UNDERSMOOTHING AND BIAS CORRECTED FUNCTIONAL ESTIMATION[I]

Abbreviated Title: Undersmoothing and Bias Correction

By Whitney Newey, Fushing Hsieh, and James Robins

Massachusetts Institute of Technology, Academia Sinica,

and Harvard University

First version October 1991; revised July 1998.

SUMMARY

There are many important example of $\sqrt{n}$-consistently estimable functionals that are interesting in econometrics, such as average derivatives and nonparametric consumer surplus. Corresponding estimators may require undersmoothing to achieve $\sqrt{n}$-consistency, due to first order bias in the expected influence function. We give a general bias correction that can be added to a plug-in estimator to remove the need for undersmoothing and improve its higher order properties. We also describe a bootstrap smoothing correction for the nonparametric estimator that achieves analogous results for the plug-in estimator and show that idempotent transformations of the empirical distribution need not require undersmoothing for $\sqrt{n}$-consistency. We find that this bias correction can lead to large efficiency improvements and lower sensitivity to bandwidth choice.

[1]Research partially supported by NSF grant SBR-9409707.

Key words and phrases. Functional estimation, nonparametric estimation, $\sqrt{n}$-consistency, undersmoothing, influence function, bias correction.

1. Introduction

Functionals of nonparametric estimators that can be $\sqrt{n}$-consistent have important applications in econometrics, including average derivatives and average consumer surplus. Most functional estimators are based on nonparametric estimation and many require "undersmoothing" of the nonparametric estimator to achieve $\sqrt{n}$-consistency, meaning the bias of the nonparametric estimator shrinks faster than its variance. In this paper we show that this requirement can be removed, by either adding a bias correction term to the functional estimator or using a smoothing correction for the nonparametric estimator, leading to $\sqrt{n}$-consistent functional estimation without undersmoothing. These modifications also lead to functional estimators with improved higher-order efficiency, that may attain $\sqrt{n}$-consistency when others do not.

The source of this improvement is a reduction in the bias of the functional estimator. Many previous functional estimators have a bias that is the same order as the bias of the density estimator. In contrast, the estimators we develop have a bias that is the same order as the *product* of the bias of the nonparametric estimator with another bias term. The other bias term is that for the influence function of the estimator, which is the mean-square derivative of the functional. We refer to the corresponding reduction in bias order as a bias complementarity, with the influence function bias term complementing the nonparametric bias term.

The properties of the influence function play a key role in our analysis. When the influence function is smooth, in certain ways to be made precise below, the influence function bias term will be small enough that the need for undersmoothing will be removed. For some of the estimators we consider, the influence function will be required to be smooth as a function of the density. This is a regularity condition that is often satisfied. For other estimators the influence function will be required to be smooth as a function of the data, a condition that does not hold for some functionals.

We consider two approaches to bias corrected functional estimation. Our first

approach is to add to the functional estimator a bias correction, consisting of the difference of integrals of an influence function estimator over the empirical distribution and the nonparametric estimate. This additional term does not affect the asymptotic distribution of the estimator in the $\sqrt{n}$-consistent case but it does lead to improved higher order properties. Our second approach is to do a bootstrap smoothing correction to the nonparametric estimator before using it in functional estimation. This correction consists of replacing a nonparametric distribution estimator $\hat{F}$ by $\tilde{F} = \hat{F} - (\hat{G} - \hat{F}) = 2\hat{F} - \hat{G}$, where $\hat{G}$ is an estimator obtained from $\hat{F}$ by the same transformation used to obtain $\hat{F}$ from the empirical distribution. We show that using this estimator removes the need for undersmoothing when the influence function is smooth enough. For kernel estimators we show that this smoothing correction leads to a particular kind of kernel, the "twicing" kernel described below. We also show that undersmoothing will not be needed for $\sqrt{n}$-consistency when $\hat{F}$ is an idempotent transformation of the empirical distribution, so that $\hat{G} = \hat{F}$ and $\tilde{F} = \hat{F}$, and the bootstrap correction is built into the original estimator. This idempotent estimator class includes orthogonal series density estimators, series estimators of conditional expectations, and sieve estimators, and thus provides an explanation for the lack of an undersmoothing requirement for these estimators.

We emphasize that the bias reduction we consider does not come from reducing the bias of the density estimator. Such higher order bias reductions (e.g. via higher-order kernels) do not remove the requirement that bias shrinks faster than variance for the nonparametric estimator. That requirement can only be removed if the bias of the functional estimator is smaller order than the bias of the nonparametric estimator. This reduction in bias is brought about by the bias complementarity we will discuss.

The bias reduction may be accompanied by some increase in the variance of the functional estimator. Although the variance still shrinks at the same rate, the size of the variance will be larger. In large samples the reduction in bias will allow adjustment of the smoothing parameters so that the variance is smaller, although the bias

reduction could increase the small sample variance of the estimator. Bias reductions for nonparametric estimators (like higher order kernels) have similar properties, except that they depend on higher order properties of the nonparametric estimator whereas our functional bias reduction depends on the properties of the influence function. We find that in a simple example the bias correction can give a large efficiency gain and reduce sensitivity to the choice of bandwidth.

The type of estimator we consider has antecedents in the literature. Bickel and Ritov (1988) developed an estimator for the average density that is $\sqrt{n}$-consistent under minimal smoothness conditions. This estimator has a similar form to the bias corrected functional discussed here, as do estimators in Pastuchova and Hasminskii (1989). Our contribution is to give a general version of this type of estimator and show the improvement in its properties. Also, we show how bootstrap corrected nonparametric estimators remove the need for undersmoothing, and hence that undersmoothing is not needed for idempotent nonparametric estimators.

Section Two of the paper describes the general form of the bias corrected functional we consider, and gives results for kernel estimators. Section Three describes a nonparametric estimator with a bootstrap correction for smoothing and its use for functional estimation. Section Four considers a special class of linear estimators where mean-square error calculations are feasible, showing more precisely the higher order effect of the bias correction, and giving exact MSE calculations for one case. Section Five analyzes semiparametric m-estimators, shows that undersmoothing is not needed when the nonparametric estimator has no effect on the limiting distribution, and derives results for kernel estimation with a smoothing correction. Section Six concludes.

2. Bias Corrected Functional Estimation

To focus on the essential features of the problem at hand it is useful to begin our discussion with a linear functional $\mu(F) = \int\delta(z)F(dz)$ where $\delta(z)$ is known and F denotes some unsigned measure (charge) on z. Although this case is relatively simple, the role of undersmoothing in achieving $\sqrt{n}$-consistency is easily understood here. Let $\hat{F}$ be some nonparametric estimator of the true distribution F_0. For example, $\hat{F}$ could correspond to a nonparametric density estimator $\hat{f}$, with $\hat{F}(z) = \int 1(u\leq z)\hat{f}(u)du$ and $\mu(\hat{F}) = \int\delta(z)\hat{f}(z)dz$. Consider the estimator $\hat{\mu} = \mu(\hat{F}) = \int\delta(z)\hat{F}(dz)$ and let $\overline{F}(z) = E[\hat{F}(z)]$. Assuming that the order of integration can be interchanged, the bias of this estimator is

$$(2.1) \qquad E[\hat{\mu}] - \mu_0 = \int\delta(z)\overline{F}(dz) - \int\delta(z)F_0(dz) = \int\delta(z)(\overline{F}-F_0)(dz).$$

If the order of this bias is the same as the order of the pointwise bias of the nonparametric density estimator, as occurs in many cases, then $\sqrt{n}$-consistency will involve the pointwise bias shrinking faster than $1/\sqrt{n}$. Furthermore, the pointwise standard deviation of a nonparametric density estimator generally shrinks no faster than $1/\sqrt{n}$, so that $\sqrt{n}$-consistency will involve the pointwise bias shrinking faster than the pointwise standard deviation. This property is referred to as undersmoothing, since less bias is generally associated with less smoothing and the fastest shrinkage of mean square error is generally associated with bias and standard deviation shrinking at the same rate.

A way to reduce the bias of $\hat{\mu}$ is to form an estimator of the bias term in equation (2.1) and subtract it off. In this simplest case there is an unbiased estimator $\int\delta(z)\hat{F}(dz) - \sum_{i=1}^{n}\delta(z_i)/n$ of the bias term, and subtracting it off gives

$$(2.2) \qquad \tilde{\mu} = \hat{\mu} + \sum_{i=1}^{n}\delta(z_i)/n - \int\delta(z)\hat{F}(dz) = \sum_{i=1}^{n}\delta(z_i)/n,$$

the usual unbiased estimator.

For nonlinear functionals the same basic idea can be applied to a first-order bias term. To describe this approach, suppose that $\mu(F)$ has some restricted domain $\mathcal{F}$ to which both $\hat{F}(z)$ and $F_0(z)$ belong. For example, $\mathcal{F}$ might be restricted to have elements that are absolutely continuous with respect to Lebesgue measure, with a density f corresponding to each $F \in \mathcal{F}$. Also suppose that there is an expansion of $\mu(F)$ on $\mathcal{F}$ such that

$$(2.3) \qquad \mu(F) = \mu(F_0) + \int \delta(z,F_0)(F-F_0)(dz) + R(F-F_0,F_0), \quad |R(F-F_0,F_0)| = o(\|F-F_0\|),$$

where $\|F\|$ denotes a function semi-norm. Let $\hat{P}$ denote the empirical distribution, $\delta(z) = \delta(z,F_0)$, and $\psi(z) = \delta(z)-E[\delta(z)]$. Then the plug-in estimator $\hat{\mu} = \mu(\hat{F})$ satisfies

$$(2.4) \qquad \sqrt{n}(\hat{\mu} - \mu_0) = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + R_n + \hat{B}_n, \quad R_n = \sqrt{n}R(\hat{F}-F_0,F_0), \quad \hat{B}_n = \sqrt{n}\int \delta(z)(\hat{F}-\hat{P})(dz).$$

The first order bias of this estimator will be $E[\hat{B}_n] = \sqrt{n}\int \delta(z)(\overline{F}-F_0)(dz)$, with $R(\hat{F}-F_0,F_0)$ including higher order terms. A bias correction can be formed by subtracting an estimate of this first order term. If $\delta(z)$ were known, an unbiased estimate of this first order term would be $\int \delta(z)\hat{F}(dz) - \sum_{i=1}^{n} \delta(z_i)/n$. A feasible version of this bias estimate can be formed by replacing $\delta(z)$ with an estimator $\hat{\delta}(z)$. Then subtracting the corresponding bias estimator gives

$$(2.5) \qquad \tilde{\mu} = \hat{\mu} + \sum_{i=1}^{n} \hat{\delta}(z_i)/n - \int \hat{\delta}(z)\hat{F}(dz) = \hat{\mu} + \int \hat{\delta}(z)(\hat{P}-\hat{F})(dz).$$

Unlike the linear functional case, this estimator will not be exactly unbiased, because of the higher order term $R(\hat{F}-F_0,F_0)$ and the estimation of $\delta(z)$. Nevertheless, because we have subtracted an estimator of the first-order bias of $\hat{\mu}$, this estimator should have smaller bias than $\hat{\mu}$. In fact, as discussed below, it has an asymptotic expansion that is the same as $\hat{\mu}$ except one of the remainder terms is of smaller order.

A simple example is the average density. Suppose that $\mathcal{F}$ is restricted to contain only absolutely continuous elements with square integrable densities and let $\mu(F) = \int f(z)^2 dz$. Let $\hat{f}$ and f_0 denote the densities corresponding to $\hat{F}$ and F_0. Note that $\mu(F) = \mu(F_0) + \int 2f_0(z)[f(z)-f_0(z)]dz + \int [f(z)-f_0(z)]^2 dz$, so that equation (2.3) is satisfied with $\delta(z,F_0) = 2f_0(z)$, and $\|F\| = \mu(F)^{1/2}$. Letting $\hat{\delta}(z) = \delta(z,\hat{F}) = 2\hat{f}(z)$, a bias corrected estimator is given by

$$(2.6) \qquad \tilde{\mu} = \int \hat{f}(z)^2 dz + \int 2\hat{f}(z)(\hat{P}-\hat{F})(dz) = 2\sum_{i=1}^n \hat{f}(z_i)/n - \int \hat{f}(z)^2 dz.$$

In this example the bias corrected estimator is a linear combination of two well known estimators, the plug-in estimator $\int \hat{f}(z)^2 dz$ and the average estimated density $\sum_{i=1}^n \hat{f}(z_i)/n$.

To see the effect of the bias correction in the general case we can compare remainder terms. We have

$$(2.7) \qquad \sqrt{n}(\tilde{\mu} - \mu_0) = \sum_{i=1}^n \psi(z_i)/\sqrt{n} + R_n + D_n, \quad D_n = \sqrt{n}\int [\delta(z)-\hat{\delta}(z)](\hat{F}-\hat{P})(dz).$$

The only difference between $\hat{\mu}$ and $\tilde{\mu}$ is that the remainder term $\hat{B}_n$ has been replaced by D_n. The remainder term D_n should be of smaller order than $\hat{B}_n$ under appropriate regularity conditions, because the fixed term $\delta(z)$ in $\hat{B}_n$ has been replaced by $\delta(z)-\hat{\delta}(z)$ that is shrinking when $\hat{\delta}(z)$ is consistent. Thus, $\tilde{\mu}$ has the same expansion as $\hat{\mu}$ except that one remainder term is smaller, and in this sense improves upon $\hat{\mu}$. In particular, the remainder term D_n is second-order, in contrast with the first-order remainder $\hat{B}_n$, and so may have smaller order than the pointwise bias of $\hat{f}$. Consequently, undersmoothing may not be needed for $\sqrt{n}$-consistency of $\tilde{\mu}$.

The form of $\tilde{\mu}$ given in equation (2.5) is like an efficient estimator that is obtained in one step from $\hat{\mu}$,

$$\tilde{\mu} = \hat{\mu} + \sum_{i=1}^n \hat{\psi}(z_i)/n, \quad \hat{\psi}(z) = \hat{\delta}(z) - \int \hat{\delta}(u)\hat{F}(du).$$

In the semiparametric efficiency literature this procedure is used to improve asymptotic efficiency (e.g. see Bickel et. al., 1990), in the sense of lowering the asymptotic variance. Here it does not affect the asymptotic variance of the estimator when R_n, $\hat{B}_n$, and D_n all converge in probability to zero. Instead, it lowers the size of one of the remainder terms, speeding up the convergence rate of the estimator when $\hat{B}_n$ dominates, improving efficiency in this sense. Although it is difficult to derive exact convergence rates, because of the nonlinearity of D_n and R_n, we can compare conditions for $\sqrt{n}$-consistency, which we do in this Section. This comparison is one important aspect of the relative convergence rates for $\hat{\mu}$ and $\tilde{\mu}$. In Section Four we consider a different class of estimators where it is possible to compare convergence rates for the original and bias corrected estimators.

To explain this estimator and its properties it is helpful to discuss the role of the expansion in equation (2.3), how to form $\hat{\delta}(z)$, and to compare conditions for $\sqrt{n}$-consistency of $\hat{\mu}$ and $\tilde{\mu}$.

2.1 The Functional Expansion

The formation of this bias corrected estimator, and its properties, depend on the expansion of equation (2.3). Equation (2.3) actually embodies two conditions; i) $\mu(F)$ is Frechet differentiable with respect to $\|F\|$; ii) there is an integral representation for the derivative. Although it is well known that Frechet differentiability does not generally hold over a domain that includes empirical distributions, our allowance of a restricted domain and for a choice of semi-norm makes this hypothesis quite general. Many functionals will have Frechet derivatives when $\mathcal{F}$ and $\|\cdot\|$ are appropriately specified. For example, we noted above that $\mu(F) = \int f(z)^2 dz$ is Frechet differentiable for $\mathcal{F}$ and $\|F\|$ as previously specified.

The other condition embodied in equation (2.3), the integral representation for the linear term, limits the scope of our results to functionals that satisfy the necessary

conditions for $\sqrt{n}$-consistent estimability. To see why, consider any parametric family $\{F_\gamma\}$ passing through F_0 at $\gamma = 0$ that is regular in the sense of Bickel et al (1990), with score $S(z) = 2f_0(z)^{-1/2}\partial f_\gamma(z)^{1/2}/\partial\gamma$ (where $f_\gamma(z)$ is a density for $F_\gamma(z)$), $\int\delta(z,F_0)^2F_\gamma(dz)$ is bounded as a function of γ, and $\|F_\gamma - F_0\| = O(\|\gamma\|)$. Then it follows by Bickel et. al. (1990) that $\partial\int\delta(z,F_0)F_\gamma(dz)/\partial\gamma = E[\delta(z,F_0)S(z)]$, so that by equation (2.3), $\mu(F_\gamma)$ is differentiable and

$$\partial\mu(F_\gamma)/\partial\gamma = E[\delta(z,F_0)S(z)], \tag{2.8}$$

where the derivatives are evaluated at zero. As shown by Van der Vaart (1991), satisfaction of this equation for all regular parametric families is necessary for existence of a (regular; see Van der Vaart, 1991) $\sqrt{n}$-consistent estimator of $\mu(F_0)$. Additional regularity conditions are often needed to attain $\sqrt{n}$-consistency (i.e. equation (2.8) is only a necessary condition) as discussed in Bickel and Ritov (1988). Often these conditions come in the form of smoothness conditions for F_0.

The $\delta(z,F_0)$ term is referred to as the influence function, motivated by the expansion of equation (2.7) where $\delta(z_i,F_0)$ gives the first-order effect, or "influence," of an observation on the estimator $\hat{\mu}$. Also, we can think of $\delta(z,F_0)$ as the first-order effect on $\mu(F)$ of changing F, with $\int\delta(z,F_0)(F-F_0)(dz)$ in equation (2.3) being analogous to the differential in multivariate calculus, and the influence function $\delta(z,F_0)$ to the gradient.

