# SMOOTHING BIAS IN DENSITY DERIVATIVE ESTIMATION

by

Thomas M. Stoker
Massachusetts Institute of Technology

WP#3336-91-EFA August 1991

# SMOOTHING BIAS IN DENSITY DERIVATIVE ESTIMATION

by

Thomas M. Stoker

August 1991

* Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA. The author wishes to thank A. Caplan, R. Carroll, K. Doksum, G. Chamberlain, W. Härdle, J. Heckman, D. Jorgenson, A. Lewbel, S. Marron, R. Matzkin, B. Nalebuff, W. Newey, J. Powell, M. Rothchild and A. Zellner for helpful comments.

Abstract

This paper discusses a fundamental feature of density estimation by smoothing, namely that estimated density derivatives and score vectors will display a downward bias. We analyze the behavior of kernel estimators with finite bandwidths, showing how the downward bias arises from Jensen's inequality, as well as from the convolution structure of the estimator. A result is shown to confirm intuition that is immediate from pictures.

We then consider the estimation of density score vectors. We motivate interest in estimating score vectors by considering average derivative estimation and adaptive estimation of location models. The bias in score vectors is characterized for normally distributed variables, as well as variables distributed via a normal mixture. For normal variables and a normal kernel, the score bias is uniformly proportional. We calculate the bias for approximately optimal bandwidth values, and note that it can be substantial. For normal mixtures, we indicate that the score bias can be approximately proportional. A simple diagnostic statistic (and/or correction) for score bias is proposed.

KEYWORDS: Nonparametric; Kernel; Density; Derivative; Bias

# SMOOTHING BIAS IN DENSITY DERIVATIVE ESTIMATION

by Thomas M. Stoker

## 1. Introduction

The study of nonparametric methods for estimating unknown functions is one of the most rapid and extensive current movements in mathematical statistics. This spectacular development has opened up the real possibility of full characterizations of the statistical structure underlying empirical data, in a fashion that is virtually free of restrictive modeling assumptions. Moreover, this development has been advantageously complemented by advances in computing power and statistical graphics, which establish the feasibility of using nonparametric methods far beyond reasonable expectations of even a decade ago.

The theoretical development of nonparametric methods has resulted in a clear understanding of the factors involved in statistical approximation of functions in large samples, as well as how optimal large sample performance can be obtained with specific procedures. Quite familiar is the tradeoff between pointwise bias and variance, how balancing them can yield optimal rates of convergence, and how the smoothness and dimensionality of the true function play essential roles. Likewise, quite familiar are methods for exploiting functional smoothness in specific procedures; for instance, the use of higher-order kernels in kernel density and regression estimation; as well as the characterization of rules for choosing bandwidths that yield optimal asymptotic approximation, such as the various versions of cross validation.[1] Further, recent work has demonstrated how "plug-in" semiparametric estimators can exhibit $\sqrt{N}$ rates of convergence, while using nonparametric estimators as basic ingredients. The asymptotic conditions for nonparametric estimation in

semiparametric problems typically differ from those of optimal approximation of unknown functions; for instance, involving "asymptotic undersmoothing" to reduce pointwise bias more rapidly than pointwise variance.[2]

At any rate, the import of this theoretical work is to ascribe substantial promise to the use of nonparametric estimators as empirical tools, at least in problems of low or moderate dimension. Whether one employs (a truncated version of) an infinite series expansion, or a method based on local averages (kernel or nearest neighbor, for instance), it is natural to expect that the results will give an accurate depiction of the density or regression function under study. In particular, such results should be free of any systematic errors that would be introduced by using a parametric model that was badly specified, namely one that cannot closely approximate the true function of interest. That is, all functional attributes should be well exhibited by a nonparametric estimator, including derivatives, extrema and other features. The work on "plug in" semiparametric estimators enhances this view, by implying that dimensionality problems of the nonparametric ingredients can be avoided when estimation is focused on a parameter, or functional of interest.

Relative to the work on asymptotic theory, there is less attention in the literature to the actual performance characteristics of nonparametric estimators in finite samples. Moreover, it is easy to imagine extreme cases where the asymptotic approximation theory would be of limited value. For one instance, suppose that a complicated relationship was to be nonparametrically approximated by polynomial regression, but that data limitations required truncation (elimination) of all terms of quadratic and higher order. In such a case, a method based on local averages could be preferable, because of a more intrinsic ability to pick up bumps or other nonlinear structure. At any rate, with small or moderate sample sizes, it is

possible that the biases of nonparametric estimators are considerable, with the standard asymptotic theory irrelevant or uninformative.

This paper points out a particular type of systematic bias in nonparametric density estimators based on local averages. In particular, because of local smoothing, pointwise estimates of density derivatives are typically too small. This feature is based on the simple notion that smoothing will tend to make a surface flatter, ergo the measured slopes will be too small. The overall purpose of the paper is show how downward derivative bias is a generic feature of density estimation by smoothing, as well as indicate how the downward bias can be substantial, even with large sample sizes.[3]

It is important to stress that the existence of bias in local averages is not surprising nor novel. However, our point is that the bias can lead to systematic mismeasurement in an important aspect, namely derivatives (and density scores). The downward bias problem depends directly on the amount of smoothing: tiny bandwidth values make the problem disappear, in accordance with the popular asymptotic theory.

We consider the estimation of a continuous density f(x) of a k-vector of predictors x, its derivative $f'(x) \equiv \partial f/\partial x$, and its (translation) score $\ell(x) \equiv -\partial \ln f/\partial x = -f'/f$. We assume that the data $\{x_i, i=1,\ldots,N\}$ is a random sample, and we focus on the standard (Rosenblatt-Parzen) kernel density estimator

$$\hat{f}(x) = N^{-1}h^{-k} \sum_{i=1}^{N} \mathcal{K}\left(\frac{x - x_i}{h}\right) , \tag{1.1}$$

where $\mathcal{K}(.)$ denotes a positive (differentiable) symmetric kernel density, where the bandwidth value h controls the amount of smoothing.

The exposition is organized as follows. The simple conceptual foundation

for downward derivative bias is discussed in Section 2, where the bias of the estimated density derivative $\hat{f}'(x) \equiv \partial \hat{f}/\partial x$ is studied. The typical downward bias is immediately evident from density diagrams consistent with Jensen's inequality. A result is given to confirm the intuition of the pictures.

The overall size and nature of the downward bias is addressed in Sections 3 and 4, where the estimated score $\hat{\ell}(x) \equiv - \hat{f}'(x)/\hat{f}(x) = -\partial \ln \hat{f}/\partial x$ is studied. Section 3 gives some basic motivation for measuring the score from two semiparametric problems, namely the estimation of average derivatives and adaptive estimation of location models, as well as some indications of the size of the bias. Section 4 begins with a basic formulation of the bias of $\hat{\ell}(x)$, and then displays the score bias when the true density is multivariate normal (with a normal kernel), and when the true density is a normal mixture. These formulations permit the bias to be computed for "optimal" bandwidth values for different sample sizes, as well as suggesting an approximate structure of the bias. Namely, the score bias is exactly proportional for each value of x in the normal case, and approximately proportional in the mixture case. This helps clarify when the score bias problem may be ignored, as well as motivates a simple level correction for it. We close Section 4 with a few general remarks on the nature of the bias in estimated scores.

The paper is intended to be thought provoking, in that it stands in somewhat striking contrast to what is expected from now standard nonparametric approximation theory. As such, it is important at the outset to point out how our posture differs from that theory. In particular, under standard conditions on the rate at which the bandwidth h shrinks with sample size, the estimators $\hat{f}(x)$, $\hat{f}'(x)$ and $\hat{\ell}(x)$ are easily shown to be pointwise consistent estimators for $f(x)$, $f'(x)$ and $\ell(x)$. Here, the bias is formed from the expectations $E[\hat{f}(x)]$ and $E[\hat{f}'(x)]$ given the bandwidth value h. These terms can be thought of as the $(\sqrt{N})$ limits of $\hat{f}(x)$ and $\hat{f}'(x)$, under asymptotic

theory that takes the bandwidth h as fixed, not shrinking with sample size N. We use this interpretation explicitly for the score $\hat{\ell}(x)$, namely characterizing the ($\sqrt{N}$) limit plim $\hat{\ell}(x) = - E[\hat{f}'(x)]/E[\hat{f}(x)]$ for given bandwidth h. As such, we treat $\hat{f}(x)$ of (1.1) as a simple sample average, with h a constant that has been set, and treat $\hat{f}'(x)$ and $\hat{\ell}(x)$ analogously. Consequently, this paper takes the posture that fixed bandwidth asymptotics may give a more accurate distributional approximation than a theory that promises that h will be shrunk when more data is obtained. Both approximation theories would coincide in a large data set with tiny bandwidth values used, but they give different results for finite bandwidth values.

## 2. Downward Bias in Density Derivative Estimates

### 2.1 Basic Framework

As indicated above, we consider a situation where the data is an i.i.d. random sample of observations on a k-vector x, distributed with density f(x). The basic structure used for the density is summarized as

Assumption 2.1: The density f(x) has convex (possibly unbounded) support $S_f \subseteq \mathbb{R}^k$, and $f(x) = 0$ for $x \in \partial S_f$, the boundary of its support. f(x) is twice continuously differentiable on $\text{int}(S_f)$.

The kernel density $\mathcal{K}(.)$ of (2.1) is structured as follows:

Assumption 2.2: The kernel $\mathcal{K}(u)$ has support $S_{\mathcal{K}} \subseteq \mathbb{R}^k$, with $\mathcal{K}(u) > 0$ for $u \in \text{int}(S_{\mathcal{K}})$ and $\mathcal{K}(u) = 0$ for $u \in \partial S_{\mathcal{K}}$, the boundary of $S_{\mathcal{K}}$. The origin $0 \in S_{\mathcal{K}}$, and if $u \in S_{\mathcal{K}}$ then $-u \in S_{\mathcal{K}}$. $\mathcal{K}(u) = \mathcal{K}(-u)$ is symmetric (with $\int u\mathcal{K}(u)du = 0$) and continuously differentiable on $\text{int}(S_{\mathcal{K}})$.

Finally, we assume the following regularity condition, in lieu of primitive conditions that guarantee it.[4]

Assumption 2.3: The integral $\int \mathcal{K}(u)f(x-hu)du$ exists for $x \in S_f$ and is differentiable in x, with derivative $(\int \mathcal{K}(u)f(x-hu)du)' = \int \mathcal{K}(u)f'(x-hu)du$.

In this section, we focus on the derivative of the density estimator $\hat{f}(x)$ of (2.1) relative to the true density derivative $f'(x)$. The derivative $\hat{f}'(x)$ is written explicitly as

$$(2.1) \qquad \hat{f}'(x) = \frac{\partial \hat{f}(x)}{\partial x} = N^{-1}h^{-k-1} \sum_{i=1}^{N} \mathcal{K}'\left(\frac{x - x_i}{h}\right)$$

In Sections 3 and 4, we study the density score estimator $\hat{\ell}(x) = -\partial \ln \hat{f}/\partial x = -\hat{f}'(x)/\hat{f}(x)$, computed from (1.1) and (2.1).

