not require any derivatives of $\gamma_0(x)$ to exist.

$\bar{\lambda} = 0$. In this case the conditions for \sqrt{n} -consistency are that $n^{-1}h^{-r}$ and nh^{4s} are bounded, which allows the order of the nonparametric variance $n^{-1}h^{-r}$ and bias h^{2s} to shrink at the same rate.

The SBP also applies to estimators that are nonlinear in $\hat{\gamma}$. Indeed, one could generalize Theorem 1 by deriving a stochastic expansion of the semiparametric estimator and the MSE of leading terms. However, this derivation would be quite complicated and so is beyond the scope of this note. Instead, we content ourselves with showing how the SBP affects \sqrt{n} -consistency.⁶ Consider a semiparametric m-estimator $\hat{\beta}$ solving

$$\sum_{i=1}^n m(z_i, \beta, \hat{\gamma}) = 0.$$

For analyzing this estimator we will adopt notation and conditions like those of Newey and McFadden (1994), to which we refer the interested reader for a fuller discussion. In particular we will assume that there is a function $D(z, \gamma)$ that is linear in γ such that

$$\|m(z, \beta_0, \gamma) - m(z, \beta_0, \gamma_0) - D(z, \gamma - \gamma_0)\| \leq b(z) \|\gamma - \gamma_0\|^2, \quad (2.3)$$

where $\|\gamma\| = \sup_{x \in X, |\lambda| \leq d} |\partial^\lambda \gamma(x)|$ for some compact set X and nonnegative integer d . In the special case of a linear kernel average $b(z) = 0$ and $D(z, \gamma) = \partial^{\bar{\lambda}} \gamma(x)y$.

We will also assume that there is a matrix of functions $v(x)$ such that

$$\bar{D}(\gamma) \stackrel{def}{=} \int D(z, \gamma) F_0(dz) = \int v(x) \gamma(x) dx. \quad (2.4)$$

For a linear kernel average, integration by parts gives this equation, with $\bar{D}(\gamma) = E[\partial^{\bar{\lambda}} \gamma(x)y] = \int \partial^{\bar{\lambda}} \gamma(x) g(x) f(x) dx = \int v(x) \gamma(x) dx$ for $v(x)$ defined above. More generally, existence of such an integral representation will be a consequence of $\bar{D}(\gamma)$ being continuous in the L_2 norm $\|\gamma\| = [\int \gamma(x)' \gamma(x) dx]^{1/2}$ and the Riesz representation theorem. Such continuity is fundamental to \sqrt{n} -consistency; e.g. see Newey and McFadden (1994). We will also impose other regularity conditions, which are stated in the Appendix. Under

⁶This derivation only requires bounding terms in a stochastic expansion rather than derivation of expectations and variances.

these conditions and the ones stated in the theorem, the asymptotic variance of $\hat{\beta}$ will be

$$V = M^{-1}E[\psi(z)\psi(z)']M^{-1'}, M = E[\partial m(z, \beta_0, \gamma_0)/\partial \beta],$$

$$\psi(z) = m(z, \beta_0, \gamma_0) + v(x)w - E[v(x)w].$$

Theorem 2: *If Assumptions A1-A6 are satisfied, $E[\|w\|^p] < \infty$ for $p > 4$, $h = h(n)$ with $n^{1-2/p}h^r/\ln(n) \rightarrow \infty$, $\sqrt{nh}^{2s} \rightarrow 0$, and either A) $v(x)$ is s times differentiable and the kernel is $K(u)$; or B) $v(x) = 0$ and the kernel is $k(u)$; then*