Estimation of the influence function is important for the bias correction, so it is useful to have a way to calculate it for a given functional. One way is to look for $\delta(z,F_0)$ that solves equation (2.8). Often, the solution to this equation can be determined by manipulating $\partial\mu(F_\gamma)/\partial\gamma$ using properties of derivatives so that it has the expected product form in equation (2.8), and recovering $\delta(z,F_0)$ by inspection. For example, for $\mu(F) = \int f(z)^2dz$, differentiating under the integral gives $\partial\mu(F_\gamma)/\partial\gamma = \partial\int f_\gamma(z)^2dz/\partial\gamma = \int 2f_0(z)[\partial f_\gamma(z)/\partial\gamma]dz = E[2f_0(z)\partial\ln f_\gamma(z)/\partial\gamma] = E[2f_0(z)S(z)]$, so $\delta(z,F_0)$

$= 2f_0(z)$ satisfies equation (2.8), (a known result).

2.2 Implementing the Bias Correction

Implementing this bias correction in practice requires an estimator $\hat{\delta}(z)$ of the influence function. One general approach is to find the formula $\delta(z,F_0)$ and then plug in the estimator $\hat{F}$ to form $\hat{\delta}(z) = \delta(z,\hat{F})$. This was the approach followed to obtain the bias corrected estimator of the integrated squared density in equation (2.6). If the formula $\delta(z,F_0)$ is very complicated (e.g. as in Hausman and Newey, 1995), it might be more feasible to form $\hat{\delta}(z)$ by another method.

A method of forming $\hat{\delta}(z)$, that bypasses an explicit formula for $\delta(z,F)$, is available for linear density estimators, where $\hat{F}(z) = \int 1(u \leq z)\hat{f}(u)du$ and $\hat{f}(z) = \sum_{i=1}^{n}\kappa(z,z_i)/n$. For example, kernel estimators have this form for $\kappa(z,z_i) = h^{-r}K((z-z_i)/h)$, where h is a bandwidth parameter with dependence on sample size suppressed for notational convenience, r is the dimension of z, and $K(u)$ is a kernel function satisfying $\int K(u)du = 1$ and other properties. In this case a simple derivative calculation can be used for the bias correction. Let $\Delta_z(u) = \int 1(t \leq u)\kappa(t,z)dt$, so that $\hat{F}(\cdot) = \sum_{i=1}^{n}\Delta_{z_i}(\cdot)/n$. Let α denote a scalar. Suppose that equation (2.3) holds uniformly in F and F_0 and replace F by $\hat{F} + \alpha\Delta_z$ and F_0 by $\hat{F}$. Dividing through by α, and assuming $\|\Delta_z\|$ is finite with probability one, as $\alpha \longrightarrow 0$,

$$[\mu(\hat{F} + \alpha\Delta_z) - \mu(\hat{F})]/\alpha = \int\delta(u,\hat{F})\kappa(u,z)du + R(\alpha\Delta_z,\hat{F})/\alpha$$

$$|R(\alpha\Delta_z,\hat{F})/\alpha| = o(\alpha\|\Delta_z\|)/\alpha = o(\alpha)/\alpha = o(1).$$

Therefore, at $\alpha = 0$, $\partial\mu(\hat{F} + \alpha\Delta_z)/\partial\alpha = \int\delta(u,\hat{F})\kappa(u,z)du$. For many choices of $\kappa(u,z)$, $\hat{\delta}(z) = \int\delta(u,\hat{F})\kappa(u,z)du$ should be close to $\delta(z,\hat{F})$ in large samples, and so $\hat{\delta}(z)$ could be used to estimate the influence function. Then inserting this estimator in equation (2.5) and interchanging the order of differentiation and integration gives

$$\tilde{\mu} = \hat{\mu} + \partial \int \mu(\hat{F} + \alpha\Delta_z)(\hat{P}-\hat{F})(dz)/\partial\alpha|_{\alpha=0}.$$

Thus, a bias corrected estimator can be calculated by differentiating the difference of the average and integrated functionals with respect to a small increment Δ_z in $\hat{F}$. This derivative could even be calculated numerically.

The influence function estimator $\hat{\delta}(z) = \partial\mu(\hat{F} + \alpha\Delta_z)/\partial\alpha$ was developed in Newey (1994) for kernel estimators and applied in Hausman and Newey (1995) to solutions to differential equations. It generalizes the delta-method estimator of a function of sample means, in the sense that if F is an unknown constant (rather than a function) and $\Delta_z(u)$ did not depend on u, so that $\hat{F} = \sum_{i=1}^{n}\Delta_{z_i}/n$ is a sample mean, then $\hat{\delta}(z) = [\partial\mu(\hat{F})/\partial F]'\Delta_z$.

2.3 $\sqrt{n}$-Consistency and Undersmoothing

Sufficient conditions for $\sqrt{n}$-consistency of each estimator are that the corresponding remainder terms in equation (2.7) are $o_p(1)$. We will consider each of these remainder terms in sequence, beginning with R_n. The following condition allows us to bound the size of R_n.

Assumption 1: There is a set of functions $\mathcal{F}$ and a constant C such that $F_0, \hat{F} \in \mathcal{F}$ with probability approaching one and for all $F \in \mathcal{F}$,

$$\mu(F) = \mu(F_0) + \int \delta(z,F_0)(F-F_0)(dz) + R(F-F_0,F), \quad |R(F-F_0,F_0)| \leq C\|F-F_0\|^2.$$

This condition formalizes equation (2.3) plus imposes a requirement that the size of the remainder be $\|F-F_0\|^2$, which will hold when $\mu(F)$ is twice Frechet differentiable and F is close enough to F_0 (e.g. see Proposition 7.3.3 of Luenberger, 1969). Under this condition $R_n = o_p(1)$ will follow from $\|\hat{F}-F_0\| = o_p(n^{-1/4})$, i.e. $\hat{F}$ being more than $n^{1/4}$-consistent. This condition helps ensure that the nonlinearity remainder term

$R(F-F_0,F_0)$ is second order.

Consider next $\hat{B}_n = \int\delta(z)(\hat{F}-\hat{P})(dz)$. As previously discussed, $E[\hat{B}_n] = \sqrt{n}\int\delta(z)(\overline{F}-F_0)(dz) \longrightarrow 0$ may require that the bias of $\hat{F}$ shrink faster than $1/\sqrt{n}$. On the other hand, the variance of $\hat{B}_n$ should go to zero under weak conditions. For instance, for a linear density estimator, we have $Var(\hat{B}_n) \leq E[\{\int\delta(u)\kappa(u,z)du - \delta(z)\}^2]$, which will go to zero when $\int\delta(u)\kappa(u,z)du$ converges in mean-square to $\delta(z)$.

The remainder term D_n of equation (2.7) is more complicated. It is helpful to decompose it as

$$D_n = \tilde{S}_n + \tilde{T}_n, \quad \tilde{S}_n = \nu_n(\hat{\delta})-\nu_n(\delta), \quad \nu_n(d) = \int d(z)(\hat{P}-F_0)(dz)/\sqrt{n},$$

$$\tilde{T}_n = -\sqrt{n}\int[\hat{\delta}(z)-\delta(z)](\hat{F}-F_0)(dz).$$

The order of $\tilde{S}_n$ depends on the order of $\hat{\delta}(z)-\delta(z)$ and the modulus of continuity of the empirical process $\nu_n(d)$. Precise conditions for $\tilde{S}_n = o_p(1)$ are available in the literature on empirical processes, e.g. see Van der Vaart and Wellner (1996). Also, it may be possible to use the structure of $\hat{\delta}(z)$ to show $\tilde{S}_n = o_p(1)$ directly, as in the kernel estimator results of Newey and McFadden (1994). Generally $\tilde{S}_n = o_p(1)$ will not require undersmoothing, because $\tilde{S}_n$ is a second-order term. Also, $\tilde{T}_n$ is a second-order term, being $\sqrt{n}$ times the product of the remainder for $\hat{\delta}$ and $\hat{F}$, and so should not require undersmoothing to be $o_p(1)$.

Combining this analysis for the various remainder terms leads to conditions for $\sqrt{n}$-consistency of the estimators. If $\hat{F}$ converges faster than $n^{-1/4}$ in the norm $\|\cdot\|$ and $\sqrt{n}E[\hat{B}_n] \longrightarrow 0$ then $\hat{\mu}$ should be $\sqrt{n}$-consistent. When the order of $E[\hat{B}_n]$ is the same as the order of the pointwise bias, these conditions will require undersmoothing. In contrast, $\tilde{\mu}$ will be $\sqrt{n}$-consistent without undersmoothing if both $\hat{\delta}$ and $\hat{F}$ converge faster than $n^{-1/4}$ and $\tilde{S}_n = o_p(I)$.

To obtain more precise results we consider kernel estimators, where $\hat{F}$ has a density $\hat{f}(z) = \sum_{i=1}^{n} K_h(z-z_i)/n$ for $K_h(u) = K(u/h)/h^r$. We also restrict the domain of

$\mu(F)$ to F that are absolutely continuous with density f and specify the norm $\|F\|$ to be a Sobolev norm in the derivatives of f. Let z be r-dimensional, let λ denote a $r \times 1$ vector of nonnegative integers, $|\lambda| = \sum_{j=1}^{r} \lambda_j$, $z^{\lambda} = \prod_{j=1}^{r} (z_j)^{\lambda_j}$, $\partial^{\lambda} f(z) = \partial^{|\lambda|} f(z) / \partial^{\lambda_1} \cdots \partial^{\lambda_r}$, $\mathcal{Z}$ denote a compact set, and

$$\|F\| = \max_{|\lambda| \leq d} \sup_{z \in \mathcal{Z}} |\partial^{\lambda} f(z)|,$$

where d is a nonnegative integer that specifies the highest order derivative of f that affects the norm.

The norm depend on derivatives of f up to order d to allow for $\mu(F)$ to depend on derivatives of f up to this order. An example where $d > 0$ would be needed is for weighted average derivatives. We specify a supremum norm to make it relatively easy to show the Frechet differentiability hypothesis of Assumption 1 and because uniform convergence rates for kernel estimators are readily available, leading to rates of convergence for the remainder terms above.

When combined with Assumption 1 the compactness condition on $\mathcal{Z}$ means that $\mu(F)$ can only depend on the values of $f(z)$ for z in a compact set. This restriction will be satisfied if $\mu(F)$ has some fixed trimming built into it, or if $f_0(z)$ is zero outside some compact set and $K(u)$ has bounded support, so that $\hat{f}(z)$ also will be zero outside some (slightly larger) compact set. For example, for $\mu(F) = \int f(z)^2 dz$, Assumption 1 will be satisfied with $d = 0$ if $f_0(z)$ has compact support, $K(u)$ has bounded support, $\mathcal{Z}$ is chosen to be a large enough compact set containing the support of $f_0(z)$ in its interior. We have chosen to impose these types of conditions because they can apply to a wide variety of examples but still lead to relatively simple results that illustrate the bias correction.

The next regularity condition concerns the kernel.

Assumption 2: $\int K(u)du = 1$, $K(-u) = K(u)$, $K(u)$ has bounded support, for $s \geq 2$, $K(u)$ is differentiable of order d with Lipschitz d^{th} derivative, and $\int K(u)u^{\lambda}du = 0$ for all λ with $|\lambda| < s$.

The symmetry condition is not needed but is convenient. The bounded support condition for the kernel is imposed here to keep the conditions relatively simple. The last condition requires that the kernel be a higher order (bias reducing) kernel of *at least* order s. Because of this higher order kernel assumption the order of the bias in the kernel estimators of up to the d^{th} derivatives of $f_0(z)$ will be no larger than h^s, if $f_0(z)$ has at least $s+d$ derivatives. The next condition imposes these smoothness restrictions on f_0.

Assumption 3: $f_0(z)$ is continuously differentiable of order $s+d$ on $\mathbb{R}^r$ with bounded $s+d^{th}$ derivatives, for some $c > 0$ and all λ with $|\lambda| = s$, $\int \sup_{\|\Delta\| \leq c} |\partial^{\lambda} f(z+\Delta)| dz < \infty$.

Under Assumptions 2 and 3 the number s can be thought of as the minimum of the order of the kernel and a degree of smoothness for $f_0(z)$. If $K(u)$ is a bias reducing kernel of order b and $f_0(z)$ has $d+a$ continuous derivatives then Assumptions 2 and 3 will be satisfied with $s = \min\{b,a\}$.

Assumptions 1-3 are sufficient to obtain the large sample properties of the estimator $\hat{\mu}$.

Theorem 2.1: If Assumption 1-3 are satisfied and $h = h_n$ *such that* $nh^{r+2d}/\ln(n) \longrightarrow \infty$ *and* $h \longrightarrow 0$, *and* $\delta(z)$ *is continuous with probability one and bounded,*

$$R_n = O_p(\ln(n)/\sqrt{n}h^{r+2d} + \sqrt{n}h^{2s}), \quad \hat{B}_n = O_p(\sqrt{n}h^s) + o_p(1).$$

Also, if $\ln(n)/\sqrt{n}h^{r+2d} \longrightarrow 0$ *and* $\sqrt{n}h^s \longrightarrow 0$ *then* $\sqrt{n}(\hat{\mu}-\mu_0) = \sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + o_p(1)$.

The hypotheses of this result require undersmoothing. The optimal bandwidth (minimum mean-square error) for estimation of the d^{th} derivative of $f_0(z)$, when $f_0(z)$ has $d+s$ continuous bounded derivatives and the kernel is s^{th} order, is $h^* = n^{1/(r+2d+2s)}$, and $\sqrt{n}(h^*)^s = n^{1/2 - s/(r+2d+2s)}$ does not go to zero. Here h must be chosen smaller than h^* to have $\sqrt{n}h^s \longrightarrow 0$.

To obtain conditions for $\sqrt{n}$-consistency of the bias corrected estimator, it is essential to be specific about $\hat{\delta}(z)$. Here we assume that $\hat{\delta}(z) = \delta(z,\hat{F})$, where $\delta(z,F)$ satisfies certain smoothness conditions in F.

Assumption 4: There is a set of functions $\mathcal{F}$ such that $F_0 \in \mathcal{F}$ and for small enough h, $\hat{F} \in \mathcal{F}$, $\int 1(u \leq z)K_h(u-z_i)du \in \mathcal{F}$ with probability one. Also, there is $b(z)$ bounded, with $b(z) = 0$ for $z \notin \mathcal{Z}$, and $D(z,F)$ that is linear in F such that for $F \in \mathcal{F}$, $|\delta(z,F)-\delta(z,F_0)-D(z,F-F_0)| \leq b(z)\|F-F_0\|^2$, and $|D(z,F)| \leq b(z)\|F\|$.

This condition follows from second order Frechet differentiability of $\delta(z,F)$ in F, with bounded derivative. It is helpful in deriving the order of both the stochastic equicontinuity term $\tilde{S}_n$ and the nonlinear term $\tilde{T}_n$ in D_n. We use this assumption to obtain the order of $\tilde{S}_n$, rather than empirical process methods, because a direct proof for kernel estimators (as in Newey and McFadden, 1994) seems to allow for a wider class of functionals. In particular, empirical process methods for showing that $\tilde{S}_n \xrightarrow{p} 0$ rely heavily on smoothness of $\hat{\delta}(z)$ in z, that can be avoided by using Assumption 4. Also, when $\hat{\delta}(z)$ is linear in $\hat{F}$ (i.e. $b(z) = 0$ as for the average density), U-statistic theory gives $\tilde{S}_n \xrightarrow{p} 0$ under very weak conditions.

These conditions lead to the following result for the bias corrected estimator:

Theorem 2.2: If Assumptions 1 - 4 are satisfied, and $h = h_n$ *such that* $nh^{r+2d}/\ln(n) \longrightarrow \infty$ *and* $h \longrightarrow 0$, *then*

$$(2.9) \qquad D_n = O_p(\ln(n)/\sqrt{n}h^{r+2d} + \sqrt{n}h^{2s} + h^s).$$

Also, if $\ln(n)/\sqrt{n}h^{r+2d} \longrightarrow 0$ *and* $\sqrt{n}h^{2s} \longrightarrow 0$ *then* $\sqrt{n}(\tilde{\mu}-\mu_0) = \sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + o_p(1)$.

The upper bound on D_n obtained here will be smaller than the upper bound obtained for $\hat{B}_n$ in Theorem 2.1, for a range of bandwidths h. Also, if the bandwidth is chosen to be the mean-square minimizing value for estimation of the d^{th} derivative of $f_0(z)$, then the estimator will be $\sqrt{n}$-consistent for any value of s that is large enough so that there exists an h satisfying the conditions of this theorem. Specifically, existence of h such that $\ln(n)/\sqrt{n}h^{r+2d} \longrightarrow 0$ and $\sqrt{n}h^{2s} \longrightarrow 0$ requires $s > d + r/2$, and in this case the bandwidth which is pointwise optimal for estimation of the d^{th} derivative of $f_0(z)$, which is $n^{1/(r+2d+2s)}$, will satisfy the conditions for $\sqrt{n}$-consistency. This result then shows that $\sqrt{n}$-consistency of $\tilde{\mu}$ does not require undersmoothing of $\hat{f}$.

The conditions for $\sqrt{n}$-consistency of $\tilde{\mu}$ are weaker than the conditions for $\sqrt{n}$-consistency of $\hat{\mu}$, in the sense that they require less smoothness. Existence of h satisfying the bandwidth conditions of Theorem 2.1 requires $s > r + 2d$, while existence of h satisfying the conditions of Theorem 2.2 for $\sqrt{n}$-consistency requires only

$$(2.10) \qquad s > (r+2d)/2.$$

Thus, with the bias-corrected estimator f(z) is only required to have half the number of derivatives, as for the original estimator, or alternatively if $f_0(z)$ has all the derivatives that are needed, the kernel for $\tilde{\mu}$ need only be half the order of the kernel for $\hat{\mu}$ in order to attain $\sqrt{n}$-consistency. This improvement is achieved at the expense of smoothness of $\delta(z,F)$ as a function of F, as in Assumption 4.

