OPTIMAL BANDWIDTH CHOICE FOR DENSITY-WEIGHTED AVERAGES

James L. Powell
Thomas M. Stoker

Massachusetts Institute of Technology

WP#3424 Revision - November 1994

Optimal Bandwidth Choice for Density-Weighted Averages

James L. Powell
Princeton University

Thomas M. Stoker*
Massachusetts Institute of Technology

May 1992, revised November 1994

Abstract

This paper characterizes the optimal bandwidth value for estimating density-weighted averages, statistics that arise in semiparametric estimation methods for index models and models of selected samples based on nonparametric kernel estimators. The optimal bandwidth is derived by minimizing the leading terms of mean squared error of the density weighted average. The optimal bandwidth formulation is developed by comparison to the optimal pointwise bandwidth of a naturally associated nonparametric estimation problem, highlighting the role of sample size and the structure of nonparametric estimation bias. The methods are illustrated by estimators of average density, density-weighted average derivatives and conditional covariances, and bandwidth values are calculated for normal designs. A simple "plug-in" estimator for the optimal bandwidth is proposed. Finally, the optimal bandwidth for estimating ratios of density-weighted averages is derived; showing that the earlier optimal formulae can be implemented directly using naturally defined "residual" values.

Keywords: smoothing; optimal bandwidth; semiparametric estimation; nonparametric estimation; density-weighting; "plug-in" estimator.

JEL classification codes: C13; C14 ; C25.

*This research was supported by National Science Foundation Grants SES 9010881 and SES 9122886. The authors wish to thank the reviewers for several valuable comments.

1. Introduction

Recent advances in the study of semiparametric methods in econometrics have yielded a number of new tools for studying empirical economic relationships. An important class of these methods involve "plug-in" estimators, where estimation of parameters of interest is facilitated by using nonparametric estimates of functions in place of the true, but unknown functions. Typical examples of such unknown functions include the density of disturbances in a model, or unknown features of the regression function of the response on the predictor variables. Examples of nonparametric estimators include kernel estimators and related local smoothing methods, or series estimators such as truncated polynomials or spline methods.

The issues of precision of nonparametric estimators are well known. In particular, suppose a function is estimated by averaging over a window of nearby data values, and consider the difference between setting a large or small window size. A large window includes more observations, thereby reducing variance, but masks subtle nonlinearity, or increases bias. Alternatively, a small window better facilitates detecting nonlinearity, or reduces bias, but involves less observations, thereby increasing variance. For estimating the function at a point, the optimal window size, or bandwidth value, is given by balancing variance with squared bias, thereby assuring the smallest mean squared error. The tradeoff between bias and variance will vary over different ranges of the function to be estimated, as well the optimal bandwidth or window size. A single, global choice of bandwidth can be based on minimizing average or integrated (pointwise) mean squared error values, or some other weighting of error across different ranges of the unknown function.[1] The literature on bandwidth choice in estimation of functions is quite extensive, and include several automatic (data-based) methods for choosing bandwidths in applications.[2]

When nonparametric estimators are used as ingredients in semiparametric estimation, the concerns regarding their precision are different. Since the parameters to be estimated are a primary focus, the relative importance of (pointwise) bias and variance of the nonparametric estimators is different than in the purely non-

[1] The same issues apply for any nonparametric method, such as choosing the degree of a polynomial expansion, or the degree of spline functions used for approximation.

[2] Textbook treatments of bandwidth choice in nonparametric estimation are given in Silverman (1986), Härdle (1991) and Hastie and Tibshirani (1990). References to more recent literature are given in Härdle, Hall and Marron (1988), Gasser, Kneip and Kohler (1991) and Nychka (1991), among others.

parametric case. For instance, if the parameter estimates are adversely affected by bias in the nonparametric estimators, then it may be sensible to lower the bandwidth size, reducing pointwise bias relative to variance. As such, semiparametric use of nonparametric estimators involves different criteria for nonparametric approximation, than optimal estimation of the unknown functions.

This feature is evident from the now standard results of asymptotic theory for semiparametric estimators. For example, procedures that employ kernel estimators typically involve "asymptotic undersmoothing"— if N denotes sample size, "asymptotic undersmoothing" refers to the notion that for parameter estimates to be $\sqrt{N}$ consistent, the bandwidth for kernel estimation must be shrunk more rapidly to zero than it would be for optimal pointwise estimation. This feature was noted for the estimators studied in Robinson (1988), Powell, Stock and Stoker (1989) and Härdle and Stoker (1989), among others, and more recently is highlighted in the unifying theory of Goldstein and Messer (1990).[3] This work does not address the issues of choosing bandwidths for particular applications, but rather just indicates how the conditions of limiting theory differ between nonparametric and semiparametric estimation.

In this paper we characterize bandwidth choice in perhaps the simplest substantive semiparametric estimation problem, namely, the estimation of density-weighted averages. This problem is interesting because it covers procedures for a wide range of semiparametric models, including situations where the precision of the nonparametric kernel estimators is a central focus.[4] We derive the optimal bandwidth by minimizing mean squared error of the estimator, and the nature of the solution is simple because the technical details of the analysis are kept to a minimum. This simplicity has the added bonus of permitting a straightforward comparison between optimal bandwidth values for pointwise estimation and for semiparametric estimation. Practical methods for bandwidth choice follow naturally from the development.

Related to our derivation is work on estimation of integrated squared density derivatives. In particular, Hall and Marron (1987) study that problem using kernel estimators,[5] and derive an optimal bandwidth formula. Our development can be

[3] Newey (1991) notes some differences between pointwise function estimation and semiparametric estimation when truncated polynomials are used.

[4] Andrews (1989) discusses situations where the precision of nonparametric estimators does not affect the asymptotic theory for "plug-in" semiparametric methods.

[5] Work on other aspects of estimating integrated squared density derivatives includes Bickel

viewed as a generalization of their results to more general estimation problems. Also related is a recent paper by Härdle, Hart, Marron and Tsybakov (1992), which studies the problem of bandwidth choice for estimation of one-dimensional unweighted average derivatives. Their results are specific to this case, and require strong conditions on the distribution of the covariates which are not imposed here.[6] Härdle and Tsybakov (1993) independently derived a result on bandwidth choice for weighted average derivatives similar to that in Section 4; our results specialize to their formulae for the weighted average derivative case.

Section 2 presents the estimator, estimand and a series of examples for motivation. Section 3 presents our assumptions, in the context of a "pointwise" nonparametric estimator that is closely associated with the density weighted average, and reviews the optimal bandwidth formula for the pointwise estimator. Section 4 derives the optimal bandwidth for the density weighted average, and spells out how the optimal bandwidth differs from the pointwise bandwidth in terms of an adjustment for sample size and an adjustment for the structure of nonparametric bias. These features are illustrated by computed bandwidth values for designs based on normal random variables. Section 4 closes with a simple "plug-in" estimator of the optimal bandwidth. Section 5 then characterizes bandwidth choice for ratios of density weighted averages, as motivated in certain examples of Section 2. Section 6 gives some concluding remarks.

2. The Estimation Problem and Examples

We assume that the data represent an i.i.d. sample of observations $\{z_i = 1, ..., N\}$, where z is the vector of responses and predictor variables as outlined below. We study estimators of "density-weighted averages" of the following form:

$$\hat{\delta}(h) = \binom{N}{2}^{-1} \sum_{i<j} p(z_i, z_j, h) \tag{2.1}$$

where the function $p(\cdot)$ is symmetric in pairs of observation – that is, $p(z_i, z_j, h) = p(z_j, z_i, h)$. As such, $\hat{\delta}(h)$ is a second order U-statistic with kernel p. The band-

and Ritov (1988) and Jones and Sheather (1991), among others.

[6] Kernel estimators of unweighted average derivatives (Härdle and Stoker (1989)) are nonlinear combinations of kernel estimators computed with trimming of the data sample. These two features substantially complicate the analysis of bandwidth choice.

width is the parameter h, and the limiting theory for $\hat{\delta}(h)$ has h decreasing with sample size, or $h = h(N) \to 0$ as $N \to \infty$.

If the expectation of $\hat{\delta}(h)$ is denoted

$$\delta(h) \equiv E\left[\hat{\delta}(h)\right] = E\left[p(z_i, z_j, h)\right], \tag{2.2}$$

then the object of estimation is

$$\delta_0 \equiv \lim_{h \to 0} \delta(h), \tag{2.3}$$

The object of the paper is to characterize the optimal bandwidth h^+ for computing $\hat{\delta}(h)$. We also characterize the optimal bandwidth for ratios of density-weighted averages, or ratios of estimators in the form (2.1).

We refer to the U-statistic in (2.1) as a "density-weighted average" because this form often arises when kernel methods are used to estimate density-weighted expectations, as in each of the following examples. Example 2.1 is a useful pedagogical device for illustrating our results, and Examples 2.2 and 2.3 arise from standard semiparametric problems in econometrics.