## 2.2 Smoothing Bias and Jensen's Inequality

The expectations of $\hat{f}(x)$ and $\hat{f}'(x)$ are found by a standard calculation (c.f. Silverman(1986)): taking the expectation of (1.1) and changing variables gives

$$(2.2) \qquad E(\hat{f}(x)) = \int \mathcal{K}(u)\ f(x-hu)du$$

and for (2.1), including integration-by-parts,

$$(2.3) \qquad E(\hat{f}'(x)) = \int \mathcal{K}(u)\ f'(x-hu)du = \partial E(\hat{f}(x))/\partial x$$

where the latter equality uses Assumption 2.3.[5]

The intuition behind the downward bias argument is that because of Jensen's inequality, $E[\hat{f}(x)]$ will tend to be "flatter" that $f(x)$, which can cause measured slopes to be too small in absolute value. In particular,

consider the univariate case where $k = 1$, and suppose that the support of $\mathcal{K}$ is compact, say $S_{\mathcal{K}} = [-1,1]$. Since $E[\hat{f}(x)] = E_u[f(x-hu)]$ ($u$ distributed with density $\mathcal{K}$), we have $E[\hat{f}(x)] < f(x)$ when $f$ is strictly concave on $(x-h,x+h)$ and $E[\hat{f}(x)] > f(x)$ when $f$ is strictly convex on $(x-h,x+h)$. In Figure 1 we have drawn a density and its derivative, as well as expectations consistent with (2.2) and (2.3) (and Jensen's inequality). Obviously, $E[\hat{f}'(x)]$ is smaller in absolute value than $f'(x)$ everywhere except for the tails. Figure 2 gives analogous pictures for a multimodal density, where again, a downward derivative bias is evident.

## 2.3 Basic Aspects of Density Smoothing, and a Simple Result

While these pictures capture the essence of the problem, it will be useful for our discussion later to reinterpret the impact of smoothing via the convolution structure of (2.2) and (2.3), and then show a result that verifies the import of the Figures 1 and 2. As well known,[6] $\hat{f}(x)$ is the density of $Z = X + hu$, where $X$ is distributed as the empirical distribution of the data $\{x_i, i=1,\ldots,N\}$, and $u$ is distributed with density $\mathcal{K}(u)$, independently of $X$. Likewise, suppose that $z = x + hu$, where $x$ is distributed with density $f(x)$, independently of $u$, which is distributed with density $\mathcal{K}(u)$. Then it follows immediately that the density $\phi_h(z)$ of $z$ is given by the formula

$$(2.4) \qquad \phi_h(z) = \int \mathcal{K}(u)\, f(z-hu)du = E[\hat{f}(z)]$$

and (from assumption 2.3) that

$$(2.5) \qquad \phi_h'(z) = \int \mathcal{K}(u)\, f'(z-hu)du = E[\hat{f}'(z)] \quad .$$

Therefore, $\hat{f}(x)$ measures the convolution $\phi_h$ evaluated at $x$. Clearly if $h \to 0$, then $\phi_h(x) \to f(x)$, but our concern is studying $\phi_h$ for given $h$.

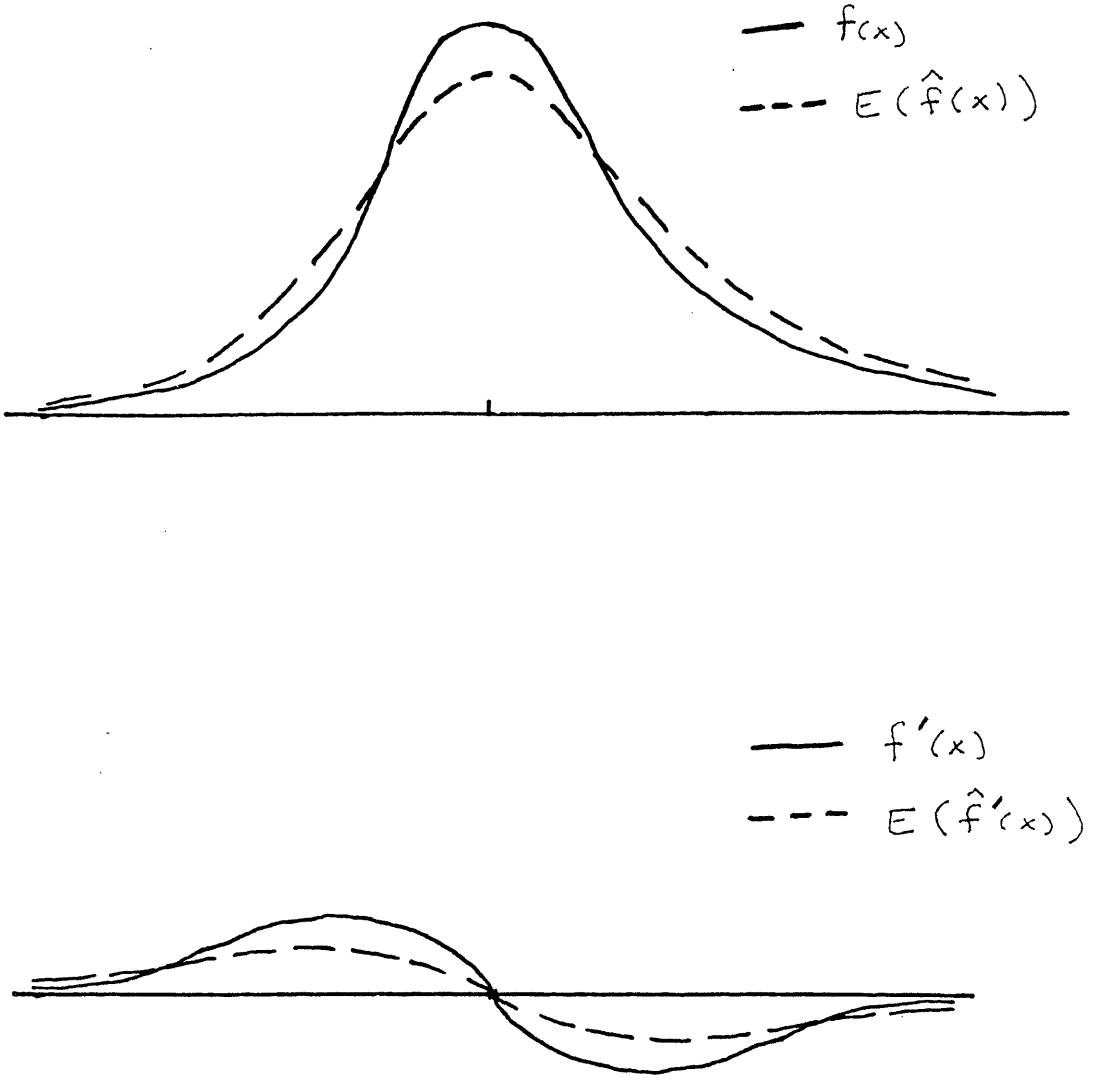
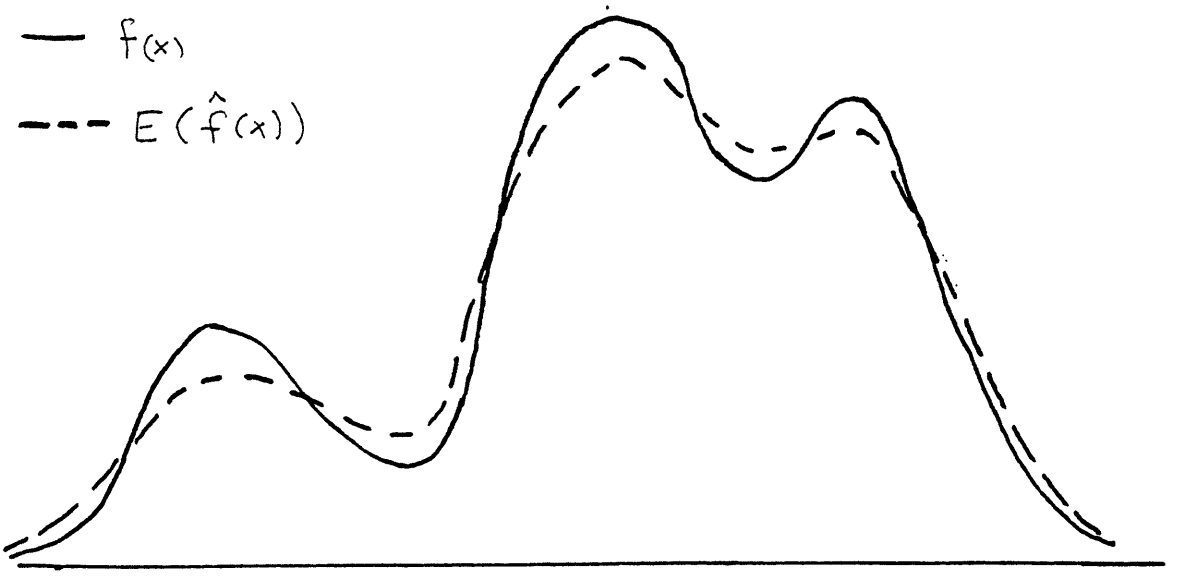



FIGURE 1:



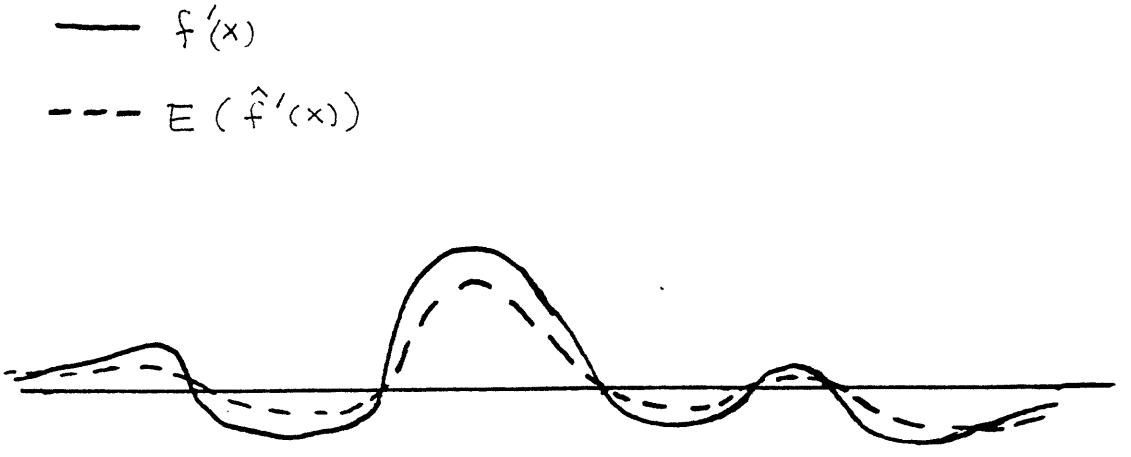




FIGURE 2.