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, V).$$

To interpret this result, note h^s is the nonparametric bias rate, and that only $\sqrt{nh}^{2s} \rightarrow 0$ is imposed rather than $\sqrt{nh}^s \rightarrow 0$. Thus, we obtain \sqrt{n} -consistency without requiring that \sqrt{n} times the nonparametric bias rate h^s converges to zero. This means that $\hat{\beta}$ has the SBP (in its stochastic expansion). The interpretation of condition A) is that the SBP holds for a twicing kernel. Condition B) shows that when $v(x) = 0$, a twicing kernel is not needed for the SBP. Since $v(x)$ accounts, in the asymptotic variance, for the presence of $\hat{\gamma}$, we interpret B) to mean that $\hat{\beta}$ automatically has the SBP when the presence of the nonparametric estimator does not affect the asymptotic variance of $\hat{\beta}$. Although we have only shown this result for kernel estimation, we conjecture that it can be shown in general, e.g. using the set up of Newey and McFadden (1994).

The result for condition B) also suggests a general approach to constructing estimators with the SBP. By adding to the original $m(z, \beta, \gamma)$ an estimator of the correction term $v(x)w - E[v(x)w]$ for estimation of γ we can construct a new $m(z, \beta, \gamma)$, where estimation of γ does not affect the asymptotic variance of $\hat{\beta}$. Hence the $\hat{\beta}$ based on the new $m(z, \beta, \gamma)$ should have the SBP. For brevity we omit further discussion of this alternative approach.⁷

⁷This approach is discussed in the previous version of this paper, including a proof that kernel average derivatives based on a twicing kernel are identical to such a modified estimator.

3 Adjoint Bias and Idempotent Transformations

The SBP can be explained by the integral form of the bias and an adjoint transformation. For simplicity, consider the case where $\gamma(x)$ is the density of x , i.e. $w = 1$. Let $\kappa(x, y) = h^{-r}k((x - y)/h)$. By a change of variables (details in the Appendix) we have $h^{-r} \int k([(x - y)/h] - t)k(t)dt = \int \kappa(x, t)\kappa(t, y)dt$. A twicing kernel density estimator is then given by

$$\hat{\gamma}(x) = \sum_{i=1}^n \{2\kappa(x, x_i) - \int \kappa(x, t)\kappa(t, x_i)dt\}/n. \quad (3.1)$$

For a function g define B as the linear mapping that gives the smoothing bias from the original kernel $k(u)$ with bandwidth h , that is

$$Bg(x) = \int \kappa(x, y)g(y)dy - g(x).$$

The pointwise bias in terms of B is given by

$$\begin{aligned} E[\hat{\gamma}(x)] - \gamma(x) &= 2 \int \kappa(x, y)\gamma_0(y)dy - \int \int \kappa(x, t)\kappa(t, y)dt\gamma_0(y)dy - \gamma_0(x) \\ &= 2 \int \kappa(x, y)\gamma_0(y)dy - \int \kappa(x, t) \int \kappa(t, y)\gamma_0(y)dydt - \gamma_0(x) \\ &= -B(B\gamma_0)(x) = -B^2\gamma_0(x). \end{aligned}$$

Thus we see that the pointwise bias from a twicing kernel is equal to the negative of the "bias of the bias" for the original kernel.

To relate the pointwise bias to the semiparametric bias, consider $B_n = E[\bar{D}(\hat{\gamma} - \gamma_0)]$, which is the bias of a linear kernel average and also is the expectation of the linear in γ term in the expansion of $\hat{\beta} - \beta_0$ for a semiparametric m-estimator. Let $\langle a, b \rangle = \int a(x)b(x)dx$ for square integrable functions a and b . Then

$$\begin{aligned} B_n &= \langle v, E[\hat{\gamma}] - \gamma \rangle = \langle v, -B^2\gamma \rangle = -\langle v, B^2\gamma \rangle \\ &= -\langle B^*v, B\gamma \rangle = -\int B^*v(x) \cdot B\gamma(x)dx. \end{aligned}$$

where $B^*g(x) = \int \kappa(y, x)g(y)dy - g(x)$ is the adjoint transformation for B .⁸ Furthermore, symmetry of the kernel gives $\kappa(x, y) = \kappa(y, x)$, so that

$$B_n = -\langle Bv, B\gamma \rangle. \quad (3.2)$$

⁸See Luenberger 1969, p. 153.