Example 2.1. *(Average Density): Here $z_i \equiv x_i \in \mathbf{R}^k$ is a continuous random vector, $x_i \sim f(x)dx$, and the object of estimation is the average density value:*

$$\delta_0 = \int f(x_i)^2 dx_i = E\left[f(x_i)\right] \tag{2.4}$$

The estimator $\hat{\delta}(h)$ of δ_0 is constructed by computing density estimates for each data point as

$$\hat{f}(x_i, h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{1}{h^k} \mathcal{K}\left(\frac{x_i - x_j}{h}\right) \tag{2.5}$$

and taking their average $\hat{\delta}(h) = N^{-1} \sum_{i=1}^{N} \hat{f}_i(x_i, h)$, which gives (2.1) with

$$p(z_i, z_j, h) = \frac{1}{h^k} \mathcal{K}\left(\frac{x_i - x_j}{h}\right) \tag{2.6}$$

Here $\mathcal{K} : \mathbf{R}^k \to \mathbf{R}$ is a (kernel) function with $\mathcal{K}(u) = \mathcal{K}(-u)$ and $\int \mathcal{K}(u)\, du = 1$. Here and elsewhere, we utilize "leave-out" kernel estimators such as (2.5); we could incorporate the "$i = j$" terms without changing any results, but unnecessarily *complicate the notation to account for the terms.*[7] *For parts of our bias discussion below, we will refer to the* order *of the kernel $\mathcal{K}$; this is defined as an integer $P(\geq 2)$ such that $\int u_1^{\ell_1}...u_k^{\ell_k}\ \mathcal{K}(u)\,du = 0$ if $\{\ell_j\}$ are nonnegative integers with $\ell_1 + ... + \ell_k < P$ and $\int u_1^{\ell_1}...u_k^{\ell_k}\ \mathcal{K}(u)\,du \neq 0$ for some $\ell_1 + ... + \ell_k = P$. (In the typical case where $\mathcal{K}$ is a positive symmetric density function, $P = 2$).*

Example 2.2. *(Density-Weighted Average Derivative): Here $z_i \equiv (y_i, x_i^{'})^{'}$, where $y_i \in \mathbf{R}^1$denotes a response or dependent variable, $x_i \in \mathbf{R}^k$ denotes a continuous predictor variable, $x_i \sim f(x)dx$, and the regression of y on x is denoted $E[y_i|x_i] = g(x_i)$. The object of estimation is the density-weighted average derivative*

$$\delta_0 = E\left[f(x_i)\frac{\partial g(x_i)}{\partial x_i}\right] = -2E\left[\frac{\partial f(x_i)}{\partial x_i}y_i\right] \tag{2.7}$$

assuming $f(x)g(x) \rightarrow 0$ as $|x| \rightarrow \infty$, and all derivatives and moments exist (Powell, Stock, and Stoker 1989). The density-weighted average derivative gives an estimator of index model coefficients up to scale – that is, when $g(x_i) = G(x_i^{'}\beta_0)$, then δ_0 is proportional to the coefficients β_0. The estimator $\hat{\delta}(h)$ is the sample analogue of the second equality, or the average of $-2y_i\partial\hat{f}_i(x_i,h)/\partial x$, where $\hat{f}_i(x_i,h)$ is the kernel estimator (2.5). This gives $\hat{\delta}(h)$ in the (vector) form (2.1), with

$$p(z_i, z_j, h) = -\frac{1}{h^{k+1}} \cdot \mathcal{V}\left(\frac{x_i - x_j}{h}\right) \cdot (y_i - y_j) \tag{2.8}$$

where $\mathcal{V}(u) = -\partial\mathcal{K}(u)/\partial u$, for $\mathcal{K} : \mathbf{R}^k \rightarrow \mathbf{R}$ assumed to have the same properties as discussed for Example 2.1. A related estimator $\hat{d}(h)$ uses $\partial\hat{f}_i(x_i,h)/\partial x$ as instruments in a linear regression of y_i on x_i; namely the sample analog of

$$d_0 = E\left[\frac{\partial f(x_i)}{\partial x_i}x_i^{'}\right]^{-1} E\left[\frac{\partial f(x_i)}{\partial x_i}y_i\right]. \tag{2.9}$$

With an index model, d_0 is likewise proportional to the coefficients β_0, and the estimator $\hat{d}(h)$ is a ratio of density-weighted averages of the form (2.1) (see Powell,

[7] The average density is covered as the integrated squared (zero order) density derivative in Hall and Marron (1987), who also discuss the omission of the "$i = j$" terms. While these terms can produce inferior asymptotic behavior of the statistic, Jones and Sheather (1991) discuss using the "$i = j$" terms to reduce finite sample mean squared error. Estimation of the average density also arises in the measurement of precision of various rank estimators (Jaeckel 1972, Jurečková 1971).

Stock and Stoker (1989) for details). We study bandwidth choice for such ratios in Section 5.

Example 2.3. *(Density-Weighted Conditional Covariances): Here* $z_i \equiv (y_i, x_i^{'}, w_i^{'})^{'}$, *where* $y_i \in \mathbf{R}^1$ *is a response,* $x_i \in \mathbf{R}^{k^{'}}$ *is a set of predictor variables (discrete or continuous), and* $w_i \in \mathbf{R}^k$ *is a set of continuous predictor variables,* $w_i \sim f(w)dw$. *The objects of estimation are density-weighted conditional covariances, given as*

$$\delta_0^y = E\left[f(w_i)\cdot(x_i - E\left[x_i|w_i\right])\cdot(y_i - E\left[y_i|w_i\right])\right], \tag{2.10}$$

and

$$\delta_0^x = E\left[f(w_i)\cdot(x_i - E\left[x_i|w_i\right])\cdot(x_i - E\left[x_i|w_i\right])\right]. \tag{2.11}$$

Density-weighted conditional covariances are relevant for estimation of the partially linear model

$$y_i = x_i^{'}\beta_0 + \theta(w_i) + u_i, \quad E\left[u_i|x_i, w_i\right] = 0 \tag{2.12}$$

where β_0 *can be written as*

$$\beta_0 = \left[\delta_0^x\right]^{-1}\delta_0^y, \tag{2.13}$$

assuming δ_0^x *is nonsingular and* $\theta(\cdot)$ *is sufficiently smooth (c.f. Powell (1987), among others). An estimator* $\hat{\delta}^y(h)$ *of* δ_0^y *is given by (2.1) with*

$$p^y(z_i, z_j, h) = \frac{1}{2h^k}\mathcal{K}\left(\frac{w_i - w_j}{h}\right)\cdot(x_i - x_j)\cdot(y_i - y_j), \tag{2.14}$$

where again $\mathcal{K}(u)$ *is a kernel function satisfying the properties discussed in Example 2.1. Analogously,* δ_0^x *is estimated by* $\hat{\delta}^x(h)$ *of (2.1), where*

$$p^x(z_i, z_j, h) = \frac{1}{2h^k}\mathcal{K}\left(\frac{w_i - w_j}{h}\right)\cdot(x_i - x_j)\cdot(x_i - x_j), \tag{2.15}$$

Again, our main focus is on bandwidth choice for $\hat{\delta}^y(h)$ *and* $\hat{\delta}^x(h)$ *in Section 4, with bandwidth choice for the ratio* $\hat{\beta}(h) = \left[\hat{\delta}^x(h)\right]^{-1}\hat{\delta}^y(h)$ *discussed in Section 5.*

3. The "Pointwise" Structure of Density Weighted Averages

At this point, we could derive the optimal bandwidth for $\hat{\delta}(h)$ directly, as we do in Section 4. However, we first develop the structure of $\hat{\delta}(h)$ by analyzing a nonparametric estimation problem that is closely associated with it. This is useful for a couple of reasons. First, the issues involved with bandwidth choice for estimating a function are well known, and provide a reasonable backdrop for discussing bandwidth choice for $\hat{\delta}(h)$. Second, one of our main aims is to spell out the differences between bandwidth choice for estimating functions and bandwidth choice in semiparametric procedures, such as $\hat{\delta}(h)$. The following development provides the relevant grounds for comparison — $\hat{\delta}(h)$ is just the average of the associated nonparametric estimator evaluated over the data sample, and our comparative analysis is based on the difference between pointwise fitting criterion and the averaged criterion appropriate for the performance of $\hat{\delta}(h)$.

Define the functions $r(z_i, h)$ and $r_0(z_i)$ as the conditional expectations[8]

$$r(z_i, h) \equiv E\left[p(z_i, z_j, h)|z_i\right], \tag{3.1}$$

$$r_0(z_i) \equiv \lim_{h\to 0} r(z_i, h) = r(z_i, 0). \tag{3.2}$$

These functions are related to $\delta(h)$ and δ_0 through

$$\delta(h) = E\left[r(z_i, h)\right], \tag{3.3}$$

and

$$\delta_0 = E\left[r_0(z_i)\right]. \tag{3.4}$$

The natural nonparametric estimator of $r_0(z_i)$ is obtained by averaging $p(z_i, z_j, h)$ over j, or

$$\hat{r}(z_i, h) \equiv \frac{1}{N-1} \sum_{\substack{j=1 \\ j\neq i}}^{N} p(z_i, z_j, h) \tag{3.5}$$

with the density-weighted average $\hat{\delta}(h)$ just the sample average of this nonparametric estimator:

$$\hat{\delta}(h) = \frac{1}{N} \sum_{i=1}^{N} \hat{r}(z_i, h). \tag{3.6}$$

[8] These functions arise from the projection of the U-statistic (2.1); c.f. Hoeffding (1948).