This convolution structure permits several immediate comparisons between $\phi_h$ and f. First, their moment structures are easily compared. Each displays the same mean $\mu = \int xf(x)dx = \int x\phi_h(x)dx$. With $\int uu^T\mathcal{K}(u)du = \Sigma_u$, we have the covariance matrices

$$\begin{aligned} \int(x-\mu)(x-\mu)^T f(x)dx &= \Sigma_x \\ \int(x-\mu)(x-\mu)^T \phi_h(x)dx &= \Sigma_x + h^2\Sigma_u \end{aligned} \tag{2.6}$$

so that $\phi_h$ displays a larger (matrix sense) covariance matrix. If we consider marginal densities implied by f and $\phi_h$, then it is easy to verify that $\phi_h$ displays the same third moment as f, and larger even moments of all orders; for the $j^{th}$ component $x_j$, we have

$$\begin{aligned} &\int(x_j-\mu_j)^3\phi_h(x)dx = \int(x_j-\mu_j)^3 f(x)dx \\ &\int(x_j-\mu_j)^r\phi_h(x)dx > \int(x_j-\mu_j)^r f(x)dx; \quad r \geq 4, \ r \text{ even.} \end{aligned} \tag{2.7}$$

provided the moments of f and $\mathcal{K}$ exist. The equality of third moments implies that the marginals of $\phi_h$ display less skewness than those of f.

Some properties are available from studies of the concavity properties of convolutions. If $\mathcal{K}$ is log-concave and f is quasi-concave,[7] then $\phi_h$ is quasi-concave (Ibragimov(1956)). If f is further assumed to be log-concave, then $\phi_h$ is log-concave (Prekopa(1973)).[8] However, if f and $\mathcal{K}$ are only assumed to be unimodal, then it is not true in general that $\phi_h$ is unimodal (Gnedenko and Kolmogorov(1954)), unless f is symmetric about its mean. For f univariate, we can establish an analogy between the extrema structure of $\phi_h$ and f by limiting the bandwidth value h. This relation then implies a simple result on the relative sizes of the derivatives of $\phi_h$ and f, that essentially verifies the intuition of the pictures of the last section.[9]

We make the following additional assumptions.

Assumption 2.4: The univariate density $f(x)$ has a finite number of local maxima $m_1, \ldots, m_d$ and local minima $b_1, \ldots, b_{d-1}$, with $m_1 < b_1 < m_2 < \ldots < b_{d-1} < m_d$.

Assumption 2.5: The kernel $\mathcal{K}(u)$ has support $S_{\mathcal{K}} = [-1.1]$.

The similarities between f and $\phi_h$ are given as:

Lemma 3.1: Under Assumptions 2.1-5,

1. The support $S_\phi$ of $\phi_h(x)$ is convex and $S_f \subset S_\phi$. $\phi_h$ is a twice continuously differentiable density on $\text{int}(S_\phi)$, and $\phi_h(x) = 0$ for x on the boundary, $x \in \partial S_\phi$.
2. If $f(x)$ is symmetric about a point $m_f$, then $\phi_h$ is symmetric about $m_f$.

Suppose $h \in (0, h_0]$ (where $h_0$ is specified in the proof).

3. If $f(x)$ is unimodal with mode $m_f$, then $\phi_h$ is unimodal with mode $m_\phi$, and $f(m_f) > \phi_h(m_\phi)$.
4. If $f(x)$ has modes $m_1, \ldots, m_d$ and local minima $b_1, \ldots, b_{d-1}$, then $\phi_h$ has associated modes $\tilde{m}_1, \ldots, \tilde{m}_d$ and local minima $\tilde{b}_1, \ldots, \tilde{b}_{d-1}$, where $\phi_h(\tilde{m}_j) < f(m_j)$ for $j = 1, \ldots, d$, and $\phi_h(\tilde{b}_j) > f(b_j)$ for $j = 1, \ldots, d-1$.

This Lemma immediately implies the following theorem on the derivatives of f and $\phi_h$:

Theorem 2.1: Under Assumptions 2.1-5, if $h \in (0,h_0]$

$$(2.8) \qquad \int_{-\infty}^{\infty} |f'(x)|dx > \int_{-\infty}^{\infty} |\phi_h'(x)|dx$$

Theorem 2.1 states in broad terms how $\phi_h'$ will be smaller than $f'$, or how smoothing causes an underestimation of density derivatives. The limitation on the bandwidth size is used in the proof of Lemma 2.1, however it seems natural that that (2.8) will obtain for larger bandwidths. In particular, smoothing away modes of f (with $\phi_h$ having a smaller number of local extrema than f) should exacerbate the typical underestimation of derivatives. If true, it would be natural to assert a similar relationship between derivatives in the multivariate case, however a proof would seemingly involve tracking the modal structure of f as its components were individually varied.

While we have given the basic logic behind the downward derivative bias implied by smoothing, the pictures and results above are not informative about how large the bias is, or whether a systematic bias of this kind would make a difference to any empirical problem. For some answers to these questions, for the remainder of the paper we consider the impact of derivative bias on measuring the score $\ell(x) = -f'(x)/f(x)$.

## 3. Background Motivation for Density Score Estimation

Our discussion of score estimation includes some general remarks on the impact of the derivative bias, as well as specific calculations for examples based on normal distributions. First, we provide some motivation for studying the density score by recalling two estimation problems that involve "plugging-in" score estimates. For this section only, we augment our notation

to include a scalar response variable y, so that the data is a random sample $(y_i,x_i)$, i=1,...,N.

3.1 Average Derivative Estimation

The (unweighted) average derivative of y on x is defined as the mean of the derivative of $E(y|x)$, namely

$$\delta = E[\partial E(y|x)/\partial x] \tag{3.1}$$

The estimation of $\delta$ is of interest to various semiparametric problems; for instance when the regression is structured as $E(y|x) = G(x^T\beta)$, then $\delta$ is proportional to $\beta$, so that a nonparametric estimator of $\delta$ measures $\beta$ up to scale.[10]

The connection of the average derivative to the density score $\ell(x)$ is seen by the following formulations of (3.1)

$$\begin{aligned} \delta &= \mathrm{Cov}[\ell(x),y] \\ &= (\mathrm{Cov}[\ell(x),x])^{-1}\mathrm{Cov}[\ell(x),y] \end{aligned} \tag{3.2}$$

These representations follows from applying integration-by-parts and the law of iterated expectation to $E[\partial E(y|x)/\partial x]$, and the latter representation follows from noting that the leading matrix is the inverse of the identity $I = E(\partial x/\partial x) = \mathrm{Cov}(\ell(x),x)$ (c.f. Stoker(1986) for details).

The sample analogs of these formulae, where $\hat{\ell}(x)$ is plugged in for $\ell(x)$, give the two average derivative estimators of interest here. Following Härdle and Stoker (1989), the first estimator is

$$\hat{\delta} = N^{-1} \sum_{i=1}^{N} \hat{\ell}(x_i)\,[y_i-\bar{y}]\,\hat{1}_i \tag{3.3}$$

or the analog of the score-covariance representation, where $\hat{1}_i = 1[\hat{f}(x_i) \geq b]$ is a trimming indicator that drops observations with small estimated density

(used in the technical technical analysis of $\hat{\delta}$). Following Stoker(1991a), the second estimator is

$$(3.4) \qquad \hat{d} = \left[ N^{-1} \sum_{i=1}^{N} \hat{\ell}(x_i)(x_i - \bar{x})^T \hat{1}_i \right]^{-1} \left[ N^{-1} \sum_{i=1}^{N} \hat{\ell}(x_i)(y_i - \bar{y}) \hat{1}_i \right]$$

which is a linear "slope" estimator because of its interpretation via instrumental variables estimation; namely $\hat{d}$ is the slope coefficient vector from estimating a linear equation

$$(3.5) \qquad y_i = \hat{c} + x_i' \hat{d} + \hat{u}_i \qquad i = 1,\ldots,N$$

using $(1, \hat{\ell}(x_i)\hat{1}_i)$ as the instrumental variable. This interpretation plays a role in the motivation below.

These estimators are useful to our discussion for two reasons. First, the estimators are based on fairly simple averages of the nonparametric components $\hat{\ell}(x_i)$, so that differences in the estimators are easy to interpret. Second, the types of differences (as illustrated below), leads one to propose corrections and/or diagnostic statistics for the presence of derivative bias.

Under an asymptotic theory involving shrinking bandwidths, these estimators can both be shown to be $\sqrt{N}$ consistent for $\delta$, and moreover they can be shown to be first-order equivalent. In particular, Härdle and Stoker(1989) show that if $\mathcal{K}$ is a kernel of order $p \geq k+2$ and some smoothness conditions obtain, then as (i) $N\to\infty$, $h\to 0$, $b\to 0$, $h^{-1}b\to 0$, (ii) for some $\epsilon > 0$, $b^4 N^{1-\epsilon} h^{2k+2} \to \infty$, and (iii) $Nh^{2p-2} \to 0$,

$$(3.6) \qquad \sqrt{N}\,(\hat{\delta} - \delta) = = N^{-1/2} \sum_{i=1}^{N} r(y_i, x_i) + o_p(1)$$

where $r(y,x) \equiv g'(x) - \delta + [y - g(x)]\,\ell(x)$, so that under standard central

limit theory, $\sqrt{N}(\hat{\delta} - \delta)$ has a limiting normal distribution with mean 0 and variance $\Sigma = E(rr^T)$. Moreover, as shown in Stoker(1991a), $\sqrt{N}(\hat{\delta} - \hat{d}) = o_p(1)$, so that $\hat{\delta}$ and $\hat{d}$ are equivalent to first order. These approximation conditions are consistent with asymptotic undersmoothing; namely the bandwidth shrinks to zero more rapidly than would be implied by optimal pointwise estimation of $\ell(x)$ by $\hat{\ell}(x)$.

While these estimators are asymptotically equivalent under the above conditions, this is no guarantee that they will be similarly behaved in finite samples.[11] In particular, substantive differences in the scale of the two estimators appear routinely in simulations (which in fact originally motivated the present study).[12] Consider the following simulation results, which are fairly typical. The first example takes the basic model to be linear:

$$(3.7) \qquad y_i = 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \epsilon_i \; ; \quad i = 1,\ldots,N$$

where the $k = 4$ predictors $x_{ji}$, and the disturbance $\epsilon_i$ are (independent) $N(0,1)$ variables. The sample size is $N = 100$, the kernel is the spherical multivariate normal density $K(u) = \Pi\, k(u_j)$ with $k(u_j) = (1/\sqrt{2\pi})\exp(-u_j^2/2)$, the bandwidth is $h = 1$ and the trimming bound b is set to drop 1% of the observations.[13] The average derivative is the vector of coefficients $\delta = (1,1,1,1)'$. Table 1 contains the means and standard errors of each of the average derivative components over 20 Monte Carlo simulations.