Thus we see that the semiparametric bias is the integral of the product of smoothing biases for $v(x)$ and $\gamma(x)$ from the original kernel $k(u)$. The SBP is a consequence of this product form for B_n , because the bias B_n will be smaller asymptotically than $B\gamma(x)$. The integral form of B_n is essential to this result. It is the integration which allows us to obtain the bias product form in place of the "bias of bias" form.

It is interesting to note that the adjoint bias formula only depends on $\kappa(x, y) = \kappa(y, x)$, and not on κ having the kernel form. Thus, any estimator of the form given in equation (3.1) will confer the SBP, because equation (3.2) is satisfied, so that the semiparametric bias B_n will be of smaller order than the nonparametric bias.

An important type of estimator that confers the SBP, because it equals the twicing estimator of equation (3.1), is

$$\hat{\gamma}(x) = \sum_{i=1}^n \kappa(x, x_i)/n, \kappa(x, y) = \int \kappa(x, t)\kappa(t, y)dt. \quad (3.3)$$

Orthogonal series density estimators are examples. Let $p(x) = (p_1(x), \dots, p_J(x))'$ be orthonormal, with $\int p(t)p(t)'dt = I$, and let $\kappa(x, y) = p(x)'p(y)$. Then $\hat{\gamma}(x) = \sum_{i=1}^n \kappa(x, x_i)/n = p(x)'\sum_{i=1}^n p(x_i)/n$ is an orthogonal series density estimator. Also,

$$\int \kappa(x, t)\kappa(t, y)dt = p(x)' \left[\int p(t)p(t)'dt \right] p(y) = p(x)'p(y) = \kappa(x, y).$$

Thus, equation (3.3) is satisfied, so that orthogonal series density estimators will confer the SBP.⁹

The condition $\int \kappa(x, t)\kappa(t, y)dt = \kappa(x, y)$ is equivalent to $\hat{\gamma}(x) = \int \kappa(x, y)\hat{F}(dy)$ being an idempotent transformation of the empirical distribution \hat{F} , in the sense that $\int \kappa(x, y)\hat{\gamma}(y)dy = \hat{\gamma}(x)$, as shown in Lemma A3 in the Appendix. Thus, we see that any nonparametric density estimator that is an idempotent linear transformation of the empirical distribution, as in equation (3.3), confers the SBP. This link between idempotent transformations and the SBP is consistent with previous results on nonparametric estimators that confer the SBP. For example, Newey (1994) showed that series estimators of conditional expectations confer the SBP. Not surprisingly, given their least squares projection form, series estimators can be interpreted as idempotent transformations of the

⁹This result was fully shown in a previous version of this paper.

empirical distribution. Also, Shen (1997) showed that a sieve density estimator, where $\hat{\gamma} = \int \ln \gamma(x) \hat{F}(dx)$ for some set of densities Γ , has the SBP. This result is consistent with the idempotency of the sieve density as a transformation of the empirical distribution, with $\tilde{\gamma} = \arg \max_{\gamma \in \Gamma} \int [\ln \gamma(x)] \tilde{\gamma}(x) dx$ holding by the information inequality. Although series and sieve estimators are nonlinear transformations of the empirical distribution, so that the preceding analysis does not directly apply, we conjecture that the SBP is a direct result of their being idempotent.

The SBP provides a simple criteria for selecting among different nonparametric estimates to use in semiparametric estimation. In the linear kernel average case of Section 2, and we conjecture in general, these estimates increase the rate at which the bias converges to zero and so lead to smaller asymptotic MSE. Of course, it is important to understand whether this asymptotic advantage translates into small sample improvements. For this purpose we turn to some Monte Carlo work.