3. Bootstrap Corrected Nonparametric Estimation

Another approach that also reduces the size of the bias in functional estimation and can remove the need to undersmooth, is to plug in a nonparametric estimator that has been corrected for smoothing. To motivate this approach, recall that the first order bias in $\mu(\hat{F})$ is the order of $E[\hat{F}]-F_0$. We could eliminate this bias if we could replace $\hat{F}$ by the empirical distribution, which is unbiased. Often this replacement is not possible because smoothing is required to bring $\hat{F}$ into the domain of $\mu(F)$ (e.g. when $\mu(F)$ depends on the density of F) and smoothing induces some bias. However, we can construct a smooth estimate of the smoothing effect and use it to partially correct for bias. To describe this correction, suppose that $\hat{F} = C\hat{P}$ for some transformation C with domain that includes the empirical distribution $\hat{P}$. Often C will be a smoothing transformation, such as convolution for kernel estimators, that depends on the sample size and gets close to the identity I as the sample size grows. Then for large samples $C\hat{F}-\hat{F}$ should be an estimate of the smoothing effect $C\hat{P}-\hat{P}$. Subtracting $C\hat{F}-\hat{F}$ from $\hat{F}$ gives

$$(3.1) \qquad \tilde{F} = \hat{F} - (C\hat{F} - \hat{F}) = 2\hat{F} - C\hat{F} = 2C\hat{P} - C^2\hat{P} = (2C-C^2)\hat{P}.$$

where C^2 is the composition of C with itself. This is a bootstrap correction for smoothing, in the sense $\hat{P}$ is replaced by $\hat{F}$ in the smoothing effect $C\hat{P}-\hat{P}$ to form the correction $C\hat{F}-\hat{F}$.

In this Section we will consider an estimator obtained by plugging in this smoothing corrected nonparametric estimator, giving $\tilde{\mu} = \mu(\tilde{F})$. It will be shown that this estimator may be $\sqrt{n}$-consistent without undersmoothing. This approach, of bootstrap correcting the distribution estimator and plugging it in, allows the same nonparametric estimator to attain the optimal nonparametric convergence rates and be plugged into functionals that attain the optimal $1/\sqrt{n}$ rate. To achieve this simultaneous optimality, the influence function will need to satisfy certain conditions that are

detailed below.

To understand the effect that the bootstrap correction has on the functional estimator, consider an expansion analogous to equation (2.4) with

(3.2) $\sqrt{n}(\tilde{\mu} - \mu_0) = \sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + \tilde{R}_n + \tilde{B}_n$, $\tilde{R}_n = \sqrt{n}R(\tilde{F}-F_0,F_0)$, $\tilde{B}_n = \sqrt{n}\int\delta(z)(\tilde{F}-\hat{P})(dz)$.

This expansion is the same as for $\mu(\hat{F})$ except that $\tilde{F}$ replaces $\hat{F}$, with $E[\tilde{B}_n]$ being the first order bias term. Suppose that C is linear, with $E[\tilde{F}] = E[(2C-C^2)\hat{P}] = (2C-C^2)F_0$. Also, suppose that there is a transformation $C^*\delta$ of $\delta(z)$ such that $\int\delta(z)CF(dz) = \int C^*\delta(z)F(dz)$ for $F = F_0$ and $F = CF_0$. With more structure C^* may be interpreted as the adjoint of C, as in examples given below. Then if the order of integration can be interchanged,

(3.3) $E[\tilde{B}_n] = \sqrt{n}\int\delta(z)[(2C-C^2-I)F_0](dz) = -\sqrt{n}\int\delta(z)[(I-C)^2F_0](dz)$

$= -\sqrt{n}\int[(I-C^*)\delta](z)[(I-C)F_0](dz)$.

Here $(I-C)F_0$ represents smoothing bias from the transformation C, small in large samples when C is close to I. The term $(I-C^*)\delta$ is an analogous term that should also be close to zero in large samples. Comparing equation (3.3) with $E[\hat{B}_n] = -\sqrt{n}\int\delta(z)[(I-C)F_0](dz)$ for $\hat{B}_n$ from equation (2.4), we see that using the bootstrap-corrected estimator $\tilde{F}$ leads to the replacement of δ by $(I-C^*)\delta$ in the integral, reducing the first order bias when $(I-C^*)\delta$ is close to zero. This is the bias complementarity effect referred to above, where the bias in $\hat{F}$ has been complemented by the influence function remainder $(I-C^*)\delta$.

Kernel estimators are an important example, where $\hat{F}$ has density $\hat{f}(z) = \int\kappa(z,u)\hat{P}(du)$ for $\kappa(z,u) = h^{-r}K((z-u)/h)$. Plugging in F for $\hat{P}$ in the formula for $\hat{f}$ leads to a transformation C where CF has density $\int\kappa(z,u)F(du)$. To describe the bootstrap corrected estimator $\tilde{F} = 2\hat{F}-C\hat{F}$, let $\tilde{K}(u) = 2K(u) - \int K(u-t)K(t)dt$ be the "twicing" kernel associated with $K(u)$. Then the density of $\tilde{F}$ will be

(3.4) $$\tilde{f}(z) = 2\hat{f}(z) - \int\kappa(z,u)\hat{f}(u)du = 2\hat{f}(z) - \int\int\kappa(z,u)\kappa(u,t)du\hat{P}(dt) = \int\tilde{\kappa}(z,u)\hat{P}(du),$$

$$\tilde{\kappa}(z,u) = 2\kappa(z,u) - \int\kappa(z,t)\kappa(t,u)dt = h^{-r}\tilde{K}((z-u)/h).$$

That is, the smoothing correction gives a kernel estimator with a twicing kernel that is constructed from the original kernel.

We use twicing kernel estimators to illustrate the bias correction, although they may not be the best choice of nonparametric estimator. A twicing kernel has order that is twice that of the original kernel and is not everywhere positive. Consequently, the density estimator $\tilde{f}(z)$ may not be everywhere positive, which may be an undesirable feature. Also, it is known that using higher order kernels may not improve mean-square error in most sample sizes of interest, e.g. see Marron and Wand (1992). This motivates a search for other bootstrap corrected estimators, as considered below.

To understand the bias formula in equation (3.3) for kernel estimators we need to find a transformation C^* with $\int C^*\delta(z)F(dz) = \int\delta(z)CF(dz)$. For any F with density f, CF has density $\int\kappa(z,u)f(u)du$, so that for $\bar{\delta}(z) = \int\kappa(u,z)\delta(z)du$,

$$\int\delta(z)CF(dz) = \int\delta(z)[\int\kappa(z,u)f(u)du]dz = \int\bar{\delta}(z)f(z)dz = \int C^*\delta(z)F(dz),$$

where $C^*\delta(z) = \bar{\delta}(z)$. Here C^* is the integral transform obtained by interchanging z and u in κ, which is known to be the adjoint under certain conditions (Luenberger, 1969, p. 153). This transformation is a convolution, with $\bar{\delta}(z) = \int K(u)\delta(z+hu)du$. Equation (3.3) now becomes

(3.5) $$E[\tilde{B}_n] = -\sqrt{n}\int[\bar{\delta}(z)-\delta(z)][\bar{f}(z)-f_0(z)]dz.$$

The bias effect of the bootstrap correction is to replace the integrated convolution bias $\int\delta(z)[\bar{f}(z)-f_0(z)]dz$ with the integral of the product of convolution biases in equation (3.5). The magnitude of the bias reduction will depend on the convolution bias

$\bar{\delta}(z)-\delta(z)$, which in turn depends on the smoothness of the influence function, as is well known from the kernel estimation literature. The following result makes these conditions precise.

Theorem 3.1: If Assumptions 1-3 are satisfied, $\delta(z)$ is continuously differentiable of order $t \leq s$ on $\mathbb{R}^r$ with bounded derivatives, $h = h_n$ such that $nh^{r+2d}/ln(n) \longrightarrow \infty$ and $h \longrightarrow 0$, then equation (3.2) is satisfied with $\hat{\mu} = \mu(\tilde{F})$ and $\tilde{B}_n = O_p(\sqrt{n}h^{s+t}+h^t)$. Also, if $ln(n)/\sqrt{n}h^{r+2d} \longrightarrow 0$ and $\sqrt{n}h^{s+t} \longrightarrow 0$ then $\sqrt{n}(\hat{\mu}-\mu_0) = \sum_{i=1}^n \psi(z_i)/\sqrt{n} + o_p(1)$.

The upper bound on $\tilde{B}_n$ is smaller than the bound on $\hat{B}_n$ in Theorem 2.1 because $\sqrt{n}h^s$ has been replaced by $\sqrt{n}h^{s+t}$. This replacement means that sufficient conditions for $\sqrt{n}$-consistency of $\mu(\tilde{F})$ are different than for the original plug-in estimator $\hat{\mu}$. If the bandwidth is chosen so the dominating remainder terms $ln(n)/\sqrt{n}h^{r+2d}$ and $\sqrt{n}h^{s+t}$ are asymptotically proportional then $\mu(\tilde{F})$ will be $\sqrt{n}$-consistent if

(3.6) $\quad s + t > r + 2d.$

This condition allows some tradeoff of smoothness of $f_0(z)$ and $\delta(z)$ for attaining $\sqrt{n}$-consistency.

This estimator also attains $\sqrt{n}$-consistency without undersmoothing if the influence function is smooth enough. Consider the case where $f_0(z)$ is only s times differentiable, so that the order of the pointwise bias in $\tilde{f}$ is h^s. Choose the bandwidth so that the squared bias order h^{2s} is proportional to the order $n^{-1}h^{-r-2d}$ of the pointwise variance. Then the conditions for $\sqrt{n}$-consistency are satisfied if

(3.7) $\quad t > r/2 + d.$

The smoothing bias correction depends on smoothness of the influence function in z, in contrast to the bias correction of Section 2, that depends on smoothness of the influence function in F. The bias correction of Section 2 may still give $\sqrt{n}$-consistency

even when $\delta(z)$ is not smooth in z. For example, the functional $\mu(F) = \int 1(a \le z \le b) f(z)^2 dz$ has $\delta(z) = 1(a \le z \le b) 2 f_0(z)$ that is discontinuous when the density is positive at a or b, and so does not satisfy the conditions of Theorem 3.1 for any t. Nevertheless, $\hat{\delta}(z) = 1(a \le z \le b) 2 \hat{f}(z)$ is an influence function estimator that would satisfy the conditions of Theorem 2.2, so that the bias corrected estimator of equation (2.5) could attain $\sqrt{n}$-consistency without undersmoothing.

It is interesting to compare the properties of a plug-in estimator based on a twicing kernel with one based on another kernel of the same order as the twicing kernel. For simplicity suppose that $d = 0$. Consider a plug-in estimator with an ordinary kernel where $s > r$ and the bandwidth is chosen to give $\sqrt{n}$-consistency. The plug-in estimator with a twicing kernel of the same order will also be $\sqrt{n}$-consistent for the same bandwidth choice, if $t = s > r/2$. That is, $\sqrt{n}$-consistency with a twicing kernel only requires $f_0(z)$ to be half as smooth, if $\delta(z)$ also is smooth enough. Also, the same bandwidth no longer has to involve undersmoothing, because only half as many derivatives of the density are needed to exist, and the optimal bandwidth for estimating a density with fewer derivatives will be smaller.

There are other nonparametric estimators that have the same functional bias reduction property as twicing kernel estimators. A particularly important class are those where the smoothing transformation is idempotent, with $C\hat{F} = \hat{F}$. Here $\tilde{F} = 2\hat{F} - C\hat{F} = 2\hat{F} - \hat{F} = \hat{F}$, so the bootstrap smoothing correction is "built into" $\hat{F}$. Our results then indicate that undersmoothing may not be needed in this case.

One idempotent nonparametric estimator is an orthogonal series density estimator. Let $(p_j(u), j = 1, 2, \ldots)$ be a sequence of functions that are orthonormal with respect to Lebesgue measure on $\mathbb{R}^r$, i.e. $\int p_j(u) p_k(u) du = 1$ if $j = k$ and equal to zero otherwise. Let $p^J(u) = (p_1(u), \ldots, p_J(u))'$ and $\hat{\alpha} = \sum_{i=1}^n p^J(z_i)/n$. Then an orthogonal series estimator is

$$\hat{f}(z) = p^J(z)'\hat{\alpha} = \int \kappa(z,u)\hat{P}(du), \quad \kappa(z,u) = p^J(z)'p^J(u).$$

Like the kernel estimator $\hat{f}$ has the linear form $\int \kappa(z,u)\hat{P}(du)$, but $\kappa(z,u)$ is now an inner product of orthonormal functions rather than a kernel. To see the effect of the bootstrap correction, plug in $\hat{F}$ for $\hat{P}$ in the density formula to obtain

$$\int \kappa(z,u)\hat{f}(u)du = \int [\int \kappa(z,u)\kappa(u,t)du]\hat{P}(dt)$$

$$= \int p^J(z)'[\int p^J(u)p^J(u)'du]p^J(t)\hat{P}(dt) = \int \kappa(z,t)\hat{P}(dt) = \hat{f}(z),$$

so that the transformation C is idempotent. Note that $\bar{f}(z) = E[\hat{f}(z)] = p^J(z)'\int p_J(u)f_0(u)du$ is the minimum integrated squared error (ISE) approximation to $f_0(z)$ and $\bar{\delta}(z) = \int \delta(u)\kappa(u,z)du = p^J(z)'\int \delta(u)p^J(u)du$ is the minimum ISE approximation to $\delta(z)$. Then by $\int \bar{\delta}(z)[f_0(z)-\bar{f}(z)]dz = 0$,

$$E[\hat{B}_n] = \sqrt{n}\int \delta(z)[\bar{f}(z)-f_0(z)]dz = -\sqrt{n}\int [\bar{\delta}(z)-\delta(z)][\bar{f}(z)-f_0(z)]dz.$$

Here the bias term that appears to be first order is actually second order, a result of the bias correction being built into $\hat{f}$. Consequently, the bias of the functional estimator can shrink faster than the pointwise bias, removing the need to undersmooth.

This example can be made precise by specifying an ISE rate of approximation for $f_0(z)$ and $\delta(z)$, leading to the following result:

Theorem 3.2: If Assumption 1 is satisfied, $f_0(z)$ is bounded, $\{\int [\bar{\delta}(z)-\delta(z)]^2dz\}^{1/2} = O(J^{-t/r})$, and $\{\int [\bar{f}(z)-f_0(z)]^2dz\}^{1/2} = O(J^{-s/r})$ then equation (2.4) is satisfied and

$$\hat{B}_n = \sum_{i=1}^n [\bar{\delta}(z_i)-\delta(z_i)]/\sqrt{n} = O_p(J^{-t/r} + \sqrt{n}J^{-s/r-t/r}).$$

Furthermore, if $\sqrt{n}J^{-s/r-t/r} \longrightarrow 0$ and $\sqrt{n}\|\hat{F}-F_0\|^2 \xrightarrow{p} 0$, then $\sqrt{n}(\hat{\mu}-\mu_0) = \sum_{i=1}^n \psi(z_i)/\sqrt{n} + o_p(1)$.

The hypotheses of this result are not very primitive, but are consistent with the known rate s/r for approximation by orthogonal polynomials, where s is the number of continuous derivatives that exist. The conditions are specific enough to see that undersmoothing will not be required for $\sqrt{n}$-consistency. Specifically, $\sqrt{n}\|\hat{F}-F_0\|^2 \xrightarrow{p} 0$ will hold if $\hat{f}$ converges slightly faster than $n^{-1/4}$, which will not require undersmoothing. Also, the mean-square bias of $\hat{f}$ is $O(J^{-s/r})$ by hypothesis, but the conditions only require that $\sqrt{n}J^{-s/r-t/r} \longrightarrow 0$. Therefore if t is large enough the bias could be allowed to shrink at the same rate as the standard deviation without affecting consistency. For instance, if $t \geq s$ then $n^{1/4}J^{-s/r} \longrightarrow 0$, meaning the bias shrinks faster than $n^{-1/4}$, will suffice for $\sqrt{n}J^{-s/r-t/r} \longrightarrow 0$, and should be implied by $\hat{f}$ converging faster than $n^{-1/4}$. We can obtain more primitive conditions in the case of specific functionals. For example, the average density estimator $\mu(\hat{f}) = \int \hat{f}(z)^2 dz$ is $\sqrt{n}$-consistent if $\{\int[\bar{f}(z)-f_0(z)]^2 dz\}^{1/2} = O(J^{-s/r})$ and $\sqrt{n}(J/n + J^{-2s/r}) \longrightarrow 0$.

So far we have only considered the case where the smoothing transformation C is linear. When C is nonlinear analogous results should also hold: A bootstrap smoothing correction may remove the need for undersmoothing when the influence function has enough derivatives, and this correction will be built into estimators that are idempotent transformations of the empirical distribution. These results could be shown by including an expansion of C in the analysis. However, this would greatly complicate the analysis, so we content ourselves with pointing out some existing examples of nonlinear, idempotent transformations where it is known that undersmoothing is not needed.

Newey (1994) showed that undersmoothing is not needed for functionals of a series estimator of a conditional expectation. This occurs because series estimators of conditional expectations are idempotent, leading to a corresponding idempotent distribution estimator. For brevity, we omit details. Shen (1997) has also shown that undersmoothing may not be needed for sieve density estimators when the influence function is smooth enough. This also corresponds to an idempotent transformation. We can describe a sieve estimator as the solution to

$$\hat{f} = \text{argmax}_{f\in\mathcal{B}}\sum_{i=I}^{n}\ln f(z_i)/n = \text{argmax}_{f\in B}\int \ln f(z)\hat{P}(dz),$$

where B is some restricted class of densities. Sieve estimators are of this form, where B is some parametric family that also imposes other restrictions, such as boundedness of higher order derivatives. It follows from the information inequality that

$$\hat{f} = \text{argmax}_{f\in B}\int [\ln f(z)]\hat{f}(z)dz,$$

i.e. if we replace $\hat{P}$ by $\hat{F}$ in the transformation that gives $\hat{F}$ we obtain $\hat{F}$ again. Thus, $C\hat{F}$ is idempotent, so the smoothing correction is built into a sieve estimator.