The "pointwise" bandwidth of interest is the optimal bandwidth choice for the estimator $\hat{r}(z_i, h)$ of $r_0(z_i)$.[9]

Further intuition can be gained from noting the role that the function $r_0(z_i)$ plays in the asymptotic theory for $\hat{\delta}(h)$. As outlined in Section 4, when $\hat{\delta}(h)$ is a $\sqrt{N}$ asymptotically normal estimator, then $2[r_0(z_i) - \delta_0]$ is the "influence" of the i^{th} observation in the asymptotic variation of $\hat{\delta}(h)$[10]. For Example 2.1, it is easy to verify that $r_0(z)$ equals $f(x)$; for Example 2.2, $r_0(z)$ equals $f(x) \cdot \partial g(x)/\partial x - [y - g(x)] \cdot \partial f(x)/\partial x$; and for of Example 2.3, $r_0(z)$ equals $(1/2) \cdot f(w) \cdot (x - E[x|w]) \cdot (y - E[y|w])$. Moreover, the sample variance of $2\hat{r}(z_i, h)$ gives a natural estimator of the asymptotic variance of $\hat{\delta}(h)$, as proposed by Powell, Stock and Stoker (1989).[11]

We state our assumptions in terms of properties of $\hat{r}(z_i, h)$ as an estimator of $r_0(z_i)$, which are easily derived for each of our examples (under standard primitive conditions). First, we assume that the bias of $\hat{r}(z_i, h)$ is of polynomial order in h:

Assumption 1. *(Rate of Convergence of Pointwise Bias of $\hat{r}(z_i, h)$): The function $r(z_i, h)$ satisfies*

$$r(z_i, h) - r_0(z_i) = s(z_i) \cdot h^{\alpha} + s^*(z_i, h) \tag{3.7}$$

for some $\alpha > 0$, where $E[s(z_i)] \neq 0$ and the remainder term $s^*(\cdot)$ satisfies

$$E \left\| s^*(z_i, h) \right\|^2 = o(h^{2\alpha}). \tag{3.8}$$

In each of our examples, the power α is the order of the kernel $\mathcal{K}(u)$ — namely $\alpha = P$ - but generally, α depends on the structure of the kernel $p(\cdot)$ of the U-statistic. Assumption 1 clearly implies a polynomial order for the bias of $\hat{\delta}(h)$ for δ_0— (3.3,3.4) and (3.7) imply that

$$\delta(h) - \delta_0 = E\left[s(z_i)\right] \cdot h^{\alpha} + o(h^{\alpha}). \tag{3.9}$$

We next structure the variance of $\hat{r}(z_i, h)$ by assuming

[9] For Example 2.1, the estimator $\hat{r}(z_i, h)$ of (3.5) is the kernel density estimator $\hat{f}(z_i, h)$ of (2.5). In this case, comparison of the optimal bandwidths for estimation of $\hat{r}(z_i, h)$ and for $\hat{\delta}(h)$ will indicate how bandwidth choice for density estimation differs from that for estimation of average density.

[10] While $\hat{\delta}(h)$ is the average of $\hat{r}(z_i, h)$ of (3.6), the fact that its asymptotic variance is given by $2\left[r_i(z_i) - \delta_0\right]$ is due to the many common components (overlaps) in $\hat{r}(z_i, h)$ for different observations $i = 1, ..., N$. Basic discussion of this overlap structure is given in Stoker (1992).

[11] This procedure was subsequently proposed in "linearized" form in the analysis of unweighted average derivative estimators by Härdle and Stoker (1989), among others.

Assumption 2. *(Series Expansion for Second Moment): The function $p(z_i, z_j, h)$ satisfies*

$$E\left[\|p(z_i, z_j, h)\|^2 |z_i\right] = q(z_i) \cdot h^{-\gamma} + q^*(z_i, h) \tag{3.10}$$

for some $\gamma > 0$, where the remainder term q^* satisfies

$$E\|q^*(z_i, h)\|^2 = o(h^{-\gamma}). \tag{3.11}$$

In Examples 2.1, 2.2 and 2.3, the coefficient γ of Assumption 2 takes the values k, $k+2$, or k, respectively. Assumption 2 clearly restricts the unconditional variance of $p(\cdot)$, as

$$E\left[\|p(z_i, z_j, h)\|^2\right] = E[q(z_i)] \cdot h^{-\gamma} + o(h^{-\gamma}). \tag{3.12}$$

All our optimal bandwidth values are derived by minimizing mean squared error, employing the leading terms of squared bias and variance. For these calculations, we take $p(\cdot)$ to be a scalar function for simplicity. For cases where $p(\cdot)$ is a vector function, our derivations apply immediately to a single component of $p(\cdot)$ and of $\hat{\delta}(h)$, and can immediately be extended to any particular linear combination $\lambda' p(.)$, by computing optimal bandwidths for $\lambda' \hat{r}$ and $\lambda' \hat{\delta}(h)$.[12]

These assumptions allow an immediate derivation of the optimal pointwise bandwidth for the estimation of the function $r_0(z)$. For a given argument value, the pointwise mean squared error of $\hat{r}(z_i, h)$ is

$$\begin{aligned} PMSE[\hat{r}(z_i, h)] &= E\left[\hat{r}(z, h) - r_0(z)|z\right]^2 \qquad (3.13) \\ &= (N-1)^{-1} Var\left[p(z, z_j, h)|z\right] + \left[r(z, h) - r_0(z)\right]^2 \\ &= N^{-1} q(z) \cdot h^{-\gamma} + s(z)^2 \cdot h^{2\alpha} + o\left(\tfrac{1}{Nh^{\gamma}}\right) + o(h^{2\alpha}), \end{aligned}$$

[12] Likewise, we could solve for the bandwidth that minimizes $E\left[\left(\hat{\delta}(h) - \delta_0\right)' W \left(\hat{\delta}(h) - \delta_0\right)\right]$ for any positive semi-definite matrix W, etc. By diagonalizing the weight matrix W, we can rewrite this problem as the minimization of the mean squared error of a linear combination $\sum_j \lambda_j' \hat{\delta}_j(h)$ of U-statistics, which is itself a scalar U-statistic. In the formulae below, we would replace the terms for the squared bias and variance with the corresponding quadratic form in bias and $E\left[\left(\hat{\delta}(h) - E\left(\hat{\delta}(h)\right)\right)' W \left(\hat{\delta}(h) - E\left(\hat{\delta}(h)\right)\right)\right]$, respectively.

and by minimizing the first two terms, we can derive the optimal bandwidth $h^*(z)$ for estimation of $r_0(z)$ at $z = z_i$ as

$$h^*(z) = \left[\frac{\gamma \cdot q(z)}{2\alpha \cdot s(z)^2}\right]^{\frac{1}{2\alpha+\gamma}} \cdot \left[\frac{1}{N}\right]^{\frac{1}{2\alpha+\gamma}} + o\left(\left[\frac{1}{N}\right]^{\frac{1}{2\alpha+\gamma}}\right), \tag{3.14}$$

where we have assumed that $s(z) \neq 0$.[13] The bandwidth $h^*(z)$ will vary with z, depending on the (local) sensitivity of bias and variance to bandwidth value, through $s(z)$ and $q(z)$. To choose a single bandwidth value for estimating $r_0(z)$ over its domain, we can minimize a global fitting criterion, such as the integrated mean squared error. For our problem, it is convenient consider the average of the pointwise mean squared error values, or

$$\begin{aligned} AMSE[\hat{r}(z_i,h)] &= E\left[\hat{r}(z,h) - r_0(z)\right]^2 \\ &= (N-1)^{-1}E\left(Var\left[p(z,z_j,h)\right]\right) + E\left[r(z,h) - r_0(z)\right]^2 \\ &= N^{-1}E\left[q(z)\right] \cdot h^{-\gamma} + E\left[s(z)^2\right] \cdot h^{2\alpha} + o\left(\tfrac{1}{Nh^\gamma}\right) + o(h^{2\alpha}). \end{aligned} \tag{3.15}$$

where we have assumed that $E[s(z_i)^2] \neq 0$. Likewise, minimizing (the leading terms of) $AMSE$ gives

$$h^* = \left[\frac{\gamma \cdot E[q(z)]}{2\alpha \cdot E[s(z)^2]}\right]^{\frac{1}{2\alpha+\gamma}} \cdot \left[\frac{1}{N}\right]^{\frac{1}{2\alpha+\gamma}} + o\left(\left[\frac{1}{N}\right]^{\frac{1}{2\alpha+\gamma}}\right) \tag{3.16}$$

We will use h^* as the "pointwise" optimal bandwidth in our comparisons later. It is easy to verify that h^* (and $h^*(z)$ for any z) displays standard rates for large sample nonparametric estimation — the orders of pointwise variance and squared bias are equated as $O(1/N(h^*)^\gamma) = O((h^*)^{2\alpha})$, with $h^* = O(N^{-1/(2\alpha+\gamma)})$. and the (best) average pointwise mean squared error rate is $AMSE[\hat{r}(z_i, h^*)] = O\left(N^{-2\alpha/(2\alpha+\gamma)}\right)$.[14]

The bandwidth h^* is approximated by its leading term, which we denote as h^{**}. For later reference, we summarize this discussion as

[13] See, for example, Silverman (1986) for a discussion of this calculation.

[14] The slow rate of convergence of the bandwidth h^* given here is well known; see, among others, Härdle, Hall and Marron (1988).