4 Monte Carlo Comparisons

The importance of the SBP in practice depends on small sample properties. We investigate these properties in a Monte Carlo experiment. We first consider the density weighted average derivative estimator of Powell, Stock, and Stoker (1989), as given in equation (2.2) for $|\bar{\lambda}| = 1$ and $w = 2$. This estimator is important because ratios of weighted average derivatives estimate the coefficients δ in an index model $E[y|x] = \tau(x'\delta)$, up to scale. Also, average derivatives can be of interest in their own right, as in Hardle, Hildenbrand, and Jerison (1991).

Our experiment was based on the same model considered by Ruud (1986), given by

$$y = \exp(x_1 + x_2 + u), \quad u \text{ distributed uniformly on } (-1/2, 1/2),$$

$$(x_1, x_2) \text{ has p.d.f. } \phi(x_1 + 1/2)\phi(x_2/2) + \phi(x_2 - 1/2)\phi(x_1/2).$$

The regressor distribution is a bivariate mixture of normal densities which leads to substantial biases in quasi-maximum likelihood estimators of the ratio of regression coefficients for x_1 and x_2 . We consider two estimators of the density weighted average deriva-

tive, one based on a standard normal kernel and the other on the corresponding twicing kernel. Figure 1 plots the MSE as a function of the bandwidth for both estimators, for sample sizes $n = 50$ and $n = 200$. We find here that the MSE for the twicing kernel has a somewhat smaller minimum and is lower than that of the original kernel over a wide range, especially for the larger sample size. In this sense the MSE function for the twicing kernel is *less* peaked than that of the original kernel, a surprising finding and a further advantage of the twicing kernel. Also, the fact that the MSE of the twicing kernel is below that of the original one over such a wide range of bandwidths may be important in practice, where the optimal bandwidth is not known.

We also obtain similar results for the ratio for $n = 200$. Although the advantages of the twicing kernel are not as pronounced here, the MSE function is still somewhat flatter and is below that of the original kernel over a very wide range of bandwidths.

Significant MSE gains from using a twicing kernel have also been reported by Newey and Ruud (2001) for the inverse density weighted least squares estimator of Ruud (1986), for the design given above. The estimator is weighted least squares, $\hat{\beta}_w = (\sum_{i=1}^n \hat{w}_i x_i x_i')^{-1} \sum_{i=1}^n \hat{w}_i x_i y_i'$ where $\hat{w}_i = \varphi(x_i, \hat{\theta}) / \hat{f}(x_i)$ and $\varphi(x_i, \hat{\theta})$ is a spherically symmetric density that can depend on estimated parameters $\hat{\theta}$. Ratios of components of $\hat{\beta}$ are consistent estimators of the corresponding ratio of components of δ . For the design given above, for a range of bandwidths the MSE of this ratio was 25 percent to 50 percent smaller with the twicing kernel than the original kernel. Slightly smaller MSE gains from using a twicing kernel were also found for a different model where $y = 1(x_1 + x_2 > 1) + u$, where u is distributed as $N(0, .01)$.

Overall, these Monte Carlo results show substantial gains from using the twicing kernel. Except for the ratio with $n = 50$, the minimum MSE for the twicing kernel is below that for the original kernel in Figures 1 and 2. Also, the MSE for the twicing kernel is below that for the original kernel over a wide range of bandwidths. This promising performance indicates potential for practice.

Appendix: Proofs

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses, and DCT a reference to the dominated convergence theorem.

ASSUMPTION A1: $\int k(u)du = 1$, $k(-u) = k(u)$, $k(u)$ has bounded support, $k(u)$ is differentiable of order $d = |\bar{\lambda}|$ with Lipschitz d^{th} derivative.

ASSUMPTION A2: For all multi-indices λ and $\bar{\lambda}$ with $|\lambda| = s$ and $|\bar{\lambda}| = s$ there is a $c > 0$ such that $\int \sup_{\|\Delta\| \leq c} |\partial^\lambda \gamma_0(x + \Delta)| \sup_{\|\Delta\| \leq c} |\partial^{\bar{\lambda}} \nu(x + \Delta)| dx < \infty$. Also, $\nu(x)$ is continuous and $\gamma_0(x)$ and $\nu(x)$ are bounded.