Higher order bootstrap smoothing corrections could also be carried out. Consider

$$\tilde{F}_L = [I-(I-C)^L]\hat{P},$$

where L is a positive integer. We have $\tilde{F}_1 = \hat{F}$ and $\tilde{F}_2 = \tilde{F}$, while $L > 2$ correspond to higher order bootstrap corrections. The estimator $\mu(\tilde{F}_L)$ will have an expansion like equation (3.2), with $\tilde{F}_L$ replacing $\tilde{F}$. Assuming C is linear as before, and letting j be any integer, $0 \le j \le L$, the first-order bias term will be

$$-\sqrt{n}\int \delta(z)[(I-C)^L F_0](dz) = -\sqrt{n}\int [(I-C^*)^j\delta](z)[(I-C)^{L-j}F_0](dz).$$

This represents a higher-order bias complementarity, where some of the smoothing bias is shifted from F_0 to δ. Of course, if C is idempotent these higher-order bootstrap corrections are built into the estimator, i.e. $\tilde{F}_L = \hat{F}$.

We also note that it is possible to bootstrap correct the functional, forming an estimator as

$$\bar{\mu} = \mu(\hat{F}) - [\mu(C\hat{F})-\mu(\hat{F})] = 2\mu(\hat{F}) - \mu(C\hat{F}).$$

If the functional were linear then $\bar{\mu} = \mu(\tilde{F})$, i.e. the estimator is the same

whether we bootstrap correct the functional or the density. In the general nonlinear case the first-order linear terms in the expansion of equation (2.3) would be the same for both estimators, so that $\hat{\mu}$ and $\bar{\mu}$ should have the same first-order bias.

4. Linear Kernel Averages

Many important estimators are averages of functions of nonparametric estimators and data observations. Examples include the well known average density estimator $\sum_{i=1}^{n}\hat{f}(x_i)/n$ and the weighted average derivative estimator of Powell, Stock, and Stoker (1989). The bias corrected estimation results can be extended to cover this case, and we do so in this Section and the following one.

Consider a parameter of interest

$$\mu_0 = E[g(z,F_0)] = \int g(z,F_0)F_0(dz),$$

where $g(z,F)$ is some known function, that may have a restricted domain as a function of F. One way to estimate μ_0 is to plug a nonparametric estimator $\hat{F}$ into $g(z,F)$ and integrate over the empirical distribution to obtain

$$\hat{\mu} = \int g(z,\hat{F})\hat{P}(dz) = \sum_{i=1}^{n} g(z_i,\hat{F})/n. \tag{4.1}$$

One could also use $\hat{F}$ in place of $\hat{P}$, although the estimator $\hat{\mu}$ is often computationally simpler.

To understand how a bias correction should be constructed for this estimator, let $\mu(F) = E[g(z,F)] = \int g(z,F)F_0(dz)$. Then

$$(4.2) \qquad \hat{\mu} = \int g(z,F_0)\hat{P}(dz) + \mu(\hat{F})-\mu(F_0) + \hat{S}_n, \quad \hat{S}_n = \int [g(z,\hat{F})-g(z,F_0)](\hat{P}-F_0)(dz).$$

Here $\hat{S}_n$ is a stochastic equicontinuity term that is second order, so to first order $\hat{\mu}$ is $\int g(z,F_0)\hat{P}(dz) + \mu(\hat{F})-\mu(F_0)$. The term $\int g(z,F_0)\hat{P}(dz)$ is unbiased for μ_0, but the expectation of $\mu(\hat{F})-\mu(F_0)$ may depart from zero due to smoothing inherent in $\hat{F}$ and to nonlinearity in $\hat{F}$. Therefore, to bias correct $\hat{\mu}$ we need to bias correct for the functional $\mu(\hat{F})$. This correction can be constructed by applying the analysis of Section 2. Suppose that $\mu(F) = \int g(z,F)F_0(dz)$ satisfies equation (2.3) and has influence function $\delta(z) = \delta(z,F_0)$ and let $\hat{\delta}(z)$ denote an estimator of $\delta(z)$. The bias correction is then $\int \hat{\delta}(z)(\hat{P}-\hat{F})(dz)$ and the corresponding estimator is

$$(4.3) \qquad \tilde{\mu} = \hat{\mu} + \int \hat{\delta}(z)(\hat{P}-\hat{F})(dz) = \hat{\mu} + \sum_{i=1}^{n} \hat{\delta}(z_i)/n - \int \hat{\delta}(z)\hat{F}(dz).$$

This correction is the same as in Section 2 except that $\hat{\delta}(z)$ estimates the influence function of $\int g(z,F)F_0(dz)$.

It is also possible to form a bias corrected estimator by applying the analysis of Section 3. Plugging in a bootstrap corrected nonparametric estimator $\tilde{F}$ in $g(z,F)$ gives

$$(4.4) \qquad \tilde{\mu} = \int g(z,\tilde{F})\hat{P}(dz) = \sum_{i=1}^{n} g(z_i,\tilde{F})/n.$$

It can be shown by results analogous to those of Section 2 and Section 3 that the need for undersmoothing can be removed by both approaches to bias correction. For brevity we omit this general analysis, that is a special case of Section 5. We focus here on the linear case, where $g(z,F)$ is linear in F and $\hat{F}$ is linear, where we can derive asymptotic mean-square error results. This allows us to quantify the variance effect of the bias reduction, and includes several interesting examples.

We consider $g(z,F) = v \cdot \partial^{\lambda}\{E_F[y|x]f(x)\}$ where v and y are not elements of x. One example is the average density, where $v = y = 1$ and $\lambda = 0$. Another is a density weighted average derivative, where λ is a unit vector and $y = 1$, so that by integration by parts $\mu_0 = E[v \cdot \partial^{\lambda} f_0(x)] = E[E[v|x]\partial^{\lambda} f_0(x)] = -E[f_0(x)\partial^{\lambda} E[v|x]]/2$. A third example is a density weighted conditional covariance, where $\mu_0 = E[vE[y|x]f_0(x)]$. The estimators are obtained by substituting a kernel estimator for $\partial^{\lambda}\{E_F[y|x]f(x)\}$ and averaging over z_i. They have the form

$$(4.5) \qquad \hat{\mu} = \sum_{i=1}^{n} v_i[\partial^{\lambda}\sum_{j=1}^{n} K_h(x_i - x_j)y_j/n]/n = \sum_{i=1}^{n}\sum_{j=1}^{n} \partial^{\lambda} K_h(x_i - x_j)v_i y_j/n^2.$$

These types of estimators have been considered by Hall and Marron (1987), Hardle and Tsybakov (1993), Powell and Stoker (1997), and others. Our contribution is to show how the bias complementarity affects the mean-square error of these estimators.

The estimator $\hat{\mu}$ has precisely the form in equation (4.1) for a certain $\hat{F}$. To describe this form suppose that $z = (x,w)$ where w includes y and v, and let

$$(4.6) \qquad \hat{F}(\bar{z}) = n^{-1}\sum_{i=1}^{n} 1(w_i \leq \bar{w}) \int 1(x \leq \bar{x}) K_h(x - x_i)dx.$$

The corresponding marginal density of x is the kernel estimator $\hat{f}(x) = \sum_{i=1}^{n} K_h(x - x_i)/n$. Also, for any function $a(w)$ the estimator of $E[a(w)|x]$ is $n^{-1}\sum_{i=1}^{n} a(w_i)K_h(x - x_i)/\hat{f}(x)$. Therefore, the estimator of $\partial^{\lambda}\{E_F[y|x]f(x)\}$ is $\partial^{\lambda}\sum_{j=1}^{n} K_h(x - x_j)y_j/n$, so that $\int g(z,\hat{F})\hat{P}(dz)$ is equal to $\hat{\mu}$ in equation (4.5).

It turns out that with a symmetric kernel and a certain $\hat{\delta}(z)$, the bias corrected estimator of equation (4.3) is identical to the bootstrap corrected estimator of equation (4.4) based on the corresponding twicing kernel. To describe this result, apply Newey (1994) to $\mu(F) = E[v \cdot \partial^{\lambda}\{E_F[y|x]f(x)\}]$ to obtain the influence function formula

$$(4.7) \qquad \delta(z) = \delta(z,F_0), \quad \delta(z,F) = \ell(x,F)y, \quad \ell(x,F) = (-1)^{|\lambda|}\partial^{\lambda}\{E_F[v|x]f(x)\}.$$

An estimator of this $\delta(z)$ obtained by plugging in the $\hat{F}$ from equation (4.6) is

$$\hat{\delta}(z) = \delta(z,\hat{F}) = \ell(x,\hat{F})y = (-1)^{|\lambda|}\partial^{\lambda}\{\sum_{i=1}^{n} K_h(x-x_i)v_i/n\}y = \partial^{\lambda}\{\sum_{i=1}^{n} K_h(x_i-x)v_i/n\}y,$$

where the last equality follows by symmetry of the kernel. Using this equality the bias corrected estimator of equation (4.3) will be

$$(4.8) \qquad \tilde{\mu} = \hat{\mu} + \sum_{i=1}^{n}\partial^{\lambda}\{\sum_{j=1}^{n} K_h(x_j-x_i)v_j/n\}y_i/n - \int \ell(x,\hat{F})y\hat{F}(dz)$$

$$= 2\hat{\mu} - n^{-1}\sum_{i=1}^{n} y_i \int \ell(x,\hat{F})K_h(x-x_i)dx = 2\hat{\mu} - n^{-1}\sum_{i=1}^{n} y_i \int \partial^{\lambda}\{\sum_{j=1}^{n} K_h(x_j-x)v_j/n\}K_h(x-x_i)dx$$

$$= 2\hat{\mu} - \sum_{i=1}^{n}\sum_{j=1}^{n}\partial^{\lambda}[\int K_h(x_j-u)K_h(u-x_i)du]y_i v_j/n^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\partial^{\lambda}\tilde{K}_h(x_i-x_j)v_i y_j/n^2.$$

where $\tilde{K}(u) = 2K(u) - \int K(u-t)K(t)dt$ is the twicing kernel corresponding to $K(u)$. This estimator has the same form as in equation (4.5), except K is replaced by $\tilde{K}$. As noted before, equation (4.5) is the same as equation (4.1) for $\hat{F}$ in equation (4.6), so equation (4.8) will be the same as equation (4.4) when $\tilde{F}$ is as given in equation (4.6) with the twicing kernel $\tilde{K}$ used in place of K. Thus, equation (4.8) shows that the bias corrected estimator of equation (4.3) is the same as the bootstrap corrected estimator of equation (4.4) obtained by using a twicing kernel.

We could proceed to derive asymptotic mean-square error expressions for the estimator in equation (4.8), but it turns out that a mean-square error improvement can be obtained by deleting the "own observation" terms in $\tilde{\mu}$ and normalizing by the total number of terms in the sum (see Jones and Sheather, 1991). This modification leads to the estimator

$$(4.9) \qquad \tilde{\mu}_c = \sum_{i\neq j}\partial^{\lambda}\tilde{K}_h(x_i-x_j)v_i y_j/n(n-1).$$

We will focus on results for this estimator because its "cross-validated" form is common in the literature and it has smaller asymptotic mean square error than $\tilde{\mu}$. Inclusion of the own observation would affect the results by adding a term

$$\partial^{\lambda}\tilde{K}_h(0)\sum_{i=1}^{n} y_i v_i/n(n-1) = [\partial^{\lambda}\tilde{K}(0)/(nh^{r+|\lambda|})](\sum_{i=1}^{n} y_i v_i/(n-1)).$$

The order of this term will be $1/(nh^{r+|\lambda|})$, which is larger than the terms that appear in the asymptotic mean-square error derived below for $\tilde{\mu}_c$.

Some additional assumptions are useful in the mean-square error calculations. Let $\mu_{yy}(x) = E[y^2|x]f_0(x)$, $\mu_{ww}(x) = E[v^2|x]f_0(x)$, and $\mu_{wy}(x) = E[vy|x]f_0(x)$.

Assumption 5: $\mu_{ww}(x)$ and $\mu_{wy}(x)$ are continuous and for some $c > 0$, $\int \sup_{\|\Delta\| \le c} \mu_{yy}(x+\Delta)\mu_{ww}(x)dx$ and $\int \sup_{\|\Delta\| \le c} |\mu_{wy}(x+\Delta)\mu_{wy}(x)|dx$ are finite.

This hypothesis is useful for controlling the variance of the estimator. The next two conditions help to determine the bias.

Assumption 6: $a(x) = \partial^{\lambda}\{E[y|x]f_0(x)\}$ and $b(x) = E[v|x]f_0(x)$ exist and are continuously differentiable of order s and $t \le s$ respectively, and for all multi-indices $\tilde{\lambda}$ and $\bar{\lambda}$ with $|\tilde{\lambda}| = s$ and $|\bar{\lambda}| = t$, there is a $c > 0$ such that $\int \sup_{\|\Delta\| \le c} |\partial^{\tilde{\lambda}} a(x+\Delta)| \sup_{\|\Delta\| \le c} |\partial^{\bar{\lambda}} b(x+\Delta)|dx < \infty$.

When combined with Assumption 2 this means s is the minimum of the number of derivatives of $a(x)$ that exist and the order of the kernel.

The next assumption imposes some further useful smoothness and moment conditions.

Assumption 7: $\ell(x) = (-1)^{|\lambda|}\partial^{\lambda}b(x)$ exists and is continuous and $a(x)$ and $\ell(x)$ are bounded, $E[y^2] < \infty$, and $E[v^2] < \infty$.

Under these conditions we can obtain the asymptotic mean-square error of the estimator $\tilde{\mu}_c$. Let $\psi(z) = a(x)v + \ell(x)y - E[a(x)v+\ell(x)y]$ be the influence function of $\int g(z,F)F(dz)$, $\zeta_{\lambda} = \int K(u)u^{\lambda}du$,

$$(4.10) \qquad Q = \{\int [\partial^{\lambda}\tilde{K}(u)]^2 du\}\int\{\mu_{yy}(x)\mu_{ww}(x)+(-1)^{|\lambda|}\mu_{wy}(x)^2\}dx,$$

$$B = \sum_{|\lambda|=s,\,|\tilde{\lambda}|=t} \zeta_\lambda \zeta_{\tilde{\lambda}} \int \partial^\lambda a(x) \partial^{\tilde{\lambda}} b(x) dx/(s!t!).$$

Theorem 4.1: If Assumptions 2 and 5 - 7 are satisfied then as $n \longrightarrow \infty$ *and* $h \longrightarrow 0$,

$$(4.11) \qquad MSE(\tilde{\mu}_c) = n^{-1} Var[\psi(z)] + n^{-2} h^{-r-2|\lambda|} Q + h^{2s+2t} B^2 + o(h^{2s+2t} + n^{-2} h^{-r-2|\lambda|} + n^{-1}).$$

The first term in this expression is the usual asymptotic variance under $\sqrt{n}$-consistency that will dominate if the other terms go to zero. The other two terms are variance and bias terms from kernel estimation. The estimator will be $\sqrt{n}$-consistent, with the usual asymptotic variance term dominating, if $n^{-1} h^{-r-2|\lambda|} \longrightarrow 0$ and $nh^{2s+2t} \longrightarrow 0$, meaning the pointwise variance of the kernel estimator goes to zero and the product of pointwise biases for $a(x)$ and $b(x)$ shrink faster than $1/\sqrt{n}$. If the bandwidth is chosen to balance the variance and bias terms, so that $n^{-1} h^{-(r/2)-|\lambda|}$ and h^{s+t} are asymptotically proportional, $\tilde{\mu}_c$ will be $\sqrt{n}$-consistent if

$$(4.12) \qquad s + t \geq r/2 + |\lambda|.$$

This condition is weaker than the requirement of Sections 2 and 3 that is made possible by the linearity of $g(z,F)$ in F and the absence of the own observation term.

Equation (4.12) allows equality rather than the strict inequality of equations (2.10) and (3.6), that is possible because uniform convergence rates for $\hat{f}$ are not used here. However $s + t = r/2 + |\lambda|$ is a knife edge case where the variance remainder term in equation (4.11) will be Qn^{-1}, the same size as the leading term. In this case $\sqrt{n}(\tilde{\mu}_c - \mu_0)$ will not be asymptotically normal with variance $Var(\psi(z))$, so the usual functional asymptotic theory does not apply.

We can compare the MSE in Theorem 4.1 to the MSE for the estimator based on the original kernel to evaluate the effect of the bias correction. Let $\overline{B} = \sum_{|\lambda|=s} \zeta_\lambda \int b(x) \partial^\lambda a(x) dx/(s!)$. Then by Powell and Stoker (1997), the estimator $\hat{\mu}_c$

obtained from equation (4.9) by replacing $\tilde{K}(u)$ by $K(u)$ has

$$(4.13) \qquad MSE(\hat{\mu}_c) = n^{-1}Var[\psi(z)] + n^{-2}h^{-r-2|\lambda|}\{\int[\partial^\lambda K(u)]^2du/\int[\partial^\lambda \tilde{K}(u)]^2du\}Q$$
$$+ h^{2s}\bar{B}^2 + o(h^{2s} + n^{-2}h^{-r-2|\lambda|} + n^{-1}).$$

The comparison of the MSE of $\hat{\mu}_c$ and $\tilde{\mu}_c$ is partly analogous to a comparison between the pointwise MSE of the corresponding kernel estimators. The ratio of kernel variance terms for $\hat{\mu}_c$ and $\tilde{\mu}_c$ is $\int[\partial^\lambda K(u)]^2du/\int[\partial^\lambda \tilde{K}(u)]^2du$, which is exactly the same as the ratio of pointwise variance terms for estimation $\partial^\lambda f_0(x)$ at a point, with $\int[\partial^\lambda \tilde{K}(u)]^2du$ known to be larger than $\int[\partial^\lambda K(u)]^2du$. On the other hand, because of the bias complementarity, the bias term is not reduced in exactly the same way as for pointwise estimation. The constant in the bias term depends on the t^{th} derivatives of $b(x)$ rather than the $2s^{th}$ order derivatives of $a(x)$. This bias reduction is better in some ways than the pointwise bias reduction from a higher order kernel. In particular, it may require no additional derivatives to exist when smoothness of $a(x)$ implies smoothness of $b(x)$ (e.g. as for the average density).