Proposition 3.1. *Given Assumptions 1 and 2, if h^* denotes the bandwidth value that minimizes $AMSE[\hat{r}(z_i, h)]$ and*

$$h^{**} = \left[\frac{\gamma \cdot E[q(z)]}{2\alpha \cdot E[s(z)^2]}\right]^{\frac{1}{2\alpha+\gamma}} \cdot \left[\frac{1}{N}\right]^{\frac{1}{2\alpha+\gamma}}, \tag{3.17}$$

then we have that $(h^{**} - h^*) = o\left(N^{-1/(2\alpha+\gamma)}\right)$.

4. Analysis of the Density-Weighted Average

4.1. Root-N Asymptotic Normality of the Density Weighted Average

The density weighted average $\hat{\delta}(h)$ has substantively different statistical properties that its associated nonparametric estimator, and involves different considerations for bandwidth choice. In particular, under certain rate conditions on the bandwidth h, $\hat{\delta}(h)$ is a $\sqrt{N}$ consistent, asymptotically normal estimator of δ_0, and the pointwise optimal bandwidth h^* does not generally satisfy those conditions.[15] In order to ensure that the bias of $\hat{\delta}(h)$ vanishes at rate $\sqrt{N}$, or $\sqrt{N}\,[\delta(h) - \delta_0] = o(1)$, the bandwidth h must satisfy

$$h = o(N^{-\frac{1}{2\alpha}}) \tag{4.1}$$

(from (3.9)), which is a condition not obeyed by h^*. The variance of $\hat{\delta}(h)$, as a U-statistic, vanishes at rate N provided that $E\left[|p(z_i, z_j, h)|^2\right] = o(N)$, (Lemma 3.1 of Powell, Stock and Stoker (1989)), which (from (3.12)) requires

$$h^{-1} = o(N^{\frac{1}{\gamma}}). \tag{4.2}$$

The expressions (4.1) and (4.2) bound the rate at which the bandwidth h converges to 0, and for both to hold simultaneously, we need $2\alpha > \gamma$. Further, $\hat{\delta}(h)$ is $\sqrt{N}$ asymptotically normally distributed if $E[|r_0(z_i)|^2] < \infty$, since we can write

$$\hat{\delta}(h) - \delta_0 = \frac{1}{N}\sum_{i=1}^{N} 2\,[r_0(z_i) - \delta_0] + o_p\left(\frac{1}{\sqrt{N}}\right), \tag{4.3}$$

so that $\sqrt{N}[\hat{\delta}(h) - \delta_0] \longrightarrow \mathcal{N}(0, V_0)$ with $V_0 \equiv 4 \cdot E\left\{[r_0(z_i) - \delta_0] \cdot [r_0(z_i) - \delta_0]'\right\}$.[16]

[15] Powell, Stock and Stoker (1989) give a rigorous derivation of the following conditions for Example 2.2, and Powell (1987) covers Example 2.3.

[16] This confirms our earlier assertion that $2\,[r_0(z_i) - \delta_0]$ represents the influence of the i^{th} observation in the asymptotic variance of $\hat{\delta}(h)$.

4.2. The Optimal Bandwidth for the Density Weighted Average

The optimal bandwidth for computing $\hat{\delta}(h)$ is computed by minimizing the leading terms of its mean squared error. Since $\hat{\delta}(h)$ is a U-statistic, its finite sample variance is found using the standard formulation in Serfling (1980, Section 5.5). With $p(\cdot)$ a scalar function as before, the finite sample variance of $\hat{\delta}(h)$ is

$$\begin{aligned} Var[\hat{\delta}(h)] &= \binom{N}{2}^{-1}[2(N-2)Var[r(z_i,h)] + Var[p(z_i,z_j,h)]] \\ &= 4N^{-1}Var[r(z_i,h)] + 2N^{-2}E[p(z_i,z_j,h)^2]^2 + o(N^{-2}). \end{aligned} \tag{4.4}$$

Consequently, we require a characterization for the variance of the conditional expectation $r(z_i,h) \equiv E[p(z_i,z_j,h)|z_i]$. From Assumption 1 it follows that

$$Var[r(z_i,h)] = Var[r_0(z_i)] + C_0 \cdot h^\alpha + o(h^\alpha), \tag{4.5}$$

where $C_0 \equiv 2Cov[r_0(z_i), s(z_i)]$.

We can now formulate the mean squared error of $\hat{\delta}(h)$ for δ_0. By combining (3.9), (3.12), (4.4) and (4.5), we have

$$\begin{aligned} MSE[\hat{\delta}(h)] &= [\delta(h) - \delta_0]^2 + Var\left[\hat{\delta}(h)\right] \\ &= \{E[s(z_i)]\}^2 h^{2\alpha} + 4N^{-1}Var[r_0(z_i)] + 2N^{-1}C_0 h^\alpha \\ &\quad + 2N^{-2}E[q(z_i)]h^{-\gamma} + o(h^{2\alpha}) + o\left(\tfrac{h^\alpha}{N}\right) + o\left(\tfrac{1}{N^2 h^\gamma}\right) + o(N^{-2}). \end{aligned} \tag{4.6}$$

For characterizing the optimal bandwidth, we subtract the variance term $4N^{-1}Var[r_0(z_i)]$ because it does not vary with the bandwidth h, writing (net) mean squared error as

$$\begin{aligned} MSE[\hat{\delta}(h)] - 4N^{-1}Var[r_0(z_i)] &= \{E[s(z_i)]\}^2 h^{2\alpha} + 2N^{-1}C_0 h^\alpha + 2N^{-2}E[q(z_i)]h^{-\gamma} \\ &\quad + \left[o(h^{2\alpha}) + o\left(\tfrac{h^\alpha}{N}\right) + o\left(\tfrac{1}{N^2 h^\gamma}\right) + o(N^{-2})\right] \\ &= T_1 + T_2 + T_3 + R. \end{aligned} \tag{4.7}$$

It is clear that T_1 and T_2 are increasing in h, while T_3 is decreasing in h. At a minimizing bandwidth sequence h^+, the term T_1 must be of larger order than T_2, since if T_2 were of larger order, the bandwidth which equates orders of squared bias and variance would be $O(N^{-1/(\alpha+\gamma)})$, which would imply T_1 is $O(N^{-2\alpha/(\alpha+\gamma)})$, which is of greater order than T_2, which would be $O(N^{(-2\alpha+\gamma)/(\alpha+\gamma)})$. Therefore, minimizing on the basis of the leading terms T_1 and T_3 gives the optimal bandwidth as

$$h^{+}=\left[\frac{\gamma\cdot E[q(z)]}{\alpha\cdot\{E[s(z)]\}^{2}}\right]^{\frac{1}{2\alpha+\gamma}}\cdot\left[\frac{1}{N}\right]^{\frac{2}{2\alpha+\gamma}}+o\left(\left[\frac{1}{N}\right]^{\frac{2}{2\alpha+\gamma}}\right). \tag{4.8}$$

This formula captures how h^{+} equates the leading orders of variance and squared bias — setting $O(1/N^{2}h^{+\gamma})=O(h^{+2\alpha})$ gives $h^{+}=O(N^{-2/(2\alpha+\gamma)})$. Moreover, the (net) mean squared error of $\hat{\delta}(h^{+})$ vanishes at rate

$$MSE[\hat{\delta}(h^{+})]-4N^{-1}Var[r_0(z_i)]=O\left(N^{-\frac{4\alpha}{2\alpha+\gamma}}\right). \tag{4.9}$$

Under the conditions for asymptotic normality, namely $2\alpha>\gamma$, the right hand remainder term is $o(N^{-1})$, but necessarily greater than $O(N^{-2})$.[17]

As before, we can approximate h^{+} by its leading term, which we denote as h^{++}. We can summarize this discussion as

Proposition 4.1. *Given Assumptions 1 and 2, if h^{+} denotes the bandwidth value that minimizes $MSE[\hat{\delta}(h^{+})]$ and*

$$h^{++}=\left[\frac{\gamma\cdot E[q(z)]}{\alpha\cdot\{E[s(z)]\}^{2}}\right]^{\frac{1}{2\alpha+\gamma}}\cdot\left[\frac{1}{N}\right]^{\frac{2}{2\alpha+\gamma}} \tag{4.10}$$

then we have that $(h^{++}-h^{+})=o\left(N^{-2/(2\alpha+\gamma)}\right)$.

4.3. Interpretation and Examples

Proposition 4.1 gives the main result of the paper. We have developed the structure of density weighted averages in some detail, to facilitate comparing the optimal bandwidth h^{+} with the pointwise optimal bandwidth h^{*}. A quick comparison

[17] Note that this result holds even if the conditions for $\sqrt{N}$ consistency of $\hat{\delta}(h)$ do not hold. In particular, in our examples, the condition $2\alpha>\gamma$ requires the use of a higher order kernel $\mathcal{K}(\cdot)$. If, on grounds of superior finite sample performance of the estimator, we took $\mathcal{K}(\cdot)$ as a standard positive density function, then our bandwidth analysis will still apply. This is relevant if a positive kernel $\mathcal{K}(\cdot)$ is used in any of our examples when $k>3$ (or $k>1$ in Example 2).

of (3.16) with (4.8) indicates that the bandwidth h^+ shrinks to 0 at twice the rate of the pointwise optimal bandwidth h^*. This feature of h^+ is referred to as "asymptotic undersmoothing," and occurs because pointwise squared bias must decrease at a faster rate than pointwise variance when estimating δ_0. Therefore h^+ must become smaller than h^* for sufficiently large samples. However, one cannot conclude that h^+ is smaller than h^* in any particular application, or for a given sample size, because of differences in the leading constants of h^+ and h^*.