While asymptotically equivalent under shrinking bandwidth theory, the covariance estimator $\hat{\delta}$ is roughly 40% of the value of the slope estimator $\hat{d}$.[14] Moreover, this simulation design ought to favor good estimator performance. The predictors are symmetrically distributed, independent and have a symmetric impact on y. The $R^2$ of the true equation is .80, which is quite large relative to survey applications, at least for economic data.

TABLE 1: SIMULATION RESULTS - LINEAR MODEL

True Value: $\delta = (1,1,1,1)$

| | $\hat{\delta}_1$ | $\hat{\delta}_2$ | $\hat{\delta}_3$ | $\hat{\delta}_4$ |
|---|---|---|---|---|
| "Covariance" | .389 | .390 | .404 | .385 |
| (3.3) | (.047) | (.078) | (.039) | (.062) |
| "Slope" | .991 | 1.01 | 1.03 | .984 |
| (3.4) | (.101) | (.154) | (.102) | (.101) |
| OLS | 1.01 | 1.01 | 1.02 | .976 |
| | (.078) | (.128) | (.111) | (.107) |

One facet of the results which might be guaranteed is the good performance of the slope estimators, because they are conditionally unbiased for the true coefficients (provided $x - \bar{x}$ and $\hat{\ell}(x)\hat{1}$ are not orthogonal). With this in mind we present simulations of a binary response model in Table 2; namely where the dependent variable is altered to

$$(3.8) \qquad y_i = 1[\ 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \epsilon_i > 0\ ]\ ; \quad i = 1,\ldots,N$$

Now the true average derivative vector is $\delta = .161\ (1,1,1,1)$. The kernel, bandwidth and trimming parameters are the same as for Table 1.

Here the same problems arise, with a substantial underestimation of $\delta$ by the covariance estimator $\hat{\delta}$, and much less bias exhibited by the slope estimator $\hat{d}$. We have included the OLS estimator (of regressing the discrete y on x) because its performance is dictated by the design: namely with normally distributed regressors, the OLS coefficients are (unconditionally) consistent for the average derivative $\delta$ (c.f. Brillinger(1983) and Stoker(1986)).

The theoretical results cited above assert that for very large data sets, with tiny bandwidth (and trimming bound) values, the differences seen in Tables 1 and 2 will disappear. As such, 100 observations may be just too small for any adherence to this approximate distributional theory. Moreover, one could approach these differences by deriving an optimal bandwidth value.[15]

However, we consider a different explanation, namely that the estimated values $\hat{\ell}(x)$ are uniformly too small, in the sense of downward bias discussed before. Specifically, suppose that for a given (fixed) bandwidth value h, we denote $\text{plim}\ \hat{\ell}(x) = \lambda_h(x)$. Using an argument analogous to that presented in Härdle and Stoker(1989), as $N \to \infty$ and $b \to 0$ (but h fixed), we can show that

$$(3.9) \qquad \text{plim}\ \hat{\delta} = \text{Cov}(\lambda_h, y)$$

$$\text{plim}\ \hat{d} = \{\text{Cov}[\lambda_h(x), x]\}^{-1}\text{Cov}[\lambda_h(x), y]$$

TABLE 2: SIMULATION RESULTS - BINARY RESPONSE MODEL

True Value: $\delta = (.161, .161, .161, .161)$

| | $\hat{\delta}_1$ | $\hat{\delta}_2$ | $\hat{\delta}_3$ | $\hat{\delta}_4$ |
|---|---|---|---|---|
| "Covariance" | .063 | .068 | .070 | .063 |
| (3.3) | (.021) | (.022) | (.015) | (.013) |
| "Slope" | .177 | .179 | .171 | .164 |
| (3.4) | (.033) | (.040) | (.036) | (.032) |
| OLS | .171 | .171 | .168 | .160 |
| | (.035) | (.033) | (.035) | (.028) |

Now, suppose that there was a proportional bias; say for each x, $\lambda_h(x) = 1/2\ \ell(x)$. Then plim $\hat{\delta} = 1/2\ \delta$ but plim $\hat{d} = \delta$. As such, the leading term of $\hat{d}$ would act to correct for systematic underestimation of $\ell(x)$ by $\hat{\ell}(x)$.

If fact, the relation $\lambda_h(x) = 1/2\ \ell(x)$ is exactly the one given in Section 4 for the normal design of Tables 1 and 2. As such, this explanation captures a large part of the differences noted above. It is also useful to note that the leading term of $\hat{d}$ corrects for any bias of matrix proportional form; if $\lambda_h(x) = A\ \ell(x)$, then plim $\hat{\delta} = A\ \delta$ and plim $\hat{d} = \delta$. This suggests an obvious diagnostic statistic (or correction) for the bias problem, which we spell out at the end of Section 4.

### 3.2 Adaptive Estimation of Location Models

While the average derivative example is closely aligned with the issues at hand, estimation of the density score plays a role in other familiar problems in statistics. Here we indicate their role in adaptive estimation of location models, as developed by Stein(1956), Stone(1975), Bickel(1982) and Manski(1984), among others.

The model of interest here is

$$y = g(x,\theta) + \epsilon \tag{3.10}$$

where $\theta$ is a finite vector of parameters of interest, and $\epsilon$ is distributed independently of x, with density $f_\epsilon$.[16] As such, the density of y conditional on x is

$$F(y|x,\theta) = f_\epsilon[y-g(x,\theta)] \tag{3.11}$$

with

$$\partial \ln F(y|x,\theta)/\partial\theta = \ell_\epsilon\ \partial g/\partial\theta \tag{3.12}$$

where $\ell_\epsilon = -f_\epsilon'/f_\epsilon$ is the translation score. The information matrix is then

$$\text{(3.13)} \qquad \begin{aligned} \mathcal{I}_\theta &= E[\partial \ln F(y|x,\theta)/\partial\theta \; \partial \ln F(y|x,\theta)/\partial\theta^T] \\ &\quad - E[\ell_\epsilon^2 \; \partial g/\partial\theta \; \partial g/\partial\theta^T] \end{aligned}$$

If the density $f_\epsilon$ is known and specified, then under standard conditions the maximum likelihood estimator $\hat{\theta}$ of $\theta$ is $\sqrt{N}$ consistent, $\sqrt{N}(\hat{\theta} - \theta) \to \mathcal{N}(0, \mathcal{I}_\theta^{-1})$, where $\mathcal{I}_\theta^{-1}$ is the Cramer-Rao lower bound.

The question of adaptive estimation is whether an estimator of $\theta$ can be constructed with this asymptotic distribution, when the density $f_\epsilon$ is unknown. From Bickel(1982) and Manski(1984), if $f_\epsilon$ is symmetric around 0, the answer here is positive. A solution is as follows: begin with any $\sqrt{N}$ consistent estimator $\tilde{\theta}$ of $\theta$, compute the residuals, $\hat{\epsilon}_i = y_i - g(x_i,\tilde{\theta})$, estimate $f_\epsilon$ and $\ell_\epsilon$ using the kernel density estimator $\hat{f}_\epsilon$, $\hat{\ell}_\epsilon = -\hat{f}_\epsilon'/\hat{f}_\epsilon$, and then update $\tilde{\theta}$ by one Newton-Raphson step:

$$\text{(3.14)} \qquad \hat{\theta} = \tilde{\theta} + \left[\sum_{i=1}^{N} \hat{\ell}_{\epsilon i}^2 \, (\partial g(x_i,\tilde{\theta})/\partial\theta) \, (\partial g(x_i,\tilde{\theta})/\partial\theta)^T \hat{1}_i\right]^{-1} \left[\sum_{i=1}^{N} \hat{\ell}_{\epsilon i} \, (\partial g(x_i,\tilde{\theta})/\partial\theta) \, (y_i - g(x_i,\tilde{\theta}) \hat{1}_i\right]$$

The natural estimator of the asymptotic variance of $\hat{\theta}$ is likewise[17]

$$\text{(3.15)} \qquad \hat{\mathcal{I}}_\theta^{-1} = \left[N^{-1} \sum_{i=1}^{N} \hat{\ell}_{\epsilon i}^2 (\partial g(x_i,\tilde{\theta})/\partial\theta)(\partial g(x_i,\tilde{\theta})/\partial\theta)^T \hat{1}_i\right]^{-1} .$$

We raise this example to just illustrate the natural role of the score $\ell_\epsilon$ in problems of location, as well as point out the implications of a uniform downward bias in the kernel estimator $\hat{\ell}_\epsilon$ in a practical

application of these ideas. Suppose that $\hat{\ell}_\epsilon$ is proportionately too small, then (since $f_\epsilon$ is symmetric) the direction of the correction (3.14) is likely accurate, but the step taken will be too large. If the full maximization of an estimated log likelihood (based on (3.11) using $\hat{f}_\epsilon$) is carried out, it is possible that the resulting adaptive estimate $\hat{\theta}$ will not be affected by the downward bias problem (again since $f_\epsilon$ and the kernel $K$ are symmetric). However, in either case, the variance estimator (3.15) will be too large, with the standard error of a component of $\hat{\theta}$ overestimated by the reciprocal of the proportion of score bias.

We now turn to an analysis of the bias in estimated density scores.

## 4. Downward Bias in Density Score Estimates

### 4.1 Basic Structure

Because of the nonlinear structure of $\hat{\ell}(x) = -\hat{f}'(x)/\hat{f}(x)$, we use a large sample approximation, but one holding the bandwidth value h fixed. This differs from the currently popular approximation theory, but does permit one to isolate and characterize the impact of local smoothing. Since the limits of $\hat{f}'(x)$ and $\hat{f}(x)$ are their expectations, we have by Slutsky's theorem that[18]

$$\text{(4.1)} \qquad \operatorname{plim} \hat{\ell}(x) = -\frac{E[\hat{f}'(x)]}{E[\hat{f}(x)]} = -\frac{\phi_h'(x)}{\phi_h(x)} = -\frac{\partial \ln \phi_h(x)}{\partial x}$$

$$= \lambda_h(x).$$

Therefore, $\hat{\ell}(x)$ estimates the score of the density $\phi_h$ evaluated at x, where $\phi_h$ is the (convolution) density of $x + hu$.

It appears difficult to establish a general result asserting that the norm of $\lambda_h(x)$ is typically smaller than that of the true score $\ell(x)$ for an arbitrary base density f(x). The end of this section contains a few

remarks on general comparisons that are available. For concreteness, we now discuss examples and calculations based on normal distributions.