The asymptotic mean-square error (MSE) given in Theorem 4.1 can be used to obtain an optimal bandwidth formula for estimation of μ_0, when $t = s$ and the kernel is order s, with $B \neq 0$. Minimizing the sum of the second and third terms in equation (4.13) over h gives

$$(4.14) \qquad h = [Q(r+2|\lambda|)/(B^2 4sn^2)]^{1/(4s+r+2|\lambda|)}.$$

This bandwidth is optimal, in the sense of minimizing the leading terms in the MSE of the estimator.

It is interesting to note that although undersmoothing is not required for $\sqrt{n}$-consistency, undersmoothing is still optimal for $\hat{\mu}$. The optimal bandwidth converges faster to zero than a bandwidth that minimizes the asymptotic mean-square error of $\hat{a}(x)$. This feature seems to be specific to linear functional estimators with the own

observation deleted. If the own observation were included then there would be an $n^{-2}h^{-2r-4|\lambda|}$ term in the MSE, which dominates the $n^{-2}h^{-r-2|\lambda|}$ term, and will make the rate for the optimal bandwidth for functional estimation the same as for nonparametric estimation. Also, when the functional is nonlinear there is an additional remainder term $R(\tilde{F}-F_0,F_0)$ as discussed in Section 3, which is of order $\|\tilde{F}-F_0\|^2$ under Assumption 1. This should make the MSE of no smaller order than $\|\hat{F}-F_0\|^2 = O_p(\max\{n^{-1}h^{-r-2|\lambda|},h^{2s}\})$, where undersmoothing would not be optimal.

The optimal bandwidth formula in equation (4.14) is potentially useful for choosing the bandwidth in practice. The constants Q and B are unknown, but could be estimated by replacing the unknown nonparametric terms in their formulae by kernel estimates to obtain $\hat{Q}$ and $\hat{B}$, which could then be used in place of Q and B in the optimal bandwidth formula to form an estimate.

An interesting example is the density weighted average derivative estimator described above, where $|\lambda| = 1$ and $y = 1$. Here the influence function for $\hat{\mu}_c$ will be $\psi(z) = \partial^\lambda f_0(x)\{v-E[v|x]\} - f_0(x)\partial^\lambda E[v|x] - \mu_0$, as derived in Powell, Stock, and Stoker (1989). Suppose that $f_0(x)$ is $s+1$ times differentiable and $E[v|x]$ is t times differentiable, that the original kernel has order at least s, and all the following integrals exist, and let

$$Q = \{\int[\partial^\lambda \tilde{K}(u)]^2du\}\int Var(v|x)f_0(x)^2dx$$

$$B = \sum_{|\bar{\lambda}|=s,|\tilde{\lambda}|=t}\zeta_{\bar{\lambda}}\zeta_{\tilde{\lambda}}\int\partial^{\bar{\lambda}}[\partial^\lambda f_0(x)]\partial^{\tilde{\lambda}}\{E[v|x]f_0(x)\}dx/s!t!.$$

Then, the conclusion of Theorem 4.1 implies that as $n \longrightarrow \infty$ and $h \longrightarrow 0$,

$$(4.15)\qquad MSE(\tilde{\mu}_c) = Var[\psi(z)]/n + n^{-2}h^{-r-2}Q + h^{2s+2t}B^2 + o(n^{-1}+n^{-2}h^{-r-2}+h^{2s+2t}).$$

In this example the optimal bandwidth with $s = t$ is

$$h = [Q(r+2)/(B^2 4sn^2)]^{1/(4s+r+2)}$$

For example, with a normal kernel, $s = t = 2$, $r < 6$, and the bandwidth chosen optimally, the bias corrected weighted average derivative will be $\sqrt{n}$-consistent with the asymptotic variance $Var(\psi(z))$. If $r = 6$ then the estimator would still be $\sqrt{n}$-consistent but the leading term in equation (4.15) would not dominate, so that the limiting distribution would not be the usual one for a semiparametric estimator. In comparison with Powell, Stock, and Stoker (1989) this result imposes weaker conditions on the smoothness of the density at the expense of requiring some smoothness of the regression $E[v|x]$.

The use of a twicing kernel can lead to an improvement even when the density is not differentiable. To illustrate, consider the case where $\lambda = 0$ and $a(x) = E[y|x]f_0(x)$ and $b(x) = \ell(x) = E[w|x]f_0(x)$ satisfy a Holder condition:

Assumption 8: There are ξ, $t > 0$, $C_a(x)$, and $C_b(x)$ such that $\int C_a(x)C_b(x)dx < \infty$ and for all $\|\tilde{x}-x\|$ small enough, $|a(\tilde{x})-a(x)| \leq C_a(x)\|\tilde{x}-x\|^{\xi}$ and $|b(\tilde{x})-b(x)| \leq C_b(x)\|\tilde{x}-x\|^{t}$.

With this condition replacing Assumption 6 the following result holds.

Theorem 4.2: If Assumptions 2, 5, 7, and 8 are satisfied then as $n \longrightarrow \infty$ and $h \longrightarrow 0$

$$MSE(\tilde{\mu}_c) = Var[\psi(z)]/n + O(n^{-2}h^{-r}) + O(h^{2\xi+2t}).$$

The order of the bias is again the product of pointwise bias orders h^{ξ} and h^{t}. If the bandwidth is chosen so that $n^{-2}h^{-r}$ and $h^{2\xi+2t}$ are asymptotically proportional then the estimator $\sqrt{n}$-consistent if $\xi + t \geq r/2$. This result is similar to the condition in equation (4.12) for the differentiable case, but only requires a Holder condition rather than derivatives. This result shows that the bias complementarity available with the

twicing kernel is not solely an artifact of the higher order property of the twicing kernel.

We can illustrate the results we have derived so far by comparing different average density estimators. If we regard $\mu(F) = \int f(z)^2 dz$ as a nonlinear functional and derive a bias correction from the influence function estimate $\hat{\delta}(z) = 2\hat{f}(z)$, we obtain the estimator of equation (2.6), $\tilde{\mu}_1 = 2\sum_{i=1}^{n} \hat{f}(z_i)/n - \int \hat{f}(z)^2 dz$. Plugging in a bootstrap corrected density estimator $\tilde{f}(z)$ gives

$$\tilde{\mu}_2 = \int \tilde{f}(z)^2 dz,$$

that will generally be different from $\tilde{\mu}_1$. Regarding $\int f_0(z)^2 dz$ as the expectation of the density, where $\mu_0 = E[g(z,F_0)]$ for $g(z,F) = f(z)$, and noting that $E[g(z,F)]$ has influence function $f_0(z)$ that can be estimated by $\hat{f}(z)$, the bias corrected estimator of equation (4.3) becomes

$$\tilde{\mu}_3 = \sum_{i=1}^{n} \hat{f}(z_i)/n + \int \hat{f}(z)(\hat{P}-\hat{F})(dz) = 2\int \hat{f}(z)\hat{P}(dz) - \int \hat{f}(z)^2 dz = \tilde{\mu}_1.$$

Also, the bias corrected estimator of equation (4.4) will be

$$\tilde{\mu}_4 = \sum_{i=1}^{n} \tilde{f}(z_i)/n.$$

It follows from equation (4.8) that $\tilde{\mu}_4 = \tilde{\mu}_1$ if $\hat{f}$ and $\tilde{f}$ are kernel estimators where $\tilde{f}$ is based on the twicing kernel corresponding to $\hat{f}$. Here we see that three of the bias corrected estimators are equal, with the different one being the nonlinear integral $\int \tilde{f}(z)^2 dz$ of a bias corrected nonparametric estimator. The conclusion of Theorem 2.2 gives $\sqrt{n}$-consistency of $\tilde{\mu}_1$ if the original kernel has order at least s, $f_0(z)$ is s times differentiable, $s > r/2$, and $\sqrt{n}h^r \longrightarrow \infty$, $\sqrt{n}h^{2s} \longrightarrow 0$.

A further reduction in MSE of the bias corrected kernel estimator $\tilde{\mu}_1$ is possible by deleting the own observation to obtain

(4.16) $\tilde{\mu}_c = \sum_{i \neq j} \tilde{K}_h(z_i - z_j)/[n(n-1)]$.

Applying Theorem 4.1, this estimator will be $\sqrt{n}$-consistent if $s \geq r/4$, and the bandwidth is chosen as in equation (4.14 . Furthermore, by Theorem 4.2, if $a(x) = b(x) = f_0(z)$ satisfies Assumption 9 then $\tilde{\mu}_c$ will still be $\sqrt{n}$-consistent if $(\xi+t)/2 = \xi \geq r/4$ and the other conditions described above are satisfied. For example, if $r = 1$ then for the bandwidth is chosen so that $n^{-2}h^{-r}$ is proportional to $h^{4\xi}$, $\tilde{\mu}_c$ is $\sqrt{n}$-consistent if $\xi \geq 1/4$. Here $\xi = 1/4$ is a knife edge case where the variance remainder term will be the same size as the leading $Var(\psi(z))/n$ term. The condition $\xi \geq 1/4$ was shown by Bickel and Ritov (1988) to be necessary for existence of a $\sqrt{n}$-consistent (regular) estimator. Thus, the twicing kernel estimator $\tilde{\mu}_c$ attains $\sqrt{n}$-consistency under minimal conditions, like the more complicated sample splitting estimator of Bickel and Ritov (1988).

To illustrate the potential efficiency gains from using twicing kernels we have done some exact MSE comparisons for different estimators of the average density when the true density is a standard normal and a standard normal kernel is used to construct the estimator. Specifically, we have calculated exact MSE for the estimator in equation (4.16) when z_i are i.i.d. with $N(0,I)$ distribution and $\tilde{K}(u)$ is either the standard normal density or the twicing kernel based on the standard normal. The bandwidths have been chosen to minimize the asymptotic MSE for the respective estimators as in equation (4.14) for the twicing kernel and the analogous equation for the standard normal kernel (with corresponding larger bias term).

Table One gives the sample sizes at which the MSE for the twicing kernel estimator becomes smaller than the MSE of the normal kernel.

Table One: MSE Crossover

Dimension	Sample size
1	18
2	17
3	19
4	21
5	25
6	31
7	39
8	52

Figure One graphs the ratio of MSE as a function of sample size for dimension $r = 1$ up to dimension $r = 4$. For dimension $r > 4$ the standard normal kernel will not be $\sqrt{n}$-consistent but the twicing kernel estimator will, up to dimension $r = 8$. We restricted attention to the $r \leq 4$ case because the MSE ratio would asymptote to zero for higher dimensions. These graphs show persistent MSE gains.

Figure Two presents graphs of the MSE as a function of sample size for dimension $r = 1$ and $r = 3$ and for sample sizes 50, 100, and 200. These graphs allow us to compare the sensitivity of the MSE to the choice of bandwidth for the standard normal and twicing kernel. Although the MSE of the twicing kernel can turn sharply upward at low bandwidths, we find that it is flat, and close to its minimum, over a wide range of bandwidths. In this sense the MSE for the twicing kernel seems less sensitive to the choice of bandwidth than the original kernel.

5. Semiparametric M-estimation

Estimators that solve an estimating equation with a nonparametric component have many interesting applications and include as special cases all the ones we have considered so far. This class also includes profile likelihood estimators and many others, e.g. see Bickel et. al. (1990). We refer to this class of estimators as semiparametric m-estimators. In this Section we develop bias-corrected versions of these estimators.

To describe a semiparametric m-estimator, let β denote a q x 1 parameter vector, F and $\hat{F}$ be as previously discussed, and $m(z,\beta,F)$ a q x 1 vector of functions. Let $\hat{\beta}$ solve

$$(5.1) \qquad \sum_{i=1}^{n} m(z_i,\beta,\hat{F})/n = 0.$$

This includes as special cases $\mu(\hat{F})$, where $m(z,\beta,F) = \beta - \mu(F)$, and $\sum_{i=1}^{n} g(z_i,\hat{F})/n$, where $m(z,\beta,F) = \beta - g(z,F)$.

To obtain bias corrections it is useful to begin by expanding the moment equation around the true value of the parameter β_0. Suppose that $m(z,\beta,F)$ is differentiable in β. Then expanding and solving for $\hat{\beta}$ gives

$$(5.2) \qquad \sqrt{n}(\hat{\beta}-\beta_0) = -\overline{M}^{-1}\sum_{i=1}^{n} m(z_i,\hat{F})/\sqrt{n},\ \overline{M} = n^{-1}\sum_{i=1}^{n} \partial m(z_i,\bar{\beta},\hat{F})/\partial\beta,\ m(z_i,F) = m(z_i,\beta_0,F),$$

where $\bar{\beta}$ is a mean-value that lies on a line joining $\hat{\beta}$ and β_0 and actually differs from element to element of m. Asymptotic normality and $\sqrt{n}$-consistency of $\hat{\beta}$ will follow from consistency of $\overline{M}$ for $M = E[\partial m(z,\beta_0,F_0)/\partial\beta]$, nonsingularity of M, and asymptotic normality of $\sum_{i=1}^{n} m(z_i,\hat{F})/\sqrt{n}$. The consistency of $\overline{M}$ is a straightforward property that is generally an implication of consistency of $\hat{\beta}$ and $\hat{F}$ and a uniform law of large numbers. Asymptotic normality of $\sum_{i=1}^{n} m(z_i,\hat{F})/\sqrt{n}$ is where undersmoothing and bias corrections may be important.

Bias corrections for $\hat{\beta}$ can be obtained by applying the analysis of Section 4 to $g(z,F) = m(z,F)$ and accounting for the Jacobian term. The first approach is by an influence function correction like that of equation (4.3). Let $\delta(z)$ be the influence function for $\mu(F) = \int m(z,F)F_0(dz)$, and let $\hat{\delta}(z)$ be an estimator of this influence function. Also, let $\hat{M} = n^{-1}\sum_{i=1}^{n} \partial m(z_i,\hat{\beta},\hat{F})/\partial\beta$. Then a bias corrected estimator is given by

$$(5.3) \qquad \tilde{\beta} = \hat{\beta} - \hat{M}^{-1}[\sum_{i=1}^{n} \hat{\delta}(z_i)/n - \int \hat{\delta}(z)\hat{F}(dz)] = \hat{\beta} - \hat{M}^{-1}\int \hat{\delta}(z)(\hat{P}-\hat{F})(dz)$$

This estimator has a form like that in equation (4.3), except that the Jacobian estimator is included to account for the presence of $\overline{M}$ in equation (5.2).

To see how the correction affects the estimator, let $\psi(z) = m(z,F_0) + \delta(z) - E[\delta(z)]$ be the influence function of $\int m(z,F)F(dz)$. Then

$$(5.4) \qquad \sqrt{n}(\tilde{\beta} - \beta_0) = -\overline{M}^{-1}[\sum_{i=1}^{n}\psi(z_i)/\sqrt{n} + \sqrt{n}(R_{1n} + R_{2n} + R_{3n})] + R_{4n}$$

$$R_{1n} = \int[m(z,\hat{F}) - m(z,F_0)](\hat{P} - F_0)(dz), \quad R_{2n} = \int[\hat{\delta}(z) - \delta(z)](\hat{P} - \hat{F})(dz),$$

$$R_{3n} = \int[m(z,\hat{F}) - m(z,F_0)]F_0(dz) - \int\delta(z)(\hat{F} - F_0)(dz). \quad R_{4n} = \sqrt{n}(\overline{M}^{-1} - \hat{M}^{-1})\int\hat{\delta}(z)(\hat{P} - \hat{F})(dz).$$

The remainder terms in equation (5.4) are all second-order, so that $\sqrt{n}$-consistency of $\tilde{\beta}$ should not require undersmoothing, although precise regularity conditions are difficult to specify at the level of generality considered here.

One important property of semiparametric estimators follows immediately from equation (5.4). Suppose that $\delta(z) = 0$, meaning that estimation of F does not affect the asymptotic distribution of $\hat{\beta}$. Then $\hat{\delta}(z) = 0$ is consistent, and the bias corrected estimator would be the original estimator. Consequently, the original estimator should not need undersmoothing. Thus, we find that if the presence of $\hat{F}$ does not affect the large sample distribution of $\hat{\beta}$ no undersmoothing will be needed.

There are many interesting semiparametric estimators where estimation of F does not affect the asymptotic variance, and so undersmoothing may not be needed. As shown in Newey (1994), if $m(z,\beta,F)$ is the gradient of a function $q(z,\beta,F)$ and $\hat{F}$ is an estimator of the "profile" distribution that maximizes $E[q(z,\beta,F)]$ then estimation of F does not affect the asymptotic variance. Hence, undersmoothing may not be needed for any of these estimators. Many known semiparametric estimators are special cases of this result, including Robinson (1988), Chen and Shiau (1991), and Ichimura (1993). Also, if $m(z,\beta,F)$ is an efficient score for β in a semiparametric model and $\hat{F}$ is an estimator that imposes the restrictions implied by the model then the limiting distribution of $\hat{\beta}$

is not affected by estimation of F when the model is true, as has been demonstrated in many examples in the literature. Hence, undersmoothing may not be needed when $m(z,\beta,F)$ is an efficient score and the model is correct.

The second approach to bias correction is to use a bootstrap corrected estimator $\tilde{F}$ in the formation of the estimator of β, choosing $\tilde{\beta}$ as the solution to

$$(5.5) \qquad \sum_{i=1}^{n} m(z_i,\beta,\tilde{F})/n = 0.$$

This estimator should not need undersmoothing because in the expansion of equation (5.2) $\tilde{F}$ will replace $\hat{F}$ in the average $\sum_{i=1}^{n} m(z_i,\tilde{F})/n$. As discussed in Section 3, any estimator $\hat{F}$ that is an idempotent transformation of the empirical distribution, such as a sieve estimator or a series estimator of a conditional expectation, has this correction built in, so undersmoothing may not be required for $\sqrt{n}$-consistency of $\hat{\beta}$.

We could derive precise results on bias-corrected m-estimation for both the estimators of equation (5.2) and (5.5). However, because the ideas here are basically the same as in Sections 2 and 3, we choose to be brief, and consider only the estimator of equation (5.5), focusing on conditions for $\sqrt{n}$-consistency when $\tilde{F}$ is a twicing kernel estimator.