We can spell this out by comparing the approximations h^{++} and h^{**}. In particular, we have that

$$h^{++} = A_N \cdot B \cdot h^{**}, \tag{4.11}$$

where

$$A_N = \left[\frac{2}{N}\right]^{\frac{1}{2\alpha+\gamma}} \tag{4.12}$$

is an adjustment factor for sample size, and

$$B = \left[\frac{E[s(z)^2]}{\{E[s(z)]\}^2}\right]^{\frac{1}{2\alpha+\gamma}} \tag{4.13}$$

is an adjustment factor for the structure of the pointwise bias.

The adjustment term A_N for sample size is less than 1 (for $N > 2$) and decreases to 0 as $N \rightarrow \infty$, reflecting the different rates of convergence discussed above. The adjustment term B for the structure of bias does not vary with sample size and is greater than one unless $s(z)$ is a constant function. In particular, $B^{2\alpha+\gamma}$ is one plus the squared coefficient of variation of $s(z_i)$. The factor B arises because the average pointwise mean square error depends on the variance of the pointwise bias of $\hat{r}(z,h)$, whereas the bias of $\hat{\delta}(h)$ depends only on the mean of the pointwise bias. Variation in the pointwise bias dictates a smaller pointwise bandwidth h^{**} relative to the bandwidth h^{++}. Whether the adjustment for sample size is larger or smaller than the adjustment for bias structure depends on the particular application.

To get a clearer notion of the sizes of the optimal bandwidths and these two effects, we compute bandwidths for Examples 2.1 and 2.2 based on normal designs.

4.3.1. Bandwidths for Average Density with Normal Variables

Recall from Example 2.1 that $z_i = x_i$ here, where $x_i \sim f(x)dx$, and the object of estimation is the average density value $\delta_0 = E[f(x_i)]$. Moreover, $\hat{r}(x_i, h)$ is the kernel density estimator $\hat{f}_i(x_i, h)$ of (2.5), and $r_0(x)$ is the density $f(x)$. For simplicity, we denote partial derivatives of f using subscripts: $f_j^{'} \equiv \partial f/\partial x_j$, $f_{jl}^{''} = \partial^2 f/\partial x_j \partial x_l$, etc., where each function is evaluated at x unless another argument value is indicated. Here $r(x,h) - r_0(x)$ is given from the familiar expression for bias of the kernel density estimator:

$$\begin{aligned} r(x,h) - r_0(x) &= E\left[h^{-k}\mathcal{K}\left(\tfrac{x_i - x_j}{h}\right)\right] - f(x) \\ &= \textstyle\int \mathcal{K}(u) f(x + hu) du. \end{aligned} \tag{4.14}$$

We take $\mathcal{K}(u)$ to be a positive density function, with $\int u\mathcal{K}(u)du = 0$ and $P = 2$, so that $\alpha = 2$ and $s(x_i) = (1/2)\partial^2 r/\partial h^2|_{h=0}$. In particular

$$s(x_i) = \left(\frac{1}{2}\right) \sum_l \sum_j [u_i u_j \mathcal{K}(u) du] f_{lj}^{''}(x_i). \tag{4.15}$$

For the variance term, we have that

$$\begin{aligned} E\left[p(x_i, x_j, h)^2 | x_i\right] &= E\left[h^{-2k}\mathcal{K}\left(\tfrac{x_i - x_j}{h}\right)^2\right] \\ &= h^{-k} \textstyle\int \mathcal{K}^2(u) f(x_i + hu) du \\ &= h^{-k} \textstyle\int \mathcal{K}^2(u) du \cdot f(x_i) + o_p(h^{-k}) \end{aligned} \tag{4.16}$$

so that

$$q(x_i) = \int \mathcal{K}^2(u) du \cdot f(x_i) \tag{4.17}$$

and $\gamma = k$.

To compute bandwidth values, we specialize the general formulae to the case where $f(x)$ is the spherical normal $\mathcal{N}(0, I)$ density, and $\mathcal{K}(u)$ is likewise chosen to be the $\mathcal{N}(0, I)$ density. These specifications imply first that

$$s(x_i) = \left(\frac{1}{2}\right) \sum_j f_{jj}^{''}(x_i). \tag{4.18}$$

If we define $c_k = (2\pi)^{-k/2}$, then some tedious arithmetic yields

$$E[s(x_i)]^2 = k^2 c_k^2 \left(\frac{1}{2}\right)^{k+4}, \tag{4.19}$$

$$E[s(x_i)^2] = k\left(k - \frac{2}{3}\right) c_k^2 \left(\frac{1}{3}\right)^{\frac{k}{2}+1}, \tag{4.20}$$

$$E[q(x_i)] = c_k^2 \left(\frac{1}{2}\right)^k, \tag{4.21}$$

and $2\alpha + \gamma = k + 4$. Therefore, the approximate pointwise optimal bandwidth (3.17) is

$$h^{**} = \left(\frac{\sqrt{3}}{2}\right)\left(\frac{4}{3k-2}\right)^{\frac{1}{k+4}}\left(\frac{1}{N}\right)^{\frac{1}{k+4}}. \tag{4.22}$$

The bias factor (4.13) is

$$B = \left(\frac{2}{\sqrt{3}}\right)\left(\frac{3k-2}{k}\right)^{\frac{1}{k+4}}, \tag{4.23}$$

and the size factor (4.12) is

$$A_N = \left(\frac{2}{N}\right)^{\frac{1}{k+4}}. \tag{4.24}$$

Consequently, the approximate optimal bandwidth (4.10) for estimating the average density is

$$h^{++} = A_N \cdot B \cdot h^{**} = \left(\frac{8}{k}\right)^{\frac{1}{k+4}}\left(\frac{1}{N}\right)^{\frac{2}{k+4}}. \tag{4.25}$$

Table 1 contains computed bandwidth values and bias and size factors for various dimensions k and sample sizes N. In terms of estimating average density versus estimating the density function, here the size factor always outweighs the bias factor, so that a smaller bandwidth should be used for estimating the average density. For increases in sample size, all bandwidths shrink, with the size effect much more pronounced in lower dimensional problems. Interestingly, for high dimension and low sample size, the optimal bandwidths for the pointwise and average density problems are nearly equal; also, for $k = 1$ and small N the pointwise bandwidths are larger than for $k = 2$, then increase monotonically in k.

4.3.2. Bandwidths for Density Weighted Average Derivatives, Linear Model With Normal Regressors

Recall from Example 2.2 that $z_i = (y_y, x_i^{'})'$ here, where $x_i \sim f(x)dx$ and $E(y_i|x_i) = g(x_i)$. We compute bandwidth values for estimation of the first component of the density weighted average derivative, $\delta_{01} = E[f(x_i)g_1^{'}(x_i)]$, where as above, $g_1^{'} = \partial g/\partial x$. We denote the first component of $\mathcal{V} = \partial\mathcal{K}/\partial u$ as $\mathcal{V}_1 = \partial\mathcal{K}/\partial u_1$. The function $r(z,h)$ is computed directly as the conditional expectation of the first component of $p(\cdot)$ of (2.8), as

$$\begin{aligned} r(z,h) &= E[p(z,z_j,h)|z] \\ &= -\int \mathcal{K}(u)\cdot y\cdot f_1^{'}(x+hu)du \qquad (4.26) \\ &\quad +\int \mathcal{K}(u)\left[g_1^{'}(h+hu)f(x+hu)+g(x+hu)f_1^{'}(x+hu)\right]du, \end{aligned}$$

where the latter equality employs integration by parts. As above, we have $\alpha = 2$, with $s(z_i) = (1/2)\partial^2 r/\partial h^2|_{h=0}$. Therefore

$$\begin{aligned} s(z) &= \left(\tfrac{1}{2}\right)\textstyle\sum_l\sum_j\{[\int u_l u_j\mathcal{K}(u)du][-y_i f_{1jl}^{'''}(x_i)+g(x_i)f_{1jl}^{'''}(x_i)+g_{1l}^{''}(x_i)f_j(x_i) \\ &\quad +g_1^{'}(x_i)f_{jl}^{''}(x_i)+g_{jl}^{''}(x_i)f_1^{'}(x_i)+g_j^{'}(x_i)f_{1l}^{''}(x_i)+g_1^{'}(x_i)f_{1j}^{''}(x_i)]\} \end{aligned} \qquad (4.27)$$

The variance term is found similarly as

$$q(z_i) = \int \mathcal{V}_1^2(u)du\cdot[y_i - 2y_i g(x_i) + E(y_i^2|x_i)], \qquad (4.28)$$

where here $\gamma = k+2$.