## 4.2 Normal Distributions and Proportional Downward Bias

When the convolution $\phi_h$ can be solved for, we can solve for $\lambda_h(x)$ and the bias explicitly. The simplest cases concern when x is normally distributed and a standard normal kernel is used to compute $\hat{f}(x)$. Suppose first that x is univariate normal, with mean $\mu$ and variance $\sigma^2$. In this case the true score $\ell(x)$ is

$$\ell(x) = \frac{1}{\sigma^2}(x - \mu) \tag{4.2}$$

The kernel estimator $\hat{f}(x)$ estimates the density $\phi_h(x)$ of (2.4), which is a normal density with mean $\mu$ and variance $\sigma^2 + h^2$. The score $\hat{\ell}(x)$ estimates $\lambda_h = -\partial \ln \phi_h / \partial x$, or

$$\lambda_h(x) = \frac{1}{\sigma^2 + h^2}(x - \mu) \tag{4.2}$$

Therefore we conclude that

$$\lambda_h(x) = \frac{\sigma^2}{\sigma^2 + h^2}\,\ell(x) \equiv a_h\,\ell(x) \quad . \tag{4.3}$$

Here the bias is uniformly proportional for all x, and downward by the factor $a_h = \sigma^2/(\sigma^2+h^2)$.

Matrix proportionality arises in the multivariate normal case. Suppose that x is distributed multivariate normally with mean $\mu$ and covariance matrix $\Sigma$, and that $K(u)$ is the spherical normal density, with mean 0 and variance I. The true score is

(4.4) $$\ell(x) = \Sigma^{-1}(x - \mu)$$

The density $\phi_h$ is normal with mean $\mu$ and covariance matrix $\Sigma + h^2 I$, so that the score $\hat{\ell}(x)$ estimates

(4.5) $$\lambda_h(x) = (\Sigma + h^2 I)^{-1} (x - \mu) = A_h \ell(x)$$

where $A_h = (\Sigma + h^2 I)^{-1}\Sigma$ is the matrix factor of proportionality.[19] If x is further assumed to be spherical normally distributed; namely with $\Sigma = \sigma^2 I$, then a symmetric componentwise underestimation occurs; namely we have $A_h = [\sigma^2/(\sigma^2+h^2)]\ I$ above, so that

(4.6) $$\lambda_h(x) = \frac{\sigma^2}{\sigma^2 + h^2} \ell(x) = a_h \ell(x)$$

This formula illustrates the potential severity of the bias problem; if h is set to the standard deviation of the components of x, then $\hat{\ell}(x)$ will estimate half the value of $\ell(x)$ for any x.[20]

To get some feeling for the size of the bias relative to dimension and sample size, we compute the bias for optimal bandwidths for estimating $\hat{f}(x)$ in the spherical normal case. Table 3 gives such values, using bandwidths computed from the formula in Silverman(1986, p. 87), based on first order approximation of pointwise bias and variance. Listed here is the standard deviation of each component of x implied by $\phi_h$, namely $(1 + h^2)^{1/2}$, and the derivative bias $1-a_h$, expressed in percentage terms.

The bias numbers are on the whole quite substantial. For instance, consider the case of estimating the univariate normal density with 100 observations. The standard deviation implied for $\phi_h$ is only marginally larger than that for f, namely 1.085 to 1.0, but there is still a 15 % downward bias in the score vector. Even with 5000 observations, the bias is not negligible.

TABLE 3: DERIVATIVE BIAS WITH APPROXIMATELY OPTIMAL BANDWIDTHS

Specification:

Normal Design: $x \sim N(0,I)$, multivariate normal k vector

Normal Kernel: $\mathcal{K}$ is the $N(0,I)$ density;

Optimal Bandwidth: $h = A(\mathcal{K})\, N^{-1/(k+4)}$, $A(\mathcal{K}) = [4/2k+1)]^{1/(k+4)}$

(Silverman (1986, p.87))

Phi Stan. Dev.: Component standard deviation from $\phi_h$; namely $(1 + h^2)^{1/2}$

Derivative Bias: $1 - a_h = h^2/(1+h^2)$

| A(K) | 1.059 | 0.963 | 0.923 | 0.904 | 0.894 | 0.888 |
|---|---|---|---|---|---|---|
| Dimension k | 1 | 2 | 3 | 4 | 5 | 10 |
| N = 25 | | | | | | |
| Bandwidth h | 0.556 | 0.563 | 0.583 | 0.604 | 0.625 | 0.706 |
| Phi Stan. Dev. | 1.144 | 1.148 | 1.157 | 1.168 | 1.179 | 1.224 |
| Derivative Bias | 23.64% | 24.10% | 25.36% | 26.75% | 28.09% | 33.25% |
| N = 50 | | | | | | |
| Bandwidth h | 0.484 | 0.502 | 0.528 | 0.554 | 0.579 | 0.672 |
| Phi Stan. Dev. | 1.111 | 1.119 | 1.131 | 1.143 | 1.155 | 1.205 |
| Derivative Bias | 19.00% | 20.13% | 21.80% | 23.49% | 25.08% | 31.09% |
| N = 100 | | | | | | |
| Bandwidth h | 0.422 | 0.447 | 0.478 | 0.508 | 0.536 | 0.639 |
| Phi Stan. Dev. | 1.085 | 1.095 | 1.108 | 1.122 | 1.134 | 1.187 |
| Derivative Bias | 15.10% | 16.67% | 18.61% | 20.52% | 22.30% | 29.01% |
| N = 500 | | | | | | |
| Bandwidth h | 0.306 | 0.342 | 0.380 | 0.416 | 0.448 | 0.570 |
| Phi Stan. Dev. | 1.046 | 1.057 | 1.070 | 1.083 | 1.096 | 1.151 |
| Derivative Bias | 8.54% | 10.47% | 12.61% | 14.72% | 16.72% | 24.51% |
| N = 1000 | | | | | | |
| Bandwidth h | 0.266 | 0.305 | 0.344 | 0.381 | 0.415 | 0.542 |
| Phi Stan. Dev. | 1.035 | 1.045 | 1.058 | 1.070 | 1.083 | 1.138 |
| Derivative Bias | 6.61% | 8.49% | 10.59% | 12.68% | 14.68% | 22.73% |
| N = 5000 | | | | | | |
| Bandwidth h | 0.193 | 0.233 | 0.273 | 0.312 | 0.347 | 0.483 |
| Phi Stan. Dev. | 1.018 | 1.027 | 1.037 | 1.047 | 1.058 | 1.111 |
| Derivative Bias | 3.59% | 5.15% | 6.96% | 8.85% | 10.74% | 18.94% |
| N = 10000 | | | | | | |
| Bandwidth h | 0.168 | 0.208 | 0.248 | 0.286 | 0.321 | 0.460 |
| Phi Stan. Dev. | 1.014 | 1.021 | 1.030 | 1.040 | 1.050 | 1.101 |
| Derivative Bias | 2.74% | 4.13% | 5.78% | 7.55% | 9.35% | 17.47% |
| N = 100000 | | | | | | |
| Bandwidth h | 0.106 | 0.141 | 0.178 | 0.214 | 0.249 | 0.390 |
| Phi Stan. Dev. | 1.006 | 1.010 | 1.016 | 1.023 | 1.030 | 1.073 |
| Derivative Bias | 1.11% | 1.96% | 3.08% | 4.39% | 5.82% | 13.22% |

For higher dimensional problems, the bias is substantial for small sample sizes, and vanishes much more slowly as the sample size is increased. For instance, the derivative bias in a 10 dimensional problem drops from 19 % with 5,000 data points to 13 % with 100,000 data points. Since the bias depends only on the value of the bandwidth, this is just a reflection of the slowness with which the optimal bandwidth shrinks with sample size, or the "curse of dimensionality."[21]

## 4.3 Normal Mixtures and Approximate Proportional Downward Bias

This matrix proportionality certainly does not exist generally, although it may provide a useful approximation. We now consider the cases where f(x) is a mixture of normals with equal covariance structures. Beginning with the univariate case, let $f_1(x)$ denote the normal density with mean $\mu_1$ and variance $\sigma^2$, and $f_2(x)$ denote the normal density with mean $\mu_2 \neq \mu_1$ and variance $\sigma^2$. With $0 < \rho < 1$, suppose that x is distributed with density $f(x) = \rho f_1(x) + (1-\rho)f_2(x)$, so that the mean and variance of x are $\mu = \rho\mu_1 + (1-\rho)\mu_2$ and $\sigma^2 + \rho(1-\rho)(\mu_2-\mu_1)^2$, respectively. Suppose, as above, that $K(u)$ is a normal density, with mean 0 and variance 1. In this example, it is easy to verify that

$$(4.7) \qquad \ell(x) = \frac{1}{\sigma^2}\,\omega(x)(x-\mu_1) + \frac{1}{\sigma^2}\,[1-\omega(x)](x-\mu_2)$$

where $\omega(x) = \rho f_1(x)/[\rho f_1(x) + (1-\rho)f_2(x)]$. If $\phi_1$ and $\phi_2$ represent the normal densities with means $\mu_1 \neq \mu_2$ and common variance $\sigma^2 + h^2$, then the density $\phi_h$ is easily seen to be

$$(4.8) \qquad \phi_h(x) = \rho\,\phi_1(x) + (1-\rho)\,\phi_2(x)$$

The score $\lambda_h(x)$ is

$$(4.9) \qquad \lambda_h(x) = \frac{1}{\sigma^2 + h^2} \tilde{\omega}(x)(x-\mu_1) + \frac{1}{\sigma^2 + h^2} [1-\tilde{\omega}(x)](x-\mu_2)$$

where $\tilde{\omega}(x) = \rho\phi_1(x)/[\rho\phi_1(x) + (1-\rho)\phi_2(x)]$. Therefore, the impact of smoothing is to induce downweighting ($\sigma^{-2}$ to $(\sigma^2 + h^2)^{-1}$), and "flatten" the relative weighting from $\omega(x)$ to $\tilde{\omega}(x)$. In particular,

$$(4.10) \qquad \lambda_h(x) - a_h \ell(x) = [\tilde{\omega}(x)-\omega(x)]\, a_h\, (\mu_2-\mu_1)/\sigma^2$$

where $a_h = \sigma^2/(\sigma^2 + h^2)$ as before. Provided the variation in weighting $\tilde{\omega}(x)-\omega(x)$ is minor, or the mean spread $\mu_2-\mu_1$ is not large relative to the (within) variance $\sigma^2$, $\lambda_h(x)$ will be approximately proportional to $\ell(x)$.