Newey and McFadden (1994) have already given general results on $\sqrt{n}$-consistency of semiparametric kernel estimators. We build on their results, deriving a corresponding result for twicing kernels, showing that undersmoothing is not needed. We adopt their specification for $m(z,\beta,F)$, where m depends on F only through $\gamma(x) = E_F[w|x]f(x)$, where w is a vector of random variables that are not elements of x. A twicing kernel estimator of the true function $\gamma_0(x) = E[w|x]f_0(x)$ would be

$$\tilde{\gamma}(x) = \sum_{i=1}^{n} \tilde{K}_h(x-x_i)w_i/n.$$

Letting $m(z,\beta,\gamma)$ be a $q \times 1$ vector of functions that depends on β and the function $\gamma(\cdot)$, a kernel semiparametric bootstrap corrected m-estimator would be $\tilde{\beta}$ solving

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

To specify regularity conditions we modify the norm $\|\cdot\|$ to apply to γ. Let $\mathcal{X}$ denote a compact set and

$$\|\gamma\| = \sup_{x\in\mathcal{X}} \sup_{|\lambda|\leq d} \|\partial^{\lambda}\gamma(x)\|.$$

This is a Sobolev norm like the one used for the kernel results of Section 2 and 3.

Assumption 9: $\gamma(x)$ is continuously differentiable to order s with bounded derivatives and for some $c > 0$ and all λ with $|\lambda| = s$, $\int \sup_{\|\Delta\|\leq c} |\partial^{\lambda}\gamma_0(x+\Delta)|dx < \infty$.

Let $m(z,\gamma) = m(z,\beta_0,\gamma)$.

Assumption 10: There are $b(z)$ and $D(z,\gamma)$ that are linear in γ such that for all γ with $\|\gamma-\gamma_0\|$ small enough, $\|m(z,\gamma)-m(z,\gamma_0)-D(z,\gamma-\gamma_0)\| \leq b(z)\|\gamma-\gamma_0\|^2$, $\|D(z,\gamma)\| \leq b(z)\|\gamma\|$, and $E[b(z)] < \infty$. Also, there is a matrix $\nu(x)$ with $E[D(z,\gamma-\gamma_0)] = \int \nu(x)(\gamma-\gamma_0)(dx)$, where $\nu(x)$ is continuously differentiable of order t with bounded derivatives. Also, for $p > 2$, $E[\|w\|^p] < \infty$ and $h = h(n)$ with $h \longrightarrow 0$ and $n^{1-(2/p)}h^r/\ln(n) \longrightarrow \infty$.

Let $\tilde{\nu}(x) = \int \tilde{K}(u)\nu(x+hu)du$, $\rho_n = \ln(n)^{1/2}/\sqrt{n}h^{r/2+d} + h^s$,

$$\tilde{B}_n = \sqrt{n}\int [\tilde{\nu}(x)-\nu(x)]w\hat{P}(dz), \quad R_n = \sqrt{n}\int [m(z,\tilde{F})-m(z,F_0)-D(z,\tilde{F}-F_0)]F_0(dz),$$

$$S_n = \sqrt{n}\int [m(z,\tilde{F})-m(z,F_0)](\hat{P}-F_0)(dz).$$

We first give a result that bounds the remainder terms in the expansion for the average of $m(z,\hat{\gamma})$.

Theorem 5.1: If Assumptions 2, 9, and 10 are satisfied then for $\psi(z) = m(z,\gamma_0) + v(x)w - E[v(x)w]$,

$$\sum_{i=1}^{n} m(z_i,\tilde{\gamma})/\sqrt{n} = \sum_{i=1}^{n} \psi(z_i)/\sqrt{n} + \tilde{B}_n + S_n + R_n,$$

$$\tilde{B}_n = O_p(\sqrt{n}h^{t+s} + h^t), \quad S_n = O_p(\sqrt{n}\rho_n^2 + h^s), \quad R_n = O_p(\sqrt{n}\rho_n^2).$$

Furthermore, if $v(x) = 0$ *then this result also holds with* $K(u)$ *replacing* $\tilde{K}(u)$.

To show a corresponding $\sqrt{n}$-consistency result for the semiparametric estimator we need to make assumptions that ensure convergence of the Jacobian term $\overline{M}$. The next condition suffices.

Assumption 11: $\hat{\beta} \xrightarrow{p} \beta_0$, for β in a neighborhood $\mathcal{B}$ of β_0 and all γ with $\|\gamma-\gamma_0\|$ small enough, $m(z,\beta,\gamma)$ is continuously differentiable and for some $c > 0$, $\|\partial m(z,\beta,\gamma)/\partial\beta - \partial m(z,\beta_0,\gamma_0)/\partial\beta\| \leq b(z)(\|\beta-\beta_0\|^c + \|\gamma-\gamma_0\|^c)$, $M = E[\partial m(z,\beta_0,\gamma_0)/\partial\beta]$ exists and is nonsingular, $m(z,\gamma_0)$ has mean zero and finite second moments.

Theorem 5.2: If Assumptions 2 and 9 - 11 are satisfied, $\sqrt{n}\rho_n^2 \longrightarrow 0$, *and* $\sqrt{n}h^{s+t} \longrightarrow 0$ *then* $\sqrt{n}(\hat{\beta}-\beta_0) \xrightarrow{d} N(0,M^{-1}Var(\psi(z))M^{-1\prime})$. *Furthermore, if* $v(x) = 0$ *then the same result holds with* $K(u)$ *replacing* $\tilde{K}(u)$.

When $v(x)$ is nonzero the conditions for $\sqrt{n}$-consistency of this estimator are exactly analogous to those discussed following Theorem 3.1. In particular, undersmoothing will not be needed if $t > r/2 + d$. The case where $v(x) = 0$ corresponds to estimation of γ having no effect on the asymptotic variance of $\tilde{\beta}$. As previously discussed, no bias correction is needed for this case: If $\sqrt{n}\rho_n^2 \longrightarrow 0$ and the kernel is the original (non-twicing one) then $\sqrt{n}$-consistency follows from Theorem 5.2.

6. Conclusion

In this paper we have shown that it is often possible to use a nonparametric estimator where the degree of smoothing is optimal (MSE minimizing) for estimation of the function to construct a $\sqrt{n}$-consistent estimator of a functional. One approach involves adding a bias correction to the estimator where the nonparametric estimate is plugged in. Another involves a smoothing correction to the nonparametric estimator. An important class of nonparametric estimators, that are idempotent transformations of the empirical distribution, have this smoothing correction built in, so that undersmoothing is not needed for $\sqrt{n}$-consistency, including orthogonal series density estimators, series estimators of conditional expectations, and sieve estimators.

The influence function plays a key role in achieving $\sqrt{n}$-consistency without undersmoothing. For the additive correction to a plug in estimator it must be possible to construct a nonparametric estimator of the influence function that converges sufficiently fast. For the smoothing adjustment to the nonparametric estimator the influence function must be smooth as a function of the data. These properties may or may not be intrinsic to the functional. A topic of future research would be to seek to identify the class of functionals for which $\sqrt{n}$-consistent estimation without undersmoothing is possible.

We have discussed bias correction and undersmoothing for an increasingly general class of estimators, beginning with functionals of a density and ending with semiparametric m-estimators. We found that for semiparametric m-estimators undersmoothing is not needed when nonparametric estimation does not affect the asymptotic variance of the estimator. Regularity conditions were given for "twicing" kernel estimators. We showed that this class of kernel estimators has a special property, being the outcome of a smoothing correction applied to the original kernel, that removes the necessity of undersmoothing for these estimators. Our numerical results show MSE gains for the twicing kernel at small sample sizes and less sensitivity to bandwidth choice in estimation of the average density, indicating some potential for use in practice.

Appendix: Proofs of Theorems

Throughout the Appendix, c and C will represent a generic positive constants, that may be different in different uses.

Proof of Theorem 2.1: By Lemma 8.10 of Newey and McFadden (1994), it follows that $\|\hat{F}-F_0\|^2 = O_p(\ln(n)/nh^{r+2d} + h^{2s})$. Therefore, for R_n in equation (2.4), $R_n = O(\sqrt{n}\|\hat{F}-F_0\|^2) = O_p(\ln(n)/\sqrt{n}h^{r+2d} + \sqrt{n}h^{2s})$. Also, for $\hat{B}_n = \sum_{i=1}^n e(z_i,h)/\sqrt{n}$ for $e(z,h) = \int K(u)[\delta(z+hu)-\delta(z)]du$. By continuity of $\delta(z)$, $\delta(z+hu)-\delta(z) \longrightarrow 0$ for almost all u as $h \longrightarrow 0$. Also, for small enough h, by $K(u)$ having bounded support $\mathcal{U}$, $\sup_u |K(u)[\delta(z+hu)-\delta(z)]| \leq 1(u\in\mathcal{U})\sup_u |K(u)| 2\sup_{\|\Delta\|\leq c}|\delta(z+\Delta)| = b(z,u)$, that is finite with probability one by $\delta(z)$ bounded and has finite integral over u by $\mathcal{U}$ compact, so by the dominated convergence theorem $e(z,h) \longrightarrow 0$ as $h \longrightarrow 0$ with probability one. Also, $|e(z,h)|^2 \leq [\int b(z,u)du]^2 \leq C$, so by the dominated convergence theorem $Var(\hat{B}_n) = Var(e(z,h)) \leq E[e(z,h)^2] \longrightarrow 0$. Also, by the usual mean-value expansion,

$$
\begin{aligned}
(A.1)\qquad & \int K(u)[f_0(z-hu)-f_0(z)]du \\
&= \int K(u)\{\sum_{j=1}^{s-1}((-h)^j/j!)\sum_{|\lambda|=j}u^\lambda \partial^\lambda f_0(x) + ((-h)^s/s!)\sum_{|\lambda|=s}u^\lambda \partial^\lambda f_0(x-\bar{h}u)\}du \\
&= (-h)^s\sum_{|\lambda|=s}\int K(u)u^\lambda \partial^\lambda f_0(z-\bar{h}u)du/s!,
\end{aligned}
$$

where $|\bar{h}| \leq |h|$ and dependence of $\bar{h}$ on z and u is suppressed for notational convenience. Then for small enough h,

$$
\begin{aligned}
|E[e(z,h)]| &= |\int\int K(u)[\delta(z+hu)-\delta(z)]duf_0(z)dz| \leq \int|\delta(z)|\,|\int K(u)[f_0(z-hu)-f_0(z)]du|dz \\
&\leq Ch^s\sum_{|\lambda|=s}[\int|K(u)|u^\lambda du]\int\sup_{\|\Delta\|\leq c}|\partial^\lambda f_0(z+\Delta)|dz = O(h^s).
\end{aligned}
$$

Then by the Markov inequality, $\hat{B}_n = O_p(\sqrt{n}h^s + o(1))$, giving the first conclusion. Furthermore, under the stated bandwidth conditions both $\hat{B}_n \xrightarrow{p} 0$ and $R_n \xrightarrow{p} 0$, giving

the second conclusion. QED.

The following Lemma is useful in the proofs to follow. Let w denote a random variable, $g(z) = E[w|z]$, and $\gamma(z)$ denote a possible value of $g(z)f_0(z)$.

Lemma A.1: If Assumption 2 is satisfied, $g(z)$ and $f_0(z)$ are continuous, $\mu(\gamma)$ is linear, $|\mu(\gamma)| \leq C\|\gamma\|$, then for $\bar{\gamma}(\cdot) = E[wK_h(\cdot-z)]$, $E[m(wK_h(\cdot-z))] = m(\bar{\gamma})$.

Proof: By $K(u)$ having bounded support and $\mathcal{Z}$ compact there is a compact set $\mathcal{C}$ such that $K_h(z-u) = 0$ for all $z \in \mathcal{Z}$ and $u \notin \mathcal{C}$. Let $\bar{\gamma}_1(z) = \int_{\mathcal{C}}\gamma(u)K_h(z-u)du$ and $\bar{\gamma}_2(z) = \int_{\mathcal{C}^c}\gamma(u)K_h(z-u)du$. For $z \in \mathcal{Z}$, $\bar{\gamma}_2(z) = 0$, so that $\|\bar{\gamma}_2\| = 0$ and $|m(\bar{\gamma}_2)| \leq C\|\bar{\gamma}_2\| = 0$. Then by linearity of $m(\gamma)$, $m(\bar{\gamma}) = m(\bar{\gamma}_1+\bar{\gamma}_2) = m(\bar{\gamma}_1)$. Also, by $\|K_h(\cdot-z)\| = 0$ for all $z \notin \mathcal{C}$, $m(\gamma(z)K_h(\cdot-z)) = 0$, for all $z \notin \mathcal{C}$, so by linearity of $m(\gamma)$, $E[m(wK_h(\cdot-z))] = E[wm(K_h(\cdot-z))] = E[g(z)m(K_h(\cdot-z))] = E[m(g(z)K_h(\cdot-z))] = \int_{\mathcal{C}}m(\gamma(u)K_h(\cdot-u))du$. Let u_J be a sequence of measures with finite support, that converge in distribution to the uniform measure on $\mathcal{C}$. Then, by continuity of $\gamma(z)$ and continuous differentiability of $K(v)$ to order d and by $\mathcal{Z}$ compact, $\gamma(z)K_h(\cdot-z)$ is continuous in z in the semi-norm $\|\cdot\|$. Hence $m(\gamma(z)K_h(\cdot-z))$ is continuous and bounded in z on $\mathcal{C}$. It follows that $\int_{\mathcal{C}}m(\gamma(z)K_h(\cdot-z))du_J \longrightarrow \int_{\mathcal{C}}m(\gamma(z)K_h(\cdot-z))du$. Also, Also, since each derivative of $\gamma(u)K_h(z-u)$ with respect to z of up to order d is bounded and continuous on $\mathcal{C}$, it follows that $\|\int_{\mathcal{C}}\gamma(u)K_h(\cdot-u)d(u_J-u)\| \longrightarrow 0$, and hence $m(\int_{\mathcal{C}}\gamma(u)K_h(\cdot-u)du_J) \longrightarrow m(\int_{\mathcal{C}}\gamma(u)K_h(\cdot-u)du)$. Furthermore, by u_J having finite support and linearity of $m(\gamma)$, $m(\int_{\mathcal{C}}\gamma(u)K_h(\cdot-u)du_J) = \int m(\gamma(u)K_h(\cdot-u))du_J$. Then by the triangle inequality,

$$m(\bar{\gamma}) = m(\int_{\mathcal{C}}\gamma(u)K_h(\cdot-u)du) = \int_{\mathcal{C}}m(\gamma(u)K_h(\cdot-u))du = E[m(wK_h(\cdot-z))]. \quad \text{QED.}$$

Proof of Theorem 2.2: For $\tilde{T}_n$ from Section 2,

(A.2) $$|\tilde{T}_n| \leq \sqrt{n}\int|\hat{\delta}(z)-\delta(z)||\hat{f}(z)-f_0(z)|dz$$

$$\leq \sqrt{n}\int b(z)|\hat{f}(z)-f_0(z)|dz\|\hat{F}-F_0\| + \sqrt{n}\int|D(z,\hat{F}-F_0)||\hat{f}(z)-f_0(z)|dz$$

$$\leq 2\sqrt{n}\int b(z)dz\|\hat{F}-F_0\|^2 = O_p(\ln(n)/\sqrt{n}h^{r+2d} + \sqrt{n}h^{2s}).$$

Next, by $b(z)$ bounded, $E[b(z)] < \infty$. Also, by Assumptions 1 and 3, $|D(z,K_h(\cdot-\tilde{z}))| \leq b(z)\|K_h(\cdot-\tilde{z})\| \leq Cb(z)h^{-r-d}$. Let $\bar{f}(z) = E[\hat{f}(z)]$ with corresponding charge $\bar{F}$, and note that $D(z,\hat{F}-F_0) = D(z,\hat{F}-\bar{F}) + D(z,\bar{F}-F_0)$. Also, by Lemma 8.10 of Newey and McFadden (1994), $\|\bar{F}-F_0\| = O(h^s)$. Then by Chebyshev's inequality,

$$\sqrt{n}\int D(z,\bar{F}-F_0)(\hat{P}-F_0)(dz) = O_p(\{E[D(z,\bar{F}-F_0)^2]\}^{1/2}) = O_p(h^s).$$

Now let $k(z,u) = D(z,K_h(\cdot-u))$, $\bar{k}_1(z) = \int k(z,u)f_0(u)dz$, and $\bar{k}_2(z) = \int k(u,z)f_0(u)dz$. By Lemma A.1, $\bar{k}_1(z) = D(z_i,\bar{F})$. Also, $E[|k(z,z)|] \leq Ch^{-r-d}$ and $\{E[k(z,z)^2]\}^{1/2} \leq Ch^{-r-d}$, so by linearity of $D(z,F)$ and a V-statistic projection result like Lemma 8.4 of Newey and McFadden (1994),

(A.3) $$\sqrt{n}\int D(z,\hat{F}-\bar{F})(\hat{P}-F_0)(dz)$$

$$= \sqrt{n}\{\int k(z,u)(\hat{P}x\hat{P})(dz,du) - \int[\bar{k}_1(z)+\bar{k}_2(z)]\hat{P}(dz) + E[\bar{k}_1(z)]\} = O_p(n^{-1/2}h^{-r-d}).$$

The triangle inequality and linearity of $D(z,F)$ in F then give $\sqrt{n}\int D(z,\hat{F}-F_0)(\hat{P}-F_0)(dz) = O_p(n^{-1/2}h^{-r-d} + h^s)$. Then by Assumption 3, for $h \longrightarrow 0$,

(A.4) $$|\tilde{S}_n| \leq \sqrt{n}|\int D(z,\hat{F}-F_0)(\hat{P}-F_0)(dz)| + [\int b(z)(\hat{P}+F_0)(dz)]\sqrt{n}\|\hat{F}-F_0\|^2$$

$$= O_p(n^{-1/2}h^{-r-2d} + h^s + \sqrt{n}h^{2s}).$$

The conclusion then follows by equation (A.2) and the triangle inequality. QED.