To compute bandwidth values, we again specialize the general formulae to the case where $f(x)$ is the spherical normal $\mathcal{N}(0,I)$ density, and $\mathcal{K}(u)$ is likewise the $\mathcal{N}(0,I)$ density. We assume the true model is linear;

$$y_i = \sum_{j=1}^{k} x_{ji} + \varepsilon_i, \qquad (4.29)$$

where ε_i is univariate normal with mean 0 and variance σ_k, and independent of x. This allows $s(z_i)$ to be simplified as

$$s(z_i) = \left(\frac{1}{2}\right)\sum_j[-f_{1jj}^{'''}(x_i) + f_{jj}^{''}(x_i) + 2f_{ij}^{''}(x_i)]. \qquad (4.30)$$

Considerably more tedious algebra gives

$$E[s(x_i)]^2 = k^2 c_k^2 \left(\frac{1}{2}\right)^{k+4}, \tag{4.31}$$

$$E[s(x_i)^2] = c_k^2 \left(\frac{1}{3}\right)^{\frac{k}{2}+2} \left[3k^2 - 1.25k + 7.25 + \sigma_k \left(\frac{k^2}{3} + \frac{k}{6} + 3\right)\right], \tag{4.32}$$

$$E[q(x_i)] = c_k \left(\frac{1}{2}\right)^{\frac{k}{2}+1} \sigma_k, \tag{4.33}$$

and $2\alpha + \gamma = k + 6$. The approximate pointwise bandwidth h^{**} , the bias factor B, the size factor A_N and the approximate optimal bandwidth h^{++} are computed from these terms as above. Tables 2A and 2B contain computed bandwidth values and bias and size factors for various dimensions k and sample sizes N. To make the values comparable across dimensions, the linear model (4.29) is assumed to have a constant value of R^2 , so that the variance of ε_i varies with k as

$$\sigma_k = k\left(\frac{1-R^2}{R^2}\right). \tag{4.34}$$

Table 2A gives values for $R^2 = .80$, and Table 2B gives values for $R^2 = .20$. These bandwidth values are qualitatively similar to those in Table 1, but there are some notable differences. As before, the sample size effect is much more pronounced in low dimensions. The bandwidths for estimating average derivatives are generally larger than those for estimating average density, and increase as the R^2 value decreases. The bias factor is more evident for average derivatives as well - for instance, the bandwidths for estimating average derivatives are larger than the pointwise bandwidths for smaller sample sizes in Table 2B. Finally, for $R^2 = .20$, several of the optimal pointwise bandwidths actually exceed their "averaged" counterparts, though this possibility vanishes as N increases relative to k.

4.4. A "Plug-In" Estimator of the Optimal Bandwidth

One approach for approximating the optimal bandwidth h^+ is to make use of cross validation or another data-based method for estimating the pointwise optimal bandwidth h^*,[18] and then applying the factorization (4.11) using the appropriate

[18] See Nychka (1991) for a recent discussion of smoothing parameter choice and references to automatic methods for kernel estimation.

A_N value and an approximate value of B. To the extent that the application at hand is similar to one of the examples above, the B values in Tables 1, 2A or 2B may give reasonable performance.[19]

Alternatively, we propose a simple "plug-in" estimator of h^+, based on empirical implementations of the bias and variance formulations above.[20] In particular, to approximate (4.8) or (4.9), we need consistent estimators of $Q_0 \equiv E[q(z_i)]$ and $S_0 \equiv E[s(z_i)]$. Denoting such estimators as $\hat{Q}$ and $\hat{S}$ respectively, the optimal bandwidth is estimated as

$$\hat{h}=\left[\frac{\gamma \cdot \hat{Q}}{\alpha \cdot \hat{S}^2}\right]^{\frac{1}{2\alpha+\gamma}} \cdot \left[\frac{1}{N}\right]^{\frac{2}{2\alpha+\gamma}}. \tag{4.35}$$

From the consistency of $\hat{Q}$ and $\hat{S}$, it follows immediately that

$$\hat{h} - h^{++} = o_p(N^{-\frac{2}{2\alpha+\gamma}}), \tag{4.36}$$

for h^{++} of (4.10), and therefore we can conclude that

$$\hat{h} - h^{+} = o_p(N^{-\frac{2}{2\alpha+\gamma}}) \tag{4.37}$$

for the optimal bandwidth h^+ of (4.8).

The variance coefficient Q_0 is estimated by empirically implementing (3.12). Using an initial bandwidth value h_0 , define

$$\hat{Q} \equiv \hat{Q}(h_0) = \binom{N}{2}^{-1} \sum_{i<j} h_0^{\gamma} \cdot p(z_i, z_j, h_0)^2. \tag{4.38}$$

Consistency of $\hat{Q}$ for Q_0 is obtained by allowing h_0 to converge to 0 at a different rate than the optimal bandwidth h^+. From the U-statistic structure of (4.38), if

$$E[p(z_i, z_j, h_0)^4] = O(h_0^{-\eta-2\gamma}), \tag{4.39}$$

[19] Hall and Marron (1987) propose an interesting "plug-in" method of estimating the pointwise optimal bandwidth for density estimation using an estimate of the integrated squared density derivative. They exploit a connection due to integration-by-parts that is not available in our general setting.

[20] Gasser, Kneip and Kohler (1991) discuss iterative "plug-in" methods for approximating optimal pointwise bandwidths.

for some $\eta > 0$, then Lemma 3.1 of Powell, Stock and Stoker (1989) implies that $\hat{Q}$ is consistent for Q_0 if $h_0 \to 0$ and $Nh_0^{\eta+2\gamma} \to 0$ as $N \to \infty$.

To estimate the bias coefficient S_0 , we exploit (3.9) in differenced form. Since S_0 is the leading term in the bias expansion of $\hat{\delta}(h)$, estimate S_0 by

$$\hat{S} = \frac{\hat{\delta}(\tau h_0) - \hat{\delta}(h_0)}{(\tau h_0)^\alpha - h_0^\alpha} \tag{4.40}$$

for some positive $\tau \neq 1$. This has expectation

$$E\left(\hat{S}\right) = \left[(\tau h_0)^\alpha - h_0^\alpha\right]^{-1} \cdot \left\{S_0\left[(\tau h_0)^\alpha - h_0^\alpha\right] + o(h_0^\alpha)\right\} = S_0 + o(1), \tag{4.41}$$

assuming $h_0 \to 0$ as $N \to \infty$. Since $\hat{S}$ is a U-statistic with kernel

$$p_S(z_i, z_j, h_0) = h_0^{-\alpha}[p(z_i, z_j, h_0) - p(z_i, z_j, \tau h_0)]/(\tau^\alpha - 1), \tag{4.42}$$

we have that

$$E[p_S^2] = O(h_0^{-2\alpha}E[p^2]) = O(h_0^{-2\alpha-\gamma}), \tag{4.43}$$

so that $\hat{S}$ is consistent for S_0 provided $h_0 \to 0$ and $Nh_0^{2\alpha+\gamma} \to \infty$ as $N \to \infty$.[21]

In summary, we have shown

Proposition 4.2. *Under Assumptions 1 and 2, suppose that (4.39) is valid for $\eta > 0$, and let $\rho = max\{\eta + 2\gamma, 2\alpha + \gamma\}$. If $h_0 \to 0$ and $Nh_0^\rho \to \infty$ as $N \to \infty$, then $\hat{Q} = \hat{Q}(h)$ and $\hat{S} = \hat{S}(h)$ are consistent estimators of $Q_0 \equiv E[q(z_i)]$ and $S_0 \equiv E[s(z_i)]$ respectively, and the "plug-in" bandwidth estimator $\hat{h}$ obeys*

$$\hat{h} - h^+ = o_p(N^{\frac{-2}{2\alpha+\gamma}}). \tag{4.44}$$

While Proposition 4.2 gives a solution to the problem of estimating h^+, we must mention one technical proviso of the result. We have not shown that our conditions will guarantee that the "plug-in" estimator $\hat{\delta}(\hat{h})$ will be asymptotically equivalent to $\hat{\delta}(h^+)$, because the mean squared error calculations used to derive the form of the optimal bandwidth held h fixed in calculating the moments of

[21] With reference to the note following (4.9), the estimator $\hat{S}$ can estimate the bias in situations where $\hat{\delta}(h)$ does not obey all the tenets of $\sqrt{N}$ consistency for δ_0. While our justification of $\hat{S}$ is asymptotic, the estimator $\hat{S}$ may provide a practical method of studying small sample bias issues, such as those discussed in Stoker (1993a,1993b).

the U-statistic in (2.1). The MSE formulae does not follow immediately if the bandwidth h were replaced by a (stochastic) bandwidth value that was constructed using the same data as appear in the U-statistic, and derivation of general large-sample properties of $\hat{\delta}(\hat{h})$ would require much stronger conditions (e.g. higher order differentiability) on the kernel function $p(z_i, z_j, h)$.

A straightforward but inelegant solution to this technical problem could be based upon a familiar "sample-splitting" device — the bandwidth $\hat{h}$ could be constructed using, say, the first $N^* = O(\ln(N))$ observations on z_i, with the remaining observations being used to form the U-statistic in (2.1). While this sample-splitting approach would ensure the equivalence of $\hat{\delta}(\hat{h})$ and $\hat{\delta}(h^+)$, it would clearly lead to very imprecise estimates of the optimal bandwidth in practice, and an approach that made use of the whole sample in both steps seems more likely to be well-behaved. Another straightforward solution would be to discretize the set of possible scaling constants, replacing the estimated constant term with the closest value in some finite set. While the optimal constant will generally not be in this set, it can be arbitrarily well approximated if the mesh of the set is small enough; the rate of convergence of the the discretized constant will be arbitrarily high in probability, so this "plug-in" bandwidth will not affect the asymptotic properties of the U-statistic.