We can develop this formula further by studying the weighting. In particular, denote

$$(4.11a) \qquad z(x) = [x - (\mu_1+\mu_2)/2]\, (\mu_2-\mu_1)/\sigma^2$$

$$(4.11b) \qquad c = \ln[(1-\rho)/\rho]$$

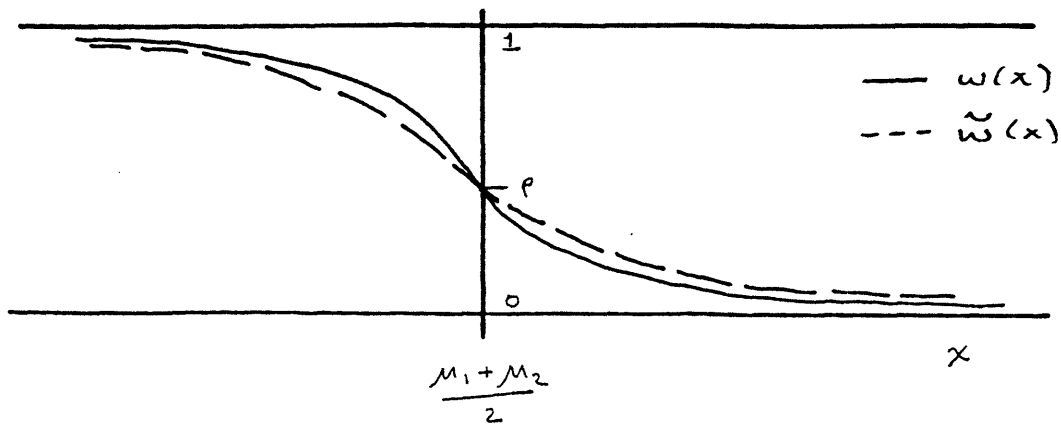
and

$$(4.12) \qquad w(x,a) = \frac{1}{1 + \exp[c + az(x)]}$$

so we have that

$$(4.13) \qquad \tilde{w}(x) = w(x,a_h), \qquad w(x) = w(x,1)$$

The function $w(x,a)$ is (one minus) a logit c.d.f, is decreasing in x (and z), with asymptotes 1 and 0, and decreasing in a for $x < (\mu_1+\mu_2)/2$ (or $w(x,a) > \rho$) and increasing in a for $x < (\mu_1+\mu_2)/2$ (or $w(x,a) > \rho$). Figure 3 plots typical versions of $w(x)$, $\tilde{w}(x)$ and $\tilde{w}(x) - w(x)$. Obviously, the difference $|\tilde{w}(x) - w(x)|$ is less than $1-\rho$ for $x < (\mu_1+\mu_2)/2$ and less that $\rho$ for $x >$



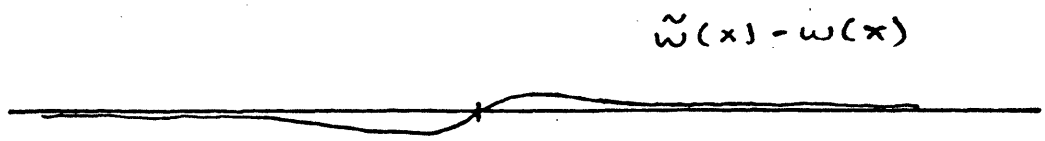




FIGURE 3.

$(\mu_1+\mu_2)/2$, and vanishes in both tails.

The difference (4.10) is now easy to characterize. By the mean value theorem, we have that

$$(4.14) \qquad \tilde{w}(x) - w(x) = z(x)\, w[x,a(x)]\{1 - w[x,a(x)]\}\,(1 - a_h)$$

where $a_h \le a(x) \le 1$. From the monotonicity properties of $w(x,a)$, we have

$$(4.15) \qquad |\tilde{w}(x) - w(x)| \le |z(x)|\,[\mathrm{Max}\{w(x),\, \tilde{w}(x)\} - w(x)\tilde{w}(x)]\,(1 - a_h)$$

Consequently, the difference (4.10) is

$$(4.16) \qquad |\lambda_h(x) - a_h \ell(x)| \le |x - (\mu_1+\mu_2)/2|\,[\mathrm{Max}\{w(x),\, \tilde{w}(x)\} - w(x)\tilde{w}(x)]$$

$$a_h(1 - a_h)\,[(\mu_2-\mu_1)/\sigma^2]^2$$

where we have inserted the formula for $z(x)$. While this bound could be computed for various specifications, it still points out how the difference clearly vanishes in the tails, and that no substantive departures are likely when the "bumps" in the distribution ($\mu_1$ and $\mu_2$) are not far apart relative to the within variance ($\sigma^2$). In other words, $\lambda_h(x) \cong a_h\, \ell(x)$ can be a good approximation.

This example is of interest because of the fairly wide range of distribution shapes that a normal mixture can represent. When the mixture variances are different, a further term is added to the difference (4.16), which does not appear to be as easy to characterize. One response to this is to consider mixtures of more than two normals with equal variances, for instance where $f = \rho_1 f_1 + \rho_2 f_2 + (1-\rho_1-\rho_2) f_3$, where $f_1$, $f_2$ are as above, and $f_3$ is a normal density with mean $\mu_3$ and variance $\sigma^2$. In this case

$$(4.17)\qquad \lambda_h(x) - a_h\ell(x) = a_h\Big([\tilde{\omega}_1(x)-\omega_1(x)](\mu_3-\mu_1)/\sigma^2 + [\tilde{\omega}_2(x)-\omega_2(x)](\mu_3-\mu_2)/\sigma^2\Big)$$

where $w_1(x) = \rho f_1(x)/f(x)$, $w_2(x) = \rho_2 f_2(x)/f(x)$, and $\tilde{w}_1(x)$, $\tilde{w}_2(x)$ are similarly defined. These weights have bivariate logit c.d.f. structure as above, and a similar analysis is possible, again showing vanishing difference in the tails, and substantive difference likely only when $\mu_1$, $\mu_2$ and $\mu_3$ are far apart relative to $\sigma^2$.

The multivariate case is quite similar. Now let $f_1(x)$ denote the multivariate normal density with mean $\mu_1$ and covariance matrix $\Sigma$, and $f_2(x)$ denote the normal density with mean $\mu_2 \neq \mu_1$ and covariance matrix $\Sigma$, and for $0 < \rho < 1$, let $f(x) = \rho f_1(x) + (1-\rho)f_2(x)$. Suppose that $\mathcal{K}(u)$ is a spherical normal density, with mean 0 and variance I. For this case we have that

$$(4.18)\qquad \lambda_h(x) - A_h\ell(x) = [\tilde{\omega}(x)-\omega(x)]A_h\Sigma^{-1}(\mu_2-\mu_1)$$

where $A_h = (\Sigma + h^2 I)^{-1}\Sigma$ as before. Further analysis is formally analogous to the univariate case, since if

$$(4.19)\qquad w(x,A) = \frac{1}{1 + \exp\{c + [x-(\mu_1+\mu_2)/2]^T A\Sigma^{-1}(\mu_2-\mu_1)\}}$$

then

$$(4.20)\qquad \tilde{w}(x) = w(x,A_h), \quad w(x) = w(x,I)$$

with the logit c.d.f. structure exploited as before. Of course, for the multivariate case, the range of possible distribution shapes is more severely restricted by the requirement of equal covariance structures across the

mixture components. Nevertheless, it raises the possibility that downward score bias is approximately matrix proportional is nonnormal examples.

## 4.4 Mean Bias Corrections

For diagnosing and/or correcting the bias problem, a couple of suggestions are immediate. First, if the data used is standardized to have unit variances, the bias proportion $1-a_h = h^2/(1 + h^2)$ appropriate for the normal case can be calculated, to give a "quick and dirty" indication of the extent of the problem.[22]

A more involved correction is suggested by the estimators discussed in Section 3.1. In particular, suppose that the bias is approximately matrix proportional, with $\lambda_h(x) \cong A\ell(x)$. In this case, we have that

$$\text{Cov}[\lambda_h(x),x] \cong A\ \text{Cov}[\ell(x),x] = A \tag{4.21}$$

But this covariance is estimated by the leading term of the slope estimator $\hat{d}$ of (3.4), namely

$$D_h = N^{-1} \sum_{i=1}^{N} \hat{\ell}(x_i)(x_i-\bar{x})^T\hat{1}_i \tag{4.22}$$

Consequently, computing $D_h$ and examining $I - D_h$ gives a method of examining the extent of the bias. Moreover, the score estimates could be corrected as

$$\bar{\ell}(x) = D_h^{-1}\ \hat{\ell}(x)\ . \tag{4.23}$$

For the normal case, this correction is consistent, with plim $\bar{\ell}(x) = \ell(x)$ regardless of whether h is treated as fixed or shrinking with sample size. The quality of the correction in general depends on the quality of the proportionality relationship $\lambda_h(x) \cong A\ell(x)$. If this relationship is not close

for certain ranges of x, the correction (4.23) only serves to fix the overall level of the score estimates, by assuring $\mathrm{Cov}[\tilde{\ell}(x),x] = I$ in large samples.

## 4.5 Further Remarks on Downward Bias in Score Estimates in the General Case

As mentioned above, it appears difficult to show a general result that the norm of $\lambda_h(x)$ is typically smaller than that of $\ell(x)$, where "general" means that f(x) can be any arbitrary base density. One type of comparison is available from efficiency theory, as follows. Pretend for the moment that the data consisted of observations $(x_i,hu_i)$, i=1,...,N, distributed with density $K(u)f(x-\mu_0)$, with true value $\mu_0=0$. Here x is sufficient for the estimation of $\mu$, and under standard regularity conditions the maximum likelihood estimator $\hat{\mu}$ $\equiv \mathrm{argmax}_\mu \sum \ln f(x_i-\mu)$ has asymptotic variance

$$\Sigma_\ell^{-1} \equiv \left[\int \ell(x)\ell(x)^T f(x)dx\right]^{-1} \tag{4.24}$$

which is the asymptotic Cramer-Rao bound. Alternatively, we could (stupidly) use the data $\{z_i = x_i+hu_i,\ i=1,...,N\}$ to estimate $\mu$, namely by $\tilde{\mu} = \mathrm{argmax}_\mu \sum \ln \phi_h(z_i-\mu)$, which has asymptotic variance

$$\tilde{\Sigma}_{\lambda h}^{-1} \equiv \left[\int \lambda_h(z)\lambda_h(z)^T \phi_h(z)dz\right]^{-1} . \tag{4.25}$$

Standard Cramer-Rao theory asserts that $\tilde{\Sigma}_{\lambda h}^{-1} - \Sigma_\ell^{-1}$ is positive semi-definite, or equivalently that $\Sigma_\ell - \tilde{\Sigma}_{\lambda h}$ is positive semidefinite. In this sense, $\lambda_h(x)$ is generally smaller than $\ell(x)$ in absolute value, but this sense is not exactly suited for the purpose at hand. In particular, since our argument is that $\lambda_h(x)$ is measured instead of $\ell(x)$, a better comparison would be of $\Sigma_\ell$ to $\Sigma_{\lambda h} = \int \lambda_h(x)\lambda_h(x)^T f(x)dx$. It is not immediately apparent how to establish that $\Sigma_\ell - \Sigma_{\lambda h}$ is positive semi-definite for arbitrary f(x).