ASSUMPTION A3: Let $\mu_{yy}(x) = E[y^2|x]f(x)$, $\mu_{ww}(x) = E[w^2|x]f(x)$, and $\mu_{yw}(x) = E[yw|x]f(x)$. Then $\mu_{ww}(x)$ and $\mu_{yw}(x)$ are continuous and for some $c > 0$,

$$\int \sup_{\|\Delta\| \leq c} \mu_{yy}(x) \mu_{ww}(x + \Delta) dx \text{ and}$$

$$\int \sup_{\|\Delta\| \leq c} |\mu_{yw}(x + \Delta) \mu_{yw}(x)| dx \text{ are finite.}$$

ASSUMPTION A4: Equation (2.3) is satisfied for all γ with $\|\gamma - \gamma_0\|$ small enough, $\|D(z, \delta)\| \leq b(z)^{1/2} \|\delta\|$, and $E[b(z) \|w\|^2] < \infty$.

ASSUMPTION A5: Equation (2.4) is satisfied and $\nu(x)$ is zero outside X .

ASSUMPTION A6: $E[m(z, \beta_0, \gamma_0)] = 0$, $E[\|m(z, \beta_0, \gamma_0)\|^2] < \infty$, and $E[\|w\|^p |x]f(x)$ is bounded. Also, for all $\|\gamma - \gamma_0\|$ small enough $m(z, \beta, \gamma)$ is continuously differentiable on a neighborhood of β_0 , $M = E[\partial m(z, \beta_0, \gamma_0)/\partial \beta]$ exists and is nonsingular, and there is $b(z)$ and $\varepsilon > 0$ such that $E[b(z)] < \infty$ and $\|\partial m(z, \beta, \gamma)/\partial \beta - \partial m(z, \beta_0, \gamma_0)/\partial \beta\| \leq b(z)(\|\beta - \beta_0\|^\varepsilon + \|\gamma - \gamma_0\|^\varepsilon)$.

Proof of Theorem 1: We have

$$\hat{\beta} = \sum_{i \neq j} \sum k(z_i, z_j) / [n(n-1)], \quad k(z_i, z_j) = \partial^\lambda K_h(x_i - x_j) w_j y_i,$$

where we suppress dependence of k on h for notational convenience. Define $\bar{\gamma}(x) = \int k(u) \gamma_0(x + hu) du$ and $\bar{v}(x) = \int k(u) \nu(x + hu) du$. Taking expectations, and integrating

by parts, it follows similarly to equation (2.7) that for $\beta_0 = \int v(x)\gamma_0(x)dx$,

$$\begin{aligned} E[\hat{\beta}] &= E[k(z_1, z_2)] = \int \int K_h(x_1 - x_2)\gamma_0(x_2)v(x_1)dx_1dx_2 - \beta_0 \\ &= - \int [\bar{v}(x) - v(x)] [\bar{\gamma}(x) - \gamma_0(x)]dx + \beta_0. \end{aligned}$$

By an expansion in h with Lagrange form of the remainder and by $\zeta_\lambda = 0$ for $|\lambda| < s$,

$$\bar{\gamma}(x) - \gamma_0(x) = (h^s/s!) \sum_{|\lambda|=s} \int k(u)u^\lambda \partial^\lambda \gamma_0(x + \bar{h}u)du, \quad |\bar{h}| < |h|. \quad (\text{A.1})$$