Proof of Theorem 3.1: For c in the statement of the Theorem, let $d_f(z) =$

$\sup_{\|\Delta\|\leq c, |\lambda|=s} |\partial^{\lambda} f_0(z+\Delta)|$. By a mean-value expansion like that in the proof of Theorem 1 and boundedness of the t^{th} derivatives of $\delta(z)$ it follows that $|\int[\delta(z+hu)-\delta(z)]\tilde{K}(u)du| \leq Ch^t$ for small enough h. Also, applying the same argument with K replacing $\tilde{K}$, for $\bar{\delta}$ and $\bar{f}$ in equation (3.5), $|\bar{\delta}(z)-\delta(z)| \leq Ch^t$ and $|\bar{f}(z)-f_0(z)| \leq Ch^t d_f(z)$. Therefore, $Var(\tilde{B}_n) \leq E[|\int[\delta(z+hu)-\delta(z)]\tilde{K}(u)du|^2] = O(h^{2t})$ and, as in equation (3.5), and $|E[\tilde{B}_n]| \leq \sqrt{n}Ch^{t+s}\int d_f(z)dz = O(h^{t+s})$. The first conclusion then follows by the Markov inequality. Also, it is easy to check that $\tilde{K}(u)$ also satisfies Assumption 2, so that by Lemma 8.10 of Newey and McFadden (1994), $\|\tilde{F}-F_0\|^2 = O_p(\ln(n)/nh^{r+2d} + h^{2s})$. The second conclusion then follows from Assumption 1. QED.

Proof of Theorem 3.2: Equation (2.4) is satisfied for $\hat{F}$ having Radon-Nikodym derivative $\hat{f}(z) = p^J(z)'\hat{\alpha}$ and $\hat{B}_n = \sum_{i=1}^n[\bar{\delta}(z_i)-\delta(z_i)]/\sqrt{n}$. Then $Var(B_n) \leq E[\{\bar{\delta}(z_i)-\delta(z_i)\}^2] \leq C\int[\bar{\delta}(z)-\delta(z)]^2dz = O(J^{-2t/r})$ by the density bounded and the approximation order given in the statement of the Theorem. Also, by the Cauchy-Schwartz inequality, $|E[\hat{B}_n]| \leq \sqrt{n}[\int\{\bar{\delta}(z)-\delta(z)\}^2dz]^{1/2}[\int\{\bar{f}(z)-f_0(z)\}^2dz]^{1/2} = O(\sqrt{n}J^{-s/r-t/r}) \longrightarrow 0$. The first conclusion then follows by the Markov inequality. The second conclusion follows by Assumption 1 and the triangle inequality. QED.

Proof of Theorem 4.1: Note that

$$\text{(A.5)} \qquad \tilde{\mu}_c = \sum\sum_{i\neq j} k(z_i,z_j)/[n(n-1)], \quad k(z_i,z_j) = \partial^{\lambda}\tilde{K}_h(x_i-x_j)y_j v_i,$$

where we suppress dependence of k on h for notational convenience. Define $\bar{a}(x) = \int K(u)a(x+hu)du$ and $\bar{b}(x) = \int K(u)b(x+hu)du$. Taking expectations, and integrating by parts, it follows similarly to equation (3.5) that for $\mu_0 = \int a(x)b(x)dx$,

$$\text{(A.6)} \qquad E[\tilde{\mu}_c] - \mu_0 = E[k(z_1,z_2)] - \mu_0 = \int\{\int\partial^{\lambda}\tilde{K}_h(x_1-x_2)E[y|x_2]f_0(x_2)dx_2\}b(x_1)dx_1 - \mu_0$$

$$= \int\int\tilde{K}_h(x_1-x_2)a(x_2)b(x_1)dx_1dx_2 - \mu_0 = -\int[\bar{a}(x)-a(x)][\bar{b}(x)-b(x)]dx.$$

By a mean value expansion in h like equation (A.1),

$$\bar{a}(x)-a(x) = (h^s/s!)\sum_{|\lambda|=s}\int K(u)u^\lambda \partial^\lambda a(x+\bar{h}u)du, \quad |\bar{h}| \leq |h|.$$

For c in the statement of the theorem let $d_a(x) = \sum_{|\lambda|=s}\sup_{\|\Delta\|\leq c}|\partial^\lambda a(x+\Delta)|$. By $K(u)$ having bounded support $\mathcal{U}$, for small enough h, $|K(u)\sum_{|\lambda|=s}u^\lambda \partial^\lambda a(x+\bar{h}u)| \leq C1(u\in\mathcal{U})d_a(x) < \infty$. Then by the dominated convergence theorem, $\int K(u)u^\lambda \partial^\lambda a(x+\bar{h}u)du \longrightarrow \zeta_\lambda \partial^\lambda a(x)$, so that $h^{-s}[\bar{a}(x)-a(x)] \longrightarrow \sum_{|\lambda|=s}\zeta_\lambda \partial^\lambda a(x)/s!$. Similarly, $h^{-t}[\bar{b}(x)-b(x)] \longrightarrow \sum_{|\lambda|=t}\zeta_\lambda \partial^\lambda b(x)/t!$. Also, noting that $|h^{-s}[\bar{a}(x)-a(x)]| \leq Cd_a(x)$ and $|h^{-t}[\bar{b}(x)-b(x)]| \leq C$ for $\int d_a(x)dx < \infty$ by Assumption 6, the dominated convergence theorem gives $h^{-s+t}(E[\hat{\mu}] - \mu_0) = h^{-2s}\int[\bar{a}(x)-a(x)][\bar{b}(x)-b(x)]dx \longrightarrow B$, so that

$$(A.7) \qquad E[\tilde{\mu}_c] = \mu_0 + h^{2s}B + o(h^{2s}).$$

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

$$(A.8) \qquad Var(\tilde{\mu}_c) = [(n-2)/n(n-1)]Var(E[k(z_1,z_2)+k(z_2,z_1)|z_1])$$

$$+ [1/n(n-1)]\{Var(k(z_1,z_2)) + Cov(k(z_1,z_2),k(z_2,z_1))\}$$

By Assumptions 2 and 5 and another application of the dominated convergence theorem, $\int\int[\partial^\lambda \tilde{K}(u)]^2\mu_{yy}(x-hu)\mu_{ww}(x)dudx \longrightarrow Q_1 = \int[\partial^\lambda \tilde{K}(u)]^2du\int\mu_{yy}(x)\mu_{ww}(x)dx$. Then by a change of variables, $u = (x_1-x_2)/h$, $x = x_1$,

$$(A.9) \qquad E[k(z_1,z_2)^2] = \int\int[\partial^\lambda \tilde{K}_h(x_1-x_2)]^2E[y_2^2|x_2]E[v_1^2|x_1]f_0(x_2)f_0(x_1)dx_2dx_1$$

$$= h^{-r-2|\lambda|}\int\int[\partial^\lambda \tilde{K}(u)]^2\mu_{yy}(x-hu)\mu_{ww}(x)dudx = h^{-r-2|\lambda|}Q_1 + o(h^{-r-2|\lambda|}).$$

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

(A.10) $[1/n(n-1)]\mathrm{Var}(k(z_1,z_2)) = [1/n(n-1)]h^{-r-2|\lambda|}Q_1 + o([1/n(n-1)]h^{-r-2|\lambda|}).$

$= n^{-2}h^{-r-2|\lambda|}Q_1 + o(n^{-2}h^{-r-2|\lambda|})$

Also, by $\partial^\lambda \tilde{K}(-u) = (-1)^{|\lambda|}\partial^\lambda \tilde{K}(u)$, it follows similarly to eq. (A.9) that

$$E[k(z_1,z_2)k(z_2,z_1)] = E[\partial^\lambda \tilde{K}_h(x_1-x_2)y_2v_1\partial^\lambda \tilde{K}_h(x_2-x_1)y_1v_2]$$

$$= (-1)^{|\lambda|}E[\partial^\lambda \tilde{K}_h(x_1-x_2)^2E[y_2v_2|x_2]E[y_1v_1|x_1]] = h^{-r-2|\lambda|}(Q-Q_1) + o(h^{-r-2|\lambda|}).$$

Therefore, analogously to equation (A.10) it follows that

(A.11) $[1/n(n-1)]\mathrm{Cov}(k(z_1,z_2),k(z_2,z_1)) = n^{-2}h^{-r-2|\lambda|}(Q-Q_1) + o(n^{-2}h^{-r-2|\lambda|}).$

Next, let $\tilde{\ell}(x) = \int \tilde{K}(u)\ell(x+hu)du$ and $\tilde{a}(x) = \int \tilde{K}(u)a(x+hu)du$. By applying the dominated convergence theorem as we have done previously it follows from Assumption 7 that $\tilde{\ell}(x) \longrightarrow \ell(x)$ and $\tilde{a}(x) \longrightarrow a(x)$ as $h \longrightarrow 0$, so that $\tilde{a}(x)v + \tilde{\ell}(x)y \longrightarrow a(x)v + \ell(x)y$ as $h \longrightarrow 0$. Since $[\tilde{a}(x)v + \tilde{\ell}(x)y]^2 \leq C(v^2+y^2)$, Assumption 7 and the dominated convergence theorem imply that $\mathrm{Var}(\tilde{a}(x)v+\tilde{\ell}(x)y) \longrightarrow \mathrm{Var}(\psi(z))$. Furthermore, by integration by parts and interchanging the order of differentiation and integration,

$$E[k(z_1,z_2)+k(z_2,z_1)|z_1] = E[\partial^\lambda \tilde{K}_h(x_1-x_2)y_2v_1|z_1] + E[\partial^\lambda \tilde{K}_h(x_2-x_1)y_1v_2|z_1]$$

$$= E[\partial^\lambda \tilde{K}_h(x_1-x_2)E[y_2|x_2]|z_1]v_1 + E[\partial^\lambda \tilde{K}_h(x_2-x_1)E[v_2|x_2]|z_1]y_1$$

$$= [\int \partial^\lambda \tilde{K}_h(x_1-x)E[y|x]f_0(x)dx]v_1 + (-1)^{|\lambda|}\int \partial^\lambda \tilde{K}_h(x_1-x)b(x)dx]y_1 = \tilde{a}(x)v + \tilde{\ell}(x)y.$$

Therefore,

(A.12) $[(n-2)/n(n-1)]\mathrm{Var}(E[k(z_1,z_2)+k(z_2,z_1)|z_1]) = \mathrm{Var}(\psi(z))/n + o(n^{-1}).$

Plugging the results from (A.10)-(A.12) into (A.8) and using $\mathrm{MSE}(\tilde{\mu}_c) = \mathrm{Var}(\tilde{\mu}_c) +$

$(E[\tilde{\mu}_c]-\mu_0)^2$ to combine equations (A.7) and (A.8) and then gives the result. QED.

Proof of Theorem 4.2: By Assumption 8 and $K(u)$ having bounded support, hu can be made as small as desired uniformly over the support of u by choosing h small enough. Then for h small enough, $|\bar{a}(x)-a(x)| = |\int K(u)[a(x+hu)-a(x)]du| \leq \int|K(u)||a(x+hu)-a(x)|du \leq \int|K(u)|C_a(x)\|hu\|^{\xi}du = h^{\xi}C_a(x)\int|K(u)|\|u\|^{\xi}du = Ch^{\xi}C_a(x)$ and similarly, $|\bar{b}(x)-b(x)| \leq Ch^{t}C_a(x)$. It then follows by equation (A.6) that

$$|E[\tilde{\mu}_c]-\mu_0| \leq \int|\bar{a}(x)-a(x)||\bar{b}(x)-b(x)|dx \leq Ch^{\xi+t}\int C_a(x)C_b(x)dx.$$

The conclusion now follows from plugging (A.10)-(A.12) into equation (A.8), similarly to the proof of Theorem 4.1. QED.

Proof of Theorem 5.1: By a change of variables,

$$\int\nu(u)\tilde{\gamma}(u)du = \int\int\nu(u)w\tilde{K}_h(u-x)du\hat{P}(dz) = \int[\int\nu(x+hu)\tilde{K}(u)du]w\hat{P}(dz) = \int\tilde{\nu}(x)w\hat{P}(dz).$$

Then by Assumption 9,

$$\int D(z,\hat{\gamma}-\gamma_0)F_0(dz) = \int\nu(u)[\tilde{\gamma}(u)-\gamma_0(u)]du = \int\nu(u)\tilde{\gamma}(u)du - E[\nu(x)w]$$

$$\int\tilde{\nu}(x)w\hat{P}(dz) - E[\nu(x)w] = \tilde{B}_n + \int\nu(x)w(\hat{P}-F_0)(dz).$$

The decomposition in the statement of the Theorem then follows. Next, by Lemma 8.10 of Newey and McFadden (1994), $\|\hat{\gamma}-\gamma_0\| = O_p(\rho_n)$. Let $\bar{S}_n = \int D(z,\tilde{F}-F_0)(\hat{P}-F_0)(dz)$. By Markov's inequality, $\int b(z)(\hat{P}+F_0)(dz) = O_p(1)$, so by Assumption 9,

$$\text{(A.12)} \qquad \sqrt{n}\|S_n-\bar{S}_n\| \leq [\int b(z)(\hat{P}+F_0)(dz)]\sqrt{n}\|\hat{\gamma}-\gamma_0\|^2 = O_p(\sqrt{n}\rho_n^2).$$

Let $\bar{\gamma}(x) = E[\hat{\gamma}(x)] = \int\tilde{K}_h(x-u)\gamma_0(u)du = \int\tilde{K}(u)\gamma_0(x-hu)du$. Then by linearity of $D(z,\gamma)$ in γ, $D(z,\hat{\gamma}-\gamma_0) = D(z,\hat{\gamma}-\bar{\gamma}) + D(z,\bar{\gamma}-\gamma_0)$. Let $k(z_i,z_j) = D(z_i,w_j\tilde{K}_h(\cdot-x_j))$, $\bar{k}_2(z) = \int k(u,z)F_0(du) = \int D(u,w\tilde{K}_h(\cdot-x))F_0(du)$, and $\bar{k}_1(z) = \int k(z,u)dF_0(du)$. By Lemma A.1, $\bar{k}_1(z)$

$= D(z,\bar{\gamma})$. Then by linearity of $D(z,\gamma)$ in γ, it follows as in equation (A.3) that $\sqrt{n}\int D(z,\hat{\gamma}-\bar{\gamma})(\hat{P}-F_0)(dz) = O_p(h^{-r-d}/\sqrt{n})$. Also, $\sqrt{n}\int D(z,\bar{\gamma}-\gamma_0)(\hat{P}-F_0)(dz) = O_p(h^s)$, so by the triangle inequality, $\bar{S}_n = \int D(z,\tilde{F}-F_0)(\hat{P}-F_0)(dz) = O_p(h^{-r-d}/\sqrt{n} + h^s)$. Then by the triangle inequality and $h^{-r-d}/\sqrt{n} \leq h^{-r-2d}/\sqrt{n}$ for n large enough,

$$(A.14) \qquad S_n = O_p(\sqrt{n}\rho_n^2) + O_p(h^{-r-d}/\sqrt{n} + h^s) = O_p(\sqrt{n}\rho_n^2 + h^s).$$

Next,

$$(A.15) \qquad |R_n| \leq \sqrt{n}\int |m(z,\hat{\gamma})-m(z,\gamma_0)-D(z,\hat{\gamma}-\gamma_0)|F_0(dz) \leq E[b(z)]\sqrt{n}\|\hat{\gamma}-\gamma_0\|^2 = O_p(\sqrt{n}\rho_n^2).$$

Next, let $\bar{\nu}(x) = \int K(u)\nu(x+hu)du$ and $\bar{\gamma}(x) = \int K(u)\gamma(x+hu)du$. Then it follows as in equation (3.5) that

$$E[\tilde{B}_n] = \sqrt{n}E[\{\tilde{\nu}(x)-\nu(x)\}w] = \sqrt{n}\int[\tilde{\nu}(x)-\nu(x)]\gamma_0(x)dx = \sqrt{n}\int[\bar{\nu}(x)-\nu(x)][\bar{\gamma}(x)-\gamma_0(x)]dx$$

Let $d_\nu(x) = \sup_{\|\Delta\|\leq c, |\lambda|=t}|\partial^\lambda \nu(x+\Delta)|$ and $d_\gamma(x) = \sup_{\|\Delta\|\leq c, |\lambda|=s}|\partial^\lambda \gamma_0(x+\Delta)|$. It follows as in the proof of Theorem 3.2 that $\|\tilde{\nu}(x)-\nu(x)\| \leq Cd_\nu(x)h^t$, $\|\bar{\nu}(x)-\nu(x)\| \leq Cd_\nu(x)h^t$, and $\|\bar{\gamma}(x)-\gamma_0(x)\| \leq Cd_\gamma(x)h^s$. Therefore, $|E[\tilde{B}_n]| = O(\sqrt{n}h^{t+s})$ and $Var(\tilde{B}_n) \leq E[\|\tilde{\nu}(x)-\nu(x)\|^2\|w\|^2] = O(h^{2t})$, so $\tilde{B}_n = O_p(\sqrt{n}h^{t+s} + h^t)$ holds by the Markov inequality, giving the first conclusion. The second conclusion follows by $\tilde{\nu}(x) = 0$, and hence $\tilde{B}_n = 0$, and the fact that the twicing kernel was only used to show $\tilde{B}_n = O_p(\sqrt{n}h^{t+s} + h^t)$. QED.

Proof of Theorem 5.2: It follows by Lemma 6 that $\sum_{i=1}^n m(z_i,\hat{\gamma})/\sqrt{n} \xrightarrow{p} N(0,Var(\psi(z)))$. Also, for $\tilde{M} = \sum_{i=1}^n \partial m(z_i,\beta_0,\gamma_0)/\partial\beta/n$, $\|\bar{M}-\tilde{M}\| \leq [\sum_{i=1}^n b(z_i)/n](\|\hat{\beta}-\beta_0\|^c + \|\hat{\gamma}-\gamma_0\|^c) \xrightarrow{p} 0$, so $\hat{M} \xrightarrow{p} M$ follows by the triangle inequality and Khintchine's law of large numbers. Then by the continuous mapping theorem, $\bar{M}^{-1} \xrightarrow{p} M^{-1}$, so the conclusion follows by Slutzky's theorem. QED.