Again, such a result is not contrary to intuition; if the components of $\lambda_h(x)\lambda_h(x)^T$ increase with $|x - \mu|$ (for instance if $\phi_h$ were log-concave), the fact that f(x) displays smaller variance that $\phi_h(x)$ would be consistent with $\Sigma_\ell - \Sigma_{\lambda h}$ positive semi-definite.

The pointwise relation between $\lambda_h$ and $\ell$ is intimately connected to the concavity properties of $\ln(\phi_h/f)$, since $\partial \ln(\phi_h/f)/\partial x = \ell(x) - \lambda_h(x)$. For instance, if f is univariate and symmetric and $\ln(\phi_h/f)$ is quasi-convex, then $|\ell(x)| \geq |\lambda_h(x)|$ for all x. The normal example given above has $\ln(\phi_h/f)$ convex, consistent with this observation. However, even when $\mathcal{K}$ and f are log-concave, $\ln(\phi_h/f)$ is the difference between two concave functions, and it is not obvious how to characterize the primitive conditions under which this difference is quasi-convex over a region of substantial probability.

We can show one result for the univariate case.

Theorem 4.1: Given Assumptions 2.1-5, if

i) f(x) is three times differentiable, symmetric and unimodal with mode $m_f$,

ii) $\ell(x)$ is an increasing convex function on $(-\infty, m_f]$.

iii) $h \in (0, h_0]$, and for $x \in [m_f - h, m_f)$, $a \in (-h, h)$,

$$-f'''(x+a) \geq \ell(x)\, f''(x+a)$$

then

$$E(|\ell(x)|) > E(|\lambda_h(x)|) \tag{4.26}$$

where $E(|\ell(x)|) \equiv \int |\ell(x)| f(x)dx$ and $E(|\lambda_h(x)|) \equiv \int |\lambda_h(x)| f(x)dx$.

Theorem 4.1 opens the possibility that $|\lambda_h|$ is smaller than $|\ell(x)|$ in general, but is based on very restrictive conditions. While the symmetry and

unimodality are needed in the proof, they are certainly not necessary for (4.26) to hold. In particular, the proof suggests that these conditions may suffice for $\ln(\phi_h/f)$ to be quasi-convex; if so then (4.26) is just a weak implication of the pointwise dominance $|\ell(x)| \geq |\lambda_h(x)|$ a.e. (f). Theorem 4.1 covers the univariate normal case above, where the score $\ell(x)$ is linear in x.

## 5. Conclusion

This paper has established how the derivatives of nonparametric density estimates may contain substantial downward biases, due to local smoothing. It is important to stress the point that the existence of bias in moderate samples is by itself not surprising, but rather that the bias causes systematic suppression of the values of derivatives. Density estimators based on local smoothing clearly represent some of the best tools for flexibly characterizing modal structure and other nonlinearity, and this paper only argues that the magnitudes of measured "bumps" and "valleys" can be understated. When the precise magnitudes are inessential to an empirical problem, the downward derivative bias is of little practical consequence.

It is also important to stress that the extent of downward derivative bias is determined solely by the amount of smoothing, or size of bandwidth used. In particular, dimensionality and/or higher order differentiability of the base density have an effect only if the bandwidth value is affected. Moreover, if the bandwidth value used is tiny, then the derivative bias is minor. Of course, if the bandwidth value chosen is smaller than that dictated by a mean squared error criterion, the pointwise variance is larger than it needs to be.

One natural response to our results is that we have just inappropriately used bandwidth values that are too large. While Table 3 is presented to address this concern, certain theory can be interpreted to say that the

bandwidth values used in that table may still be too large. For instance, Goldstein and Messer (1990) suggest implementing the "asymptotic undersmoothing" requirement of estimating smooth functionals by choosing bandwidths that are smaller than those that are optimal for pointwise estimation. While the sense of this analogy is clear, it is an empirical issue as to whether the smaller bandwidths will work in practice for realistic sample sizes. The author's experience with average derivative estimators (discussed in Section 3.1) is contrary to this, leading in part to the present study. At any rate, in any given empirical problem, a bandwidth value must be chosen, and the downward bias will be present. Whether tiny bandwidths can be used successfully in small samples, for either nonparametric or semiparametric problems, is a central question of future research.

While the derivative bias is a generic problem, its structure may likewise permit generic corrections. In this spirit we have advanced the idea of proportionality in the bias of estimated scores. This coincides with the author's experience that the slope versions of average derivative estimators (such as (3.4)) give good estimates under a wide variety of simulation designs. Moreover, the correction suggested is simple and interpretable. A natural question of future research is whether such a simple correction has practical value in general. In other words, are there standard empirical settings, remote from our examples, that dictate more involved corrections as a general rule?

This point, as well as our earlier conclusions, stress the need for studying the performance of nonparametric and semiparametric methods in realistically sized samples. The well-developed theoretical paradigm for these methods is in need of empirical confirmation, as the highest priority for future work.

## Appendix: Proofs of Theorems

Proof of Lemma 2.1:

1. That $\phi_h$ is twice differentiable follows from the twice differentiability of f and Assumption 2.4. The remaining points are immediate, with $S_f = [a,b]$ implying $S_\phi = [a-h,b+h]$.

2. f(x) is symmetric about $m_f$ if $f(x) = f(2m_f-x)$. We have $\phi_h(x) = \int \mathcal{K}(u)f(x-hu)du = \int \mathcal{K}(u)f(2m_f-x+hu)du = \int \mathcal{K}(u)f(2m_f-x-hu)du = \phi_h(2m_f-x)$ by the symmetry of $\mathcal{K}(u)$.

3. Suppose $I = \{x;\ w\in(x,m_f],\ f''(w) < 0\} = [w_1,w_2]$, where $w_1<m_f<w_2$ by continuity of f". Set $h_0 = (1/2) \min \{m_f-w_1, w_2-m_f\}$. From (3.7), we have $\phi_h'(x) \geq 0$ for $x \in (-\infty, m_f-h]$ and $\phi_h'(x) \leq 0$ for $x \in [m_f+h,\infty)$. Suppose $\phi_h$ has two modes $m_1,m_2 \in [m_f-h, m_f+h]$, $m_1 \leq m_2$, $m_1$. Then $0 = \int \mathcal{K}(u)[f'(m_1-uh)-f'(m_2-uh) = \int \mathcal{K}(u)[(m_2-m_1)f''(\xi(u))du = (m_2-m_1)f''(\xi)$, where the second equality is the mean value theorem, with $\xi(u)\in[m_1-h,m_2+h]$ and the third inequality is the mean value theorem for integrals, with $\xi\in[m_f-2h,m_f+2h]$. But since $h \leq h_0$, $f''(\xi) < 0$, so that $m_1 = m_2$, with $\phi_h$ unimodal. Since $f(m_f) = \sup f(x)$ and $m_f$ is unique, by (3.6) we have $\phi_h(m_\phi) < f(m_f)$.

4. For each local mode, define $I_j = \{x;\ w\in(x,m_j],\ f''(w) < 0\} = [w_{j1},w_{j2}]$, $h_j = (1/2) \min \{m_j-w_{j1}, w_{j2}-m_f\}$. For each local minimum, define $I^j = \{x;\ w\in(x,b_j],\ f''(w) > 0\} = [w^{j1},w^{j2}]$ and $h^j = (1/2) \min \{b_j-w^{1j}, w^{2j}-b_j\}$. Define $h_0 = (1/2) \min \{h_j, j=1,\ldots,d,\ h^j,\ j=1,\ldots,d-1\}$, and note that $h_0 > 0$ by the continuity of f". By an argument analogous to that given in 3 for unimodality, we conclude that $\phi_h$ has a unique mode $\tilde{m}_j$ in $[m_j-h,m_j+h]$, for each $j=1,\ldots,d$, and a unique local minimum $\tilde{b}_j$ in $[b_j-h,b_j+h]$, $j=1,\ldots,d-1$. The remaining properties follow from the fact that the supremum of f(x) over

$[m_j-2h, m_j+2h]$ is $f(m_j)$, and that the infimum of $f(x)$ over $[b_j-2h, b_j+2h]$ is $f(b_j)$. QED Lemma 2.1

Proof of Theorem 2.1: $f'(x)$ and $\phi_h'(x)$ alternate in sign: $f'(x) \geq 0$, $x \in (-\infty, m_1)$; $f'(x) \leq 0$, $x \in [m_1, b_1]$; ...; $f'(x) \leq 0$, $x \in [m_d, -\infty)$; and $\phi_h'(x) \geq 0$, $x \in (-\infty, \tilde{m}_1)$; $\phi_h'(x) \leq 0$, $x \in [\tilde{m}_1, \tilde{b}_1]$; ...; $f'(x) \leq 0$, $x \in [\tilde{m}_d, -\infty)$. Consequently,

$$\int_{-\infty}^{\infty} (|f'(x)| - |\phi_h'(x)|)dx = 2 \sum [f(m_j) - \phi_h(\tilde{m}_j)] + 2 \sum [\phi_h(\tilde{b}_j) - f(b_j)] > 0$$

by Lemma 2.1. QED Theorem 2.1.

Proof of Theorem 4.1: Since $f$ is symmetric and unimodal (with mode $m_f$), $\phi_h$ is symmetric and unimodal with mode $m_f$, by Lemma 3.1. Consequently, $f'(x)$ and $\phi_h'(x)$ are asymmetric about $m_f$ (namely $f'(x) = -f'(2m_f - x)$, etc.), as are $\ell(x)$ and $\lambda_h(x)$. Thus

$$E(|\ell(x)|) - E(|\lambda_h(x)|) = \int_{-\infty}^{\infty} |\ell(x)| - |\lambda_h(x)| \, dx$$

$$= 2 \int_{-\infty}^{m_f} \left( \frac{\partial \ln f}{\partial x} - \frac{\partial \ln \phi_h}{\partial x} \right) f(x)dx$$

For $b < m_f$, apply integration by parts as

$$\int_b^{m_f} \left( \frac{\partial \ln f}{\partial x} - \frac{\partial \ln \phi_h}{\partial x} \right) f(x)dx = \ln \frac{f(m_f)}{\phi_h(m_f)} f(m_f) - \ln \frac{f(b)}{\phi_h(b)} f(b)$$

$$- \int_b^{m_f} \ln \frac{f(x)}{\phi_h(x)} f'(x) \, dx$$

$$= \left( \ln \frac{f(m_f)}{\phi_h(m_f)} - \ln \frac{f(\xi)}{\phi_h(\xi)} \right) f(m_f) + \left( \ln \frac{\phi_h(b)}{f(b)} - \ln \frac{\phi_h(\xi)}{f(\xi)} \right) f(b)$$

where the latter equality follows from the mean value theorem for integrals,

with $\xi \in [b, m_f]$. If $\phi_h(x)/f(x)$ is a monotonically decreasing function of x, then the theorem follows from taking the limit as b " $-\infty$ (note that the inequality is strict; since $\phi_h(m_f)/f(m_f) < 1$, $\phi_h(x)/f(x)$ constant would violate the fact that $\phi_h(x)$ is a density function).