For c in Assumption A2 let $d_\gamma(x) = \sum_{|\lambda|=s} \sup_{\|\Delta\| \leq c} |\partial^\lambda \gamma_0(x + \Delta)|$. By $k(u)$ having bounded support \mathcal{U} , for small enough h , $|k(u) \sum_{|\lambda|=s} u^\lambda \partial^\lambda \gamma_0(x + \bar{h}u)| \leq C1(u \in \mathcal{U})d_a(x) < \infty$. Then by the dominated convergence theorem, $\int K(u)u^\lambda \partial^\lambda \gamma_0(x + \bar{h}u)du \rightarrow \zeta_\lambda \partial^\lambda \gamma_0(x)$, so that $h^{-s}[\bar{\gamma}(x) - \gamma_0(x)] \rightarrow \sum_{|\lambda|=s} \zeta_\lambda \partial^\lambda \gamma_0(x)/s!$. Similarly, $h^{-s}[\bar{v}(x) - v(x)] \rightarrow \sum_{|\lambda|=s} \zeta_\lambda \partial^\lambda v(x)/t!$. Also, noting that $|h^{-s}[\bar{\gamma}(x) - \gamma_0(x)]| \leq Cd_\gamma(x)$ and $|h^{-t}[\bar{v}(x) - v(x)]| \leq Cd_v(x)$ for $d_v(x)$ defined analogously to $d_\gamma(x)$, and for $\int d_\gamma(x)d_v(x)dx < \infty$ by Assumption A2 and eq. (A.1), the dominated convergence theorem gives $h^{-2s}(E[\hat{\beta}] - \beta_0) \rightarrow P$, so that

$$E[\hat{\beta}] = \beta_0 + h^{2s}P + o(h^{2s}). \quad (\text{A.2})$$

Next, note that $\hat{\beta}$ is a U -statistic with kernel $[k(z_i, z_j) + k(z_j, z_i)]/2$. Then by Serfling (1980),

$$\begin{aligned} \text{Var}(\hat{\beta}) &= [(n-2)/n(n-1)] \text{Var}(E[k(z_1, z_2) + k(z_2, z_1)|z_1]) \\ &\quad + [1/n(n-1)]\{\text{Var}(k(z_1, z_2)) + \text{Cov}(k(z_1, z_2), k(z_2, z_1))\} \end{aligned}$$

By Assumptions A1 and A3 and another application of the dominated convergence theorem, $\int \int [\partial^\lambda K(u)]^2 \mu_{yy}(x - hu)\mu_{ww}(x)dudx \rightarrow Q_1 = \int [\partial^\lambda K(u)]^2 du \int \mu_{yy}(x)\mu_{ww}(x)dx$. Then by a change of variables, $u = (x_1 - x_2)/h$, $x = x$,

$$\begin{aligned} E[k(z_1, z_2)^2] &= \int \int [\partial^\lambda K_h(x_1 - x_2)]^2 E[w_2^2|x_2]E[y_1^2|x_1]f_0(x_2)f_0(x_1)dx_2dx_1 \\ &= h^{-r-2|\lambda|} \int \int [\partial^\lambda K(u)]^2 \mu_{yy}(x - hu)\mu_{ww}(x)dudx = h^{-r-2|\lambda|}Q_1 + o(h^{-r-2|\lambda|}). \end{aligned}$$

Also, since $E[k(z_1, z_2)]$ converges, $E[k(z_1, z_2)]^2 = o(h^{-r-2|\lambda|})$, so $\text{Var}(k(z_1, z_2)) = E[k(z_1, z_2)^2] + o(h^{-r-2|\lambda|})$. Therefore,

$$\begin{aligned} [1/n(n-1)] \text{Var}(k(z_1, z_2)) &= [1/n(n-1)] h^{-r-2|\lambda|} Q_1 + o([1/n(n-1)] h^{-r-2|\lambda|}) \\ &= n^{-2} h^{-r-2|\lambda|} Q_1 + o(n^{-2} h^{-r-2|\lambda|}), \end{aligned}$$