Acknowledgement: Peter Bickel and Yaacov Ritov provided helpful comments, as did seminar participants at the Harvard/MIT econometrics workshop.

References

Andrews, D.W.K. (1994). Empirical process methods in econometrics. *Handbook of Econometrics, Vol 4,* Amsterdam: North-Holland.

Bickel, P.J. and Y. Ritov (1988). Estimating integrated squared density derivatives: Sharp best order of convergence results. *Sankhya* 50A 381-393.

Bickel P., C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1990). *Efficient and Adaptive Inference in Semiparametric Models*. Baltimore: John Hopkins University Press.

Chen, H. and J.J.H. Shiau (1991). A two-stage spline smoothing method for partially linear models. *J. of Stat. Planning and Inference* 27 187-201.

Hall, P. and J.S. Marron (1987). Estimation of integrated squared density derivatives. *Statistics and Probability Letters* 6 109-115.

Hardle, W., J. Hart, J.S. Marron and A.B. Tsybakov (1992). Bandwidth choice for average derivative estimation. *Journal of the American Statistical Association* 87 227-233.

Hardle, W. and T.M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives 84 986-995.

Hardle, W. and A.B. Tsybakov (1993). How sensitive are average derivatives. *Journal of Econometrics* 58 31-48.

Hausman, J.A. and W.K. Newey (1995). Nonparametric estimation of exact consumer surplus and deadweight loss. *Econometrica* 63, 1445-1476.

Ichimura, H. (1993). Estimation of single index models. *Journal of Econometrics* 58, 71-120.

Jones, M.C. and S.J. Sheather (1991). Using non-stochastic terms to advantage in kernel based estimation of integrated squared density derivatives. *Statistics and Probability Letters* 11 511-514.

Marron, J.S. and M.P. Wand (1992). Exact mean integrated squared error. *Annals of Statistics* 20, 712-736.

Newey, W.K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* 62 1349-1382.

Newey, W.K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory* 10 233-253.

Newey, W.K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics, Vol 4,* Amsterdam: North-Holland.

Y.I. and R.Z. Hasminskii (1989). Estimation of nonlinear functionals from the regression function with the possibility of the regressor's design. *Problems of Control and Information Theory.* 18 65-77.

Powell, J.L., J.H. Stock, and T.M. Stoker (1989). Semiparametric estimation of Index coefficients. *Econometrica.* 57 1403-1430.

Powell, J.L. and T.M. Stoker (1997). Optimal bandwidth choice for density weighted averages. *Journal of Econometrics*

Robinson, P.M. (1988). Root-N consistent semiparametric regression. *Econometrica* 56 931-954.

Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics,* New York: Wiley and Sons.

Shen, X. (1997). On methods of sieves and penalization. *Annals of Statistics.*

Van der Vaart, A. (1991). On differentiable functionals. *Annals of Statistics.* 19 178-204.

Van der Vaart, A. and J.A. Wellner (1996). *Weak Convergence and Empirical Processes.* Springer, New York.

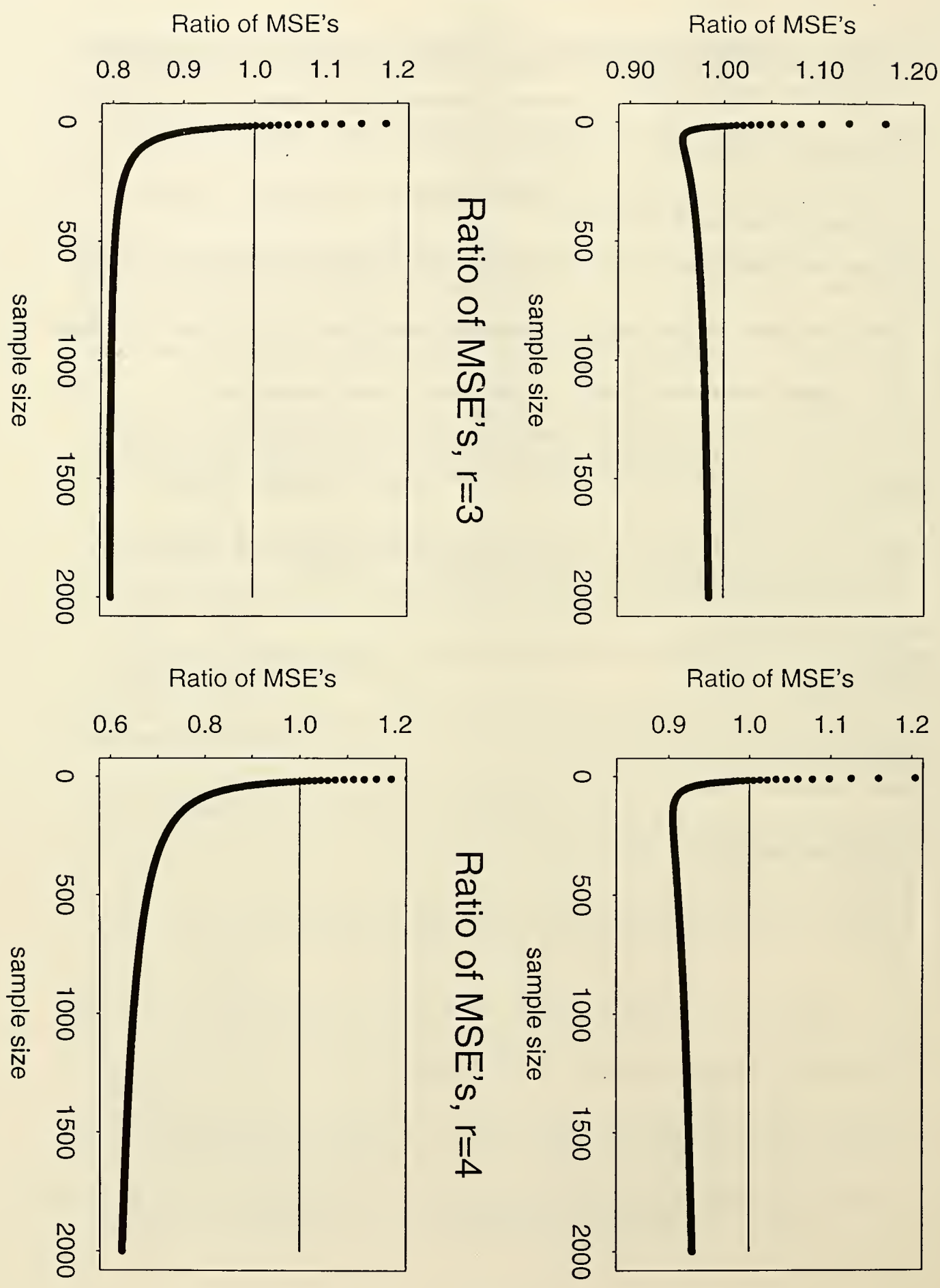
Wong, W.H. and T.A. Severini (1991). On maximum likelihood estimation in infinite dimensional parameters spaces. *Annals of Statistics.* 19 603-632.

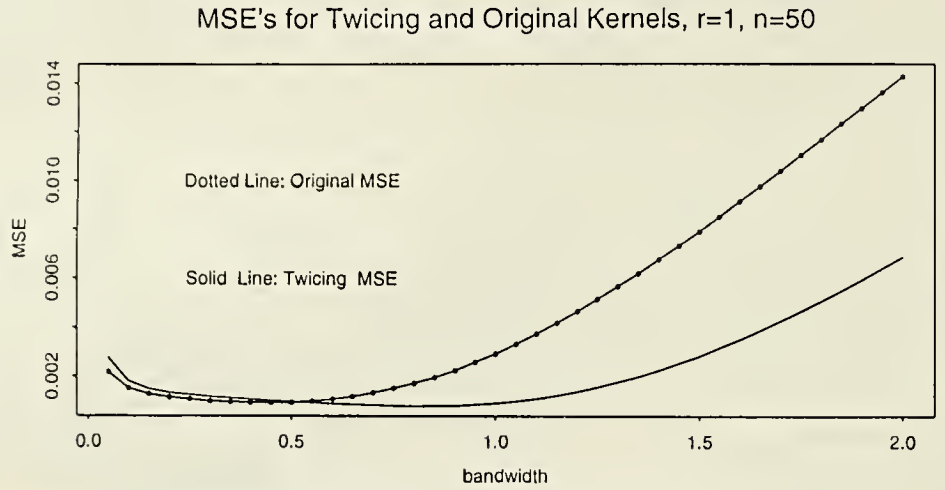
Department of Economics
MIT
Cambridge, MA 02139
wnewey@mit.edu

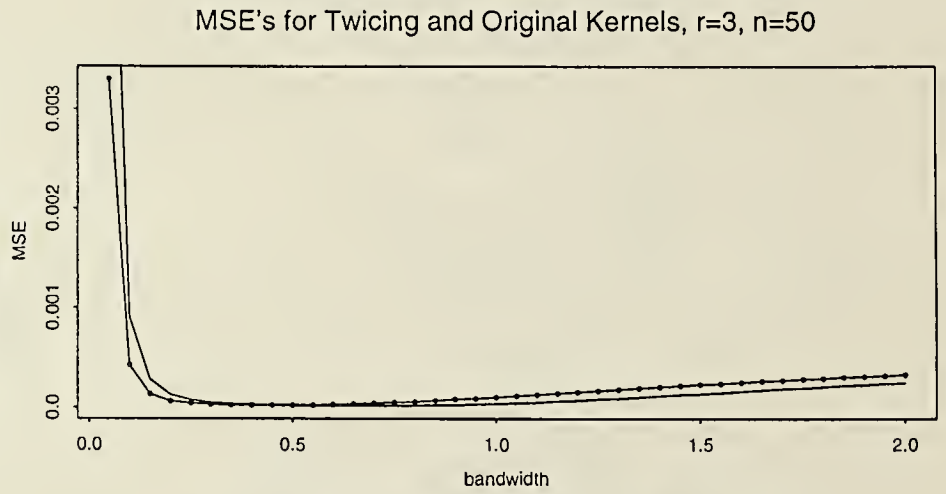
Institute of Statistical Science
Academia Sinica
Taipei, Taiwan

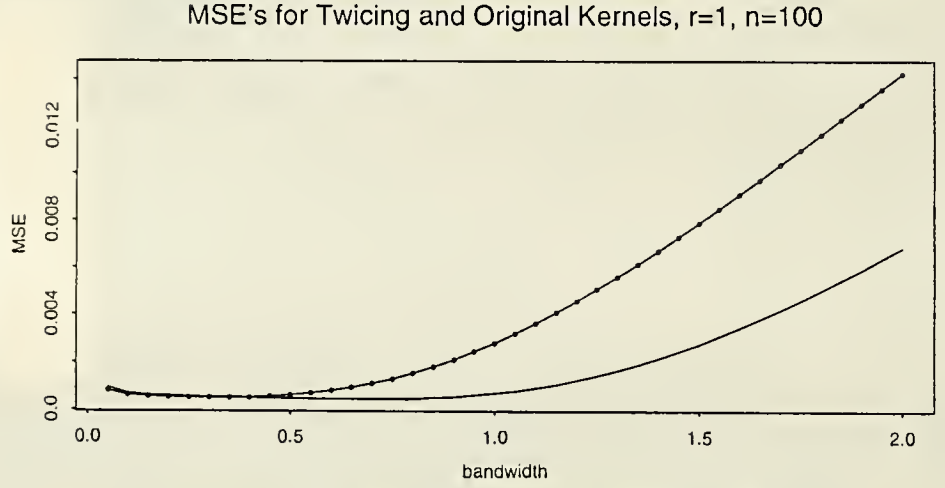
Harvard School of Public Health
Harvard University
Cambridge, MA 02138

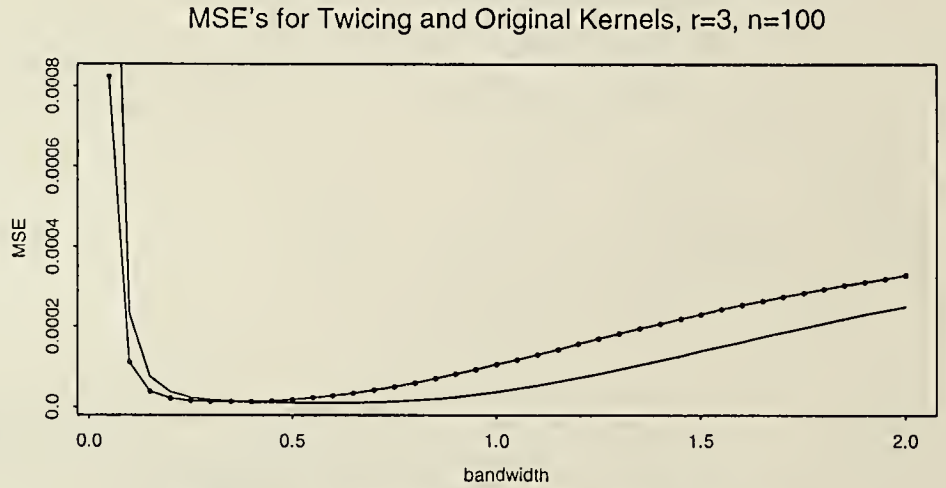
FIGURE ONE

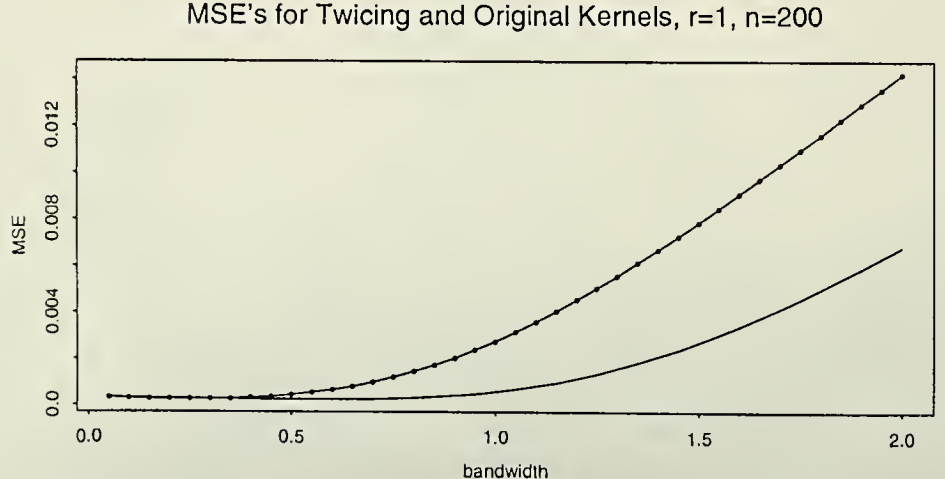

Ratio of MSE's, r=1
Ratio of MSE's
0.90
1.00
1.10
1.20
0
500
1000
1500
2000
sample size
Ratio of MSE's, r=2
Ratio of MSE's
0.9
1.0
1.1
1.2
0
500
1000
1500
2000
sample size
Ratio of MSE's, r=3
Ratio of MSE's
0.8
0.9
1.0
1.1
1.2
0
500
1000
1500
2000
sample size
Ratio of MSE's, r=4
Ratio of MSE's
0.6
0.8
1.0
1.2
0
500
1000
1500
2000
sample size


MSE's for Twicing and Original Kernels, r=1, n=50
Dotted Line: Original MSE
Solid Line: Twicing MSE
MSE
0.002
0.006
0.010
0.014
0.0
0.5
1.0
1.5
2.0
bandwidth


MSE's for Twicing and Original Kernels, r=3, n=50
MSE
0.0
0.001
0.002
0.003
0.0
0.5
1.0
1.5
2.0
bandwidth


MSE's for Twicing and Original Kernels, r=1, n=100
MSE
0.0
0.004
0.008
0.012
0.0
0.5
1.0
1.5
2.0
bandwidth


MSE's for Twicing and Original Kernels, r=3, n=100
MSE
0.0
0.0002
0.0004
0.0006
0.0008
0.0
0.5
1.0
1.5
2.0
bandwidth


MSE's for Twicing and Original Kernels, r=1, n=200
MSE
0.0
0.004
0.008
0.012
0.0
0.5
1.0
1.5
2.0
bandwidth

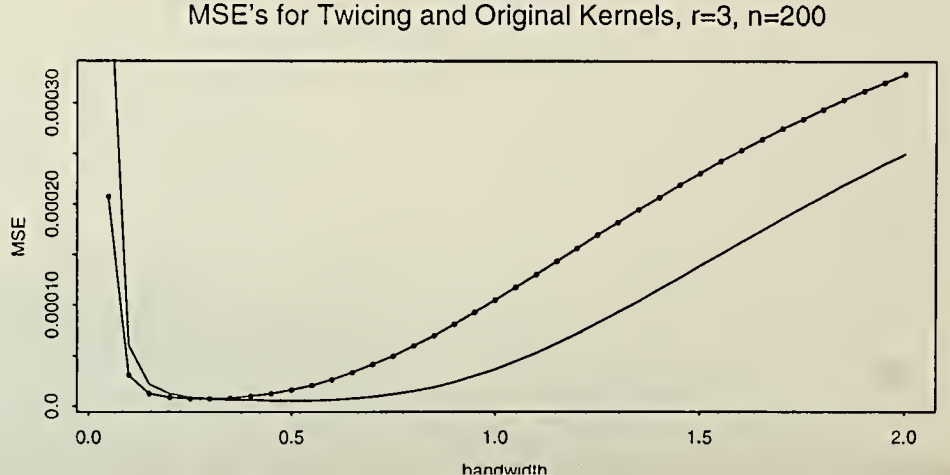

MSE's for Twicing and Original Kernels, r=3, n=200
MSE
0.0
0.00010
0.00020
0.00030
0.0
0.5
1.0
1.5
2.0
bandwidth

FIGURE TWO