To show that $\phi_h(x)/f(x)$ is decreasing, differentiate as

$$\frac{\partial(\phi_h/f)}{\partial x} = \frac{1}{f(x)} \int_{-1}^{1} \mathcal{K}(u) \left( f'(x-hu) - \frac{f'(x)}{f(x)} f(x-hu) \right) du$$

$$= \frac{1}{f(x)} \int_{0}^{1} \mathcal{K}(u) \left( f'(x-hu) + f'(x+hu) - \frac{f'(x)}{f(x)} [f(x-hu) + f(x+hu)] \right) du$$

by the symmetry of $\mathcal{K}$. If the term in brackets is less than or equal to zero, then so is the integral, and the result follows. This occurs if

$$(*) \qquad f'(x-hu) + f'(x+hu) \leq \frac{f'(x)}{f(x)} [f(x-hu) + f(x+hu)]$$

For $x \leq m_f - h$, define $w(x,hu) \equiv f(x-hu)/[f(x-hu)+f(x+hu)]$, and note that $w(x,hu) \leq (1/2)$ since f(x) is decreasing for $x \leq m_f$. Therefore

$$\frac{f'(x-hu) + f'(x+hu)}{f(x-hu) + f(x+hu)} = w(x,hu) \frac{f'(x-hu)}{f(x-hu)} + (1-w(x,hu)) \frac{f'(x+hu)}{f(x+hu)}$$

$$\leq \frac{f'[x + (1-2w)hu]}{f[x + (1-2w)hu]} \leq \frac{f'(x)}{f(x)}$$

since $\ell(x) = - f'(x)/f(x)$ is convex, and increasing in x, proving (*). Now suppose $x \in [m_f - h, m_f)$. Condition (iii) implies that

$$v(a) = f'(x+a) - \frac{f'(x)}{f(x)} f(x+a)$$

is a concave function in $a \in (-h,h)$, and (*) is implied by

$$(1/2)v(-hu) + (1/2)v(hu) \leq v(0) = 0.$$

QED Theorem 4.1.

## Notes

[1] Textbook treatments of this work can be found in Silverman (1986), Prakasa-Rao (1983) and Härdle (1991) among many others; for a recent survey on work on nonparametric methods in econometrics, see Delgado and Robinson (1991).

[2] A recent general treatment is given by Goldstein and Messer(1990). For such results in a specific problem see Powell, Stock and Stoker (1989) and Härdle and Stoker (1989), among others.

[3] Stoker(1991b) discusses downward derivative bias with kernel regression estimators. While the mathematics of this problem is similar, the structure of the bias is quite different. Preliminary version of some results from the current paper as well as Stoker (1991b) were previously reported in a manuscript entitled "Smoothing Bias in Derivative Estimation," revised July 1990.

[4] See Ibragimov, I.A. and Has'minskii (1981), among others.

[5] Provided the variances of the components of (1.1) and (2.1) exist, standard laws of large numbers and central limit theory (with the bandwidth h fixed) imply that $\operatorname{plim} \hat{f}(x) = E[\hat{f}(x)]$ and $\operatorname{plim} \hat{f}'(x) = E[\hat{f}'(x)]$, and that $\sqrt{N}(\hat{f}(x) - E[\hat{f}(x)])$ and $\sqrt{N}(\hat{f}'(x) - E[\hat{f}'(x)])$ have limiting normal distributions.

[6] See Silverman (1986) and Manski (1988), among many others.

[7] A function R(x) is quasi-concave if $R[ax_1+(1-a)x_2] \geq \min[R(x_1),R(x_2)]$, and is log-concave if ln R is concave.

[8] These properties were suggested to the author by A. Caplan and B. Nalebuff, and are reviewed in Prekopa(1980) and Caplan and Nalebuff(1990). Some related properties can be found in the theory of majorization; c.f. Marshall and Olkin(1979).

[9] The following lemma and theorem are quite basic, however I could not find similar results in the literature.

[10] See Stoker (1986) and Härdle and Stoker (1989) for many examples of index structure of this type.

[11] For instance, in the simulation study of "density weighted" average derivative estimators in Powell, Stock and Stoker(1989), it was shown that slope estimators (with positive kernels) gave somewhat better small sample performance than moment estimators. However, their simulations were based on normalized estimators, so that a uniform derivative bias would not be detectable.

[12] The following results will be revised as part of a large simulation study to be reported as Stoker and Villas-Boas (1991).

[13] While the positive kernel is not consistent with the asymptotic normality results indicated above, Powell, Stock and Stoker(1989) and Stoker and Villas-Boas(1990) find that a positive kernel gives superior small sample performance. Moreover, a positive kernel is used in the remainder of the exposition, so that the results reported are relevant.

[14] The bandwidth value h = 1 was not chosen subject to an optimality criterion, but rather was roughly the minimum value for which the results on $\hat{\delta}$ of (3.3) were not erratic. Tables 1 and 2 are mainly illustrative; many more than 20 Monte Carlo simulations called for (see Note 12).

[15] A preliminary study of this problem is given in Härdle, Hart, Marron and Tsybakov (1991), for the one dimensional case.

[16] We depart from our previous notation in this section only, namely with $f_\epsilon$ the density to be estimated nonparametrically.

[17] We have included the "low density" trimming indicator $\hat{1}_i$ as in Stone (1975), but abstracted from the "sample-splitting" feature of Bickel's(1982) analysis. This is because our concern here is not with the technical issues of estimation, but just to illustrate uses of the density score. Of course, if basing density estimation on a split sample necessitated using a larger bandwidth, then the score bias problems discussed next would be exacerbated.

[18] As before, a standard application of central limit theory (and the delta method) shows that for fixed h, $\sqrt{N}[\hat{\ell}(x) - \lambda_h(x)]$ has a limiting normal distribution, provided x is in the interior of the support of f(x).

[19] Here $\lambda_h(x)$ is a "matrix weighted average" of $\ell(x)$ and 0 in the sense of Chamberlain and Leamer(1976), namely $\lambda_h(x) = A_h\, \ell(x) + (I-A_h)\, 0$.

[20] This verifies the assertion at the end of Section 3.1 that $\lambda_h(x) = 1/2\ \ell(x)$ applied to the simulation design, where $\Sigma = I$ and $h = 1$.

[21] There are at least three reasons to believe that the bandwidth values of Table 3 are too small, so that Table 3 understates the downward bias. First, the bandwidth formula is for estimating the density, and not the derivative or score, which could be expected to give larger values. Second, the bandwidth formula is based on the leading terms of the Taylor series of integrated mean squared error, which is an asymptotic approximation. Steve Marron has told me about some of his joint work calculating exact optimal bandwidth values for the normal design here, that indicates that the bandwidth values in Table 3 are too small. Unfortunately, at the time of writing, I haven't been able to locate a reference to this work. Finally, as stated in note 14, erratic behavior of the average derivative estimator (3.3) was exhibited for bandwidths less that 1.0, so that the "optimal value" of .508 in Table 3 did not give good performance. Consequently, this table may have bandwidth values that are too small for realistic situations, especially those values given for very small samples and moderate to large dimension.

[22] This ratio overstates the bias in the normal mixture case, because of the "between variance;" in the two mixture case, the data would be standardized for the variance $\sigma^2 + \rho(1-\rho)(\mu_2-\mu_1)^2$, not just $\sigma^2$ as it appears in $a_h = \sigma^2/(\sigma^2+h^2)$.

## References


Bickel, P. (1982): "On Adaptive Estimation," *Annals of Statistics*, 10, 647-671.

Brillinger, D.R. (1983), "A Generalized Linear Model with 'Gaussian' Regressor Variables," in P.J. Bickel, K.A. Doksum and J.l. Hodges, eds., *A Festschrift for Erich L. Lehmann*, Woodsworth International Group, Belmont, CA.

Caplan, A. and B. Nalebuff(1990), "Aggregation and Social Choice: A Mean Voter Theorem," Cowles Foundation Discussion Paper.

Chamberlain, G. and E. Leamer(1976), "Matrix Weighted Averages and Posterior Bounds," *Journal of the Royal Statistical Society*, *B*, 73-84.

Delgado, M.A. and P.M. Robinson (1991), "Nonparametric Methods and Semiparametric Methods for Economic Research," draft, London School of Economics.

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

Gnedenko, B.V. and A.N. Kolmogorov(1954), *Limit Distributions for Sums of Independent Random Variables*, Addison Wesley, Cambridge, Mass.

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

Härdle, W., J. Hart, J.S. Marron and A.B. Tsybakov (1991), "Bandwidth Choice for Average Derivative Estimation," draft, Universität Bonn.

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.

Ibragimov, I.A.(1956), "On the Composition of Unimodal Functions," *Theor. Probability Appl.*, 1, 255-260.

Ibragimov, I.A. and R.Z. Has'minskii (1981), *Statistical Estimation: Asymptotic Theory*, New York: Springer-Verlag.

Manski, C.F. (1984), "Adaptive Estimation of Nonlinear Regression Models," *Econometric Reviews*, 3, 145-194.

Manski, C.F. (1988) *Analog Estimation in Econometrics*, London, Chapman Hall.

Marshall, A.W. and I. Olkin(1979), *Inequalities: Theory of Majorization and Its Applications*, Academic Press, New York.

Prakasa-Rao, B.L.S.(1983), *Nonparametric Functional Estimation*, Academic Press, New York.

Prekopa, A.(1973), "On Logarithmic Concave Measures and Functions," *Acta. Sci. Math.*, 34, 355-343.

Prekopa, A.(1980), "Logarithmic Concave Measures and Related Measures," in M.A.H. Dempster, ed., *Stochastic Programming*, Academic Press, New York.

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

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, London, Chapman Hall.

Stein, C. (1956), "Efficient Nonparametric Testing and Estimation," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, Berkeley, University of California Press.

Stoker, T.M. (1986), "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54, 1461-1482.

Stoker, T.M. (1991a), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W.A. Barnett, J.L. Powell and G. Tauchen, Cambridge University Press.

Stoker, T.M. (1991b), "Smoothing Bias in the Measurement of Marginal Effects," MIT Sloan School of Management Working Paper, August.

Stoker, T.M. and J.M. Villas-Boas (1991), "Monte Carlo Simulation of Average Derivative Estimators", draft.

Stone, C.J. (1975): "Adaptive Maximum Likelihood Estimators of a Location Parameter," *Annals of Statistics*, 3, 267-284.