Also, by $\partial^\lambda K(-u) = (-1)^{|\lambda|} \partial^\lambda K(u)$, it follows similarly that

$$\begin{aligned} E[k(z_1, z_2)k(z_2, z_1)] &= E[\partial^\lambda K_h(x_1 - x_2)w_2y_1 \partial^\lambda K_h(x_2 - x_1)w_1y_2] \\ &= (-1)^{|\lambda|} E[\partial^\lambda K_h(x_1 - x_2)^2 E[w_2y_2|x_2] E[w_1y_1|x_1]] \\ &= h^{-r-2|\lambda|} (Q - Q_1) + o(h^{-r-2|\lambda|}). \end{aligned}$$

so that

$$[1/n(n-1)] \text{Cov}(k(z_1, z_2), k(z_2, z_1)) = n^{-2} h^{-r-2|\lambda|} (Q - Q_1) + o(n^{-2} h^{-r-2|\lambda|}).$$

Next, let $\tilde{\nu}(x) = \int K(u)\nu(x+hu)du$ and $\tilde{a}(x) = \int K(u)a(x+hu)du$. By applying the dominated convergence theorem as we have done previously it follows from Assumption A2 that $\tilde{\nu}(x) \rightarrow \nu(x)$ and $\tilde{a}(x) \rightarrow a(x)$ as $h \rightarrow 0$, so that $\tilde{a}(x)y + \tilde{\nu}(x)w \rightarrow a(x)y + \nu(x)w$ as $h \rightarrow 0$. Since $[\tilde{a}(x)y + \tilde{\nu}(x)w]^2 \leq C(y^2 + w^2)$, Assumption A2 and the dominated convergence theorem imply that $\text{Var}(\tilde{a}(x)y + \tilde{\nu}(x)w) \rightarrow V$. Furthermore, by integration by parts and interchanging the order of differentiation and integration,

$$\begin{aligned} &E[k(z_1, z_2) + k(z_2, z_1)|z_1] \\ &= E[\partial^\lambda K_h(x_1 - x_2)E[w_2|x_2]|z_1]y_1 + E[\partial^\lambda K_h(x_2 - x_1)E[y_2|x_2]|z_1]w_1 \\ &= \left[\int \partial^\lambda K_h(x_1 - x)E[w|x]f_0(x)dx \right]y_1 + (-1)^{|\lambda|} \int \partial^\lambda K_h(x_1 - x)b(x)dx w_1 \\ &= \tilde{a}(x)y + \tilde{\nu}(x)w. \end{aligned}$$

Therefore,

$$[(n-2)/n(n-1)] \text{Var}(E[k(z_1, z_2) + k(z_2, z_1)|z_1]) = V/n + o(n^{-1}).$$

Then $MSE(\hat{\beta}) = \text{Var}(\hat{\beta}) + (E[\hat{\beta}] - \beta_0)^2$ gives the result.

QED.

Proof of Theorem 2: It follows exactly as in Newey and McFadden (1994) that for any $\bar{\beta} \xrightarrow{p} \beta_0$ we have $\sum_{i=1}^n \partial m(z_i, \bar{\beta}, \hat{\gamma}) / \partial \beta / n \xrightarrow{p} M$, and for $\tilde{\gamma}(x) = E[\hat{\gamma}(x)]$,

$$\begin{aligned} & \sum_{i=1}^n [m(z_i, \beta_0, \hat{\gamma}) - m(z_i, \beta_0, \gamma_0) - D(z_i, \hat{\gamma} - \gamma_0)] / \sqrt{n} \xrightarrow{p} 0, \\ & \sum_{i=1}^n [D(z_i, \hat{\gamma} - \gamma_0) - \bar{D}(\hat{\gamma} - \gamma_0)] / \sqrt{n} \xrightarrow{p} 0, \\ & \sqrt{n} \bar{D}(\hat{\gamma} - \tilde{\gamma}) - \sum_{i=1}^n \{v(x_i)w_i - E[v(x)w]\} / \sqrt{n} \xrightarrow{p} 0. \end{aligned}$$