5. The Optimal Bandwidth for Estimating Ratios of Density Weighted Averages

As outlined for Examples 2.2 and 2.3, there are various estimators of interest that take the form of ratios of density weighted averages. We can solve the optimal bandwidth problem for estimators of this type by a straightforward modification of the derivations above. We now spell out this modification.[22]

In particular, as motivated by (2.9) and (2.13) of Examples 2.2 and 2.3, we are often interested in the estimation of a parameter of the form

$$\beta_0 = [\delta_0^x]^{-1}\,\delta_0^y. \tag{5.1}$$

[22] We do not consider the possibility of using different bandwidths for the "numerator" and "denominator" of the ratio, but rather a single bandwidth for each. Our version is directly motivated by certain problems, such as the instrumental variables estimator of Example 2.2. Different bandwidths would imply that different instruments are used for the "x" moments than for the "y" moments.

The corresponding estimator takes the form

$$\hat{\beta}(h) = \left[\hat{\delta}^x(h)\right]^{-1} \hat{\delta}^y(h). \tag{5.2}$$

The optimal bandwidth for the estimator (5.2) can be studied by manipulations similar to those familiar from linear regression analysis. In particular, we focus the variation of $\hat{\beta}(h)$ on the "residuals" of the problem by defining

$$\hat{\delta}^u(h) \equiv \hat{\delta}^y(h) - \hat{\delta}^x(h)\beta_0. \tag{5.3}$$

$\hat{\delta}^u(h)$ is a second-order U-statistic of the form (2.1) with kernel

$$u(z_i, z_j, h) \equiv p^y(z_i, z_j, h) - p^x(z_i, z_j, h)\beta_0, \tag{5.4}$$

where $p^y(\cdot)$ and $p^x(\cdot)$ are the kernels of the U statistics $\hat{\delta}^y(h)$ and $\hat{\delta}^x(h)$ respectively. By construction, $\hat{\delta}^u(h)$ is an estimator of

$$\delta_0^u = \delta_0^y - \delta_0^x\beta_0 = 0. \tag{5.5}$$

Returning to $\hat{\beta}(h)$, we have that

$$\begin{aligned}\hat{\beta}(h) - \beta_0 &= \left[\hat{\delta}^x(h)\right]^{-1} \hat{\delta}^u(h) \\ &= \left[\delta_0^x\right]^{-1} \cdot \hat{\delta}^u(h) + \left(\left[\hat{\delta}^x(h)\right]^{-1} - \left[\delta_0^x\right]^{-1}\right) \hat{\delta}^u(h).\end{aligned} \tag{5.6}$$

As long as $\hat{\delta}^x(h)$ is N^η-consistent for some $\eta > 0$, the second term on the right-hand-side of (5.6) is of smaller order (in N) than the first, so that the rate of convergence of $\hat{\beta}(h)$ to β_0 is the same as that of $\hat{\delta}^u(h)$ to zero. Thus the optimal bandwidth for $\hat{\beta}(h)$ is the same as that for the U-statistic based on the scaled residuals $\left[\delta_0^x\right]^{-1} u(z_i, z_j, h)$.

It is easy to check that the use of an estimated "residual" does not change any of the conclusions above, provided the rate of convergence of $\hat{\delta}^x(h)$ is the same as for $\hat{\delta}^u(h)$. To be more precise, suppose that $\tilde{\delta}^x$ and $\tilde{\beta}$ are consistent estimators of δ_0^x and β_0, say based on an initial bandwidth value, and define

$$\left[\tilde{\delta}^x\right]^{-1} \tilde{\delta}^u(h) \equiv \left[\tilde{\delta}^x\right]^{-1} \left[\hat{\delta}^y(h) - \hat{\delta}^x(h)\tilde{\beta}\right]. \tag{5.7}$$

Then we have that

$$\begin{aligned}\left[\tilde{\delta}^x\right]^{-1}\tilde{\delta}^u(h) &= [\delta_0^x]^{-1}\tilde{\delta}^u(h) - \left([\delta_0^x]^{-1} - \left[\tilde{\delta}^x\right]^{-1}\right)\tilde{\delta}^u(h) - \left[\tilde{\delta}^x\right]^{-1}\hat{\delta}^x(h)\left[\tilde{\beta} - \beta_0\right] \\ &= [\delta_0^x]^{-1}\tilde{\delta}^u(h) + o_p\left([\delta_0^x]^{-1}\tilde{\delta}^u(h)\right) + o_p\left(\hat{\delta}^x(h)\right). \end{aligned} \tag{5.8}$$

Replacement of $\left[\tilde{\delta}^x\right]^{-1}\tilde{\delta}^u(h)$ by $[\delta_0^x]^{-1}\tilde{\delta}^u(h)$ does not affect the leading terms of mean squared error, nor does it alter the optimal bandwidth value.

In summary, we have shown

Proposition 5.1. *Under Assumptions 1 and 2 applied to $\hat{\delta}^y(h)$ and $\hat{\delta}^x(h)$, the optimal bandwidth for estimating the ratio $\hat{\beta}(h) = [\hat{\delta}^x(h)]^{-1}\hat{\delta}^y(h)$ is the same as that for estimating $[\delta_0^x]^{-1}\hat{\delta}^u(h)$, the density weighted average with scaled "residual" kernel $[\delta_0^x]^{-1}u(z_i, z_j, h)$ of (5.4). This bandwidth is consistently estimated using consistent estimates $\tilde{\delta}^x$ and $\tilde{\beta}$ to construct the scaled residual kernel.*

It is worthwhile noting one special case of our results, because of its appearance in common model designs. In particular, there are settings where the biases in $\hat{\beta}(h)$ and $\hat{\delta}^u(h)$ are identically zero. For instance, consider the instrumental variables estimators for Example 2.2 (implementing (2.7)). If the true model is linear, since the "instruments" $\partial\hat{f}_i(x_i)/\partial x$ are solely functions of x, the coefficient estimator $\hat{d}(h)$ is conditionally unbiased for the linear coefficients. This implies that the optimal bandwidth calculation only contains the terms for the variance of $\hat{d}(h)$, which are decreasing in h. Consequently, one will want to set the bandwidth to a "large" value in this setting.[23] This is reflected in the optimal bandwidth formula (4.11) by noting that $E[s(x_i)]^2 = 0$ in this case. At any rate, our "plug-in" estimation method of Section 4 permits determining whether this is the case empirically, in that a small estimate of the bias in $\hat{\delta}^u(h)$ will translate to the large bandwidth value.

6. Conclusion

In this paper we have characterized the optimal bandwidth for estimating density-weighted averages, and ratios of density-weighted averages. Our main purpose was

[23] More precisely, the optimal bandwidth value may not shrink with sample size in this case. However, while larger bandwidth values decrease variance, at the limit $h = \infty$, $\partial\hat{f}_i(x_i)/\partial x_i$ is everywhere zero, and therefore fails to be a proper instrument.

to provide a guide for choice of bandwidth for estimators that employ kernel estimators in the form of density weighted averages. Most of the existing asymptotic theory for these estimators had little to say about how to set bandwidth values in applications, so our results give more specific help for this problem. A natural next step is to study the performance of fully automatic methods (namely estimators that use estimates of optimal bandwidths), to see whether approximating an optimal bandwidth gives rise to real practical benefits.

One concern raised by the computed bandwidth values of Section 4.3 is the quality of the asymptotic approximations we have employed. In particular, if the bandwidth is of the same order (say one half) of the standard error of the data components, then one could question the quality of our analysis based on leading terms in a series expansion. Further research is indicated to see whether the remainder terms substantially affect the optimal bandwidth value. Nevertheless, computing the estimators of Section 4.4 (of the leading coefficients of bias and variance) will be informative in applications, for indicating how sensitive the estimated results are to the bandwidth values used.

Another object of the paper was to give a concrete comparison of bandwidth settings for optimal function approximation versus optimal performance of a derived semiparametric estimator. The simple framework above gave rise to an immediate relationship of this type, and permitted us to compare a pointwise optimal bandwidth with an optimal bandwidth for the semiparametric problem. Since the nonparametric estimator for this exercise is intrinsically connected to the density-weighted average of interest, our results somewhat beg the question of bandwidth choice in other semiparametric contexts. However, the simple structure of our results may provide some general insight in how to adjust for different uses of nonparametric estimators, as well as how to practically implement "asymptotic undersmoothing."

References

[1] Andrews, D.W.K. (1989), "Asymptotics for Semiparametric Econometric Models: I. Estimation," Working Paper, Cowles Foundation, Yale University.

[2] Bickel, P.J. and Y. Ritov (1988), "Estimating Integrated Squared Density Derivatives: Sharp Best Order of Convergence Estimates," *Sankhyā*, 50, Se-

ries A, 381-393.

[3] Gasser, T., A. Kneip and W. Kohler (1991), "A Flexible and Fast Method for Automatic Smoothing," *Journal of the American Statistical Association*, 86, 643-653.

[4] Goldstein, L. and K. Messer (1990), "Optimal Plug-in Estimators for Nonparametric Functional Estimation," Technical Report No. 277, Stanford University.

[5] Hall, P. and J.S. Marron (1987), "Estimation of Integrated Squared Density Derivatives," *Statistics and Probability Letters*, 6, 109-115.

[6] Härdle, W. (1991), *Applied Nonparametric Regression*, Cambridge, Cambridge University Press, Econometric Society Monographs.