Then by the central limit theorem and the triangle inequality, for $\sum_{i=1}^n m(z_i, \beta_0, \hat{\gamma}) / \sqrt{n} \xrightarrow{d} N(0, E[\psi(z)\psi(z)'])$ it then suffices to show that $\sqrt{n} \bar{D}(\tilde{\gamma} - \gamma_0) \rightarrow 0$. It follows similarly to eqs. (3.2) and (A.2) that

$$\bar{D}(\tilde{\gamma} - \gamma_0) = - \int [\bar{v}(x) - v(x)][\tilde{\gamma}(x) - \gamma_0(x)] dx = O(h^{2s}),$$

so that the conclusion follows by $\sqrt{n}h^{2s} \rightarrow 0$. Q.E.D.

Lemma A3: If x_1, \dots, x_n are i.i.d. with support S for x_i and for all $x \in S$, $E[\kappa(x, x_i)^2] < \infty$ and $E[\{\int \kappa(x, y)\kappa(y, x_i) dy\}^2] < \infty$, then $\kappa(x, y) = \int \kappa(x, t)\kappa(t, y) dt$ for all $(x, y) \in S \times S$ if and only if $\int \kappa(x, y)\hat{\gamma}(y) dy = \hat{\gamma}(x)$ for all $x \in S$.

Proof: The only if part is easy, so we show just the if statement. We have $e'U = 0$ where e and U are $n \times 1$ vectors with $e = (1, \dots, 1)'$, $U_i = \kappa(x, x_i) - \int \kappa(x, y)\kappa(y, x_i) dy$. Then by U_i i.i.d with finite second moment,

$$0 = E[e'U] = nE[U_i], 0 = Var(e'U) = e'Var(U)e = nVar(U_i),$$

By $E[U_i] = Var(U_i) = 0$, $U_i = 0 = \kappa(x, x_i) - \int \kappa(x, y)\kappa(y, x_i) dy$ for all $x_i \in S$. Q.E.D.

References

- [1] BICKEL, P.J. AND Y. RITOV (1988): "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Results," *Sankhya* 50A, 381-393.
- [2] HARDLE, W., W. HILDENBRAND, AND M. JERISON (1991): "Empirical Evidence on the Law of Demand," *Econometrica* 59, 1525-1549.

- [3] HAUSMAN, J.A. AND W.K. NEWEY (1995): "Nonparametric Estimation of Exact Consumer Surplus and Deadweight Loss," *Econometrica*, 63, 1445-1476.
- [4] KAUERMANN, G., M. MUELLER, AND R.J. CARROLL (1997): "The Efficiency of Bias-Corrected Estimators for Nonparametric Kernel Estimation Based on Local Estimating Equations," preprint.
- [5] LEWBEL, A. (1998): "Semiparametric Latent Variable Model Estimation With Endogeneous or Mismeasured Regressors," *Econometrica* 66, 105-121.
- [6] LUENBERGER, D.G., (1969): *Optimization By Vector Space Methods*, New York: John Wiley and Sons.
- [7] NEWEY, W.K. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica* 62, 1349-1382.
- [8] NEWEY, W.K. AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, Vol 4, Amsterdam: North-Holland.
- [9] POWELL, J.L., J.H. STOCK, AND T.M. STOKER (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403-1430.
- [10] POWELL, J.L. AND T.M. STOKER (1996): "Optimal Bandwidth Choice for Density-Weighted Averages," *Journal of Econometrics* 75, 291-316.
- [11] RÜUD, P.A. (1986): "Consistent Estimation of Limited Dependent Variable Models Despite Misspecification of Distribution," *Journal of Econometrics* 32, 157-187.
- [12] SERFLING, R.J. (1980): *Approximation Theorems of Mathematical Statistics*, New York: Wiley and Sons.
- [13] SHEN, X. (1997): "On Methods of Sieves and Penalization," *Annals of Statistics* 25, 2555-2591.

- [14] STOKER, T.M. (1993): "Smoothing Bias in Density Derivative Estimation," *Journal of the American Statistical Association* 88, 855-863.