[7] Härdle, W., P. Hall, and S. Marron (1988), "How Far are Automatically Chosen Regression Smoothing Parameters from Their Optimum?" *Journal of the American Statistical Association*, 83, 86-95.

[8] Härdle, 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.

[9] Härdle, W. and T. M. Stoker(1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986-995.

[10] Härdle, W. and A.B. Tsybakov (1993), "How Sensitive are Average Derivatives?," *Journal of Econometrics*, 58, 31-48.

[11] Hastie, T.J. and R.J. Tibshirani (1990), *Generalized Additive Models*, London, Chapman and Hall.

[12] Hoeffding, W. (1948), "A Class of Statistics with Asymptotically Normal Distribution," *Annals of Mathematical Statistics,* 19, 293-325.

[13] Jaeckel, L.A. (1972), "Estimating Regression Coefficients by Minimizing the Dispersion of the Residuals," *Annals of Mathematical Statistics*, 43, 1449-1458.

[14] 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.

[15] Jurečková, J. (1971), "Nonparametric Estimate of Regression Coefficients," *Annals of Mathematical Statistics*, 42, 1328-1338.

[16] Newey, W.K. (1991), "The Asymptotic Variance of Semiparametric Estimators," Working Paper No. 583, Department of Economics, MIT, revised July.

[17] Nychka, D. (1991), "Choosing a Range for the Amount of Smoothing in Nonparametric Regression," *Journal of the American Statistical Association,* 86, 653-665.

[18] Powell, J.L. (1987), "Semiparametric Estimation of Bivariate Latent Variables Models," draft, University of Wisconsin.

[19] Powell, J.L., J.H. Stock and T.M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403-1430.

[20] Robinson, P.M. (1988b), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931-954.

[21] Serfling, R.J. (1980), Approximation Theorems of Mathematical Statistics, New York, Wiley.

[22] Stoker, T.M. (1992), *Lectures on Semiparametric Econometrics*, CORE Lecture Series, CORE Foundation, Louvain-la-Neuvre.

[23] Stoker, T.M. (1993a), "Smoothing Bias in the Estimation of Density Derivatives," forthcoming *Journal of the American Statistical Association.*

[24] Stoker, T.M. (1993b), "Smoothing Bias in the Measurement of Marginal Effects," MIT Sloan School of Management Working Paper, revised January.

TABLE 1: OPTIMAL BANDWIDTHS FOR ESTIMATING DENSITY AND AVERAGE DENSITY

Specification:

Spherical Normal Regressors: $f(x) = \mathcal{N}(0,I)$ Density

Spherical Normal Kernel: $\mathcal{K}(u) = \mathcal{N}(0,I)$ Density

Dimension k	1	2	3	4	5	10
Bias Factor	1.155	1.296	1.303	1.295	1.284	1.243
N = 50						
Bandwidth (PW)	0.523	0.451	0.457	0.474	0.492	0.570
Size Factor	0.525	0.585	0.631	0.669	0.699	0.795
Bandwidth (Ave)	0.317	0.342	0.376	0.410	0.442	0.563
N = 100						
Bandwidth (PW)	0.455	0.402	0.414	0.434	0.455	0.542
Size Factor	0.457	0.521	0.572	0.613	0.647	0.756
Bandwidth (Ave)	0.240	0.271	0.309	0.345	0.379	0.510
N = 500						
Bandwidth (PW)	0.330	0.307	0.329	0.355	0.381	0.483
Size Factor	0.331	0.398	0.454	0.501	0.541	0.674
Bandwidth (Ave)	0.126	0.159	0.195	0.231	0.265	0.405
N = 1000						
Bandwidth (PW)	0.287	0.274	0.298	0.326	0.353	0.460
Size Factor	0.289	0.355	0.412	0.460	0.501	0.642
Bandwidth (Ave)	0.096	0.126	0.160	0.194	0.227	0.367
N = 5000						
Bandwidth (PW)	0.208	0.209	0.237	0.266	0.295	0.410
Size Factor	0.209	0.271	0.327	0.376	0.419	0.572
Bandwidth (Ave)	0.050	0.074	0.101	0.130	0.159	0.292
N = 10000						
Bandwidth (PW)	0.181	0.187	0.214	0.244	0.273	0.390
Size Factor	0.182	0.242	0.296	0.345	0.388	0.544
Bandwidth (Ave)	0.038	0.058	0.083	0.109	0.136	0.264
N = 100000						
Bandwidth (PW)	0.114	0.127	0.154	0.183	0.211	0.331
Size Factor	0.115	0.165	0.213	0.259	0.301	0.462
Bandwidth (Ave)	0.015	0.027	0.043	0.061	0.082	0.190

TABLE 2: OPTIMAL BANDWIDTHS FOR DENSITY WEIGHTED AVERAGE DERIVATIVES

Specification:

Spherical Normal Regressors: $f(x) = \mathcal{N}(0,I)$ Density

Spherical Normal Kernel: $\mathcal{K}(u) = \mathcal{N}(0,I)$ Density

Linear Model: $y_i = \sum_{j=1}^{k} x_{ji} + \varepsilon_i \; ; \; i = 1,\ldots,N$

$$\varepsilon \sim \mathcal{N}(0,\sigma_k^2); \quad \sigma_k^2 = k \left[\frac{1 - R^2}{R^2} \right]$$

TABLE 2A: $R^2 = .80$

Dimension k	1	2	3	4	5	10
Bias Factor	1.537	1.354	1.302	1.279	1.265	1.235
N = 50						
Bandwidth (PW)	0.472	0.621	0.743	0.852	0.953	1.355
Size Factor	0.631	0.669	0.699	0.725	0.746	0.818
Bandwidth (Ave)	0.458	0.563	0.677	0.790	0.900	1.368
N = 100						
Bandwidth (PW)	0.428	0.570	0.688	0.795	0.895	1.297
Size Factor	0.572	0.613	0.647	0.676	0.701	0.783
Bandwidth (Ave)	0.376	0.473	0.580	0.688	0.793	1.254
N = 500						
Bandwidth (PW)	0.340	0.466	0.575	0.677	0.773	1.173
Size Factor	0.454	0.501	0.541	0.576	0.605	0.708
Bandwidth (Ave)	0.237	0.316	0.406	0.498	0.592	1.026
N = 1000						
Bandwidth (PW)	0.308	0.427	0.533	0.632	0.726	1.123
Size Factor	0.412	0.460	0.501	0.537	0.568	0.678
Bandwidth (Ave)	0.195	0.266	0.348	0.434	0.522	0.941
N = 5000						
Bandwidth (PW)	0.245	0.349	0.445	0.538	0.627	1.016
Size Factor	0.327	0.376	0.419	0.457	0.491	0.613
Bandwidth (Ave)	0.123	0.178	0.243	0.315	0.390	0.769
N = 10000						
Bandwidth (PW)	0.222	0.320	0.412	0.502	0.589	0.973
Size Factor	0.296	0.345	0.388	0.427	0.461	0.587
Bandwidth (Ave)	0.101	0.150	0.208	0.274	0.343	0.705
N = 100000						
Bandwidth (PW)	0.159	0.240	0.319	0.399	0.478	0.842
Size Factor	0.213	0.259	0.301	0.339	0.374	0.509
Bandwidth (Ave)	0.052	0.084	0.125	0.173	0.226	0.529

TABLE 2B: $R^2 = .20$

Dimension k	1	2	3	4	5	10
Bias Factor	1.734	1.543	1.475	1.440	1.418	1.359
N = 50						
Bandwidth (PW)	0.622	0.771	0.893	0.999	1.094	1.463
Size Factor	0.631	0.669	0.699	0.725	0.746	0.818
Bandwidth (Ave)	0.681	0.796	0.921	1.042	1.158	1.627
N = 100						
Bandwidth (PW)	0.563	0.707	0.827	0.932	1.027	1.401
Size Factor	0.572	0.613	0.647	0.676	0.701	0.783
Bandwidth (Ave)	0.559	0.669	0.789	0.908	1.021	1.492
N = 500						
Bandwidth (PW)	0.448	0.579	0.691	0.793	0.887	1.267
Size Factor	0.454	0.501	0.541	0.576	0.605	0.708
Bandwidth (Ave)	0.353	0.448	0.552	0.658	0.762	1.220
N = 1000						
Bandwidth (PW)	0.405	0.531	0.640	0.740	0.833	1.214
Size Factor	0.412	0.460	0.501	0.537	0.568	0.678
Bandwidth (Ave)	0.289	0.376	0.473	0.573	0.672	1.119
N = 5000						
Bandwidth (PW)	0.322	0.434	0.535	0.630	0.720	1.097
Size Factor	0.327	0.376	0.419	0.457	0.491	0.613
Bandwidth (Ave)	0.183	0.252	0.331	0.415	0.501	0.915
N = 10000						
Bandwidth (PW)	0.292	0.398	0.495	0.588	0.676	1.051
Size Factor	0.296	0.345	0.388	0.427	0.461	0.587
Bandwidth (Ave)	0.150	0.212	0.284	0.361	0.442	0.839
N = 100000						
Bandwidth (PW)	0.210	0.298	0.384	0.467	0.548	0.910
Size Factor	0.213	0.259	0.301	0.339	0.374	0.509
Bandwidth (Ave)	0.078	0.119	0.170	0.228	0.291	0.629